FEATURE SELECTION AND CLASSIFICATION MODELS FOR NETWORK INTRUSION DETECTION SYSTEM- A SURVEY

Dhrisya K[1], Maya Mohan[2], and Sruthy Manmadhan[3]
[1]M.Tech First Year Student, [2,3]Assistant Professor
Department of Computer Science and Engineering,
N.S.S College of Engineering, Palakkad
Email: [1]dhrisyachandra@gmail.com, [2]mayajeevan@gmail.com, [3]sruthym.88@gmail.com

***Abstract*—Network intrusion detection system (NIDS) is a hardware or software component designed to monitor traffics on a network and detect anomalies and attack in the network. These systems gather information from the network and detect an attack by examining the content and header information of all packets that are moving across the network. Network intrusion detection can be implemented through different approaches including machine learning, data mining and many more. Intrusion detection system highly depends on features of input data. These features give information to learning algorithms used for intrusion detection. Feature selection is the procedure of selecting the subset of the accessible components to decrease the dimensionality of dataset. Which remove the irrelevant and insignificant feature from the dataset leads to increase the performance and reduce the time complexity for the entire detection process. Mainly the algorithms are used for effective feature selection in order to obtain optimized detection rate. The aim of this paper is to present a survey of various feature selection methods and classifiers for NIDS on KDD CUP 99 datasets.**

Keywords—Intrusion detection system, feature selection, Dataset, Classification.

I. INTRODUCTION

An intrusion detection system is a device or software application that monitors a network or systems for malicious activity. The major catagories of intrusion detection systems are network intrusion detection system and host based intrusion detection system. This survey paper is based on feature selection methods for network intrusion detection system.In recent times, network security has been an important topic of many researches. With the tremendous growth of network, there is a gradual increase in the number of intruders and new attacks. Traditional and current security policies are not sufficient to detect attack efficiently. Technologies like encryption, firewalls, authorization mechanism offer security, but they still sensitive for attacks from hackers. Network intrusion detection system is the one of the solution to detect the intrusions happens on the network. The purpose of NIDS is to help the computer system to deal with the attack.

Current network traffic data, which are often huge in size, present. These big data slow down the entire detection process and may lead to unsatisfactory classification accuracy due to the computational difficulties in handling such data. Classifying a huge amount of data usually causes many mathematical difficulties which then lead to higher computational complexity. As a well-known intrusion evaluation dataset, KDD Cup 99 dataset is a typical example of large-scale datasets. This dataset consists of more than five million of training samples and two million of testing samples respectively. This large feature space contains irrelevant, redundant and noisy features. This leads to increase the false positive rate and misclassification rate of the detection system. So it is important to remove these ambiguous features from the dataset and selecting key features for the better classification.

Feature selection is frequently used techniques in data preprocessing for selecting a subset of relevant features to build effective IDS. The efficient feature subset can improve the training and testing time that helps to build IDS with guaranteed high detection rates and makes it suitable for real time and on-line detection of attacks. Methods of feature selection are generally classified into filter and wrapper methods. In comparison with filter methods, wrapper methods are often much more computationally expensive when dealing with large-scale data. This survey focuses on different feature selection methods used for the feature selection and performance evaluation those methods.

II. RELATED WORKS

The accuracy of classifiers used in network intrusion detection system is not only depends on the classification algorithm but also on the feature selection method. The effective feature selection method improves the performance of the detection system. The following sections discuss about different feature selection methods implemented in recent years.

A. Filter based feature selection algorithm

Muhammed A.Ambusaidi and Priyadarsi Nanda proposes a filter based feature selection algorithm based intrusion detection system[1]. This algorithm works purely based on mutual information concept.This detection system is evaluated by using the standard KDD dataset. The proposed feature selection method does

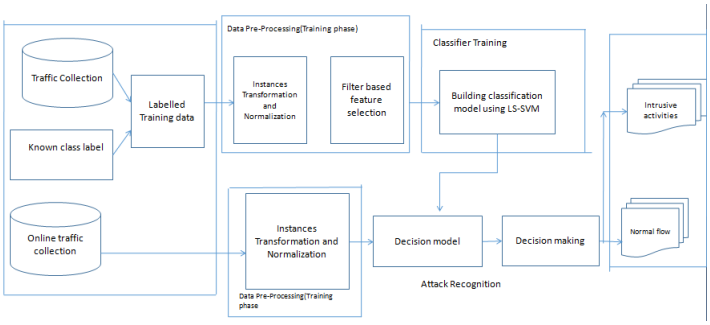

Figure 1. The framework of the LS-SVM-based intrusion detection system[1]

not contain any parameter such as β in MIFS(Mutual Information Feature Selection) and MMIFS(Modified Mutual Information based Feature Selection)[2, 3]. This feature reduce the computational complexity of the system. This new intrusion detection works based on filter based faeture selection method instead of computationally expensive wrapper method.The proposed system classify the attack using advanced version of SVM(Support Vector Machine) classifier called LS-SVM(Least Square Support Vector Machine classifier). Five LS-SVM classifier used in the system distinguishes one class of record from the other.

