False Alarm Rate Reduction using Hybrid Model in Network Anomaly Detection

Shaimon Rahman Noman
Ahsanullah University of Science & Technology

Munawara Saiyara Munia
Ahsanullah University of Science & Technology

Samira Samrose
Ahsanullah University of Science & Technology

ABSTRACT
Network based intrusion causes predominantly to reveal network and service vulnerabilities. And that is why network based intrusion detection system execute thoroughly packet inspection. For faster execution with better detection accuracy, of the overall procedure while facing new dataset, we are representing a hybrid intrusion detection system in this paper. The hybridized algorithms are Triangle Inequality based k-means clustering algorithm and k-nearest neighbor classifier. Basically a combination of clustering and classification algorithms is studied in this paper. The dataset we used is the refined version of KDD'99 dataset and it is NSL KDD dataset. Some ingrained problems are solved in NSL KDD dataset. This paper work mainly focuses on the reduction of the false alarm rate. But the system is capable of detecting U2R, R2L, probe and Dos with high accuracy.

Keywords
Hybrid intrusion detection system, data mining, Triangle Inequality based k-means, k nearest neighbor, NSL-KDD dataset, accuracy, false alarm rate.

1. INTRODUCTION
The network based technology, system and communication creates a huge revolutionary reorganized mankind. Now-a-days people are more comfortable dealing with network based systems. Not only because of the painless and simple procedure but the high-speed connectivity makes the network based system more popular. And therefore it is highly important to secure the system to support the overall system maintenance. An intrusion detection system (IDS) is responsible for inspecting all incoming and outgoing network activity, along with identifying apprehensive patterns which may specify a network or system intrusion intended to violate the systems integrity. Hence, an IDS monitors the computer systems and network traffic for possible hostile attacks initiating from outside the organization as well as for a system abuse or attacks originating from inside the organization.

Network-based IDS observes data exchange between hosts and performs an analysis of passing traffic on the entire subnet and checks for the similarity between passed traffic and known attacks stored in database. The administrator will get an alert if there found any abnormal behavior. An example of network based intrusion detection is installing it on the subnet in the presence of firewall. The obtainable data for IDS is typically high dimensional, with a combination of categorical and continuous attributes.

In this work we have used clustering and classification algorithms to create a hybrid model. And for the clustering algorithm we are using triangle inequality based K-means or fast K-means and for classification algorithm, K nearest neighbor algorithm. In this hybrid detection system we mainly studied on the accuracy and false alarm rate.

2. RELATED WORKS
Machine learning methods have been employed in the field of anomaly detection to identify whether the behavior of data is normal or abnormal [1]. Moreover, the reasonable accuracy and detection rate can be earned by employing the combinational approach, when as a minimum two algorithms of machine learning various clustering and classification procedures are gathered to perform anomaly detection [2][3]. Author use K-Means and DB-Scan to efficiently identify a group of traffic behaviors that are similar to each other using cluster analysis. Author [4] proposed Linear Discriminant Analysis (LDA) to reduce features on the NSL-KDD dataset to 4 features only this gives 97% reduction in the input data and approximately 94% reduction in the training time.

3. PROPOSED SYSTEM ARCHITECTURE:
We have used data mining techniques for the detection purpose. Signature based learning techniques show high detection rates, but with equally high false alarm rate though [5, 6, 7]. For building a more efficient anomaly detection system, Triangle Inequality Based K-Means (a variant of general K-Means, also denoted as Fast K-Means (FKM) in this work) [8] and K-Nearest Neighbor algorithms are hybridized which can accomplish high detection accuracy with very minimal false alarm rate. We have applied the hybrid classifier on NSL KDD dataset for checking accuracy of the model.

While building the model, we have preprocessed the data for further manipulation in the first place. Then we have selected the most relevant features for reducing the dataset for efficient usage of the algorithms. We have applied Fast K-Means to the preprocessed data instances to fragment the data into five clusters: one normal cluster and four anomalous clusters named DoS, U2R, R2L and Probe. Each cluster is given an index number and we have labeled the records with the cluster indices. During training phase, we have applied the labeled training records to the K-nearest neighbor classifier. Then the trained KNN classifier model was evaluated using 10-fold cross validation technique. Finally, we have applied the unlabeled testing data to the trained classifier for classification. K-NN classifier eventually classifies the unlabeled record into anomalous and normal clusters.

