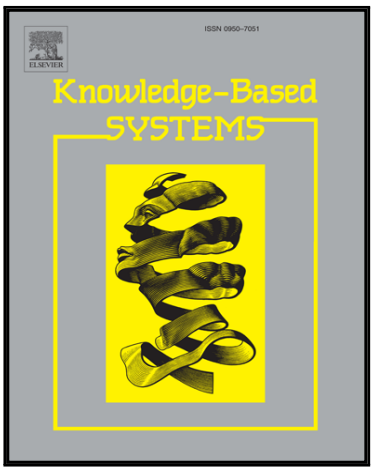

Adaptive density peak clustering based on k-nearest neighbors with aggregating strategy

Liu Yaohui, Ma Zhengming, Yu Fang

PII: S0950-7051(17)30326-X
DOI: 10.1016/j.knosys.2017.07.010
Reference: KNOSYS 3974

To appear in: *Knowledge-Based Systems*

Received date: 5 September 2016
Revised date: 30 June 2017
Accepted date: 8 July 2017

Please cite this article as: Liu Yaohui, Ma Zhengming, Yu Fang, Adaptive density peak clustering based on k-nearest neighbors with aggregating strategy, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.07.010

Adaptive density peak clustering based on k-nearest neighbors with aggregating strategy

Liu Yaohui[a,b], Ma Zhengming[a,], Yu Fang[b]

[a]*School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, Guangdong, China*
[b]*School of Software and Communication Engineering, Xiangnan University, Chenzhou 423000, Hunan, China*

Abstract

Recently a density peaks based clustering algorithm (dubbed as DPC) was proposed to group data by setting up a decision graph and finding out cluster centers from the graph fast. It is simple but efficient since it is noniterative and needs few parameters. However, the improper selection of its parameter cutoff distance d_c will lead to the wrong selection of initial cluster centers, but the DPC cannot correct it in the subsequent assignment process. Furthermore, in some cases, even the proper value of dc was set , initial cluster centers are still difficult to be selected from the decision graph. To overcome these defects, an adaptive clustering algorithm (named as ADPC-KNN) is proposed in this paper. We introduce the idea of K-nearest neighbors to compute the global parameter d_c and the local density ρ_i of each point, apply a new approach to select initial cluster centers automatically, and finally aggregate clusters if they are density reachable. The ADPC-KNN requires only one parameter and the clustering is automatic. Experiments on synthetic and real-world data show that the proposed clustering algorithm can often outperform DBSCAN, DPC, K-Means++, Expectation Maximization (EM) and single-link.

Keywords: Clustering algorithm, Density peaks, K-nearest neighbors, aggregating

1. Introduction

Clustering is the task to find a set of groups that similar objects are in the same group but different objects are separated into different groups. Since clustering can uncover the inherent, potential and unknown knowledge, principles or rules in the real-world, it has been widely used in many fields, including data mining, pattern recognition, machine learning, information retrieval, image analysis and computer graphics [1, 8, 13, 16, 24, 32, 33]. Several different clustering strategies such as the partitioning, the hierarchical, the density-based, the distribution-based have been proposed [13, 21, 24, 33], but no consensus has been reached even on the definition of a cluster [24].

The k-means clustering algorithm is the popular one of the partitioning methods. It starts with k initial cluster centers and then assigns each object iteratively to the "closest" cluster by optimizing an objective function [8, 15, 21]. However, assigning each object to its nearest center makes the k-means algorithm fail to detect non-spherical clusters [15]. k-means++ [2] provides a method to select initial cluster centers and improves the accuracy of k-means.

Density-based clustering is a nonparametric approach where the clusters are considered to be high-density areas and separated from each other by contiguous regions with low density of objects [1, 7, 8, 19, 24, 25, 32]. In density-based spatial clustering of applications with noise (DBSCAN) [8], points are classified as core objects or outliers with the density thresholds and the core objects are assigned to a cluster if they are closely packed together. However, choosing an appropriate threshold can be nontrivial [8, 19].

*School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, Guangdong, China. Tel: +86 13925128057. email: issmzm@mail.sysu.edu.cn (Z. M. Ma)

Email addresses: liuyh28@mail2.sysu.edu.cn (Liu Yaohui), issmzm@mail.sysu.edu.cn (Ma Zhengming), yfjammy@qq.com (Yu Fang)