The figure 1 shows the framework of proposed intrusion detection system. Which consist of four phases of operations namely Data collection, Data processing, Classifier training and attack recognition. The proposed Flexible Mutual Information Feature Selection algorithm shows promising results in terms of low computational cost and high classification results.

B. Hadoop based parallel binary bat algorithm

P.Netesan proposed an another feature selection algorithm for network intrusion detection system called hadoop based parallel binary bat algorithm[4]. This proposed system is designed using hadoop framework. It support parallel computing model, due to which the system handle big data processing. This intrusion detection system uses extended version of MapReduce model namely iterative MapReduce model to select optimal feature subset from the search space. This model helps to improve computational complexity of the system. The distributed classification model using naive bayes algorithm increase the accuracy of attack classification and detection time.

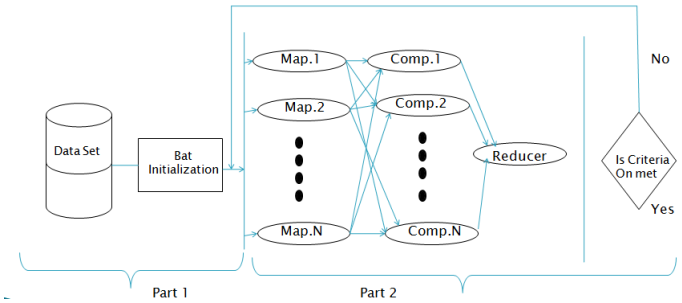

Figure 2. Mapper-Chainer model for Parallel Binary Bat Algorithm[5]

Basically, the bat algorithm[5] is purely based on the echolocation behaviour of natural bats. The feature selection algorithm, creates artificial bats with some parameters such as velecity, frequency, position etc are initialized on the search space. The best bat with fitness value is selected and which is given to the iterative MapReduce programming model.The proposed feature selection methodology shown in figure 2 adopts the Mapper-Chainer model.The first part consists of a single MapReduce responsible for bat initialization and the second part is used for iterative MapReduce function to determine the optimal solution. The iterative process gives an output of feature subset. The overall performance of the system is less time consuming and it also handle huge data processing.The performance of the proposed detection system evaluated by KDD dataset. This system have Naive bayes classifier[6] for the classification of attacks.

C. Attack's feature selection based intrusion detection system

Attack's feature selection based network intrusion detection system using fuzzy control language is proposed by S.Ramakrishnan[7, 8, 9]. This proposed system introduces an entropy based feature selection to select optimal relevant set of features from the search space. Similarly the above discussed system, this system is also evaluated by KDD dataset. The system generate new fuzzy rule[10, 11, 12] using features on the selected subset of features.

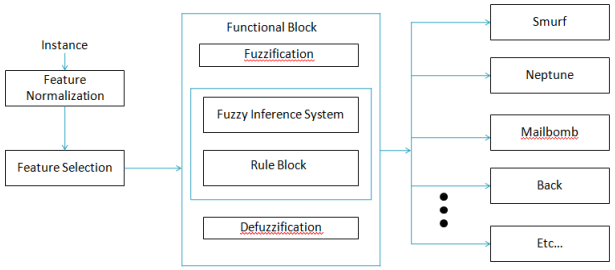

Figure 3. Architecture of proposed layered fuzzy control language[8]

figure 3 shows the proposed intrusion detection system's archtecture. Most of the feature selection methods select same subsets of features for classification of various attacks in intrusion detection system. In order to solve this problem, fuzzy-based layered classifier is proposed for classifying various attacks in the intrusion detection system. This newly introduced layered classifier approach is major difference of this system from the other system. This classifier detect various type of attack and improves the performance as well as reduce the computational time.

D. Effective combining classifier approach using tree algorithm

Jasmin Kervic proposed an effective combining classifier approach using tree algorithms for network intrusion detection system[13]. This Combining classifier approach combines the classifiers used in the intrusion detection system without implementing any feature selection methods on the dataset[4]. This method builds an effective intrusion detection system

by removing feature selection complexities. First each trained classifier over the same trained set is used and independently to perform attack detection. Then the evidences are combined in order to produce final decision. The approach based on classifier combination may also attain effective attack detection as the combination of multiple evidences usually exhibit higher accuracies due to lower false positive rate.

