International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 07
June 2017

A Study of Feature Selection Methods in Intrusion Detection System: A Survey

[1]A.MOUNIKA [2]DR. K NAGESWARARAO

[1]Pg Scholar, Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally, anumolumounika789@gmail.com

[2] Professor & HOD, Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally, nageswararaokapu@yahoo.com

ABSTRACT— Redundant and irrelevant features in data have caused a long-term problem in network traffic classification. These features not only slow down the process of classification but also prevent a classifier from making accurate decisions, especially when coping with big data. In this paper, we propose a mutual information based algorithm that analytically selects the optimal feature for classification. This mutual information based feature selection algorithm can handle linearly and nonlinearly dependent data features. Its effectiveness is evaluated in the cases of network intrusion detection. An Intrusion Detection System (IDS), named Least Square Support Vector Machine based IDS (LSSVM-IDS), is built using the features selected by our proposed feature selection algorithm. The performance of LSSVM-IDS is evaluated using three intrusion detection evaluation datasets, namely KDD Cup 99, NSL-KDD and Kyoto 2006+ dataset. The evaluation results show that our feature selection algorithm contributes more critical features for LSSVM-IDS to achieve better accuracy and lower computational cost compared with the state of-the-art methods.

1INTRODUCTION

DESPITE increasing awareness of network security, the existing solutions remain incapable of fully protecting internet applications and computer networks against the threats from ever-advancing cyber attack techniques such as DoS attack and computer malware. Developing effective and adaptive security approaches, therefore, has become more critical than ever before. The traditional security techniques, as the first line of security defence, such as user authentication, firewall and data encryption, are insufficient to fully cover the entire landscape of network security while facing challenges from ever-evolving intrusion skills and techniques. Hence, another line of security defence is highly recommended, such as Intrusion Detection System (IDS). Recently, an IDS alongside with anti-virus

software has become an important complement to the security infrastructure of most organizations. The combination of these two lines provides a more comprehensive defence against those threats and enhances network security. A significant amount of research has been conducted to develop intelligent intrusion detection techniques, which help achieve better network security. Bagged boosting-based on C5 decision trees and Kernel Miner are two of the earliest attempts to build intrusion detection schemes. Methods proposed in and have successfully applied machine learning techniques, such as Support Vector Ma. M. A. Ambusaidi, X. He and P. Nanda are with the School of Computing and Communications, Faculty of Engineering and IT, University of Technology, Sydney, chine (SVM), to classify network traffic patterns that do not match normal network traffic. Both systems were equipped with five distinct classifiers to detect normal traffic and four different types of attacks (i.e., DoS, probing, U2R and R2L). Experimental results show the effectiveness and robustness of using SVM in IDS. Mukkamala et al. investigated the possibility of assembling various learning methods, including Artificial Neural Networks (ANN), SVMs and Multivariate Adaptive Regression Splines (MARS) to detect intrusions. They trained five different classifiers to distinguish the normal traffic from the four different types of attacks. They compared the performance of each of the learning methods with their model and found that the ensemble of ANNs, SVMs and MARS achieved the best performance in terms of classification accuracies for all the five classes. Toosi et al. combined a set of neuro-fuzzy classifiers in their design of a detection system, in which a genetic algorithm was applied to optimize the structures of neuro-fuzzy systems used in the classifiers. Based on the pre-determined fuzzy inference system (i.e., classifiers), detection decision was made on the incoming traffic. Recently, we proposed an anomaly-based scheme for detecting DoS attacks. The system has been evaluated on KDD Cup 99 and ISCX 2012 datasets and achieved promising detection accuracy of 99.95% and 90.12% respectively.

