International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Performance Analysis of the Clustering in k-Means Algorithms based on K Values

Mesele Mohammed[1], Aderajew Kassie[2], Mrs Sunita Yadwad[3]
[1] *Dept. of computer engineering arul Institute of Engineering and Technology – Parul University Vadodara, Gujarat, India*
[2] *Dept. of Information Technology, Parul Institute of Engineering and Technology – Parul University Vadodara, Gujarat, India*
[3] *Dept. of Computer Science and Engineering Parul Institute of Technology – Parul University Vadodara, Gujarat, India*

Abstract*: Clustering is one of unsupervised learning algorithms which is used to cluster the objects into different groups or precisely the partitioning of dataset into sub small groups called cluster and the collection of cluster is called clustering. The dataset in each group contains the share some common traits according to some distance measurement. Data is grouped in value of k in case of simple k-means algorithms. K is number of group or cluster in k-means clustering. In this paper researcher tries to discuss the values of k and measure the quality or performance of clustering based on k values. The value of k plays a great role in clustering. K value plus the iteration of clustering and square errors of clustering are other factors during clustering. So the main objective of this paper is discussing the performance of cluster by using different k value in clustering.
Key words*: K-means; squared errors; supervised learning.

I. INTRODUCTION

k-means algorithms is one of unsupervised learning algorithms and the objectives of k-means algorithms is portioning of an objects into k cluster based on k values which is less than number of objects in which each object belongs to cluster or sub group with the nearest distance value to the one of the sub cluster. K-means algorithm is useful for undirected knowledge discovery and is relatively simple. K-means has found wide spread usage in lot of fields, ranging from supervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligence, image processing, machine vision, and many others The idea of this paper is discusses the k values influences on clustering when the value of k decrease and increase in clustering.

II. DATA MINING TECHINIQUES AND METHODOLOGY

Different functions of data mining are mainly classified as classification, clustering feature selection and association rule mining [2]. For this paper the researcher used different data mining techniques and algorithms. Data is collected from UK-Bank and preprocessed for experiment.

A. K-Means Algorithm

Clustering analysis or clustering is process of partitioning a set of observation in to subsets [9]. K-means is one of clustering algorithms. K-means algorithms assign each point to the cluster which is nearest to center called centroid [5]. Is an algorithm to cluster 'n' objects based on attributes into k partitions, where $k < n$. n is number of observations, k is positive integer. This proposition makes the center closer to some points and apart from the other points, in points that become center to the center will stay in that center, there is no need to find its distance to other cluster centers [12]. It is similar to expectation maximization algorithms for mixture of Gaussians incase of them both tries to deals with centroid of natural groups in the dataset [1]. Simple k-means algorithms is an algorithms for segmenting 'n' data points into k disjoint sub groups S_j containing data points to minimize the sum of squares criteria. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

1) Flow charts of how k-Means Algorithms Works

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

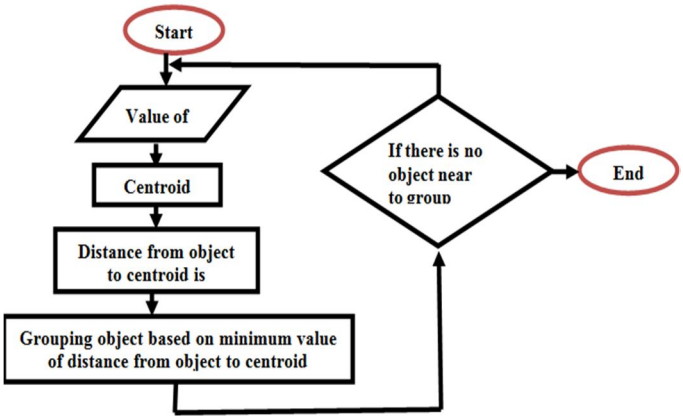

Figure 1 Flow chart that show how k-means algorithm Work

B. Data Clustering

When we talking about clustering rather than classification the output takes in the form of diagram that shows how the instance fall into the sub cluster [10]. Clustering widely used in diverse area today. The financial data in banking and financial industry is generally reliable and of high quality facilitate systematic analysis and data mining [3]. Data clustering is grouping a set of objects in a homogenous group based on distance from centroid to the same group called cluster. It is main task of explanatory data mining and a common techniques for statistical data analysis used in many field including machine learning, pattern recognition, image analysis, information retrieval, data compression and computer graphics. Popular notion of clustering include groups with small distance between cluster members dense area of data space intervals or particular statistical distributions. Clustering can be formulated multi-objective optimization problem. The appropriate clustering algorithms and parameter setting such as distance function depend on individual dataset and intended use of results. Clustering is not automatic process it is iterative process that involves trial and error. It is necessary to modify data preprocessing and modeling parameter until the result achieves good product or clustering [11].