Fig 1 shows proposed system architecture which can be divided into four major modules:

A) Data preprocessing module
B) Feature selection module
C) Clustering module
D) Classification module

3.1 Data Preprocessing Module:
The early data sources intended for the intrusion detection process may be different, with significantly different data

structures, which may provide unorganized and redundant early data, with partial information. Also, not all the algorithms can perform properly with categorical data. The necessity for data preprocessing is that redundant and insignificant features may often puzzle the classification algorithm, which may result in the discovery of erroneous or unproductive knowledge with significantly higher processing time due to the use of all features. Hence we must carry out data preprocessing in order to improve the efficiency, removing redundant and incomplete data and get reliable results by altering the data into an unvarying format.

3.2 Feature Selection Module:

Feature selection plays a vital role in reducing the irrelevant attributes, thus improving the efficiency of the clustering module. So, we have used Information Gain (IG) technique as a ranking process of the features. It can be separated into two steps: in step one, we have ranked the features in descending order using the algorithm, based on their degree of relationship to the target class, measuring their information gain. Then we have chosen top 8 relevant attributes. In the second step we remove the rest of the extraneous features with low estimation ability to the target class, thus reducing the dimensionality of the feature space.

3.3 Clustering Module:

Clustering is an unsupervised method for isolating unlabeled data into sets of analogous substances using a simple distance-based metric. The members from the identical cluster are analogous and are unlike the members of different cluster. It is unsupervised because the data points are not classified previously.

In the proposed approach, we have used Triangle Inequality Based K-Means algorithm, a variant of general K-Means, with very high processing speed. In this module, the preprocessed data records are clustered into five major groups, which increases the detection accuracy. It basically works as a pre classification module as clustering is appropriate for detecting novel attacks without any prior training data.

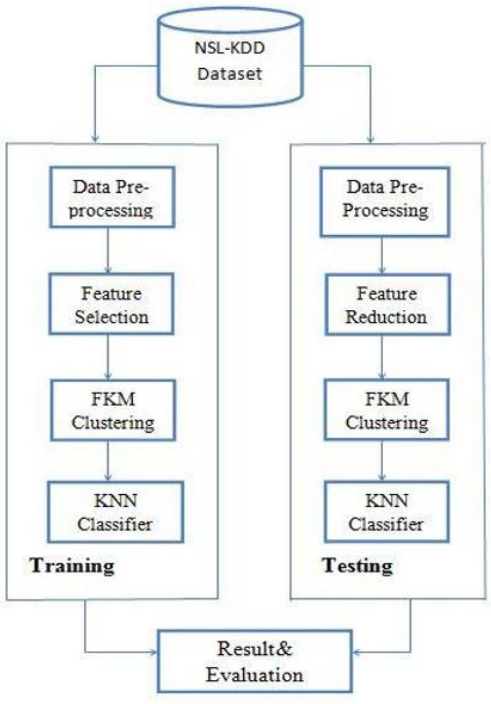

Fig 1: Fast K-Means & KNN Hybrid Evaluation

3.4 Classification Module:

A classification algorithm is generally used for guessing the class label of unknown records. From sample training dataset, a classification model is built using a methodical approach from the data set to decide the class to which the unknown data belongs. First, a training data records with known class labels should be provided, which is used to form the classifier model. This is done by applying a learning algorithm which finds the association among the set of attributes and the class of input data records and the trained classifier should be able to appropriately predict the class labels of unknown data records. Then the unlabeled test data records are passed through the trained classifier successively. The performance of the classifier is evaluated based on how many test records were classified correctly and incorrectly by the model. Examples include SVM, Neural Network, decision tree classifiers etc.

Proposed algorithm

Input: Training Dataset TR, Test Dataset TS, Anomalous cluster A, Normal Cluster N.

Output: TSi is normal or anomalous.

Procedure of the Hybrid Algorithm:

Training phase:

- Input original training data set TR which includes feature set F = {F1,F2,….,F41} and target classes C
- Convert symbolic features into numeric in Training Dataset TR using Arbitrary Assignment technique.
- Normalize TR to change the attribute values range into [0,1]. The new attribute values can be measured using the following formulas:

$$NewV = \frac{V - MinX}{MaxX - MinX}$$

- Removing irrelevant features as follows:

  i. Calculate Information Gain for each class.