2 RELATED WORKS

2.1 Feature Selection

Feature selection is a technique for eliminating irrelevant and redundant features and selecting the most optimal subset of features that produce a better

characterization of patterns belonging to different classes. Methods for feature selection are generally classified into filter and wrapper methods . Filter algorithms utilize an independent measure (such as, information measures, distance measures, or consistency measures) as a criterion for estimating the relation of a set of features, while wrapper algorithms make use of particular learning algorithms to evaluate the value of features. In comparison with filter methods, wrapper methods are often much more computationally expensive when dealing with high-dimensional data or large-scale data. In this study hence, we focus on filter methods for IDS. Due to the continuous growth of data dimensionality, feature selection as a pre-processing step is becoming an essential part in building intrusion detection systems . Mukkamala and Sung [14] proposed a novel feature selection algorithm to reduce the feature space of KDD Cup 99 dataset from 41 dimensions to 6 dimensions and evaluated the 6 selected features using an IDS based on SVM. The results show that the classification accuracy increases by 1% when using the selected features. Chebrolu et al. investigated the performance in the use of a Markov blanket model and decision tree analysis for feature selection, which showed its capability of reducing the number of features in KDD Cup 99 from 41 to 12 features. Chen et al proposed an IDS based on Flexible Neural Tree (FNT). The model applied a pre-processing feature selection phase to improve the detection performance. Using the KDD Cup 99, FNT model achieved 99.19% detection accuracy with only 4 features. Recently, Amiri [12] proposed a forward feature selection algorithm using the mutual information method to measure the relation among features. The optimal feature set was then used to train the LS-SVM classifier and build the IDS. Horng et al. proposed an SVM-based IDS, which combines a hierarchical clustering and the SVM. The hierarchical clustering algorithm was used to provide the classifier with fewer and higher quality training data to reduce the average training and testing time and improve the classification performance of the classifier. Experiment on the corrected labels KDD Cup 99 dataset, which includes some new attacks, the SVM-based IDS scored an overall accuracy of 95.75% with a false positive rate of 0.7%.

2.2 Performance Evaluation

All of the aforementioned detection techniques were evaluated on the KDD Cup 99 dataset. However, due to some limitations in this dataset, which will be discussed in Subsection some other detection methods were evaluated using other intrusion detection datasets, such as NSL-KDD and Kyoto 2006.A dimensionality reduction method proposed in was to find the most.This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI important features involved in building a naive Bayesian classifier for intrusion detection. Experiments conducted on the NSL-KDD dataset produced encouraging results. Chitrakar et al. proposed a Candidate Support Vector based Incremental SVM algorithm (CSV-ISVM in short). The algorithm was applied to network intrusion detection. They evaluated their CSV-ISVM-based IDS on the Kyoto 2006+ [25] dataset. Experimental results showed that their IDS produced promising results in terms of detection rate and false alarm rate. The IDS was claimed to perform realtime network intrusion detection. Therefore, in this work, to make a fair comparison with those detection systems, we evaluate our proposed model on the aforementioned datasets.

3 INTRUSION DETECTION FRAMEWORK BASED ON LEAST SQUARE SUPPORT VECTOR MACHINE

The framework of the proposed intrusion detection system is depicted in Figure 1. The detection framework is comprised of four main phases: (1) data collection, where sequences of network packets are collected, (2) data preprocessing, where training and test data are preprocessed and

important features that can distinguish one class from the others are selected, (3) classifier training, where the model for classification is trained using LS-SVM, and (4) attack recognition, where the trained classifier is used to detect intrusions on the test data. Support Vector Machine (SVM) is a supervised learning method . It studies a given labeled dataset and constructs an optimal hyperplane in the corresponding data space to separate the data into different classes. Instead of solving the classification problem by quadratic programming, Suykens and Vandewalle suggested re-framing the task of classification into a linear programming problem. They named this new formulation the Least Squares

SVM (LS-SVM). LS-SVM is a generalized scheme for classification and also incurs low computation complexity in comparison with the ordinary SVM scheme . One can find more details about calculating LS-SVM in Appendix B. The following subsections explain each phase in detail.

3.1 Data Collection

Data collection is the first and a critical step to intrusion detection. The type of data source and the location where data is collected from are two determinate factors in the design and the effectiveness of an IDS. To provide the best suited protection for the targeted host or networks, this study proposes a network-based IDS to test our proposed approaches. The proposed IDS runs on the nearest router to the victim(s) and monitors the inbound network traffic. During the training stage, the collected data samples are categorised with respect to the transport/Internet layer protocols and are labeled against the domain knowledge. However, the data collected in the test stage are categorized according to the protocol types only.

3.2 Data Preprocessing

The data obtained during the phase of data collection are first processed to generate the basic features such as the ones in KDD Cup 99 dataset . This phase contains three main stages shown as follows.

