Information Sciences 436–437 (2018) 13–30

Contents lists available at ScienceDirect

Information Sciences

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

RECOME: A new density-based clustering algorithm using relative KNN kernel density

Yangli-ao Geng[a], Qingyong Li[a,*], Rong Zheng[b], Fuzhen Zhuang[c,d], Ruisi He[e], Naixue Xiong[f]

[a] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[b] Department of Computing and Software, McMaster University, Hamilton, Canada
[c] Key Lab of Intelligen Information Processing of Chinese Academy of Sciences (CAS), ICT, CAS, Beijing 100190, China
[d] University of Chinese Academy of Sciences, Beijing 100049, China
[e] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
[f] Department of Mathematics and Computer Science, Northeastern State University, Tahlequah, USA

ARTICLE INFO

Article history:
Received 7 April 2017
Revised 10 December 2017
Accepted 8 January 2018

Keywords:
Density-based clustering
Density estimation
K nearest neighbors
Graph theory

ABSTRACT

Discovering clusters from a dataset with different shapes, densities, and scales is a known challenging problem in data clustering. In this paper, we propose the RElative COre MErge (RECOME) clustering algorithm. The core of RECOME is a novel density measure, i.e., Relative *K* nearest Neighbor Kernel Density (RNKD). RECOME identifies core objects with unit RNKD, and partitions non-core objects into atom clusters by successively following higher-density neighbor relations toward core objects. Core objects and their corresponding atom clusters are then merged through α-reachable paths on a KNN graph. We discover that the number of clusters computed by RECOME is a step function of the α parameter with jump discontinuity on a small collection of values. A fast jump discontinuity discovery (FJDD) method is proposed based on graph theory. RECOME is evaluated on both synthetic datasets and real datasets. Experimental results indicate that RECOME is able to discover clusters with different shapes, densities, and scales. It outperforms six baseline methods on both synthetic datasets and real datasets. Moreover, FJDD is shown to be effective to extract the jump discontinuity set of parameter α for all tested datasets, which can ease the task of data exploration and parameter tuning.

1. Introduction

Clustering, also known as unsupervised learning, is a process of discovery and exploration for investigating inherent and hidden structures within a large dataset [10]. It has been extensively applied to a variety of tasks [11,17,18,20,21,30,32,41,45–47]. Many clustering algorithms have been proposed in different scientific disciplines [13], and these methods often differ in the selection of objective functions, probabilistic models or heuristics adopted. Nonetheless, two difficulties, how to choose appropriate clustering number and how to discover clusters of an arbitrary shape, are faced by most methods. Density-based clustering approaches are characterized by aggregating mechanisms based on density [28]. They can handle data with

* Corresponding author.
E-mail addresses: gengyla@bjtu.edu.cn (Y.-a. Geng), liqy@bjtu.edu.cn (Q. Li), rzheng@mcmaster.ca (R. Zheng), zhuangfz@ics.ict.ac.cn (F. Zhuang), ruisi.he@bjtu.edu.cn (R. He).

https://doi.org/10.1016/j.ins.2018.01.013
0020-0255/

irregular shapes and determine clustering number automatically. Ester et al. [9] and Sander et al. [35] pioneered two density-based methods, Density Based Spatial Clustering of Applications with Noise (DBSCAN) and Generalizing DBSCAN, to detect clusters in a spatial database according to density differences. Although both methods can detect clusters with different shapes, they face the challenge of choosing appropriate parameter values. Subsequently, many improved methods have been proposed [3,7,24,26,29]. Recently, a novel density based clustering method, named Fast search-and-find of Density Peaks (FDP) [33], was proposed. This algorithm assumes that cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any point with higher density. FDP can recognize clusters regardless of their shape and of the dimensionality of the space in which they are embedded, but it lacks an efficient quantitative criterion for judging cluster centers. Accordingly, approaches such as 3DC [23] and STClu [42] have been proposed to improve FDP.

Density-based clustering methods have the advantages of discovering clusters with arbitrary shapes and dealing with noisy data, but they face two challenges. First, traditional density measures are not adaptive to clusters with different densities. Second, performances of traditional methods (e.g., DBSCAN and FDP) are sensitive to parameters, and it is non-trivial to set these parameters properly for different datasets.

