www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 11 Nov. 2016, Page No. 18853-18855

Implementation of clustering of uncertain data on probability distribution similarity

Nikhatparvin Ahamad[#1], Shahid Nadeem[*2],Shyam Dubey[#3],Nusrat Anjum[#4]

[1]M tech (CSE) Nuva college of Enginerring & technology,Nagpur Maharashtra,India
nikhat.ahamad@gmail.com

[2]Professor (CSE) Nuva College of Engineering & technology,Nagpur Maharashtra,India
Shahid4sam@gmail.com

[3] Professor (CSE) Nuva College of Engineering & technology,Nagpur Maharashtra,India
Shyam.nuva@rediffmail.com

[4] Professor (CSE) Anjuman College of Engineering & technology,Nagpur Maharashtra,India
Anjum.nusrat72@gmail.com

***Abstract*— Clustering on uncertain data, one of the essential tasks in mining uncertain data, posts significant challenges on both modeling similarity between uncertain objects and developing efficient computational methods. The previous methods extend traditional partitioning clustering methods like k-means and density-based clustering methods like DBSCAN to uncertain data, thus rely on geometric distances between objects. Such methods cannot handle uncertain objects that are geometrically indistinguishable, such as products with the same mean but very different variances in customer ratings. Surprisingly, probability distributions, which are essential characteristics of uncertain objects, have not been considered in measuring similarity between uncertain objects. In this project, we systematically model uncertain objects in both continuous and discrete domains, where an uncertain object is modeled as a continuous and discrete random variable, respectively. We use the well-known Kullback-Leibler divergence to measure similarity between uncertain objects in both the continuous and discrete cases, and integrate it into partitioning and density-based clustering methods to cluster uncertain objects.**

***Keywords*— Cluster,Uncertain data, probability density function**

I. INTRODUCTION

The previous studies on clustering uncertain data are largely various extensions of the traditional clustering algorithms designed for certain data. As an object in a certain data set is a single point, the distribution regarding the object itself is not considered in traditional clustering algorithms. Thus, the studies that extended traditional algorithms to cluster uncertain data are limited to using geometric distance-based similarity measures, and cannot capture the difference between uncertain objects with different distributions. Specifically, three principal categories exist in literature, namely partitioning clustering approaches, density-based clustering approaches, and possible world approaches. The first two are along the line of the categorization of clustering methods for certain data, the possible world approaches are specific for uncertain data following the popular possible world semantics for uncertain data. As these approaches only explore the geometric properties of data objects and focus on instances of uncertain objects, they do not consider the similarity between uncertain objects in terms of distributions.

II. LITRATURE SURVEY

Our mental representations of the world are formed by processing large numbers of sensory inputs including, for example, the pixel intensities of images, the power spectra of sounds, and the joint angles of articulated bodies. While complex stimuli of this form can be represented by points in a high-dimensional vector space, they typically have a much more compact description. Coherent structure in the world leads to strong correlations between inputs (such as between neighboring pixels in images), generating observations that lie on or close to a smooth low-dimensional manifold. To compare and classify such observations in effect, to reason about the world depends crucially on modeling the nonlinear geometry of these low-dimensional manifolds.

Scientists interested in exploratory analysis or visualization of multivariate data (*1*) face a similar problem in dimensionality reduction. The problem involves mapping high-dimensional inputs into a low dimensional "description" space with as many coordinates as observed modes of variability. Previous approaches to this problem, based on multidimensional scaling (MDS), have computed embeddings that attempt to preserve pairwise distances or generalized disparities between data points; these distances are measured along straight lines or, in more sophisticated usages of MDS such as Isomap, along shortest paths confined to the manifold of observed inputs. Here, we take a different approach, called locally linear embedding (LLE) that eliminates the need to estimate pair wise distances between widely separated data points.

Unlike previous methods, LLE recovers global nonlinear structure from locally linear fits. The LLE algorithm, is based on simple geometric intuitions. Suppose the data consist of *N* real-valued vectors *X*Wi, each of dimensionality *D*, sampled from some underlying manifold. Provided there is sufficient data (such that the manifold is well-sampled), we expect each data point and its neighbors to lie on or close to a locally linear

patch of the manifold. We characterize the local geometry of these patches by linear coefficients that reconstruct each data point from its neighbors.

The purpose of any clustering technique is to evolve a K partition matrix of a data set X in RN, representing its partitioning into a number, say K, of clusters (C1; C2; . . . ; CK). The partition matrix may be represented as U the membership of pattern xj to clusters Ck. Clustering techniques broadly fall into two classes, partitional and hierarchical. K-Means and single linkage are widely used techniques used in the domains of partitional and hierarchical clustering, respectively.

