RESEARCH ARTICLE OPEN ACCESS

An Efficient Method for Intrusion Detection System Using a Filter Based Feature Selection Algorithm

K.Sandhya Rani[1], Tunga Urmila[2]

*[1] Assistant Professor, Department of MCA, Vignan's Lara Institute of Technology & Science, Vadlamudi, Guntur, Andhra Pradesh, India

[2]MCA Student, Department of MCA, Vignan's Lara Institute of Technology & Science, Vadlamudi, Guntur, Andhra Pradesh, India

Abstract:

Redundant and insignificant features in data have caused a whole deal issue in mastermind action gathering. These features back off the technique of course of action and also shield a classifier from settling on exact decisions, especially when adjusting to gigantic data. In this paper, we propose a mutual information based estimation that symptomatically picks the perfect segment for order. This regular information based component decision count can manage specifically and nonlinearly subordinate data features. Its ampleness is evaluated in the examples of system intrusion revelation. An Intrusion Detection System (IDS), named Least Square Support Vector Machine based IDS (LSSVM-IDS), is created using the features picked by our proposed incorporate decision count. The execution of LSSVM-IDS is surveyed using three interference acknowledgment appraisal datasets, particularly KDD Cup 99, NSL-KDD and Kyoto 2006+dataset. The evaluation comes to fruition exhibit that our component assurance estimation contributes more essential features for LSSVM-IDS to achieve better precision 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.**

I. INTRODUCTION

Despite extending commonality with sort out security, the present plans remain unequipped for totally guaranteeing web applications and PC systems against the risks from reliably advancing computerized ambush methods, for instance, DoS strike and PC malware. Making fruitful and adaptable security approaches, thusly, has ended up being more essential than whenever in late memory. The traditional security techniques, as the vital line of security protection, for instance, customer confirmation, firewall and data encryption, are insufficient to totally cover the entire scene of system security while standing up to challenges from reliably propelling intrusion aptitudes and strategies. Along these lines, an alternate line of security watch is significantly recommended, for instance, Intrusion Detection System (IDS). Starting late, an IDS close by with unfriendly to contamination programming has transformed into a basic supplement to the security establishment of for the most part affiliations. The blend of these two lines gives a more extensive shield against those risks and updates sort out security. A great deal of research has been coordinated to make sharp interference area techniques, which help achieve better system security. Stowed boosting-in light of C5 decision trees and Kernel Miner are two of the most reliable undertakings to gather intrusion area designs. Systems proposed in and have successfully associated machine learning techniques, for instance, 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 gather mastermind action plans that don't facilitate run of the mill system development. The two systems were outfitted with five unmistakable classifiers to recognize run of the mill development and four interesting sorts of ambushes (i.e., DoS, looking at, U2R and R2L). Exploratory results exhibit the sufficiency and quality of using SVM in IDS. Mukkamala et al. inspected the probability of gathering distinctive learning methods, including Artificial Neural Networks (ANN), SVMs and Multivariate Adaptive Regression Splines (MARS) to distinguish interferences. They arranged five one of a kind classifiers to perceive the common movement from the four particular sorts of strikes. They took a gander at the execution of every one of the learning procedures with their model and found that the troupe of ANNs,

SVMs and MARS achieved the best execution to the extent gathering exactnesses for all the five classes. Toosi et al. merged a course of action of neuro-feathery classifiers in their arrangement of an area system, in which a genetic computation was associated with enhance the structures of neuro-soft systems used as a piece of the classifiers. In perspective of the pre-chosen cushioned inferring structure (i.e., classifiers), acknowledgment decision was made on the moving toward movement. Starting late, we proposed a peculiarity based arrangement for recognizing DoS attacks. The system has been evaluated on KDD Cup 99 and ISCX 2012 datasets and achieved promising acknowledgment precision of 99.95% and 90.12% exclusively.