Data transferring

4.2.2 Data normalisation

An essential step of data preprocessing after transferring all

symbolic attributes into numerical values is normalisation.

Data normalisation

Feature selection

3.3 Classifier Training

Once the optimal subset of features is selected, this subset is then taken into the classifier training phase where LS-SVM is employed. Since SVMs can only handle binary classification problems and because for KDD Cup 99 five optimal feature subsets are selected for all classes, five LS-SVM classifiers need to be employed. Each classifier distinguishes one class of records from the others. For example the classifier of Normal class distinguishes Normal data from non-Normal (All types of attacks). The DoS class distinguishes DoS traffic from non-DoS data (including Normal, Probe, R2L and U2R instances) and so on. The five LS-SVM classifiers are then combined to build the intrusion detection model to distinguish all different classes.

4.4 Attack Recognition

In general, it is simpler to build a classifier to distinguish between two classes than considering multiclasses in a problem. This is because the decision boundaries in the first case can be simpler. The first part of the experiments in this paper uses two classes, where records matching to the normal class are reported as normal data, otherwise are considered as attacks. However, to deal with a problem having more than two classes, there are two popular techniques: \One-Vs-One" (OVO) and \One-Vs-All" (OVA). Given a classification problem with M classes (M > 2), the OVO approach on the one hand divides an M-class problem into M_(M□1) 2 binary problems. Each problem is handled by a separate binary

Algorithm Intrusion detection based on LS-SVM Distinguishing intrusive network traffic from normal network traffic in the case of multiclassg

Input: LS-SVM Normal Classifier, selected features (normal class), an observed data item x

Output: Lx - the classification label of x

begin

Lx classification of x with LS-SVM of Normal class

if Lx == \Normal" **then**

Return LX

else

do: Run Algorithm 4 to determine the class of attack

end

end

classifier, which is responsible for separating the data of a pair of classes. The OVA approach, on the other hand, divides an Mclass problem into M binary problems. Each problem is handled by a binary classifier, which is responsible for separating the data of a single class from all other classes. Obviously, the OVO approach requires more binary classifiers than OVA. Therefore, it is more computationally intensive. Rifkin and Klautau demonstrated that the OVA technique was preferred over OVO. As such, the OVA technique is applied to the proposed IDS to distinguish between normal and abnormal data using the LS-SVM method. After completing all the aforementioned steps and the classifier is trained using the optimal subset of features which includes the most correlated and important features, the normal and intrusion traffics can be identified by using the saved trained classifier. The test data is then directed to the saved trained model to detect intrusions. Records matching to the normal class are considered as normal data, and the

other records are reported as attacks. If the classifier model confirms that the record is abnormal, the subclass of the abnormal record (type of attacks) can be used to determine the record's type. describe the detection processes.

Algorithm Attack classification based on LS-SVM

Input: LS-SVM Normal Classifier, selected features (normal
class), an observed data item x

Output: Lx - the classification label of x

begin

Lx classification of x with LS-SVM of DoS class

if Lx==\DoS" **then**

Return LX

else

Lx classification of x with LS-SVM of Probe class

if Lx == \Probe" **then**

Return LX

else

Lx classification of x with LS-SVM of R2L class

if Lx == \R2L" **then**

Return LX

else

Lx == \U2R";

Return LX

end

end

end

end

6 CONCLUSION

Recent studies have shown that two main components are essential to build an IDS. They are a robust classification method and an efficient feature selection algorithm. In this paper, a supervised filter-based feature selection algorithm has been proposed, namely Flexible Mutual Information Feature Selection (FMIFS). FMIFS is an improvement over MIFS and MMIFS. FMIFS suggests a modification to Battiti's algorithm to reduce the redundancy among features. FMIFS eliminates the redundancy parameter _ required in MIFS and MMIFS. This is desirable in practice since there is no specific procedure or guideline to select the best value for this parameter.

