A Hybrid Data Mining Model for Intrusion Detection

Mahreen Nasir
School of Computer Science, University of Windsor
ON, N9B 3P4

ABSTRACT
Network intrusion detection requires analysis of network data streams for identification of possible attacks. An Intrusion Detection System (IDS) is used to analyse such attacks and prevent future attacks. Main categories of IDS are anomaly detection and misuse detection. The limitation of anomaly based detection is high false positive rate whereas misuse detection based systems can only deal with known attack types. To address these, the main contribution of this paper is to propose a framework using hybrid approach based on clustering and classification methods for Intrusion Detection (CCID).

General Terms
Computer Science, Security, Computer Networks

Keywords
Intrusion Detection, Classification, Clustering, Data Mining

1. INTRODUCTION
Due to substantial increase in internet usage, it is essential to secure the network and avoid potential attacks. To get access to the private data, illegal users perform scanning of the systems and networks to find out the vulnerabilitiesand then break into the system. To prevent this type of situations, Intrusion Detection Systems (IDS) were designed which detect and respond to intrusions and generate alerts so that appropriate measures can be taken [9], [12-13]. An IDS investigates various data sources such as user behaviour, network traffic or logs to find out traces of computer misuse. Many detection schemes for IDS exist. Misuse Detection is a detection scheme used by many intrusion detection systems in which known bad behaviors are represented in the form of signatures. IDS based on this technique are good in detecting only well-known attacks.

An alternative to this is anomaly detection in which users' normal behavior is represented [1, 6]. For detecting such anomalies various learning techniques are used including data mining and machine learning. The detection is done bymatching the new data in opposition to the normality model, and any variations are considered as anomalies. Such systems are good in detecting unseen attacks because the normality model's training does not require any prior knowledge of attacks. However, there are short comings to both approaches. For example, unknown attacks may be missed in misuse detection because it is impossibleto define all possible attacks. In case of anomaly detection, it is difficult toseparate boundaries between normal and abnormal which leads to high falsepositive rate which means normal instances are classified as abnormal. To benefit from anomaly and misuse detection methods and address their limitations, this paper proposes hybrid approach based on Clustering and Classification forIntrusion Detection (CCID). The framework will consist of two stages. Stage 1 for anomaly detection and stage 2 for misuse detection. Anomaly analysis will use K-Means clustering [3] and use cluster centroid to detect the networkconnections as normal or abnormal. The connections labeled as normal from stage 1 will be further tested for false negatives by misuse detection stage. It will use inter class distance and K-Nearest Neighbor(K-NN) [8] to evaluate theinstances from stage 1. They will be considered normal if no similarity is found with the attacks in the training data otherwise they are abnormal connections misclassified as normal from stage 1.

The rest of the paper is organized as follows: Section 2 highlights the related work. Proposed framework is discussed in section 3. Section 4 presents the experiments with section 5 containing experimental analysis. Section 6 concludes the paper and provide details for future implementation.

2. RELATED WORK
Data mining is used to discover patterns of interest in data and can complement the process of intrusion detection. This can be achieved by analyzing the data for many interesting patterns such as different attack types using statistical methods and data mining algorithms. One widely used unsupervised approach is Clustering which finds patterns of interest from high dimensional unlabeleddata [11]. Clustering techniques use different criteria to group data points in a cluster. For instance, hierarchical clustering uses distance measure between items to form clusters, statistical distribution (Expectation Maximization) creates clusters which are compliant with a statistical distribution, centroid based such as K-Means represents a cluster by its mean value and graph-based techniques considers data points as connected nodes where data points are connected to each other with at least one data point.

Various previous studies have used clustering methods for the detection of anomalous traffic in the network [7]. Authors in [7] used two clustering schemes. First, to identify the attack and second to determine normal traffic in a supervised way. Their main idea is to perform these two tasks in parallel. The output of these stage is used to extract signatures which can later be used by thesecurity professionals. They used KDD data set in experiments. Accuracy and cluster integrity were used as performance metric. The results were reported to achieve 70% to 80% detection rate for unknown attacks. The authors in [2] usedDBSCAN clustering to categorize normal and anomalous traffic. Clustering method threshold was used to control the FAR of system. They preprocessed KDD data set and used correlation analysis to select features. Furthermore, an attack to no attack ratio of 10% was set during data preprocessing. Data of nine users from Purdue University comprising of 500 sessions was used by [10].To differentiate between a regular user and intruder, they used user command level data. Users' commands in a session were represented as sequence of tokens. The authors used longest common subsequence metric as a similarity measure for sequence matching.

