Enhanced Clustering Algorithm for Classification of Datasets

Shalini Singh[1], Ravi Sheth[2], Anand Pandya[3]

1Information Technology Dept., A.D Patel Institute of Technology, New V V nagar-388120, Gujarat, India
2Information Technology Dept., A.D.Patel Institute of Technology, New V V nagar-388121, Gujarat, India
3Information Technology Dept., A.D.Patel Institute of Technology, New V V nagar-388121, Gujarat, India

Abstract

Clustering is one of the most important research areas in the field of data mining. In simple words, clustering is a division of data into different groups. Data are grouped into clusters in such a way that data of the same group are similar and those in other groups are dissimilar. It aims to minimize intra-class similarity while to maximize interclass dissimilarity. Clustering is an unsupervised learning technique. Clustering is useful to obtain interesting patterns and structures from a large set of data. Clustering can be applied in many areas, such as marketing studies, DNA analyses, city planning, text mining, and web documents classification. Large datasets with many attributes make the task of clustering complex. Many methods have been developed to deal with these problems. Here we have represented, partitioning based method – k-means, and enhanced clustering algorithm are implemented which will overcome the limitation of k-means and comparison is done between k-means and enhanced clustering algorithm applied on large datasets. The study given here explores the behaviour of this clustering algorithm with low processing cost and efficient result and less time complexity.

***Keywords*— data mining, clustering, incremental/enhanced k-means clustering algorithm, classification.**

I. INTRODUCTION

In Nowadays, information has grown and importance has increased as the web has spread out and communications to improve global market is also increasing day by day in which information technology is taking participation through different domain. Facing this new world and information extraction from data has become a major goal for all sorts of applications. The major lack of information has forced gathering data to be one of the most difficult tasks in traditional data mining applications[1]. The amount of data available from a given source is so high that traditional batch systems, based on memory storage and multiple reading of the same data, are unable to prove themselves for better results. Moreover, in recent real-world applications, data owns continuously from a data stream at high speed, producing examples over time, usually one at a time. There are different algorithms have been developed in recent time age that will process data in real –time for different application. The algorithm is useful in updating newly arrived data in the datasets in many ways by following different criteria [2]. One of the usual data mining problems is data clustering. Clustering will find the clusters of similar datasets by whole dataset clustering by separating similar group in one and dissimilar in other group by some similarity function. Data clustering techniques that work in real-time must allow a clear update of the clusters based on some large datasets. Moreover, more than just a definition of clusters, one could be interested in inspecting the structure of clusters and their relations with different data in the large datasets. Amongst the different techniques known in the literature, partitioning method propose better versatility as they require an a priori definition of the number of clusters to be define. From these, another issue emerges when we consider one of the major drawbacks of using static models, which is the assumption that the process generating the data follows a stationary distribution. , which can detect changes in the distribution producing the data. By definition, incremental methods lack the power of multiple use of information. One of the most visible criteria introduced by incremental algorithms is the

necessity to determine the minimum number of observations needed to assure convergence. These techniques have already been successfully applied to letter recognition patterns datasets using enhanced k-means.

II. METHODOLOGY

K-means

K-Means is one of the simplest unsupervised learning methods among all partitioning based clustering methods. It classifies a given set of n data objects in k clusters, where k is the number of desired clusters and it is required in advance. A centred is defined for each cluster. All the data objects are placed in a cluster having centroid nearest (or most similar) to that data object. After processing all data objects, k-means, or centroids, are re-calculated, and the entire process is repeated [3]. All data objects are bound to the clusters based on the new centroids. In each iteration centroids change their location step by step. In other words, centroids move in each iteration. This process is continued until no any centroid move. As a result, k clusters are found representing a set of n data objects.

