Modified Dynamic Algorithm of Data Clustering Using Fuzzy C Mean Algorithm

Priyanka Sharma, Anu Aggarwal

ABSTRACT*: *Clustering is a division of data into group of similar objects. Each group, called a cluster, consists of objects that are similar between themselves and dissimilar compared to objects of other group. Dynamic-means is a widely used clustering method. While there are considerable research efforts to characterize the key features of K-means clustering, further investigation is needed to reveal whether the optimal number of clusters can be found. This paper presents a modified Dynamic-means algorithm with the intension of improving cluster quality. In dynamic mean algorithm each data elements can be a member of one and only one cluster at a time. The proposed works apply Fuzzy c means algorithm over dynamic-means algorithm to improve the membership grade i.e. each data element can show their membership in each and every clusters.

Keywords: clustering, dynamic mean clustering and fuzzy c mean clustering.

I. INTRODUCTION

Recently various algorithms for clustering large data sets and streaming data sets have been proposed The focus has been primarily either on sampling or incrementally loading partial data, as much as can fit into memory at one time. The incremental approach generally keeps sufficient statistics or past knowledge of clusters from a previous run of a clustering algorithm in some data structures and uses them in improving the model for the future. Various algorithms for speeding up clustering have also been proposed. While many algorithms have been proposed for large and very large data sets for the crisp case, not as much work has been done for the fuzzy case.

As pointed out in the crisp case may not be easily generalized for fuzzy clustering. This is due to the fact that in fuzzy methods an example does not belong to a cluster completely but has partial membership values in most clusters. More about clustering algorithms can be found in. Clustering large amounts of data takes a long time. Further, new unlabeled data sets, which will not fit in memory, are becoming available. To cluster them, either sub sampling is required to fit the data in memory or the time will be greatly affected by disk accesses making clustering an unattractive choice for data analysis. Another source of large data sets is streaming data where you do not store all the data, but process it and delete it. There are some very large data sets for which a little labelled data is available and the rest of the data is unlabeled i.e. for example, computer intrusion detection. Semi-supervised clustering might be applied to this type of data. In general, clustering algorithms, which can process very large data sets, are becoming increasingly important. The work in this thesis is related to clustering algorithm. So it becomes important to have an overview of concept of data clustering. The basic steps of data clustering are:

1. Pattern representation (optimally including feature extraction and/or selection)
2. Defination of pattern proximity measure appropriate to the data domain
3. Clustering or grouping
4. Data abstraction(if needed)
5. Assessment of output(if needed)

Clustering technique is used for combining observed objects into clusters (groups), which satisfy tow main criteria:

- Each group or cluster is homogeneous; objects that belong to the same group are similar to each other.
- Each group of cluster should be different from other clusters, that is, objects that belong to one cluster should be different from the objects of other clusters.

Depending on the clustering technique, clusters can be expressed in different ways:

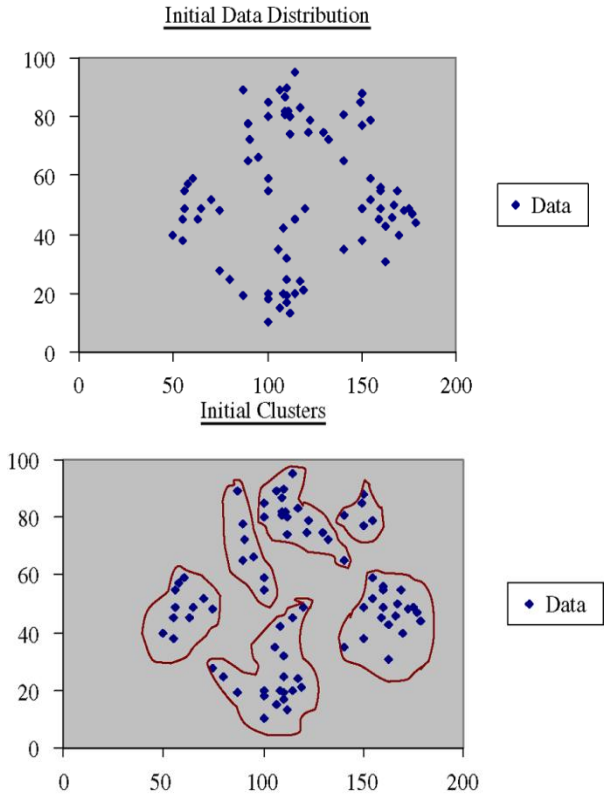
- Identified clusters may be exclusive, so that any object belongs to only one cluster.
- They may be overlapping; an object may belong to several clusters.
- They may be probabilistic, whereby an object belongs to each cluster with a certain probability.
- Clusters might have hierarchical structure, having crude division of objects at highest level of hierarchy, which is then refined to sub-clusters at lower levels.