Another approach comes under classification such as Baysian networks which is a probabilistic graphical model representing the variables and the relationships between them

[5]. It can be used for classification of network streams. A study by [4] proposed a framework using Baysian network classifiers. For anomaly detection stage, an inference junction tree was used to make a decision. They reported a performance of 88% on normal and 89% on attack categories. Next,an anomaly detection module was used to recognize different attack types from the attack data. For attack types of DoS, Probe or Scan, R2L, U2R, and other classes, a performance of 89%, 99%, 21%, 7%, and 66% was reportedrespectively.

3. PROPOSED METHODOLOGY

The proposed framework will consist of two stages. Fig. 1 shows an overview of the proposed framework. Stage 1 for anomaly detection and stage 2 for misuse detection. Anomaly analysis will use K-Means clustering [3] and use cluster centroid to detect the network connections as normal or abnormal. The connections labelled as normal from stage 1 will be further tested for false negatives by misuse detection stage. It will use interclass distance and K-NN [8] to evaluate the instances from stage 1. They will be considered normal if no similarity is found with the attacks in the training data otherwise they are abnormal connections misclassified as normal from stage 1.First, a cluster consisting of training instances from normal class will be created using K-Means. The centroid of this cluster will be recorded and set as external threshold C_e. This external threshold will be used by anomaly detection module to declare connections as normal or abnormal. The details for both sections are provided below.

3.1 Anomaly Detection Module

The purpose of this module is to declare connections as normal or abnormal.The module works by partitioning the training data set consisting of normal and abnormal connection into K clusters$\{C_1, C_2, \ldots\ldots, C_k\}$. The centroid of each cluster C_i will be computed and compared with the external threshold. If cluster centroid is greater than the threshold, the cluster will be separated and all the instances in that cluster will be labelled as abnormal. The instances for which the centroid is less will be labelled as normal. The steps are listed below:

(i) Partition the training data set consisting of normal and abnormal connection into K clusters $\{C_1, C_2, \ldots\ldots, C_k\}$.

(ii) Compute centroid of each cluster C_i from (i) using K-Means.

(iii) Compare each$C_i$with the centroid set as external threshold C_e, If$C_i < C_e$ then label instance as \normal" otherwise label instance as \abnormal".

(iv) For clusters where instances are labelled as abnormal, compute interclass distance d_i between these clusters using (Eq.1) and select the maximum distance. Interclass distance is the distance between two clusters which defines the level of isolation between instances in each cluster. This maximum inter class distance will be set as internal threshold d_l which shows themaximum distance between classes representing attack type instances. This will be used in the next stage of misuse detection.

$$d_i = | C_i - C_{i+1} | \qquad Eq.1$$

Where d_i is the interclass difference between two cluster based on their centroids and C_i represents centroid.

3.2 Misuse Detection Module

The purpose of this module is to further analyze instances classified as normal by anomaly detection module. It will investigate whether the instance is a false positive (when an instance is normal but misclassified as attack) or actually anormal instance. This module will be based on (i) using K-Nearest Neighbor (K-NN) which uses cosine similarity to measure similarity between instances and (ii) using internal threshold. Initially, cosine similarity will be computed between each training instance x_i and the instance labeled as normal $x_n$from anomaly detection module. Higher the similarity means more closeness to attack type. Next, average of top k similarity scores will be computed and compared with the internal threshold to predict the label as attack or normal. Steps for this module are below:

(i) For each instance in training set with attack types, calculate $cossim(x_n, x_i)$ where x_n is the instance labeled as normal from anomaly detection module and x_i is the instance in training data.If$cossim(x_n, x_i)$ =1 then label x_n as attack

(ii) Find top K score of $cossim(x_n, x_i)$ =1 and calculate the averagescore.

(iii) If average score is less than internal threshold then label x_n as attackotherwise label x_n as normal.

4. EXPERIMENTS

The KDD cup99 data set was used for testing the proposed methodology. The kddcup99 dataset is widely used dataset forintrusion detection and was first given by Massachusetts Institute of Technology. The dataset contains 24 kinds of attacks that can be categorized as four types to be named as denial of service attack (DOS), user to root attack (U2R), remote to local(R2L) attack and probe attack (PR). It contains 41 attributes divided into 34 nominal and 7 numeric attributes[14]. The KDD's dataset is divided into two categories as training set and test set, both of which contains a large number of connection records. Table 1. shows data set details. More details of data set can be found at [14].