II. RELATED WORKS

2.1 Feature Selection

Highlight determination is a procedure for wiping out insignificant and repetitive highlights and choosing the most ideal subset of highlights that deliver a superior portrayal of examples having a place with various classes. Techniques for include choice are for the most part characterized into channel and wrapper strategies . Channel calculations use a free measure, (for example, data measures, remove measures, or consistency measures) as a standard for assessing the connection of an arrangement of highlights, while wrapper calculations make utilization of specific learning calculations to assess the estimation of highlights. In examination with channel techniques, wrapper strategies are frequently considerably more computationally costly when managing high-dimensional information or extensive scale information. In this investigation consequently, we center around channel techniques for IDS. Because of the ceaseless development of information dimensionality, include determination as a pre-handling step is turning into a fundamental part in building intrusion detection systems . Mukkamala and Sung [14] proposed a novel component determination calculation to decrease the element space of KDD Cup 99 dataset from 41 measurements to 6 measurements and assessed the 6 chose highlights utilizing an IDS based on SVM. The outcomes demonstrate that the arrangement precision increments by 1% when utilizing the chose highlights. Chebrolu et al. examined the execution in the utilization of a Markov cover model and choice tree investigation for include choice, which demonstrated its capacity of decreasing the quantity of highlights in KDD Cup 99 from 41 to 12 highlights. Chen et al proposed an IDS based on Flexible Neural Tree (FNT). The model connected a pre-handling highlight choice stage to enhance the detection execution. Utilizing the KDD Cup 99, FNT show accomplished 99.19% detection exactness with just 4 highlights. As of late, Amiri [12] proposed a forward element determination calculation utilizing the common data strategy to quantify the connection among highlights. The ideal list of capabilities was then used to prepare the LS-SVM classifier and fabricate the IDS. Horng et al. proposed a SVM-based IDS, which consolidates a various leveled grouping and the SVM. The various leveled bunching calculation was utilized to furnish the classifier with less and higher quality preparing information to decrease the normal preparing and testing time and enhance the characterization execution of the classifier. Trial on the amended names KDD Cup 99 dataset, which incorporates some new assaults, the SVM-based IDS scored a general exactness 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.

III. LITERATURE SURVEY

3.1 INTEROPERABILITY OF PERSONAL HEALTH RECORDS

AUTHORS: J. L ahteenm€ aki, J. Lepp anen, and H. Kaijanranta,
The establishment of the Meaningful Use criteria has created a critical need for robust interoperability of health records. A universal definition of a personal health record (PHR) has not been agreed upon. Standardized code sets have been built for specific entities, but integration between them has not been supported. The purpose of this research study was to explore the hindrance and promotion of interoperability standards in relationship to PHRs to describe interoperability progress in this area. The study was conducted following the basic principles of a systematic review, with 61 articles used in the study. Lagging interoperability has stemmed from slow adoption by patients, creation of disparate systems due to rapid development to meet requirements for the Meaningful Use stages, and rapid early development of PHRs prior to the mandate for integration among multiple systems. Findings of this study suggest that deadlines for implementation to capture Meaningful Use incentive payments are supporting the creation of PHR data silos, thereby hindering the goal of high-level interoperability.

3.2 APPLYING CLOUD COMPUTING MODEL IN PHR ARCHITECTURE

AUTHORS: S. Kikuchi, S. Sachdeva, and S. Bhalla,
In recent years, some practical and commercial Personal Health Records and some related services such as Google Health [1] and Microsoft HealthVault [2] have been launched. On the other hand, Cloud Computing has matured more and become the major streams to realize a more effective operational environment. However so far, there have been few studies in regards to applying Cloud architecture in the PHR explicitly despite generating volume data. In this paper, we review our trial on the general architecture design by applying the Cloud components for supporting healthcare record areas and clarify the required conditions to realize it.

3.3 HEALTH INFORMATION PRIVACY, SECURITY, AND YOUR EHR

AUTHORS: M. Bellare
If your patients lack trust in Electronic Health Records (EHRs) and Health Information Exchanges (HIEs), feeling that the confidentiality and accuracy of their electronic health information is at risk, they may not want to disclose health information to you. Withholding their health information could have life-threatening consequences. To reap the promise of digital health information to achieve better health outcomes, smarter spending, and healthier people, providers and individuals alike must trust that an individual's health information is private and secure.

3.4 A SECURE ANTI-COLLUSION DATA SHARING SCHEME FOR DYNAMIC GROUPS IN THE CLOUD

AUTHORS: C. Ng and P. Lee. Revdedup
Benefited from cloud computing, users can achieve an effective and economical approach for data sharing among group members in the cloud with the characters of low maintenance and little management cost. Meanwhile, we must provide security guarantees for the sharing data files since they are outsourced. Unfortunately, because of the frequent change of the membership, sharing data while providing privacy-preserving is still a challenging issue, especially for an untrusted cloud due to the collusion attack. Moreover, for existing schemes, the security of key distribution is based on the secure communication channel, however, to have such channel is a strong assumption and is difficult for practice. In this paper, we propose a secure data sharing scheme for dynamic members. Firstly, we propose a secure way for key distribution without any secure communication channels, and the users can securely obtain their private keys from group manager. Secondly, our scheme can achieve fine-grained access control, any user in the group can use the source in the cloud and revoked users cannot access the cloud again after they are revoked. Thirdly, we can protect the scheme from collusion attack, which means that revoked users cannot get the original data file even if they conspire with the

untrusted cloud. In our approach, by leveraging polynomial function, we can achieve a secure user revocation scheme. Finally, our scheme can achieve fine efficiency, which means previous users need not to update their private keys for the situation either a new user joins in the group or a user is revoked from the group

3.5 ADVANCE SECURITY TO CLOUD DATA STORAGE

AUTHORS: P. Lee, and W. Lou

The proposed system is an effective and flexible distributed Scheme with explicit dynamic data support to ensure the correctness of user's data in the cloud. To fully ensure the data integrity and save the cloud users computation it is of critical importance to enable public auditing service for cloud data storage, so that users may depend on independent third party auditor to audit the outsourced data. The Third party auditor can periodically check the integrity of all the data stored in the cloud .which provides easier way for the users to ensure their storage correctness in the cloud.

