

# Uncertain Data Mining: An Example in Clustering Location Data

Michael Chau<sup>1</sup>, Reynold Cheng<sup>2</sup>, Ben Kao<sup>3</sup>, and Jackey Ng<sup>1</sup>

<sup>1</sup> School of Business, The University of Hong Kong, Pokfulam, Hong Kong  
mchau@business.hku.hk, jackeyng@hkusua.hku.hk

<sup>2</sup> Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong  
cscckcheng@comp.polyu.edu.hk

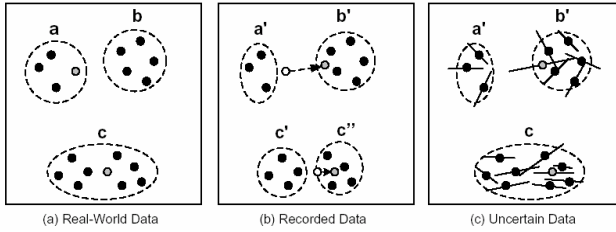
<sup>3</sup> Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong  
kao@cs.hku.hk

**Abstract.** Data uncertainty is an inherent property in various applications due to reasons such as outdated sources or imprecise measurement. When data mining techniques are applied to these data, their uncertainty has to be considered to obtain high quality results. We present UK-means clustering, an algorithm that enhances the K-means algorithm to handle data uncertainty. We apply UK-means to the particular pattern of moving-object uncertainty. Experimental results show that by considering uncertainty, a clustering algorithm can produce more accurate results.

## 1 Introduction

In applications that require interaction with the physical world, such as location-based services [6] and sensor monitoring [3], data uncertainty is an inherent property due to measurement inaccuracy, sampling discrepancy, outdated data sources, or other errors. Although much research effort has been directed towards the management of uncertain data in databases, few researchers have addressed the issue of mining uncertain data. We note that with uncertainty, data values are no longer atomic. To apply traditional data mining techniques, uncertain data has to be *summarized* into atomic values. Unfortunately, discrepancy in the summarized recorded values and the actual values could seriously affect the quality of the mining results. Figure 1 illustrates this problem when a clustering algorithm is applied to moving objects with location uncertainty. If we solely rely on the recorded values, many objects could possibly be put into wrong clusters. Even worse, each member of a cluster would change the cluster centroids, thus resulting in more errors.

We suggest incorporating uncertainty information, such as the probability density functions (pdf) of uncertain data, into existing data mining methods so that the mining results could resemble closer to the results obtained as if actual data were used in the mining process [2]. In this paper we study how uncertainty can be incorporated in data mining by using data clustering as a motivating example. In particular, we study one of the most popular clustering methods – K-means clustering.



**Fig. 1.** (a) The real-world data are partitioned into three clusters (a, b, c). (b) The recorded locations of some objects (shaded) are not the same as their true location, thus creating clusters  $a'$ ,  $b'$ ,  $c'$  and  $c''$ . (c) When line uncertainty is considered, clusters  $a'$ ,  $b'$  and  $c$  are produced. The clustering result is closer to that of (a) than (b) is.

## 2 Related Work

There is significant research interest in data uncertainty management in recent years. Most work has been devoted to “imprecise queries”, which provide probabilistic guarantees over correctness of answers. For example, in [4], indexing solutions for range queries over uncertain data have been proposed. The same authors also proposed solutions for aggregate queries such as nearest-neighbor queries in [3]. Notice that all these works have applied the study of uncertain data management to simple database queries, instead of to the more complicated data analysis and mining problems.

Clusterization has been well studied in data mining research. However, only a few studies on data mining or data clustering for uncertain data have been reported. Hamdan and Govaert have addressed the problem of fitting mixture densities to uncertain data for clustering using the EM algorithm [5]. However, the model cannot be readily applied to other clustering algorithms and is rather customized for EM. Clustering on interval data also has been studied. However, the pdf of the interval is not taken into account in most of the metrics used. Another related area of research is fuzzy clustering. In fuzzy clustering, a cluster is represented by a fuzzy subset of a set of objects. Each object has a “degree of belongingness” for each cluster. In other words, an object can belong to more than one cluster, each with a different degree. The fuzzy  $c$ -means algorithm was one of the most widely used fuzzy clustering method [1].

## 3 Clustering on Data with Uncertainty

**Problem Definition:** Let  $S$  be a set of  $V$ -dimensional vectors  $\mathbf{x}_i$ , where  $i = 1$  to  $n$ , representing the attribute values of all the records in the clustering application. Each record  $o_i$  is associated with a probability density function (pdf),  $f_i(\mathbf{x})$ , which is the pdf of  $o_i$ 's attribute values  $\mathbf{x}$  at time  $t$ . The clustering problem is to find a set  $C$  of clusters  $C_j$ , where  $j = 1$  to  $K$ , with cluster means  $\mathbf{c}_j$  based on similarity. Different clustering algorithms have different objective functions, but the general idea is to minimize the distance between objects in the same cluster while maximizing the distance between objects in different clusters. Minimization of intra-cluster distance can also be viewed

as the minimization of the distance between each data  $\mathbf{x}_i$  and the cluster means  $\mathbf{c}_j$  of the cluster  $C_j$  that  $\mathbf{x}_i$  is assigned to.

