Technique For Clustering Uncertain Data Based On Probability Distribution Similarity

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

Vandana Dubey

PBCOE, Nagpur,Maharashtra, India

Mrs A A Nikose

Assistant Professor

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. 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.***

Keywords: Uncertain data clustering, Probability distribution, KL divergence, Initial medoid.

1. Introduction

In generally Data Mining deals with the difficulty of extracting patterns from the information by paying suspicious attention to computing, communication and human-computer interface issues. Clustering is one of the major data mining tasks to group the similar information or data. All clustering algorithms aim of dividing the collection all data objects into subsets or similar clusters. A cluster is a collection of objects which are 'similar' between them and are 'dissimilar' to the objects belonging to other clusters and a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set. In data mining Clustering certain data have been well studied in the various areas such as data mining, machine learning, Bioinformatics, and pattern recognition.

However, there is only preliminary research on clustering uncertain data. In this study clustering uncertain data object problem have been solved with probability distribution function.

Clustering uncertain data

In many applications, data contain intrinsic uncertainty. Numerals of factors contribute the uncertainty such as the random nature of the physical data creation and collection procedure, measurement of error, and data staling. One purpose of the clustering is the selection of a device as the leader for each cluster. A leader's role is to collect data (such as location data) from its cluster members and to communicate with a server or a base station with batched updates. In this way, most communications are short-ranged messages among the cluster members and their leaders.

The previous studies on clustering uncertain data are largely various extensions of the traditional clustering algorithms considered for certain data. Here the object in certain dataset is considered as a single point and distribution concerning the object itself is not considered in traditional clustering algorithms. The study that extends conventional algorithms to cluster uncertain data that are restricted to using geometric distance-based similarity measures and cannot capture the dissimilarity between uncertain objects with diverse distributions.

ISSN: 2278 – 7798

The main areas of research are:
Modeling of uncertain data: A key issue is the process of modeling the uncertain data. Hence, the fundamental complexities have been captured while keeping the data helpful for database management applications.

Uncertain data mining: The results of data mining applications are affected by the underlying uncertainty in the data or objects. Therefore, it is difficult to design data mining techniques that can take such uncertainty into account during the computations.

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.

3. PROPOSED WORK

In dataset collection, we will study weather data set. 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.

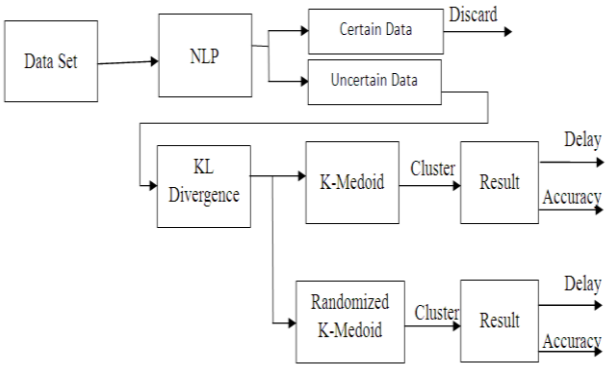


Fig: Block diagram for clustering uncertain data

MODULES

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

- **Dataset Collection:**

In dataset collection, we will use weather dataset. Weather dataset contains temperature dataset, humidity dataset and sound datasets. Datasets contains records up to five years. Temperature dataset has attributes like year, month and day. Humidity datasets contains attributes year, month and day.

- **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.

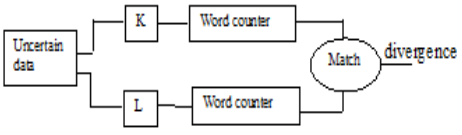
Example:
Word: Paper, Tag: Noun
Word: Go, Tag:Verb
Word: Famous, Tag:Adjective
Chunking
Chunking is also called shallow parsing and it's basically the identification of parts of speech and short phrases (like noun phrases). Part of speech tagging tells you whether words are nouns, verbs, adjectives, 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.

Fig: Finding divergence between lines of data



ISSN: 2278 – 7798

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.

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.

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.

Uncertain K-mediod Method

The K-medoid method consists of two phases, the building phase and the swapping phase.

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.

