A literature study on clustering the uncertainty data

S.SATHAPPAN[1], Dr.D.C.TOMAR[2]

[1] Research Scholar, Sathyabama University, Chennai, India
[2] Professor (IT), Jerusalem College of Engineering, Chennai, India

Abstract - Clustering is one of the important topics in data mining. The purpose of clustering is to group the similar data items. Clustering the uncertainty data is not an easy task but an essential task in data mining. Uncertain data represents the unexpected outcome. It is mostly found in the area of sensor networks. The uncertain data may have numerical and categorical data. For numerical clustering, the distance measure is based on geometric concepts such as Euclidean distance or Manhattan distance. Since the categorical data contains nominal values like [good, *bad*], [*low, medium, high*],the geometric distance measures are not applicable for categorical or nominal data. We used the numerical data (i.e.) Gas sensor data. From the literature in this field, it is noticed that very few attempts had been made for clustering gas sensor dataset using few methods and discuss the same in this paper.

Key words: *Clustering, data mining applications, uncertain data, k-means*

Introduction

Clustering problem is partitioning the dataset into different groups where one cluster consists of similar points whereas the other cluster consists of different points. In the real world, data mining applications are affected by uncertain data. Due to uncertainty, during the computations, designing of data mining technique has become critical. Uncertainty is caused by error measuring, sensor effects and other external factors such as Humidity, temperature etc. There has been a rapid awareness with uncertainty data while handling the database system and processed data. In recent years, many advanced technologies were developed for storing and recording continuous data in large quantities. These stored data contains numerical records as well as categorical data. Also Clustering the uncertain data have some significant challenges. To overcome these challenges engineers must apply effective methods to identify efficient way to cluster Uncertain data. Therefore, Priority R-Tree with k-Mode algorithm is proposed to improve the robustness and accuracy of the clustering outcome to a great extent. By minimizing the expected error with respect to the optimal classifier, experimental results display the cluster using the Gas sensor array drift Dataset. The motivation for Clustering of uncertain data is got from business perspective and a major concept in data mining since more and more applications such as sensor database, location database, biometric information systems . Clustering of uncertain data objects is a challenge in spatial databases. We intend to design a priority model for efficient clustering of uncertain data. Another motivation is to improve the clustering algorithm for producing quality clusters that can be used for any dataset which takes less time and cost. In addition, we use different datasets and compare them to produce best results.

Clustering and classification

Clustering and classification is a well studied area in data mining. Numerous Clustering algorithms have been proposed in the literature such as k-means, global kernel k-means, u-rule, u k-means algorithm, Fuzzy c-means algorithm. Initially Martin Ester, Hans-Peter Kriegel Jörg Sander and Xiaowei Xu (1996) proposed density based clustering algorithm called as **DBSCAN** (for *density-based spatial clustering of applications with noise*) for clustering uncertain data. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes to handle uncertain data. This was modified by Hans-Peter Kriegal and Martin Pfeifle (2005). Hans-Peter Kriegel and Martin Pfeifle.(2005) proposed density based clustering(DB Clustering) for clustering the data. The main difference of DB clustering is that there is no need to specify total number of clusters in advance. Here to find a cluster, DB clustering starts with an arbitrary point p and retrieves all points density reachable from p with respect to NEps (Neighbourhood points with maximum radius) and MinPts (Minimum number of points in an NEps neighbourhood). If p is a core point, this procedure yields a cluster with respect to NEps and MinPts. If p is a border point, no points are density reachable from p and DB visits the next point of the database . There is a need to specify NEps and MinPts, which can be difficult in practice.

Uncertainty –Lineage database

Benjelloun et al (2006) proposed Uncertainty-Lineage Database (ULDB) which uses database with both uncertainty and lineage for clustering uncertain data. In data management applications, the influence of data source is an important factor that should be accounted. S.D. Lee, B. Kao, and R. Cheng (2007) proposed a novel method for computing the EDs efficiently. It only works for a certain form of distance function. Hae-Sang Park and Chi-Hyuck Jun (2008) proposed K-medoid algorithm. It is a classical partitioning method to cluster the

data. A partitioning clustering method organizes a set of uncertain data into K number of clusters. Charu C Aggarwal and Philip S Yu (2008) discussed various methodologies for processing and mining uncertain data. They used a probabilistic database. A probabilistic-information database is a finite probability space whose outcomes are all possible database instances consistent with a given schema. This probabilistic database is also called "possible world's model" .The specification of such databases is unrealistic, since an exponential number of instances would be needed to represent the table. Therefore, the natural solution is to use a variety of simplified models which can be easily used for data mining and data management purposes.

