IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 4, Ver. V (Jul.-Aug. 2016), PP 12-16
www.iosrjournals.org

An Efficient Methodology for Clustering Uncertain Data Based on Similarity Measure

Manisha Padole[1], Prof. Sonali Bodkhe[2]
[1](Department of computer science & Engineering, G. H. Raisoni Academy of Engineering & Technology,Nagpur, India)
[2](Department of computer science & Engineering, G. H. Raisoni Academy of Engineering & Technology,Nagpur, India)

Abstract: *In data mining clustering is one of the most important and crucial task. There exists various clustering techniques like partitioning and density based techniques. The previous methods worked on traditional partitioning methods which are found unreliable to handle uncertain data as uncertain objects are geometrically indistinguishable. Data uncertainty can be occurred in various data collecting systems like sensors, weather information collecting systems, etc.... Clustering gained much interest in clustering of uncertain data because the clustering algorithms works on the certain data so there is a need to do work on the algorithms to be as useful for uncertain data. The proposed work introduces the well-known Skew Divergence to measure the similarity between uncertain data. Previous methodology worked on KL-Divergence as well as on Janson-Shannon Divergence as the similarity measure. As KL-Divergence is very costly as compared to Skew Divergence or even infeasible. Skew Divergence with FCM creates a good combination and proves to be the best method of clustering uncertain data.*

Keywords: *Clustering, FCM, KL-Divergence, Skew Divergence, Uncertain Data*

I. Introduction

In recent years various methodologies came in research while there is discussion about clustering data. The new research moved toward the interesting concept of Uncertain Data which attracted so many researchers towards itself. Uncertain data is of certain interest because it cannot be handled by the traditional clustering approaches without any modification. Uncertainty in data can be found in big databases where data from various resources came. Databases like weather information, sensor databases, online customer reviews on particular products and so on are of very big interests as they give predictions about future possibilities.

Sometimes it can happen that there is problem while gathering weather information due to some environmental fluctuations. This can make uncertainty in database. In other example consider online customer review on mobile. Different users have different reviews, so, this does not cause uncertainty. Uncertainty occurs when different users give review on its different features. For example a mobile may have good battery backup, good camera quality but RAM is less. Some user may be happy with this RAM but some found it very less. This can happen with other features as well. So many researches has been done on uncertain data because the uncertainty in collected information cannot be ignored. Uncertain data is also as important as certain data because it can contain important information.

In another example, a weather station monitors different weather conditions. It contains different measures like precipitation, humidity, wind speed, temperature, pressure and direction. If daily records are compared they will vary from day to day. This dissimilarity in data can be considered as uncertain object represented by a distribution over the space formed by several measurements. For grouping the weather conditions of one month from different station then there is a need to cluster the uncertain data according to their distributions.

Limitations in Existing Clustering Methodologies

The traditional approaches for clustering of uncertain data uses the traditional clustering approaches and their techniques. The traditional clustering methods works on geometric distances of uncertain objects which is not the case with uncertain objects as they are distributed over different measures. So there is need to calculate their distributions which is the basic characteristics of uncertain data.

In partitioning approaches, if the well-known k-means approach has been considered, it uses the expected distance approach to calculate the difference between uncertain objects. The expected distance between cluster centre c and the object P is calculated using the formula, $ED(P,c) = \int pfp(x)dist(x,c)dx$, fp is the probability density function of P and dist(x,c) is the square of Euclidean distance. But [14] proved that, expected distance is equal to distance between the cluster centre and uncertain data centre i.e. P.c plus the variance of P as follows,

$$ED(P,c) = dist(P.c,c) + var(P) \quad (1)$$

So, P can be assigned to the centre of the cluster as argminc{ ED(P,c)}= argminc{ dist(P.c,c) }. Thus in this uncertain version of k-means only cluster centres are considered. In this case every object has the same distance which is not distinguishable by the algorithms based on expected distance. They will not find the different distributions of the object.

In Density-Based approaches, the objects in geometrically dense regions are grouped together in a cluster while they are separated from the sparse regions. The DBSCAN method [32] and OPTICS method [36] worked on uncertain objects in probabilistic way and does not change the basic idea. In this case, objects heavily overlap. There is no clear sparse region through which can be separate out different clusters. Thus, these approaches cannot be considered for clustering of uncertain data.

