

Intrusion Detection using unsupervised learning

Kusum bharti

M.Tech (C.S.E.)
M.A.N.I.T.
Bhopal

Sanyam Shukla

Assistant Professor
M.A.N.I.T.
Bhopal

Shweta Jain

Assistant Professor
M.A.N.I.T.
Bhopal

Abstract— Clustering is the one of the efficient datamining techniques for intrusion detection. In clustering algorithm k-mean clustering is widely used for intrusion detection. Because it gives efficient results incase of huge datasets. But sometime k-mean clustering fails to give best result because of class dominance problem and no class problem. So for removing these problems we are proposing two new algorithms for cluster to class assignment. According to our experimental results the proposed algorithm are having high precision and recall for low class instances.

Keywords- Feature selection, k-mean clustering, fuzzy k mean clustering, and KDDcup 99 dataset


Introduction (Heading 1)

Intrusion is the sequence of the set of related activity which perform unauthorized access to the useful information and unauthorized file modification which causes harmful activity. 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.

Various techniques have been used for intrusion detection. Datamining is one of the efficient techniques for intrusion detection. Datamining uses two learning, supervised learning and unsupervised learning. Clustering is unsupervised learning which characterize the datasets into subparts based on observation. Datapoint which belong to the clusters same clusters share common property. To find similarity between data points distance measure are used. In many papers Euclidean distance measure is used for deciding the similarity between the datapoints.

This paper is organized as follow: Section I give some over view of related works, section II gives basic concept of k-mean clustering, the section III presents the architecture of the proposed model. Section IV summarizes the obtained results with comparison and discussions. Section V concludes the paper along with future works.

I. RELATED WORKS

First, Authors [1-3] have used k-mean clustering for intrusion detection. The performance of k-mean clustering affected initial cluster center and number of cluster centroid. Zhang Chen et.al[4] has proposed a new concept for selecting the number of clusters. According author [4] the number of Initial cluster for a datasets is  and after that combine or

divide the sub cluster based on the defined measures. Mark Junjie Li troids et al. [5] has proposed an extension to the standard fuzzy K-Means algorithm by introducing a penalty term to the objective function to make the clustering process not sensitive to the initial cluster centers Which make clustering to insensitive to initial cluster center. Mrutyunjaya Panda et.al [6] has used k-mean and fuzzy k-mean for intrusion detection. Sometimes k-mean clustering does not gives best results for large datasets. So for removing this problem Yu Guan et. al. [7] have introduced a new method Y- mean which is variation of k-mean clustering it removes the dependency and degeneracy problem of k-mean clustering. Sometime single clustering algorithm doesnt gives best result so for removing this problem , Fangfei Weng et.al.[8] has used k-mean clustering with new concepts which is called Ensemble K-mean clustering. Cuixiao Zhang et.al [9] have used KD clustering for intrusion detection. Some of the authors have used k-mean clustering along with the other method for improving the detection rate of intrusion detection system. Authors [9-13] have used k mean clustering along with the other datamining techniques for intrusion detection. Authors [14] have used ANN along with the fuzzy k-mean clustering for intrusion detection which removes the problem related to the ANN. All of these techniques improve the detection rate for intrusion detection but no able to solve the class dominance problem of k-mean clustering So for removing this problem we are proposing two new algorithm which removes the class dominance problem along with the no class problem. In class dominance problem low instance classes (i.e. R2L and U2R) are dominated by high instances classes. In no class problem some of the clusters are assigned to no class.

II. EXISTING TECHNIQUES

K-mean clustering is a unsupervised machine learning techniques [17], It was first proposed by James MacQueen in 1967.

Algorithm

Input: Datasets to be clustered which contains N number of instances, k=number of clusters needed, randomly select k centroids from datasets.

Outputs: datasets in form of k clusters which have achieved the convergence criteria.

Step1 (Initialization): First of initialize k number of clusters along with k number of centroids.

Step 2 (Assignment): Assign each datapoints to the corresponding cluster based upon the distance measures (Mostly Euclidean distance is used [18]).

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

Where p and q are two points in Euclidean distance.