Parameters of K-means

K-means algorithm requires three user-specified parameters: number of clusters K,cluster initialization, and distance metric. The most critical choice is K. While no mathematical criterion exists, a number of heuristics are available for choosing K.Typically, K-means is run independently for different values of K and the partition that appears the most meaningful to the domain expert is selected. Different initializations can lead to different final clustering because K-means only converges to local minima. One way to overcome the local minima is to run the K-means algorithm, for a given K, with several different initial partitions and choose the partition with the smallest value of the squared error. K-means is typically used with the Euclidean metric for computing the distance between points and cluster canters. As a result, K-means finds spherical or ball-shaped clusters in data. The algorithm steps for k-means algorithm are given below.

Algorithm of K-means [4]

Input : 'k', the number of clusters to be partitioned; 'n', the number of objects.

Output: A set of 'k' clusters based on given similarity function.

Steps:

(1) Arbitrarily choose 'k' objects as the initial cluster center.
(2) Repeat,
    (a) (Re)assign each object to the cluster to which the object is the most similar; based on the given similarity function;
    (b) Update the centroid (cluster means), i.e., calculate the mean value of the objects for each cluster;
(3) Until no change.

Explanation of k-means algorithm represents that firstly for any datasets the input is taken as the number of clusters represented as k and then from that datasets randomly any center or centriod point is taken for clustering on the basis of distance as a similarity function to form a clusters. Each and every time whenever the clusters is form in every iteration the new centroid is formed by mean of all the clusters. The process keeps on going till the stagnation point occurs in the algorithm. K-means algorithm works by selecting random centriod and finally process continues till the end.

Enhanced K-means Clustering Algorithm

The enhanced algorithm is necessary to increase the maximum number of similar cluster groups with the different types of data sets. As the study of dissertation is based on k-means algorithm some enhanced feature is carried out which will do the clustering on the updates coming each and every time to the datasets. Thus clustering on every increments and k-means algorithm will gives result of Incremental k-means clustering algorithm is considered as enhanced k-means algorithm.

Problem with K-means

K-means algorithm is applied on any large data sets which results in the clusters. But ,problem with k-means is that it will not provide efficient results if overhead increases with more numbers of data coming every time because it will form clusters by doing clustering from the base, so processing cost will be more with more number of iteration[5]. Thus need arise to develop enhanced clustering algorithm (incremental clustering).The different k-means problem is to provide good quality of clusters. The execution time taken by k-means is very high which in return will degrade cluster quality. In other case the initialization of initial centroid will also take more time by doing mean of clusters on each iteration. The different problem with k-means can be overcome by enhanced k-means clustering. Thus enhanced clustering monitors the quality of clusters

as program executes. The enhanced clustering algorithm is any update sequence in datasets which finds k clusters by initial clustering algorithm and clustering on updates on each updates such that as each it is presented, then the newly datasets arrived it is assigned to one of the current k clusters[6] .
The algorithm for enhanced clustering algorithm is represented below.

Enhanced K-means Clustering

Input : Existing 'n' data objects clustered into 'k' clusters; newly added data objects to be clustered.
Output: A set of 'k' clusters based on given similarity function.
Steps:

(1) Clusterize / adjust newly added objects
    (a) Assign each newly added object to one of the existing clusters to which the object is the most similar, based on the given similarity function;
    (b) Update the centroid (cluster means), i.e. calculate the mean value of the objects for each cluster.
(2) Clusterize all objects
(3) repeat,
    (a) Reassign each object to the cluster to which the object is the most similar; based on the given similarity function;
    (b) Update the centroid (cluster means), i.e., calculate the mean value of the objects for each cluster; until no change.

