International Journal of Computer and Communication Technology

Volume 1 | Issue 4 Article 5

October 2010

Intrusion detection using clustering

Kusum Kumari Bharti
M.Tech (C.S.E) 1, Assistant Professor M.A.N.I.T. . Bhopal ,India, bharti.kkusum@gmail.com

Sanyam Shukla
M.Tech (C.S.E) 1, Assistant Professor M.A.N.I.T. . Bhopal ,India, sanyamshukla@manit.ac.in

Sweta Jain
M.Tech (C.S.E) 1, Assistant Professor M.A.N.I.T. . Bhopal ,India, shweta_j82@yahoo.co.in

Intrusion detection using clustering

Kusum Kumari Bharti [1], Sanyam Shukla [2], Sweta Jain [3]
M.Tech (C.S.E) [1], Assistant Professor [2, 3]
M.A.N.I.T. .
Bhopal ,India
bharti.kkusum@gmail.com [1], sanyamshukla@manit.ac.in[2], shweta_j82@yahoo.co.in [3]

Abstract

In increasing trends of network environment every one gets connected to the system. So there is need of securing information, because there are lots of security threats are present in network environment. A number of techniques are available for intrusion detection. Data mining is the one of the efficient techniques available for intrusion detection. Data mining techniques may be supervised or unsuprevised.Various Author have applied various clustering algorithm for intrusion detection, but all of these are suffers form class dominance, force assignment and No Class problem. This paper proposes a hybrid model to overcome these problems. The performance of proposed model is evaluated over KDD Cup 1999 data set.

Keywords: K-Mean, J48 decision tree, Random Forest, CFSSubEval Feature selection (FS), KDDCUP 1999, IDS, DoS, U2R, R2L, Probe

I. Introduction

Intrusion detection system deal with supervising the incidents happening in computer system or network environments and examining them for signs of possible events, which are infringement or imminent threats to computer security , or standard security practices [1].

1.1 Intrusion detection system can be categories as:

A. Host based intrusion detection System

Host based intrusion detection system mainly deal with single computer and perform intrusion detection based on the system call, kernel, firewall and system logs. Based upon these concepts they can be categories as [3].

- File system monitor
- Log file analyser
- Connection analyzer
- Kernel based analyzer

1.1.2 Network Based Intrusion Detection System:

Network intrusion detection system works on large scale. It monitors network traffic and examines the traffic, and based upon the observation it categorizes the traffic into normal or suspicious. Traffic monitoring is done at firewall, hub and switch etc [2].

1.2 Clustering

Clustering is an unsupervised learning technique which divides the datasets into subparts, which share common

properties. For clustering data points, there should be high intra cluster similarity and low inter cluster similarity. A clustering method which results in such type of clusters is considered as good clustering algorithm.

1.2.1 Clustering *methods can be classified as* [3]

1.2.1.1 Hierarchical clustering

In hierarchical clustering data are not gets clustered at ones instead stepwise procedures is followed for clustering the datasets

Hierarchical clustering can be further classified as

A. Division clustering

In division clustering formation of clustering whole data point is considered as a single cluster and formation of new clusters starts from the whole data point to single datapoint. It starts form root to leave.

.

B. *Agglomerative Clustering*

In this type of clustering consider each data point as a cluster, and formation of the clusters starts by combining two instances based upon the certain criteria. It starts form leave to root.

1.2.1.2 Partitional clustering

In this data points are divided into k subparts based upon certain relevance criteria.

***A*. K-Mean Clustering**

In K-Mean clustering, assignment of the data points to clusters is depend upon the distance between cluster centroid and data point.mainlyThere are three variation of k-mean clustering 1

k-mean: which is used for numerical data sets.2 k-mediod : It is used for categorical datasets and 3 k-prototype: It is used for both categorical and numerical dataset.

B. Fuzzy C Mean Clustering

Another variation of K-mean clustering algorithm is Fuzzy C Mean. In this clustering algorithm along with the calculation of distance, membership of the data points with the cluster are also considered.

