International Journal of Scientific Research and Management Studies (IJSRMS)
ISSN: 2349-3771 **Volume 1 Issue 5, pg: 183-191**

A NOVEL HYBRID INTRUSION DETECTION FRAMEWORK BASED ON DATA MINING TECHNIQUES

Maryam Hajizadeh and Marzieh Ahmadzadeh
School of Computer Engineering & IT, Shiraz University of Technology, Shiraz, Iran

ABSTRACT

With rapid expansion of computer networks in recent years, network security has gained a growing importance. In addition, almost all computer systems suffer from security vulnerabilities. Therefore Intrusion Detection Systems are designed in order to protect system or network from potential attacks. To decide whether events and activities occurring in computer systems are intrusive or legitimate, Intrusion Detection System widely relies on data mining techniques which are defined as methods for extracting useful knowledge from large amount of network data. In this paper, a new hybrid approach is proposed for intrusion detection framework by combining two data mining algorithms called K-Means clustering and Naïve Bayes classification. NSL-KDD dataset is utilized for building a model and verifying it. Since input dataset consists of noisy, irrelevant and redundant features, a Gain Ratio measure is employed for reducing dataset and achieving proper data. This paper employs SSE measure to specify proper number of clusters for this experiment. Using K-Means clustering, the whole dataset is divided into corresponding clusters, afterwards each group is fed to Naïve Bayes algorithm as input datasets in order to classify each instance into one of five general categories which are Normal, DoS, Probe, R2L and U2R. As illustrated in this article, Results and analyses show that the proposed framework performed better in terms of accuracy and achieved high true positive rate and very low false positive rate for five class values in comparison with a simple Naïve Bayes approach.

KEYWORDS: *K-Means, Naïve Bayes, Accuracy, True Positive Rate, False Positive Rate*

I. INTRODUCTION

Since network attacks have increased in number and severity in the recent years, Intrusion Detection Systems (IDSs) are becoming an important impartible part of network and information security architecture, which monitor the network traffic and analyze them for detecting security attack. An IDS is a sensor that raises an alarm if specific things occur [7]. While an IDS is also capable of sending early alarms upon risk exposure caused by any attack, at the same time it has the potential to generate high volume of false alarms. Hence, the objective of IDS is to identify potential attack with high true positive rate and less false positive rate [5]. The existing IDS systems based on data mining have focused on the feature selection techniques because of their high efficiency in reducing the false positive rate [2].

Data mining techniques have been extensively applied to intrusion detection in recent years to detect normal and abnormal behaviors in large network datasets. Two important techniques of data mining which have been widely used are clustering and classification. In general, classification attributes each instance to a set of predefined classes. An example of it is the Naïve Bayes which is a heavily simplified version of Bayesian probability model and due to simple structure; it can produce highly competitive results in detecting anomaly-based network intrusion [4]. Clustering is an unsupervised learning method which tries to divide dataset into sub sets based on their common properties. Clustering tools like K-Means could be used to efficiently identify a group of traffic behaviors that are similar to each other using cluster analysis [3].

The KDDCUP'99 is an intrusion detection contest dataset [6] first given by Massachusetts Institute of Technology. The huge number of redundant instances is one of the most important deficiencies in the

KDDCUP'99 dataset. By selecting specific instances of KDDCUP'99, this problem is solved in NSL-KDD [13] which is considered as a modified version of KDDCUP'99. More details about the inherent problems found in KDDCUP'99 dataset can be obtained from [12]. The NSL-KDD dataset contains one type of normal data and 21 different types of attacks which fall into one of four categories. These are DoS, Probe, R2L, and U2R. Denial of Service (DoS) is an attack type in which the attacker makes specific machine resources unavailable or too busy to answer to legitimate users requests. Probing attack is an attempt to collect useful information about the target host. Remote to Local (R2L) occurs when an attacker tries to gain access because he does not have an account on the victim machine and User-to-Root (U2R) attacks are attempts by a non-privileged user to gain administrative privileges [12].
In this paper, a novel hybrid framework is presented which is a combination of K-Means clustering and Naïve Bayes classifier. Because K-Means requires proper number of clusters, we describe how it is estimated by utilizing SSE measure. Moreover, Gain ratio measure is used to find an optimal subset of features from NSL-KDD dataset instances. Finally, proposed framework can classify instances into one of the five major classes: Normal, DOS, Probe, R2L and U2R.
The rest of the paper is structured as follows: Section II summarizes the related work of different researchers. In Section III the proposed framework is introduced in details. The experimental analysis is shown in Section IV. Finally the conclusion is presented in section V.