Algorithm *K*-medoid clustering based on *KL* divergence method

Input:
D = {d_1, d_2...,d_n} set of *n* data items.
K, Number of desired clusters
Output:
A set of *K* clusters.
Steps:
Phase 1: Determine the initial medoids of the clusters by using Algorithm 1.
Phase 2: Assign each data point to the appropriate clusters by using Algorithm 2.

Algorithm 1

Input:
D = {d_1, d_2,.....,d_n} set of n data items
K = number of desired cluster
Output:
A set of *K* initial medoids
Steps:
1. Set p = 1
2. Compute the probability distribution similarity between each data and all other data in the set D
3. Find the most similar pair of data from the set D and form a data set Am which contains these two data , Delete these two data from the set D
4. Find the data in D that is similar to the data set A_m, Add it to Am and delete it from D
5. Repeat step 4 until the number of data in A_m reaches 0.75*(n/k)
6. If p<k, then p = p+1, find another pair of data from D between which the highest similarity, form another data set Am and delete them from D, Go to step 4
7. For each data set Am find the arithmetic mean of the vectors of data in Am, these means will be the initial medoids.

Algorithm 1 describes the method for finding initial mediod of the clusters effectively. Initially, compute the probability distribution similarity between each data and all other data in the set of data. Then find out the most similar pair of data and form a set A_1 consisting of these two data, and delete them from the data set D. Then determine the data which is similar to the set A_1, add it to A_1 and delete it from D. Repeat this procedure until the number of elements in the set A1 reaches a threshold. At that point go back to the second step and form another data set A_2. Repeat this till '*K*' such sets of data are obtained. Finally the initial medoids are obtained by averaging all the data in each data set. The *KL* divergence method is used for determining the probability distribution similarity between each data.

These initial medoids are given as input to the second phase ,for assigning data to appropriate clusters. The steps involved in this phase are outlined as Algorithm 2.

Algorithm 2

Input:
D = {d_1, d_2,…...,d_n} set of n data items
A set of k initial medoids
Output:
A set of k clusters
Steps:
1. Associate each data point to the most similar medoid. ("Similar" here is defined using *KL* divergence)
2. For each medoid m for each non medoid data o swap m and o and compute the total cost of the swapping
3. Select the swapping with the lowest cost
4. Repeat steps 2 to 5 until the clusters are not changed

ISSN: 2278 – 7798

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.

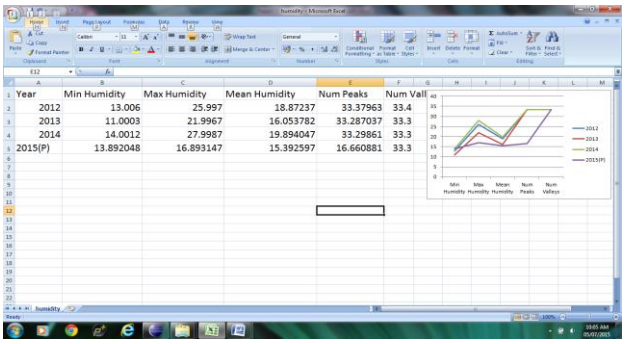
In each iteration, instead of finding the optimal non-representative object for swapping in the uncertain k-medoids method, a non-representative object P is randomly selected to replace the representative C to which P is assigned.

4. ANALYSIS

The outputs will be compared and show using graph.

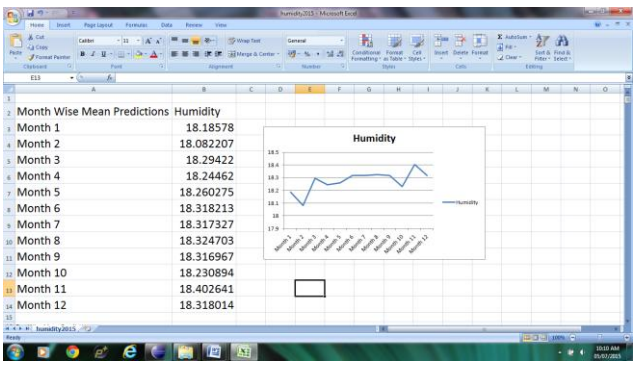
EXPERIMENTAL RESULTS

Fig 1: Prediction of the humidity of 2015 year by analysing the records of the previous three years i.e. 2012, 2013 and 2014.



