A Novel Density Based K-Means Clustering Algorithm for Intrusion Detection

Randeep Brar
Department of Computer Science and Engineering, Chandigarh Engineering College, Landran, India.

Dr. Neeraj Sharma
Head of Department, Department of Computer Science and Engineering
Chandigarh engineering College, Landran, Mohali, India.

Abstract - Expanded utilization of web and information trade has prompted more assaults hence brought about expanded prerequisite of intrusion detection system. The analysts' interest is expanded in intrusion identification as the volume of information to be secured is rising to an exponent with the year which is straightforwardly corresponding to the sorts of intrusion. Intrusion or assaults represent a serious threat in systems administration and information sharing. In this way, there is urgent need to create perfect or close to perfect intrusion detection. In this paper, we have built up an intrusion detection system utilizing a novel methodology of k-means clustering algorithm for a signature based Intrusion detection system. The proposed strategy of utilizing k- means clustering for density based model is a modified version of k-means clustering which is a novel approach in the field of intrusion detection. Experimental results are completed on 10% of KDD Cup 99 data set. The result are measured as accuracy, detection rate, and false positive rates and are displayed in the graphical structure for normal queries and every sort of attack, namely Denial of Service, Remote to Local, User to Root, Probe. The finished up result demonstrate the viability of the proposed Intrusion detection system.

Index Terms – K-means clustering, Density-based clustering, Accuracy, Detection rate, False positive Rate, DBK-means Clustering, Confusion matrix

1. INTRODUCTION

Information security is an important issue following the evolution of the data exchange facilities. A unit of noxious activity may prompt complete deterioration. Intrusion detection is a principal stride for the security of data on the system can be disclosed as to identify and locate the error or malicious activity in the system. Working of IDS (Intrusion Detection System) is a joint effort of various steps which incorporates examining the data and distinguishing attacks followed by the classification of attacks based on some attributes. Further, IDS's are classified on the premise of analysis technique and on the basis of the placement of IDS. On the basis of analysis technique used, it can be parted as an anomaly-based IDS and signature-based IDS [1]. An anomaly-based IDS is a conventional framework to detect dynamic attacks. It detects attacks without the prior information about the patterns and detects vitality query which have developed properties from ordinary queries. The focal points of anomaly-based IDS are its ability to detect the new attack and its low dependence on stored attack patterns; however, the hindrances are the high false alarm rates. Though, signature-based IDS work on the basis of pre-stored attack framework. It maintains a database of previous attack pattern. It is straightforward as compared to anomaly-based IDS. The points supporting signature based IDS are its lower false positive rates whereas on other side, it oblique high upkeep of the database and cannot identify new attacks. Based on placement approach, IDS is classified as host-based IDS and network-based IDS. Host-construct IDS is situated with respect to a specific framework and secures it. The host-based IDS monitors the system, store data to log files and conducts detection. The points supporting host-based IDS are its well mulled over detailed pattern and enhanced results. On the flip-side, high maintenance, and increased workload are some of the hindrances confronted by utilizing host-based IDS. The network-based IDS detect the attacks within the data that moves over the network using the NIDS machine. The data requires passing through NIDS machine before casing to the system. NIDS is placed using a network segment, consequently it protects entire network segment. The merits of using IDS are the requirement of fewer assets and less system overhead whereas demerits are that it can't overcome every time and constrained obstruction with a host computer.

2. CLUSTERING

Clustering is a procedure of collection of objects on the premises of similarity among them. The objects or the data points in the same cluster have more degree of comparability when contrasted to another data points outside the cluster. Clustering is an essential technique of data mining. The clusters are formed by conveying calculations on the data set. In accordance to the necessity of particular application, diverse parameters and techniques can be used for clustering. Cluster investigation is not made by any specific programming; then again it is an iterative process of introducing information. In these iterations, preprocessing and parameters are adjusted until the outcomes are obtained as required. Clustering is an unsupervised learning issue. It

manages to discover an accumulation of unlabeled information. Clustering methods describe the criteria for making the cluster. Some of the clustering methods can be portrayed in brief as:

- Connectivity based clustering[12]

This clustering, also known as leveled clustering, considers the data points more identified with nearby data points than to consider other data points. The nearby points make a cluster and distant points make distinct clusters.