Step3 (Recalculation): After assign each datapoints to the corresponding clusters recalculate the centroid of the cluster (mean of the clusters).

Step4 (Repeat): Repeat steps 2 and 3 until convergence criteria are not met

An [19] algorithm for partitioning (or clustering) N data points into k disjoint subsets S_j containing N_j data points so as to minimize the sum-of-squares criterion

$$J = \sum_{j=1}^k \sum_{n \in S_j} |x_n - \mu_j|^2$$

Where x_n is a vector representing the n^{th} data point and μ_j is the geometric centroid of the data points in S_j .

Along with considering the minimizing the sum of square criterion two more criteria is also considered. Inter and Intra cluster distance.

A. Existing class to cluster mapping which is used in weka

Cluster to class mapping, No class, and class dominance is a key problem in k-mean clustering. Machine learning tool WEKA [20] uses number of instances to assign a cluster to a particular class. The algorithm used by weka for cluster to class mapping is as follows

Weka_Cluster_Class_Mapping_Algorithm:

Step1. Class-wise analysis: Search the cluster, for each class which contains majority of instances of that class. After this step for each class we know the cluster number which contains maximum number of instances of that class.

Step 2. Cluster-wise analysis: In this step we analyze each cluster on the basis of results obtained in previous step.

- a. If a cluster contains maximum number of instances of only a particular class then the cluster is assigned to that class.
- b. If a cluster contains maximum number of instances of more than one class then cluster is assigned to the class with greater number of instances.
- c. The cluster which does not contain maximum number of instances for any class is assigned to no class.

In the above approach not more than one cluster can be assigned to a single class. This approach has class dominance and no class problem.

III. PROPOSED MODEL

A. Algorithm 1 Percentage wise class to cluster assignment

The Algorithm 1_Cluster_Class_Mapping_Approach:

Input: Confusion Matrix, where column contain various clusters and rows contains classes.

Step1. Calculate the purity of each class corresponding to each cluster.

Purity of class $p_{ij} =$

$$\frac{\text{Number of instances of a class belong to cluster}}{\text{Total number of instances of a class}}$$

Replace the contents of confusion matrix with purity of the class corresponding to each cluster.

Step2. Class-wise analysis: Search the cluster, for each class having highest purity for a particular class. After this step, we know the cluster number which contains high purity corresponding to each class. After this step we know the cluster number for each class which contains highest purity for a particular class.

Step3. Cluster-wise analysis: In this step we analyze each cluster on the basis of results obtained in previous step.

- a. If a cluster contains high purity corresponding to only one class then the cluster is assigned to that class only.
- b. If a cluster is having highest purity for more than one class then cluster is assigned to that class which have highest purity.
- c. The cluster which does not contain highest purity for any class that cluster assigned to no class.

This approach only removes the class dominance problem. For removing no class problem approach2 has been introduced.

A cluster can be assigned to only one class. This approach only removes the class dominance problem. For removing No class problem approach has proposed.

B. Alogorithm 2

Headings Algorithm 2_Cluster_Class_Mapping_Approach:

Input: Confusion Matrix, where column contain various clusters and rows contains classes.

Step1. Calculate the purity of each class corresponding to each cluster.

$\frac{\text{Number of instances of a class}_i \text{ belonging to cluster}_j}{\text{Total number of instances of a class}_i}$

Purity for cluster $P_{ij} =$

Replace the contents of confusion matrix with purity of the class corresponding to each cluster.

Step2. Cluster-wise analysis: for each cluster find the class with highest purity value. Allot the cluster to that class.

This approach removes class dominances problem along with the no class problem. Because in this algorithm a class can have more than one clusters which was not possible in algorithm 1.

IV. EXPERIMENTAL RESULT AND ANALYSIS

For our experiments we are using KDD CUP 99 datasets. The class attributes of original train and test datasets of KDD CUP 1999 has 42 labels. The 41 labels can be generalized as only 5 labels U2R, R2L, Probe, DoS, Normal and these five labels again can be characterize into 2 labels. Precision and recall are used as performance metric [21]:

Recall: The percentage of the total relevant documents in a database retrieved by your search.