C. QT Clustering

QT (Quality Threshold) Clustering is an algorithm that groups datapoint into clusters. Quality is ensured by finding large cluster whose diameter does not exceed a given user-defined diameter threshold value [4].In this paper, a we propose a cluster classifier combinationl for removing clustering related problems. In Section 2, we present theclustering related works: in Section 3, we discuss problem statement and proposed model: in Secetion4 we evaluate the performance of proposed model and summarized: and Section 5 concludes the paper.

II. CLUSTERING RELATED WORK:

The following section explores the use of clustering in the field of intrusion detection system.

2.1 Optimized Sampling with Clustering Approach for Large Intrusion Detection Data

In [5] authors have been used K-Mean evolution clustering method for intrusion detection. This method decreases the overhead of performing the detection over whole datasets. As it work on partition of data sets and this results in increase in the processing speed of the clustering method. From the experimental result it is shown that using this method 620 clusters are formed from 5000 data points with standard deviation of .03.But sometimes it gives wrong result because it performs detection on the sample datasets.

2.2 Clustering Based Method for Unsupervised Intrusion Detection

In [6] a new clustering algorithm for intrusion detection CBUID has been proposed. It uses outlier factor for checking the deviation degree of a cluster. Data classification is performed by using an improved nearest neighbour Clustering method. The clustering algorithm used is incremental in nature. CBUID consist of three models: setting the model, detection model and update model. In setting the model, clustering is performed on the training set. After that clusters are arranged according to the outlier factor and finally summarizing the cluster. In detection model cluster corresponding to the new data points is determined. Lastly the model can be updated if new type of attack is discovered and relabelling of the clusters can be done through setting model.
They has selected, r in the range [EX-0.25DX, EX+0.25DX], and selecting parameter for outlier factor has been taken as 1.
They have used two datasets KDD99 and DARPA. Performance metric are total detection rate; detection rate for new attack and false alarm rate number of the cluster for KDDCUP99 is 97,61,42,26 and 14. It is observed that as r increases the number of clusters decreases.

Total detection rate for week 1 with r=05 is 94.82 and for r=0.6 is 90.28, for week 2 with r=0.5 is 88.98 and for week 2 with r=.6 is 90.28.flase alarm rate for week 1 with r=0.5 is 3.83 and with r=0.6 is 0.07 and for week 2 with r=0.5 is 3.75 and with r=0.6 is 0.07. For KDD1999 dataset, detection rate for all types of attacks is 98.53-98.65, false alarm rate lies in range 0.05-1.30 and detection rate for new attack is 32.44-42.12But there is need of improvement for back, pod, rootkit, and warezclient and warezmaster attacks.

2.3 Clustering Algorithm to Enhance the Performance of the Network Intrusion Detection System.

In [7] authors have used two types of clustering algorithm, soft clustering and hard clustering. K-mean clustering is an example of hard clustering and fuzzy K-Mean is the example of the soft clustering in which data point may belong to more than one cluster. In k-mean method first of all k is selected which is the number of the clusters, this also represents the number of centroids. Next step is to calculate distance between the centroids and the data points. There are three method of calculating the distance as mentioned in [8]. Data points are then allocated to the cluster based upon the shortest distance criteria. The process of calculation of the distance and assignment of the data points is repeated until centroid gets stabilized.
In fuzzy c mean a modified K-Mean method is used, where membership values of each data point corresponding to that cluster is also calculated. They have applied their clustering algorithm on KDD datasets 99 datasets and they have used Roc curve in order to show results. One of the major limitations of this approach is deciding the membership value of the data points. This approach does not give good results for the large datasets.

