Information Systems 36 (2011) 476–497

Contents lists available at ScienceDirect

Information Systems

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

Metric and trigonometric pruning for clustering of uncertain data in 2D geometric space

Wang Kay Ngai [a], Ben Kao [a,*], Reynold Cheng [a], Michael Chau [b], Sau Dan Lee [a], David W. Cheung [a], Kevin Y. Yip [c,d]

[a] Department of Computer Science, The University of Hong Kong, Hong Kong
[b] School of Business, The University of Hong Kong, Hong Kong
[c] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[d] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States

ARTICLE INFO

Article history:
Received 16 August 2009
Received in revised form
19 August 2010
Accepted 16 September 2010
Recommended by P. Pucheral

Keywords:
Clustering
Data uncertainty

ABSTRACT

We study the problem of clustering data objects with location uncertainty. In our model, a data object is represented by an uncertainty region over which a probability density function (pdf) is defined. One method to cluster such uncertain objects is to apply the UK-means algorithm [1], an extension of the traditional K-means algorithm, which assigns each object to the cluster whose representative has the smallest *expected distance* from it. For arbitrary pdf, calculating the expected distance between an object and a cluster representative requires expensive integration of the pdf. We study two pruning methods: pre-computation (PC) and cluster shift (CS) that can significantly reduce the number of integrations computed. Both pruning methods rely on good bounding techniques. We propose and evaluate two such techniques that are based on metric properties (Met) and trigonometry (Tri). Our experimental results show that Tri offers a very high pruning power. In some cases, more than 99.9% of the expected distance calculations are pruned. This results in a very efficient clustering algorithm.[1]

1. Introduction

Clustering is a technique that has been widely studied and used in real applications. Many efficient algorithms, including the well-known and widely applied K-means algorithm, have been devised to solve the clustering problem efficiently. Traditionally, clustering algorithms deal with a set of objects whose positions are accurately known, and do not address situations in which object locations are uncertain. Data uncertainty is, however, inherent in many real-life applications due to factors such as the random nature of the physical data generation and collection processes, measurement errors, and data staling. Recent works (e.g., [3–5]) have also suggested to protect location privacy by lowering the precision of a user's location, which poses problems for traditional clustering algorithms.

In this paper we study the problem of clustering spatial objects with location uncertainty. In our model, an object's location is represented by a spatial probability density function (pdf). Our objective is to study the computational issues in adapting the traditional K-means algorithm to clustering uncertain objects, and to devise efficient algorithms for solving the clustering problem.

* Corresponding author.
E-mail addresses: wkngai@cs.hku.hk (W.K. Ngai), kao@cs.hku.hk (B. Kao), ckcheng@cs.hku.hk (R. Cheng), mchau@business.hku.hk (M. Chau), sdlee@cs.hku.hk (S.D. Lee), dcheung@cs.hku.hk (D.W. Cheung), kevinyip@cse.cuhk.edu.hk (K.Y. Yip).

[1] Part of this paper appears in Ngai et al., 2006 [2], in which the algorithms PC and CS were described. The idea of trigonometric pruning (Section 6), most of the empirical performance study (Section 7), discussion (Section 8), and all proofs of theorems (Appendixes), have not been previously published. The new materials amount to about $\frac{3}{4}$ of the current paper.

0306-4379/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2010.09.005

As a motivating example, let us consider the problem of clustering mobile devices. In many wireless network applications, mobile devices report their locations periodically to a remote server [6]. Each device can make low-power short-ranged communication to neighboring devices, or high-power long-ranged communication with the remote server directly. To reduce power consumption, batching protocols have been proposed. Under these protocols, certain devices are elected as leaders, whose job is to collect messages from neighboring devices through short-ranged communication. The leaders then send the collected messages in batch to the server through long-ranged communication [7,8] (Fig. 1). By batching messages, many long-ranged messages are replaced by short-ranged ones. The election of local leaders can be formulated as a clustering problem. The goal is to minimize the distance between every device and its corresponding local leader. This clustering problem differs from the traditional setting in the existence of data uncertainty:

- The physical instruments used for determining the device locations are accurate only up to a certain precision.
- The current locations of the mobile devices can only be estimated based on their last reported values, i.e., the data are always stale. Other practical problems, such as packet loss, could also increase the degree of uncertainty.
- Data uncertainty may also be introduced by the user to protect his location privacy. Particularly, the idea of *location cloaking* has been investigated [4,5], where the actual location of a user is converted to a larger region, before it is sent to the service provider.