II. RELATED WORK

Since evaluation results of the IDSs based on KDDCUP'99 was unreliable due to its problems, Tavallaee et al. [12] conducted a statistical analysis on this dataset and proposed a new dataset named NSL-KDD. They also demonstrated the use of the seven data mining algorithms such as J48 decision tree, Naïve Bayes and etc. on their proposed dataset. Although result indicates better performance of these algorithms when applied to NSL-KDD in comparison with KDDCUP'99 in term of accuracy, they have not tried further any feature selection measures to improve the accuracy.
Mukherjee et al. [10] has proposed a feature vitality based reduction method (FVBRM) which is applied on NSL-KDD dataset in order to identify a reduced set of important input features. In this work, by using Naïve Bayes algorithm each instance of the reduced dataset is classified into one of the five categories: Normal, DOS, Probe, R2L and U2R. FVBRM has increased Naïve Bayes classifier accuracy compared to Tavalaee's attempt, which is an indication of the importance of feature selection in building effective and computationally efficient IDSs.
Another methodology that is applied on NSL-KDD is Decision Table/Naïve Bayes (DTNB) proposed by Azad et al. [1]. DTNB is a hybrid classifier that is used to identify possible intrusions which also utilized feature selection procedure. Unlike FVBRM, DTNB classifier considered each instance in the natural specific class values instead of considering five general categories. However, DTNB achieved competitive accuracy when compared to the FVBRM method.
In order to provide the analysis of NSL-KDD, Kumar et al [9] categorized 21 different type of attacks into four groups using simple K-Means clustering. The main benefit of this task is the ability to detect novel type of attacks without any prior notice and capability to find natural grouping of data, based on similarities among the patterns.

III. PROPOSED FRAMEWORK

Proposed hybrid framework is based on the combination of two methods of data mining i.e. clustering and classification. Hence, we utilized K-Means clustering and Naïve Bayes classifier, respectively. These two steps are the core of proposed framework. Evaluation tests on proposed approach are repeated 20 times to ensure the reliability of obtained results. This section described each step of the proposed framework in details.

3.1 Choosing Dataset

First step to perform each data mining task is choosing the proper dataset. Although KDDCUP'99 is widely used in the field of network intrusion detection, but in this paper, NSL-KDD dataset is used due to its certain advantages [12]. Unlike approaches which use KDDCUP'99, our proposed framework will not be biased toward more frequent instances and its true positive rate will be ensured because of

the fact that NSL-KDD does not include redundant instances. The evaluation experiments are conducted on 20% of the NSL-KDD dataset which consisted of 25192 instances among which 13449 are normal and 11743 are attacks. Each instance of mentioned dataset represents a separate connection.

3.2 Preprocessing

3.2.1 Generalization of Class Value

NSL-KDD dataset contains 41 fields as features and 42nd field as a class variable that labels the instance as normal or one type of attack. Since NSL-KDD have 21 different attacks type, it was very inconvenient to assess the performance of the proposed framework. On the other hand each attack type falls into one of the four major categories of networking attacks. Hence the attack labels are generalized to their respective categories for the ease of analysis. Finally five general categories are formed as the class values i.e. DoS, Probe, R2L, U2R and Normal.

Input: T = {t_1, t_2, t_3… t_n}
DoS = {Neptune, Smurf, Back, teardrop, Pod, Land}
Probe = {Ipsweep, Satan, Portsweep, Nmap}
R2L = {Warezclient, Imap, Guess_Passwd, Warezmaster, Multihop, Phf, Ftp_write, Spy}
U2R = {Buffer_overflow, Rootkit, Loadmodule}
Begin
For each instance $t_i \in$ T
if (t_i.class ∈(one of the four attack sets))
assign that set's name to t_i.class as a value
else
assign Normal value to t_i.class
End

3.2.2 Discretization

In order to improve the performance of the proposed framework, discretization technique is employed to turn numeric features into nominal features. Discretization is the supervised feature filtering which converts the numeric values into a small number of distinct ranges for the sake of producing a better model [8].

