International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Feature Selection Based Supervised Learning Method for Network Intrusion Detection

Ch. Mallikarjuna Rao, G. Ramesh, D. V. Lalitha Parameswari
Karanam Madhavi, K. Sudheer Babu

Abstract*: Supervised learning is one of the data mining phenomena where a knowledge model is built for artificial intelligence. Learning from training samples has its advantages in predictive solutions. Such solution is essential for network intrusion detection problems. Networks of all kinds do have problem of intrusions as they are exposed to public communications in one way or other. Intrusions over a network are in the form of network flows that need to be analyzed. Manual observation of the flows and detecting intrusions is very time taking. Therefore it is essential to have an automated system for quickly detection of intrusions to safeguard network systems. There are many intrusion detection systems found in the literature. However, there is need for faster algorithm that makes sense in helping network administrators with accurate knowledge presented. Towards this end we proposed a framework with a feature subset selection mechanism to speed up detection process and improve accuracy of the same. The feature subset selection algorithm and Support Vector Machine (SVM) work together in order to have a faster detection system. Benchmark datasets like KDD and NSL-KDD are used for experiments. The empirical results showed that the proposed SVM-FSS framework shows better performance over the state of the art framework.

***Index Terms*: Data mining, feature selection, intrusion detection, Support Vector Machine, machine learning**

I. INTRODUCTION

To Data mining is widely used in real world applications. It is the discipline where historical data is analyzed to obtain hidden information. In other words, it is the process of extracting or discovering latent trends or patterns that are not known earlier. These trends or patterns uncovered from the databases are used to take expert decisions. The process of mining is essential for any enterprise in different domains. Knowledge discovery helps domain experts to have interpretation of knowledge and take decisions. Models are built in order to have solutions to different problems. The general steps involved in knowledge discovery from databases (KDD) are visualized in Figure 1.

There are many steps in KDD. First of all a problem is defined. Then data is gathered in order to solve the problem. Then data mining algorithms are used to built a model and evaluate it. This gives rise to knowledge needed. This knowledge is used to make expert decisions that result in business growth and profits. There are many algorithms related to data mining. They include association rule mining, decision trees, clustering and classification. These algorithms take time and resources to complete mining process. When high dimensional data is taken, these algorithms take long time to execute and consume more resources. To overcome this problem, it is important to reduce dimensions.

Many existing data mining based intrusion methods do not use feature selection method. For instance neural networks and SVM based approach [7], ANN and fuzzy clustering [10], SVM based approach [12], and fuzzy logic based approach [17] and Hidden Naive Bayes method [18]. There are some methods found with feature selection. They include Naive Bayes based method [15], Mutual information based intrusion detection [22] and [24] where many feature selection algorithms are reviewed. However, it is understood that feature selection is still an optimization problem which leads to further enhancement in accuracy and performance of data mining techniques for intrusion detection. Our contributions are as follows.

1. We proposed a framework named SVM-FSS for feature selection based intrusion detection that enhances the capability of SVM.
2. We proposed an algorithm named FSS for effective feature selection prior to employing classification technique on intrusion datasets like KDD and NSL-KDD.
3. We built an application to show the effectiveness of the framework and evaluated with the two datasets.

The remainder of the paper is structured as follows. Section 2 provides literature review on data mining techniques that are used for detecting network intrusions. Section 3 covers the proposed methodology for intrusion detection. Section 4 presents experimental results and evaluation while Section 5 provides conclusions and gives possible scope for future work.

Revised Manuscript Received on 22 May 2019.
***** Correspondence Author
Dr. Ch. Mallikarjuna Rao*, Professor, Department of CSE, GRIET, Hyderabad, Telangana, India.
Dr. G. Ramesh, Associate Professor, Department of CSE, GRIET, Hyderabad, Telangana, India.
Dr. D. V. Lalitha Parameswari, Sr.Asst. Professor, Department of CSE, GNITS, Hyderabad, Telangana, India.
Dr. Karanam Madhavi, Professor, Department of CSE, GRIET, Hyderabad, Telangana, India.
Mr. K. Sudheer Babu, Assistant Professor, Department of CSE, GRIET, Hyderabad, Telangana, India.

