International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 4 Issue 4, April-2015, pp: (1-8), Impact Factor: 1.296, Available online at: www.erpublications.com

A Weather Forecasting Model using Clustering Approach

Ms. Ruchi Modi[1], Prof. Ketan Sarvakar[2]
[1]ME (CSE Student), UVPCE, Kherva, Gujarat, India
[2]Asst. Professor, UVPCE, Kherva, Gujarat, India

Abstract: Cluster analysis is one of the primary data analysis methods and k-means is one of the most well known popular clustering algorithms. The k-means algorithm is one of the frequently used clustering method in data mining, due to its performance in clustering massive data sets. The final clustering result of the k-means clustering algorithm greatly depends upon the correctness of the initial centroids, which are selected randomly. The original k-means algorithm converges to local minimum, not the global optimum. Many improvements were already proposed to improve the performance of the k-means, but most of these require additional inputs like threshold values for the number of data points in a set. In this paper for generating clustering objects we developed efficient mining approach and result analysis with ahmweather, rajweather and iris datasets. Our efficient approach is compare with K-Means algorithm and time required for generating clustering objects is less.

I. Introduction

Mining is a technique used to extract and mine the invisible, meaningful information from mountain of data. The term data mining is also relevantly used as Knowledge Discovery in Database, Knowledge engineering. Based on the patterns we look for the Data Mining models and tasks are divided into two main categories Predictive models and Descriptive Models. Whereas the Predictive Model is used to predict the feasibility of outcome, the other Descriptive model is used to describe the important features of dataset. The types of Predictive model are classification, regression, prediction and time series analysis. The various models included in descriptive model are clustering, summarization, Association rules and sequence discovery.

Clustering an unsupervised learning technique established in the area of data mining. Clustering or cluster analysis can be defined as a data reduction tool used to create subgroups that are more manageable than individual datum. Generally, clustering is defined as a process used for organizing/grouping a large amount of data into meaningful groups or clusters based on some similarity between data. Clusters are the groups that have data similar on basis of common features and dissimilar to data in other clusters. The applications areas where clustering plays an important role are machine learning, image processing, data mining, marketing, text mining. The terms clustering and classifications are always confused with each other, since they are two separate terms. Whereas Clustering is unsupervised learning process because the resulting clusters are not known before the execution which implies the absence of predefined classes in clustering. On the other hand classification is a supervised learning process due to presence of predefined classes. The high quality clustering is to obtain high intracluster similarity and low intercluster similarity.

One of the most popular data mining approaches which are adequate for Data numerous applications is clustering. The major reason for its wide range of application is the capability of clustering technique to work on datasets with least or no previous knowledge. This enables clustering as a convenient tool for many real world applications. Clustering is a technique of grouping comparable objects that are similar to each other and dissimilar to the data objects belonging to other clusters based on certain features [1]. Clustering is exploited to assemble items that appear to come naturally together [2]. Several kinds of clustering techniques are available, namely hierarchical vs. partitioned, exclusive vs. overlapping vs. fuzzy and complete vs. partial [3]. Clustering is a kind of unsupervised learning technique that separates data in such a way that comparable data items are assembled together in a set which are referred to as clusters. This technique is essential for condensing and recognizing patterns in data [4].

In recent times, high dimensional data has stimulated the attention of database researchers because of its significant challenges brought to the research community. In huge dimensional space, the distance between a record and its nearest neighbor can approach its distance to the outermost record [5]. In the framework of clustering, the difficulty causes the

distance among two records of the same cluster to move toward the distance among two records of various clusters. Conventional clustering approaches possibly will be unsuccessful to recognize the accurate clusters.

Clustering is obviously in need of several techniques to the categorization and association learning approach. Subspace clustering and projected clustering are current research topics in the field of high dimensional space clustering. On the other hand, in high dimensional datasets, conventional clustering approaches are likely to fail based on both accuracy and efficiency [6].

Clustering is the process of organizing data objects into a set of disjoint classes called clusters. Clustering is an example of unsupervised classification. Classification refers to a procedure that assigns data objects to a set of classes. Unsupervised means that clustering does not depends on predefined classes and training examples while classifying the data objects. Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. Therefore, a cluster is a collection of objects that are similar among themselves and dissimilar to the objects belonging to other clusters. Clustering is an crucial area of research, which finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics, etc.

Cluster analysis is a one of the primary data analysis tool in the data mining. Clustering algorithms are mainly divided into two categories: Hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion. A partition clustering algorithm partition the data set into desired number of sets in a single step.

