Clustering Uncertain Data Via K-Medoids

Francesco Gullo, Giovanni Ponti, and Andrea Tagarelli

DEIS, University of Calabria, Via P.Bucci 41c, Rende (CS) I87036, Italy
{fgullo,gponti,tagarelli}@deis.unical.it

Abstract. Uncertain data are usually represented in terms of an uncertainty region over which a probability density function (pdf) is defined. In the context of uncertain data management, there has been a growing interest in clustering uncertain data. In particular, the classic K-means clustering algorithm has been recently adapted to handle uncertain data. However, the centroid-based partitional clustering approach used in the adapted K-means presents two major weaknesses that are related to: *(i)* an accuracy issue, since cluster centroids are computed as deterministic objects using the expected values of the pdfs of the clustered objects; and, *(ii)* an efficiency issue, since the expected distance between uncertain objects and cluster centroids is computationally expensive.

In this paper, we address the problem of clustering uncertain data by proposing a K-medoids-based algorithm, called *UK-medoids*, which is designed to overcome the above issues. In particular, our UK-medoids algorithm employs distance functions properly defined for uncertain objects, and exploits a K-medoids scheme. Experiments have shown that UK-medoids outperforms existing algorithms from an accuracy viewpoint while achieving reasonably good efficiency.

1 Introduction

Handling uncertainty in data management has been requiring more and more importance in a wide range of application contexts. Indeed, data uncertainty naturally arises from, e.g., implicit randomness in a process of data generation/acquisition, imprecision in physical measurements, and data staling. Various notions of uncertainty have been defined depending on the application domain (e.g., [2,3,4,5,6,7,8]). In general, uncertainty can be considered at table, tuple or attribute level [9], and is usually specified by fuzzy models [10], evidence-oriented models [11,12], or probabilistic models [13].

In this paper, we focus on data containing attribute-level uncertainty, which is modeled according to a probabilistic model. We hereinafter refer to this data as *uncertain objects*. An 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 [14,15,1], rather than by a traditional vectorial form of deterministic values.

Attribute-level uncertainty expressed by means of probabilistic models is present in several application domains. For instance, sensor measurements may be imprecise at a certain degree due to the presence of various noisy factors

S. Greco and T. Lukasiewicz (Eds.): SUM 2008, LNAI 5291, pp. 229–242, 2008.

(e.g., signal noise, instrumental errors, wireless transmission) [16,14]. To address this issue, it is advisable to model sensor data as continuous pdfs [17,18]. Another example is given by data representing moving objects, which continuously change their location so that exact positional information at a given time instant may be unavailable [19]. Further examples come from distributed applications, privacy preserving data mining, and forecasting or other statistical techniques used to generate data attributes [20].

Dealing with uncertain objects has raised several issues in data management and knowledge discovery. In particular, organizing uncertain objects is challenging since the intrinsic difficulty underlying the various notions of uncertainty. As a major exploratory task of data mining, *clustering* is organizing a collection of objects (whose classification is unknown) into meaningful groups (clusters), based on interesting relationships discovered in the data. Objects within a cluster will be each other highly similar, but will be very dissimilar from objects in other clusters. One of the most popular clustering approaches is represented by partitional (or partitioning) clustering [21], which iteratively assigns objects to the clusters according to a certain distance/similarity function. A major cruciality in partitional clustering is how to devise a notion of cluster prototype. In particular, a cluster prototype can be defined as a *centroid*, which is the "mean" object in the cluster, or as a *medoid*, which is an actual object that is nearest to all the other objects in the cluster. The K-means [22] and K-medoids [23] algorithms are the exemplary methods of centroid-based and medoid-based partitional clustering, respectively.

In a recent work [1], the K-means algorithm has been adapted to the uncertain data domain. However, the resulting algorithm, named UK-means, has two major weak points. First, cluster centroids are defined as deterministic objects and computed as the mean of the expected values over the pdfs of the uncertain objects in the cluster; defining centroids in this way may result in loss of accuracy, since only the expected values of the pdfs of the uncertain objects are taken into account. Second, the computation of the Expected Distance (ED) between cluster centroids and uncertain objects is computationally expensive, as it requires non-trivial numerical integral estimations; this represents an efficiency bottleneck at each iteration of the algorithm.

In this paper, we present *UK-medoids*, an algorithm for clustering uncertain objects based on the K-medoids clustering scheme. The proposed algorithm exploits a distance function for uncertain objects, which is not limited to consider only scalar values derived from the pdfs associated to the objects (e.g., pdf expected values). This allows for better estimating the real distance between two uncertain objects, leading to significant improvement of the clustering quality. Also, our algorithm does not require any expensive operation to be repeated at each iteration; indeed, the computation of the distances between uncertain objects in the dataset is performed only once, thus guaranteeing a significant improvement of the efficiency w.r.t. UK-means. Experiments have shown that our method outperforms existing algorithms from an accuracy viewpoint while achieving reasonably good efficiency.