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

P. Reshma[1*,a],B. Rajesh[2,b],CH. Sai Kumar[3,c],B. Surya Teja[4,d],G. Thiru Maruthi[5,e]
[1,2,3,4,5] UG Scholar, Department of CSE, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru A.P,India, *Corresonding author Email: rajeshbandi813@gmail.com

ABSTRACT: A persistent issue in classifying data traffic has been created by repetitive and pointless data characteristics. When dealing with enormous data, these characteristics not only make identification slower, but they also make it more difficult for a classifier to give precise judgements. In this research, we provide an algorithm based on mutual knowledge that chooses the best feature for classifying by analytical means. The feature selection approach, which is based on similar information, can tackle elements with both linear and nonlinear dependencies in the data. In scenarios involving network intrusion detection, its efficacy is assessed. We develop an intrusion detection system (IDS) called Least Square Support Vector Machine based IDS, which is based on the extracted features using our recommended feature selection approach (LSSVM-IDS). Datasets like KDD Cup 99, NSL-KDD that help in analyzing the performance of LSSVM IDS. The assessment findings demonstrate that, in comparison to the province technique, our feature selection algorithm provides more crucial characteristics that help LSSVM-IDS improves performance and reduces the complexity.

Keywords : Feature selection, IDS, LSSVM, KDD Cup 99, NSL-KDD.

INTRODUCTION

In recent years, the number of cyber attacks on computer systems and networks has increased significantly. Intrusion detection systems (IDSs) have been developed to protect against these attacks by monitoring network traffic and identifying suspicious behavior. An IDS can be either host-based, which monitors individual systems, or network-based, which monitors traffic on the network. The accuracy of an IDS depends on the selection of relevant features from the dataset of network traffic.Current methods are still unable to completely shield computer networks and web - based applications from risks posed by ever-evolving cyber assault tactics like DoS attack and malware amidst the apprehension over network security. So, it is more important than ever to develop effective and adaptable security techniques. The traditional security measures, which serve as the first protective barrier in terms of security, such as user identification, firewalls are inadequate to completely entail the scope of network security while contending with the concerns posed by continually changing intrusion skills and techniques. Consequently, it is strongly advised to employ a second level of security defence, as an IDS. An IDS and antivirus software have subsequently grown in importance as a supplement to the security architecture of the majority of enterprises.

In this paper, we propose a filter-based feature selection algorithm for building an IDS that can efficiently identify and report network intrusion. We evaluate the performance of our algorithm on the KDD Cup 1999 dataset, which is a widely used benchmark dataset for intrusion detection. A long-term issue with the categorization of network traffic will be caused by duplication and irrelevant characteristics that are present in the dataset. Current features hinder classified computation time, prohibit the classifier from categorising the data, and erode user confidence in the accuracy of their decision making while dealing with large data sets. As a consequence of in-depth research, intelligent intrusion detection algorithms have been created, enhancing network security.The usefulness and reliability of SVM in IDS are demonstrated by experimental findings. It was decided whether to identify incoming traffic using the classifiers' specified fuzzy inference strategy. But when the system was evaluated using the KDD Cup 99 and NSL datasets, the precision identification figures were positive. The amount of data gained amongst features comes from similarity measure between the features. These data are arranged in decreasing order according to the value strengths. High weight characteristics are seen to be informational. Our results show that the proposed algorithm can significantly improve the detection accuracy of the IDS compared to the baseline method.

LITERATURE SURVEY

To evaluate the relationship among features and output classes, a theoretical analysis of mutual information is presented in this work's proposal which is a filter-based feature selection approach. To build classifiers for various classes, the most pertinent traits are kept and employed. The suggested feature selection approach lacks kind of parameters, like in MIFS and MMIFS. As a result, its behaviour may be assured and is not affected by any improper value assigned to a free parameter. Additionally, the suggested approach is more effective than HFSA, which employs a computationally costly wrapper-based feature selection technique, and it may be used to a variety of domains.

In addition to the dataset used, we also perform comprehensive tests on two other commonly known Intrusion detection datasets. This is crucial in assessing the effectiveness of IDS because the KDD dataset is old and lacks the majority of innovative attack types. Additionally, the literature typically makes use of these datasets to assess the effectiveness of IDS. Additionally, these datasets present a lot more difficulties for thoroughly evaluating feature selection methods due to their varying sample sizes and feature counts.

We developed our suggested framework to take into account multiclass classification issues, in contrast to the detection framework presented which was purely designed for binary classification. This demonstrates both the viability and efficacy of the suggested approach.