2.4 Y-Mean: A Clustering Method for Intrusion Detection

In [9] intrusion detection another modification to K-Mean clustering algorithm has been proposed. This modified K-Mean clustering algorithm is called as Y-Mean clustering. It overcomes the shortcoming of K-mean clustering mainly number of cluster dependency and degeneracy Y-Mean clustering is summarized as follows

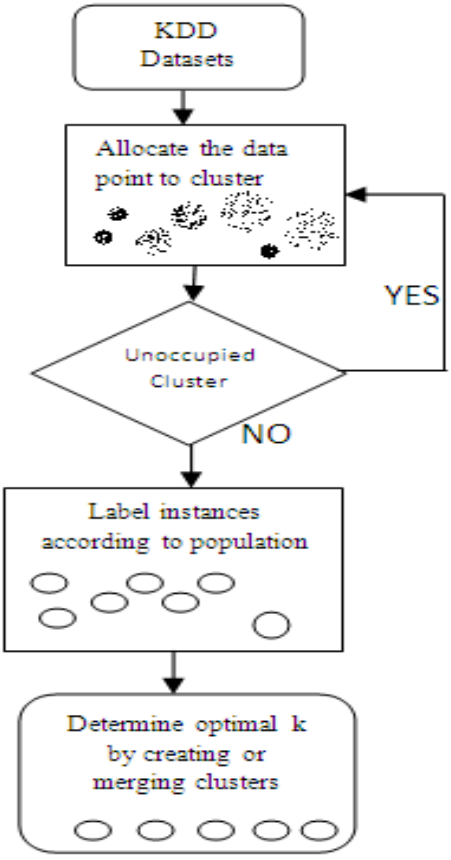

Fig: 1 Y-Mean

Y-Mean is tested on KDD CUP-99 data sets. They have used 2.32 SD as the threshold. On an average, the Y-|Mean algorithm has a detection rate of 86.63% and false alarm rate of 1.53%.The algorithm gives best results for k=20 with detection rate of 89.89% and false alarm rate 1.00%.Traning over unlabeled datasets and testing over labelled datasets results in detection rate of 82.32% and false alarm rate 2.5%. However it has a detection rate of less than 90% for every dataset

2.4 Data Clustering for Anomaly Detection in Network Intrusion Detection

In [10] authors have used hierarchical clustering methods for intrusion detection. In hierarchical clustering method, from creation of the clusters start form either top to bottom or bottom to top. The author has used bottom to top approach .In bottom to top approach, aggregation of the data points start form the single data point, it then clusters the data point according to the distance. Firstly the number of the clusters and degree of the membership are decided. Next step is to identify normality of the data points and calculation of the deviation degree from the signature are set the threshold values for checking the deviation degree, if deviation degree exceed from these values then that point is called anomalous. These steps are repeated until entire data sets get clustered. Detection rate and false alarm rate are used as performance metrics. For creating the ROC curve cubic interpolation and optimum theory are used. For new data points, the distance between the data points are calculated for every labelled cluster, allocate the data points according to the minimum distance and label according to the cluster label. Author has used DARPA 1998 datasets in which degree of the membership varies with the number of the clusters. They have achieved high detection rate using 40 clusters and degree of membership is 10.Overall Detection rate is .99 and false alarm rate is .25. Detection rate attack is 50%. for known attack is 70% and detection rate for unknown

2.6 K-mean+ID3: A novel methods for intrusion detection

In [11] Author has used k-mean followed by ID3 decision tree, author is mainly focusing on removing the shortcoming of k-mean clustering which is force assignment and class dominance problem. For removing these problem authors have used K-Mean followed by ID3 decision tree. Proposed model is divided into 2 phase 1.Candidate Selection 2.Candidate Combination. In candidate selection f cluster that are nearest are chosen and in candidate combination use nearest consensus or nearest neighbour for combining the candidates.

2.7 Mixed Unsupervised Clustering Based Intrusion Detection Model

In [12] modified K-mean clustering algorithm called KD clustering has been used for intrusion detection. In this clustering algorithm, Set S is initialized to null where S is the collection of clusters. For allocation of the data points, checked , if S is null then a new cluster is built and added it to S. Otherwise the cluster which gives minimum distance between the data point and the cluster is found, next the distance is compared the result with the cluster radius threshold, if it is less than the defined threshold then the data point is added to the cluster otherwise new cluster is created with the data point and is centroid of the cluster the recalculated the until cluster center does not change.