- Distribution based clustering[12]

The clusters formed in this mode of clustering are characterized as items having a place destined to the same circulation.

- Centroids based clustering[12]

As the name, centroids are the prime part of such clustering. The specific numbers of centroids for the dataset are selected. The clustering procedure progresses iteratively until the data points in clustering stay unaltered.

- Density-based clustering[12]

Application of density based clustering results in the formation of the clusters, which behave as the zones with higher density than other data points in the dataset.

2.1. Clustering algorithm

A clustering algorithm can be clarified as the collaboration of some steps to carry out clustering [8]. Every clustering algorithm bears the properties.

- The algorithm may handle big data or data with variations.
- The clustering algorithm must be time efficient and data efficient as much as possible.
- The outcome obtained must be interpretable so that it can further be demonstrated in the form of any parameters.

Some of the well-known clustering algorithms are k-means clustering, FCM clustering, and hierarchical clustering etc. The clustering algorithm can also be proposed from the combination of the two or more clustering algorithm, but the satisfying criteria are the properties described above.

In this paper, we have proposed the density based k-means clustering which is used for intrusion detection properties. The challenge is to have high accuracy, high detection rate, and low false alarm rate.

3. RELATED WORK

Intrusion detection is an area of concern for the researchers. Various researchers have carried out experiments with clustering algorithms based IDS to overcome the hindrances and broaden the scope of IDS. Brief overview of some of them is discussed as:

Sharma et al. [9] proposed an intrusion detection system based on k-means clustering followed by naïve Bayes classification. The proposed technique is compared with naïve Bayes classification of data. The intrusion detection procedure starts with the acquisition of the dataset followed by feature selection and normalization. Then the main algorithm K-means clustering is applied in addition to naïve Bayes classification. The results are given in the form of accuracy, detection rate, and false alarm rate. The result demonstrates that proposed methodology has 99% detection rate and 4% FPR, which is the restriction in the proposed methodology.

DBSCAN (Density-Based Spatial Clustering of Application with Noise) is used in an improved means by Yong et al. in [14]. The proposed clustering algorithm is named as IIDBG. The IIDBG is applied to intrusion detection system, in the same way as DBSCAN is applied. The results declared in the form of detection rate and false negative rate shows that proposed methodology improves the performance. On the flipside, consideration of numerous parameters is the limitation.

In addition to this, clustering heuristic named as Y- means clustering is proposed by Guan et al. in their research paper entitled 'Y-means: A clustering method for intrusion detection'[16]. The proposed algorithm is based on K-means clustering algorithm and other related clustering algorithms. This algorithm works on basic steps of K-means algorithm namely, initialization, assignment, updating and iteration. The main aim of proposing Y-means algorithm is to overcome two limitations of K-means 'number of clusters dependency' and 'degeneracy'. The proposed algorithm partitions the dataset into an appropriate number of clusters that is, initial no of clusters does not depend upon cluster results. Also, raw data can be used as training data without labeling it. It is better clustering algorithm for intrusion detection without supervision.

An improved K-means clustering algorithm for NIDS is proposed by Tian et al. in [17]. The k-algorithm is improved in the way that it overcomes the limitations of K-means clustering. The limitations taken in consideration are the dependence of algorithm on starting value selection and the time expenses of the k-means clustering algorithm. This paper introduced the optimized dynamic central point cyclic method. The algorithm resulted in the reduction of false detection rate and enhancement of detection rate.

4. PROPOSED MODELLING

The proposed model for intrusion detection is meant for signature-based NIDS.

4.1. The dataset description

The KDD cup 1999 dataset is considered for carrying out the experiments as it is generally utilized for intrusion detection system. This dataset is discovered by Stlofo et al. and is based in the perspective of the information caught in DARPA 98' intrusion detection system [14]. It contains large set intrusions reproduced in the military system environment. It contains 4,900,000 datasets, and every dataset has the feature which is part of the feature set. Only 10% of the dataset is used for carrying out experiments. The queries in the dataset are labeled as normal queries and some forms of the attacks. The re-instituted attacks fall in four classifications:

a. Denial of service(DoS)

This attack aims to hinder any assistance given to user or system. The assistance here means the use of the resources. Fundamental system to do such is to obstruct the pathway of the solicitation. For example apache, smurf and so on [12]

b. Probe

In this attack type, the data is assembled around the system for gaining its security controls with the aim to intrude the system in such a way that the information about the system is collected. Notwithstanding this, one may enter the system and its records through a known point in a system. Port scan is an example for this type of attacks.

c. User to Root

User to Root (U2R) attack is any activity carried out by the user who access the system as an ordinary user to gain the ability to get root access to the system. The client level user gets access to the root level system. This attack aims to get the credential information about the system.

d. Remote to Local

In (Remote to local) R2L the gatecrasher tries to abuse the framework vulnerabilities with a specific end goal to control the remote machine through the system as a neighborhood client.