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2\left(\frac{s_i}{s}\right)$$

  ii. Calculate Entropy of the Feature set F.

$$E(F) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} \times I(s_{1j}, \ldots, s_{mj})$$

  iii. Then the Information Gain for feature set F can be calculated using the below formula :

  Gain (F) = I (s1,…,sm) –E(F)

- Select top 8 attributes with highest Information Gains and remove rest of the redundant features. New relevant Feature set is Fnew.
- Select 5 centroids randomly and apply Triangle Inequality Based K-Means Algorithm for producing data clusters and divide the data into five groups.
- Obtain cluster indices and these cluster indices serve as the new training class labels for the training

data. Then a separate copy of the training data set (New-TR) file is updated.

- Data records from dataset New-TR are now applied to the Classification algorithm (KNN) to train the Classifier. Apply 10fold cross validation as validation process.

Testing phase:

- Take Test Dataset TS
- For each record TSi in TS, do:

  Find distances (TSi, d), for all (TSi, d) which is the element of TS, where d is the other data record.

- Then the distances are arranged in ascending order.
- Find the first K shortest distances and select first k nearest neighbors.

  i. If (vote (TSi, N) <vote (TSi, A)) TSi is Normal

  ii. Else If (vote (TSi, N)> vote (TSi, A)) TSi is Anomalous

  iii. Else TSi is undefined.

4. EXPERIMENTAL SETUP AND RESULT EVALUATION

4.1 Evaluation Dataset:

For evaluating the performance of most of the IDS, the most widely used data set is KDDCUP'99. But due to various shortcomings of the KDDCUP'99 dataset, in this work, we have used the NSL-KDD data set (which is the refined version of its predecessor KDDCUP'99 data set and includes 41 features) to study the effectiveness of the proposed system in detecting the anomalies in the network traffic patterns.

4.2 Experimentation:

The computer used to implement the experiments is equipped with Core-i5 6th generation processor, with clock speed 2.30 GHz, 8 core CPU and 8 GB RAM and a 64 bit Windows operating System. Experimental work is carried out using MATLAB data mining software and MATLAB Statistics toolbox. We have used 10 fold cross validation for training and validation and used test dataset for evaluation. In our work we used 125973 records as training data, where 53% are normal records and 47% are distributed among the different attack types. We used 22544 test records where 43% data are normal record, 17% are unknown and the rest are distributed among the different attack types. The data distribution ratio is shown in Fig. 2 and 3.

