Clustering Uncertain Data Based on Fuzzy C-Means Clustering

Sobin Mathew[1]Mr.R.Hariharan[2]
[1]SecondYear M.E (Computer Science andEngineering) [2]M.E Assistant Professor (Computer Science and Engineering)
[1,] Maharaja Prithvi Engineering College, Avinashi - 641654

Abstract---Data clusteringis the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible. Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed.In many applications, data contains inherent uncertainty. A number of factors contribute to the uncertainty such as the random nature of the physical data generation and collection process and measurement error. As an example, consider the problem of clustering mobile devices continuously according to the periodic updates of their locations. One application of the clustering is the selection of a device as the leader for each cluster.Clustering on uncertain data, one of the essential tasks in mining uncertain data, posts significant challenges on both modeling similarity between uncertain objects and developing efficient computational methods. Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method developed by is frequently used in pattern recognition.

FCM provides more value for validity index.Using FCM, more accurate clustering can be performed.

I. Introduction

A. Clustering Uncertain Data:

Probabilistic data is becoming more and more common through various analysis and acquisition techniques. This is problematic because many of the existing methods for classifying and clustering data are meant to work with n-dimensional points. Instead of certain points, we are now faced with uncertain regions in n-dimensional space. This provides to be a more difficult problem. New methods have slowly been introduced for learning from uncertain data; however, it is far from a solved problem.The well-known Kullback-Leibler divergence to measure similarity between uncertain objects in both the continuous and discrete cases, and integrate it into partitioning and density-based clustering methods to cluster uncertain objects. To tackle the problem, estimate KL divergence in the continuous case by kernel density estimation and employ the fast Gauss transform technique to further speed up the computation.In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzylogic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster.

Extensive experiment results verify the effectiveness, efficiency, and scalability of approaches. Here we are going to compare the K-Medoids and FCM clustering methods.

B. Challenges In Clustering Uncertain Data:

The previous studies on clustering uncertain data are largely various extensions of the traditional clustering algorithms designed for certain data. As an object in a certain data set is a single point, the distribution regarding the object itself is not considered in traditional clustering algorithms. Thus, the studies that extended traditional algorithms to cluster uncertain data are limited to using geometric distance-based similarity measures, and cannot capture the difference between uncertain objects with different distributions.

II. Problem Statement

The previous studies on clustering uncertain data are largely various extensions of the traditional clustering algorithms designed for certain data. As an object in a certain data set is a single point, the distribution regarding the object itself is not considered in traditional clustering algorithms. Thus, the studies that extended traditional algorithms to cluster uncertain data are limited to using geometric distance-based similarity measures, and cannot capture the difference between uncertain objects with different distributions. Specifically, three principal categories exist in literature, namely partitioning clustering approaches, density-based clustering approaches, and possible world approaches. The first two are along the line of the categorization of clustering methods for certain data, the possible world approaches are specific for uncertain data following the popular possible world semantics for uncertain data. As these approaches only explore the geometric properties of data objects and focus on instances of uncertain objects, they do not consider the similarity between uncertain objects in terms of distributions.

Let us examine this problem in the three existing categories of approaches in detail. Suppose have two sets AA and IB of uncertain objects. The objects in AA follow uniform distribution, and those in IB follow Gaussian distribution. different distributions. Partitioning clustering approaches extend the k-means method with the use of the expected distance to measure the similarity between two uncertain objects. The expected distance between an object P and a cluster center c (which is a certain point) where FP denotes the probability density function of P and the distance measure distance is the square of Euclidean distance. In , it is proved that cP is equal to the distance between the center (i.e., the mean) P:c of P and c plus the variance of P. That is, Accordingly, P can be assigned to the cluster center Thus, only the centers of objects are taken into account in these uncertain versions of the k-means method. In this case, as every object has the same center, the expected distance-based approaches cannot distinguish the two sets of objects having different distributions.