Explanation of the enhanced k-means clustering algorithm state that the input is taken as the n number of objects to be clustered and the number of cluster required for clustering the data sets object. As we know in k-means we are randomly selected centroid from the objects and the process continuously till the clusters from given datasets is formed. In same way enhanced k-means will do the clustering on the updates or the increments coming to the data sets. So initially the process is to select random centroid from the newly generated data as we are doing in k-means. Once the randomly centroid is selected then comparison on the basis of similarity function that is distance is done and nearest situated objects with distance is grouped in the cluster. Once the clustering is done with newly generated data then the next step is to adjust that clusters in to existing groups. Thus enhanced k-means clustering do the clustering on the updates and adjust it in existing groups. The process will continue whenever the new updates are coming to the datasets. The enhanced k-means clustering overcome all the problems occurring in the k-means algorithm. It reduces the complexity and also increases the accuracy and gives optimal solution. It will give accurate result for large as well as small data sets. It also reduces the overhead of number of iteration that is occurring in k-means algorithm. Time consumption for obtaining result is also less for enhanced k-means algorithm. An incremental algorithm is arrived from k-means so the name enhanced clustering algorithm is given to satisfy the research done in clustering. The experiment and results of k-means and incremental k-means which is the enhanced form of k-means is experimented on the numbers of small and large data sets which will result in good precision value and speed up factor providing accurate results in comparison with k-means and show the completeness of the enhanced clustering algorithm. The experiments taken on data sets shows that the number of iteration carried out by k-means will be much more iterations than the proposed enhanced clustering algorithm [9]. After the clustering done by enhanced clustering algorithm on the datasets the comparison of the results are carried out between the original clustering results given with data sets and clustering results of enhanced clustering algorithm. The comparison between k-means and enhanced clustering is done in terms of number of iterations and precision value. The statistical characteristics of data sets and evaluation methodology are represented by RMSE (Root Mean Square Error) error metrics [10]. In both the algorithm the starting step is to randomly initializing the centroids from the given data sets. But after the iterations started and increments from the datasets start joining the existing data sets, in that case k- means will execute with total numbers of data sets including the increments and early data and start doing the clustering while enhanced k-means algorithm start doing clustering directly on the updates and adjust the newly generated data in the existing clusters. Thus incremental clustering is far better than k-means in terms of speed and accuracy of results.

III. DATA SET DISCRIPTION

The testing was done on different datasets to test the results for incremental k-means algorithm and is compared with original cluster data in the datasets given to verify the results of proposed algorithm for clustering. The clustering was done on the large as well as small data sets to show that the enhanced algorithm is compatible with both types of data ranging from 150 data sets to 58,000 data sets with numbers of different attributes.
The descriptions of all the data sets that is used for result testing is listed below with each and every small details.

Iris Data sets [11]

Characteristics of Data sets :	Multi variates
Tasks to be formed :	Classification, Clustering
Characteristics of Attributes :	Real
Numbers of Data :	150
Number of Attributes :	4
Clusters required :	3

This is the famous data sets used by many scholars in research area in the field of pattern recognition literature. The data sets contain 3 clusters which are grouped by 50 data in each cluster. The detailed description of data is given below for the different four types of attributes.

1). Sepal length in cm
2). Sepal width in cm
3). Petal length in cm
4). Petal width in cm

Clusters class can be grouped into 3 groups after clustering as follows: (1) Iris Setosa, (2) Iris Versicolour, and (3) Iris Virginica. Thus these are the important feature related to the Iris data sets are given in summary form to verify the results of clustering.

Table 1: Summary Statistics

	Min	**Max**	**Mean**	**SD**	**Class Correlation**
Sepal length	4.3	7.9	5.84	0.83	0.7826
Sepal width	2.0	4.4	3.05	0.43	-0.4194
Petal length	1.0	6.9	3.76	1.76	0.9490 (high!)
Petal width	0.1	2.5	1.20	0.76	0.9565 (high!)

Shuttle Data Sets [13]

Characteristics of Data sets :	Multi variates
Tasks to be formed :	Classification, Clustering
Characteristics of Attributes :	Integer
Numbers of Data :	58000
Number of Attributes :	9
Clusters required :	3

The shuttle data sets contains approximately 80% of the data belongs to class 1. Therefore the default accuracy is about 80%. The examples in the original dataset were in time order, and this time order could presumably be relevant in classification and clustering. However, this was not deemed relevant for StatLog purposes, so the order of the examples in the original dataset was randomized, and a portion of the original dataset removed for validation purposes. The attribute information of shuttle datasets is listed below with detailed description. The shuttle dataset contains 9 attributes all of which are numerical. The first one being time. The last column is the class which has been coded as follows:

1) Rad Flow
2) Fpv Close
3) Fpv Open
4) High
5) Bypass
6) Bpv Close
7) Bpv Open

The cluster description of the shuttle data sets are displayed which is used for the comparison the clusters with the clusters obtain by incremental clustering.