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects, which are “similar” between them and are “dissimilar” to the objects belonging to other clusters fig shows the process of clustering with a simple graphical example:

Manuscript received on August, 2012

Ms.Priyanka Sharma,Department of computer science and engg., KurukshetraUniversity,Doon Valley Institute of Engg And Technology , Karnal, India.

MsAnu Aggarwal, Department of computer science and engg.,Kurukshetra university, Lecturer in Doon Valley Institute Of Engg. And Technology, Karnal, India.



A. The Goals of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But now to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion, which would be independent of the final aim of the clustering. Consequently, it is the user, which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties, in finding useful and suitable groupings or in finding unusual data objects (outlier detection). By grouping the objects in a cluster, a significant improvement in search and in further analysis can be obtained. The following general example illustrates the idea.

II. D-M CLUSTERING

Dynamic Means clustering approach is the new methodology to cluster the data objects into number of groups, which is unknown initially. Number of groups (clusters) is some positive integer. The grouping is done by measuring the distance between object and centroid. Objects are iteratively grouped into the existing clusters or a new cluster formation is done with those objects based up on the threshold limit. Thus the purpose of dynamic clustering is to classify the data. It could improve the chances of finding the global optima with careful selection of initial cluster. In this algorithm data objects are stored in secondary memory and transferred to main memory one at a time. Only the cluster representatives are stored permanently in main memory to alleviate space limitations.

Therefore, a space requirement of this algorithm is very small, necessary only for the centroids of the clusters. This algorithm is non-iterative and therefore its time requirement is also small.

Threshold limit allows maximum permissible distance between data object and centroid of any cluster. It can vary according to the density of database.

A. D-M Clustering Steps

To perform Dynamic Means clustering data objects are either stored in a data structure like array or then can be transferred to main memory one at a time. Then the Dynamic Means algorithm does the following steps until all objects are selected.

1. Assign the first data object to the first cluster.
2. Iterate until all objects are selected

i) Select next object. Determine the minimum distance between selected object and each centroid of existing clusters.

ii) Compare the distance with the threshold limit, group the object into existing cluster or a new cluster formation is done with that object.

The algorithm which elaborates this concept in detail as well as shows all above steps performed by dynamic clustering algorithm with the help of diagrams. Loop terminates after fixed number of steps equal to n-1. Its time complexity is lesser than time complexity of k-means. Time complexity of this algorithm is O(n) in best case and $O(n^2)$ in worst case.

B. D-M Clustering Flow Chart

The Dynamic Means clustering method clusters n data objects automatically into k cluster. It asks from user to enter data objects and a threshold limit. Threshold limit allows maximum permissible distance between a data object and centroid. Initially, from a cluster from first data object. Next step is to find the distance between next data object and centroid of existing clusters. Assign data object to same cluster or make a new cluster based upon the threshold limit. Repeat this procedure until all objects are selected.

Algorithm:

D-m clustering algorithm (D)

1 letk=1
2 K_k={dK}
3 K={K_k}
4 $C_k=d_k$
5 Assign n some constant valve to T_{th}
6 for I=2 to n do
7 Determine distance (m) between d_i and each centroid c_i of any k_i in k such that, m is minimum.(1<=j<=k)
8 if(m<=T_{th}) then
9 $k_j=k_j U\ d_i$
10 Calculate new mean (Centroid c_j) for cluster k_j;
11 else k=k+1
12 $k_k=d_i$
13 K=K U k_k
14 $C_k=d_i$

III. MODIFIED-DYNAMIC ALGORITHM

Dynamic mean clustering approach is the new methodology to cluster the data objects into number of groups, which is unkown initially .Number of groups (clusters) is some positive integer.

The grouping is done by measuring the distance between objects and centroid .Objects are iteratively grouped into existing clusters or a new cluster formation is done with those objects based on the threshold limit. This method clusters n data objects automatically into k cluster.It asks from user to enter data objects and threshold limit but in this method each data element can be a member of one and only one cluster at a time. In other words we can say that the sum of membership grades of each data point in all clusters is equal to one and in all the remaining clusters its membership grade is zero. The proposed work is to modified the dynamic mean algorithm .The proposed method finds the membership of each data element in every clusters means each data point show their membership in each and every cluster.