4.1 Data set Pre-processing

The 10% training and test data set were further split into 4 files each comprising of approx. 250-350 instances. The main focus is on the connection records with DOS attack type represented by labels as smurf, back and neptune along with the class label normal. The instances approximately represent equal distribution of all the above mentioned types. The attributes dimension for each record was reduced from 41 to 10 as proposed by [15] in order to increase classifier speed and accuracy. Table 1 shows distribution of normal and attack type data in the training and test data and Table 2 provides details about the data set attributes used in the experiments.

Table 1. Distribution of Normal and Attack Data inTraining and Test Files

Category	No. of Connections	% Split	Class
Training Data	5 Milllion	10% 494,020	97,277 Normal 39,6743 Attacks
Testing Data	2 Million	10% 311,029	60,593 Normal 250,436 Attacks

Table 2. Selected Attributes used in the experiments

Sr. No.	Attribute order based on Information Gain Ratio	Attribute	Type	Description
1.	3	service	discrete	network service on the destination, e.g., http, telnet, etc.
2.	4	flag	discrete	normal or error status of the connection
3.	5	src_bytes	continuous	number of data bytes from source to destination
4.	6	dst_bytes	continuous	number of data bytes from destination to source
5.	8	Wrong_fragment	continuous	number of ``wrong'' fragments
6.	10	hot	continuous	number of ``hot'' indicators
7.	13	num_compromised	continuous	number of ``compromised'' conditions
8.	23	count	continuous	number of connections to the same host as the current connection in the past two seconds
9.	24	srv_count	continuous	number of connections to the same service as the current connection in the past two seconds
10	37	dst_host_srv_diff_host_rate	continuous	% of connection to the different hosts

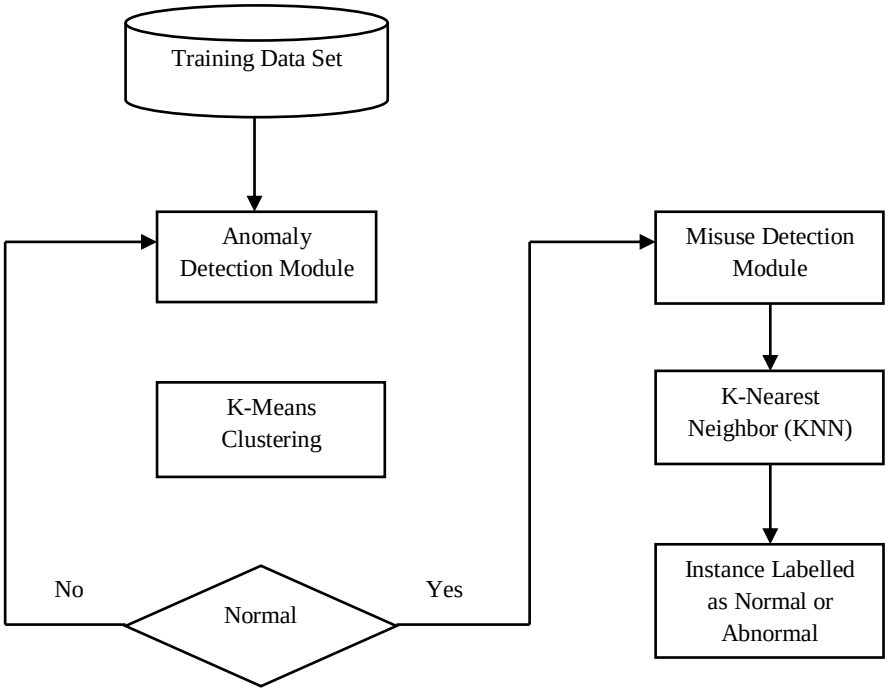

Fig 1: Overview of the proposed Framework

5. RESULTS AND ANALYSIS

After performing K-Means clustering the instances are grouped into four clusters according to the data distribution.The following clusters results are shown after loading second file. Here the rectangle shows the instances that are misclustered. A total of five iterations were performed and sum of cluster centroids and inter-class distance were recorded in each iteration. Refer to Table. 3 for detailed results.

Initially the classifier is supplied with the training data for the construction of base model and afterwards the classifier was retrained according to the results of clustering process. Classifiers error rate shown to be an indicative measure to represent that change of concept has been successfully detected or not. After applying K-NN classifier to the data set the instances are classified according to the following classes. The rectangles (in red presenting the smurf class) show the misclassified instances. The classifier's error-rate in Iteration 1 was 32%.After updating the classifier by using clustering results the classifier's accuracy has been increased and error-rate had been reduced from 32% to 28%.A total of five iterations were performed and the classifier's error-rate was recorded in each iteration**.** Refer to Table. 3 for detailed results.

