A Survey of Clustering Uncertain Data Based Probability Distribution Similarity

S.Geetha
Dept.of Computer Science
Sri Ramakrishna College of Arts & Science for women
Coimbatore, India

E. Mary shyla,MCA
Dept.of Computer Science
Sri Ramakrishna College of Arts & Science for women

Abstract
Clustering is one of the major tasks in the field of data mining .The main aim of the clustering is grouping the data or similar objects into one group based on their data find the similarity between the objects. Clustering of uncertain data have been becoming the major issues in the mining uncertain data for data mining or applications. Clustering uncertain data problems have been solved in many ways with the help of data mining techniques or algorithm. In recent work many data mining algorithms solve the issues of the uncertain data object. Generally uncertain data objects can be solved in two ways: measuring the similarity between the data objects or clustered data, measuring the similarity with data objects with Probability Distribution functions. Measuring the similarity between the data objects is based on a similarity distance measure and further clustered with density based clustering or hierarchical clustering methods. In recent years, a numeral of indirect data gathering methodologies has led to the propagation of uncertain data and developing efficient clustering methods. In recent work several datamining methods model uncertain data object. In this work, the uncertain data object has been represented by probability distribution similarity function. Generally the problem of uncertain data objects according to probability distribution happens in many ways. First the probability distribution method for model uncertain data object then after that measure the similarity between data objects using distance metrics, then finally best clustering methods such as partition clustering, density based clustering. This study focus on partition based clustering methods .The survey discusses different methodologies to process and mine uncertain data in a diversity of forms.

Index Terms
Clustering, Clustering uncertain data, Mining methods and algorithms, density based clustering, partition clustering .

1. INTRODUCTION

In generally Data Mining deals with the difficulty of extracting patterns from the information by paying suspicious attention to computing, communication and human-computer interface issues. Clustering is one of the major data mining tasks to group the similar information or data. All clustering algorithms aim of dividing the collection all data objects into subsets or similar clusters. A cluster is a collection of objects which are 'similar' between them and are 'dissimilar' to the objects belonging to other clusters [1]; and a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set. In data mining Clustering certain data have
been well studied in the various areas such as data mining, machine learning, Bioinformatics, and pattern recognition. However, there is only preliminary research on clustering uncertain data. In this study clustering uncertain data object problem have been solved with probability distribution function.

Clustering uncertain data

In many applications, data contain intrinsic uncertainty. Numerals of factors contribute the uncertainty such as the random nature of the physical data creation and collection procedure, measurement of error, and data staling. One purpose of the clustering is the selection of a device as the leader for each cluster. A leader's role is to collect data (such as location data) from its cluster members and to communicate with a server or a base station with batched updates. In this way, most communications are short-ranged messages among the cluster members and their leaders.
The previous studies on clustering uncertain data are largely various extensions of the traditional clustering algorithms considered for certain data. Here the object in certain dataset is considered as a single point and distribution concerning the object itself is not considered in traditional clustering algorithms. The study that extends conventional algorithms to cluster uncertain data that are restricted to using geometric distance-based similarity measures and cannot capture the dissimilarity between uncertain objects with diverse distributions.
The main areas of research are:
Modeling of uncertain data: A key issue is the process of modeling the uncertain data. Hence, the fundamental complexities have been captured while keeping the data helpful for database management applications.
Uncertain data mining: The results of data mining applications are affected by the underlying uncertainty in the data or objects. Therefore, it is difficult to design data mining techniques that can take such uncertainty into

Manuscript received September 5, 2014
Manuscript revised September 20, 2014

account during the computations. Generally the clustering of the data is categorized in three ways: Partitioning clustering approaches [5], [6], [7], Density-based clustering approaches [8], [9], and possible world approaches [10]. The first two are along the line of the categorization of clustering methods for certain data [11], the possible world approaches are specifically for uncertain data following the popular possible world semantics for uncertain data [12], [13], [14].

2. Related Work

In previous years several methods have been proposed to be mined the uncertain data objects [15-16]. The uncertain data object problem occurs in many applications , due to limitations of the underlying equipment (e.g., unreliable sensors or sensor networks), use of attribution, interruption or extrapolation techniques (to estimate the position of moving objects). Techniques such as the density based clustering algorithm in [8] and hierarchical clustering algorithm in [9] are useful for working with a specific application such as clustering or classification of data objects. A method of this nature has been proposed in [17] , a relaxed assumption is used that only the errors (in terms of standard deviation) of the records are known rather than the entire probability density function. This is a more realistic assumption of many scenarios, since it may often be possible to measure the standard deviation of an uncertain record, whereas the probability density function may be obtained only by more extensive theoretical modeling.