Density-based clustering approaches extend the DBSCAN method and the OPTICS method in a probabilistic way. The basic idea behind the algorithms does not change objects in geometrically dense regions are grouped together as clusters and clusters are separated by sparse regions. However, in case, objects heavily overlap. There is no clear sparse regions to separate objects into clusters. Therefore, the density-based approaches cannot work well. Possible world approaches follow the possible world semantics. A set of possible worlds are sampled from an uncertain data set. Each possible world consists of an instance from each object. Clustering is conducted individually on each possible world and the final clustering is obtained by aggregating the clustering results on all possible worlds into a single global clustering. The goal is to minimize the sum of the difference between the global clustering and the clustering of every possible world. Clearly, a sampled possible world does not consider the distribution of a data object since a possible world only contains one instance from each object. The clustering results from different possible worlds can be drastically different. The most probable clusters calculated using possible worlds may still carry a very low probability. Thus, the possible world approaches often cannot provide a stable and meaningful clustering result at the object level, not to mention that it is computationally infeasible due to the exponential number of possible worlds.

A. Disadvantages:

- Problems due to accuracy of the Output.
- Waste of resource for recalculation of the Clusters.
- Not Suitable for uncertain data.

III. PROPOSED SYSTEM

Fuzzy c-means (FCM) is very famous and representative method in clustering algorithms. The FCM is based on hard c-means (HCM) and has been constructed by fuzzi classification of HCM. Some FCMs is used in the field of clustering. Each FCM corresponds with the way to fuzzify the HCM. Particularly, the entropy regularized FCM is known as effective in FCMs. By the way, there are many cases that data has some errors in clustering. Until now, the errors have been represented by interval values [3, 4] in these case. But the way is not adequate because only the boundary of interval values are considered and calculated frequently in these algorithms for the data with the errors. Therefore, try to formulate these error problems into the optimization problems with inequality constraints and construct new clustering algorithms through solving the problems in it. The rest will denote the tolerance "k which means the permissible range of the error, introduce the tolerance into the optimization problems and formulate the problems. The next, will solve the problems by using Kuhn-Tucker conditions. The last, will construct new algorithms based on the solutions.

A. Advantages:

- Validity of the output is increased.
- Low computational cost as compared to Kl-Divergence Method.
- Suitable For Uncertain Data.

IV. CLUSTERING IMPLEMENTATION

In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

One of the most widely used fuzzy clustering algorithms is the Fuzzy C Means (FCM) Algorithm.

The lists of modules used in the project are

- Data Acquisition
- K-Medoids Clustering
- FCM
- Result
- Comparison

A. Data Acquisition:

Data acquisition module extract dataset available for processing the data mining applications. One is synthetic and other is real time dataset. The process of acquiring the dataset is carried on this module. Once the data set is acquired.it has to be converted into suitable structure for further processing by the algorithm. Java collections are used to represent the data from the dataset.

B. K-Medoids:

The uncertain k-medoids method consists of two phases, the building phase and the swapping phase. Building phase. In the building phase, the uncertain k-medoids method obtains an initial clustering by selecting k representatives one after another. The first representative Ci is the one which has the smallest sum of the KL divergence to all other objects in the set.

$Ci = argmin_P(\sum_P D\ (P1 P)$----------(1)

Ci is cost between points P1 and P.

For each object P which has not been selected, test whether it should be selected in the current round. For any other nonselected object P0, P0 will be assigned to the new representative P if the divergence D(P1 || P) is smaller than the divergence between P0 and any previously selected representatives. Therefore, calculate the contribution of P0 to the decrease of the total KL divergence by selecting P as

$\max(0, min_j \left(D(P1 C_j)\right) - D(P1 P))$----------(2)

Calculate the total decrease of the total KL divergence by selecting P as the sum over the contribution of all nonselected object, denoted by DEC(P). Then, the object to be selected in the i^{th} iteration is the one that can incur the largest decrease that is,

$Ci = argmax_P(DEC(P))$----------(3)

The cost between points is calculated using KL divergence.

C. Fcm:

In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. An overview and comparison of different fuzzy clustering algorithms is available.

Any point x has a set of coefficients giving the degree of being in the k^{th} cluster wk(x). With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

The degree of belonging, wk(x), is related inversely to the distance from x to the cluster center as calculated on

the previous pass. It also depends on a parameter m that controls how much weight is given to the closest center. The fuzzy c-means algorithm is very similar to the k-means algorithm.

- Choose a number of clusters.
- Assign randomly to each point coefficients for being in the clusters.
- Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than , the given sensitivity threshold) :
  $c_{k=}(\sum_x w_k(x)x)/(\sum_x w_k(x))$----------(4)
- Compute the centroid for each cluster, using the formula (4)
- For each point, compute its coefficients of being in the clusters, using the formula above.
- The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means; the minimum is a local minimum, and the results depend on the initial choice of weights.

Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes. Another algorithm closely related to Fuzzy C-Means is Soft K-means.

D. Result:

The result module displays the output of the clustering, The output are shown as tabular data. The result is shown as clustered data which represents the weather details with different attributes. Each time the updated result may be obtained.

V. COMPARISON

In Comparison module the algorithm is compared based on different techniques. The basic techniques included time complexity,space complexity and cluster validity. Each of them are explained below.

A. Time Complexity:

The total time required for the algorithm to run successfully, and produce an output. T(A) = End Time – Start Time.

B. Space Complexity:

The space complexity is denoted by the amount of space occupied by the variables or data structures while running the algorithm.

S(A) = End { Space(Variables) + Space(Data Structures) } – Start { Space(Variables) + Space(Data Structures)}

C. Cluster Validity:

The validity of the cluster is measured based on validity index. The validity measure indicates the correctness of the cluster. The higher the validity measure , the more valid the cluster. Two types of validity measures are used here :

Xie-Beni's Index

Fukuyama-Sugeno's Index

Xie-Beni'sIndex(XB) is given by:

$$X1 = \sum_{k=1}^{n}\sum_{i=1}^{c}(Uki)^{m}\|x_k - v_i\|^2$$

$$X2 = n\, min_{i,j}\|v_i - v_j\|^2$$

XB=X1/X2

Where X1=Xie-BeniNumarator

X2=Xie-Beni Denominator

XB=Xie-Beni Index

VI. CONCLUSION

In the existing system uncertain data is clustered using K-Medoids clustering. Kullback-Leibler divergence method is used for the similarity measurement, and systematically define the KL divergence between objects in both the continuous and discrete cases. The integration of KL divergence into the partitioning and density based clustering methods was take place.The proposed system is clustering uncertain data using FCM.It is a soft clustering type of implementation. FCM provides high validity index measure than that of K-Medoids.FCM minimizes intra-cluster variance. The comparison between the K-Medoids and FCM will take place in terms of validity index measures. The more accurate clustering of uncertain data can be implemented using FCM. As future enhancement Soft K-Means algorithm can implement to obtain more accurate results.

VII. ACKNOWLEDGMENTS

I express my sincere and heartfelt thanks to our chairman Thiru. K.PARAMASIVAM B.Sc., and our Correspondent Thiru. P.SATHIYAMOORTHY B.E., MBA., MS., for giving this opportunity and providing all the facilities to carry out this project work.

I express my sincere and heartfelt thanks to our Respected Principal Dr. A.S.RAMKUMAR, M.E., Ph.D., MIE., for provided me this opportunity to carry out this project work.

I wish to express my sincere thanks toMrs.A.BRINDA M.E., Assistant Professor, and Head of the Department of Computer Science and Engineering for all the blessings and help rendered during the tenure of my project work.

I am indebted to my project guide,Mr R.HARIHARAN M.E., Assistant Professor for her constant help and creative ideas over the period of project work.

I express my sincere words of thankfulness to members of my family, friends and all staff members of the Department of Computer Science and Engineering for their support and blessings to complete this project successfully.

REFERENCES

[1] Bin Jiang, Jian Pei, Yufei Tao, XueminLin,"Clustering Uncertain Data Based on Probability Distribution Similarity", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 4, APRIL 2013.

[2] A.K.Jain,M.NMurty,P.JFlynn,"Data Clustering: A Review",A ACM Computing Surveys, Vol. 31,No.3 , September 1999.

[3] Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences", J. Machine Learning Research, vol. 6, pp. 1705-1749, 2005.

[4] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent DirichletAllocation",J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[5] I.S. Dhillon, S. Mallela, and R. Kumar, "A Divisive Information- Theoretic Feature Clustering Algorithm

for Text Classification", J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.

[6] M. Ankerst, M.M. Breunig, H.-P. Kriegel, and J. Sander, "Optics:Ordering Points to Identify the Clustering Structure", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 1999.

[7] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proc.Second Int'l Conf. Knowledge Discovery and Data Mining (KDD), 1996.

[8] N.N. Dalvi and D. Suciu, "Management of Probabilistic Data: Foundations and Challenges", Proc. ACM SIGMOD-SIGACTSIGART Symp. Principles of Database Systems (PODS), 2007.

[9] R. Cheng, D.V. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2003.