Numerous methods have been proposed to solve clustering problem. One of the most popular clustering method is k-means clustering algorithm. The easiness of k-means clustering algorithm made this algorithm used in several fields. The k-means clustering algorithm is a partitioning clustering method that separates data into k groups. The k-means clustering algorithm is more prominent since its intelligence to cluster massive data rapidly and efficiently. However, k-means algorithm is highly precarious in initial cluster centers. Because of the initial cluster centers produced arbitrarily, k-means algorithm does not promise to produce the peculiar clustering results. Efficiency of the original k-means algorithm heavily rely on the initial centroids. Initial centroids also have an influence on the number of iterations required while running the original k-means algorithm. The computational complexity of the original k-means algorithm is very high, specifically for massive data sets. Various methods have been proposed in the literature to enhance the accuracy and efficiency of the k-means clustering algorithm. In this paper, section 2 discuss the related work of the clustering analysis; section 3 discuss the methodology for clustering objects mining; section 4 discuss result analysis; Finally section 5 concludes the paper.

II. Related Work

Fast rescue of the related information from the databases has always been a major issue. Different techniques have been developed for this purpose, one of them is Data Clustering. Clustering is a technique which divides data objects into groups based on the information establish in data that illustrates the objects and relationships between them, their mark values which can be used in many applications, such as knowledge discovery, vector quantization, pattern recognition, data mining, data dredging and etc. A categorization of major clustering methods:

1. Partitioning methods
2. Hierarchical methods
3. Density-based methods
4. Grid-based methods
5. Model-based methods
6. High-dimensional data clustering methods
7. Constraint-based clustering methods

Mining of association rules is a field of data mining that has received a lot of attention in recent years. Data mining researchers often try to find most feasible and efficient methods for extraction of useful patterns from stock data. In the paper titled frequent patterns mining of stock data using hybrid clustering association algorithm proposed a methodology where stock data is divided into clusters using K-means algorithm and then to generate the frequent patterns based on Most Frequent pattern Mining algorithm [7]. But K-means algorithm has some limitations the major being the cluster results heavily depend on the selection of initial centroids which caused to converge at local optimum. In another paper titled a mid-point based k-means clustering algorithm for data mining a new enhanced method for K-means algorithm is proposed where a systematic method to determine the initial centroid is explained.

In this research we applied this enhanced method for clustering instead of the proposed K-means algorithm in the earlier paper. After clustering is done using the enhanced method then it is applied on for mining patterns of huge stock data to show factors affecting the sale of products to find the frequencies of property values of the corresponding items. The extensions of K-means Algorithm improved method to overcome the various problems of original k-means algorithm [8]. There are lot of scope for the proposed Dbkmeans clustering algorithm in different application area such as medical image segmentation and medical data mining [9].

The filtering algorithm needs to perform a large number of nearest-neighbour queries for the points in the dataset [10]. The global k-means algorithm, to improve the local convergence properties of k-means algorithm. But the greedy k-means algorithm requires the pre-estimated the number of clusters, k, which is the same to the standard k-means algorithm [11].

The simple to implement requiring only two variations of kd-trees as the major data structure. With the assumption of a global clustering for k-1 centers, we introduce an efficient method to compute the global clustering for k clusters [12]. Therefore these algorithms are not sufficient for mining clusters from dataset. Based on given research gap we have list out the problems which are not solved in the above paper. Complexity of the standard K-means algorithm is very high. In K-means algorithm the initial centroids are selected randomly. K-means algorithm used Euclidean distance to measure the distance between data points and centroid of each cluster. But it calculates the distance in each iteration which leads to the poor performance of the algorithm and also increases the complexity of the algorithm.

III. Methodology

We have proposed efficient mining approach. In this approach, two steps denoted as follows:

Procedure :

Input: D= Set of n data points
K=desired number of clusters
Output: k number of clusters

Steps:

Part A : Determine the initial centroids of the clusters by using Part A.
Part B : Assign each data point to the appropriate clusters by using Part B.
In the first part, the initial centroids are determined systematically so as to produce clusters with better accuracy. The second part assigns each data point to the appropriate clusters. The two parts of the efficient approach are described below as Part A and Part B.

