

**Land cover mapping of one specific protected
habitat under the requirements of the
European Union Habitats Directive using
remote sensing.**

Carolina Sanchez Hernandez

Ph.D

Kingston University

2006

Abstract

Habitat loss is considered to be one of the greatest challenges currently facing society. An important part of the European Union's (EU) response to this problem is the Habitats Directive, the main aim of which is to protect biodiversity through the conservation and protection of natural habitats. Consequently, accurate mapping of specific habitats is of high importance in order to monitor ecosystem changes. Within this context, remote sensing has enormous potential as a source of land cover information which has been used widely for land cover mapping. In many instances this land cover mapping is only concerned with one particular habitat. However, in standard classification analysis, training data for all of the land cover classes contained in the area is typically required which means a wasteful use of resources. This thesis aims to address these issues by investigating advanced classification methods that focus on the accurate mapping of one specific protected habitat. The data used for this purpose is a Landsat ETM+ image of East Anglia, UK, acquired in June 2000 and ground truth data in the form of aerial photography. The habitat of interest chosen for this investigation is fens, a habitat protected by the EU Habitats Directive, whose diverse and dynamic nature is a particular challenge for its mapping and monitoring. A second protected habitat, saltmarshes, will be used for comparison purposes in order to determine any bias within the results.

The methods considered to map the selected habitat consist of binary classifiers and one class classifiers. The binary classifiers chosen were Support Vector Machines (SVMs) and Decision Trees (DTs) which are two new methods very recently applied to land cover classification and remote sensing research. They are still very much the focus of current research for multiclass classification. In this thesis they are used in its binary form to classify the class of interest against all the other classes. Both classifiers perform very well when compared with a classic parametric Maximum Likelihood Classification (MLC). Narrowing down the idea of classifying just one habitat of interest, one-class classifiers are put to the test. They have been explored in pattern recognition research but not yet within remote sensing image classification and land cover mapping. Specifically, the Support Vector Data Description (SVDD)

classifier is considered particularly suitable for land cover classification as it is based upon the basis of SVMs which have already been applied in this area with success. When the results of the SVDD classification are compared against those obtained by the other classifiers these show an improvement in overall classification and a reduction in the errors of commission. Furthermore, another method is also put to the test to improve the accuracy of the classification of the class of interest. This method is the ensemble of classifiers, which in many research studies within pattern recognition has proven to improve accuracy of single classifiers. The results in this thesis also show an improvement in accuracy, although further investigation is needed.

In conclusion, DT, SVM and SVDD classification methods offer clear advantages over standard classification analysis when concentrating on the classification and mapping of a particular habitat. All three classifiers obtained higher accuracies than the ML classifier with the use of significantly less training data. Furthermore, in the case of one-class classification only data from the class of interest was needed. Also in both binary and one-class classification approaches the attention was focused on separating the class of interest from all the other classes and therefore training efficiency was bigger than in a standard multiclass classification where efforts are directly to achieve a high overall accuracy. All three methods were found to be highly suitable for classifying and mapping a specific habitat and its application should be considered in future work involving the accurate mapping of protected habitats.

Declaration

This thesis has been composed by myself and the work is my own.

Carolina Sanchez Hernandez

April 2006

Acknowledgements

First of all, I would like to express my sincere gratitude and thanks to my supervisors, Professor Guy Robinson and Dr. Doreen Boyd. I will always be grateful for their continuous support, help and guidance throughout the period of this study. Without all his help and suggestions, this work would not have been possible.

I am very grateful to Kingston University for giving me the opportunity and awarding me with the scholarship to pursue a PhD degree. I would like to convey my thanks to all the staff members of the School of Earth Sciences and Geography and other staff from different areas of the university, for their cooperation and support. I am also thankful to David Livingstone for helping me in the initial stage of my work.

I wish to extend my thanks to Professor Giles Foody from the School of Geography, University of Southampton, for providing invaluable help and guidance for this research and for his important collaboration in different publications. I also would like to thank Dr. David Tax from The Delft Institute in the Netherlands for providing help with the DD_tools programming and one-class classification issues.

I wish to extend my thanks to NERC for providing the necessary satellite data to carry out this research, the Environment Agency and the Broads Authority for providing aerial photographs for the areas of interest and for their kind help and advice. Also thanks to Salford Systems for allowing me to use the CART software for the purpose of this research and the Image, Speech and Intelligent Systems research group, School of Electronics and Computer Science, University of Southampton, for the SVM used in this study

Special recognition is due to my friends and family for their continuous encouragement. Special thanks are due to my partner Edward Sell for his endless patience and understanding during all long hours of work and for his endless support.

TABLE OF CONTENTS

1 INTRODUCTION	9
1.1 BACKGROUND	9
1.2 AIMS AND OBJECTIVES OF THIS THESIS.....	17
1.3 AREA OF STUDY AND HABITATS OF INTEREST.....	18
1.4 THESIS STRUCTURE.....	19
 2 REMOTE SENSING AND IMAGE CLASSIFICATION: REVIEW AND SELECTION OF SUITABLE METHODS FOR CLASSIFYING A HABITAT OF INTEREST	 22
2.1 REMOTE SENSING METHODS	24
2.1.1 BACKGROUND	24
2.1.2 THE IMAGE CLASSIFICATION PROCESS	26
2.1.3 CHOOSING A CLASSIFICATION METHOD AND TRAINING THE CLASSIFIER.....	30
2.1.4 CLASSIFICATION OUTPUT: LABELLING.....	39
2.1.5 ACCURACY ASSESSMENT	40
2.2 CLASSIFYING A CLASS OF INTEREST.....	42
2.2.1 BINARY CLASSIFIERS FOR THE CLASSIFICATION OF ONE HABITAT OF INTEREST.....	43
2.2.2 ONE-CLASS CLASSIFIERS FOR THE CLASSIFICATION OF ONE HABITAT OF INTEREST.....	46
2.3 SUMMARY	49
 3 MAPPING ONE SPECIFIC HABITAT OF INTEREST: CASE STUDY AND METHODS	 51
3.1 THEMATIC CONTENT: HABITAT OF INTEREST AND AREAS	53
3.1.1 FENS	53

3.1.2	SALTMARSHES.....	56
3.2	REMOTE SENSING DATA AND GROUND DATA SOURCES	59
3.3	DATA SELECTION. TRAINING AND TESTING DATASETS	64
3.3.1	TRAINING AND TESTING DATASETS.....	66
3.4	ACCURACY ASSESSMENT	71
3.5	SUMMARY	76

4 BINARY CLASSIFICATION FOR LAND COVER MAPPING OF A CLASS OF INTEREST: SUPPORT VECTOR MACHINES AND DECISION TREES VERSUS MAXIMUM LIKELIHOOD CLASSIFIERS **77**

4.1	MAXIMUM LIKELIHOOD CLASSIFICATION: PRINCIPLES AND APPLICATION TO THE CASE STUDY.....	80
4.1.1	CLASSIFICATION OF A HABITAT OF INTEREST USING ML CLASSIFICATION. CASE STUDY. 83	
4.2	SUPPORT VECTOR MACHINES: PRINCIPLES AND CASE STUDY	91
4.2.1	SUPPORT VECTOR MACHINES: PRINCIPLES.....	93
4.2.2	CLASSIFICATION OF A HABITAT OF INTEREST USING THE SVM CLASSIFIER. CASE STUDY. 100	
4.3	DECISION TREES CLASSIFIERS.....	108
4.3.1	CLASSIFICATION OF A HABITAT OF INTEREST USING THE DT CLASSIFIER. CASE STUDY. 117	
4.4	SUMMARY AND CONCLUSIONS.....	123

5 ONE-CLASS CLASSIFICATION METHODS AND THEIR APPLICATION TO ONE CLASS LAND COVER MAPPING **126**

5.1	ONE-CLASS CLASSIFICATION. PRINCIPLES AND METHODS	129
5.1.1	ONE-CLASS CLASSIFICATION METHODS	131
5.1.2	COMPARISON OF ONE-CLASS CLASSIFICATION METHODS	136

5.2	SUPPORT VECTOR DATA DESCRIPTION (SVDD), PRINCIPLES AND CASE STUDY	143
5.3	CASE STUDY. SVDD AND CLASSIFICATION AND MAPPING OF A CLASS OF INTEREST.	148
5.3.1	SVDD A	150
5.3.2	SVDD B.....	152
5.4	SUMMARY AND CONCLUSIONS.....	155
6	ENSEMBLE OF CLASSIFIERS FOR THE CLASSIFICATION OF A HABITAT OF INTEREST	157
6.1	ENSEMBLE OF CLASSIFIERS. THEORY AND METHODS.....	160
6.2	ENSEMBLES OF BINARY SVM AND DT CLASSIFIERS	165
6.2.1	SVM ENSEMBLES.....	165
6.2.2	DECISION TREES ENSEMBLES.....	170
6.3	ENSEMBLES OF ONE-CLASS CLASSIFIERS	176
6.4	COMBINING ALL CLASSIFIERS.....	182
6.5	SUMMARY AND CONCLUSIONS.....	185
7	FINAL DISCUSSION, CONCLUSIONS AND FURTHER RESEARCH	188
7.1	FINAL DISCUSSION: COMPARISON SVDD, SVM AND DT CLASSIFIERS	189
7.2	CONCLUSIONS AND FURTHER RESEARCH.....	206

**ANNEX A CONFUSION MATRICES FOR SVM, DT AND SVDD
CLASSIFICATIONS**

**ANNEX B PARAMETER CHOICE FOR SVM AND SVDD BY CROSS-
VALIDATION**

**ANNEX C SVM SUPPORT VECTORS AND SEPARATING
HYPERPLANES**

ANNEX D DECISION TREES STRUCTURES

ANNEX E LIST OF PUBLICATIONS

ANNEX F. CD WITH RAW DATA. TRAINING AND TESTING DATA SETS

Reference list

TABLE OF FIGURES

<i>Figure 1.1 Thesis structure. The first three chapters show the background of the research The research chapters shift towards the right showing the progression of the research towards the final discussion and conclusions.</i>	<i>19</i>
<i>Figure 2.1 Electromagnetic Spectrum. Based upon Lillesand and Kiefer (2004).....</i>	<i>24</i>
<i>Figure 2.2 Typical spectral reflectance curve. Based upon Lillesand and Kiefer (2004).</i>	<i>25</i>
<i>Figure 2.3 Representation of a pixel (feature vector) in a 2 dimensional feature space. After Jensen (1996).....</i>	<i>26</i>
<i>Figure 2.4. Digital image analysis. Based upon Campbell (2002). If there is ancillary data available the classification is normally supervised.....</i>	<i>27</i>
<i>Figure 2.5. Minimum Distance to mean (MD) classification strategy. Based on Lillesand and Kiefer (2004).....</i>	<i>32</i>
<i>Figure 2.6. Parallelepiped classification strategy. Based on Lillesand and Kiefer (2004).....</i>	<i>33</i>
<i>Figure 2.7. Equiprobability contours defined by a maximum likelihood classifier. Based on Lillesand and Kiefer (2004).</i>	<i>34</i>
<i>Figure 2.8. Example of k-NN classifier. Based on Lillesand and Kiefer (2004).</i>	<i>36</i>
<i>Figure 2.9. Example of an Artificial Neural Network classifier.....</i>	<i>37</i>
<i>Figure 2.10. Example of a Decision Tree Classifier</i>	<i>38</i>
<i>Figure 2.11. Example of optimal hyperplane or boundary between two classes using a SVM. Based upon Burges (1998).....</i>	<i>39</i>
<i>Figure 3.1 Case study methodology. The sections with dotted line are dealt with in different chapters.</i>	<i>52</i>
<i>Figure 3.2. Landsat ETM+ 19th of June 2000 provided by NERC and selected subset of the image used for further analysis</i>	<i>61</i>
<i>Figure 3.3 Map of the Norfolk Fens and the test area chosen within the River Yare National Nature Reserve to illustrate the results of different classification methods (Map based upon the Broads Authority information).....</i>	<i>63</i>
<i>Figure 3.4 ETM+ Band 2 image for the test area</i>	<i>70</i>
<i>Figure 3.5 NDVI image for the test area. Values in green denote bigger NDVI values.....</i>	<i>70</i>
<i>Figure 4.1 Chapter 4 structure.....</i>	<i>78</i>
<i>Figure 4.2 Representation of a hyperplane separating two classes.</i>	<i>79</i>
<i>Figure 4.3. Maximum likelihood classification. Based upon Lillesand and Kiefer (2004).</i>	<i>81</i>
<i>Figure 4.4 Local minima and global solution. If a classifier starts with a weight set corresponding to point P the first solution that it encounters is at M_l. This is called local minima and corresponds to a partial solution in response to the training data. M_g is the global minimum or global solution. In neural networks unless measures are taken to escape from the local minima the global solution will never be reached (Based upon Burges, 1998).</i>	<i>93</i>
<i>Figure 4.5. Representation of separating Hyperplane($w \cdot x + b = 0$). Based upon Burges (1998).</i>	<i>94</i>

<i>Figure 4.6. Optimal separating hyperplane. The circled points represent support vectors. Based upon Burges (1998).</i>	95
<i>Figure 4.7 Mapping into feature spaces.</i>	98
<i>Figure 4.8. Accuracy results for class fen for different training sizes.</i>	103
<i>Figure 4.9. Accuracy results for class saltmarsh for different training sizes.</i>	104
<i>Figure 4.10 Number of support vectors used by the SVM classifier for fen and saltmarsh</i>	105
<i>Figure 4.11 Comparison of accuracies SVM v MLC. Fen as class of interest.</i>	106
<i>Figure 4.12 Comparison of accuracies SVM v MLC. Saltmarsh as class of interest.</i>	107
<i>Figure 4.13. Decision tree. A, B and C denote the final classification of training data into these three classes. After Pal and Mather (2003).</i>	110
<i>Figure 4.14. Axis parallel decision boundaries of a univariate decision tree. Based upon Friedl and Brodley (1997).</i>	112
<i>Figure 4.15. Decision boundaries for a multivariate decision tree classifier. Based upon Friedl and Brodley (1997).</i>	113
<i>Figure 4.16 Binary Decision Tree classification. Fen as class of interest.</i>	118
<i>Figure 4.17 Binary Decision Tree classification. Saltmarsh as class of interest</i>	119
<i>Figure 4.18 Decision Tree structure resultant with the binary classification in CART fen and saltmarsh</i>	120
<i>Figure 4.19 Decision Tree output for saltmarsh after cost-complexity pruning</i>	121
<i>Figure 4.20 Comparison of accuracies DT v MLC. Fen as class of interest</i>	122
<i>Figure 4.21 Comparison of accuracies DT v MLC. Saltmarsh as class of interest.</i>	122
<i>Figure 5.1 Chapter 5 structure.</i>	128
<i>Figure 5.2 Different possibilities of boundary fit around the target data being B1 and B2 two input variables.</i>	130
<i>Figure 5.3 Overall accuracy results for the one-class classifiers Fen as class of interest.</i>	138
<i>Figure 5.4 Overall accuracy results for the one-class classifiers Saltmarsh as class of interest.</i>	138
<i>Figure 5.5 Training data for fen in 2 dimensional feature space.</i>	139
<i>Figure 5.6 Testing data for fen in 2 dimensional feature space</i>	139
<i>Figure 5.7 Training data for saltmarsh in 2 dimensional feature space</i>	140
<i>Figure 5.8 Testing data for saltmarsh in 2 dimensional feature space</i>	140
<i>Figure 5.9 Producer's and user's accuracy for the one-class classifiers. Fen as class of interest....</i>	141
<i>Figure 5.10 Producer's and user's accuracy for the one-class classifiers. Saltmarsh as class of interest.</i>	142
<i>Figure 5.11 The hypersphere containing the target data, described by the center a and radius R. Based upon Tax (2001).</i>	144
<i>Figure 5.12 Training sizes impact on overall classification accuracy (fen)</i>	151
<i>Figure 5.13 Training sizes impact on overall classification accuracy (saltmarsh).</i>	151
<i>Figure 5.14 Training set with outliers. 75% target data 25% outliers. Fen as class of interest.</i>	153
<i>Figure 5.15 Comparison of accuracies SVDD v MLC. Fen as class of interest</i>	154
<i>Figure 5.16 Comparison of accuracies SVDD v MLC. Saltmarsh as class of interest</i>	154

<i>Figure 6.1. Reasons why best individual classifiers do not guarantee an optimal solution. C_1 C_2, C_3 represent the individual classifiers. C_e is the classifier obtained by an ensemble and C_{opt} is the optimal classifier. H is the search space where the classifiers are looking for the optimal solution. In 1) the ensemble C_e is closer to the optimal classifier 2) the search path of each classifier might not go near the optimal classifier 3) the optimal classifier might be outside the search space. Based upon Günter and Bunke (2004).</i>	158
<i>Figure 6.2 Chapter 6 structure</i>	160
<i>Figure 6.3 Different combination methods for ensembles of classifiers</i>	161
<i>Figure 6.4 Schematic representation of the SVM Ensemble of classifiers</i>	168
<i>Figure 6.5 SVM ensemble producer's accuracy results</i>	170
<i>Figure 6.6 Schematic representation of the DT Ensemble of classifiers</i>	173
<i>Figure 6.7 Misclassified fen pixels in the feature space</i>	175
<i>Figure 6.8 Schematic representation of the One-class classifiers Ensemble</i>	178
<i>Figure 6.9 Schematic representation of the SVDD Ensemble of classifiers</i>	180
<i>Figure 6.10 Schematic representation of the All classifiers Ensemble</i>	184
<i>Figure 7.1 Comparison of accuracies for ML, SVDD, SVM and DT classifiers. Fen as class of interest</i>	189
<i>Figure 7.2 Comparison of accuracies for ML, SVDD, SVM and DT classifiers. Saltmarsh as class of interest</i>	190
<i>Figure 7.3 Comparison of overall accuracies SVM-DT-SVDD for fen as class of interest</i>	191
<i>Figure 7.4 Comparison of producer's accuracies SVM-DT-SVDD for fen as class of interest</i>	192
<i>Figure 7.5 Comparison of user's accuracies SVM-DT-SVDD for fen as class of interest</i>	192
<i>Figure 7.6 McNemar's test. Z values for SVDD-DT.</i>	193
<i>Figure 7.7 McNemar's test. Z values for SVDD-SVM.</i>	193
<i>Figure 7.8 McNemar's test. Z values for DT-SVM.</i>	194
<i>Figure 7.9 Test area River Yare NNR</i>	195
<i>Figure 7.10 Comparison of overall accuracies between SVDD, SVM and DT classifiers and respective ensembles.</i>	205
<i>Figure 7.11 Comparison of user's accuracies between SVDD, SVM and DT classifiers and respective ensembles.</i>	205
<i>Figure 7.12 Comparison of producer's accuracies between SVDD, SVM and DT classifiers and respective ensembles.</i>	206
 <i>Table 2.1. Supervised multiclass classifications methods</i>	 31
<i>Table 2.2. Advantages and disadvantages of parametric and non parametric Multiclass classifiers.</i>	44
<i>Table 2.3. One-class classification methods</i>	48
<i>Table 3.1. Based upon information from Nasa archives. Updated April 2, 1999. (http://www.nasa.gov).</i>	60
<i>Table 3.2. Error matrix example</i>	72
<i>Table 3.3 Confusion matrix for McNemar's test. Based upon Foody (2004).</i>	75

<i>Table 4.1 ML classification. Training set 1. 15 pixels per class</i>	<i>84</i>
<i>Table 4.2 ML classification. Training set 2. 150 pixels per class</i>	<i>84</i>
<i>Table 4.3 ML classification confusion matrix using 150 pixels per class. Fen as the class of interest (Testing set formed 50% fen, 50% others)</i>	<i>86</i>
<i>Table 4.4 ML classification confusion matrix using 15 pixels per class. Fen as the class of interest (Testing set formed 50% fen, 50% others)</i>	<i>87</i>
<i>Table 4.5 ML classification confusion matrix using 150 pixels per class. Saltmarsh as the class of interest (Testing set formed 50% saltmarshes, 50% others)</i>	<i>88</i>
<i>Table 4.6 ML classification confusion matrix using 15 pixels per class. Saltmarsh as the class of interest (Testing set formed 50% saltmarshes, 50% others)</i>	<i>89</i>
<i>Table 6.1 Error matrix for the SVM ensemble using fen as the target class</i>	<i>169</i>
<i>Table 6.2 Error matrix for the SVM using fen as the target class with a training size of 150 pixels .</i>	<i>169</i>
<i>Table 6.3 Error matrix for the DT ensemble using fen as the target class.....</i>	<i>174</i>
<i>Table 6.4 Error matrix for the DT using fen as the target class with a training size of 150 pixels....</i>	<i>174</i>
<i>Table 6.5 Coordinates and details of misclassified pixels in the SVM and DT ensembles.....</i>	<i>176</i>
<i>Table 6.6 Error matrix for the one-class classifiers ensemble using fen as the target class.....</i>	<i>178</i>
<i>Table 6.7 Error matrix for the SVDD Ensemble using fen as the target class</i>	<i>181</i>
<i>Table 6.8 Error matrix for the SVDD using fen as the target class with a training size of 150 pixels</i>	<i>181</i>
<i>Table 6.9 Coordinates and details of misclassified pixels in the SVDD ensemble.....</i>	<i>182</i>
<i>Table 6.10 Error matrix for the all classifiers ensemble using fen as the target class.....</i>	<i>185</i>
<i>Table 7.1 SVM, DT and SVDD advantages and disadvantages</i>	<i>204</i>
<i>Equation 2.1 Posterior probability.....</i>	<i>34</i>
<i>Equation 3.1 Equation used to calculate NDVI values</i>	<i>69</i>
<i>Equation 3.2 KHAT statistic.....</i>	<i>73</i>
<i>Equation 3.3 Z statistic test</i>	<i>74</i>
<i>Equation 3.4 McNemar's test.....</i>	<i>74</i>

1 Introduction

"Do you suppose I could buy back my introduction to you?"

Groucho Marx

The purpose of this chapter is to introduce the framework within which this thesis has been developed. It explains the background and motivations that justify the aims and objectives of the research undertaken. Finally, it also describes the structure that this thesis will follow.

1.1 Background

The purpose of this thesis is to investigate and evaluate methods for the accurate mapping of one particular habitat of interest with the aid of remote sensing. The reason for focusing all the attention on one particular habitat is that specific habitats are being fragmented and lost worldwide with major negative impacts on biodiversity. This is widely considered to be one of the greatest challenges currently facing society and constitutes a global problem with economic, biological, societal and ethical consequences (Van Kooten *et al.*, 2000, Harris, 2004). Therefore, the mapping and monitoring of these specific habitats is of utmost importance.

This general concern for habitat conservation turned into political commitment through the signing of the Biodiversity Convention at the Rio "Earth Summit"

Conference in 1992 by representatives of 168 countries (Ledoux *et al.*, 2000). Progress has occurred since then, but there are still various issues regarding which habitats to protect and how much land is required for their conservation (Jones, 2004). For example, the island ecology theory (Macarthur and Wilson, 1967) contends that species richness increases with an increase in the size of the habitat. However, the mosaic-concept developed by Duelli (1997) considers that species richness increases the more habitat types there are and the more heterogeneous these habitats are.

Apart from this debate about the optimal size of habitats for biodiversity conservation, the concept of habitat itself has been the focus of much research. Since the early 1900s there have been various attempts to define the term habitat. For example, Clements (1916) associated habitat with a “community of plant species in a phytosociological sense”; Lindeman (1942) went beyond that point and defined habitat as a “functioning ecosystem”. Mills (1969) directly associated community of species with a particular environment (the habitat) and concluded that the interaction of these species with each other and with the habitat makes them distinctive communities separable from other groups. In the late 1980s the European Union (EU) Habitats Directive defined natural habitats as “terrestrial or aquatic areas distinguished by geographic, abiotic and biotic features, whether entirely natural or semi-natural” (Council Directive 92/43/EEC). It is this latter definition that this thesis will adopt.

Finally, there are also concerns regarding the definition of boundaries and energy flows within a habitat which have repercussions for its protection and conservation (Griffiths, 1999). Habitats are dynamic entities. In this sense, since the 1980s attention has been called towards the instability and chaotic fluctuations that characterise many environmental systems (Pickett *et al.*, 1989). This is a particular challenge for habitat protection since habitats are systems opened to external as well as internal exchanges of materials, energy and organisms and in constant evolution (Zimmerer, 1994). In this sense the EU Habitats Directive also feels the need to describe conservation status of a natural habitat. This conservation status is defined

as the sum of the influences acting on a natural habitat and its typical species that may affect its long-term natural distribution, structure and functions (Council Directive 92/43/EEC).

The above mentioned EU Habitats Directive is part of the European Union's response to the problem of biodiversity loss as a signatory of the Biodiversity Convention. Its main aim is to protect biodiversity through the conservation and protection of natural habitats and of wild fauna and flora (Council Directive 92/43/EEC). The implementation of the Directive establishes the designation of Special Areas of Conservation (SACs) and Special Protection Areas (SPAs) which form the Natura 2000 network of protected areas. The aim of Natura 2000 is to provide the framework for the conservation of the 169 habitat types and 623 species identified in Annexes I and II of the Directive. These habitat types and species are those considered to be most in need of conservation at a European level because they are particularly vulnerable and are mainly, or exclusively, found within the European Union. In the UK the Directive has been transposed into legislation by The Conservation (Natural Habitats, etc.) Regulations 1994 and The Conservation (Natural Habitats, etc.) (Northern Ireland) Regulations 1995 as amended (informally known as 'The Habitats Regulations'). Once designated, Natura 2000 sites are to be protected from deterioration and damage and so in effect the Directive is underpinned by a no-net-loss policy (Ledoux *et al.*, 2000). Consequently, any loss of protected habitats must be compensated by restoration or creation of new ones of at least the same surface area and providing the same ecological value. Therefore, sustained accurate mapping of habitats is of high importance in order to keep track of any habitat losses (Turner *et al.*, 1998).

To meet the reporting requirements for Natura 2000, local authorities have to produce protected habitat mapping which has to be updated every 6 years. Due to financial budgets of local and regional authorities, the need for innovative methods that aim to optimise their resources are essential (Weiers *et al.*, 2004). In this sense, these authorities are normally interested in a sub-set of the habitats or just one habitat. This means that an approach which directs all the available resources to map

habitats of interest could be of great importance for those authorities in charge of mapping and monitoring protected habitats under the EU Habitats Directive.

In this sense, satellite remote sensing data are important sources for producing valuable land cover information which is essential to resource management and monitoring programmes. Land cover information derived from satellite remote sensing is also an important input in a number of ecological models (Kasetkasem *et al.*, 2005) and it is particularly useful when repeated measurements at frequent intervals are needed for habitat monitoring (Fassnacht *et al.*, 2006). Furthermore, it is well-suited to mapping and monitoring land cover at a range of scales (e.g. Roy and Tomar, 2000, Amarnath *et al.*, 2003, Kerr and Ostrovsky, 2003, Rouget, 2003), including those associated with the demands of the EU Habitats Directive. In particular, satellite systems such as Landsat, le Systeme pour l'Observation de la Terre (SPOT) and National Oceanic and Atmospheric Administration (NOAA) have the capacity of observing the Earth's land cover at various scales and time intervals supplying vital information that previously would have been impossible to acquire. Nevertheless, it is important to take into consideration that the quality of remote sensing imagery can vary considerably due to different atmospheric and technical conditions during the acquisition process (Campbell, 2002). Also, the information recorded is restricted to the energy returned in one or more wavebands which in some cases could result in some land cover classes being not identified properly. These issues are normally addressed by using ancillary data, multiple processing methods or combined with other analysis such as GIS and other data sources (Fassnacht *et al.*, 2006).

Another issue regarding land cover mapping is that the no-net loss policy of the EU Habitats Directive means that the accuracy with which the habitat of interest is classified is critical. Furthermore, in management and policy applications high accuracies are needed due to the implications that this information might have in decision making (Fassnacht *et al.*, 2006). Although there are many research studies using supervised classification to map land cover, there are still many issues that limit the classification and mapping accuracy (Foody, 2002). These problems range

from the spectral and spatial resolution of the imagery available (which might not be the most appropriate for the particular case) through to the nature of the classes on the ground. As a result, there is extensive research to increase the accuracy with which land cover information acquired from remote sensing imagery is translated into land cover maps. In particular, a large amount of effort has been directed towards improving classification methods (Foody, 2002).

The most common method of producing land cover maps from satellite images is through supervised image classification. This method consists of assigning each pixel that forms the image to a land cover class defined by the analyst. It is referred to as being supervised because the analyst provides examples of each class to the classifier and these are used to outline decision rules in order to label all the other pixels in the image (Mather, 2004). This is what is called a crisp or hard classification. However, it could be the case that a single pixel does not correspond to a single class and it could be that two or more classes are present within the same pixel. This is a recurrent problem in extracting accurate land cover information from remote sensing imagery and unmixing classifiers has been used in order to resolve it. These classifiers calculate the proportion of land cover classes inside the pixel and can be very informative in that sense. However, they fail to account for the spatial distribution of each of these classes within the pixel (Verhoeve and Wulf, 2002). Normally, the output consists of fraction images equal to the number of classes with each fraction image defining the proportion of that class present within each pixel.

Taking all the above into account, in this thesis it will be assumed that (i) a single pixel is assigned to a single land cover class (although the limitations of crisp classification will be taken into consideration) and (ii) a single habitat of interest corresponds to a single land cover class and that the terms habitat and class may, therefore, be used synonymously.

In the case that concerns this thesis, mapping one specific habitat of interest, the use of standard image classification approaches could result in very low accuracies. The reason for this is that standard classification analyses assume that the classes are

discrete, mutually exclusive and have been exhaustively defined. Therefore, all the classes have to be included in an image classification to ensure the assumption of an exhaustively defined set is satisfied (Foody, 2004a, 2004b). Furthermore, the size of the training set required for a classification is linked to the number of classes as well as the dimensionality of the data used (e.g. number of wavebands). But in the case of focusing on only one habitat of interest, most of the classes may be of little or no interest. Consequently, the requirement to acquire training data for all the classes, including those that are not of interest is a waste of resources and effort.

In addition to the above, standard image classifiers consider all classes in their analysis, paying equal attention to all the classes. For example, in standard probabilistic approaches such as maximum likelihood (ML) classification, the aim is to maximize the overall probability that a pixel is allocated to a class correctly. The aim is to have a high overall accuracy rather than a high accuracy for the specific class of actual interest (Lark, 1995c). Therefore, when the emphasis of the classification lies upon a specific habitat as with the requirements from the Habitats Directive, it may be more appropriate and efficient to focus on the single class of interest, producing an alternative and more efficient approach to those responsible for habitat management (Pullin *et al.*, 2004).

The idea of focusing on the classification of a class of interest has proved very valuable in other research areas of pattern recognition such as document classification (distinguishing one specific category from other categories) (Manevitz and Yousef, 2001), texture segmentation (distinguishing one specific texture from other textures) (Tax and Duin, 2002), and image retrieval (retrieving a subset of images based on the similarity between given query images) (Lai *et al.*, 2002), but not yet explored within remote sensing for land cover mapping. This thesis aims to address this issue by assessing a variety of methods that could be used to accurately map a single class from remotely sensed imagery. In theory a standard probabilistic classifier could be optimized to minimize error associated with the class of interest (Lark, 1995c). This approach, however, typically trades one type of error with another and may require further analyses to produce a final map of the class of

interest (Foody *et al.*, 2005). Alternatively, a soft classification may be used. This approach is also not without problems such as the decision of a threshold membership value to apply in separating members of the class from non-members. Moreover, with both approaches, assumptions of the data are also commonly made with the classifiers used. Consequently, this research will focus on classification methods which aim to map and monitor a specific class of interest. These methods are:

1) A non-parametric binary classification analysis that simply seeks to separate the class of interest from all others. With this approach only a small sample of all the other classes would be required and the classifier would be concentrating on separating two classes. This would address the two problems described above: (i) wasteful use of training data and (ii) wasteful consideration of all the classes by the classifier. Attractive means to achieve this are through the adoption of decision tree (DT) classifiers and support vector machine (SVM) classifiers.

A DT classifier learns from a data set and is able to classify previously unseen cases through the formulation of explicit rules (Goel *et al.*, 2003). They have long been popular in machine learning, statistics and other disciplines and only recently, DTs have become popular for the classification of remotely sensed data and the production of land cover maps (Pal and Mather, 2003, Brown de Colstoun *et al.*, 2003, Joy *et al.*, 2003). DT algorithms are less demanding than the conventional ML classifier due to their non-parametric nature, conceptual simplicity and computational efficiency (Friedl and Brodley, 1997, Pal and Mather, 2003).

As with the DT approach, the potential of SVM for the classification of remotely sensed data has been recognized recently. Comparative studies have also demonstrated that a SVM may be used to classify land cover more accurately than conventional approaches, such as the ML classifier as well as popular alternatives like Artificial Neural Networks (ANNs) (Huang *et al.*, 2002, Foody and Mathur, 2004a). A SVM is a binary classifier that seeks to fit an optimal separating decision boundary between the classes. It is therefore, also well-suited to the mapping of a

single class of interest, by separating it from all others. The advantage of the SVM is the potential for accurate classification from small training sets, particularly if intelligently defined (Foody and Mathur, 2004b). Obviously some effort is required in training the non-habitat class in both DT and SVM but the degree of precision required is less than in training a conventional classifier and therefore the classification can be considerably less demanding and less wasteful than standard classifications (Foody and Mathur, 2004b).

2) Another approach would be to concentrate completely on the class of interest disregarding the other classes present in the image which yet again could deal with both problems. Concentrating just on one class and using only data from such a class of interest can be achieved through one-class classification methods. These methods have been extensively applied within the area of pattern recognition but have not yet been applied within the remote sensing community. The advantage of one-class classifiers is that they can focus totally upon the class of interest without the need of data from any other class present in the image.

3) Finally, as mentioned earlier in this Chapter, the accuracy with which the class of interest is classified is essential. In recent years, research into improving classification accuracy has been carried out within remote sensing proposing the ensemble of classifiers to produce a single classification. The basis of an ensemble of classifiers is that if there are different classifiers that can be applied within a research project, it would be reasonable to consider using them in combination in the hope of increasing the overall accuracy (Jain *et al.*, 2000). This approach has received a lot of attention in pattern recognition and machine learning but it is only very recently that some research has started in the area of remote sensing. This new approach will also be explored.

1.2 Aims and objectives of this thesis

The motivation of the present research is directly related to the classification problem of one class of interest which is exemplified by the current protection of habitats legislation designed by the EU Habitats Directive, with a view to consider how remote sensing could help to achieve a more accurate monitoring of such protected habitats. As seen in the previous section, to classify a particular habitat of interest is a challenging task that requires approaches different from the traditional multiclass classification. Taking all this into account and as mentioned at the very beginning of this Chapter the main aim of this thesis is:

To investigate and evaluate methods for mapping one particular habitat of interest with the aid of remote sensing.

As already mentioned in the previous section, sub-aims include increasing the classification accuracy when focusing on a class of interest by:

- 1) Optimising the use of training data.
- 2) Optimising the use of remote sensing methods by applying suitable classifiers to the specific task of classifying a class of interest.

Specific objectives in order to achieve the above aim and sub-aims include the assessment of different classifiers and their specific associated issues. These main objectives consist of:

- Evaluating the potential of binary classification for the mapping of a specific habitat of interest using SVM and DT classifiers.
- Evaluating the potential of one-class classification for the mapping of a specific habitat of interest.

- Evaluating the potential ensemble of classifiers in order to obtain a higher classification accuracy.

Each of these three objectives will be addressed within the three main research chapters of this thesis.

1.3 Area of study and habitats of interest

To carry out the evaluation of the above mentioned classifiers, it was decided to perform the different classifications on a particular habitat of interest protected by the EU Habitats Directive. A second habitat of interest was also chosen to determine whether the classification results were biased by the type of habitat. The first habitat chosen to test the different approaches described in the previous sections was fen. There are several reasons for choosing this habitat. For example, there are no examples in the literature of mapping and monitoring fens using satellite remote sensing methods. The reason for this could be that its rich composition (which could make it difficult to define spectrally) and its dynamic nature have encouraged field and aerial photography studies as opposed to satellite remote sensing analysis. This makes fens a particular challenge for testing the capabilities of the classifiers used within this thesis. The other habitat chosen for comparison purposes was saltmarsh. Together with fens, this is a very dynamic habitat protected by the EU Habitats Directive. However, in this case, it has been the focus of extensive and successful studies for mapping and monitoring using satellite remote sensing imagery. Therefore, the results of its classification could serve as a valid point of reference against those obtained for fens and determine any bias of the classifiers towards a particular type of habitat.

The area chosen for the study of both habitats was East Anglia, United Kingdom. Here, the Norfolk Broads is one of two sites selected as SACs under the EU Habitats Directive for alkaline fens in East Anglia, where a main concentration of lowland fens occurs. Also East Anglia includes the area of the North Norfolk coast which has

been selected as a SAC for its saltmarshes which are unparalleled among coastal sites in the UK for their diversity and are amongst the most important in Europe.

1.4 Thesis structure

In order to achieve the previously mentioned aims and objectives, this thesis has been structured in seven chapters (see Figure 1) including this introductory chapter which describes the context within which this research has been developed.

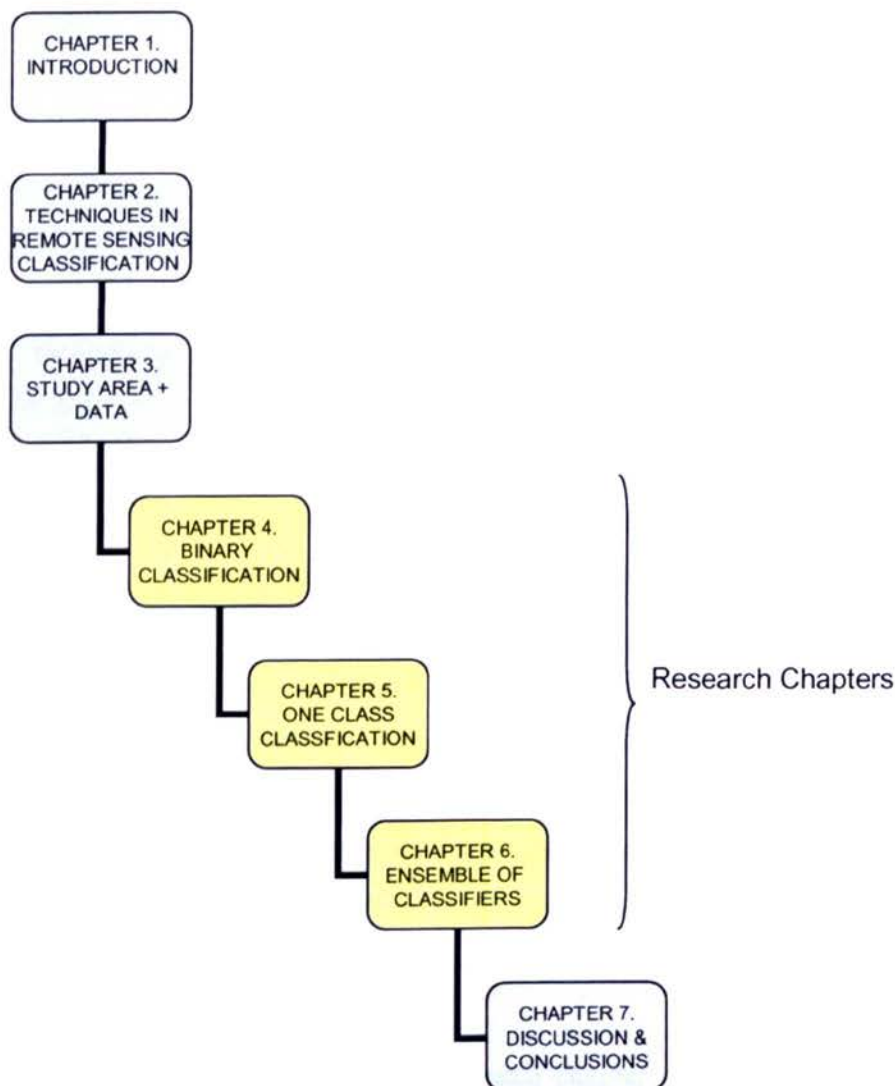


Figure 1.1 Thesis structure. The first three chapters show the background of the research. The research chapters shift towards the right showing the progression of the research towards the final discussion and conclusions.

The first three chapters provide background information about the research context, different classification methods and study area and methodology. In particular Chapter 2 reviews the current classification approaches in remote sensing which are mainly orientated towards multiclass classification. The basics and process of remote sensing classification are explained. Different parametric and non-parametric approaches are described and compared and new approaches such as SVMs and DTs and one-class classifiers are introduced.

Chapter 3 focuses on the methodology followed to carry out the present research. It describes the area of study chosen and the data used in terms of remote sensing imagery and ground data. Furthermore, it describes the procedure followed to acquire the training and testing sets that will be used within the research chapters and the measure of the accuracy of classification outputs.

Chapter 4 is the first of the three research chapters within this thesis. It approaches the issue of classification of a class of interest through binary classification which corresponds with the first objective described in section 1.2. It contains a detailed description of the two main classifiers used for the binary approach: SVMs and DTs. The performance of both classifiers is measured against that of a classic parametric approach: ML classification.

Chapter 5 is the second research chapter and investigates the idea of one-class classification and its application within remote sensing and land cover classification. It explores different one class classifiers addressing the second objective of this thesis. The results obtained are also measured against the benchmark of the ML classification.

Chapter 6 is the third and final research chapter of this thesis which addresses the issue of the ensemble of classifiers (third objective) and assesses whether the combination of the classifiers used in the present research will increase the final classification accuracy.

Finally, Chapter 7 will discuss the results obtained by the different approaches described in the previous research chapters and assess whether they have met the aims and sub-aims of this thesis. It also summarises the findings of the research and presents the conclusions drawn from them. Ultimately it provides recommendations for future research.

2 Remote Sensing and Image Classification: Review and selection of suitable methods for classifying a habitat of interest

"Human beings, for all their pretensions, have a remarkable propensity for lending themselves to classification somewhere within neatly labelled categories.

Even the outrageous exceptions may be classified as outrageous exceptions!"

W.J. Reichmann

It is a well-known fact that land cover information is a critical variable linking human and physical processes (Boyd and Foody, 2004). However, the lack of up-to-date information on type, location, size and quantity of natural habitats has been identified as a major constraint for the implementation of the EU Habitats Directive (Weiers *et al.*, 2004).

In this sense, remote sensing technology is a valuable source of land cover information which is able to acquire data at various spatial and temporal scales (Huang *et al.*, 2002). The advantages of remote sensing over alternative forms of environmental data gathering are that large surface areas can be mapped and monitored. It is also less costly than aerial and ground surveys for long-term studies and for large and/or inaccessible areas (Wilkinson, 2000). This is very important because the common problem for mapping and monitoring programmes is to be able

to update existing data in a cost and time effective way. Also, the acquisition, checking and update of data take a great part of the mapping budget (Weiers *et al.*, 2004).

However, a concern about land cover maps derived from remotely sensed data is that there are normally disagreements between them and other maps derived from field surveys. Technical difficulties can occur in the acquisition process of the image such as adverse weather conditions or sensor characteristics. These difficulties are inherent to collecting data remotely and could have impacts on the quality of the final land cover map. Pre-processing techniques aim to correct many of these technical setbacks. A great part of the current research that aims to correct the disagreements between land cover maps derived from remotely sensed data and ground data is directed to decrease error in image classification (Foody, 1999, Mather, 1999) as well as assessing the suitability of remotely sensed data for certain mapping applications (Estes *et al.*, 1999; Wilkinson, 2000).

In order to understand the current issues relating to image classification methods and their applicability to the mapping of one habitat of interest, this chapter will describe the main steps in the process of land cover mapping from the image acquisition by the sensor to the accuracy assessment of the resulting land cover map. It will also address the advantages and disadvantages of current standard image classification methods and why new approaches are needed in order to classify a specific habitat of interest.

2.1 Remote sensing methods

2.1.1 Background

Remote sensing is based upon the interpretation of electromagnetic radiation (EMR) reflected or emitted by a target which is measured by a sensor that is distant from such a target (Mather, 2001). Each feature on the Earth's surface reflects or emits differently EMR. EMR is a form of energy transfer from one target to another through space and media and behaves in two inseparable ways; as regular waves of energy and as rapidly moving and indivisible particles. In remote sensing, electromagnetic waves are categorised by their wavelength location within the electromagnetic spectrum (Figure 2.1).

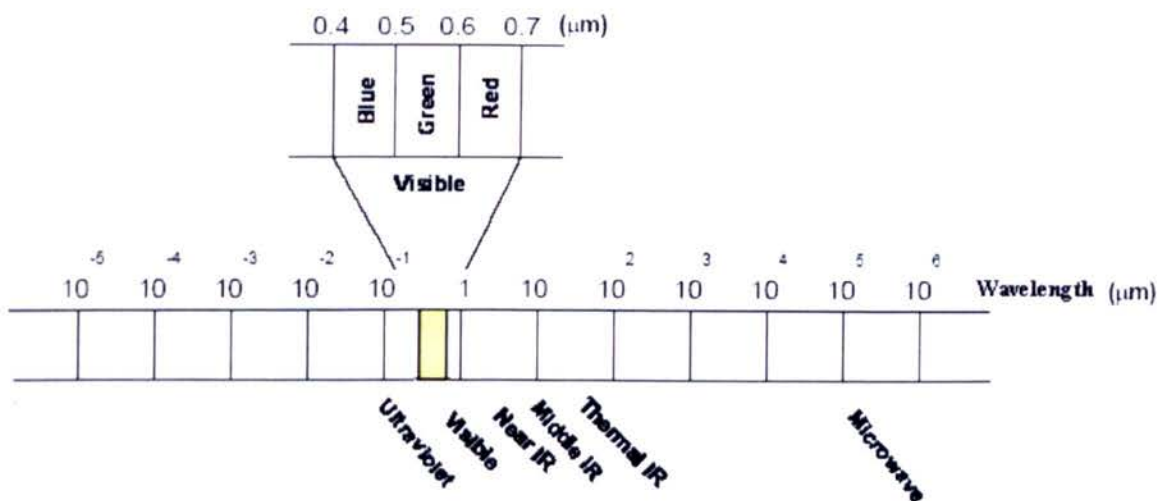


Figure 2.1 Electromagnetic Spectrum. Based upon Lillesand and Kiefer (2004)

Normally, the spectral sensitivity of the sensors used in remote sensing are within the ultraviolet, visible, infrared and microwave part of the spectrum and they capture the reflectance and emission of EMR from Earth surface features in these portions of the spectrum. Each type of feature has a typical spectral response which is normally known as the spectral signature. This is a very important characteristic of Earth surface features because this signature provides a specific description of the response

of this feature to EMR which could determine which sensor to choose for a particular application (Lillesand and Kiefer, 2004). A typical average spectral signature for three main Earth surface features (water, soil and vegetation) could be represented as in Figure 2.2.

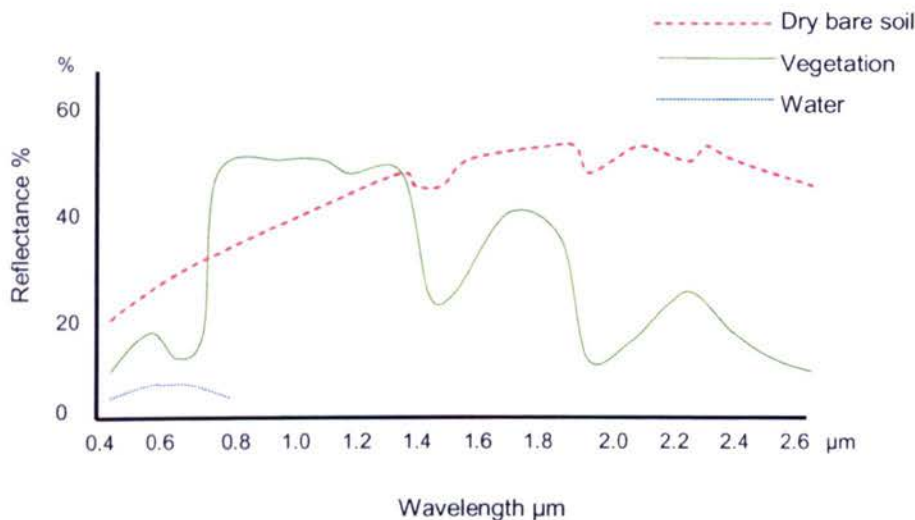


Figure 2.2 Typical spectral reflectance curve. Based upon Lillesand and Kiefer (2004).

When the data acquisition process is complete the result is a digital image that contains raw data. This digital image is a numerical record of the radiance reflected or emitted from each of the features on the ground in each of the spectral bands that are being recorded (Mather, 2001). A pixel is the cell of an output device to which a measurement is allocated. Each pixel is assigned with a digital number (DN) corresponding to the average radiance measured in this pixel. Typically, the DNs constituting a digital image are recorded over numerical ranges driven by the number of bits used to record the image. In such numerical formats, the image can be readily analysed with the aid of a computer (Lillesand and Kiefer, 2004).

The graphic representation of these pixels is normally carried out in a feature space which is the space where each pixel is represented as a point. Each measurement (feature or variable) about the pixel gives a coordinate along one axis of the space.

The dimensionality of the feature space is equal to the number of variables used (for example, if two features are used, the space will be a plane, with the first feature on the X axis, and the second feature on the Y axis). When a pixel is represented in a feature space it becomes a vector which is formed by the values of that pixel for each of the variables or features represented in such feature space (see Figure 2.3).

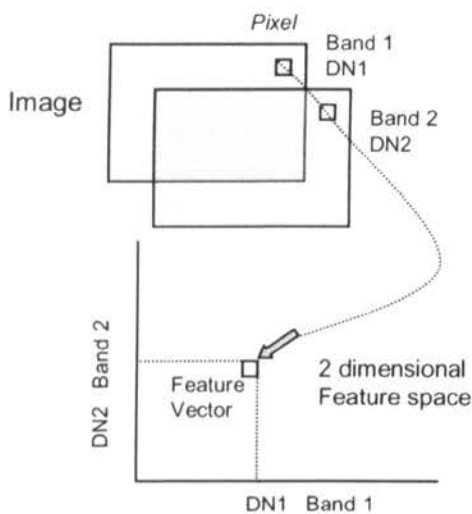


Figure 2.3 Representation of a pixel (feature vector) in a 2 dimensional feature space. After Jensen (1996).

Therefore pixels will also be referred to as vectors in many of the following sections within this thesis. Finally, the distance between two points or vectors within the feature space is expressed as Euclidean distance.

2.1.2 The image classification process

Image classification is one of the most common methods for image interpretation in remote sensing. The overall objective of image classification is to categorise all pixels in an image into land cover classes or themes. Normally, the DN for each pixel is used as the basis for this classification and it is allocated to a specific class which matches its spectral signature. This is known as spectral classification. There are also other types of classification methods used in image interpretation such as (i) spatial pattern recognition which consists of the classification of image pixels on the basis of

their spatial relationship with pixels surrounding them (Lillesand and Kiefer, 2004) and (ii) temporal pattern recognition which uses change in spectral reflectance over time as the basis of feature identification (Pal, 2002). Nevertheless, spectrally oriented classification is the most used procedure for land cover mapping (Lillesand and Kiefer, 2004) and it is also the one chosen for the purposes of this thesis. Generally, image analysis involves several steps (Figure 2.4 below):

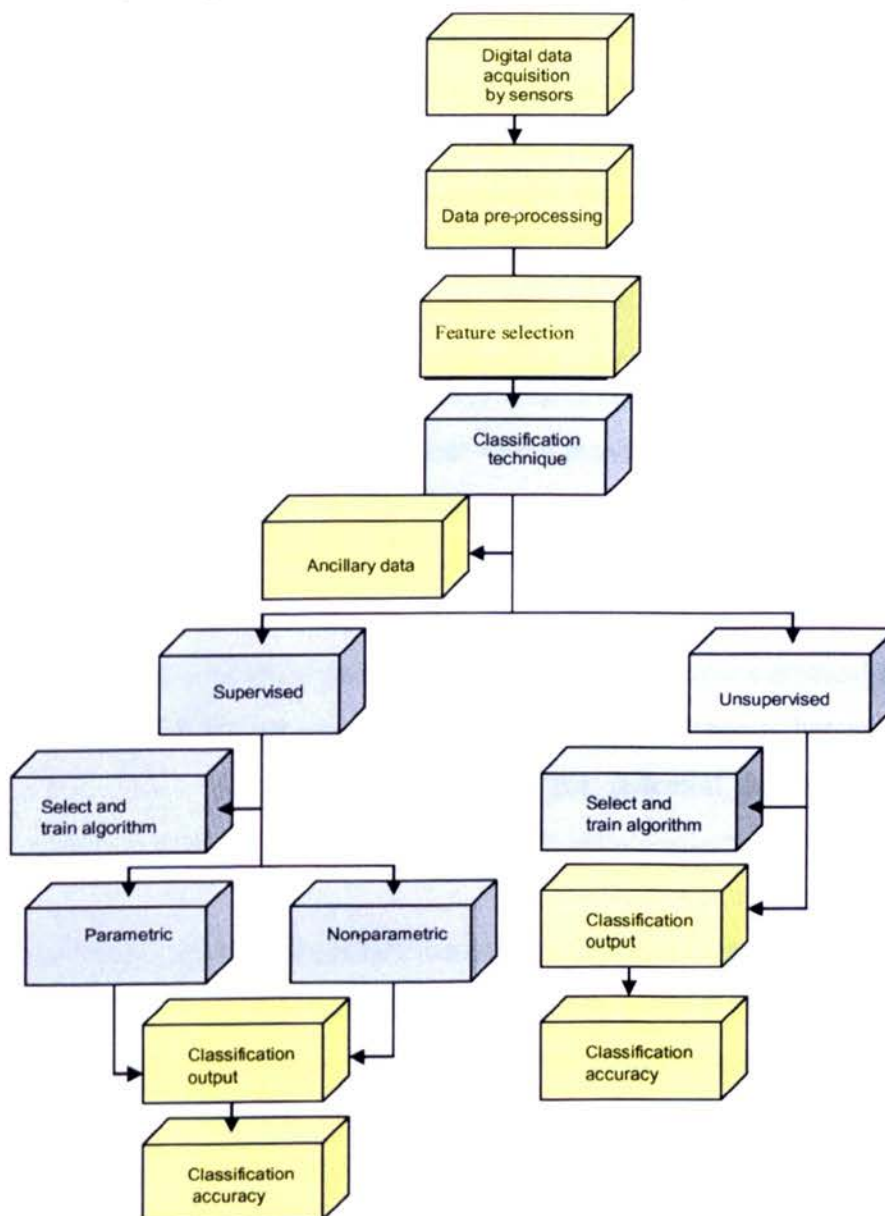


Figure 2.4. Digital image analysis. Based upon Campbell (2002). If there is ancillary data available the classification is normally supervised.

Pre-processing of the digital image consists of operations that prepare the data for further analysis and where the aim is to correct or compensate for random and systematic errors that might occur during the image acquisition such as defects in sensor operation, atmospheric absorption and scattering, variations in illumination, system noise and geometric corrections (Campbell, 2002). Once this is complete, it might be necessary to perform a feature extraction in order to reduce data dimensionality. Feature extraction optimises the use of the available data getting rid of any redundancy within this data. A discrimination technique can involve graphical and/or statistical analysis which will determine the degree of separability between classes. Graphical methods of feature selection include: (i) bar graph spectral plots, (ii) co-spectral mean vector plots and (iii) feature space plots (Jensen, 1996):

(i) Bar graph spectral plots. In this graphical representation, the means for each band are displayed in a bar graph format which provides an effective visual representation of the degree of separability between classes for one band at a time. However, the display provides no information on how well any two bands would perform.

(ii) Co-spectral mean vector plots can be used to present statistical information about at least two bands at one time. The greater the distance is between numbers in the feature space distribution, the greater the potential for accurate discrimination between classes.

(iii) Feature space plots in two dimensions represent the distribution of all the pixels in the scene using two bands at a time. The brighter the pixel is in the feature space plot display, the greater the number of pixels having the same values in the two bands of interest.

On the other hand, statistical methods are used to quantitatively select which subset of bands provides the greatest degree of statistical separability between any two classes. Separability measurements between two spectral classes may greatly reduce the occurrence of errors of commission and omission. The probability of these errors

is related to the statistical separability between spectral classes. This may be quantified by their relative divergence (d'Urso and Menenti, 1996). Divergence addresses the problem of deciding what the best subset of bands is. It is computed using the mean and covariance matrices of the class statistics collected in the training phase of the supervised classification. A problem with this technique is that it could highlight easily separable classes which will weight the average divergence upward in a misleading way, with the consequence that sub-optimal feature subsets might be indicated as best (Richards, 1993). To avoid this, a transformed divergence is used instead. Transformed divergence gives an exponentially decreasing weight to increasing distances between the classes (Jensen, 1996).

Having addressed the subjects of pre-processing and feature extraction, the next step in the image classification process is to choose which classification technique is going to be adopted (Figure 2.4). First of all, it is necessary to decide whether the classification is going to be supervised or unsupervised. Supervised classification methods are based upon ancillary data that provide the researcher with some sort of knowledge about the area to be classified. This ancillary data can be provided by field work, aerial photography, maps or reports about the area (Mather, 1999). On the contrary, in unsupervised classification no prior information about the land cover types or their distribution is known or required. This method organises the data into classes sharing similar spectral characteristics. Unsupervised classification methods divide the scene into more or less pure spectral clusters, which are constrained by pre-defined parameters that describe their statistical properties and their relationships with neighbouring clusters (Cihlar, 2000). However, in most studies there are usually some ground data available. Therefore from here on all the following subsections will be referring to supervised classification and the steps required to perform this classification.

2.1.3 Choosing a classification method and training the classifier

In supervised classification, in order to classify an image into categories or land cover classes, the classification algorithm needs to be trained to distinguish those categories. Areas that have similar spectral characteristics are labelled and called class signatures. However, it is important to take into account that within an area of particular land cover several spectral classes can occur, resulting in a heterogeneous spectral signature whose characteristics depend on the proportions of each of the component land-cover types (Schowengerdt, 1997). Standard classification algorithms produce a likelihood function in order to assign a class label to each pixel. As mentioned in Chapter 1, when a class label is assigned to each pixel a hard classification is produced. When allowing for multiple labels at each pixel a soft classification is created (Schowengerdt, 1997). In recent years numerous variants of these two basic classification methods have been developed. These include Decision Trees (DTs) (Hansen *et al.* 1996), Artificial Neural Networks (ANNs) (Carpenter *et al.*, 1997, Foody *et al.*, 1997, Bischof and Leonardini, 1998, Yool, 1998), fuzzy classification (Foody, 1996, 1998, Mannan *et al.*, 1998), Support Vector Machines (SVMs) (Vapnik, 1998) and mixture modelling (van der Meer, 1995) for supervised classification; and classification by progressive generalisation (Cihlar *et al.*, 1998), classification through enhancement (Beaubien *et al.*, 1999) and post-processing adjustments (Lark 1995a, Lark 1995b) for unsupervised classification. Table 2.1 summarises the different types of hard and soft supervised classification methods.

Supervised	Parametric	Hard classification	Soft classification
		Minimum Distance to Mean Classifier	
		Parallelepiped Classifier Maximum Likelihood Classifier	
	Non Parametric	K-nearest neighbour Decision Trees	
		Neural Networks Support Vector Machines	
			Mixture modelling Fuzzy classification

Table 2.1. Supervised multiclass classifications methods

As already established in the introduction, this thesis is not dealing with the issue of mixed pixels classification. Therefore the following sections will be concentrating on supervised hard classification methods.

The supervised classification approach involves training the classifier with a number of sites where the class signature is known. It is very important to define training areas which represent the spectral characteristics of each class because the quality of the training set has a significant effect on the classification process and its accuracy (Chuvieco and Congalton, 1988). It is also important that the training areas are a homogeneous sample of the respective class, but at the same time include the range of variability for the class. In many cases it is impossible to obtain homogeneous sites. One technique for improving training data under these conditions is to “clean” the sites of outlying pixels before developing the final class signatures. When classifying the training pixels according to their given signatures some training pixels are likely to be misclassified. These pixels could be excluded from the training set and the class signatures are recalculated from the remaining pixels (Schowengerdt, 1997).

Commonly, supervised image classification constitutes the basis of land cover and land cover change assessments. Supervised classification algorithms may be grouped into one of two types: (a) parametric classifiers, if they assume the existence of an underlying probability distribution of the data, and (b) non-parametric classifiers, if they do not assume anything about the probability distribution (Cortijo and Perez de la Blanca, 1998).

Parametric Classifiers

Generally, parametric classifiers assume a gaussian distribution of the data and are based on the mean vector and the covariance matrix of learned normal distributions (Hubert-Moy *et al.*, 2001). One of the simplest approaches to supervised parametric classification is the Minimum Distance (MD) to mean classifier (Figure 2.5). This classifier utilises the Euclidean distances in spectral feature space between (i) the pixels to be classified and (ii) the class means (obtained from training data). Each pixel in the remainder of the image may then be allocated to the class mean to which it is nearest (minimum distance) in multivariate feature space (Atkinson and Lewis, 2000). The main disadvantage of this classifier is that it assumes that classes are symmetric in multispectral space (Atkinson and Lewis, 2000).

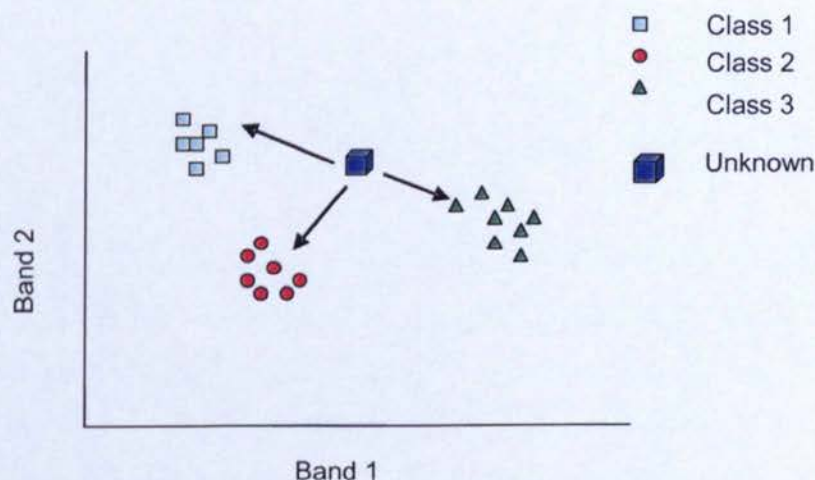


Figure 2.5. Minimum Distance to mean (MD) classification strategy. Based on Lillesand and Kiefer (2004).

The Parallelepiped classifier introduces sensitivity to each class variance by considering the range of values in each category (Lillesand and Kiefer, 2004). An unknown pixel is classified according to the category range, or decision region, in which it lies or it is classified as “unknown” if it lies outside all regions. These regions are then of a rectangular shape and are referred to as parallelepipeds (Figure 2.6).

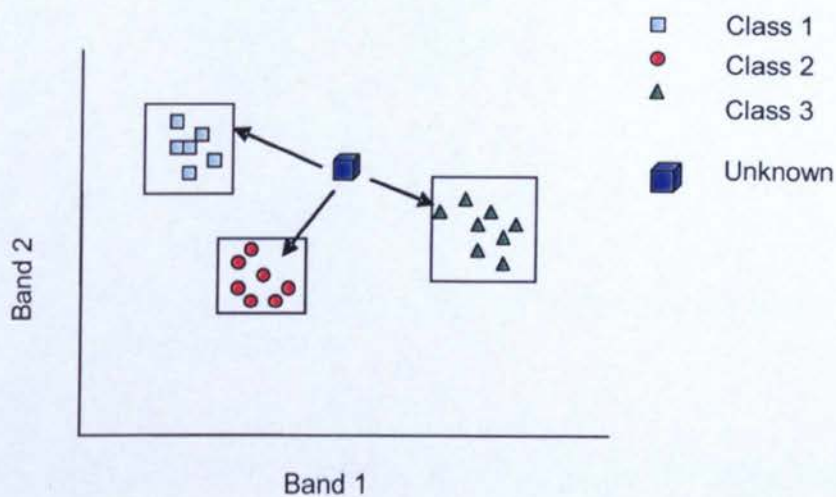


Figure 2.6. Parallelepiped classification strategy. Based on Lillesand and Kiefer (2004).

However, a disadvantage of this classification becomes obvious when category ranges overlap which occurs with the presence of covariance. Pixels that occur in the overlap areas will be classified as “not sure” or will be arbitrarily placed in one or both of the two overlapping classes (Lillesand and Kiefer, 2004).

The Maximum Likelihood (ML) classifier is the most popular parametric method. It quantitatively evaluates both the class variance and covariance when classifying an unknown pixel (Lillesand and Kiefer, 2004). The data likelihood for each class can be weighted with some estimate of *a priori* probability (frequency) that this class occurs (Strahler, 1980, Hubert-Moy *et al.*, 2001). These *a priori* probabilities may be estimated from external information sources such as ground surveys, existing maps or historical data (Schowengerdt, 1997). The probability density functions are used to

classify an unidentified pixel by computing the maximum likelihood of the pixel value belonging to each category. After evaluating the probability of each category, the pixel would be assigned to the most likely class (Figure 2.7).

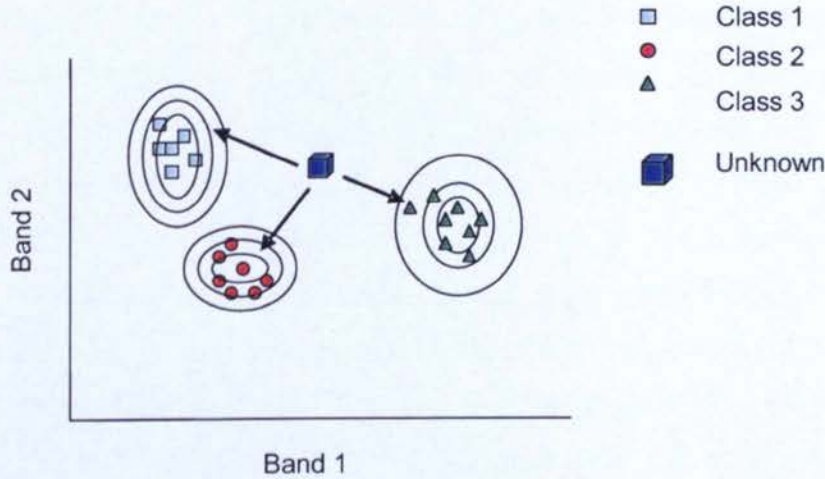


Figure 2.7. Equiprobability contours defined by a maximum likelihood classifier. Based on Lillesand and Kiefer (2004).

The likelihood (L_k) is defined as the posterior probability of a pixel belonging to class k :

$$L_k = P(k / \mathbf{x}) = \frac{P(k) \cdot P(\mathbf{x} / k)}{\sum P(i) \cdot P(\mathbf{x} / i)}$$

Equation 2.1 Posterior probability

Where $P(k)$ is the prior probability of class k and $P(\mathbf{x}/k)$ is the conditional probability to observe \mathbf{x} from class k or probability density function. Usually $P(k)$ are assumed to be equal to each other and $\sum P(i) \cdot P(\mathbf{x} / i)$ is also common to all classes. Therefore L_k depends on $P(\mathbf{x}/k)$ or the probability density function. If the probability values are all below a threshold set by the analyst \mathbf{x} will be labelled unknown (Lillesand and Kiefer, 2004).

Applications of Maximum Likelihood (ML) classification are well established in the literature of remote sensing (Swain and Davis, 1978, Estes *et al.*, 1983, Schowengerdt, 1983, Sabins, 1997, Lillesand and Kiefer, 2004, Jensen, 1996).

Summarising, statistical procedures require that data must be based on some pre-defined model (usually the gaussian normal distribution). Consequently the performance of a these classification methods will depend on how well the data match the pre-defined model. If the data are complex in structure then to model those in an appropriate way can become a problem. Other drawbacks include that each sample is tested against all classes, which leads to a relatively high degree of inefficiency. Out of the three classifiers, the ML classifier addresses the problems of MD to mean and Parallelipiped classifiers and in spite of its disadvantages it has been adopted as a standard classification technique for many land cover mapping applications (Pal, 2002).

Non-parametric Classifiers

To try to solve the problems of parametric classification non-parametric classifiers are being introduced into image classification methods. Non parametric classifiers include k-Nearest Neighbour (k-NN) (Cover and Hart, 1967, Devijver and Kittler, 1982), Artificial Neural Networks (ANNs) (Bishop, 1995), Decision Trees (DTs) (Breiman *et al.*, 1984) and very recently kernel methods and Support Vector Machines (SVMs) (Vapnik, 1998). Due to their flexibility, these non-parametric methods are appealing alternatives to parametric ones. However, the training remains a critical issue. Most of them usually require large training sets to be properly trained (Hubert-Moy *et al.*, 2001).

The k-NN method is widely used in pattern recognition. This method is effective for estimation of densities and for classification. As a classifier, this algorithm carries out the following three steps: (1) calculate the distances between an unknown sample and all training samples, (2) choose the k-nearest training samples to the sample, and

(3) assign a class label by applying the majority rule to the k -nearest samples (see Figure 2.8). The disadvantage of this method is the large amount of computation required to calculate the distances of a given sample to all training samples. Therefore, various methods for reducing the amount of computation have been proposed. One group of methods aims to reduce the size of the training sample set to be referred in the search stage. Another group aims to reduce the number of distance calculations in the search stage by adopting an efficient search procedure (Kudo *et al.*, 2003).

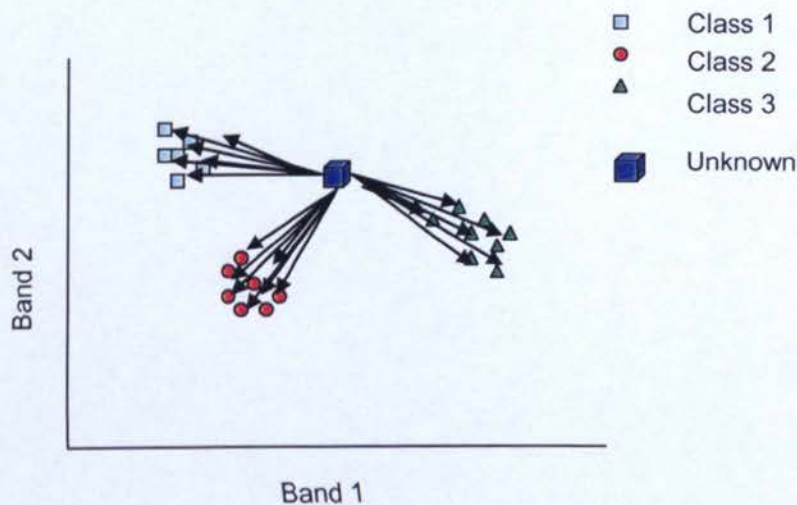


Figure 2.8. Example of k -NN classifier. Based on Lillesand and Kiefer (2004).

The ANN classifier is a recent popular non-parametric approach to classification. ANNs have been successfully used in remote sensing and image analysis including supervised classification (Benediktsson *et al.*, 1990, Hepner *et al.*, 1990, Heerman and Kahzenie, 1992, Foody and Arora, 1997, Mas, 2004, Ingram *et al.*, 2005) and unsupervised classification (Baraldi and Parmiggiani, 1995, Schaale and Furrer, 1995, Tso, 1997, Kurnaz *et al.*, 2004). The method with the widest reported application in image classification in recent years is back-propagation (BP), with several variants available. These employ many processing iterations to arrive at a solution (Brown *et al.*, 1998). The decision boundaries are not fixed by a deterministic rule applied to the training signatures but are determined in an iterative fashion by minimising an error criterion on the labelling of the training data

(Schowengerdt, 1997). In essence, a neural network may be considered to comprise a relatively large number of simple interconnected neurons or units that work in parallel to categorise input data into output classes (see Figure 2.9) (Hepner *et al.*, 1990, Schalkoff, 1992).