Algorithm has been simulated on KDD 1999 datasets. For DOS cluster radius has been taken as 1.75 which gives the detection rate of 99% and false alarm rate 1.67%. For R2L with cluster radius 2.4, the algorithm gives a detection rate of 84% and false alarm rate 0.88%. For probing attacks with cluster radius 2.1 that gives the detection rate of 99% and false alarm rate is 0.36% .For mix type of attacks with cluster radius equal to 2.1 gives detection rate 90% and false alarm rate2.2. From the experimental result it is clear that detection rate for R2L is not so good so there is need of improvement for R2L types of attack. Detection rate for R2L is better in k-mean clustering than KD.

2.8 Labelling Cluster in an Intrusion Detection System Using a Combination of Clustering Evaluation Techniques.

Author Slobodan petrovic et.al [13] have used Davies Bouldin index for clustering and centroid diameters for labeling the datasets. The architecture they have used consists of three main components: Sensor, in which clustering algorithm is implemented, assessor in which labeling algorithm is implemented and manager which adjusts sensor and assessor

for giving optimized results. K-Mean clustering algorithm has been used in the sensors which uses Euclidean distance for calculating the distance. There are many different types of the cluster evaluation methods as mentioned in [14] but here Davies-Bouldin index has been used for clustering evaluation measure there are different methods for calculating the inter-cluster distance and intra-cluster distance as given in [15]. In addition of using the Davies--Bouldin index author has they have used centroid linkage for calculating the inter cluster distance. For labelling the datasets they partition the datasets into clusters and assign a label .Then they, calculate the Davies-Bouldin index threshold and centroid diameter difference threshold. These values are then used for relabeling the cluster.

Algorithm has been tested on KDD 1999 datasets and Roc curve for the performance metric. ROC curve depicts the relationship between False Positive Rate and True positive Rate.

The algorithm gives best results on 0% filtered datasets for FPR and TPR over ROC curve. It gives same results as other clustering algorithm on 98% and 99% filtered datasets. Centroid diameter has been taken as 500 for first cluster and zero for second cluster. Davies—Bouldin threshold is between 0.2 and 0.45 and it performs well in heavily attacked environments. Nothing has been mentioned about the network parameters which should be used.

2.9 An Intrusion Detection System Based on the Clustering Ensemble

In [16] a new clustering algorithm called ensemble algorithm has been introduced which combines different clustering algorithms. The author has tested already exiting algorithms using different parameters. EAIDS pre-process the datasets and then create the classifier based on EA algorithm. The algorithm is shown in fig. 2

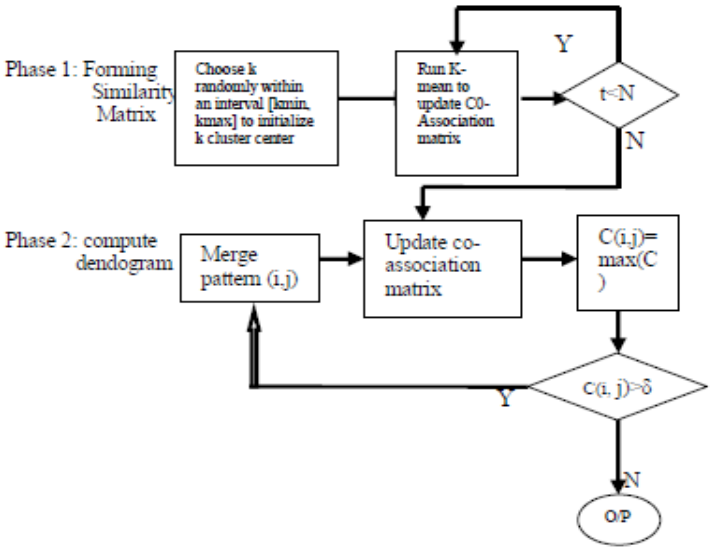

Fig:2 Ensemble Clustering

Algorithm has been experiment on KDD99 dataset has been used for testing the performance of proposed clustering algorithm .Performance metrics used are Detection rate and False alarm rate. threshold value has been taken between values of 0.5 to 0.7 .It achieves the best performance on threshold 0.6.For DoS detection rate is 100% and false alarm rate is 0.779%, for probing detection rate is 100% and false alarm rate is 1.0250%, for R2L detection rate is 100% an false alarm rate is 1.0104% and for U2R detection rate is 100% and For mixed attack false alarm rate is 0.8928%. For mixed attack DR is 100% and FA is 0.7736%. Further improvement is needed for low false alarm rate.