4.2. Architecture of the proposed model

a. Acquisition of dataset

The basic step is to acquire the dataset. The dataset acquired is KDD cup 99' dataset. For carrying out the experiment, 10% of the whole dataset is considered.

b. Partitioning dataset into training and testing data

The dataset is partitioned as training dataset and testing dataset. In dataset a training set is executed to develop a model, while a test (or approval) set is to accept the model fabricated The training dataset is 80% of the provided dataset whereas testing data is the 20% of the provided dataset. The final results are evaluated on the basis of the experiments on the testing dataset.

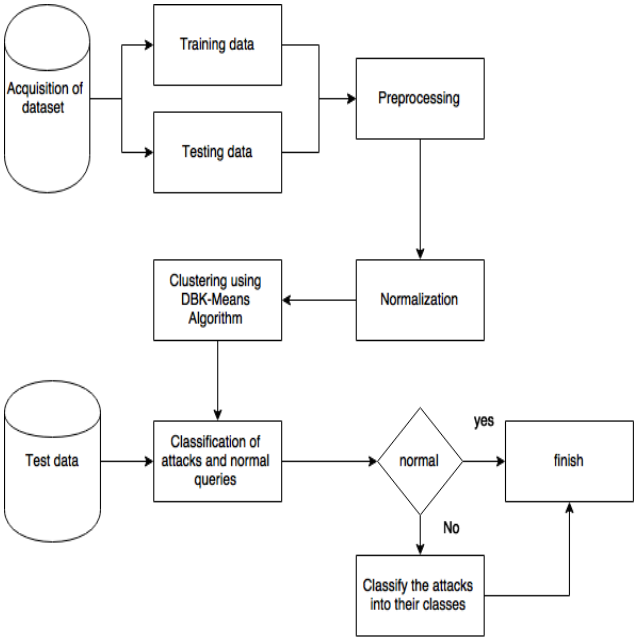

Fig.1. Architecture of DBK-Means Clustering Based IDS

Various steps in the architecture are explained as:

c. Preprocessing and normalization

In preprocessing or feature selection the basic features of the dataset are filtered out of the numerous features. The features may have discrete or continuous values. The 41 features are selected, which portrays the classes of the abnormal queries and distinguish them from abnormal queries. The values of the features in the dataset are not scaled. Some features have value of very high scale while some may have very less scale. To curb this menace, normalization is done which scales the value between any scales. Normalization helps to get the result in the desired way. For this experiment, we scale the values for features between 0 and 1

d. Clustering using DBK-Means clustering

DBK-Means clustering algorithm is a combination of DBSCAN and K-mean clustering. Basically, DBK-Means clustering algorithm defeats the downsides of DBSCAN and K-Means clustering algorithms. This algorithm performs better then DBSCAN when handling clusters of circularly dispersed data points and marginally overlapped clusters. The no of data points in the cluster are based on the number of expected points in a cluster or the expected density of the cluster (derived by using the number of points in a cluster and the area of the cluster).

Let A = {a_1, a_2, a_3... a_n} be the set of data points in Dataset D, Euclidean "ε" (eps).

Suppose the clusters to discover are counted by labeling them K and base numbers of neighbors required in ε neighborhood to form a cluster are supposed as MinPts and N is a set of points in ε neighborhood.

Step 1: Start with a discretionary starting point (not visited).

Step 2: Select arbitrary centroids with the number of centroids being equal to the quantity of obliged clusters.

Step 3: Calculate Euclidean distance of all points from the centroids or the starting data point within a dataset.

Step 4: Calculate the distance of all neighbors around a point and find the density.

Step 5: Keeping density under consideration locate the cluster where the point fits in and label it.

