International Journal of Engineering & Technology, 7 (2.33) (2018) 734-738

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Clustering of faculty by evaluating their appraisal performance by machine learning algorithms

Ravinder Ahuja [1] *, Alisha Chopra [1], Omanshi [1], Dhruv Sharma [1]

[1] Jaypee Institute of Information Technology, Noida
**Corresponding author E-mail: ahujaravinder022@gmail.com*

Abstract

Machine learning is a method which is mainly concerned with the design of the algorithm and with its development. It allows the computer to work according to the given data, mostly in the form of a database; Its major purpose is to automatically make intelligent decisions based on data and to recognize complex patterns. In this paper, we will group the data into multiple clusters on the basis of their similarities and dissimilarities. [5] Clustering is basically dividing the dataset or the given information into the subset (called clusters) so those same properties are classified in the same clusters. In every cluster, observations are similar in some senses. In this research paper, we are considering 15 factors related to the level of their teaching to help evaluate the performance of the staff members. On the basis of the feedback given by the students, the performance level is computed. It helps in assessing the annual increment and other promotion.In this research paper; we divide the staff member into three Group1, Group2, and Group3. Group1 has scored between 25 and 30, Group2 has scored between 20 and 25 and Group3 has scored between 15 and 20. These groups are a divide on the bases of the Points which is the average of all the 15 characteristics.

Keywords: *Clustering; Fuzzy Grouping; Similarities; Unsupervised Algorithms*

1. Introduction

Cluster analysis is the process in which data is divided into meaningful segments that share common characteristics. It is a study in which the machine automatically learns from the training set provided and works on the testing data. All the web pages for the same topic have to be grouped. The clustering of these different groups is a step moving ahead towards the automation process, which includes fields, includes web search engines, web robots, and data analysis.

Cluster analysis is a form of classification that labels which consist of various class labels. After classification, these are derived from the data on its own. Data mining has its own functionalities and these are characterization and discrimination. it also includes the frequent mining patterns, association, correlation, classification, and prediction. Some analysis is also their cluster analysis, outlier analysis, and evolution analysis are mainly its portion. Clustering is a vivid and a simplifying method. There is no predefined structure clustering always provides clusters or groups. So, the results of clustering should never be generalized.

2. Related work

There are many different approaches for clustering, but in this paper, we use [5] Algorithms. All [5] Algorithms are a part of unsupervised learning. All the 5 algorithms work upon the Points which is average of 15 characteristics that are Regularity, Presentation, Syllabus Coverage, Discussion, Availability, Curriculum, Punctuality, Create Interest, Coverage, Critical Thinking, Testing Student, Evaluation, Time Utilization, Subject Knowledge, Subject Depth.

In this we use 5 different clustering Algorithms that are a) K-means clustering b) Fuzzy c-mean clustering c) Self Organizing Mapping d) Agglomerative clustering e) Hierarchical K-means. Here all the algorithms are applied to the sample data set which consists the information of 330 teachers on the basis of 15 different aspects. At the end of this training resource allocation, network learns a various type of different type of data set for respective algorithms using the different functions. The network has been tested various times using the given data set on every possible aspect which includes accuracy, error and the performance. It has been found that the result of the data is quite accurate and the network produces perfect classified results. [16] The complexity of K-means clustering is O (n^2). It works upon the Euclidean Distance. The main findings are that fuzzy c-means clustering is better than Self Organizing Mapping and Hierarchical K-means clustering. Hierarchical K-mean clustering is better than Agglomerative and K-means clustering. If we use fuzzy C-means algorithm, then the given training samples will be clustered and the data, which is inappropriate, will be detected and will move to another dataset and will be used differently in the classification phase.

