Modelling Uncertain Spatial Data Sets using Uncertain Partitioning Clustering

[1]Ramachandra Rao Kurada
Assistant Professor, Department of Computer Applications
Shri Vishnu Engineering College for Women, Bhimavaram
ramachandrarao.kurada@gmail.com
[2]Aruna Kumari M
Department of Computer Applications
Shri Vishnu Engineering College for Women, Bhimavaram
aruna.marii@gmail.com

Abstract- Uncertainty in spatial data set is due to various causes like imprecision, inconsistency and inaccuracy in the information acquired through data collection and amalgamation from the instruments and infrastructures on Geo-Informatics. Such uncertain data is usually represented in terms of uncertain regions over which the Probability Density Function (PDF) is defined. In the present paper, the problem of clustering uncertain data has been addressed by proposing UK-Medoids algorithm. The proposed algorithm employs a similarity function DIE deviation to find the distance between uncertain objects. Experiments have shown that UK-Medoids outperforms conventional algorithm from an accuracy view point while achieving reasonably good and efficient results.

Keywords: Information Entropy, UK-Medoids clustering, Uncertain Spatial Data.

1. Introduction

Handling uncertainly [9] in data management is of paramount importance in a wide range of application contexts. Indeed, data uncertainty [1], [6] naturally arises from implicit randomness in a process of data generation/ acquisition, imprecision in physical measurements, and loss of data freshness.

The uncertainty-based spatial data mining [5] is to extract knowledge from the vast repositories of practical spatial data under the umbrella of uncertainties with the given perspectives and parameters. With different granularities, scales, mining-angles, and uncertain parameters, it discovers the collective attribute distribution of spatial entities by perceiving various variations of spatial data and their combinations in the data space.

Various notations of uncertainty have been defined depending on the application domain. Attribute – level uncertainty [4], [1] has been considered in this paper, to model according to a probability model. The uncertain object is usually represented by means of probability density functions PDFs, which describe the likelihood that the object appears at each position in a multidimensional space rather than by a traditional vector form of deterministic values.

Clustering is useful in exploratory data analysis. Cluster analysis organizes data by grouping individuals into a population in order to discover structure or clusters. Many partitioned algorithms [8] have been proposed, some based on k-centroid, some based on K-Medoid, some based on fuzzy analysis, etc. K-Medoid clustering is similar to k-Centroid clustering. Both of them attempt to partition the data by assigning each object to a representative and then optimizing a statistical homogeneity criterion - namely, the total expected squared dissimilarity. However, K-Medoid clustering only allows objects to be chosen as representatives. In comparison with K-Centroid, the use of Medoids for clustering has several advantages. Firstly, this method has been shown to be robust to the existence of noise or outliers and generally produces clusters of high quality. Secondly, this method can be used not only on points or vectors for which the mean is defined but also on any objects for which a similarity measure between two objects is given.

This paper aims to apply uncertain K-medoids partitioning clustering to model uncertain spatial data and evaluate the pattern after clustering is applied. The remained sections will be organized as follows. Section 3 presents the Literature survey. The accessible method is presented in Section 4, Section 5 gives the experimental results. Conclusion is drawn finally in Section 6.

2. Literature Survey

Uncertain data [6] is ubiquitous in real world applications due to various causes. In recent years, clustering uncertain data has been paid more attention by the research community and the classical clustering algorithm based on partition, density and hierarchy have been extended to handle the uncertain data.

Clustering is known as very useful tool in many fields for data mining. The structure of data sets is found through the clustering methods [10]. Now, the more the ability of computers increase, the more works for uncertainties have been studied. In the past, data handled by the computers was approximately represented as one point or value because of poor ability of the computers. However, the ability is now enough to handle uncertain data.

Therefore, a lot of researchers have tried to handle original data from the viewpoint that the data should been represented as not one point approximately but some range exactly in a data space. Uncertainty may have an influence on the confidential level, supportable level, and interesting level of spatial data mining [11].

K-Medoids [8] is a clustering algorithm that is very much like K-means. The main difference between the two algorithms is the cluster center being used. K-means uses the average of all instances in a cluster, while K-Medoids uses the instance that is the closest to the mean, i.e. the most 'central' point of the cluster. Using an actual point of the data set to cluster makes the K-Medoids algorithm more robust to outliers than the K-means algorithm.