Cluster Distribution:

Rad Flow	45586
Fpv Close	50
Fpv Open	171
High	8903
Bypass	3267
Bpv Close	10
Bpv Open	13

In the above cluster description the first group consist of maximum data and all other clusters are having very minimum number of data. This was the reason that the results coming from shuttle data sets is only limited to 80% so this was the limitation of data sets.

IV. RESULTS

The results are taken by applying clustering algorithm k-means and incremental k-means on the different data sets. Here the results are carried out on different criteria which will show that the enhanced algorithm overcomes the limitation of k-means. The performance of the algorithm is measured on the basis of Precision value and RMSE value. The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated. RMSD is a good measure of precision. These individual differences are also called residuals, and the RMSD

serves to aggregate them into a single measure of predictive power. Precision value is the also the repeatability or reproducibility is the degree to which repeated measurements under unchanged conditions show the same results. Here the mismatch calculation is also done to know the results that how many numbers of data are not in proper clusters so mismatch will give the proper idea. The proper functionality of algorithm is checked by the number of iterations taken during execution on the particular data sets.

Results for Iris Data sets

Table 2: Comparison of K-means and Incremental clustering on Iris data sets

Increment Number	(K-means) Static Iteration	(Enhanced K-means) Dynamic Iteration	Precision	RMSE	Speed - Up
1	3	3	89.23%	0.33	0.00%
2	6	5	88.89%	0.33	16.67%
3	6	2	89.29%	0.33	66.67%
4	6	2	89.66%	0.32	66.67%
5	5	3	88.67%	0.34	40.00%
Frequency of data in cluster	**50**	**61**		**39**	

Average static Iterations: 5.20
Average dynamic Iterations: 3.00
Average Precision: 89.15%
Average RMSE: 0.33
Average Speed Up: 38.00%

The above is the results of enhanced clustering algorithm which show the comparison between the static as well as the dynamic clustering algorithm. The increment number shows the number of updates coming to the fixed data sets which give rise to the total number of new dates with old data in the data sets .Thus on each number of data in the data sets increases with number of increments. The static as well as the dynamic iterations show that number of iteration for static is more than dynamic .The precision and RMSE value for dynamic clustering algorithm is far better than static clustering algorithm. The speed-up factor shows that the execution time taken by enhanced clustering algorithm more and execute the clusters in less time .Thus running time for proposed algorithm is very good. In this first figure the graph shows that the iteration for execution of iris data sets in k-means is more compared to incremental clustering. As the red line is for K-means iteration and blue line is for incremental which is executed in less iteration. The execution time is also less for incremental clustering.

Here, Graph showing the advantage of enhanced k-means clustering algorithm which represents that incremental algorithm is the more speedily execute the algorithm to form clusters and this was the reason to prove that it has less processing cost with less complexity.

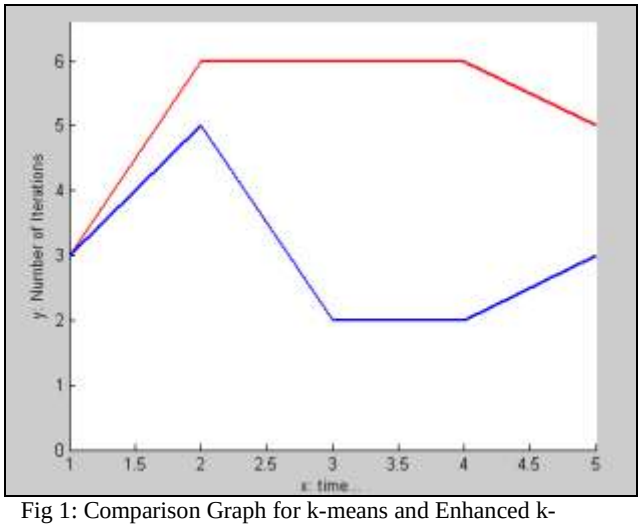

Fig 1: Comparison Graph for k-means and Enhanced k-means iterations for Iris Data sets.