College computes the appraisal performance of the staff members based on these following 15 features:

1) The regularity of the teacher
1) Good
2) Average
3) Poor
2) Presentation in the class
1) Highly Impressive
2) Impressive
3) Not Impressive
3) Coverage of the syllabus within the class
1) More than 95%

2) Between 85 to 95 %
3) Less than 85 %
4) Discussions and questioning sessions within the class
1) Highly Supportive
2) Supportive
3) Less Supportive
5) Participation of the teacher after class
1) Easily Available
2) Occasionally Available
3) Hardly Available
6) Extra Activities with the students
1) Highly Supportive
2) Supportive
3) Less Supportive
7) Punctuality
1) Punctual
2) Fairly punctual
3) Not punctual
8) The interest created by the teacher
1) Highly Motivating
2) Occasionally Motivating
3) Never Motivating
9) Course Completion Speed
1) Average
2) Not so Fast / Slow
3) Extreme fast / slow
10) Critical views encouragement
1) Good
2) Fair
3) Poor
11) The relevance of tests and other evaluations
1) Very Helpful
2) Helpful
3) Never Helpful
12) Evaluation and assessment quality
1) Highly Fair
2) Good
3) Poor
13) Time Management
1) Good
2) Fair
3) Poor
14) Teacher’s knowledge for the subject
1) Highly Satisfying
2) Satisfying
3) Never Satisfying
15) Deep Study of the Subject
1) More than adequate
2) Adequate
3) Inadequate

3. Clustering algorithms

3.1. K-means clustering: k-means clustering is a simple method for unsupervised hard clustering. [12]

The K-mean algorithm operates as follows:

1) Initialize cluster centroids C.
2) For each iteration
a) Recalculate distance from data item to centroids (C1, C2, .Ck), and find the closest centered Cmin.
b) Further moving this cluster Ck into a new cluster Cmin. Repeat the same calculations to find centroid for Ck and Cmin.
3) Now repeat the above step till either of the two conditions is
a) Iteration limit reached is to the maximum.
b) There are no changes in the cluster assignments.

K k args min∑ ∑ ||x-μ□||^2 =args min ∑ |S□| Var S□

j=1 x ∈s j=1

c) Fuzzy C-mean Clustering: This clustering is a way of clustering data in an unsupervised manner.

The Fuzzy C mean clustering model is a best solution problem J m:

N K Jm (U, V; X) = ∑ ∑ uijm ||xi-vj||A2 (1)

I=1 j=1

where the following variable X represents the set of data X = {Xi, i = 1 ⋯ N} ⊆Rq number of clusters is represented by K, fuzzy degree by m, number of data n, membership of degree as Uij center of cluster j as Vj and distance between Vj(object and Xi as ||Xi-Vj||.Now consider :

ᴪKN = {U є RNK: 0≤uij≤1, □i, □j; □i □ j uij>0} (2)

K K

Mfc = {U є ᴪKN: ∑ uij =1, □i є {1, N}; ∑uij>0, □jє {1, N}} (3)

J=1 i=1

Theorem

If D ijA = ||x i−v j||A > 0, for all i, j, m > 1, and at least K different number patterns is included in data set X (U, V) ∈ M fc × ℜ K×q and minimization of J m is possible only if: K

Uij=(∑ (||xi-vj||A2 / ||xs-vj||A2)1/(m-1))-1,
i∈{1,……,N},j∈{1,……,K}, (4)

s=1 N N

Vj = (∑ uijm xi)/ (∑ uijm), □j є {1, K} (5)

d) Self-Organizing Map / Self-Organizing feature map:

This is an artificial neural network(ANN), in which a low-dimensional (especially two dimensional) and training is done on bases of unsupervised based learning, and therefore is a method to do dimensionality reduction because of the competitive learning, SOM maps are different from rest of the artificial neural network.

Algorithm:

1) Initially, the weight of the node is randomized.
2) An input vector is chosen randomly.
3) All nodes (with no exception) in the map are visited.
a) The input vectors and weight vector of the map’s node are evaluated on the bases of similarities using the formula of Euclidean distance.
b) Track the best matching unit (BMU) which is the node having the smallest distance.
4) The weight of the BMU neighborhood nodes is updated (including BMU by keeping them nearer to the input vector).