Mining Applications for Uncertain Data

Recently, a numeral of mining applications has been studied for the case of uncertain data. It includes clustering and classification with presence of uncertainty; it can affect the results of data mining applications significantly. For example classification application, an attribute which has lesser uncertainty is further useful than an attribute which has a superior level of uncertainty. Similarly a clustering function the attributes which have a superior level of uncertainty need to be treated in a different way from those which have a lesser level of uncertainty.

Clustering based uncertain data

Attribute space, somewhat than as a single point as usual unsaid when uncertainty is neglected. Mining technique that has been proposed for such data include clustering algorithms [5], [6], density estimation techniques [8], outlier detection [20]. Beside this recent body of work several methods have been proposed to analysis of interval-valued or fuzzy data, in which not well-known attributes are represented by intervals [21] and possibility distributions [22], [23].

Probability distributions, intervals and possibility distributions may be seen as three instances of a more general model, in which data uncertainty is articulated by means of belief functions. A belief function is also known as Dempster-Shafer presumption or Evidence presumption, was developed by Dempster [27] and Shafer [28], and was further elaborated by Smets [29]. A belief function may be seen both as a generalized set and as a non-additive measure, i.e., a generalized probability distribution. A belief function thus includes extension of set-theoretic operations, such union, intersection and extensions of probabilistic operation such as conditioning distribution and marginalization distribution for representing the data uncertainty has been mostly confined to classification.

In [30], a k-nearest neighbor rule based on Dempster-Shafer theory was introduced. In this method the generated rule make it possible to handle partially supervised data, in which uncertain class labels are repress by belief feature selection function. This rule was applied to regression problems with uncertain dependent variable. Methods for building decision trees from partially supervised data were proposed in [31], [32], [33]. An extension of the k-mode clustering algorithm for data with uncertain attributes was introduced in [34].

3. METHODOLOGY

Clustering is a primary data mining task. Clustering certain data has been considered for years in machine learning, data mining, pattern identification, Bioinformatics and other field's .However, there is only preliminary research on clustering uncertain data. Data uncertainty brings new challenges to clustering, since clustering uncertain data difficulty in the measurement of similarity between uncertain data objects. The majority of studies clustering uncertain data used distance-based similarity measures and few theoretical studies considered using divergences to measure the similarity between objects.

DBSCAN distance measure

Kriegel and Pfeifle [8] proposed the FDBSCAN algorithm which is a probabilistic extension of the deterministic DBSCAN algorithm [35] for clustering certain data. The fuzzy version of the DBSCAN algorithm (referred to as FDBSCAN) works in a similar way to the DBSCAN algorithm, except for not the density at a given point is uncertain because of the underlying uncertainty of the data points. It corresponds to the fact that the numeral of data points within the ε-neighborhood of a given data point can be estimated only probabilistically and is fundamentally an uncertain variable.

The goal is to define an implicit output in terms of ordering data points, so the DBSCAN is applied to this ordering of the data points, one can obtain the hierarchical clustering at any level for different values of the density parameter. The solution is to ensure that the clusters at different levels of the hierarchy with one to each other at the desired consistency level. If anyone of the observation is that clusters defined over a lower value an are completely contained in clusters defined over a higher value of b if the value of Min Pts is not varied. Therefore, the data points are prearranged based on the value of c required in order to obtain Min Pts in the ϵ-neighborhood. Here the data points with smaller values are processed first and then it is assured that higher density regions are always processed before lower density regions.

Hierarchical clustering techniques

DBSCAN is extended to a hierarchical density-based clustering method referred to as OPTICS [36] by Kriegel. An effective (deterministic) density based hierarchical clustering algorithm is OPTICS [36]. Here, the core idea in OPTICS is quite similar to DBSCAN and it is based on the concept of reachability distance between data points. While the method in DBSCAN defines a large-scale density parameter which is used as a threshold in order to define reachability. It ensures the DBSCAN algorithm is used for different values with this ordering, then a consistent result is obtained. The output of OPTICS algorithm is not the cluster membership, but it is the orders of data points are processed. OPTICS algorithm shares so many characteristics with the DBSCAN algorithm, it is comparatively easy to extend the OPTICS algorithm to the uncertain case using the same approach as that was used for extending the DBSCAN algorithm. It is referred to as the FOPTICS algorithm. In the uncertain case, this value is defined probabilistically, and the consequent expected values are used to order the data points.