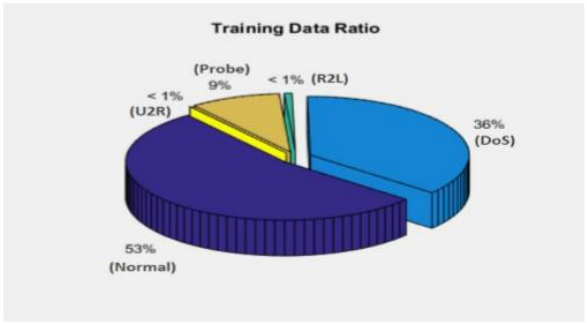

Fig 2: Training data distribution ratio

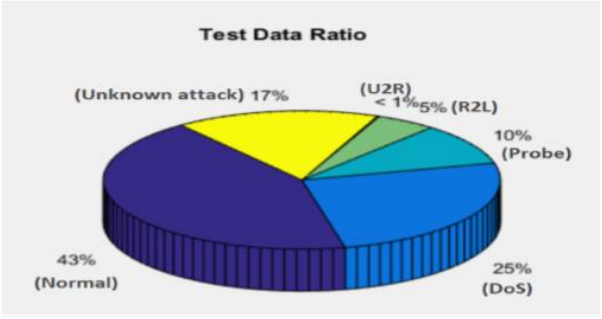

Fig 3: Testing data distribution ratio

To find a more efficient classifier, two steps hybrid model is employed. We took the advantage of K-means to classify 38 different types of attack and normal classes into five major classes – Normal, U2R, R2L, Probe & DoS, which played a vital role in increasing the efficiency of the hybrid classifier. As dataset preprocessing is very necessary for getting the proper accuracy as different algorithms and also the feature selection technique takes into account only the discrete attributes, not the continuous ones, hence we need to convert the continuous features to discrete ones before feature selection. So our first motto was to change the symbolic or categorical features into numeric features. We have used arbitrary assignment technique for this purpose which maps each category into sequential integer values. For emphasizing equal importance to each feature value, we have normalized the feature values from 0 to 1. Then the most important features are selected by computing information gain in which feature number 3 is having highest rank: 0.9101. The feature selection technique has reduced the number of features from 41 to 8 (attribute 3, 4, 23, 29, 30, 33, 34, 35).We then applied Fast K-means clustering on training dataset. We have chosen the initial centroids randomly and after clustering dataset into 5 clusters, the clusters are investigated and similar attacks are found into same clusters. On the next step, a field is appended to dataset which shows the desirable cluster label from '1', '2', '3', '4', '5' for each instance, while the original field class label was removed. After running the classification algorithm, a classifier model for prediction of '1' to '5' was created. After the classifier model was built, we evaluated the detection accuracy using test dataset. The confusion matrix of different data groups is depicted in Table 5. For example, table 5 shows that our approach can predict normal data group with an accuracy of 99.3%. It means that we can predict a new instance belongs to a group 'normal' which doesn't belong to other groups with an accuracy of 99.3%. We evaluated the effectiveness of the hybrid classifiers by comparing them with the single classifiers and by means of accuracy, detection rate (DR), false positive rate (FPR), sensitivity and specificity which can be calculated as follows:

Accuracy = (TP+TN) / (TP+TN+FP+FN)
Detection Rate = (TP) / (TP+FP)
False Alarm Rate = (FP) / (FP+TN)
Sensitivity = (TP) / (TP+FN)
Specificity = (TN) / (TN + FP)

Where,
True positive (TP): an attack data identified as an attack.
True negative (TN): a normal data identified as normal.
False positive (FP): a normal data identified as an attack.
False negative (FN): an attack data identified as normal.

The evaluation parameters of the classifiers are calculated and presented in table 6, the same is depicted in graphical form in Fig. 4-6. The next experiment demonstrates the strength of the hybrid models in improving the data classification processing time by introducing triangular inequality based K-Means clustering approach. To show the ability of the FKM-KNN to reduce the processing time, this hybrid approach is compared with general KM– KNN model. Table 7 is showing the execution time of proposed the technique on nearly 150000 data (including training and testing data). As the execution time of the algorithms differ from one execution to another execution, the algorithms were executed a couple of times and the average times are counted for analysis. The total elapsed time to cluster all the 125973 data points into 5 clusters using Fast K-Means is 1.6-1.7 seconds (including data preprocessing and feature selection) whereas for 22544 test data, it is 0.15 seconds, whereas the general K-Means showed a very poor result, 67.489s in total for clustering both the training and test data.

Table 1. Classification Result for KNN Using Training Dataset

Actual	Predicted Normal	Predicted Attack
Normal	66470	873
Attack	1161	57412

Table 2. Classification Result for FKM-KNN Using Training Dataset

Actual	Predicted Normal	Predicted Attack
Normal	54019	44
Attack	106	71663

Table 3. Classification Result for KNN Using Test Dataset

Actual	Predicted Normal	Predicted Attack
Normal	9916	515
Attack	2293	6711

Table 4. Classification Result for FKM-KNN Using Test Dataset

Actual	Predicted Normal	Predicted Attack
Normal	11605	9
Attack	144	10628

Table 5. Confusion Matrix for FKM-KNN on Test Dataset

Predicted / Actual	Normal	U2R	R2L	Probe	DoS	Accuracy (%)
Normal	11605	0	3	6	0	99.3%
U2R	0	660	5	0	0	99.6%
R2L	5	78	1084	0	26	99.33%
Probe	139	0	19	4325	1	99.2%
DoS	0	1	15	13	4559	99.7%

Table 6. Comparison of Single and Hybrid Classifiers using Different Measures

Dataset	Training		Testing	
Classifiers	**KNN**	**FKM-KNN**	**KNN**	**FKM-KNN**
Accuracy (%)	98.34	99.7	84.64	98.6
Detection Rate (%)	98.5	99.9	92.87	98.78
False Alarm Rate (%)	1.29	0.149	5.3	0.0774
Sensitivity (%)	98.0	99.9	74.5	99.9
Specificity (%)	98.7	99.8	94.7	98.7

