Physica A 460 (2016) 188–200

Contents lists available at ScienceDirect

Physica A

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

Spatial clustering with Density-Ordered tree

Qing Cheng [a,*], Xin Lu [b,c,d,e], Zhong Liu [a], Jincai Huang [a], Guangquan Cheng [a]

[a] *Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, PR China*

[b] *College of Information System and Management, National University of Defense Technology, Changsha 410073, PR China*

[c] *Flowminder Foundation, Stockholm 17177, Sweden*

[d] *Department of Public Health Sciences, Karolinska Institutet, Stockholm 17177, Sweden*

[e] *Division of Infectious Disease, Key Laboratory of Surveillance and Early-Warning on Infectious Disease, Chinese Centre for Disease Control and Prevention, Beijing 102206, PR China*

HIGHLIGHTS

- We develop a Density-Ordered tree to represent the original data. It efficiently integrates information on not only distance but also density between data points.
- Our method can effectively identify clusters with diverse shapes and densities for spatial dataset.
- Our method provides an innovative way to identify noise and cluster center.
- Experiments demonstrate that our method is more effective than DBSCAN and Chameleon.

ARTICLE INFO

Article history:
Received 26 October 2015
Received in revised form 28 March 2016
Available online 14 May 2016

Keywords:
Spatial clustering
Diverse shapes and densities
Density-Ordered tree
Partition-and-merge strategy

ABSTRACT

Clustering has emerged as an active research direction for knowledge discovery in spatial databases. Most spatial clustering methods become ineffective when inappropriate parameters are given or when datasets of diverse shapes and densities are provided. To address this issue, we propose a novel clustering method, called SCDOT (Spatial Clustering with Density-Ordered Tree). By projecting a dataset to a Density-Ordered Tree, SCDOT partitions the data into several relatively small sub-clusters with a box-plot method. A heuristic method is proposed to find the genuine clusters by repeatedly merging sub-clusters and an iteration strategy is utilized to automatically determine input parameters. Moreover, we also provide an innovative way to identify cluster center and noise. Extensive experiments on both synthetic and real-world datasets demonstrate the superior performance of SCDOT over the baseline methods.

1. Introduction

Spatial data mining becomes more and more important in the analysis of spatial databases since increasingly large amounts of data obtained from satellite images, X-ray crystallography or other automatic equipment are stored in spatial databases. Spatial clustering, which groups similar spatial objects into classes, is an important component of spatial data mining. Spatial clustering can be used as a tool to get insight into the distribution of data, to observe the characteristics of

[*] Corresponding author.
E-mail address: sgggps@163.com (Q. Cheng).

http://dx.doi.org/10.1016/j.physa.2016.05.041

each cluster, and to focus on a particular set of clusters for further analysis. It has been applied in diverse fields, such as spatial epidemiology, landscape ecology, crime analysis, disease surveillance and population genetics [1–4].

Due to its wide applications in various areas, many studies on spatial clustering have been conducted [5,6], but most of them become ineffective when inappropriate parameters are given or when spatial datasets with diverse shapes, sizes and densities are provided. For example, partition clustering methods, such as K-means [7] and K-medoids [8], are very sensitive to noise and cannot detect arbitrary shaped clusters. Some advanced density-methods are applied in spatial clustering, such as DBSCAN [9]. DBSCAN relies on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. However, the performance of DBSCAN depends on two specified parameters and it does not perform well for datasets with varying densities. Several variants of hierarchical clustering methods were proposed to effectively handle spatial dataset with different shapes and densities, such as Chameleon [10] and SAM [11], however, similar to DBSCAN, all these methods still require some pre-determined parameters as inputs. In fact, without enough prior knowledge, appropriate parameters are often not known in advance when dealing with large databases. This drawback also exists in some shared nearest neighbor approaches, such as SNN [12]. Thus, without specifying any parameters related to respective datasets, it remains a challenge to identify clusters of different shapes and densities in spatial clustering.

