Fuzzy K-mean Clustering Via J48 For Intrusiion Detection System

Kusum Bharti[#1], Shweta Jain [*2], Sanyam Shukla [#3]

C.S.E., M.A.N.I.T.
Bhopal(Indaia)
[1]kkusum.bharti@gmail.com
[2]*shweta_j82@yahoo.co.in*
[3]*sanyamshukla@manit.ac.in*

***Abstract*— Due to fast growth of the internet technology there is need to establish security mechanism. So for achieving this objective NIDS is used. Datamining is one of the most effective techniques used for intrusion detection. This work evaluates the performance of unsupervised learning techniques over benchmark intrusion detection datasets. The model generation is computation intensive, hence to reduce the time required for model generation various feature selection algorithm. Various algorithms for cluster to class mapping have been proposed to overcome problem like, class dominance, and null class problems. From experimental results it is observed that for 2 class datasets filtered fuzzy random forest dataset gives the better results. It is having 99.2% precision and 100% recall, So it can be summarize that proposed percentage is assignments and statistical model is giving better performance.**

Keywords*— Feature selection, k-mean clustering, fuzzy k mean clustering, J48 clustering, and KDDcup 99 dataset.

Introduction

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. Most of the times distance measures are used for deciding the membership of the clusters. In many papers

Euclidean distance measure is used for deciding the similarity between the datapoints.

This paper is organized as follow: Section I gives over view of related works, section II contains framework of proposed model , section III contains experimental results and analysis, and finally Section IV 6 concludes the paper along with future works.

I. RELATED WORK

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 doesnot 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 [10-14] have used k mean clustering along with the other datamining techniques for intrusion detection. Authors [15] 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. FRAMEWORK OF PROPOSED MODEL

Redundant attributes increases the time requirements so for removing this problem in this work we have used feature selection algorithm. Main problem with k-mean clustering is it uses hard assignment for assigning the datapoints to the corresponding clusters. So for removing this problem and calculating the membership of every datapoint corresponding to every cluster we have used fuzzy k mean clustering. Another problem with clustering algorithms is cluster to class assignments. For this we have used J48 classification techniques for assigning a cluster to a particular class.

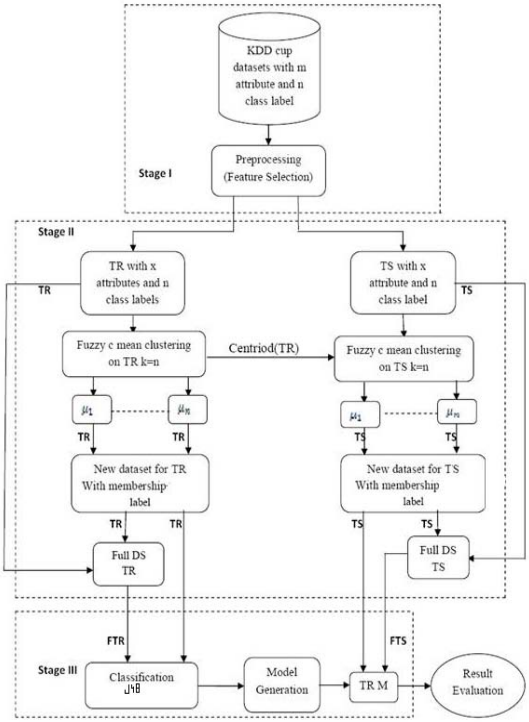

Figure 1: Proposed Model

A. *Feature Selection*

Step 1 consists of preprocessing of kddcup datasets. In preprocessing remove the redundant attribute which nis done by various feature selection algorithm. In this work we have used 3 feature selection algorithm: CFSSubSetEval, ConsistencySubSetEval, and FilteredSubSetEval[23,24].

B. *Fuzzy k mean clustering*

Fuzzy k mean is variation of k mean clustering in which a datapoint belongs o every cluster with some membership[22].

. It is based on minimization of the following objective function:

#

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} \mu_{ij}^{m}\|x_i - c_j\|^2 , \quad 1 \le m < \infty \quad \ldots\ldots\ldots\ldots 1$$

Where m is any real number is the degree of membership of x_i in the cluster j c_j is the d-dimension center of the cluster

Algorithm

Input: Set of data points , number of clusters
Output: Set of datapoonts in form of cluster along with there membership

1. Initialize membership of datapoints based upon the initial centroid U=[u_{ij}] matrix, $U^{(0)}$.
2. At k-step: calculate the centers vectors $C^{(k)}$=[c_j] with $U^{(k)}$.

#

$$c = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}} \quad \ldots\ldots\ldots\ldots 2$$

3. Update $U^{(k)}$, $U^{(k+1)}$

$$\mu_i(x_k) = \frac{1}{\sum_{j=1}^{c} \left(\frac{d^2(x_k, V_i)}{d^2(x_k, V_j)} \right)^{\frac{1}{(m-1)}}} \quad \ldots\ldots\ldots\ldots (3$$

This iteration will stop when max $\{|\mu_i(x^{k+1})-\mu_i(x^k)|\}<\varepsilon$, where ε is a termination criterion between 0 and 1, whereas *k* are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

A is a symmetric positive definite matrix, N_s is total number of pattern vectors, m is Fuzziness Index (m >1). Membership of training datasets is calculated by fuzzy c mean clustering and for test dataset use the same centroid as used in training datasets. Number of centroid for train and test datasets is equal to number of classes.

C. *J48*

A [21] decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable.