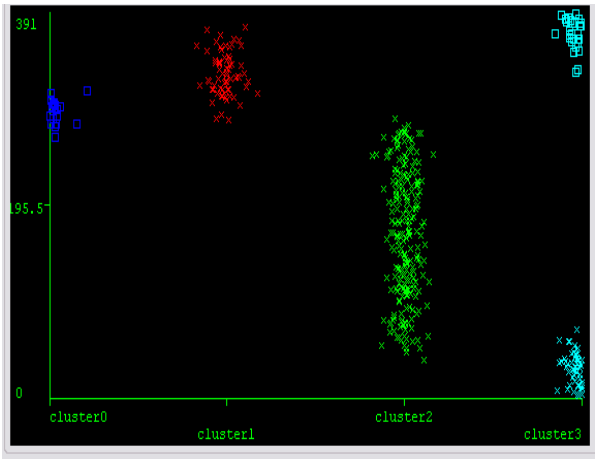

Fig.2: Clustering Results in Iteration II (With Jitter)

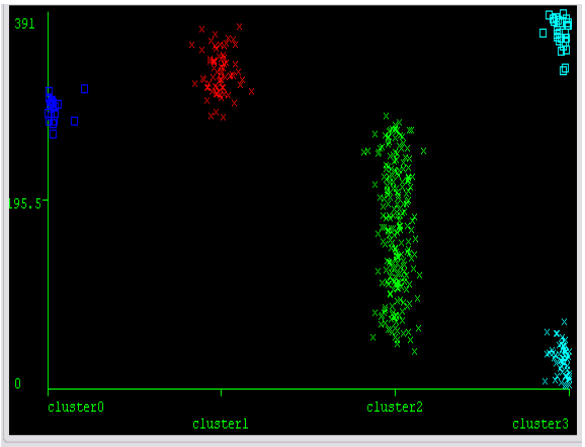

Fig.3: Clustering Results in Iteration II (With Jitter)

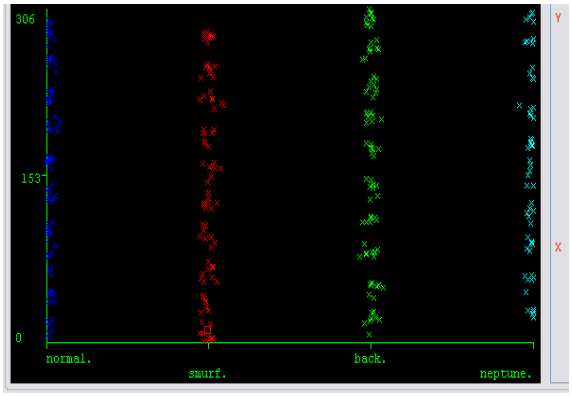

Fig. 4: Classification Results (With Jitter)

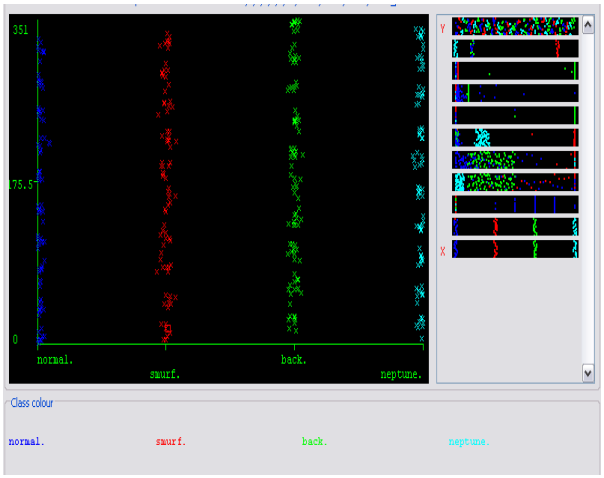

Fig. 5: Classification Results after Re-training (With jitter)

Table 3. Representing the Max Inter Class Differences & Classifier's Error-Rate in each iteration

Iteration No.	Max. Inter-Class Difference	Classifier Error-Rate
1	0	32
2	0.844418	28
3	2.501788	25
4	3.930608	22
5	5.080278	20

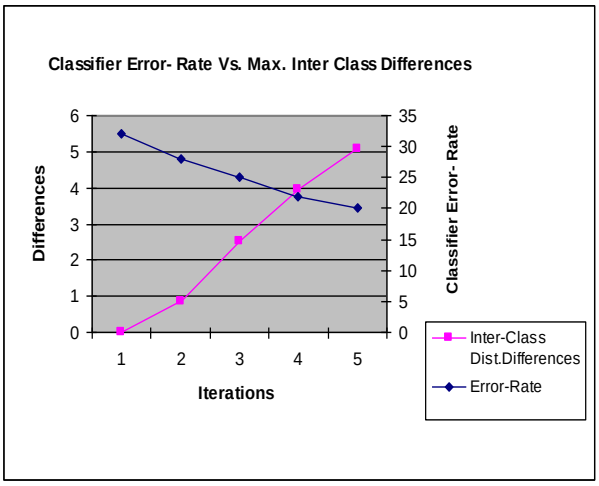

Fig. 6: Showing relationship among Max. Inter Class Distance Differences &Classifier's Error-Rate

