Knowledge-Based Systems 66 (2014) 112–125

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A heuristic approach to effective and efficient clustering on uncertain objects

Lei Xu [a,*], Qinghua Hu [b], Edward Hung [a], Chi-Cheong Szeto [a]

[a] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[b] School of Computer Science and Technology, Tianjin University, Weijin Road No. 92, Nankai District, Tianjin, PR China

ARTICLE INFO

Article history:
Received 1 September 2013
Received in revised form 3 April 2014
Accepted 17 April 2014
Available online 24 April 2014

Keywords:
Clustering
Uncertain objects
UK-means
Expected Euclidean distance
Expected squared Euclidean distance

ABSTRACT

We study the problem of clustering uncertain objects whose locations are uncertain and described by probability density functions (pdf). We analyze existing pruning algorithms and experimentally show that there exists a new bottleneck in the performance due to the overhead of pruning candidate clusters for assignment of each uncertain object in each iteration. In this article, we will show that by considering squared Euclidean distance, UK-means (without pruning techniques) is reduced to K-means and performs much faster than pruning algorithms, however, with some discrepancies in the clustering results due to using different distance functions. Thus, we propose Approximate UK-means to heuristically identify objects of boundary cases and re-assign them to better clusters. Three models for the representation of cluster representative (certain model, uncertain model and heuristic model) are proposed to calculate expected squared Euclidean distance between objects and cluster representatives in this paper. Our experimental results show that on average the execution time of Approximate UK-means is only 25% more than K-means and our approach reduces the discrepancies of K-means' clustering results by up to 70%.

1. Introduction

While there has been a large amount of research on mining and queries on relational databases [1], the focus has been on databases that store data in exact values. In many real-life applications, however, the raw data (for example, in the case of sensor data) are not precise or accurate when they were collected or produced. The clustering of such "uncertain" data can be illustrated by the following simple realistic example. Considering sensors on wild animals that update their locations periodically, the sample locations of an animal over a period generate a (discrete) probability distribution function (PDF) describing the possible locations of the animal. Clustering results on those animals may reveal the possible groups and interactions between them. In our work, we consider the problem of clustering objects with multidimensional uncertainty where an object is represented by an uncertain region over which a discrete probability distribution function (PDF) or a probability density function (pdf) is defined. Formally, there is a set of n objects $\{o_1, \ldots, o_i, \ldots, o_n\}$, $1 < i < n$ in an m-dimensional space. An object o_i is represented by a pdf $f_i : \boldsymbol{R}^m \rightarrow \boldsymbol{R}$ ($\boldsymbol{R}$ represents real number space) that specifies the probability density of each possible location of object o_i. The method discussed in this paper does not rely on any special forms of f_i, but only requires that for each object o_i, the uncertain region A_i of object o_i is finite, i.e. $\forall x \notin A_i,\ f_i(x) = 0$. Thus, each object can be bounded by a finite bounding box (A_i is bounded by a finite region, but it does not mean that the number of samples is finite). This assumption is convincing because in practice the probability density of an object is high only within a very small region of concentration.

The goal of UK-means is to group n objects into K clusters so that the sum of *expected Euclidean distances* (EED) [2] between the uncertain objects and their cluster centers is minimized. Thus, $C(o_i) = c_j$ is used to represent that object o_i is assigned to cluster c_j, and $p_{C(o_i)}$ is the cluster's representative point. K cluster representatives should be found such that the objective function $\sum_{i=1}^{n} EED(o_i, p_{C(o_i)}) = \sum_{i=1}^{n} \left(\int f_i(x) ED(x, p_{C(o_i)}) \, dx \right)$ is minimized where ED is Euclidean distance function based on a metric d (i.e. Euclidean distance in UK-means and pruning algorithms, squared Euclidean distance in our approach).

In UK-means, the calculation of expected Euclidean distance is time consuming. Pruning UK-means [3,4] improves the efficiency of UK-means by pruning candidate clusters. However, existing algorithms are still undesirably slow due to the overhead of

* Corresponding author. Tel.: +852 90633012.
E-mail address: cslxu@comp.polyu.edu.hk (L. Xu).

http://dx.doi.org/10.1016/j.knosys.2014.04.027
0950-7051/

pruning candidate clusters for each object in each iteration. In this paper, UK-means is improved by using squared Euclidean distance and heuristically reduce the discrepancy generated by using different distance functions. Three models for the representation of cluster representative (certain model, uncertain model and heuristic model) are also proposed to calculate expected squared Euclidean distance between objects and cluster representatives. Heuristic model is a combination of certain model and uncertain model for clustering uncertain objects heuristically.

Our contributions of this work include:

(i) After applying the analytic solution [5,6] to reduce UK-means to K-means, we experimentally show that K-means performs much faster than existing pruning algorithms proposed in [3,4] with some discrepancies in the clustering results due to using different distance functions.
(ii) We propose Approximate UK-means to heuristically identify objects of boundary cases and re-assign them to better clusters. Our experimental results show that on average the execution time of Approximate UK-means is only 25% more than K-means (while pruning algorithms are 300% more) and our approach reduces the discrepancies of K-means' clustering results up to 70% with assuming the results of pruning algorithms [3,4] to be the ground truth.
(iii) Different from previous work in [7], we propose three models for the representation of cluster representative (certain model, uncertain model and heuristic model) to calculate expected squared Euclidean distance between objects and cluster representatives. From the experiments, Approximate UK-means based on the uncertain model can additionally reduce the discrepancy of the certain cluster representative model up to 77% with only a little execution time increased.