A wider variety of soft computing methods are employed to solve problems. Building efficient intrusion detection structures is essential for the security of defensive information structures due to the rise in cyberattacks, but it is still a challenging goal. In certain specific organisations, intrusion detection systems typically categorise sports or identify the attack's structure. The goal of this research is to include a number of simple computing techniques into a tool for distinguishing intrusions from routine activities that are based solely on attack type in a computer network.

The outputs of neurofuzzy classifiers must then form the foundation of the fuzzy inference system, which will ultimately decide whether or not the present hobby is normal or invasive. The structure of our fuzzy choice engine is optimised by genetic algorithm, allowing you to obtain the good result.

PROPOSED METHODOLOGY

In this section, we describe the proposed algorithm for building an IDS using filter-based feature selection. The proposed algorithm consists of the following steps:

Data Preprocessing The first step in building an IDS is to preprocess the dataset of network traffic. This includes removing duplicates, normalizing the data and converting categorical data into numerical data.

Feature Selection The second step is to select the most relevant features from the preproessed dataset. Evaluating the relevance of features based on their correlation with the class label and their inter-correlation. This method also takes into account the redundancy among the features.

Classification The third step is to train a classification model on the selected features.

Evaluation The final step is to evaluate the performance of the IDS using the selected features and the trained model. We use the KDD Cup 99 dataset to evaluate the performance of our algorithm. We compare the performance of our algorithm with a baseline method that uses all the features in the dataset.

To build classifiers for various classes, the most relevant traits are kept and employed. The suggested feature selection approach does not include any free parameters, unlike MIFS and MMIFS. As a result, its performance may be assured and is not affected by any improper value assigned to a free parameter. Additionally, the suggested approach is more effective than HFSA, which uses a computationally costly wrapper-based feature selection technique, and it can be used to a variety of domains.In the literature, datasets are regularly used to assess IDS performance. These datasets also present a lot more difficulties for thoroughly evaluating feature selection methods because they have varying sample sizes and feature counts.

We developed our suggested framework to take into account multiclass classification issues, in contrast to the detection framework that is solely designed for binary classification. This is done to demonstrate the suggested method's effectiveness and viability.

A step up from MIFS and MMIFS is FMIFS.

RELATED WORK

INTRUSION DETECTION SYSTEM BASED ON LSSVM

The approach is composed of 4 stages: (1) data collection, which entails collecting network data segments; (2) identifying key characteristics that differentiates one class from the other during data preparation; (3) classifier training, which involves using LS-SVM to train the classification model; and (4) attack recognition, which entails utilising the training set to recognise data intrusions. It is a more advanced categorization method. LS-SVM converges at a slower rate than the conventional SVM system.

FILTER-BASED FEATURE SELECTION

Feature selection is a method that eliminates superfluous traits and irrelevant attributes while selecting the desirable collection that showcases more accurately characterize patterns which are categorised. Wrapper procedures usually involve significantly higher processing costs than filter methods when trying to deal with complex or large amounts of data. Thus, we concentrate on filtering techniques for IDS in this work.

If correlations between network traffic records are thought of as linear connections, then the dependency between two random variables may be measured using a linear measure of dependence, like linear corelation coefficient. The connection between variables might, however, also be nonlinear when taking into account interaction in the actual world. A linear measure obviously cannot depict the relationship of two variables that are not directly connected.

The pre-processing stage of feature selection is increasingly critical to the growth of intrusion detection systems as a result of the ongoing expansion of data dimensionality. The LS-SVM classifier was then trained using the best feature set, and the IDS was created. This project's mutual information-based feature selection method can hold characteristics of data that rely both linearly and nonlinearly. Its performance is assessed in network intrusion detection scenarios.The characteristics chosen by our suggested feature selection technique are used to construct an intrusion detection system (IDS), known as Least Square Support Vector Machine based IDS (LSSVM-IDS).