3.3 Feature Selection

Since most of the typical dataset consists of noisy features, data mining tasks require to remove features from the original dataset that don't follow certain criteria, otherwise, knowledge discovery during the training phase is more difficult. Feature selection which is proposed for this goal is a subset of dimension reduction. Hence in this framework, Gain Ratio has been employed to rank the attributes of high dimensional NSL-KDD datasets. The low ranking attributes are filtered to form new reduced dataset. In order to reach a proper dimension reduction, first step is to calculate Gain Ratio for each of the 41 features, and then based on the results they are rated. Finally features which have higher rank than the rest of others have been chosen which leads to increase in model accuracy. Table 1 shows the 30 features selected using Gain Ratio measure.

Table 1. Optimum subset selection using Gain Ratio

duration	serror_rate	srv_diff_host_rate
service	rerror_rate	num_failed_logins
urgent	diff_srv_rate	dst_host_srv_count
count	protocol_type	dst_host_serror_rate
root_shell	is_guest_login	dst_host_diff_srv_rate
logged_in	wrong_fragment	dst_host_same_srv_rate
src_bytes	same_srv_rate	dst_host_srv_rerror_rate
dst_bytes	dst_host_count	dst_host_srv_serror_rate
flag	srv_rerror_rate	dst_host_srv_diff_host_rate
hot	srv_serror_rate	dst_host_same_src_port_rate

3.4 Clustering

In this framework clustering is an interesting approach for our purpose to partition a NSL-KDD into several groups according to a similarity with larger value in a group than other groups. Hence, K-Means clustering is employed in order to provide better distribution of instances to their categories as presented in the NSL-KDD. K-Means treats each instance in dataset as a point having a location in space. Each cluster in the grouping is defined by its member points, x, and by its center. The center for each cluster, m_i, is the point to which the sum of distances from all points in that cluster is minimized.

3.4.1 Proper Value Selection for K

The advantage of the K-Means clustering is its favorable execution time. But one of the greatest challenges related to using it is specifying the number of clusters (k). If the whole dataset is divided into very small number of clusters, instances existing in each cluster are far apart. On the other hand, very large values for k results in instances which are close to each other placed in different clusters. Hence, estimating appropriate number of clusters has significant importance. Therefore, we use common measure called Sum of Squared Error (SSE) to solve this problem, according to mentioned in [11]. SSE is measured by the following formula:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

In order to find the best grouping of dataset with corresponding proper SSE, K-Means clustering with various values of k parameter as an input is applied on the reduced dataset for several times. At each run the number of clusters has been increased to specify whether K-Means can find a better grouping of the dataset. SSE is evaluated for each grouping corresponding to initial k values, as illustrated in Figure 1. Lower SSE means better clustering. As the number of clusters increases, the SSE should decrease because clusters are smaller. Hence in this framework the knee of the SSE curve is considered as an appropriate value for k i.e. it indicates that 10 clusters are better separated.