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

Precision: The percentage of relevant documents in relation to the number of documents retrieved.

$$\text{Precision} = \frac{TP}{(TP + TN)}$$

For measuring the performance of proposed model we have created the confusion matrix in which column corresponds to the class and rows corresponds to the class. We have evaluated the performance of the proposed model over 4, 5, 6 and 7 clusters.

Experiment with varying clusters

Table 1 Distribution of Instances Over 4 Clusters

Attack	Cluster 0	Cluster 1	Cluster 2	Cluster 3
U2R	0	92	0	136
Normal	6	59555	900	132
Dos	17839	7300	164157	40557
Probe	379	817	103	2867
R2L	2	16163	1	23

Table 2 Cluster Assignment Based on Proposed Techniques

Cluster number	Class (instance Based)	Class (percentage based)	Class2 (more than on 1 cluster based on perc.)
Cluster 0	R2L	normal	Probe
Cluster 1	Normal	R2L	R2L
Cluster 3	DoS	DoS	DoS
Cluster 4	Probe	Probe	Probe

Table 3 Comparison Over Precision and Recall of Proposed Techniques

Attack	Precison	Recall	Preci sion1	Recall 1	Precisi on 2	Reca ll 2
Normal	0.710	0.983	0.0003	9.90213 E-	NA	NA
R2L	0.0001	0.0001	0.193	0.998	0.193	0.998
DOS	0.994	0.714	0.993	0.714	0.994	0.714
probe	0.0656	0.688	0.065	0.688	0.052	0.779
U2R	0	0	0	0	0	0