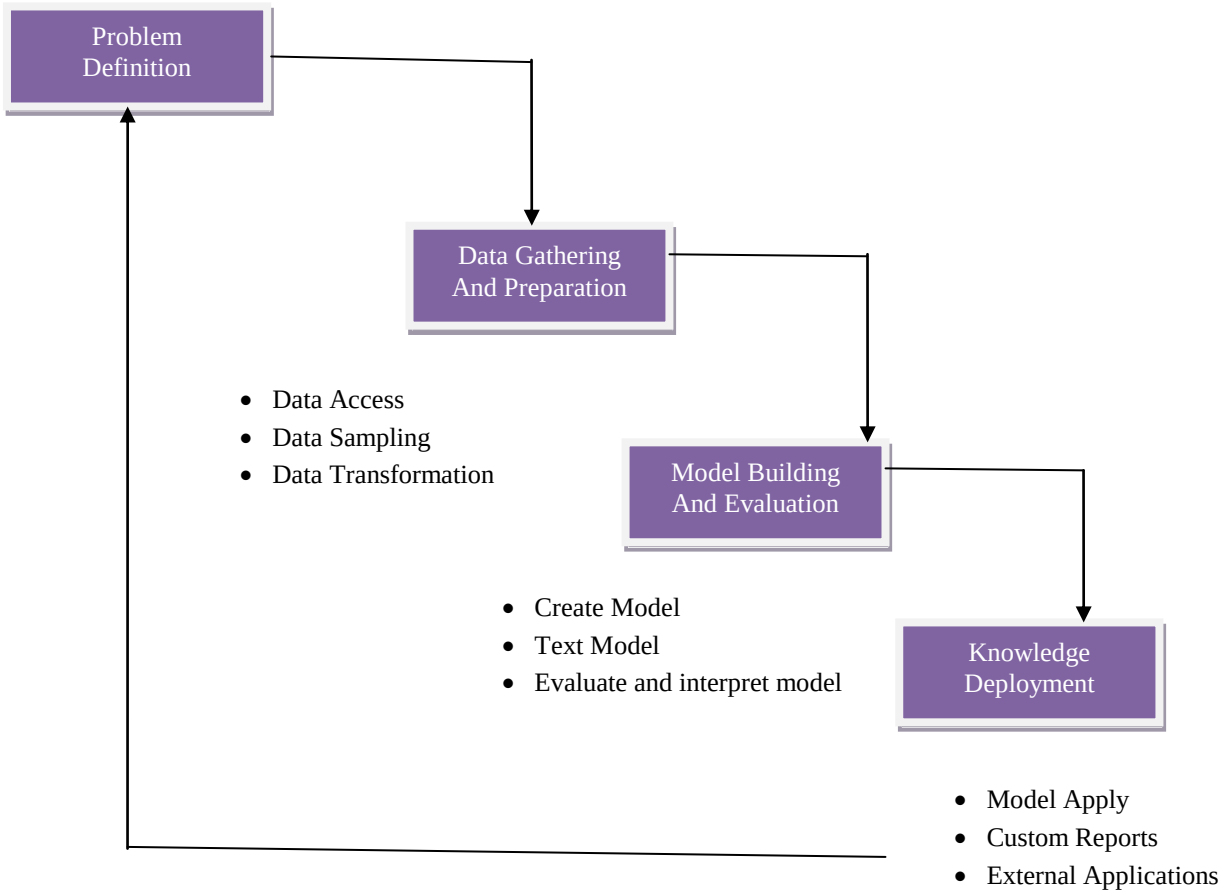

Figure 1: Steps in data mining

II. RELATED WORK

This section reviews literature on intrusion detection techniques. Garcia-Teodoro *et al*. [1] reviewed anomaly based mechanisms for network intrusion detection besides focusing on the demerits of the techniques. Dokas *et al*. [2] and Bloedorn *et al*. [4] on the other hand employed data mining techniques to detect intrusions and found they are to be efficient when compared with SNORT. Intrusion detection with fusion method meant for protecting cyberspace is the main focus in [3]. Lee *et al*. [5] explored real time approach for intrusion detection using data mining techniques with a novel architecture. Their research was limited to network level intrusion detection and does not consider application level.

Barbara *et al*. [6] proposed a testbed for data mining techniques for intrusion detection. It is named as Audit Data Analysis and Mining (ADAM). It could be used for rapid development of intrusion detection methods. Mukkamala *et al*. [7] used the combination of Support Vector Machine (SVM) and Neural Networks for developing an intrusion detection system and evaluated with KDD dataset. Classification models with supervised learning are used in [8] for intrusion detection. Intrusion detection research associated with cloud computing is made in [9] and [20] while Wang *et al*. [10] focused on the fuzzy clustering and Artificial Neural Network (ANN) for making intrusion detection mechanism. It is called FC-ANN.

