ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

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

[1] B. Avanthi, [2] Mr.N.Srinivas
[2] Associate Professor, Head of the Department
[1][2] Department of computer Science Engineering, Vignana Bharathi Institute of Technology, Telangana, India.

***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.**

INTRODUCTION

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 [1]. 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 [2] and Kernel Miner [3] are two of the earliest attempts to build intrusion detection schemes. Methods proposed in [4] and [5] have successfully applied machine learning techniques, such as Support Vector Ma 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. [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 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.

EXISTING SYSTEM

- 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

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

Miner are two of the earliest attempts to build intrusion detection schemes.

- 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.

DISADVANTAGES OF EXISTING SYSTEM

- 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.
- Current network traffic data, which are often huge in size, present a major challenge to IDSs. 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.
- Large-scale datasets usually contain noisy, redundant, or uninformative features which present critical challenges to knowledge discovery and data modeling.

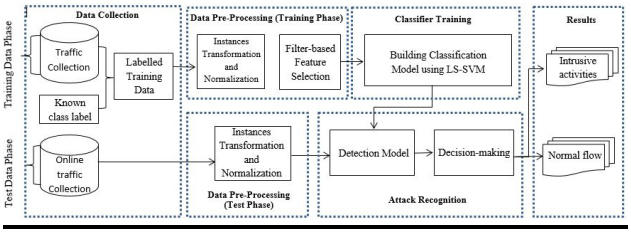
PROPOSED SYSTEM

- We have proposed a hybrid feature selection algorithm (HFSA). HFSA consists of two phases.
- The upper phase conducts a preliminary search to eliminate irrelevant and redundancy features from the original data. This helps the wrapper method (the lower phase) to decrease the searching range from the entire original feature space to the pre-selected features (the output of the upper phase). The key contributions of this paper are listed as follows.
- This work proposes a new filter-based feature selection method, in which theoretical analysis of mutual information is introduced to evaluate the dependence between features and output classes.
- The most relevant features are retained and used to construct classifiers for respective classes. As an enhancement of Mutual Information Feature Selection (MIFS) and Modified Mutual Information based Feature Selection (MMIFS), the proposed feature selection method does not have any free parameter, such as in MIFS and MMIFS. Therefore, its performance is free from being influenced by any inappropriate assignment of value to a free parameter and can be guaranteed. Moreover, the proposed method is feasible to work in various domains, and more efficient in comparison with HFSA, where the computationally expensive wrapper-based feature selection mechanism is used.
- We conduct complete experiments on two well known IDS datasets in addition to the dataset used. This is very important in evaluating the performance of IDS since KDD dataset is outdated and does not contain most novel attack patterns in it. In addition, these datasets are frequently used in the literature to evaluate the performance of IDS. Moreover, these datasets have various sample sizes and different numbers of features, so they provide a lot more challenges for comprehensively testing feature selection algorithms.
- Different from the detection framework proposed that designs only for binary classification, we design our proposed framework to consider multiclass classification problems. This is to show the effectiveness and the feasibility of the proposed method.

ADVANTAGES OF PROPOSED SYSTEM

- 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.

System Architecture



ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

Modules
a) Data Preprocessing
b) Filter primarily based function selection
c) Attack type & Recognition
d) Performance Evaluation

MODULES DESCRIPTION

a) Data Preprocessing
i. The statistics received at some level within the section of records collection are first processed to generate the fundamental features including those in KDD Cup 99 dataset. The educated classifier calls for each report inside the enter statistics to be represented as a vector of real huge range. Thus, each symbolic feature in a dataset is first transformed right into a numerical price. For instance, the KDD CUP ninety nine dataset carries numerical similarly to symbolic functions. These symbolic capabilities include the kind of protocol (i.E., TCP, UDP and ICMP), carrier kind (e.G., HTTP, FTP, Telnet and so forth) and TCP recognition flag (e.G., SF, REJ and so forth). The method actually replaces the values of the explicit attributes with numeric values.
ii. An essential step of facts preprocessing after transferring all symbolic attributes into numerical values is normalization. Data normalization is a way of scaling the charge of each attribute right right into a well-proportioned variety, in order that the unfairness in want of capabilities with more values is removed from the dataset.

b) Filter based definitely function preference
i. If one considers correlations among community visitors statistics to be linear institutions, then a linear degree of dependence which include linear correlation coefficient can be used to degree the dependence between two random variables. However, thinking about the actual international communication, the correlation among variables can be nonlinear as properly. Apparently, a linear degree can't display the relation between nonlinearly based variables. Thus, we need a degree able to studying the relation amongst two variables irrespective of whether they will be linearly or nonlinearly based. For these motives, this art work intends to find out a manner of selecting most applicable features from a feature space irrespective of the sort of correlation amongst them.

ii.We growth algorithms for function selection way. There are: Flexible mutual information based feature selection and Feature Selection Based on Linear Correlation Coefficient.