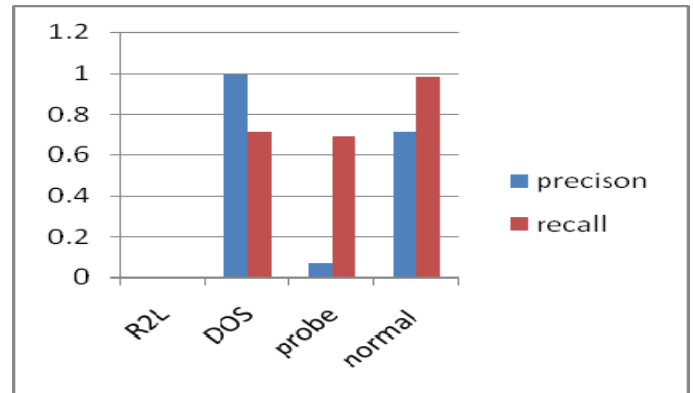


Figure 1 Instanced Based Assignment

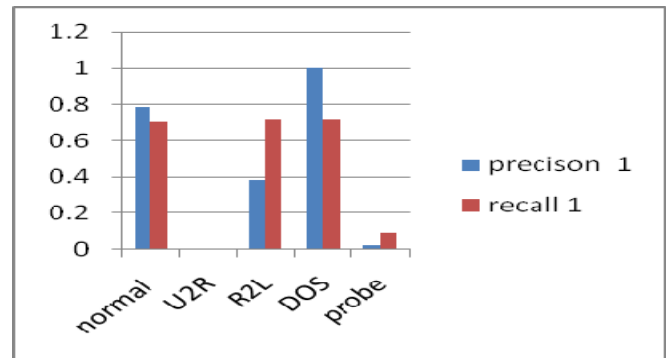


Figure 2 Assignment based on algorithm 1

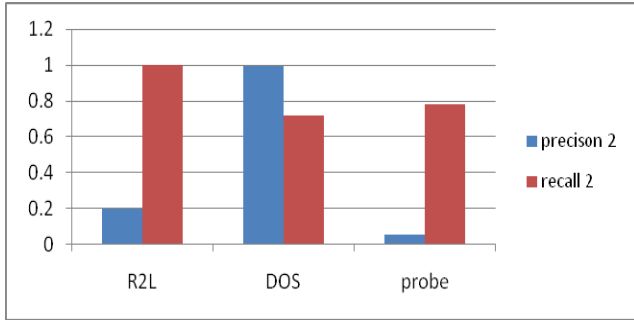


Figure 3 Percentage Based Assignment with Using More than one Cluster

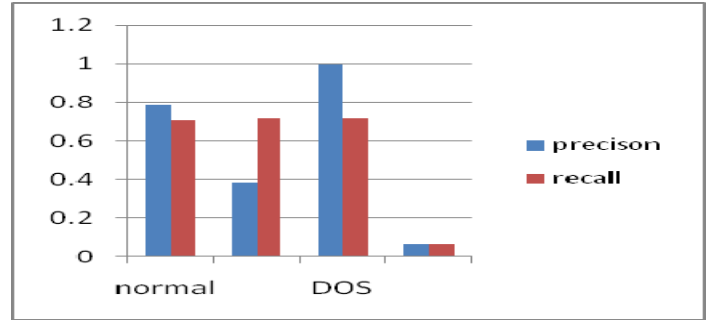


Figure 4 Instanced Based Assignment

Table 4 Distributions of Instances Over 5 Clusters

Attack	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4
U2R	0	55	0	136	37
Normal	6	42651	667	124	17145
Dos	17865	6962	164083	40567	376
Probe	379	6	3	2861	917
R2L	2	4622	0	26	11539

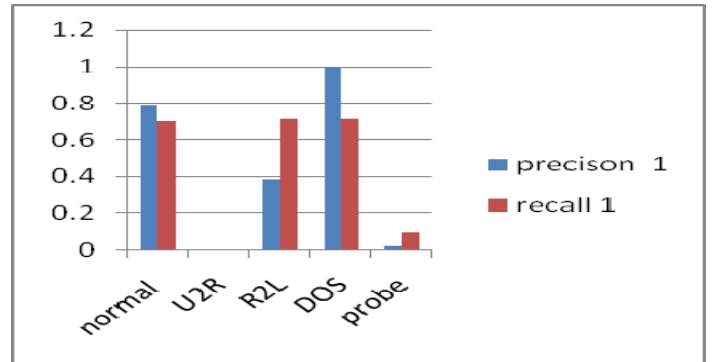


Figure 5 Assignment based on algorithm 1

Table 5 Cluster Assignment Based on Proposed Techniques

Cluster number	Class (instance Based)	Class (percentage based)	Class2 (more than on 1 cluster based on perc.)
Cluster 0	No class	Probe	Probe
Cluster 1	Normal	normal	Normal
Cluster 2	DoS	DoS	DoS
Cluster 3	Probe	U2R	U2R
Cluster 4	R2L	R2L	R2L

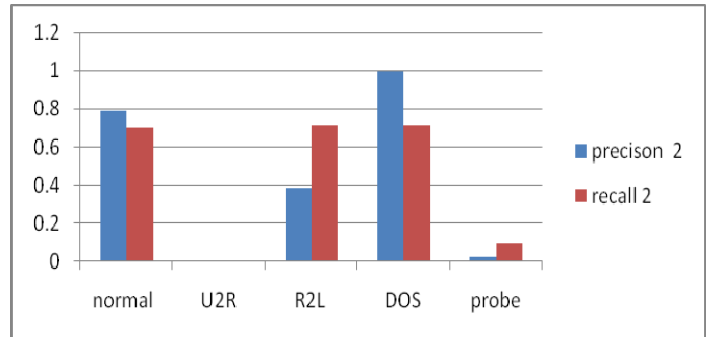


Figure 6 Assignment using algorithm 2

Table 6 compariosn of recesion and recall

Attac k	Precisio n	Recal l	Precisio n1	Recall 1	Precisi on 2	Recall 2
Norm al	0.786	0.704	0.786	0.704	0.786	0.704
R2L	0.3842	0.713	0.384	0.713	0.384	0.713
DOS	0.996	0.714	0.996	0.714	0.996	0.714
probe	0.065	0.065	0.021	0.091	0.021	0.091
U2R	0	0	0.003	0.002	0.003	0.002

Table 7 Distributions of Instances Over 5 Clusters

Attack	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster4	Cluster 5
U2R	0	28	0	135	10	55
Normal	6	16716	607	124	488	42652
Dos	17865	334	164076	40567	83	6928
Probe	379	512	3	2861	402	9
R2L	2	10566	0	26	982	4613

