An Evaluation of Feature Selection and Reduction Algorithms for Network IDS Data

Therese Bjerkestrand, Dimitris Tsaptsinos, Eckhard Pfluegel

*Kingston University, Faculty of SEC, School of CIS, Penhryn Road, Kingston Upon Thames, KT1 2EE, United Kingdom*

**Abstract**

Intrusion detection is concerned with monitoring and analysing events occurring in a computer system in order to discover potential malicious activity. Data mining, which is part of the procedure of knowledge discovery in databases, is the process of analysing the collected data to find patterns or correlations. As the amount of data collected, stored and processed only increases, so does the significance and importance of intrusion detection and data mining.

A dataset that has been particularly exposed to research is the dataset used for the Third International Knowledge Discovery and Data Mining Tools competition, KDD99. The KDD99 dataset has been used to identify what data mining techniques relate to certain attack classes and employed to demonstrate that decision trees are more efficient than the Naïve Bayes model when it comes to detecting new attacks. When it comes to detecting network intrusions, the C4.5 algorithm performs better than SVM.   
 The aim of our research is to evaluate and compare the usage of various feature selection and reduction algorithms against publicly available datasets. In this contribution, the focus is on feature selection and reduction algorithms. Three feature selection algorithms, consisting of an attribute evaluator and a test method, have been used. Initial results indicate that the performance of the classifier is unaffected by reducing the number of attributes.

*Keywords: Data mining, KDD dataset, intrusion detection, knowledge discovery, feature selection and reduction*

**INTRODUCTION**

There are two types of intrusion detection: misuse detection and anomaly detection. Misuse detection uses a database consisting of signatures of recognized intrusions to match and identify unlabelled data. The signatures are used to define abnormal behaviour. Any behaviour that is not recognized is considered normal (Panda & Patra, 2008).  
 Anomaly detection uses an opposite approach. A model is built based on what is considered normal behaviour. Any variation from the model will be classified as abnormal (Panda & Patra, 2008). Compared to misuse detection, anomaly detection is considered more powerful. This is due to its theoretical potential for discovering novel attacks (Amudha & Abdul Rauf, 2011).  
 Data mining is part of the knowledge discovery in databases (KDD) procedure and is the process of analysing collected data to discover patterns or correlations. The KDD procedure can be seen as five steps: Data selection, data cleaning/pre-processing, data reduction, data mining and interpretation/evaluation (Jensen & Shen, 2008).

**KDD99 DATASET**

The KDD99 dataset is a collection of simulated raw TCP dump data collected over a period of nine weeks (Amudha & Abdul Rauf, 2011), and it has been particularly exposed to research. The dataset consists of 41 attributes and instances of normal data and 22 different attacks. The attacks are divided into four categories: Denial-of-Service (DoS), Probe, User-to-Root (U2R) and Remote-to-Local (R2L) (Kdd.ics.uci.edu, 2015). There are approximately five million instances in the KDD99 dataset (Tavallaee, Bagheri, Lu & Ghorbani, 2009).

|  |  |  |
| --- | --- | --- |
| **KDD99 attributes** | | |
| duration | protocol\_type | srv\_serror\_rate |
| service | src\_bytes | srv\_diff\_host\_rate |
| dst\_bytes | flag | dst\_host\_srv\_count |
| land | wrong\_fragment | dst\_host\_diff\_srv\_rate |
| urgent\_ | hot | srv\_rerror\_rate |
| num\_failed\_logins | logged\_in | dst\_host\_count |
| lnum\_compromised | lroot\_shell | dst\_host\_same\_srv\_rate |
| lsu\_attempted | lnum\_root | dst\_host\_same\_src\_port\_rate |
| lnum\_file\_creations | lnum\_shells | dst\_host\_srv\_diff\_host\_rate |
| lnum\_access\_files | lnum\_outbound\_cmds | dst\_host\_srv\_serror\_rate |
| is\_host\_login | is\_guest\_login | dst\_host\_srv\_rerror\_rate |
| count | serror\_rate | dst\_host\_serror\_rate |
| rerror\_rate | same\_srv\_rate | dst\_host\_rerror\_rate |
| diff\_srvrate | srv\_count | *label (class)* |

For this experiments, three new datasets were created, consisting of randomly selected instances from the original KDD99 dataset. No changes were made to the attributes or the class, meaning the new datasets all have 41 attributes and uses the attribute ‘label’ as the class.   
 The training set created contains of the DoS attacks *pod, teardrop, land, back* and an equal amount of *normal* instances. Duplicates in the datasets were removed, resulting in 5,536 instances in the training set. The first test set created contains unseen instances of the DoS attack *neptune* and has a total of 239,615 instances. The second test set contains unseen *normal* instances. To create a common label and to be able to classify data as either DoS or normal, all attacks were renamed and labelled *dos*.