To develop the proposed system, the authors conducted experiments on three algorithm based classifiers namely random tree, NBTree, C4.5. They applied the learned models on two test sets, namely KDDTest+ and KDDTest-21. The performances of the classification are expressed in terms of sensitivity and specificity. Sensitivity measure shows how good our proposed algorithms are when detecting network attacks. Specificity, on the other hand, represents the measure of false detections, or the performance of a classifier in detecting normal attacks. Three detection algorithms achieving highest accuracies were selected: random tree, C4.5, and NBTree.

In case of individual classifiers, random tree performed the best although C4.5 had least number of false alarms and NBTree performed the worse based on any evaluation criteria. In case of combining classifiers approach,random tree + NBTree combination provides better detection performance than the individual random tree classifier or any other combination of classifiers.

E. An IWD-based feature selection method for intrusion detection system

Neha Acharya and Shailendra Singh introduced intelligent water drops (IWD) algorithm[14]. This naturally inspired IWD algorithm is used for the creation of feature subset along with support vector machine as a classifier for evaluation of the features selected. The experiments are conducted using KDD-CUPâ99 dataset. The proposed detection algorithm goes through four phases of operations namely Initialization, Solution building, Reconstructing and Termination condition

- Initialization: This step initializes the values of static and dynamic parameters
- Solution building: This phase develops the solution for every water drop for a single iteration.
- Restructuring: From all the solutions found out by every IWD, the iterationâs best solution is calculated.
- Termination condition: Phase2 and 3 are repeated until the maximum number of iterations is reached.

The figure 4 shows the model of proposed detection system. Like hadoop based binary bat algorithm,this IWD algorithm is also a naturally inspired one. The algorithm concept is purely based on the flow water through the river. This Intelligent Water Drop algorithm is conducted with four phases of operation. That are Initialization, Solution binding, Restructuring and Termination

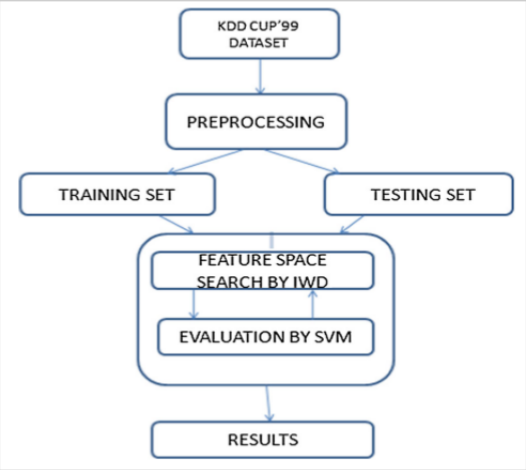

Figure 4. Proposed model for IDS[14]

condition. Here artificial water drops are generated on the search space with some parameters. This algorithm first generate a graph stucture with node represents the features on the search space and edges represent the mutual information between the features. The final output of the algorithm gives a path with some features. These features are considered as the relevant subset of features. This proposed algorithm is simple to execute compared with the other algorithms discussed above.

III. Discussion and Findings

From the analysis, it if clear that most of the IDS was created and evaluated based on the standard KDD dataset.One of the major disadvantage of the system is that KDD dataset does not contain new type of attack features. So, it is impossible to detect new attack. The discussed feature selection and classification techniques provide better detection and false positive rate compared to existing technologies. Commonly used classification models are probability based naive bayes classifier, extended version of LS-SVM classifier and classification algorithms like RBTree,decision tree and randomtree also used in the detection sytems.

The table I clearly shows that the effectiveness of each techniques used for the IDS design.The Fuzzy control language supported layered classifier approach suggested by the author Ramakrishnan gives more efficient detection of various type of attack with high accuracy. The detection rate of IWD based feature selection have highest rate of detection with high false positive rate. The combining classifier technique provide high attack detection rate and reduced false positive rate. When compared with the other intrusion detection system these rate is very less. But this value is more better due the reason that combining classifier approach only classify the attack without implementing feature selection algorithm.This technique also shows that the combined classifier algorithms such as RBTree and random tree combines the advantages of both algorithms and gives a better output.The hadoop based parallel binary bat algorithm handle massive data in the network with low computational complexity.