A. Modified-Dynamic Clustering Algorithm

The modified dynamic clustering algorithm is as follow:

Input: k: number of clusters (for dynamic clustering initialize k=2)
Fixed number of clusters = yes or no (Boolean).
D: a data set containing n objects.
V:set center cluster
U: partition matrix

Output: A set of k clusters with membership grade.

Method:

1. Arbitrarily choose k objects from D as the initial cluster centers.
2. Repeat step1 until D is empty
3. (re)assign each object to the cluster to which the object is most similar, based on mean value of the objects in the cluster.
4. Update the cluster means i.e. calculate mean value of the objects for each cluster.
5. Until no change.
6. Fixed-no-of-clusters=no
7. Compute inter-cluster distance.
8. Compute intra-cluster distance.
9. If intra Dist < oldintra Dist and inter Dist >oldinter Dist Then
10. K=K+1
11. Choose number of cluster given by Dynamic clustering
12. Calculate C center[vi] for each step.
13. Calculate the distance matrix.
14. Update the partition matrix.
15. If $\|U^{(k+1)} - U^{(k)}\| < \delta$ otherwise goto step 12
16. STOP.

IV. RESULT AND ANALYSIS

More and more efforts have been done to improve scalability of Dynamic mean algorithm, which can be applied to a variety of databases of different sizes, in effective and efficient manner. Effective means the good quality clusters are produced while efficient means clusters are producing in optimal time. But only a few efforts have been done to improve Dynamic mean algorithm, which produce good clusters automatically. In Dynamic mean algorithm each data element can be a member of one and only one cluster at a time. In other words we can say that the sum of membership grades of each data point in all clusters is equal to one and in all the remaining clusters its membership grade is zero .In our thesis dynamic algorithm is modified using fuzzy algorithm. By applying fuzzy algorithm over Dynamic algorithm we can show the membership of each data element in all clusters.

Clustering can be at an incredibly faster rate when the proposed algorithm is applied. It is applicable to a large amount of data stored in repositories . The overall results are significant in showing that modified Dynamic algorithm show membership of each data element in every clusters using MATLAB.

MATLAB is a numerical computing environment and fourth-generation programming language Developed by Math Work, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C,C++ Java, and Fortran.

Matlab is a case sensitive. Typical uses include :

- Math and computation
- Algorithm development
- Data acquisition Modeling, simulation, and prototyping
- Data analysis, exploration, and visualization
- Scientific and engineering graphics
- Application development, including graphical user interface building

MATLAB is an interactive system whose basic data element is an array that does not require dimensioning.

A. Comparison Between Dynamic Aad Modified Dynamic Algorithm

DYNAMIC MEAN ALGORITHM	MODIFIED DYNAMIC MEAN ALGORITHM
In dynamic algorithm each data element can be a member of one and only one cluster at a time.	In modified dynamic algorithm each data element can show a membership in all clusters.
In dynamic algorithm time complexity increased for large dataset.	In modified dynamic algorithm time complexity can be decreased.
Dynamic algorithm produce clusters in which membership grade is zero.	Modified dynamic algorithm produce clusters in which membership grade is shown in each clusters.