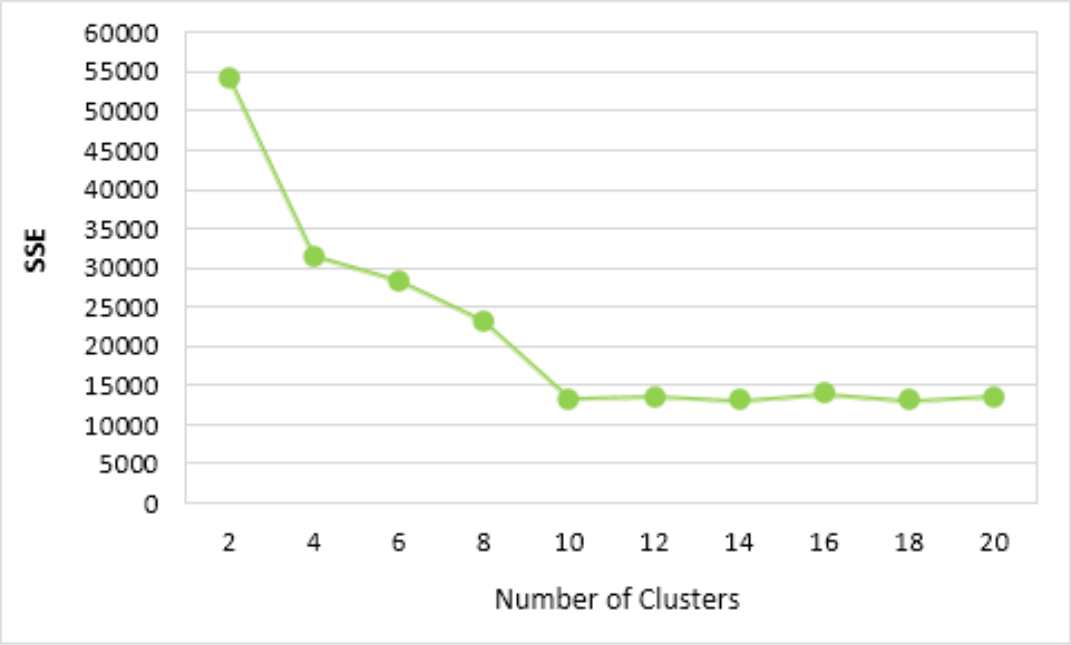

Figure 1. SSE versus number of clusters for the reduced NSL-KDD

3.4.2 Apply K-Means 20 Times

The first cluster center is chosen randomly with uniform distribution from the points that are being clustered which is another challenge for K-Means. Due to randomness basis of the K-Means clustering and the need to ensure the accuracy of obtained results, this phase is repeated 20 times. In each repetition the experiment runs with k=10 and reduced data as its inputs. So original dataset is converted into ten clusters i.e. ten distinct datasets included instances that are more similar compared with original dataset.

3.5 Classification

In this framework each of ten datasets are given to Naïve Bayes in order to build a learning model and validate it. 66% of the dataset is allocated to the training set and the remaining is allocated to the testing set. Since experiment in previous step was repeated 20 times, subsequently this step is repeated 20 times as well.

3.6 Performance Evaluation

In order to check the performance of our proposed approach and compare it with Naïve Bayes, three metrics are being used: Accuracy, True Positive Rate (TPR), and False Positive Rate (FPR). They are measured by the following formulas:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

Where True Positive (TP) is the number of test instances which are correctly classified as positive, False Negative (FN) indicates the number of test instances incorrectly classified as negative; False Positive (FP) is the number of instances which are incorrectly classified as positive and finally, True Negative (TN) which is the number of test instances correctly classified as negative.

IV. EXPERIMENTAL ANALYSIS

The IDS requires high accuracy and TPR as well as low FPR. Hence the proposed framework (Proposed FW) is compared against Naïve Bayes which is applied on NSL-KDD with 41 features (NB+41Feature) and Naïve Bayes which is applied on reduced NSL-KDD with 30 features (NB+30Features) based on the previously mentioned metrics. To guarantee the reliability of the achieved results, experiment is repeated 20 times and a 95% confidence interval is constructed around the mean value for Accuracy, TPR and FPR metrics. As illustrated in Figure 2 and Table 2, experimental results indicate improvement of accuracy for the Proposed FW by 13.65% and 9.45% in comparison with a NB+41F and NB+30F, respectively.