FMIFS is then combined with the LSSVM method to build an IDS. LSSVM is a least square version of SVM that works with equality constraints instead of inequality constraints in the formulation designed to solve a set of linear equations for classification problems rather than a quadratic programming problem. The proposed LSSVMIDS + FMIFS has been

evaluated using three well known intrusion detection datasets: KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets. The performance of LSSVM-IDS + FMIFS on KDD Cup test data, KDDTest+ and the data, collected on 1, 2 and 3 November 2007, from Kyoto dataset has exhibited better classification performance in terms of classification accuracy, detection rate, false positive rate and F-measure than some of the existing detection approaches. In addition, the proposed LSSVM-IDS + FMIFS has shown comparable results with other state-of the-art approaches

when using the Corrected Labels sub-dateset of the KDD Cup 99 dataset and tested on Normal, DoS, and Probe classes; it outperforms other detection models when tested on U2R and R2L classes. Furthermore, for the experiments on the KDDTest□21 dataset, LSSVM-IDS + FMIFS produces the best classification accuracy compared with other detection systems tested on the same dataset. Finally, based on the experimental results achieved on all datasets, it can be concluded that the proposed detection system has achieved promising performance in detecting intrusions over computer networks. Overall, LSSVM-IDS + FMIFS has performed the best when compared with the other state-of the art models. Although the proposed feature selection algorithm FMIFS has shown encouraging performance, it could be further enhanced by optimizing the search strategy. In addition, the impact of the unbalanced sample distribution on an IDS needs to be given a careful consideration in our future studies.

REFERENCES

[1] S. Pontarelli, G. Bianchi, S. Teofili, Traffic-aware design of a highspeed fpga network intrusion detection system, Computers, IEEE Transactions on 62 (11) (2013) 2322–2334. 0018-9340 (c) 2015 IEEE. Personal use is permitted

[2] B. Pfahringer,Winning the kdd99 classification cup: Bagged boosting, SIGKDD Explorations 1 (2) (2000) 65–66.

[3] I. Levin, Kdd-99 classifier learning contest: Llsoft's results overview, SIGKDD explorations 1 (2) (2000) 67–75.

[4] D. S. Kim, J. S. Park, Network-based intrusion detection with support vector machines, in: Information Networking, Vol. 2662, Springer, 2003, pp. 747–756.

[5] A. Chandrasekhar, K. Raghuveer, An effective technique for intrusion detection using neuro-fuzzy and radial svm classifier, in: Computer Networks & Communications (NetCom), Vol. 131, Springer, 2013, pp. 499–507.

[6] S. Mukkamala, A. H. Sung, A. Abraham, Intrusion detection using an ensemble of intelligent paradigms, Journal of network and computer applications 28 (2) (2005) 167–182.

[7] A. N. Toosi, M. Kahani, A new approach to intrusion detection based on an evolutionary soft computing model using

neurofuzzy classifiers, Computer communications 30 (10) (2007) 2201–2212.
[8] Z. Tan, A. Jamdagni, X. He, P. Nanda, L. R. Ping Ren, J. Hu, Detection of denial-of-service attacks based on computer vision techniques, IEEE Transactions on Computers 64 (9) (2015) 2519– 2533.
[9] A. M. Ambusaidi, X. He, P. Nanda, Unsupervised feature selection method for intrusion detection system, in: International Conference on Trust, Security and Privacy in Computing and Communications, IEEE, 2015.
[10] A. M. Ambusaidi, X. He, Z. Tan, P. Nanda, L. F. Lu, T. U. Nagar, A novel feature selection approach for intrusion detection data classification, in: International Conference on Trust, Security and Privacy in Computing and Communications, IEEE, 2014, pp. 82– 89.
[11] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks 5 (4) (1994) 537–550.
[12] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, N. Yazdani, Mutual information-based feature selection for intrusion detection systems, Journal of Network and Computer Applications 34 (4) (2011) 1184–1199.
[13] A. Abraham, R. Jain, J. Thomas, S. Y. Han, D-scids: Distributed soft computing intrusion detection system, Journal of Network and Computer Applications 30 (1) (2007) 81–98.
[14] S. Mukkamala, A. H. Sung, Significant feature selection using computational intelligent techniques for intrusion detection, in: Advanced Methods for Knowledge Discovery from Complex Data, Springer, 2005, pp. 285–306.
[15] S. Chebrolu, A. Abraham, J. P. Thomas, Feature deduction and ensemble design of intrusion detection systems, Computers & Security 24 (4) (2005) 295–307.
[16] Y. Chen, A. Abraham, B. Yang, Feature selection and classification flexible neural tree, Neurocomputing 70 (1) (2006) 305–313.