International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Review on Clustering of Uncertain Data

Manisha Padole[1], Sonali Bodkhe[2]

[1]Department of Computer Science and Engineering, G.H.R.A.E.T. College of Engineering, Nagpur, India

[2]Professor & Head of Computer Science and Engineering, G.H.R.A.E.T. College of Engineering, Nagpur, India

Abstract: ***Clustering is a crucial task in data mining. There are so many techniques for data mining which are continuously emerging since the concept of data mining has been taken into account. All the clustering methodologies focus on certain data or the known data. But very few methodologies focus on the clustering of uncertain data. By considering the uncertain data for clustering, the results of clustering algorithm get affected and can show any unwanted results. For example, in the sensor based applications the output can be uncertain if there is any errors in input (e.g. noise occurred during capturing information). In such situations, the algorithm must be strong enough to consider the uncertain data because the uncertain data is also important. As there is continuous increase in data accumulation in the databases we need to be more aware about to handle uncertain data. In this paper we will discuss about the data uncertainty and the various approaches to handle and cluster the uncertain data.***

Keywords: Clustering, Uncertain data, FCM, KL-Divergence, clustering algorithm

1. Introduction

From the last few years there are so many researches has been done on uncertain data. As there is continuously increase in the information so there is also increase in data accumulation in the data bases and data warehouses. So many methodologies are came into account to clear the data uncertainty problem. Various data mining tools have been discovered to solve the uncertainty problem.

As there is huge data in the database and data warehouses which is continuously increasing day by day the various strategies are applied on these data to manage and handle it. Data mining tools access the information from these data reservoirs to generate the pattern to get insight into the data. As the amount of data is very large to understand and is not possible for a user to read it thoroughly and make decisions according to it. This can make a user to get into confusion and may sometimes create a headache for the user. The data mining applications uses various techniques to extract the exact information from the huge data set. The data mining tools generates the patterns which can be easily understood by the users. It helps the user to make good decisions and makes his task easy. The mining applications has techniques to extract the information. The patterns are generated using the similarity between the objects collected in the extraction. Various Classification or Clustering techniques are applied on the data to get the result. The mining applications extract the complete and certain information. Following is the figure showing process of cluster formation.

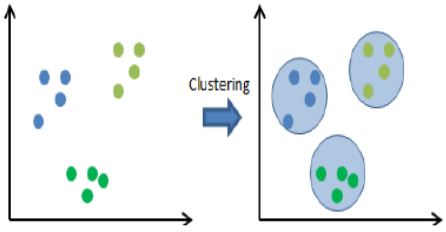

Figure 1: Formation of Clusters [20]

As the data is very large there is also a large amount of uncertain data in the reservoirs. The uncertainty in the data can be the result of any of the following situations described below:

- The uncertainty may be due to wrong input given by the wireless sensor. As it may happen that there is noise in input [1].
- In the online marketing research, e.g. users are supposed to rate the cameras. He / She will rate the camera according to his/her comfort. Different users rate the camera on its different features (e.g. battery backup, mega pixels, image quality, user friendliness and so on). The market analysis task [2] is to find the best camera among these by using clustering method. But this type of data creates uncertainty.
- In the weather forecasting system, various data is collected at the same time based on various features in the atmosphere [2]. (E.g. temperature, humidity, precipitation, wind speed, etc.) The uncertainty is due to the day to day variation in the record.

The data mining application faces various challenges while managing and clustering these uncertain data. The problem occurs when the results of data mining applications are affected if the uncertain data is considered for clustering. The uncertain data causes problem due to its uncertainty. In the next section we will discuss the various issues related to uncertain data and their clustering techniques. We will also discuss the problems occurred during applying various techniques and their solutions.

The paper is organized into following sections: In Section 2, we will discuss various clustering methods. In section 3, we suggested some Methodologies. In Section 4, we will go through the conclusion. We will examine the clustering problem and the mining approach to cluster uncertain data.

2. Literature Survey

In the recent years there are so many researches on clustering uncertain data. So many authors has proposed their different algorithms for clustering uncertain data. Uncertain data

clustering is a very hot topic in the field of clustering as various algorithms does not consider uncertain data for clustering. Many of the times uncertain data affects the results of proposed algorithms as they are not robust to uncertain data.

The traditional clustering approach focuses on the geometric distance based similarity measures for clustering. In [14], it has mentioned that by calculating simple geometric distances between data objects cannot be applied to uncertain data as well. In [2], it has mentioned that previous approaches only focus on instances of uncertain data but does not find the similarity between them if we focus on their distribution.