To consider data uncertainty in the clustering process, we propose a clustering algorithm with the goal of minimizing the *expected* sum of squared errors  $E(\text{SSE})$ . Note that a data object  $\mathbf{x}_i$  is specified by an uncertainty region with an uncertainty pdf  $f(\mathbf{x}_i)$ . Given a set of clusters,  $C_j$ 's the expected SSE can be calculated as follow:

$$E\left(\sum_{j=1}^k \sum_{i \in C_j} \|\mathbf{c}_j - \mathbf{x}_i\|^2\right) = \sum_{j=1}^k \sum_{i \in C_j} \int \|\mathbf{c}_j - \mathbf{x}_i\|^2 f(\mathbf{x}_i) d\mathbf{x}_i \quad (1)$$

where  $\|\cdot\|$  is a distance metric between a data point  $\mathbf{x}_i$  and a cluster mean  $\mathbf{c}_j$ . Cluster means are given by:

$$\mathbf{c}_j = E\left(\frac{1}{|C_j|} \sum_{i \in C_j} \mathbf{x}_i\right) = \frac{1}{|C_j|} \sum_{i \in C_j} \int \mathbf{x}_i f(\mathbf{x}_i) d\mathbf{x}_i \quad (2)$$

We propose a new K-means algorithm, called UK-means, for clustering uncertain data:

1. Assign initial values for cluster means  $\mathbf{c}_1$  to  $\mathbf{c}_K$
2. **repeat**
3.     **for**  $i = 1$  to  $n$  **do**
4.         Assign each data  $\mathbf{x}_i$  to cluster  $C_j$  where  $E(\|\mathbf{c}_j - \mathbf{x}_i\|)$  is the minimum
5.     **end for**
6.     **for**  $j = 1$  to  $K$  **do**
7.         Recalculate cluster mean  $\mathbf{c}_j$  of cluster  $C_j$
8.     **end for**
9. **until** convergence
10. **return**  $C$

The main difference between UK-mean clustering and the traditional K-means clustering lies in the computation of distance and clusters. In particular, UK-means compute the *expected* distance and cluster centroids based on the data uncertainty model. Convergence can be defined based on different criteria.

In Step 4, it is often difficult to determine  $E(\|\mathbf{c}_j - \mathbf{x}_i\|)$  algebraically. In particular, the variety of geometric shapes of uncertainty regions (e.g., line, circle) and different uncertainty pdf imply that numerical integration methods are necessary. We propose to use the squared expected distance  $E(\|\mathbf{c}_j - \mathbf{x}_i\|^2)$ , which is much easier to obtain.

## 4 UK-Means Clustering for Moving Objects with Uncertainty

The UK-means algorithm presented in the last section is applicable to any uncertainty region and pdf. In this section, we describe how the proposed algorithm can be applied to uncertainty models specific to moving objects that are moving in a two-dimensional space. According to [4] and [6], there are two types of moving-object uncertainty, namely line-moving uncertainty and free-moving uncertainty. In line-

moving uncertainty, an object moves at a velocity vector, which is smaller than  $V_{max}$ , along a fixed direction. Line-moving uncertainty can be unidirectional or bidirectional. The free-moving uncertainty model assumes that an object cannot move beyond a certain speed,  $V_{max}$ . Given that the current position of the object is  $(h, k)$  at time  $t_0$ , the object's location is uniformly distributed within a circle of radius  $V_{max} \times (t - t_0)$ .

Suppose we have a centroid  $\mathbf{c} = (p, q)$  and a data object  $\mathbf{x}$  specified by a line uncertainty region with a uniform distribution. Let the end points of the line segment uncertainty be  $(a, b)$  and  $(c, d)$ . The line equation can be parameterized by  $(a + t(c - a), b + t(d - b))$ , where  $t$  is between  $[0, 1]$ . Let the uncertainty pdf be  $f(t)$ . Also, let the distance of the line segment uncertainty be  $D = \sqrt{(c - a)^2 + (d - b)^2}$ . We have:

$$E(\|\mathbf{c} - \mathbf{x}\|^2) = \int_0^1 f(t)(D^2 t^2 + Bt + C) dt \quad (3)$$

where  $B = 2[(c - a)(a - p) + (d - b)(b - q)]$ ,  $C = (p - a)^2 + (q - b)^2$

If  $f(t)$  is uniform, then  $f(t) = 1$ , and the above becomes:

$$E(\text{distance of line uncertainty from centroid}^2) = \frac{D^2}{3} + \frac{B}{2} + C \quad (4)$$