Liao *et al*. [11] studied data mining techniques used in different applications including intrusion detection. They found it to be useful as learning new possibilities is done with data mining. Bhavsar and Waghmare [12] developed an intrusion detection system using SVM. A survey of anomaly detection mechanisms based on data mining is found in [13]. Sommer and Paxson [14] focused on network intrusion detection using machine learning methods. Naive Bayes classification techniques is employed in [15] with feature selection for effectiveness in intrusion detection. Fuzzy logic is employed in [16] for network intrusion detection with the help of a set of rules. Hoque *et al*. [17] employed Genetic Algorithm (GA) for reducing complexity in the detection process.

Hidden Naive Bayes technique with multiple classes is employed in [18] for intrusion detection. The concept of neural visualization method is employed in [19] for visualizing network traffic and find intrusions with ease. Berthier *et al*. [21] developed an intrusion detection mechanism for advanced metering infrastructures. Amiri *et al*. [22] proposed information-based feature selection method for improving performance of LS-SVM. Many feature selection methods are explored in [24]. A review of Artificial Intelligence (AI) methods for intrusion detection is made in [23]. Hajian *et al*, Y.A.Siva Prasad et.al [25][27][28] studied the mechanisms to prevent discrimination in the research of intrusion and crime detection. From the review of literature it is found that feature selection is an optimization problem that is open for new avenues. In this paper we proposed a new

feature selection method for enhancing accuracy of data mining based intrusion detection.

III. PROPOSED INTRUSION DETECTION METHODOLOGY

This section provides the proposed intrusion detection methodology based on machine learning and feature subset selection. Well known intrusion datasets such as KDD and NSL-KDD are used for empirical study. The following sub sections provide the methodology in detail.

A. *Problem Definition*

High dimensional datasets throw the problem of curse of dimensionality. There might be some redundant features or irrelevant features that are not compatible for given objective function. Therefore, it is essential to have an effective feature subset selection to minimize computational complexity and also enhance performance of machine learning algorithms in terms of accuracy. If this problem is not solved with feature subset selection, it may render drastically deterioration of machine learning algorithms or even failure of them to perform intended functionality. This is the challenging problem considered.

B. *Framework*

The methodology includes a framework with training and testing phases as shown in Figure 2. Since the methodology is based on supervised learning, learning a classifier needs training set. Therefore training data sets of KDD and NSL-KDD are used. They are pre-processed in order to handle any missing values. Then the training sets are subjected to feature subset selection. Section 3.3 provides the proposed algorithm for the same. The Feature Subset Selection (FSS) algorithm works on the training set and finds a subset of features that are essential for the objective in hand. It returns the features that are required for learning classifier. Therefore, the algorithm hands over the selected features to SVM classification algorithm. This algorithm learns and builds a classifier (knowledge model) that is used to perform classification of unlabelled instances (testing set). The whole process is therefore divided into training and testing phases. In the former, a classifier is trained with the training set. In the latter, the classifier that has been built in training phase is used to perform classification. Hence it is known as supervised learning method. The proposed framework includes SVM and FSS. It is known as SVM-FSS framework.