6. CONCLUSION AND FUTURE WORK

This paper proposed a hybrid model based on K-means clustering and K-NN classification for Intrusion Detection. It uses cluster centroids and interclass distance as external and internal thresholds. First stage in the framework dealswith anomaly detection and labels instances as normal or attack. To further investigate the instances identified as normal to see if they are misclassified, second stage of misuse detection employs K-NN and use interclass distancemeasure.

The experiments performed leads to the result that Inter-Class distance is a strong measure to segregate instances of normal and attack type data. Also, the graph showing differences between maximum inter-class distances and the classifier's error rate clearly indicates that an increase in the distance significantly decreases the classifier's error-rate. The various cluster statistics represent information regarding the clusters made, including the data elements diversity, data distribution etc. Similarly, interclass distance represents the difference between the two clusters, i.e., it shows the maximum difference that can be among the clusters. This implies that the distance increases only if data elements have diversity in them, as elements in one cluster are grouped based on some similarity and they differ from the elements in other clusters. So, greater the interclass difference, more the change in the concept (e.g., attack type instance or normal instance).Future research will focus on evaluating this framework on more instances by considering other attack types and also determining which other cluster statistics along with interclass distance can be helpful in the detection of intrusive concepts in network data streams.Furthermore, the proposed approach will be tested on NSL-KDD data set to evaluate the effectiveness of the proposed method on a variety of data sets.

7. REFERENCES

[1] Monowar H Bhuyan, Dhruba Kumar Bhattacharyya, and Jugal K Kalita. \Network anomaly detection: methods, systems and tools". In: *Ieee communications surveys & tutorials* 16.1 (2014), pp. 303-336.

[2] Misty Blowers and Jonathan Williams. \Machine learning applied to cyber operations". In: *Network science and cybersecurity*. Springer, 2014,pp. 155-175.

[3] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classication*.John Wiley & Sons, 2012.

[4] Farah Jemili, Montaceur Zaghdoud, and Mohamed Ben Ahmed. "A Framework for an Adaptive Intrusion Detection System using Bayesian Network." In: *ISI*. 2007, pp. 66-70.

[5] Finn V. Jensen. Bayesian Networks and Decision Graphs. Berlin, Heidelberg*: Springer-Verlag*, 2001. isbn: 0387952594.

[6] Suleman Khan. Network forensics Review, taxonomy, and open challenges". en. In: *Journal of Network and Computer Applications* (2016),p. 22.

[7] Kingsly Leung and Christopher Leckie. "Unsupervised anomaly detection in network intrusion detection using clusters". In: *Proceedings of the Twenty-eighth Australasian conference on Computer Science*-Volume 38.Australian Computer Society, Inc. 2005, pp. 333-342.

[8] Yihua Liao and V Rao Vemuri. "Use of k-nearest neighbor classifier for intrusion detection". In: *Computers & security* 21.5 (2002), pp. 439-448.

[9] Safaa O Al-mamory and Firas S Jassim. “Evaluation of different data mining algorithms with KDD CUP 99 Data Set". In: *Journal of University of Babylon* 21.8 (2013), pp. 2663-2681.

[10] Karlton Sequeira and Mohammed Zaki. \ADMIT: anomaly-based data mining for intrusions". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, pp. 386-395.

[11] Richard Zuech, Taghi M Khoshgoftaar, and Randall Wald. “Intrusion detection and Big Heterogeneous Data: a Survey". en. In: (2015), p. 41.

[12] Salo, Fadi, et al. "Data mining techniques in intrusion detection systems: A systematic literature review." *IEEE Access* 6 (2018): 56046-56058.

[13] Agrawal, Diptee, and Chetan Agrawal. "A Review on Various Methods of Intrusion Detection System." *Computer Engineering and Intelligent Systems* 11.1 (2020).

[14] KDD Cup 1999, [online] Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (October, 2007)

[15] Wang, Wei, Xiaohong Guan, and Xiangliang Zhang. "Processing of massive audit data streams for real-time anomaly intrusion detection." Computer communications 31.1 (2008): 58-72.