A study on Cluster Uncertain Data based on Probability Distribution

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 on uncertain data, one of the essential tasks in data mining. The traditional algorithms like K-Means clustering, UK-Means clustering, density based clustering etc, to cluster uncertain data are limited to using geometric distance based similarity measures,and cannot capture the difference between uncertain data with their distributions. Such methods cannot handle uncertain objects that are geometrically indistinguishable, such as products with the same mean but very different variances in customer ratings [6]. In the case of K-medoid clustering of uncertain data on the basis of their KL divergence similarity, they cluster the data based on their probability distribution similarity. Several methods have been proposed for the clustering of uncertain data. In this paper, Some of these methods are reviewed. Compared to the traditional clustering methods, K-Medoid clustering algorithm based on KL divergence similarity is more efficient.

Keywords: Uncertain data clustering, Probability distribution.

Introduction

Clustering is one of the most important research areas in the field of data mining. In simple words, clustering is a division of data into different groups. Data are grouped into clusters in such a way that data of the same group are similar and those in other groups are dissimilar. Clustering is a method of unsupervised learning. Uncertainty in data arises naturally due to random errors in physical measurements, data staling, as well as defects in the data collection models. The main characteristics of uncertain data are, they change continuously, we cannot predict their behavior, the accurate position of uncertain objects is not known and they are geometrically indistinguishable. Because of these reason it is very difficult to cluster the uncertain data by using the traditional clustering methods .Clustering of uncertain data has recently attracted interests from researchers. This is driven by the need of applying clustering techniques to data that are uncertain in nature, and a lack of clustering algorithms that can cope with the uncertainty.

Related Research Work

Various algorithms have been proposed for the clustering of uncertain data. Researchers are always trying to improve the performance and efficiency of the clustered data.Dr.T. Velmurugan.[1] 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.

Improved K-Means algorithm

Samir N. Ajani[2] proposes an improved K-means algorithm to cluster the uncertain data effectively called UK means (Uncertain K means) clustering algorithm. UK-means basically follows the well-known K-means algorithm except that it uses Expected distance (ED) calculation instead of using Euclidian distance. Initially, k arbitrary point's c1. . . ck are chosen as the cluster representatives. Then, UK-means repeats the following steps until the result converges. First, for each data di, Expected Distance (di, cj) is computed for all centroids and data. Data di is then assigned to cluster cj that minimizes the Expected Distance. Here the computation of Expected distance involves numerically integrating functions, it is difficult to calculate. The Expected distance calculation is one of the main problem in UK-means clustering. The efficiency of UK-means clustering is improved if the ED calculation is reduced.

ED Calculation in UK Means method

Ben Kao [3] proposes a new method to reduce the ED calculation in UK-means method. Here cluster the uncertain data by using UK means method with voronoi diagrams. Martin Ester.[4] proposes a density based clustering(DB Clustering) for clustering the data. The main difference of DB clustering is that here we do not 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 wrt. 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 wrt. 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 [5]. Need to specify NEps and MinPts, which can be difficult in practice.

Probability distribution

Bin Jiang and Jian Pei.[6] 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, which are essential 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.

Uncertain Objects and Probability Distributions

Consider an uncertain object as a random variable following a probability distribution. We consider both the discrete and continuous cases. If the data is discrete with a finite or countable infinite number of values, the object is a discrete random variable and its probability distribution is described by a probability mass function (pmf). Otherwise, if the domain is continuous with a continuous range of values, the object is a continuous random variable and its probability distribution is described by a probability density function (pdf). For example, the domain of the ratings of cameras is a discrete set and the domain of temperature is continuous real numbers. After finding the probability distribution we have to find the probability distribution similarity between the data. Kullback-Leibler divergence (KL divergence) is one of the main method to calculate the probability distribution similarity between the data [7]. We show that distribution differences cannot be captured by the previous methods based on geometric distances. We use KL divergence to measure the similarity between distributions, and demonstrate the effectiveness of KL

divergence using K-medoid clustering method.

Applying KL divergence into K-medoid algorithm

K-medoid is a classical partitioning method to cluster the data[8]. A partitioning clustering method organizes a set of uncertain data into K number of clusters. Using KL divergence as similarity, Partitioning clustering method tries to partition data into K clusters and chooses the K representatives, one for each cluster to minimize the total KL divergence. K-medoid method uses an actual data in a cluster as its representative. Here use K-medoid method to demonstrate the performance of clustering using KL divergence similarity. The K-medoid method consists of two phases, the building phase and the swapping phase. Clustering uncertain data based on their probability distribution similarity is very efficient clustering method compare to other methods. But in the building phase the algorithm select initial medoids randomly that affect the quality of the resulting clusters and sometimes it generates unstable clusters which are meaningless. Also here the initial partition is based on the initial medoids and the initial partition affect the result and total number of iterations. If the initial medoids are selected in an efficient way then it does not produce any empty clusters and also we can reduce the total number of iterations.

CONCLUSION

Several works on clustering uncertain data are studied in detail. Clustering uncertain data based on their probability distribution similarity is more efficient. Random selection of initial medoid is the main drawback of probability distribution based clustering method.

ACKNOWLEDGEMENTS

The authors thank Dr.S.Sridhar, Director, RVCT, R V College of Engineering, Bangalore for communicating this paper for publication in the form of literature study article.

References

[1] Dr. T. Velmurugan “Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points ”. IJCTA, 2012
[2] Samir Anjani and Prof. Mangesh Wangjari. “Clustering of uncertain data object using improved K-Means algorithm” IJARCSSE, 2013
[3] Ben Kao Sau Dan Lee Foris K. F. Lee David W. Cheung and Wai-Shing Ho.” Clustering Uncertain Data using Voronoi Diagrams and R-Tree Index” IEEE, 2010
[4] Hans-Peter Kriegel and Martin Pfeifle.” Hierarchical Density-Based Clustering of Uncertain Data” IEEE, 2005
[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu.” A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise” International Conference on Knowledge Discovery and Data Mining.
[6] Bin Jiang, Jian Pei, Yufei Tao and Xuemin Lin. “Clustering Uncertain Data Based on Probability Distribution Similarity”IEEE, 2013
[7] Fernando Perez Cruz.”Kullback-Leibler Divergence Estimation of Continuous Distributions”
[8] Hae-Sang Park and Chi-Hyuck Jun.” A simple and fast algorithm for K-medoids clustering”Elsevier, 2008

[9] Mrs. S. Sujatha and Mrs. A. Shanthi Sona. ” New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method” IJERT, 2013
[10] Nick Larusso.” A Survey of Uncertain Data Algorithms and Applications”. IEEE Transaction On Knowledge And Data Engineering, 2009
[11] Patrick M, Ozsu, S.Sridhar, "Principles of Distributed Database Systems", Pearson Education, 2006.
[12] Dunham, S.Sridhar, Introduction to Datamining, Pearson Education, 2006