Pfeifle et al [9] developed a probabilistic version of OPTICS called FOPTICS for clustering uncertain data objects. FOPTICS results a hierarchical categories in which data objects, as a replacement of the determined clustering membership for each object, and uncertain data objects are clustered. Volk et al. [10] followed the possible world semantics using Monte Carlo sampling [14]. This approach finds the clustering of a set of sampled possible worlds using existing clustering algorithms for certain data. Then, the final clustering is aggregate from individual sample clustering's.

Partitioning clustering

K-means & K-medoids are two partioning methods. K-means algorithm in order to cluster the data. This method is referred to as the UK-means algorithm. Ngai et al. [5] proposed the UK-means method extends the k-means method. The UK-means technique measures the distance between an uncertain object and the cluster center (which is a certain point) by the expected distance. Recently, Lee et al. [7] showed that the UK-means method can be reduced to the k-means method on certain data points. In UK-means, an object is assigned to the cluster whose representative has the smallest expected distance to the object. Hence, that the estimated distance computation is an expensive task. Therefore, the technique in [5] uses a number of pruning operations in order to reduce the computational consignment. The idea here is to use branch-and-bound (BB) techniques in order to minimize the number of expected distance computations between data points and cluster representatives. The wide idea is that once an upper bound on the minimum distance of a particular data point to some cluster representative has been quantified, it is essential to perform the calculation between this point and another cluster represent, if it can be proved that the consequent distance is greater than this band. This branch and bound (BB) approach is used to design an efficient algorithm for clustering uncertain location data.

Clustering Based on Distribution Similarity

Aware of the clustering distributions has appeared in the area of information retrieval when clustering documents [37], [38]. The major difference in the work does not assume any knowledge on the types of distributions of uncertain data objects. While clustering documents, every document is modeled as a multinomial distribution in the language model. To measure the similarity between the clustering, Xu and Croft [37] discussed a k-means clustering method with KL divergence as the similarity measurement between multinomial distributions of documents. Multinomial distributions KL divergence can be estimated using the numeral of occurrences of terms in documents.

Banerjee et al. [40] theoretically analyzed the k-means like iterative relocation clustering algorithms based on Bregman divergences which is an all-purpose case of KL divergence. They summarize a comprehensive iterative relocation clustering framework for a variety of similarity measures from the previous work from an information theoretical viewpoint. They showed that finding the best clustering is equivalent to minimize the loss function in Bregman information corresponding to the selected Bregman divergence used as the underlying similarity measure. In terms of effectiveness, their algorithms have a linear complication in every iteration with respect to the number of objects. However, they did not present methods for proficiently evaluating Bregman divergence nor calculating the mean of a set of distributions in a cluster. For uncertain objects problem which can have arbitrary

discrete or continuous distributions, it is critical to solve the above mentioned problems in large data sets.

4. Conclusion

The field of uncertain data object has seen a revival in recent years because of new ways of collecting data which have resulted in the need for uncertain representations. This paper surveys, the broad areas of probability based distribution similarity measurement techniques in the field along with the key representational issues in uncertain data management. It represents both the partitioning and density-based clustering methods with better clustering quality when using KL divergences and other divergence as similarity than using distance metric. The results confirm that KL divergence can naturally capture the distributional difference which geometric distance cannot capture best uncertain data object results. This paper proposes a plan to measure the clustering similarity with distribution function and apply the different divergence based similarity for the uncertain data object.

References

[1] M. Matteucci. "A Tutorial on Clustering Algorithms",http://home.dei.polimi.it /matteucc / Clustering /tutorial_html/.,2008.

[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Elsevier, 2000.

[3] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data, Prentice Hall, 1988.

[4] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, 1990.

[5] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "Efficient Clustering of Uncertain Data," Proc. Sixth Int'l Conf. Data Mining (ICDM), 2006.

[6] B. Kao, S.D. Lee, D.W. Cheung, W.-S. Ho, and K.F. Chan, "Clustering Uncertain Data Using Voronoi Diagrams," Proc. IEEE Int'l Conf. Data Mining (ICDM), 2008.

[7] S.D. Lee, B. Kao, and R. Cheng, "Reducing Uk-Means to k-Means," Proc. IEEE Int'l Conf. Data Mining Workshops (ICDM), 2007

[8] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), 2005.

[9] H.-P. Kriegel and M. Pfeifle, "Hierarchical Density-Based Clustering of Uncertain Data," Proc. IEEE Int'l Conf. Data Mining (ICDM), 2005.

[10] P.B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner, "Clustering Uncertain Data with Possible Worlds," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2009

[11] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Elsevier, 2000.