c) Attack class & Recognition
i. In modern, it's far less complicated to construct a classifier to distinguish between instructions than considering multiclass in a hassle. This is due to the fact the decision obstacles within the first case can be simpler. The first part of the experiments in this paper makes use of two training, wherein statistics matching to the ordinary magnificence are said as normal data, otherwise are taken into consideration as assaults. However, to deal with a problem having greater than commands, there are famous techniques: One-Vs- One" (OVO) and One-Vs-All" (OVA).

ii. After completing all the aforementioned steps and the classifier is knowledgeable using the best first-class subset of capabilities which includes the maximum correlated and critical abilities, the regular and intrusion traffics may be identified via the usage of the stored trained classifier. The check statistics is then directed to the saved skilled model to discover intrusions. Records matching to the normal beauty are considered as regular information, and the opportunity information are pronounced as attacks. If the classifier version confirms that the record is odd, the subclass of the atypical record (type of attacks) may be used to decide the document's type

d) Performance Evaluation
i. The majority of the IDS experiments have been finished at the KDD Cup ninety nine datasets. In addition, those datasets have awesome facts sizes and various numbers of features which provide comprehensive assessments in validating feature selection strategies.
ii. The KDD Cup 99 dataset is one of the most popular and whole intrusion detection datasets and is widely performed to assess the overall overall performance of intrusion detection systems. It consists of 5 special schooling, which is probably everyday and four sorts of attack (i.E., DoS, Probe, U2R and R2L). It incorporates schooling data with approximately 5 million connection statistics and test information with approximately million connection information. Each report in those datasets is labeled as both everyday or an assault, and it has 41 unique quantitative and qualitative functions.
iii. Several experiments had been accomplished to evaluate the overall performance and effectiveness of the proposed LSSVMIDS. For this motive, the accuracy

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

price, detection fee, fake terrific rate and F-diploma metrics are implemented.

LITERATURE SURVEY

1. Traffic-conscious layout of a excessive velocity fpga community intrusion detection device, Computers

AUTHORS: S. Pontarelli, G. Bianchi, S. Teofili

Security of today's networks intently relies on community intrusion detection structures (NIDSs). The functionality to proper away replace the supported rule devices and stumble upon new rising attacks makes concern-programmable gate arrays (FPGAs) a completely appealing generation. An important trouble is a way to scale FPGA-based NIDS implementations to ever faster community links. Whereas a trivial approach is to balance site visitors over a couple of, however functionally identical, hardware blocks, every implementing the whole rule set (numerous thousand rules), the plain cons is the linear increase inside the useful resource career. In this paintings, we promote a selected, site visitors-aware, modular method in the layout of FPGA-primarily based NIDS. Instead of in primary terms splitting website visitors in some unspecified time in the future of equal modules, we classify and company homogeneous traffic, and dispatch it to in some other manner capable hardware blocks, each supporting a (smaller) rule set tailored to the specific traffic elegance. We implement and validate our technique using the rule of thumb of thumb set of the well-known Snort NIDS, and we experimentally look into the emerging alternate-offs and benefits, showing useful useful resource financial savings as an awful lot as eighty percent based totally on real-global internet site website online traffic facts accrued from an operator's backbone.

2. Network-based totally actually intrusion detection with help vector machines

AUTHORS: D. S. Kim, J. S. Park

This paper proposes a way of applying Support Vector Machines to network-based Intrusion Detection System (SVM IDS). Support vector machines(SVM) is a studying technique which has been effectively finished in masses of software regions. Intrusion detection can be taken into consideration as -elegance class trouble or multi-beauty kind problem. We used dataset from 1999 KDD intrusion detection contest. SVM IDS became found out with triaing set and tested with take a look at gadgets to evaluate the overall performance of SVM IDS to the unconventional attacks. And we also examine the importance of each function to enhance the overall standard overall performance of IDS. The effects of experiments showcase that utilising SVM in Intrusion Detection System may be an effective and green manner for detecting intrusions

3. An powerful technique for intrusion detection using neuro-fuzzy and radial svm classifier

AUTHORS:A.Chandrasekhar,K. Raghuveer

Intrusion detection is not however a outstanding era. This has given records mining the opportunity to make numerous important contributions to the arena of intrusion detection. In this paper, we've were given proposed a latest approach through using statistics mining strategies such as neuro-fuzzy and radial foundation assist vector device (SVM) for the intrusion detection tool. The proposed approach has four major steps in which, first step is to carry out the Fuzzy C-way clustering (FCM). Then, neuro-fuzzy is informed, such that each of the statistics point is skilled with the corresponding neuro-fuzzy classifier related to the cluster. Subsequently, a vector for SVM magnificence is original and within the fourth step, kind the use of radial SVM is completed to come upon intrusion has befell or now not. Data set used is the KDD cup 99 dataset and we have used sensitivity, specificity and accuracy due to the fact the evaluation metrics parameters. Our method need to gain better accuracy for all kinds of intrusions. It carried out approximately 98.94 % accuracy in case of DOS assault and reached heights of ninety seven.11 % accuracy in case of PROBE assault. In case of R2L and U2R assaults it has attained ninety seven.Seventy 8 and ninety seven.80 % accuracy respectively. We in assessment the proposed method with the opportunity contemporary state of artwork techniques. These comparisons proved the effectiveness of our technique