3. Proposed Work

a. Uncertain Partitioning based Clustering Approaches: In UK-Medoids, as every object has the same cluster, the expected distance based approach in UK-Means [12], [13] cannot distinguish the two sets of objects having different distributions. In UK-Mediods clustering of uncertain objects is made according to the similarity between their probability distribution. In information theory, the similarity between two distributions can be measured by the Information Entropy deviation (IE). This IE is used to measure the similarity between distributions, and demonstrate the effectiveness of similarity. The PDFs [12], [13] is used over the entire data domain and the difference is captured using the IE deviation.

b. Similarity using IE Deviation: In general, IE between the two probability density functions is defined as follows: In the discrete case, let A and B be two probability distribution functions in a discrete domain D with a finite number of values. The IE deviation between A and B is $IE(A||B) = \sum_{x\in D} F(x) log \frac{A(x)}{B(x)}$. In the continuous case, let A and B be two probability density functions in a continuous domain D with a continuous range of values. The IE deviation between A and B is $IE(A||B) = \sum_D F(x) log \frac{A(x)}{B(x)} dx.$ In both discrete and continuous cases, IE deviation is defined only in the case where for any x in domain D if A(x) > 0 then B(x) > 0, by convention, $0 \log \frac{0}{p} = 0$, for any p ≠ 0 and the base of log is 2.

c. Partitioning Clustering Methods on Uncertain data: A partitioning clustering method organizes a set of k uncertain objects O into n clusters $C_1, \dots, C_k \; such \; that \; C_i \subseteq O \; (1 \leq i \leq k), C_i \neq \phi, \bigcup_{i=1}^{k} C_i = 0, and \; C_i \cap C_j = \phi \; for \; any \; i \neq j.$ Using IE deviation as similarity, a partitioning clustering method tries to partition objects into k clusters and chooses the best n representatives, one for each cluster. To minimize the total IE deviation the computation is $DIE = \sum_{i=1}^{n} \sum_{p\in C_i} IE\,(x||\,c_i)$. For an object x in cluster $C_i (1 \leq i \leq k)$, the IE deviation $IE(x||c_i)$ between p and the representative c_i measures the extra information required to construct p given c_i. Therefore, $\sum_{p\subset C_i} IE(x||c_i)$ captures the total extra information required to construct the whole cluster C_i using its representative c_i. Summing over all n clusters, the total IE deviation thus measures the quality of the partitioning clustering. The smaller the value of TIE, the better the clustering is performed.

The pseudo code of core uncertain K-Medoids is presented initially with IE deviations as a distance function to calculate the distances between any two uncertain objects. Further refinement is made by incorporate Min-Max bounding box pruning technique into UK-Mediods and final sophistication to the algorithm is made by integrating R*-Tree indexing in to UK-Mediods.

d. 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 representatives C_1 is the one which has the smallest sum of the IE deviation to all other objects in O. That is, $C_1 = argmin_x(\sum_{x'\in O\{x\}} IE(x'||x)$.

The rest of the n-1representatives are selected iteratively. In the i^{th} ($2 \leq i \leq n$) iteration, the algorithm selects the representative C_i which decreases the total IE deviation as much as possible. For each object x which has not been selected, a test is made to select the current round. For any other non-selected object x', x' will be assigned to the new representative x if the divergence $IE(x'||x)$ is smaller than the divergence between x' and any previously selected representatives. Therefore, a calculation is made for the contribution of p' to the decrease of the total IE deviation by selecting x as

$$max\left(0, min_{j=1}^{i-1}\left(IE\left(x'||C_j\right)\right) - IE\left(x'||x\right)\right).$$

The total decrease of the IE deviation is calculated by selecting x as the sum over the contribution of the non-selected objects, denoted by DIE(x). Then, objects to be selected in the i^{th} iteration is the one that can incur the largest decrease that $C_i = argmax_{x\in O\{c_1,\dots,c_{i-1}\}}(DIE(x))$. In the swapping phase, the uncertain K-Medoids method iteratively improves the clustering by swapping a non-representative object with the representative to which it is assigned. For a non-representative

object x, suppose it is assigned to cluster C whose representative is c. The effect of swapping x and c in two cases for all non-selected object x' other than x is considered. If x' currently belongs to c, when c is replaced by x, a reassignment of x' to x or one of the other n − 1 existing representatives, to which x' is the most similar is considered. If x' currently belongs to a representative c' other than c, and $IE(x'||x) < IE(x'||c')$, x' is reassigned to x.

When a reassignment happens, the decrease in the total KL deviation by swapping x and c is recorded. After all non-representative objects are examined; the object is selected by swapping x and c. Then, object is selected by object x_{max} which can make the largest decrease. That is, $x_{max} = argmax_{p\in O\{c_1,\dots,c_k\}}(DIE(x))$. A check is made on swapping x_{max} to improve the clusters, i.e.,$DIE(x_{max}) > 0$. If so, the swapping is carried into execution. Otherwise, the method terminates and reports the final clustering. The following algorithm presents the pseudo code of the uncertain K-Medoids method.