$W_x(s+1) = W_x(s) + \Theta$ (u, v, s).α(s). (D (t)-W_x(s))

5) Now s is increased and step two is repeated while (s<λ).
e) Agglomerative Clustering

This is that kind of clustering method in which every cluster has sub-clusters, which further have sub-clusters and so on. Species taxonomy is one of the examples. This type of hierarchical quality is also exhibited by Gene expression data (example neurotransmitter gene families). It generally starts with a single object in a given cluster and then in the following iterations it agglomerates the nearest pair of the cluster by satisfying the same criterion(not necessary all) until the data is collected in one single cluster.[15]

The following characteristics are present in the hierarchy with in the final cluster:

1) Nesting of the cluster obtained in the later stages is done which generates in the early stages.
2) Cluster in the tree with different sizes can be valuable for discovery.

The algorithm of Agglomerative hierarchical clustering is:

a) Preparing the data.
b) Dissimilarities in all the pair of an object belonging to the data set are computed.
c) Using linkage function to group objects into small hierarchical cluster tree, based on distance information which was generated in step 1. On the bases of linkage function, the cluster having a close proximity are linked.
d) Cut the hierarchical tree into the cluster and create partition data.
e) ierarchical K-Means Clustering:

This clustering is used to accelerate the clustering and feature vector construction and lookup.

Algorithm:

The algorithm is summarized as follow:

1) Compute the hierarchical clustering, cut the tree (node) into k-clusters.
2) Find the center (i.e. mean) of each cluster.
3) Compute K-means by using these set of cluster centers (above defined) as initial cluster centers.

luster analysis in educational contexts has been evidently used in the literature due to the need for researchers to discover characteristics common to different groups of students.

The problem is that there are many clustering methods but few guidelines on which algorithm to use. The ideal choice is dependent on the nature of the data and can rarely be found directly without any comparison between different methods. In Hämäläinen et al. [10], the authors evaluated the main clustering methods from this perspective. On the basis of their work, they found the most promising methods according to different situations.

The work of Lopez et al. [11] showed a classification from the use of clustering to predict the final grades of beginning college students. The article analyzed whether the participation of students in the course forum can be a good predictor of the final grade and if the classification proposed by grouping can obtain the grade with similar accuracy to traditional classification algorithms.

The comparison of several cluster algorithms using the proposed approach was done with traditional classification algorithms to predict the outcome of student’s performances in the course, based on their forum usage data in Moodle.

In Dominguez et al. [12], the authors presented a proposal for a tool that generates tips for students who are completing programming exercises. These tips may be links to topics that are relevant to the problem you are experiencing and may include preventative tips to avoid future errors. From previous year’s data, the tasks of grouping and classification were used and analysis is done to generate the tips. The system analyzes the patterns that affect students' performance during their interaction with the system.

The work of France and Amaral [13] focused on the performance of students and presented the use of grouping techniques, aiming at the formation of similar groups of students with learning difficulties in Object Oriented Programming. Peckham and McCalla [14] conducted an experiment in a learning environment designed to simulate hypermedia courses in order to identify patterns of student behavior in a reading comprehension task. K-means clustering algorithm was used for this identification.

Although there are many studies where the comparison of main clustering methods was done, such as Hämäläinen et al. [10], it is noticeable that there are only a few papers that give practical applications that describe characteristics of comparison between hierarchical, non-hierarchical and other methods of clustering.

The complete analysis of student data and comparison between different clustering algorithms has been done using R language. R is statistical and graphical analysis tool which is widely used for linear and nonlinear modeling, classical statistical test, time series analysis, classification, and clustering. R GUI (R development tool) can be easily downloaded and installed on the computer since it is open source [15]. It is a very interactive language in which well-designed quality graphical plots can be generated with ease. To run various algorithms and to plot certain graphs in R we need to download packages which are easily available on the internet [16].