One of the key advantages of ANNs is the distribution-free nature. Prior knowledge about the statistical distribution of classes is not required (Brown *et al.*, 1998). However, previous work with ANNs has been adversely affected by a lack of reliable procedures for developing optimum network architectures, training data and sampling considerations, difficulty in reaching the global minima of the error curve, variable output depending on weight initialisation, and slow processing speeds (Foody *et al.*, 1995, Blamire, 1996, Brown *et al.*, 1998).

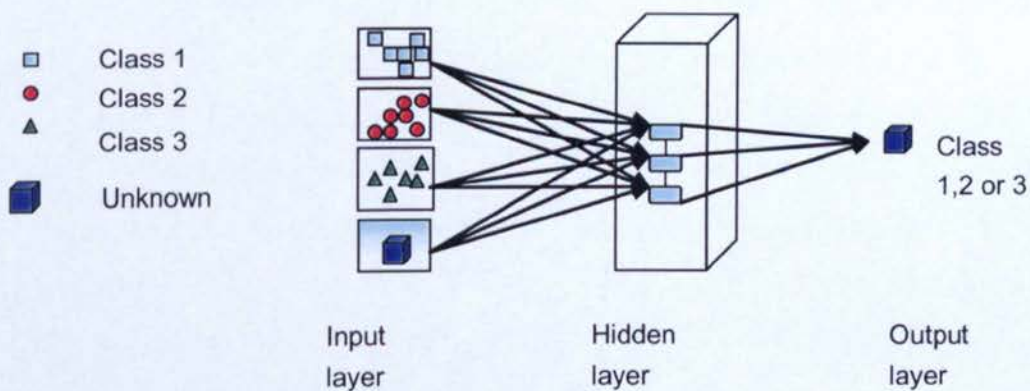


Figure 2.9. Example of an Artificial Neural Network classifier

A DT classifier takes a different approach to classification. It breaks an often very complex classification problem into multiple stages of simpler decision-making processes (Safavian and Landgrebe, 1991). The tree is composed of a root node formed from all of the data, a set of internal nodes or splits, and a set of terminal nodes or leaves (see Figure 2.10).

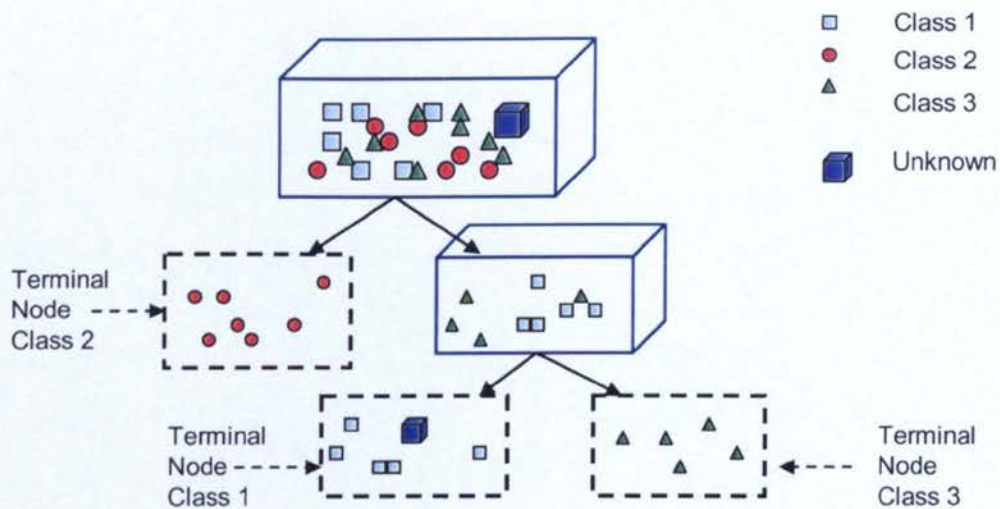


Figure 2.10. Example of a Decision Tree Classifier

Each node has only one parent node and two or more descendant nodes. A data set is classified by subdividing it according to the decision framework defined by the tree and a class label is assigned to each observation according to the terminal node into which the observation falls (Friedl and Brodley, 1997). Decision tree classification methods have significant advantages for remote sensing classification problems because of their flexibility, simplicity and computational efficiency. Also decision tree algorithms are generally fast and insensitive to noise in input data and therefore have substantial utility for classifying the large volumes of data inherent in remote-sensing land cover mapping (Friedl and Brodley, 1997).

Finally, SVM classifiers are relatively new to the remote sensing community compared with other popular classifiers such as ML, ANN and DT (Huang *et al.*, 2002). A SVM employs optimisation algorithms to locate the optimal hyperplane that separates two classes (Figure 2.11). Statistically the optimal boundaries should be generalised to unseen samples with least errors among all possible boundaries separating the classes, therefore minimising the confusion between classes.

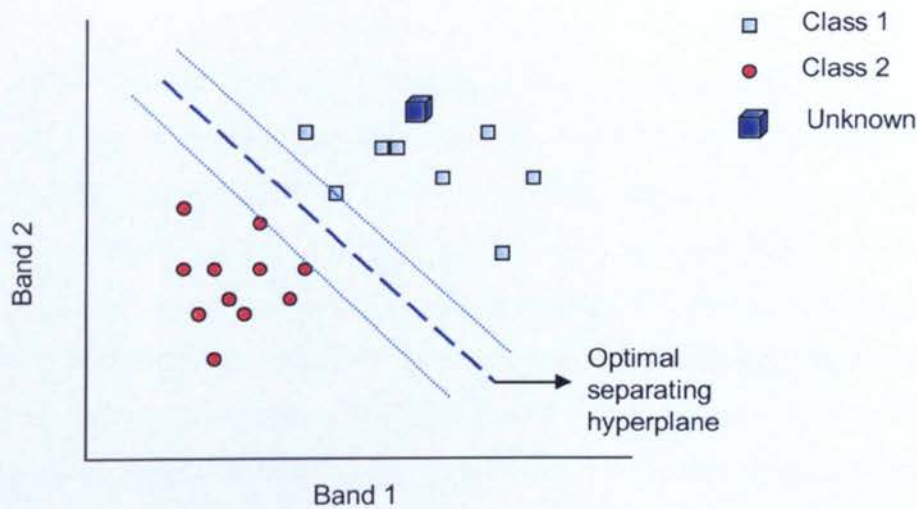


Figure 2.11. Example of optimal hyperplane or boundary between two classes using a SVM. Based upon Burges (1998).

The potential of SVM for the classification of remotely sensed data has only been recognised recently (Zhu and Blumberg, 2002, Huang *et al.*, 2002, Gualtieri and Crompt, 1998, Chapelle *et al.*, 1999). Comparative studies have demonstrated that a SVM may be used to classify land cover more accurately than the conventional ML classifier and popular alternatives such as DTs and ANNs (Huang *et al.*, 2002, Foody and Mathur, 2004a).

2.1.4 Classification output: Labelling

After the chosen classifiers have been trained, the third stage of the classification process is called labelling. Labelling is the process of allocating individual pixels to their most likely class. This process of labelling can be approached in one of two ways.

(i) When the number of separable pixels that exist in the area covered by the image is known, and the estimation of the statistical properties of the values of each of these pixels is possible (in statistical classifiers), then individual pixels (test pixels) can be labelled as belonging to the classes based on these statistical properties.

(ii) When the number and character of the land cover classes present in the images is unknown a method of allocating and reallocating each pixel to one of an initial set of randomly-chosen pixels is used. At each stage, each pixel in turn is given the label of one of these randomly chosen pixels using some classifier. At the end of first iteration, when every pixel has been labelled, the randomly chosen pixels can be altered in character (either by combining, splitting, and removing some of the pixels) according to the nature of the pixels which have been associated with them. This process of pixel labelling is repeated until the process converges. At this stage the user can relate these pixels to some land cover class (Schowengerdt, 1997).

2.1.5 Accuracy Assessment

Finally, no classification process is complete without the assessment of its accuracy. Many methods of accuracy assessment have been discussed in the remote sensing literature (e.g., Aronoff, 1985, Rosenfield and Fitzpatrick- Lins, 1986, Kalkhan *et al.*, 1995, Koukoulas and Blackburn, 2001). In general, there are two components of accuracy within the context of remote sensing. These are positional accuracy and thematic accuracy (Janssen and van der Wel, 1994). Positional accuracy determines how closely the positions of objects shown on a rectified image (map) agree with the true position on the ground. Positional accuracy is achieved by registering an image to a map with the use of ground control points as reference. Once enough control points have been identified the image is resampled and a root-mean-square (RMS) error is calculated. There are different recent studies that look in detail into this type of accuracy (Fonseca and Manjunah, 1996, Vieira *et al.*, 2004).

On the other hand, thematic accuracy or classification accuracy refers to the agreement of assigned class labels of a set of samples against those classes observed on the ground. The most widely used thematic accuracy is derived from a confusion or error matrix. An error matrix is a cross-tabulation of the assigned class label of a

set of samples against that observed in the ground or reference data at specified locations (Foody, 2002). It provides the basis to describe overall classification accuracy and characterise errors such as errors of omission and commission. Overall accuracy is obtained by dividing the total number of correctly classified pixels by the total number of reference pixels. It also measures misclassification errors such as an omission from the correct class but also a commission into another class (Story and Congalton, 1986).

A key issue about the measure of accuracy described above is that the basic assumptions underlying the assessment of classification accuracy may not be satisfied. For example the data might present problems due to the presence of mixed pixels making the generalisation of class definitions problematic. Also, it could be the case that a misregistration of the ground and remotely sensed data sets is present. This could be due to the fact that the ground data available could not be an accurate representation of the ground conditions at the time that the remotely sensed image was taken. Furthermore, the information on the sampling design used in their acquisition is normally not provided. Obtaining a reliable confusion matrix is, therefore, not always possible (Smits *et al.*, 1999), but it remains central to most accuracy assessment and reporting.

Finally, another problem with error matrices for some users is that particular cases may have been allocated to the correct class purely by chance (Turk, 1979, Rosenfield and Fitzpatrick-Lins, 1986, Congalton, 1991, Pontius, 2000). To accommodate for the effects of chance agreement, Cohen's kappa coefficient has often been used (Smits *et al.*, 1999). The kappa coefficient makes some compensation for chance agreement and a variance term may be calculated for it enabling the statistical testing of the significance of the difference between two coefficients (Rosenfield and Fitzpatrick-Lins, 1986). This is often important, as frequently, there is a desire to compare different classifications and matrices. A more detailed account of the different methods to assess classification accuracy will be given in Chapter 3.

2.2 Classifying a class of interest

As mentioned earlier in this chapter, standard classification approaches assume discrete, mutually exclusive and exhaustively defined classes. If a class is present in the region to be mapped but absent from the training stage of the classification (i.e. an untrained class) cases of that class will be commissioned by the trained classes in the analysis leading to error. This, however, may not actually be measured and reflected in the accuracy statement. Therefore, it is very important that all classes be included in a classification training stage to ensure the assumption of an exhaustively defined set is satisfied.

This has obvious implications for training the classifier such as the size of the training set required for a classification linked to the number of classes as well as the properties of the data. Furthermore, some of the classes may be of little or no interest to the study. As seen previously, the EU Habitats Directive interest commonly focuses on one or on a sub-set of the classes or land covers under consideration. Also, in standard classifications, the aim is often to maximize the overall probability that a pixel is allocated correctly. This may not be appropriate for a specific study as the focus is on overall accuracy rather than on the sub-set of classes or a class of actual interest (Lark, 1995c). Therefore, a more logical approach would be to focus only on data gathering for the class of interest and in addition minimise and optimise the amount of this data needed to train a classifier.

This approach could have a great potential when considered from the point of view of the European Habitats Directive and the protection and monitoring of specific habitats. The possibility of discriminating one priority habitat within an image could potentially reduce monitoring costs and result in a more time efficient approach to image classification.

This thesis will explore two main alternatives to the standard multiclass classification methods in order to concentrate on the classification of a class of interest. The first one is the use of a binary classification where the class of interest is classified against all the other classes present in the image. The second one is one-class classification methods where the classifier concentrates completely on defining the class of interest and classifies it from any other possible class.

2.2.1 Binary classifiers for the classification of one habitat of interest

Rifkin and Klautau (2004) have recently reviewed the concept of standard multiclass classification. The authors opted for going back to the basic problem of binary classification. In their work they trained N different binary classifiers, each one trained to distinguish the examples of a single class from the examples of all the remaining classes. This process was repeated for all the classes. To classify a new example the N classifiers were run and the classifier with the most positive outputs was chosen. This technique is known as “one-versus-all” (OVA) classification. Their results confirmed that the OVA approach is as valid as other approaches that aim to achieve a higher multiclass classification accuracy and they argue it should be preferred due to its computational and conceptual simplicity. The idea of an OVA approach seemed quite appropriate for this thesis. In this case the OVA approach would be adopted to perform a binary classification in order to separate the class of interest from all the other classes. This directs resources towards gathering enough data to define the class of interest and less so in order to define all the other individual classes.

With the purpose of assessing which methods would be useful for this binary classification approach, Table 2.2 below summarises the advantages and disadvantages of the classifiers described in the previous section:

Classifiers	Advantages	Disadvantages
Parametric		
Multiclass classifiers		
Minimum Distance	Simplicity. It only takes into account the class mean	It assumes that classes are symmetric in multispectral space
Parallelepiped	Sensitivity to class variance	When overlapping of different classes occurs decision regions fit the training data poorly
Maximum Likelihood	Takes into account both variance and covariance of the spectral classes and uses probability density functions	It still label pixels as unknown if they don't fit into any category
Non Parametric		
Multiclass classifiers		
K-nearest neighbour	Widely used in pattern recognition. Effective when estimating densities and for classification	Large amount of computation required
Decision Trees	Breaks complicated classification problems into simpler stages. Flexibility. Computational efficiency.	Problems of overfitting. When an attribute has a large number of possible values it may potentially be a problem. Continuous valued attributes may also be a problem as they may contain a large or infinite set of values.
Neural Networks	Successfully applied in remote sensing. No deterministic. Distribution free nature	Complicated network architectures. Variable output depending on weight initialisation, slow processing speeds.
Support Vector Machines	Finds optimal separating decision boundaries between classes. No problems with small sets of data. Few parameters to choose.	Optimisation of the algorithm can be time consuming.

Table 2.2. Advantages and disadvantages of parametric and non parametric Multiclass classifiers.

SVM is the only classifier that was originally designed as a binary classifier and therefore it would be expected to perform better in a binary classification problem. It is also a very interesting algorithm as it has not been explored in depth by the remote sensing community and even less in its quality as a binary classifier. It has obvious advantages over the other classifiers in terms of few parameters to choose and having no problems with small datasets (Table 2.2). Bennett and Campbell (2000) reviewed different applications of SVM to date and came to the conclusion that SVMs eliminate problems experienced with other methods such as neural networks. For example, SVMs have few parameters to pick and the final results are stable, reproducible and largely independent of the specific algorithm used to optimize the SVM. However, as mentioned earlier, the potential of SVM for the classification of remotely sensed data has only been recognised recently and always in the context of multiclass classification. Chapter 4 assesses the application of such classifier in the context of classification of a particular priority habitat.

Of the other non-parametric classifiers, the k-NN and ANN classifiers normally require a large amount of calculations and are computationally demanding. In particular, ANNs usually require complicated network structures which are often very subjective (Foody, 2002) and that could collide with the simplicity of an OVA classification design. Furthermore, both k-NN and ANN classifiers need large amounts of data in order to train the classifier properly. It has been widely documented that in order to train a ANN a sufficiently large sample needs to be obtained, which also has to be unbiased towards the population it is to represent. However, it may be the case that there is insufficient data of a satisfactory quality and thus further data will be required and cost issues then become a further worry. Various studies have incorporated data that have not been representative and have consequently had difficulties when using the network (Spellman, 1999, Foody, 2002). After considering the problems that the k-NN and ANN classifiers presented with regards to training data, it was decided not to use these two classifiers for the purpose of this thesis.

On the other hand, DTs were an attractive choice. They have not been used by the remote sensing community as extensively as statistical or neural/connectionist methods (Pal and Mather, 2003). DTs offer advantages over the above mentioned methods such as handling data at different scales, flexibility and can be trained quickly and they are rapid in execution (Friedl and Brodley, 1997). The construction of DTs involves splitting of the dataset into increasingly homogeneous subsets. At each level of the tree, a test is applied to one or more attribute values that may have one of two outcomes. This structure could very appropriate for a binary approach where at each split the decision tree has to decide if a pixel belongs to the class of interest or the “other” class. In terms of amount of training data Pal and Mather (2003) concluded that (i) the accuracy of DTs got higher as the size of the training data set increased but just up to a certain point and (ii) DTs did not require a large amount of training data to be effective.

Taking into account all the above, SVM and DT classifiers were selected as the most suitable methods for the classification of a specific habitat of interest using a binary approach. For comparison purposes the results from these classifications were assessed against those obtained by a standard ML parametric classifier in the corresponding research chapter.

2.2.2 One-class classifiers for the classification of one habitat of interest

Another approach to the issue of concentrating on a habitat of interest during the classification process is one-class classification methods. The problem with the binary OVA scheme is that it is still necessary to have some sort of data from the other classes present in the image to perform the classification and therefore the training stage still involves the collection of data of classes that are of no interest for the research. One-class classification methods could resolve this issue as only data

from the class of interest are necessary to train the classifiers. However, they have not been applied yet to remote sensing classification (as far as the author is aware).

In practice, one-class classification is a special type of binary classification problem where each of the two classes has a special meaning. The two classes are called the target and the outlier class respectively. The target class is the equivalent to the class of interest. The outlier class is the equivalent to “all the other classes” in the OVA classification scheme. The important difference between a typical binary classification and a one-class classification is that in a binary classification it is the analyst who knows which one is the target class and the classifier treats both target and non-target as two classes. In one-class classification the classifier is trained to recognise the target class and any further classification is based on the description of the target class. The outlier class can be sampled very sparsely or can be totally absent (Tax, 2004).

There are several one-class classification methods. For example, the reconstruction methods are designed to model the data. These methods use prior knowledge about the data and make assumptions about the generating process. Then a model is chosen and fitted to the data and new data can be described in terms of a state of this model. In this sense they are very similar to parametric approaches. With the application of the reconstruction methods, it is assumed that outlier objects do not satisfy the assumptions about the target class distribution. However, as the outliers are usually badly represented their reconstruction error is normally high. Therefore, this method is not optimised for outlier detection. It requires a classification task to be solved and this can be computationally expensive (Tax, 2001).

Another one-class classifier method is to estimate the density of the training data (Tarassenko *et al.*, 1995) and to set a threshold on this density. They achieve this by assuming a uniform outlier distribution. Only the prior probabilities of the target and outlier class should be chosen beforehand. This directly influences the choice where the probability should be thresholded to obtain a target and an outlier region. Several

distributions can be assumed, such as a gaussian or a poisson distribution, and numerous tests, called discordancy tests, are then available to test new objects (Barnett and Lewis, 1994). When the sample size is sufficiently high and a flexible density model is used (for example a parzen density estimation), this approach works very well. Unfortunately, it requires a large number of training samples and it assumes that the training data are a typical sample from the true data distribution. This makes the application of the density methods problematic (Tax, 2001).

According to Vapnik (1998) the disadvantage of the estimation of the complete density of the data is that it might require too much data and could result in bad descriptions. Therefore, boundary methods are proposed where only a boundary around the target set is optimised. In most cases distances or weighted distances to a set of objects in the training set are computed. So, although the required sample size for the boundary methods is smaller than the density methods, a part of the burden is now put onto well-defined distances. Also, due to their focus on the boundary, the threshold on the output is always obtained in a direct way. The output of these boundary methods cannot be interpreted as a probability (Tax, 2001).

The advantages and disadvantages of these three methods are summarised in Table 2.3 below:

One class classifiers	Advantages	Disadvantages
Reconstruction methods	Simplest solution for outlier detection	Performs poorly in high dimensional scales. Computationally demanding.
Density methods	Use probability densities to describe the target class	The estimation of the complete density of the target class can result in poor descriptions
Boundary methods	Avoids estimation of the total density of the target class and focus on the boundary around it	It could inherit the disadvantage of neural networks. When based on Support Vector classifiers the data descriptions are more flexible

Table 2.3. One-class classification methods

When comparing the advantages and disadvantages of the three methods described above, boundary methods seemed to be appropriate for this research as they do not make any assumptions regarding the data distribution and they do not require too much training data. When they are based upon the SVM theory they are more flexible than the other two approaches as the emphasis is put upon the boundary that separates the class of interest. For example Tax (2001) uses a Support Vector Data Description (SVDD) classifier that is inspired by Vapnik's SVM which aims to calculate an optimum hyperplane boundary that closes around the target class. All these classifiers will be analysed in detail in the respective research chapters.

2.3 Summary

This chapter has reviewed the process of classification in remote sensing where the principles of remote sensing and data acquisition are briefly explained. In particular, the advantages of satellite remote sensing over other ways of data gathering are that large global surface areas can be mapped and monitored at different spatial and temporal scales. In many instances, it is also less costly than aerial and ground surveys. This is a very important issue for local and regional authorities in charge of mapping and monitoring because the acquisition, verification and update of data take great part of the mapping budget. Therefore, the use of satellite remote sensing data could be a suitable choice for habitat mapping and monitoring.

So far, the typical approach to remote sensing classification has been a multiclass classification that assumes discrete, mutually exclusive and exhaustively defined classes. It is very important that all classes be included in a classification training stage. However, this has obvious implications for training the classifier such as the size of the training required for a classification linked to the number of classes as well as the properties of the data. Furthermore, some of the classes may be of little or no interest to the study as with the requirements of the EU Habitats Directives which focuses on particular habitats of interest. To explore the different possibilities to deal with this problem, section 2.2 described the different stages of standard image

classification process including feature extraction, training, labelling and accuracy assessment. In particular, it focused upon the training stage and the parametric and non-parametric classification methods used by the remote sensing community. Section 2.3 concentrated on how these methods could be applied to a classification scheme such as OVA (one-versus-all). After assessing the advantages and disadvantages of different classifiers, it was decided that SVMs and DTs were the most suitable methods for mapping a particular class of interest using an OVA approach. SVM is a classifier that has not been explored in depth yet by the remote sensing community. However, it has obvious advantages over the other classifiers in terms of few parameters to choose and having no problems with small datasets. This is a very interesting characteristic within the context of this thesis in terms of optimising the amount of training data required. DTs are also chosen as another classifier to perform an OVA classification. Also binary in nature DTs offer advantages such as handling data at different scales, flexibility and can be trained fairly quickly.

Finally, it was decided to go one step forward and explore one-class classifiers in order to totally concentrate on the class of interest. One-class classifiers are normally based upon data from a specific class and they are trained on this data in order to separate this class from all the possible other classes (outliers). One-class classifiers have been extensively used in pattern recognition but they are a totally novel approach within remote sensing image classification. Of the three main methods used in one-class classification, boundary classifiers present more advantages in terms of needing small training datasets and having good generalisation. In particular, the SVDD designed by Tax (2001) is based upon the principles of SVM and shares all the benefits of the SVMs. Therefore, it would be sensible to compare the results obtained by the binary SVM and the one-class SVDD.

Having reviewed different classification methods and identified the classifiers that are more appropriate for the aim of this thesis, the following chapter will describe the area of study, data and methods used to investigate these classifiers.

3 Mapping one specific habitat of interest: Case study and methods

*"The true method of knowledge is experiment."
William Blake*

The aim of the present chapter is to describe the area of study, data and methods chosen for the purpose of assessing the performance of the classifiers selected in Chapter 2. Choosing the appropriate datasets and processing methods is one of the most important factors that contributes to the successful application of remote sensing (Phinn *et al.*, 2000). In order to do this, it is important to take into consideration a few principles that determine the characteristics of the data needed for the particular research project. According to different authors (Cihlar, 2000, Phinn *et al.*, 2000) these important considerations include:

- 1) Purpose of the research.
- 2) Suitable analytic methods.
- 3) Thematic content. Environment type, land cover classes to classify.
- 4) Selection of suitable remote sensing data.
- 5) Appropriate training and testing data sets.

The purpose and objectives of this research have already been established in Chapter 1. Also, different classification methods were assessed in Chapter 2 in order to choose the most suitable ones in order to achieve these aims and objectives.

Therefore this chapter will be concentrating on the other three data considerations: (i) the thematic content of the case study (area and habitats chosen), (ii) the remote sensing data appropriate for this case study (and ground data availability) and (iii) the procedure followed to acquire the training and testing datasets (see Figure 3.1). Finally the measure the accuracy of the classification outputs will also be considered. The classification stage and comparison of different classification methods will be looked in detail in the corresponding research chapters.

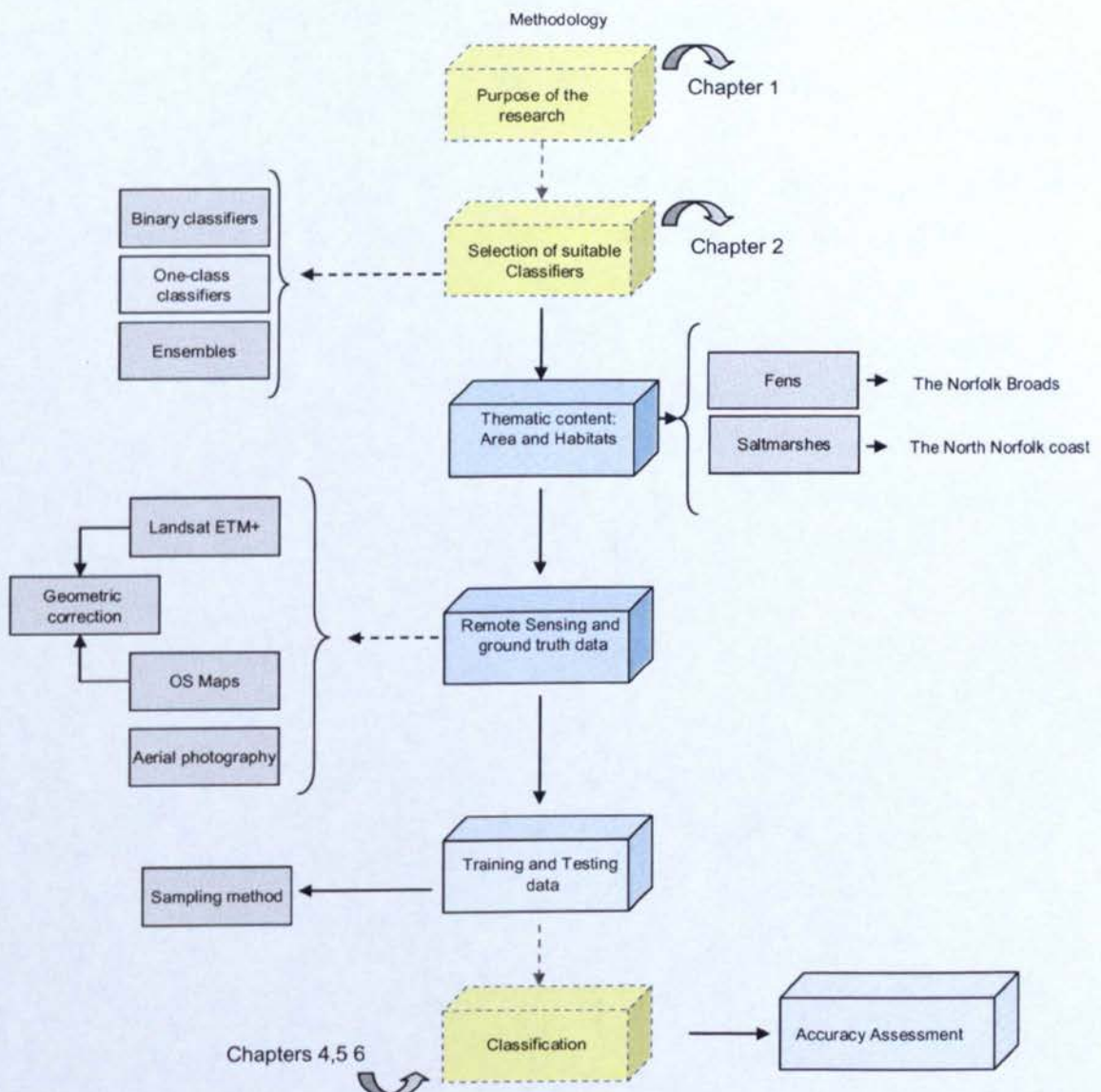


Figure 3.1 Case study methodology. The sections with dotted line are dealt with in different chapters.

3.1 Thematic content: Habitat of interest and areas

As mentioned in Chapter 1, it was decided to choose a main habitat of interest to assess the classifiers but also to perform the different classifications on a second habitat protected by the EU Habitats Directive. The reason for this was to assess whether the classification results were biased by the type of habitat and area chosen. After conversations with staff from English Nature and the Environment Agency it was decided that the two habitats chosen to perform the analysis should be fen as the first habitat of interest and saltmarsh as the second habitat used for comparison purposes. The majority of these habitats in the UK are notified as Sites of Special Scientific Interest (Areas of Special Scientific Interest in Northern Ireland (SSSIs/ASSIs), Wetlands of International Importance under the Ramsar Convention, SPAs under the EC Birds Directive and SACs under the EC Habitats Directive. It is clear that the importance of the monitoring of these habitats is crucial for many areas and vital for authorities across the United Kingdom. Also both habitats are extremely dynamic and difficult to map accurately. This is definitely a challenge for the remote sensing community. This research will try to demonstrate that using remote sensing methods can aid the process of mapping these land cover classes and their monitoring.

3.1.1 Fens

The UK is thought to host a large proportion of the fens surviving in the EU. As in other parts of Europe fen vegetation has declined dramatically in the past century (The Broads Authority, 2004). According to the Joint Nature Conservation Committee (JNCC), the UK government's wildlife advisor, fens are "peatlands which receive water and nutrients from the soil, rock and ground water as well as from rainfall: they are minerotrophic". In terms of vegetation fens can also be described as 'poor-fens' or 'rich-fens'. Poor-fens, where the water is derived from base-poor rock such as sandstones and granites occur mainly in the uplands, or are associated with lowland heaths. They are characterised by short vegetation with a

high proportion of bog mosses *Sphagnum* spp. and acid water (pH of 5 or less). Rich-fens, are fed by mineral-enriched calcareous waters (pH 5 or more) and are mainly confined to the lowlands and where there are localised occurrences of base-rich rocks such as limestone in the uplands and they are characterised by tall-herb communities.

Fen habitats support a great diversity of plant and animal communities. Some can contain up to 550 species of higher plants, a third of the UK's native plant species; up to and occasionally more than half the UK's species of dragonflies, several thousand other insect species, as well as being an important habitat for a range of aquatic beetles. They are dynamic semi-natural systems and in general, careful management is needed to maintain open-fen communities and their associated species richness. Without appropriate management (e.g. mowing, grazing, burning, peat cutting, scrub clearance), natural succession will lead to scrub and woodland forming. According to the UK Biodiversity Action Plan (UKBAP) current factors affecting this habitat type are:

- Past loss of area by drainage and conversion to intensive agriculture.
- Small total area of habitat and critically small population sizes of several key species dependent on the habitat.
- Lack of or inappropriate management of existing fens leading to drying, scrub encroachment and succession to woodland.
- Enrichment or hypertrophication resulting in changing plant communities.
- They are particularly exposed to the impacts of climate change such as a rise in temperature, sea levels and changes in precipitation.

Regardless of all the above, there are not many studies that focus on fens and their mapping and monitoring. Penny Anderson Associates (PAA) were appointed by English Nature (EN) in the year 2000 to review and evaluate what research is currently being done for this and other wetlands within the United Kingdom (from 1995 onwards). The evaluation was based upon database searches, on bibliography and responses from representatives of individuals and organisations outside EN. The bibliography from research publications was concerned only with the hydrology of

the habitat and the studies were purely descriptive. Applied research in the form of projects concerned with fen habitats included one project into fen hydrology, one on habitat creation/restoration and one concerned with the management of fen vegetation. This information came mainly from the Broads Authority that had produced a Fen Management Strategy. Maybe the response from individuals and organisations was low but more worryingly is the thought that very little research has been carried out on this habitat (PAA, 2001).

In terms of management and monitoring a Common Standards and Monitoring guide was published by the JNCC in August 2004. When recommending methods of monitoring for habitat extent and habitat composition in the Common Standards and Monitoring guide (JNCC, 2004), the JNCC mentions the use of aerial photography for this purpose but never mentions the aid of satellite remote sensing. It is therefore very important to demonstrate that satellite remote sensing data can definitely help in the monitoring of these important habitats, especially when it comes to address habitat loss.

Further recommendations from the JNCC in the Common Standards and Monitoring guide (JNCC, 2004) include monitoring times for these habitats. Generally the best time to carry out monitoring in wetland systems is between early June and the end of September, when sedges are flowering or fruiting and their identification is easiest. However, other times of year may be more appropriate for some investigations. They recommend that at least one visit should be made to each site within a single six-year reporting cycle. These considerations reinforce the arguments in favour of the use of satellite remote sensing data as a suitable option since in the summer there is a higher probability of cloud free images and the monitoring of these areas can be done every year with satellites such as Landsat or SPOT at a low cost.

When it comes to selecting an area of study, the Norfolk Broads seems to be an appropriate option. The Norfolk Broads are located in eastern England (see Figure 3.3) and it is Britain's largest nationally protected wetland. Also it is one of two sites selected as SACs under the EU Habitats Directive for alkaline fens in East Anglia,

where a main concentration of lowland fens occurs, and a great part of it is designated as SPA, also under the EU Directives. There are areas of short sedge fens (both *Schoenus nigricans* – *Juncus subnodulosus* mire and *Carex rostrata* – *Calliergon cuspidatum/giganteum* mire), which in places form a mosaic with *Phragmites australis* – *Peucedanum palustris* fens. There are complex zonations present and many differences exist between the individual fens that comprise the site. The fens are principally of the flood plain mire type. This site contains a range of rare and local plant species, including the Annex II from the Habitats Directive Fen orchid *Liparis loeselii*, lesser tussock-sedge *Carex diandra* and slender sedge *C. lasiocarpa* (<http://www.jncc.gov.uk>).

The Broads Authority has developed habitat-based restoration and conservation strategies. These working strategies are based on a thorough evaluation of the natural resources. Although they actively use aerial photography for habitat monitoring, other satellite or airborne remote sensing methods are not being taken into account. Furthermore, according to the Broads Action Plan 2004, the management of the Broads will consist of a cyclical process with six stages, of which the first one is a report of the state of the site. However, currently there is no such information and therefore no benchmarks against which to assess its condition and monitor change (The Broads Authority, 2004). All these fully validate the choice of fen as the primary class of interest for this research.

3.1.2 Saltmarshes

As mentioned earlier, saltmarsh was the habitat chosen in order to test any bias of the classifiers towards the characteristics of the class fen. Saltmarshes have been the focus of extensive research within the remote sensing community using multispectral images (Dale *et al.*, 1986, Donoghue and Shennan, 1987, Phinn *et al.*, 1999) and hyperspectral airborne data (Bajjouk *et al.*, 1996, Eastwood *et al.*, 1997, Smith *et al.*, 1998, Silvestri *et al.*, 2002). Coastal saltmarshes are usually restricted to sheltered locations in five main situations: estuaries, saline lagoons, behind barrier islands, at the heads of sea lochs, and on beach plains. Saltmarsh vegetation consists of a

limited number of halophytic (salt tolerant) species adapted to regular immersion by the tides. A natural saltmarsh system shows a clear zonation according to the frequency of inundation. At the lowest level the pioneer glassworts *Salicornia* spp can withstand immersion by as many as 600 tides per year, while transitional species of the upper marsh can only withstand occasional inundation (<http://www.jncc.gov.uk>).

The communities of stabilised saltmarshes can be divided into species-poor low-mid marsh, and the more diverse communities of the mid-upper marsh. On traditionally grazed sites, saltmarshes vegetation is shorter and dominated by grasses. At the upper tidal limits, true saltmarshes communities are replaced by driftline, swamp or transitional communities which can only withstand occasional inundation. Saltmarshes communities are additionally affected by differences in climate, the particle size of the sediment and, within estuaries, by decreasing salinity in the upper reaches (<http://www.jncc.gov.uk>).

Furthermore, saltmarshes are an important resource for wading birds and wildfowl. They act as high tide refuges for birds feeding on adjacent mudflats, as breeding sites for waders, gulls and terns and as a source of food for passerine birds particularly in autumn and winter. In winter, grazed saltmarshes are used as feeding grounds by large flocks of wild ducks and geese. Areas with high structural and plant diversity, particularly where freshwater seepages provide a transition from fresh to brackish conditions, are particularly important for invertebrates. Saltmarshes also provide sheltered nursery sites for several species of fish (<http://www.jncc.gov.uk>).

Current factors affecting the habitat according to the UK Biodiversity Action Plan (UKBAP) are:

- Land claim. This practice continued until very recently. As a consequence, many saltmarshes now adjoin arable land, and the upper and transitional zones of saltmarshes have become quite scarce in England.

- Erosion and coastal squeeze. Erosion of the seaward edge of saltmarshes occurs widely and it is exacerbated by climate change. Furthermore, many saltmarshes are being 'squeezed' between an eroding seaward edge and fixed flood defence walls.
- Sediment dynamics. Local sediment budgets may be affected by coast protection works, or by changes in estuary morphology caused by land claim, dredging of shipping channels and the impacts of flood defence works over the years.
- Grazing. Intensive grazing is considered to be a problem in some areas by reducing the height of the vegetation and the diversity of plant and invertebrate species.
- Other human influences. Waste tipping, pollution, military activity, oil pollution, recreational pressure, and eutrophication due to sewage effluent and agricultural fertiliser run-off. (<http://www.ukbap.org.uk>)

In terms of monitoring this habitat, the suggested visiting period recommended by the JNCC is May to October. However, in areas of coastal squeeze, where low-marsh communities dominate, annuals are relatively more abundant and the assessment will need to take this into account (April to August is suggested). In addition to the basic six-yearly monitoring cycles, a more frequent monitoring is recommended.

The area of the North Norfolk coast has been selected as a SAC for the extensive Atlantic salt meadows that form part of a sequence of vegetation types that are unparalleled among coastal sites in the UK. Furthermore, for their diversity they are amongst the most important in Europe. Saltmarsh swards dominated by sea-lavenders *Limonium* spp. are particularly well-represented on this site. In addition to typical lower and middle saltmarshes communities, in North Norfolk there are transitions from upper marsh to freshwater reedswamp, sand dunes, shingle beaches and mud/sandflats. The area is also designated as a SPA due to the large numbers of waterbirds occurring throughout the year. In summer, the site holds large breeding populations of waders, four species of terns, Bittern *Botaurus stellaris* and wetland raptors such as Marsh Harrier *Circus aeruginosus*. In winter, the coast is used by

very large numbers of geese, sea-ducks, other ducks and waders. The coast is also of major importance for staging waterbirds in the spring and autumn migration periods (<http://www.jncc.gov.uk>).

3.2 Remote sensing data and ground data sources

Over recent decades, remote sensing data have become one of the primary sources for obtaining information about the Earth's land cover. At a global scale, this has been achieved primarily from data acquired by the Advanced Very High Resolution Radiometer (AVHRR) onboard of NOAA meteorological satellites (DeFries *et al.*, 1998, Hansen *et al.*, 2000). At regional and local scales, Landsat and SPOT satellites have been extensively used to extract information about particular locations (DeFries and Chan, 2000).

The Landsat satellite programme in particular represents the world's longest continuously acquired collection of space-based land remote sensing data. For over 30 years, the Landsat satellite series has collected and produced low-cost, moderate-resolution multispectral data for researchers and decision-makers worldwide. This provides an invaluable source of information for tracking changes in the environment over the last 30 years (Lillesand and Kiefer, 2004).

The ETM+ instrument on the Landsat 7 spacecraft contains sensors to record Earth scene radiation in three specific bands:

- visible and near infrared (VNIR) bands - bands 1,2,3,4,and 8 (PAN) with a spectral range between 0.4 and 1.0 micrometres.
- short wavelength infrared (SWIR) bands - bands 5 and 7 with a spectral range between 1.0 and 3.0 micrometres.
- thermal long wavelength infrared (LWIR) band - band 6 with a spectral range between 8.0 and 12.0 micrometres.

Landsat satellite images are acquired every 16 days with a spatial resolution of 30 m.

There are numerous applications of Landsat satellite images in which remote sensing data have been translated into useful ecological information. This information has been used to describe the status of habitats by land cover mapping and the dynamics of these habitats by change detection analysis (Cohen and Goward, 2004). Some of them are highlighted in Table 3.1.

Agriculture, Forestry and Range Resources	Hydrology	Coastal Resources	Environmental Monitoring
Discriminating vegetative, crop and timber types	Determining water boundaries and surface water areas	Determining patterns and extent of turbidity	Monitoring deforestation
Measuring crop and timber acreage	Mapping floods and flood plain characteristics	Mapping shoreline changes	Monitoring volcanic flow activity
Precision farming land management	Determining area extent of snow and ice coverage	Mapping shoals, reefs and shallow areas	Mapping and monitoring endangered habitats
Monitoring crop and forest harvests	Measuring changes and extent of glacial features	Mapping and monitoring sea ice in shipping lanes	Determining effects of natural disasters
Determining range readiness, biomass and health	Measuring turbidity and sediment patterns	Tracking beach erosion and flooding	Assessing drought impact
Monitoring desert blooms	Monitoring lake inventories and health	Determining coastal circulation patterns	Assessing and monitoring grass and forest fires
Assessing wildlife habitat	Estimating snow melt runoff	Measuring sea surface temperature	Mapping and monitoring lake eutrophication

Table 3.1. Based upon information from Nasa archives. Updated April 2, 1999.

(<http://www.nasa.gov>).

There are obvious advantages and disadvantages with the use of satellite data. Some limitations of acquisition and processing of satellite information include: spatial resolution which would not be suitable to detailed studies of particular habitats; cloud cover; satellite passover time and tide which could have implications for coastal habitat mapping and monitoring. However advantages in the use of satellite data for habitat mapping greatly surpass the disadvantages. They are: cost effectiveness, timeliness, monitoring capability, large area mapped quickly, quantitative information obtainable, potential reduction in field work and

cartographic products easily produced (<http://www.nasa.gov>).

Taking into account all the above, it was decided to use a Landsat satellite image to carry out the present research. The Landsat ETM+ image was dated 19th of June 2000 and provided by the NERC Earth Observation Centre. The date when the image was taken fits in with the recommendations from the JNCC for the optimal time for monitoring these two habitats as mentioned in the previous section. The area covers the East Anglia study site (centred on 53.104 Latitude, 1.078 Longitude covering an area of 34,235.2879 km² in total). The six spectral bands (bands 1-5, and 7) with a spatial resolution of 30 m were selected for use in the analysis. The image was geometrically corrected using the Transverse Mercator Airy Projection, Ordnance Survey 1936 with an rms error of 0.12 pixels. Because the areas of interest described in the previous section both lay within the area of Norfolk and to simplify further analysis, a subset of the image was selected so that it included both areas of interest (North Norfolk coast and the Norfolk Broads) (see Figure 3.2).



Figure 3.2. Landsat ETM+ 19th of June 2000 provided by NERC and selected subset of the image used for further analysis

Aerial photography is normally used as ground data and as a reference for accuracy assessment in studies involving remote sensing. However, aerial photography is usually expensive to acquire so existing photography is often used. Whenever possible, it is advisable that this aerial photography is as close to the date of the satellite image as possible (Congalton and Green, 1999). The aerial photography used as ground data for this research was acquired close to the time of the Landsat ETM+ image. Access to aerial photography, scale 1:10,000 acquired in June 1999 for the Norfolk Broads, was provided by the Broads Authority. Access to aerial photography acquired in July 2000 for the North Norfolk area, also 1:10,000, was provided by the Environment Agency. Both sets had been geocorrected to OS 1936 and were used to select the two classes of interest and six other broadly defined land cover types featured across the areas of interest. The aerial photographs used as ground data were not available in digital form. Therefore, they had to be manually traced, scanned, digitised and co-registered to the Landsat image. The training and testing data sets were subsequently acquired using these aerial photographs as reference using a stratified random sampling as will be explained in the following section.

In order to validate and illustrate the results of the classifications, an area of the Norfolk Broads was chosen as a sample site for producing a land cover map of fens (Figure 3.3). This area belongs to the Mid River Yare National Nature Reserve (NNR). It is a key site in the Norfolk Broads and includes tall fens, botanically-rich fens meadows and areas of willow-alder carr on fenland peats.

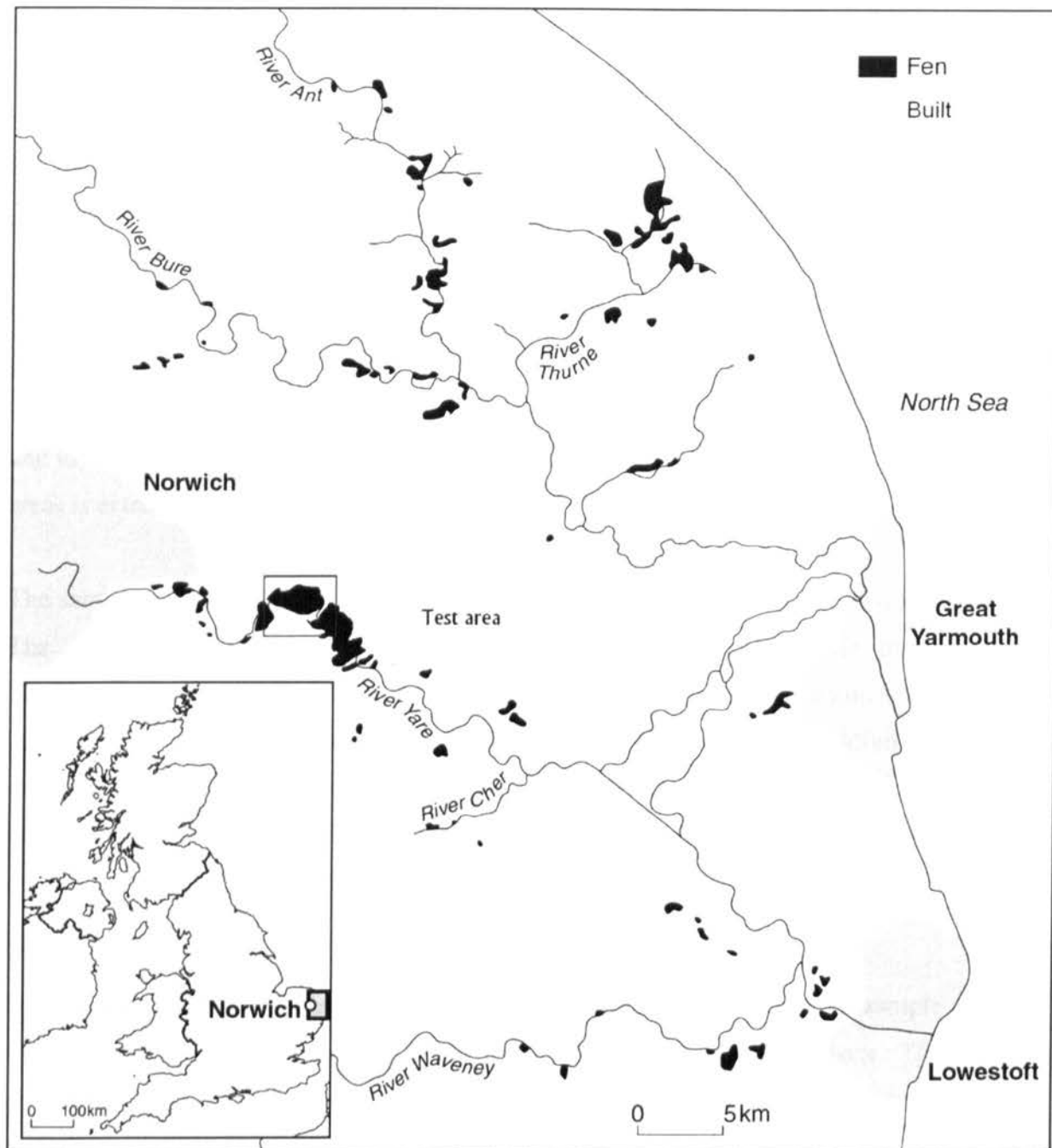


Figure 3.3 Map of the Norfolk Fens and the test area chosen within the River Yare National Nature Reserve to illustrate the results of different classification methods (Map based upon the Broads Authority information).

3.3 Data selection. Training and testing datasets

Training data selection is one of the major factors determining to what degree the classification rules can be generalised to unseen samples (Paola and Schowengerdt, 1995). The characteristics of the data used to train the classifier have a considerable influence on the accuracy of the resulting classification. In supervised classification pixels of known identity are used to classify pixels of unknown identity. These known pixels are located within the training areas selected to characterise the classes and to train the classifier (Campbell, 2002). Therefore the selection of these training areas is extremely important in order to train the classifier properly.

The sample design is a critical part of the image classification accuracy assessment. The standard form for reporting overall error for each class of the image classification is by the use of an error matrix. Error matrices represent the number of correctly mapped pixels by comparing ground data with corresponding results of computer-assisted classification. Designing a poor sampling scheme can easily result in significant biases being introduced into the error matrix, which will then affect the classification accuracy (e.g. Richards, 1993, Congalton, 1991, Lunetta *et al.*, 1991).

One sampling technique is systematic sampling. This is a method where the sample units (here pixels) are selected at some equal interval over time or space. The advantage of systematic sampling is the uniform spread of the sampled observations over the entire population. The major disadvantage, on the other hand, is that the selection procedure implies that each unit in the population does not have an equal chance of being included in the sample (Berry and Baker, 1968). Systematic sampling can either be random systematic or stratified systematic. A random systematic sampling design is when the population elements are arranged in some order. The first sample is randomly located and thereafter each unit is selected systematically from around that single sampled unit using a fixed interval (Clark and Hosking, 1986). Stratified systematic sampling combines the advantages of

randomization and stratification with the useful aspects of systematic sampling, while avoiding the possibilities of bias due to the presence of periodicity (Berry and Baker, 1968).

Another sampling technique is simple random sampling. This method selects sample units out of the population, such that each element in the population has an equal chance of being selected (Cochran, 1977). However, simple random sampling tends to under-sample small but potentially important areas (Congalton, 1988, 1991). That problem can be overcome by using stratified random sampling, which is a sampling method in which the elements of the population are allocated into sub-populations (e.g. strata) before the sample is taken, and then each stratum is randomly sampled. This sampling approach can be used when specific information about certain sub-populations and increasing precision of the estimates for the entire population is desired (Cochran, 1977, Clark and Hosking, 1986). After performing sampling simulations on three spatially diverse areas, Congalton (1988) came to the conclusion that simple random and stratified random sampling provided satisfactory results in all cases. But as mentioned above, simple random sampling tends to undersample small, but possibly important, areas.

The habitats of interest selected for this research are located in very specific areas within the image. Therefore stratified random sampling provides the better choice and it was the option chosen to select the training and testing sets used by the different classifiers. Areas of high occurrence of the classes of interest were identified with the aid of the aerial photography used as ground data. These sub-populations were then randomly sampled. Also, there was no pre-selection of core pixels or boundary pixels as these could have biased the classification in some way.

3.3.1 Training and testing datasets

Supervised classification requires data to train and validate classification algorithms. Training data are examples presented to supervised classification algorithms that are to be recognized. Validation or testing data are an independent set of data that are used to assess the performance or accuracy of the classification algorithm. In the remote sensing context, calibration and validation are generally data from areas on the ground that are defined by pixels. Calibration and validation sites should be as homogeneous as possible (Muchoney and Straher, 2002). In this sense, the classes that form the training and testing sets typically occupy a relatively discrete area of feature space. The size and location of this area in feature space is determined by the spectral variation of the class. Therefore, the training samples should be a representation of such illustration of the class within the feature space.

Furthermore, the impact of the amount of training data used to train a classifier has been recently reviewed by different authors and the general findings show that classification accuracy tends to be positively related to training set size (Arora and Foody, 1997, Foody and Mathur, 2004b, Foody *et al.*, 1995, Huang *et al.*, 2002, Pal and Mather, 2003 and Zhuang *et al.*, 1994). The recommended size of a training set is often linked to the degree of complexity of a classifier and the dimensionality of the data to train such classifiers (Kavzoglu and Mather, 2003 and Tso and Mather, 2001). For example, some of the literature suggests the use of a minimum of $10-30p$ cases per-class for training, where p is the number of wavebands used (Piper, 1992; Mather, 2004; van Niel *et al.*, 2005). However, acquiring large datasets is normally costly in terms of time and finance (Buchheim and Lillesand, 1989 and Jackson and Landgrebe, 2001). This has had as a consequence that many classification analyses have been performed with training sets that may be smaller than that which might be expected to be required for an accurate classification (Bishop, 1995, Jackson and Landgrebe, 2002, Tadjudin and Landgrebe, 1999 and Tadjudin and Landgrebe, 2000).

The performance of classification algorithms normally degrades in situations such as: (i) data with high noise content, (ii) small sample sizes relative to number of features or variables and (iii) irrelevant or redundant information (Kumar *et al.*, 2005). To solve these problems, research has been based upon (i) investigation into defining an efficient sampling design for training sample acquisition (Atkinson, 1991, Campbell, 1981 and Webster *et al.*, 1989); (ii) use of feature selection and feature extraction methods to reduce the dimensionality of the dataset to be classified (Kuo and Landgrebe, 2002) and (iii) use of unsupervised classifications to help guide the analysis (Huang, 2002, Jackson and Landgrebe, 2002 and Tadjudin and Landgrebe, 2000). Foody and Mathur (2004b) take another approach in order to reduce the need for large training datasets by focusing on only those training samples that are helpful in fitting the decision boundary that can separate the classes accurately. This assumption is founded on the use of non-parametric classifiers such as SVMs which might not need a full and representative description of each class in order to classify it accurately (Foody and Mathur, 2004b). The potential for intelligent training is obvious for SVM-based classification as the process is based on the notion that only the training samples that lie on the class boundaries are necessary for discrimination (Brown *et al.*, 1999), leading to the definition of smaller training sets (Huang *et al.*, 2002).

The approach taken in the present research will concentrate on reducing data dimensionality. Although Landsat ETM+ imagery are not particularly high dimensional datasets, there are still features that can be redundant and it has been proven that the efficiency of learning algorithms decreases with irrelevant and redundant features (John, 1997, Kohavi and John, 1997, Kohavi and Sommerfield, 1995, Koller and Sahami, 1996, Langley, 1996, Langley and Sage, 1994, Liu and Motoda, 1998). Also it is commonly accepted that as the number of variables increases the number of training samples needed to train the classifiers also grows (Duda and Hart, 1973, Jain and Chandrasekaran, 1982). Consequently, the performance of a classifier and the need for large training datasets can be optimised by removing such noisy, irrelevant or redundant information. Finally, selecting

which subset of bands provides the greatest degree of statistical separability between any two classes may greatly reduce the occurrence of errors of commission and omission. This is mainly achieved by feature extraction and feature selection methods (Kumar *et al.*, 2005).

Feature extraction consists of linear (and non-linear transformation) of the data and projections to a lower dimensional space (e.g. principal components analysis, linear discriminant analysis, Isomap). Feature selection, however, is a special case of feature extraction and selects a subset of features that describes the data as well as the original set, getting rid of irrelevant or redundant features. The benefits of feature selection include the reduction of the amount of data needed for training a classifier, improving predictive accuracy and reducing execution time (Kumar *et al.*, 2005).

In remote sensing, feature selection involves graphical and/or statistical analysis to determine the degree of between class separability. For this thesis's feature selection analysis, a set of 1,000 pixels, 500 belonging to each class of interest (fen / saltmarsh) and 500 belonging to the rest of the land cover types amalgamated to one class, were extracted from the Landsat ETM+ image in each of the 6 non-thermal spectral wavebands and the Normalized Difference Vegetation Index (NDVI) derived from the data acquired in ETM+ wavebands 3 and 4. NDVI is a band-ratio technique which produces a raster model that estimates the degree of photosynthetically active vegetation within each pixel and that is extensively used for vegetation studies. Values in this dataset range between -1 and 1, where -1 represents no photosynthetically active vegetation and 1 represents a high degree of photosynthetically active vegetation. Two bands are needed to calculate this index: one containing reflectance values for the visible red (VR) spectrum, and the second containing reflectance values for the near infrared (NIR) portion of the spectrum. The NDVI model is the quotient of the difference and sum of these two datasets (Equation 3.1).

$$NDVI = \frac{(NIR) - (VR)}{(NIR) + (VR)}$$

Equation 3.1 Equation used to calculate NDVI values

The transformed divergence (TD) statistic was calculated for every possible pair of features. The formula for computing the transformed divergence is as follows:

$$D_{ij} = \frac{1}{2} \text{tr}((C_i - C_j)(C_i^{-1} - C_j^{-1})) + \frac{1}{2} ((C_i^{-1} - C_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T)$$

$$TD_{ij} = 2000(1 - \exp\left(\frac{-D_{ij}}{8}\right))$$

Where:

i and j = the two classes being compared

C_i = the covariance matrix of signature i

μ_i = the mean vector of signature i

tr= the trace function (matrix algebra)

T= the transposition function

D_{ij} = Divergence

Evaluation of the derived transformed divergence statistics indicated that the data acquired in Landsat ETM+ Band 2 and the NDVI offered the highest average separability. Figure 3.4 and Figure 3.5 illustrate the Landsat ETM+ Band 2 and the NDVI image respectively for the test area within Mid River Yare National Nature Reserve (Figure 3.3).



Figure 3.4 ETM+ Band 2 image for the test area

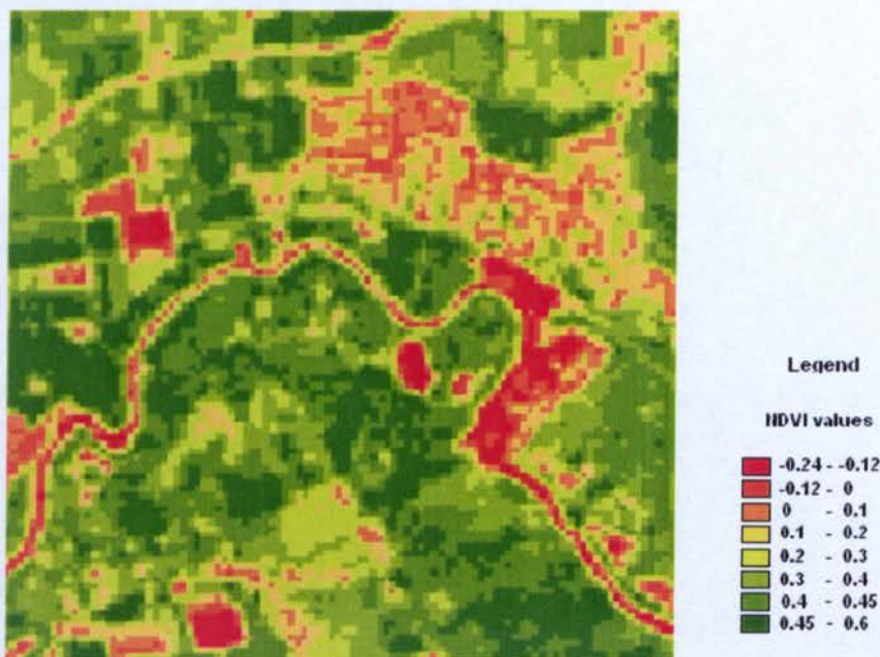


Figure 3.5 NDVI image for the test area. Values in green denote bigger NDVI values

As mentioned earlier, some of the literature suggests a heuristic approach that uses a minimum of $10-30p$ cases per-class for training, where p is the number of wavebands used. In this particular case study two input variables (NDVI and ETM+Band 2) were used. Therefore, the minimum recommended training set size in this case study

would be between 20 to 60 pixels per class. It was decided that the training sizes used for the different classifications should be performed starting from a very small size below the recommended minimum size to test whether this rule also applied to these classification methods. A detailed description of these training sets is given for each of the classifiers in the corresponding sections within the research chapters. It is important to highlight that the same testing set was used to assess the accuracy of all the classifiers. This testing set was formed by 50% pixels from the class of interest and the other 50% by a mixture of pixels from all the other classes merged into the “other” class, forming a total of 250 pixels. The pixels were totally independent from those used in the training sets to avoid any biases in the confidence level of accuracy (Fitzpatrick-Lins, 1981, Dicks and Lo 1990). The complete training and testing datasets for the different classifiers can be found in Annex F of this thesis.

3.4 Accuracy assessment

In order to compare the results obtained by the different classifiers it was necessary to measure in some way the accuracy of each classifier. Measures of accuracy are also important in order to analyze sources of error of a particular classification strategy. However, classification accuracy is not straightforward. Individual measures of map accuracy are well established in the literature (e.g., Congalton, 1991, Congalton and Green, 1999, Stehman, 1997), but considerable ambiguity remains about the implementation and interpretation of thematic map accuracy assessment. Uncertainties include the selection of which accuracy measures to report, how to interpret them, and the nature and quality of reference samples. As a result, map quality remains a difficult variable to consider objectively (Foody, 2002).

As mentioned in Chapter 2, the most widely used method for classification accuracy assessment in remote sensing is a confusion or error matrix. If Table 3.2 is taken as an example of an error matrix, overall accuracy is obtained by dividing the total number of correctly classified pixels (diagonal units) by the total number of reference pixels.

		Predicted				
Actual	Class	A	B	C	Σ	Producer's accuracy
	A	15	10	0	25	$(15/25)*100=60.00\%$
	B	2	20	3	25	$(20/25)*100=80.00\%$
	C	5	2	18	25	$(18/25)*100=72.00\%$
	Σ	22	32	21	75	
	User's accuracy	$(15/22)*100=61.20\%$	$(20/32)*100=62.50\%$	$(18/21)*100=85.70\%$		Overall accuracy: $(15+20+18/75)*100=70\%$

Table 3.2. Error matrix example

Although overall accuracy is the most commonly reported, it is not enough as it does not indicate how the classifier performs for each of the classes. Producer's and user's accuracies are ways of calculating individual category accuracies. The producer's accuracy relates to the probability that a reference pixel will be correctly mapped and measures the errors of omission. In contrast, the user's accuracy indicates the probability that a sample from land cover map actually matches what it is from the reference data and measures the error of commission (Congalton and Green, 1999). Producer's accuracy can be calculated by dividing the correct number of pixels classified for a particular class by the row total. In Table 3.2 the producer's accuracy of class A would be obtained by dividing 15 correctly classified pixels for class A by 25 pixels that belong to this class. This would give a 60.00% producer's accuracy. The user's accuracy is calculated by dividing the total pixels classified as class A by the column total, (this is, 15/22) which result would be 61.20%. This means that only 61.20% of those 100% pixels identified as A are actually class A on the ground. Having a high producer's accuracy is probably the most important in the context of the present research as this will show the potential of the classifier to identify the class of interest on the ground. If a low user's accuracy is obtained this means that there is a high percentage of error of commission, this is, there are more

pixels identified on the ground as a class of interest than the actual amount. But this error could be easily corrected by the use of available ancillary data and ground surveys.

Also as mentioned in Chapter 2, there are a few issues regarding this measure of accuracy. For example the problem of mixed pixels is not addressed in an error matrix. However this is not dealt with in this research and therefore it will not be studied in detail. Another important issue is the effect of chance agreement; this is, when some cases might be allocated correctly by chance (Congalton, 1991, Pontius, 2000, Rosenfield and Fritzpatrick-Lins, 1986). One measure that has been widely used in order to address this matter is Cohen's kappa coefficient (Cohen, 1960, Congalton and Mead 1983, Stehman 1996, Smits *et al.*, 1999). It is also called the KHAT statistic and indicates whether the confusion matrix is different from a random result. It is also used to compare different matrices from different classifiers and to determine if one is significantly better than the other.

The KHAT statistic is computed as follows:

$$\hat{K} = \frac{N \sum_{i=1}^r x_{ii} - \sum_{i=1}^r (x_{i+} x_{+i})}{N^2 - \sum_{i=1}^r (x_{i+} x_{+i})}$$

Equation 3.2 KHAT statistic

Where r is the number of rows in the matrix, x_{ii} is the number of observations in row i and column i , (x_{i+}) and (x_{+i}) are the marginal totals of row i and column i and N is the total of observations.

Two results can be compared by using a test for significant difference or Z statistic test:

$$Z \approx \frac{\hat{K}_1 - \hat{K}_2}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}$$

Equation 3.3 Z statistic test

Although the above method is extensively used, Foody (2002) points out that accuracy assessment is still a topic of considerable research within the remote sensing community, partly because methods such as Kappa statistics have become standard practice but their use is not always appropriate. Some authors argue that the K coefficient overestimates the chance of agreement and underestimates the classification accuracy (Foody, 1992, Ma and Redmond, 1995) or that it is not appropriate because it is non-probability based (Stehman and Czaplewski, 1998). Most importantly, various measures of accuracy evaluate different components of accuracy making different assumptions of the data (Lark, 1995c) and therefore there is no single universally acceptable measure of accuracy but a variety of indices sensitive to different features (Stehman, 1997).

While the Z statistics are in principle used to compare different accuracy results, in the case of this thesis it would not be appropriate as the samples used for comparison are not independent. Therefore, it would be better to look at specific pairwise analysis techniques.

Of all the paired tests available, McNemar's test seems to be the most appropriate for this research (Foody, 2004b). It uses matched-pairs of labels (A , B) and it determines whether the proportion of A and B labels is equal for both classifiers. It is a variation of the Chi-square test with the difference that in the Chi-square test the two datasets are independent and in the McNemar's test the datasets are the same. McNemar's test should be used when comparing the same data which have been processed in two different ways. This test is calculated as follows:

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}}$$

Equation 3.4 McNemar's test

The elements of this formula are given by the following confusion matrix:

	Classification 2	
Classification 1	Correct	Incorrect
Correct	f_{11}	f_{12}
Incorrect	f_{21}	F_{22}

Table 3.3 Confusion matrix for McNemar's test. Based upon Foody (2004).

Therefore, the results of all the classifications will be presented using an error matrix in which overall, user's and producer's accuracies will be compared and also a McNemar's test will be performed to assess the chance of agreement of those error matrices. With these accuracy measures, the results obtained by the different classifiers (using the same testing set, as explained earlier) will be compared as follows:

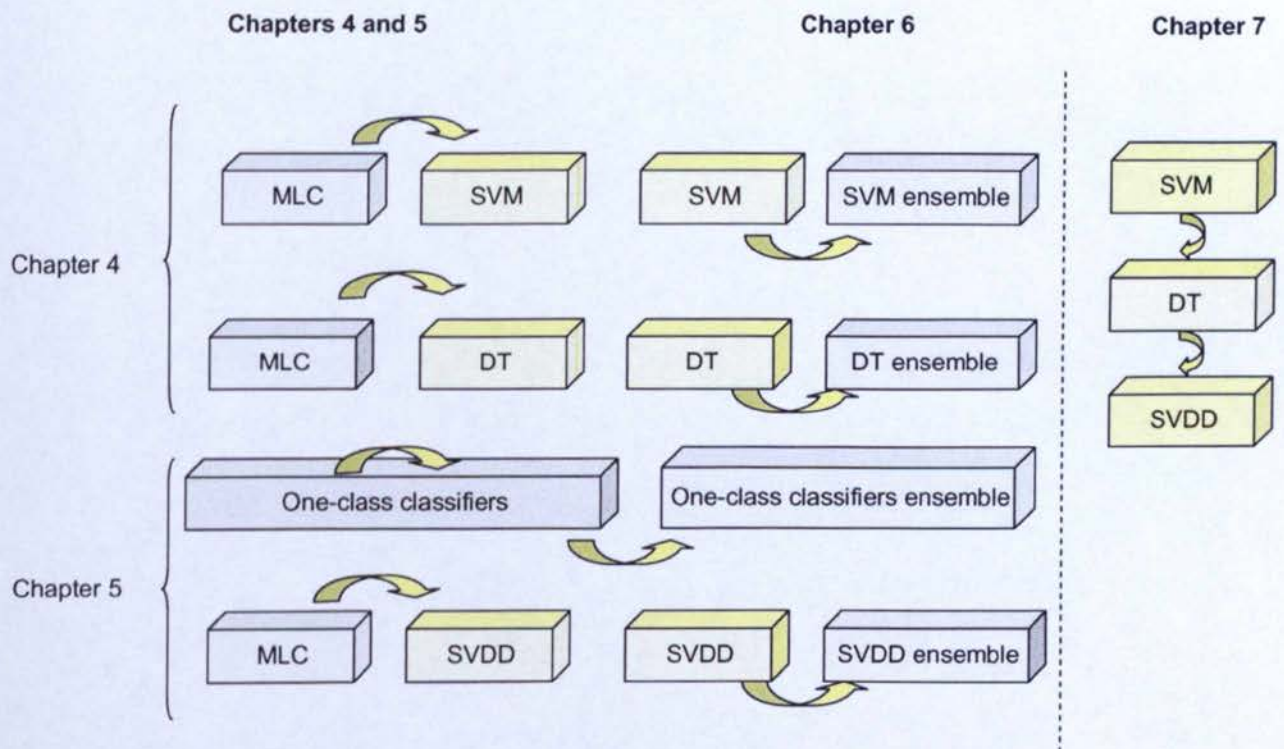


Figure 3.6 Comparison of classifier's accuracies in different chapters. The arrows represent the different comparisons

3.5 Summary

As mentioned already, the purpose of this research is the investigation of methods for mapping one particular class (habitat) using remote sensing. It is therefore necessary to carefully select the training and testing datasets necessary to carry out the two different approaches that have been selected in the previous chapter: (i) a binary classification using SVM and DT classifiers and (ii) a one-class classification using boundary methods. It was decided that the main habitat of interest to perform the classification should be fen and that the second habitat of interest should be saltmarsh. The latter was chosen in order to compare the results obtained by the classifiers with the class fen as this would show whether the classifiers were biased towards the spectral characteristics of the class. The areas of interest selected for each of them were The Norfolk Broads and the North Norfolk Coast. The remote sensing data used in the research was a Landsat ETM+ image dated 19th of June 2000 provided by NERC. The ground data were provided by English Nature and the Environment Agency in the form of aerial photography taken close to the date of the satellite image. Finally, it was decided that the results from the classifications performed by the different algorithms selected in Chapter 2 should be compared using a confusion matrix. The McNemar's test would be used to assess whether the accuracies were statistically different.

All the above considerations about data and how the data have been selected to train the classifiers are of the utmost importance in order to have an accurate output. Having defined the purpose of this research, the main classification methods to achieve this purpose and the data to be used in these classifications, the following chapter will investigate SVM and DT classifiers for the classification and mapping of a specific habitat of interest.

4 Binary classification for land cover mapping of a class of interest: Support Vector Machines and Decision Trees versus Maximum Likelihood classifiers

*"The world is divided into two classes, those who believe the incredible,
and those who do the improbable."
Oscar Wilde*

The present chapter is the first of the three research chapters that form the core of this thesis. The objective of these research chapters is to evaluate the potential of different classification methods in order to classify accurately a particular habitat of interest. In doing this, these chapters will contribute towards the overall aim of this thesis: to investigate and evaluate classification methods for mapping one particular habitat of interest with the aid of remote sensing data. These chapters also intend to meet the sub-aims of this thesis increasing the classification accuracy when focusing on a habitat of interest by (i) optimising the use of training data and (ii) optimising the use of remote sensing by applying suitable classifiers to the specific task of classifying a class of interest.

