ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm

Pushp Raj[1], Sana Ansari[1], Narendra Varude[1], Ashwini Sagade[2]

B. E Student, Department of Computer Engineering, Dr D Y Patil School of Engineering Computer Science, Pune, India[1]

Professor, Department of Computer Engineering, Dr D Y Patil School of Engineering Computer Science, Pune, India[2]

ABSTRACT: With the development of Internet, there has been a huge increment in the quantity of attacks and in this manner Intrusion detection Systems (IDS's).unessential segments in data have achieved a whole deal issue in framework action gathering. These segments back off the method of course of action and in addition shield a classifier from settling on exact decisions, especially when adjusting to huge data. In this paper, we propose common information based figuring that methodically picks the perfect part for game plan. This mutual information based segment assurance computation can manage straightforwardly and nonlinearly subordinate data features. Its ampleness is surveyed in the cases of framework intrusion disclosure. An Intrusion Detection System (IDS), is manufactured using the components picked by our proposed incorporate assurance estimation. The execution of Intrusion Detection is surveyed using two intrusion recognizable proof evaluation datasets, to be particular KDD Cup 99 and NSL KDD.The evaluation comes to fruition show that our component decision figuring contributes more fundamental components for Intrusion Detection to fulfill better accuracy and lower computational cost differentiated and the best in class methods.

KEYWORDS: Intrusion detection, Feature selection, Mutual information, linear correlation coefficient, least square support vector machine, naive Bayes classifier.

I. INTRODUCTION

The growing awareness of network security, existing solutions are not yet able to fully protect Internet applications and computer networks from the threats of cyber-attack techniques in constant progress such as DoS attack and 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 defense, 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 [1]. Hence, another line of security defense is highly recommended, such as Outlier Detection System (IDS). Recently, an ODS 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 defense against those threats and enhances network security. A significant amount of research has been conducted to develop intelligent outlier detection techniques, which help achieve better network security. Bagged boosting-based on C5 decision trees [2] and Kernel Miner [3] are two of the earliest attempts to build Outlier detection schemes. Methods proposed in [4] and [5] have successfully applied machine learning techniques, such as Support Vector Machine (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 ODS. Mukkamala et al. [6] 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

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

MARS achieved the best performance in terms of classification accuracies for all the five classes. Toosi et al. [7] 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 [8]. 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. However, current network traffic data, which are often huge in size, present a major challenge to ODSs [9]. These "big data" slow down the entire detection process and may lead to unsatisfactory classification accuracy due to the computational difficulties in handling such data. Classifying a huge amount of data usually causes many mathematical difficulties which then lead to higher computational complexity. As a well-known Outlier evaluation dataset, KDD Cup 99 dataset is a typical example of large-scale datasets.
This dataset consists of more than five million of training samples and two million of testing samples respectively. Such a large scale dataset retards the building and testing processes of a classifier, or makes the classifier unable to perform due to system failures caused by insufficient memory. Furthermore, large-scale datasets usually contain noisy, redundant, or uninformative features which present critical challenges to knowledge discovery and data modeling.

II. LITERATURE SURVEY

1] F. Amiri, M. RezaeiYousefi, 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.

Refer Points:
This Paper 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.
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.

2] 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.

Refer Points:

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.

3] 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.

Refer Points:

This Paper 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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

4] S. Chebrolu, A. Abraham, J. P. Thomas, Feature deduction and ensemble design of intrusion detection systems, Computers & Security 24 (4) (2005) 295–307.

Refer Points:

This Paper 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.

5] Y. Chen, A. Abraham, B. Yang, Feature selection and classification flexible neural tree, Neurocomputing 70 (1) (2006) 305–313. [7] S.-J. Horng, M.-Y.Su, Y.-H.Chen, T.-W.Kao, R.-J.Chen, J.-L. Lai, C. D. Perkasa, A novel intrusion detection system based on hierarchical clustering and support vector machines, Expert systems with Applications 38 (1) (2011) 306–313.

Refer Points:

This Paper 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.

6] S.-J. Horng, M.-Y.Su, Y.-H.Chen, T.-W.Kao, R.-J.Chen, J.-L. Lai, C. D. Perkasa, A novel intrusion detection system based on hierarchical clustering and support vector machines, Expert systems with Applications 38 (1) (2011) 306–313.

Refer Points:

This Paper 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. Experimented 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%. All of the aforementioned detection techniques were evaluated on the KDD Cup 99 dataset.

7] R. Chitrakar, C. Huang, Selection of candidate support vectors in incremental svm for network intrusion detection, Computers & Security 45 (2014) 231–241.

Refer Points:

This Paper 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+ [11] dataset. Experimental results showed that their IDS produced promising results in terms of detection rate and false alarm rate.

8] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, K. Nakao, Statistical analysis of honeypot data and building of kyoto 2006+ dataset for nids evaluation, in: Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, ACM, 2011, pp. 29–36.

