ijecs open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 1 January 2015, Page No. 10148-10151

A Novel And Improved Technique For Clustering Uncertain Data

Vandana Dubey[1], Mrs A A Nikose[2]

Vandana Dubey

PBCOE, Nagpur,Maharashtra, India

vandana.dubey611@gmail.com

Mrs A A Nikose

PBCOE , Nagpur,Maharashtra, 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. 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. 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, KL divergence, Initial medoid.

1. Introduction

Data mining is the process of extracting or mining knowledge from large amount of data. Data mining tools and techniques helps to predict business trends those can occur in near future such as Clustering, Classification, Association rule, Decision trees. As an important research direction in the field of data mining, clustering has drawn more and more attention to researchers in the data mining. Clustering on uncertain data, one of the essential tasks in mining uncertain data, posts significant challenges on both modelling similarity between uncertain objects and developing efficient computational methods. It used to place data elements into related groups without advance knowledge of the group definitions.

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 behaviour, 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.

For example, in a shop the users are asked to evaluate a camera on the basis of various aspects such as quality, battery performance, image quality etc. Each camera may be scored by many users. Thus, the user satisfaction to a camera can be modelled as an uncertain object. There are often a good number of cameras under a user study. A frequent analysis task is to
cluster the cameras according to user satisfaction data.

2. PROBLEM DEFINITION

Clustering on uncertain data, one of the essential tasks in mining uncertain data,posts significant challenges on both modelling similarity between uncertain objects and developing efficient computational methods. The previous methods extend traditional partitioning clustering methods. The Kullback-Leibler divergence to measure similarity between uncertain objects and apply k-mediod & randomized k-mediods to cluster uncertain objects. The two output parameters delays and accuracy will be compared to get the best algorithm out of mediod and randomized mediods.

3. PROPOSED WORK

In dataset collection, we will study data sets from online resources such as twitter dataset. After that apply some natural language processing technique to find certain and uncertain data. Once the uncertain data is found apply the KL divergence and k mediods method to cluster this data into various categories. Once the data is cluster we will evaluate the output parameter delays and accuracy. Now apply randomized k-mediods method for clustering and evaluate its efficiency. The two outputs will be compared to get the best algorithm out of mediod and randomized mediods.

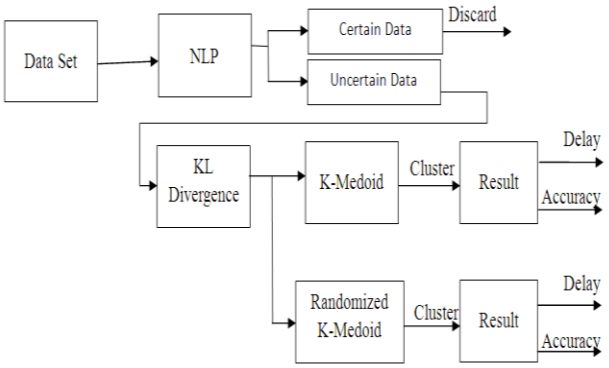

Fig: Block diagram for clustering uncertain data

MODULES

- Dataset
- NLP
- KL Divergence
- Clustering
- Comparing result

- **Dataset Collection:**

In dataset collection, we will use UCI dataset.UCI is machine learning repository.

- **Natural Language Processing(NLP):**

NLP has two steps:

i. **POS tagging:** Parts of speech tagging, in this the data is tagged into various parts of speech like noun, pronoun, verbs etc.

ii. **Chunking:** The POS data is chunk and unwanted tags are removed.

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

What is Parts-Of-Speech Tagging?

The process of assigning one of the parts of speech to the given word is called Parts Of Speech tagging. It is commonly referred to as POS tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories.

Example:

Word: Paper, Tag: Noun
Word: Go, Tag:Verb
Word: Famous, Tag:Adjective

Architecture of POS tagger

1. Tokenization: The given text is divided into tokens so that they can be used for further analysis. The tokens may be words, punctuation marks, and utterance boundaries.

2. Ambiguity look-up: This is to use lexicon and a guessor for unknown words. While lexicon provides list of word forms and their likely parts of speech, guessors analyze unknown tokens.

3. Ambiguity Resolution: This is also called disambiguation. Disambiguation is based on information about word such as the probability of the word.