In order to do this, the specific objective of this chapter is to evaluate the potential of binary classification for the mapping of a specific habitat of interest using SVM and DT classifiers. These two classifiers have been recently introduced to remote sensing land cover mapping and provide an alternative to the standard multi-class classifiers when focusing on a specific land cover class of interest (as explained in Chapter 2). The results obtained by the SVM and DT classifiers will be compared against those of a ML classifier which as mentioned in Chapter 2, will be used as benchmark. In order to achieve this objective, this chapter will be structured as shown in Figure 4.1:

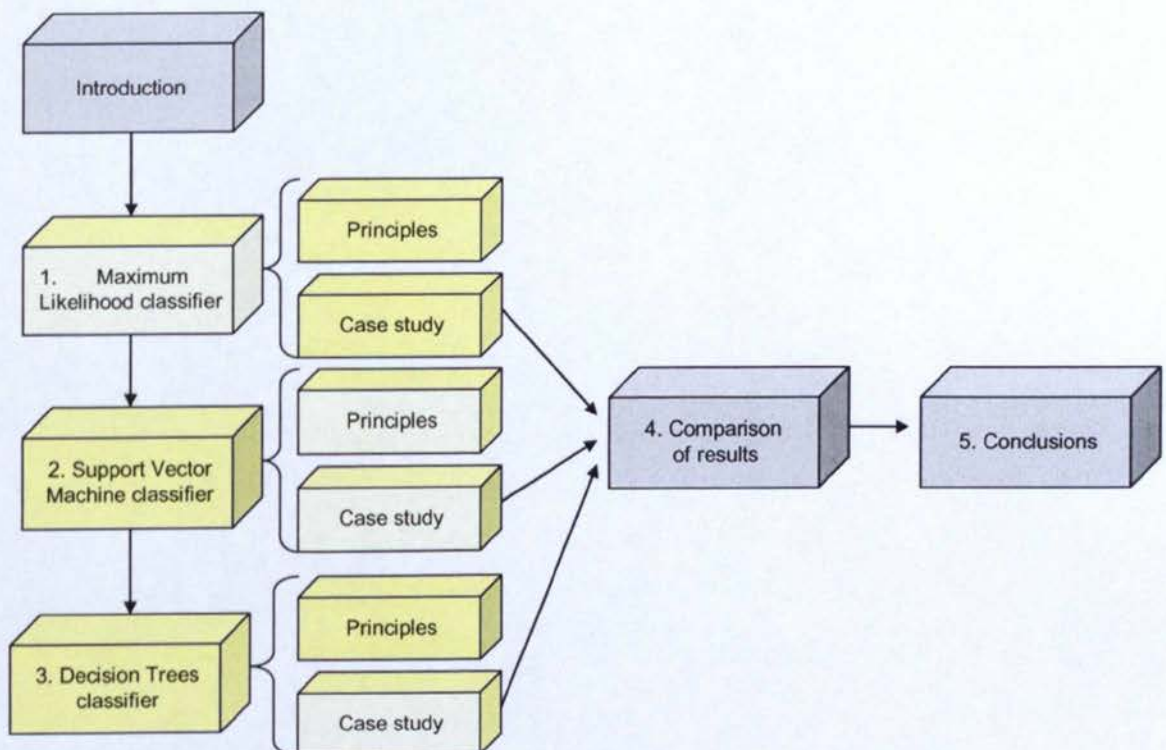


Figure 4.1 Chapter 4 structure

To be able to understand the following sections within this chapter it is necessary to describe a few basic concepts about binary classification. In binary classification, a training set X^r used to train the classifier is formed by the set of pixels $X^r = \{(\mathbf{x}_i, y_i) | i = 1, \dots, N\}$, in which \mathbf{x}_i is a vector formed by the values of that pixel for each of the variables represented in feature space and will be labelled with y_i

being $y_i \in \{+1, -1\}$. Those pixels belonging to the class of interest will be labelled as $y_i = +1$ and those pixels that belong to the other class will be labelled as $y_i = -1$. In order to perform this OVA binary classification, a function has to be derived from the training set in such a way that for a given pixel \mathbf{x}_i an estimate of the label $y_i \in \{+1, -1\}$ can be obtained (Tax, 2001). In most classification methods the type of function f is chosen before hand and just a few parameters of the function have to be determined. The function can be represented by $f(\mathbf{x}; \mathbf{w})$ which states the dependence on the parameters or weights \mathbf{w} . Examples of these functions have been described in Chapter 2 and include those used by parametric and non-parametric classifiers. The separating function is normally represented by a separating hyperplane in the feature space. A hyperplane is an N -dimensional analogy of a line or plane, which divides an ' $N + 1$ ' dimensional space into two (Figure 4.2).

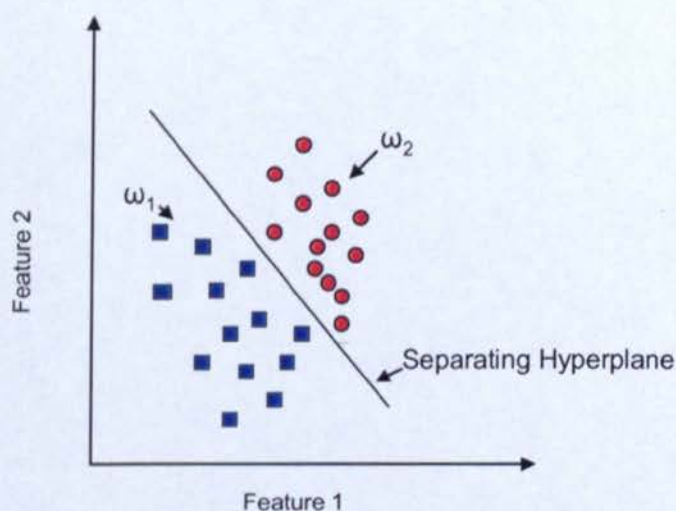


Figure 4.2 Representation of a hyperplane separating two classes.

The following sections will describe the principles behind ML, SVM and DT classifiers and how these classifiers can be applied to the classification of a class of interest. It is important to highlight that this is not an exhaustive description of the three classifiers and that only the most important and relevant aspects will be explained. The results obtained by the three classifiers will be compared and in the final section conclusions will be derived from the results of the three classifications.

4.1 Maximum Likelihood classification: Principles and application to the case study

Discriminant analysis based on ML classification algorithms is still widely applied and it is traditionally used as a baseline for the classification of remotely sensed data. Furthermore, ML is still the standard in the routine work of the space agencies and remote sensing research and it is normally used as a benchmark when assessing the performance of any other classifiers (Arbia *et al.*, 1999).

ML classification is based upon the assumption that there exist statistical models describing the distribution of the classes in the feature space. Given these models, the class of a new object is determined by calculating which of the models is more likely to describe that object. ML classification usually assumes normal (Gaussian) models. The normal distribution is specified by two parameters, the mean and the variance. A characteristic property of the normal distribution is that 68% of all of its observations fall within a range of ± 1 standard deviation from the mean, and a range of ± 2 standard deviations includes 95% of the scores. In other words, in a normal distribution, observations that have a standardized value of less than -2 or more than +2 have a relative frequency of 5% or less. (standardised value means that a value is expressed in terms of its difference from the mean, divided by the standard deviation) (Hill and Lewicki 2006). The mean controls the location of the distribution and the variance controls the spread of the data. When more than one feature is involved, there is one mean for each feature making up a mean vector. The multivariate equivalent of the variance is the variance-covariance matrix, representing the variability of pixel values for each feature within a particular class and the correlations between the features.

Given these two parameters, the statistical probability of a pixel being a member of a particular land cover class is computed. The results are probability density functions which are used to classify the unknown pixel by computing the probability of the

pixel belonging to each category by a discriminant analysis (Figure 4.3) (Lillesand and Kiefer, 2004).

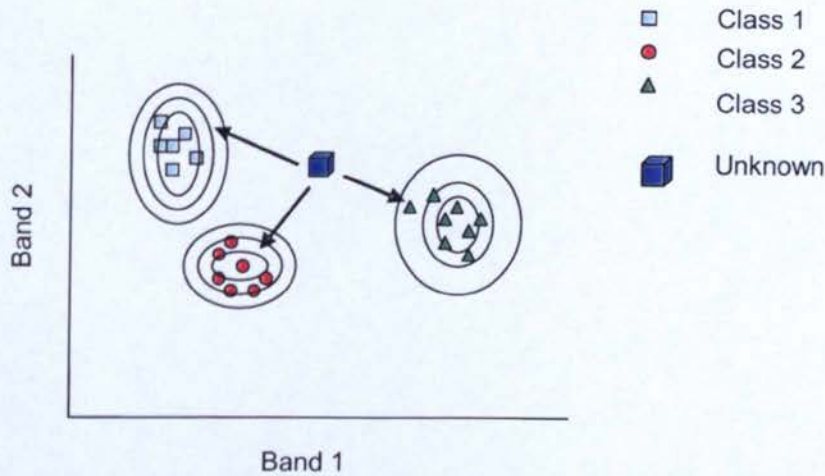


Figure 4.3. Maximum likelihood classification. Based upon Lillesand and Kiefer (2004).

The unknown pixel is assigned to the class for which the probability of membership is the highest. Although in practice the assumption of normally distributed data is not always met, the classifier generally outputs an acceptable result (Pal, 2002).

For the multivariate case, statistical theory describes the probability that an observation vector $\mathbf{x} = (x_1, x_2, x_3, \dots, x_n)$ belongs to class k $j = 1, 2, 3, \dots, c$, based on the following formula:

$$P_{k_j}(\mathbf{x}) = (2\pi)^{-1/2\rho} \left| \sum_{k_j} \right|^{-1/2} \times e^{-1/2(\mathbf{x}-\mu_{k_j})^T \cdot \sum_{k_j}^{-1} (\mathbf{x}-\mu_{k_j})} \quad (1)$$

Where $P_{k_j}(\mathbf{x})$ is the probability density value associated with the observation vector \mathbf{x} quantified for class k_j , \sum_{k_j} is the covariance matrix of the class k_j with dimension $\rho \times \rho$, μ_{k_j} is the mean vector of the class k_j and $| |$ represents the determinant of the given matrix. As applied in maximum likelihood decision rule, this equation allows the calculation of separate probabilities that an observation is a member of each of k

classes. The individual is then assigned to the class for which the probability value is greatest.

In an operational context this equation can be reduced to the following one by taking logarithms to the base e :

$$\ln[P_{k_j}(\mathbf{x})] = -\frac{1}{2} \rho \ln(2\pi) - \frac{1}{2} \ln|D_{k_j}| - \frac{1}{2} (\mathbf{x} - m_{k_j})^T \cdot D_{k_j}^{-1} \cdot (\mathbf{x} - m_{k_j}) \quad (2)$$

where D_{k_j} is the estimate of matrix \sum_{k_j} and m_{k_j} is the estimate of μ_{k_j} . These estimates are computed from the training data. As $\rho \ln(2\pi)$ is the same for all classes it can be regarded as a constant and omitted. The remainder of this equation could be written as follows:

$$-2 \ln[p_{k_j}(\mathbf{x})] = \ln|D_{k_j}| + (\mathbf{x} - m_{k_j})^T \cdot D_{k_j}^{-1} (\mathbf{x} - m_{k_j}) \quad (3)$$

where the expression

$$(\mathbf{x} - m_{k_j})^T \cdot D_{k_j}^{-1} (\mathbf{x} - m_{k_j}) \quad (4)$$

is the measure of the distance of one observation vector from the class mean m_{k_j} , corrected for the variance and covariance of the class k_j and is known as the Mahalanobis distance. An observation vector will be assigned to the class for which the value $-2 \ln[p_{k_j}(\mathbf{x})]$ is the smallest.

Furthermore, a ML classification normally assumes a linear discriminant analysis (LDA). However, it is important to know that quadratic discriminant analysis (QDA) and regularized discriminant analysis (RDA) methods are also used as classification methods to determine to which class an unknown pixel belongs. LDA assumes that

the variability within each group follows a normal distribution with the same matrix of covariance. If the covariance matrices of the categories are not equal, then the separation surface is achieved by a quadratic function which has greater curvature the greater the difference between the covariance matrices. Using QDA each of the covariance matrices is estimated separately, which requires a larger sample of data than that used in LDA to reach the same level of reliability in the predictions (Sanchez and Sarabia, 1995). After experiments carried out with these three discriminant methods and different type of data, Sanchez and Sarabia (1995) concluded that the dominant component is the size of the training data while changing the method is not significant. The increase in the percentage of correct classifications can only be associated with the degree of separability between classes.

One of the important drawbacks of the ML classifier is that the reliability of the results obtained declines when the distribution of the data differs from normality. Furthermore, the reliability of the estimates of mean vector and variance-covariance matrix is affected by the relationship between sample size and the number of features. In general, the bigger the sample and the bigger the number of features the better the classification (Pal, 2002). But this leads to another important drawback which is the computational cost required in order to classify each pixel. To address these issues new classifiers such as the SVMs and DTs are being introduced into remote sensing research. Nevertheless, as mentioned at the beginning of this section, ML classification is normally used as a benchmark against which the results of other classifiers are measured and it was used for that purpose within this thesis.

4.1.1 Classification of a habitat of interest using ML classification.

Case study.

ML classifiers are normally used to perform standard multiclass classifications where all the classes present in the image are taken into account. As mentioned in Chapter 3, some of the literature suggests the use of a minimum of 10-30p cases per-class for

training, where p is the number of wavebands used (Mather, 2004). As a feature selection was performed with the result of having ETM+Band 2 and NDVI as input variables, the minimum recommended training size for each class should be of approximately between 20 and 60 pixels. In order to assess the impact of training data size in the performance of the ML classifier, it was decided to perform the ML classification with two training sets. One of them was formed by less than the minimum amount recommended of pixels per class (Table 4.1) and another training set with more than the minimum examples for each class (150 pixels per class) (Table 4.2). The rest of the classifiers were subsequently trained within this range of values for comparison purposes.

Training set 1	
Class	Pixels
Saltmarsh	15
Fen	15
Agriculture	15
Forest	15
Grazing marshes	15
Sand	15
Urban	15
Water	15

Table 4.1 ML classification. Training set 1. 15 pixels per class

Training set 2	
Class	Pixels
Saltmarsh	150
Fen	150
Agriculture	150
Forest	150
Grazing marshes	150
Sand	150
Urban	150
Water	150

Table 4.2 ML classification. Training set 2. 150 pixels per class

Moreover, as mentioned in Chapter 3, the same testing data set was used to test the accuracy of all the classifiers within this thesis. The testing set is formed by 250 pixels, 50% belonging to the class of interest and 50% belonging to the other class. The confusion matrices obtained for each of the training and testing sets are as follows:

Chapter 4

Binary classification for land cover mapping of a class of interest: Support Vector Machines and Decision Trees versus Maximum Likelihood classifiers.

	Training set 150 pixels per class	Predicted									
	Class	FE	SM	A	FO	G	S	U	W	Σ	Producer's accuracy (%)
Actual	Fen (FE)	97	14	0	14	0	0	0	0	125	77.60
	Saltmarsh (SM)	3	14	0	0	0	0	3	0	20	70.00
	Agriculture (A)	0	0	1	8	5	3	3	0	20	0.00
	Forest (FO)	11	0	0	3	0	0	0	0	14	21.43
	Grazing marsh (G)	0	0	2	1	18	0	0	0	21	85.71
	Sand (S)	0	0	0	0	0	15	0	0	15	100.00
	Urban (U)	0	0	1	0	2	1	14	0	18	77.78
	Water (W)	0	0	0	0	0	0	0	17	17	100.00
	Σ	111	28	4	26	25	19	20	17	250	
	User's accuracy (%)	87.39	50.00	0.00	11.54	72.00	78.95	70.00	100.00		Total accuracy 71.60%

Table 4.3 ML classification confusion matrix using 150 pixels per class. Fen as the class of interest (Testing set formed 50% fen, 50% others)

Chapter 4

Binary classification for land cover mapping of a class of interest: Support Vector Machines and Decision Trees versus Maximum Likelihood classifiers.

	Training set pixels per class	Predicted										
	Class	FE	SM	A	FO	G	S	U	W	Σ	Producer's accuracy (%)	
Actual	Fen (FE)	91	11	0	19	4	0	0	0	125	72.80	
	Saltmarsh (SM)	8	9	0	0	0	0	3	0	20	45.00	
	Agriculture (A)	0	0	1	0	13	3	3	0	20	5.00	
	Forest (FO)	3	0	0	10	1	0	0	0	14	71.43	
	Grazing marsh (G)	0	0	0	1	20	0	0	0	21	95.24	
	Sand (S)	0	0	0	0	0	12	3	0	15	80.00	
	Urban (U)	0	3	1	0	0	0	14	0	18	77.78	
	Water (W)	0	0	0	0	0	0	0	17	17	100.00	
	Σ	102	23	2	30	38	15	23	17	250		
	User's accuracy (%)	89.22	39.13	0.00	33.33	52.63	80.00	60.87	100.00		Total accuracy 69.60%	

Table 4.4 ML classification confusion matrix using 15 pixels per class. Fen as the class of interest (Testing set formed 50% fen, 50% others)

Chapter 4

Binary classification for land cover mapping of a class of interest: Support Vector Machines and Decision Trees versus Maximum Likelihood classifiers.

	Predicted										
Training set 150 pixels per class	SM	FE	A	FO	G	S	U	W	Σ	Producer's accuracy (%)	
Actual	Saltmarsh(SM)	78	22	1	1	0	23	0	125	62.40	
	Fen (FE)	4	16	0	0	0	0	0	20	80.00	
	Agriculture (A)	1	0	1	8	5	3	2	20	0.00	
	Forest (FO)	0	11	0	3	0	0	0	14	21.43	
	Grazing marsh (G)	0	0	2	1	18	0	0	21	85.71	
	Sand (S)	0	0	0	0	0	15	0	15	100.00	
	Urban (U)	0	0	1	0	2	1	14	18	77.78	
	Water (W)	0	0	0	0	0	0	17	17	100.00	
	Σ	83	49	5	13	25	19	39	17	250	
	User's accuracy (%)	93.98	32.65	0.00	23.08	72.00	78.95	35.90	100.00	Total accuracy 64.80%	

Table 4.5 ML classification confusion matrix using 150 pixels per class. Saltmarsh as the class of interest (Testing set formed 50% saltmarshes, 50% others)

Chapter 4

Binary classification for land cover mapping of a class of interest: Support Vector Machines and Decision Trees versus Maximum Likelihood classifiers.

	Training set pixels per class	Predicted										
	Class	SM	FE	A	FO	G	S	U	W	Σ	Producer's accuracy (%)	
Actual	Saltmarsh (SM)	65	27	2	0	1	0	30	0	125	52.00	
	Fen (FE)	4	14	0	2	0	0	0	0	20	70.00	
	Agriculture (A)	0	0	1	0	13	3	2	0	20	0.00	
	Forest (FO)	3	0	0	10	1	0	0	0	14	71.43	
	Grazing marsh (G)	0	0	0	1	20	0	0	0	21	95.24	
	Sand (S)	0	0	0	0	0	12	3	0	15	80.00	
	Urban (U)	0	3	1	0	0	0	14	0	18	77.78	
	Water (W)	0	0	0	0	0	0	0	17	17	100.00	
	Σ	72	44	4	13	35	15	49	17	250		
	User's accuracy (%)	90.28	31.82	0.00	76.92	57.14	80.00	28.57	100.00		Total accuracy 61.20%	

Table 4.6 ML classification confusion matrix using 15 pixels per class. Saltmarsh as the class of interest (Testing set formed 50% saltmarshes, 50% others)

The confusion matrices shown in Table 4.4 and Table 4.6 were obtained using 15 pixels per class which is less than the recommended range of 10-30 pixels per class. As it can be observed, the overall accuracy obtained when using fen and saltmarsh as classes of interest were 69.60% and 61.20% respectively. When focusing on the results of a class of interest (e.g. fen) the producer's accuracy obtained was 72.80% and user's accuracy 89.22%. For saltmarsh, the result for the producer's accuracy was 52.00% which is quite low when compared with the producer's accuracy obtained for fen but a very high user's accuracy of 90.28%.

Increasing the training dataset to 150 pixels per class produced an increase in overall accuracy for fen and saltmarsh (71.20% and 64.80% respectively). The main class of interest to this thesis, the class fen, also showed an increase in producer's and user's accuracy (77.60% and 87.39% respectively) (see Table 4.3). This increase was also shared by the class saltmarsh which producer's accuracy increased up to 62.40% and the user's accuracy to 93.98% (Table 4.5). As mentioned before, the ML classifier obtains better classification accuracies the bigger the sample and the bigger the number of features (Pal, 2002). Consequently, the overall accuracy and the producer's and user's accuracy for each of these classes could be potentially higher by using all the features available or by adding more pixels to the training datasets. However, this would increase enormously the amount of training data needed (e.g. 150 pixels multiplied by 6 non-thermal bands per class). This would result in a very large training dataset and a huge computational effort. Also, as already mentioned, this parametric classifier focuses on obtaining a high overall accuracy without paying attention to any particular class. These are precisely two of the issues that this thesis is addressing.

As said at the beginning of this section, the purpose of performing a ML classification was to use the results obtained by this classifier as a benchmark for those obtained by other classification methods. Therefore, these results shown above

were taken as a benchmark to measure the ones obtained by the SVM and DT classifiers in the following sections of this chapter.

4.2 Support Vector Machines: Principles and Case study

The basis of SVMs started in the 1970's with the theories published by Vapnik and Chervonenkis on statistical machine learning (Vapnik and Chervonenkis, 1971, 1979) and later on with the development of the support vector machine algorithm (Vapnik, 1995, 1998). Machine learning non-parametric classifiers were developed in order to give a better solution to the problems that could not be solved by conventional parametric classifiers. Its use has been increasing in several research areas. For example handwritten digit recognition (Cortes and Vapnik, 1995, Le Cun *et al.*, 1995), text categorisation (Drucker *et al.*, 1999, Joachims, 1998), face detection (Osuna *et al.*, 1997), pharmaceutical data analysis (Burbidge *et al.*, 2001, Czerminski *et al.*, 2001), time series forecasting (Tay and Cao, 2001, Cao, 2003) and computational neuroscience (Eghbalnia and Asadi, 2001).

It is only very recently that SVMs have started to be applied in remote sensing (Huang *et al.*, 2002, Zhu and Blumberg, 2002, Pal and Mather, 2003, 2005, Foody and Mathur, 2004a). Huang *et al.* (2002) compared the performance of SVMs for land cover classification against DT, ML and ANN classifiers using data from Landsat Thematic Mapper. Of the four classifiers ML had lower accuracies than the three non-parametric ones. Their results showed that the SVM was more accurate than DTs and also gave higher accuracies than ANNs when more variables were used in the calculations. Other comparative studies include Pal and Mather (2003, 2005) where the performance of SVMs are compared against DTs and ANNs using data from ETM+, SAR and DAIS hyperspectral data. Their results show that the performance and accuracy obtained by a SVM is comparable to that of ANNs and DTs. More advanced research into the nature of SVMs includes Foody and Mathur (2004b) in which the potential for intelligent training of SVMs is investigated. This

is based on the idea that only a specific part of a training data set is necessary to train the classifier.

The wide range of applications and success of the SVMs are mainly due to the following characteristics (Bennett and Campbell, 2000):

- 1) The general methodology is very flexible. It can be customized to meet particular application needs.
- 2) They eliminate problems experienced with other methods such as neural networks and decision trees:
 - a. There are few parameters to pick.
 - b. The final results are stable, reproducible and largely independent of the specific algorithm used to optimize the SVM.
- 3) The method is relatively simple to use compared to classifiers such as neural networks.
- 4) They have proven to be robust to noise and perform well on small training samples.
- 5) SVMs always find a global solution. A global solution means that there exists no other point in the feasible region at which the objective function takes a lower value (Burgess, 1998). This is a distinctive characteristic of SVMs in contrast to the case of ANNs where many local minima solutions exist. Both concepts are exemplified in Figure 4.4.

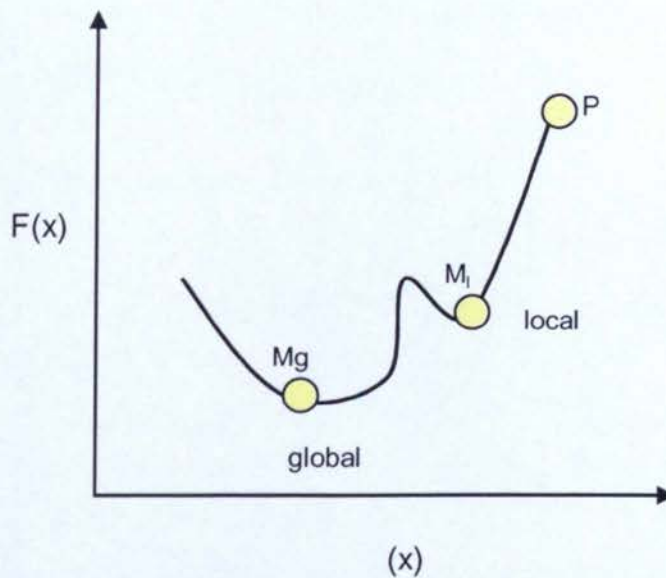


Figure 4.4 Local minima and global solution. If a classifier starts with a weight set corresponding to point P the first solution that it encounters is at M_i . This is called local minima and corresponds to a partial solution in response to the training data. M_g is the global minimum or global solution. In neural networks unless measures are taken to escape from the local minima the global solution will never be reached (Based upon Burges, 1998).

It is all these characteristics and its binary nature that makes the SVM a very suitable approach for its application to classifying a specific habitat of interest. Therefore, the following sections will describe the principles behind the SVMs and the application of this classifier in the case study.

4.2.1 Support Vector Machines: Principles

In the two-class classification problem where $\{\mathbf{x}_i, y_i\}$, $i = 1, \dots, l$, $y_i \in \{-1, 1\}$ it is necessary to find an optimal separation between those two classes. When operating in a multi-dimensional feature space this separation is normally achieved by a separating hyperplane. In this case a hyperplane would separate positive from negative examples and would satisfy $w \cdot \mathbf{x} + b = 0$, where w is the normal to the

hyperplane, $|b|/\|w\|$ is the perpendicular distance from the hyperplane to the origin and $\|w\|$ is the Euclidean norm of w .

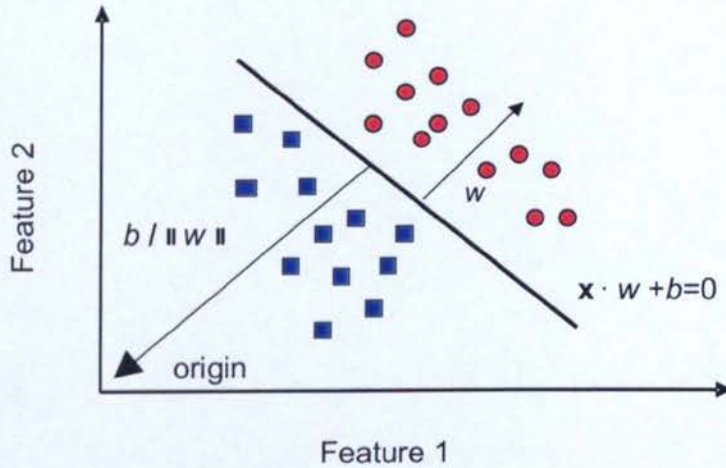


Figure 4.5. Representation of separating Hyperplane($w \cdot x + b = 0$). Based upon Burges (1998).

If d_+ (d_-) are the shortest distance from the hyperplane to the closest positive (negative) example, the margin of a separating hyperplane would be $d_+ + d_-$. For the linearly separable case the support vector machine looks for the hyperplane with the largest margin (see Figure 4.6). If all the training data satisfy the following constraints:

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} + b &\geq +1 \quad \text{for} \quad y_i = +1 \\ \mathbf{x}_i \cdot \mathbf{w} + b &\leq -1 \quad \text{for} \quad y_i = -1 \end{aligned} \quad (1)$$

Then these can be combined into:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \quad \forall_i \quad (2)$$

If the first inequality holds the points that lie in H_1 : $\mathbf{x}_i \cdot \mathbf{w} + b = 1$ and if the second holds H_2 : $\mathbf{x}_i \cdot \mathbf{w} + b = -1$. These points are called **support vectors**. H_1 and H_2 are parallel (they have the same normal w) and no training points fall between them (see

Figure 4.6). Therefore it is possible to find a pair of hyperplanes which give maximum margin subject to constraints (2).

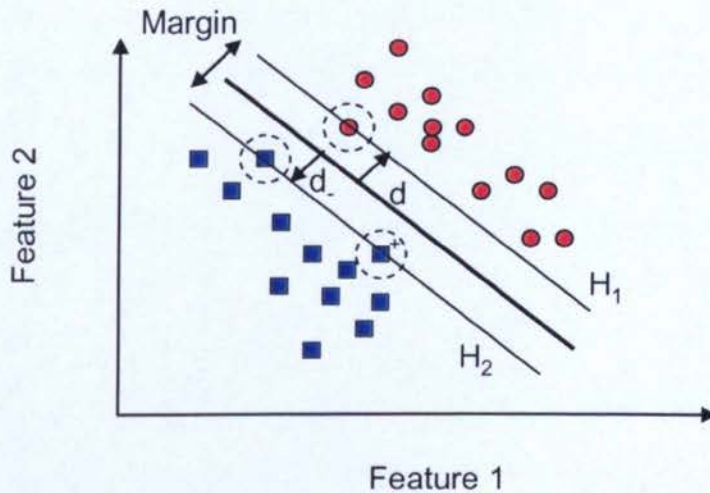


Figure 4.6. Optimal separating hyperplane. The circled points represent support vectors. Based upon Burges (1998).

As said before, those points that lie on one of the hyperplanes H_1 or H_2 are called support vectors. They are critical elements of the training set. They lie closest to the decision boundary. If all the other training points were removed or moved around and the training was repeated the same separating hyperplane would be found. This is a fundamental advantage of SVMs with respect to other classifiers such as ANNs. If some of the training data were to be removed or values moved around, the ANN's architecture would have to be redesigned and the same solution would be difficult to achieve again.

All the above applies when dealing with linear separable cases. However, it is often the case that the classifier has to deal with non-linearly separable cases. In order to be able to generalise the above method to a nonlinear case it is necessary to switch this problem to a Lagrangian formulation. A Lagrangian formulation is an algebraic term within the context of problems of mathematical optimization subject to constraints. The aim of a Lagrangian is that the constraints are placed in the

Lagrangian multiplier themselves of the form α_i , $i = 1, \dots, l$. Therefore, the constrained problem can be converted to an unconstrained problem by forming a Lagrangian formulation and this would facilitate its application to non-linear cases. To form a Lagrangian the constraint equations are multiplied by positive Lagrangian multipliers and subtracted from the objective function. The objective function in this case is the maximisation of the margin and is given by $\frac{1}{2}\|w\|^2$. Lagrange multipliers are therefore introduced for each of the inequalities of constraints (2).

$$L_P = \frac{1}{2}\|w\|^2 - \sum_{i=1}^l \alpha_i y_i (\mathbf{x}_i \cdot w + b) + \sum_{i=1}^l \alpha_i \quad (3)$$

Having done this, it is possible to establish a dual problem. Minimise L_P with respect to w and b and simultaneously all derivatives of L_P vanish subject to $\alpha_i \geq 0$ (this is called constraints C_1) or maximise L_P subject to the constraints that the gradient of L_P with respect to w and b vanish and also $\alpha_i \geq 0$ (called constraints C_2). This is a quadratic problem called Wolfe dual (Fletcher, 1987) and it has the property that the maximum of L_P subject to constraints C_1 occurs at the same values of the w , b and α as the minimum of L_P subject to constraints C_2 .

The dual formulation will be expressed as:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \quad (4)$$

In this solution the points that are $\alpha_i > 0$ are the support vectors and they are essential for the calculation of the optimal hyperplane. All other training points have $\alpha_i = 0$ and are not necessary for any of the calculations performed. The dual formulation is still feasible when applied to non-separable data but the dual could grow quite large. To further relax the constraints on the Lagrange multipliers when necessary, slack variables are introduced (Cortes and Vapnik, 1995).

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} + b &\geq +1 - \xi_i & \text{for } y_i = +1 \\ \mathbf{x}_i \cdot \mathbf{w} + b &\leq -1 + \xi_i & \text{for } y_i = -1 \end{aligned} \quad (5)$$

Being $\xi_i \geq 0 \quad \forall_i$

For an error to occur ξ_i must be bigger than 1, so $\sum_i \xi_i$ is an upper bound on the number of training errors. So we can assign an extra cost for errors to the objective function $\|\mathbf{w}\|^2/2$ in the form of: $\|\mathbf{w}\|^2/2 + C(\sum_i \xi_i)$ where C is a parameter to be chosen by the user, a larger C corresponding to a higher penalty to errors.

a) Non-linear Support Vector Machines. The Kernel space

As seen in the previous section, a non-linear problem is much more difficult to solve than linear cases and real world applications normally require more flexible solutions than linear functions. A further step in the search for simplification of non-linear problems is the introduction of kernel representations. These kernel representations were introduced by Boser, Guyon and Vapnik (1992), based upon the work of Aizerman *et al.* (1964), and they consist of projecting the data into a high dimensional Euclidean space H in which the linear learning machines can be implemented (see Figure 4.7).

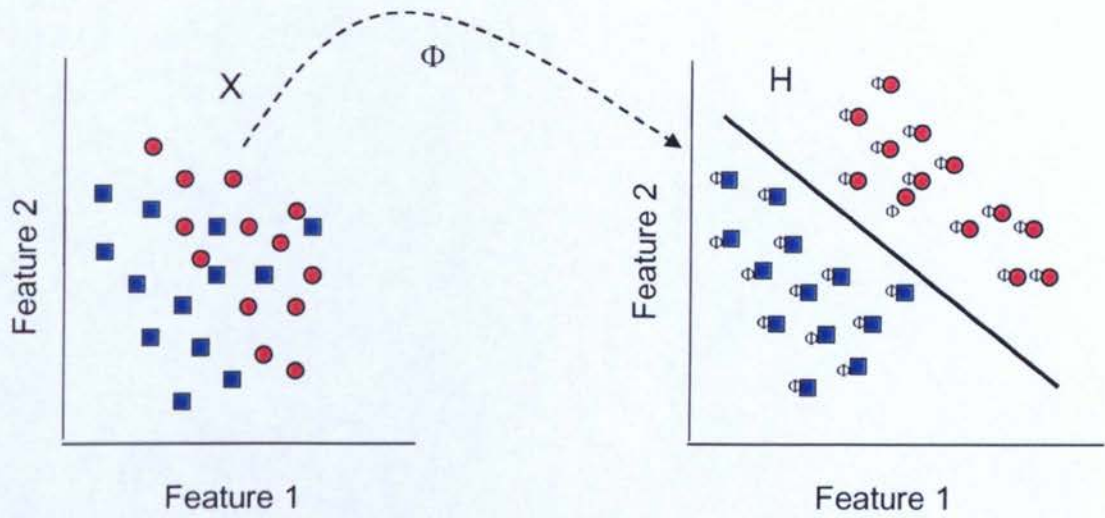


Figure 4.7 Mapping into feature spaces.

For that a mapping ϕ is performed so that $\phi: R^d \rightarrow H$. The kernel function would be expressed in the form of:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \quad (6)$$

The use of kernels makes it possible to map the data implicitly into a feature space and to train a linear machine in such space. The key is finding a kernel function that can be evaluated efficiently. There are many valid functions with kernels as described by Smola *et al.* (1998). For the aim of this research we will be focusing on three of the most important ones: (i) polynomial, (ii) gaussian radial basis function and (iii) exponential radial basis function.

(i) Polynomial kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1)^n \quad (7)$$

(ii) Gaussian radial basis function (RBF):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2s^2}\right) \quad (8)$$

(iii) Exponential radial basis function:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2s^2}\right) \quad (9)$$

Where s is the width of the gaussian and the exponential kernels. The kernel function plays a major role in locating decision boundaries between classes. Therefore, the selection of kernel function and appropriate values for corresponding kernel parameters might greatly affect the performance of the SVM (Huang *et al.*, 2002). In fact, according to Huang *et al.* (2002), some of the most important factors that affect the classification accuracy of SVM classifiers are:

1. Choice of kernel used.
2. Choice of the parameters related to a particular kernel.
3. Choice of parameter C .

However, there is little guidance in the literature on the criteria to be used to choose a kernel and its specific parameters (Pal and Mather, 2005). According to Cherkassky and Ma (2004) practical approaches to choose the most appropriate kernel parameters and value of C can be summarized as follows:

1. Kernel free parameters and C can be selected by users based on a priori knowledge and/or user expertise (Cherkassky and Mulier, 1998; Schölkopf *et*

al., 1999, Vapnik, 1998, 1999). Obviously, this approach is not appropriate for non-expert users.

2. Kwok (2000) and Smola *et al.* (1998) proposed asymptotically optimal values of the free parameter which are proportional to noise variance, in agreement with general sources on SVM (Cherkassky and Mulier, 1998, Vapnik, 1998, 1999). The main practical drawback of such a proposal is that it does not reflect sample size.

3 Selecting parameter C equal to the range of output values (Mattera and Haykin, 1999). This proposal does not take into account possible effects of outliers in the training data.

4 Using cross-validation for parameter selection (Cherkassky and Mulier, 1998, Schölkopf *et al.*, 1999, Hsu *et al.*, 2003). The only drawback of this approach could be very computation and data-intensive.

Considering all the above the most straightforward option for parameter selection seems to be cross-validation. It is therefore this approach that will be used to calibrate the kernels used in this research.

4.2.2 Classification of a habitat of interest using the SVM classifier. Case study.

After reviewing the basics of SVMs, this section describes the different classification experiments that were carried out., The SVM used to perform this research was the Support Vector Machine Toolbox developed by Steve Gunn (Gunn, 1998), a member of the Image Speech and Intelligent Systems Group at the University of Southampton. The main reason for this selection was that this particular SVM had a user-friendly graphical interface specially designed to analyse binary classifications.

As mentioned at the end of the previous section, the most straightforward method for selecting the optimal parameters to be used by the SVM is cross-validation. In v-fold cross-validation, the training set is divided into v subsets of equal size. Sequentially, one subset is tested using the classifier trained on the remaining subsets. Therefore, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data that are correctly classified (Hill and Lewicki 2006). In previous studies using the polynomial kernel (Cortes and Vapnik, 1995, Huang *et al.*, 2002) degrees of 1 to 8 were tested. These researchers found that depending on the input variables the degree of polynomial would have a major or minor impact in the final accuracy. In this case only two input variables were used but it was still decided to test the polynomial with values 1 to 10 to cover a slightly bigger range than previous studies. For the RBF kernel the default value of 1 for the free parameter was used in previous research (Vapnik 1995, Joachims, 1998). Other authors (Huang *et al.*, 2002) decided to use values 1 to 20. However in this latter case the overall accuracy only changed slightly when the value increased beyond 7.5. Based upon these studies it was decided to test the kernel using values 1 to 10. There were no studies found on the assessment of the better values to be used for the exponential kernel. It was decided to test the exponential kernel between the same values of the other two for comparison purposes.

These three main kernels were cross-validated in a 5-fold cross validation test using values 1 to 10 for their respective free parameters. For values of C exponentially growing sequences of C seems to be a practical method to identify the optimal value (Hsu *et al.*, 2003). Consequently, it was decided to use values 1, 10, 10^2 , 10^3 in order to calibrate the different kernels. The detailed results of this cross-validation are described in Annex B. The results for fen as the class of interest showed that the polynomial kernel failed to separate the two classes and classified all the pixels as the “other” class. The other two kernels gave positive results. As a general tendency for the gaussian and exponential kernels the accuracy results were higher when using free parameter values 1 to 3 with a clear decrease in accuracy as the parameter value

gets higher. Although the accuracy results were quite similar with both kernels the accuracy obtained by the gaussian kernel was always a bit higher than the exponential kernel. After averaging the results of the cross-validation, the highest overall accuracy was obtained by the gaussian kernel when using a value of $C = 100$ and free parameter value of 1.

To test whether the above results were biased in any way by the data provided or the spectral characteristics of the class of interest, another cross-validation was performed using saltmarsh as the class of interest. The results (Annex B) showed that the general tendencies are the same as before. The polynomial kernel continued classifying all the data as “others”. Also as with the class fen, the accuracy for the other two kernels seemed to decrease when the free parameter value increased and in general the best accuracies were obtained by the gaussian kernel. In terms of the value for misclassification errors, $C = 10$ and $C = 100$ gave the higher accuracies. In the case of saltmarsh the overall best accuracy was obtained by the gaussian kernel when using a penalty value of $C = 10$ and free parameter value of 2 and the same result was obtained by the Exponential kernel $C = 10$ and free parameter value of 1.

These parameters were consequently used by the SVM classifier when it was trained with the datasets of varying sizes. The effect of different training sizes on the overall accuracy of SVMs has been tested by previous research but always on multiclass classification (Huang *et al.*, 2002, Pal and Mather 2003). In this thesis, this was assessed for the binary classification. As described in Chapter 3 and earlier in the ML classification, it is suggested that a minimum of $10-30p$ cases per-class is used for training, where p is the number of wavebands used. In order to test whether this was also the minimum size recommended for this particular case study it was decided to go below this limit and above this limit. The sizes in this experiment ranged from 30, 50, 100, 150, 200, 250 to 300 pixels where each set contained the pixels from the previous one so that $30 \in 50 \in 100 \in 150 \in 200 \in 250 \in 300$. Following a binary (OVA) classification approach these training sets were divided 50/50 between the target class and the “other” class. As already mentioned, the testing set for all the

classifiers was 250 pixels that was divided in the same 50/50 proportion (see Annex F for full training and testing datasets).

The classification performed for fen as the class of interest gave the following results:

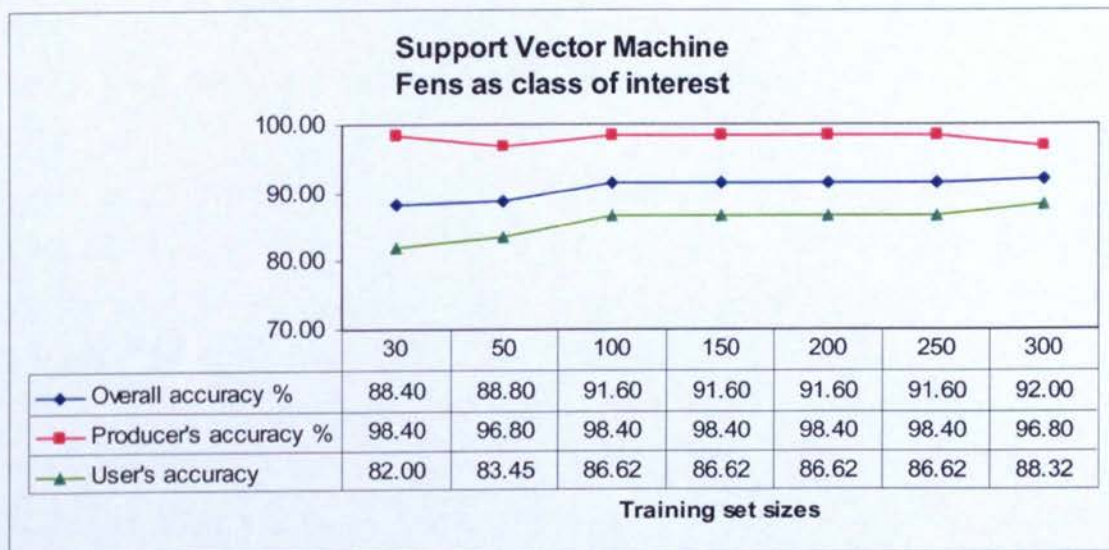


Figure 4.8. Accuracy results for class fen for different training sizes

Figure 4.8 shows that the small training datasets of 30 and 50 pixels achieved a lower overall accuracy than sizes 100 and above. However, the producer's accuracy for fen was similarly high for all the training sets. That means that the capacity of the SVM to differentiate the class of interest from all the other classes was quite high even when using very small training sizes. The user's accuracy stayed in the range of 82.00-88.00% which denoted a relatively high error of commission. With a training size of 100 pixels the SVM achieved an overall accuracy (91.60%) and a producer's accuracy of 98.40%. The increase on training set size to 150, 200, 250 and 300 did not change this result. Adding more pixels to the training set did not seem to add any new information into the classifier. So it could be concluded that for this particular case, the SVM found the optimal solution for the classification of the class of interest

with a training set of 100 pixels which is within the minimum recommended size (10-30p pixels).

To make sure that these results were not achieved by the SVM classifier because it was using this particular class of interest, the same classification was performed using saltmarsh as the class of interest. The results obtained were:

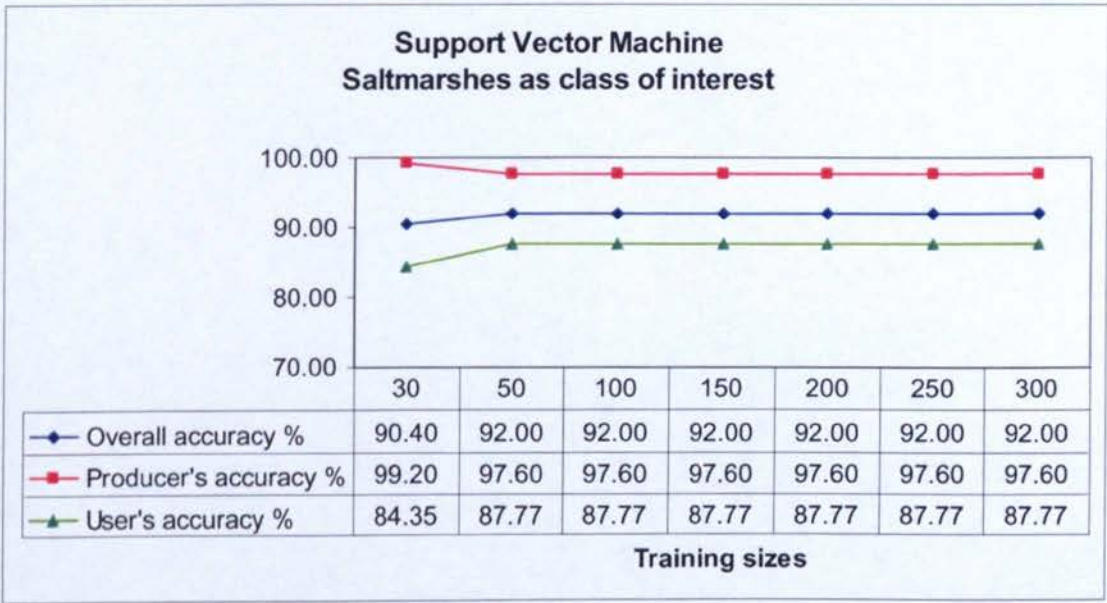


Figure 4.9. Accuracy results for class saltmarsh for different training sizes

For saltmarsh as class of interest (see Figure 4.9 above) the SVM found an optimal solution using a training size of 50 pixels with an overall accuracy of 92.00% and a producer's accuracy of 97.60%. After this, the accuracy values stayed the same independently of how much more data were added. Once more, this highlights the capacity of a SVM to find an optimal solution with a very small training dataset. As with the class fen, the producer's accuracy showed values of 97.60% for most of the training sizes and the user's accuracy stayed in the range of 84.00-88.00%. As mentioned in Chapter 3, this low user's accuracy could be rectified in a post-classification analysis if enough field data were available for the area. Confusion matrices for both classes and all the sizes can be found in Annex A.

Finally, one important characteristic of SVM classification is that it only uses a percentage of the training data to find the optimal hyperplane. The more data it uses the more it relies on the training dataset and the more difficult it would be to generalise to unseen samples. Therefore, the number of support vectors used in each of the above cases could give us an indication of how much the SVM relied on the training data for its calculations and if this was producing over-fitting. For that reason, the percentage of support vectors used in each training dataset was calculated.

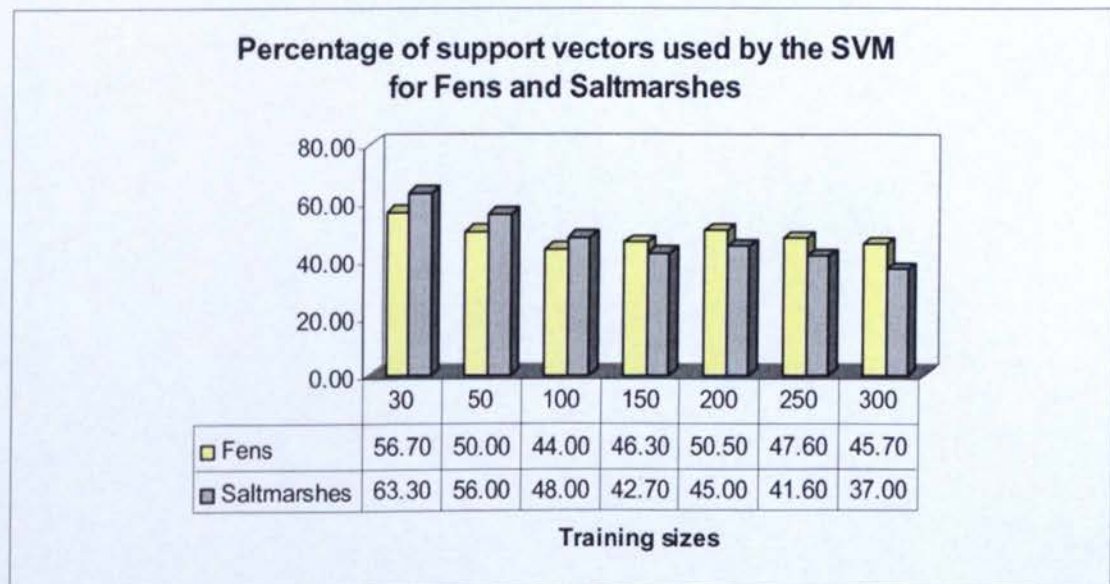


Figure 4.10 Number of support vectors used by the SVM classifier for fen and saltmarsh

As seen in Figure 4.10, it was in the small training sets where the SVM relied more on the data, with 56.70% of the data being considered as support vectors for fen (training size 30) and 63.30% and for saltmarshes (training size 30). After this, the SVM used 50.00% or less of the data as support vectors in the rest of the training sizes. This is a clear advantage over other classifiers where the calculations rely on the whole training dataset, which could in some cases over-fit the classification. Here, all the data that were not used as support vectors could be absent from the training set and the results still would be the same. Graphical representations of these support vectors and the separating hyperplane for each training data set can be found

in Annex C (please note that the decision boundaries shown in the graphics are only indicative).

Finally the ML classification results were used as a benchmark to assess the above SVM classification results. For that the ML results obtained when using a training set of 150 pixels per class were compared against the results obtained by the SVM using a training set of 100 pixels (50 belonging to the class of interest and 50 belonging to the other class).

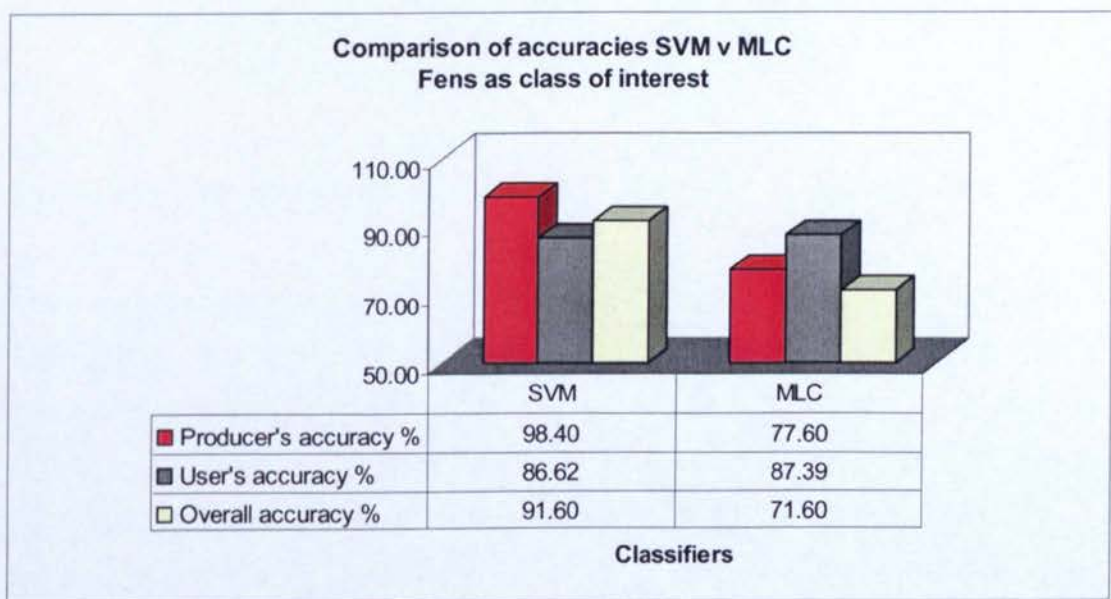


Figure 4.11 Comparison of accuracies SVM v MLC. Fen as class of interest

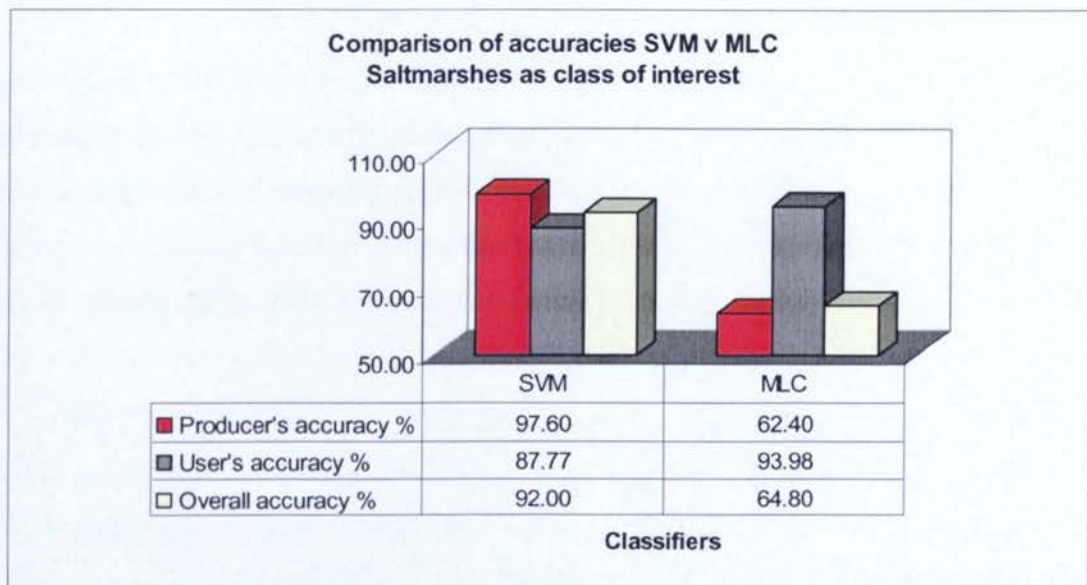


Figure 4.12 Comparison of accuracies SVM v MLC. Saltmarsh as class of interest

The results in Figure 4.11 and Figure 4.12 showed clearly that the SVM obtained much higher overall and producer's accuracies for both fen and saltmarsh as classes of interest. Performing the McNemar's test confirmed that this difference is significant ($Z = \sim 7$) at the 95% confidence interval. However, the user's accuracy was higher for both classes when using the ML classification. This could be due to the fact that the error of commission was spread across the 8 different classes in the image. As already mentioned in Chapter 3, for the purpose of this thesis the producer's accuracy is the most important one as it indicates the capacity of the classifier for identifying the class of interest from all the others. The errors of commission pointed out by the user's accuracy can be easily corrected with the use of ancillary data. Therefore, it can be concluded that the binary SVM classifier is suitable for the application to land cover mapping when focusing on a specific habitat surpassing greatly the accuracy of a standard ML classification. Also, the SVM achieved these results with a fraction of the data used by the ML. In the comparison illustrated above in Figure 4.11 and Figure 4.12 the ML classifier used a training set of 1,200 pixels in total as supposed to the SVM with a training set of only 100 pixels which highlights the capacity of the SVM to obtain high accuracies with very small training datasets.

Having reviewed the principles behind SVMs and demonstrated its suitability to the classification of a class of interest, the following section describes the DT classifier and its application to the case study.

4.3 Decision Trees classifiers

Decision Tree (DT) classifiers have long been popular in machine learning, statistics and other disciplines for solving classification problems. The beginning of DTs dates from work in the social sciences by Morgan and Sonquist (1963) and Morgan and Messenger (1973). Breiman *et al.* (1984) had a decisive influence both in bringing the work to the consideration of statisticians and in proposing new algorithms for constructing trees. At around the same time DT induction was beginning to be used in the field of machine learning (Quinlan, 1979, 1983, Patterson and Niblett 1983, Kononenko *et al.*, 1984, Cestnik *et al.*, 1987). Since then, DT classifiers have been used successfully for a wide range of classification problems but have not been tested in detail by the remote sensing community in spite of their advantages over other non-parametric classifiers (Pal and Mather, 2003). These advantages include the ability to handle data at different scales and do not require assumptions regarding the distributions of the input data (Friedl and Brodley, 1997). In addition they handle nonlinear relations between features and classes, allow for missing values and are capable of handling both numeric and categorical inputs (Fayyad and Irani, 1992, Hampson and Volper, 1986). Therefore they are becoming increasingly popular due to their conceptual simplicity and computational efficiency (Pal, 2002).

There are a few studies that have investigated the use of DT for the classification of remotely sensed data. Lees and Ritman (1991) looked at the use of DTs for mapping vegetation species using Landsat and other spatial data. Byungyong and Landgrebe (1991) used DTs to classify AVIRIS data. Eklund *et al.* (1994) used a DT approach to assess the effect of incremental data layers on groundwater recharge estimates.

Friedl and Brodley (1997) showed that DT algorithms consistently outperformed ML methods when classifying spectral data. They noted, however, that DT algorithms tend to optimise overall classification accuracy at the expense of smaller classes. For this reason, the methods used to assess the accuracy of a classifier must be carefully selected. Pal and Mather (2003) assessed the effectiveness of DT methods for land cover classification and found that the performance of DT classifiers is acceptably good when classifying data from Landsat ETM+ data in comparison with other classifiers such as ANN and ML classifiers. Brown de Colstoun *et al.* (2003) applied DT classifiers to vegetation mapping using multitemporal Landsat ETM+ data. They have also been applied to hyperspectral imagery (Lawrence and Labus, 2003, Yang *et al.*, 2003), incorporating ancillary data with multispectral imagery for increased classification accuracy (Lawrence and Wright, 2001), and change detection analysis (Rogan *et al.*, 2003).

The advantages that DTs can offer to the remote sensing community include an ability to handle data measured on different scales, lack of any assumptions concerning the frequency distributions of the data in each of the classes, flexibility, and capability to handle non-linear relationships between features and classes (Friedl and Brodley, 1997). In contrast to other classifiers such as ANNs, DTs can be trained quickly, and are rapid in execution (Gahegan and West, 1998). Furthermore, they can be used for feature selection/reduction as well as for classification purposes (Borak and Strahler, 1999).

The terminology of trees is graphic. The root is the top node and examples are passed down the tree, with decisions being made at each node until a terminal node or leaf is reached. Each non-terminal node contains a question on which a split is based (see Figure 4.13).

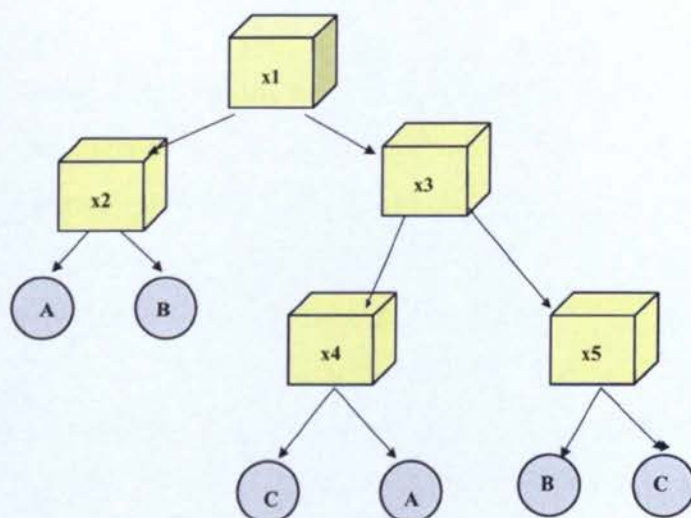


Figure 4.13. Decision tree. A, B and C denote the final classification of training data into these three classes. After Pal and Mather (2003).

A range of factors need to be considered with regards to how the tree is constructed depending on the type of test and the nature of the data that are used to build the DT (Brodley and Utgoff, 1992). These factors include (i) how the estimation procedure handles different types of data (Fayyad and Irani, 1992) (ii) how missing data are handled (Quinlan, 1986) (iii) the partition merit criteria used to measure the goodness of a split (Safavian and Landgrebe, 1991) and (iv) the specific algorithm to perform feature selection at internal nodes when multivariate tests are employed (Kittler, 1986).

Therefore, the design of a DTs can be divided into the following stages:

- a) the appropriate choice of tree structure
- b) the choice of feature subsets to be used at each internal node
- c) the choice of the decision rule or strategy to be used at each internal node.

a) Appropriate choice of tree structure

In general there are two approaches to the design of DTs: Manual design methods and heuristic search methods (Swain and Hauska, 1997). Manual methods use statistics such as the mean vector and covariance matrix, which are calculated for all classes. Afterwards a graph is derived in which the means and variances for all the classes are plotted for each feature. This graph is called a coincident spectral plot. It is often possible to estimate suitable decision boundaries from this graph such that all classes are separated in a number of decision steps. This method is not suitable when two or more features are to be used in a given stage of the tree because the graph does not show how the interactions between features can be used. Also it is not suitable if the data are not normally distributed, thus making it difficult to estimate the covariance matrices in an unbiased way. Therefore, if the difficulty of discriminating the classes requires the use of a combination of several features, the manual design approach based on the spectral plot is severely limited (Pal, 2002).

With the heuristic approach it is assumed that a training data set consisting of feature vectors and their corresponding class labels is available. The DT is then constructed by recursively partitioning the training dataset into purer, more homogenous, subsets on the basis of a set of tests applied to one or more attribute values at each branch or node in the tree (Figure 4.13). At each node a decision algorithm decides how the data are split. DT classification algorithms can be distinguished according to whether these algorithms are homogeneous or heterogeneous (Friedl and Brodley, 1997). Traditional approaches are based on homogeneous classification models for which a single algorithm is used to estimate each split. Two types of DTs that are based on such homogeneous classification are: (i) univariate DTs and (ii) multivariate DTs. A hybrid hypothesis space on the other hand would allow the combination of different algorithms but they are only applied in complex classification problems (Friedl and Brodley, 1997).

(i) Univariate Decision Trees (UDT)

In a Univariate Decision Tree (UDT) the decision boundaries at each node of the tree are defined by a single feature of the input data. The data are split into two or more subsets on the basis of a test of a single feature of the input data and each test is required to have a discrete number of outcomes (Friedl and Brodley, 1997). The specific values of the decision boundaries in a UDT are estimated empirically from training data. A boolean test of the form $\mathbf{x}_i > b$ is estimated at each internal node where \mathbf{x}_i is a feature in the data space and b is a threshold in the observed range of \mathbf{x}_i . The value b can be estimated by using some objective measure that maximises dissimilarity or minimises similarity of the descendent nodes. (Friedl and Brodley, 1997). As each test in UDT is based on one of the input variables, it is restricted to representing that variable axis, as show in Figure 4.14.

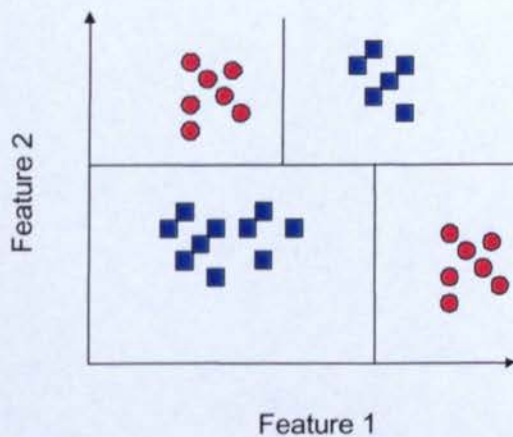


Figure 4.14. Axis parallel decision boundaries of a univariate decision tree. Based upon Friedl and Brodley (1997).

(ii) Multivariate Decision Trees (MDT)

Multivariate Decision Trees (MDT) are similar to UDTs except that the splitting test at each node is based on more than one feature of the input data. The test at each node has the form:

$$\sum_i a_i x_i \leq c \quad (1)$$

Where x_i represent the features in the data space, a is the vector of coefficients of the linear discriminant functions, and c is the threshold value. The decision boundaries are represented graphically in Figure 4.15.

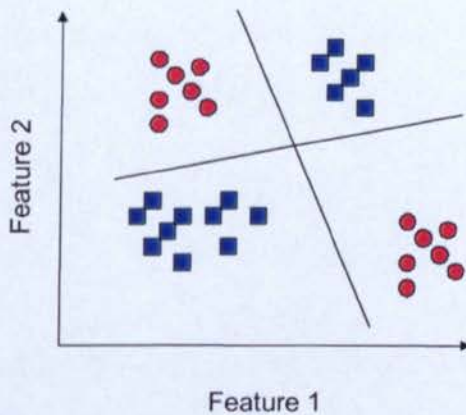


Figure 4.15. Decision boundaries for a multivariate decision tree classifier. Based upon Friedl and Brodley (1997).

However, the higher complexity of MDT implies that different algorithms have to be used to estimate the splitting rule at internal nodes which can vary depending on the type of data and classification problem. Also several different algorithms have to be performed for feature selection at each node. This feature selection is made on the basis of the data in a particular node and it does not apply a uniform set of features for the entire tree. In general MDT are more difficult to interpret than UDT and should be applied only when necessary (Friedl and Brodley, 1997).

Regardless of the structure of tree chosen (univariate, multivariate or hybrid), at each node in the tree the DT tries to minimize the number of features included in the test. The two basic approaches are Sequential Backward Elimination (SBE) and Sequential Forward Selection (SFS) (Broadley and Utgoff, 1995). Sequential Backward Elimination is a top down method that starts with all the features and tries to remove the feature that will cause the smallest decrease of some partition-merit criterion that reflects the amount of classification information conveyed by the feature (Kittler, 1986). Sequential Forward Selection (SFS) is a bottom up search that starts with zero features and tries to add the feature that will cause the largest increase of some partition-merit criterion.

Of the two, the most popular method is the Sequential Backward Elimination. There are two choices that must be made to implement the SBE algorithm: the choice of (i) partition-merit criterion and (ii) the stopping and pruning criteria. A partition-merit criterion may measure the accuracy of the test when applied to the training data or measure the entropy, as with the Information Gain Criteria (Quinlan, 1986) and the Gini index (Breiman *et al.*, 1984). The stopping criterion determines when to stop eliminating features from the linear combination test.

(i) The Partition merit criterion.

The Information Gain Criteria was developed by Quinlan (1986) in the following way. When applied to a set of training objects, $\text{info}(T)$ gives the average amount of information needed to identify the object of a class in T . This amount is also known as the entropy of the set T .

$$\text{info}_x(T) = \sum_{i=1}^n \frac{|T_i|}{|T|} \cdot \text{info}_x(T_i) \quad (2)$$

The quantity

$$gain(\mathbf{x}) = info(T) - info_{\mathbf{x}}(T) \quad (3)$$

measures the information that is gained by partitioning T in accordance with the test \mathbf{x} . The *gain criterion* (Quinlan, 1993) selects a test to maximise this information gain. However, the gain criterion has one significant disadvantage in that it is biased towards tests with many outcomes. The *gain ratio criterion* (Quinlan, 1993) was developed to avoid this bias. The information generated by dividing T into n subsets is given by

$$split\ info(\mathbf{x}) = - \sum_{i=1}^n \frac{|T_i|}{|T|} \cdot \log_2 \left(\frac{|T_i|}{|T|} \right) \quad (4)$$

The proportion of information generated by the split that is useful for classification is

$$gain\ ratio(\mathbf{x}) = gain(\mathbf{x}) / split\ info(\mathbf{x}) \quad (5)$$

This compares with CART's *impurity function* approach (Breiman *et al.*, 1984), where impurity is a measure of the class mix of a subset and splits are chosen so that the decrease in impurity is maximised. This approach led to the development of the Gini index (Breiman *et al.*, 1984).

$$i(\mathbf{p}) = \sum_{i \neq j} \mathbf{p}_i \mathbf{p}_j = 1 - \sum_j \mathbf{p}_j^2 \quad (6)$$

The impurity function approach considers the probability of misclassifying a new sample from the overall population, given that the sample was not part of the training sample, T . This probability is called the *misclassification rate* and is estimated using either the re-substitution *estimate* (or training set accuracy), or the *test sample estimate* (test set accuracy). The node assignment rule selects i to minimise this misclassification rate. In addition, the Gini index promotes splits that minimise the

overall size of the tree (Evans, 1998). Gini attempts to separate classes by focusing on one class at a time. It will always favour working on the largest or the most "important" class in a node. This means that when classifying the binary DT can identify the class of interest out of the classes present in the classification. This is extremely important in this thesis as the focus of the classification is on one specific class. Therefore, the logical choice for the present case study will be choosing a DT with the Gini index splitting criteria.

(ii) The stopping and pruning criteria

The result of dividing the training data into subsets may produce a very large and complex tree. Also, fitting a decision tree until all leaves contain data for a single class may over-fit to the noise in the training data, as the training samples may not be representative of the population they are intended to represent. According to Breiman *et al.* (1984) there are two ways in which a DT classifier can be modified to address these problems:

1. Deciding not to divide a set of training data any further (also called stopping criterion), and
2. To remove retrospectively some part of the tree structure built by recursive partitioning (known as pruning).

The first approach looks at the best way of splitting a dataset and to assess the split from the point of view of a factor such as information gain or error reduction. If this assessment falls below some threshold, the division is rejected. The problem with this approach is to identify an acceptable stopping rule (Breiman *et al.*, 1984). If the threshold value is too high it can terminate division before the benefits of subsequent splits become evident, while too low a value results in little simplification of the tree. In the second

approach, the tree is allowed to grow in full. Then, this over-fitted tree is pruned. This method needs more computation in building parts of the tree that are

subsequently discarded, but this cost is offset against benefits due to more thorough exploration of possible partitions.

4.3.1 Classification of a habitat of interest using the DT classifier. Case study.

Of all the DTs available, it was decided to use the CART method implemented by Salford Systems to perform the classification. This method is based upon Breiman's approach to DT classifiers and has been extensively applied in different disciplines such as meteorology (Burrows *et al.*, 2001, Dunsmuir *et al.*, 2003, Firth *et al.*, 2005), clinical studies (Chang *et al.*, 2002, Bolt *et al.*, 2004, Chavanet *et al.*, 2004), ecological analysis (Fabricius and De'ath, 2000, Feicht *et al.*, 1998, Koh and Sodhi, 2004). It has also been tested with remote sensing data (Hansen *et al.*, 1996, Frield and Brodley, 1997). The results of this work showed that CART's DT performed comparably to other established classifiers such as ML and ANNs. They also showed that DTs have advantages in terms of feature selection and handling missing data (Frield and Brodley, 1997).