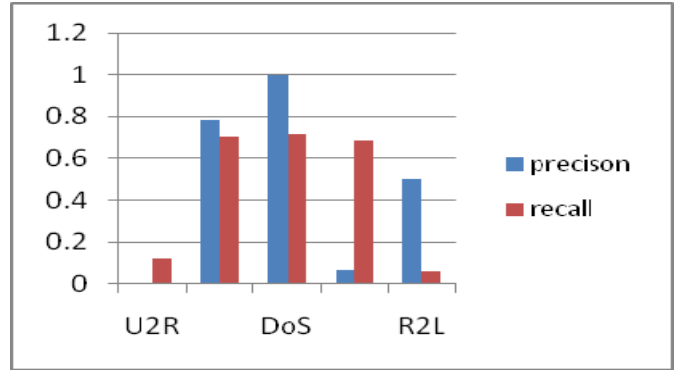


Figure 7 Instance Based Assignment

Table 8 Cluster Assignment Based on Proposed Techniques

Cluster number	Class (instance Based)	Class percentage based)	Class2 (more than on 1 cluster based on perc.)
Cluster 0	No class	No class	Probe
Cluster 1	U2R	R2L	R2L
Cluster 2	DoS	DoS	DoS
Cluster 3	Probe	Probe	Probe
Cluster 4	R2L	U2R	Probe
Cluster 5	Normal	normal	normal

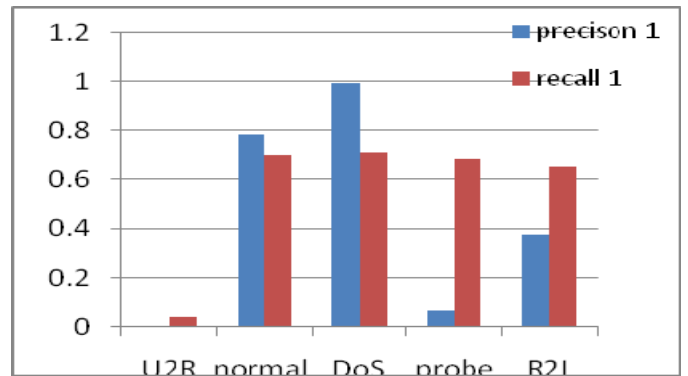


Figure 8 Assignment based on algorithm 1

Table 9 comparision of precesion and recall

Attack	Precisi on	Recal l	Precisio n1	Recall 1	Precisio n 2	Recall 2
Normal	0.786	0.704	0.786	0.704	0.786	0.704
R2L	0.500	0.061	0.375	0.653	0.375	0.653
DOS	0.996	0.712	0.996	0.714	0.996	0.714
probe	0.065	0.687	0.065	0.687	0.057	0.874
U2R	0.0010	0.123	0.005	0.044	NA	NA

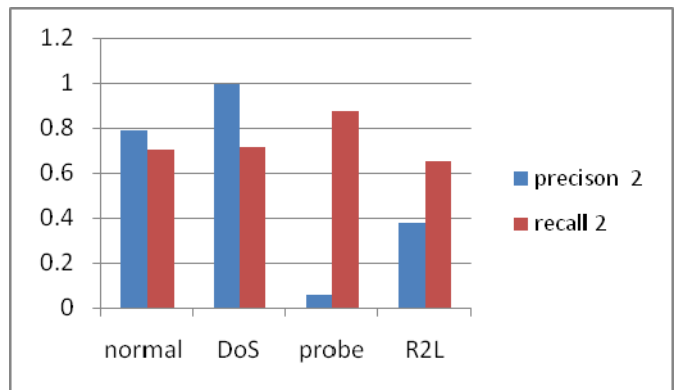


Figure 9 Assignment using algorithm 2

It is clearly shown from above experiments that the Performance of the proposed algorithm are vary from cluster to cluster. It is giving best performance for Probe, U2R, and for R2L classes. For cluster 4 proposed algorithm gives 77.9% recall for probe attack which the existing algorithm gives 6.5% recall for probe. For cluster 5 proposed algorithm are having 0.091% recall while existing algorithm are having only 0.065% recall for Probe. It also improves the recall of U2R. In cluster 6 proposed algorithms are giving best performance for probe and R2L attack. For probe it gives 87.4% recall for probe attack while the existing algorithm giving only 68.7% recall.