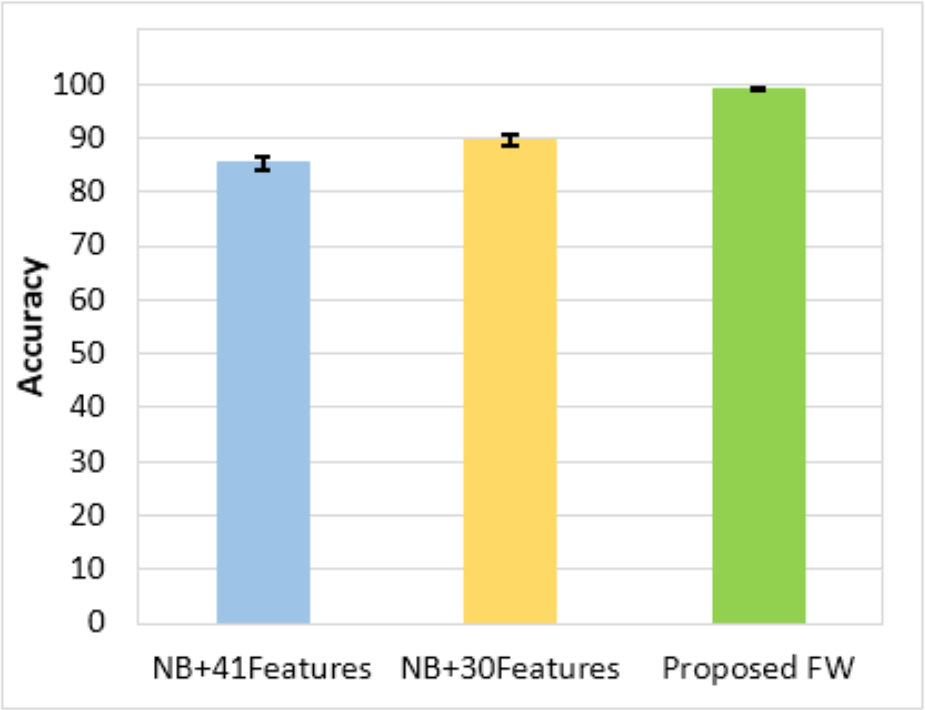

Figure 2. Accuracy comparison

Table 2. Accuracy comparison in detail

	Accuracy Mean ± CI	Max	Min
Proposed FW	98.96 ± 0.15	99.32	98.21
NB+30Features	89.51 ± 1.08	89.76	81.54
NB+41Features	85.31 ± 1.13	88.87	81.32

Figure 3 and Table 3 represent TPR for five class values which indicate better performance of Proposed FW in detecting Normal, DoS, R2L and Probe instances in comparison with NB+41Features and NB+30Features. However Proposed FW detected only 39.89% of U2R attacks. Small number of U2R attacks in total dataset instances (11 instances) and the fact that Proposed FW is based on dividing NSL-KDD into 10 clusters cause U2R instances to became sparse and lead to low TPR for them.

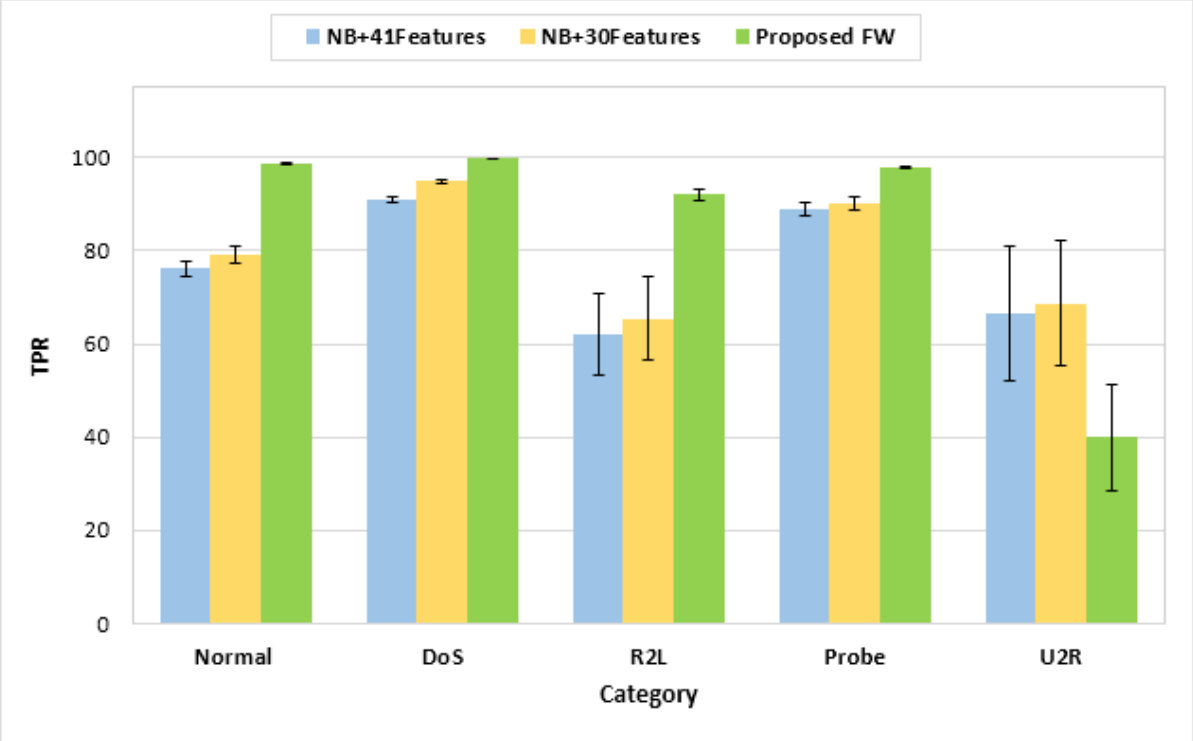

Figure 3. TPR comparison

Table 3. TPR comparison in detail

	TPR Mean ± CI		
	NB+41Features	**NB+30Features**	**Proposed FW**
Normal	76.2 ± 1.71	79.1 ± 1.71	98.77 ± 0.23
DoS	90.91 ± 0.51	94.9 ± 0.43	99.85 ± 0.05
R2L	62.14 ± 8.81	65.47 ± 8.9	91.99 ± 1.17
Probe	88.9 ± 1.32	90.1 ± 1.41	97.96 ± 0.32
U2R	66.54 ± 14.41	68.76 ± 13.49	39.89 ± 11.51