In this paper, we propose a novel clustering algorithm, namely, Spatial Clustering with Density-Ordered Tree (SCDOT), which is capable of detecting clusters of different shapes and densities, in detail, we construct a Density-Ordered tree (DOT) to represent original data by combining density and distance. Then, a split-and-merge strategy is applied to DOT for identifying clusters with diverse shapes and densities, and an iteration optimization is utilized to automatically determine input parameters. To the best of our knowledge, this is the first work that employs the density-based clustering, partition and hierarchal clustering principle. The main contributions of this paper are summarized as follows:

1. We develop a Density-Ordered tree method to represent the original data. The method efficiently reflects not only distance relation but also density relation among data points. Based on this method, we convert the cluster discovery problem into a Density-Ordered tree partitioning and merging problem.
2. Our method provides an innovative way to identify noise and cluster center. After Density-Ordered tree being divided into many sub-trees, cluster centers are recognized as roots of sub-trees, and noises are recognized as the anomalous leaves.
3. The new clustering method can effectively identify clusters with diverse shapes and densities for spatial dataset. Given the number of clusters to be detected, there is no need to tune other parameters by the user.

2. Related work

The problem of detecting clusters has been extensively studied for years, in this paper we focus on the research of spatial clustering. An distinct attribute of spatial data is that it contains spatial information and are composed of clusters with diverse shapes, sizes and densities. In this section, a brief review of spatial clustering methods is given.

The partition clustering method splits a dataset at once using an objective function, K-means [7] is one of the most popular examples of partition clustering, it employs mean-squared-error as its objective function: find partitions such that the sum of square error between empirical mean of a cluster and points in that cluster is minimized. Many variations of K-means were proposed, such as K-medoids [8], K-medians clustering [13], K-means++ [14], Fuzzy c-means [15]. CLARANS (clustering large applications based on randomized search) [5] is another type of partition clustering method which works well on large dataset. Partition clustering produces inaccurate results when the objective function used does not capture the intrinsic structure of the data [11]. This limitation makes partition clustering incapable of handling clusters of arbitrary shapes, distinct sizes and densities. Similar limitation exists for distribution-based methods, such as EM algorithm for clustering [16], one attempts to reproduce the observed realization of data points as a mix of predefined probability distribution functions, the accuracy of such methods depends on the capability of the trial probability to represent the data. In particular, for many real datasets, there may be no concisely predefined probability.

Density-based clustering methods assume clusters as dense regions of objects in the data space and are separated by regions of low density. They are much powerful for filtering outliers and discovering arbitrary shaped clusters compared with partition and distribution-based clustering methods. DBSCAN [9] is a well-known density-based clustering algorithm. It judges the density around the neighborhood of an object to be sufficiently dense if the number of data points within a distance *eps* of the object is greater than *MinPts* number of points. It can discover clusters of arbitrary shapes and sizes. The main weakness of DBSCAN is that the performance is poor when clusters have greatly varied densities and users need to set suitable parameters in order to get qualified result. Additionally, many varieties of DBSCAN, such as OPTICS (Ordering Points to Identify the Clustering Structure) [17], l-DBSCAN [18], rough-DBSCAN [19], S-T DBSCAN [20], DENCLUE [21], are proposed to improve clustering performance in different perspectives, but pre-defined parameters, such as *MinPts* and *eps*, are still needed from users.

More recently, Rodriguez and Laio [22] proposed a clustering method based on the idea that cluster centers are characterized by higher density than their neighbors and by relatively large distance from points with higher densities. Similar to the K-medoids method, it has its basis only in the distance between data points. Like DBSCAN, it is able to detect nonspherical clusters and to automatically find adequate number of clusters. It provided not only a new way for combining density and distance to identify clusters with diverse shapes and densities but also a novel criterion for the automatic choice of cluster centers. We have used this method to solve the problem of community discovery [23].