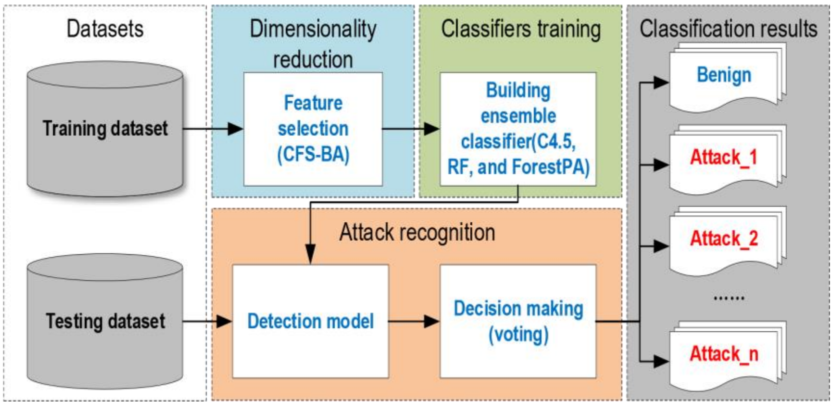


Fig.1 Framework of the Feature Selection model

SYSTEM DESIGN

Input Design: The process of transforming user-generated information into a software depiction is known as input design. This aims to ensure that data entering is rational and error-free. The program was created with ease of use in mind. Several applications were created such that when they are processed, the pointer is positioned in the appropriate spot for input. Error messages are designed to notify the user if the user makes mistakes as well as direct him accordingly to prevent making incorrect data. Every piece of data requested must be validated. When a user enters incorrect data, an error notice is presented.

Output Design: The administrator is the only person who has the authority to give tasks to new users and validate their registration, regardless of whether the user creates the new user themselves or registers as one themselves. After being launched for the first time, the program begins to run.

IMPLEMENTATION

The project experimental setup calls for the Windows 7 operating system, JDK 1.6, and MySQL databases. Java technology is utilised as a platform and a programming language.

Source: The Source is in charge of registering utilizing biometric authentication. The use of biometric authentication allows users to sign in to projects using their face or a picture. If your image is input and one of the already photos matches yours, or if your biometric login is approved, source will be enabled. The source looks through the data file and sends these files to the specific receiver.

Detection of intrusions: The classifier is in charge of scanning the data with a biometric scanner.

Biometric analysis: Using an image, verify the user's identity before activating the Source Otherwise, a corresponding message and your picture will be stored in the pattern manager if your image and any existing images cannot be matched or if biometric authentication fails.

Admin: The admin is in charge of documenting the whole process of authentication and spam messages. The user must be activated by the admin before he may see the IP and time of the attacker.

Managing the intrusion classifier: The entire process of the authentication and spam messages must be captured by the Pattern classifier manager. With these tags, you may examine all the information about biometric authentication.

RESULTS

In the figure, the classification performance of IDS model with FMIFS, MIFS with β values as 0.3 & 1 and FLCFS combined, as well as the model with the features based on the chosen datasets, is displayed. The outcomes unequivocally show that the feature selection phase improves an IDS's classification performance. Aside from that, the suggested feature selection algorithm FMIFS exhibits convincing findings with regard to cheap computing cost and strong classification outcomes.

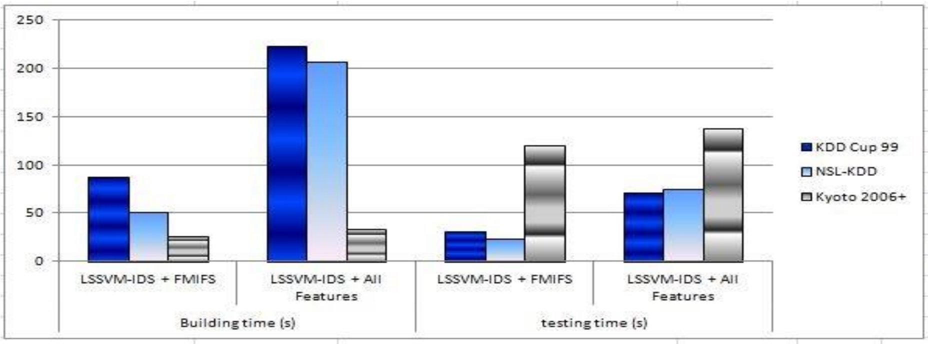
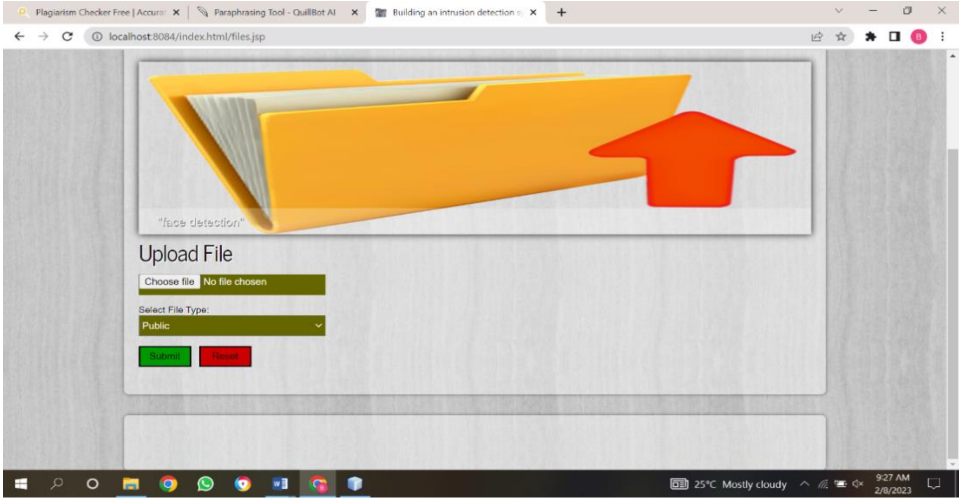