Algorithm 1: Uncertain K-Medoids Algorithm

Input: a set $O(o_1, \dots, o_n)$ of uncertain objects, the number of clusters n; Output: n clusters$c_1, \dots, c_n$;

1: $C_1 = argmin_x(\sum_{x'\in O\{x\}} IE(x'||x)$

2: i=2;

3: while $i \le k$ do

4: for $x \in O\backslash\{c_1, \dots, c_{i-1}\}\ do$

5:$IE(x) = \sum_{x'\in O\{x,c_1,\dots,c_{i-1}\}} \quad max \left(0, min_{j=1}^{i-1}\left(IE(x'||C_j)\right) - IE(x'||x)\right)$

6: $c_i = argmax_{p\in O\{c_1,\dots,c_{i-1}\}}(DIE(x))$

7: $i = i + 1$

8: $c_i = \phi(1 \le i \le k)$

9: repeat

10: for $x \in O\ do$

11: $j = argmin_{i\in[1,k]}(IE(x||c_i))$.

12: $x.IE = IE(x||c_j), c_j = c_j \cup \{x\}$

13: for $x \in O\backslash\{c_1, \dots, c_{nk}\}$ do

14: assume $x \in C; DIE(x) = 0;$

$for\ x' \in C\ do$

15: $j = argmin_{i\in[i,k]}(IE(x||c_i))$

16:$DIE(x) = DIE(x) + x'.IE - IE(x'||c_j)$

17: for $x' \in O\backslash C\backslash\{c_1, \dots, c_{nk}\}$ do

18: if $IE(x'||x) < x'.IE$ then

19:$DIE(x) = DIE(x) + x'.IE - IE(x'||x)$

20: $x_{max} = argmax_{x\in O\{c_1,\dots,c_n\}}(DIE(x))$

21: if $DIE(x_{max}) > 0$ then

22: replace the center to which x_{max} is assigned in the last iteration by x_{max}

23: until $DIE(x_{max}) \le 0$

24: return $c_1, \dots, c_n$

4. Experimental Results

The algorithm described in the previous section have been implemented on java using JDK 1.6.0 and a series of experiments were performed on a PC with Intel(R) corei3, 2.93 GHz and 4GB of main memory, running on windows 7 operating system. This section focuses on the algorithm runtime performance and cluster generation in generation of a pattern.

The spatial dataset Forest Cover Type is used for all the experiments to evaluate the performance of data uncertainty. Forest Cover Type is a benchmark quandary extracted from http://kdd.ics.uci.edu /databases/covertype/covertype.html UCI KDD Archive. This problem relates to the actual forest cover type for given observation that was determined from US Forest Service (USFS) Region to Resource Information System (RIS). Forest Cover Type includes 581,012 samples represented in point valued data with 7 cover type; each sample has 54 attributes including 10 remotely sensed data and 44 cartographic data. Data preprocessing is applied to this data set and is transformed into many uncertain data sets by replacing each data point with an MBR and also generates the PDF. The proposed algorithm in the previous section is experimented with only ten percent of the available data object due to main memory constriction.

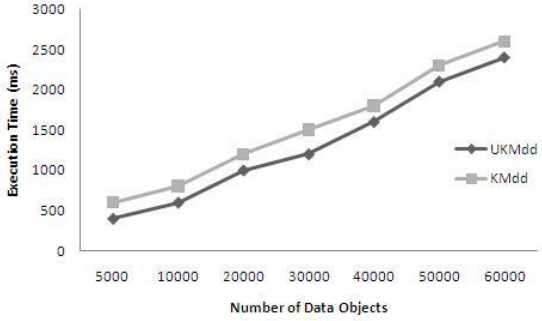

Fig. 1 Execution Time Effectiveness

a. Execution Time: One of an obligatory constraint for any algorithm was to assess its execution time recital when applied over its implanted datasets or databases. To evaluate the algorithms competence proposed in this paper, Fig. 1 is symbolized with the number of uncertain spatial data objects on horizontal axis and execution time measured in milliseconds on vertical axis. It is evident from Fig. 1 that as the number of data objects are increased the execution time also increases. It is one of substantiation that UKMdd is praiseworthy to model the spatial data objects.

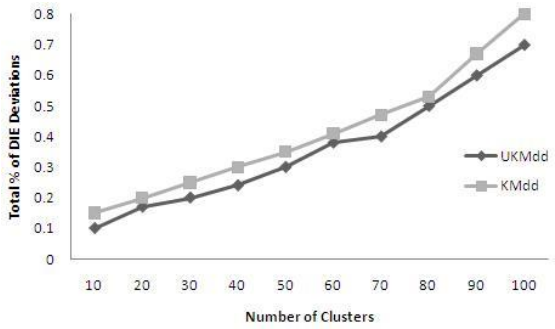

Fig. 2 Cluster Generation on Uncertain Spatial Data Objects

b. Cluster Generation: It is observed from Fig. 2 that the cluster arrangement in UKMdd seen to be consistently growing with the whole range of total percentage of DIE deviations. This is because with a large number of clusters, cluster representatives are generally less spread out. Hence total DIE deviations will have to be computed to determine the cluster assignment. The total percentage of DIE deviations in the algorithms increases when there is more number of clusters.