For free-moving uncertainty, suppose we have a centroid  $\mathbf{c} = (p, q)$  and a data object  $\mathbf{x}$  specified by a circle uncertainty region with a uniform distribution. Suppose the circle uncertainty has center  $(h, k)$  and radius  $R$ . Let the uncertainty pdf of the circle be  $f(r, \theta)$ . Then we have:

$$E(\|\mathbf{c} - \mathbf{x}\|^2) = \int_0^R \int_0^{2\pi} f(r, \theta)(A \cos \theta + B \sin \theta + C) r dr d\theta \quad (5)$$

where  $A = 2r(h - p)$ ,  $B = 2r(k - q)$ ,  $C = r^2 + (h - p)^2 + (k - q)^2$

We are thus able to compute the expected squared distance easily for line-moving and free-moving object uncertainty. The use of uniform distribution is only a specific example here. When the pdf's are not uniform (e.g., Gaussian), sampling techniques can be used to estimate  $E(\|\mathbf{c}_j - \mathbf{x}_j\|)$ .

## 5 Experiments

In our experiments, we simulate a scenario in which a system that tracks the locations of a set of moving objects has taken a snapshot of these locations [2]. This location data is stored in a set called **recorded**. Each object assumes an uncertainty model captured in **uncertainty**. We compare two clustering approaches: (1) apply K-means to **recorded** and (2) apply UK-means to **recorded + uncertainty**. We first generated a set of random data points in a 100 x 100 2D space as **recorded**. For each data point, we then randomly generated its uncertainty according to a chosen uncertainty model. We also generated **actual** — the *actual locations* of the objects based on **recorded** and **uncertainty**, simulating the scenario that the objects have moved away from their original locations as registered in **recorded**. We remark that *ideally*, a system should know **actual** and apply K-means on the actual locations. Hence, we compute and compare the cluster outputs of the following data sets:

- (1) **recorded** (using classical K-means)
- (2) **recorded + uncertainty** (using UK-means)
- (3) **actual** (using classical K-means)

We use the Adjusted Rand Index (ARI) to measure the similarity between the clustering results [7]. A higher ARI value indicates a higher degree of similarity between two sets of clusters. We compare the ARI between the sets of clusters created in (2) and (3) and the ARI between those created in (1) and (3). Due to limited space, only the results of unidirectional line uncertainty are reported here.

The number of objects ( $n$ ), number of clusters ( $K$ ), and the maximum distance an object can move ( $d$ ) were varied during the experiment. Table 1 shows the different experiment results by varying  $d$  while keeping  $n = 1000$  and  $K = 20$ . Under each set of different parameter settings, 500 rounds were run and the results were averaged. In each round, the sets of **recorded**, **uncertainty**, and **actual** were first generated and the same set of data was used for the three clustering processes. The same set of initial centroids were also used in each of the three processes in order to avoid any bias.

The UK-means algorithm consistently showed a higher ARI than the traditional K-means algorithm applied on the recorded data. Pairwise  $t$ -tests were conducted and the results showed that the difference in the ARI values of the two methods was significant ( $p < 0.000001$  for all cases). The results demonstrated that the UK-means algorithm can give a set of clusters that could be a better prediction of the clusters that would be produced if the real-world data were available.

**Table 1.** Experiment results

$d$	1.5	2.5	5	7.5	10	20	50
ARI (UK-means)	0.740	0.733	0.689	0.652	0.632	0.506	0.311
ARI (K-means)	0.715	0.700	0.626	0.573	0.523	0.351	0.121
% of improvement	3.58%	4.77%	10.03%	13.84%	20.82%	44.34%	155.75%

## 6 Conclusions and Future Work

In this paper we present the UK-means algorithm, which aims at improving the accuracy of clustering by considering the uncertainty associated with data. Although in this paper we only present clustering algorithms for uncertain data with uniform distribution, the model can be generalized to other distribution (e.g., by using sampling techniques). We also suggest that our concept of using expected distance could be applied to other clustering approaches (such as nearest neighbor clustering and self-organizing maps) and other data mining techniques (such as data classification).

## Acknowledgement

We thank David Cheung (University of Hong Kong), Edward Hung (Hong Kong Polytechnic University) and Kevin Yip (Yale University) for their helpful comments.

## References

1. Bezdek, J. C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981).
2. Chau, M., Cheng, R., and Kao, B.: Uncertain Data Mining: A New Research Direction. In Proc. Workshop on the Sciences of the Artificial, Hualien, Taiwan (2005).
3. Cheng, R., Kalashnikov, D., and Prabhakar, S.: Querying Imprecise Data in Moving Object Environments. *IEEE TKDE*, 16(9) (2004) 1112-1127.
4. Cheng, R., Xia, X., Prabhakar, S., Shah, R. and Vitter, J.: Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data. In Proc. VLDB, 2004.
5. Hamdan, H. and Govaert, G.: Mixture Model Clustering of Uncertain Data. *IEEE International Conference on Fuzzy Systems* (2005) 879-884.
6. Wolfson, O., Sistla, P., Chamberlain, S. and Yesha, Y.: Updating and Querying Databases that Track Mobile Units. *Distributed and Parallel Databases*, 7(3), 1999.
7. Yeung, K. and Ruzzo, W.: An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics* 17(9) (2001) 763-774.