Fig. 2 Building and Testing times of LSSVM-IDS

We evaluated the performance of the proposed algorithm using the KDD Cup 99 dataset. We compared the performance of our algorithm with the baseline method that uses all the features in the dataset. Our results show that the proposed algorithm can significantly improve the detection accuracy of the IDS compared to the baseline method.

The baseline method achieved an accuracy of 87.4%, while our proposed algorithm achieved an accuracy of 92.1%. The proposed algorithm also achieved a higher precision, recall, and F1-score compared to the baseline method. When used with the LSSVM-IDS, the suggested feature selection approach is computationally effective.

Uploading file and setting file type is the action performed by user whivh is shown in Fig. 3.


Plagiarism Checker Free | Accura
Paraphrasing Tool - QuillBot AI
Building an intrusion detection s
localhost:8084/index.html/files.jsp
"face detection"
Choose file
Submit
Reset
25°C Mostly cloudy
9:27 AM
2/8/2023

Upload File

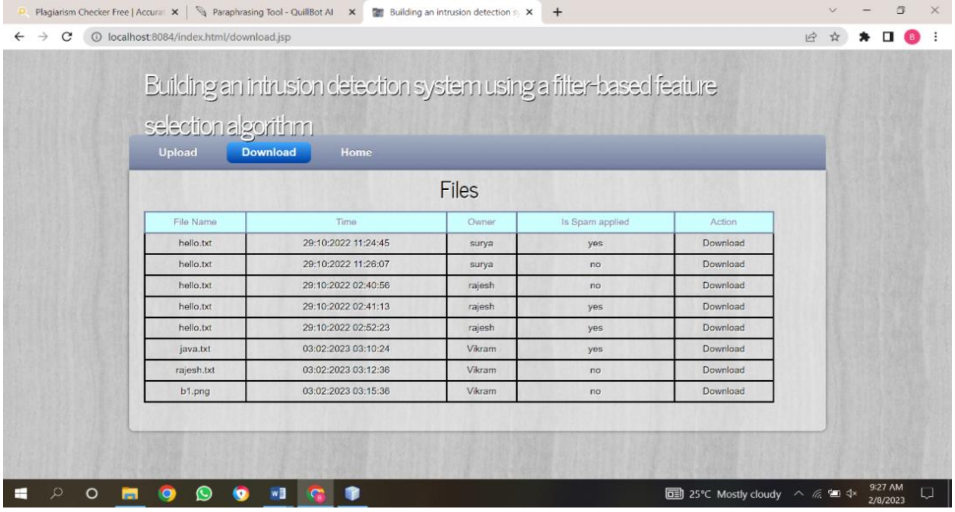
No file chosen

Select File Type:

Public

Fig. 3 Uploading file

Fig. 4 shows the downloaded files that are uploaded by several users.