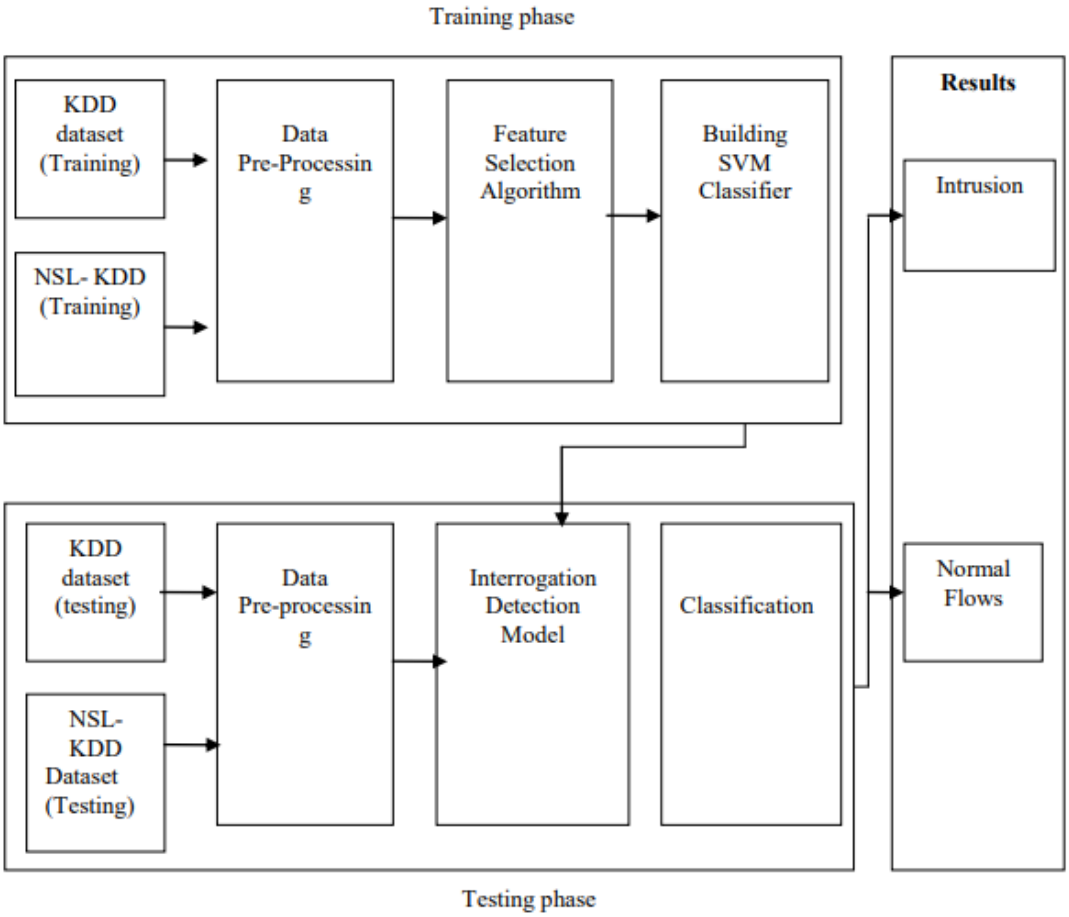

Figure 2: Proposed intrusion detection methodology

The classification results in identifying class labels to unlabelled instances. The two important class labels are intrusion and normal flows. SVM is the binary classifier by default which makes two classes based on the training given to it. Normal flows indicate that the network flows that did not carry any malicious attacks.

Intrusion flows mean that, the network traffic carried malicious content that could lead to damage of information systems or cause potential risk to critical digital infrastructure of a company or country.

Thus the proposed framework helps in identifying cyber attacks and provides possible prevention to it once business intelligence is garnered with the enhanced functionality of the machine learning algorithm SVM with the proposal of FSS algorithm.

C. *Feature Subset Selection Algorithm*

This algorithm is crucial to improve intrusion detection performance of SVM. With this algorithm, the proposed machine learning framework is known as SVM-FSS. The algorithm takes given dataset as input and produces chosen features that are used by the SVM algorithm for building a classifier.

Algorithm: Feature Subset algorithm
Inputs: Dataset ***D***
Outputs: Selected Features F

Relevant Feature Extraction
01 For each feature ***f*** in ***D***
02 Calculate entropy
03 Calculate gain
04 IF ***entropy*** and ***gain*** satisfy threshold
05 Add ***f*** to ***F***
06 END IF
07 End For
Tree construction
08 For each feature ***f*** in the selected features ***F***
09 Add ***f*** to ***T***
10 End For
Feature Selection
11 Empty ***F***
12 For each feature ***f*** in the tree ***T***
13 Get feature pair
14 IF there is correlation among them THEN
15 Add feature ***f*** to ***F***
16 END IF
17 End For
18 Return ***F***

Algorithm 1: Feature subset selection algorithm

This algorithm is meant for obtaining relevant features from the given dataset. By finding correlations among the features, it reduces dimensionality and final selection of features result in the feature set that satisfies given objective for which mining is carried out.

Reduction of dimensionality makes the proposed SVM-FSS framework superior to its predecessors in terms of classification accuracy. The gain and entropy computed based on the network flows in the dataset are used to have decision making with a threshold. Entropy is computed as in Eq. (1) while the gain is derived from the Eq. (2).