Step 6: Mark starting point as visited if 'N' is greater than or equal to 'minPts', otherwise mark it as unvisited. Then new unvisited points are recalled until all points are marked as visited. After that, we have 'm' clusters and then find cluster centers 'Cm' by taking the mean of a total number of points in each cluster.

Step 7: If 'm' cluster is greater than 'K' clusters, then join two or more cluster based on density and find the new cluster center. Rehash Step 7, until achieving K clusters with 'Ck' centers.

Step 8: Otherwise if 'm' is greater than or equal to the number of clusters initially found by density based clustering algorithm 'l'; select a cluster based on density and number of points split it using K-means clustering algorithm. Repeat Step 8, until achieving K clusters with 'Ck' centers.

Step 9: Apply iteration of k-mean clustering with k and new 'Ck' centers as the initial parameters and label all the clusters with k labels.

Step 10: End.

e. Calculation of normal queries and attack types

After obtaining the results of the algorithm, they are interpreted as normal queries or some specific attacks. If the queries belong to the normal clusters then finish else classify the attacks according to their class that is depicted by the particular cluster.

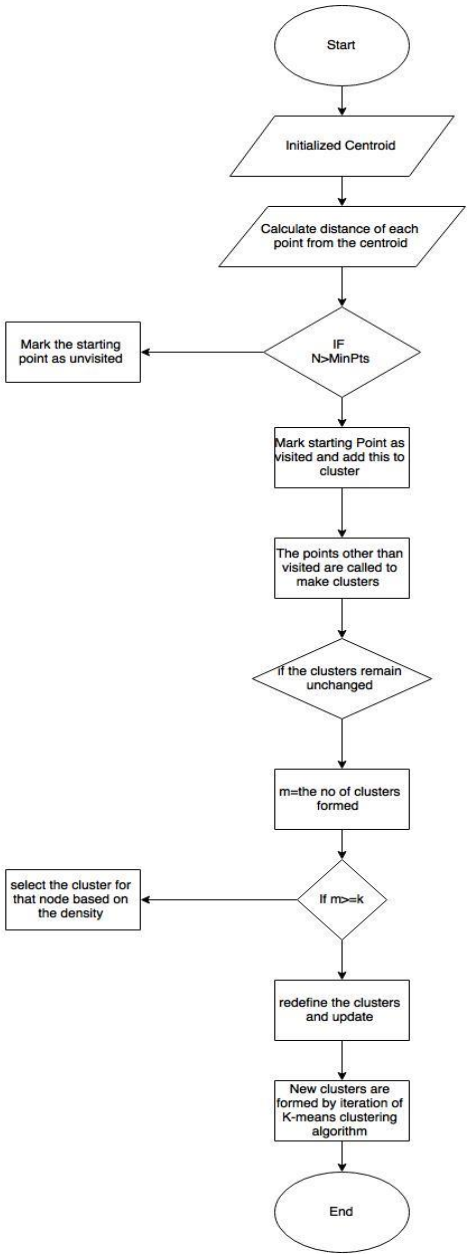

Fig.2. Algorithm for DBK-Means Clustering

5. RESULTS AND DISCUSSIONS

Examination chip away at this paper is done in a manner that to the point result will be given with the assistance of enhanced parametric calculation which would help to give high exactness, accuracy, and audit.

5.1. Result evaluation parameters

The evaluation parameters for measuring the performance of the proposed clustering algorithm are Accuracy, Detection Rate, and FPR.

The symbols used as TP, TN, FP and FN stands for True Positive, True Negative and False Positive and False Negative respectively

The values for TP, TN, FP and FN can be calculated from the confusion matrix for the testing dataset. The confusion matrix for the testing dataset is given as:

Actual class	Predicted class				
	Normal	DoS	U2R	R2L	Probe
Normal	178	0	0	1	0
DoS	9	236	0	0	0
U2R	4	1	5	0	1
R2L	8	0	0	94	0
Probe	0	0	0	0	275

Table1. The confusion matrix for DBK-means clustering

Filtering the values for TP, TN, FP and FN from the confusion matrix is done as shown in Table 2.

Actual class	Predicted Class		
	Class 1	Class 2	Class 3
Class 1	TP	FN	FN
Class 2	TN	FP	FP
Class 3	TN	FP	FP

Table2. Predicting TP, TN, FP and FN from confusion matrix

a. Accuracy

It demonstrates the aggregate no. of genuine results got out of the aggregate results got and implies an execution level which adds to the achievement rate of a calculation. Higher the accuracy; more enhanced is the method.