4 .Intrusion detection using an ensemble of practical paradigms

AUTHORS: S. Mukkamala, A. H. Sung, A. Abraham

Soft computing strategies are an increasing extensive range of getting used for hassle solving. This paper addresses the use of an ensemble method of numerous gentle computing and tough computing techniques for intrusion detection. Due to growing incidents of cyber attacks, constructing effective intrusion detection structures are vital for defensive information structures safety, and but it stays an elusive purpose and a wonderful venture. We studied the overall performance of Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and Multivariate Adaptive Regression Splines (MARS). We show that an ensemble of ANNs, SVMs and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 1, January 2018

MARS is superior to individual methods for intrusion detection in terms of category accuracy.

5. A new technique to intrusion detection based on an evolutionary tender computing version the usage of neurofuzzy classifiers

AUTHORS: A. N. Toosi, M. Kahani

An intrusion detection tool's important goal is to classify sports activities of a gadget into two crucial education: everyday and suspicious (intrusive) sports activities activities. Intrusion detection systems normally specify the shape of attack or classify sports in some precise businesses. The purpose of this paper is to include severa easy computing techniques into the classifying tool to find out and classify intrusions from everyday behaviors based totally at the assault kind in a laptop network. Among the severa gentle computing paradigms, neuro-fuzzy networks, fuzzy inference technique and genetic algorithms are investigated on this paintings. A set of parallel neuro-fuzzy classifiers are used to do an initial kind. The fuzzy inference system need to then be primarily based definitely mostly on the outputs of neuro-fuzzy classifiers, making very last preference of whether or no longer the current hobby is ordinary or intrusive. Finally, that allows you to acquire the high-quality surrender result, genetic algorithm optimizes the shape of our fuzzy choice engine. The experiments and evaluations of the proposed technique were completed with the KDD Cup ninety nine intrusion detection dataset.

CONCLUSION

Recent studies have validated that principal additives are crucial to assemble an IDS. They are a robust kind method and an efficient function choice algorithm. In this paper, a supervised filter-based totally feature choice algorithm has been proposed, especially Flexible Mutual Information Feature Selection (FMIFS). FMIFS is an improvement over MIFS and MMIFS. FMIFS shows a change to Battiti's algorithm to reduce the redundancy among capabilities. FMIFS gets rid of the redundancy parameter _ required in MIFS and MMIFS. This is acceptable in exercise for the cause that there can be no specific gadget or guiding principle to select the satisfactory price for this parameter.

FMIFS is then combined with the LSSVM method to assemble an IDS. LSSVM is a least square model of SVM that works with equality constraints rather than inequality constraints inside the components designed to solve a fixed of linear equations for category troubles instead of a quadratic programming hassle. The proposed LSSVMIDS + FMIFS has been evaluated the use of three well known intrusion detection datasets: KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets.

The basic overall performance of LSSVM-IDS + FMIFS on KDD Cup test records, KDDTest+ and the statistics, collected on 1, 2 and three November 2007, from Kyoto dataset has exhibited higher type performance in terms of type accuracy, detection price, false extraordinary rate and F-diploma than a number of the prevailing detection methods. In addition, the proposed LSSVM-IDS + FMIFS has proven comparable outcomes with different nation-of the-artwork processes when the use of the Corrected Labels sub-dateset of the KDD Cup 99 dataset and examined on Normal, DoS, and Probe classes; it outperforms other detection models whilst examined on U2R and R2L instructions. Furthermore, for the experiments at the KDDTest 21 dataset, LSSVM-IDS + FMIFS produces the best category accuracy in comparison with unique detection structures examined on the identical dataset.

Finally, based totally at the experimental results completed on all datasets, it may be concluded that the proposed detection gadget has carried out promising performance in detecting intrusions over computer networks. Overall, LSSVM-IDS + FMIFS has executed the first-rate while in evaluation with the alternative united states of america-of the- artwork models.Although the proposed characteristic choice algorithm FMIFS has verified encouraging performance, it may be further advanced by using using optimizing the quest method. In addition, the effect of the unbalanced sample distribution on an IDS desires to take shipping of a cautious consideration in our future studies.

REFERENCES

[1] S. Pontarelli, G. Bianchi, S. Teofili, Traffic-aware design of a high speed fpga network intrusion detection system, Computers, IEEE Transactions on 62 (11) (2013) 2322–2334.

[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 eurofuzzy classifiers, Computer communications 30 (10) (2007) 2201–2212.