The reason for choosing CART was that it uses the Gini Index as the main splitting criteria. As mentioned earlier, one very important characteristic of this index is that the class of interest can be pre-defined and the splitting rule will prioritise the classification of this class over any other. This was extremely important in this case study and a clear advantage of using this classifier in an OVA binary classification scheme. Another advantage of CART is that the testing and selection of the optimal tree are an integral part of CART. It calculates the classification for both multivariate and univariate trees. The best linear combination found by CART for a multivariate approach is added to the set of possible univariate tests and the best of this new set is chosen as a test at the current node. In terms of pruning, CART introduces the notion of over-growing trees and then pruning back; this idea guarantees that important tree structures are not overlooked by stopping too soon. In other words, CART tries to

achieve the global solution instead of stopping at a local minima as explained in the previous section.

The range of training sets used to train the DT classifier were the same as those used previously to train the SVM. This is, training sets of 30, 50, 100, 150, 200, 250 to 300 pixels where each set contained the pixels from the previous one so that $30 \in 50 \in 100 \in 150 \in 200 \in 250 \in 300$ and the same testing set as used before with 250 pixels in the same 50/50 proportion. In previous studies regarding training set sizes impact on DT classifiers, Oates and Jenson (1997) suggested that increasing the number of training data only affected the size of the tree but had a little effect on the classification accuracy. However, Pal and Mather (2003) showed that the size of training datasets affected the final accuracy of the DT and that this accuracy increased with increase of training data up to a point where the DT got stabilised and adding more training data did not increase the accuracy.. The results obtained for the present case study were the following:

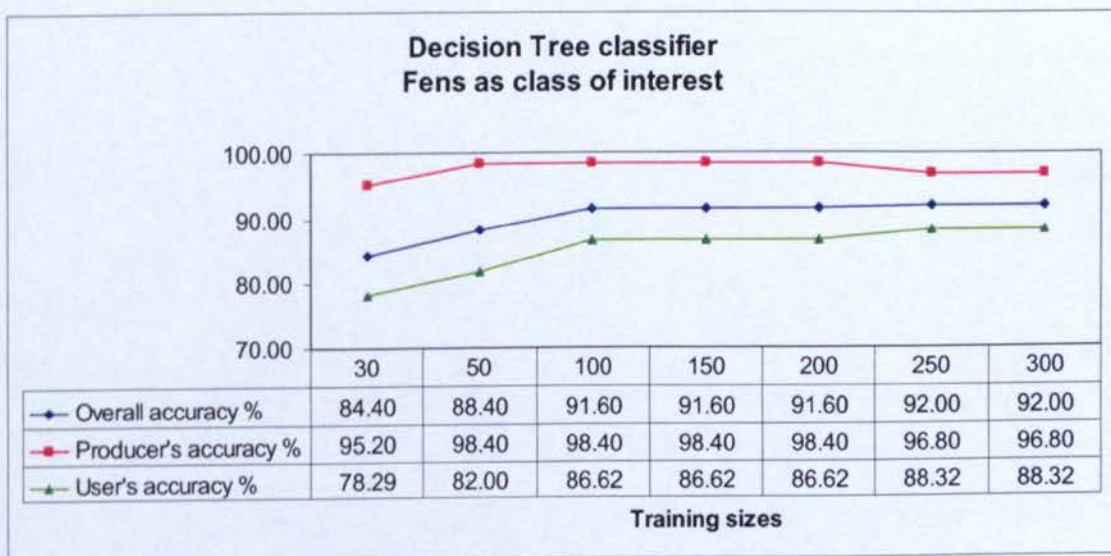


Figure 4.16 Binary Decision Tree classification. Fen as class of interest

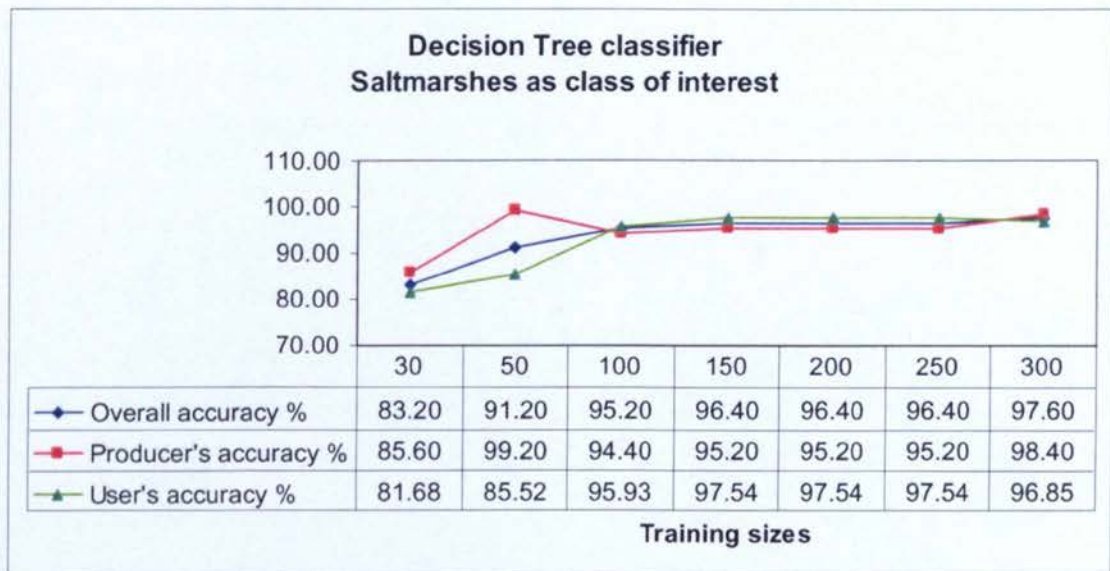


Figure 4.17 Binary Decision Tree classification. Saltmarsh as class of interest

With regards to different training sizes the results agreed with those of Pal and Mather (2003). For fen the DT achieved a high accuracy using a training set of 100 pixels which did not get significantly higher after that. With saltmarsh as the class of interest the same trend was observed, although here the accuracies obtained were much higher with 97.60% overall accuracy for the training set with 300 pixels as supposed to 92.00% for the same amount of training data for fen. This could be due to the higher separability of the class saltmarsh from all the other classes. As with the class fen, the DT achieved high accuracies with sizes as small as 100 pixels. Confusion matrices for each of the training sizes can be found in Annex A.

With regards to the tree structure, the higher accuracy of saltmarsh was obtained by a more complicated tree than the one for fen. A detailed account of trees for different training sizes can be found in Annex D. Figure 4.18 below compares the trees obtained by the classifiers with a medium training size of 150 pixels.

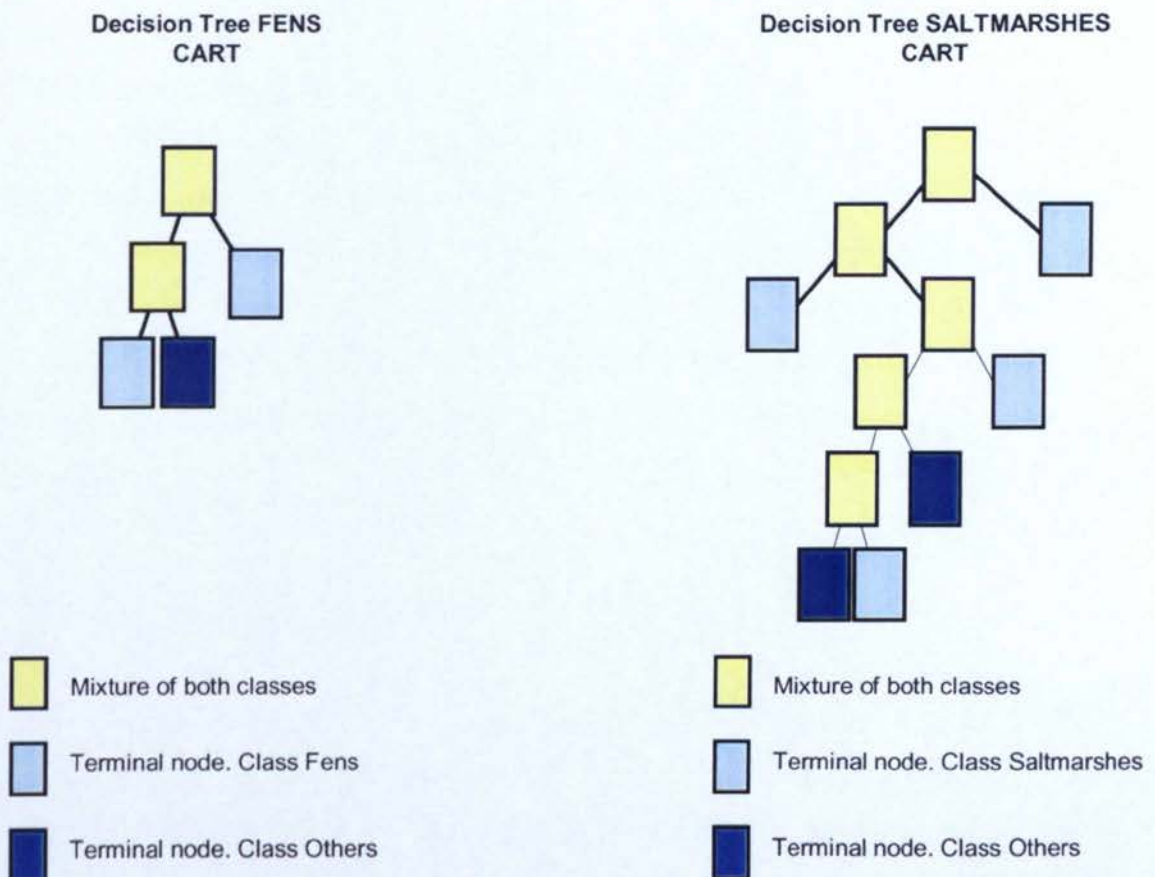





Figure 4.18 Decision Tree structure resultant with the binary classification in CART fen and saltmarsh

The difference between both trees could be due to the different spectral signature of each class of interest and the degree of separability between that class and all the others. As seen in previous sections, the DT aims to refine the training sample into subsets which have only a single class. CART performs the cost-complexity pruning automatically and produces the optimal tree. However if the researcher thinks that the tree is still quite complex this can be pruned back further. As an experiment it was decided to prune again the saltmarsh's tree. The result obtained for the pruned tree can be seen in Figure 4.19 below.

-  Mixture of both classes
-  Terminal node. Class Saltmarshes
-  Terminal node. Class Others

The overall accuracy for this new tree was slightly lower than the original (93.20% as supposed to 94.40% respectively). This confirms the fact that pruning a DT can cause it to misclassify more of the training data. The reason for this is that the leaves of the pruned tree will not automatically contain training data from a single class. Instead, there will be a class distribution specifying, for each class, the probability that a training data sample at the leaf belongs to that class (Pal and Mather, 2003).

- (i) Accuracy increased with increase of training data up to a point. For example for fen the accuracy seemed to get to a maximum at 100 pixels and there was no major increase using bigger sizes. For saltmarsh this was also the case.
- (ii) The DT classifier did not need a great amount of training data to be effective. In this particular case 100 pixels seemed to be sufficient.

Page 121

ML classification using a training set of 150 pixels per class with those obtained by the DT classifier using a training set of 100 pixels (50 pixels belonging to the class of interest and 50 belonging to the other class). The results were:

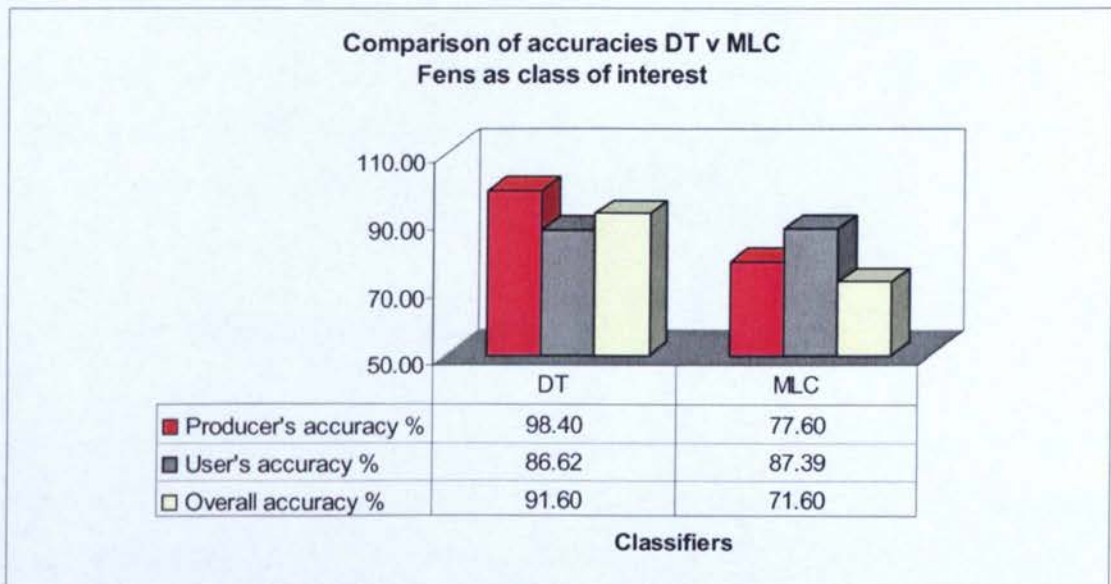


Figure 4.20 Comparison of accuracies DT v MLC. Fen as class of interest

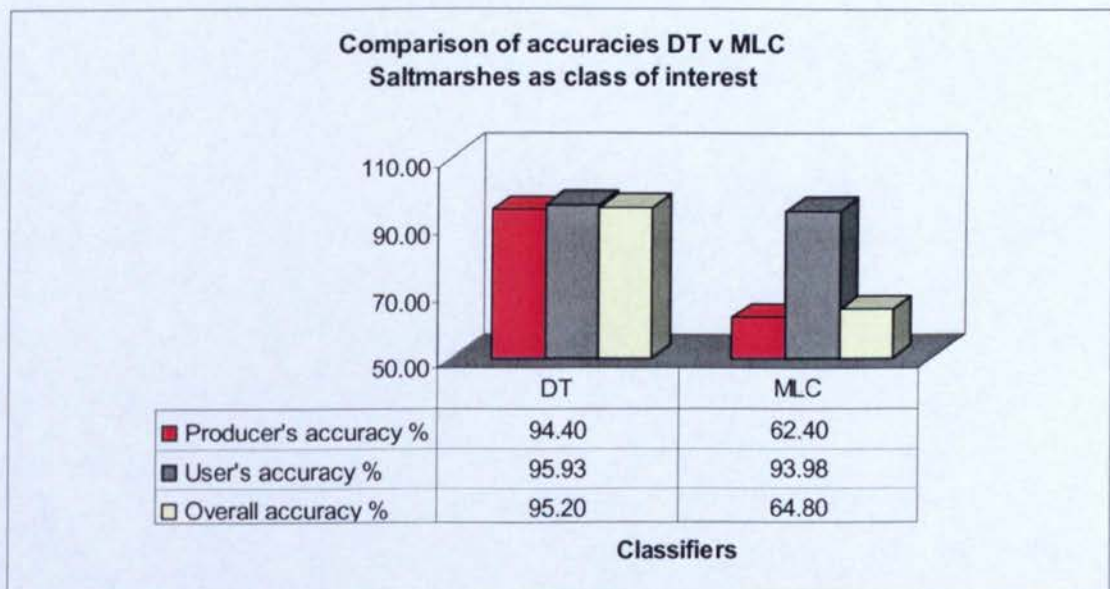


Figure 4.21 Comparison of accuracies DT v MLC. Saltmarsh as class of interest

The accuracies obtained by the DT classifier were significantly higher than those obtained by the multiclass ML classifier ($Z = \sim 7$) at 95% confidence interval. For fen as class of interest the ML classification still obtained a higher user's accuracy than the DT classifier (see Figure 4.20). However, for saltmarsh this was not the case, and the DT obtained a user's accuracy of 95.93% as supposed to 93.98% obtained by the ML classification (Figure 4.21). Also it is once again important to highlight that the ML classifier used a training set of 1,200 pixels in total as supposed to the DT with a training set of only 100 pixels.

4.4 Summary and Conclusions

Satellite remote sensing has enormous potential as a source of land cover information and in particular, for monitoring land cover and its dynamics at the scales associated with the demands of the EU Habitats Directive. In standard statistical supervised image classifications, the aim is to maximize the overall probability that a pixel is allocated to a class correctly. This requires that each class within the area to be mapped is included in the analysis to satisfy the assumption of an exhaustively defined set of classes. This approach treats all classes equally, including those of no interest, rather than focus on a class that is of real interest.

The results obtained by the classifiers showed that significant increases in accuracy can be achieved through the use of binary classifications that aim to separate the class of interest from all others. In the classifications performed using the binary SVM and DT, the classes of interest (fen and saltmarsh) were classified with much higher accuracies than those obtained by the ML classifier. This clearly highlights the potential of these classifiers for land cover mapping of specific habitats that may aid environmental monitoring as part of obligations linked to the EU Habitats Directive. In particular, the use of these classifiers could reduce problems associated with having to exhaustively define and train all classes. As demonstrated in this chapter a conventional statistical classifier such as a multiclass ML requires a large

training sample for each class present in the image in order to exhaustively define each class and obtain acceptable accuracies. However, the reliability of the results obtained with this classifier declines when the frequency distribution of the data departs from normality. It should also be noted that all features are used to discriminate between classes, rather than the minimum effective set.

For the binary classifiers, high accuracies were obtained with only 50 pixels per class. Furthermore, even the accuracies obtained by these classifiers with only 15 pixels per class surpassed the accuracy obtained by the ML classifier using a training set of 1,200 pixels. The accuracies were higher when selecting the class saltmarsh as the class of interest which could indicate that the separability between this class and all the other classes is higher than the one for fen and the other classes.

Furthermore, both classifiers seemed to show a high error of commission. This could be important in some cases because the EU Habitats Directive requires that relevant authorities carry out very precise mapping and monitoring of protected habitats. This problem could be addressed if other ancillary data and field surveys are available to correct the classification in the areas that have been misclassified as the class of interest. In conclusion, the ability to focus on the class of specific interest and to reduce the amount of effort wastefully directed on other classes may help realise the potential of remote sensing as a viable source for land cover mapping of specific protected habitats to the relevant authorities. This chapter has demonstrated that both SVM and DT classifiers have the ability of optimising the training process by using very small training data sets in order to achieve high classification accuracies. This optimisation could be further refined in the case of SVM as only the data selected as support vectors are needed for the separation of the class of interest from all the other classes. Furthermore, the suitability of SVM and DT binary classification for the classification of a habitat of interest under the requirements of the EU Habitats Directive has also been confirmed.

However, although binary classifications using SVM and DT classifiers can be successfully used for classifying a class of interest, training data for the ‘other’ class are still required. It could be that in some cases these data are very difficult to acquire. In this sense, the following chapter will address the problem of the classification of a particular class by focusing solely on this class of interest by using one-class classification methods.

5 One-class classification methods and their application to one class land cover mapping

"It's not that I'm so smart, it's just that I stay with problems longer".

Albert Einstein

The previous research chapter investigated a binary classification approach in order to classify a habitat of interest from all the other habitats present in an image. However, as already mentioned, one disadvantage of binary classifiers is that they still need to have some information about the other classes present in the image in order to amalgamate them into the "other" class. It could be that in some instances these data are not available or very difficult to acquire. Also, it is reasonable that if the focus of the research is one class of interest, all the classification efforts should be concentrating on this class. These issues could be addressed with the application of one-class classifiers which only require training data from the class of interest in order to classify it accurately. This would also be in line with the EU Habitats Directive requirements and the need of the relevant authorities to optimise their resources in order to comply with these requirements.

The problem of remote sensing classification and land cover mapping when focusing on a class of interest is very similar to those problems encountered in pattern recognition. It is virtually impossible to have a description of all the possible data that a classifier could encounter in real life situations. For example, it is impossible

to have a description of all the document categories in document classification or all the possible textures in texture segmentation. However, it is possible to describe the category or texture of interest and separate this from all other possible occurrences. Likewise, in remote sensing, there is a limit to the number of classes that can be differentiated in an image and the amount of data that are possible to be collected for each of them. Sometimes the lack of ground data or appropriate spatial and spectral resolution makes it impossible to exhaustively define all the classes present in the image. A method that allows concentrating on the description of a class of interest regardless of how many other classes are present in one image is therefore extremely attractive.

One-class classification has proved very valuable to solve these issues in areas of pattern recognition such as document classification (distinguishing one specific category from other categories) (Manevitz and Yousef, 2001), texture segmentation (distinguishing one specific texture from other textures) (Tax and Duin, 2002), and image retrieval (retrieving a subset of images based on the similarity between given query images) (Lai *et al.*, 2002). It has also recently being used in ecological models coupled with GIS to predict the potential distribution of a new forest disease called Sudden Oak Death in California (Guo *et al.*, 2005). However, it has not yet been applied to remote sensing image classification and land cover mapping (as far as the author is aware).

Therefore, taking into account all the above, the specific objective of this chapter is to evaluate the potential of one-class classification for the mapping of a specific habitat of interest using remote sensing data and will be structured as follows:

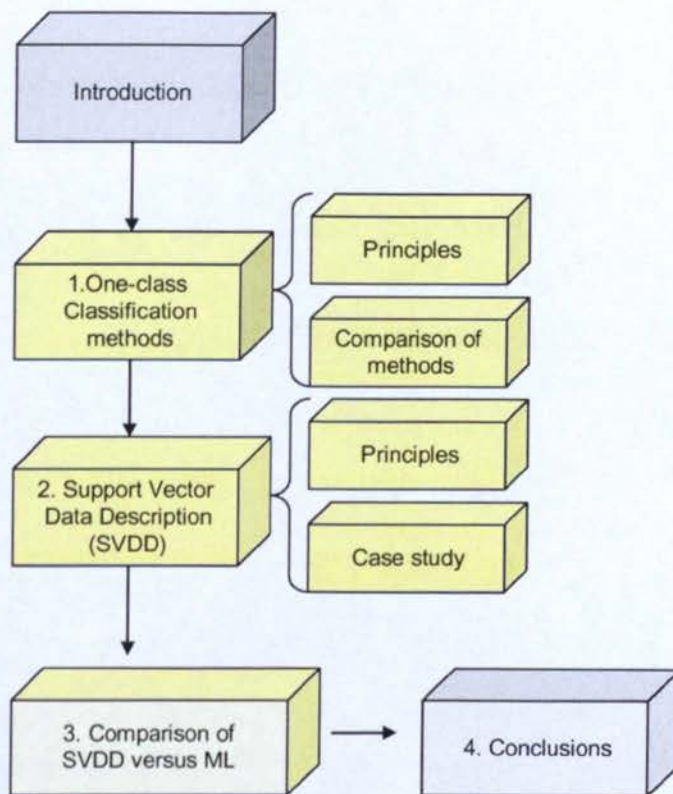


Figure 5.1 Chapter 5 structure

Following Figure 5.1, the first section of this chapter will describe the principles of one-class classification methods and compare the different methods in order to assess which one is most appropriate for ultimately achieving the aim of this thesis. Following the conclusions of this section, the SVDD one-class classifier is investigated and applied to the case study. Finally, these results are compared against those obtained by the standard ML classifier and conclusions about the applicability of one-class classifiers to remote sensing land cover classification will be drawn.

5.1 One-class classification. Principles and methods

The term one-class classification is believed to originate from Moya (Moya *et al.*, 1993), but also outlier detection (Ritter and Gallegos, 1997), novelty detection (Bishop, 1994, Markou and Singh, 2003) or concept learning (Japkowicz *et al.*, 1995) are used. These different terms normally derive from the different applications of this type of classification. Within this thesis the term used is one-class classification as the aim of this research is to classify one habitat or class of interest.

As mentioned in Chapter 1, one-class classification is a special type of binary classification problem with two classes: (i) the target and (ii) the outlier class.

(i) Target class refers to the class of interest and it is assumed to be sampled well and that enough training data are available. It does not necessarily mean that the sampling of the training set is done completely but that enough samples are provided to characterise the class in the feature space. For the purpose of the present thesis, the target class and class or habitat of interest are used as equivalent terms.

(ii) Outlier class refers to any other class different from the class of interest. In the binary classification problem this would be the equivalent to the 'other' class. It can be sampled very sparsely or can be totally absent. There could be many reasons for this. For example, in remote sensing it might be that this class is very hard to measure, or it might be very expensive to do the measurements on these types of habitats, or that the outliers (other classes) are so abundant that a good sampling is not possible or simply, the attention is focused upon a class of interest and the sampling of outliers is considered wasteful.

In one-class classification only target class data (ω_1 in Figure 5.2) are used in the training stage. However, in the testing and validating stage the classifier will encounter outlier data that were not present in the training stage. This means that the

classifier has to have the capacity to distinguish if the data in a testing set belong to the target class or are unknown and as such belong to the outlier class. The success of the application of one-class classification depends upon the type of method used and the statistical properties of the target class data (Markou and Singh, 2003). Consequently, there are different approaches to one-class classification. The main ones are reconstruction methods, density methods and boundary methods. Reconstruction methods and density estimation are not always easy to achieve, as they rely on the density and description of the data and as such require a large amount of training data as will be explained later in this section. To deal with this problem, boundary classifiers concentrate on fitting a separating boundary around the target class. The problem of finding a boundary in one-class classification is harder than the problem of two-class classification. The reason for this is that in two-class problems the decision boundary to separate one class from another is supported by samples of both classes ω_1 (target class) and ω_2 (other classes) as seen in Chapter 4. However, in the case of one-class classification, only data from the target class ω_1 are available and therefore only one side of the boundary is supported. It is therefore difficult to decide how tight the boundary should be and the nature of such boundary (see Figure 5.2) (Tax and Duin, 2001).

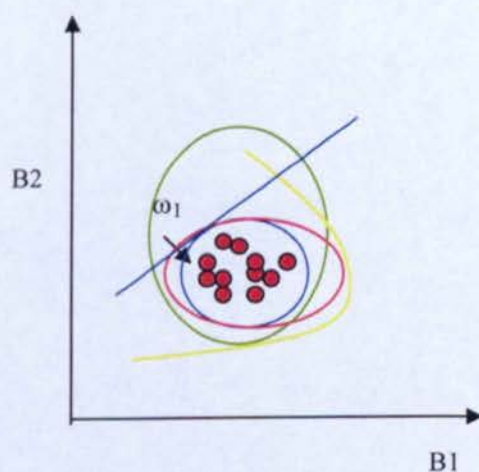


Figure 5.2 Different possibilities of boundary fit around the target data being B1 and B2 two input variables

In that sense, the support vector novelty detectors (SVNDs) have been recently developed and they are largely based upon the theory of statistical machine learning and support vector machines. The first SVND was proposed by Tax and Duin (1999), to estimate a boundary in the form of a sphere that contains all the target data patterns within the smallest radius. They called it Support Vector Data Description (SVDD). Another alternative SVND is proposed by Schölkopf *et al.* (2001). Instead of a sphere, a function is estimated to separate the region of normal data patterns from that of outliers with maximal margin, thus detecting the outliers from the normal data patterns. As demonstrated in Schölkopf *et al.* (2001), the two SVNDs are very similar. The advantage of Tax's approach is that a closed boundary is obtained around the target class. This characteristic could be important in the case of remote sensing when a class of interest could be surrounded by all the other classes in the feature space and therefore the more effective separation of this class from all the others would be a closed boundary.

The following section describes the different one-class classification methods and compares their performance when applied to remote sensing classification using data from the case study.

5.1.1 One-class classification methods

In the area of remote sensing it is practically impossible to train a classifier on all possible classes that it is likely to come across. In this sense, one-class classification is a particularly suitable approach as only data from one class are used to train the classifier (Markou and Singh, 2003). As mentioned in the introduction to this section, the three main methods in order to approach the problem of one-class classification are: (i) density methods, (ii) reconstruction methods and (iii) boundary methods.

(i) Density methods

Density methods are based upon the estimation of the density of the target data. The three main density methods used for one-class classification are Gaussian, Mixture of Gaussians and Parzen density.

The Gaussian is a simple method that imposes a normal density model of the data. Chow (1970) studied the trade-off between recognition rate for the target class and proportion of data rejected based on a threshold determined by the user. Developing the work of Chow, Hanssen *et al.* (1997) introduced the concept of classifier confidence in its decisions. Fumera *et al.* (2000) also built upon Chow's research by suggesting the use of multiple thresholds, one for each class. However, the Gaussian density model is very rigid because in most of the cases real data are not normally distributed and according to Markou and Singh (2003) this technique has therefore little importance in practical applications

To remedy some of the problems with the Gaussian approach, the normal distribution is extended to a Mixture of Gaussians (MoGs) which was defined by Bishop (1995) as a combination of normal distributions. Roberts and Tarassenko (1994) developed a method in which the number of gaussians was defined beforehand by the user and the means and covariance could be estimated by an algorithm that ensured that every gaussian had seen each sample in the training data at least once. This method is very similar to those of Barnett and Lewis (1994), Bishop (1994), Tarassenko (1995), Parra *et al.* (1995), Desforges *et al.* (1998), Brotherton *et al.* (1998), Yeung and Chow (2002) and others. The main disadvantage with the use of MoGs is that normally a large number of samples are needed to train the model (Markou and Singh, 2003).

Finally, Parzen density estimation (Parzen, 1962) uses diagonal covariances matrices for its calculations. A good description of the model depends on the representativeness of the training set. During testing distances to all training objects

have to be calculated and sorted which limits greatly the applicability of the method. Studies applying this method can be found in Yeung and Chow (2002) Bishop (1994), Tarassenko (1995) Tarassenko *et al.* (1999), Desforges *et al.* (1998), Brotherton *et al.* (1998) and others. But like the other two methods it presents two main drawbacks: (i) in general, a large number of samples are required and (ii) they will not be very efficient with training data which do not represent the complete density distribution (Tax and Duin, 2004).

(ii) Reconstruction methods

These methods use prior knowledge about the data and make assumptions about the generating process therefore producing a model of the data. Most of them assume the clustering characteristics of the data or their distribution in subspaces. There are quite a few examples of reconstruction methods: (i) clustering methods such as K-means clustering, Self Organising Maps (SOM), (ii) Principal Components Analysis (PCA), Mixture of Principal Components Analysis, (iii) Diabolo Networks and Autoencoder Networks.

Clustering approaches partition the data into a number of clusters where each data point is assigned a degree of membership to each of the clusters. In the K-means method the degree of membership is thresholded and a new data point belongs or not to a cluster depending on the threshold. Outliers can be detected when a point does not belong to any of the clusters (Markou and Singh, 2003). SOM were proposed by Kohonen (2001) and it is an unsupervised approach. Some sort of cluster membership value is thresholded to determine whether a sample belongs to a cluster or not. However, in all the clustering methods the Euclidean distance is used in the definition of the error and the computation of the distance. Therefore they are very sensitive to scaling of the features.

Another reconstruction method is Principal Components Analysis (PCA). PCA tries to capture the variance of the data as best as possible. For that PCA is a technique

that stretches the range of the data along an orthonormal axis. The preserved direction in which the data are distributed is called an eigenvector of the transformation and the associated amount by which it has been stretched is an eigenvalue. The PCA is relatively sensitive to the scaling of the features as this directly influences the feature variances. To get a more efficient performance, PCA can be extended to a mixture of PCAs. This introduces several principal components bases and a mixture of probability models is optimised. However the number of free parameters in this method is very high and consequently the sample size for training is also very high (Markou and Singh, 2003).

Finally, auto-encoders and diabolos networks are neural network approaches to describe the distribution of the data. The difference between the two of them comes from the number of hidden layers and the sizes of the layers. The number of free parameters in both of them is very high and they inherit the same problems as conventional neural networks requiring a predefined network structure that can vary in complexity and therefore make their use and implementation difficult (Tax, 2001).

In conclusion, reconstruction methods also present several drawbacks. Some of them are quite sensitive to scaling of the features (clustering methods), others also need a great amount of training data in order to be properly trained (PCA) and finally the diabolos networks and autoencoders need to predefined a network structure that can be quite complicated. To try and address these problems, boundary methods will use a totally different approach.

(iii) Boundary methods

In boundary methods only the boundary around the target set is optimised. Therefore, a smaller sample size than for density or reconstruction methods is required. Also, due to their focus on the boundary, the threshold on the output is always obtained in a direct way and consequently the outputs of boundary methods are not interpreted as a probability (Tax, 2001). The main boundary methods are the K-centres method, the

nearest neighbour (NN) method and the support vector data description (SVDD) method.

K-centre method covers the dataset with k small balls with equal radii. The ball centres are placed in training objects. It resembles the K-means method but they differ in the error which is minimised. K-centre tries to optimise the centres and radii of the balls to accept all the data. The method is very sensitive to outliers in the training set but it will work well when clear clusters are present in the data.

In the NN method a test object or pixel z is accepted depending on its distance to the nearest training data points. A hypersphere is centered on the test object or pixel z and the volume grows until it captures k objects from the training set. The number k is determined beforehand by the researcher. The problem of this technique is that for large datasets a large number of calculations have to be performed (Markou and Singh, 2003). It has no free parameters so it relies completely on the training dataset. Because it uses distances to the training samples it is scale sensitive. Also it can reject parts of the feature space which are within the target distribution.

As an alternative to the above one-class classifiers, a SVDD was proposed by Tax and Duin (1999), where a boundary in the form of a sphere contains all the target data within the smallest radius and all the outliers will lie outside this sphere. These outliers are identified by calculating the distance of a new object z to the centre of the sphere. Another boundary method was proposed by Schölkopf *et al.* (2001). Here a function is estimated to separate the region of normal data patterns from that of outliers with maximal margin. As said before, both classification methods are very similar. The only difference is that the SVDD always finds a closed boundary around the target class.

In order to assess the performance of each of the above one-class classifiers, Tax (2001) compared all the above methods using different artificial datasets. He arrived to the following conclusions:

- 1) When the sample sizes are very small the density and reconstruction methods perform poorly and the best performing methods are the boundary ones. For larger samples most of the methods seem to perform well.
- 2) When the distribution of the target dataset and testing dataset are the same the best classification methods are the density methods. In particular, the Parzen density method appears to perform very well on the considered data.
- 3) When the training data distribution is different from the testing data distribution the boundary methods are preferred and the SVDD achieves very good generalization. This is quite important as it is the most likely situation with real data and the SVDD is still able to find a good description of the data.

All the above shows that the SVDD has clear advantages over other one-class classifiers when the sample sizes are small and when the test data distribution is different from the training data and this would make it especially suitable for remote sensing applications. However, before concentrating on the advantages of the SVDD it was decided to assess the performance of the other one-class classification methods using remote sensing data to see if the results will share the same conclusions as those described above.

5.1.2 Comparison of one-class classification methods

The Data Description MATLAB toolbox named DD_tools version 1.12 developed by Tax (2004) was used for this experiment. To compare the different one-class classifiers the training data collected for the class of interest were used and these training data consisted of pure pixels. This means that the data were very clearly defined in the feature space and had limited complexity. Also the dimensionality was very low with only 2 variables. These characteristics agreed with those of the

training data set created by Tax (2001) where the data contained a maximum of 3 clusters and the dimensionality ranged from 2 to 10. Therefore, it was decided to use the same free and user defined parameters than those used by Tax (2001) to avoid exhaustive experimentation. The purpose of the test was to give an indication of how well each method performed using remote sensing data even if the parameters had not been entirely-tuned.

Therefore, the parameters used for the experiment were as follows: the number of clusters in the K-means and K-center methods was set to 10, which is enough to find 2 or 3 clusters in these experiments. The dimensionality of the SOM was set to 2. Because the SOM uses $k = 10$ neurons for each dimension in the map, in total 10×2 neurons are required. The auto-encoder network had 10 hidden units. For the SVDD a gaussian kernel was used with $C = 10$.

The experiment was carried out with a training set of 100 pixels (bigger than the minimum size recommended (10-30p) to ensure that the classifiers had enough data to achieve a description of the target class) and 2 input variables (NDVI and ETM+ Band 2). The testing set was the same one that is being used throughout this thesis for all the experiments consisting of 250 pixels of which 50% belong to the target class and 50% to the 'other' class. These experiments were performed for both fen and saltmarsh as classes of interest. The results were as follows:

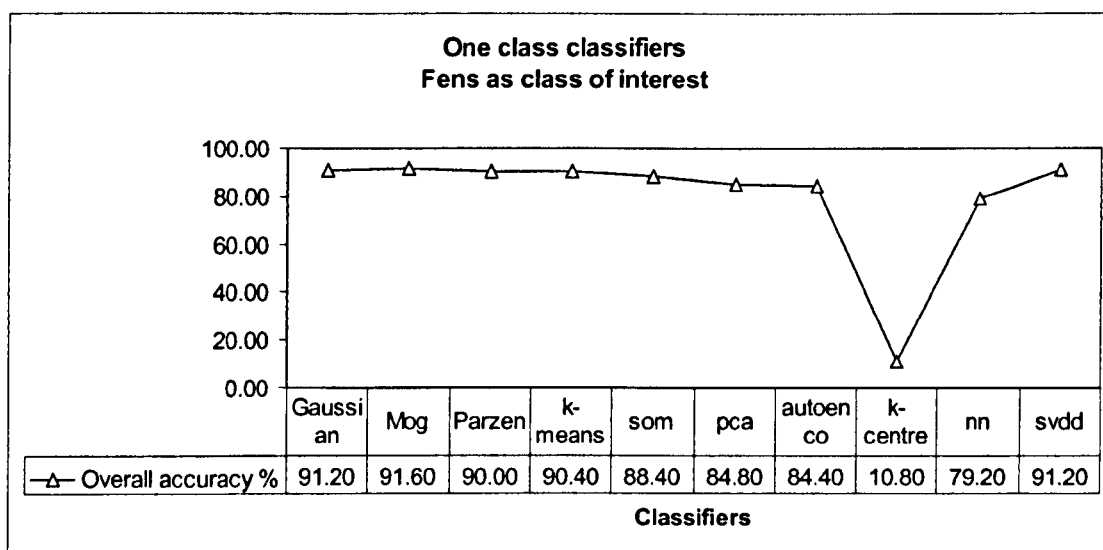


Figure 5.3 Overall accuracy results for the one-class classifiers Fens as class of interest

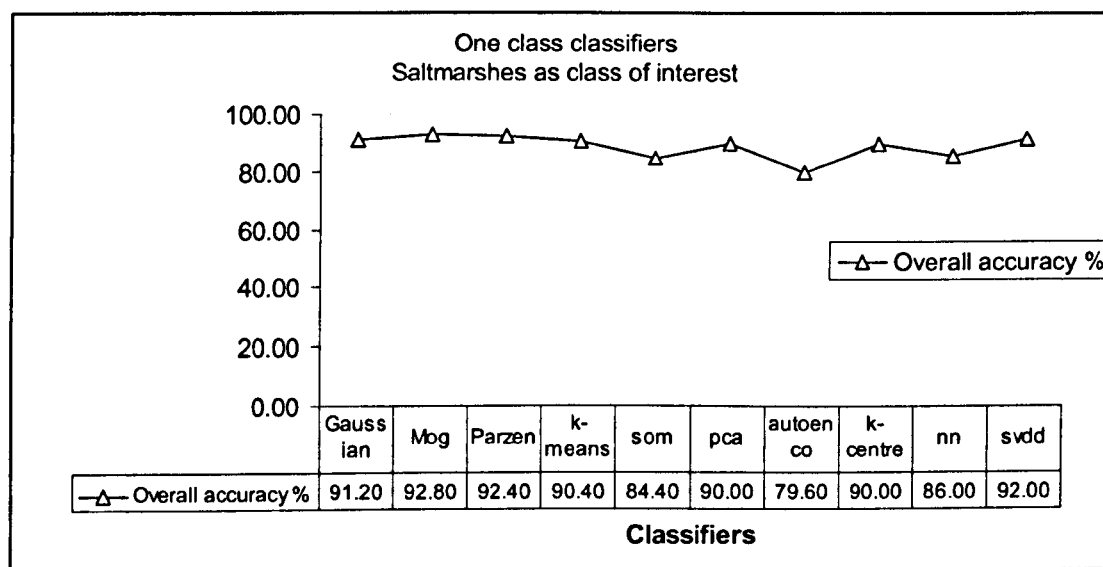


Figure 5.4 Overall accuracy results for the one-class classifiers Saltmarsh as class of interest

Looking at the overall accuracy obtained by the different classifiers (see Figure 5.3 and Figure 5.4 above) all the density classifiers showed high accuracies for both fen and saltmarsh which is shared by the K-means reconstruction classifier and the SVDD boundary classifier. Although the training dataset was relatively small, the density classifiers performed well. The reason for this could be that the training data

for the class of interest and the testing data distribution in the feature space were very similar (see Figure 5.5 and Figure 5.6 for fen as and Figure 5.7 and Figure 5.8 for saltmarsh), which as seen previously is one of the conditions for these type of classifiers to perform well.

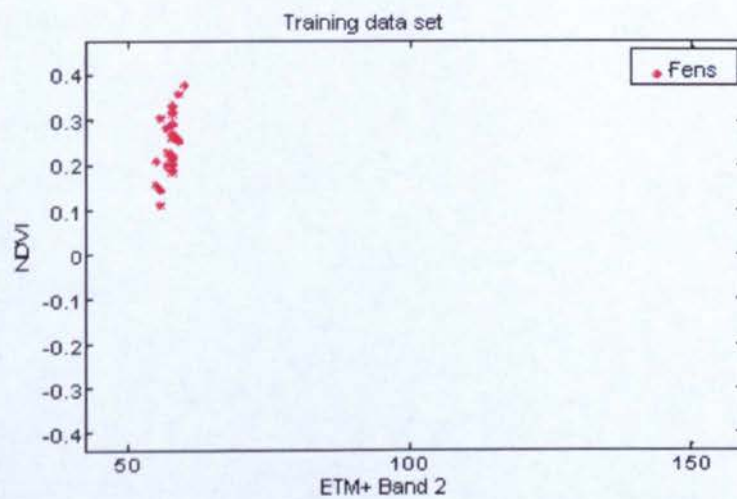


Figure 5.5 Training data for fen in 2 dimensional feature space

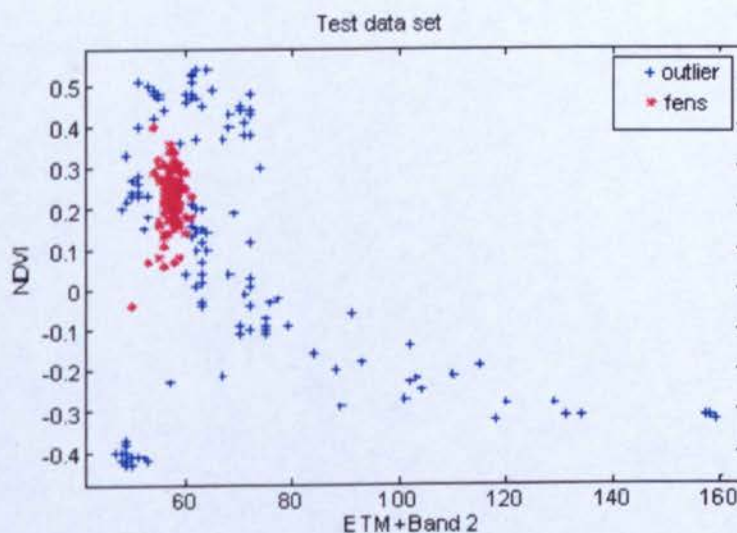


Figure 5.6 Testing data for fen in 2 dimensional feature space

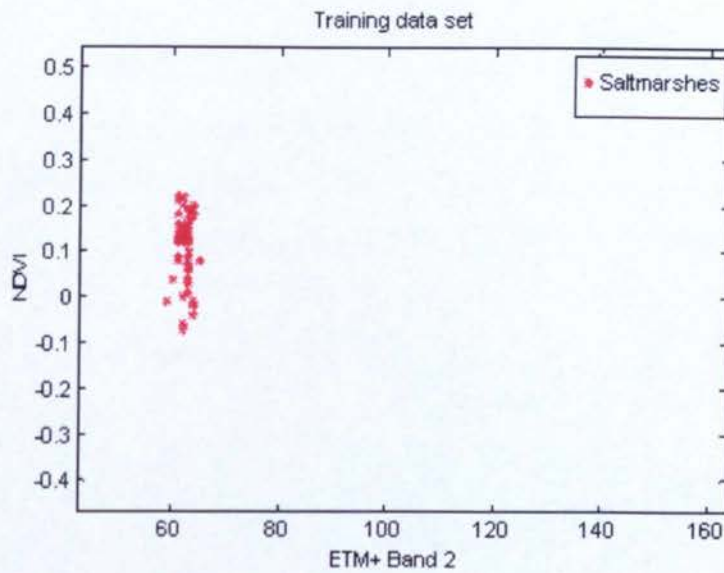


Figure 5.7 Training data for saltmarsh in 2 dimensional feature space

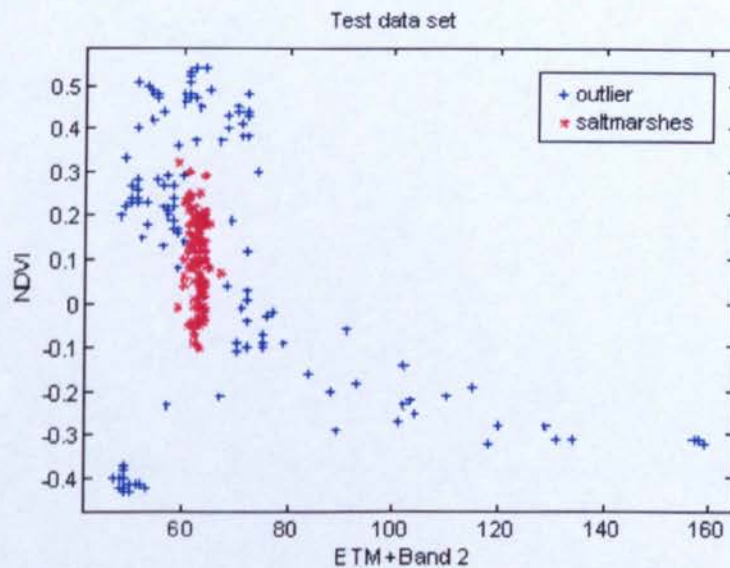


Figure 5.8 Testing data for saltmarsh in 2 dimensional feature space

When using fen as the class of interest, apart from the K-means classifier, the reconstruction methods did not seem to respond as well as the other methods and the overall accuracies were lower than those obtained with the density methods. Although the K-means classifier performed well with these particular data, it relies

on calculations of Euclidean distance to the unknown sample to the different clusters so it could be computationally very expensive. With saltmarsh, the PCA method obtained also high overall accuracy. Finally, of the boundary methods, the SVDD gave the best overall accuracy for both fen and saltmarsh. To have a closer look at these results producer's and user's accuracies were calculated.

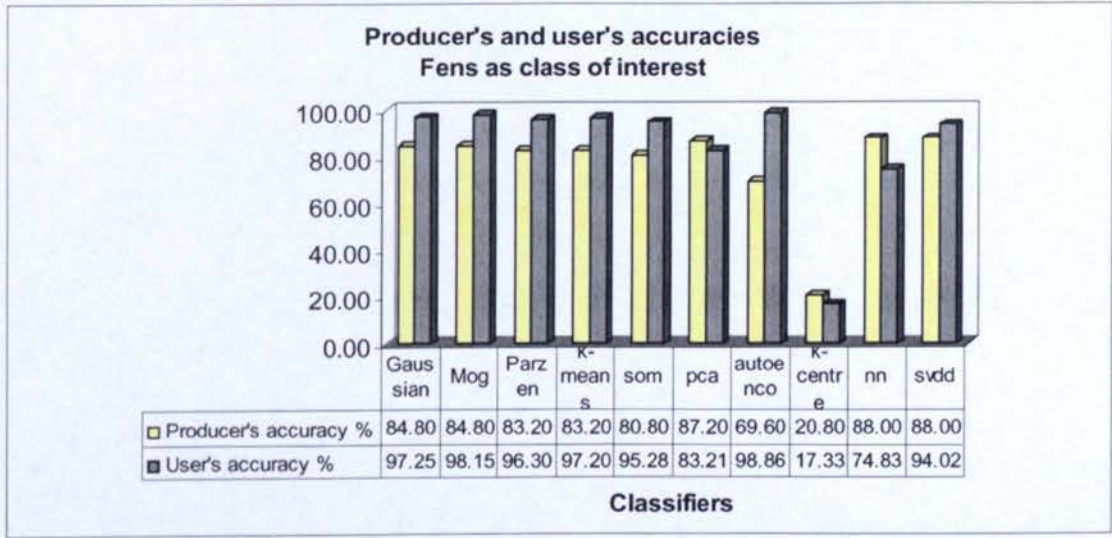


Figure 5.9 Producer's and user's accuracy for the one-class classifiers. Fen as class of interest

When using fen as the class of interest (see Figure 5.9 above) the SVDD obtained the highest producer's accuracies for both fen and saltmarsh (88.00% for both classes). In the case of fen, this was the highest producer's accuracy of all the classifiers.

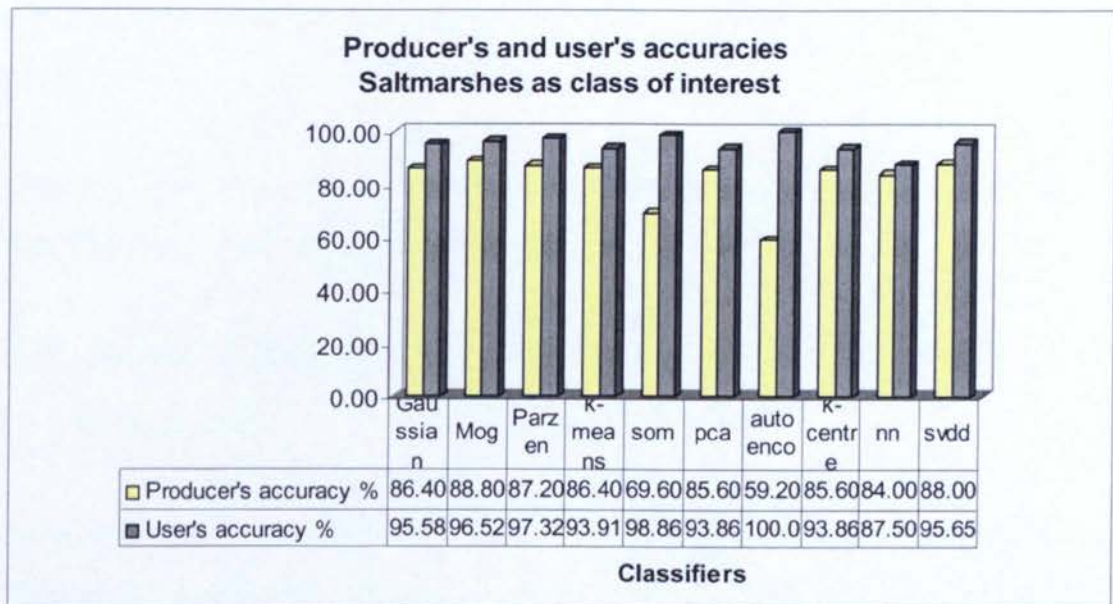


Figure 5.10 Producer's and user's accuracy for the one-class classifiers. Saltmarsh as class of interest

For saltmarsh (see Figure 5.10 above) this high producer's accuracy was shared by the MoG and Parzen density classifiers. Obtaining high producer's accuracy is important because the aim of this thesis is to accurately classify and map a particular habitat of interest. A high producer's accuracy means that a big percentage of the data have been identified as the class of interest. As seen in earlier chapters, high user's accuracies could also be important in order to minimise the errors of commission. As it can be observed in Figure 5.9 and Figure 5.10, SVDD offered high results for both accuracies.

In conclusion, although the density classifiers and the K-means reconstruction method showed high accuracies, it was decided that the SVDD classifier offered the best option for the classification of remote sensing data as it does not depend totally on the characteristics of the whole training and testing datasets and therefore has potentially a higher degree of flexibility. Further advantages of using the SVDD boundary method over density classifiers also include: (i) works very well with small training datasets that do not necessarily have to have the same density as the test data

and (ii) being based upon the theory of support vectors it is likely to show good generalisation.

Therefore, the following section will concentrate on describing this one-class classifier and its application to the case study.

5.2 Support Vector Data Description (SVDD). Principles and case study

As already mentioned, Tax and Duin (1999) developed a model based upon the Support Vector Machine classifier (Vapnik, 1995) called the Support Vector Data Description (SVDD). The sphere definition in Tax's SVDD has been related to the Minimum Enclosing Ball (MEB) problem in computational geometry (Badoiu *et al.*, 2002, Kumar *et al.*, 2003). This theory establishes that given a set S of points, the MEB of S , denoted by $MEB(S)$, is the unique minimum radius ball that contains all of S . The MEB is required to enclose all data points in S , including even outliers. SVDD on the other hand, leaves outliers outside the sphere as it will be explained later on in this section. Moreover, MEB can only handle low-dimensional data, whereas SVDD can function in the possibly infinite-dimensional kernel-induced feature space (Markou and Singh, 2003)

In the SVDD, the hypersphere around the target class is characterised by a centre a and radius $R > 0$ (Figure 5.11). The volume of the sphere is minimised by minimising R^2 with the function F :

$$F(R, a) = R^2 \tag{1}$$

With the constraints that all the training data are within R^2

$$\|x_i - a\|^2 \leq R^2, \quad \forall_i \tag{2}$$

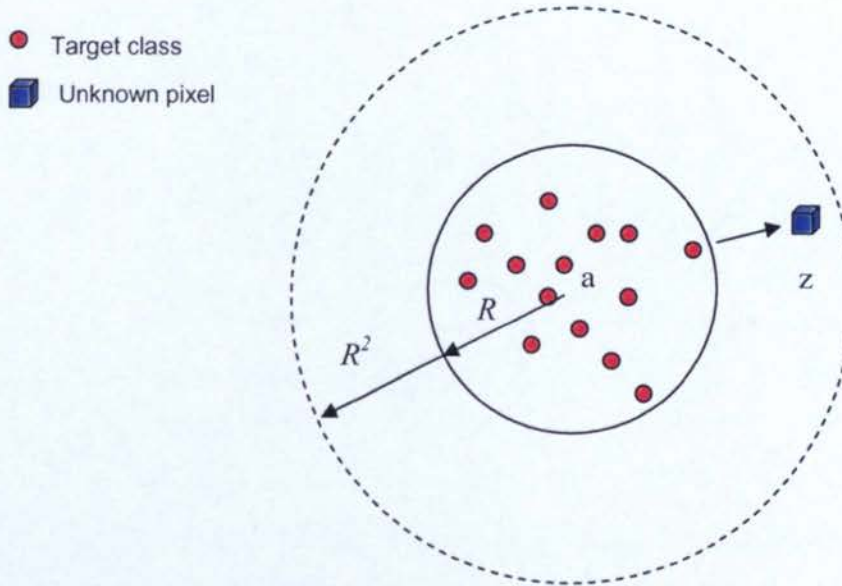


Figure 5.11 The hypersphere containing the target data, described by the center a and radius R . Based upon Tax (2001).

Considering R^2 instead of just simply R avoids overtraining including the possibility that the training set is not a total representation of the class of interest and therefore data belonging to the target class can be outside the hypersphere defined by R but inside R^2 .

Because it is possible to calculate the centre of the hypersphere a , it is easy to test if a new object (in our case a pixel) z is accepted by the sphere description and therefore belongs to the class of interest (Figure 5.11). For that, the distance from the pixel z to the centre of the hypersphere is calculated. A test pixel z is accepted within the hypersphere when this distance is smaller than or equal to the square radius. This can be formulated as:

$$f_{SVD}(\mathbf{z}; R) = I(\|\mathbf{z} - \mathbf{a}\|^2 \leq R^2) \quad (3)$$

Where I is defined as:

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

In the SVDD the pixels that are support vectors lie on the border of the hypersphere. As well as with the support vectors in the SVM, these support vectors are essential for the calculation of the optimal hypersphere. The difference between both methods is that SVM normally uses a percentage of the data as support vectors (as seen in Chapter 4) whereas SVDD normally only needs very few pixels in order to define the hypersphere (Tax and Duin, 2004). This is because when defining a rigid sphere only the coordinates of the centre of the sphere and the radius of the sphere are required. Therefore, in theory, only two support vectors are enough to determine the sphere (independent of dimensionality). For d -dimensional data, the required number of objects can increase up to $d+1$ (Tax and Duin, 2004).

However, this model can be too rigid when applied to real data when it is possible that the training dataset is not completely representative of the variability of the target class, and also it could be possible that some outliers could be within the description of the hypersphere. To relax this model and allow for these possibilities Tax (2001) introduces the idea of slack variables that measure the distance of a point to the boundary and determine if this point is outside the description. Also, the parameter C is introduced so that it gives the trade off between the volume of the description and the misclassification errors. C is a parameter to be chosen by the user, a larger C corresponding to a higher penalty to misclassification errors. This gives the following error:

$$\varepsilon(R, a, \xi) = R^2 + C \sum_i \xi_i \quad (5)$$

with constraints that (almost) all objects are within the sphere:

$$\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad \forall_i \quad (6)$$

As with SVMs to relax even more these constraints and to make this method suitable for non-linear cases the formulation can be transformed into a Lagrangian (as explained in Chapter 4). Constraints (6) can be incorporated into equation (5) giving the following formulae:

$$L(R, \mathbf{a}, \xi, \alpha, \gamma) = R^2 + C \sum_i \xi_i - \sum_i \alpha_i \{R^2 + \xi_i - (\mathbf{x}_i \cdot \mathbf{x}_i - 2\mathbf{a} \cdot \mathbf{x}_i + \mathbf{a} \cdot \mathbf{a})\} - \sum_i \gamma_i \xi_i \quad (7)$$

Where the Lagrange multipliers are $\alpha_i \geq 0$ and $\gamma_i \geq 0$. L has to be minimized with respect to R , \mathbf{a} and ξ . To minimize L :

$$\sum_i \alpha_i = 1 \quad (8)$$

$$\mathbf{a} = \sum_i \alpha_i \mathbf{x}_i \quad (9)$$

$$C - \alpha_i - \gamma_i = 0 \quad \forall_i \quad (10)$$

Resubstituting (8) and (10) into equation (7) gives:

$$L = \sum_i \alpha_i (\mathbf{x}_i \cdot \mathbf{x}_i) - \sum_{i,j} \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad (11)$$

With an upper bound

$$0 \leq \alpha_i \leq C \quad (12)$$

being α the Lagrangian multipliers as described in Chapter 4.

This upper bound means that when objects obtain $\alpha_i = C$ these objects are considered as outliers. Furthermore, in the minimization of L , a large fraction of $\alpha_i = 0$. So there are only a few objects \mathbf{x}_i with $\alpha_i > 0$. These objects are the support vectors.

The one-class classifier Support Vector Data Description can then be expressed as:

$$f_{SVDD}(\mathbf{z}; \alpha, R) = I(\|\mathbf{z} - \mathbf{a}\|^2 \leq R^2) = I\left((\mathbf{z} \cdot \mathbf{z}) - 2 \sum_i \alpha_i (\mathbf{z} \cdot \mathbf{x}_i) + \sum_{i,j} \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \leq R^2\right) \quad (13)$$

However, it can be argued that it is still unlikely that this model will fit the data well in real life situations. If the data were to be mapped in a different feature space, it would be possible to obtain a better fit between the actual data boundary and the hypersphere model. As with Vapnik's model (1995) Tax assumes a mapping ϕ of the data using kernel functions.

When a kernel function maps the target data into the feature space the hypersphere model fits the data in a better way and a better classification performance is obtained. Then equations 8 and 9 will become:

$$L = \sum_i \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_i) - \sum_{i,j} \alpha_i \alpha_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \quad (14)$$

and

$$f_{SVDD}(\mathbf{z}; \alpha, R) = I\left(\phi(\mathbf{z}) \cdot \phi(\mathbf{z}) - 2 \sum_i \alpha_i \phi(\mathbf{z}) \cdot \phi(\mathbf{x}_i) + \sum_{i,j} \alpha_i \alpha_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \leq R^2\right) \quad (15)$$

One of the strengths of SVMs and the SVDD is that they allow the researcher to have control of the analysis and to choose an appropriate kernel given the data and the problem specific knowledge (Smola *et al.*, 1998). Moreover, according to Tax (2001), when outliers are available they can be used during the training to obtain a more precise data description and to obtain a tighter boundary around the data in the areas where outliers are present. In this scenario it could be argued that it is possible to train both a SVDD and a traditional (two-class) classifier on these data. This is true when both a representative sample from the target class and a large amount of example outliers are available. In that case the conventional classifier could work as well or better than the SVDD. The choice between a SVDD and an ordinary classifier is therefore influenced by both the number of outlier objects available for training and how well they represent the target and the outlier distributions (Tax, 2001). The advantage of the SVDD approach for remote sensing classification is that these outliers do not need to be as well defined as in a binary classification. Also, samples of outliers should be easy to find within an image.

Having reviewed the principles behind the SVDD classifier, the following section will show the application and results of the classification performed by the SVDD using the case study data.

5.3 Case study. SVDD and classification and mapping of a class of interest.

It was decided to compare the performance of the one-class classifier SVDD using a) only the class of interest to train the classifier (SVDD A) and b) incorporating some outliers (SVDD B). In order to do this it was necessary to define (i) the training datasets and (ii) the parameters used by the SVDD classifier.

(i) Training and testing datasets

One of the advantages of SVMs and SVDDs over other classifiers is their capacity to use smaller training datasets to get similar levels of accuracy. To assess the impact of training dataset size and to establish a minimum training set to train the SVDD classifier it was decided to use a series of very small training datasets. As mentioned already in Chapter 3 and Chapter 4, some of the literature suggests the use of a minimum of $10-30p$ cases per-class for training (p being the number of wavebands used) (Mather, 2004). As with the previous binary classifiers, it was decided to use a range of training sets that included smaller training sizes than the recommended to assess the performance of the classifier. The training datasets ranged from 5, 10, 20, 25, 50, 75, 100, 125, to 150 pixels in which each of the datasets included the previous one so that $5 \in 10 \in 20 \in 25 \in 50 \in 75 \in 100 \in 125 \in 150$. In SVDD B the training sets were composed of 75% of the same pure pixels used in the SVDD A, adding in each set 25% of data composed by “outliers” or a group of pixels collected from the other 7 classes identified in the image.

The testing set was formed by 50% pixels from the class of interest and 50% of outliers forming a total of 250 pixels. As already established, this is the same testing set that was used for the binary classifications performed in Chapter 4. All the training and testing sets can be found in Annex F.

ii) SVDD parameters

As with the SVM, the use of kernels makes it possible to map the data implicitly into a feature space and to train a linear machine in such space. The key is finding the kernel function that makes the calculations of hypersphere in the feature space more efficient. As with the binary SVM, it was decided to assess the three most important kernel functions as described by Smola *et al.* (1998): (i) polynomial, (ii) gaussian radial basis function and (iii) exponential radial basis function. The best performing

kernel was selected by a 5-fold cross validation. It was also decided to perform the kernel selection using the same values for the free parameters as those used when selecting the kernel for the binary SVM classification. This is, for the polynomial kernel it was decided to test the polynomial with values 1 to 10 based upon previous studies (Cortes and Vapnik, 1995, Huang *et al.*, 2002). For the RBF kernel, also based upon earlier studies (Vapnik 1995, Joachims, 1998, Huang *et al.*, 2002), it was decided to test the kernel using free parameter values of 1 to 10. And finally, due to the lack of references regarding the exponential kernel it was decided to keep within the above range of values. Each kernel was tested with different values of C (1, 10, 100, 1000). The criteria used to select both kernel and value of C was the overall accuracy obtained in the classification. The detailed results of this 5-fold validation can be found in Annex B. The results showed that the polynomial kernel (free parameter 1 to 3) and the RBF kernel (free parameter 8 to 10) provided the best accuracies for both fen and saltmarsh. For both the highest overall accuracy was obtained by the polynomial kernel free parameter 2, so it was decided to use these parameters for the application of the SVDD to the case study. In terms of the value of C , the value $C = 1$ gave computational errors and consequently the cross-validation was only effective when using values 10, 100 and 1000. Bigger values did not change the accuracy results obtained by $C = 1000$. The highest results were obtained when using $C = 100$ and the polynomial kernel so consequently these were the parameters used for the experiments that followed.

5.3.1 SVDD A

The SVDD A was trained using the parameters described in the previous section and training datasets composed exclusively of the class of interest with sizes ranging from 5 to 150 as described earlier. The results obtained for fen and saltmarsh can be observed in Figure 5.12 and Figure 5.13 below:

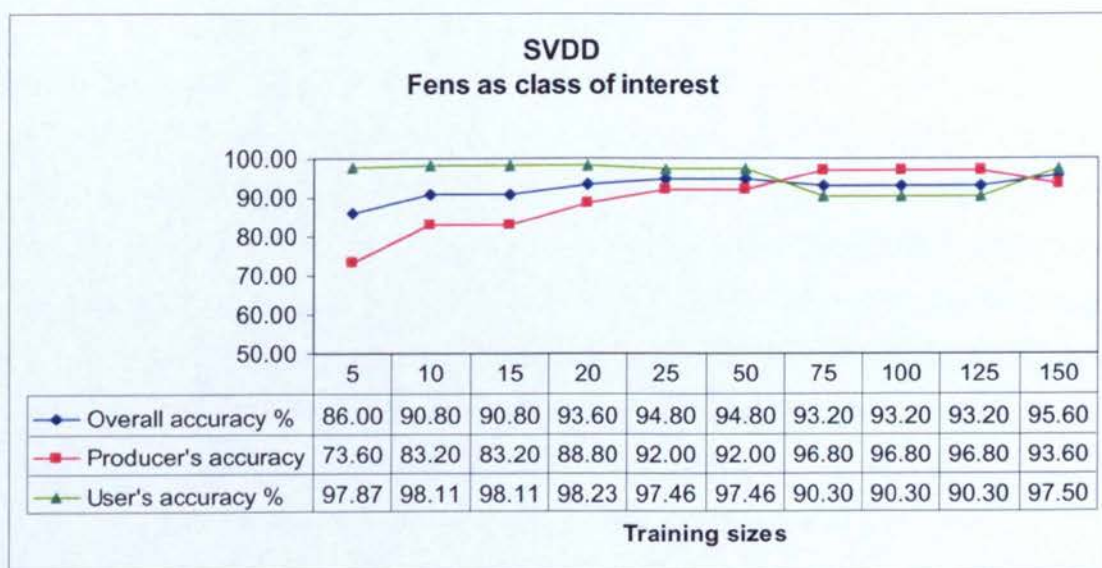


Figure 5.12 Training sizes impact on overall classification accuracy (fen)

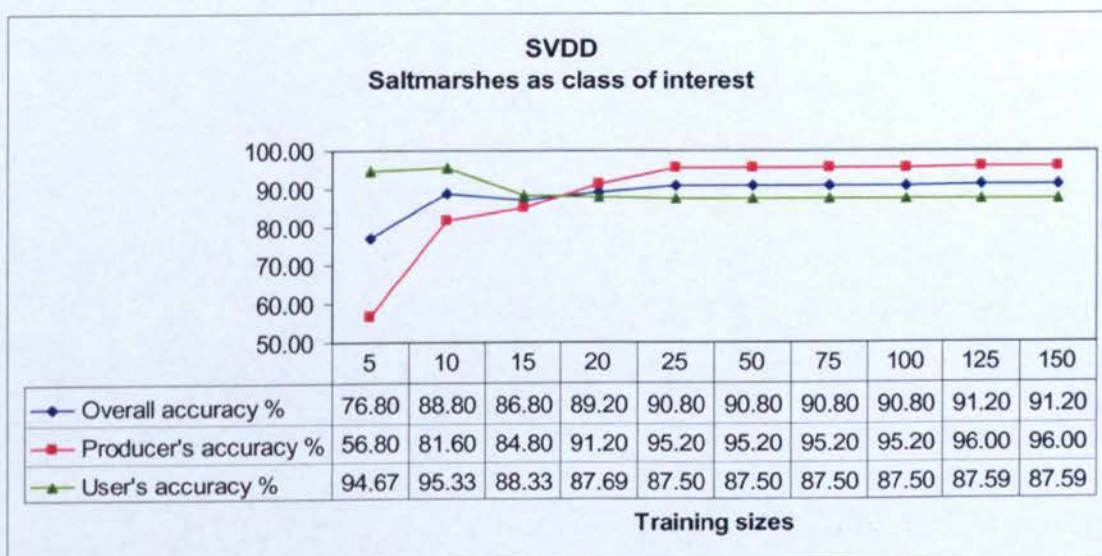


Figure 5.13 Training sizes impact on overall classification accuracy (saltmarsh)

When training the SVDD for fen as target class, there was a general increase in overall accuracy with the increase of training size. The lower accuracies were obtained with the very small training datasets (5 to 15 pixels). The training size of 25 pixels obtained high overall, producer's and user's accuracies (94.80%, 92.00% and 97.46% respectively) (see Figure 5.12). Increasing the training set size to bigger training sizes 50, 75, 100, 125 and 150 did not produce a significant increase in

accuracies ($Z < \sim 1.96$) at 95% confidence interval. In the case of saltmarsh the same trend occurred (see Figure 5.13).

This shows one of the major advantages of classifiers based upon support vector machines over all the other classifiers. This is, its capacity to obtain high accuracies using only a small amount of training data. In particular, with SVDD these accuracies are obtained with only data from the class of interest so the researcher can optimise the training process focusing only on this class. Confusion matrices for each of the training sizes can be found in Annex A.

5.3.2 SVDD B

As stated earlier on, it was decided to test the SVDD using a sample of outliers to see if this would alter in any way the result obtained when using only the class of interest when training the classifier. This classification was performed on training sets formed by 75% of the target class and 25% of data composed by “outliers”. However, in this case, the results obtained were exactly the same as the ones obtained with no outliers. When this result was obtained, it was decided to increase the number of outliers in the different datasets with a composition of 60% of pure pixels and 40% outliers. This also had no impact on the final result. The conclusion drawn from this could be that in this case study outliers are pure pixels that occupy small remote areas in the feature space and consequently do not affect the definition of the boundary around the target class (see Figure 5.14 below).

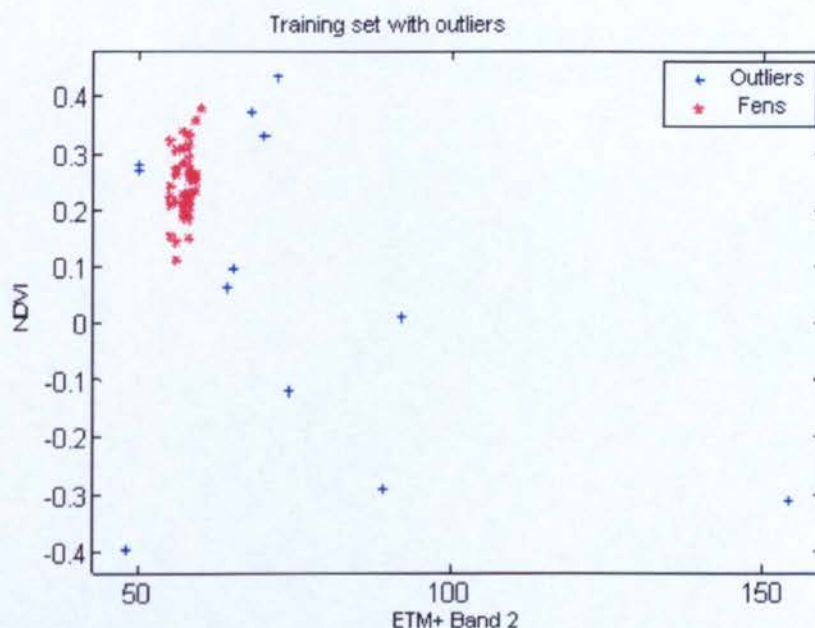


Figure 5.14 Training set with outliers. 75% target data 25% outliers. Fen as class of interest.