C. Centroid Based Clustering(K-Means)

A simple k-means algorithm is one of the unsupervised learning algorithms which centroid is based clustering. During centroid based clustering groups or cluster are represented by a central vector which may not necessary be a member of dataset when a number of cluster or k value is fixed to k. The k clustered center such that the squared distance from the cluster are minimized and the point in a given subset are closer to that center than to any other center [8].

D. Application of Data Mining in Bank System

Data mining is becoming strategically important area for much business organization including bank system. It is process of analyzing the data from different perspective and summarizing it into variable pattern that are important to decision making. Data mining assists bank system to look for pattern in a group and discover unknown relationship among dataset. Today customer has many ideas with regard to where they can choose to their business. Early data analysis techniques were oriented toward extracting quantitative statistical data characteristics. These techniques facilitate useful data interpretation for the bank sector to avoid customer attrition. The focus of this paper is by using bank dataset quality of k-means algorithms by increasing the value of k and what is the output of clustering in different k value [4]. At least more than 3 value of k is tested and output is checked in bank dataset. For this paper k value are 6, 5, and 4 are tested and output are discussed below and output is measured based on expert. There are many algorithms that are used for calculating distance. For this paper Manhattan distance formula is selected. The Manhattan distance function computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components. The formula for this distance between a point *X*= (*X1*, *X2*, etc.) And a point *Y*= (*Y1*, *Y2*, etc.) is:

$$D = \sum_{i=0}^{n} |xi - yi| \qquad (1)$$

Where n is the number of variables, and ***Xi*** and ***Yi*** are the values of the i^{th} variable, at points ***X*** and ***Y*** respectively.

E. Weka Tools

Weka stands for Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Found only on the islands of New Zealand, the weak is a flightless bird with an inquisitive nature. Weka is open source software issued under the GNU General Public License. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions for this paper Weka tool is selected for data analysis.

G. Error Sum of Squares (SSE)

SSE is the sum of the squared differences between each observation and its group's mean. The purpose of SSE is to measure the performance of each method used [6]. It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the SSE would then be equal to 0. As k goes to infinite the mean square error within cluster approaches to zero [7]. The formula for SSE is as follows.

$$SSE = \sum_{i=1}^{n} (xi - \bar{x})^2 \qquad (2)$$

Where 'n' is the number of observations xi is the value of the i^{th} observation and $\bar{x}$ is the mean of all the observations

III. EXPJERIMENTAL WORK AND RESULTS

For this paper clustering using simple k-means algorithms by using personal bank dataset in Weka tool. Weka tool provide algorithms for simple k-means algorithms. The following three steps are listed and discussed and outputs are evaluated in each cluster. For the first experimentation k-6 used, for the second experimentation k=5 and for third experimentation k=4 are used. In last all outputs are evaluated based SSE value and number of iteration.

A. Results for k-Means algorIthms When k=6

For this study personal bank dataset used which contains 8 attributes and 3000 records. Data is collected from UK-Banks and different websites then integrated. Many preprocessing are undertaken for dataset to change dataset from raw data type to a format which s acceptable by Weka tool for experiment.