Part A:
Input: D = Set of n data items
K number of desired clusters
Output: A set of k initial centroids
Steps:

1. Set x = 1.
2. Compute the distance between each data object and all other data- objects in the set D.
3. Find the closest pair of data objects from the set D and form a data-object set Ax (1<= x <= k) which contains these two data- objects, Delete these two data objects from the set D.
4. Find the data object in D that is closest to the data object set Ax, Add it to Ax and delete it from D.
5. Repeat step 4 until the number of data objects in Ax reaches 0.75*(n/k).
6. If x<k, then x = x+1, find another pair of data objects from D between which the distance is the shortest, form another data-object set Ax and delete them from D, Go to step 4.

For each data-object set Ax (1<=x<=k) find the arithmetic mean of the vectors of data objects in Ax, these means will be the initial centroids. Part A describes the method for finding initial centroids of the clusters. Initially, compute the distances between each data point and all other data points in the set of data points. Then find out the closest pair of data points and form a set A1 consisting of these two data points, and delete them from the data point set D. Then determine the data point which is closest to the set A1, add it to A1 and delete it from D. Repeat this procedure until the number of elements in the set A1 reaches a threshold. At that point go back to the second step and form another data-point set A2. Repeat this till 'k' such sets of data points are obtained. Finally the initial centroids are obtained by averaging all the vectors in each data-point set. The initial centroids of the clusters are given as input to the second phase, for assigning data-points to appropriate clusters. The steps involved in this phase are outlined as Part B.

Part B:

Input: D = Set of n data items
A set of k initial centroids
Output: A set of k clusters
Steps:

1. Compute the distance of each data-object di (1<=i<=n) to all the centroids cj (1<=j<=k) as d(di, cj).
2. For each data-object di, find the closest centroid cj and assign di to cluster j.
3. Set ClusterId[i] = j.
4. Set Nearest Dist[i] = d(di, cj).
5. For each cluster j (1<=j<=k), recalculate the centroids.
6. Repeat.
7. For each data-object di, Compute its distance from the centroid of the current nearest cluster; If this distance is less than or equal to the present nearest distance, the data object stays in the cluster; Else (1) For every centroid cj (1<=j<=k) Compute the distance d(di, cj) (2) Assign the data-object di to the cluster with the nearest centroid cj (3) Set ClusterId[i]=j (4) Set Nearest_Dist[i]= d(di, cj).
8. For each cluster j (1<=j<=k), recalculate the centroids; until the convergence.

The first step in Part B is to determine the distance between each data-point and the initial centroids of all the clusters. The data-points are then assigned to the clusters having the closest centroids. This results in an initial grouping of the data-points. For each data-point, the cluster to which it is assigned and its distance from the centroid of the nearest cluster are noted. Inclusion of data-points in various clusters may lead to a change in the values of the cluster centroids. For each cluster, the centroids are recalculated by taking the mean of the values of its data-points. Up to this step, the procedure is almost similar to the original k-means algorithm except that the initial centroids are computed systematically.

The next stage is an iterative process which makes use of a heuristic method to improve the efficiency. During the iteration, the data-points may get redistributed to different clusters. The method involves keeping track of the distance between each data-point and the centroid of its present nearest cluster. At the beginning of the iteration, the distance of each data-point from the new centroid of its present nearest cluster is determined. If this distance is less than or equal to the previous nearest distance, that is an indication that the data point stays in that cluster itself and there is no need to compute its distance from other centroids. These results in the saving of time required to compute the distances to k-1 cluster centroids. On the other hand, if the new centroid of the present nearest cluster is more distant from the data-point than its previous centroid, there is a chance for the data-point getting included in another nearer cluster. In that case, it is required to determine the distance of the data-point from all the cluster centroids. Identify the new nearest cluster and record the new value of the nearest distance. The loop is repeated until no more data-points cross cluster boundaries, which indicates the convergence criterion. The heuristic method described above results in significant reduction in the number of computations and thus improves the efficiency.

Example :

The example denoted below using the proposed efficient approach. The input data set contains 16 entities which are described by X and Y values denoted in Table 4.1. The input parameter k is taken as 4. i.e. all the 16 entities have to be categorized into 4 clusters based on their efficiency.

Table 4.1: Dataset

Data Object	X	Y
1	15	58
2	50	93
3	25	130
4	40	130
5	25	165
6	50	170
7	25	225
8	60	220
9	40	250
10	43	270
11	50	280
12	50	320
13	43	360
14	60	360
15	60	405
16	60	540

Part A:

Step 1, 2, 3, 4, 5, 6, 7: After calculating the distance of each data object and all other data- objects, closest pair of data objects from the set D and form a data-object set Ax. Each data-object set Ax (1<=x<=k) find the arithmetic mean of the vectors of data objects in Ax, these means will be the initial centroids. The initial centroids of each group denoted in Table 4.2.