However, in cases when the class of interest has been selected through a process of intelligent training where only those pixels likely to be support vectors are selected, the addition of outliers does seem to produce a positive effect and the accuracy results are higher (Foody *et al.*, in press)

To have a closer look at the performance of the SVDD it was decided to compare it against the standard ML classification. The comparison was done between the ML results obtained with a training dataset of 150 pixels per class against those obtained by the SVDD with a training dataset of 100 pixels belonging to the class of interest. These results are illustrated in Figure 5.15 and Figure 5.16 below.

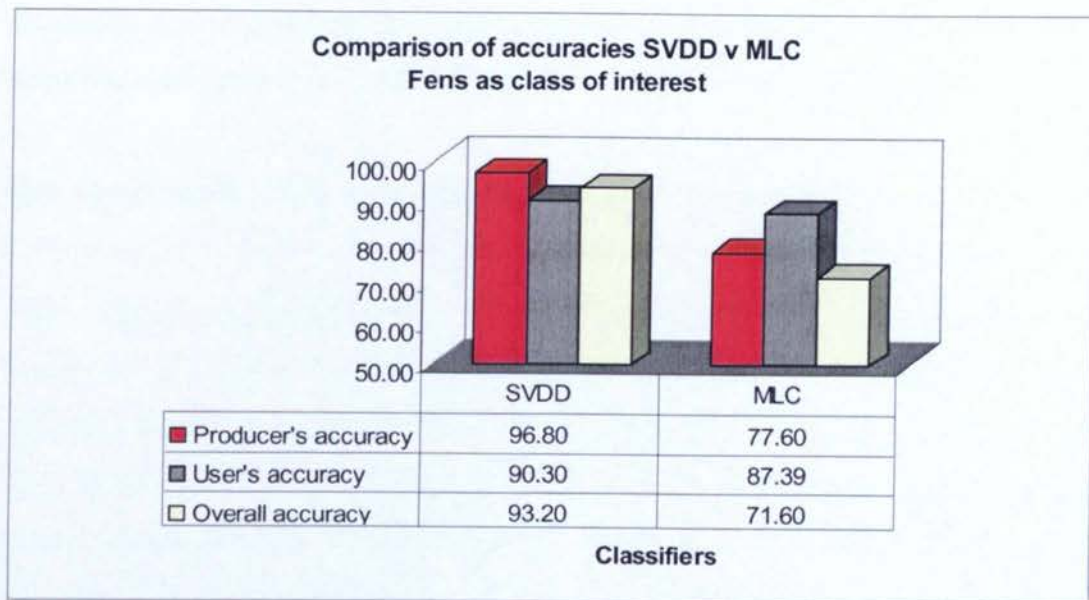


Figure 5.15 Comparison of accuracies SVDD v MLC. Fen as class of interest

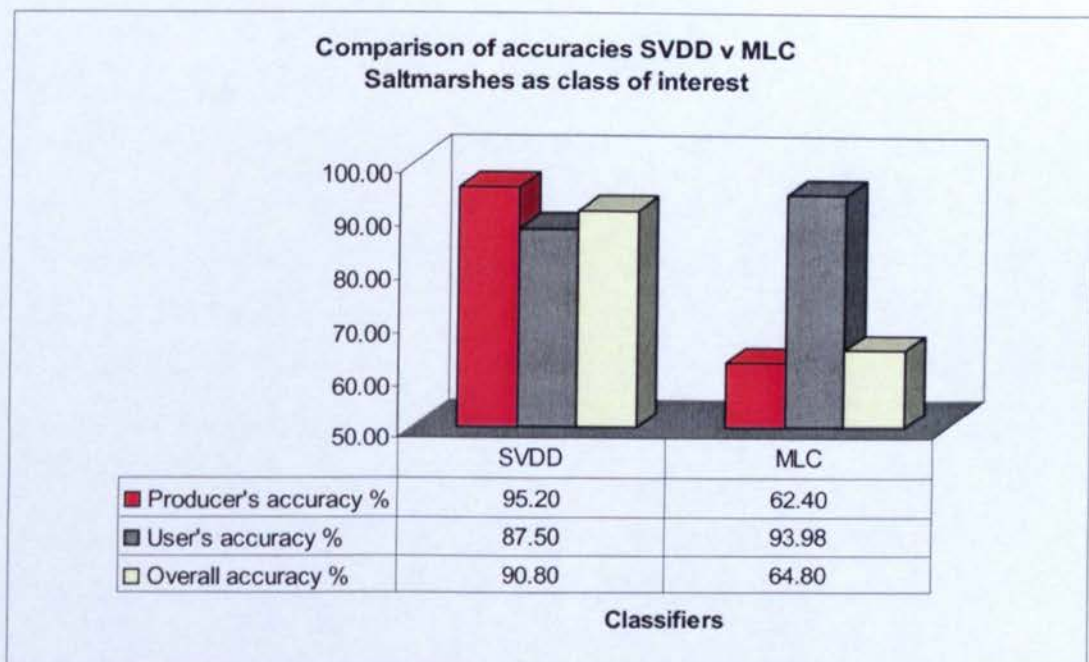


Figure 5.16 Comparison of accuracies SVDD v MLC. Saltmarsh as class of interest

These results highlight the suitability of the SVDD to classify of a habitat of interest using remote sensing data with results that significantly surpassed those obtained by the standard ML classifier ($Z = \sim 8$ at 95% confidence interval). They clearly

emphasize the potential of this one-class classifier for remote sensing applications when concentrating on the classification and mapping of a particular class.

5.4 Summary and Conclusions

This chapter has reviewed the principles behind one-class classification and compared the applicability of different one-class classifiers to remote sensing classification using a training dataset of 100 pixels. The results showed that in this case, density classifiers performed quite well due to the fact that training and testing datasets shared the same distribution in the feature space. Of the other classifiers, the classifier K-centre (reconstruction method) showed a high accuracy for both fen and saltmarsh and so did the SVDD. However, the SVDD was considered to be more suitable for remote sensing classification due to its advantages in terms of flexibility and ability for generalisation using small training datasets that do not necessarily have to share the same distribution in the feature space as the test data. Also it is important to highlight that the SVDD showed the highest producer's and user's accuracies which could be extremely important in cases where ground data are not readily available. Furthermore, the SVDD achieved these high accuracies with training sizes as small as 25 pixels. As a reference, these results could be compared against those obtained by the Land Cover Map of Great Britain 2000 (LCM2000). Here the class fen had been aggregated into the class "seminatural grass". For this aggregated class the producer's accuracy was 41.00% and the user's accuracy was 48.00%. However, the general conclusions of the report stated that the average classification accuracy for target classes is of 80.00-85.00% (Fuller *et al.*, 2002). Either way, the results obtained by the SVDD clearly surpass those of the LCM2000.

One clear advantage of the SVDD is that it does not require data from any of the other classes in the image. Furthermore, being based upon the theory of support vector machines the ability to generalise to unseen data is much higher than other one-class classifiers. Moreover, the SVDD only needs only a few support vectors ($d + 1$), where d is the dimensionality of the data, to be able to define an optimal

hypersphere around the target data. This could have additional implications for further research as the training process could be optimised as far as identifying these few pixels and discard the rest of the data. In this sense, however, it is very important to remember that the SVDD is a one-class classifier and as such it requires a relatively good description of the target class in order to be able to find these key pixels that act as support vectors.

The findings of this chapter demonstrate the suitability of the SVDD one-class classifier for classification of remotely sensed data and land cover mapping depicting a specific class of interest. Moreover, the ability of the SVDD classifier to operate with minimal training data means that a competitive and efficient approach to the classification can be adopted. These results fully contribute towards the aim and sub-aims of this thesis. The full implications of these findings are also of real value in many studies, particularly those where resources are scarce, time is limited and where there is a particular habitat of interest that needs to be classified and mapped. This is fully in line with the needs of relevant authorities that have to comply with the mapping and monitoring requirements of the EU Habitats Directive.

Having addressed the issues of binary and one-class classification for the classification of a particular habitat of interest, the following chapter will attempt to achieve better results than the ones achieved by these classifiers by applying a new trend in pattern recognition that is being recently introduced into remote sensing: ensemble of classifiers. This last research chapter will build upon the results obtained by the SVM, DT and SVDD with the purpose of achieving even higher classification accuracies in order to satisfy the EU Habitats Directive mapping and monitoring requirements.

6 Ensemble of Classifiers for the classification of a habitat of interest

"You build trust through the ensemble work."

Eric Bass

The main aim of this last research chapter is to investigate advanced methods in order to obtain a higher accuracy than the binary and one-class classifiers explored previously in this thesis.

As seen in chapters 4 and 5, the traditional approach to solving a classification task is to compare the performance of several classifiers in order to select the most suitable one for a particular case. However such an approach does not always guarantee that the optimal solution has been found (Roli and Giadito, 2002). There are different reasons for this. Individual classifiers try to find hypotheses within the search space H that will provide the optimal solution to the classification problem. If there are enough available data it is possible for individual classifiers to find the optimal solution for the classification problem. In this case, different classifiers can obtain similar solutions and it is possible to select just one of them which offers more simplicity or higher generalization potential (Valentini and Masulli, 2002). However, it could be the case that the training dataset is not a very representative sample and that individual classifiers that perform very well with this training data do not perform that well when confronted with real data. So, although they perform well with the training data, the classifiers are far from being the optimal classifier for the

case study (see case 1 in Figure 6.1 below) and therefore their generalization is very poor. Another case (case 2 in Figure 6.1 below) could be that the search paths of the different algorithms of individual classifiers stop the search when they arrive at a solution which is far from the optimal solution to the classification problem. Finally it could happen that the individual classifiers are not the ideal ones for the case study and that the optimal classifier for that case is outside the search space (case 3 in Figure 6.1).

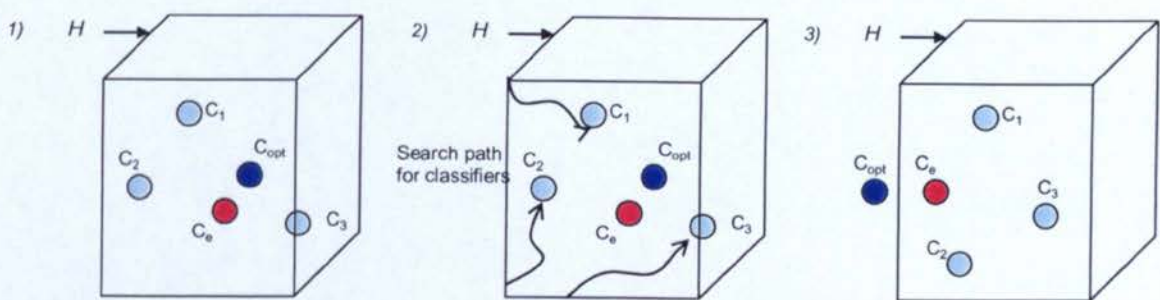


Figure 6.1. Reasons why best individual classifiers do not guarantee an optimal solution. C_1 , C_2 , C_3 represent the individual classifiers. C_e is the classifier obtained by an ensemble and C_{opt} is the optimal classifier. H is the search space where the classifiers are looking for the optimal solution. In 1) the ensemble C_e is closer to the optimal classifier 2) the search path of each classifier might not go near the optimal classifier 3) the optimal classifier might be outside the search space. Based upon Günter and Bunke (2004).

Therefore, an alternative approach has emerged over recent years and it is based upon the idea of combining different classifiers (Windeatt, 2003). The basis of an ensemble of classifiers is that instead of choosing the classifier that gives the highest accuracy for a particular case there is the possibility that using them in a combination could increase the overall accuracy of the classification and find a result closer to the optimal solution (C_e in Figure 6.1) (Jain *et al.*, 2000, Windeatt, 2003). This idea has been studied by mathematicians since the 18th century with the Condorcet Jury Theorem (1785). This theorem stated that the judgment of a group is superior to that of an individual, with the condition that the members of the group have reasonable competence (a probability of being correct higher than 0.5). Within the context of

machine learning the theory of combining classifiers can be traced back as far as the 1960s when they were investigated by Nilsson (1965). But it was the work on neural networks by Hansen and Salamon (1990) that generated a new research interest on combining classifiers and it has been during the last decade that this area of research has greatly evolved within statistical pattern recognition and machine learning.

In remote sensing, the application of an ensemble of classifiers is very recent and there are few examples in the literature linking an ensemble of classifiers and remote sensing. A few studies include multivariate land cover detection change (Bruzzone *et al.*, 2004), neural network ensembles for image classification (Giacinto and Roli, 2002) enhanced classification algorithms for land cover classification (Chan *et al.*, 2001, 2003) and multiple classifiers applied to multisource remote sensing data (Briem *et al.*, 2002), all of which presented encouraging results although the authors are aware that more research is still needed in this area.

Taking all the above into account, the present chapter's objective is to assess the method of an ensemble of classifiers to evaluate whether this approach would enhance the performance of binary and one-class classifiers to classify and map a class of interest. The different ensembles of classifiers were constructed using fen as the only class of interest. It was not considered necessary to construct the different ensembles using also the class saltmarsh because the investigation in the previous research chapters demonstrated that the performance of these classifiers was not habitat specific.

The following sections look in more detail into different ensemble methods by describing the principles behind the ensemble of classifiers and different methods available. Following this, it is shown how the ensemble of classifiers was formed and tested using the binary and the one-class classifiers studied in chapter 4 and 5. Finally, an overall ensemble with both binary and one-class was built and final conclusions were drawn. The structure of this chapter is illustrated in Figure 6.2 below.

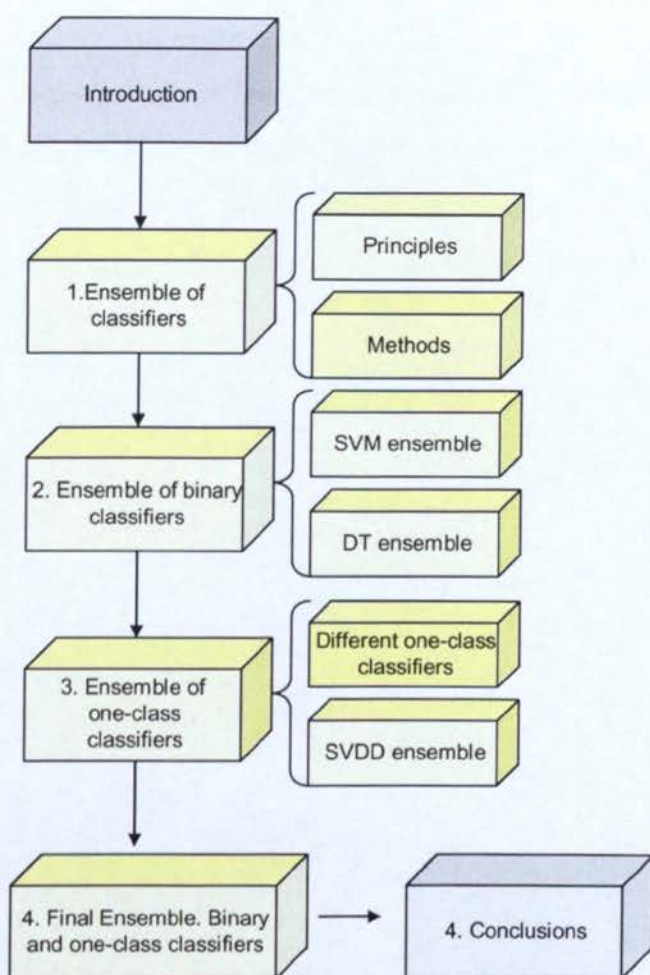


Figure 6.2 Chapter 6 structure

6.1 Ensemble of classifiers. Theory and methods

In the extensive literature that has been produced since 1990 (Battiti and Colla, 1994, Drucker *et al.*, 1994, Fillipi *et al.*, 1994, Lam and Sue, 1995, Woods *et al.*, 1997, Xu *et al.*, 1992, Kittler *et al.*, 1998, Jain *et al.*, 2000, Sharkey *et al.*, 2000, Dietterich, 2000), a variety of terms has been used to define the idea of combining classifiers. Such terms include committee of classifiers, classifier fusion, ensemble of classifiers, combination of classifiers and aggregation of classifiers. Of all of them, the term ensemble seems to have the widest meaning including a wide range of combining methods. Consequently this term will be the one adopted within this thesis.

The construction of an ensemble of classifiers has become an important trend of research within machine learning (Kittler and Roli, 2000, Kuncheva *et al.*, 2001) and several methods have been proposed (Wolpert, 1992, Tumer and Ghosh, 1996, Benediktsson *et al.*, 1997). Depending on the main classification criterion adopted combination methods can be grouped and analysed in different ways (see Figure 6.3 below).

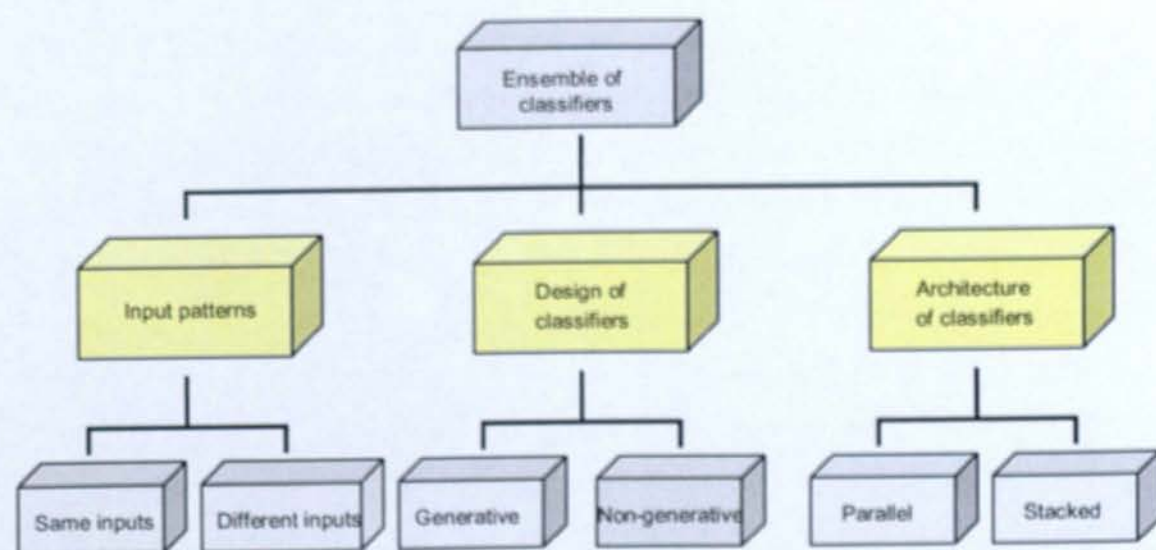


Figure 6.3 Different combination methods for ensembles of classifiers

According to Kittler (1998) if considering the nature of the input patterns as the main criterion it is possible to identify two main groups: (i) one that uses the same input patterns (i.e. features and training datasets) and (ii) one that uses different representations of the inputs. Heavily studied modifications of these inputs are bootstrapping (bagging) or re-weighting the data (boosting). If considering the nature of the classifiers of the ensemble as the main criterion it is also possible to distinguish between two groups: (i) non-generative and (ii) generative ensemble methods. Non-generative ensemble methods combine a set of given classifiers. Generative ensemble methods generate sets of base classifiers based on the learning algorithms or on the structure of the data set and try to actively enhance diversity and accuracy of these base classifiers (Valentini and Masulli, 2002). Finally, if considering the architecture as the main criterion, the construction of an ensemble

can be done in two main ways: (i) parallel combining of classifiers and (ii) stacked combining. Parallel combining might be especially useful if the objects are represented by different feature sets, and they are often, but not necessarily, of the same type. In stacked combining the different classifiers are computed for the same feature space. Stacked classifiers are normally of a different nature (Duin, 2002).

Whatever the method chosen for combining the classifiers, the diversity of these classifiers is of great importance. If the classifiers were identical there would be no gain or any progress by combining them. The hypothesis that an ensemble of classifiers is more accurate than any of its individual component classifiers if and only if the component classifiers are accurate and diverse was first introduced within the machine learning community by Hansen and Salamon (1990). In their work they demonstrated how using individual classifiers that are independent and with error rates less than 50%, the error rate of the ensemble classifier will decrease with the number of individual classifiers. Diversity has been recognised as a key issue in ensembles of classifiers (Lam, 2000, Cunningham and Carney, 2000). However, other authors (Shipp and Kuncheva, 2002) found that the correlation between diversity and combination methods is not very high or consistent and that the issue of diversity measures in designing ensembles is still very much open.

Furthermore, the classifiers used in the ensemble should be different but they should also be comparable, i.e. their outputs should be represented such that a combining classifier can use them as inputs (Duin, 2002). In this sense, a reliable set of different classifiers might be generated in the following ways: (i) different initializations, (ii) different parameter choices, (iii) different architectures, (iv) different classifiers, (v) different training sets and (vi) different feature sets. Of all of these, the last three make a bigger contribution towards diversity of classifiers (Duin, 2002).

However, it is the manipulation of the training sets that has captured a great part of the research on ensemble diversity. This manipulation consists of the classifier running several times using each time a different partition of the training set. It works well for learning algorithms whose output predictions have changes in response to a

small change in the training samples. Among all these methods, cross-validation, boosting and bagging are the most successful and representative methods. In particular, bagging and boosting have demonstrated their superiority when compared against other ensemble methods by different researchers (Dietterich, 2000, Quinlan, 1996, Bauer and Kohavi, 1999). The difference between the three methods is that in cross-validation the training dataset is divided into k subsets which are used to train the classifier k times. Then the average error across all k trials is computed. However, bagging and boosting modify the original training dataset, building classifiers on these modified training sets and then combine them into a final decision (Skurichina *et al.*, 2002).

In the case of bagging, which was first proposed by Breiman (1996), the classifier is run several times on training samples, which are obtained randomly based upon the original training samples by sampling with replacement with the same size as the original training size. Some training samples may appear in the produced training sets while others may not. Such a training set is called a bootstrap replicate of the original training set, and this technique is called Bootstrap Aggregating, from which the name of Bagging is derived. These methods have been shown to reduce the variance of the classification (Gislason *et al.*, 2004). Boosting was first proposed by Freund and Schapire (1996) and it was aimed to enhance the performance of weak classifiers. The training sets are obtained in a deterministic way as opposed to bagging where the training sets are obtained randomly and independently from the previous step of the algorithm. In each training set a higher weight is assigned to cases that are incorrectly classified in a present trial and so they have a higher probability of being chosen in a new training set. The aim of boosting is to maximize the margins of the training data in a similar way to the SVM. Boosting margins are maximized locally and subject to a particular training set whilst SVM looks for a global optimization. Boosting reduces both the variance and the bias of the classification and in general is a very accurate classification technique (Gislason *et al.*, 2004). But it also has a few drawbacks: it is a computationally demanding and slow process, it can overtrain and therefore jeopardize its generalization capacity and it is very sensitive to any noise present in the training data (Briem *et al.*, 2002). Well-

known algorithms for boosting include Adaboost (Freund and Schapire, 1996) and Arc-4x (Breiman, 1998). A comparison of ensembles using bagging, Adaboost and Arc-4x boosting for a case of land cover classification is shown in Chan *et al.* (2003). In their research they compare the performance of the above ensemble methods for the classification and mapping of logged forest in tropical Africa. Their results show that in terms of overall classification accuracy, bagging, Adaboost and Arc-4x are similar and also that the accuracy enhancements between them are marginal.

Bagging is normally useful for linear classifiers constructed on small training samples and reducing data dimensionality (Skurichina *et al.*, 2002). Also it is effective when a learning algorithm is unstable and a small change in training samples leads to a large change in accuracy (Chan *et al.*, 2003). Boosting on the other hand, is efficient for low complexity classifiers that are constructed on large training samples (Skurichina, 2001, Freund and Schapire, 1997).

Once a number of diverse classifiers have been obtained by any of the approaches described above, a method for combining such classifiers is required. An extensive account of such methods can be found in Shipp and Kuncheva (2002) and Kittler *et al.* (1998). Recent studies have reported that voting methods used in ensembles of classifiers are useful to increase accuracy in land cover classification from remote sensing data (Chan *et al.* 2001, DeFries and Chan, 2000). The simplest one of the voting methods uses the majority rule. Although this is extremely simple it has been regarded as a very robust combination compared to more sophisticated ones (Yu, 2003). This method is based on an ensemble of classifiers which combines the outputs of a set of classifiers (Hansen and Salamon, 1990, Benediktsson and Swain, 1992, Wolpert, 1992). Voting considers only the output by each classifier and regards the output which appears most often when counting the output of the classifiers as output of the combined classifier. According to Brodley and Friedl (1996) a majority vote ensemble classifier will outperform each individual base-level classifier on a dataset if two conditions hold: (1) the probability of a correct classification by each individual classifier is greater than 0.5 and (2) if the errors in predictions of the classifiers are independent (Hansen and Salamon, 1990). More

elaborate schema used weight voting rules where each component is associated with weight during the training stage. In weighted voting schemes, each vote receives a weight, which is usually proportional to the estimated generalization performance of the corresponding component classifier (Bauer and Kohavi, 1999). The weighted majority voting is similar to weighted voting; the main difference is how the weights are generated. It makes predictions by taking a weighted vote among a pool of classification algorithms and learns by altering the weight associated with each prediction algorithm.

To conclude this section, it is important to note that, although all the above studies support the use of ensembles over a simple classification method, some classification and pattern recognition problems can actually be solved by a single classification. Furthermore, according to Skurichina *et al.* (2002) it is normally in cases where the data distribution is very complex or high dimensional when an ensemble of classifiers is likely to perform better.

Taking all the above into account, the following sections of this chapter will address the application of ensemble methods to the classifiers that have been studied in the previous research chapters, this is, binary SVMs and DTs classifiers, one-class classifiers and within these the SVDD classifier. The purpose of these experiments will be to assess whether an ensemble could obtain a higher accuracy than the one already achieved by these classifiers and if their application would be appropriate to meet the aim of this thesis.

6.2 Ensembles of binary SVM and DT classifiers

6.2.1 SVM Ensembles

The idea of forming SVM ensembles in order to increase classification accuracy was suggested by Vapnik (1999). As seen in Chapter 4, the SVM classifier showed good generalization and the parameters were easy to learn for a global optimum (Burges, 1998). Because of this, SVM ensembles have not been considered as a method for

improving classification accuracy until very recently. SVM algorithms are normally implemented using approximate parameters and sometimes this is not enough to classify all unknown test examples. Therefore, despite the high performance of SVM several researchers have sought to achieve better results with ensemble methods (Derbeko *et al.*, 2002, Kim *et al.*, 2003, Pavlov *et al.*, 2000, Valentini *et al.*, 2003, Valentini *et al.*, 2004).

Kim *et al.* (2003) proposed to use an SVM ensemble using (i) bagging and boosting methods to construct the ensembles and (ii) majority voting, weighted majority voting and stacked architectures as combination methods. Their results showed that these three methods produced higher accuracies than the SVM that was in principle chosen as the single best classifier. Another approach can be found in Ma *et al.* (2004). They used SVM ensembles also using bagging and boosting in order to reduce the size of training datasets and consequently increase SVM training speed. Their results showed that the SVM trained with the original dataset and with the bagged and boosted training sets were the same as long as the parameter C was calibrated accordingly. Their conclusions were that duplicating a sample n times was equivalent to increasing its parameter C n times when training the SVM so the increase in accuracy was relative.

As seen in the previous section, bagging is the most effective method when the training sample size is smaller or comparable with the data dimensionality. In the case of this thesis the calibration of different classifiers had been obtained by using the minimum training size set possible with which high accuracy results were obtained. In this sense, it was logical to say that the appropriate technique in order to create an ensemble of diverse SVMs was bagging. Moreover, because the results of the SVM were shown as labels the voting measure to apply to the ensemble was majority voting (Kimura and Shridhar, 1991, Franke and Mandler, 1992). If the outputs were based on a posteriori probabilities then an average or a linear combination method would be applied (Hashem and Schmeiser, 1995, Kittler *et al.*, 1997). When the classifier outputs are fuzzy membership values then other methods

such as belief functions or fuzzy rules are used (Tresp and Taniguchi, 1995, Rogova, 1994).

In order to construct an ensemble of classifiers it was also necessary to determine how big the ensemble should be. Early work on ensembles suggested that ensembles with as few as ten members were adequate to sufficiently reduce test-set error (Hansen and Salamon, 1990). To confirm this suggestion Opitz and Maclin (1999) evaluated 23 different datasets using neural networks and decision trees. For neural networks both bagging and boosting showed that much of the reduction in error occurs after ten classifiers, in some cases 15. A similar conclusion was reached for bagging and decision trees, which is consistent with Breiman (1996). Boosting for decision trees also showed this tendency although the reduction of error continued at a very small rate until the 25th iteration was reached. Yuan and Cho (2006) use a 10-member SVM ensemble in their research as an optimal size for obtaining a significant error reduction. Taking into account all the above it was concluded that to test whether this approach could obtain a higher classification accuracy than the one already obtained by the SVM, an ensemble of 10 members should be sufficient. The starting training set consisted of 150 pixels dataset which is above the recommended minimum training size ($10-30p$). In order to compare the results with those obtained by a simple SVM classifiers such as the one used in Chapter 4, the same testing set of 250 pixels was used.

Therefore, the method followed to construct the ensemble is as shown in Figure 6.4 below:

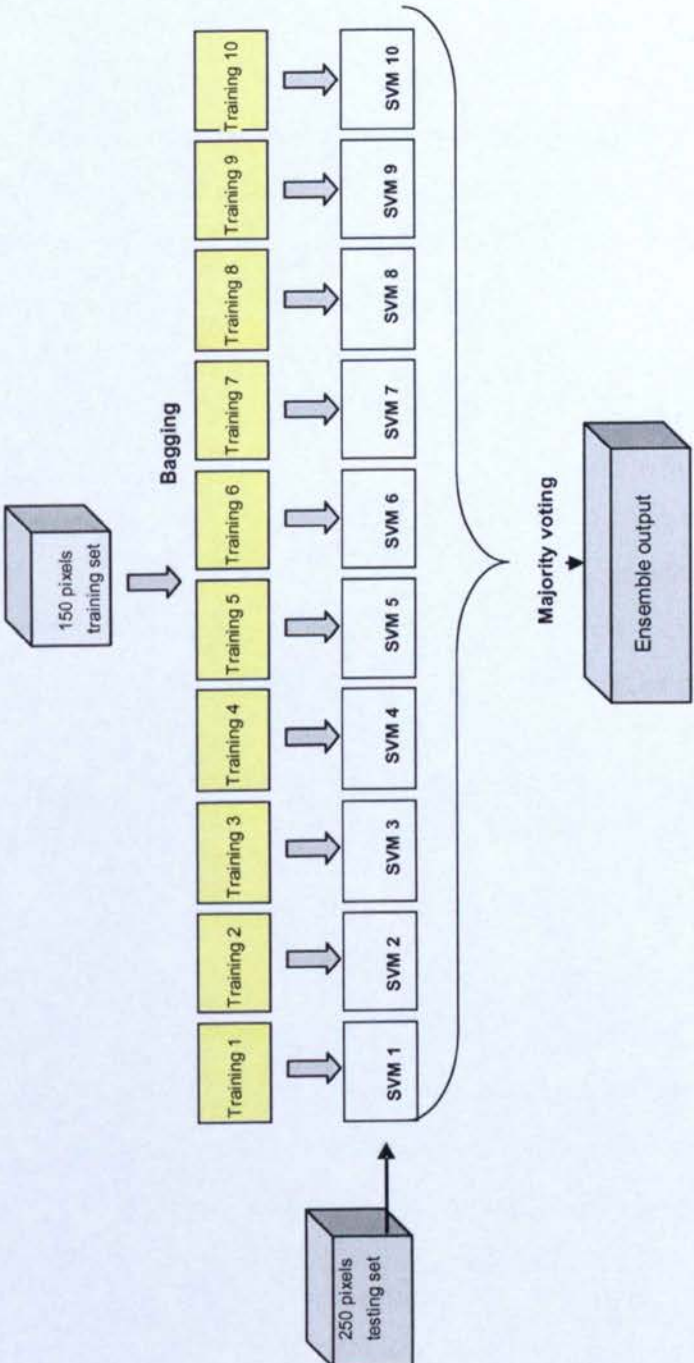


Figure 6.4 Schematic representation of the SVM Ensemble of classifiers

When comparing the error matrix of the SVM and the SVM ensemble (see Table 6.1 and Table 6.2 below) the user's accuracy for fen was definitely higher when using the ensemble (94.40% as opposed to 86.60%). However, the producer's accuracy was reduced from 98.40% to 94.40% in the ensemble.

SVM Ensemble		Predicted			
		Fen	Other	Σ	Producer's accuracy
Actual	Fen	118	7	125	94.40%
	Other	7	118	125	94.40%
	Σ	125	125	250	
	User's accuracy	94.40 %	94.40 %		Overall 94.40%

Table 6.1 Error matrix for the SVM ensemble using fen as the target class

SVM		Predicted			
		Fen	Other	Σ	Producer's accuracy
Actual	Fen	123	2	125	98.40%
	Other	19	106	125	84.80%
	Σ	142	108	250	
	User's accuracy	86.60 %	98.14 %		Overall 91.60%

Table 6.2 Error matrix for the SVM using fen as the target class with a training size of 150 pixels

The reason for this reduction in producer's accuracy could be due to the fact that bagging forms the new datasets from the original one by sampling with replacement. This means that some support vectors can be omitted from the training datasets which could contribute to lower producer's accuracy when the classifier is confronted with testing data. To illustrate this, the results for the producer's accuracy of the different bagged training datasets are shown in Figure 6.5:

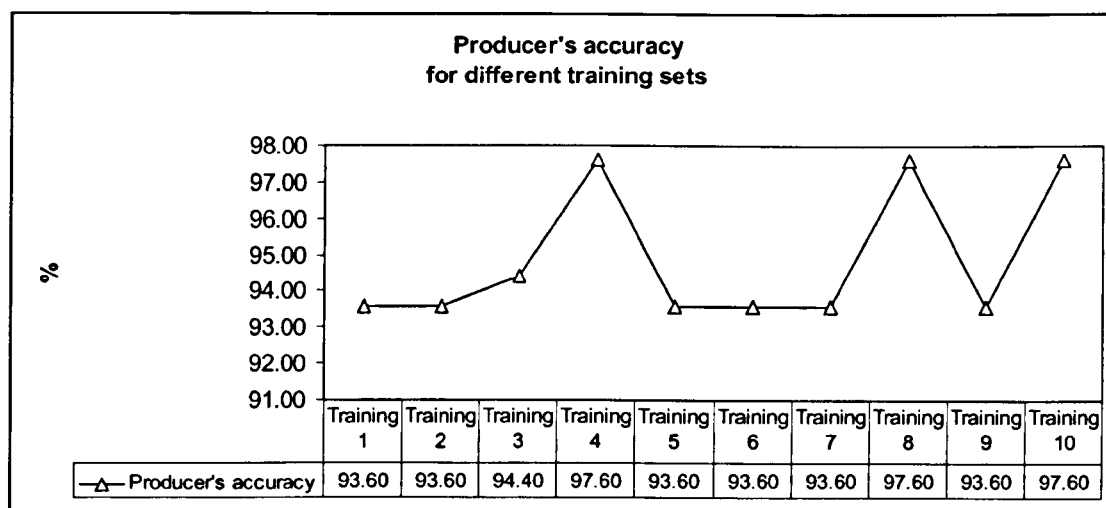


Figure 6.5 SVM ensemble producer's accuracy results

As it can be seen, there were a few cases where the producer's accuracy was as high as 97.60%. However, the major parts of the training sets generate lower accuracies and consequently when performing the majority voting the final producer's accuracy for the ensemble ended up being lower than that of the simple SVM classifier. Although the increase of overall accuracy could be seen as an advantage, the reduction on producer's accuracy is a clear disadvantage.. As explained in Chapter 3, producer's accuracy is very important as it determines the capacity of the classifier for identifying the class of interest on the ground.

6.2.2 Decision Trees Ensembles

The flexible nature of decision trees has supported recent advances in the field of machine learning such as ensemble of classifiers. Widely applied methods to create an ensemble of DT are boosting and bagging (Freund and Schapire, 1997, Quinlan, 1996). In bagging, several decision trees are created from random subsets of the training data and the final result is produced from a majority vote by all the trees. Boosting creates a series of decision trees in an iterative way, with each successive tree focusing on the errors of the previous tree. A boosted tree is then produced by voting among the different trees that have been created (Friedl *et al.*, 1999).

As seen in Chapter 4, in remote sensing classification and land cover mapping decision trees have been applied to images generated from MODIS data (Friedl *et al.*, 2002), AVHRR (DeFries *et al.*, 1998; Hansen *et al.*, 2000), and Landsat Thematic Mapper (TM) and Enhanced Thematic Mapper Plus (ETM+) (Lawrence and Wright, 2001, Pal and Mather, 2003, Lawrence *et al.*, 2004). The benefits of decision trees have also been demonstrated with multitemporal Landsat ETM+ data (Brown de Colstoun *et al.*, 2003).

Their use within land cover mapping is further supported due to the flexibility for handling continuous or categorical variables, ancillary or missing data and for obtaining higher accuracies due to the use of ensemble methods (DeFries and Chan, 2000). Therefore decision trees are increasingly being used for analysis and classification of remotely sensed imagery (Brown de Colstoun and Wathall, 2006). However, they present a few problems that have activated further research in the area in order to obtain higher accuracies and this has lead to the development of ensemble methods. These problems include : (i) DT classification do not necessarily produce the optimal tree (ii) inaccuracies in the training data can greatly affect the decision tree as they can represent a great portion of the variability of the data and (iii) an unbalanced dataset can also affect the performance of a DT (Friedman, 2001). Therefore, methods such as boosting and bagging, have recently been developed to address these limitations (Bauer and Kohavi, 1999, DeFries and Chan, 2000, Friedl *et al.*, 1999).

Boosting increases classification accuracy in many cases and in others the results are comparable to other ensemble methods (Freund and Schapire, 1996, 1999, Opitz and Maclin, 1999) but it does not address the issues of inaccurate training data or unbalanced datasets (Bauer and Kohavi, 1999). Bagging also shows higher accuracies when compared with a simple DT classifier and as mentioned in the previous section it is particularly effective when dealing with small datasets.

Very recently, a new approach to ensembles of decision tree classifiers has been proposed by Breiman (1999). This approach is called Random Forests and they use a

similar but enhanced method of bootstrapping (Gislason *et al.*, 2004). Ham *et al.*, (2005) applied Random Forests to the classification of an hyperspectral remote sensing image. Their results were good for a study that had a very limited training dataset. Gislason *et al.* (2004) applied this ensemble method to the classification of remote sensing data using multisource remote sensing and geographical data and its accuracy was comparable with that of other ensemble methods such as bagging and boosting. It is however specially designed for high dimensional data so it was not applied within this thesis.

Considering all the above, the approach taken to construct the DT ensemble was based upon bagging and followed the same structure as the SVM ensemble (see Figure 6.6):

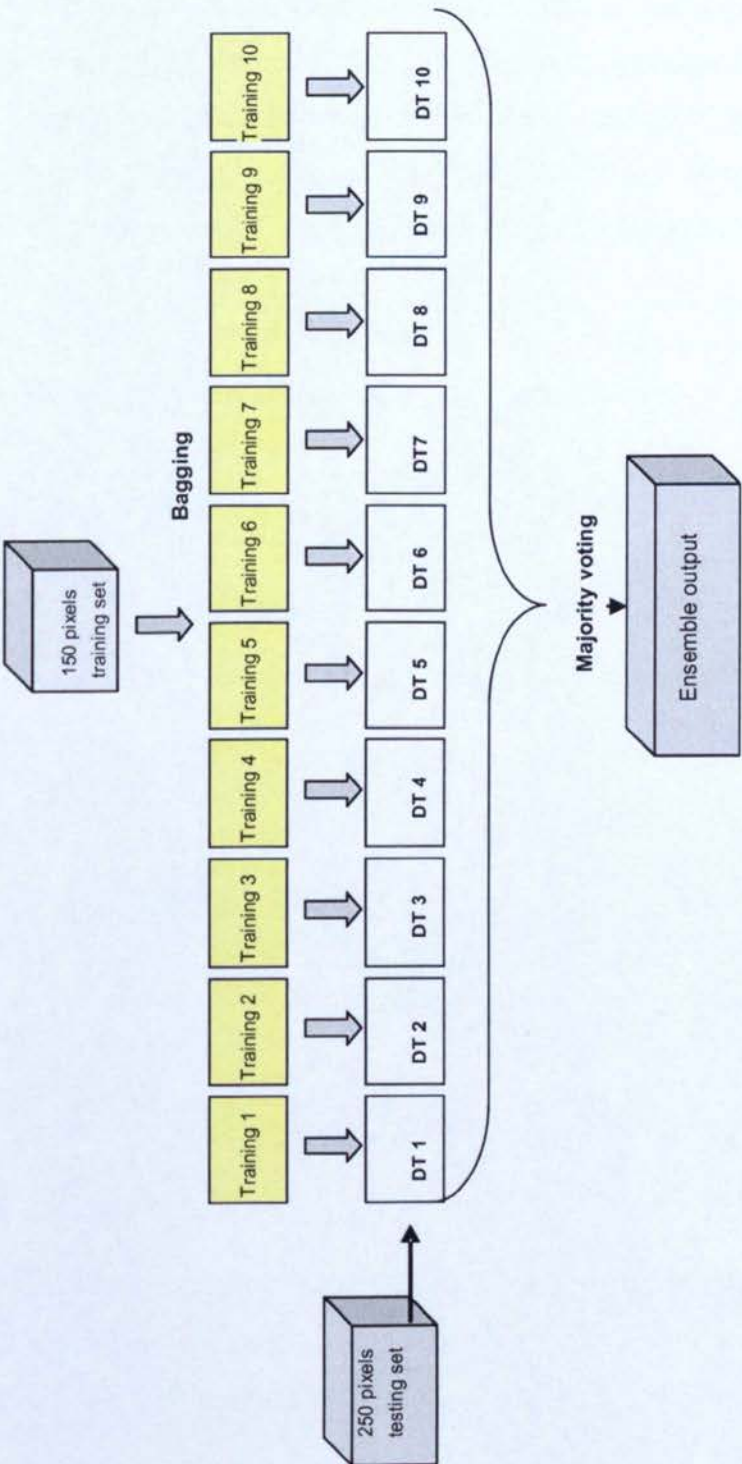


Figure 6.6 Schematic representation of the DT Ensemble of classifiers

After performing the majority voting of all the results the overall accuracy was 94.4% which showed an increase when compared with the simple DT classifier used in Chapter 4 with an accuracy for the same training data size of 91.60%. Comparing the two error matrices (**Error! Reference source not found.** and Table 6.4 below) it can be observed that the ensemble obtained an error matrix with the same values for producer's and user's accuracies of 94.4%. This shared the same trends than the SVM classifier with an increase in the user's accuracy and a decrease in the producer's accuracy of the DT ensemble with respect to the simple DT classifier.

DT Ensemble		Predicted			
		FE	Other	Σ	Producer's accuracy
Actual	FE	118	7	125	94.4%
	Other	7	118	125	94.4%
	Σ	125	125	250	
	User's accuracy	94.4%	94.4%		94.40%

Table 6.3 Error matrix for the DT ensemble using *fen* as the target class

DT		Predicted			
		FE	Other	Σ	Producer's accuracy
Actual	FE	123	2	125	98.40%
	Other	19	106	125	84.80%
	Σ	142	108	250	
	User's accuracy	86.60 %	98.14 %		91.60%

Table 6.4 Error matrix for the DT using *fen* as the target class with a training size of 150 pixels

In the case of DTs this decrease in producer's accuracy could be again due to the nature of the bagging method. This is, DTs classifiers are quite dependant on the characteristics of the training dataset. In this sense, bagging could be providing datasets in which the variability of the target class is not well represented and therefore affects the capacity of the DT for identifying such class.

Furthermore, the above results showed that both SVM and DT ensembles obtained exactly the same solution. When checking the results in detail, exactly the same pixels were being misclassified. When plotting these points in a feature space (Figure 6.7 below) it became apparent that they were in the borders of the spectral signature of the class fen.

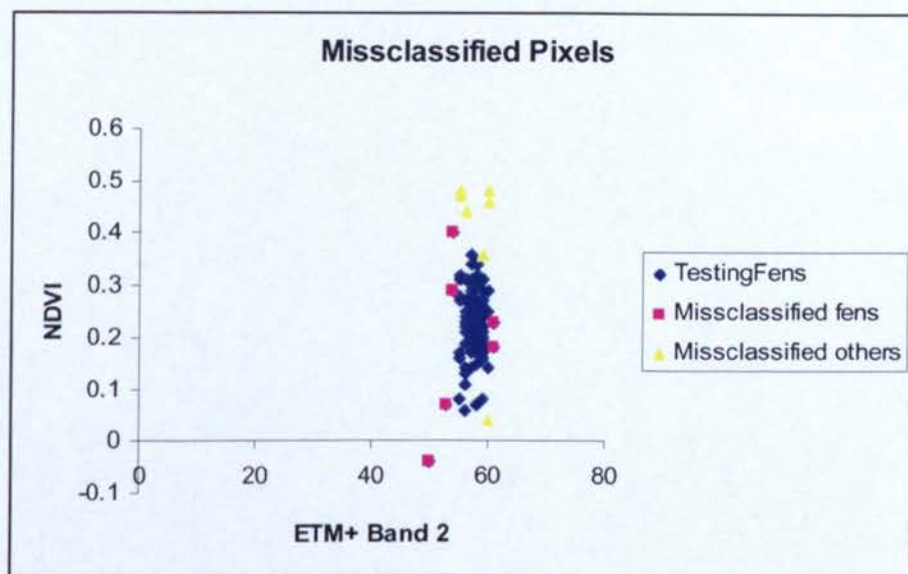


Figure 6.7 Misclassified fen pixels in the feature space

As it is possible to associate these pixels to their coordinates on the ground correctly it would be easy to get ground data and confirm whether they belong to the class of interest or not (see Table 6.5 below).

X	Y	ETM+ B2	NDVI	Misclassified as class
636491	323042	53	0.07	Fen
636865	321036	54	0.4	Fen
641557	321206	50	-0.04	Fen
636967	319846	54	0.29	Fen
636559	319608	61	0.23	Fen
636593	319608	61	0.18	Fen
636661	319540	61	0.18	Fen
591917	334228	55	0.47	Others
570973	338002	56	0.44	Others
577501	344666	60	0.04	Others
632921	306178	60	0.46	Others
632717	306076	60	0.48	Others
587089	341912	55	0.48	Others
608271	329366	59	0.36	Others

Table 6.5 Coordinates and details of misclassified pixels in the SVM and DT ensembles

6.3 Ensembles of one-class classifiers

As with the binary classification problems, a single one-class classifier might not be sufficient to exploit the discriminative characteristics of the data and as a result the ensemble of classifiers has been the focus of recent investigation (He *et al.*, 2004). Although the amount of research regarding ensembles of classifiers is quite abundant in binary and multiclass classification problems as seen in previous sections, one-class classification ensembles have not been explored in depth.

Two different ways of combining one-class classifiers have been proposed by Tax and Duin (2004). The first consists of combining classifiers trained on different datasets; the second is to combine different classifiers trained on a common training dataset. Their research indicated that the first option gave better results than the second and this could be due to the fact that different datasets contain more information than different views of the individual classifiers on one dataset. They

based their research upon estimated posterior probabilities and therefore, did not use the majority voting approach in their comparison of different methods.

As with the previous classifiers, using labels as outputs (class of interest, other class), the ensemble of one-class classifiers will be combined using majority voting. Furthermore, in the case of one-class classifiers, outlier distribution is unknown and prior probabilities and posterior probabilities are difficult to estimate. Therefore, the ensembles of classifiers that use combination rules based upon these estimated probabilities do not get better results than those of the individual classifiers (Tax 2001).

Taking into account all the above, the ensemble of one-class classifiers used the output labels of the classifiers and performed the majority voting method. This kept the analysis in line with the ones carried out for the binary classifiers in the previous section. Also following the approach of Tax and Duin (2004) the ensembles were constructed using two approaches: (i) combining different one-class classifiers trained with the same data and (ii) combining the same one-class classifier trained with different datasets.

(i) Combining different classifiers trained on a common training set

The first ensemble of classifiers was formed by the one-class classifiers that were assessed in Chapter 5 with the exception of the K-centre classifier that did not achieve an overall accuracy of 50% for fen as class of interest and as such would not add anything in order to increase the accuracy of the ensemble.

When constructing the ensemble and performing a majority voting with the rest of the classifiers (Figure 6.8) the outcome was an overall accuracy of 91.2% with a producer's accuracy for fen of 88% and user's accuracy of 94.02% (see Table 6.6).

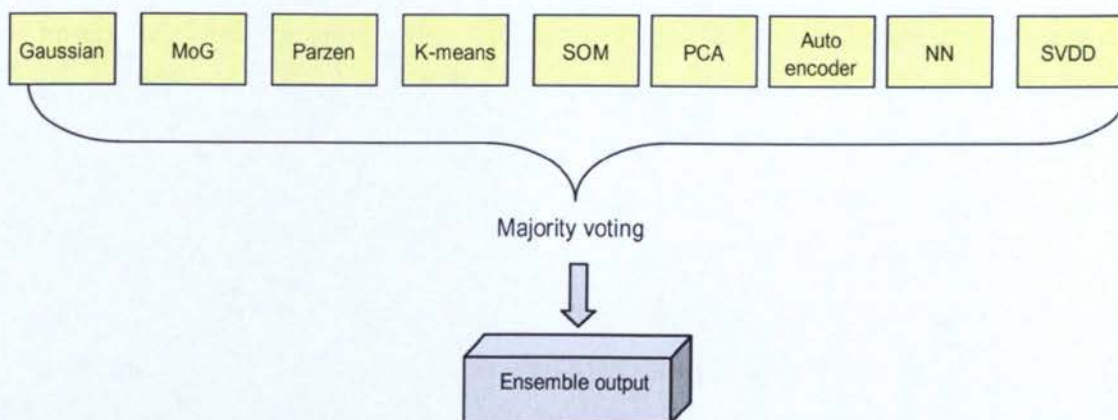


Figure 6.8 Schematic representation of the One-class classifiers Ensemble

One class classifiers ensemble		Predicted			
		FE	Other	Σ	Producer's accuracy
Actual	FE	110	15	125	88%
	Other	7	118	125	94.40%
	Σ	117	133	250	
	User's accuracy	94.02%	88.72%		Overall 91.20%

Table 6.6 Error matrix for the one-class classifiers ensemble using *fen* as the target class

The overall accuracies obtained by the ensemble were very close to those obtained by the density classifiers and were exactly the same as the one obtained using only the SVDD classifier which confirms the findings of Tax and Duin (2004) that the ensemble of different classifiers trained with the same dataset normally equals the performance of the strongest classifier.

(ii) Combining one type of classifiers trained on different training datasets

In order to test whether using different training sets would make any difference upon the accuracy obtained by a one-class classifier, the technique of bagging was used once more due to the small datasets involved in this thesis. The classifier chosen was the focus of much of Chapter 5, the Support Vector Data Description. Once more the

ensemble's structure was the same as the one used in the previous sections (see Figure 6.9).

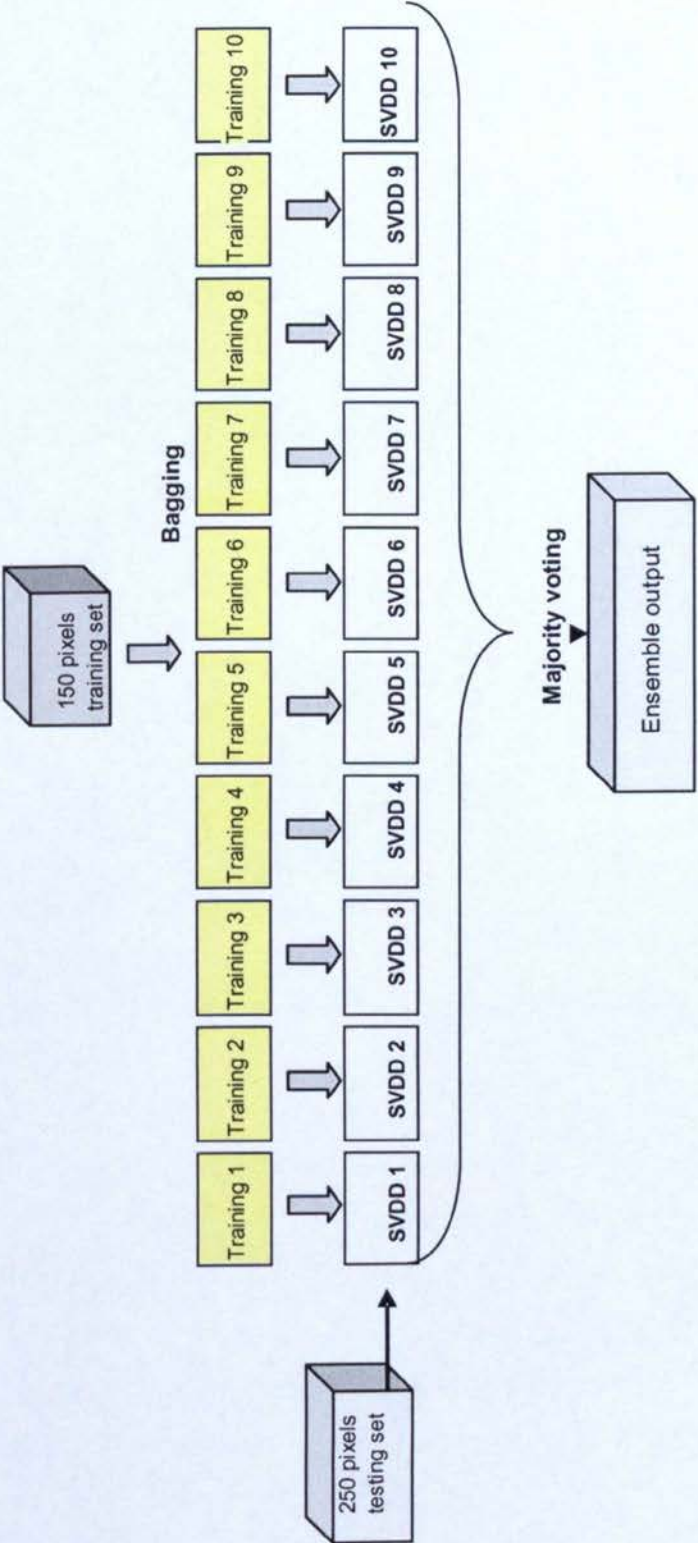


Figure 6.9 Schematic representation of the SVDD Ensemble of classifiers

When performing the majority voting the final overall accuracy was 96.9%, which meant a very slight increase in accuracy compared with the one obtained by the simple classifier with 95.2% (Table 6.8 and **Error! Reference source not found.** below).

SVDD Ensemble		Predicted			
		Fen	Other	Σ	Producer's accuracy
Actual	Fen	118	7	125	94.4%
	Others	2	123	125	98.4%
	Σ	120	130	250	
	User's accuracy	98.3%	94.6%		Overall 96.9%

Table 6.7 Error matrix for the SVDD Ensemble using fen as the target class

SVDD		Predicted			
		Fen	Other	Σ	Producer's accuracy
Actual	Fen	117	8	125	93.6%
	Others	3	122	125	97.6%
	Σ	120	130	250	
	User's accuracy	97.5%	93.8%		Overall 95.6%

Table 6.8 Error matrix for the SVDD using fen as the target class with a training size of 150 pixels

Looking in detail at the results, the 7 pixels belonging to the class fen misclassified as 'others' were the same than the ones misclassified by the SVM and the DT ensembles. However, the SVDD ensemble obtained a much better result when classifying the other class with only 2 misclassified pixels which also were misclassified by the other two ensembles (see Table 6.5 below).

X	Y	B2	NDVI	Missclassified as class
636491	323042	53	0.07	Fen
636865	321036	54	0.4	Fen
641557	321206	50	-0.04	Fen
636967	319846	54	0.29	Fen
636559	319608	61	0.23	Fen
636593	319608	61	0.18	Fen
636661	319540	61	0.18	Fen
608271	329366	59	0.36	Others
570973	338002	56	0.44	Others

Table 6.9 Coordinates and details of misclassified pixels in the SVDD ensemble

In conclusion to this section it can be said that the both ensembles (different one-class classifiers using the same training data and SVDD using a bagging technique on the training data) did not obtain significantly higher accuracies than the ones already obtained by simple classifiers.

6.4 Combining all classifiers

Going back to the first section of this chapter, it was established that an ensemble of classifiers has to be formed by a group of different classifiers and that their outputs should be comparable (Duin, 2002). It was also recognized that there were different methods of generating different classifiers such as: (i) different initializations (ii) different parameter choices (iii) different architectures (iv) different classifiers (v) different training sets and (vi) different feature sets (Duin, 2002). Although a great part of the research has concentrated on producing ensembles with different training sets, the combination of different classifiers can also contribute greatly towards the creation of an ensemble that could obtain higher accuracies than the single classifiers.

Over the past three chapters very different classifiers has been used in order to classify a class of interest. Some are multiclass classifiers (MLC), some are binary in

nature (SVM, DT) and finally, one-class classifiers. Because the training and testing of all of the classifiers had been carried out with the same datasets (only taking the data for the class of interest in the case of one-class classifiers) the results obtained by them were comparable. As a final experiment it was decided to combine all of them in one single ensemble and in that way test whether the advantages of binary and one-class classification combined could get higher classification accuracies. The ensemble of all these classifiers is represented in Figure 6.10:

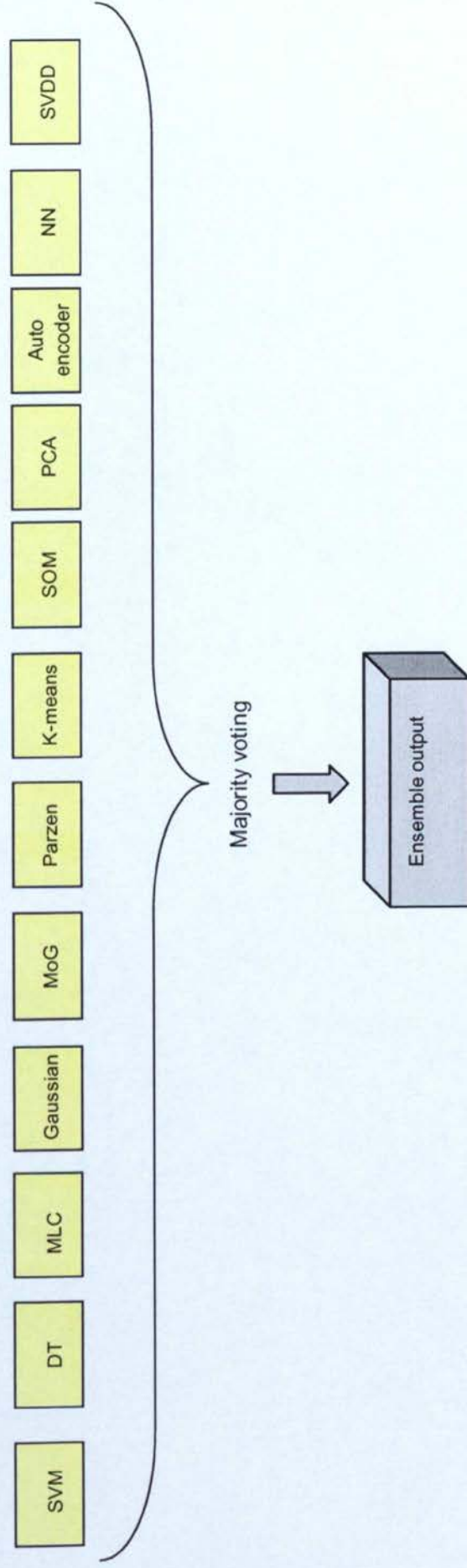


Figure 6.10 Schematic representation of the All classifiers Ensemble

Once again, the results of this ensemble gave an accuracy of 94.4% (see Table 6.10) which was not bigger than the accuracy obtained by the simple SVDD. The user's and producer's accuracy were also very similar to those obtained by the SVDD.

All classifiers ensemble		Predicted			
		Fen	Other	Σ	Producer's accuracy
Actual	Fen	115	10	125	92.00%
	Others	4	121	125	96.80%
	Σ	119	131	250	
	User's accuracy	96.64%	92.37%		Overall 94.4%

Table 6.10 Error matrix for the all classifiers ensemble using fen as the target class

6.5 Summary and Conclusions

The objective of this chapter was to test several ensembles of classifiers in order to assess whether this method could achieve higher accuracies than those obtained by the classifiers that had been already explored in Chapters 4 and 5. In section 6.2 the binary classifiers SVM and DT were used to create two ensembles in which the classifier diversity was guaranteed by using the bagging technique on the training set. The results obtained showed that

- (i) The overall accuracy of the SVM ensemble was 94.4% as opposed to 91.6% obtained when using the simple SVM classification in Chapter 4. The error matrices of both the SVM ensemble and the SVM (see Table 6.1 and Table 6.2) showed a higher user's accuracy of the class of interest fen when using the ensemble (94.4% as supposed to 86.6%). However, the producer's accuracy for fen was reduced from 98.4% to 94.4% in the ensemble.

- (ii) The overall accuracy of the DT ensemble was 94.40% which showed an increase when compared with the simple DT classifier used in Chapter 4 (91.60% overall accuracy). The two error matrices (Table 6.4 and **Error! Reference source not found.**) showed producer's and user's accuracies of 94.40%. This meant quite an increase in the user's accuracy but a decrease in the producer's accuracy with respect to the simple DT classifier.

These results demonstrate the efficiency of an ensemble of classifiers in improving the overall accuracy compared with that of the simple SVM and DT classifiers. However, the question is which results are more suitable to use for this particular land cover classification. As discussed in Chapter 4, ultimately it all depends on whether the final user considers a high producer's accuracy as the main objective and whether a low user's accuracy can be subsequently corrected with the use of ground data. Much of the decision on which classifier to adopt will probably be based upon the availability of these ground data. If these are not available, the ensemble option offers similar values for both accuracies and therefore a greater certainty that the pixels allocated to the class of interest are actually that class on the ground.

Regarding the ensemble of one-class classifiers the main results were:

- (i) When constructing an ensemble using different one-class classifiers training with the same dataset the overall accuracy obtained was 94.80% with a producer's accuracy for fen of 91.20% and user's accuracy of 98.28%, which were results very close to those obtained by a simple SVDD used in Chapter 5 .
- (ii) When constructing the ensemble using the SVDD with a bagging training set the results showed an overall accuracy of 96.90% which was slightly higher than the accuracy obtained by the simple classifier with 95.20% (Table 6.8 and **Error! Reference source not found.**).

- (iii) A final ensemble using all the classifiers described in Chapters 4 and 5 did not show a higher accuracy than that obtained by the simple SVDD.

As a general conclusion, the results of this chapter concur in part with the findings of many other researchers that outline the better performance of an ensemble of classifiers as opposed to a single classifier. The main reason for this success is believed to be due to the degree of diversity within the ensemble, which is a line of research very much alive with many researchers currently studying this idea in depth within pattern recognition (Kuncheva, 2005). It is true that for the binary classifiers there was an increase in overall accuracy and user's accuracy. However, the decrease in producer's accuracy could be a drawback when the objective of the classification is to obtain an accurate land cover map of a class of interest. In the case of one-class classifiers the ensemble technique failed to obtain a higher overall accuracy than the obtained by the SVDD. Furthermore, using the SVDD with the bagging technique did not produce a significant increase in any of the accuracies.

In this chapter, very simple experiments were carried out in order to test an ensemble of classifiers for land cover classification. Only bagging was used as a diversity measure using a small training dataset and only one ensemble size was assessed. Also only output labels was taken into account and consequently majority voting was the method used when combining the classifiers within the ensemble. Finally, the performance of the ensembles could have been affected by the small training dataset used and the low dimensionality of the data as highlighted by Skurichina *et al.* (2002), which definitely means that a lot more research is needed in this area.

To conclude, the following chapter will address a final discussion of the findings of this thesis, derive conclusions from these findings and identify future areas of research.

7 Final discussion, conclusions and further research

"Finally, in conclusion, let me say just this".

Peter Sellers

This thesis has investigated several classification methods in order to accurately classify a particular class of interest using remote sensing data. Focusing on a class of interest has been the focal point in other research areas such as those in pattern recognition but not yet commonly explored within remote sensing classification and land cover mapping. In this sense, this research was carried out within the context of current habitat conservation concerns in order to assess whether these classification methods would be suitable for mapping specific habitats protected by the EU Habitats Directive. The classification methods used for this purpose aimed to increase classification accuracy when concentrating on a particular habitat and to optimize the training process and avoid the waste of resources by (i) eliminating or significantly reducing the collection of data for classes that are of no interest (ii) optimising the performance of the classifiers by concentrating the classification on the class of interest using binary and one-class classification approaches.

The purpose of this final chapter is to assess whether the aims and objectives of this thesis were met. For that, the first part of the chapter consists of a final discussion comparing the performance of the different classifiers analysed in the previous research chapters for the classification and mapping of a specific habitat. Conclusions are drawn from this discussion and identification of areas of further research will conclude this thesis.

7.1 Final discussion: Comparison SVDD, SVM and DT classifiers

In Chapter 4, the outcomes of the classification performed by the binary SVM and DT classifiers were compared against that of a standard ML classification with the conclusion that both binary classifiers attained higher accuracy results than the ML classifier. In Chapter 5, different one-class classifiers were compared in order to assess their suitability for the classification of a habitat of interest with the conclusion that the SVDD classifier was the one that provided more advantages. The results obtained by this classifier were also compared against the performance of the standard ML classifier with the conclusion that the SVDD results also surpassed that of the parametric classifier (see Figure 7.1 and Figure 7.2 below).

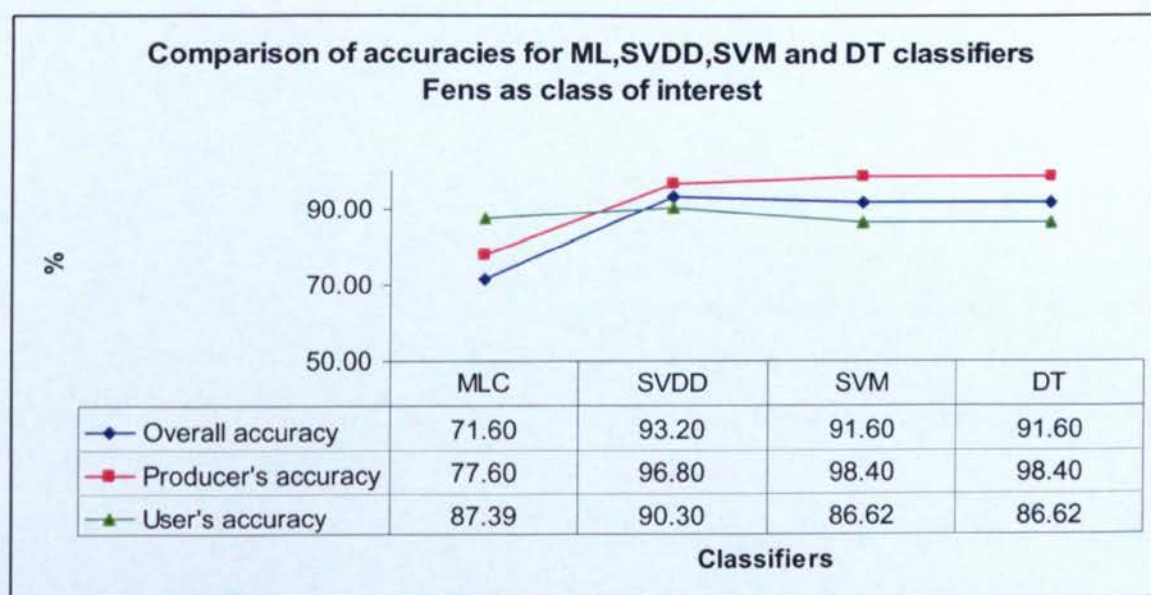


Figure 7.1 Comparison of accuracies for ML, SVDD, SVM and DT classifiers. Fen as class of interest

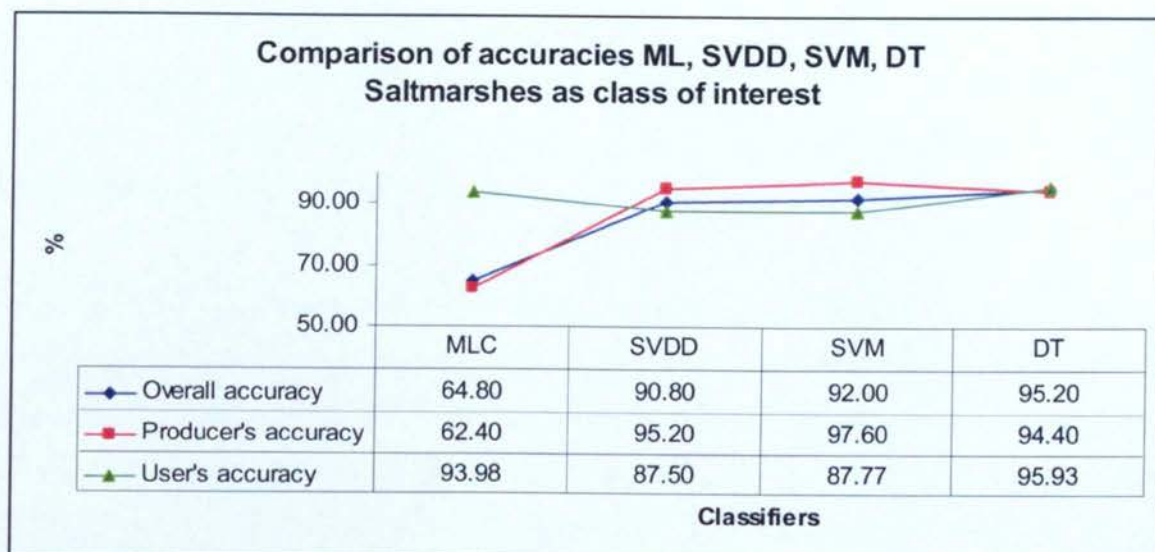


Figure 7.2 Comparison of accuracies for ML, SVDD, SVM and DT classifiers. Saltmarsh as class of interest

These results clearly stated that:

- 1) Both approaches (binary and one-class classification) were suitable for its application to land cover classification using remote sensing data
- 2) For the particular case of focusing on a class of interest, both approaches surpassed the results obtained by the ML standard classification
- 3) Both approaches used significantly less training data than the ML classifier

Therefore, having clearly demonstrated that these two approaches can be successfully applied to the classification and mapping of a particular habitat of interest under the requirements of the EU Habitats Directive, this final discussion compares the performances of binary classifiers against the one-class classifier and highlights advantages and disadvantages in the application of each of these methods.

For that, it was decided to carry out a comparison between overall, producer's and user's accuracies for the SVM, DT and SVDD classifiers with fen as the class of interest. It was not considered necessary to compare the results of the three classifiers for the class saltmarsh as the correspondent research chapters already demonstrated that the performance of these classifiers was not habitat specific although the high degree of

separability of the class of interest from all the other classes could produce higher accuracies as in the case of saltmarsh when using a binary classification. When performing the classification using the one-class classification approach the accuracies obtained for both classes were very similar.