Accuracy can be computed as:

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

b. Detection Rate

Detection rate can be defined as the no of the detected attacks of a particular class out of the total attacks of that class. The enhancement of the detection rate favors the result orientation of the intrusion detection system

$$\text{Detection rate} = \frac{TP}{TP+FP}$$

c. False Positive Rate

It gives the number of normal queries which are mistakenly detected as the queries belonging to a particular class

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$

The values of the parameters obtained by using the proposed techniques for each class of the query are given as:

DBK-Means Clustering algorithm					
Query Parameters	**Normal**	**Dos**	**U2R**	**R2L**	**Probe**
Accuracy	99.44	96.33	45.45	92.16	100
DR	0.894	0.995	1.0	0.989	0.996
FPR	0.033	0.0017	0.0	0.001	0.001

Table 3: Results Obtained from DBK-means clustering

Graphical representation of the values for accuracy, detection rate and FPR are given as:

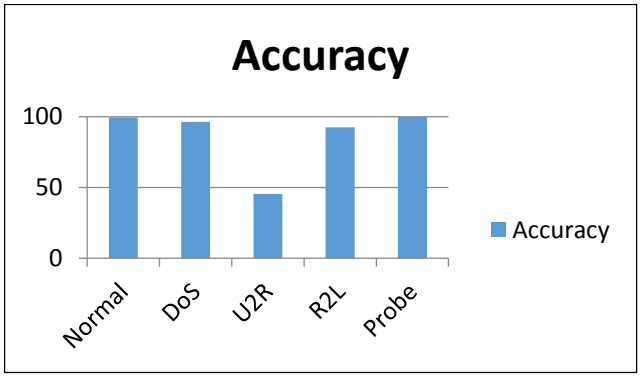

Fig.3. Graph depicting the values of accuracy for classes of the attack

As the graph depicts, for the probe attack class and for normal queries, it gives the higher values whereas it behaves in opposite for U2R attack class.

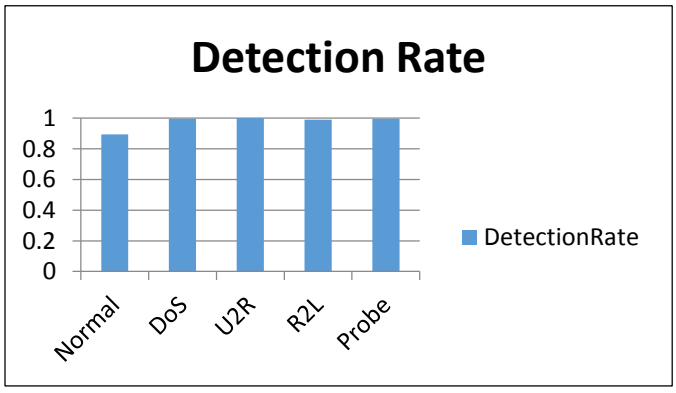

Fig.4. Graph depicting the values of Detection Rate for classes of the attack

The graph above (Fig. 4) evidences the fact that the proposed model for IDS gives the excellent detection rate for all the attack types. For normal queries, the detection rate is less but it is considerable.

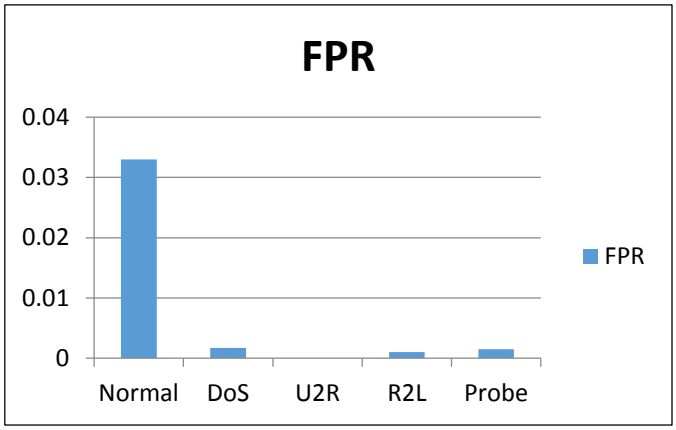

Fig.5. Graph depicting the values of False Positive Rate for classes of the attack

For all the four classes of attack types, values of the false positive rate are less than 0.002. Hence, the False alarm rate is less for the or near to null for attack classes. However, for normal queries, it is 0.03.