Aiming to address these challenges, we propose the RElative COre MErge (RECOME) clustering algorithm, which is based on two density measures: the K nearest Neighbor Kernel Density (NKD) and Relative K nearest Neighbor Kernel Density (RNKD). RECOME firstly identifies *core objects* corresponding to objects with RNKD equal 1. A core object and its descendants, which are defined by a directed relation (i.e., *higher density nearest-neighbor*) based on the NKD, form an *atom cluster*. These atom clusters are then merged using a novel notion of α-connectivity on a KNN graph. RECOME has been evaluated using both synthetic datasets and real world datasets. Experimental results demonstrate that RECOME outperforms six baseline methods. Furthermore, we find that the clustering results of RECOME can be characterized by a step function of its parameter α, and therefore devise a fast jump discontinuity discovery (FJDD) algorithm to extract the small collection of jump discontinuity values. In summary, this work makes the following contributions.

1. We give a formal analysis showing that the density measure NKD enjoys some desirable properties. Furthermore, based on the NKD, we propose a new density measure RNKD, which is instrumental in detecting clusters with different densities.
2. RECOME can avoid the "decision graph fraud" problem [23] of FDP and can handle clusters with different shapes, densities, and scales. Furthermore, RECOME has nearly linear computational complexity if the K nearest neighbors of each object are computed in advance.
3. FJDD can extract all jump discontinuity values of parameter α for any dataset in $\mathcal{O}(n \log n)$ time, where n is the number of objects. It will greatly benefit parameter selection in real-world applications.

This paper is organized as follows. Section 2 introduces the related work. Section 3 presents the new density measure RNKD and discusses the robustness of NKD and RNKD. Section 4 describes the proposed clustering method RECOME. Section 5 presents the auxiliary algorithm *FJDD*. Section 6 demonstrates experimental results. Finally, we conclude the paper in Section 7.

2. Related work

Existing clustering methods can be categorized into partitional methods, hierarchical methods, grid-based methods, graph-based methods, density-based methods, etc [10]. Partitional methods such as K-means [27] and K-medoids [16], divide data to a number of partitions and a certain quantitative measure of the "goodness" of the resulting clusters is maximized iteratively. Hierarchical clustering methods can be agglomerative (bottom-up) or divisive (top-down). An agglomerative clustering (e.g., AGNES [14]) starts with one object for each cluster and recursively merges two or more of the most appropriate clusters. A divisive clustering (e.g., DIANA [15]) starts with the dataset as one cluster and recursively splits the most appropriate cluster. The process continues until a stopping criterion is reached. Grid-based methods such as STING [43] and CLIQUE [1], divide the original data space into grids, and then group the grids according to the statistical characters of objects in each grid. Graph-based methods, such as SCAN [44] and spectral clustering [37], first construct a similarity graph from a dataset, and then utilize the notion of structural-context similarity or the eigenvalues of Laplacian matrix to generate clusters. Density-based methods (e.g., DBSCAN [9] and DENCLUE [12]) first estimate the distribution density of objects in a feature space, and then recognize clusters as regions of high density separated by regions of lower density. In this paper, we focus on density-based methods because they are highly relevant to the proposed algorithm.

In [9], Ester et al. proposed the first density-based method DBSCAN. In DBSCAN, a *cut-off density* of an object o is defined as the number of objects falling inside a ball of radius ϵ centered at o. If the cut-off density of o is higher than a threshold, *MinPts, o* is regarded as a key object. When the distance between two key objects is less than ϵ, they are called density-reachable. Density-reachable key objects form basic clusters. A non-key object is assigned to a basic cluster if it is within ϵ distance to a key object in the respective cluster; otherwise, the non-key object is treated as noise. DBSCAN is sensitive to the choice of parameters ϵ and *MinPts*, and can hardly handle clusters with heterogeneous densities. To overcome these drawbacks, Ankerst et al. [3] proposed an enhanced density-connected algorithm OPTICS. OPTICS provides a visual tool to help users find the cluster structure and determine the parameters. Although OPTICS reduces the subjectivity in a parameter estimation, when dealing with a complex dataset, it is also difficult to determine how many ϵ's are needed to find potential clusters [7].