V CONCLUSION AND FUTURE WORK

This work explores new cluster to class mapping algorithm which increases the recall for Probe, U2R and R2L attacks. Proposed algorithms are giving the best result for Probe, U2R and R2L classes which is not possible with existing algorithm. I give best recall for cluster 6 for probe attack which is 84.4%.

The proposed algorithm is not having any significance on high instance classes. So there is need of implementing such algorithm which increases the detection rate for low instances as well as for high instances classes.

REFERENCES

- [1]. Jose F.Nieves "Data clustering for anomaly detection in Network intrusion detection", Research Alliance in Math and Science August 14, 2009, pp.1-12
info.ornl.gov/sites/rams09/j_nieves_rodrigues/Documents/report.pdf
- [2]. 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
doi.ieeecomputersociety.org/10.1109/IFITA.2009.34
- [3]. 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
asnugroho.net/papers/rict2009_clustering.pdf
- [4]. Zhang Chen, Xia Shixiong, "K-means Clustering Algorithm with improved Initial Center", Second International Workshop on Knowledge Discovery and Data Mining, 2009 IEEE, pp790-793
ieeexplore.ieee.org/iel5/4771854/4771855/04772054.pdf?arnumber
- [5]. Mark Junjie Li, Michael K. Ng, Yiu-ming Cheung, Senior Member, IEEE, and Joshua Zhexue Huang, "Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters", *IEEE transactions on knowledge and data engineering*, vol. 20, no. 11, november 2008, pp
ieeexplore.ieee.org/iel5/69/4358933/04515866.pdf?arnumber=4515866
- [6]. Mrutyunjaya Panda, Manas Ranjan Patra, "Some Clustering intrusion detection system", *Journal of Theoretical and Applied technology*, 2005-2008, pp.710-716
www.jatit.org/volumes/research-papers/Vol4No9/5Vol4No9.pdf
- [7]. Yu Guan and Ali A. Ghorbani, Nabil Belacel, "Y-Mean: A Clustering method For Intrusion Detection", *1CCECE 2003*, pp.1-4
www.jatit.org/volumes/research-papers/Vol4No9/5Vol4No9.pdf
- [8]. 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
ieeexplore.ieee.org/iel5/4244765/4244766/04244796.pdf?arnumber..
- [9]. 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
doi.ieeecomputersociety.org/10.1109/WGEC.2009.72
- [10]. 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.
doi.ieeecomputersociety.org/10.1109/TKDE.2007.44
- [11]. Mrutyunjaya Panda1 and Manas Ranjan Patra2. *Network Intrusion Detection Using Naive Bayes*. *IJCSNS International Journal of Computer Science and Network Security*, VOL.7 No.12, December 2007
citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.936&rep.
- [12]. K.S.Anil Kumar, and |Dr V.NandaMohan, "Novel anomaly intrusion detection using neuro-fuzzy interference system", *IJCSNS International journal of computer science and network security*, Vol 8 No. 8 August 2008. pp. 6-11
paper.ijcsns.org/07_book/200808/20080802.pdf
- [13]. Krishnamoorth Makkithaya, N.V.Subba reddy and dinesh acharya, "Intrusion detection system using modified c-fuzzy decision tree classifier" *IJCSNS International journal of computer science and network security*, Vol 8 No. 11 November 2008. pp. 29-35
paper.ijcsns.org/07_book/200811/20081105.pdf
- [14]. Gang Wang, Jinxing Hao, Jian Ma and Lihua Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering",
linkinghub.elsevier.com/retrieve/pii/S0957417410001417
- [15]. 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
- [16]. www.waset.org/journals/ijci/v4/v4-3-26.pdf
- [17]. http://en.wikipedia.org/wiki/K-means_clustering
- [18]. http://en.wikipedia.org/wiki/Euclidean_distance
- [19]. Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation". www.csse.monash.edu.au/~rosset/papers/cal99.pdf
- [20]. http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-Ex3.html
- [21]. http://en.wikipedia.org/wiki/Precision_and_recall