As illustrated in Figure 4 for five class values, the Proposed FW has low FPR which is significantly lower when compared to NB+41Feature and NB+30Feature. In addition, the obtained narrow confidence interval for FPR indicates Proposed FW results are more precise. The Proposed FW achieved 0.61% FPR for Normal, 0.18% for DoS, 0.23% for R2L, 0.42% for Probe and 0.17% for U2R which is the lowest as compared to others, as shown in Table 4.

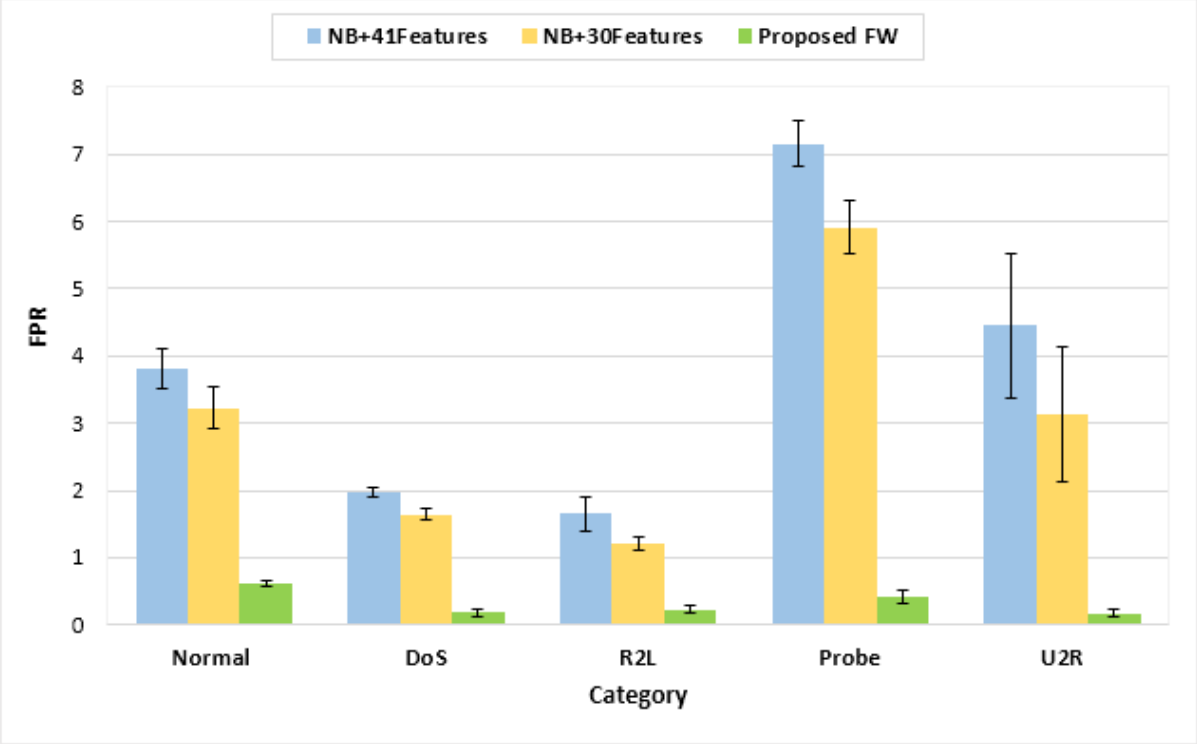

Figure 4. FPR comparison

Table 4: FPR comparison in detail

	FPR Mean ± CI		
	NB+41Features	**NB+30Features**	**Proposed FW**
Normal	3.81 ± 0.29	3.22 ± 0.31	0.61 ± 0.05
DoS	1.96 ± 0.07	1.64 ± 0.08	0.18 ± 0.05
R2L	1.65 ± 0.26	1.21 ± 0.09	0.23 ± 0.05
Probe	7.16 ± 0.34	5.91 ± 0.4	0.42 ± 0.09
U2R	4.45 ± 1.08	3.13 ± 1.01	0.17 ± 0.05

