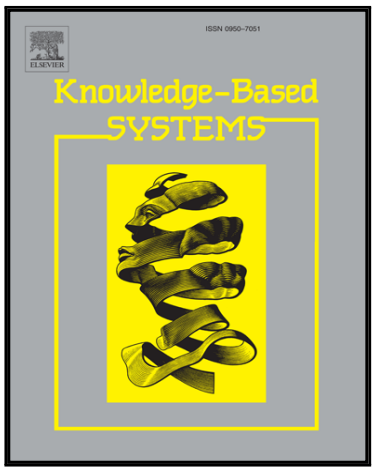

Density Core-based Clustering Algorithm with Dynamic Scanning Radius

Xie Jiang, Zhong-Yang Xiong, Yu-Fang Zhang, Yong Feng, Jie Ma

PII: S0950-7051(17)30557-9
DOI: 10.1016/j.knosys.2017.11.025
Reference: KNOSYS 4120

To appear in: *Knowledge-Based Systems*

Received date: 23 July 2017
Revised date: 14 November 2017
Accepted date: 22 November 2017

Please cite this article as: Xie Jiang, Zhong-Yang Xiong, Yu-Fang Zhang, Yong Feng, Jie Ma, Density Core-based Clustering Algorithm with Dynamic Scanning Radius, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.11.025

Density Core-based Clustering Algorithm with Dynamic Scanning Radius

Jiang Xie, Zhong-Yang Xiong, Yu-Fang Zhang, Yong Feng, Jie Ma

(College of Computer Science, Chongqing University, Chongqing 404100, China)

Abstract

Clustering analysis has been widely used in many fields such as image segmentation, pattern recognition, data analysis, market researches and so on. However, the distribution patterns of clusters are natural and complex in many research areas. In other words, most real data sets are non-spherical or non-elliptical clusters. For example, face images and hand-writing digital images are distributed in manifolds and some biological data sets are distributed in hyper-rectangles. Therefore, it is a great challenge to detect clusters of arbitrary shapes in multi-density datasets. Most of previous clustering algorithms cannot be applied to complex patterns with large variations in density because they only find hyper-elliptical and hyper-spherical clusters through centroid-based clustering approaches or fixed global parameters. This paper presents DCNaN, a clustering algorithm based on the density core and the natural neighbor to recognize complex patterns with large variations in density. Density cores can roughly retain the shape of clusters and natural neighbors are introduced to find dynamic scanning radiuses rather than fixed global parameters. The results of our experiments show that compared to state-of-the-art clustering techniques, our algorithm achieves better clustering quality, accuracy and efficiency, especially in recognizing extremely complex patterns with large variations in density.

Keywords:
Clustering, Density core, Dynamic Scanning Radius, Natural Neighbor

1. Introduction

Nowadays, a large number of data continuously generate from application such as consumer clicks, telephone usage flows, multimedia data mining, computer network intrusion detection and sensor network data clustering. Clustering data has become one of the most important issues since a majority of unlabeled data come in the age of Big Data[1]. Along with the emergence of Big Data, there is an ever-increasing interest in clustering algorithms that can automatically understand, process and summarize the data [9]. Clustering analysis is the process of dividing a dataset into groups or clusters according to their similarities, therefore the objects in different clusters have few similarities while the objects in the same cluster have many similarities. There are many clustering algorithms such as density-based methods[3], partitioning methods[10], hierarchical methods[12][13], grid-based methods, fuzzy clustering methods[15], mean shift clustering methods[14] and combinations of these. Density-based methods and partitioning methods are the most popular two.

As the representative of density-based methods, Density-based Spatial Clustering of Applications with Noise(DBSCAN)[3] is widely used in clustering. It can detect clusters in any arbitrary shapes through a fixed scanning radius eps and a density threshold MinPts[16][18]. However, DBSCAN cannot be applied to high-dimensional data with large variations in density because of the fixed global parameter and the curse of dimensionality. Some clustering algorithms[17][19][20] are introduced to eliminate the limitations of DBSCAN.

Partitioning methods, such as K-Means[4], K-Medoids and K-Center[11], have been widely used to cluster data in Gaussian-like distribution for half a century. Restrictions in the selection of right initial clusters lead that K-means algorithm must run repeatedly to achieve better results. A data point is always assigned to the nearest center, therefore partitioning methods are neither able to detect non-spherical clusters nor robust to outliers and noises.

It is crucial to find clusters of arbitrary shapes in the application of clustering algorithms. For example, in the process of geographic information data, mountains, rivers and other terrains often present various irregular shapes that can be recognized by clustering algorithms. Moreover, in medicinal fields, clustering algorithms can effectively identify irregular spatial structures of biological proteins, which helps to cognize the composition and function of proteins. Therefore, in recent years, a lot of clustering algorithms have been presented for the discovery of arbitrary shapes. Clustering by Fast Search and Find of Density Peaks (DPeak)[8] is based on the assumption that whatever the shape of a cluster is, centers are surrounded by neighbors with lower local densities and these neighbors are at relatively large distance from any points with higher local densities. DPeak first identify cluster centers, then assign the remaining points to their appropriate clusters and detect outliers as well. However, DPeak also has some drawbacks. Firstly, the selection of the cutoff distance dc has great influence on the clustering results; Secondly, as a centroid-based method, DPeak may lead up to shape loss, false distance and false peak when applied to complex patterns [2]. In the elimination of the high dimensional curse, DPC-KNN-PCA[23] combines PCA, DPeak and k-nearest neighbor to avoid

Email address: xiejiang@cqu.edu.cn (Jiang Xie, Zhong-Yang Xiong, Yu-Fang Zhang, Yong Feng, Jie Ma)