Table 4.2: Initial centroids of each set Ax

Set Ax	X	Y	Initial centroids
A1	50 50 25	93 170 165	(41.67, 142.67)
A2	43 50 40	270 280 250	(44.33, 266.67)
A3	25 40 15	130 130 58	(26.67, 106)
A4	43 60 50	360 360 320	(51, 346.67)

Part B:

Now using the calculated initial centroids for each set Ax as the initial 4 centroids, apply the step 1, 2, 3, 4, 5, 6, 7 and 8 on the input data. After three iterations, stability was achieved. The resulting clusters are denoted in Table 4.3.

Table 4.3: Clustering results

Data Object	X	Y	Resulting Cluster
1	15	58	1
2	50	93	1
3	25	130	1
4	40	130	1
5	25	165	2
6	50	170	2
7	25	225	2
8	60	220	2
9	40	250	2
10	43	270	3
11	50	280	3
12	50	320	3
13	43	360	3
14	60	360	3
15	60	405	4
16	60	540	4

IV. Result Analysis

In our experiments we choose different datasets with different with different number of records to prove the efficiency of the algorithm. Table 4 shows the different datasets characteristics.

Table 4: The characteristics of Dataset

Dataset	Number of Records
ahmweather.data.txt	1813
rajweather.data.txt	1820
iris.data.txt	150

As a result of the experimental study, revealed the performance of clustering approach with the K-Means algorithm. The run time is the time to mine the clustering objects. The experimental result of time is shown in Table 5 to Table 7 reveals that the algorithm outperforms the K-Means algorithm. The experimental result is also shown in Figure 1 to Figure 3. As it is clear from the comparison clustering approach performs well for the higher cluster value for all datasets shown in Table 4. But at the lower cluster its performance small reduces compare to K-Means algorithm. Difference between execution time of K-Means algorithm and K-Means algorithm are increases in later stages.

Table 5: Execution Time for K-Means algorithm and Clustering Approach using Ahmweather dataset

Clusters	Total Execution time in second	
	K-Means algorithm	Clustering Approach
4	36.42	32.57
5	42.16	35.09
6	54.08	44.77

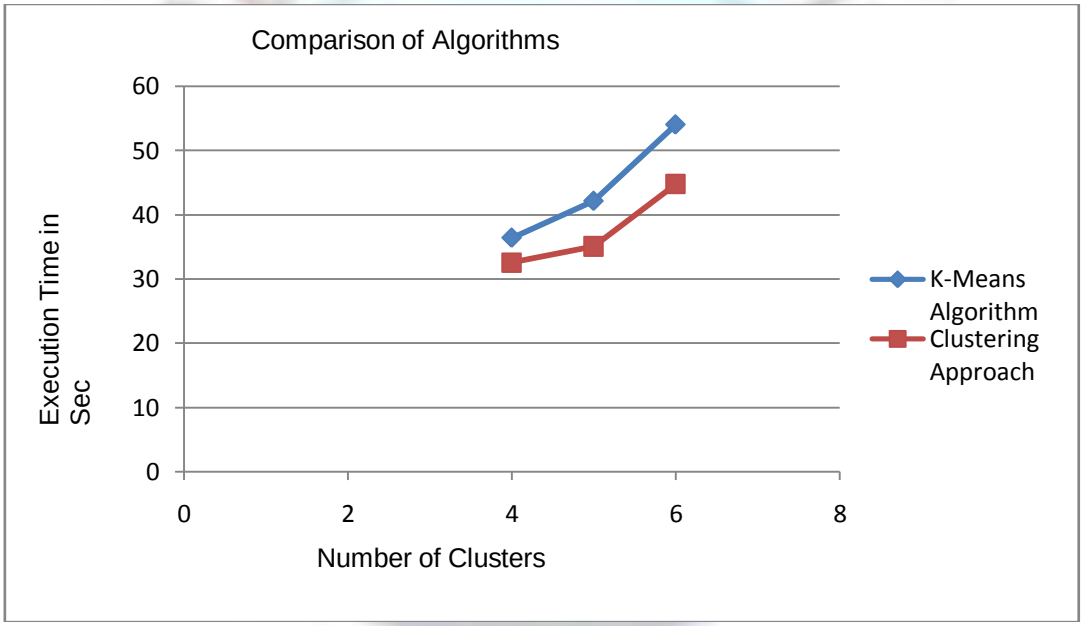

Figure 1: Total Execution Time for K-Means algorithm and Clustering Approach using Ahmweather dataset