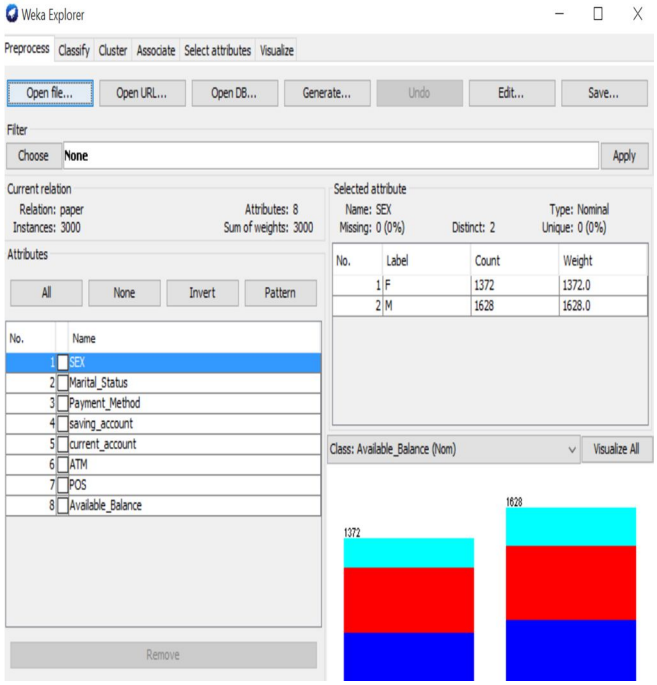

Figure 2 Weka 3-7 Interfaces with the dataset opened to start the first clustering run

For the first experimentation the value of k is 6 and default seed value is 10 and number of iteration is 500.