IV. INTRUSION DETECTION FRAMEWORK BASED ON LEAST SQUARE SUPPORT VECTOR MACHINE

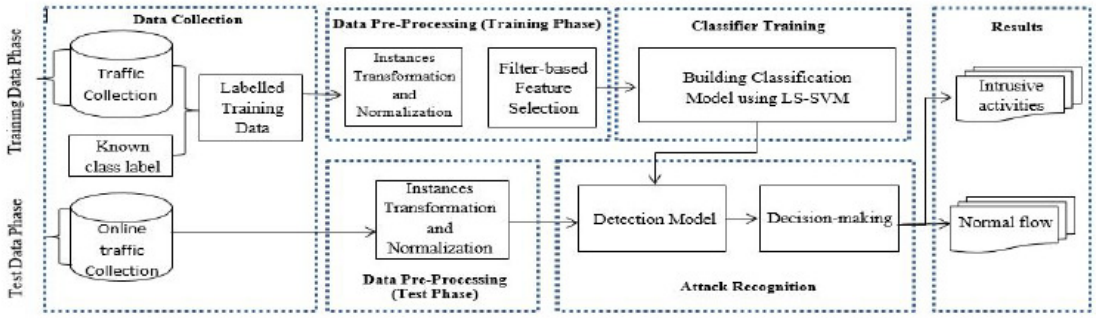

Fig 1: The framework of the LS-SVM-based Intrusion Detection System.

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.

4.1 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

Data normalization

An essential step of data preprocessing after transferring all symbolic attributes into numerical values is normalisation.

In Feature Selection the values which system have got are compared with trained dataset and only some features are selected based on the algorithm flexible mutual information based feature selection and flexible linear correlation coefficient based feature selection.

4.3 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.

An intrusion detection system (IDS) is a device or software application that monitors a network or systems for malicious activity or policy violations. Any detected activity or violation is typically reported either to an administrator or collected centrally using a security information and event management (SIEM) system.

Intrusion Detection Framework on Least Square Vector Machine The framework of the proposed intrusion detection system is depicted in figure 1. The detection framework is comprised of four phases: (1) data collection (2) data preprocessing (3) classifier training, and (4) attack recognition.

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: L_x - 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
```

V. CONCLUSION

Late examinations have shown that two guideline parts are fundamental to collect an IDS. They are an intense request technique and a compelling segment assurance estimation. In this paper, a controlled channel based segment decision count has been proposed, specifically Flexible Mutual Information Feature Selection (FMIFS). FMIFS is a change over MIFS and MMIFS. FMIFS proposes an acclimation to Battiti's figuring to lessen the redundancy among features. FMIFS takes out the redundancy parameter _ required in MIFS and MMIFS. This is appealing for all intents and purposes since there is no specific system or govern to pick the best a motivator for this parameter.FMIFS is then joined with the LSSVM methodology to develop an IDS. LSSVM is a smallest square type of SVM that works with reasonableness prerequisites as opposed to difference objectives in the arrangement proposed to comprehend a course of action of direct conditions for gathering issues rather than a quadratic programming issue. The proposed LSSVMIDS + FMIFS has been surveyed using three comprehended interference recognizable proof datasets: KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets. The execution of LSSVM-IDS + FMIFS on KDD Cup test data, KDDTest+ and the data, accumulated on 1, 2 and 3 November 2007, from Kyoto dataset has demonstrated better course of action execution to the extent arrange precision, acknowledgment rate, false positive rate and F-measure than a bit of the present area approaches. Additionally, the proposed LSSVM-IDS + FMIFS has exhibited comparative results with other best in class approaches while using the Corrected Labels sub-dateset of the KDD Cup 99 dataset and attempted on Normal, DoS, and Probe classes; it beats other acknowledgment models when attempted on U2R and R2L classes. Besides, for the examinations on the KDDTest 21 dataset, LSSVM-IDS + FMIFS produces the best

portrayal precision differentiated and other area systems attempted on the same dataset. Finally, in perspective of the test comes to fruition achieved on all datasets, it can be assumed that the proposed revelation structure has achieved promising execution in distinguishing intrusions over PC systems. All things considered, LSSVM-IDS + FMIFS has played out the best when differentiated and the other best in class models. Notwithstanding the way that the proposed incorporate decision computation FMIFS has shown enabling execution, it could be also enhanced by propelling the chase system. In like manner, the impact of the unbalanced case transport on an IDS ought to be given a wary idea in our future examinations.

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.