Since clustering algorithms are based on unsupervised learning [17], there is no need for training and testing dataset. All the entries in the dataset are used for grouping and then the analysis can be done to extract useful information and pattern from these groups. In the case of hierarchical clustering, grouping has been done by using Ward's method and Euclidean distance because they are widely used and give much better results. In the case of non-hierarchical clustering, we have used KMeans, KMeans++ (advanced version of KMeans), and CMeans algorithms for our analysis. These algorithms can collectively be used to optimize our results and to verify the output of one algorithm with that of other. For the sake of convenience, the dataset has been divided into 4 groups in all the algorithms and these groups are then analyzed to find the relation between grades and other attributes.

Four groups can be chosen in hierarchical clustering by visualizing and cutting the obtained dendrogram at a particular height as shown in figure 1. In non-hierarchical clustering, grouping is comparatively easy because we have to initialize the number of clusters beforehand. The output of non-hierarchical clustering is shown in figure 2.

Firstly, we run the algorithms on our dataset to identify the groups and students in each group, and then compare the groups of hierarchical clustering with that of non-hierarchical clustering with the help of comparison matrix. For each algorithm, we then calculate the mean of each attribute of students from the same group and analyze the dependency of student’s grade on these attributes.

After the process of mining, we will evaluate, interpret and use the extracted information to visualize the outcomes.

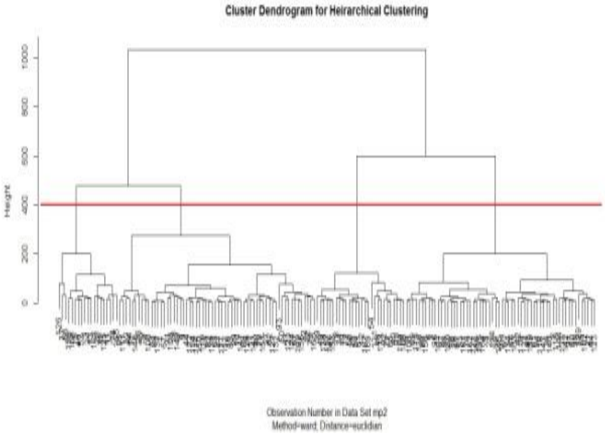

Fig. 1: Dendrogram Obtained in Hierarchical Clustering.

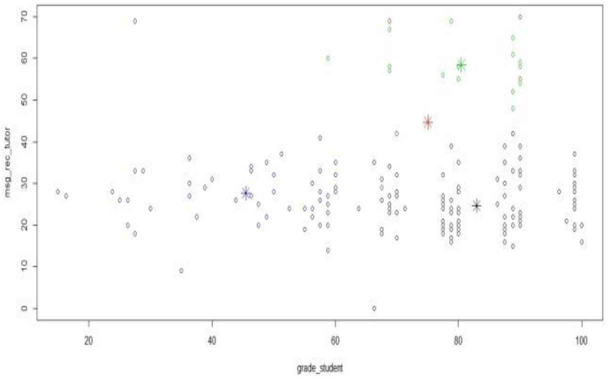

Fig. 2: Non-Hierarchical Clustering (K-Means).

4. Results

4.1. K-means clustering: we need to consider all the 15 points and the final score and using the r tool plot the graph

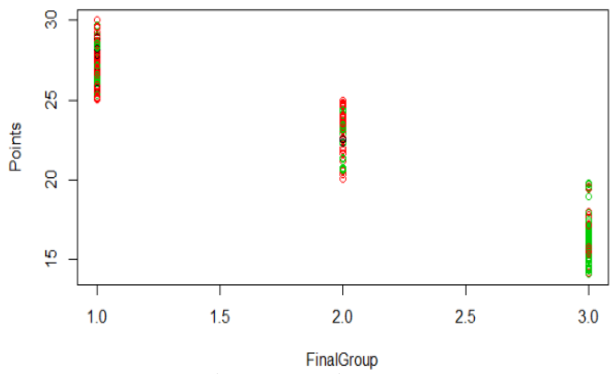

Fig 1: The Confusion Matrix for K-Means Clustering.