**RELATED WORK**

Panda and Patra (Panda & Patra, 2008) used the KDD99 dataset when the performances of three well-known data mining classifier algorithms, ID3, J48 and Naïve Bayes, were evaluated. Their results demonstrated that Naïve Bayes is one of the most effective algorithms while decision trees, ID3 and J48, are more interesting when it comes to detecting novel attacks (Panda & Patra, 2008). Nguyen and Choi (Nguyen & Choi, 2008) used the KDD99 dataset to evaluate performances of a comprehensive set of classifier algorithms. An extensive performance comparison is presented and they propose two models for algorithm selection. Ektefa *et al* (Ektefa, Memar, Sidi & Affendey, 2010) used the same dataset and data mining techniques for intrusion detection. They discovered that the C4.5 algorithms performs better than SVM in detecting network intrusions and false alarm rates.

**EXPERIMENTS**

The decision tree learning algorithm C4.5, known as J48 in Weka, was applied to the training set. The accuracy of J48 was also tested against both test sets. All tests classified the training and testing sets correctly.   
 Three feature selection algorithms were chosen and applied to the training sets. Each feature selection algorithm (FSA) consists of an attribute evaluator and a search method.

|  |  |  |
| --- | --- | --- |
| No. | Attribute evaluator | Search method |
| FSA1 | Fuzzy Rough Subset Evaluation | Hill Climber |
| FSA2 | Correlation Attribute Evaluation | Ranker |
| FSA3 | Cf Subset Evaluation | Best First |

The results showed that the different algorithms chose different attributes. FSA1 selected *service*, *flag* and *src\_bytes*. FSA2 produced a list ranking all attributes from 0-1, and attributes with a ranking of zero were excluded (l*num\_root, lnum\_file\_creations, lsu\_attempted, is\_guest\_login, lnum\_outbound\_cmds, urgent, num\_failed\_logins* and *is\_host\_login)*. FSA3 selected four attributes: *src\_bytes, dst\_bytes, dst\_host\_srv\_diff\_host* and *dst\_host\_rerror\_rate*.

The training set and testing sets were reduced to contain only the attributes chosen by the FSAs. The reduced datasets were then used to test the accuracy of J48.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Performance using training set | Performance using testing set 1 (unseen *dos*) | Performance using testing set 2 (unseen *normal*) |
| FULL | 100% | 100% | 89% |
| FSA1 | 99% | 82% | 97% |
| FAS2 | 100% | 100% | 89% |
| FAS3 | 100% | 100% | 86% |

The initial results indicate that reducing the number of attributes, and hence resulting in a faster decision and less of storage, has no detrimental effect on the performance of the classifier.

**FUTURE WORK**

To be able to present a comprehensive review and comparison of different feature selection and reduction algorithms, more tests will have to be performed. Other tests will involve different learning algorithms and different feature selection and reduction algorithms consisting of various attribute evaluators and search methods.

REFERENCES

Amudha, P., & Abdul Rauf, H. (2011). Performance Analysis of Data Mining Approaches in Intrusion Detection. *2011 International Conference On Process Automation, Control And Computing*. doi:10.1109/pacc.2011.5978878

Ektefa, M., Memar, S., Sidi, F., & Affendey, L. (2010). Intrusion detection using data mining techniques. *2010 International Conference On Information Retrieval & Knowledge Management (CAMP)*. doi:10.1109/infrkm.2010.5466919.

Jensen, R., & Shen, Q. (2008). *Computational intelligence and feature selection*. Piscataway, NJ : IEEE Press: John Wiley & Sons.

Kdd.ics.uci.edu,. (2015). *KDD Cup 1999 Data*. Retrieved 7 April 2015, from <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

Nguyen, H., & Choi, D. (2008). Application of Data Mining to Network Intrusion Detection: Classifier Selection Model. *Challenges For Next Generation Network Operations And Service Management*, 399-408. doi:10.1007/978-3-540-88623-5\_41

Panda, M., & Patra, M. (2008). A Comparative Study of Data Mining Algorithms for Network Intrusion Detection. *2008 First International Conference On Emerging Trends In Engineering And Technology*. doi:10.1109/icetet.2008.80

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. (2009). A detailed analysis of the KDD CUP 99 data set. *2009 IEEE Symposium On Computational Intelligence For Security And Defense Applications*. doi:10.1109/cisda.2009.5356528