Consider an example [2] in which we are taking two sets A and B. Both sets are collection of uncertain objects. Objects in A follows Uniform distribution and objects in B follows Gaussian distribution. If we are finding mean value [2] and suppose we get same mean value for all objects in both the sets, then it is obvious that both sets forms two clusters because of their different distributions.

Consider partitioning approach K-Means. In this technique, [14] extended the k-means method for measuring the similarity between two uncertain objects by using expected distance. Expected distance is the distance between the object P and cluster center c as:

$$ED(P,c) = \int_p f_p(x) dist(x,c) dx$$

where fp is the probability density function of P and dist(x,c) is the square of euclidean distance. In [14], it is proved that expected distance is equal to distance between center of P and c plus the variance of P, i.e.

$$ED(P,c) = dist(P.c,c) + Var(P).$$

where, P.c is the center of object P.

In this way only center of uncertain objects are taken into consideration. But this is not the case with us, as our all uncertain data will have the same center. So the traditional approach will not be able to find the dissimilarity between the objects.

There is another version of k-means called UK-means proposed by Ngai et al. which is specially designed for uncertain data and very specifically known as Uncertain K Means. This is the extension of K-means method. This algorithm calculates the expected distance between the uncertain object and the cluster center. In [14] S.D Lee Ben Kao and Reynold Cheng showed that the UK-means can be reduced to K-means on certain data points. In [5] Mr. V. V. Kulkarni and Prof. V.V. Bag performed clustering on multi-attribute uncertain data. They used Jenson-Shannon Divergence and Kullback-Leibler Divergence as a similarity measure. They considered data only for discrete case. In this approach for discrete data, Jenson-Shannon Divergence showed good results.

In [6] 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 Farook and Balachandran KP [13] proposed Algorithm for modified *K*-medoid clustering based on *KL* divergence method. They modified K-medoid method to find the best cluster. They showed that using modified K-medoid the efficiency of clustering can be improved.

3. Proposed Plan of Work

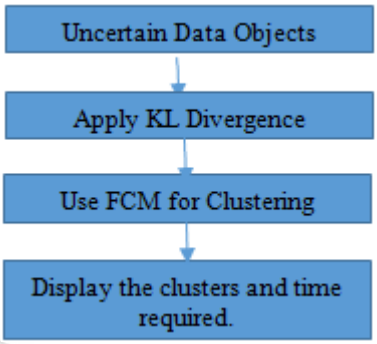

Figure 2: Proposed Work Approach

In proposed work we will use KL-Divergence for calculating the similarity measure. The Kullback–Leibler divergence, also called discrimination information divergence, KL divergence, is a measure of the difference between two probability distributions P and Q. It is not symmetric in P and Q. In applications, P typically represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while Q typically represents a theory, model, description, or approximation of P. Although it is often intuited as a way of measuring the distance between probability distributions, the Kullback–Leibler divergence is not a true metric. It does not obey the triangle inequality, and in general DKL(P‖Q) does not equal DKL(Q‖P).

After finding the similarity we will use FCM for calculating the clusters. Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. The algorithm is composed of the following steps:

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
2. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$

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

4. *If* $|| U^{(k+1)} - U^{(k)}|| < \varepsilon$ *then STOP; otherwise return to step 2.*

In the FCM approach, instead, the same given datum does not belong exclusively to a well-defined cluster, but it can be placed in a middle way. In this case, the membership function follows a smoother line to indicate that every datum may belong to several clusters with different values of the membership coefficient. So FCM shows good results as per our observation and creates accurate clusters better than K-Means.

4. Conclusion

In this paper various clustering techniques are well discussed. Also this paper discusses the pros and cons of various techniques. FCM shows the better results in clustering as compared to other partitioning methods. If FCM can be implemented with KL-Divergence then it will improve the performance of the FCM Algorithm and also will be even better for uncertain data objects. From various discussions it has been observed that if KL Divergence is used with FCM it can give better results. In future this proposed plan can be implemented to prove the increased performance of FCM.

References

[1] Charu C.Aggarwal, Senior Member, IEEE, and Philip S. Yu, Fellow, IEEE, "A Survey of Uncertain Data Algorithms and Applications", IEEE Transactions On Knowledge And Data Engineering, VOL 21, No. 5, May 2009.