Table 6: Execution Time for K-Means algorithm and Clustering Approach using Rajweather dataset

Clusters	Total Execution time in second	
	K-Means algorithm	Clustering Approach
4	36.28	32.45
5	43.21	35.97
6	52.66	41.74

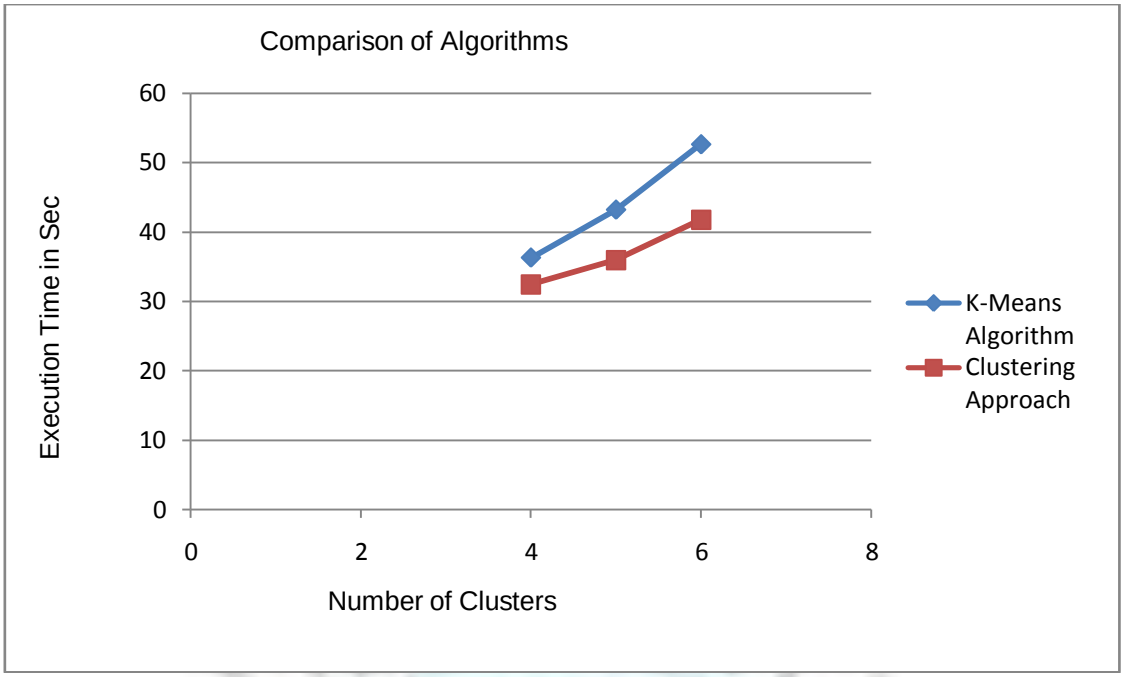

Figure 2: Total Execution Time for K-Means algorithm and Clustering Approach using Rajweather dataset

Table 7: Execution Time for K-Means algorithm and Clustering Approach using Iris dataset

Clusters	Total Execution time in second	
	K-Means algorithm	Clustering Approach
4	3.020	2.784
5	3.413	3.078
6	4.179	3.775