The two fundamental questions that need to be addressed in any typical clustering system are: How many clusters are actually present in the data and how real or good is the clustering itself. That is, whatever the clustering method may be, one has to determine the number of clusters and also the goodness or validity of the clusters formed. The measure of validity of the clusters should be such that it will be able to impose an ordering of the clusters in terms of its goodness. Milligan and Cooper have provided a comparison of several validity indices for data sets containing distinct non overlapping clusters while using only hierarchical clustering algorithms. Meil_a and Heckerman provide a comparison of some clustering methods and initialization strategies. Some more clustering algorithms may be found. In this paper, we aim to evaluate the performance of four validity indices, namely, the Davies-Bouldin index, Dunn's index, Calinski- Harabasz index, and a recently developed index I, in conjunction with three clustering algorithms viz. the well-known K-means and single linkage algorithms, as well as a recently developed simulated annealing (SA) based clustering scheme. The number of clusters is varied from Kmin to Kmax for K-means and the simulated annealing-based clustering algorithms, while, for single linkage algorithm (which incorporates automatic variation of number of clusters), the partitions in this range are considered.

III. PROPOSED SYSTEM

In this project, we consider uncertain objects as random variables with certain distributions. We consider both the discrete case and the continuous case. In the discrete case, the domain has a finite number of values, for example, the rating of a camera can only take a value. In the continuous case, the domain is a continuous range of values, for example, the temperatures recorded in a weather station are continuous real numbers. Directly computing KL divergence between probability distributions can be very costly or even infeasible if the distributions are complex. Although KL divergence is meaningful, a significant challenge of clustering using KL divergence is how to evaluate KL divergence efficiently on many uncertain objects. To the best of our knowledge, this project is the first to study clustering uncertain data objects using KL divergence in a general setting. We make several contributions. We develop a general framework of clustering uncertain objects considering the distribution as the first class citizen in both discrete and continuous cases. Uncertain objects can have any discrete or continuous distribution. 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 in both partitioning and density-based clustering methods.

IV. MODULES

There are the five modules Uncertain probability distribution,KL divergence similarity, partitioning clustering method,K-Medoids Mathod,Randomized K-Medoids method.

1)Uncertain Probability Distributions*:*

We consider an uncertain object as a random variable following a probability distribution in a domain ID. We consider both the discrete and continuous cases. If the domain is discrete with a finite or 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. In many case, the accurate probability distributions of uncertain objects are not known beforehand in practice. Instead, the probability distribution of an uncertain object is often derived from our observations of the corresponding random variable. Therefore, we associate each object with a sample of observations, and assume that the sample is finite and the observations are independent and identically distributed (IID). By overloading the notation, for an uncertain object P, we still use P to denote the corresponding random variable, the probability mass/density function, and the sample. For discrete domains, the probability mass function of an uncertain object can be directly estimated by normalizing the number of observations against the size of the sample.

2)Using KL Divergence Similairty *:*

It is natural to quantify the similarity between two uncertain objects by KL divergence. Given two uncertain objects P and Q and their corresponding probability distributions, evaluates the relative uncertainty of Q given the distribution of P, which is the expected likelihood ratio of the two distributions and tells how similar they are. The KL divergence is always nonnegative, and satisfies Gibbs' inequality. Therefore, the smaller the KL divergence, the more similar the two uncertain objects. In the discrete case, it is straightforward to calculate the KL divergence between two uncertain objects P and Q from their probability mass functions.

3). Partitioning Clustering Methods

k-means and k-medoids are two classical partitioning methods. The difference is that the k-means method represents each cluster by the mean of all objects in this cluster, while the k-medoids method uses an actual object in a cluster as its representative. In the context of uncertain data where objects are probability distributions, it is inefficient to compute the mean of probability density

functions. k-medoids method avoids computing the means. For efficiency, we adopt the k-medoids method to demonstrate the performance of partitioning clustering methods using KL divergence to cluster uncertain objects.

4) Uncertain K-Medoids Method

The uncertain k-medoids method consists of two phases, the building phase and the swapping phase. In the building phase, the uncertain k-medoids method obtains an initial clustering by selecting k representatives one after another. The first representative C1 is the one which has the smallest sum of the KL divergence to all other objects The rest k representatives are selected iteratively. In the ith iteration, the algorithm selects the representative Ci which decreases the total KL divergence as much as possible. For each object P which has not been selected, we test whether it should be selected in the current round. For any other non selected object will be assigned to the new representative P if the divergence is smaller than the divergence between P and any previously selected representatives. Therefore, we calculate the contribution to decrease of the total KL divergence. We calculate the total decrease of the total KL divergence by selecting P as the sum over the contribution of all non selected object. Then, the object to be selected in the ith iteration is the one that can incur the largest decrease.

