Information Sciences 000 (2016) 1–17

Contents lists available at ScienceDirect

Information Sciences

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

Decentralized Clustering by Finding Loose and Distributed Density Cores

Yewang Chen [a,*], Shengyu Tang [a], Lida Zhou [a], Cheng Wang [a], Jixiang Du [a], Tian Wang [a], Songwen Pei [b,c]

[a] College of Computer Science and Technology of Huaqiao University, 668 Jimei Avenue, Xiamen 362021, China
[b] Shanghai Key Laboratory of Modern Optical Systems, University of Shanghai for Science and Technology, Shanghai 200093, China
[c] Parallel Systems and Computer Architecture Lab, University of California, Irvine, CA 92697, USA

ARTICLE INFO

Article history:
Received 5 February 2016
Revised 18 July 2016
Accepted 5 August 2016
Available online xxx

Keywords:
Local density peaks
Shape loss
False peaks
False distances
Density cores

ABSTRACT

Centroid-based clustering approaches fail to recognize extremely complex patterns that are non-isotropic. We analyze the underlying causes and find some inherent flaws in these approaches, including *Shape Loss*, *False Distances* and *False Peaks*, which typically cause centroid-based approaches to fail when applied to complex patterns. As an alternative to current methods, we propose a hybrid decentralized approach named DCore, which is based on finding density cores instead of centroids, to overcome these flaws. The underlying idea is that we consider each cluster to have a shrunken density core that roughly retains the shape of the cluster. Each such core consists of a set of loosely connected local density peaks of higher density than their surroundings. Borders, edges and outliers are distributed around the outsides of these cores in a hierarchical structure. Experiments demonstrate that the promise of DCore lies in its power to recognize extremely complex patterns and its high performance in real applications, for example, image segmentation and face clustering, regardless of the dimensionality of the space in which the data are embedded.

1. Introduction

Cluster analysis is the formal study of the grouping or clustering of objects according to their measured or perceived intrinsic characteristics or similarity. The goal is to discover the natural grouping(s) of a set of patterns, points, or objects. Thousands of clustering algorithms have been published, and more will continue to appear. The two most popular paradigms are the centroid-based method represented by the k-means algorithm [35] and the density-based method represented by DBSCAN [13].

The *k*-means algorithm is still widely used despite having been proposed over 50 years ago. In the *k*-means, *k*-medioids [32], and *k*-center [23] algorithms, clusters are defined as groups of data characterized by a small distance to the cluster center. An objective function, typically the sum of the distances to a set of potential cluster centers, is optimized [14,24,47] until the best cluster center candidates are found. A data point is always assigned to its nearest cluster center. However, these approaches are not able to detect non-spherical clusters. DBSCAN [13] is designed to discover clusters of arbitrary shape

* Corresponding author.
E-mail addresses: ywchen@hqu.edu.cn (Y. Chen), swpei@usst.edu.cn (S. Pei).

http://dx.doi.org/10.1016/j.ins.2016.08.009
0020-0255/

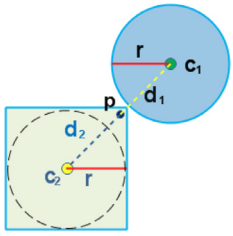

Fig. 1. Two clusters with different shapes: a square and a circle. c_1 and c_2 are the centroids of cluster 1 and cluster 2, respectively; d_1 and d_2 are the distances from p to c_1 and c_2, respectively.

with a fixed scanning radius *eps* and a density threshold *MinPts*. However, this algorithm can be rendered nearly useless when applied to high-dimensional data because of the so-called "curse of dimensionality". Many researchers have proposed various techniques based on DBSCAN in attempts to improve its performance, such as Fast-DBSCAN [22] and others [3,45]. In addition to these, many other important density-based approaches also exist, including [5,34,44,46].

DPeak [42] is based on the concept that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from other points with higher densities. This method has the ability to intuitively identify the number of clusters; outliers are automatically recognized and excluded from the analysis. Although the authors claim that clusters are recognized regardless of their shape and the dimensionality of the space in which they are embedded, there are still some complex shapes that the method cannot recognize. Essentially, DPeak is also a centroid-based method because each density peak defined in this method is still a type of centroid, although not the centroid of a sphere.

Lu and Fu [31] proposed an iterative procedure called the sentence-to-sentence clustering procedure, in which proximity plays a key role; each unlabeled pattern is assigned to the cluster of its nearest labeled neighbor pattern, provided that the distance to that labeled neighbor is below a certain threshold. The process continues until all patterns are labeled or no additional labeling operations can occur. However, this method has poor noise immunity. In mean shift clustering [7], a cluster is defined as a set of points that converge to the same local maximum of the density distribution function. This method is able to find non-spherical clusters but is only applicable to data defined with respect to a set of coordinates.

In some studies, such as [10], [11] and [37], the authors have suggested representing a cluster of points by a set of distant points in the cluster, namely, a boundary. However, capturing the boundary of a cluster with high dimensionality is a non-trivial problem, and the number of points that must be used to represent a cluster increases as the complexity of its shape increases [27]. AUTOCLUST [12] automatically extracts boundaries based on Voronoi modeling [19] and Delaunay diagrams [28]; however, it is only suitable for two-dimensional data.

In recent years, many works have focused on clustering for large-scale data of high dimensionality. Ewa Nowakowska [38] used a dimensionality reduction technique to preserve certain characteristics of data, with a focus on data originating from a mixture of Gaussian distributions, and proposed a clustering method for data of an unknown cluster structure. Jin-Yin Chen [6] proposed a fast density-based data stream clustering algorithm with self-determined cluster centers for mixed data and an efficient distance evaluation method. Junhao Gan and Yufei Tao [21] proposed a novel method named ρ-approximate DBSCAN; based on extending the grid technics used in Fast-DBSCAN and a new tree-structure, ρ-approximate DBSCAN has a computation time that scales only linearly with n.

In addition to all of the methods introduced above, many other clustering approaches have emerged in the past 2 years. For example, Ferrari [15] proposed new ways to obtain meta-knowledge for clustering tasks, Huang [25] proposed a new k-means-type smooth subspace clustering algorithm called Time Series k-means (TSkmeans) for the clustering of time-series data, Ozturk [39] proposed an improved binary artificial bee colony algorithm (IDisABC) for dynamic clustering, Peralta [40] proposed a variation of a non-parametric Bayesian modeling approach for supervised clustering, Gagolewski [17] proposed a new hierarchical clustering linkage criterion called Genie, and Xie [48] proposed a robust clustering algorithm to overcome the deficiencies of DPeak [42].

In summary, centroid-based clustering and other methods typically fail when the patterns of interest are complex or contain high-dimensional data. In this paper, we analyze the underlying causes of their failure and propose a decentralized clustering method, named DCore, to solve these problems.

2. The inherent flaws of centroid-based methods

As mentioned above, centroid-based methods do not perform well on complex patterns. To find the underlying causes of this poor performance, we analyzed various cases and discovered the existence of several inherent flaws, which we call *centralized flaws*, as follows.

(1) ***Shape loss***: Consider the example shown in Fig. 1, in which there are two clusters with different shapes. Cluster 1 is a circle with radius r; c_1 is its centroid. Cluster 2 is a square with side length $2r$, and c_2 is its centroid. In centroid-based methods, c_1 and c_2 are used to represent cluster 1 and cluster 2, respectively. However, the centroid itself does