The SVDD training sizes used for comparison were 15, 25, 50, 75, 100, 125 and 150. These sizes were compared against training sets of 30, 50, 100, 150, 200, 250 and 300 pixels for the SVM and DT. The reason for this is that in this way the comparison between the two types of classifier was done using the same amount of information about the class of interest. The results are shown in Figure 7.3, Figure 7.4 and Figure 7.5 below:

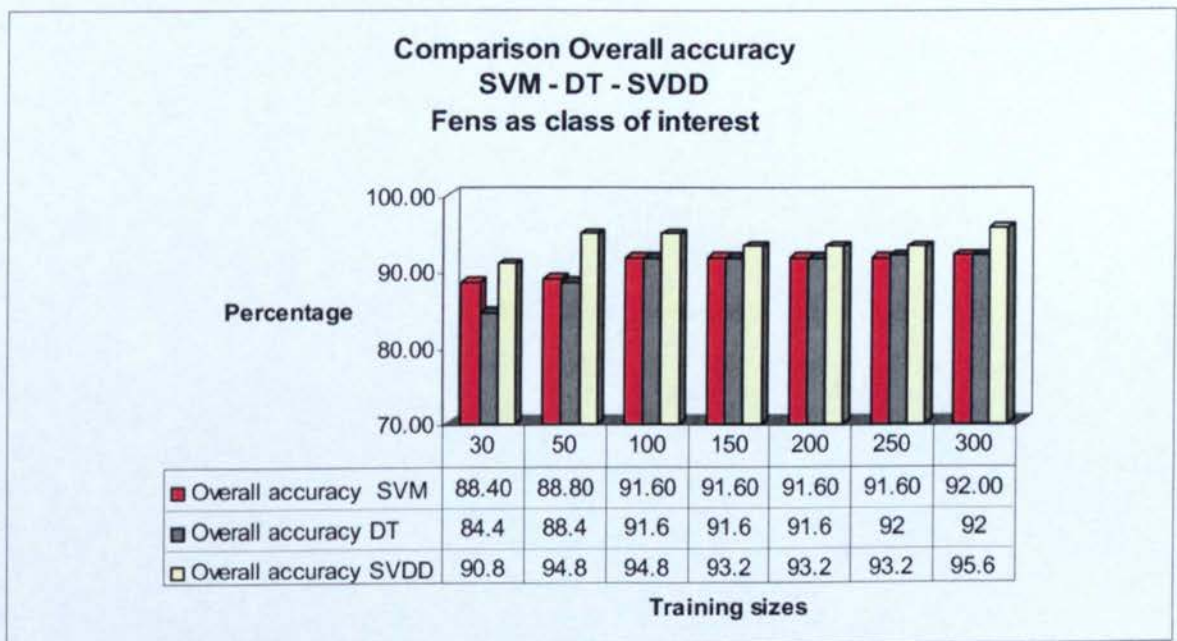


Figure 7.3 Comparison of overall accuracies SVM-DT-SVDD for fen as class of interest

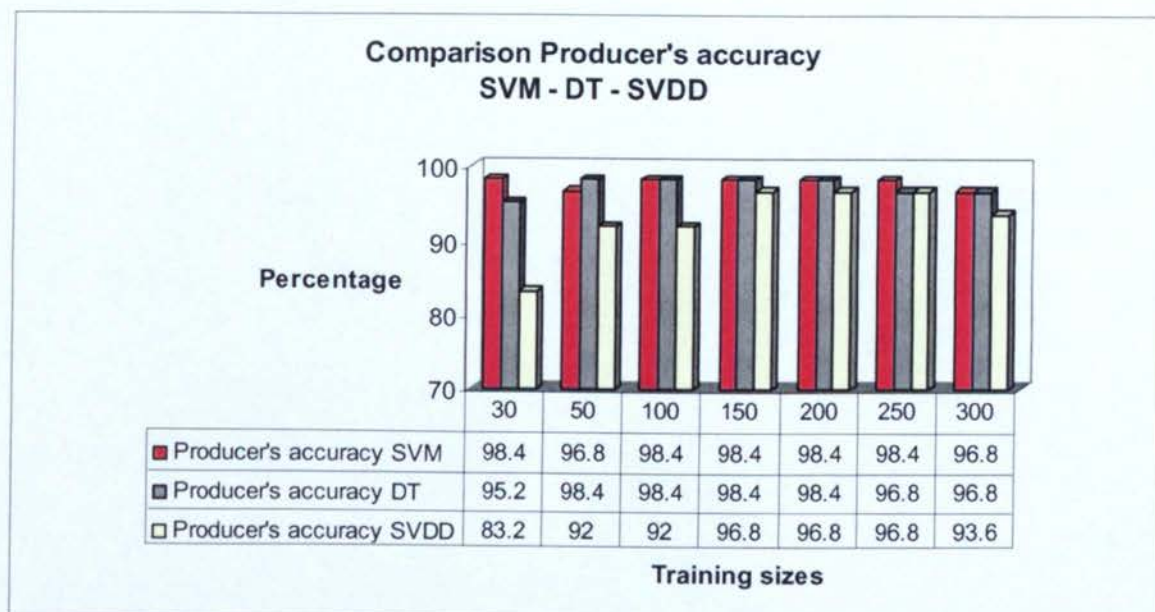


Figure 7.4 Comparison of producer's accuracies SVM-DT-SVDD for fen as class of interest

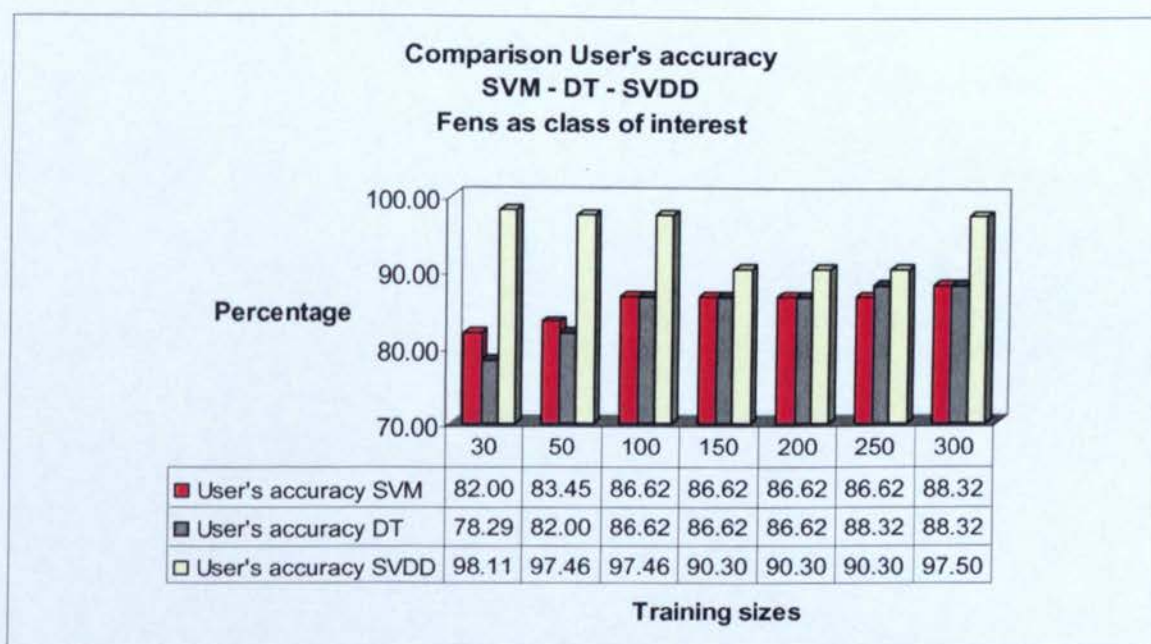


Figure 7.5 Comparison of user's accuracies SVM-DT-SVDD for fen as class of interest

Furthermore, the McNemar's test was performed in order to see whether the differences in accuracy were statistically significant ($Z > \sim 1.96$ at 95% interval)

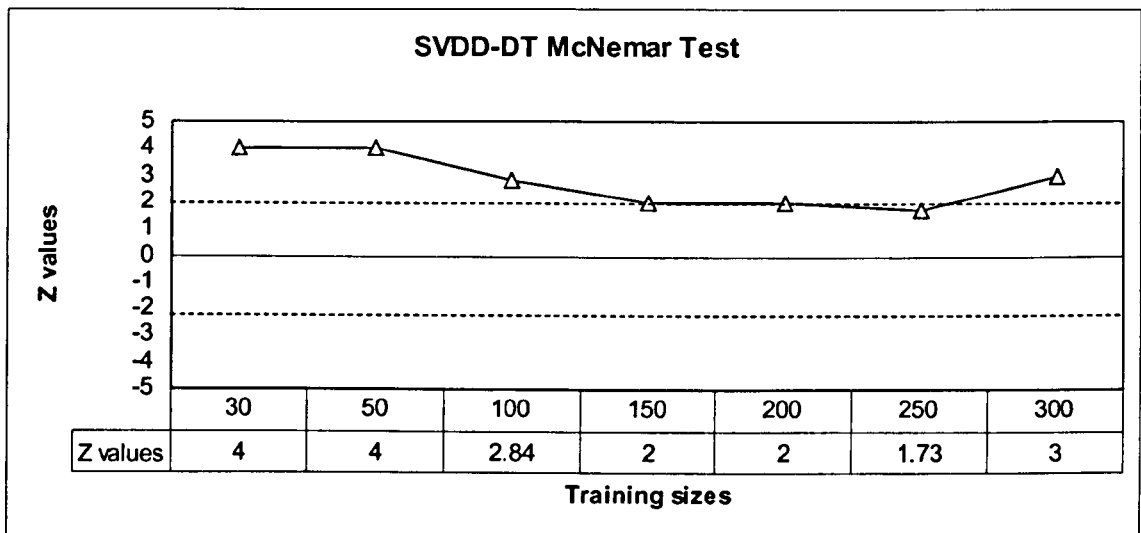


Figure 7.6 McNemar's test. Z values for SVDD-DT.

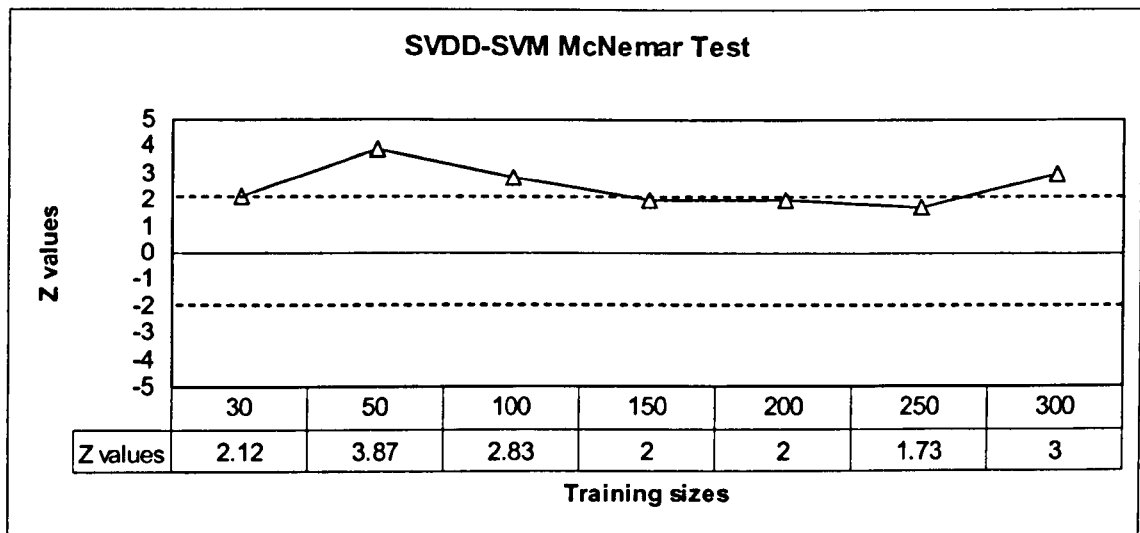


Figure 7.7 McNemar's test. Z values for SVDD-SVM.

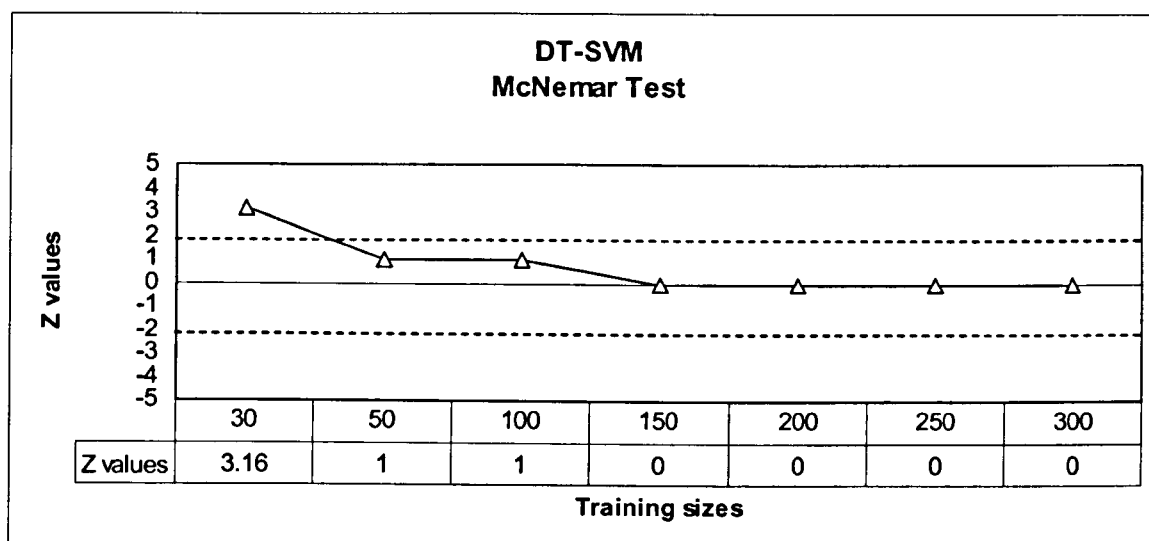
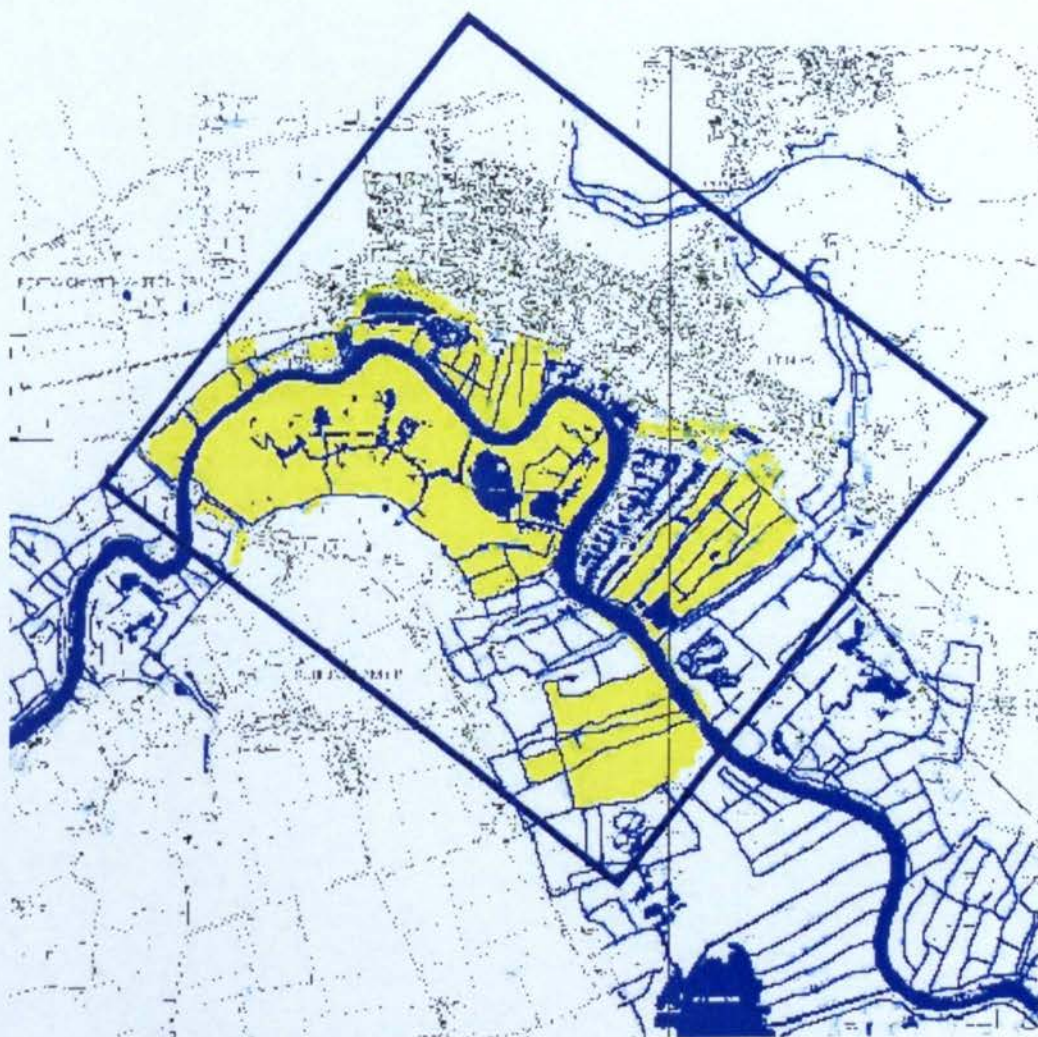


Figure 7.8 McNemar's test. Z values for DT-SVM

In terms of overall accuracy, the highest value was obtained by the SVDD when using the 150 training dataset with 95.60%. This compares to the highest accuracy of 91.60% obtained by the SVM and DT with 100 training pixels. Furthermore, when using 50 pixels as training data the SVDD still produced an overall accuracy of 94.80%. These were very encouraging results as the SVDD achieved these high overall accuracies only using target data to train the classifier. Having a closer look at the other accuracies, the producer's accuracy was higher when using the binary approach, which means that the classifiers actually identified a higher proportion of pixels as the target class which is of the utmost importance for this research. However, the user's accuracy showed that the SVDD was the one that classified the highest percentage of pixels being actually fens on the ground. The McNemar's test showed that these differences between the SVDD and the two binary classifiers were statistically significant at the 95% confidence interval (see Figure 7.6 and Figure 7.7) which reflected the different nature of these two approaches. The SVM and DT classifiers presented very similar results for all the accuracies and only showed a significant difference for the small training set of 30 pixels as seen in Figure 7.8.

When these classifiers were compared in the validation area of the River Yare NNR (see Figure 7.10 and Figure 7.11) the differences between the two approaches were obvious.

Test area River Yare NNR



Scale: 1:10000

- Fens*
- Water (derived from GIS data provided by OS @ Crown Copyright)*
- Others*
- Area covered by aerial photography*

Figure 7.9 Test area River Yare NNR

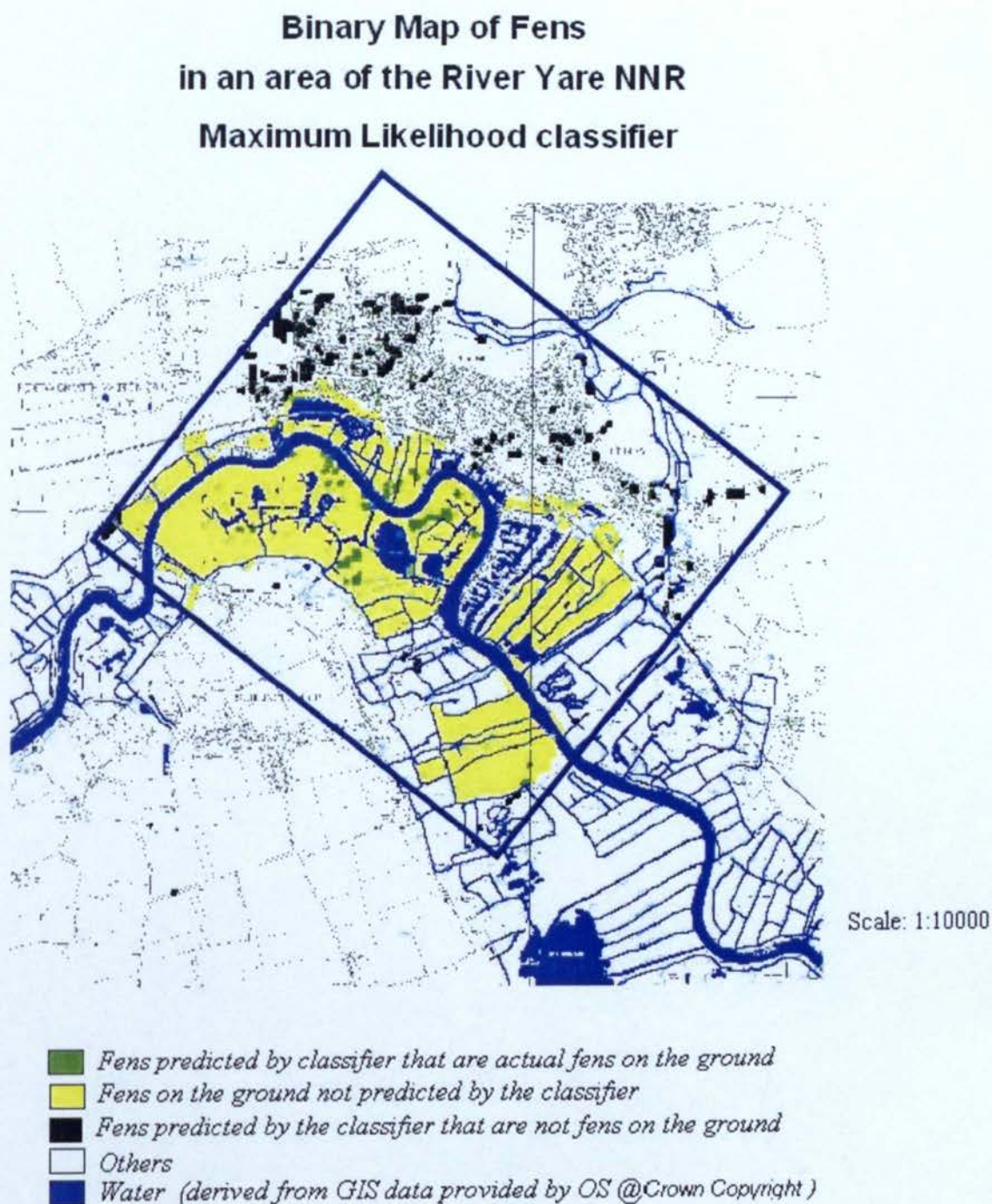


Figure 7.10 ML classification result for fens in the River Yare NNR test area

Binary Map of Fens in an area of the River Yare NNR Support Vector Machine classifier

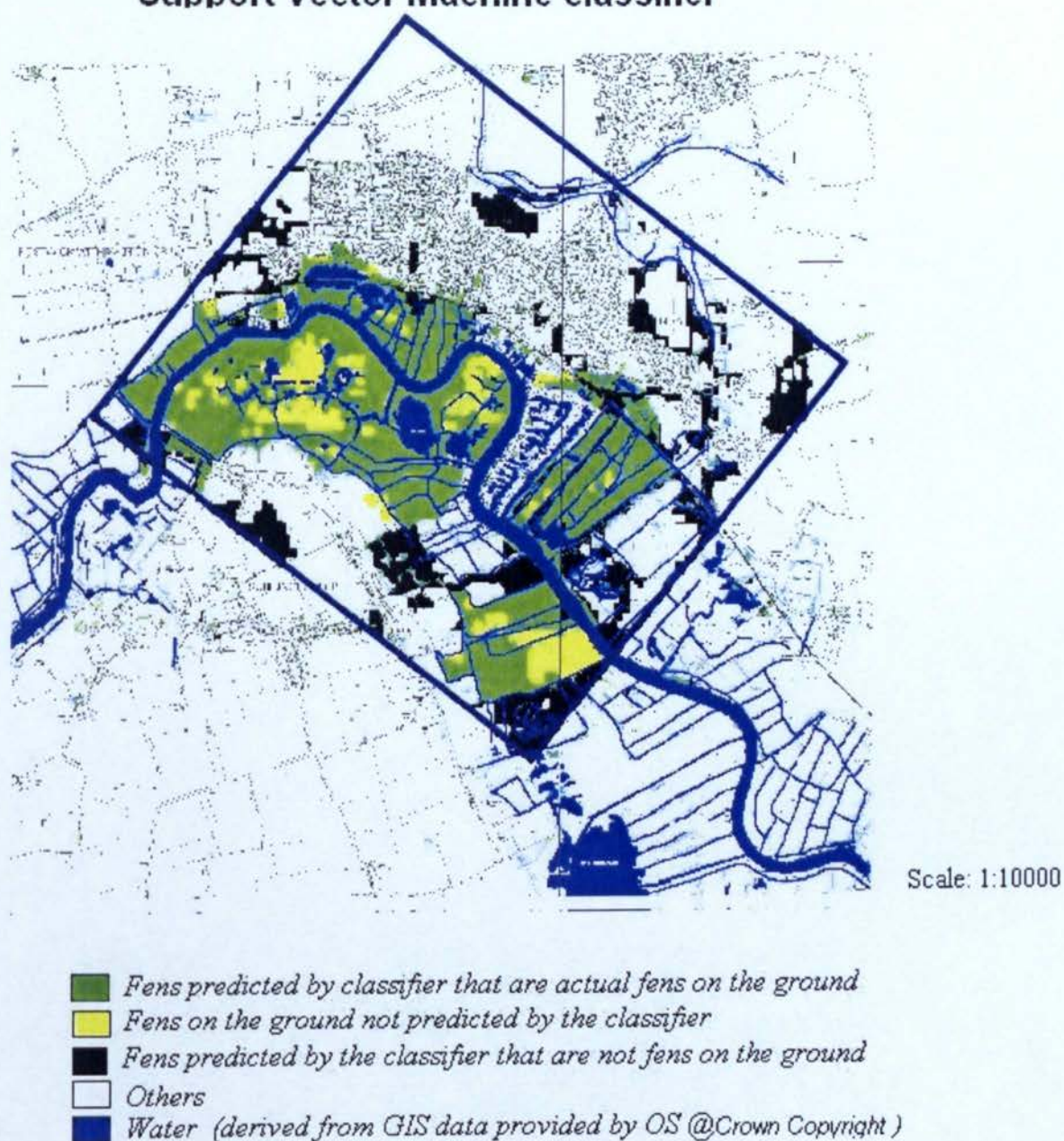
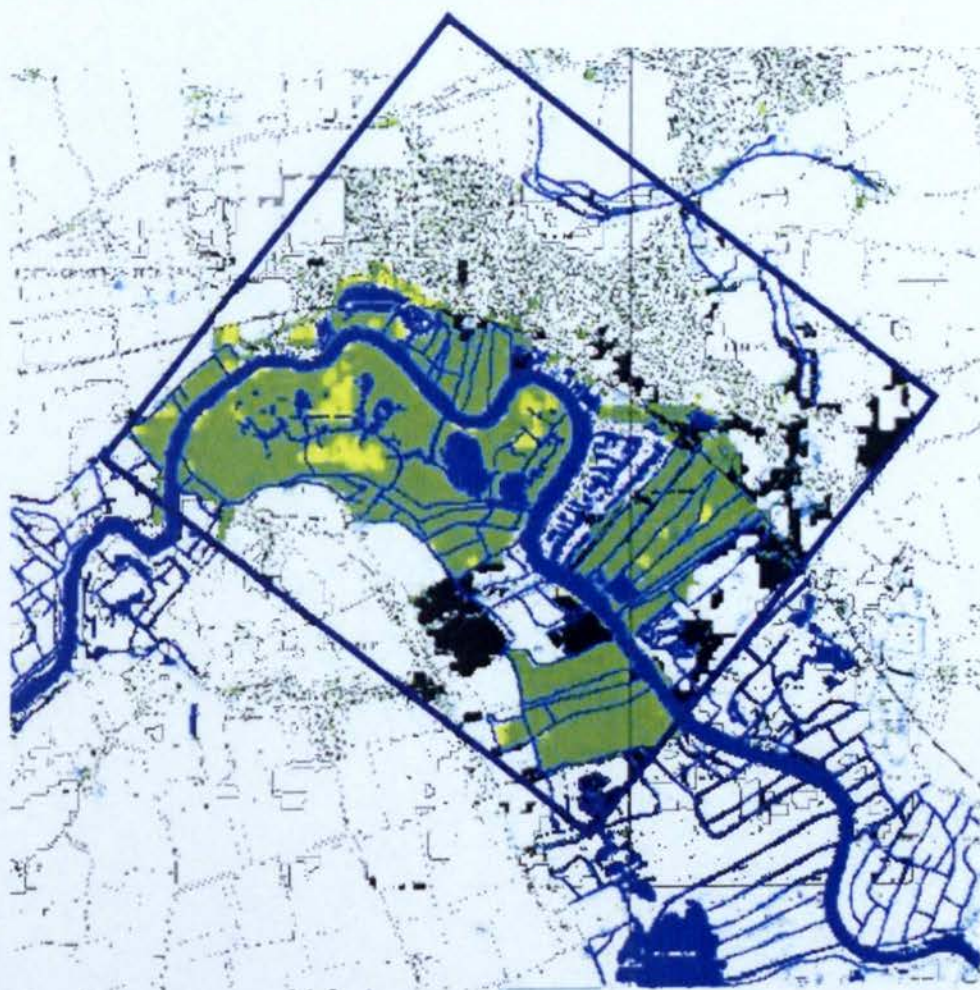


Figure 7.11 SVM classification result for fens in the River Yare NNR test area

**Binary Map of Fens
in an area of the River Yare NNR
Decision Tree classifier**



Scale: 1:10000

- Fens predicted by classifier that are actual fens on the ground
- Fens on the ground not predicted by the classifier
- Fens predicted by the classifier that are not fens on the ground
- Others
- Water (derived from GIS data provided by OS @Crown Copyright)

Figure 7.12 DT classification result for fens in the River Yare NNR test area

**Binary Map of Fens
in an area of the River Yare NNR
Support Vector Data Description classifier (A)**

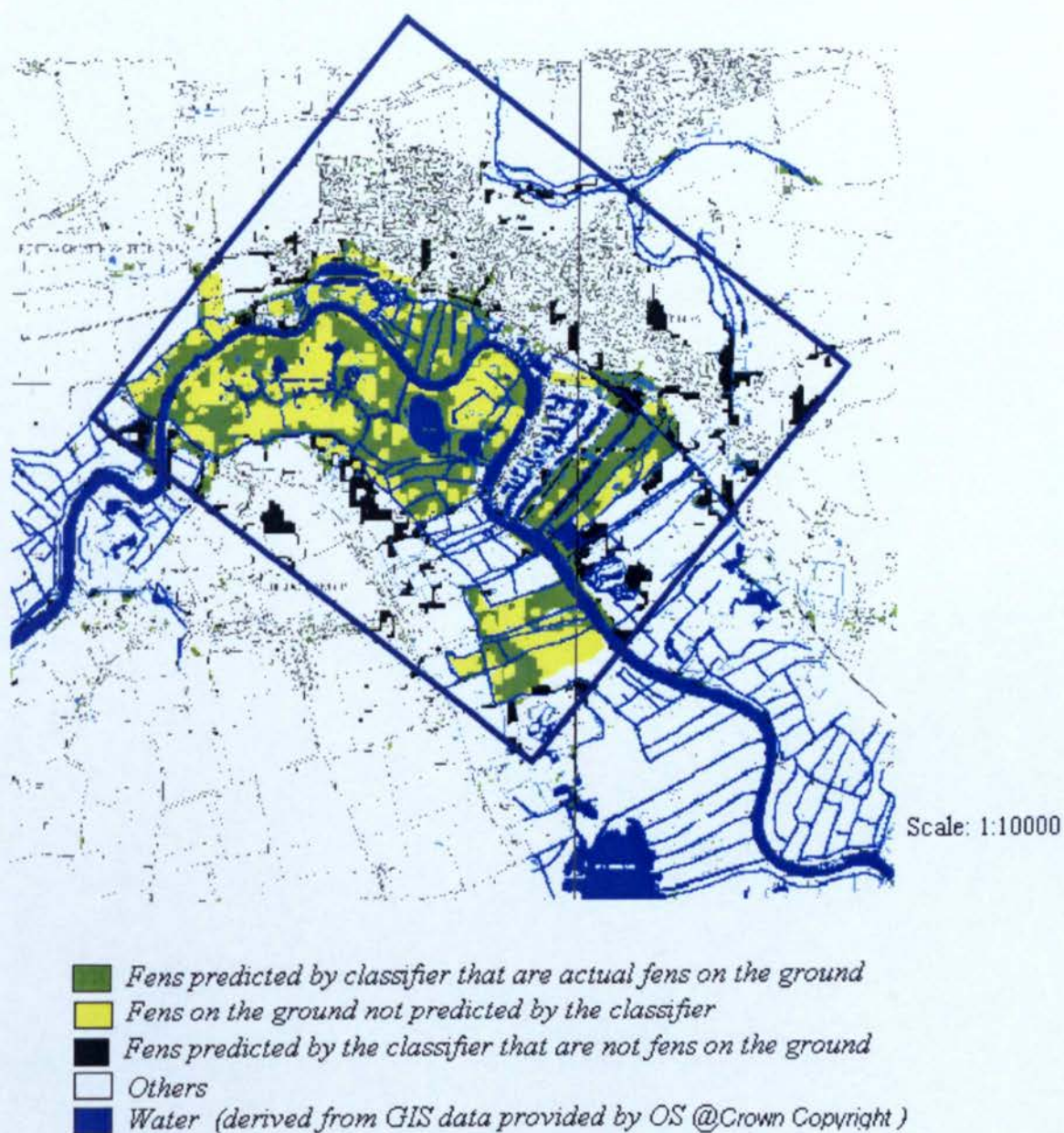


Figure 7.13 SVDD (A) classification result for fens in the River Yare NNR test area

**Binary Map of Fens
in an area of the River Yare NNR
Support Vector Data Description classifier (B)**



Scale: 1:10000

- Fens predicted by classifier that are actual fens on the ground
- Fens on the ground not predicted by the classifier
- Fens predicted by the classifier that are not fens on the ground
- Others
- Water (derived from GIS data provided by OS @Crown Copyright)

Figure 7.14. SVDD (B) classification result for fens in the River Yare NNR test area

The SVDD classifier performed much better than the standard ML classifier but the area correctly identified as fen was smaller than that of the SVM and the DT (See Figure 7.10, Figure 7.11, Figure 7.12 and Figure 7.13 above). The SVM, DT and SVDD maps were classified using a training set of 100 pixels for SVM and DT and 100 pixels for SVDD (A). The reason for the performance of the SVDD classifier when mapping the class fen in this area could be that the SVDD did not have enough training data about the class fen even though this training size produced high accuracies when testing it against a 250 pixel dataset. This could be due to the fact that this habitat is highly heterogeneous and consequently a 100 pixel training dataset might not define fully the variability of this class which can result in the definition of a very small radius for the optimal hypersphere. This could end up in overfitting and this could explain the results obtained for the area of the Mid River Yare NNR (See Figure 7.13). As described in Chapter 5, the SVDD has the advantage of being based upon the support vector theory and consequently the amount of data needed to describe the target class is not as large as with other one-class classifiers but still a good description of the target class is needed in order to find the optimal hypersphere without overfitting the training data.

To put the above to the test another map was produced using a training dataset of 300 pixels. The resultant map (see Figure 7.14) achieves a highly accurate classification of the area identified as fens by the aerial photography. This confirms the high potential of this one-class classifier for its application to land cover mapping focusing on a class of interest and further research is definitely recommended.

Regarding the other two classifiers, although they did not show any significant differences when tested with the 250 pixels testing set, when it came to the validation area the results showed that the DT classified the area more accurately than the SVM classifier (Figure 7.11 and Figure 7.12). Both of them had commission errors but the DT identified accurately more fens than the SVM. Maybe this is due to the fact that the DT software CART has the ability to focus on the class of interest when

performing the classification with a very strong binary split search which might produce better results when tested with the validation data. Also the training dataset might reflect very well the variability of the class of interest which is one of the factors affecting the performance of the DT classifier in order to discriminate this class from all the other classes. However, one very important advantage of SVMs over DT and other standard classification methods is that only a percentage of the training data is taken into account for the calculation of the optimal hyperplane that separates the class of interest from all the other classes. It is also possible to identify which are the exact pixels that are support vectors. Each of these pixels have associated X and Y coordinates, which means that it is possible to locate accurately in the field those important areas in order to train the classifier. Also it indicates which pixels from the “other” class are to be taken into account in order to find the optimal hyperplane. This potentially allows the researcher to:

1. discard the classes and set of pixels within the “other” class that do not contribute towards to final solution
2. concentrate on those locations of the class of interest and other classes that seem to have important characteristics in terms of training the classifier.

This has enormous implications for future research and refinement of the training and classification process as recently studied by Foody and Mathur (2004b) where the training can be directed to those specific pixels that act as support vectors optimising even more the training process and obtaining better accuracies.

In terms of simplicity and computational efficiency, the DT was the fastest and most user-friendly method of the three. The SVM and SVDD required much more time and effort in order to find the optimal parameters. These conclusions were also shared by Pal and Mather (2003) in their comparison of SVMs and DTs. Furthermore, in the case of this thesis, CART uses a Windows interface that was easy to use and that could be quickly mastered by any user. In the case of the SVM and the SVDD they both operate in the MATLAB environment which programming language has to be learned previously. The advantage of MATLAB is that it can

solve computing problems faster than with traditional programming languages, such as C, C++, and Fortran. Furthermore, the SVM Toolbox developed by Steve Gunn (Image Speech and Intelligent Systems Group at the University of Southampton) offered a graphical user interface within MATLAB specially designed for binary classification problems. The DD_tools MATLAB toolbox used for the SVDD did not provide any graphical user interface but, as it was based upon the principles of the SVM, it was easier to understand once the SVM and MATLAB programming had been mastered. In this sense, these characteristics could have implications for relevant authorities in order to adopt one method or another. In certain applications the use of SVMs and SVDDs could be more advantageous even though the training of personnel on these methods could take longer. In this sense, the decision as to which classifier to use for a particular application depends mainly on the advantages and disadvantages described in Table 7.1. The DT used in this thesis (CART) has the advantage that the analyst does not need to learn any programming language. Also, if the training data are well sampled and reflects the variability of the class it can produce very good results. SVM however gives an opportunity for development and the on-going mapping of a particular habitat can be greatly refined by the selection of appropriate training data that can be used as support vectors and therefore can be more appropriate for on-going monitoring programmes. Furthermore, using a SVDD has the great advantage of concentrating solely on a class of interest and training data are only needed for this class.

Summarising, the advantages and disadvantages of each of the above methods are as follows:

	Advantages	Disadvantages
SVM	<ul style="list-style-type: none"> • Performs well with small training datasets • It finds the global optimal solution • Generalises well • Only uses part of the training data for the calculations of the optimal hyperplane, the support vectors. Any other data can be absent and the results are the same. • Potential for refinement of the training process and increase in accuracy when focusing on specific classes 	<ul style="list-style-type: none"> • Classification obtained depends on the choice of parameter values such as kernel, kernel parameters and value of C. This choice is normally approximate. • Results show high error of commission that could be an issue if not enough ancillary data are available to correct them • It requires skills such as understanding MATLAB programming language before being able to train and test the classifier
DT	<ul style="list-style-type: none"> • Performs well with small training datasets and increases accuracy with training data set up to a point • It is statistically simple and easy to use (Windows interface) with no need of knowledge of programming language. • CART finds the optimal global solution by over-growing the trees and pruning them back • The time to train and test the classifier is very short compared to the other two classifiers 	<ul style="list-style-type: none"> • The classification obtained is mainly affected by the splitting rule and pruning method chosen • It relies on the characteristics of training dataset to train the classifier, especially the representation of the variability of the class by the training dataset • Results show high error of commission that could be an issue if not enough ancillary data are available to correct them
SVDD	<ul style="list-style-type: none"> • Only uses data from the class of interest, therefore optimising the training stage • It is based on the theory of support vectors and as such only uses part of the data for calculating the optimal hypersphere • It also finds the optimal global solution • The results obtained are not probabilities 	<ul style="list-style-type: none"> • The producer's accuracy values are not as high as those obtained by the binary classifiers • As with the SVM the classification obtained depends on the choice of kernel, kernel parameter and value of C. • It also requires skills from the analyst such as understanding the MATLAB programming environment

Table 7.1 SVM, DT and SVDD advantages and disadvantages

Finally, as discussed in Chapter 6, ensembles of classifiers were formed for each of these methods in order to assess whether this approach could increase the accuracies obtained by these classifiers. The conclusions for fen as the class of interest showed that overall accuracies were definitely higher (see Figure 7.10 below).

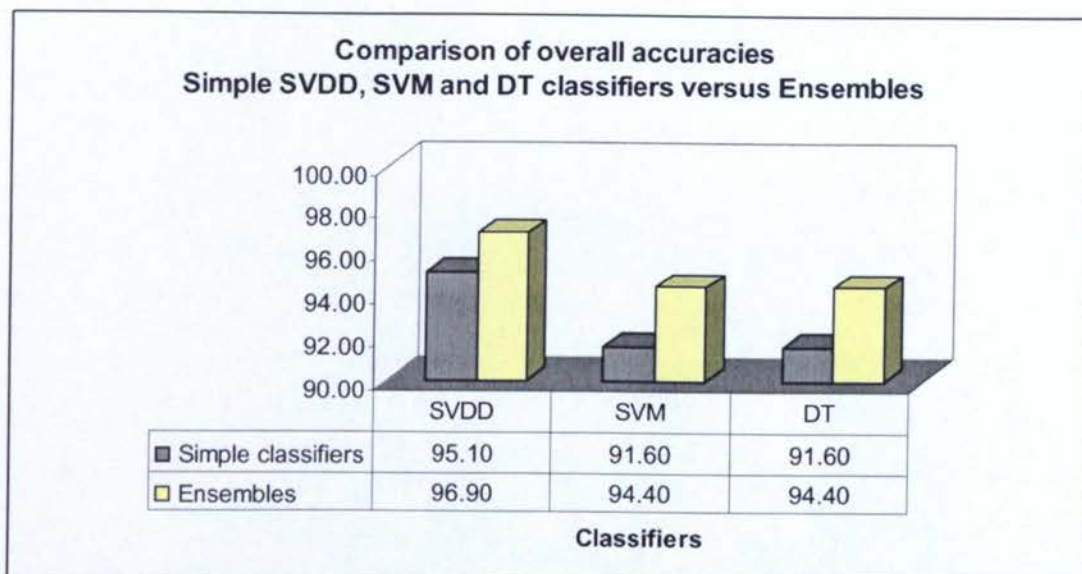


Figure 7.10 Comparison of overall accuracies between SVDD, SVM and DT classifiers and respective ensembles.

The ensembles also produced higher user's accuracies for the DT and SVM and slightly lower for the SVDD (see Figure 7.11 below).

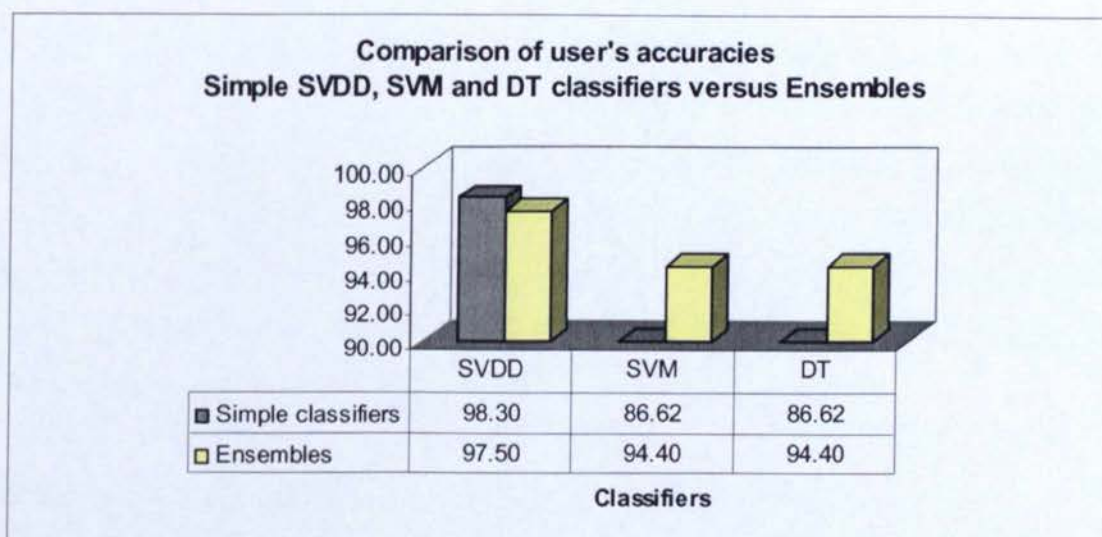


Figure 7.11 Comparison of user's accuracies between SVDD, SVM and DT classifiers and respective ensembles.

However, when comparing the producer's accuracies between the simple classifiers and the ensembles, the latter obtained lower producer's accuracies than the original

SVM and DT classifiers and the increase was marginal for the SVDD classifier (see Figure 7.12 below).

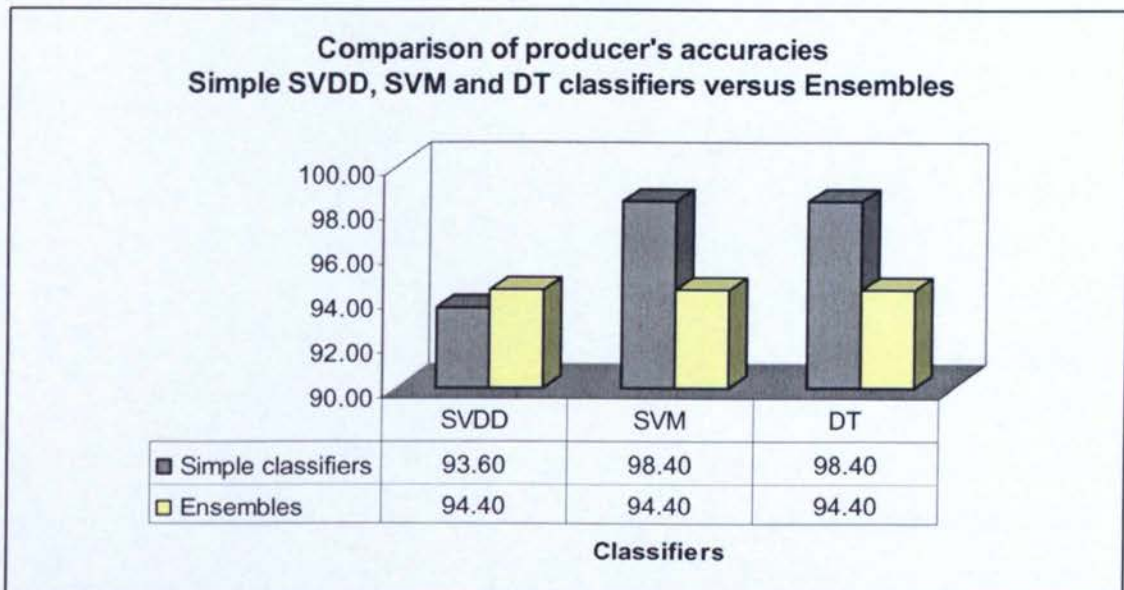


Figure 7.12 Comparison of producer's accuracies between SVDD, SVM and DT classifiers and respective ensembles.

It was therefore concluded that for this particular case study, the option of using ensembles of classifiers was not the most appropriate due to the decrease in producer's accuracy as this is considered to be of utmost importance for the purpose of this thesis. Also the small increase in producer's accuracy for the SVDD classifier did not compensate the computational effort involved. However, it could be an option suitable for other remote sensing studies and consequently further research in this area is recommended in the following section.

7.2 Conclusions and further research

As stated at the beginning of the Introductory chapter, the main aim of this thesis was to investigate and evaluate suitable methods for the accurate mapping of one particular habitat of interest with the aid of remote sensing. It also included the following sub-aims to increase the classification accuracy when focusing on a class of interest (i) optimising the use of training data and (ii) optimising the use of remote

sensing by applying suitable classifiers to the specific task of classifying a class of interest. These aims and sub-aims were addressed through three specific objectives that consisted of investigating the application of three alternative methods. These alternative methods consisted of a binary classification approach using SVM and DT classifiers and a one-class classification approach using specifically the SVDD classifier and the higher accuracy obtained by both approaches using the ensemble technique. These approaches were investigated in Chapters 4, 5 and 6 respectively, using remote sensing data from a Landsat ETM+ satellite image of East Anglia and focusing on two different habitats of interest. The habitat fen was the main focus of the investigation and saltmarsh was a second class of interest used to assess whether the performance of the classifiers was biased towards the specific characteristics of one class. The results showed that the behaviour of the classifiers was the same when using both classes, although results for saltmarsh were generally better than those obtained for fen as the class of interest. The reason for this could be that the degree of separability of the class saltmarsh against all the other classes is higher than the class fen. These results obtained by the binary and one-class classifiers were compared against those obtained by a standard parametric ML classification with the outcome of all the classifiers performing considerably better than the ML classifier. Consequently, these classifiers are not habitat specific and are perfectly suitable for its application to classifying and mapping a particular habitat of interest. This is particularly important in the case of the SVDD classifier as this was the first time that this was applied to remote sensing classification for land cover mapping (far as the author is aware).

Furthermore, the application of these classifiers aimed to optimise the use of training data. By choosing a binary approach the classification would only require very little data belonging to the other classes present in the image. And in the case of one-class classification only data from the class of interest were needed. Moreover, the binary nature of the SVM and DT classifiers meant that the attention was focused on separating the class of interest from all the other classes and therefore training efficiency was bigger than in a standard multiclass classification where efforts are directly to achieve a high overall accuracy. In the case of the SVDD it is only

necessary to provide enough information about the target class so that this classifier can generate a description of this class and consequently be able to distinguish it from any other possible class. The application of this classifier was particularly important because one-class classification has not been applied before to remote sensing classification. The choice of one or another for a particular application depends on the advantages and disadvantages of each of them as discussed in the previous section.

Finally, the technique of using an ensemble of classifiers was put to the test. This technique is also a very novel method that has only very recently been applied in remote sensing classification but has not yet been studied in detail. In this particular case, the results showed that the ensemble of classifiers obtained higher overall accuracy compared with that of the simple SVM, DT and SVDD classifiers. In the case of the SVM and DT ensembles this increase in overall accuracy was accompanied by a decrease in producer's accuracy. As producer's accuracy is considered to be of great importance because it reflects the capacity of the classifier to identify the class of interest on the ground, it was concluded that the ensemble was not the appropriate option for this case study. In the case of the SVDD the increase of accuracy was very small and not worth all the computational effort involved in constructing the ensemble.

All the above have practical implications from the point of view of authorities that have to comply with the requirements of the EU Habitats Directive. The protection of habitats listed in the Directive is closely related to obligatory monitoring of these habitats which implies accurate mapping for monitoring and impact assessment. Specifically, the Directive requires that within Natura 2000 sites measures are taken to maintain and restore these habitats to a "favourable conservation status" which roughly means that a species or habitats have to be in a stable or increasing state. Authorities at different levels dealing with these issues have an obvious challenge and attention is increasingly focused on the issue of management in accordance with the provisions of Article 6 of the Habitats Directive. However, proposed monitoring and management options have to deal with the ecological requirements of the

different protected habitats (which can vary significantly from one to another) and which in turn are also influenced by the economic, social and cultural requirements of the area (Keramitsoglou et al., 2005). For example, in the particular case of the Norfolk Broads, the Broads Authority's statutory duties try to balance navigation, nature conservation and recreation/amenity interests. This is a complex political, economic and environmental context that now has to incorporate the requirements of the EU Habitats Directives with a rather rigid regulatory interpretation of nature protection.

For the Broads Authority and many others, to incorporate the EU Habitats Directive into their statutory obligations means that a cost-effective and time consistent practice has to be developed. In this sense, for many of these authorities, traditional approaches for habitat mapping based upon field work are still thought to provide high accuracy at local level applications. However, Cherrill and McClean (1999) found that although standard methods of surveying are widely used by different researchers, agreement between pairs of maps in the UK averaged only 25.6% which represents a huge problem in terms of quality assurance of habitat mapping. Only recently, it has been recognised that integrating remote sensing data with field survey can increase the precision of habitat mapping (Cherrill and McClean, 1999). In that sense, aerial photographs offer the advantages of generally good availability, high quality and resolution and potential regional-scale coverage and they are now widely used by authorities at local and regional level such as the Broads Authority and the Environment Agency. However, they have some disadvantages such as their high cost and the temporal gap between different photographs which can vary from 3 to 6 years or more. Satellite imagery has been so far less used (or not at all) for terrestrial habitat classification and mapping because of cost, poor availability (e.g., in regions prone to regular cloud cover) and because resolutions maybe not appropriate for the mapping of specific habitats. However, technology has advanced in recent years and costs and availability issues have been overcome. Satellite remote sensing has the clear advantage of a high temporal resolution (16 days as in the case of Landsat) and different spatial resolutions that can meet local and regional monitoring needs. Therefore, there is definitely a great potential for satellite imagery to contribute more

and more to monitor habitat conservation (Mehner *et al.*, 2004, Kerr and Ostrovsky, 2003, Turner *et al.*, 2003, Read *et al.*, 2003, Mumby and Edwards, 2002, Nagendra and Gadgil, 1999).

Within this context, the binary and one-class classification methods studied within this thesis have shown a high suitability for mapping specific habitats in need of careful monitoring using satellite remote sensing data. The accuracies obtained by all of them surpass those of the standard parametric classifiers such as Maximum Likelihood classification. However, the ML classifier is still widely used as exemplified by the Land Cover Map of Great Britain (LCM2000). Based upon the results of this thesis, the use of DT, SVM and SVDD classifiers is highly recommended as an alternative to standard classifiers such as ML. Furthermore, these classifiers have proven to be highly suitable for classifying and mapping a specific habitat and its application should definitely be considered in future work involving the accurate mapping of protected habitats. These methods could be used to support and complement existing monitoring methods already in place meeting the needs of relevant authorities and ultimately the requirements of the EU Habitats Directive.

Taking all the above into account, this thesis has definitely opened up new areas for future research for classifying and mapping a particular habitat of interest. In terms of the different methods that have been explored further research could address:

- 1) Investigation in the potential increase of classification accuracy by the optimisation of the training process. In the case of SVM classifiers this could be done by identifying specific training data used for the calculation of the optimal separating hyperplane. This could mean that this classifier could be finely tuned to the particular characteristics of a specific habitat under study and be a very cost-effective method for the relevant authority as only specific data would have to be collected regularly. DT and SVDD classifiers on the other hand depend more on a good description of the class of interest in order to perform well. Although SVDD is based upon the support vectors method,

this thesis has found that it is very much dependant on a very good description of the habitat to be classified and mapped as it does not count with any other data regarding the other classes present in the image. For DT this is also the case as the training data should reflect as much as possible the variability of the classes to be classified.

- 2) Further investigation into the use of outliers to optimise the classification using the SVDD classifier. As seen in chapter 5, the incorporation of outliers in the training set when performing the one-class classification using the SVDD did not have any impact in the final classification accuracy. However, a second application of this classifier in Foody *et al.*, (2006, in press) shows that the use of outliers can make a difference in the final classification accuracy. This second study uses very specific training datasets that have been acquired in order to intelligently train a SVM. These classes are also more homogeneous than the classes fen and saltmarsh upon which this thesis has focused. This highlights again the importance of further research into training data used for training these classifiers and the importance of the degree of spectral variability of the class under study.
- 3) More in detail investigation of other one-class classifiers. As seen in Chapter 5, there are different types of one-class classifiers (density methods, reconstruction methods and boundary methods) that have been used in pattern recognition studies but have not yet been applied within the remote sensing community. When comparing the performance of these classifiers in Chapter 5, the results showed that some of these classifiers (in particular the density methods) were comparable to those obtained by the boundary classifiers and in particular the SVDD classifier. However, these one-class classifiers are very dependant upon the density and of training data sharing the disadvantages of parametric classifiers. But, in certain cases, these classifiers could be useful and their investigation within remote sensing applications is highly recommended.

- 4) Future work could also include further research regarding ensemble techniques. Chapter 6 illustrated how the use of ensemble of classifiers could obtain higher accuracies using bagging techniques and majority voting. In this particular case, the results showed a marginal increase of accuracy that in most of the cases was not worth all the computational effort involved. However, further research could include the use of boosting methods, bigger training sizes and bigger ensemble sizes, and use of a posteriori probabilities and other combining rules in order to assess whether this technique can significantly increase classification accuracy for specific cases.
- 5) Application of all the methods for the classification of a particular habitat using hyperspectral remote sensing imagery. The present thesis has based the assessment of binary and one-class classifiers using multispectral remote sensing data acquired by Landsat sensors. Further assessment of these classifiers could include studies where detailed spatial resolution is needed and where hyperspectral data are used. One particular application within the requirements of the EU Habitats Directive could be the mapping and monitoring of specific protected species that are key for the conservation of protected habitats.
- 6) Finally, another area of future research could be based upon the classification of a particular habitat addressing the problem of mixed pixels and soft classification. This is a research area that this thesis has not explored. However, when dealing with heterogeneous habitats such as fen and saltmarsh there could be cases where certain ambiguity exists as to what class a pixel belongs. This is in turn translated into errors of omission and/or commission. In order to address the accuracy requirements to comply with legislations such as the EU Habitats Directive, it will be necessary to resolve these ambiguities. In this sense, SVMs are a useful tool that has been recently applied to unmix the class proportions in a pixel (Brown *et al.*, 2000) and that definitely require further research for the mapping of specific habitats.

This thesis has addressed the issue of classification of one particular habitat of interest using remote sensing data in order to comply with the requirements of the EU Habitats Directive. For that, advanced classification methods from pattern recognition and machine learning have been applied to remote sensing in order to answer a scientific problem within the environmental sciences community. This confirms that in order to address a pressing challenge such as biodiversity conservation (and monitoring) advances in environmental science, computation, technology and legislation have to come together appropriately.

ANNEX A

Confusion Matrices for SVM, DT and SVDD classifications

1) Error matrices for different training data sets using the one class SVM classifier. Fens (FE) as class of interest

SVM		Predicted			
Training size 30		FE	Other	Σ	Producer's accuracy
Actual	FE	123	2	125	98.40%
	Other	27	98	125	78.40%
	Σ	150	100	250	
	User's accuracy	82.00%	98.00%		88.4%

SVM		Predicted			
Training size 50		FE	Other	Σ	Producer's accuracy
Actual	FE	121	4	125	96.80%
	Other	24	101	125	80.80%
	Σ	145	105	250	
	User's accuracy	83.45%	96.19%		88.80%

SVM		Predicted			
Training size 100		FE	Other	Σ	Producer's accuracy
Actual	FE	123	2	125	98.40%
	Other	19	106	125	84.80%
	Σ	142	108	250	
	User's accuracy	86.62%	98.15%		91.6%

SVM		Predicted			
Training size 150		FE	Other	Σ	Producer's accuracy
Actual	FE	123	2	125	98.40%
	Other	19	106	125	84.80%
	Σ	142	108	250	
	User's accuracy	86.62%	98.15%		91.6%

SVM		Predicted			
Training size 200		FE	Other	Σ	Producer's accuracy
Actual	FE	123	2	125	98.40%
	Other	19	106	125	84.80%
	Σ	142	108	250	
	User's accuracy	86.62%	98.15%		91.6%

SVM		Predicted			
Training size 250		FE	Other	Σ	Producer's accuracy
Actual	FE	121	4	125	96.80%
	Other	16	109	125	87.20%
	Σ	137	113	250	
	User's accuracy	88.32%	96.46%		91.6%

SVM		Predicted			
Training size 300		FE	Other	Σ	Producer's accuracy
Actual	FE	121	4	125	96.80%
	Other	16	109	125	87.20%
	Σ	137	113	250	
	User's accuracy	88.32%	96.46%		92.00%

2) Error matrices for different training data sets using the one class SVM classifier. Saltmarshes (SA) as class of interest

SVM		Predicted			
Training size 30		SA	Other	Σ	Producer's accuracy
Actual	SA	124	1	125	99.20%
	Other	23	102	125	81.60%
	Σ	147	103	250	
	User's accuracy	84.35%	99.03%		90.40%

SVM		Predicted			
Training size 50		SA	Other	Σ	Producer's accuracy
Actual	SA	124	1	125	99.20%
	Other	21	104	125	83.20%
	Σ	145	105	250	
	User's accuracy	85.52%	99.05%		92.00%

SVM		Predicted			
Training size 100		SA	Other	Σ	Producer's accuracy
Actual	SA	122	3	125	97.60%
	Other	17	108	125	86.40%
	Σ	139	111	250	
	User's accuracy	87.77%	97.30%		92.00%

SVM		Predicted			
Training size 150		SA	Other	Σ	Producer's accuracy
Actual	SA	122	3	125	97.60%
	Other	17	108	125	86.40%
	Σ	139	111	250	
	User's accuracy	87.77%	97.30%		92.00%

SVM		Predicted			
Training size 200		SA	Other	Σ	Producer's accuracy
Actual	SA	122	3	125	97.60%
	Other	17	108	125	86.40%
	Σ	139	111	250	
	User's accuracy	87.77%	97.30%		92.00%

SVM		Predicted			
Training size 250		SA	Other	Σ	Producer's accuracy
Actual	SA	122	3	125	97.60%
	Other	17	108	125	86.40%
	Σ	139	111	250	
	User's accuracy	87.77%	97.30%		92.00%

SVM		Predicted			
Training size 300		SA	Other	Σ	Producer's accuracy
Actual	SA	122	3	125	97.60%
	Other	17	108	125	86.40%
	Σ	139	111	250	
	User's accuracy	87.77%	97.30%		92.00%

3) Error matrices for different training data sets using the one class DT classifier.
Fens (FE) as class of interest

DT		Predicted			
Training size 30		FE	Other	Σ	Producer's accuracy
Actual	FE	119	6	125	95.20%
	Other	33	92	125	73.60%
	Σ	152	98	250	
	User's accuracy	78.29%	93.88%		84.4

DT		Predicted			
Training size 50		FE	Other	Σ	Producer's accuracy
Actual	FE	123	2	125	98.40%
	Other	27	98	125	78.40%
	Σ	150	100	250	
	User's accuracy	82.00%	98.00%		88.4

DT		Predicted			
Training size 100		FE	Other	Σ	Producer's accuracy
Actual	FE	123	2	125	98.40%
	Other	19	106	125	84.80%
	Σ	142	108	250	
	User's accuracy	86.62%	98.15%		91.6

DT		Predicted			
Training size 150		FE	Other	Σ	Producer's accuracy
Actual	FE	123	2	125	98.40%
	Other	19	106	125	84.80%
	Σ	142	108	250	
	User's accuracy	86.62%	98.15%		91.6

DT		Predicted			
Training size 200		FE	Other	Σ	Producer's accuracy
Actual	FE	123	2	125	98.40%
	Other	19	106	125	84.80%
	Σ	142	108	250	
	User's accuracy	86.62%	98.15%		91.6

DT		Predicted			
Training size 250		FE	Other	Σ	Producer's accuracy
Actual	FE	121	4	125	96.80%
	Other	16	109	125	87.20%
	Σ	137	113	250	
	User's accuracy	88.32%	96.46%		92

DT		Predicted			
Training size 300		FE	Other	Σ	Producer's accuracy
Actual	FE	121	4	125	96.80%
	Other	16	109	125	87.20%
	Σ	137	113	250	
	User's accuracy	88.32%	96.46%		92

4) Error matrices for different training data sets using the one class DT classifier.
Saltmarshes (SA) as class of interest

DT		Predicted			
Training size		SA	Other	Σ	Producer's accuracy
30					
Actual	SA	107	18	125	85.60%
	Other	24	101	125	80.80%
	Σ	131	119	250	
	User's accuracy	81.68%	84.87%		83.2

DT		Predicted			
Training size		SA	Other	Σ	Producer's accuracy
50					
Actual	SA	124	1	125	99.20%
	Other	21	104	125	83.20%
	Σ	145	105	250	
	User's accuracy	85.52%	99.05%		91.2

DT		Predicted			
Training size 100		SA	Other	Σ	Producer's accuracy
Actual	SA	118	7	125	94.40%
	Other	5	120	125	96.00%
	Σ	123	127	250	
	User's accuracy	95.93%	94.49%		95.2

DT		Predicted			
Training size 150		SA	Other	Σ	Producer's accuracy
Actual	SA	119	6	125	95.20%
	Other	3	122	125	97.60%
	Σ	122	128	250	
	User's accuracy	97.54%	95.31%		96.4

DT		Predicted			
Training size 200		SA	Other	Σ	Producer's accuracy
Actual	SA	119	6	125	95.20%
	Other	3	122	125	97.60%
	Σ	122	128	250	
	User's accuracy	97.54%	95.31%		96.4

DT		Predicted			
Training size 250		SA	Other	Σ	Producer's accuracy
Actual	SA	119	6	125	95.20%
	Other	3	122	125	97.60%
	Σ	122	128	250	
	User's accuracy	97.54%	95.31%		96.4

DT		Predicted			
Training size 300		SA	Other	Σ	Producer's accuracy
Actual	SA	123	2	125	98.40%
	Other	4	121	125	96.80%
	Σ	127	123	250	
	User's accuracy	96.85%	98.37%		97.6

5) Error matrices for different training data sets using the one class SVDD classifier. Fens (FE) as class of interest

SVDD		Predicted			
Training size 5		FE	Other	Σ	Producer's accuracy
Actual	FE	92	33	125	73.60%
	Other	2	123	125	98.40%
	Σ	94	156	250	
	User's accuracy	97.87 %	78.85 %		

SVDD		Predicted			
Training size 10		FE	Other	Σ	Producer's accuracy
Actual	FE	104	21	125	83.20%
	Other	2	123	125	98.40%
	Σ	106	144	250	
	User's accuracy	98.11%	85.42%		

SVDD		Predicted			
Training size 15		FE	Other	Σ	Producer's accuracy
Actual	FE	104	21	125	83.20%
	Other	2	123	125	98.40%
	Σ	106	144	250	
	User's accuracy	98.11%	85.42%		

SVDD		Predicted			
Training size 20		FE	Other	Σ	Producer's accuracy
Actual	FE	109	16	125	87.20%
	Other	2	123	125	98.40%
	Σ	111	139	250	
	User's accuracy	98.20%	88.49%		

SVDD		Predicted			
Training size 25		FE	Other	Σ	Producer's accuracy
Actual	FE	115	10	125	92.00%
	Other	3	122	125	97.60%
	Σ	118	132	250	
	User's accuracy	97.46%	92.42%		Overall 94.80%

SVDD		Predicted			
Training size 50		FE	Other	Σ	Producer's accuracy
Actual	FE	115	10	125	92.00%
	Other	3	122	125	97.60%
	Σ	118	132	250	
	User's accuracy	97.46%	92.42%		Overall 94.80%

SVDD		Predicted			
Training size 75		FE	Other	Σ	Producer's accuracy
Actual	FE	121	4	125	96.8%
	Other	13	112	125	89.6%
	Σ	134	116	250	
	User's accuracy	90.30%	96.55%		Overall 93.20%

SVDD		Predicted			
Training size 100		FE	Other	Σ	Producer's accuracy
Actual	FE	121	4	125	96.8%
	Other	13	112	125	89.6%
	Σ	134	116	250	
	User's accuracy	90.30%	96.55%		Overall 93.20%

SVDD		Predicted			
Training size 125		FE	Other	Σ	Producer's accuracy
Actual	FE	121	4	125	96.8%
	Other	13	112	125	89.6%
	Σ	134	116	250	
	User's accuracy	90.30%	96.55%		Overall 93.20%

SVDD		Predicted			
Training size 150		FE	Other	Σ	Producer's accuracy
Actual	FE	117	8	125	93.6%
	Other	3	122	125	97.6%
	Σ	120	130	250	
	User's accuracy	97.50%	93.85%		Overall 95.60%

SVDD		Predicted			
Training size 200		FE	Other	Σ	Producer's accuracy
Actual	FE	117	8	125	93.6%
	Other	3	122	125	97.6%
	Σ	120	130	250	
	User's accuracy	97.50%	93.85%		Overall 95.60%

6) Error matrices for different training data sets using the one class SVDD classifier. Saltmarshes (SA) as class of interest

SVDD		Predicted			
Training size 5		SA	Other	Σ	Producer's accuracy
Actual	SA	71	54	125	57%
	Others	4	121	125	96.80%
	Σ	75	175	250	
	User's accuracy	99.01%	83.22%		Overall 76.80%

SVDD		Predicted			
Training size 10		SA	Other	Σ	Producer's accuracy
Actual	SA	102	23	125	82%
	Others	5	120	125	96.00%
	Σ	107	143	250	
	User's accuracy	99.01%	83.22%		Overall 88.80%

SVDD		Predicted			
Training size 15		SA	Other	Σ	Producer's accuracy
Actual	SA	106	19	125	85%
	Others	14	111	125	88.80%
	Σ	120	130	250	
	User's accuracy	99.01%	83.22%		Overall 86.80%

SVDD		Predicted			
Training size 20		SA	Other	Σ	Producer's accuracy
Actual	SA	114	11	125	91%
	Others	16	109	125	87.20%
	Σ	130	120	250	
	User's accuracy	99.01%	83.22%		Overall 89.20%

SVDD		Predicted			
Training size 25		SA	Other	Σ	Producer's accuracy
Actual	SA	100	25	125	80%
	Others	1	124	125	99.2%
	Σ	101	149	250	
	User's accuracy	99.01%	83.22%		Overall 89.60%

SVDD		Predicted			
Training size 50		SA	Other	Σ	Producer's accuracy
Actual	SA	102	23	125	81.6%
	Others	1	124	125	99.2%
	Σ	103	147	250	
	User's accuracy	99.03%	84.35%		Overall 90.40%

SVDD		Predicted			
Training size 75		SA	Other	Σ	Producer's accuracy
Actual	SA	104	21	125	83.2%
	Others	1	124	125	99.2%
	Σ	105	145	250	
	User's accuracy	99.05%	85.52%		Overall 91.20%

SVDD		Predicted			
Training size 100		SA	Other	Σ	Producer's accuracy
Actual	SA	116	10	125	92.8%
	Others	2	123	125	98.4%
	Σ	118	133	250	
	User's accuracy	98.31%	92.48%		Overall 95.60%

SVDD		Predicted			
Training size 125		SA	Other	Σ	Producer's accuracy
Actual	SA	116	10	125	92.8%
	Others	2	123	125	98.4%
	Σ	118	133	250	
	User's accuracy	98.31%	92.48%		Overall 95.60%