K-means Algorithm

Shamir N Ajani (2013) proposed the k-means algorithm to handle the uncertainty data. Using the synthetic dataset, it clusters the data first to find the mean value and then applying to the nearer mean value in the cluster but it takes the large computation time and its variation depends on the initial k-value. Hence the indexing techniques are applied to the k-means algorithm and then the cluster generation time is significantly reduced. Grigorios F Tzortzis and Aristidis C Likas (2009) proposed the kernel k-means to handle the non-linearly separable cluster data and independent of initial k-value but it does not support the large dataset. The fast global kernel k-means algorithm supports the large dataset. Ben Kao and Sau Dan Lee (2010) used the uk-mean algorithm with pruning technique i.e. R-Tree. This method reduces the computation time. He proposed the pruning techniques that are based on Voronoi diagrams to reduce the number of expected distance calculations, but it handles only the linear separable data. Biao Qin and Yuni Xia (2010) used a new rule based on classification and prediction technique for classifying the uncertain data. This algorithm introduces new measures for generating, pruning, and optimizing the rules. But it is not efficiently pruning the data.

Simple matching measure

Joshua Zhexue Huang and Aranganayagi S (2015) used simple matching dissimilarity measure for categorical objects using k-mode algorithm with a heuristic approach which allows the use of the *k*-mode paradigm to obtain a cluster with strong intra-similarity, and to efficiently cluster large categorical data sets. It compares the two words and then finds the similarity and dissimilarity measures. The main aim of our paper is to derive rigorously the updating formula of the *k*-mode clustering algorithm with the new dissimilarity measure. reuse evolved from development of function calls in early programming languages and libraries of software routines for performing scientific calculations. In modern-day approaches of reuse covers the entire software life cycle and all software artifacts. Software reuse relies on preplanning to reuse a software component that meets the needs of the organizations involved in software development in new context. Le Li Zhiwen et al (2011) describes the automatic classification technique by soft classifier for the classification of uncertain data which appears in databases such as sensor ,location biometrics information databases with uncertainties. This data is generally imprecise in nature. This soft classifier technique is based on fuzzy c-means method with a fuzzy distance function to classify uncertain objects. The advantage of this method is that it works very well in uncertain data objects but not in certain data objects.

K-means algorithm & clustering

Velmurugan (2012) proposed a K-means algorithm to cluster the data. Here the given set of data is grouped into K number of disjoint clusters, where the value of K is to be fixed in advance. The algorithm consists of two separate phases: the first phase is to define K initial centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Generally Euclidean distance is used as the measure to determine the distance between data and the centroids. Then the centroids are recalculated and clustering is done with these new centroids. The process is repeated till the clusters are not changed . K-means method is not that much efficient to cluster the uncertain data, that is the main disadvantage. Carson Kai-Sang (2013) describes the Naive approach. It is for finding constraint frequent pattern from uncertain data and to find all the frequent patterns first approach , and checks these frequent patterns against the user constraints as a post-processing step – to filter out the patterns that do not satisfy the constraints.

Probability similarity

Bin Jiang and Jian Pei (2013) proposed a new method for clustering uncertain data based on their probability distribution similarity. The previous methods extend traditional partitioning clustering methods like K-means, UK means and density-based clustering methods to uncertain data, thus rely on geometric distances between data. Probability distributions, are essential for characteristics of uncertain objects. Here systematically model uncertain objects in both continuous and discrete domains, where an uncertain object is modelled as a continuous and discrete random variable, respectively. Then use the well-known Kullback-Leibler (KL) divergence to measure similarity between uncertain objects in both the continuous and discrete cases, and integrate it into K-medoid method to cluster uncertain data. Compared to the traditional clustering methods, K-Medoid clustering algorithm based on KL divergence similarity is more efficient.

Conclusions

In all existing methods , we see that they cannot handle uncertain objects that are geometrically indistinguishable, such as products with the same mean but very different variances in customer ratings. Though, significant progress has been made on clustering uncertain data , many important problems remain. In short, some of the key findings are:- Calculating the expected distance between an object and a cluster representative requires expensive integration computation; The expected distance calculation takes lot of time and cost to produce efficient clusters; Accuracy of producing clustering results is not good; Lack of methods for efficient clustering of uncertain data for gas sensor dataset. The traditional algorithms were focused neither categorical nor numerical data, but our proposed method is suitable for all kinds of data. The above experiments are proved our enhanced k-Mode algorithm efficiency with Gas Sensor's numerical values. The K-Mode algorithm and Probability density calculations are only used. So the complexity is also reduced. Moreover the computational cost is very low. The accuracy in producing the resulting clusters is good. The uncertainties caused by external factors are predicted to the corresponding cluster. It has been observed that if the clustering algorithm is combined with indexing method, then the clustering of uncertain data objects can be done very easily. There are scopes for further research in this direction. we are working on to find out efficient methods to initialize the modes. In future, we will be able to expand the work by using a hypothesis as keeping aggregate of clusters as mode for clustering uncertain data in a more optimal manner. The long term goal of this approach is to provide new algorithms that use combination of B Trees or B+ trees for improving the efficiency of clustering results. We intend to apply this model in a wider variety of applications for different real time datasets.