2.10 The application on intrusion detection based on K-means Cluster Algorithm

In [17] author have been used K-mean algorithms for performing the clustering and mean standard deviation has been used for converting the data sets into standard format. Algorithm has been tested on KDD CUP 99 datasets for intrusion detection system. Proposed algorithm performs well for all types of attacks .Detection rate for different attacks are 99.06%-99.88% and false alarm rate are 0.12%-3.94%.False alarm rate for some attacks is greater than 2% which is not tolerable.

2.11 A new approach to intrusion detection using Artificial Neural Network and fuzzy clustering.

Main [18] problem of ANN based IDS is lower detection precision for low frequent attack and weaker stability. For removing these shortcoming of ANN author have introduce. In model which is based on FC-ANN. Author has divided their model in three phases: 1 Fuzzy Clustering Module, 2 ANN modules and 3 Fuzzy Aggregation Module. They successively achieve the high recall and precision for low frequent attacks (U2R, R2L).

III. Problem statement

There are two Techniques for Intrusion Detection: Signature based intrusion detection and Statistical Anomaly based intrusion detection. Signature based intrusion is also known as misuse detection. Signature based IDS detect the intrusion by analyzing the network traffic or host system logs and compare it with some known predefined or preconfigured patterns. If captured pattern matches with these pre-established pattern then it will characterized as intrusive otherwise as normal. Main drawback of misuse detection, it is not able to characterize new pattern as intrusive or normal because of new and unknown signature this increases which increases the false alarm rate.

While in Statistical Anomaly based intrusion detection which is also known as anomaly based intrusion detection system. It removes the shortcoming of misuse detection. It is not only able to identify the known pattern but also the unknown pattern. Anomaly based intrusion detection is a system to identify the attack by comparing the traffic and system logs

with normal profile and classify it as normal or suspicious. But one of the major drawbacks of Statistical Anomaly based intrusion is high false alarms rate.

For removing these shortcomings of misuse detection and anomaly detection profiles should be updated with large amount the datasets at regular interval of time [19].But large amount the datasets also increases the problem of inconsistency, redundancy and ambiguity.

Several datamining techniques have been applied for intrusion detection. K-Mean Clustering is unsupervised datamining techniques for intrusion detection. K-Mean clustering is easy to implement. Three major drawback of K-mean clustering is: 1 class dominance problem, 2.force assignment problem, and 3. No class Problem. It has been observed that single model cannot give better result in terms of recall and precision.

3.1 Proposed Model

For removing all these problems are proposing a new model which is based on feature selection as a first phase, K-Mean clustering model generation as a second phase, classification of this new dataset which is generated by second phase as third phase, and finally evaluating the performance of this proposed model in terms of precision and recall.

Feature selection: In machine learning, feature selection techniques can be characterized into filter and wrapper. Filter method select the subset of features by using relevance criteria and totally ignoring the effect of the selected feature subset on the performance of induction algorithm. Filter method may be further characterized into Focus and Relief algorithm. In FOCUS algorithm it selects the minimal subset of the feature which tightly satisfies the relevance criteria. In RELIEF algorithm it select the maximum subset of feature which may be satisfy the relevance criteria up to the some extends. Feature Wrapper approach have higher accuracy than filter approach because when it conducting the search for good subset of feature it assess induction algorithm as a part of evaluation functions. For measuring the accuracy of selected feature of wrapper method cross validation is used. Finally subsets of feature are ranked based upon performance measures.