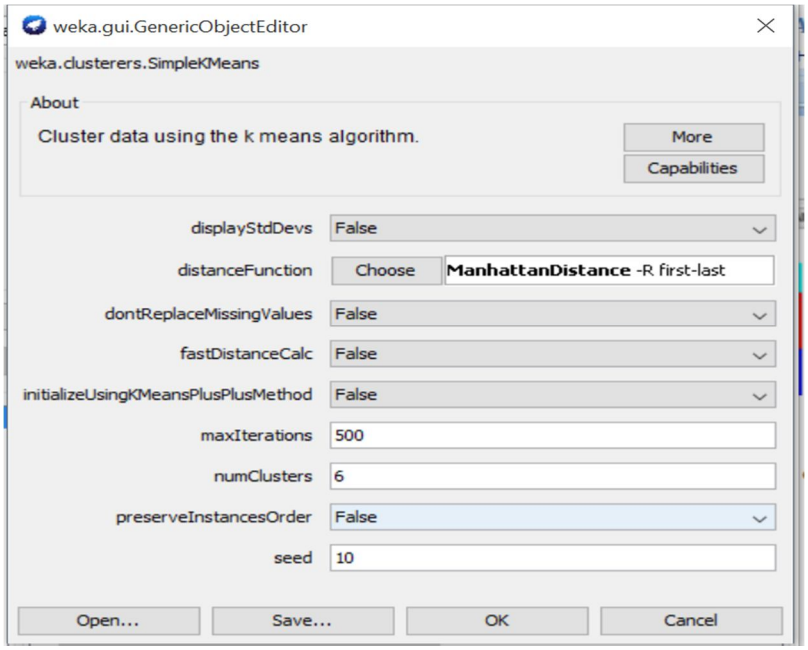

Figure 3 Simple k-means algorithm dialog box

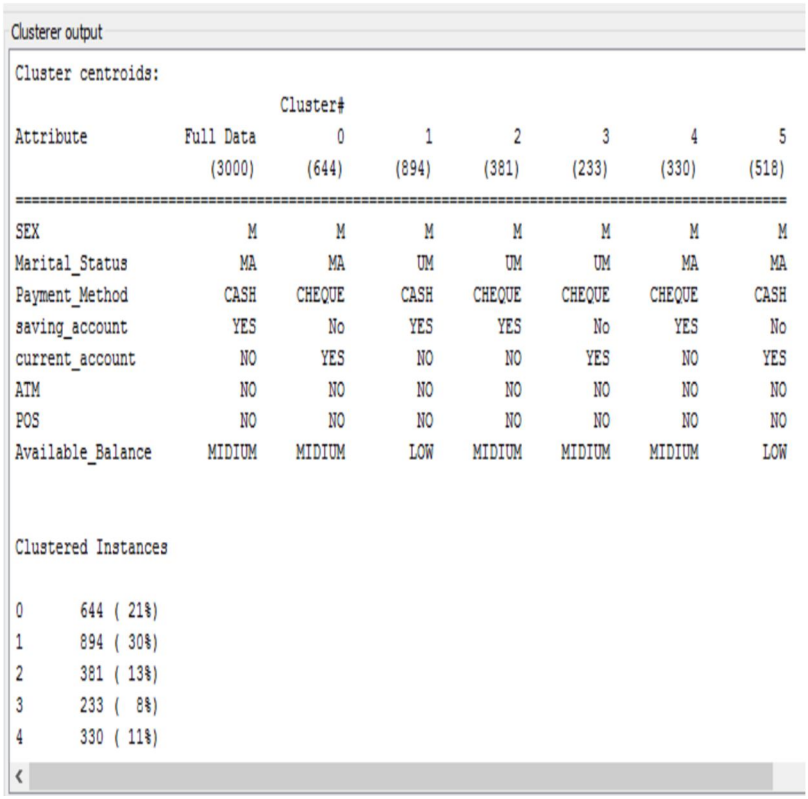

Figure 4 The first Cluster output with k=6 and with the default seed value= 10

The above output shows that the training result of the clustering model, including the number of attributes are used for clustering, the number of instances used, the clustering algorithm used, the test mode and other additional information. The above output is also shown in table format below

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Cluste	Freq-recor	Sex	Marit-	Paym ent	Savin	Curre nt acc	ATM	POS	Balan ce
1	644 (21%)	M	M A	Chequ e	Ye s	Yes	No	No	Mediu m
2	894(30%	M	U M	Cash	No	No	No	No	Low
3	381 (13%)	M	U M	Chequ e	Ye s	No	No	No	Mediu m
4	233 (8%)	M	U M	Chequ e	No	Yes	No	No	Mediu m
5	330 (11%)	M	M A	Chequ e	Ye s	No	No	No	Low
6	518 (17%)	M	MA	Cash	No	Yes	No	No	Mediu m

Table 1 Clustering result of the first experiment with k=6 and default seed =10

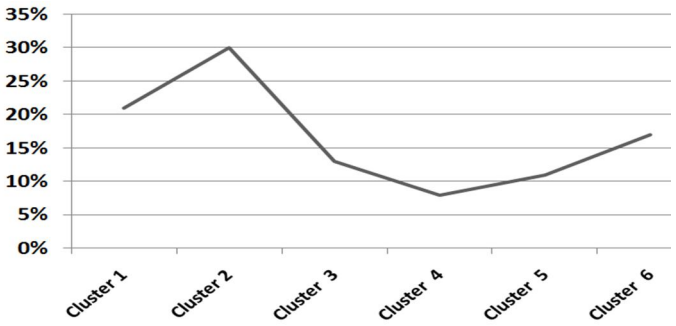

Figure 5 Diagrammatic representation of cluster k=6

IV. RESULTS FOR K-MEANS ALGORITHMS WHEN K=5

The below output describes the training result of the clustering model with k value 5 including the number of attributes are used for clustering the number of instance used, the clustering algorithms used the test mode and other additional information.

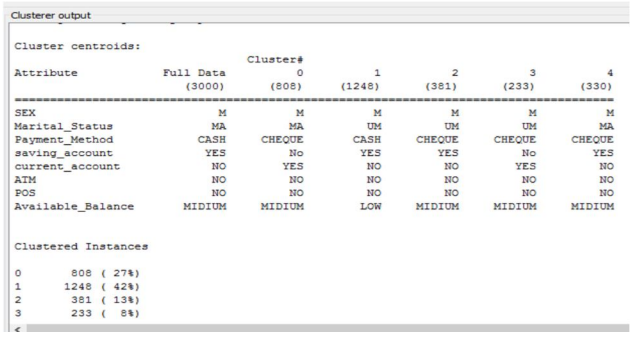

Figure 6 Cluster distribution with k=5 and with the seed value= 10

The above output is also shown in table format below

Cluste	Freq-record	Sex	Marita	Payme nt	Saving	Curren t acc	ATM	POS	Balanc
1	808 (27%	M	M A	Chequ e	No	Yes	No	No	Me diu m
2	1248 (42%)	M	U M	Cash	Ye s	No	No	No	Lo w
3	381 (13%)	M	U M	Chequ e	Ye s	No	No	No	Me diu m
4	233 (8%)	M	U M	Chequ e	Ye s	Yes	No	No	Me diu m
5	330 (11%)	M	M A	Chequ e	No	No	No	No	Me diu m