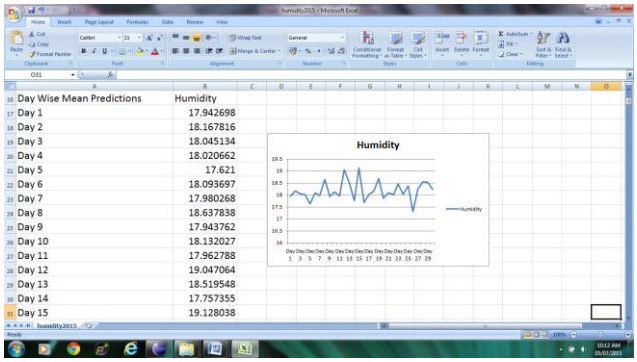
Year	Min Humidity	Max Humidity	Mean Humidity	Num Peaks	Num Valleys
2012	13.006	25.997	18.87237	33.37963	33.4
2013	11.0003	21.9967	16.053782	33.287037	33.3
2014	14.0012	27.9987	19.894047	33.29861	33.3
2015(P)	13.892048	16.893147	15.392597	16.660881	33.3

Fig 2:Month wise prediction of humidity of 2015 year.



Month Wise Mean Predictions	Humidity
Month 1	18.18578
Month 2	18.082207
Month 3	18.29422
Month 4	18.24462
Month 5	18.260275
Month 6	18.318213
Month 7	18.317327
Month 8	18.324703
Month 9	18.316967
Month 10	18.230894
Month 11	18.402641
Month 12	18.318014

Fig 3:Day wise prediction of humidity of 2015 year record.



Day Wise Mean Predictions	Humidity
Day 1	17.942698
Day 2	18.167816
Day 3	18.045134
Day 4	18.020662
Day 5	17.621
Day 6	18.093697
Day 7	17.980268
Day 8	18.637838
Day 9	17.943762
Day 10	18.132027
Day 11	17.962788
Day 12	19.047064
Day 13	18.519548
Day 14	17.757355
Day 15	19.128038

Fig 4: Hour wise prediction of temperature of 2015 year record.

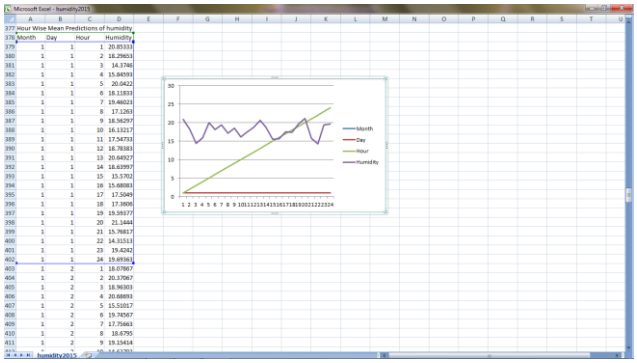
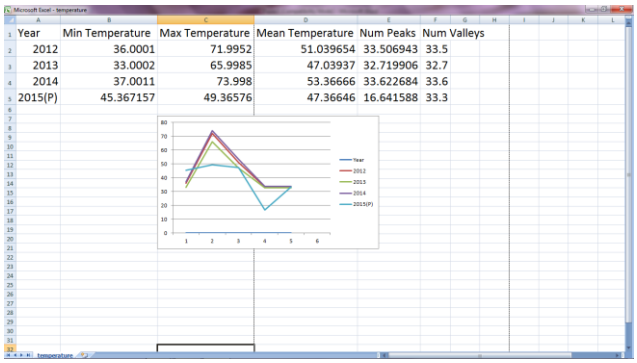
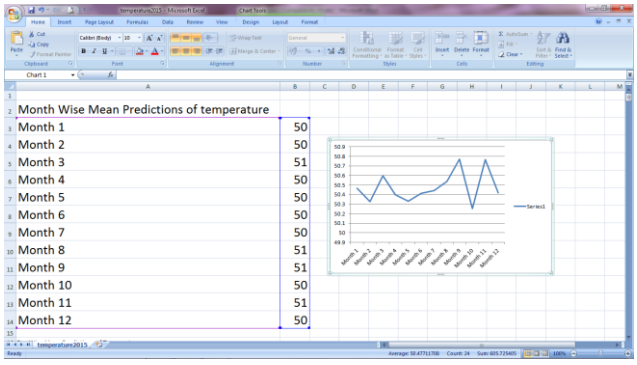


Fig 5: Prediction of the temperature of 2015 year by analysing the records of the previous three years i.e. 2012, 2013 and 2014.



