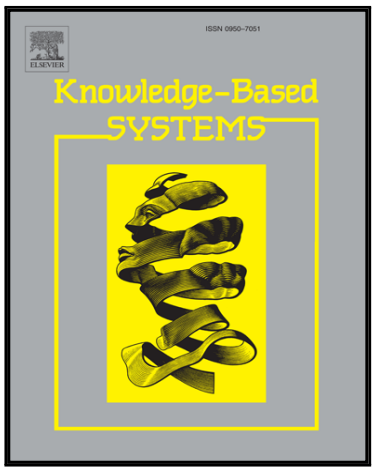
Accepted Manuscript



An improved Density Peaks Clustering Algorithm with Fast Finding Cluster Centers

Xiao Xu , Shifei Ding , Zhongzhi Shi

PII: S0950-7051(18)30264-8
DOI: 10.1016/j.knosys.2018.05.034
Reference: KNOSYS 4354

To appear in: *Knowledge-Based Systems*

Received date: 28 November 2017
Revised date: 22 May 2018
Accepted date: 24 May 2018

Please cite this article as: Xiao Xu , Shifei Ding , Zhongzhi Shi , An improved Density Peaks Clustering Algorithm with Fast Finding Cluster Centers, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.05.034

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

An improved Density Peaks Clustering Algorithm with Fast Finding Cluster Centers

Xiao Xu[1], Shifei Ding[1,2], Zhongzhi Shi[2]

[1](School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

[2](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China)

Abstract: Fast and efficient are common requirements for all clustering algorithms. Density peaks clustering algorithm (DPC) can deal with non-spherical clusters well. However, due to the difficulty of large-scale data set storage and its high computational complexity, how to conduct effective data mining has become a challenge. To address this issue, we propose an improved density peaks clustering algorithm with fast finding cluster centers, which improves the efficiency of DPC algorithm by screening points with higher local density based on two novel prescreening strategies. The first strategy is based on the grid-division (GDPC), which screens points according to the density of corresponding grid cells. The second strategy is based on the circle-division (CDPC), which screens the points according to the uneven distribution of data sets in the corresponding circles. Theoretical analysis and experimental results show that both the prescreening strategies can reduce the calculation complexity, and the proposed algorithm not only more satisfied than DPC algorithm, but also superior than well-known Nyström-SC algorithm on the large-scale data sets. Moreover, due to the different theories of the two prescreening strategies, the first strategy is faster and the second strategy is more accurate on the large-scale data sets.

Keywords: density peaks clustering algorithm; prescreening strategy; large-scale data set; decision graph; computational complexity

1 Introduction

The rapid development of information technology and the spread of the Internet make data update faster, diversify data sources and lead to the data volumes increase at an unprecedented rate. Faced with difficulty of large-scale data storage, high computational complexity and a series of problems, how to carry on the effective data mining of large-scale data sets and fast access to valuable information has become the focus of recent study [1]. Clustering learning is an important data analysis technique. In order to find useful information from the complex data, we can do data clustering first, that is group similar objects in the same class according to the characteristics of data objects, and the objects that difference between them is larger are divided into different classes, to find the internal relation between data. Clustering learning provides support for decision making and has a certain application value in market analysis, pattern recognition, genetic research, image processing, and other fields [2].

Density peaks clustering algorithm is a state-of-the-art density-based clustering algorithm proposed by Rodriguez and Laio in 2014, which is abbreviated as DPC [3]. DPC algorithm is based on two characteristics of the cluster centers: 1) the local density of the cluster center is larger, that is, the density of its neighbors is not more than itself; 2) the distance between the

Corresponding author. Tel.: +8615852494396; e-mail: dingsf@cumt.edu.cn

Download English Version:

https://daneshyari.com/en/article/6861277

Download Persian Version:

https://daneshyari.com/article/6861277

Daneshyari.com