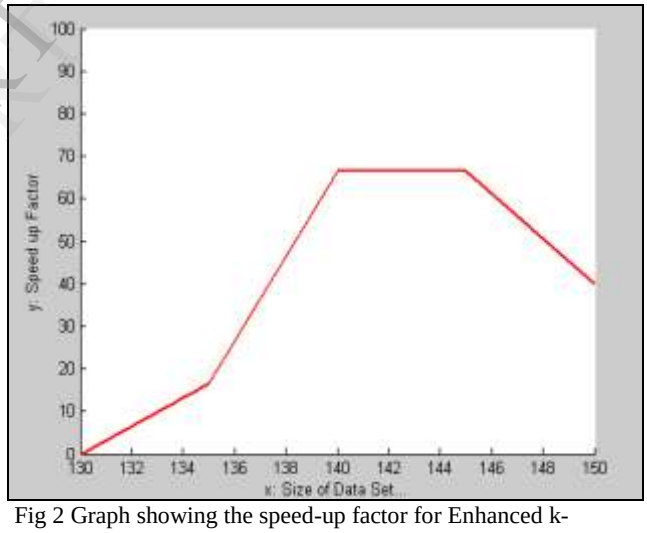

Fig 2 Graph showing the speed-up factor for Enhanced k-means clustering algorithm for Iris data sets

Results for Shuttle Data sets

In below table the comparison is done static approach and dynamic approach for 58,000 0f shuttle data sets the comparison shows that the speed-up factors in terms of execution to form clusters is more for incremental clustering. The number of iteration in incremental clustering is very less compared to k-means. It also shows the average result for both static as well as dynamic. Here in this graph the comparison is shown between K-means and enhanced clustering algorithm for shuttle data sets. Here it is very clear that Clustering done by k-

means forms the clusters in large iteration while in case of iteration cause by incremental for clusters is much less.

Table 3: Comparison of K-means and Incremental on Shuttle Data sets

Increment Number	(K-means) Static Iteration	(Enhanced K-means) Dynamic Iteration	Precision	RMSE	Speed – Up	
1	20	4	65.72%	0.59	80.00%	
2	20	3	65.74%	0.59	85.00%	
3	20	4	65.74%	0.59	80.00%	
4	20	3	65.72%	0.59	85.00%	
5	21	8	65.77%	0.59	61.90%	
Frequency of data sets occurring in clusters						
36728	**10972**	**5**	**7**	**4772**	**8**	**8**

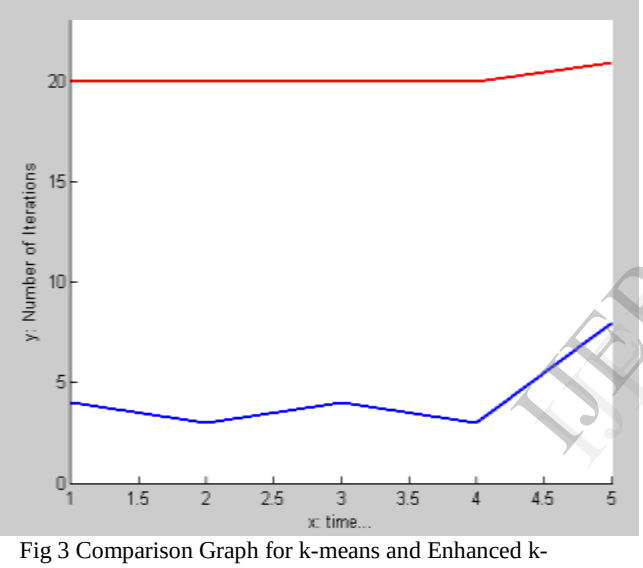

Fig 3 Comparison Graph for k-means and Enhanced k-means iterations for shuttle data sets.