Year	Min Temperature	Max Temperature	Mean Temperature	Num Peaks	Num Valleys
2012	36.0001	71.9952	51.039654	33.506943	33.5
2013	33.0002	65.9985	47.03937	32.719906	32.7
2014	37.0011	73.998	53.36666	33.622684	33.6
2015(P)	45.367157	49.36576	47.36646	16.641588	33.3

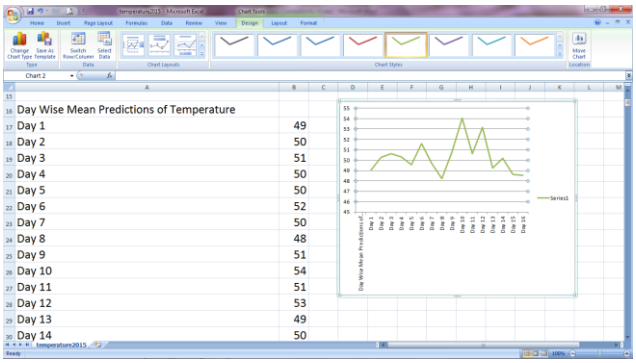
Fig 6: Month wise prediction of temperature of 2015 year.



Month Wise Mean Predictions of temperature	
Month 1	50
Month 2	50
Month 3	51
Month 4	50
Month 5	50
Month 6	50
Month 7	50
Month 8	51
Month 9	51
Month 10	50
Month 11	51
Month 12	50

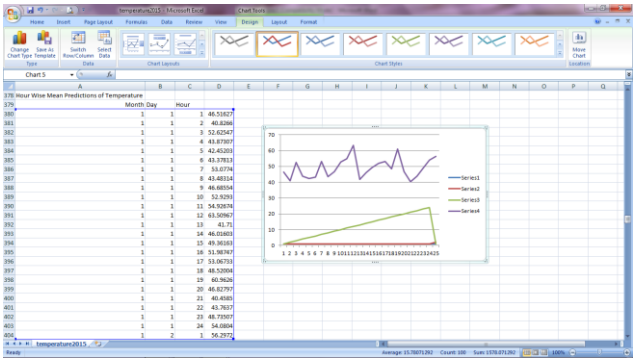
ISSN: 2278 – 7798

Fig 7: Day wise prediction of temperature of 2015 year record.



Day Wise Mean Predictions of Temperature	
Day 1	49
Day 2	50
Day 3	51
Day 4	50
Day 5	50
Day 6	52
Day 7	50
Day 8	48
Day 9	51
Day 10	54
Day 11	51
Day 12	53
Day 13	49
Day 14	50

Fig 8: Hour wise prediction of temperature of 2015 year record.



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 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 and after that final clustering is done.

As we are using weather datasets in our project various tables and related graphs according to the information is display. In weather datasets we have humidity dataset and temperature dataset containing records of last 3 years data. By analysing records of last 3 years record we can predicate the coming year i.e. 2015 year record. We are display records year wise, month wise, day wise and hour wise also.

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

[4] Dr. T. Velmurugan "Efficiency of K-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points ". IJCTA,2012

[5] Nick Larusso." A Survey of Uncertain Data Algorithms and Applications". IEEE Transaction On Knowledge And Data Engineering, 2009

[6] 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

[7] Hae-Sang Park and Chi-Hyuck Jun." A simple and fast algorithm for K-medoids clustering"Elsevier, 2008

[8] http://home.dei.polimi.it /matteucc / Clustering /tutorial_html/.,2008. M. Matteucci. "A Tutorial on Clustering Algorithms",

[9] "An Efficient Distance Calculation Method for Uncertain Objects", Lurong Xiao, Edward Hung Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

[10] F. Gullo, G. Ponti, and A. Tagarelli. Clustering uncertain data via k-medoids. In Scalable Uncertainty Management, pages 229{242, 2008.

ISSN: 2278 – 7798