Table 1: Confusion Matrix for K-Means

	ONE	TWO	THREE
1.	51	26	82
2.	26	6	5
3.	65	36	33

4.2. Fuzzy c-means clustering

We consider the entire 15 characteristics to plot the multi-dimensional graph on C-mean clustering:

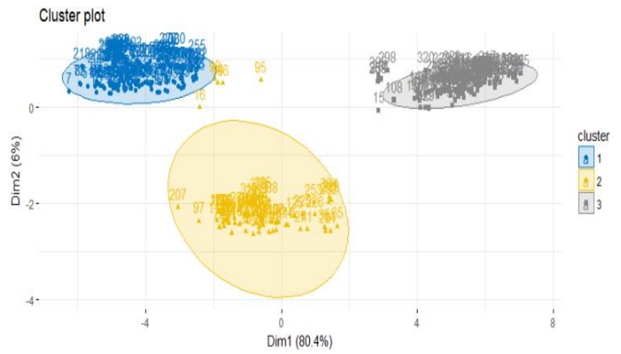

Fig. 2: The Confusion Matrix for C-Means Clustering.

Table 2: Confusion Matrix for C-Means

	One	Two	Three
1	139	0	0
2	3	68	0
3	0	0	120

4.3. Self-organizing mapping

Graph between mean distances at various iterations can be plotted as follows:

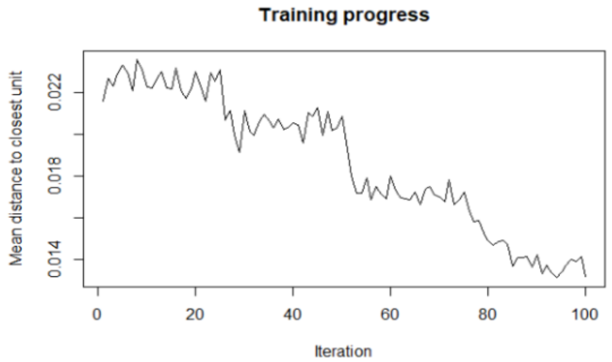

Fig. 3: The Confusion Matrix for Self Organizing Map.

Table 3: Confusion Matrix for SOM

	ONE	TWO	THREE
1.	140	3	0
2.	2	65	0
3.	0	0	120

4.4. Agglomerative clustering

Dendrogram of Agglomerative clustering is:

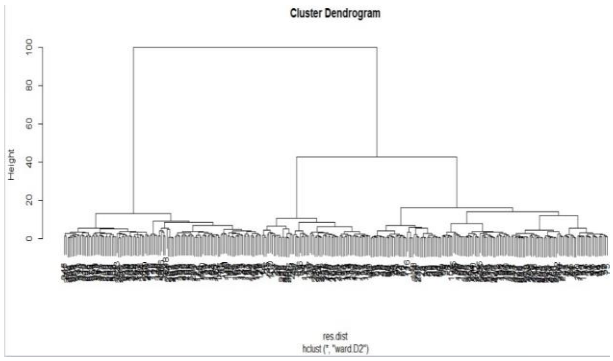

Fig. 4: Graph.

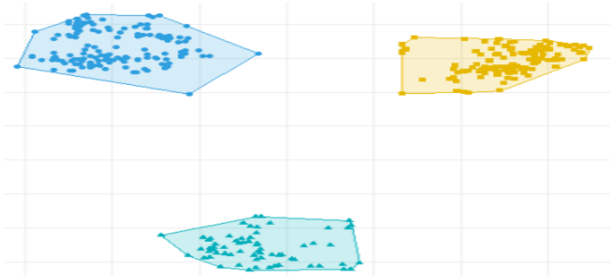

Fig. 5: The Confusion Matrix of Agglomerative Clustering.