Chunking

Chunking is also called shallow parsing and it's basically the identification of parts of speech and short phrases (like noun phrases). Part of speech tagging tells you whether words are nouns, verbs, adjectives, etc.

- **Uncertain Data**

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. For discrete domains, the probability mass function of an uncertain data can be directly estimated by normalizing the number of observations against the size of the sample. The pmf of data P is expressed in eq.1

$$P(X) = \Sigma\, Px\,(x)\,dx \qquad (1)$$

For continuous domains, the probability density function of an uncertain data can be calculated by using the following eq.2

$$P(X) = \int Px(X)\,dx \qquad (2)$$

- **KL Divergence**

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. 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. In the discrete case, let f and g are two probability mass functions in a discrete domain with a finite or countably infinite number of values. The Kullback-Leibler diverge between f and g is defined in eq.2

$$D(f||g) = \Sigma f(x) \log \frac{f(x)}{g(x)} \quad (3)$$

In the continuous case, let f and g be two probability density functions in a continuous domain with a continuous range of values. The Kullback-Leibler divergence between f and g is defined in eq.4

$$D(f||g) = \int f(x) \log \frac{f(x)\,dx}{g(x)} \quad (4)$$

- **Clustering Algorithm**

Applying KL divergence into K-mediod algorithm

K-mediod is a classical partitioning method to cluster the data. 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 as shown in fig.

We apply some clustering methods using KL divergence to cluster uncertain objects in two categories. First, the uncertain k-mediods method which extends a popular partitioning clustering method k-mediods by using KL divergence. Then, we develop a randomized k-medoids method based on the uncertain k-mediods method to reduce the time complexity.

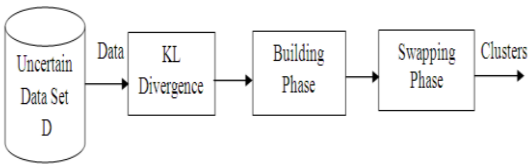

Fig: Uncertain data clustering process

Building phase: In the building phase, the k-medoid method obtains an initial clustering by selecting initial medoids randomly.

Swapping Phase: In the swapping phase the uncertain *k*-medoid method iteratively improves the clustering by swapping a no representative data with the representative to which it is assigned.

I. Partitioning Clustering Methods

Using KL divergence as similarity, a partitioning clustering method tries to partition objects into k clusters and chooses the best k representatives, one for each cluster, to minimize the total KL divergence. K-medoids is one of the classical partitioning methods. We first apply the uncertain k-medoids method which integrates KL divergence into the original k-medoids method and we develop a randomized k-medoids method in to reduce the complexity of the uncertain one.

Randomized Clustering 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.

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 no representative objects.

- **Comparing Result**

The two outputs will be compared to get best algorithm out of k-medoid clustering and randomized medoid clustering using delay and accuracy values and graphs.

Conclusion

The field of uncertain data management has seen a revival in recent years because of new ways of collecting data which have resulted in the need for uncertain representations. We presented the important data mining and management techniques in this field along with the key representational issues in uncertain data management. Nearest neighbor search on uncertain data based on distribution similarity has been evaluated.

We explore clustering uncertain data based on the similarity between their distributions. We advocate using the Kullback-Leibler divergence as the similarity measurement. Apply some clustering methods using KL divergence to cluster uncertain objects in two categories. First, the uncertain k-medoids method which extends a popular partitioning clustering method k-medoids by using KL divergence. Then, we develop a randomized k-medoids method based on the uncertain k-medoids method to reduce the time complexity.The two outputs will be compared to get best algorithm out of k-medoid clustering and randomized medoid clustering using delay and accuracy values and graphs.

References

[1] Pei,jein,tao, " Clustering Uncertain Data Based on Probability Distribution Similarity",IEEE Transactions on knowledge and data Engineering, Volume: 25, issue_4, Publication Year: 2013, Page(s): 721 –733.

[2] 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

[3] 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.

[4] Dr. T. Velmurugan "Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points ". IJCTA,2012
[5] Samir Anjani and Prof. Mangesh Wangjari. "Clustering of uncertain data object using improved K-Means algorithm" IJARCSSE, 2013
[6] Mrs. S. Sujatha and Mrs. A. Shanthi Sona. " New Fast KMeans Clustering Algorithm using Modified Centroid Selection Method" IJERT, 2013