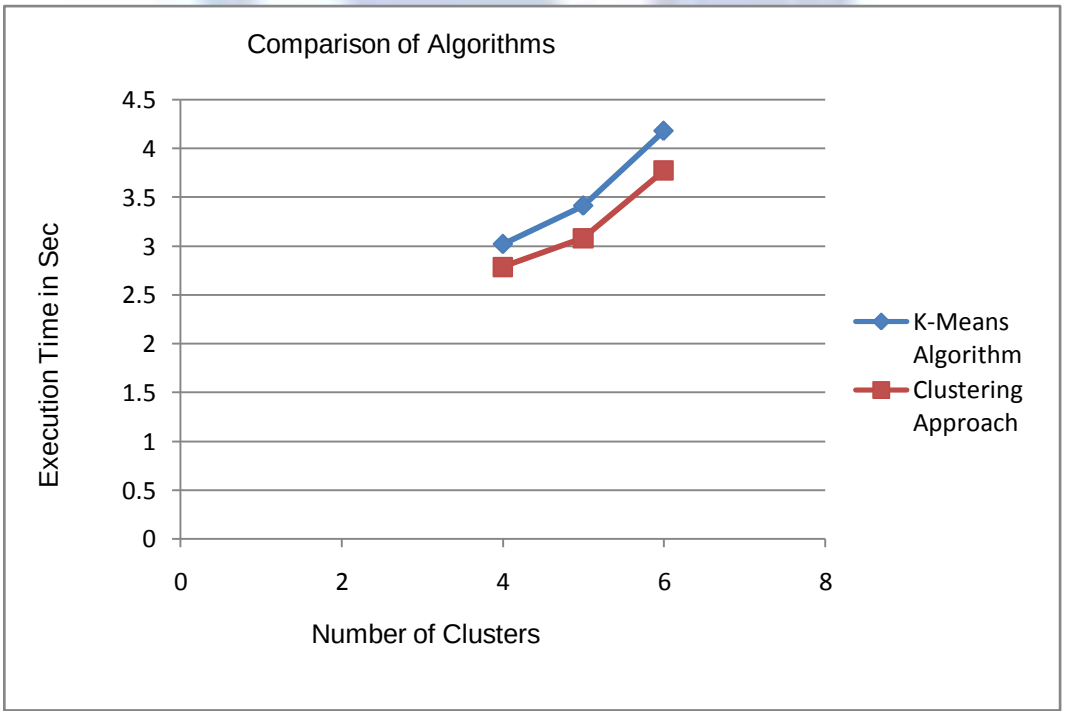

Figure 3: Total Execution Time for K-Means algorithm and Clustering Approach using Iris dataset

Conclusion

Clustering plays a crucial role in many applications. The commonly used efficient clustering algorithm is k-means clustering. K-means clustering is an important topic of research now a days in data mining. We choose different datasets with different number of records to prove the efficiency of the algorithm. We choose three datasets such as, ahmweather, rajweather, and iris. Clustering problem has been studied extensively with alternative by considered the following factor for creating our clustering approach, which are the time consumption, these factor is affected by the approach for finding the clustering objects. Work has been done to develop a clustering approach which is an improvement over K-Means algorithm. For different datasets the running time consumption of our clustering approach outperformed K-Means algorithm. Whereas the running time of clustering approach performed well over the K-Means algorithm on the collected dataset at the higher cluster value and also running time of clustering approach performed well at lower cluster value. Thus it saves much time and considered as a clustering approach as proved from the results.

References

[1]. Ali Alijamaat, Madjid Khalilian and Norwati Mustapha, "A Novel Approach for High Dimensional Data Clustering", Third International Conference on Knowledge Discovery and Data Mining, pp. 264-267, 2010.
[2]. Witten, Ian H and Eibe Frank, "Data Mining–Practical Machine Learning Tools and Techniques", 2nd Edition, Morhan Kaufmann, San Fransisco, 2005.
[3]. Tan, Pang Nin, Michael Steinbach and Vipin Kumar, "Introduction to Data Mining", Pearson International Edition, Boston, 2006.
[4]. Poncelet, Pascal, Maguelonne Teisseire and Florent Masseglia, "Data Mining Patterns: New Method and Application", London, 2008.
[5]. K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft, "When is nearest neighbor meaningful?", Lecture Notes in Computer Science, Vol. 1540, Pp. 217-235, 1999.
[6]. Gabriela Moise and Jorg Sander, "Finding Non- Redundant, Statistically Significant Regions in High Dimensional Data: A Novel Approach to Projected and Subspace Clustering", Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA.
[7]. Aurangazeb Khan, Khairulllah Khan,behram B.Baharuddin "Mining Frequent Patterns Minning Of Stock Data Using Hybrid Clustering Association Algorithm" International Conference 2009.
[8]. Zhexu Huang, "Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery 2, 283–304 (1998).
[9]. K. Mumtaz and Dr. K. Duraiswamy, "A Novel Density based improved k-means Clustering Algorithm - Dbkmeans", Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids,"Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.
[10]. T. Kanungo, D.M. Mount, N.S. Netanyahu, C. Piatko, R. Silverman and A.Y. Wu: An Efficient k-means clustering algorithm: analysis and implementation. In IEEE Transactions On Pattern Analysis And Machine Intelligence Vol. 24, No. 7, July 2002, pp. 881-892.
[11]. Aristidis Likas, Nikos Vlassis and Jacob J. Verbeek, "The global k-means clustering algorithm", In Pattern Recognition Vol 36, No 2, 2003.
[12]. Eena Gilhotra and Priyanka Trikha, "A Fast K-Means Algorithm", The International Journal of Computer Science & Applications, pages 87–94, February 2013.