6. CONCLUSION

In this paper, the density based k-means clustering (DBK-means) based model of IDS is proposed. The proposed model works on the dataset and makes specific patterns of network attacks and normal queries by labeling the dataset. Using these specific patterns, the proposed model runs over the dataset and detects the attack using the novel density based K-means clustering. The results show the improved accuracy and detection rate with lower false positive rate and support the fact that the proposed methodology can be used for intrusion detection system. The future scope for the methodology is to use the corrected dataset, improving the results for normal queries and also the model can be tested for the parameters other than the listed in the paper.

REFERENCES

[1] Alan Bivens, Mark Embrechts, Chandrika Palagiri, Rasheda Smith, and Boleslaw Szymanski, "Network-Based Intrusion Detection Using Neural Networks", *Artificial Neural Networks In Engineering, St. Louis, Missouri, November 2002*.

[2] Ayman I. Madbouly, Amr M. Gody, Tamer M. Barakat, "Relevant Feature Selection Model Using Data Mining for Intrusion Detection System", *International Journal of Engineering Trends and Technology (IJETT) Vol. 9, No. 10, March 2014.*

[3] Chang, Ray-I, Liang-Bin Lai, Wen-De Su, Jen-Chieh Wang, and Jen-Shaing Kouh. "Intrusion Detection by Back propagation Neural Networks with Sample-Query and Attribute-Query", *International Journal of Computational Intelligence Research, Vol. 3, No.1, 2007.*

[4] Evgeniya Petrova Nikolova & Veselina Gospodinova Jecheva "An Adaptive Approach of Clustering Application in the Intrusion Detection Systems", *Open Journal of Information Security and Applications, Vol. 1, No. 3, December 2014.*

[5] Moradi , M., Zulkernine, M.,,"A Neural Network Based System for Intrusion Detection and Classification of Attacks", *Natural Sciences and Engineering Research Council of Canada (NSERC).*

[6] Omar Al-Jarrah & Ahmed Arafat, "Network Intrusion Detection System Using Attack Behaviour Classification", 5th International *Conference on Information and Communication Systems (ICICS), 2014.*

[7] Riad,Elhenaway, Hassan, and Awadallah. "Visualize Network Anomaly Detection by Using K-means Clustering Algorithm " *International Journal of Computer Networks & Communications (IJCNC) Vol.5, No.5, September 2013*

[8] Sharma S. K., Pandey P., Tiwari S. K., Sisodia M. S., "An Improved Network Intrusion Detection Technique based on K-means Clustering via Naïve Bayes Classification ",*Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on [proceedings] : date, 30-31 March 2012*. Piscataway, NJ: IEEE, 2012.

[9] Sumit More, Maru Mathews, Anupam Joshi Tim Finin, "A Knowledge-Based Approach To Intrusion Detection Modeling", *IEEE CS Security and Privacy Workshops, 2012.*

[10]Tawallee M., Bagheri E., Lu W., Ghorbani A.,"A detailed analysis of KDD cup 99 data set",*Proceedings of 2009 IEEE Symposium on computational intelligence in Security and Defence Applications(CIDSA),2009.*

[11]<http://home.deib.polimi.it/matteucc/Clustering/tutorial_html>

[12]http://www.draw.io

[13]Xue-Yong, L., Guo-Hong, G., & Jia-Xia, S. (n.d.). "A New Intrusion Detection Method Based on Improved DBSCAN" WASE International Conference on Information Engineering,2010

[14]Paliwal, swati.,Gupta,Ravinder.,."Denial-of-service, Probing and Remote-to-user (R2L) Attack determining using Genetic Algorithm" International Journal of Computer Applications (IJCA) Vol. 60, No.19, Dec2012.

[15]Guan, Y., Ghorbani, A., & Belacel, N. (n.d.). "Y-means: A clustering method for intrusion detection" Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (CCECE),2003

[16]Tian, L., & Jianwen, W. (n.d.). "Research on Network Intrusion Detection System Based on Improved K-means Clustering Algorithm". International Forum on Computer Science-Technology andApplications.,2009