ACKNOWLEDGEMENTS

The authors thank Dr.S.Sridhar, Professor and Director, Cognitive & Central Computing of R.V.College of Engineering, Bangalore, India for reviewing this article and communicating to this Journal for publication.

REFERENCES

[1] Achtert E., Goldhofer S., Kriegel H.P., Schubert E. and Zimek A. (2012) , "Evaluation of clusterings – metrics and visual support", In Proc. ICDE, pp. 1285-1288.
[2] Achtert E., Kriegel H.P., Reichert L., Schubert E., Wojdanowski, R. and Zimek A. (2010), "Visual evaluation of outlier detection models", in Proc. DASFAA.
[3] Ackermann M.R., Blomer J. and Sohler C. (2008), "Clustering for Metric and Non-Metric Distance Measures", Proc. Ann. ACM-SIAM Symp. Discrete Algorithms (SODA).
[4] Aggarwal C.C. (2008), "On Unifying Privacy and Uncertain Data Models," Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE).
[5] Aggarwal C. and Yu P. S. (2009), "A survey of uncertain data algorithms and applications", IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 5, pp. 609-623.
[6] Aggarwal C. C. (2009), "On High-Dimensional Projected Clustering of Uncertain Data Streams," in ICDE Conference, 2009 (poster version). Full version in IBM Research Report.
[7] Aggarwal C.C. (2007), "On Density Based Transformations for Uncertain Data Mining", Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE.
[8] Aggarwal C.C. and Yu P.S. (2008), "A Framework for Clustering Uncertain Data Streams", Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE).
[9] Aggarwal C.C. and Yu P.S. (2008), "Outlier Detection with Uncertain Data", Proc. SIAM Int'l Conf. Data Mining (SDM).
[10] Antova L., Jansen T., Koch C. and Olteanu D. (2008) , "Fast and Simple Relational Processing of Uncertain Data", Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE).
[11] Aranganayagi S.and Thangavel K. (2009), "Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure", International Journal of Information and Mathematical Sciences.
[12] Banerjee S. Merugu, Dhillon I.S. and Ghosh J. (2005), "Clustering with Bregman Divergences," J. Machine Learning Research, Vol. 6, pp. 1705-1749.
[13] Ben Kao Sau, Dan Lee David W., Cheung Wai-Shing and Ho K. F. Chan (2010), "Clustering Uncertain Data using Voronoi Diagrams", IEEE.
[14] Benjelloun O., Das Sarma A., Halevy A. and Widom J. (2006), "ULDBs: Databases with Uncertainty and Lineage", Proc. 32nd Int'l Conf. Very Large Data Bases (VLDB).
[15] Bernecker T., Kriegel H.P., Renz M., Verhein F. and Zufle A. (2009), "Probabilistic frequent itemset mining in uncertain databases", in Proc. KDD.
[16] Bi J. and Zhang T. 2005, "Support vector classification with input data uncertainty", in Advances in Neural Information Processing Systems 17, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, pp. 161-168.
[17] Bin Jiang G, Jian Pei, Yufei Tao and Xuemin Lin (2013), "Clustering Uncertain Data Based on Probability Distribution Similarity" IEEE.
[18] Bin Wang, Gang Xiao, Hao Yu and Xiaochun Yang (2011), "Distance-Based Outlier Detection on Uncertain Data", IEEE Eleventh International Conference on Computer and Information Technology.
[19] Burdick D., Deshpande P., Jayram T., Ramakrishnan R. and Vaithyanathan S. (2005), "OLAP Over Uncertain and Imprecise Data," in VLDB Conference Proceedings.
[20] Carson Kai-Sang (2013) , "An Approach for clustering uncertain data objects: A Survey", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol. 2, Issue 6.
[21] Charu C. Aggarwal (2009), "A Survey of Uncertain Data Algorithms and Applications", IEEE Transactions on knowledge and data engineering.
[22] Cheng R., Kalashnikov D. and Prabhakar S (2004), "Querying Imprecise Data in Moving Object Environments," IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, pp. 1112-1127.
[23] Cheng R., Kalashnikov D. and Prabhakar S. (2003), "Evaluating Probabilistic Queries over Imprecise Data," Proceedings of the ACM SIGMOD International Conference on Management of Data, June 2003.

