An Enhanced Clustering Algorithm to Analyze Spatial Data

Dr. Mahesh Kumar, Mr. Sachin Yadav

***Abstract*— Cluster analysis is one of the basic tools for exploring the underlying structure of a given data set and is being applied in a wide variety of engineering and scientific disciplines such as medicine, psychology, biology, sociology, pattern recognition, and image processing. The primary objective of cluster analysis is to partition a given data set of multidimensional vectors (patterns) into so-called homogeneous clusters such that patterns within a cluster are more similar to each other than patterns belonging to different clusters. Cluster analysis aims to group data on the basis of similarities and dissimilarities among the data elements. The process can be performed in a supervised, semi-supervised or unsupervised manner.**

In the paper we have proposed an enhanced algorithm for clustering using K Means technique of spatial data by which results through WEKA GUI Chooser tool are quite satisfactory. Though the results of the enhanced algorithm are compared to the most common known technique such as K Means, Euclidian distance algorithm and DBSCAN algorithm etc. In future it could be beneficiary for searching of the huge data in various sector of life as data is increasing day by day.

***Index Terms*— Cluster, Pattern recognition, Spatial, Supervised and Unsupervised Manner.**

I. INTRODUCTION

Clustering involves identifying a finite set of categories (clusters) to describe the data. The clusters can be mutually exclusive, hierarchical or overlapping. Each member of a cluster should be very similar to other members in its cluster and dissimilar to other clusters. Techniques for creating clusters include partitioning (often using the k-means algorithm) and hierarchical methods (which group objects into a tree of clusters), as well as grid, model, and density-based methods that subset. It also called characterization or generalization [1].

II. FEW TERMS ASSOCIATED WITH CLUSTERING

A cluster is an ordered list of objects, which have some common characteristics. The objects belong to an interval [a, b], in our case [0, 1].

A. Distance between two Clusters :

The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed.

B. Similarity :

A similarity measure similar (d_i, d_j) can be used to represent the similarity between the documents. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement.

C. *Average Similarity* :

If the similarity measure is computed for all pairs of documents (d_i, d_j) except when i=j, an average value average similarity is obtainable. specifically, average similarity = constant similar (d_i, d_j), where i=1,2,….n and j=1,2,….n and i < > j

D.Threshold :

The lowest possible input value of similarity required to join two objects in one cluster.

E. Similarity Matrix :

Similarity between objects calculated by the function similar ($d_{i,}$,d_j), represented in the form of a matrix is called a similarity matrix.

III. CLUSTERING TECHNIQUES

There are so many clustering technique evolved till now but we will see few algorithms among them.

A. K Means Clustering Algorithm :

K-means is a typical non supervised clustering algorithm in data mining and which is widely used for clustering large set of data. It is a partitioning clustering algorithm, this method is to classify the given data objects into k different clusters through the iterative, converging to a local minimum. So the results of generated clusters are compact and independent. The algorithm consists of two separate phases. The first phase selects k centers randomly, where the value k is fixed in advance [2].

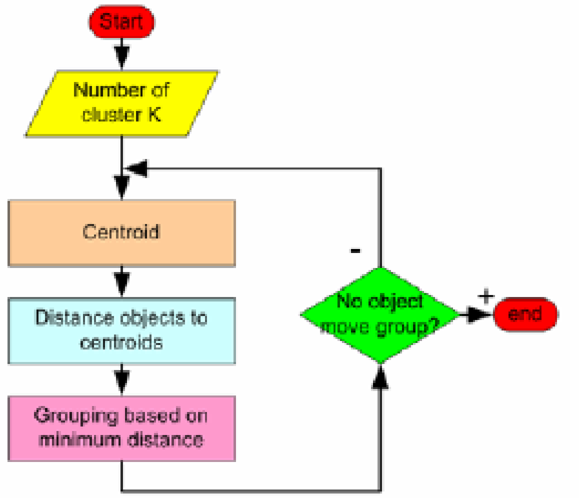

Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

Figure 1 K-means clustering

Manuscript received July 16, 2014.