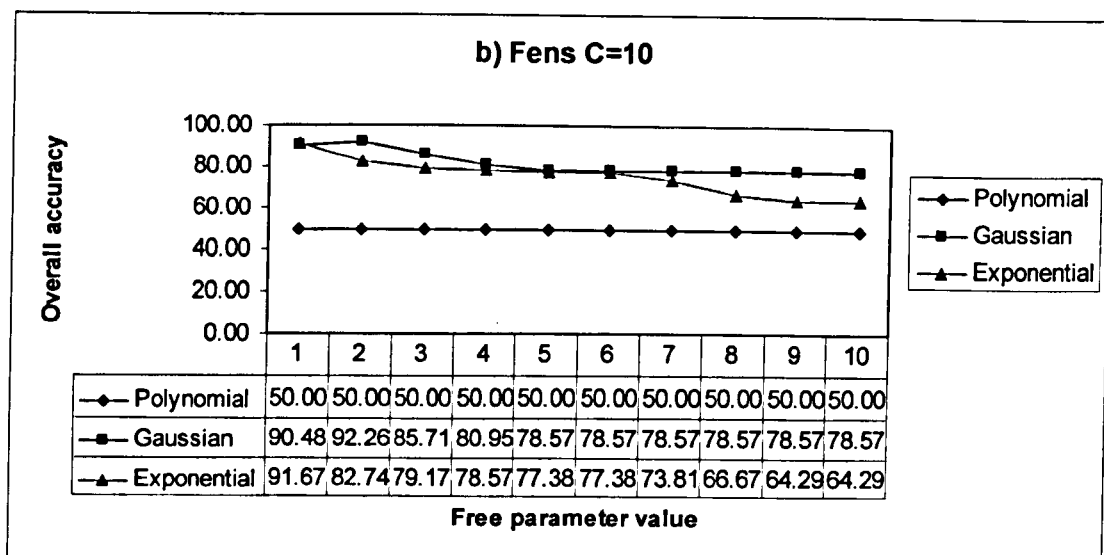
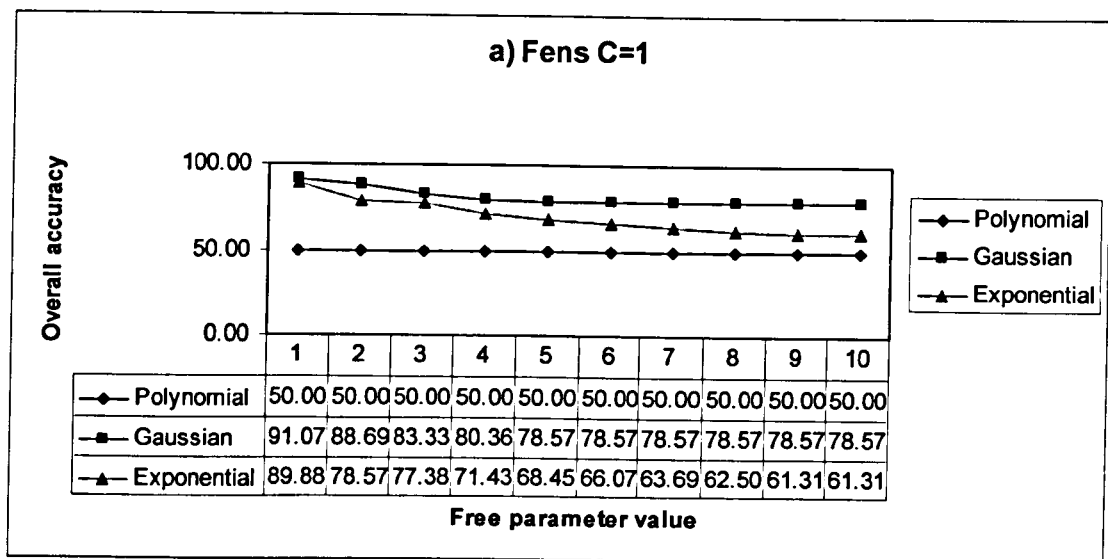
SVDD		Predicted			
Training size 150		SA	Other	Σ	Producer's accuracy
Actual	SA	115	10	125	92%
	Others	2	123	125	98.4%
	Σ	117	133	250	
	User's accuracy	98.29%	92.48%		Overall 95.20%

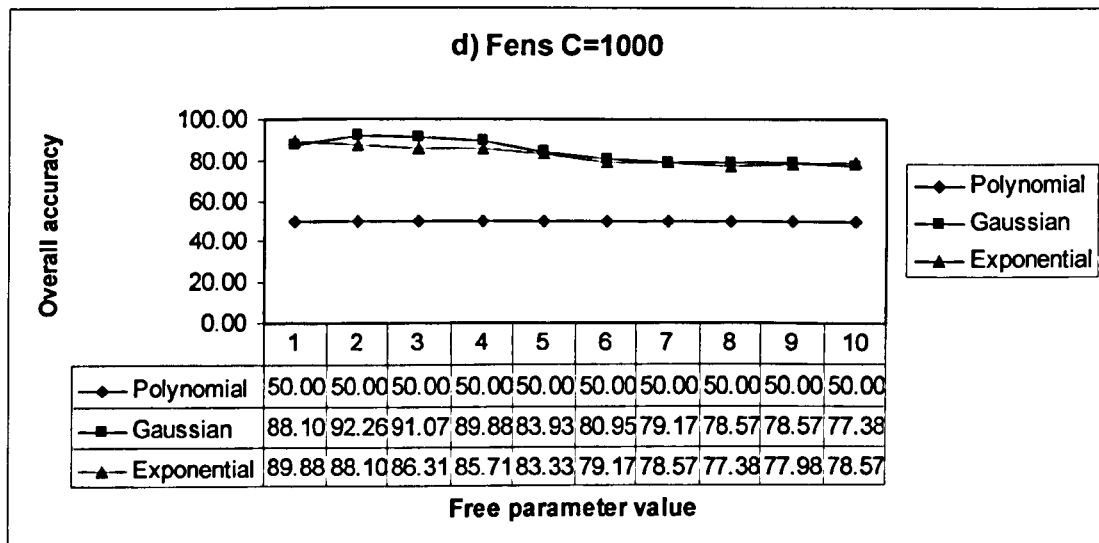
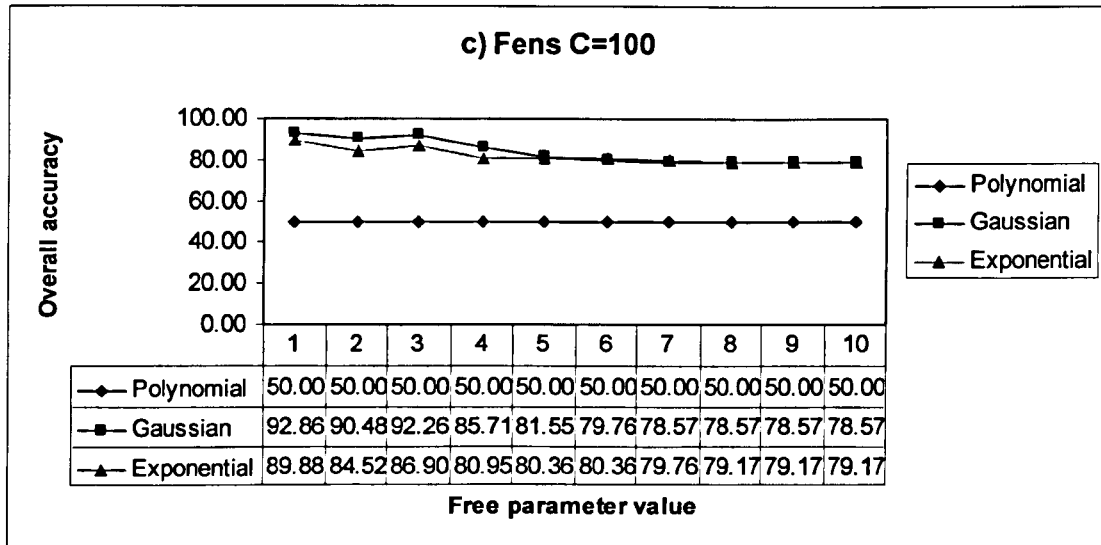
SVDD		Predicted			
Training size 200		SA	Other	Σ	Producer's accuracy
Actual	SA	115	10	125	92%
	Others	2	123	125	98.4%
	Σ	117	133	250	
	User's accuracy	98.29%	92.48%		Overall 95.20%

ANNEX B

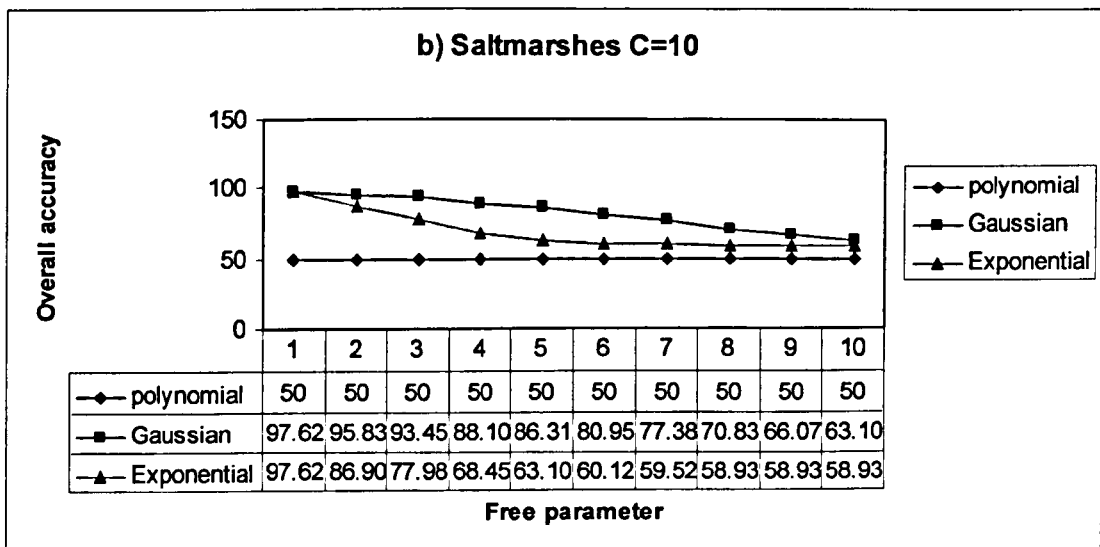
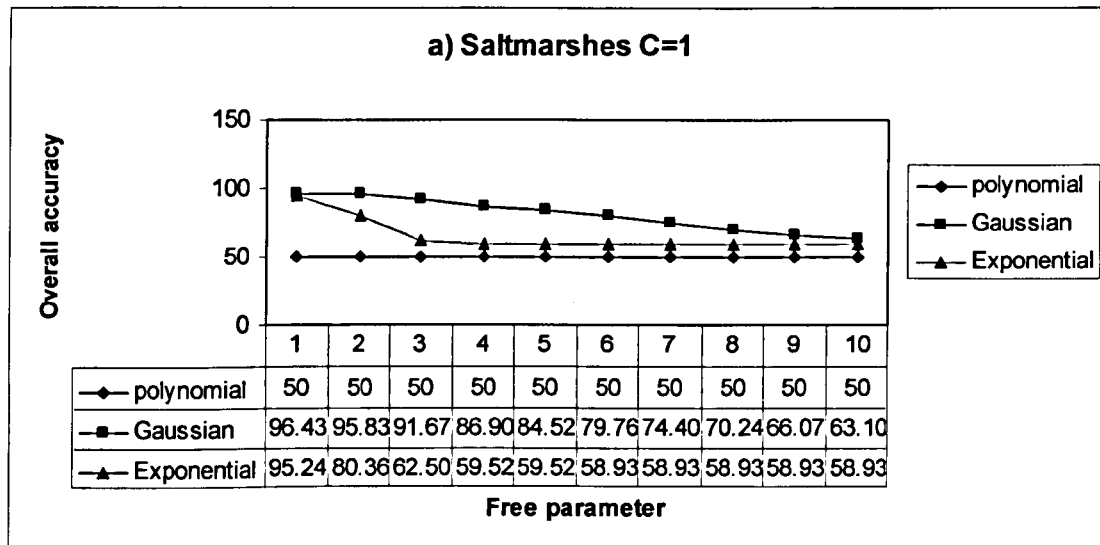
Parameter choice for SVM and SVDD by cross-validation

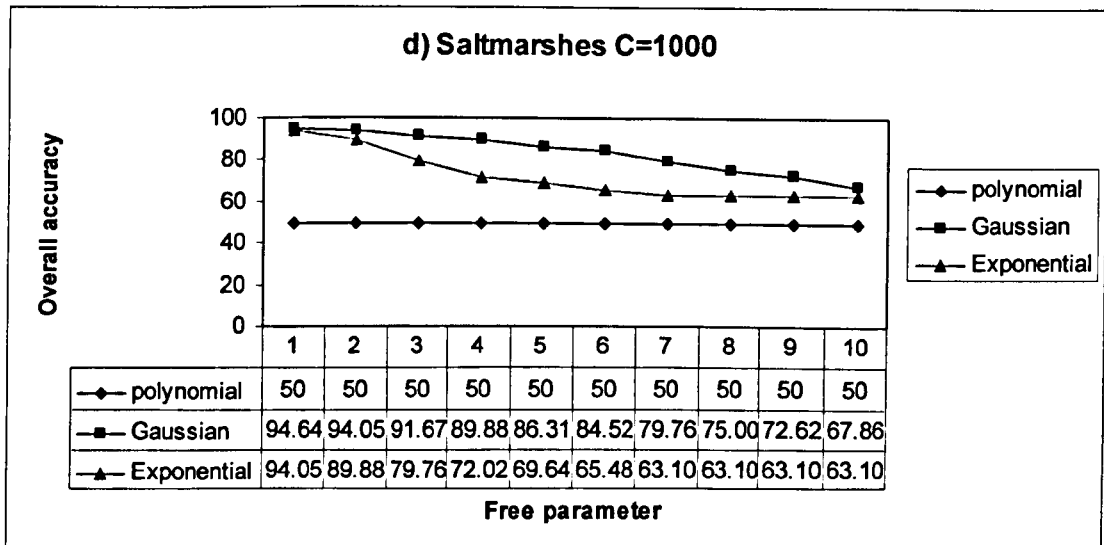
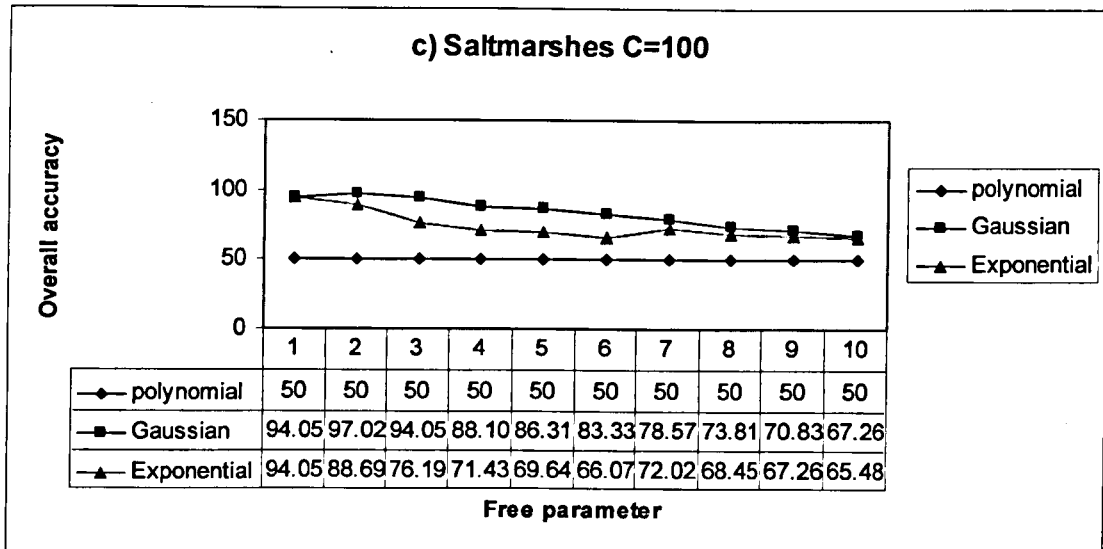
1) SVM parameters. Cross-validation. Fens as class of interest



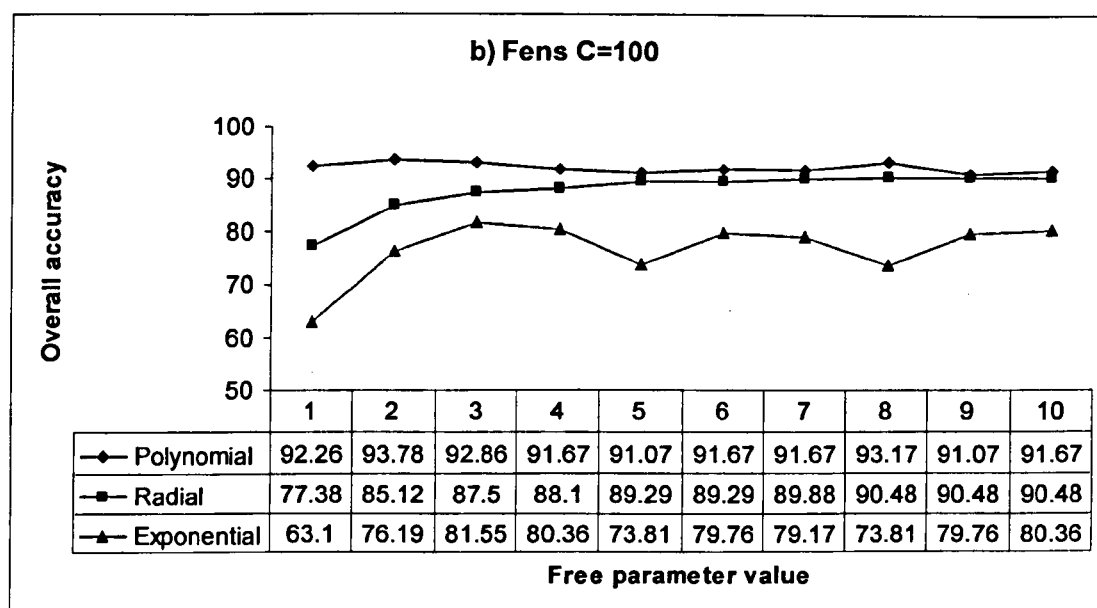
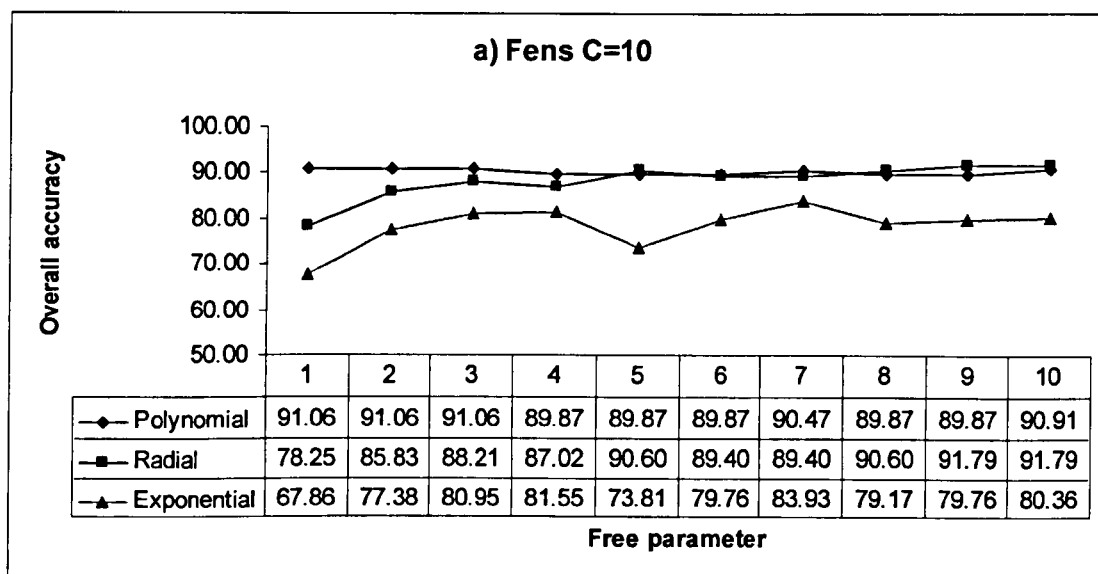


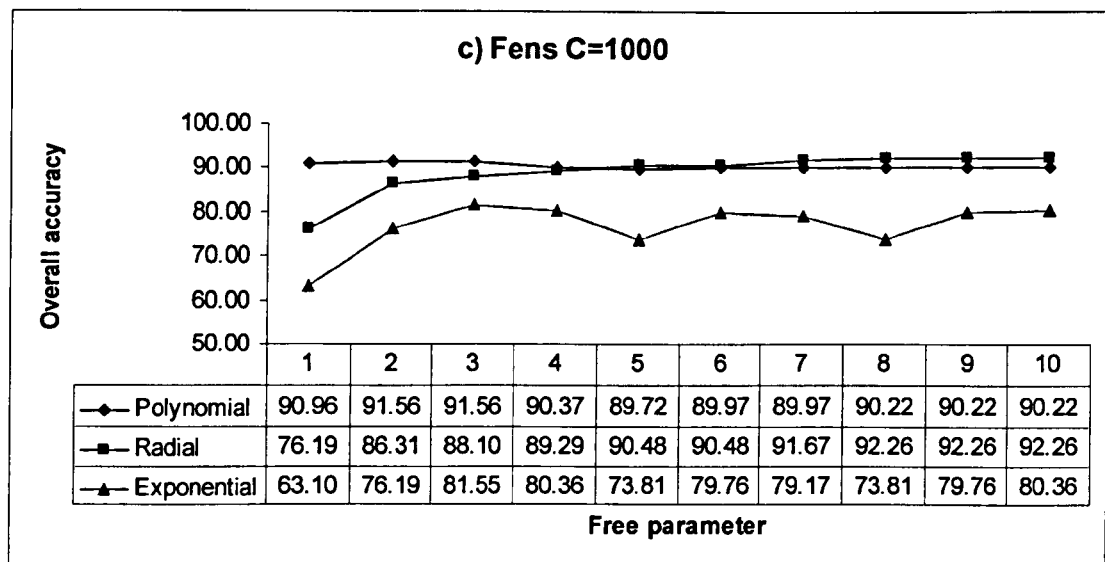
2) SVM parameters. Cross-validation. Saltmarshes as class of interest



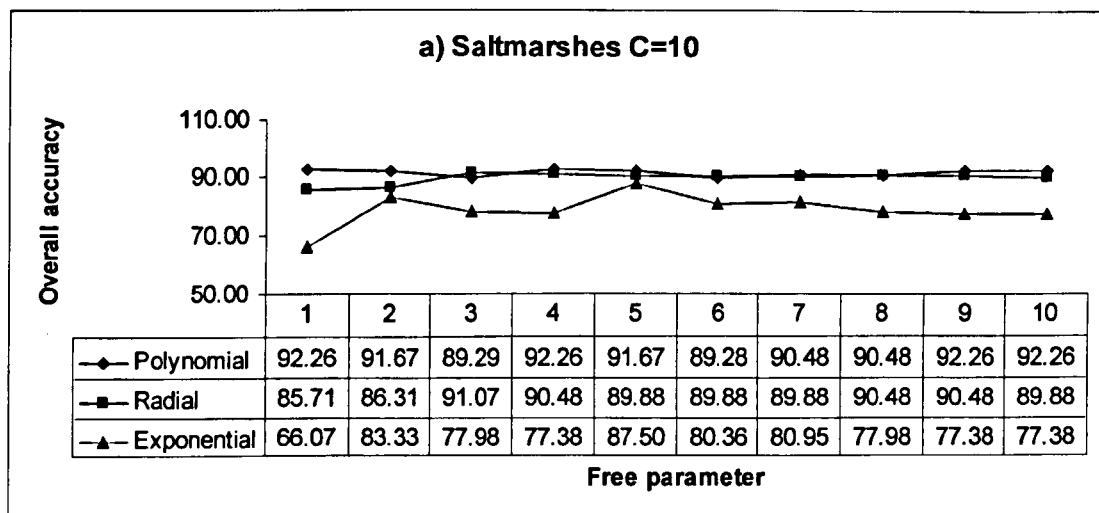


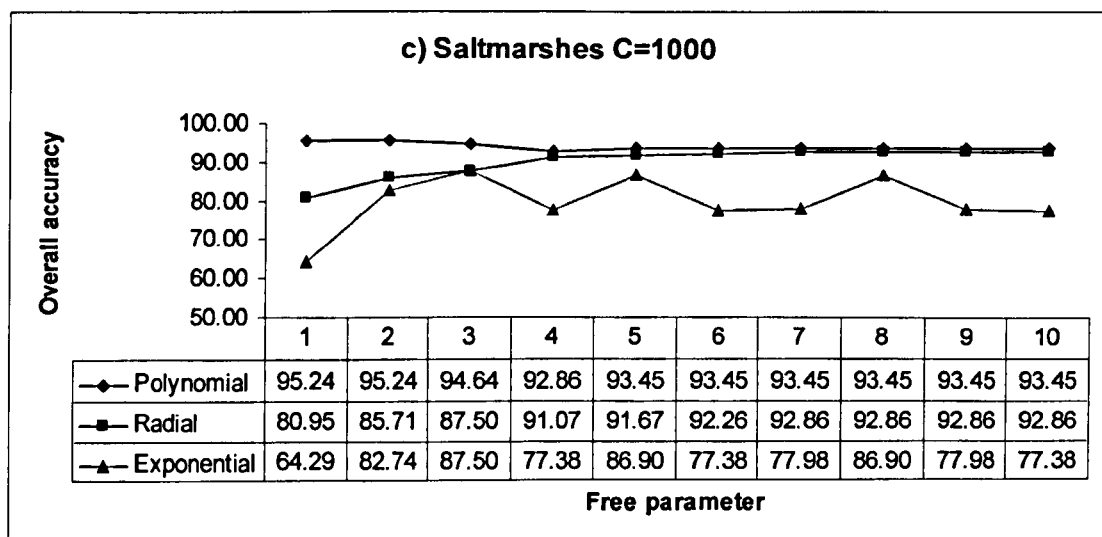
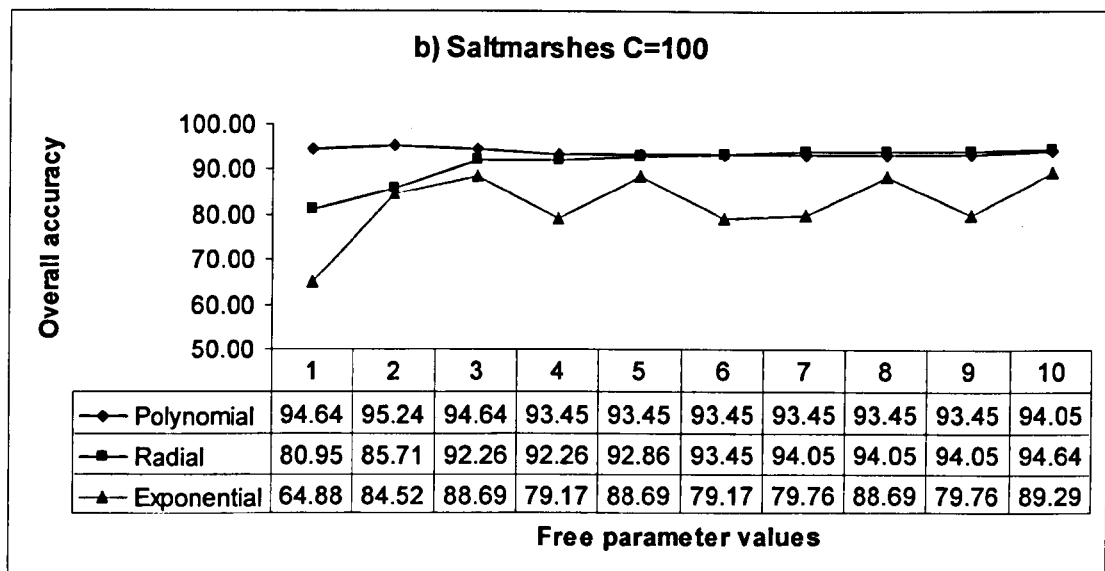
3) SVDD parameters. Cross-validation. Fens as class of interest





4) SVDD Cross-validation. Saltmarshes as class of interest

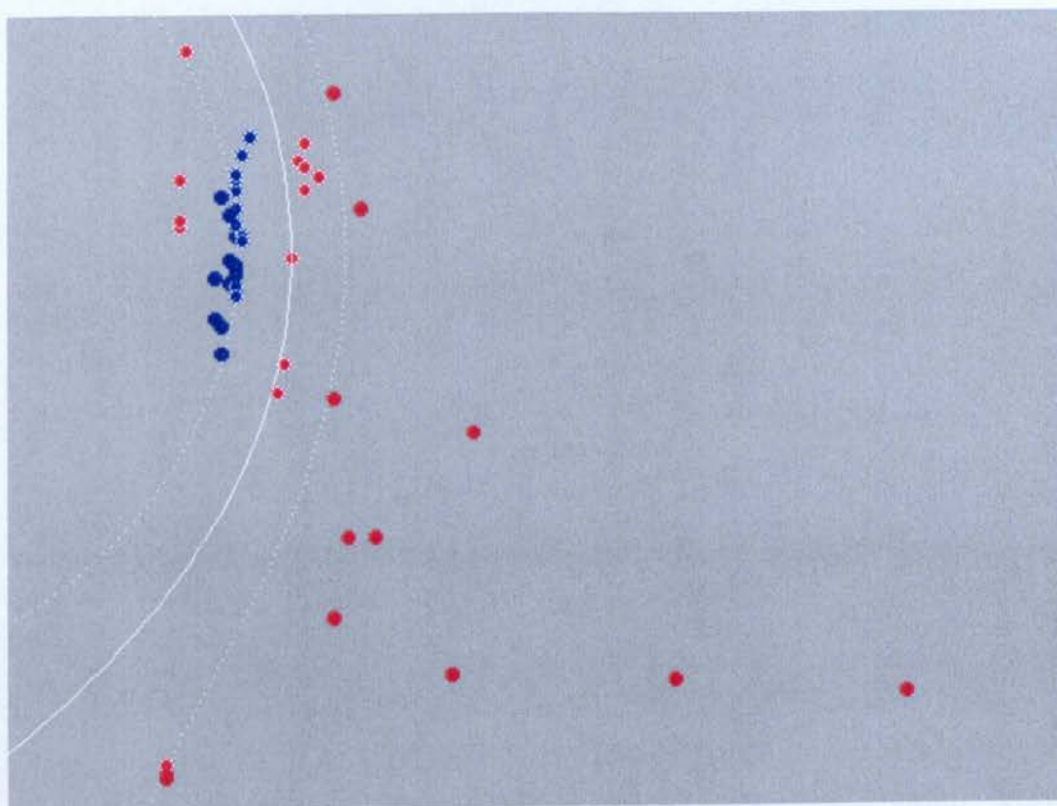




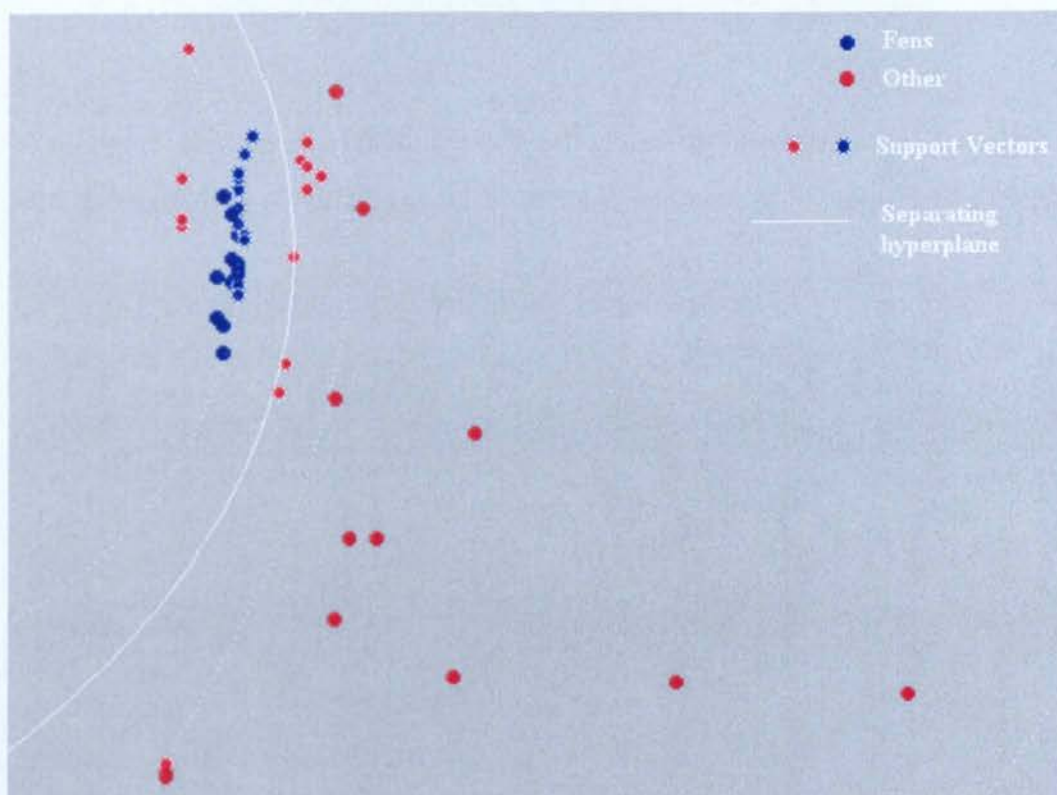
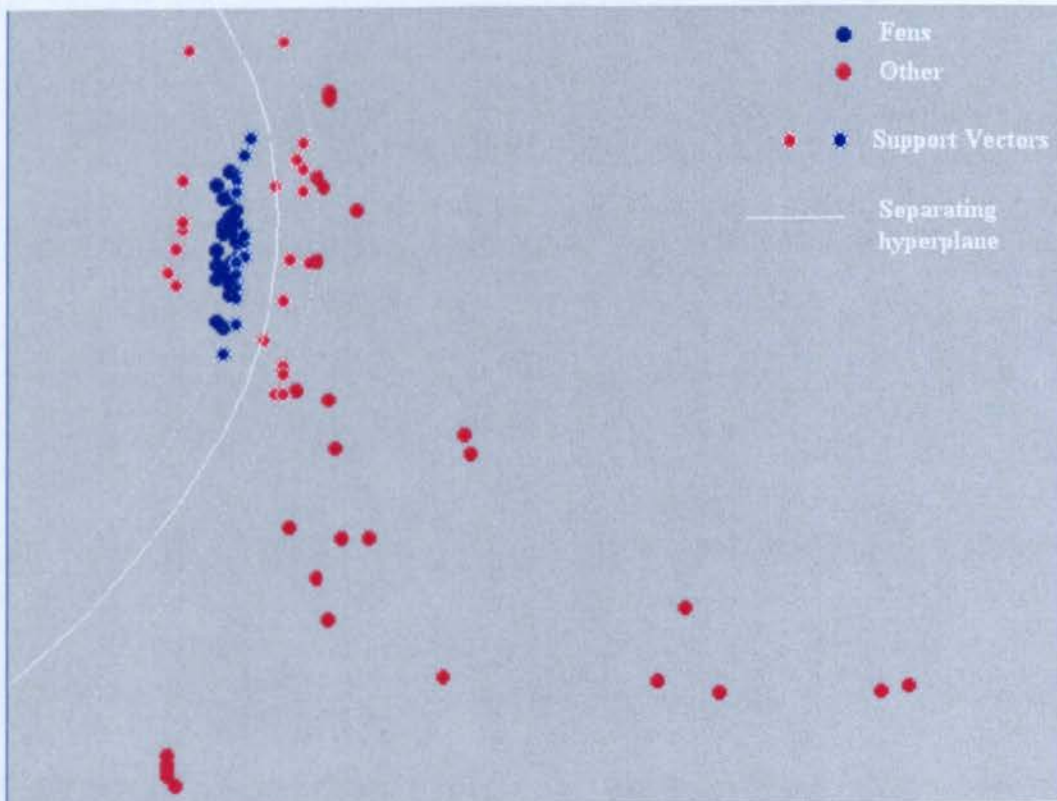
ANNEX C

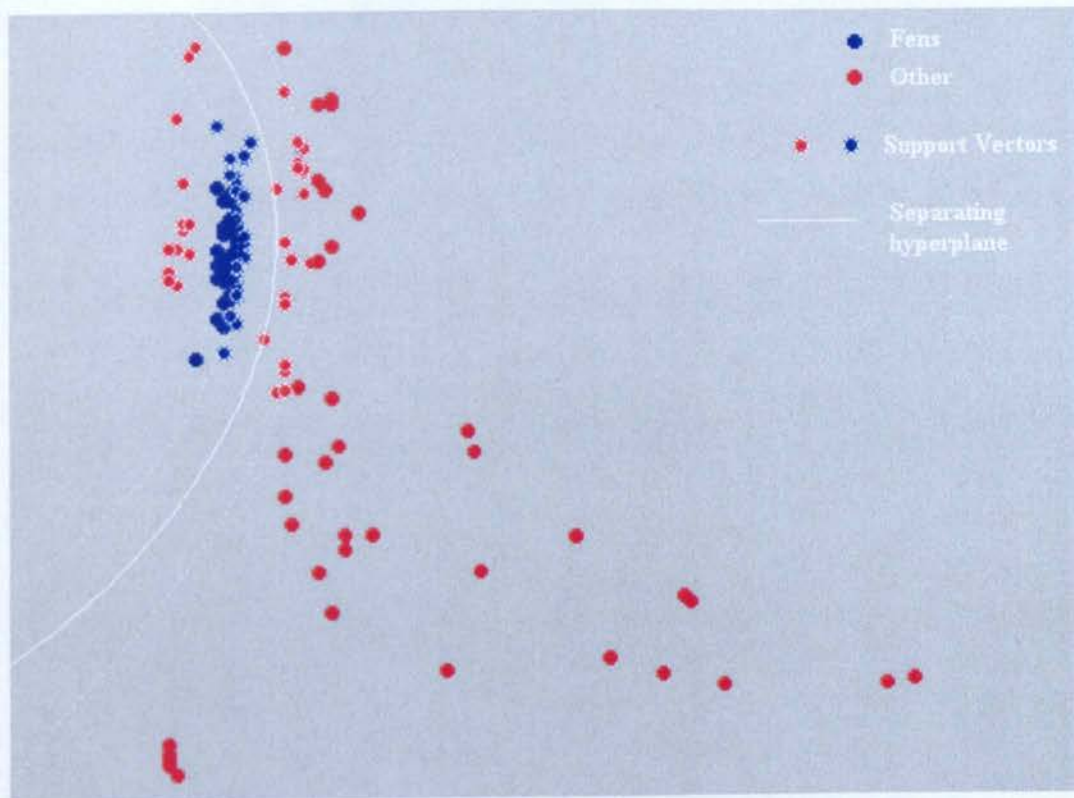
SVM Support Vectors and Separating Hyperplanes

1) Support Vectors and separating hyperplanes for the SVM per training size. Fens as class of interest

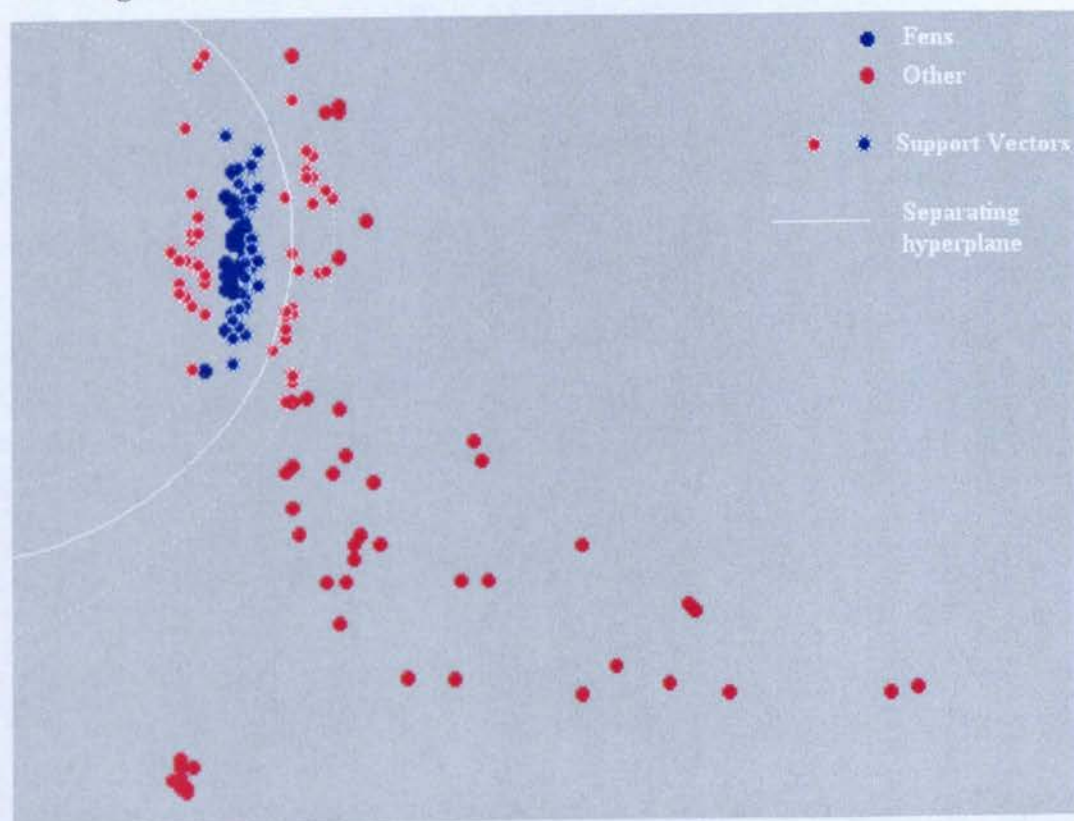


Training size 30

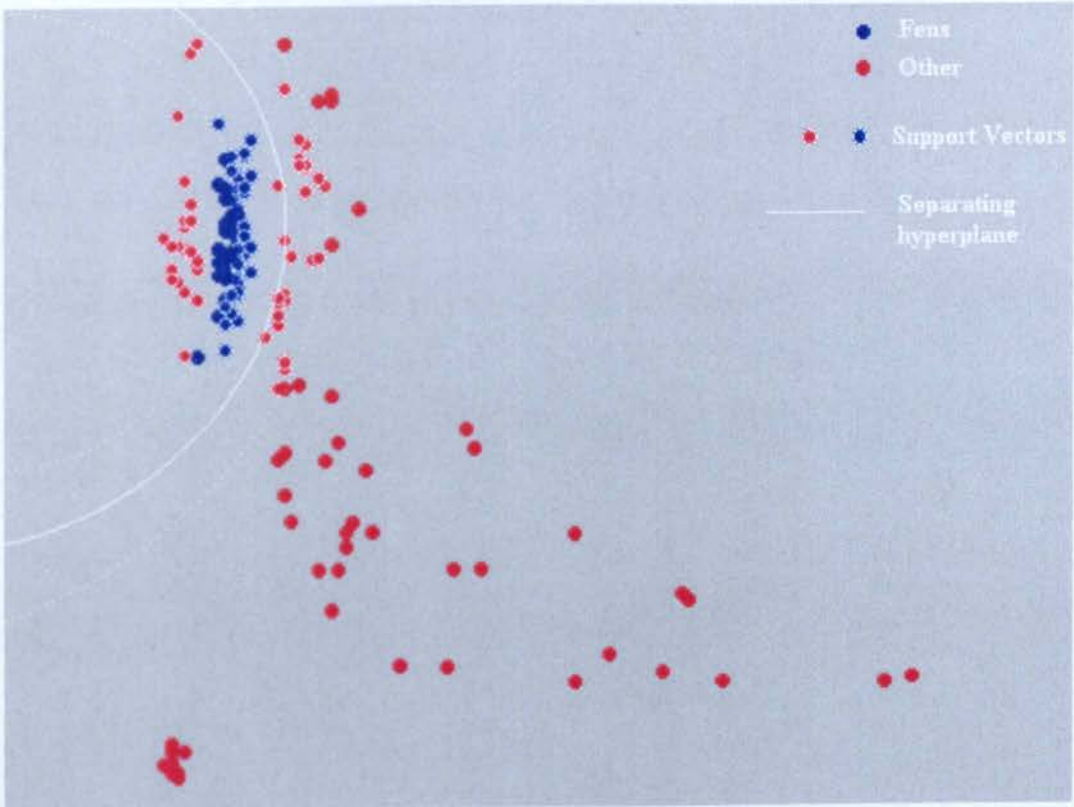
**Training size 50****Training size 100**



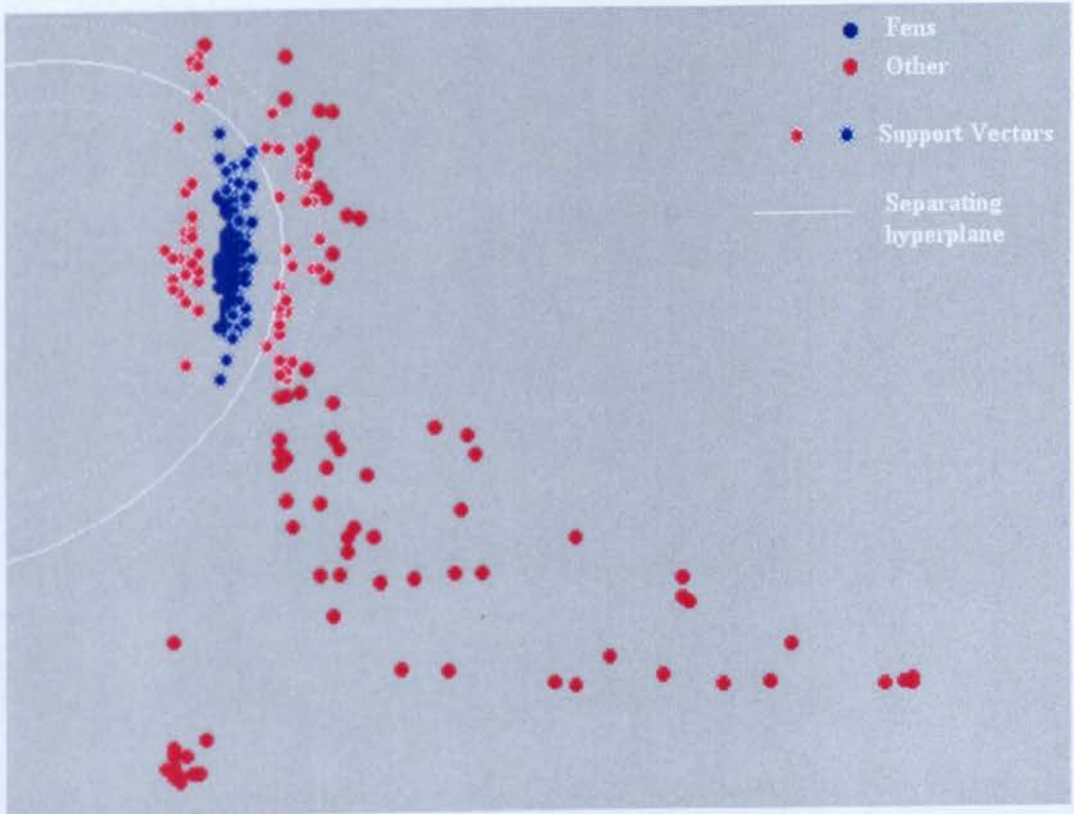
Training size 150



Training size 200

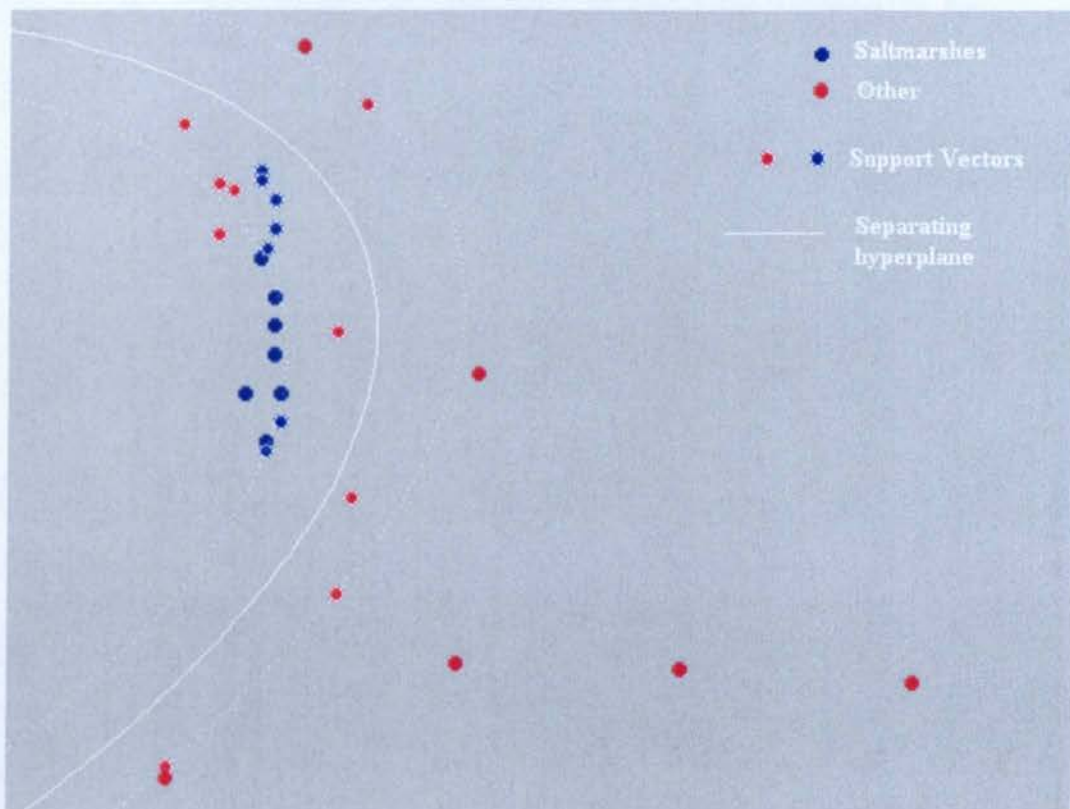


Training size 250

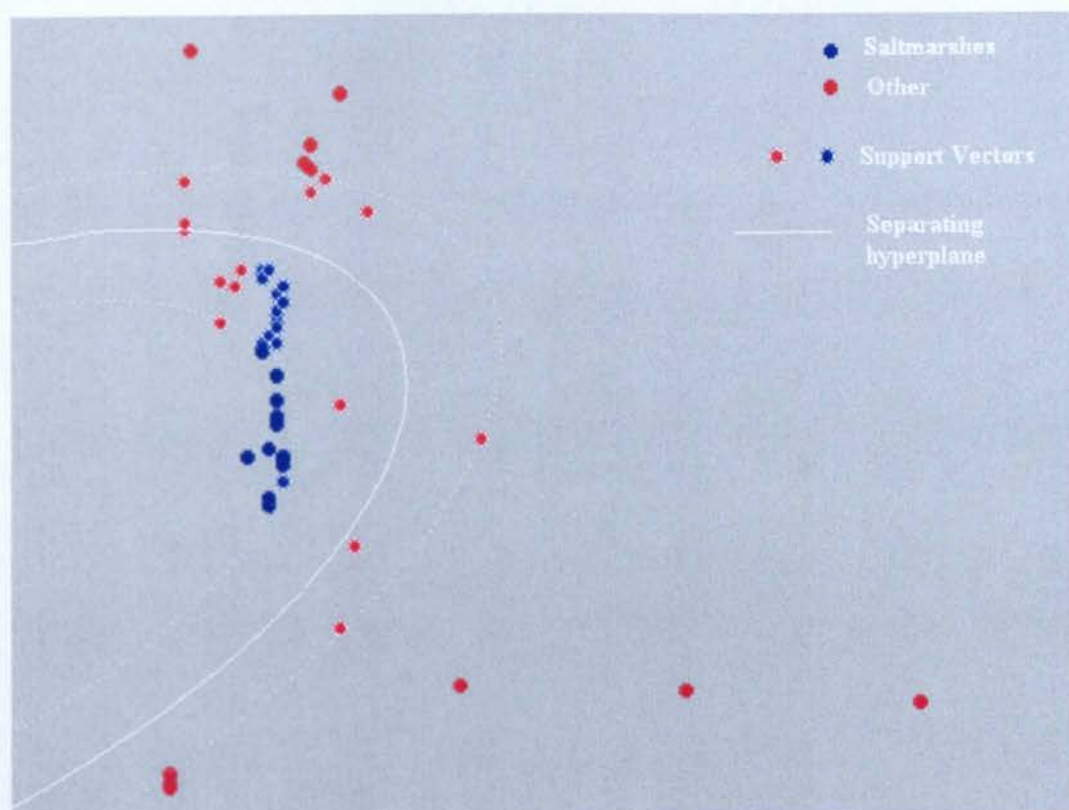


Training size 300

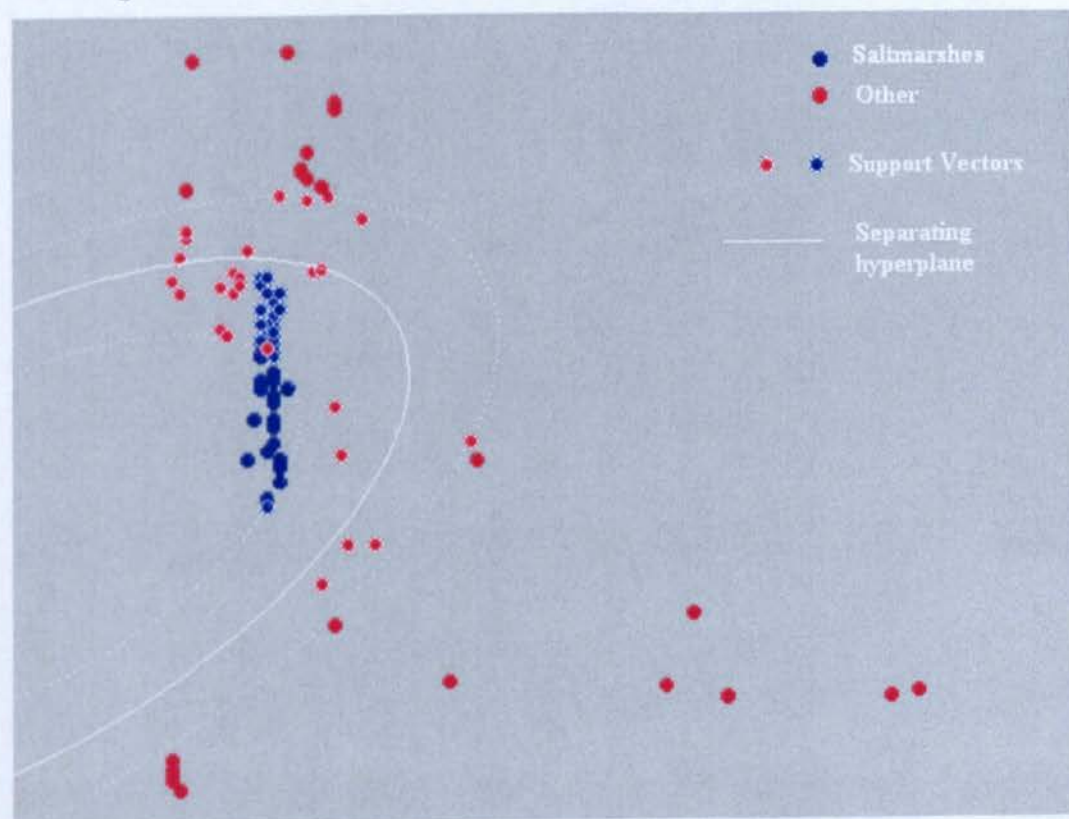
2) Support Vectors and separating hyperplanes for the SVM per training size. Saltmarshes as class of interest.



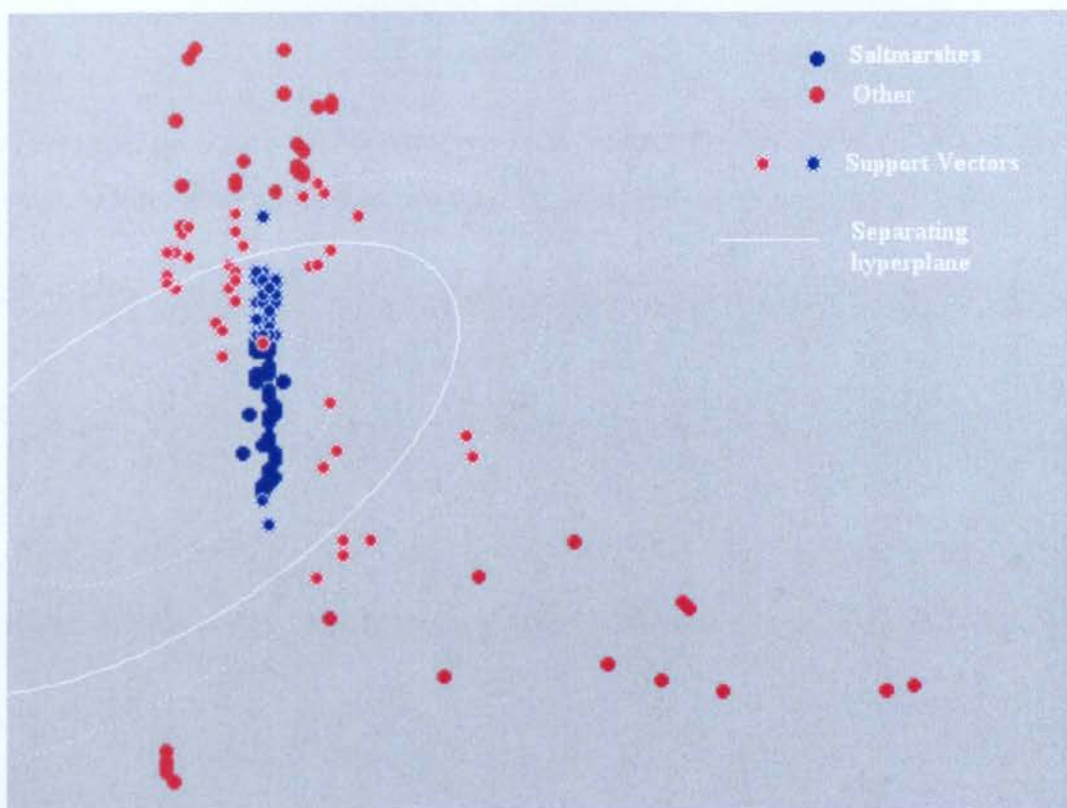
Training size 30



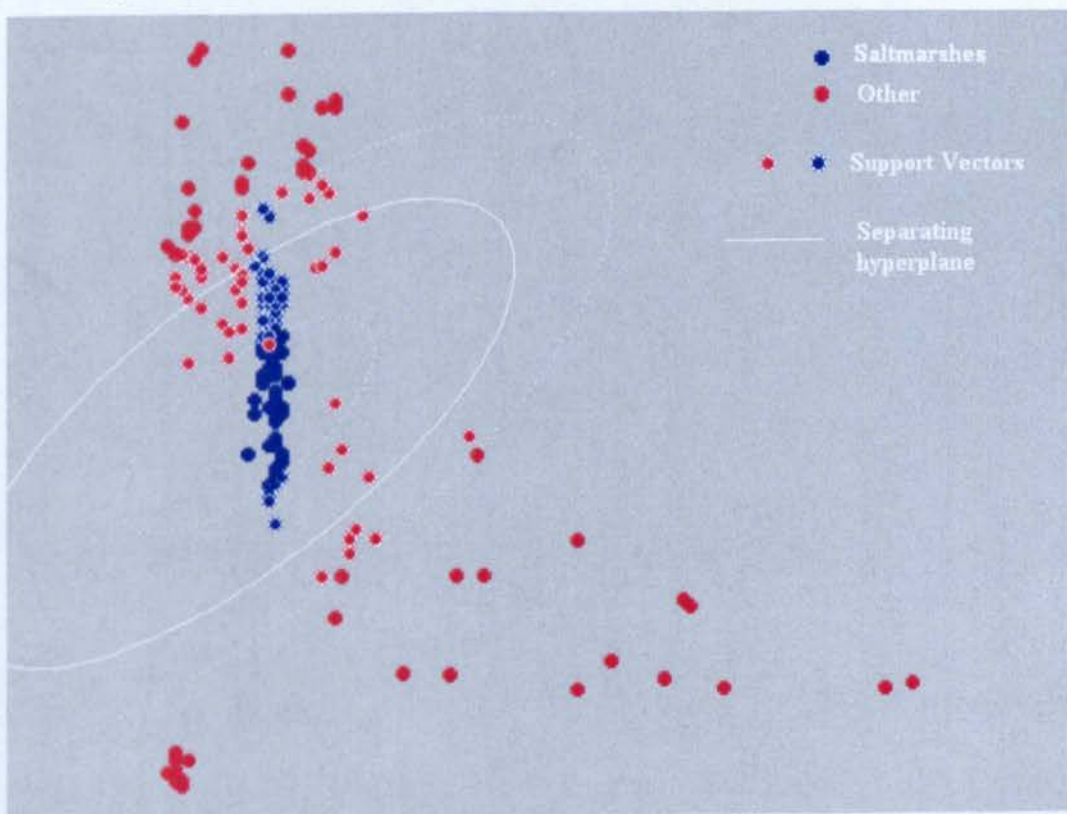
Training size 50



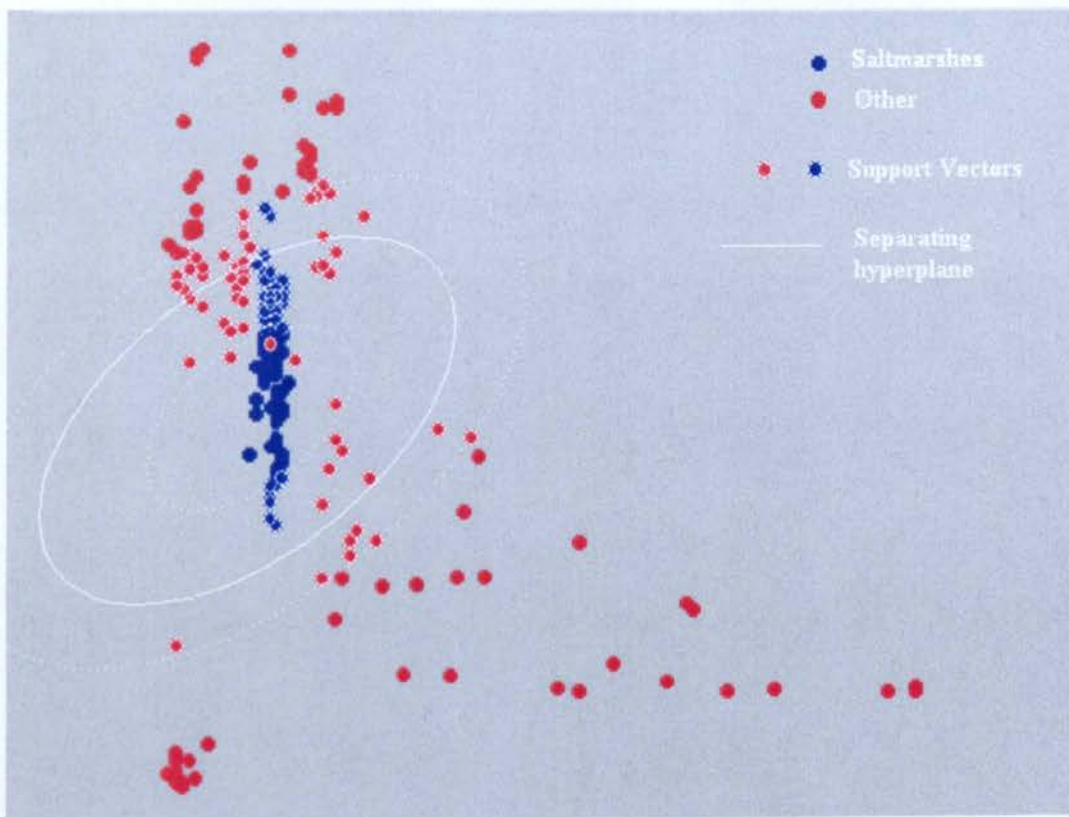
Training size 100



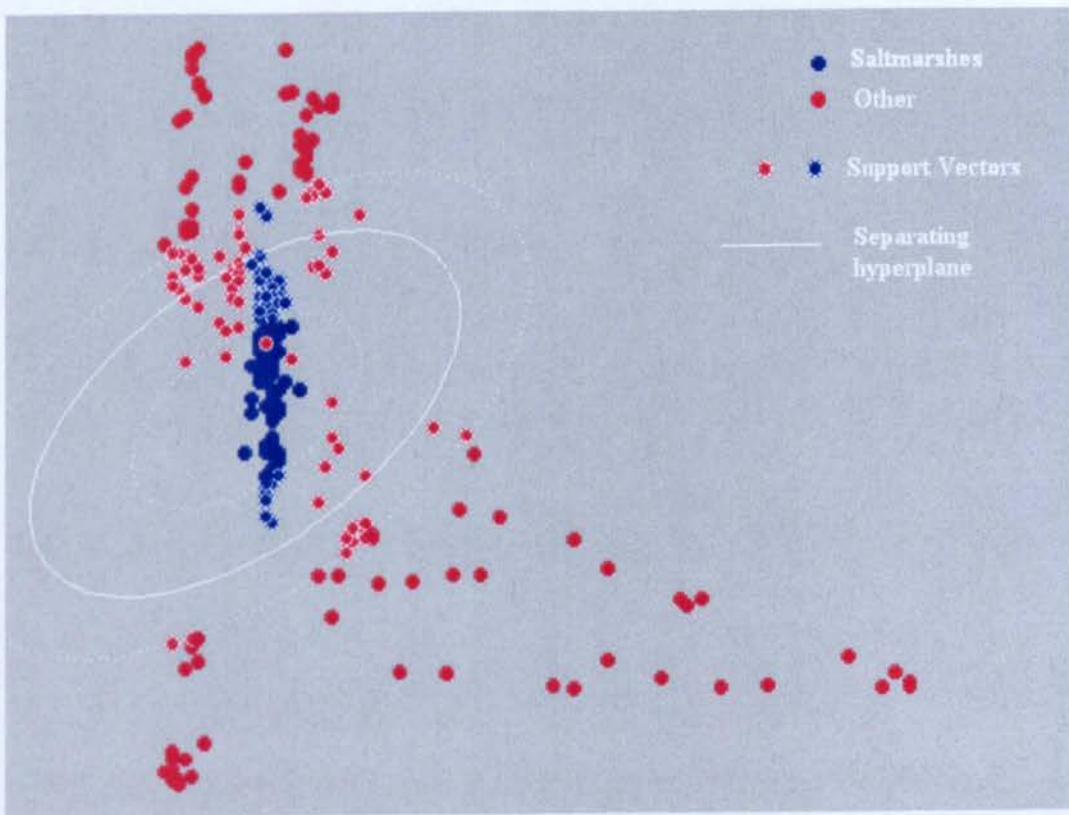
Training size 150



Training size 200



Training size 250



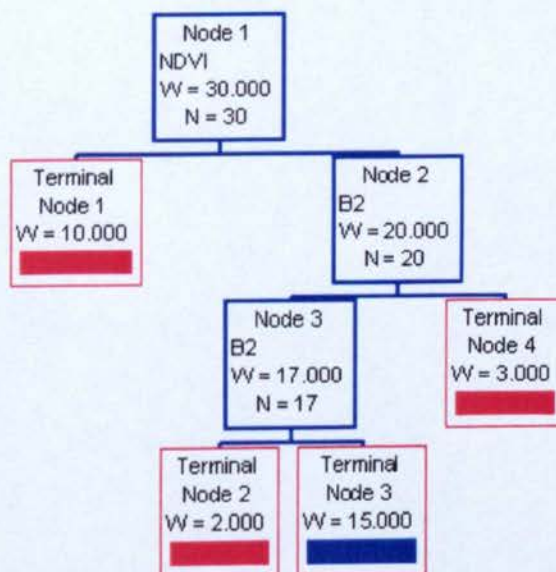
Training size 300

ANNEX D
Decision Trees structures

1) Decision Trees. Optimal tree structure by training set size

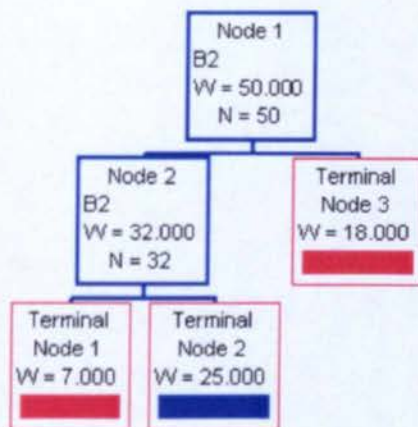
Fens as class of interest


Training size 30 pixels




■ Fens
■ Other class

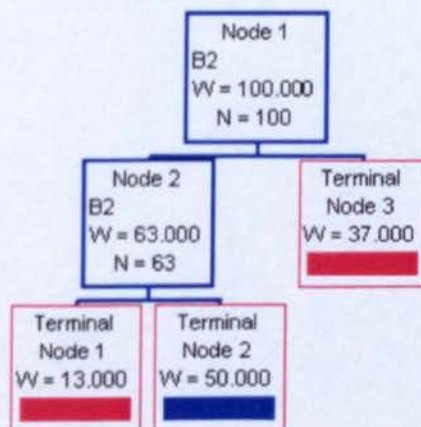
Training size 50 pixels





 Fens

 Other class

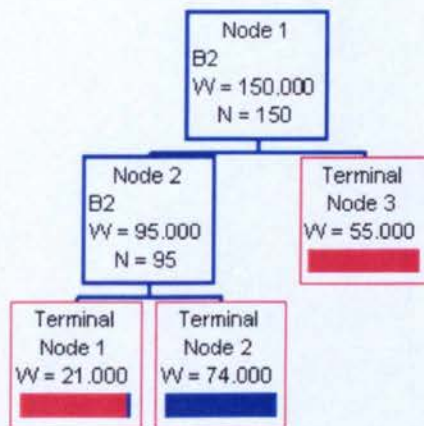
Training size 100 pixels





 Fens

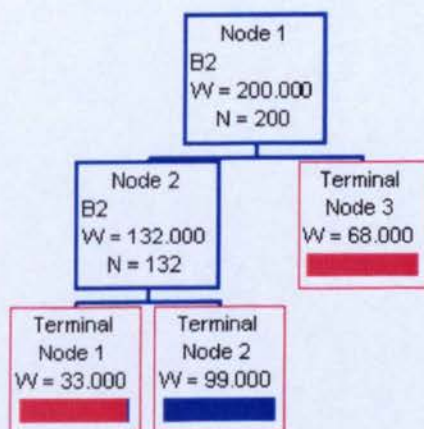
 Other class



Training size 150 pixels



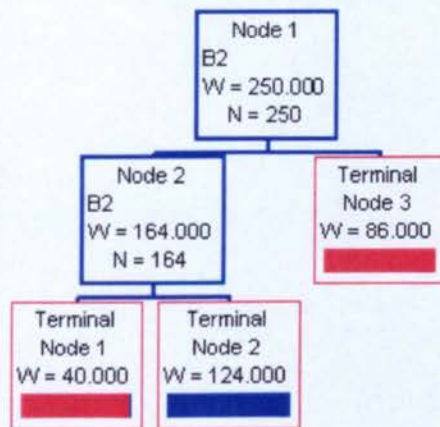
 Fens
 Other class

Training size 200 pixels



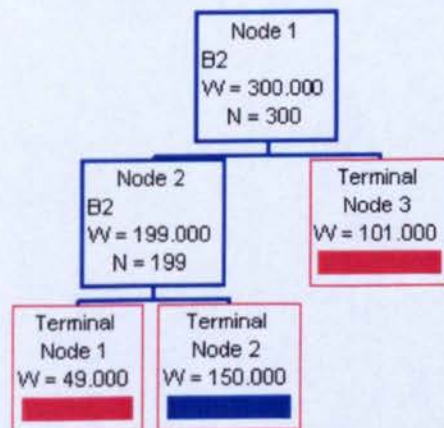
 Fens
 Other class

Training size 250 pixels



■ Fens
■ Other class

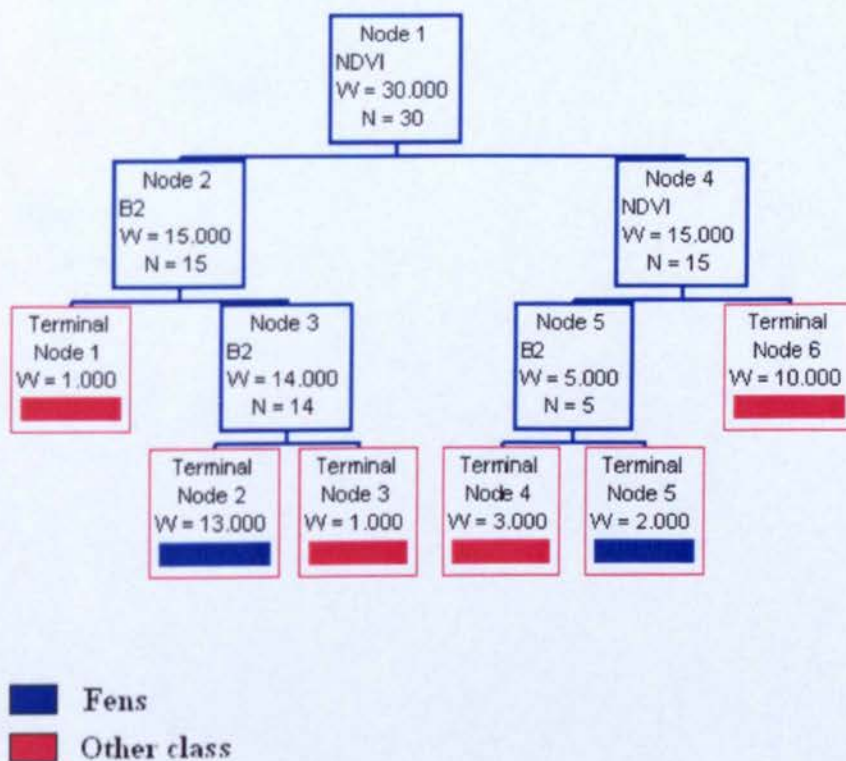
Training size 300 pixels



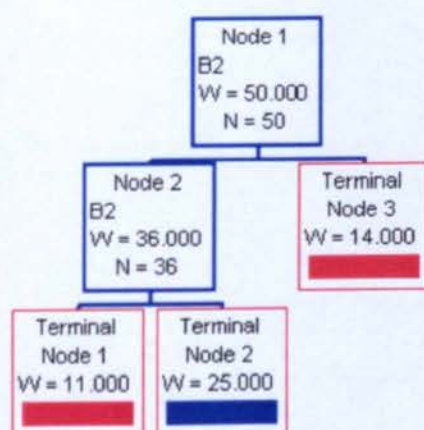

■ Fens
■ Other class

Decision Trees. Optimal tree structure by training set size**Saltmarshes as class of interest**