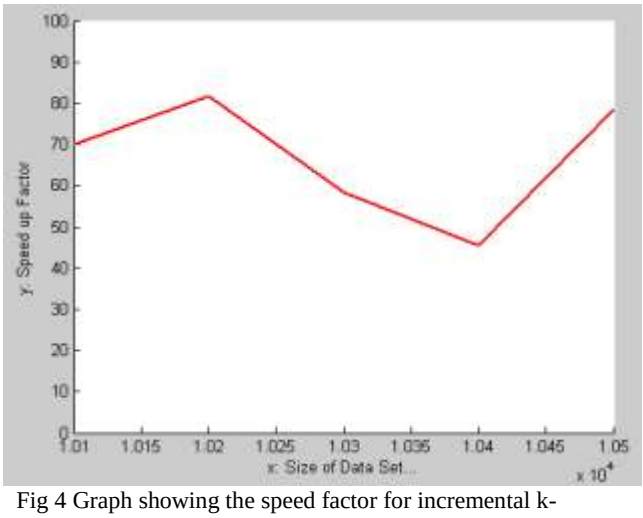

Fig 4 Graph showing the speed factor for incremental k-means for Shuttle Data Sets

Here, Graph shows the speed factor for shuttle data sets which increases the performance of the clustering algorithm though the shuttle data sets consist of 5000 datas and on every increments the number of data goes on increases but then also the speed up factor for such large data is also high.

V. CONCLUSION

The final conclusion over which we had come across is based upon the comparison between static and dynamic approach for clustering algorithm. Here the static algorithm for clustering is K-means which will do the initial clustering on data sets, but when the increments come to the data k-means will do clustering on all the existing data and new increments .This causes the large number of iterations for clustering to form clusters .On other side the clustering with enhanced algorithm will do clustering on newly coming increments and adjust into existing data .Thus iterations of proposed algorithm is very less compared to K-means. The other factor to check performance level is based on RMSE metric. The speed up factor to know process time taken by enhanced algorithm to get results is also high. The proposed algorithm is applied on small and large data sets and it has been proved that enhanced k-means clustering algorithm is better than k-means in terms of performance, iterations and speed-up factor for every data sets. Thus enhanced k-means clustering algorithm is far better than classic approach.

References

[1] L. Wanner, “Introduction to Clustering Techniques”, International Union of Local Authorities, July, 2004.

[2] T. Velmurugan,and T. Santhanam, “A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach”, Information. Technology journal, Vol 10, No 3, pp478-484,2011

[3] J. Han and M. Kamber. “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, August 2000.

[4] Jiawei Han and Micheline Kamber, “Data Mining Techniques”, Morgan Kaufmann Publishers, 2000.

[5] A.K. Jain, “Data Clustering: 50 Years Beyond K-Means” ,Pattern Recognition Letters, Vol 31 Issue 8 : pp.651-666 , June 2010

[6] B. S. Everitt, “Cluster Analysis”,3rd Edition, Edwar Publishers, 1993.

[7] A. A. K. Jain, M. N. Murty, and P. J. Flynn” Data ustering: a review”. ACM Computing Surveys, Vol .31No 3,pp.264–323, 1999.

[8] Hartigan, John A, “Clustering Algorithms”. John Wiley. New York,1975.

I.K.Rao, “Data mining and clustering

[9] Techniques"Workshop on Semantic web(DRTC), Banglore,8th-10th December ,2003.

[10] C. Ding, X.He, H.Zha, and H.Simon.,"Adaptive dimension reduction for clustering high dimensional data". In Proceedings of the 2nd IEEE International Conference on Data Mining, 2002.

[11] R.A.Fisher,"Iris Data Sets",UCI repository of machine learning databases",1936.

[12] D.J.Slate,"Letter recognition Data Sets", UCI repository of machine learning Databases, Odesta Corporation, Maple Ave; Suite 115; Evanston, IL 60201 ,1890.

[13] J.C.Basser,"Shuttle Data Sets",UCI repository of machine learning databases", Department of Computer Science, University of Sydney, N.S.W, Australia.