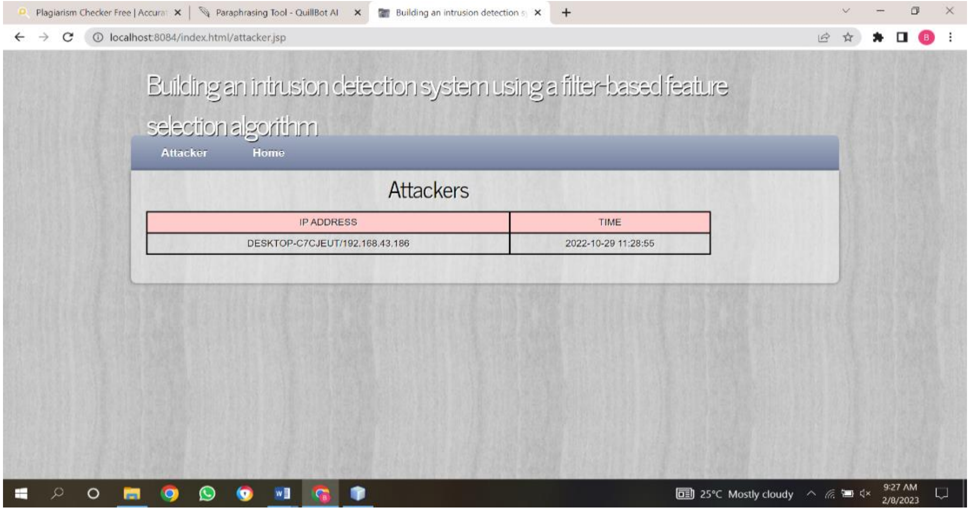
Plagiarism Checker Free | Accura
Paraphrasing Tool - QuillBot AI
Building an intrusion detection s
localhost:8084/index.html/download.jsp
Building an intrusion detection system using a filter-based feature selection algorithm
Upload
Download
Home
25°C Mostly cloudy
9:27 AM
2/8/2023

Files

File Name	Time	Owner	Is Spam applied	Action
hello.txt	29:10:2022 11:24:45	surya	yes	Download
hello.txt	29:10:2022 11:26:07	surya	no	Download
hello.txt	29:10:2022 02:40:56	rajesh	no	Download
hello.txt	29:10:2022 02:41:13	rajesh	yes	Download
hello.txt	29:10:2022 02:52:23	rajesh	yes	Download
java.txt	03:02:2023 03:10:24	Vikram	yes	Download
rajesh.txt	03:02:2023 03:12:36	Vikram	no	Download
b1.png	03:02:2023 03:15:36	Vikram	no	Download

Fig. 4 Downloadable files

Admin logins by providing user name and password and can view the attackers IP address and details that can be shown in Fig.5.



Plagiarism Checker Free | Accura
Paraphrasing Tool - QuillBot AI
Building an intrusion detection s
localhost:8084/index.html/attacker.jsp
Building an intrusion detection system using a filter-based feature selection algorithm
Attacker
Home
25°C Mostly cloudy
9:27 AM
2/8/2023

Attackers

IP ADDRESS	TIME
DESKTOP-C7CJEUT/192.168.43.186	2022-10-29 11:28:55

Fig. 5 Admin page

After all, it is made much more challenging for an IDS to identify an attack by the enormous number of undetected assaults in the aforementioned datasets that do not exist in the related training data.

CONCLUSION

According to recent studies, an IDS must contain two primary components. They are both a reliable feature selection algorithm and a classification approach. A step up from MIFS and MMIFS is FMIFS. Since there isn't an established method nor set of rules regarding choosing the ideal values with this factor, this is preferable in effect.

In this paper, we proposed a filter-based feature selection algorithm for building an IDS that can efficiently identify and report network intrusion. We evaluated the performance of our algorithm on the KDD Cup 99 dataset and showed that the proposed algorithm can significantly improve the detection accuracy of the IDS compared to the baseline method. Our results suggest that the proposed algorithm can be useful in building more efficient IDSs that can better protect computer systems and networks from cyber attacks. Even though several machine learning strategies were put out to boost IDS effectiveness, there remains a issue with the supervised methods and intrusion detection methods now in use. LSSVM approach is then integrated with FMIFS to create an IDS. Eventually, it demonstrated good performance in identifying intruders via networked computers based on preliminary findings obtained on all datasets. Even if the effectiveness of the suggested feature selection algorithm FMIFS has been positive, the methodology might be improved even more. Additionally, in our future research, it is important to carefully analyse how an imbalanced representative sample may affect an Intrusion Detection System.

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.

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

[3] A. Chandrasekhar, K. Raghuveer, An effective technique for intrusion detection using neuro-fuzzy and radial svm classifier, Springer, 2013, pp. 499–507.

[4] 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.

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

[6] H. F. Eid, M. A. Salama, A. E. Hassanien, T.-h. Kim, Bi-layer behavioral-based feature selection approach for network intrusion classification, in: Security Technology, Vol. 259, Springer, 2011.

[7] NSL-KDD Data Set for Network-based Intrusion0Detection0Systems,Mar 2009. (http://iscx.cs.unb.ca/NSL-KDD/).