Due to uncertainty, the whereabouts of a mobile device can only be estimated by imposing an uncertainty model on its last reported location [9]. A typical uncertainty model requires knowledge about the moving speed of the device and whether its movement is restricted (such as a car moving in a road network) or unrestricted (such as a tracking device mounted on an animal moving on plains). Typically, a 2D probability density function is defined over a bounded region to model such uncertainty.

Let us now formally define our uncertain data clustering problem. We consider a set of n objects o_i $(1 \le i \le n)$ in a 2D space. Each object o_i is represented by a probability density function (pdf) $f_i : \mathbb{R}^2 \to \mathbb{R}$ that specifies the probability density of each possible location of the object. The goal is to partition the objects into k clusters, such that each object o_i is assigned to a cluster c_i and that o_i is *close* to a cluster representative point p_{c_i} of c_i. To measure *closeness*, we define a distance function between an uncertain object and a cluster representative point as the expected distance between them:

$$ED(o_i,p_{c_i}) = \int f_i(x)d(x,p_{c_i})\,dx, \tag{1}$$

where d is the Euclidean distance and the integration is taken over the *uncertainty region* (which is assumed to be bounded, as we will discuss below) in which the pdf integrates to one. Given a cluster c_i, its representative p_{c_i} is given by the mean of the centers of mass of all the objects assigned to c_i. The clustering goal is then to find c_i's (and thus p_{c_i}'s) such that the following objective function is minimized:

$$G = \sum_{i=1}^{n} ED(o_i,p_{c_i}) = \sum_{i=1}^{n} \int f_i(x)d(x,p_{c_i})\,dx. \tag{2}$$

We assume that the pdfs can take any arbitrary form, which is important when the possible locations of a device are constrained by its dynamic environment, such as the road structure. The only additional requirement we impose on the pdfs is that each of them should integrate to one within a bounded region. This is a reasonable requirement for many applications. For example, the current location of a mobile device is restricted by its last reported location, its maximum speed, and the duration between two location reports [10,11]. In location cloaking, the actual coordinates of a user's location were replaced by a uniform distribution over a bounded region [4,5].

In a separate study [1], it was shown that the quality of clustering results could be improved by explicitly considering data uncertainty. An algorithm called UK-means (Uncertain K-means) was proposed to take data uncertainty into account during the clustering process. Experimental results showed that UK-means consistently produced better clusters than the traditional K-means algorithm. Yet for arbitrary pdfs, expected distance calculations require costly numerical integrations. A straightforward implementation of UK-means for arbitrary pdfs would require a lot of such expected distance calculations, which are computationally impractical.

In this paper we study two pruning algorithms, namely pre-computation (PC) and cluster-shift (CS), which can significantly reduce the number of expected distance calculations of UK-means. The effectiveness of both algorithms relies on good bounds of expected distances. We propose and evaluate two bounding techniques that are based on metric properties (`Met`) and trigonometry (`Tri`). `Met` bounds are derived using the triangle inequality, and `Tri` bounds are obtained by a number of trigonometric rules. Our experimental results show that while the simple `Met` bounds are already powerful in pruning expected distance calculations, the more advanced `Tri` bounds provide further pruning power. In some of our experiments, more than 99.9% of the expected

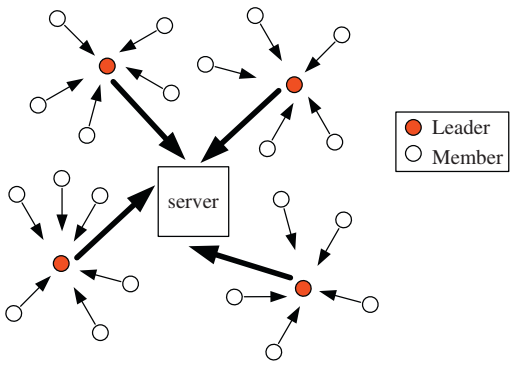

Fig. 1. Reporting locations to cluster leaders using short-ranged communication has a much lower power consumption than making long-ranged communication with the server directly.