Table 2 First cluster with k=5 and default seed value=10

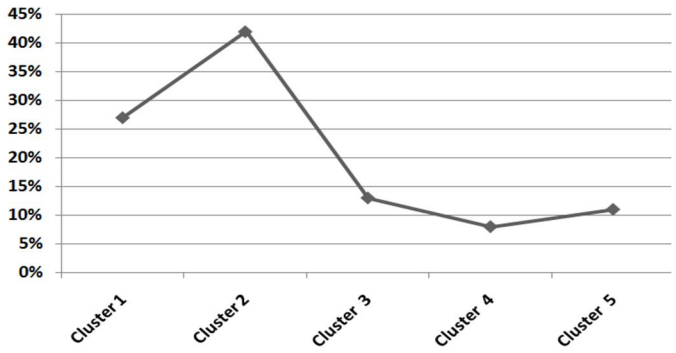

Figure 7 Diagrammatic representation of cluster k=5

V. RESULTS FOR K-MEANS ALGORITHMS WHEN K=4

The researcher has conducted experimentation with a cluster with k value 4 and with default seed value 10. The following output generates the summarized results of the 3rd and the detailed description of this result is depicted in table below

```
Clusterer output

                          Cluster#
Attribute            Full Data         0         1         2         3
                        (3000)    (816)    (1386)    (565)    (233)
=============================================================================
SEX                         M         M         M         M         M
Marital_Status             MA        MA        UM        UM        UM
Payment_Method           CASH   CHEQUE      CASH   CHEQUE   CHEQUE
saving_account            YES        No       YES       YES        No
current_account            NO       YES        NO        NO       YES
ATM                        NO        NO        NO        NO        NO
POS                        NO        NO        NO        NO        NO
Available_Balance      MIDIUM    MIDIUM       LOW    MIDIUM    MIDIUM


Clustered Instances

0       816 ( 27%)
1      1386 ( 46%)
2       565 ( 19%)
3       233 (  8%)
```

Figure 8 Cluster output with k=4 and the seed value= 10

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Cluste	Freq-recor	Sex	Marit-	Paym ent	Savin	Curre nt acc	ATM	POS	Balan ce
1	816 (27%)	M	M A	Chequ e	No	Yes	No	No	Mediu m
2	1386 (46%)	M	U M	Cash	Ye s	No	No	No	Low
3	565 (19%)	M	U M	Chequ e	Ye s	No	No	No	Mediu m
4	233 (8%)	M	U M	Chequ e	Ye s	Yes	No	No	Mediu m

Table 3 Result of the cluster with k=4 and default seed value =10

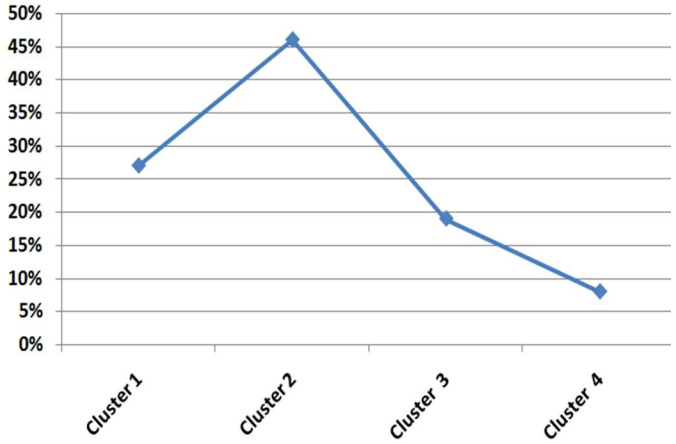

Figure 9 Diagrammatic representation of cluster k=4

In this experimentation four cluster are created. The result and interpretation of this clustering run with above experimentation is presented

VI. COMPARISON OF THE CLUSTERING MODEL BASED ON INCREASING K VALUE

In the above three different clusters model are done in order to find goodness of cluster model. This paper has taken different experimentation value of k=6, 5 and 4 with default seed value of 10. At experimentation with k value 6 all the created clusters were with different behavior. Different seed sizes are tested on each of this cluster formation to see whether the distribution of the segment could be improved. All the cluster with k=6,5 and 4 and the default seed value has not shown a significant difference in the segment data distribution.