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (1)$$

$$\text{Gain}(X/Y) = H(X) - H(X/Y) \quad (2)$$

$$= H(Y) - H(Y/X)$$

Uncertainty in the dataset (network flows) is determined using entropy. Gain is derived from entropy as in Eq. (2) which reflects expected reduction of entropy. These two are related measures that help in making well informed decisions related to correlation of features before making a subset of features that satisfy given objective.

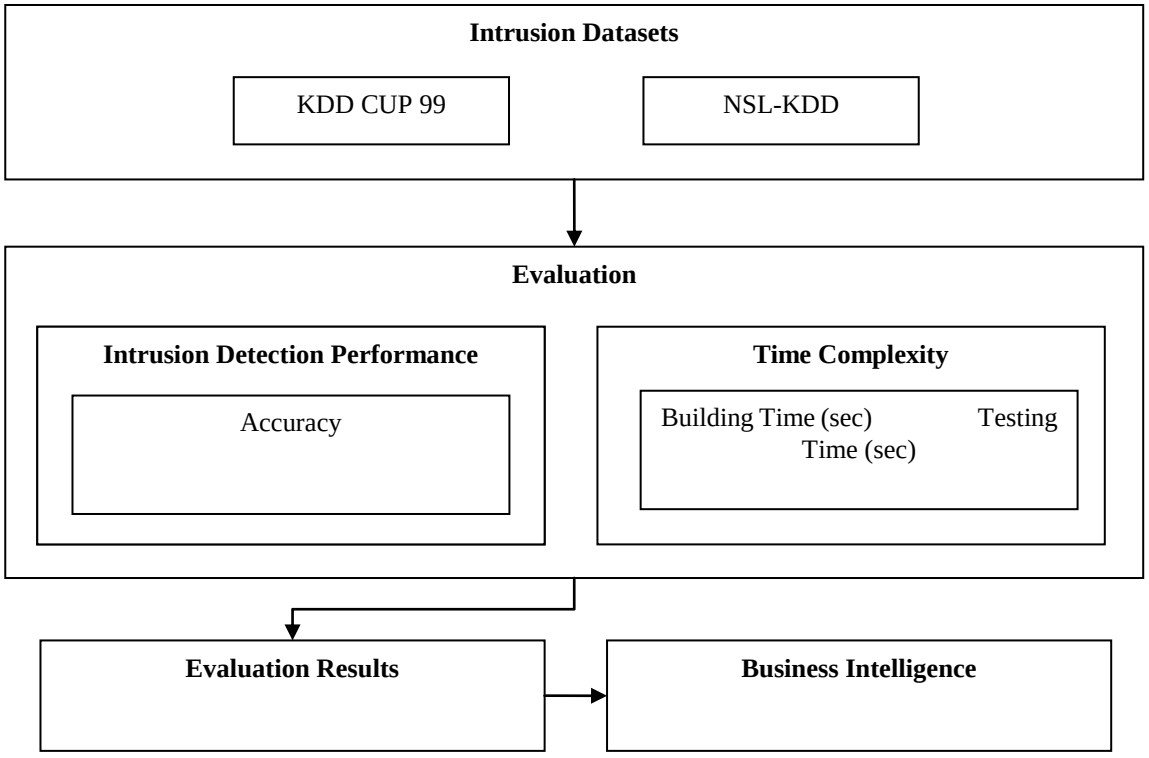

Figure 3: Evaluation procedure

D.*Evaluation*

Evaluation procedure followed in this paper is presented in Figure 3. Time complexity and intrusion detection accuracy are two important measures used for evaluation. The evaluation results obtaining business intelligence that may help network administrators to make strategic decisions.

The evaluation procedure includes the usage of accuracy measure. This measure needs values related to True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The meaning of these values is presented in Table 1 in the form of confusion matrix.

Table 1: Illustrates confusion matrix

	Ground Truth (correct prediction of intrusions)	Ground Truth (wrong prediction of intrusions)
Result of SVM-FSS (correct prediction of intrusions)	True Positive (TP)	False Positive (FP)
Result of SVM-FSS (wrong prediction of intrusions)	False Negative (FN)	True Negative (TN)

Based on the results of the proposed framework, it is possible to derive a measure known as accuracy or detection accuracy. It is computed as in Eq. (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (3)$$

Accuracy of classification algorithm SVM-FSS is recorded for the two datasets and compared with the results presented for state of the art classification method explored in [26]. It does mean that the proposed framework SVM-FSS is compared with the intrusion detection framework presented in [26].