Algorithm
Training instances

1. Tree is constructed in top down recursive divide and conquer manner
2. Feature that is having the highest information gain is selected as root node of the decision tree.
3. Select the attribute that gives us the next highest information gain.
4. Repeat step 2 and 3 until reach from the root node to the leaf node

Test instances

1. Classify the new instance based upon the this decision tree

Stopping criteria

1. All sample for a given node belong to the same class
2. If there are no remaining attribute for further partitioning

In several cases, it was seen that J48 Decision Trees had a higher accuracy than either Naïve Bayes, or Support Vector Machines

III. EXPERIMENTAL RESULTS AND ANALYSIS

For our experiments we are using KDD CUP 99datasets. The class attributes of original train and test datasets of KDD CUP 1999 has 42 labels. The 41 labels can be generalized as only 2 labels Attacks and Normal. The performances of each method are measured according to the Precision and recall using the following expressions:

A. *Evaluation Criteria*

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

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

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

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

B. Results and discussion

TABLE I
LIST OF PROPOSED MODEL

Proposed Model	***labelling***
K-mean	1
CFS_K-mean2	2
CFS-KM-J48	3
CF-FZ-J48	4
FL-CF-FZ-J48	5
CONS-KM	6
CONS-KM-J48	7
CON-FZ-J48	8
FL-CONS-FZ-J48	9
FILTERED_K_MEAN2	10
FILT_KM_J48	11
FILT-FZ-J48	12
FL-FILT-FZ-J48	13

TABLE 2
COMPARIOSN OF RESULTS

A. U.	**Normal**		**Attack**	
	Precision	**Recall**	**Precision**	**recall**
1	0.004279	0.004423	0.757143	0.750978
2	0.764534	0.730101	0.935402	0.945595
3	0.758	0.985	0.996	0.924
4	0.947	0.002	0.805	1
5	0.757	0.99	0.997	0.923
6	0.210835	0.337069	0.812432	0.69474
7	0.739	0.995	0.999	0.915
8	0	0	0.805	1
9	0.733	0.991	0.998	0.913
10	0.418728	0.935108	0.977622	0.685924
11	0.433	0.988	0.996	0.687
12	0.947	0.002	0.805	1
13	0.74	0.914	0.978	0.922

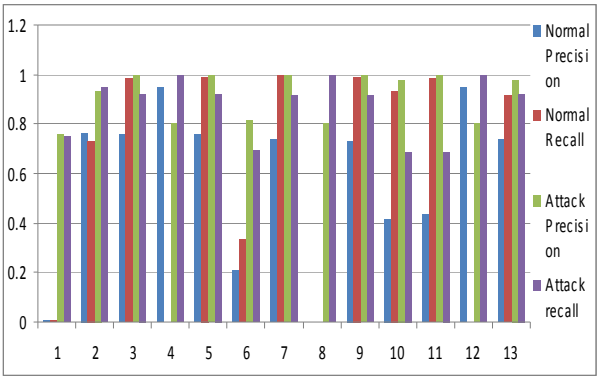

Figure2: Rresult comparison over proposed models

From above table 2 and figure 2 it can be depicted that model 5, 4, 7, 12 and 3 is giving best result. Precision is 0.433-94.7% and recalls are 0.002-0.999% for attack class precision is 0.805-0.999% and recalls are 0.687-1%.

Among this model 12 and model 4 are giving the best result. 12 is having 0.947% precisions and 0.002% recall for normal and 0.805% precision and 1% recall for attack class.

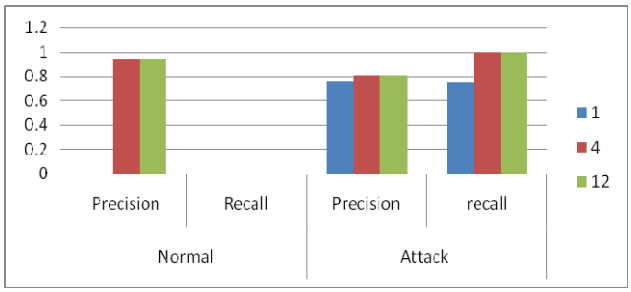

Figure3: Comparison of proposed model and k-mean over J48

IV. CONCLUSION AND FUTURE WORK

The main focus of this thesis was to eliminate the problems of class dominance and no class problem found in existing clustering algorithms. Proposed model is based on statistical model approach. From the different experiments it has been found that for 2 class the new algorithm gives better results than existing algorithm. For 2 class k-mean is having 75% recall for attack and for normal, precision is approximately 0. In the proposed statistical model, for 2 class datasets, filtered fuzzy J48 gives better result. It is having 99.2% precision and 100% recall.

Combination of clustering and only 5 classifier has been used for model generation. In future for model generation other clustering and classifiers can be used to improve the detection rate of intrusion detection system. Experiments can be carried out on 41 class labels datasets and 5 class labels datasets. And also multiclassifier can be used to improve the performance of intrusion detection system.

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?arnumbe

[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.12 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]. Mark Junjie Li troids et al. [31] 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.

[13]. K.S.Anil Kumar, and |Dr V.NandaMohan," Novel anomaly intrusion detection using neuro-fuzzy interference system",IJCNS 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

[14]. Krishnamoorthe Makkithaya,N.V.Subba reddy and dinesh acharya,"Intrusion detection system using modified c-fuzzy decision tree classifier" IJCNS 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

[15]. Gang Wang, Jinxing Hao, Jian Ma and Lihua Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering", inkinghub.elsevier.com/retrieve/pii/S0957417 410001417

[16]. 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 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/~roset/papers/cal99.pdf

[20]. http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-Ex3.html

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

[22]. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/ cmeans.html

[23]. Mark A. Hall, Lloyd A. Smith, FeatureSubset Selection: A Correlation Based Filter Approach 1997

[24]. Manoranjan Dash, and Huan Liu, "Consistency-based search in feature selection", Elsevier, Volume 151, Issues 1-2, December 2003, pp.155-176 linkinghub.elsevier.com/retrieve/pii/S0004370203 000791