A. Lastly, the best Cluster Model With Best Cluster Distribution Has Been Evaluated Based On

1) Number of iteration the algorithm uses{this shows the algorithms has moved and all misplaced data items in their correct classes within a few looming and the minimum value shows k-means algorithms converged very soon)

2) Within cluster sum of squared errors (This is the measure of the goodness of the clustering and tells how tight the clustering is overall that means lower values of squared errors are better) and

3) The jug dement of the domain expert based on the focus stated in the first experiment Therefore, based on these criteria's the three cluster models are k=6, 5 and 4 are compared

4) The cluster model at k=6 consists of number of iteration=3 and within the cluster sum of squared error =5804.

5) The cluster model at k=5 consists of number of iteration=3 and within the cluster sum of squared error =6486.0 an

6) The cluster model at k=4 consists of: Number of iterations= 3 and within the cluster sum of squared errors= 6816.0

Consequently, as stated above, the cluster model at k=6 shows the least value within the cluster sum of squared errors and in a number of iterations than cluster model at k=5 and 4. Also squared error of k=5 is less than cluster of k=4. Min squared error for

above problem at k=6 is 5804.0 and max value is 6816. As researcher discussed above in Error Sum of Squares (SSE) if the value of SSE is small the goodness of clustering is high.

VII. CONCLUSION

To conclude that clustering is very important in bank system for segmentation of customers in their homogeneity. The new coming customer's are simply registered to one of cluster based on their domain expert. K-means is one unsupervised learning algorithms which is important to group the customer any business or organization based on their similarity. Based on this the value of k has great role in clustering. As we have seen in above experimentation the performance of clustering is along with increasing the value of k. because the square error of cluster is decrease down when the value of k increases this indicates that the goodness of clustering is best when the squared error is less. So square error and k values are inversely proportional to each other as discussed in above experimentation. In other case seed value and iteration are other influential factors in goodness of clustering. In general when the value of k increases the quality of clustering is good.

REFERENCES

[1] Walter A. Shewhart And Samuel S. Wilks,"Wiley Series In Probability and Mathematical StatisticS" Yale University 1975.
[2] Tina R. Patil, Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification" International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013.
[3] N. Kavithasri, R. Porkodi, "Survey of Different Data Clustering Algorithms" Science and Technology Science and Technology 2018 IJSRST Volume 4
[4] Dr. K. Chitra1, B. Sabatini, "Data Mining Techniques and its Applications in Banking Sector", International Journal of Emerging Technology and Advanced Engineering August 2013.
[5] Giri Virati, Bewoor Mirunal, and Apte S."k-means driven single document summarization", intenational journal of computer application and bussness intelligence April-June, 2010.
[6] Bain Khusul Khotimah, Firli Irhamni, and Tri Sundarwati, "A genetic algorithm for optimized initial centers k-means clustering in smes", Journal of Theoretical and Applied Information Technology August 2016
[7] John Wiley and Sons "Series In Probability and Mathematical Statistics" 1975.
[8] K. Mumtaz and Dr. K. Duraiswamy," An Analysis on Density Based Clustering of Multi Dimensional Spatial Data", Indian Journal of Computer Science and Engineering Vol 1 No 1 8-12.
[9] Jiawei, Micheline Kamber and Jian Pei "Data Mining Concepts and Tecnques" Third Edition 2012
[10] Ian H. Witten and Eibe Frank, "data mining Practical Machine Learning Tools and Techniques," Second Edition 2005.
[11] Rajesh Wadhvani, R. K. Pateriya and Devshri Roy "A Topic-driven Summarization using K-mean Clustering and Tf-Isf Sentence Ranking" International Journal of Computer Applications (0975 8887), Volume 79 - No. 8, October 2013
[12] Miss Dhanshree Hadawale, Miss Prajakta Mande, Miss Sushama Patil, "An Efficient Analysis For Competitive Algorithm to Find
[13] Best Clusters" International Journal of Advanced Research in Computer Engineering & Technology (Ijarcet) Volume 4 Issue 3, March 2015