IV. EXPERIMENTAL RESULTS

This section provides experimental results in terms of building time and testing time of the proposed framework against KDD and NSL-KDD datasets. The result also includes accuracy of the SVM-FSS and the other framework LSSVM-IDS-FMIFS [26].

Table 2: Building time and testing time comparison

	Algorithms	KDD	NSL-KDD
Building Time (s)	LSSVM-IDS + FMIFS	80	50
	SVM-FSS	73.7	46.5
Testing Time (s)	LSSVM-IDS + FMIFS	30	25
	SVM-FSS	28.9	23.8

As shown in Table 2, the performance of the proposed framework in training and testing phases is presented along with that of state of the art. The building time and testing time of the proposed and existing systems are presented for KDD and NSL- KDD.

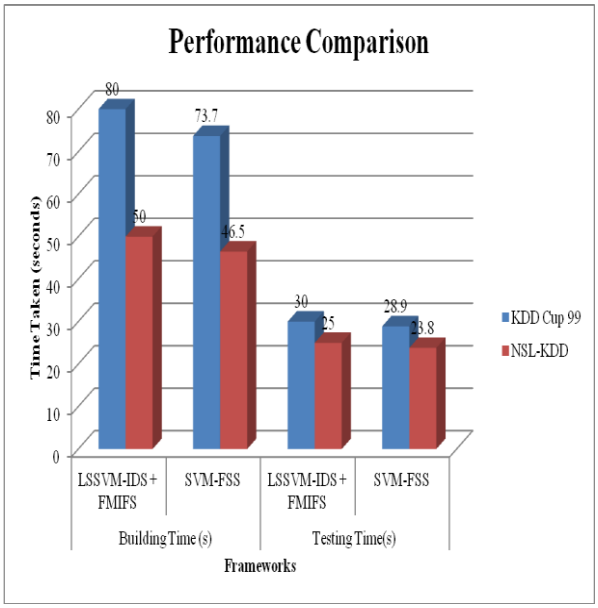

Figure 3: Performance comparison

As presented in Figure 3, it is evident that the proposed framework SVM-FSS and its predecessor are shown in X axis while the time taken by them for building and testing is shown in Y axis. The results reveal for both KDD Cup 99 and NSL-KDD. It is understood that for both datasets the proposed SVM-FSS has shown better performance over LSSVM-IDS+FMIFS. The feature subset selection based on entropy and gain extracted from the network flows is the important contribution to the speed of the classification algorithm.

Table 3: Performance in terms of accuracy

DATASETS	ACCURACY (%)	
	SVM-FMIFS	SVM-FSS
KDD	99.79	99.81
NSL-KDD	99.911	99.93

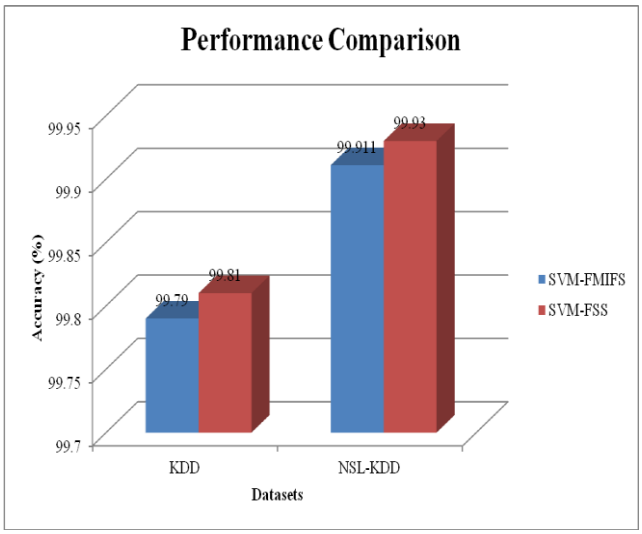

Figure 4: Performance in terms of classification accuracy

As presented in Figure 4, it is evident that the proposed framework SVM-FSS and its predecessor are compared in terms of classification accuracy.

X axis shows benchmark datasets used for experiments while the accuracy shown by the algorithms is shown in Y axis. The results reveal for both KDD Cup 99 and NSL-KDD. It is understood that for both datasets the proposed SVM-FSS has shown better performance over LSSVM-IDS+FMIFS. The feature subset selection based on entropy and gain extracted from the network flows is the important contribution to the accuracy of the classification algorithm.