5)Randomized K-Medoids Method

The randomized k-medoids method, instead of finding the optimal non representative object for swapping, randomly selects a non representative object for swapping if the clustering quality can be improved. We follow the annealing technique to prevent the method from being stuck at a local optimal result. The randomized k-medoids method follows the building swapping framework. At the beginning, the building phase is simplified by selecting the initial k representatives at random. Non selected objects are assigned to the most similar representative according to KL divergence. Then, in the swapping phase, we iteratively replace representatives by non representative objects. In each iteration, instead of finding the optimal non representative object for swapping in the uncertain k-medoids method, a non representative object P is randomly selected to replace the representative C to which P is assigned. After all non representative objects are examined, the total decrease of the total KL divergence by swapping P and C is recorded.

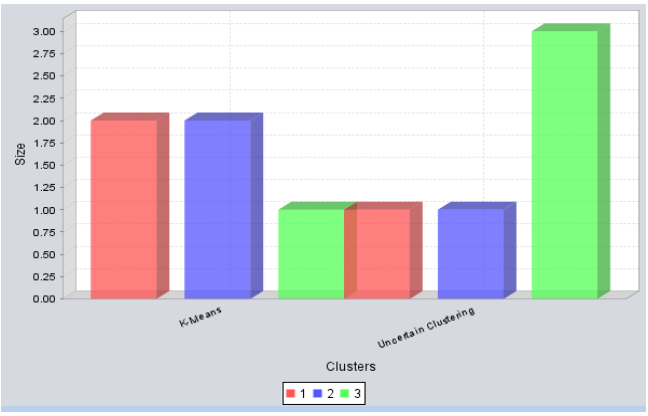

Chart -1: Comparison graph between different clustering algorithm

V. Conclusions

In this project, we explore clustering uncertain data based on the similarity between their distributions. We advocate using the Kullback-Leibler divergence as the similarity measurement, and systematically define the KL divergence between objects in both the continuous and discrete cases. We integrated KL divergence into the partitioning and density-based clustering methods to demonstrate the effectiveness of clustering using KL divergence. To tackle the computational challenge in the continuous case, we estimate KL divergence by kernel density estimation and employ the fast Gauss transform technique to further speed up the computation. The extensive experiments confirm that our methods are effective and efficient. The most important contribution of this project is to introduce distribution difference as the similarity measure for uncertain data. Besides clustering, similarity is also of fundamental significance to many other applications, such as nearest neighbor search. In the future, we will study those problems on uncertain data based on distribution similarity.

References

[1] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science, vol. 290, no. 5500, pp. 2323-2326, 2000.

[2] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1650-1654, Dec. 2002.

[3] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the Number of Clusters in a Dataset via the Gap Statistics," J. Royal Statistical Soc. B, vol. 63, pp. 411-423, 2001.

[4] P. Guo, C. Chen, and M. Lyu, "Cluster Number Selection for a Small Set of Samples Using the Bayesian Ying-Yang Model," IEEE Trans. Neural Networks, vol. 13, no. 3, pp. 757-763, 2002.

[5] R. Xu and D. Wunsch II, "Survey of Clustering Algorithms," IEEE Trans. Neural Networks, vol. 16, no. 3, pp. 645-678, 2005.

[6] J.C. Bezdek, R.J. Hathaway, and J. Huband, "Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices," IEEE Trans. Fuzzy Systems, vol. 15, no. 5, pp. 890-903, 2007.

[7] J.C. Bezdek and R. Hathaway, "VAT: A Tool for Visual Assessment of (Cluster) Tendency," Proc. Int'l Joint Conf. Neural Networks (IJCNN '02), pp. 2225-2230, 2002.

[8] J. Huband, J.C. Bezdek, and R. Hathaway, "bigVAT: Visual Assessment of Cluster Tendency for Large Data Sets," Pattern Recognition, vol. 38, no. 11, pp. 1875-1886, 2005.

[9] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," Proc. Advances in Neural Information Processing Systems (NIPS), 2002.

[10] D. Cai, X. He, and J. Han, "Spectral Regression for Efficient Regularized Subspace Learning," Proc. 11th Int'l Conf. Computer Vision (ICCV), 2007.

[11] N. Pal, J. Keller, M. Popescu, J. Bezdek, J. Mitchell, and J. Huband, "Gene Ontology-Based Knowledge Discovery through Fuzzy Cluster Analysis," J. Neural, Parallel and Scientific Computing, vol. 13, pp. 337-361, 2005.

[1] I. Sledge, J. Huband, and J.C. Bezdek, "(Automatic) Cluster Count Extraction from Unlabeled Datasets," Joint Proc. Fourth Int'l Conf. Natural Computation (ICNC) and Fifth Int'l Conf. Fuzzy Systems and Knowledge Discovery (FSKD), 2008.