Information Sciences 354 (2016) 19–40

Contents lists available at ScienceDirect

Information Sciences

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

Robust clustering by detecting density peaks and assigning points based on fuzzy weighted *K*-nearest neighbors

Juanying Xie[a,*], Hongchao Gao[a], Weixin Xie[b], Xiaohui Liu[c], Philip W. Grant[d]

[a] School of computer science, Shaanxi Normal University, Xi'an 710062, PR China
[b] School of Information Engineering, National Key Laboratory of ATR, Shenzhen University, Shenzhen 518006, PR China
[c] Department of Computer Science, Brunel University, London UB8 3PH, UK
[d] Department of Computer Science, College of Science, Swansea University, Singleton Park, Swansea SA2 8PP, UK

ARTICLE INFO

Article history:
Received 14 May 2015
Revised 4 February 2016
Accepted 6 March 2016
Available online 12 March 2016

Keywords:
Local density
Density peaks
Clustering
K-nearest neighbors
Fuzzy weighted *K*-nearest neighbors

ABSTRACT

Clustering by fast search and find of Density Peaks (referred to as DPC) was introduced by Alex Rodríguez and Alessandro Laio. The DPC algorithm is based on the idea that cluster centers are characterized by having a higher density than their neighbors and by being at a relatively large distance from points with higher densities. The power of DPC was demonstrated on several test cases. It can intuitively find the number of clusters and can detect and exclude the outliers automatically, while recognizing the clusters regardless of their shape and the dimensions of the space containing them. However, DPC does have some drawbacks to be addressed before it may be widely applied. First, the local density ρ_i of point i is affected by the cutoff distance d_c, and is computed in different ways depending on the size of datasets, which can influence the clustering, especially for small real-world cases. Second, the assignment strategy for the remaining points, after the density peaks (that is the cluster centers) have been found, can create a "Domino Effect", whereby once one point is assigned erroneously, then there may be many more points subsequently misassigned. This is especially the case in real-word datasets where there could exist several clusters of arbitrary shape overlapping each other. To overcome these deficiencies, a robust clustering algorithm is proposed in this paper. To find the density peaks, this algorithm computes the local density ρ_i of point i relative to its *K*-nearest neighbors for any size dataset independent of the cutoff distance d_c, and assigns the remaining points to the most probable clusters using two new point assignment strategies. The first strategy assigns non-outliers by undertaking a breadth first search of the *K*-nearest neighbors of a point starting from cluster centers. The second strategy assigns outliers and the points unassigned by the first assignment procedure using the technique of fuzzy weighted *K*-nearest neighbors. The proposed clustering algorithm is benchmarked on publicly available synthetic and real-world datasets which are commonly used for testing the performance of clustering algorithms. The clustering results of the proposed algorithm are compared not only with that of DPC but also with that of several well known clustering algorithms including Affinity Propagation (AP), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and *K*-means. The benchmarks used are: clustering accuracy (Acc), Adjusted Mutual Information (AMI) and Adjusted Rand Index (ARI). The experimental results demonstrate that our proposed clustering algorithm can find cluster centers, recognize clusters regardless of their shape and dimension of the space in which they are embedded, be unaffected by outliers, and can often outperform DPC, AP, DBSCAN and *K*-means.

* Corresponding author. Tel.: +86 13088965815.
E-mail address: xiejuany@snnu.edu.cn, juanyingxie@gmail.com (J. Xie).

http://dx.doi.org/10.1016/j.ins.2016.03.011
0020-0255/

1. Introduction

Clustering is the process of grouping objects together according to their similarities, so that objects in the same cluster are similar to one another and dissimilar to objects in any other cluster [12,15,18,27,30,33]. Clustering is important for uncovering the inherent, potential and unknown knowledge, principles or rules in the real-world, and has been widely used in scientific and engineering applications [12,15,18,27,30,33]. With the emergence of *big data*, there is an ever increasing interest in clustering algorithms that can automatically understand, process and summarize the data [9,29].