The experiments made with the two benchmark datasets provide indications that feature subset selection has its advantages. However, there might be many threats to validity of this proposition. The first one is that the datasets are collected from Internet sources and that may have limitations when compared with the novel approaches being followed by intruders. The second threat to validity is that the proposed mechanisms are tested for intrusion dataset. It may work for other domains as well. However it cannot be generalized to be a common classification mechanism for different domains unless it is further evaluated.

V.CONCLUSIONS AND FUTURE WORK

Data mining algorithms with supervised learning are very useful for prediction of class labels. Such algorithms need two phases of functionality known as training and testing. Generally training is made offline while testing is made online. In this paper we considered network intrusion detection as the case study to demonstrate proof of the concept for the proposed supervised learning based framework. We proposed a framework for intrusion detection. The framework is known as SVM-FSS. A feature selection algorithm known as Feature Subset Selection (FSS) is proposed to work in tandem with the well known classification algorithm SVM. Our contribution to improve speed and accuracy of classification is the introduction of FSS which has potential to improve performance of any classification algorithm. FSS selects important and relevant features to reduce time and space complexity that will reflect in the speed and accuracy of classification mechanism.

We used benchmark datasets like KDD and NSL-KDD for the empirical study. The proposed framework SVM-FSS showed better performance over a state of the art classification framework found in the literature. In future we intend to improve our feature subset selection mechanism with a hybrid approach where both filter and wrapper approaches are employed for performance enhancement.

REFERENCES

1. P. Garcıa-Teodoroa, J. Dıaz-Verdejoa , G. MaciaFernandeza , E. Vazquezb. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. computers& security. 28, p18-28.
2. Paul Dokas, LeventErtoz, Vipin Kumar, AleksandarLazarevic, Jaideep Srivastava, Pang-Ning Tan. (2002). Data Mining for Network Intrusion Detection . IEEE, p1-10.
3. Tim Bass. (2000). Intrusion Detection Systems &Multisensor Data Fusion: Creating Cyberspace Situational Awareness. Communications Of The Acm - Accepted For Publication (Draft), p1-6.
4. Eric Bloedorn, Alan D. Christiansen, William Hill, Clement Skorupka, Lisa M. Talbot, Jonathan Tivel. (2001). Data Mining for Network Intrusion Detection: How to Get Started. IEEE, p1-9.
5. Wenke Lee , Salvatore J. Stolfo , Philip K. Chan , EleazarEskin , Wei Fan , Matthew Miller , ShlomoHershkop , and Junxin Zhan. (2001). Real Time Data Mining-based Intrusion Detection. IEEE, p1-13.
6. Daniel Barbard, Julia Couto, SushilJajodia, NingningWu . (2001). Adam: A Testbed for Exploring the Use of Data Mining in Intrusion Detection . IEEE. 30 (4), p1-10.
7. SrinivasMukkamala, Guadalupe Janoski, Andrew Sung . (2002). Intrusion Detection Using Neural Networks and Support Vector Machines . IEEE, p1-6.
8. Ramesh Agarwal, Mahesh V. Joshiy. (2008). PNrule: A New Framework for Learning Classier Models in Data Mining (A Case-Study in Network Intrusion Detection). IEEE, p1-22.
9. ChiragModi, Dhiren Patel, Hiren Patel, BhaveshBorisaniya, Avi Patel, MuttukrishnanRajarajan. (2013). A survey of intrusion detection techniques in Cloud . IEEE, p1-14.
10. Gang Wang, Jinxing Hao, Jian Ma, LihuaHuang . (2010). A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. Expert Systems with Applications. 37, p6225-6232.
11. Shu-HsienLiao , Pei-Hui Chu, Pei-Yuan Hsiao. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. Expert Systems with Applications. 39, p11303-11211.
12. Yogita B. Bhavsar, KalyaniC.Waghmare. (2013). Intrusion Detection System Using Data Mining Technique: Support Vector Machine . International Journal of Emerging Technology and Advanced Engineering. 3 (3), p1-6.
13. Shikha Agrawal, JitendraAgrawal . (2015). Survey on Anomaly Detection using Data Mining Techniques . 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems . 60, p708-715.
14. Robin Sommer,VernPaxson. (2010). Outside the Closed World: On Using Machine Learning For Network Intrusion Detection. IEEE Symposium on Security and Privacy, p1-12.
15. Dr. SaurabhMukherjeea ,NeelamSharmaa . (2012). Intrusion Detection using Naive Bayes Classifier with Feature Reduction .ProcediaTechnology .4 , p119-128.
16. R. Shanmugavadivu ,Dr.N.Nagarajan . (2011). Network Intrusion Detection System Using Fuzzy Logic . Indian Journal of Computer Science and Engineering. 2 (1), p1-11.
17. Mohammad SazzadulHoque, Md. Abdul Mukit and Md. Abu NaserBikas. (2012). An Implementation Of Intrusion Detection System Using Genetic Algorithm. International Journal of Network Security & Its Applications . 4 (2), p1-12.
18. LeventKoc, Thomas A. Mazzuchi, ShahramSarkani. (2012). A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier. Expert Systems with Applications. 39, p13492-13500.
19. Emilio Corchado , And Álvaro Herrero. (2011). Neural Visualization of Network Traffic Data for Intrusion Detection . IEEE, p1-53.
20. Ahmed Patel, Mona Taghavi ,KavehBakhtiyari , JoaquimCelestinoJúnior. (2013). An Intrusion Detection And Prevention System In Cloud Computing: A Systematic Review. Journal of Network and Computer Applications. 3 (1), p25-41.
21. Robin Berthier, William H. Sanders, and HimanshuKhurana . (2010). Intrusion Detection for Advanced Metering Infrastructures: Requirements and Architectural Directions, IEEE, p1-6
22. FatemehAmiri, MohammadMahdiRezaeiYousefi , Caro Lucas , AzadehShakery, Nasser Yazdani . (2011). Mutual information-based feature selection for intrusion detection systems. Journal of Network and Computer Applications. 34, p1184-1199.
23. GulshanKumar ,Krishan Kumar , Monika Sachdev. (2010). The use of artificial intelligence based techniques for intrusion detection: a review. IEEE, p1-20.
24. HuanLiu,HiroshiMotoda,RudySetiono,Zheng Zhao. (2010). Feature Selection: An Ever Evolving Frontier in Data Mining. IEEE, p1-10.
25. Sara Hajian, Josep Domingo-Ferrer and AntoniMart´ınez-Balleste. (2011). Discrimination Prevention in Data Mining for Intrusion and Crime Detection. IEEE, p1-8.
26. Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi Nanda and Zhiyuan Tan (2014). Building an intrusion detection system using a filter-based feature selection algorithm. IEEE Transactions on Computers, p1-13.
27. Y.A.Siva Prasad, and G. Ramakrishna. "A NOVEL PROBABILISTIC BASED FEATURE SELECTION MODEL FOR CREDIT CARD ANOMALY DETECTION." Journal of Theoretical & Applied Information Technology 94.2 (2016).