One of the major drawbacks of the wrapper methods is more time consumptions than filter method. So it becomes impractical for applying it on large datasets. Filter methods have less time consummation as compared to wrapper methods. It is mostly used for large datasets or real time datasets. In this paper we are using CfsSubSetEval methods along with BestFirst search. BestFirst search method uses the greedy hill climbing along with the backtracking capability. In CfsSubSetEval [19, 20, 21, 26] it selects the subset of the feature which is highly correlated to class and less correlated to each other. We have used conditional entropy for measuring the correlation between the feature and class and between the features an also measured the correlation and dependency features. If x and y are variables of the datasets along with range R_x and R

$$H(Y) = -\sum_{Y \in Ry} p(y) \log ((y)) \quad 1$$

$$(|) = - \sum_{Y \in Rx} p(x) \sum_{Y \in Ry} p (y|x) \log (p (y|x)) \quad 2$$

Uncertainty coefficient which is used for measuring the dependency and correlation is defined as follows.

$$C (Y|X) = \frac{H (Y) - H (Y|X)}{H(Y)} \quad 3$$

The output of the feature selection algorithm is used as input in the second phase.

Second phase of proposed model is K-Mean which is an unsupervised clustering algorithm-mean clustering algorithm was developed by J.MacQueen (1967) and then by J.A Hartigan and M.A.wong around 1975[22].K-Mean clustering involves 4 steps.

Input:
1. Training and test datasets
2. Number of clusters k

Output:
Datapoints in form of k clusters

Initialization: First all define the number of clusters, and randomly initial centroid of the clusters

Assignment: Assign the data point to their corresponding cluster based upon the least distance between the datapoint and cluster centroid

Recalculation: After assignment of all datapoint to their corresponding cluster recalculate the clusters centroid.

Repeat step 2 which is an assignment of the datapoint to the clusters.Untill there is no further variation of the cluster centroids.

Finally: Assign the datapoint to their corresponding cluster. We have created the model on training datasets and then apply this model on test datasets. For cluster to class mapping we have used cluster as a new attribute in the current datasets.

Final phase of the proposed model is classification phase which uses J48 and Random Forest. Both of the classification method is parts of decision tree main difference between J48 and Random Forest [23, 24, 26] is Random Forest is use multiple tree and J48 uses one tree. Root node is selected based upon the information gain of the attribute. Random Forest and J48 are used model generation on the new datasets with n+1 attributes. Finally evaluating the accuracy of the proposed algorithm by applying the classification model on test datasets.

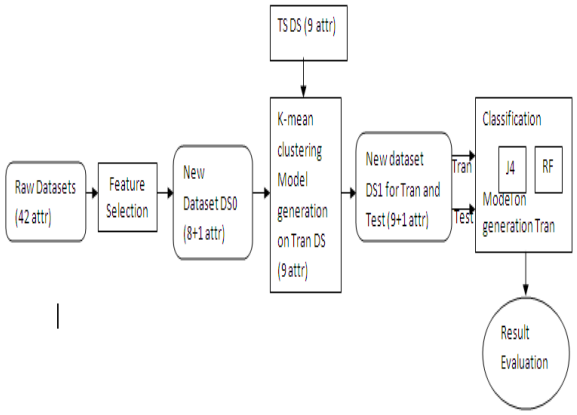

Fig.3. Proposed Model

IV. Evaluation of Proposed Model

To evaluate the performances of proposed model we have used KDD cup 1999 datasets which consists of around 4 lacks of instances for training and around 3 lacks record for testing. The datasets consists of 42 attributes. Each of the instances of trainings data sets label by normal or specific types of attacks .There are four types of attacks in network environment.1 DoS(Denial of service), Attackers makes system resources unavailable the for authorized users Ex. Mail bomb,SYN Flood, Ping of death, Teardrop and Smurf. 2. Probe: In this type of attack, attacker uses software or hard ware for gathering the information about target system.Ex.Lpsweep, MScan, Saint and Nmap.3. U2R. This is one of the serious type of attack in which authorized user try to access the privilege of root user for which they are not authorized.Ex.Ps,Xterm,Perl and Fdformat.4.R2L.External intruders try to access the privilege of the for internal users.Ex.Ftp_write,Phf, and Guest.