This substantiation is evidently revealed in Fig. 2 with number of clusters on horizontal axis and percentage of total DIE deviations in vertical axis. It is also verified in the anticipated algorithms as the number of cluster patterns increases the total percentage of DIE deviations.

Justification for Results: All the selected datasets originally contain deterministic values; hence uncertainty was synthetically generated for each object in the dataset. It is prominent that anticipated algorithms performed similarly for all the PDFs computations, since they employ similar clustering scheme. The new techniques bring down the execution of UKMdd and also reduce the computation time of PDF. A consistent demise pattern is observed in DIE deviations which substantiate that the integration being practical on uncertain spatial data and affirms the completeness in assortment of the anticipated algorithms.

Conclusion

This paper proposes a realistic way to model uncertainties in spatial data under the umbrella of uncertain spatial data mining with given perspective and parameter. The novel approach proposed in this paper accommodates tuples having numerical attributes with uncertainty described by arbitrary PDF. Performance is an issue to provoke privileged number of cluster progression even if more complicated computations of Information entropy deviations are involved.

Uncertain K-Medoids algorithm is primarily designed to handle tremendous uncertain spatial data. This algorithm is experientially verified to be highly effective and empirically good in lessening the execution time. This anticipated algorithm is pragmatic in exploiting spatial data uncertainties with remarkable higher accuracies. The execution times are of an order of magnitude comparable to classical algorithms.

Future work may address the issues involved in modeling uncertain data with soft computing based partitioned clustering.

References

1. B. Kao, S.D. Lee, D.W. Cheung, W.-S. Ho, and K.F. Chan, "Clustering Uncertain Data Using Voronoi Diagrams," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 333-342, Dec. 2008.
2. Beckmann, N., Kriegel, H.P., Schneider, R., Seeger, B.: The R*-tree: An efficient and robust access method for points and rectangles. In: SIGMOD. (1990) 322–331.
3. C. Bohm, P. Kunath, A. Pryakhin, and M. Schubert, "Querying Objects Modeled by Arbitrary Probability Distributions," Proc. 10th Int'l Symp. Spatial and Temporal Databases (SSTD), 2007.
4. C.C. Aggarwal and P.S. Yu, "A Framework for Clustering Uncertain Data Streams," Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE), 2008.
5. Ester, M., Kriegel, H.P., Xu, X.: Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In: SSD. (1995) 67–82.
6. Francesco Gullo, Giovanni Ponti, Andrea Tagarelli: Clustering Uncertain Data Via K-Medoids. SUM 2008: 229-242.
7. G. Cormode and A. McGregor, "Approximation algorithms for clustering uncertain data," in *PODS*, M. Lenzerini and D. Lembo, Eds. Vancouver, BC, Canada: ACM, 9th–11th Jun. 2008, pp. 191–200.
8. Jaiwei Han, Micheline Kamber, Book Title ―Data Mining Concept and Techniques, Morgan Kaufmann (An Imprint of ELSEVIER) Publication, ISBN: 1-55860-489-8, 2003.
9. M. Chau, Reynold Cheng, B. Kao and J. Ng. Data with uncertainty Mining: An Example in Clustering Location Data. In the Methodologies for Knowledge Discovery and Data Mining, Pacific-Asia Conference (PAKDD 2006), Singapore, 2006.
10. Ng R.T. and Han J. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining, Proc. 20th Int. Conf. on Very Large Data Bases, 144-155. Santiago, Chile.
11. Ng, R.T., Han, J.: Efficient and effective clustering methods for spatial data mining. In: VLDB. (1994) 144–155.
12. Ramachandra Rao Kurada, Durga Sri B, "A Novel Approach by Applying Partitioning Clustering over Uncertain Spatial Data Objects using Pruning Techniques and R*-tree Indexing" in ICCCE 2012, 12 & 13 April 2012, Dr. MGR University, Chennai, Published by Coimbatore Institute of Information Technology, ISBN No: 978-1-4675-2248-9.
13. Ramachandra Rao Kurada, Durga Sri B, "Unsupervised Classification of Uncertain Data Objects in Spatial Databases Using Computational Geometry and Indexing Techniques", IJERA, Vol.2 Issue 2, March-April 2012, pp. 806-814.ISSN No.: 2248-9622.