Information Sciences 450 (2018) 200–226

Contents lists available at ScienceDirect

Information Sciences

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

Shared-nearest-neighbor-based clustering by fast search and find of density peaks☆

Rui Liu[a], Hong Wang[a,b,c,*], Xiaomei Yu[a,b,c]

[a] School of Information Science and Engineering, Shandong Normal University, Jinan, Shandong 250358, China
[b] Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan, Shandong 250014, China
[c] Institute of Life Science, Shandong Normal University, Jinan, Shandong 250014, China

ARTICLE INFO

Article history:
Received 10 October 2017
Revised 31 January 2018
Accepted 11 March 2018
Available online 20 March 2018

Keywords:
Clustering
Shared nearest neighbor
Density peaks
Fast search clustering
Local density

ABSTRACT

Clustering by fast search and find of density peaks (DPC) is a new clustering method that was reported in Science in June 2014. This clustering algorithm is based on the assumption that cluster centers have high local densities and are generally far from each other. With a decision graph, cluster centers can be easily located. However, this approach suffers from certain disadvantages. First, the definition of the local density and distance measurement is too simple; therefore, the DPC algorithm might perform poorly on complex datasets that are of multiple scales, cross-winding, of various densities, or of high dimensionality. Second, the one-step allocation strategy is not robust and has poor fault tolerance. Thus, if a point is assigned incorrectly, then the subsequent allocation will further amplify the error, resulting in more errors, which will have a severe negative impact on the clustering results. Third, the cutoff distance d_c is generally difficult to determine since the range of each attribute is unknown in most cases. Even when being normalized or using the relative percentage method, a small change in d_c will still cause a conspicuous fluctuation in the result, and this is especially true for real-world datasets. Considering these drawbacks, we propose a shared-nearest-neighbor-based clustering by fast search and find of density peaks (SNN-DPC) algorithm. We present three new definitions: SNN similarity, local density ρ and distance from the nearest larger density point δ. These definitions take the information of the nearest neighbors and the shared neighbors into account, and they can self-adapt to the local surroundings. Then, we introduce our two-step allocation method: inevitably subordinate and possibly subordinate. The former quickly and accurately recognizes and allocates the points that certainly belong to one cluster by counting the number of shared neighbors between two points. The latter assigns the remaining points by finding the clusters to which more neighbors belong. The algorithm is benchmarked on publicly available synthetic datasets, UCI real-world datasets and the Olivetti Faces dataset, which are often used to test the performance of clustering algorithms. We compared the results with those of DPC, fuzzy weighted K-nearest neighbors density peak clustering (FKNN-DPC), affinity propagation (AP), ordering points to identify the clustering structure (OP-

☆ The source code of this paper is available at https://github.com/liurui39660/SnnDpc

* Corresponding author at: School of Information Science and Engineering, Shandong Normal University, Jinan, Shandong 250358, China.
E-mail addresses: xxliuruiabc@163.com (R. Liu), wanghong106@163.com (H. Wang).

https://doi.org/10.1016/j.ins.2018.03.031

TICS), density-based spatial clustering of applications with noise (DBSCAN), and K-means. The metrics used are adjusted mutual information (AMI), adjusted Rand index (ARI), and Fowlkes–Mallows index (FMI). The experimental results prove that our method can recognize clusters regardless of their size, shape, and dimensions; is robust to noise; and is remarkably superior to DPC, FKNN-DPC, AP, OPTICS, DBSCAN, and K-means.

1. Introduction

Clustering, also known as unsupervised classification, divides objects into subsets or clusters according to the similarity measure of the object (physical or abstract) such that the objects within the cluster have a high degree of similarity and that the objects belonging to different clusters have a high degree of dissimilarity [36]. Cluster analysis plays an important role in the fields of social sciences, psychology, biology, statistics, pattern recognition and information retrieval as an important basis for other problems.

Cluster analysis is a challenging problem in data mining and machine learning. Over the past few decades, a number of clustering algorithms have been developed for different types of applications. Typical algorithms include K-means [25] and K-medoids [23] based on partitioning, CURE [18] and BIRCH [45] based on hierarchy, DBSCAN [11] and OPTICS [2] based on density, WaveCluster [30] and STING [41] based on grids, statistical clustering [8] based on models, and spectral clustering [39] based on graph theory. In recent years, with the advancements in cluster analysis, some new clustering methods, such as subspace clustering [1], ensemble clustering [35], and deep embedded clustering [43], have been proposed. The performances of these algorithms are different. The classical K-means clustering algorithm achieves good clustering results on datasets with convex spherical structures. Although DBSCAN provides good clustering results on irregular clusters and coiled clusters and has a strong anti-noise capability, for variable-density clusters and high-dimensional data, the clustering result is poor [36,42]. Moreover, selecting the radius and threshold also represents a difficult problem for DBSCAN.

In June 2014, Rodriguez et al. reported the DPC algorithm (clustering by fast search and find of density peaks) [28] in the well-known scientific journal Science. DPC is a new clustering algorithm based on density and distance. Compared with traditional clustering algorithms, the DPC algorithm has many advantages, including the following:

1. The algorithm is simple and efficient, and it can quickly find the high density peak point (cluster center) without iteratively calculating the objective function.
2. The DPC algorithm is suitable for cluster analysis on large-scale data.

Because of the above advantages, in a short period of three years, the DPC algorithm has become widely used in computer vision [32], image recognition [7], text mining [46] and other fields.

Although the DPC algorithm has obvious advantages over other clustering algorithms, it has the following shortcomings:

1. The definition of the local density and distance measurement is too simple; therefore, the clustering result of the DPC algorithm might be poor when working with complex datasets that are of multiple scales, cross-winding, of various densities, or of high dimensions.
2. The allocation strategy is sensitive and has poor fault tolerance. Thus, if a point is assigned incorrectly, then the subsequent allocation will further amplify the error, resulting in more errors that will have a serious negative impact on the clustering results.
3. The cutoff distance, d_c, is generally difficult to determine since the range of each attribute is unknown in most cases. Moreover, even if being normalized or using the relative percentage method, a small change in d_c will still cause a conspicuous fluctuation in results.

To solve the above problems, this paper proposes the shared-nearest-neighbor-based clustering by fast search and find of density peaks (SNN-DPC) algorithm. The main innovations of the SNN-DPC algorithm include the following:

1. A similarity measurement based on shared neighbors is proposed. This criterion can be used to calculate the similarity between points according to the shared neighbor information.
2. A local density metric of points based on shared neighbors is proposed. This criterion can be applied not only to simple datasets but also to complex datasets that are of multiple scales, cross-winding, of various densities, or of high dimensions.
3. An adaptive metric of distance from the nearest larger density point is proposed. This metric can be adjusted according to the local density information to ensure that the correct points are more easily chosen as the centroid.
4. A novel and fast density peak clustering algorithm based on the new density and similarity measure is proposed. This algorithm can quickly and accurately find the density peak (center) of each cluster.
5. A two-step point allocation algorithm based on shared neighbors is proposed to improve the probability that the non-central points are correctly allocated and to avoid further errors when a point is incorrectly assigned.