4.1 Performance metric

Following measure have been used for measuring the performances of Proposed IDS: True Positive: Number of connection that Was correctly classified as attack True Negative: Number of connections that were incorrectly classified as attack. False Positive: It is also known as Type I error, error of the first kind and α error in this number of attack connection that were classified as normal. False Negative: It is also known as Type II error, error of the second kind, and β error, It is more riskier than false positive ,in this type of attack number of normal connection that were classified as attack. Finally measuring the accuracy of proposed we have used precision and recall as a performance metric. Recall is used for determine how many miss classification are there and precision is used for determining how may correctly classified [25]

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

4.2 Evaluation of Result

Table 1.1 with Number of Cluster k=4

Attack	K-Mean 4		Hybrid (J48)		Hybrid Random Forest	
	Precision	Recall	Precision	Recall	Precision	Recall
Normal	0.7096	0.9829	0.684	0.994	0.675	0.995
DoS	0.9939	0.7142	0.995	0.946	0.999	0.946
Probe	0.0656	0.6882	0.926	0.859	0.935	0.855
U2R	NA	NA	1	0.018	0.545	0.026
R2L	0.0001	0.0001	0.927	0.029	0.713	0.005

Table 1.2 with Number of Cluster k=5

Attack	K-Mean5		Hybrid (J48)		Hybrid Random Forest	
	Precision	Recall	Precision	Recall	precision	Recall
Normal	0.7431	0.7380	0.677	0.995	0.677	0.994
DoS	0.9993	0.6049	0.999	0.946	0.999	0.946
Probe	0.0659	0.0659	0.936	0.853	.924	0.855
U2R	NA	NA	0.364	0.018	0.545	0.026
R2L	0.381	0.6115	0.926	0.026	0.932	0.906

Table 1.1 with Number of Cluster k=6

Attack	K-Mean6		Hybrid (J48)		Hybrid Random Forest	
	Precision	Recall	Precision	Recall	Precision	Recall
Normal	0.7861	0.7039	0.676	0.995	0.677	0.994
DoS	0.9966	0.7138	0.999	0.946	0.998	0.947
Probe	0.0654	0.6867	0.934	0.829	0.921	0.81
U2R	0.00099	0.1228	1	0.018	0.667	0.026
R2L	0.4997	0.0607	0.927	0.027	0.943	0.024

Proposed model gives better results than k-mean clustering over KDD cup 1999 datasets for all types of attack (DoS,Probe,U2R,and R2L). For normal class, k-mean clustering gives better result than proposed model as the number of clusters increases precision of k-mean clustering increases and at the same time recall decrease for k-mean clustering precision is from 70.96%-98.51% and recall is from 98.29%-38.12%.Value of precision and recall varies with the value the value of k. For Dos attack hybrid model with Random forest gives the best result. Precision of DoS class using this model is 99.9% and recall is from 94.6%-94.8%. Hybrid J48 and Random forest both gives the better result for Probe ,in J48 precision increases and decreases as varying the increasing the of k and recall decreases. For Random forest both precision and recall both decreases with the increasing the value of k. For U2R and R2L class hybrid model with Random forest gives best result. Precision for U2R is 54.5%-66.7% and recall is 2.6%. Precision for R2L is 71.3%-95% and recall is 0.5%-2.6%.

V. Conclusion and future works

From the above discussion it has shown that proposed model gives better results for DoS, Probe, U2R and R2L all types of attacks .Our proposed algorithm have higher precision and recall rate. This paper has introduced the new methods for cluster to class mapping which increases the accuracy of the model for all types of attacks. It has been also observed from results that as value of k changes than corresponding precision and recall also increases and decreases. Accuracy of k-mean clustering depends upon the value of k. determining the appropriate number of clusters is challenging area for researchers. Another clustering algorithm like fuzzy k mean clustering and other classification methods can be used for or other data mining techniques and it can also be used for determining the value of cluster (k).

VI. References

[1] http://en.wikipedia.org/wiki/Intrusion_detection_system.

[2] Pieter de Boer & Martin Pels, Host-based Intrusion Detection Systems, Revision 1.10 – February 4, 2005. pp.19-20.