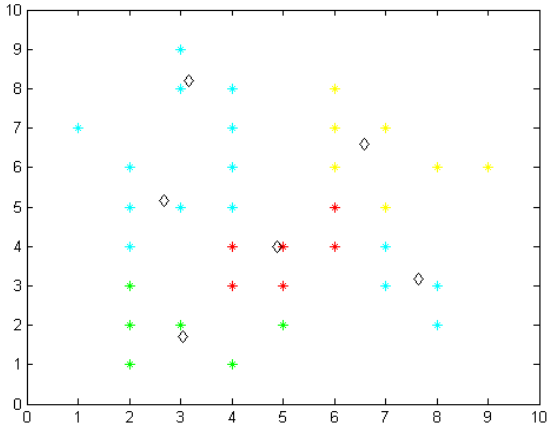


Fig: Modified dynamic clustering for 1st dataset

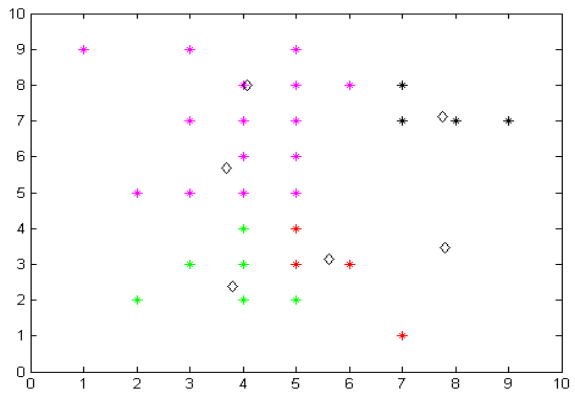


Fig: Modified dynamic clustering for 2nd dataset

V. CONCLUSION

Clustering determines the relationships between data objects in the database. The objects are clustered or grouped based on the principle of "maximizing the infraclass similarity and minimizing the interclass similarity". It finds out something valuable from database. Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning etc. Clustering methods can be divided into various types: Partitioning methods, Hierarchical methods, Density based methods, Grid-based methods; Model based methods, Probabilistic techniques, and Graph theoretic and Fuzzy methods. The Dynamic mean algorithm are the major focus of this thesis work.

Dynamic mean algorithm produce good clusters automatically because there is no need to defined the number of clusters beforehead but in Dynamic mean algorithm each data element can be a member of one and only one cluster at a time. In other words we can say that the sum of membership grades of each data point in all clusters is equal to one and in all the remaining clusters its membership grade is zero .In our thesis dynamic algorithm is modified using fuzzy algorithm. By applying fuzzy algorithm over Dynamic algorithm we can show the membership of each data element in all clusters .By applying Fuzzy algorithm over Dynamic algorithm clustering can be at an incredibly faster rate. It is applicable to a large amount of data stored in repositories. The overall results are significant in showing that Dynamic algorithm show membership of each data element in every clusters.

VI.FUTURE SCOPE

Clustering has been an active research area and most of the research has focused on effectiveness and scalability of algorithms. Many clustering algorithms exits in the literature from partition based, non-parametric density estimation based methods, graph theoretical based, to empirical and hybrid approaches. They all are underlying some concept about data organization and cluster characteristics to find interesting patterns or cluster in the given dataset. The Dynamic mean algorithm are the major focus of this thesis work, which is widely used in many applications still there, exits some drawbacks in this algorithm.

The proposed work in this thesis is to modified dynamic algorithm using fuzzy algorithm. One main drawback in Dynamic mean algorithm is that the sum of membership grades of each data point in all clusters is equal to one and in all the remaining clusters its membership grade is zero. To overcome this drawback fuzzy c mean algorithm is applied over dynamic mean algorithm. Future work can focus on how to develop more scalable and computationally algorithms for large datasets and develop clustering algorithms that can provide concise explanation on the characteristics of the objects that were assigned to each clusters.

REFERENCE

1. Ahamad Shafeeq and Hareesha"Dynamic clustering of data with modified K-mean algorithm", 2012.
2. Ran Vijay Singh and M.P.S Bhatia , "Data Clustering with Modified K-means Algorithm", IEEE International Conference on Recent Trends in Information Technology, ICRTIT 2011, pp 717-721.
3. Shi Na., Liu Xumin, Guan Yon , "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics(IITSI), pp.63-67, 2-4 April 2010.
4. Grigorios F. Tztzis and Aristidis C. Likas, "The Global Kernel k-Means Algorithm for Clustering in Feature Space", IEEE Trans. On Neural Networks, Vol. 20, No. 7, July 2009, pp. 1181-1194.
5. Wei Li, "Modified K-means clustering algorithm", IEEE computer society Congress on Image and Signal Processing, 2008, pp. 618-621.
6. Mohamad Alata, Mohamad Molhim and Abdullah Ramini"Produce and optimization of Fuzzy C mean clustering algorithm using GA", 2008.
7. David Altman, Efficient Fuzzy Clustering of Multi-spectral Images, FUZZ-IEEE , 1999.
8. Vicenc Torra, 2004" Fuzzy C-means For fuzzy hierarchical clustering".

AUTHOR PROFILE

Ms.Priyanka Sharma doing mtech in Department of Computer Science and Engg from Doon Valley Instituteof Engg & Technology, Karnal-132001, Kurukshetra University.published her paper titled Modified Dynamic Algorithm Of Data Clustering Using Fuzzy Algorithm..In this paper we are trying to modify dynamic algorithm using fuzzy algorithm.

MsAnu Aggarwal,Department of computer science and engg.,Kurukshetra university, Lecturer in Doon Valley Institute Of Engg. And Technology, Karnal, India