Refer Points:

A dimensionality reduction method proposed in [8] was to find the most important features involved in building a naive Bayesian classifier for intrusion detection. Experiments conducted on the NSL-KDD dataset produced encouraging results.

III. PROPOSED SYSTEM / SYSTEM OVERVIEW

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,

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

(3) Classifier training, where the model for classification is trained and
(4) Attack recognition, where the trained classifier is used to detect intrusions on the test data.

Proposed System Algorithm:

1. Start
2. First we collect the training labeled data from training dataset.
3. Second we process the collected data, in pre-processing for feature extraction we use flexible mutual information based feature selection or flexible linear correlation coefficient based feature selection algorithm.
4. After feature selection we classify the data.
5. After classification recognize the attack using Attack classification algorithms.
6. Finally results demonstrate that the intrusion detected or not.
7. Stop

IV. SYSTEM ARCHITECTURE

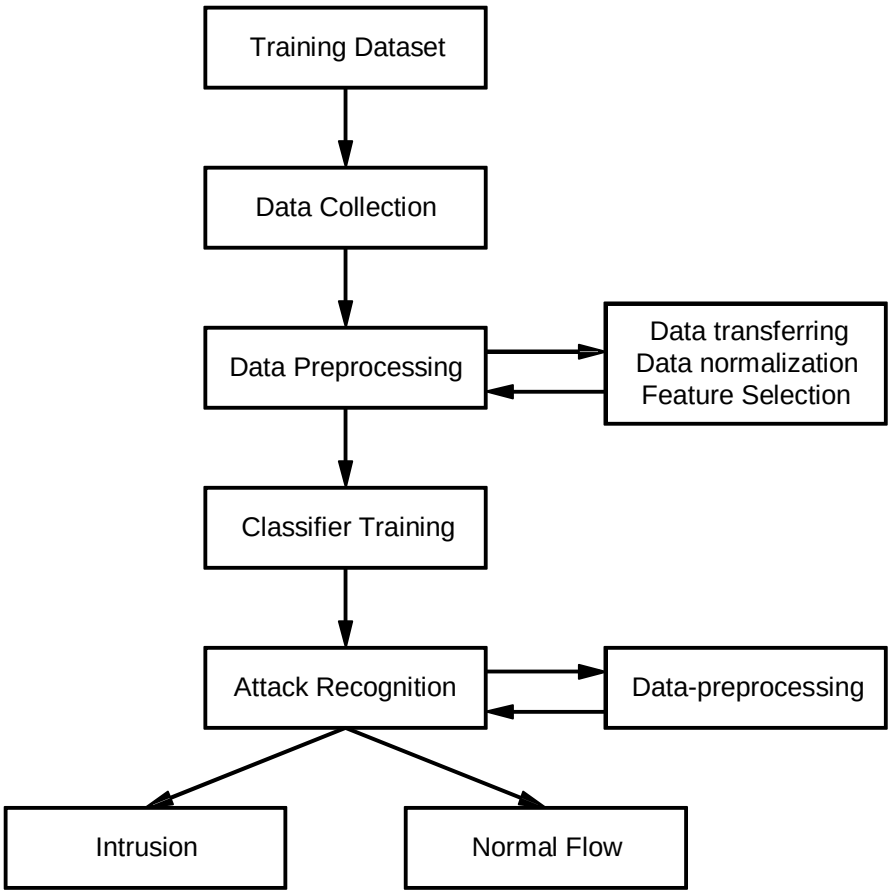

Fig.1 Proposed System Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

V. 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. This is desirable in practice since there is no specific procedure or guideline to select the best value for this parameter. The proposed Intrusion Detection has been evaluated using twowell known intrusion detection datasets: KDD Cup 99 and NSL KDD. In addition, the proposed IDShas shown comparable results with other state-of the-art approaches when using the Corrected Labels sub-dataset of the KDD Cup 99 dataset. Furthermore, for the experiments on the KDDTest, IDS 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, IDS has performed the best when compared with the other state-of-the- art models.

REFERENCES

[1] F. Amiri, M. RezaeiYousefi, 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.
[2] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks 5 (4) (1994) 537–550.
[3] S. Chebrolu, A. Abraham, J. P. Thomas, Feature deduction and ensemble design of intrusion detection systems, Computers & Security 24 (4) (2005) 295–307.
[4] 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.
[5] 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.
[6] Y. Chen, A. Abraham, B. Yang, Feature selection and classification flexible neural tree, Neurocomputing 70 (1) (2006) 305–313
[7] S.-J. Horng, M.-Y.Su, Y.-H.Chen, T.-W.Kao, R.-J.Chen, J.-L. Lai, C. D. Perkasa, A novel intrusion detection system based on hierarchical clustering and support vector machines, Expert systems with Applications 38 (1) (2011) 306–313.
[8] G. Kim, S. Lee, S. Kim, A novel hybrid intrusion detection method integrating anomaly detection with misuse detection, Expert Systems with Applications 41 (4) (2014) 1690–1700.