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

[13] A.D. Sarma, O. Benjelloun, A.Y. Halevy, and J. Widom, "Working Models for Uncertain Data," Proc. Int'l Conf. Data Eng. (ICDE), 2006

[14] R. Jampani, F. Xu, M. Wu, L.L. Perez, C.M. Jermaine, and P.J. Haas, "Mcdb: A Monte Carlo Approach to Managing Uncertain Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2008.

[15] C. Aggarwal and P. S. Yu, "A survey of uncertain data algorithms and applications," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, pp. 609–623, 2009.

[16] R. Cheng, M. Chau, M. Garofalakis, and J. X. Yu, "Guest editors' introduction: Special section on mining large uncertain and probabilistic databases," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 9, p. 1201, 2010.

[17] C.C. Aggarwal, "On Density Based Transformations for Uncertain Data Mining," Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE), 2007.

[18] C.C. Aggarwal and P.S. Yu, "A Framework for Clustering Uncertain Data Streams," Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE), 2008.

[19] G. Cormode and A. McGregor, "Approximation Algorithms for Clustering Uncertain Data," Proc. 27th ACM SIGMOD-SIGACTSIGART Symp. Principles of Database Systems (PODS), 2008.

[20] C. C. Aggarwal and P. S. Yu, "Outlier detection with uncertain data," in Proceedings of the SIAM International Conference on Data Mining (SDM 2008), Atlanta, Georgia, USA, 2008, pp. 483–493.

[21] L. Billard and E. Diday, Symbolic Data Analysis. Chichester, England: Wiley, 2006.

[22] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," Fuzzy Sets and Systems, vol. 1, pp. 3–28, 1978.

[23] J. Gebhardt, M. A. Gil, and R. Kruse, "Fuzzy set-theoretic methods in statistics," in Fuzzy sets in decision analysis, operations research and statistics, R. Slowinski, Ed. Boston: Kluwer Academic Publishers, 1998, pp. 311–347.

[24] P. Cazes, A. Chouakria, E. Diday, and Y. Schektman, "Extension de l'analyse en composantes principales `a des donn´ees de type intervalle," Revue de Statistique Appliqu´ee, vol. 14, no. 3, pp. 5–24, 1997.

[25] T. Denoeux and M.-H. Masson, "Principal component analysis of fuzzy data using autoassociative neural networks," IEEE Transactions on Fuzzy Systems, vol. 12, no. 3, pp. 336–349, 2004.

[26] P. Giordani and H. A. L. Kiers, "A comparison of three methods for principal component analysis of fuzzy interval data," Computational Statistics and Data Analysis, vol. 51, no. 1, pp. 379–397, 2006

[27] G. Shafer, A mathematical theory of evidence. Princeton, N.J.: Princeton University Press, 1976.

[28] P. Smets, "The combination of evidence in the Transferable Belief Model," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 5, pp. 447–458, 1990.

[29] P. Smets and R. Kennes, "The Transferable Belief Model," Artificial Intelligence, vol. 66, pp. 191–243, 1994.

[30] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," IEEE Trans. on Systems, Man and Cybernetics, vol. 25, no. 05, pp. 804–813, 1995.

[31] T. Denoeux and M. Skarstein-Bjanger, “Induction of decision trees for partially classified data,” in Proceedings of SMC’2000. Nashville, TN: IEEE, October 2000, pp. 2923–2928.
[32] Z. Elouedi, K. Mellouli, and P. Smets, “Belief decision trees: Theoretical foundations,” International Journal of Approximate Reasoning, vol. 28, pp. 91–124, 2001.
[33] S. Trabelsi, Z. Elouedi, and K. Mellouli, “Pruning belief decision tree methods in averaging and conjunctive approaches,” International Journal of Approximate Reasoning, vol. 46, no. 3, pp. 568–595, 2007.
[34] S. Ben Hariz, Z. Elouedi, and K. Mellouli, “Clustering approach using belief function theory,” in Artificial Intelligence: Methodology ,Systems, and Applications, ser. Lecture Notes in Computer Science, J. Euzenat and J. Domingue, Eds. Springer Berlin / Heidelberg, 2006, vol. 4183, pp. 162–171.
[35] 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.
[36] 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.
[37] J. Xu and W.B. Croft, “Cluster-Based Language Models for Distributed Retrieval,” Proc. 22nd Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 1999.
[38] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation,” J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.
[39] 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.
[40] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, “Clustering with Bregman Divergences,” J. Machine Learning Research, vol. 6, pp. 1705-1749, 2005.