[2] Bin Jiang, Jian Pei, Senior Member, IEEE, Yufei Tao, Member, IEEE, and Xuemin Lin, Senioe Member, IEEE, "Clustering Uncertain Data Based on Probability Distribution Similarity", IEEE Transactions On Knowledge And Data Engineering, VOL 25, No. 4, April 2013.

[3] Graham Cormode, AT&T Labs-Research and Andrew McGregor, UC San Diego, "Approximation Algorithms for Clustering Uncertain Data", PODS'08, June 9-12, 2008, Vancouver, BC, Canada.

[4] Kurada Ramachandra Rao, PSE Purnima, M Naga Sulochana, B Durga Sri, "Unsupervised Classification of Uncertain Data Objects in Spatial Databases Using Computational Geometry and Indexing Techniques", International Jouranal of Engineering Research and Applications(IJERA), ISSN:2248-9622, Vol. 2, Issue 2, Mar-Apr 2012, pp.806-814.

[5] Mr. V. V. Kulkarni and Prof. V. V. Bag, N. K. Orchid College of Engineering and Technology, Solapur-413002, "Clustering Multi-Attribute Uncertain Data Using Jenson-Shannon Divergence", International Journal of Application or Innovation in Engineering & Management(IJAIEM), Volume 3, Issue 8, August 2014.

[6] 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.

[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] Samir N. Anjani, Prof. Mangesh Wanjari, "An Approach for clustering uncertain data objects: A Survey", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013.

[9] Hans-Peter Kriegel, Martin Pfeifle, Institute for Computer Science University of Munich, Germany, "Hierarchical Density-Based Clustering of Uncertain Data", Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05), 1550-4786/05, 2005.

[10] 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.

[11] Xiancho Zhang and Han Liu and Xiaotong Zhang and Xinyue Liu, School of Software Technology Dalian University of Technology, Dalian 116620, China, "Novel Density-Based Clustering Algorithms for Uncertain Data", Association for the Advancement of Artigicial Intelligence, 2014.

[12] C.Deepika, R.Rangaraj, "An Efficient Uncertain Data Point Clustering Nased On Probability-Maximization Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2014.

[13] 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.

[14] 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".

[15] T. Soni Madhulatha, Associate Professor, Alluri Institute of Management Sciences, Warangal, "An Overview on Clustering Methods", IOSR Journal of Engineering, Apr. 2012, Vol. 2(4) pp:719-725.

[16] Ajit B. Patil, Prof. M.D.Ingle, "A Review of Clustering Algorithms for Clusteringn Uncertain Data", International Journal on Recent and Innovation Trends in Computing and Communication, Volume:2, Issue:11, ISSN:2321-8169,3643-3646.

[17] S.Geetha, E. Mary shyla, Dept. of Computer Science, Sri Ramakrishna College of Arts & Science for women Coimbatore, India, "A Survey of Clustering Uncertain Data Based Probability Distribution Similarity", IJCSNS International Journal of Computer Science and Network Security, VOL.14 No.9, September 2014.

[18] Pramod Patil, Ashish Patel, Parag Kulkarni, "Density-Based Clustering Based on Probability Distribution for Uncertain Data", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-3, Issue-5, June 2014.

[19] Mr. V.V.Kulkarni, Prof V.V. Bag, "Clustering Multi-Attribute Uncertain Data Using Jenson-Shannon Divergence", International Journal of Application or Innovation in Engineering & Management (IJAIEM) ISSN 2319 - 4847 Volume 3, Issue 8, August 2014.
[20] Shraddha K.Popat, Emmanuel M., Pune Institute of Computer Technology University of Pune India, "Review and Comparative Study of Clustering Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 805-812, ISSN:0975-9646.
[21] Lovely Sharma, Prof. K. Ramya, "A Review on Density based Clustering Algorithms for Very Large Datasets", International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 12, December 2013).
[22] Francesco Gullo, Andrea Tagarelli, "Uncertain Centroid based Partitional Clustering of Uncertain Data".
[23] Wang Kay Ngai, Ben Kao, Chun Kit Chui, Michael Chau, Reynold Cheng, Kevin Y. Yip, "Efficient Clustering of Uncertain Data", HKU 7134/06E.
[24] Miss Pragati Pandey, Miss Prateeksha Pandey, Mrs. Minu Choudhary, " Uncertain Data Algorithms and Applications", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 7, July 2012, ISSN: 2277 128X.
[25] 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.

Author Profile

Manisha Padole, Computer Science & Engineering Department, G.H.R.A.E.T., Nagpur, India.

Prof. Sonali Bodkhe, Computer Science & Engineering Department, India.