28. Y.A.Siva Prasad, and G. Ramakrishna" Distributed differential privacy preserving mechanism on real time datasets", International Journal of Applied Engineering Research,(2015).

AUTHORS PROFILE

Dr. Ch. Mallikarjuna Rao received the B. Tech degree in Computer Science and Engineering from Dr. Baba Sahib Ambedkar Marathwada University, Aurangabad, Maharastra in 1998, M.Tech Degree in Computer Science and Engineering from JNTU Ananthapuramu, Andhrapradesh in 2007 and Ph.D in Computer Science and Engineering from JNTU , Ananthapuramu, Andhrapradesh in 2016 Currently he is working in " Gokaraju Rangaraju Institute of Engineering and Technology", Hyderabad. Telangana , India. His area of Interests are Data Mining, Bigdata and Software Engineering.

Dr. G. Ramesh is currently working as an Associate Professor in Department of Computer Science & Engineering, GRIET, Hyderabad. He received his B.Tech degree in information technology from RGMCET, Nandyal, Kurnool Dist. Andhra Pradesh, and He received his M. Tech degree in software engineering from JNTUA College of Engineering, Ananthapuramu, Andhra Pradesh, India, He received his PhD from JNTUA, Anatapuramu, Andhra Pradesh, India. His main research interest includes software engineering and big data. He is in teaching since 2008. He has published a several papers in various international journals/ conferences. He is life member in IAENG.

Dr. K.Madhavi, working as a Professor in Computer Science and Engineering Department, Gokaraju Rangaraju Institute of Engineering and Technology. She has completed her B.E in 1997 from Kuvempu University, M.Tech from JNTUA in 2003 and awarded Ph.D from JNTUA in 2013. She has 20 years of teaching experience. She has published several papers in reputed international journals and international conferences. Her research interest includes software engineering, Model Driven Engineering, Data Mining, and Mobile Software Engineering.