TABLE I
COMPARISON TABLE

Technique	Dataset	Classifier	Detection Rate
FBFS	KDD dataset	LS-SVM	98.93
HBPBBA	KDD dataset	Naive Bayes Classifier	93.54
FCLFS	KDD dataset	Layered Classifier	99.15
Combining Classifier	NSL-KDD	NBTree+ Random tree	89.24
IWD	KDD Cup99	SVM	99.40

The realistic intrusion detection system dataset based on fuzzy qualitative modeling is a latest dataset[15] developed in 2017.These dataset solves the problems of existing dataset used for the evaluation network intrusion detection. For the generation of this future dataset, the authors firstly proposed a metric using a fuzzy logic system based on the Sugeno fuzzy inference model for evaluating the quality of the realism of existing intrusion detection system datasets. Secondly, based on the proposed metric results, a synthetically realistic next generation intrusion detection systems dataset is designed and generated, and a preliminary analysis conducted to assist in the design of future intrusion detection systems. This generated dataset consists of both normal and abnormal reflections of current network activities occurring at critical cyber infrastructure levels in various enterprises.So, I infer that, Implementing HPBBA using NGIDS-DS instead of KDD Cup99 Will increase the performance with high detection rate. Though which we can easily handle the intrusions on the current network traffic. So, this technique reduce the computational complexity of the detection system. Through the analysis of above discussed systems, it can conclude that each techniques are good compared to the other existing techniques and those have its own shortcomings.

IV. CONCLUSION

Nowadays the network is growing tremendously. The intruders spread across the network is a major threat in the present world. The network intrusion detection system is an effective technique compared to the traditional techniques.

In recent years, the network intrusion detection system is an emerging technique for the detection of attack on the network. The main part of the detection system is data preprocessing part and classification model. The heart of preprocessing part is feature selection method. We can improve the performance of the intrusion detection system through introducing effective feature selection algorithms and classification algorithms. The all techniques are evaluated by the standard KDD dataset. From the survey, it is clear that most of the intrusion detection system uses naÃ¯ve bayes and least square support vector machine. The discussed techniques are good compared to the other existing techniques and those have its own shortcomings.

REFERENCES

[1] Mohammed A. Ambusaidi, Member, Xiangjian , Priyadarsi Nanda, and Zhiyuan Tan,2016, âBuilding an intrusion detection system using a filter-based feature selection algorithmâ,IEEE transactions on computer,Vol.65,Issue.10,pp.2986-2998.

[2] T. M. Cover, J. A. Thomas,2012, Elements of information theory, John Wiley Sons.

[3] R.Battiti,1994,Using mutual information for selecting features insupervised neural net learning, IEEE Transactions on Neural Networks,pp.537-550.

[4] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, N. Yazdani,2011, Mutualinformation-based feature selection for intrusion detection systems, Journal of Network and Computer Applications,Vol.34.

[5] P. Natesan, R. R. Rajalaxmi ,G. Gowrison ,P. Balasubramanie,2016, âHadoop Based Parallel Binary Bat Algorithm for Network Intrusion Detectionâ, International Journal of Parallel Programming Springer,Vol.45,Issue.5,pp.1194-1213.

[6] Yang,X.S,2010,A new meta heuristic bat inspired algorithm .In:Gonzalez,J.R.,etal.(eds.)NatureInspired Cooperative Strategies for Optimization (NICSO 2010), vol.284, pp.65â74. Springer.

[7] Good IJ,1965,The estimation of probabilities: an essay on modern Bayesian methods. MIT Press.

[8] S. Ramakrishnan, S. Devaraju,2016,âAttackâs Feature Selection-Based Network Intrusion Detection System Using Fuzzy Control Languageâ,International Journal of fuzzy system, Springer, Vol.19,Issue.2,pp.316-328.

[9] Gupta, K.K., Nath, B., Kotagiri,2012, R.: Layered approach using conditional random fields for intrusion detection. IEEE Trans. Dependable Sec. Comput.

[10] Timothy, 2010,J.: Ross: Fuzzy Logic with Engineering Applications,John Wiley Sons Ltd, Hoboken.

[11] Tanaka.K,2010,An introduction to Fuzzy Logic for Practical Application.John Wiley Sons Ltd,Hoboken.

[12] Cingolani, P., Alcala-Fdez,J,2010,FuzzyLogic: a robust and flexible fuzzy-logic inference system language implementation, IEEE World Congr. Comput. Intell.

[13] Jasmin Kevric, Samed Jukic, Abdulhamit Subasi,2016,â An effective combining classifier approach using tree algorithms for network intrusion detectionâ,Neural computation and application, Springer,Vol.28.

[14] Neha Acharya, Shailendra,2017,"An IWD-based feature selection method for intrusion detection system".

[15] W.Haider,J.Hu,J.Slay,2017,Generating realistic intrusion detction system dataset based on fuzzy qualitative modelling,Journal of network and computer applications, Elsevier.