International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 14
November 2017

Intrusion Detection System Using a Filter Based Feature Selection Algorithm

[1]G.Ramya, [2]A.Swetha,[3]Prof.V.Janaki , [4]Prof.P. Prakash

1.Pg Scholar,Department of CNIS, Vaagdevi College of Engineering

2.Assistant professor, Deparment of CSE, Vaagdevi College of Engineering

3.Professor,HOD, Deparment of CSE, Vaagdevi College of Engineering

4.Professor,Principal, Vaagdevi College of Engineering

Abstract— *Repetitive and unimportant highlights in information have caused a long haul issue in arrange activity grouping. These highlights back off the procedure of arrangement as well as keep a classifier from settling on precise choices, particularly when adapting to enormous information. In this paper, we propose a shared data based calculation that diagnostically chooses the ideal component forclassification. This common data based element choice calculation can deal with directly and nonlinearly subordinate information highlights. Its adequacy is assessed in the instances of system interruption discovery. An Intrusion Detection System (IDS), named Least Square Support Vector Machine based IDS (LSSVM-IDS), is fabricated utilizing the highlights chose by our proposed include choice calculation. The execution of LSSVM-IDS is assessed utilizing three interruption recognition assessment datasets, specifically KDD Cup 99, NSL-KDD and Kyoto 2006+ dataset. The assessment comes about demonstrate that our element determination calculation contributes more basic highlights for LSSVM-IDS to accomplish better exactness and lower computational cost contrasted and the best in class techniques.*

1INTRODUCTION

Regardless of expanding familiarity with organize security, the current arrangements stay unequipped for completely ensuring web applications and PC systems against the dangers from consistently progressing digital assault procedures, for example, DoS assault and PC malware. Creating successful and versatile security approaches, along these lines, has turned out to be more basic than any time in recent memory. The conventional security strategies, as the principal line of security resistance, for example, client verification, firewall and information encryption, are deficient to completely cover the whole scene of system security while confronting challenges from consistently advancing interruption aptitudes and procedures. Thus, a different line of security guard is profoundly suggested, for example, Intrusion Detection System (IDS). As of late, an IDS close by with hostile to infection programming has turned into a critical supplement to the security foundation of generally associations. The mix of these two lines gives a more far reaching safeguard against those dangers and upgrades organize security. A lot of research has been directed to create keen interruption location strategies, which help accomplish better system security. Stowed boosting-in light of C5 choice trees and Kernel Miner are two of the most punctual endeavors to assemble interruption location plans. Strategies proposed in and have effectively connected machine learning procedures, for example, 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 group arrange activity designs that don't coordinate typical system movement. The two frameworks were furnished with five unmistakable classifiers to distinguish typical movement and four unique sorts of assaults (i.e., DoS, examining, U2R and R2L). Exploratory outcomes demonstrate the adequacy and strength of utilizing SVM in IDS. Mukkamala et al. examined the likelihood of amassing different learning techniques, including Artificial Neural Networks (ANN), SVMs and Multivariate Adaptive Regression Splines (MARS) to identify interruptions. They prepared five unique classifiers to recognize the ordinary activity from the four distinct sorts of assaults. They looked at the execution of each of the learning strategies with their model and found that the troupe of ANNs, SVMs and MARS accomplished the best execution as far as grouping exactnesses for all the five classes. Toosi et al. consolidated an arrangement of neuro-fluffy classifiers in their plan of a location framework, in which a hereditary calculation was connected to improve the structures of neuro-fluffy frameworks utilized as a part of the classifiers. In view of the pre-decided fluffy surmising framework (i.e., classifiers), recognition choice was made on the approaching activity. As of late, we proposed an oddity based plan for distinguishing DoS assaults. The framework has been assessed on KDD Cup 99 and ISCX 2012 datasets and accomplished promising recognition exactness of 99.95% and 90.12% individually.

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 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 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
```

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

Late investigations have demonstrated that two principle parts are basic to assemble an IDS. They are a powerful order strategy and an effective component determination calculation. In this paper, an administered channel based component choice calculation has been proposed, in particular Flexible Mutual Information Feature Selection (FMIFS). FMIFS is a change over MIFS and MMIFS. FMIFS proposes an adjustment to Battiti's calculation to diminish the repetition among highlights. FMIFS takes out the repetition parameter _ required in MIFS and MMIFS. This is alluring practically speaking since there is no particular system or rule to choose the best an incentive for this parameter.

FMIFS is then joined with the LSSVM strategy to construct an IDS. LSSVM is a slightest square form of SVM that works with fairness requirements rather than disparity imperatives in the plan intended to understand an arrangement of direct conditions for grouping issues as opposed to a quadratic programming issue. The proposed LSSVMIDS + FMIFS has been assessed utilizing three understood interruption identification datasets: KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets. The execution of LSSVM-IDS + FMIFS on KDD Cup test information, KDDTest+ and the information, gathered on 1, 2 and 3 November 2007, from Kyoto dataset has shown better arrangement execution as far as order exactness, recognition rate, false positive rate and F-measure than a portion of the current location approaches. Also, the proposed

LSSVM-IDS + FMIFS has demonstrated similar outcomes with other best in class approaches

when utilizing the Corrected Labels sub-dateset of the KDD Cup 99 dataset and tried on Normal, DoS, and Probe classes; it beats other recognition models when tried on U2R and R2L classes. Moreover, for the examinations on the KDDTest□21 dataset, LSSVM-IDS + FMIFS produces the best characterization exactness contrasted and other location frameworks tried on the same dataset. At long last, in view of the test comes about accomplished on all datasets, it can be presumed that the proposed discovery framework has accomplished promising execution in identifying interruptions over PC systems. By and large, LSSVM-IDS + FMIFS has played out the best when contrasted and the other best in class models. In spite of the fact that the proposed include choice calculation FMIFS has indicated empowering execution, it could be additionally improved by advancing the hunt system. Likewise, the effect of the lopsided example conveyance on an IDS should be given a cautious thought in our future investigations.

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.

AUTHOR'S DETAILS:

G.Ramya, Pg Scholar,Department of CNIS, Vaagdevi College of Engineering.

A.Swetha, Assistant professor, Deparment of CSE, Vaagdevi College of Engineering