Table 7. Processing Time of Single and Hybrid Classifiers in Seconds

Method	Processing Time (s)	Training Time (S)	Testing Time (s)	Total Time (s)
FKM-KNN	1.833	8.4448	7.2035	17.4812
KM-KNN	67.489	7.789	7.034	82.311

4.3 Result Analysis:

The classified data records as a measure of TP, TN, FP and FN are shown as confusion matrices in table 1-4. No. of attributes has significant role in the performance of the model, and with 8 attributes with top information gain, our model depicts remarkable efficiency. Table 1 and 2 represent the confusion matrices obtained from KNN and FKM-KNN against training dataset. FKM-KNN outperforms the KNN in identifying attack and normal data more accurately for testing dataset, where merely 9 normal instances identified as attack and merely 144 attacks instances were detected as normal, whereas KNN produces 515 false positives and 2293 false negatives which is almost 14 times more than the hybrid model, as illustrated in Table 3 and 4. It can be depicted from these results that single classifier contributes in increasing false alarm rate compared to hybrid approach.

Table 5 provides the detailed Confusion Matrix for FKMKNN on test dataset and as we can see, FKM-KNN can detect all types of attacks (U2R, Probe, R2L, DoS) very efficiently, with accuracies of more than 99%. It misses many R2L attacks though, but the overall accuracy is highly increased and the false alarm rate is decreased to a great extent.

By using training dataset, KNN produced almost the same accuracy with FKM-KNN, with similar detection rate and higher false alarm. In testing environment, hybrid approach increased the accuracy by +14.2%, detection rate by 6%, whereas lowering down false alarm rate up to -5.2%. In contract, single classifier obtained 84.64%, 92.87% and 5.3% respectively. To sum up, KNN undergoes in high false alarm compared to FKM-KNN. In general, the excellence and efficiency of anomaly based detection is assessed by the false alarm value. The smaller amount of false alarm values, the

higher the proficiency of the anomaly based detection model. This can be seen in Table 6.

Table 7 shows the total processing time of this experiment. The total execution time of this experiment is 17.4812s and 82.311s for FKM-KNN and KM-KNN respectively.

So, it can be said that our proposed hybrid technique shows superior result than the general KNN classifier. The clustering module plays vital role in the impressive improvements of the result and feature selection can make significant increases in processing speed.

5. CONCLUSION

The main motivation of NIDS system is to detect intrusions. So there are many studies on different algorithms and hybrid models. In this paper we represented different algorithms based hybrid model applied on an updated dataset. The reduction of false alarm rate and get more accuracy is the main concern here. To create a hybrid model with less complexity and get better result is our next goal.

6. REFERENCES

[1] SundusJuma,ZaitonMuda and WarusiaYasin "Reducing False Alarm Using Hybrid Intrusion Detection Based On XMeans Clustering And Random Forest Classification .''

[2] C.F. Tsai and C.Y. Lin, "A triangle area based nearest neighbors approach to intrusion detection," Pattern Recognition, Vol. 43, 2010, pp. 222-229.

[3] C.H. Tsang, S. Kwong and H. Wang, "Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection," Pattern Recognition , Vol. 40, 2007, pp.2373–2391.

[4] RupaliDatti,BhupendraVerma,"Feature Reduction for Intrusion Detection Using Linear Discriminant Analysis." (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 1072-1078.

[5] Chetan R &Ashoka D.V. "Data Mining Based Network Intrusion Detection System: A Database Centric Approach" 2012 International Conference on Computer Communication and Informatics (ICCCI 2012), Jan. 10 – 12, 2012, Coimbatore, INDIA

[6] VirendraBarot and DurgaToshniwal "A New Data Mining Based Hybrid Network Intrusion Detection Model" IEEE 2012. [3] Wang Pu and Wang Jun-qing "Intrusion Detection System with the Data Mining Technologies" IEEE 2011.

[7] Wang Pu and Wang Jun-qing "Intrusion Detection System with the Data Mining Technologies" IEEE 2011.

[8] Charles Elkan "Using the Triangle Inequality to Accelerate K-Means" Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.