The rest of the paper is organized as follows. Section 2 briefly describes related work on uncertain objects. In Section 3, we introduce expected squared Euclidean distance by using three different models of cluster representative. Section 4 proposes a heuristic Approximate UK-means to identify objects of boundary cases and re-assign them to better clusters. Section 5 demonstrates the efficiency and effectiveness of Approximate UK-means by extensive experiments. Finally, Section 6 concludes the paper.

2. Related work

There has been significant research interest in uncertain database [8,9]. Data uncertainty are classified into two types. One is existential uncertainty, which is caused by the fact that we are not sure the object or data tuple exists or not [10–13]; The other is value uncertainty caused by not knowing the value precisely. In this paper, we focus on the second case (value uncertainty). The relationship uncertainty is first proposed by Jiang et al. in [14] recently. The task of summarizing the relationship uncertainty [14] between objects is learning the order of the values on a dimension of the domain. For example, a traveler gives a higher score to a hotel whose location is close to the central, and a lower score to a hotel that is far from the central, which likely means that the user prefers the hotel near the central. Learning the order of values can infer more knowledge of domain.

2.1. Clustering on uncertain objects

UK-means [2] is a generalization of traditional K-means to handle uncertain objects whose locations are represented by pdfs. In K-means, object o is assigned to cluster c such that the Euclidean distance between o and c's representative is the smallest among all clusters. The only difference between UK-means and K-means is distance calculation method between an object and a cluster representative. Expected Euclidean distance is calculated in UK-means while K-means uses Euclidean distance. For arbitrary pdfs, the bottleneck of UK-means is the calculation of expected distance, which is computationally expensive. Thus, pruning techniques are proposed to remove the candidate clusters from consideration, which are certainly not closest to the object, reducing a large amount of expected distance calculations.

The basic idea of pruning techniques is using simple distance calculation to identify cluster representatives that cannot be the closest one to a given uncertain object. Hence, the expected distance calculations between those cluster representatives and the object can be skipped. In MinMax-BB [3], each object o_i has a minimum bounding rectangle (MBR)[1] outside which the object has zero (or negligible) probability of occurrence. The minimum distance ($MinDist_{i,j}$) and the maximum distance ($MaxDist_{i,j}$) are calculated to prune unnecessary expected distance calculations. Among all the maximum distances, the smallest one is called the minmax distance which is used to prune unnecessary expected distance calculations. The overhead of MinMax-BB pruning includes the time of $MinDist_{i,j}$ and $MaxDist_{i,j}$ calculations.

VDBi [4] is another pruning method using VORONOI diagrams [15] to consider the spatial relationships among cluster representatives. VDBi [4] is more efficient than MinMax-BB by using VORONOI-cell pruning and bisector pruning. VDBi pruning includes VORONOI-cell pruning and bisector pruning. For VORONOI-cell pruning, given K cluster representatives, the VORONOI diagrams divide the space $\boldsymbol{R}^m$ into k cells called $V(p_{c_1})$, $V(p_{c_2})$,..., $V(p_{c_j})$,..., $V(p_{c_k})$ with the properties of $d(x,p_{c_j}) < d(x,p_{c_k})$ $\forall x \in V(p_{c_j})$, $p_{c_j} \neq p_{c_k}$. Therefore, if the MBR of object o_i lies completely inside any VORONOI-cell $V(p_{c_j})$, object o_i can be assigned to cluster c_j directly without any expected distance calculation. Bisector pruning considers the case of distinct cluster representative pair (i.e. p_{c_j} and p_{c_k} from a set of cluster representatives C). The boundary of a cell $V(p_{c_j})$ and its adjacent cell $V(p_{c_k})$ consists of points in the perpendicular bisector, denoted by $p_{c_j}|p_{c_k}$ between the points p_{c_j} and p_{c_k}. The bisector is the hyperplane that is perpendicular to the line segment joining p_{c_j} and p_{c_k} that passes through the mid-point of the line segment. The space $\boldsymbol{R}^m$ is divided into two halves. $H_{j/k}$ denotes the half containing p_{c_j} (excluding the hyperplane). Thus, the following properties are obtained:

(i) $\forall p_{c_j},\ p_{c_k} \in C,\ d(x,p_{c_j}) < d(x,p_{c_k})\ \forall x \in H_{j/k}$;
(ii) $d(x,p_{c_j}) = d(x,p_{c_k})\ \forall x \in p_{c_j}|p_{c_k}$.

If the MBR of o_i lies completely in $H_{j/k}$, p_{c_k} can be pruned from C. Thus, VDBi can be more efficient than MinMax-BB. If a candidate cluster c_j cannot be pruned by VDBi, neither does MinMax-BB. However, VDBi may prune a candidate cluster c_j that cannot be pruned by MinMax-BB [4]. The overhead of VDBi includes the time of VORONOI diagrams construction, VORONOI-cell pruning and bisector pruning. The pruning methods can be more efficient with the use of cluster-shift technique [3,4]. Because it is likely that the cluster representatives shift by small distance in the next iteration, the tighter bound can be made to prune candidate clusters more efficiently. Cluster-shift (SHIFT) technique can be applied to MinMax-BB and VDBi pruning techniques. The additional overhead of SHIFT technique includes the time of cluster representative shift calculation between two consecutive iterations. Although the

[1] The pruning techniques [3,4] require that for each object o_i, the uncertain region A_i of each object o_i is finite, but it does not mean that the number of samples should be finite. We can just say that the probability of the possible location x will be 0 if the possible location x falls outside A_i, i.e. $\forall x \notin A_i$, $f_i(x) = 0$. Thus, each object can be bounded by a finite bounding box.