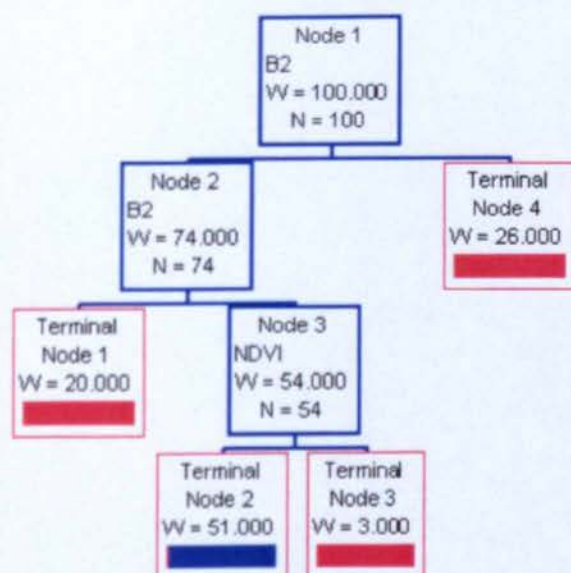
Training size 30 pixels



Training size 50 pixels

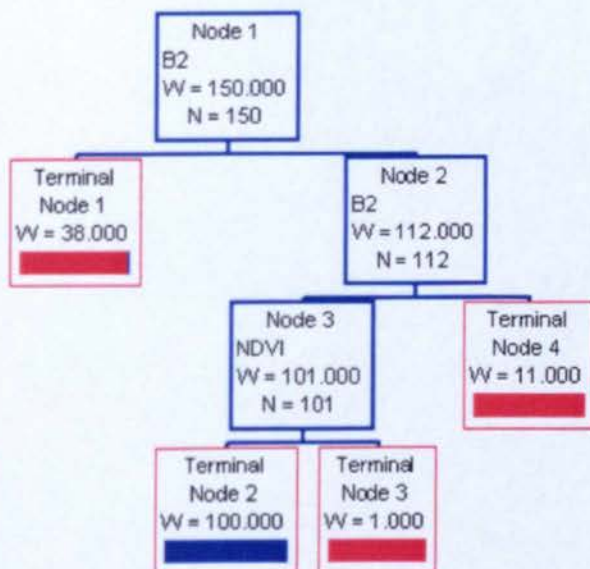

 Fens Other class

Training size 100 pixels

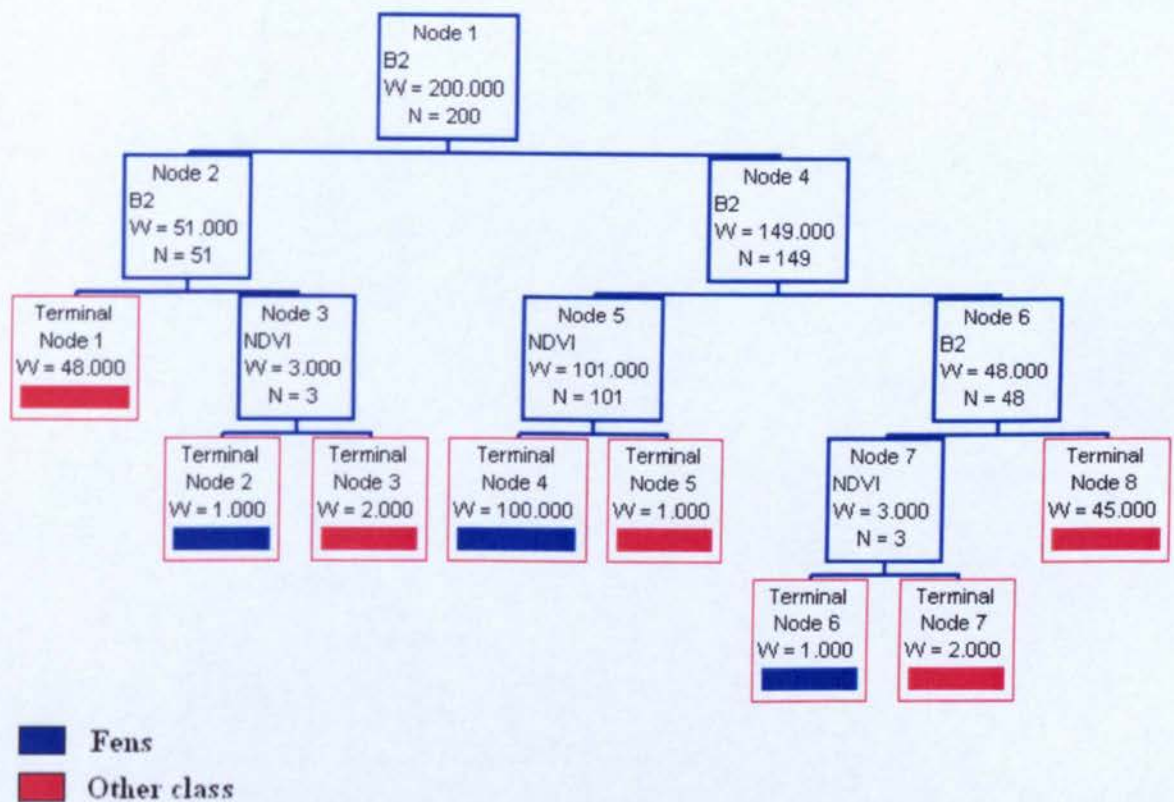


■ Fens
■ Other class

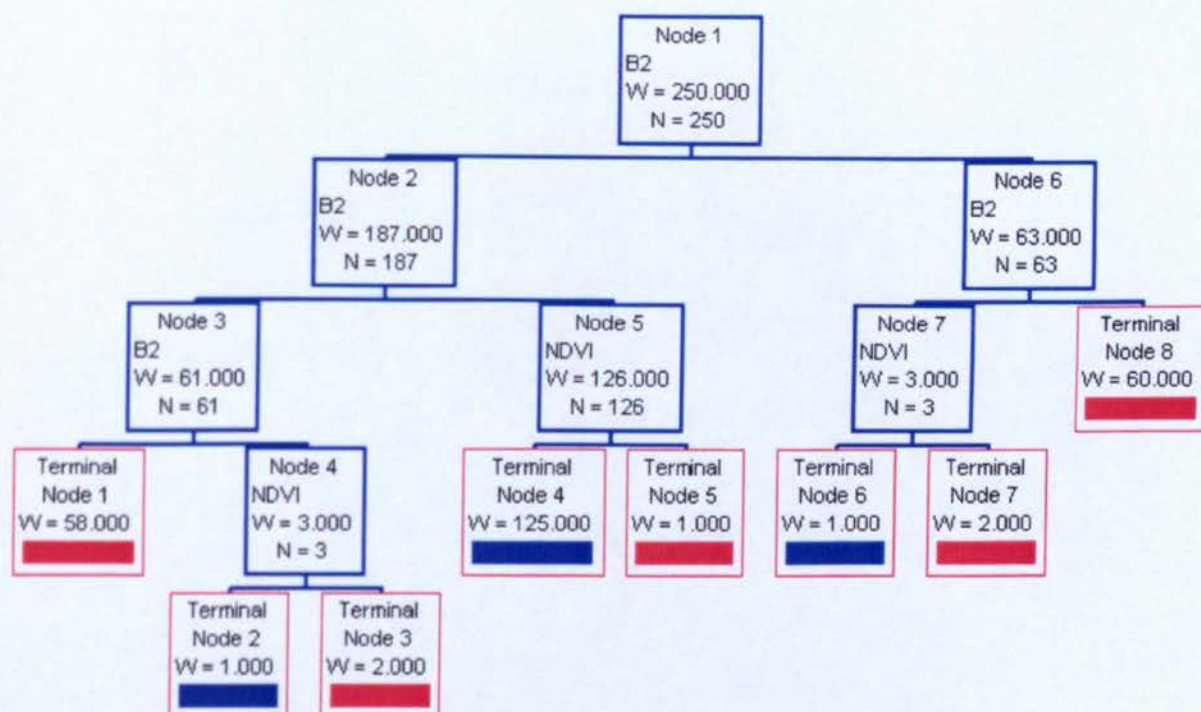
Training size 150

 Fens Other class

Training size 200 pixels

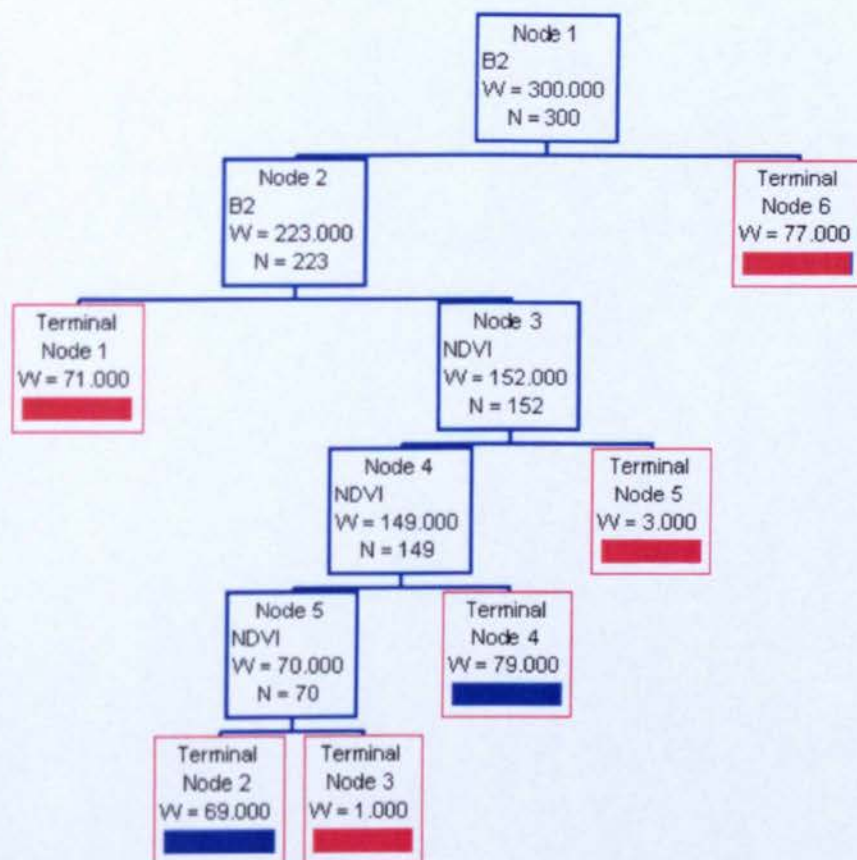



Training size 250 pixels




■ Fens
■ Other class

Training size 300 pixels



 Fens

 Other class

ANNEX E
List of publications

List of publications

Sanchez-Hernandez C., Boyd D., Foody, G. M. (2004), One Class Classification for EU priority Habitats Monitoring, *Proceedings of the RSPSoc Annual Conference 2004*, Aberdeen

Boyd D. S., Sanchez-Hernandez C. and Foody G. M., Mapping a specific class for priority habitats monitoring from satellite sensor data, *International Journal of Remote Sensing* (in press)

Foody G. M., Mathur A., Sanchez-Hernandez C. and Boyd D., Reducing training set size requirements for the classification of a specific class, *Remote Sensing of the Environment* (in press)

Sanchez-Hernandez C., Boyd D. and Foody G. M., One class classification versus binary classification for EU habitat monitoring, *IEEE Transactions in Geosciences and Remote Sensing* (accepted)

Foody G. M., Mathur A., Sanchez-Hernandez C. and Boyd D., Mapping a specific class with an ensemble of classifiers, *International Journal of Remote Sensing*, (accepted)

ANNEX F

Cd-rom with raw data. Training and testing data sets

- Aizerman, M. A., Braverman, E. M. and Rozonoer L. I. (1964).** Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25: 821 – 837.
- Amarnath, G., Murthy, M. S. R., Britto, S. J., Rajashekar, G., and Dutt, C. B. S. (2003).** Diagnostic analysis of conservation zones using remote sensing and GIS techniques in wet evergreen forests of the Western Ghats- An ecological hotspot, Tamil Nadu, India. *Biodiversity and Conservation*, 12: 2331 – 2359.
- Arbia, G., Benedetti, R. and Espa, G. (1999).** Contextual classification in image analysis: an assessment of accuracy of ICM, *Computational Statistics & Data Analysis*, 30: 443 – 455.
- Aronoff, S. (1982).** Classification Accuracy: A User's Approach. *Photogrammetric Engineering & Remote Sensing*, 48(8): 1299 – 1307.
- Arora, M. K. and Foody, G. M. (1997).** Log-linear modelling for the evaluation of the variables affecting the accuracy of probabilistic, fuzzy and neural network classifications. *International Journal of Remote Sensing*, 18 (4): 785 – 798.
- Atkinson, P. M. (1991).** Optimal ground-based sampling for remote sensing investigations: estimating the regional mean. *International Journal of Remote Sensing*, 12: 559 – 567.
- Atkinson, P. M. and Lewis, P. (2000).** Geostatistical classification for remote sensing: an introduction. *Computers and Geosciences*, 26: 361 – 371.
- Badoiu, M., Har-Peled, S. and Indyk, P. (2002).** Approximate clustering via core-sets. In *Proceedings 34th Annual ACM Symposium on Theory of Computing*, Montréal, Québec, Canada, May 19-21.
- Bajjouk, T., Guillaumont, B. and Populus, J. (1996).** Application of airborne imaging spectrometry system data to intertidal seaweed classification and mapping. *Hydrobiologia*, 326/327: 463 – 471.
- Baraldi, A. and Parmiggiani, F. (1995).** A neural network for unsupervised categorisation of multivalued input patterns: an application to satellite image clustering. *IEEE Transactions on Geoscience and Remote Sensing*, 33: 305 – 316.

Barnett V. and Lewis T., (1994). Outliers in Statistical Data. New York: Wiley.

Battiti, R. and Colla, A. M. (1994). Democracy in neural nets: Voting schemes for classification. *Neural Networks*, 7: 691 – 707.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning*, 36(1/2): 525 – 536.

Beaubien, J., Cihlar J., Simard, G. and Latifovic, R. (1999). Land cover from multiple Thematic Mapper scenes using a new enhancement - classification methodology. *Journal of Geophysical Research*, 104(D22): 27909 – 27920.

Benediktsson, J. A. and Swain, P. H. (1992). Consensus theoretic classification methods. *IEEE Transactions on Systems, Man and Cybernetics*, 22: 688 – 704.

Benediktsson, J. A., Sveinsson, J., Ersoy, O. and Swain, P. H. (1997). Parallel consensual neural networks. *IEEE Transactions on Neural Networks*, 8: 54 – 65.

Benediktsson, J. A., Swain, P. H. and Ersoy, O. K. (1990). Neural network approaches versus statistical-methods in classification of multisource remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 28: 540 – 552.

Bennett, K. P. and Campbell, C. (2000). Support Vector Machines: Hype or Hallelujah? *SIGKDD* (Special Interest Group on Knowledge Discovery and Data Mining). *Explorations*, 2 (2): 1 – 13.

Berry, B. J. L. and Baker, A. M. (1968). Geographical sampling. Spatial Analysis: A Reader. In Statistical Geography, Berry, B. J. L., and Marble, D. F. (eds.), Englewood Cliffs, N. J.: Prentice-Hall.

Bischof, H. and Leonardis, A. (1998). Finding optimal neural networks for land use classification. *IEEE Transactions on Geoscience and Remote Sensing* 36: 337 – 341.

- Bishop, C. M. (1994).** Novelty detection and neural network validation. *IEE Proceedings on Vision, Image and Signal Processing. Special Issue on Applications of Neural Networks*, 141(4): 217 – 222.
- Bishop, C. M. (1995).** Neural Networks for Pattern Recognition. Oxford University Press.
- Blamire, P. A. (1996).** The influence of relative sample size in training artificial neural networks. *International Journal of Remote Sensing*, 17: 223 – 230.
- Bolt, H. M., Ickstadt, K., Zucknick, M. and Schwender, H. (2004).** A pilot study on the application of statistical classification procedures to molecular epidemiological data. *Toxicology Letters*, 151(1): 291 – 299.
- Borak, J. S. and Strahler, A. H. (1999).** Feature selection and land cover, classification of a MODIS-like data set for semi-arid environment. *International Journal of Remote Sensing*, 20: 919 – 938.
- Boser, B., Guyon, I. and Vapnik, V. N. (1992).** A training algorithm for optimal margin classifiers. *Proceedings of 5th Annual Workshop on Computer Learning Theory*, Pittsburgh, PA: ACM, 144 – 152.
- Boyd, D. and Foody, G. (2004).** Changing Land Cover, In Global Environmental issues, Harris F. (ed.), Chichester: Wiley.
- Breiman L. (1998).** Arcing classifiers. *The Annals of Statistics*, 26(3): 801 – 849.
- Breiman, L. (1996).** Bagging predictors. *Machine Learning*, 24(2): 123 – 140.
- Breiman, L. (1999).** Prediction games and arcing algorithms. *Neural Computation*, 11: 1493 – 517.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984).** Classification and Regression Trees, Belmont, CA: Wadsworth.

Briem, G. J., Benediktsson, J. A. and Sveinsson J. R. (2002). Multiple Classifiers Applied to Multisource Remote Sensing Data, *IEEE Transactions on Geoscience and Remote Sensing*, 40 (10): 2291 – 2299.

Brodley, C. E. and Friedl, M. A. (1996). Improving automated land cover mapping by identifying and eliminating mislabeled observations from training data. In *Proceedings of the 1996 International Geoscience and Remote Sensing Symposium*, Lincoln, Nebraska, May 27-31.

Brodley, C. E. and Utgoff, P. E. (1992). Multivariate versus univariate decision trees. *Technical Report 92-8*. Department of Computer Science, University of Massachusetts, Amherst, Massachusetts, USA.

Brotherton, T., Johnson, T. and Chadderdon, G. (1998). Classification and novelty detection using linear models and a class dependent— elliptical basis function neural network. In *Proceedings of the IJCNN Conference*, Anchorage, May 4-5.

Brown de Colstoun, E. C. and Wathall, C. L. (2006). Improving global scale land cover classifications with multi-directional POLDER data and a decision tree classifier. *Remote Sensing of Environment*, 100 (4): 474 – 485.

Brown de Colstoun, E. C., Story, M. H., Thompson, C., Commisso, K., Smith, T. G. and Irons, J. R. (2003). National Park vegetation mapping using multitemporal Landsat 7 data and a decision tree classifier. *Remote Sensing of Environment*, 85: 316 – 327.

Brown, M., Gunn, S. R. and Lewis, H. G. (1999). Support vector machines for optimal classification and spectral unmixing. *Ecological Modelling*, 120: 167 – 179.

Brown, M., Lewis, H. G. and Gunn, S. R. (2000). Linear spectral mixture models and support vector machines remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 38: 2346 – 2360.

Bruzzone, L. and Cossu, R. G. (2004). Detection of land-cover transitions by combining multirate classifiers. *Pattern Recognition Letters*, 25(13): 1491 – 1500.

Buchheim, M. P. and Lillesand, T. M. (1989). Semi-automated training field extraction and analysis for efficient digital image classification. *Photogrammetric Engineering and Remote Sensing*, 55: 1347 – 1355.

Burbidge, R., Trotter, M., Buxton, B. and Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computational Chemistry*, 26 (1): 5 – 14.

Burges, C. J .C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):955-974.

Byungyong, K. and Landgrebe, D. (1991). Hierarchical decision tree classifiers in high dimensional and large class data. *IEEE Transactions on the Geosciences and Remote Sensing*, 29 (4): 518 – 528.

Campbell, J. B. (1981). Spatial correlation effects upon accuracy of supervised classification of land cover. *Photogrammetric Engineering and Remote Sensing*, 47: 355 – 363.

Campbell, J. B. (2002). Introduction to Remote Sensing (third ed.): London, Taylor and Francis.

Cao, L. (2003). Support vector machines experts for time series forecasting. *Neurocomputing*, 51: 321 – 339.

Carpenter, G., Gjaja, M., Gopal, S. and Woodcock, C. (1997). ART networks in Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 35(2): 308 – 325.

Cestnik G., Kononenko I., and Bratko I., (1987). Assistant-86: A knowledge - elicitation tool for sophisticated users. In *Progress in Machine Learning*, Bratko, I. and Lavrac, N. (eds.), 31 – 45, U.K.: Wilmslow, Sigma.

Chan, J. C.-W., Huang, C. and Defries, R. S. (2001). Enhanced algorithm performance for land cover classification from remotely sensed data using bagging and boosting. *IEEE Transactions of Geoscience and Remote Sensing*, 39: 693 – 695.

Chan, J. C.-W., Laporte, N. and Defries, R. S. (2003). Texture classification of logged forests in tropical Africa using machine-learning algorithms. *International Journal of Remote Sensing*, 24 (6): 1401 – 1407.

Chang, M. S., Ding, P. Y. A., Wang, S. P., Charng, M. J., Pan, J. P., Hsu, N. W., Chan, W. L., Lin, S. J., Chen, Y. H. and Chen, L. C. (2002). Clinical and angiographic determinants of adverse cardiac events in patients with stent restenosis. *Catheterization and Cardiovascular Interventions*, 55 (3): 331 – 337.

Chapelle, O., Haffner, P. and Vapnik, V. N. (1999). Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10: 1055 – 1064.

Chavanet, P., Portier, H., Lequeu, C., Bergoin, E., Piroth, L., Etienne, M. and Croisier, D. (2004). In vivo pharmacodynamic efficacy of gatifloxacin against *Streptococcus pneumoniae* in an experimental model of pneumonia: impact of the low levels of fluoroquinolone resistance on the enrichment of res. *Journal of Antimicrobial Chemotherapy*, 54: 640 – 647.

Cherkassky, V. and Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks*, 17 (1): 113 – 126.

Cherkassky, V. and Mulier, F. (1998). Learning from Data - Concepts, Theory, and Methods. USA: John Wiley and Sons.

Cherrill, A. J. and McClean, C. (1999). Between-observer variation in the application of a standard method of habitat mapping by environmental consultants in the UK. *Journal of Applied Ecology*, 36: 989 – 1008.

Chow, C.K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1): 41 – 46.

Chuvieco, E. and Congalton, R. G. (1988). Using cluster analysis to improve the selection of training statistics in classifying remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, 54(9): 1275 – 1281.

Cihlar, J. (2000). Land cover mapping of large areas from satellites: status and research priorities. *International Journal of Remote Sensing*, 21(6/7):1093 – 1114

Cihlar, J., Xia, Q. H., Chen, J., Beaubien, J., Fung, K. and Latifovic, R. (1998). Classification by progressive generalization: A new automated methodology for remote sensing multichannel data. *International Journal of Remote Sensing*, 19 (14): 2685 – 2704.

Clark, W. A. V. and Hosking, P. L. (1986). Statistical Methods for Geographers. New York: John Wiley and Sons.

Clements, F. E. (1916). Plant Succession: an analysis of the development of vegetation. Carnegie Institution of Washington D.C. Publication 242.

Cochran, W. G. (1977). *Sampling techniques*. New York: John Wiley and Sons.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37 – 40.

Cohen, W. B. and Goward, S. N. (2004). Landsat's Role in Ecological Applications of Remote Sensing. *BioScience*, 54 (6): 535 – 545.

Condorcet N.C. Marquis de. (1785), *Essai sur l' application de l' analyse `a la probabilit'e des decisions rendues `a la pluralit'e des voix*. Imprimerie Royale, Paris.

Congalton, R. G. (1988). Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, 54: 587 – 592.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37: 35 – 46.

Congalton, R. G. and Green, K. (1999). Assessing the accuracy of remotely sensed data: principles and practices. Boca Raton: Lewis Publishers.

Congalton, R. G. and Mead, R. A. (1983). A quantitative method to test for consistency and correctness in photo-interpretation. *Photogrammetric Engineering and Remote Sensing*, 49(1): 69 – 74.

Cortes, C. and Vapnik, V. N. (1995). Support vector networks. *Machine Learning*, 20: 273 – 297.

Cortijo, F. J. and Perez de la Blanca, N. (1998). Improving classical contextual classifications. *International Journal of Remote Sensing*, 19(8) 1591 – 1613.

Council Directive 92/43/EEC (1992). Of the conservation of natural habitats and wild flora and fauna, the Council of the European Communities.

Cover, T. and Hart, P. (1967). Nearest Neighbour pattern classification. *IEEE Transactions on Information Theory*, 13: 21 – 27.

Cristianini, N. and Taylor, J. (2000). An Introduction to Support Vector Machines. Cambridge University Press.

Cunningham, P. and Carney, J. (2000). Diversity versus Quality in Classification Ensembles Based on Feature Selection. *Technical Report TCD-CS-2000-02*, Dept. of Computer Science, Trinity College, Dublin.

Czerminski, R., Yasri, A. and Hartsough, D. (2001). Use of support vector machines in pattern classification: application to QSAR studies. *Quantitative Structure-Activity Relationships*, 20: 227 – 240.

D' Urso, G. and Menenti, M. (1996). Performance indicators for the statistical evaluation of digital image classifications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 51(2): 78 – 90.

Dale, P. E. R., Hulsman, K. and Chandica, A. L. (1986). Classification of reflectance on colour aerial photographs and sub-tropical saltmarsh vegetation types. *International Journal of Remote Sensing*, 7 (12): 1783 – 1788.

DeFries, R. S. and Chan, J. C.-W. (2000). Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sensing of Environment*, 74: 503 – 515.

DeFries, R. S., Hansen, M., Townshend, J. R. G. and Sohlberg, R. (1998). Global land cover classifications at 8 km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers. *International Journal of Remote Sensing*, 19: 3141 – 3168.

Derbeko, P., El-Yaniv, R. and Meir, R. (2002). Variance optimized bagging. In *Proceedings of the 13th European Conference on Machine Learning*, Helsinki, Finland, August 19 – 23.

Desforges, M.J., Jacob, P.J. and Cooper J.E. (1998). Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institute of Mechanical Engineers*, 212: 687 – 703.

Devijver, P. and Kittler, J. (1982). Pattern Recognition. A Statistical Approach. London: Prentice Hall International.

Dicks, S. E. and Lo, T. H. C. (1990). Evaluation of thematic map accuracy in a land-use and land-cover mapping program. *Photogrammetric Engineering and Remote Sensing*, 56: 1247 – 1252.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, 40(2): 139 – 158.

Donoghue, D. N. M. and Shennan, I. (1987). A preliminary assessment of Landsat TM imagery for mapping vegetation and sediment distribution in the Wash estuary. *International Journal of Remote Sensing* 8 (7): 1101 – 1108.

Drucker H., Cortes C., Jackel L., LeCun Y. and Vapnik, V. (1994). Boosting and other ensemble methods. *Neural Computation*, 6(6): 1289 – 1301.

Drucker, H., Wu, D. and Vapnik, V. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10 (5): 1048 – 1054.

Duda, R. O. and Hart, P. E. (1973). Pattern Classification and Scene Analysis, New York: Wiley and Sons.

Duelli, P. (1997). Biodiversity evaluation in agricultural landscapes: an approach at two different scales. *Agriculture Ecosystems and Environment*, 62: 81 – 91.

Duin, R. P. W. (2002). The combining classifier: to train or not to train. In: *Proceedings 16th International Conference on Pattern Recognition (ICPR 2002)*, Quebec, Canada, August 11-15.

Dunsmuir, W. T. M., Spark, E. and Connor, G. J. (2003). Statistical forecasting techniques to describe the surface winds in Sydney Harbour. *Australian Meteorological Magazine*, 52 (2): 117 – 126.

Eastwood, J. A., Yates, M. G., Thomson, A. G. and Fuller, R. M. (1997). The reliability of vegetation indices for monitoring saltmarsh vegetation cover. *International Journal of Remote Sensing*, 18: 3901 – 3907.

Eghbalnia, H. and Assadi, A. (2001). An application of support vector machines and symmetry to computational modeling of perception through visual attention. *Neurocomputing*, 38–40: 1193 – 1201.

Eklund, P. W., Kirkby, S. D. and Salim, A. (1994). A framework for incremental knowledge Base Update from Additional Data Coverages, in *Proceedings of the 7th Australasian Remote Sensing Conference*, 367 – 374, Melbourne, March 1-4.

Estes, J. A., Hajic, E. J. and Tinney, L. R. (1983). Fundamentals of image analysis: analysis of visible and thermal infrared data, in Colwell R. N., (ed.) *Manual of remote sensing*. (second edition), Vol. 1. American Society of Photogrammetry, Falls Church, Va.

Estes, J., Belward, A., Loveland, T., Scepan, J., Strahler, A., Townshend, J. and Justice, C. (1999). The Way Forward. *Photogrammetric Engineering and Remote Sensing*, 65 (9): 1089 – 1093.

Estes, J., Sailor, C. and Tinney, L. (1986). Applications of artificial intelligence techniques to remote sensing. *Professional Geographer*, 38: 133 – 141.

Evans, F. (1998). An Investigation into the Use of Maximum Likelihood Classifiers, Decision Trees, Neural Networks and Conditional Probabilistic Network for Mapping and Predicting Salinity. M. Sc. Thesis, Department of Computer Science, Curtin University of Technology, Australia.

Fabricius, K.E. and De'ath, G. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, 81: 3178 – 3192.

Fassnacht, K. S., Cohen, W. B. and Spies, T. A. (2006). Key issues in making and using satellite-based maps in ecology: A primer. *Forest Ecology and Management*, 222 (1-3): 167 – 181.

Fayyad, U. and Irani, K. (1992). The Attribute Selection Problem in Decision Tree Generation. In *AAAI-92 Proceedings of the 10th National Conference on Artificial Intelligence*, 104 – 110, AAAI Press/MIT Press, Cambridge, Massachusetts. July 12-16.

Feicht, D. L., Colbert, J. J. and Gottschalk, K. W. (1998). Tree mortality risk of oak due to gypsy moth. *European Journal of Forest Pathology*, 28: 121 – 132.

Filippi, E., Costa, M. and Pasero, E. (1994). Multi-layer perceptron ensembles for increased performance and fault-tolerance in pattern recognition tasks. In *IEEE International Conference on Neural Networks*, 2901–2906, Orlando, Florida, June 28- July 2.

Firth, L. Hazelton, M.L. and Campbell, E. P. (2005). Predicting the onset of Australian winter rainfall by nonlinear classification. *Journal of Climate*, 18 (6): 772 – 781.

Fitzpatrick-Lins, K. (1981). Comparison of sampling procedures and data analysis for a land use and land cover map. *Photogrammetric Engineering and Remote Sensing*, 47: 343 – 351.

Fletcher, R. (1987). Practical Methods of Optimization. New York: John Wiley and Sons.

Fonseca, L. M. G. and Manjunath, B. S., (1996). Registration techniques for multisensor remotely sensed imagery. *Photogrammetric Engineering and Remote Sensing*, 62(9): 1049 – 1056.

Foody, G. M. (1992). On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, 58: 1459 – 1460.

Foody, G. M. (1995). Cross-entropy for the evaluation of the accuracy of a fuzzy land cover classification with fuzzy ground data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 50 (5): 2 – 12.

Foody, G. M. (1998). Sharpening fuzzy classification output to refine the representation of sub-pixel land cover distribution. *International Journal of Remote Sensing*, 19 (13): 2593 – 2599.

Foody, G. M. (1999). The continuum of classification fuzziness in thematic mapping, *Photogrammetric Engineering and Remote Sensing*, 65: 443 – 451.

Foody, G. M. (2002). Status of land cover classification accuracy assessment, *Remote Sensing of Environment*, 80: 185-201.

Foody, G. M. (2004a). Supervised classification by MLP and RBF neural networks with and without an exhaustively defined set of classes. *International Journal of Remote Sensing*, 25: 3091 – 3104.

Foody, G. M. (2004b). Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing*, 70(5): 627 – 633.

Foody, G. M. and Arora, M. K. (1997). An evaluation of some factors affecting the accuracy of classification by an artificial neural network. *International Journal of Remote Sensing*, 18: 799 – 810.

Foody, G. M. and Mathur, A. (2004a). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42: 1335-1343.

Foody, G. M. and Mathur, A. (2004b). Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of Environment*, 93: 107-117.

Foody, G. M., Atkinson, P. M., Gething, P., Ravenhill, N. A. and Kelly, C. K. (2005). Identification of specific tree species in ancient semi-natural woodland from digital aerial sensor imagery. *Ecological Applications* (in press).

Foody, G. M., Lucas, R. M., Curran, P. J. and Honzak, M. (1997). Non-linear mixture modelling without end-members using an artificial neural network. *International Journal of Remote Sensing*, 18 (4): 937 – 953.

Foody, G. M., McCulloch, M. B. and Yates, W. B. (1995). The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing*, 16: 1707 – 1723.

Foody, G. M., Muslim, A. M. and Atkinson, P. M. (2005). Super-resolution mapping of the waterline from remotely sensed data. *International Journal of Remote Sensing*, 26 (24): 5381 – 5392.

Franke, J. and Mandler, E. (1992). A comparison of two approaches for combining the votes of cooperating classifiers. In *Proceedings 11th International Conference on Pattern Recognition*, The Hague, The Netherlands, August 30-September 3.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, Bari, Italy.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and Systems Sciences*, 55(1): 119 – 139.

Freund, Y. and Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14: 771 – 780.

Friedl, M. A. and Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61: 399-409.

Friedl, M. A., Brodley, C. E. and Strahler, A. H. (1999). Maximizing land cover classification accuracies produced by decision trees at continental to global scales. *IEEE Transactions on Geoscience and Remote Sensing*, GE-37: 969 – 977.

Friedl, M. A., McIver, D. K., Hodges, J. C. F., Zhang, X. Y., Muchoney, D., Strahler, A. H., Woodcock, C. E., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F. and Schaaf, C. (2002). Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, 83: 287 – 302.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5): 1189 – 1232.

Fuller, R. M., Smith, G. M., Sanderson, J. M., Hill, R. A., Thomson, A. G., Cox, R., Brown, N. J., Clarke, R. T., Rothery, P. and Gerard, F. F. (2002). Land Cover Map 2000, Module 7 Final Report. Centre for Ecology and Hydrology, Monks Wood, Abbots Ripton, Huntingdon, available from http://www.cs2000.org.uk/mod7_info.htm

Fumera, G., Roli, F. and Giacinto, G. (2000). Reject option with multiple thresholds. *Pattern Recognition*, 33: 2099 – 2101.

Gahegan, M. and West, G. (1998). The classification of complex data sets: an operational comparison of artificial neural networks and decision tree classifiers, *Proceedings of the 3rd International Conference on Geocomputation*, University of Bristol, UK, September 17-19.

Giacinto, G. and Roli, F. (2002). An approach to the automatic design of multiple classifier Systems. *Pattern Recognition Letters*, 22: 25 – 33.

Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R. (2004). Random forest classification of multisource remote sensing and geographic data, *Proceedings 2004 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2004)*. 1049 – 1052, Anchorage, Alaska.

Goel, P. K., Prasher, S. O., Patel, R. M., Landry, J. A., Bonnell, R. B. and Viau, A. A. (2003). Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn. *Computers and Electronics in Agriculture*, 39: 67 – 93.

Griffiths, G. H. (1999). Integrated species and habitat data for nature conservation in Great Britain, *Geographical paper no 130*, Department of Geography, University of Reading.

Gualtieri, J. A. and Crompton, R. F. (1998). Support vector machines for hyperspectral remote sensing classification. In *Proceedings of the SPIE, 27th AIPR Workshop: Advances in Computer Assisted Recognition*, Washington, DC, October 14-16.

Gunn, S. R. (1998). Support Vector Machines for Classification and Regression, (*Technical Report*), Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, Southampton University.

Günter, S. and Bunke, H. (2004). An Evaluation of Ensemble Methods in Handwritten Word Recognition Based on Feature Selection. In *17th International Conference on Pattern Recognition (ICPR 2004)*, Cambridge, UK, 388-392. August 23-26.

Guo, Q., Kelly, M. and Graham, C. H. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling*, 182: 75 – 90.

Ham, J., Chen, Y., Crawford, M. M. and Gosh, J. (2005). Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geosciences and Remote Sensing*, 43: 492 – 501.

Hampson, S. and Volper, D., (1986). Linear function neurons: Structure and training. *Biological Cybernetics*, 53: 203 – 217.

Hansen, L. K. and Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12: 993 – 1001.

Hansen, M. C., DeFries, R. S., Townshend, J. R. G. and Sohlberg, R. (2000). Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, 21: 1331 – 1364.

Hansen, M., Dubayah, R. and Defries, R. (1996). Classification trees: An alternative to traditional land cover classifiers. *International Journal of Remote Sensing*, 17: 1075 – 1081.

Hanssen, L.K., Liisberg, C. and Salamon P. (1997). The error-reject tradeoff. *Open Systems Inform. Dynamics*, 4: 159 – 184.

Harris, F. (2004). Conserving biodiversity resources. In Harris F. (editor), *Global Environmental Issues*, 95 – 113, Chichester: Wiley.

Hashem, S. and Schmeiser B. (1995). Improving Model Accuracy Using Optimal Linear Combinations of Trained Neural Networks. *IEEE Transactions on Neural Networks*, 6 (3): 792 – 794.

He, C., Girolami, M. and Ross G. (2004). Employing optimized combinations of one-class classifiers for automated currency validation. *Pattern Recognition*, 37: 1085 – 1096.

Heerman, P. D. and Khazenie, N. (1992). Classification of multispectral remote sensing data using a back propagation neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 30: 81 – 88.

Hepner, G. F., Logan, T., Ritter, N. and Bryant, N. (1990). Artificial neural network classification using a minimal training set: comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing*, 56: 469 – 473.

Hill, P. and Lewicki, P. (2006), Statistical Methods and applications, Statsoft, Tulsa. Available from <http://www.stat.ufl.edu/>.

Hsu, C. W., Chang, C. C. and Lin, C. J. (2003). A practical guide to support vector classification. Available in <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>, 2003.

<http://www.jncc.gov.uk/>. Joint Nature Conservation Committee. Accessed February 2004.

<http://www.nasa.gov/>. Accessed 12th June 2005

<http://www.stat.ufl.edu/>. University of Florida. Department of Statistics. Accessed 23th May 2004.

<http://www.ukbap.org.uk/>. UK Biodiversity Action Plan, accessed June 2004.

Huang, C., Davis, L. S. and Townshend, J. G. R. (2002). An assessment of support vector machines land cover classification. *International Journal of Remote Sensing* 23: 725 – 749.

Huang, K. Y. (2002). The use of a newly developed algorithm of divisive hierarchical clustering for remote sensing image analysis. *International Journal of Remote Sensing*, 23: 3149 – 3168.

Hubert-Moy, L., Cotonnec, A., Le Du, L., Chardin, A. and Perez, P., (2001). A comparison of parametric classification procedures of remotely sensed data applied on different landscape units, *Remote Sensing of Environment*, 75: 174 – 187.

Ingram J. C., Dawson T. P. and Whittaker R. J. (2005). Mapping tropical forest structure in southeastern Madagascar using remote sensing and artificial neural networks. *Remote Sensing of Environment*, 94 (4): 491 – 507.

Irahama, K., and Furukawa, Y. (1995). Gradient descent learning of nearest neighbor classifiers with outlier rejection. *Pattern Recognition*, 28 (5): 761 – 768.

Jackson, Q. and Landgrebe D. A. (2001). An Adaptive Classifier Design for High-Dimensional Data Analysis with a Limited Training Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, 39 (12): 2664 – 2679.

Jackson, Q. and Landgrebe D. A. (2002). An Adaptive Method for Combined Covariance Estimation and Classification. *IEEE Transactions on Geoscience and Remote Sensing*. 40 (5): 1082 – 1087.

Jain, A. K. and Chandrasekaran, B. (1982). Dimensionality and sample size considerations in pattern recognition practice. In Krishnaiah P. R. and Kanal L. N. (eds.), *Handbook of Statistics*, 2, 835—855 Amsterdam: North-Holland.

Jain, A. K., Duin, R. P. W. and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 22(1): 4 – 37.

Janssen, L. L. F. and van der Wel, F. J. M. (1994). Accuracy assessment of satellite derived land-cover data: A review. *Photogrammetric Engineering and Remote Sensing*, 60: 419 – 426.

Japkowicz, N. (1999). Concept learning in the absence of counter examples: An auto association-based approach to classification. PhD thesis, Graduate School, New Brunswick, Rutgers, The State University of New Jersey

Japkowicz, N., Myers, C. and Gluck, M. (1995). A novelty detection approach to classification. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Montréal, Québec, Canada, August 20-25.

Jensen, J. R. (1996). *Introductory Digital Image Processing - A Remote Sensing Perspective*. London: Prentice Hall.

JNCC, 2004, Common Standards Monitoring Guidance for Lowland Wetland, available in http://www.jncc.gov.uk/pdf/CSM_lowland_wetland.pdf

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features, in: *Proceedings of the European Conference on Machine Learning (ECML)*, Chemnitz, Germany

John, G. H. (1997). Enhancements to the Data Mining Process, PhD Thesis, Computer Science Department, School of Engineering, Stanford University.

Jones, G. (2004). People and Environment, Harlow: Prentice Hall.

Joy, S. M., Reich, R. M. and Reynolds, R. T. (2003). A non-parametric, supervised classification of vegetation types on the Kaibab National Forest using decision trees. *International Journal of Remote Sensing*, 24: 1835 – 1852.

Kalkhan, M. A., Reich, R. M. and Czaplewski, R. L. (1995). Statistical properties of five indices in assessing the accuracy of remotely sensed data using simple random sampling. Presented at the *1995 ACSM/ASPRS Annual Convention and Exposition*, Charlotte, North Carolina

Kasetkasem, T., Arora, M. K. and Varshney, P. K. (2005). Super-resolution land cover mapping using a Markov random field based approach. *Remote Sensing of Environment*, 96: 302 – 314.

Kavzoglu, T. and Mather, P. M. (2003). The Use of Backpropagating Artificial Neural Networks in Land Cover Classification. *International Journal of Remote Sensing*, 24: 4907 – 4938.

Keramitsoglou, I., Kontoes C., Sifakis N., Mitchley J. and Xofis P. (2005). Kernel based re-classification of Earth observation data for fine scale habitat mapping. *Journal for Nature Conservation*, 13: 91 – 99.

Kerr, J. T. and Ostrovsky, M. (2003). From space to species: Ecological applications for remote sensing. *Trends in Ecology and Evolution*, 18: 299 – 305.

Kim, H. C., Pang, S., Je , H-M., Kim, D. and Bang, S. Y. (2003). Constructing support vector machine ensemble. *Pattern Recognition*, 36: 2757 – 2767.

Kimura, F. and Shridhar, M. (1991). Handwritten numerical recognition based on multiple algorithms. *Pattern Recognition*, 24 (10): 969 – 983.

Kittler, J. (1986). Feature selection and extraction, In Young T., and Fu K.-S. (eds.), *Handbook of pattern recognition and image processing*, 203-217, New York: Academic Press.

Kittler, J. (1998). Combining classifiers: a theoretical framework. *Pattern Analysis and Applications*, (1): 18 – 27.

Kittler, J., Hatef, M., Duin, R. P. W. and Matas, J. (1998). On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3): 226 – 239.

Kittler, J., Matas, J., Jonsson, K. and Ramos Sánchez, M. U. (1997). Combining Evidence in Personal Identity Verification Systems. *Pattern Recognition Letters*, 18(9): 845 – 852.

Koh, L. P. and Sodhi, N. S. (2004). Importance of reserves, fragments and parks for butterfly conservation in a tropical urban landscape. *Ecological Applications* 14: 1695 – 1708.

Kohavi R. and John G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97: 273 – 324.

Kohavi R. and Sommerfield, A. (1995). Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD 95)*, Montreal, Quebec, Canada, August 20-21.

Kohonen, T. (2001). *Self-Organizing Maps*. Berlin: Springer-Verlag.

Koller, D. and Sahami, M. (1996). Toward Optimal Feature Selection. In *Proceedings of the 13th International Conference on Machine Learning*, July 3-6, Bari, Italy

- Kononenko, I., Bratko, I. and Roskar, E. (1984).** Experiments in automatic learning of medical diagnostic rules (*Technical report*). Jozef Stefan Institute, Ljubljana, Yugoslavia.
- Koukoulas, S. and Blackburn, G.A. (2001).** Introducing new indices for accuracy evaluation of classified images representing semi-natural woodland environments. *Photogrammetric Engineering and Remote Sensing*, 67(4): 499 – 510.
- Kudo, M., Naoto, M., Jun, T. and Masaru, S. (2003).** Simple termination conditions for k-nearest neighbour method, *Pattern Recognition Letters*, 24 (9-10): 1203 – 1213.
- Kumar, P., Mitchell, J. and Yildirim, A. (2003).** Computing core-sets and approximate smallest enclosing hyperspheres in high dimensions. In *5th workshop on Algorithm Engineering and Experiments*, Baltimore, Maryland, USA; January 11.
- Kumar, R., Jayaraman, V. K. and Kulkarni, B. D. (2005).** A SVM classifier incorporating simultaneous noise reduction and feature selection: illustrative case examples. *Pattern Recognition*, 38: 41 – 49.
- Kuncheva, L. I., (2005).** Diversity in multiple classifier systems. *Information Fusion*, 6: 3 – 5.
- Kuncheva, L. I., Bezdek, J. C. and Duin, R. P. W. (2001).** Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2): 299 – 314.
- Kuo, B. C. and Landgrebe, D. A. (2002).** A covariance estimator for small sample size classification problems and its application to feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 40: 814 – 819.
- Kurnaz M. N., Dokur Z. and Ölmez T. (2004).** Segmentation of remote-sensing images by incremental neural network. *Pattern Recognition Letters*, 26 (8): 1096 – 1104.
- Kwok J. (2000).** The evidence framework applied to support vector machines. *IEEE Transactions on Neural Networks*, 11(5):1162 – 1173.

Lai C., Tax D. M. J., Duin R. P. W., Pekalska E. and Paclik P. (2002). On combining one-class classifiers for image database retrieval. *Lecture Notes in Computer Science*, 2364: 212 – 221.

Lam, L. (2000). Classifier combinations: Implementations and theoretical issues. In *Proceedings of the 1st International Workshop, MCS 2000*, 77–86, Cagliari, Italy.

Lam, L. and Sue, C. (1995). Optimal combination of pattern classifiers. *Pattern Recognition Letters*, 16: 945 – 954.

Langley P. and Sage S. (1994). Induction of selective bayesian classifiers. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 399 - 406. Seattle, WA. USA.

Langley, P. (1996). Elements of machine learning. San Francisco: Morgan Kaufmann.

Lark, R. M. (1995a). A reappraisal of unsupervised classification. I Correspondence between spectral and conceptual classes, *International Journal of Remote Sensing*, 16: 1425 – 1443.

Lark, R. M. (1995b). A reappraisal of unsupervised classification. II Optimal adjustment of the map legend and a neighbourhood approach for mapping legend units, *International Journal of Remote Sensing*. 16: 1445 – 1460.

Lark, R. M. (1995c). Components of accuracy of maps with special reference to discriminant analysis of remote sensor data. *International Journal of Remote Sensing*, 16, pp. 1461 – 1480.

Lawrence, R. L. and Labus, M., (2003). Early detection of douglas-fir beetle infestation with subcanopy resolution hyperspectral imagery. *Western Journal of Applied Forestry*, 18: 202 – 206.

- Lawrence, R. L. and Wright, A. (2001).** Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering and Remote Sensing*, 67: 1137 – 1142.
- Lawrence, R. L., Bunn A., Powell S. and Zambon M. (2004).** Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90: 331 – 336.
- Le Cun Y., Jackel L. D., Bottou L., Cortes C., Denker J., Drucker H., Guyon I., Muller U. A, Sackinger E., Simard P. and Vapnik V. (1995),** Learning algorithms for classification: a comparison on handwritten digit recognition. In: Kwon J. H. and Cho, S. (Eds.), *Neural Networks: The Statistical Mechanics Perspective*, 261–276, Singapore: World Scientific.
- Ledoux, L., Crooks, S., Jordan, A. and Turner, K. (2000).** Implementing EU biodiversity policy UK experiences. *Land Use Policy*, 17: 257 – 268.
- Lees, B. G. and Ritman, K. (1991).** Decision-Tree and Rule-Induction Approach to Integration of Remotely Sensed and GIS Data in Mapping Vegetation in Disturbed or Hilly Environments. *Environmental Management*, 15 (6): 823 – 831.
- Lillesand, T. M. and Kiefer, R. W. (2004).** Remote sensing and image interpretation (5th ed.). New York: Wiley.
- Lindeman, R. L. (1942).** The trophic-dynamic aspect of ecology. *Ecology*, 23: 399 – 418.
- Liu, H. and Motoda, H. (eds.) (1998).** Feature extraction, construction and selection: A data mining perspective. Kluwer Academic Publishers.
- Lunetta, R. S., Congalton, R. G., Lynn, F. K., Jensen J. R., McGwire K. C. and Tinney L. R. (1991).** Remote sensing and geographic information systems data integration: Error sources and research issues. *Photogrammetric Engineering and Remote Sensing*, 57: 677 - 687.
- Ma, J., Krishnamurthy, A. and Ahalt S. (2004).** SVM training with duplicated samples and its application in SVM-based ensemble methods. *Neurocomputing*, 61: 455 – 459.

Ma, Z. and Redmond, R. L. (1995). Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Engineering and Remote Sensing*, 61: 435 – 439.

Macarthur, R. and Wilson, E. O. (1967). The Theory of Island Biography. Princeton: Princeton University Press.

Manevitz, L. M. and Yousef, M. (2001). One-class SVMs for document classification. *Journal of Machine Learning Research*, 2: 139 – 154.

Mannan, B., Roy, J., and Ray, A. K. (1998). Fuzzy ARTMAP Supervised Classification of Multispectral Remotely-sensed Images, *International Journal of Remote Sensing*, 19 (4): 767 – 774.

Markou M. and Singh S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 83: 2481 – 2497.

Mas, J. F. (2004). Mapping land use/cover in a tropical coastal area using satellite sensor data, GIS and artificial neural networks. *Estuarine, Coastal and Shelf Science*, 59 (2): 219-230.

Mather, P. M. (2004). *Computer-Processing of Remotely-Sensed Images*, Third edition, Chichester: John Wiley and Sons.

Mattera, D. and S. Haykin, (1999). Support vector machines for dynamic reconstruction of a chaotic system. In Schölkopf, B., Burges, C. J. C. and Smola, A. J., (eds.), *Advances in Kernel Methods*, 211- 241, Cambridge, MA: MIT Press,

Mehner, H., Cutler, M., Fairburn, D. and Thompson, G. (2004). Remote sensing of upland vegetation: The potential of high spatial resolution satellite sensors. *Global Ecology and Biogeography*, 13: 359 – 369.

Mills, E. L. (1969). The community concept in marine zoology, with comments on continua and instability in some marine communities: a review. *Journal of Fisheries Research Board of Canada* 26: 1415 – 1428.

Morgan, J. and Sonquist, J. A. (1963). Problem in the analysis of survey data, and a proposal. *Journal of American Statistical Association*, 58: 415 – 435.

Morgan, J. N. and Messenger, R. C. (1973). THAID: A sequential analysis program for the analysis of nominal scale dependent variables. *Technical report*, Institute of Social Research, University of Michigan: Ann Arbor Press.

Mowrer, H. T. and Congalton, R. G. (2000). Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and remote sensing. Chelsea, Michigan: Ann Arbor Press.

Moya, M. and Hush, D.R. (1996). Network constraints and multi-objective optimization for one-class classification. *Neural Networks* 9(3): 463 – 474.

Moya, M., Koch, M. and Hostetler, L. (1993). One-class classifier networks for target recognition applications. In *Proceedings of the World congress on neural networks*, Portland, OR. International Neural Network Society, INNS.

Muchoney, D. M. and Strahler, A. H. (2002). Pixel- and site-based calibration and validation methods for evaluating supervised classification of remotely sensed data. *Remote Sensing of Environment*, 81: 290 – 299.

Mumby, P. J. and Edwards, A. J. (2002). Mapping marine environments with IKONOS imagery: Enhanced spatial resolution can deliver greater thematic accuracy. *Remote Sensing of Environment*, 82: 248 – 257.

Nagendra, H. and Gadgil, M. (1999). Using remote sensing to assess biodiversity. *International Journal of Remote Sensing*, 22: 2377 – 2400.

Nilsson, N. J. (1965). *Learning machines*. New York: McGraw-Hill.

Oates, T. and Jensen, D. (1997). The effects of training set size on decision tree complexity. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 254-262, Fisher D. (ed). Morgan Kaufmann.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, 11: 169 – 198.

Osuna, E., Freund, R. and Girosi, F., (1997). Training support vector machines: An application to face detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR '97*, 130 – 136. Puerto Rico.

PAA (Penny Anderson Associates), (2001), Information review for wetland habitat action plans. English Nature, UK.

Pal, M. (2002), Factors influencing the accuracy of remote sensing classifications: a comparative study. Unpublished PhD Thesis. University of Nottingham

Pal, M. and Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, 86: 554 – 565.

Pal, M. and Mather, P. M. (2005). Support Vector classification in remote sensing. *International Journal of Remote Sensing, Remote Sensing Letters*, 26 (5): 1007 – 1011.

Paola, J. D. and Schowengerdt, R. A. (1995). A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification. *IEEE Transactions in Geosciences and Remote Sensing*, 33: 981 – 996.

Parra L., Deco G. and Miesbach S. (1995). Statistical independence and novelty detection with information preserving non-linear maps. *Neural Computation*, 8 (2): 260 – 269.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33: 1065 – 1076.

Patterson, A. and Niblett, T. (1983). ACLS user manual. Glasgow: Intelligent Terminals Ltd.

Pavlov, D., Mao, J. and Dom B. (2000). Scaling-up support vector machines using the boosting algorithm. In *Proceedings of the International Conference on Pattern Recognition*, 19–22, Barcelona, Spain, September 3–7.

Phinn S. R., Menges, C., Hill G. J. E. and Stanford M. (2000). Optimizing Remotely Sensed Solutions for Monitoring, Modeling, and Managing Coastal Environments. *Remote Sensing of Environment*, 73: 117 – 132.

Phinn S. R., Stow D. A. and van Mouwerick D. (1999). Remotely Sensed Estimates of Vegetation Structural Characteristics in Restored Wetlands, Southern California. *Photogrammetric Engineering & Remote Sensing*, 65 (4): 485 – 493.

Pickett, S. T. A., Kolasa, J., Armesto, J. J. and Collins, S. L. (1989). The ecological concept of disturbance and its expression at various hierarchical levels. *Oikos*, 54: 129 – 136.

Piper, J. (1992). Variability and bias in experimentally measured classifier error rates. *Pattern Recognition Letters*, 13: 685 – 692.

Pontius, R. G. (2000). Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing*, 66: 1011 – 1016.

Pullin, A. S., Knight, T. M., Stone, D. A. and Charman, K. (2004). Do conservation managers use scientific evidence to support their decision making? *Biological Conservation*, 119: 245 – 252.

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (ed.), *Expert systems in the micro electronic age*. Edinburgh University Press.

Quinlan, J. R. (1986). Induction of Decision Trees, *Machine Learning*, 1 (1): 81 – 106.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann.

Quinlan, J. R. (1996). Bagging, boosting and C4.5. In *Proceedings of the 13th national conference on artificial intelligence*, 725 – 730. Portland, OR, USA, August.

Read, J. M., Clark, D. B., Venticinque, E. M. and Moreira, M. P. (2003). Application of merged 1-m and 4-m resolution satellite data to research and management in tropical forests. *Journal of Applied Ecology*, 40: 592 – 600.

Richards, J. A. (1993). Remote Sensing Digital Image Analysis- An Introduction. Berlin : Springer-Verlag.

Rifkin, R. M. and Klautau, A. (2004). In Defense of One-Vs-All Classification. *Journal of Machine Learning Research*, 5: 101 – 141.

Ritter, G. and Gallegos, M. (1997). Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters*, 18: 525–539.

Roberts, S. and Tarassenko, L. (1994). A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6: 270 – 284.

Rogan, J., Miller, J., Stow, D., Franklin, J., Levien, L. and Fisher, C. (2003). Land-cover change monitoring with classification trees using Landsat TM and ancillary data. *Photogrammetric Engineering and Remote Sensing*, 69: 793 – 804.

Rogova, G. (1994). Combining the results of several neural network classifiers. *Neural Networks*, 7 (5): 777 – 781.

Roli, F. and Giacinto, G. (2002). Design of multiple classifier systems, in: *Hybrid Methods in Pattern Recognition*, World Scientific Publishing, 199 – 226.

Rosenfield G.H. and Fitzpatrick-Lins K. (1986). A Coefficient of Agreement as a Measure of Thematic Classification Accuracy. *Photogrammetric Engineering and Remote Sensing*, 52 (2): 223 – 227.

- Rosenfield, G. H. and Fitzpatrick-Lins, K. (1986).** A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing*, 52: 223 – 227.
- Rouget, M. (2003).** Measuring conservation value at fine and broad scale: implications for diverse and fragmented regions. *Biological Conservation*, 112: 217–232.
- Roy, P. S. and Tomar, S. (2000).** Biodiversity characterization at landscape level using geospatial modeling technique. *Biological Conservation* 95(1): 95 – 109.
- Sabins, F.F. (1997).** Remote Sensing. Principles and Interpretation. Third Edition. New York: W.H. Freeman.
- Safavian, S. R. and Landgrebe, D. (1991).** A survey of decision tree classifier methodology. *IEEE Transactions of Systems, Man, and Cybernetics*, 21: 660 – 675.
- Sánchez, M. S. and Sarabia L. A. (1995).** Efficiency of multi-layered feed-forward neural networks on classification in relation to linear discriminant analysis, quadratic discriminant analysis, and regularized discriminant analysis. *Chemometrics and Intelligent Laboratory Systems*, 28: 287 – 303.
- Schaale, M. and Furrer, R. (1995).** Land surface classification by neural networks. *International Journal of Remote Sensing*, 16: 3003 – 3031.
- Schalkoff, R. J. (1992).** Pattern Recognition: Statistical, Structural and Neural Approaches. New York: Wiley.
- Schölkopf, B., Burges, J. C. and Smola, A. J. (1999).** Advances in Kernel Methods. Cambridge, MA: MIT Press.
- Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. and Williamson, R. C. (2001).** Estimating the support of a high-dimensional distribution. *Neural Computation*. 13: 1443 – 1471.

Schölkopf, B., Williamson, R., Smola, A. and Shawe-Taylor, J. (1999). SV estimation of a distribution's support. In *13th Neural Information Processing Systems Meeting Proceedings*, Denver, USA.

Schowengerdt, R. A. (1983). Techniques for image processing and classification in remote sensing. New York: Academic Press.

Schowengerdt, R. A. (1997). Remote Sensing Models and Methods for Image Processing. New York: Academic Press.

Sharkey, N., Gerecke, U. and Chandroth G. (2000). The test and select approach to ensemble combination. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems. First International Workshop, MCS 2000, Cagliari, Italy*, volume 1857 of *Lecture Notes in Computer Science*, 30–44. Springer-Verlag.

Shipp, C.A. and Kuncheva, L. I. (2002). Relationship between combination methods and measures of diversity in combining classifiers. *Information Fusion*, 3 (2): 135 – 148.

Silvestri, S., Marani, M., Settle, J., Benvenuto, F. and Marani, A. (2002). Salt marsh vegetation radiometry: data analysis and scaling. *Remote Sensing of the Environment*, 2: 473 – 482.

Skurichina, M. (2001). Stabilizing Weak Classifiers. PhD thesis, Delft University of Technology, Delft, The Netherlands

Skurichina, M., Kumcheva, L. I. And Duin R. P. W. (2002). Bagging and Boosting for the Nearest Mean Classifier: Effects of Sample Size on Diversity and Accuracy. In *Proceedings of the 3rd International Workshop Multiple Classifier Systems*, Cagliari, Italy, *Lecture notes in computer science*, 62-71, Springer-Verlag.

Smith, G. M., Spencer, T., Muray, A. L. and French, J. R. (1998). Assessing seasonal change in coastal wetlands with airborne remote sensing. *Mangroves and SaltMarshes*, 2: 15 – 28.

Smits, P. C., Dellepiane, S. G. and Schowengerdt, R. A. (1999). Quality assessment of image classification algorithms for land-cover mapping: a review and proposal for a cost-based approach. *International Journal of Remote Sensing*, 20: 1461 – 1486.

Smola, A. J., Schölkopf, B. and Müller, K. (1998). The connection between regularization operators and support vector kernels. *Neural Networks*, 11: 637 – 649.

Spellman, G. (1999). An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom, *Applied Geography* 19 (2): 123 – 136.

Stehman, S. V. (1996). Estimating the kappa coefficient and its variance under stratified random sampling. *Photogrammetric Engineering and Remote Sensing*, 62: 401 – 407.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy, *Remote Sensing of Environment*, 62: 77 – 89.

Stehman, S. V. and Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: fundamental principles. *Remote Sensing of Environment*, 64: 331 – 344.

Story, M. and Congalton, R. G. (1986). Accuracy assessment: a user's perspective. *Photogrammetric Engineering and Remote Sensing*, 52: 397 – 399.

Strahler, A. H. (1980). The use of prior probabilities in maximum likelihood classification of remotely sensed data. *Remote Sensing of the Environment*, 10:135 – 143.

Swain, P. H. and Davis, S. M. (1978). Remote Sensing: The Quantitative Approach, New York: McGraw-Hill.

Swain, P. H. and Hauska, H. (1997). The Decision Tree Classifier: Design and Potential. *IEEE Transactions on Geoscience Electronics*, 15: 142 – 147.

Tadjudin, S. and Landgrebe, D. A. (1999). Covariance estimation with limited training samples. *IEEE Transactions on Geosciences and Remote Sensing*, 37: 2113 – 2118.

Tadjudin, S. and Landgrebe, D. A. (2000). Robust parameter estimation for mixture model. *IEEE Transactions on Geoscience and Remote Sensing*, 38: 439 – 445.

Tarassenko, L. (1995). Novelty detection for the identification of masses in mammograms. In *Proceedings of the Fourth International IEE Conference on Artificial Neural Networks*, Cambridge, UK.

Tarassenko, L., Hayton, P. and Brady, M. (1995). Novelty detection for the identification of masses in mammograms. In *Proceedings of the Fourth International IEE Conference on Artificial Neural Networks*, 442–447, Cambridge, UK.

Tarassenko, L., Nairac, A. and Townsend, N. (1999). Novelty detection in jet engines, *IEEE Colloquium on Condition Monitoring, Imagery, External Structures and Health*, 41 – 45.

Tax, D. M. J. (2001). One-class classification. Un published Ph.D. Thesis, Delf University of Technology. Available from <<http://www.ph.tn.tudelft.nl/~davidt/papers.html>>.

Tax, D. M. J. (2004). Data description toolbox. DD tools 1.12, A Matlab toolbox for data description, outlier and novelty detection. Available from <<http://www.ph.tn.tudelft.nl/~davidt>>.

Tax, D. M. J. and Duin, R. P. W. (1999). Data domain description using support vectors. In *Proceedings of the European Symposium on Artificial Neural Networks*, Bruges, Belgium, April, 21-23, 251–256.

Tax, D. M. J. and Duin, R. P. W. (2001). Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, 2 155 – 173.

Tax, D. M. J. and Duin, R. P. W. (2002). Uniform Object Generation for

Optimizing One-class Classifiers, *Journal of Machine Learning Research, Special Issue on Kernel Methods*, 2 (2): 155 – 173.

Tax, D. M. J. and Duin, R. P. W. (2004). Support Vector Data Description, *Machine Learning*, 54: 45 – 66.

.

Tay, F. E. H. and Cao, L. J., (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing*, 48 (1–4): 847 – 861.

The Broads Authority, (2004). Broads Action Plan 2004. Available from <http://www.broads-authority.gov.uk/broads/pages/Publica.html>

Tresp, V. and Taniguchi, M. (1995). Combining Estimators Using Non-Constant Weighting Functions. In *Advances in Neural Information Processing Systems 7*. Tesauro, G., Touretzky, D.S., and Leen, T.K., (eds). Cambridge, Mass.: MIT Press.

Tso, B. C. K. (1997). An Investigation of Alternate Strategies for Incorporating Spectral, Textural, and Contextual Information in Remote Sensing Image Classification. PhD Thesis. School of Geography, The University of Nottingham, Nottingham, UK.

Tso, B. C. K. and Mather, P. M. (2001). Classification Methods for Remotely sensed Data. London: Taylor and Francis.

Tumer, K. and Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3/4): 385 – 404.

Turk R. K. (1979). GT index: a measure of the success of prediction. *Remote Sensing of Environment*, 8: 65–75.

Turner, R. K., Adger, N. and Brouwer, R. (1998). Ecosystems services value, research needs and policy relevance. *Ecological Economics*, 25: 61-65.

Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E. and Steininger, M. (2003). Remote sensing for biodiversity science and conservation. *Trends in Ecology and Evolution*, 18: 306 – 314.

Valentini, G. and Masulli, F. (2002). Ensembles of Learning Machines. *Lecture Notes in Computer Science*, 2486/2002: 3 – 19.

Valentini, G., Muselli, M. and Rufino, F. (2003). Bagged ensembles of SVMs for gene expression data analysis. In *Proceeding of the International Joint Conference on Neural Networks*, 1844–1849, Portland, OR, USA.

Valentini, G., Muselli, M. and Rufino, F. (2004). Cancer recognition with bagged ensembles of support vector machines. *Neurocomputing*, 56 (1): 461 – 466.

van der Meer, F. D. (1995). Spectral unmixing of LANDSAT thematic mapper data. *International journal of Remote Sensing*, 16: 3189 – 3194.

van Kooten, G. C., Bulte, E. H. and Sinclair, A. E. R. (editors) (2000). Conserving Nature's Diversity: Insights from Biology, Ethics and Economics. Aldershot, UK: Ashgate Publishing.

van Niel, T. G., McVicar, T. R. and Datt, B. (2005). On the relationship between training sample size and data dimensionality of broadband multi-temporal classification, *Remote Sensing of Environment*, 98: 468-480.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. New York: Springer-Verlag.

Vapnik, V. (1998). Statistical Learning Theory. New York: Wiley.

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transactions of Neural Networks*, 10: 988 – 999.

Vapnik, V. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 17: 264 – 280.

Vapnik, V. and Chervonenkis, A. Y. (1979). *Theory of Pattern Recognition*. Berlin: Akademie-Verlag.

Verhoeve, J. and de Wulf, R. (2002). Land cover mapping at sub-pixel scales using linear optimization techniques. *Remote Sensing of Environment*, 79: 96 – 104.

Vieira, C. A. O., Mather, P. M. and Aplin, P. (2004). Assessing the positional and thematic accuracy of remotely sensed data, *24th ISPRS Congress*, July 12-23, Istanbul, Turkey.

Walmsley, J. L., Barthelmie, R. J. and Burrows, W.R. (2001). The Statistical Prediction Of Offshore Winds From Land-Based Data For Wind-Energy Applications. *Boundary-Layer Meteorology*, 101 (3): 409 – 433.

Webster, R., Curran, P. J. and Munden, J. W. (1989). Spatial correlation in reflected radiation from the ground and its implications for sampling and mapping by ground-based radiometry. *Remote Sensing of Environment*, 29: 67 – 78.

Weiers, S., Bock, M., Wissen, M. and Rossner, G. (2004). Mapping and indicator approaches for the assessment of habitats at different scales using remote sensing and GIS methods, *Landscape and Urban Planning*, 67: 43–65.

Wilkinson, G. G. (2000). Processing and classification of satellite images. In *Encyclopaedia of Analytical Chemistry*, 8679 – 8693, Meyers R. A. (editor). John Wiley and Sons.

Windeatt, T. (2003). Vote counting measures for ensemble classifiers. *Pattern Recognition*, 36(12): 2743 – 2756.

Wolpert, D. H. (1992). Stacked generalization. *Technical Report LA-UR-90-3460*, Complex Systems Group, Theoretical Division, and Center for Non-linear Studies, MS B213, LANL, Los Alamos, N.M.

Woods, K., Kegelmeyer, W.P., and Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(4): 405 – 410.

Xu, L., Krzyzak, C. and Suen, C. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3): 418 – 435.

Yang, C.-C., Prasher, S. O, Enright, P., Madramootoo, C., Burges, M., Goel, P. K. and Callum, I. (2003). Application of decision tree technology for image classification using remote sensing data. *Agricultural Systems*, 76: 1101 – 1117.

Yeung, D.Y. and Chow, C. (2002). Parzen window network intrusion detectors. In *Proceedings of the International Conference on Pattern Recognition*, Quebec, Canada.

Yool, S. R. (1998). Land cover classification in rugged areas using moderate-resolution multispectral data and an artificial neural network. *International Journal of Remote Sensing*, 19(1): 85 – 96.

Yu, S. (2003). Feature Selection and Classifier Ensembles: A Study on Hyperspectral Remote Sensing Data. Unpublished PhD thesis. The University of Antwerp, The Netherlands.

Yuan, E. and Cho, S. (2006). Constructing response model using ensemble based on feature subset selection. *Expert Systems with Applications*, in press

Zhu, G. and Blumberg, D. G. (2002). Classification using ASTER data and SVM algorithms; The case study of Beer Sheva, Israel. *Remote Sensing of Environment*, 80: 233 – 240.

Zhuang, X., Engel, B. A., Lozanogarcia, D. F., Fernandez, R. N. and Johannsen, C. J. (1994). Optimization of training data required for neuro-classification. *International Journal of Remote Sensing*, 15: 3271 – 3277.

Zimmerer, K. S. (1994). Human geography and the "new ecology": The prospect and promise of integration. *Annals of the Association of American Geographers*, 84: 108 – 25.