Dr. Mahesh Kumar, Department of Computer Science & Engineering, MRKIET Rewari, India

Mr. Sachin Yadav, Department of Computer Science & Engineering, MRKIET Rewari, India

The next phase is to take each data object to the nearest centre. Euclidean distance is generally considered to determine the distance between each data object and the cluster centres. When all the data objects are included in some clusters, the first step is completed and an early grouping is done by recalculating the average of the early formed clusters. This iterative process continues repeatedly until the criterion function becomes the minimum.

B. Hierarchical Clustering Algorithm

Hierarchical clustering algorithms according to the method that produce clusters can be further divided into:

- *Agglomerative algorithms*. They produce a sequence of clustering schemes of decreasing number of clusters at east step. The clustering scheme produced at each step results from the previous one by merging the two closest clusters into one.
- *Divisive algorithms*. These algorithms produce a sequence of clustering schemes of increasing number of clusters at each step. Contrary to the agglomerative algorithms the clustering produced at each step results from the previous one by splitting a cluster into two.

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds. The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds [3].

C. Density based Clustering

Density-based clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. DBSCAN and OPTICS are two typical algorithms of this kind. Subspace clustering methods look for clusters that can only be seen in a particular projection (subspace, manifold) of the data. These methods thus can ignore irrelevant attributes.

The key idea of density based clustering is that for each object of a cluster the neighbourhood of a given radius ε has to contain at least a minimum number of μ objects, i.e. the cardinality of the neighbourhood has to exceed a given threshold [4].

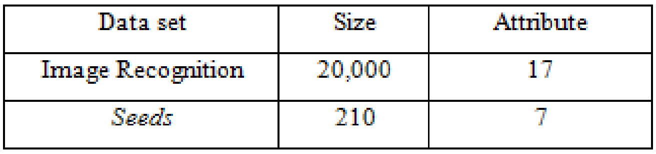
IV. EXPERIMENTAL SETUP AND RESULTS

For clustering using different algorithms as mentioned earlier, WEKA GUI Chooser tool has been used. WEKA is a landmark system in the history of the data mining and machine learning research communities and is freely available for download. WEKA offers many powerful features (sometimes not found in commercial data mining software), it has become one of the most widely used data mining systems. WEKA also became one of the favorite vehicles for data mining research and helped to advance it by making many powerful features available to all.

The WEKA project aims to provide a comprehensive collection of machine learning algorithms and data pre-processing tools to researchers and practitioners alike. It allows users to quickly try out and compare different machine learning methods on new data sets.

To perform the enhanced algorithm we have used the following data set as shown below in the table.

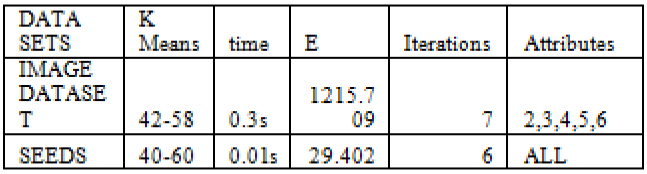
Table 1 DATA SET USED



Data set	Size	Attribute
Image Recognition	20,000	17
Seeds	210	7

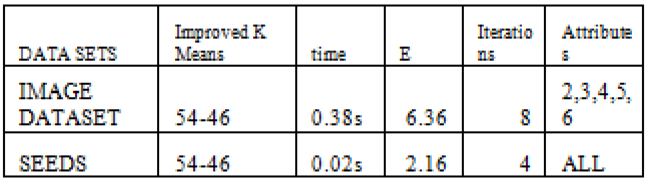
This data sets is described and used in WEKA tool. It can be downloaded from *archive.ics.uci.edu/ml/datasets.html.* The image recognition dataset has 17 attributes and 20,000 instances and Seeds data set has 7 attribute and 210 instances. When we clustered these datasets in the WEKA tool we get the following results:

Table 2 Results using K Means using s=10



DATA SETS	K Means	time	E	Iterations	Attributes
IMAGE DATASE T	42-58	0.3s	1215.7 09	7	2,3,4,5,6
SEEDS	40-60	0.01s	29.402	6	ALL

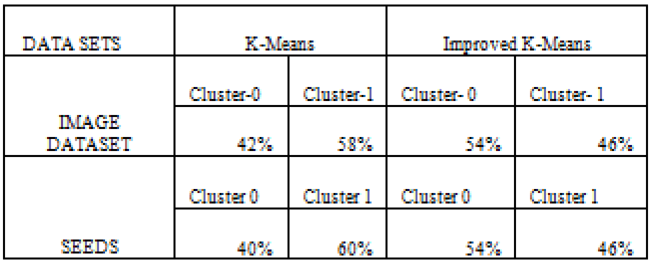
Table 3 Results using Improved K Means using s=10



DATA SETS	Improved K Means	time	E	Iteratio ns	Attribute s
IMAGE DATASET	54-46	0.38s	6.36	8	2,3,4,5, 6
SEEDS	54-46	0.02s	2.16	4	ALL

Now using the result shown above, we can draw the distribution for two clusters in the table:

Table 4 Cluster-0 and cluster-1distribution



DATA SETS	K-Means		Improved K-Means	
IMAGE DATASET	Cluster-0	Cluster-1	Cluster- 0	Cluster- 1
	42%	58%	54%	46%
SEEDS	Cluster 0	Cluster 1	Cluster 0	Cluster 1
	40%	60%	54%	46%

As here in table E stands for error measured.

Now using the result shown below, we can draw the distribution in the form of a pie chart:

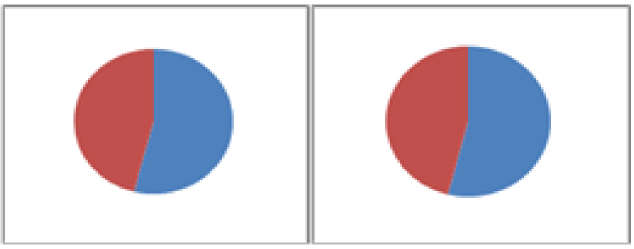


Fig 2: Image Recognition Dataset

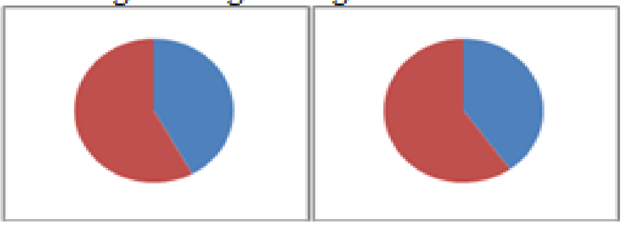


Fig 3: Seeds Dataset

As from above figures and table it can be easily concluded that the results shown by the improved K-Means Algorithm results are better, thus can be used for the data set used in different segment of life as Medical, transportation etc.

V. CONCLUSION

From the comparative study it can be easily concluded that instances of a data set are more closely clustered as there is less difference in the two clustered formed using the modified algorithm. Thus the proposed algorithm can be used further for data mining in the medical, transportation, social work analysis, pattern matching.

REFERENCES

[1] Anil K. Jain, "*Data Clustering: 50 Years Beyond K-Means*" 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, December 8, 2008.

[2] Prof. Pier Luca Lanzi,"*Clustering Partationing*" (Spring 2009)

[3] Halkidi M., Batistakis Y., Vazirgiannis M., "*On Clustering Validation Techniques*"**,** Journal of Intelligent Information Systems, 17:2/3, 107–145, 2001.

[4] [52] Hand D., Mannila H., Smyth P., "*Principles of Data Mining*", MIT Press, Cambridge, MA. ISBN 0-262 - 08290-X, 2001.

[5] http://www. en.wikipedia.org

[6] http://www.google.com

[7] http://www.archive.ics.uci.edu/ml/datasets.html