Table 4: Confusion Matrix of Agglomerative Clustering

	ONE	TWO	THREE
1.	140	4	0
2.	2	64	0
3.	0	0	120

4.5. Hierarchical k-means

Dendrogram for Hierarchical K-Means is:

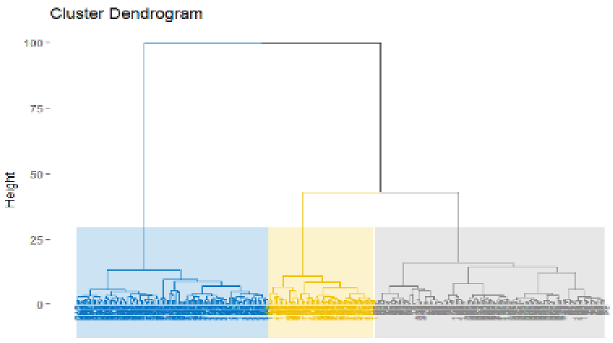

Fig. 6: Cluster Plot for Hierarchical K-Means is.

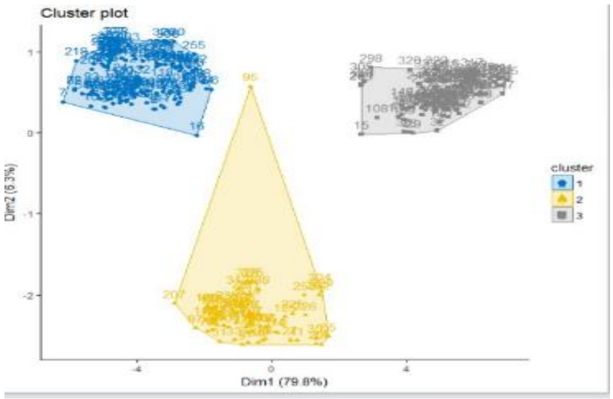

Fig. 7:

Table 5: Confusion Matrix for Hierarchical Clustering

	ONE	TWO	THREE
1.	140	3	0
2.	2	65	0
3.	0	0	120

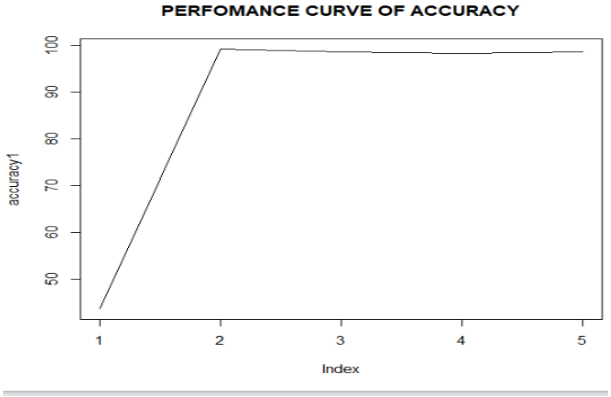

Fig. 8:

Table 6: Comparison of Accuracy and Error of All the Algorithms

Sr. No.	Clustering	Accuracy	Error
1.	K-means Clustering	27.2727	72.7273
2.	Fuzzy C-means Clustering	99.0909	.9091
3.	Self- Organizing Map	98.4848	1.5152
4.	Agglomerative Clustering	98.1818	1.8182
5.	Hierarchical Clustering	98.4848	1.5152

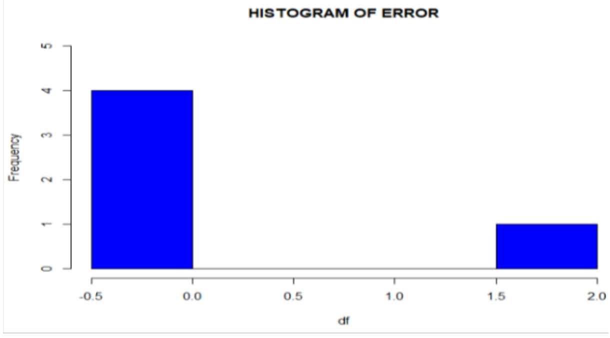

Fig. 9:

5. Conclusions and future work

We have perceived a comparative study of [5] different algorithms that are K-means clustering, Fuzzy C-Means clustering, Hierarchical K-Means, Self Organized Mapping and Agglomerative clustering using Siddaganga Institute of Information Technology teacher's performance database. The result of the overall comparative study is that fuzzy C-mean clustering gives much more accurate result than all the other algorithms its accuracy is 99.0909% which is better than all other algorithms. On the other hand accuracy of self-organizing mapping and hierarchical clustering is same 98.4848% and is greater than Agglomerative clustering which is 98.1818% and K-mean clustering which is 27.2727%. It was proved experimentally that fuzzy c means is the best algorithm in terms of accuracy (i.e. have less error). This analysis can be further improved by testing the algorithms for large data and keeping in mind the accuracy.

References

[1] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufman Publishers, Second Edition, (2006).

[2] Zakrzewska D and Murlewski J. Clustering algorithms for bank customer segmentation. In Proceedings of fifth International conference on intelligent systems Design and Applications (ISDA), pp. 197–202, 2005.

[3] Jain A. K. & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.

[4] T. N. Nagabhushana, Y.S. Nija gunaryao.―An Effective Data mining in Symbolic data Using Incremental learning Neural Networks‖, Elsevier Science, June 2005.

[5] Sunitha Chittineni, Raveendra Babu Bhogapathi "Determining Contribution of Features in Clustering Multidimensional Data Using Neural Network, IJITCS, Vol.4, No.10, September 2012.

[6] Data Mining with R: learning by case studies Luis Torgo.

[7] Timothy C.Havens, James C.Bezdek, Marimuthu Palaniswami. ―Fuzzy c-Means Algorithms for Very Large Data ―IEEE Transactions on Fuzzy Systems, Vol.20, No.6, December 2012.

[8] E cient and E active Clustering Methods for Spatial Data Mining.

[9] Rakesh Agrawal, Tomasz Imieliński, Arun Swami Mining association rules between sets of items in large databases.

[10] Arun Kumar, Jug Yanq-Data Management in Machine Learning: Challenges, Techniques and Techniques.

[11] Kianmehr, K. "Calling communities analysis and identification using machine learning techniques", Expert Systems with Applications, 200904

[12] Mohammad H. Nassralla, Mohammad M. Mansour, Louay M. A. Jalloul, "A Low-Complexity Detection Algorithm for the Primary Synchronization Signal in LTE", *Vehicular Technology IEEE Transactions on*, vol. 65, pp. 8751-8757, 2016, ISSN 0018-9545.

[13] Rajini, N. Hemi, and R. Bhavani. "Enhancing k means and kernelized fuzzy c-means clustering with cluster center initialization in segmenting MRI brain images", 2011 3rd International Conference on Electronics Computer Technology, 2011.

[14] Passelergue, J.-C., G. Foggia, M. Biserica, and E. Chanzy. "Transaction areas for local voltage control in distribution networks", 22nd International Conference and Exhibition on Electricity Distribution (CIRED 2013), 2013.

[15] Manjusha, R., and R. Ramachandran. "Web mining framework for security in e-commerce", 2011 International Conference on Recent Trends in Information Technology (ICRTIT), 2011.

[16] Chandraiah, Bhanuprakash & Nijagunarya, Y.S. & M.A, Jayaram. (2017). Clustering of Faculty by Evaluating their Appraisal Performance by using Feed Forward Neural Network Approach. International Journal of Intelligent Systems and Applications. 9. 34-40. 10.5815/ijisa.2017.03.05.