ISSN 2348 – 8034
Impact Factor- 3.155

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

VARIOUS METHODS FOR CLUSTERING UNCERTAIN DATA

Vandana Dubey[*1] and Mrs A A Nikose[2]
[*1]PBCOE, Nagpur,Maharashtra, India
[2]Assistant Professor, PBCOE, Nagpur, Maharashtra, India

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. 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.

Keywords- *Clustering, Clustering uncertain data, Mining methods and algorithms, partition clustering .*

I. 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.

II. PROPOSED WORK

In dataset collection, we will use weather data sets. 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.

MODULES

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

ISSN 2348 – 8034
Impact Factor- 3.155

- **Dataset Collection:**
  In dataset collection, we are use weather dataset.

- **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

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.

- **KL Divergence**

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.

The measure of the difference between two inputs containing K and L values respectively is called as KL divergence.

The divergence is an inverse factor to similarity and is calculated using the following technique.

STEPS
Step1:Loop through each entry of the current sentence.
Step2:Set a divergence value to the length of the entries.
Step3:For each entry if the K^{th} entry is matching with the L^{th} entry then reduce the divergence value and loop through the entied lines of database.
Step4:The final obtained value is the value of KL divergence.

- **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.

ISSN 2348 – 8034
Impact Factor- 3.155

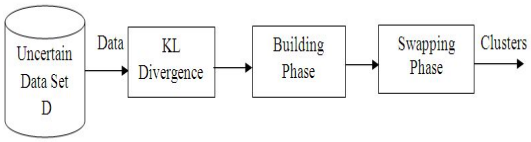


Fig: Uncertain data clustering process

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.

III. 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

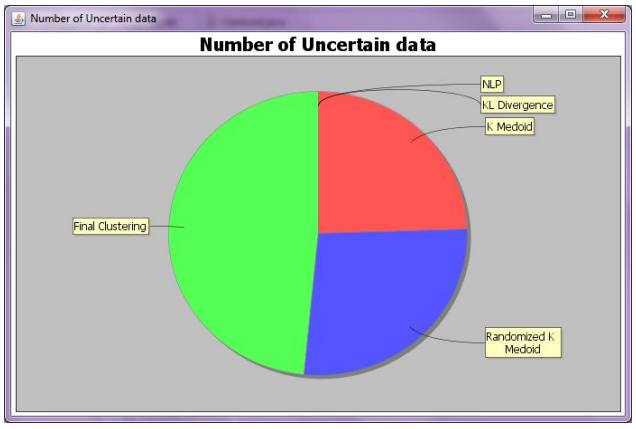
The randomized k-medoids method follows the building-swapping framework. At the beginning, the building phase is simplified by selecting the initial k representatives at random. Non-selected objects are assigned to the most similar representative according to KL divergence. Then, in the swapping phase, we iteratively replace representatives by no representative objects.

Final Clustering

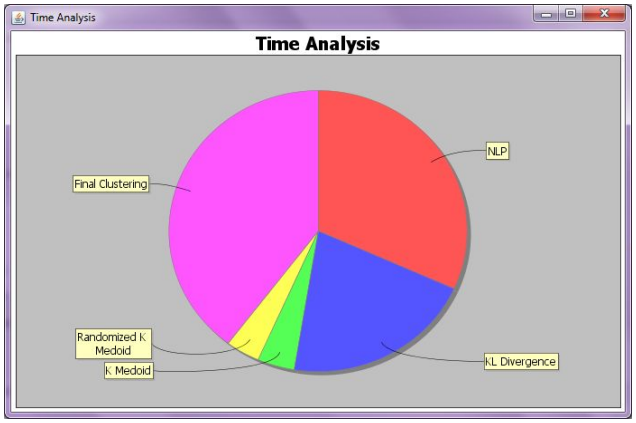
Final clustering follows the k-mediod method. The output of randomized clustering is taken as input for the final clustering and we found majority of uncertain data with very few certain data.

IV. RESULT

The time required for NLP,KL Divergence ,K mediod,randomized k mediod and final clustering are shown in the pie chart below.
The possibility of uncertain data found during various method is shown in the piechart.



ISSN 2348 – 8034
Impact Factor- 3.155



V. 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 and show the results using some graphs.

REFERENCES

1. *C.C. Aggarwal and P.S. Yu, "A Framework for Clustering Uncertain Data Streams," Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE), 2008.*
2. *M.R. Ackermann, J. Blo¨mer, and C. Sohler, "Clustering for Metric and Non-Metric Distance Measures," Proc. Ann. ACM-SIAM Symp.Discrete Algorithms (SODA), 2008.*
3. *E. Achtert, H.-P. Kriegel, E. Schubert, and A. Zimek.Interactive data mining with 3D-Parallel-Coordinate-Trees. In Proc. SIGMOD,pages 1009{1012, 2013.*
4. *F. Gullo, G. Ponti, and A. Tagarelli. Clustering uncertain data via k-medoids. In Scalable Uncertainty Management, pages 229{242, 2008.*
5. *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.*
6. *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.*