Many different methods of clustering exist including partitioning, hierarchical, density-based, grid-based, or combinations of these [15,33]. One very popular partitioning method is the *K-means* clustering algorithm [18,24]. It is well known that *K*-means is heavily dependent on the initial cluster centers or initial partitions, and is unable to find clusters with arbitrary shape, with clustering easily affected by noise or outliers. Furthermore, the value of K must be pre-specified [15,18,33]. The *Global K-means* algorithm and its variations were proposed to overcome the disadvantages of *K*-means [21,32], and other algorithms developed to remedy the sensitivity of *K*-means to the outliers [4,16,20]. However, because a data point is always assigned to the nearest center, these approaches are not able to detect nonspherical clusters. Clusters with arbitrary shape are easily detected by approaches based on the local density of data points. Density-Based Spatial Clustering of Applications with Noise (*DBSCAN*) [8] is the typical density-based clustering algorithm which can detect clusters with any arbitrary shape only if the density thresholds are specified, such as ϵ the neighborhood radius and *MinPts* the minimum number of points included in the neighborhood with radius ϵ [15,17,33]. However, choosing the appropriate thresholds may be nontrivial. A relative density based *K*-nearest neighbors clustering algorithm was introduced in [22] to remedy these limitations of DBSCAN.

Affinity Propagation (*AP*) [12] is another well-known clustering algorithm which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerge. AP takes as input a real number $s(k, k)$ (where $s(i, k)$ is the similarity indicating how well the data point with index k is suited to be the exemplar for data point i) for each data point k so that data points with larger values of $s(k, k)$ are more likely to be chosen as exemplars. These values are referred to as *preferences*. The number of identified exemplars (number of clusters) is influenced by the values of the input preferences. However, by simultaneously considering all data points as candidate exemplars and gradually identifying clusters by a message-passing procedure, AP is able to identify the exemplars and detect the clusters of a dataset. AP was tested as a simple and efficient clustering algorithm, especially for cases where the number of the identified exemplars is relatively large [31], but it does not work well for clusters with arbitrary shape. A new hierarchical clustering technique based on synchronization was proposed by Huang et al. [17]. It was reported that the clustering algorithm can detect the clusters of any size and shape. However, the drawbacks of hierarchical clustering cannot be avoided.

Recently, a novel clustering algorithm was proposed by Alex Rodríguez and Alessandro Laio [27], based on the assumption that cluster centers are surrounded by the neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density. We refer to this algorithm as *DPC* (Density Peak Clustering) in this paper. It was demonstrated on several test cases that DPC can efficiently find the cluster centers (i.e., the density peaks) and assign the remaining points to their appropriate clusters as well as detect outliers. However, DPC does have some shortcomings. First, the density metric is different for large and small datasets, and the arbitrarily selected cutoff distance d_c can greatly influence the clustering of a small dataset, and furthermore no criterion is given for determining whether a dataset is small or large. Secondly, after the density peaks have been found, the strategy of assigning the remaining points to the same cluster as its nearest neighbor of higher density can cause propagation of errors. Once a point i is assigned to a wrong cluster, then there may be several points with lower density than point i being assigned to incorrect clusters, producing poor clustering. This is specially true for real-world datasets which usually contain highly overlapped clusters of arbitrary shape. Such is the case for the commonly used test dataset Iris, contained in the UCI machine learning repository [1], on which DPC performs poorly.

In order to remedy these problems with DPC, we introduce several modifications to the basic algorithm. Our new algorithm defines the local density ρ_i of point i based on its *K*-nearest neighbors, so that the deficiencies of DPC in defining the local density of a point can be avoided. Two new assignment strategies are proposed based respectively on the idea of *K*-nearest neighbors and fuzzy weighted *K*-nearest neighbors. We refer to our proposed algorithm as *FKNN-DPC* (Fuzzy weighted *K*-Nearest Neighbors Density Peak Clustering).

The new features of our FKNN-DPC are (i) a uniform local density metric is proposed based on the *K*-nearest neighbors, so that the local density ρ_i of point i can be computed by one density metric independent of the size of the dataset, and the density peaks (cluster centers) will be found efficiently and correctly; and (ii) two new strategies for assigning the remaining points to their most likely clusters are introduced. These new strategies are applied consecutively. The points assigned by strategy one comprise the core of the clusters, and the other points assigned by strategy two are the *halo* of the clusters. The core together with halo make up the whole cluster. The power of our FKNN-DPC to find the density peaks and detect the clusters of a dataset will be demonstrated in Section 4 of this paper.

This paper is organized as follows: Section 2 briefly describes the idea of DPC and analyzes its deficiencies by displaying its clustering on a synthetic dataset which is often used to test the performance of a clustering algorithm. Section 3 introduces our robust clustering algorithm and gives a detailed analysis. Section 4 tests our proposed algorithm on several synthetic and real-world datasets, and compares its performance with DPC, AP, DBSCAN and *K*-means in terms of several