[3] Paul Dokas,Levent Ertoz,Vipin Kumar,Aleksandar Lazarevic, Jaideep Srivastava, Pang- Ning Tan, Data Mining for Network Intrusion Detection

[4] http://www.chem.agilent.com/cag/bsp/products/gsgx/Downloads/pdf/qt_clustering.pdf

[5] Nani Yasmin1, Anto Satriyo Nugroho2, Harya Widiputra3," Optimized Sampling with Clustering Approach for Large Intrusion Detection Data", International Conference on Rural Information and Communication Technology 2009 Pp.56-60

[6] Sheng Yi Jiang, Xiaoyu Song, Hui Wang, Jian-Jun Han Qing-Hua Li, "A clustering based method for unsupervised intrusion detection", Elsevier, vol-27, Issue 7 (May 2006) 2006, pp. 802-810

[7] Mrutyunjaya Panda, Manas Ranjan Patra,"Some Clustering intrusion detection system", Journal of Theoretical and Applied technology, 2005-2008,pp.710-716

[8] http://en.wikipedia.org/wiki/Cluster_analysis

[9] Yu Guan and Ali A. Ghorbani, Nabil Belacel,"Y-Mean: A Clustering method For Intrusion Detection", 1CCECE 2003, pp.1-4

[10] Jose F.Nieves "Data clustering for anomaly detection in Network intrusion detection", Research Alliance in Math and Science August 14, 2009,pp.1-12

[11] Shekhar R. Gaddam, Vir V. Phoha, Kiran S. Balagani, "K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods," IEEE Transactions on Knowledge and Data Engineering, vol.19, no. 3, Mar. 2007 pp. 345-354.

[12] Cuixiao Zhang; Guobing Zhang; Shanshan Sun, **"**A Mixed Unsupervised Clustering-based Intrusion Detection Model ", Third International Conference on Genetic and Evolutionary Computing, 2009, pp.426-428

[13] Slobodan Petrovic, Gonzalo Alvarez, Agustin Orfila and Javier Carbo," Labeling Cluster in Intrusion detection System Using a Combination of Clustering Evaluation Techniques", Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06) Track 6, 2006, vol. 6, pp.129b

[14] machaon.karanagai.com/validation_algorithms.html

[15] http://machaon.karanagai.com/distances.html#INTRA

[16] Fangfei Weng, Qingshan Jiang, Liang Shi, and Nannan Wu,"An Intrusion Detection System Based on the Clustering Ensemble", IEEE International workshop on 16-18 April 2007,pp.121 - 124

[17] Meng Jianliang Shang Haikun Bian Ling, "The Application on Intrusion Detection Based on K-Means Cluster Algorithm", International Forum on Information Technology and Application, 15-17 May 2009 ,pp. 150 - 152

[18] Gang Wang, Jinxing Hao, Jian Ma and Lihua Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering**",** Elsevier 2010 pp.6225-6232

[19] Mark A. Hall, Lloyd A. Smith, "Feature Subset Selection: A Correlation Based Filter Approach 1997

[20] T. S. Chou, K. K. Yen, and J. Luo,"Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms", International Journal of Computational Intelligence 4; 3 2008, pp.196-208

[21] K. Selvakuberan, M. Indradevi, Dr. R. Rajaram "Combined Feature Selection and classification – A novel approach for the categorization of web pages", Journal of Information and Computing Science, Vol. 3, No. 2, 2008, *pp*. 083-089

[22] http://people.revoledu.com/kardi/tutorial/kMean / WhatIs.htm

[23] http://www.d.umn.edu/~padhy005/Chapter5.html

[24] http://en.wikipedia.org/wiki/Random_forest

[25] http://en.wikipedia.org/wiki/Precision_and_recall

[26] Li Tian, Wang Jianwen, "Research on Network Intrusion Detection System Based on Improved K-means Clustering Algorithm," International Forum on Computer Science-Technology and Applications, 2009, vol. 1, pp.76-79

[27] Ian H. Witten & Eiben Frank," Data Mining Pratical Machine Learning Tools and Techniques", 500 Sansome Street, Suite 400 Francisco, CA 94111, Elsevier 2005