II. Literature Survey

Samir N. Ajani [1] proposes an improved K-means algorithm. They proposed the method improved k-means in which they combined k-means algorithm with Indexing and vornoi clustering. First indexing is applied on the input set then k-means clustering is applied to find the clusters. Indexing decreases the iteration time of k-means algorithm which will increase the efficiency of k-means algorithm. Again vornoi clustering is applied on the output set to refine the clustering results of k-means.

In [2], it has mentioned that by calculating simple geometric distances between data objects cannot be applied to uncertain data as well. UK-mean [3] is an extension to the traditional K-mean algorithm to handle uncertain data object. UK- mean algorithm require to compute expected distance between each object and to obtaining expected distance is very costly because computation of ED function involves probability function. Probability density functions are different and arbitrary. The major computational cost of the UK-mean algorithm is the evaluation of Expected distance(ED).

In [4] Geetha and Mary Shyla proposed kernel skew divergence as the similarity measure for both the continuous and discrete cases. The KSD method showed the best result as compared to KL Divergence as a similarity measure. Kernel skew divergence proved to be the time reducing and increase the speed of clustering as compared to KL-Divergence. Aliya Edathadathi, Syed Farooq and Balachandran KP [5] proposed Algorithm for modified K-medoids clustering based on KL divergence method. They modified K-medoids method to find the best cluster. They showed that using modified K-medoids the efficiency of clustering can be improved.

The fuzzy version of DBSCAN [6] algorithm is known as FDBSCAN [7] algorithm. It is similar to DBSCAN with minor changes. In [8], another algorithm called OPTICS, unable to produce clusters of data set as it creates augmented ordering of the database to represent the density based clustering structures. In FOPTICS [9], uses probability function to show the similarity between fuzzy objects. It used fuzzy distance function for measuring the similarity between uncertain objects. This algorithm has integrated fuzzy distance function with OPTICS hence named as FOPTICS.

III. Proposed Plan of Work

In the proposed plan of work it uses Fuzzy C Means as the Clustering algorithm and Skew Divergence as the similarity measure. FCM clustering method allows one piece of data to belong to two or more clusters. That is why FCM is a very common algorithm used in clustering methodologies. It assigns a membership to each data point on the basis of the distance between the data point and cluster centre. More the data is near to the cluster centre more is its membership towards the cluster centre. So the summation of membership of each data point is one. It is based on minimization of following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad (2)$$

where m (the Fuzziness Exponent) is any real number greater than 1, N is the number of data, C is the number of clusters, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the i^{th} of d-dimensional measured data, c_j is the d-dimension centre of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the centre. Following are the steps of FCM algorithm:

Step 1: Randomly select cluster centre from given N points

Step 2: Calculate the u_{ij} using:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \quad (3)$$

Step 3: At k-step, calculate the centre vectors $C^{(k)} = [c_j]$ with $U^{(k)}$, i.e. calculate the new cluster centres:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m}{\sum_{i=1}^{N} u_{ij}^m} \quad (4)$$

Step 4: Update the member function as:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (5)$$

Step 5: If $(U_{ij})_{current} - (U_{ij})_{previous} < \varepsilon$ or the minimum j is achieved, then STOP; otherwise return to Step 2.

In the above algorithm the distance between x_i and c_j is calculated using Skew Divergence and not by using the Euclidean distance. The Skew Divergence is calculated using the following formula:

$$\begin{aligned} SD(q,r) &= D(r \,||\, \propto q + (1-\propto)r) \\ &= D(r||z) \\ &= r * \frac{\log r}{\log z} \end{aligned} \quad (6)$$

where α controls the degree to which the function approximates D(r||q), r is the datapoint, q is the centroid, α must be closest to 1 to achieve smooth distribution i.e. $0.5 < \alpha < 1$.The flowchart in Fig.(1) shows proposed plan of work. It shows the work done step by step by following the algorithm which was described above:

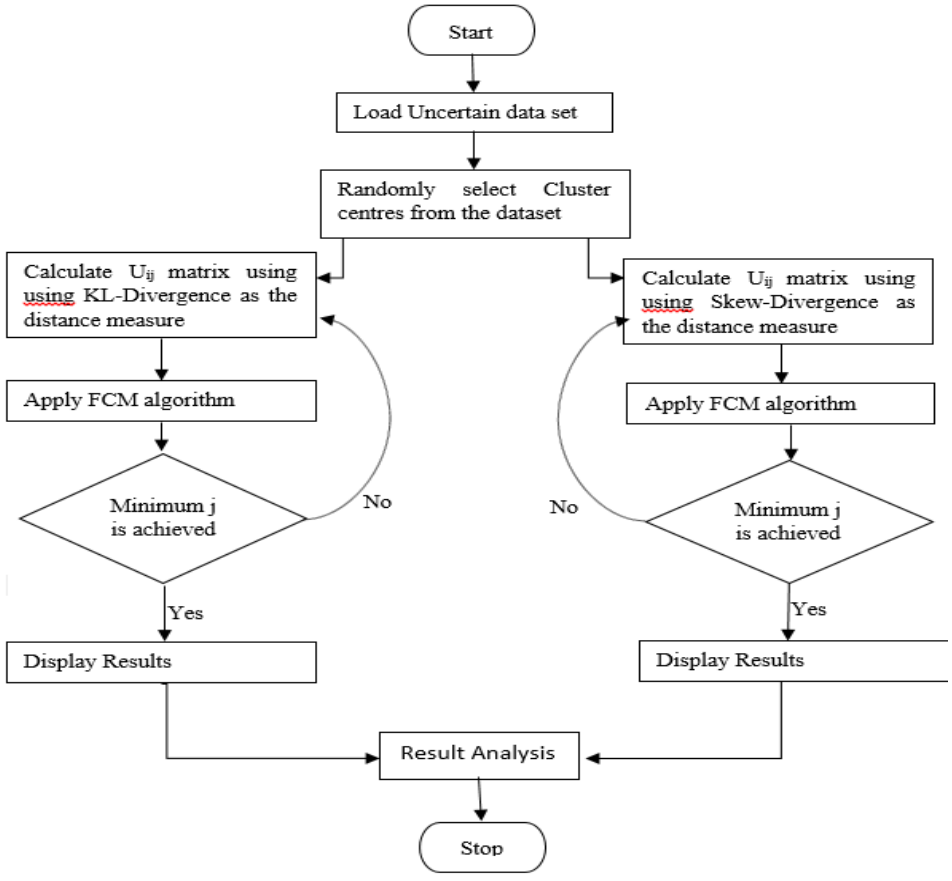

Fig. 1 Flowchart of Proposed Plan of Work

IV. Experimental Results

As it has performed clustering using FCM using KL-Divergence as well as Skew Divergence, it shows better results using Skew-Divergence. FCM with Skew-Divergence requires less time to calculate the clusters, which is displayed in the following graph. The table shows the time requirement for different number of clusters. Then graphs showing the calculation time when different number of clusters are considered. Clearly from the graph it can be predicted that Skew-Divergence is better than KL-Divergence. It can also be seen from Table (1). The table shows the different reading with both of the algorithms. The time is compared for both of the algorithms by considering the different number of clusters. Basically, it considers three, four, five and six number of clusters just to simplify the results, so that user can easily distinguish the difference between these two algorithms.

Sr. No	No. of Clusters	FCM-KL Time	FCM-Skew Time
1	3	60,668	44,398
2	4	39,578	28,329
3	5	46,098	37,097
4	6	55,630	37,706

Table 1 Time requirement for different number of clusters**.**

Graph showing time requirement when three centroids are considered. The first reading is for FCM with KL Divergence and the second one is for FCM with Skew Divergence. It is clear that the time requirement for FCM with Skew Divergence is very less comparatively to the FCM with KL Divergence. Hence respectively it considers four, five and six number of centroids for getting the better results. The results are same for each condition i.e. for all four different readings FCM with Skew Divergence prove to be better than FCM with KL Divergence. Fig. 2 shows the calculated readings in the graphical form:

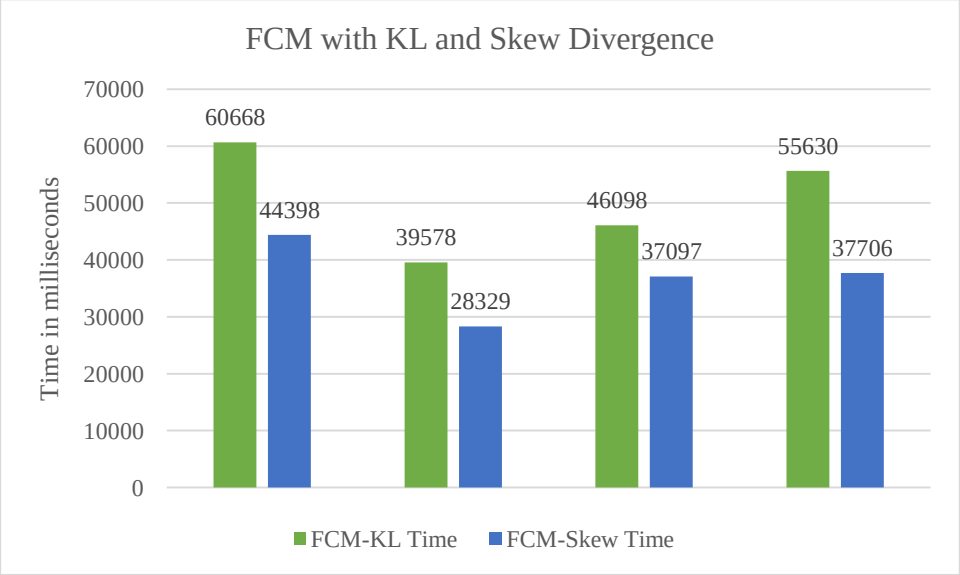

Fig. 2 Graph showing results with different number of clusters

V. Conclusion

This paper introduces the concept of FCM with Skew Divergence for clustering the uncertain dataset. The proposed measure is quite different from conventional ones from the viewpoint of evaluating clustering results with uncertain data. It uses Skew-Divergence as the similarity measure instead of Euclidean Distance which is the conventional method to calculate the distance between cluster center and data point. The combination of FCM with Skew-Divergence shows better results as compared to the conventional methods.

VI. Future Work

` In the future, besides the simple time calculations, there are another factors which help to reduce the time like fast gauss transformation, kernel estimation, etc. If FCM with Skew Divergence is combined with the above time reducing methods then the results can be again better.

References

[1] Samir N. Ajani, Prof. Mangesh Wanjari, "Clustering of Uncertain Data Objects using Improved K-means Algorithm", Volume 3, Issue 5, May 2013, ISSN: 2277 128X.

[2] S.D. Lee Ben Kao, Department of Computer Science, The University of Hong Kong, Reynold Cheng, Department of Computing, Hong Kong Polytechnic University, "Reducing UK-means to K-means".

[3] Priyadarshini J., Akila Devi.S, Askerunisa.A, "Kullback-Leibler Divergence Measurement for Clustering Based On Probability Distribution Similarity", International Journal of Innovative Research in Science, Engineering and Technology, volume 3, Special Issue 3, March 2014.

[4] Geetha and Mary Shyla, "An Efficient Divergence and Distribution Based Similarity Measure for Clustering Of Uncertain Data", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

[5] Aliya Edathadathil, Syed Farook, Balachandran KP, "A Modified K-Medoid Method to Cluster Uncertain Data Based on Probability Distribution Similarity", International Journal Of Engineering And Computer Science Issn:2319-7242, Volume 3, Issue 7, July 2014, Page No. 6871-6875.

[6] 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.

[7] Reshma MR and Suchismita Sahoo, Student and Asst. Prof at KMEA Engineering College, Kerala, India, "HANDLING UNCERTAINTY AND CLUSTERING IN UNCERTAIN DATA BASED ON KL DIVERGENCE TECHNIQUE", IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555, Vol. 3, No. 5, October 2013.

[8] 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.

[9] Mr. V. V. Kulkarni, Prof. V. V. Bag, "Clustering Multi-Attribute Uncertain Data Using Jenson-Shannon Divergence", International Journal of Application of Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 8, August 2014, ISSN 2319 - 4847.