V. CONCLUSION AND FUTURE ASPECT

In this paper, a suitable framework is proposed for analyzing large number of network logs based on the combination of K-Means clustering and Naïve Bayes classifier. This framework was compared and evaluated using NSL-KDD dataset which is reduced to 30 features by using Gain Ratio technique. The fundamental solution is to group instances into 10 different clusters for providing better distribution of instances. Subsequently, each cluster is fed to Naïve Bayes to classify instances into one of the five general categories: Normal, DoS, Probe, R2L, U2R. As numerical results indicate, key points of this framework are achieving lower than 0.7% FPR for all of five class values, while keeping the average accuracy higher than 98.9% threshold. Moreover, the proposed framework improves the TPR for all class values except for U2R attacks. The TPR value for U2R attack can be further enhanced by efficient method of clustering which is defined as our future work.

REFERENCES

[1]. Azad, C., & Jha, V. K., "Data Mining based Hybrid Intrusion Detection System", Indian Journal of Science and Technology, Vol. 7, No. 6, pp.781-789, 2014.
[2]. Bhatti, D. G., & Virparia, P.V., "Data Preprocessing for Reducing False Positive Rate in Intrusion Detection", International Journal of Computer Applications, Vol. 57, No. 5, pp. 15-19, 2012.
[3]. Celebi, M. E., Kingravi, H. A., & Vela, P. A., "A comparative study of efficient initialization methods for the k-means clustering algorithm", Expert Systems with Applications, Vol. 40, No. 1, pp. 200-210., 2013.

[4]. Farid, D. M., & Rahman, M. Z., "Anomaly network intrusion detection based on improved self-adaptive bayesian algorithm", Journal of computers, Vol. 5, No. 1, pp. 23-31, 2010.
[5]. Gu, G., Fogla, P., Dagon, D., Lee, W., & Skoric, B., "Measuring intrusion detection capability: An information-theoretic approach", In Proceedings of ASIACCS, pp. 90-101. 2006.
[6]. KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html Retrieved 2014/7/10
[7]. Kemmerer, R. A., & Giovanni V., "Intrusion detection: A brief history and overview", Computer, Vol. 35, No. 4, pp. 27-30, 2002.
[8]. Kotsiantis, S., & Kanellopoulos, D., "Discretization techniques: A recent survey", GESTS International Transactions on Computer Science and Engineering, Vol. 32, No. 1, pp. 47-58, 2006.
[9]. Kumar, V., Chauhan, H., & Panwar, D., "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset", International Journal of Soft Computing and Engineering, Vol. 3, No. 4, pp. 1-4, 2013.
[10]. Mukherjee, S., & Sharma, N., "Intrusion detection using naive Bayes classifier with feature reduction", In Proceedings of C3IT, Vol. 4, pp. 119-128, 2012.
[11]. Tan, P. N., Steinbach, M., & Kumar, V. "Introduction to Data Mining", Addison-Wesley, 2005.
[12]. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A., "A detailed analysis of the KDD CUP 99 data set", In Proceedings of CISDA, pp. 1-6, 2009.
[13]. The NSL-KDD Dataset, http://nsl.cs.unb.ca/NSL-KDD Retrieved 2014/7/10
[14]. Gunwanti S. Mahajan & Kanchan S. Bhagat, "Survey on Medical Image Segmentation using Enhanced K-Means and Kernelized Fuzzy C- Means", International Journal of Advances in Engineering & Technology, Vol. 6, Issue 6, pp. 2531-2536, Jan. 2014.

AUTHORS SHORT BIOGRAPHY

Maryam Hajizadeh holds BSc degree in Information Technology received from Shiraz University of Technology, Shiraz, Iran. She is currently pursuing MSc degree in Information Technology (Computer Network) at Shiraz University of Technology, Shiraz, Iran. Her research interests include Data Mining, Security, Software Defined Networking and Ad hoc Network.

Marzieh Ahmadzadeh holds a PhD in Computer Science and MSc. in Information Technology, both received from the University of Nottingham, UK, and a first class BSc in Software Engineering received form Isfahan University, Iran. Since September 2006, she has been a lecturer and assistant professor at the school of computer Engineering and IT, Shiraz University of Technology, responsible for conveying a variety of modules to undergraduates and postgraduate, supervising dissertations and thesis. Amongst her main research interest is data mining in general and the application of data mining in detection and prevention of system security problems including intrusion to systems.