[24] Cheng R., Chau M., Garofalakis M. and Yu J. X. (2010), "Guest editors' introduction: Special section on mining large uncertain and probabilistic databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 9, pp. 1201.
[25] Cormode G. and McGregor A. (2008), "Approximation Algorithms for Clustering Uncertain Data", Proc. Symp. Principles of Database Systems (PODS), M. Lenzerini and D. Lembo, eds., pp. 191-200.
[26] Das Sarma A., Benjelloun O., Halevy A. and Widom J. (2006), "Working Models for Uncertain Data," in ICDE Conference Proceedings.
[27] Dhillon I.S., Mallela S. and Kumar R. (2003), "A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification", J. Machine Learning Research, Vol. 3, pp.1265-1287.
[28] Dunham H. and Sridhar S. (2006), Introduction to Data mining, Pearson Education.
[29] Giordani P. and Kiers H. A. L. (2006), "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.
[30] Hae-Sang Park & Chi-Hyuck Jun (2008), "A simple and fast algorithm for K-medoids clustering", Elsevier.
[31] Hans-Peter Kriegal and Martin Pfeifle (2005), "Density-based clustering of uncertain data", Proceeding of the eleventh ACM SIGKDD International conference on Knowledge discovery in data mining , pp.672 .
[32] Jampani R., Xu F., Wu M., Perez L.L., Jermaine C.M. and Haas P.J. (2008), "Mcdb: A Monte Carlo Approach to Managing Uncertain Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD.
[33] Kao B., Lee S. D., Lee F. K. F., Cheung D. W. L., & Ho W. S. (2010), "Clustering uncertain data using Voronoi diagrams and R-tree index", IEEE TKDE, Vol. 22, No. 9, pp. 1219-1233.
[34] Kriegel H.P. and Pfeifle M. (2005), "Hierarchical Density-Based Clustering of Uncertain Data," Proc. IEEE Int'l Conf. Data Mining (ICDM.
[35] Le Li Zhiwen, Yul Zijian and Fengl Xiaohang Zhangl (2011), "Automatic Classification of Uncertain Data by Soft Classifier", Proceedings of the 2011 International Conference on machine Learning and Cybernetics, GuiJin.
[36] Lee S.D., Kao B. and Cheng R. (2007), "Reducing UK-Means to KMeans", Proc. First Workshop Data Mining of Uncertain Data (DUNE), in Conjunction with the Seventh IEEE Int'l Conf. Data Mining (ICDM).
[37] Lurong Xiao and Edward Hung (2007), "An Efficient Distance Calculation Method for Uncertain Objects", Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007) .
[38] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu (1996) " A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" , International Conference on Knowledge Discovery and Data Mining.
[39] Ngai W. K., Kao B., Cheng R., Chau M., Lee S. D., Cheung D. W. and Yip K. Y. (2011), "Metric and trigonometric pruning for clustering of uncertain data in 2D geometric space", Information Systems, Vol. 36, No. 2, pp. 476–497.
[40] Nick Larusso (2009), "A Survey of Uncertain Data Algorithms and Applications", IEEE Transaction on Knowledge and Data Engineering.
[41] Reynold Cheng, Xike Xie, Man Lung Yiu, Jinchuan Chen and Liwen Sun (2010), " UV-Diagram: A Voronoi Diagram for Uncertain Data", ICDE Conference IEEE.
[42] Samir Anjani H and Mangesh Wangjari (2013), "Clustering of uncertain data object using improved K-Means algorithm", IJARCSSE.
[43] Sarma A.D., Benjelloun O., Halevy A.Y. and Widom J., (2006), "Working Models for Uncertain Data", Proc. Int'l Conf. Data Eng. (ICDE).
[44] Sonia Manhas, Rajeev Vashisht, Parvinder S. Sandhu and Nirvair Neeru (2010), "Reusability Evaluation Model for Procedure Based Software Systems", International Journal of Computer and Electrical Engineering, Vol.2, No.6, pp. 1107-1110.
[45] Sujatha S. and Shanthi Sona A. (2013), " New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method", IJERT.
[46] Velmurugan T. (2012), "Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points", IJCTA.
[47] Volk Habich, Clemens Utzny, Ralf Dittmann and Wolfgang Lehner (2007), "Error-Aware Density-Based Clustering of Imprecise Measurement Values", Seventh IEEE International Conference on Data Mining Workshops, ICDM Workshops, IEEE.
[48] Volk P.B., Rosenthal F., Hahmann M., Habich D. and Lehner W. (2009), "Clustering Uncertain Data with Possible Worlds", Proc. IEEE Int'l Conf. Data Eng. (ICDE).
[49] Xu J. and Croft W.B. (1999), "Cluster-Based Language Models for Distributed Retrieval," Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR).
[50] Z¨ufle A., Emrich T., Schmid K. A., Mamoulis N., Zimek A. and Renz M. (2014), " Representative clustering of uncertain data", in Proc. KDD, pp: 243–252.