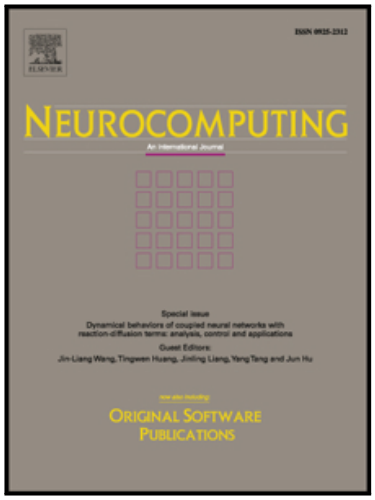
Communicated by Dr. Haijun Zhang

Accepted Manuscript



Clustering by defining and merging candidates of cluster centers via independence and affinity

Wang Gaochao, Wei Yiheng, Tse Peter

PII: S0925-2312(18)30878-6
DOI: https://doi.org/10.1016/j.neucom.2018.07.043
Reference: NEUCOM 19796

To appear in: *Neurocomputing*

Received date: 5 March 2018
Revised date: 10 July 2018
Accepted date: 17 July 2018

Please cite this article as: Wang Gaochao, Wei Yiheng, Tse Peter, Clustering by defining and merging candidates of cluster centers via independence and affinity, *Neurocomputing* (2018), doi: https://doi.org/10.1016/j.neucom.2018.07.043

Clustering by defining and merging candidates of cluster centers via independence and affinity

Wang Gaochao[a], Wei Yiheng[b], and Tse Peter[a,*]

[a] *Department of Systems Engineering and Engineering Management, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China.*

[b] *Department of Automation, University of Science and Technology of China, Hefei, China.*

[*] *Corresponding author.*

E-mail: meptse@cityu.edu.hk

Abstract

Clustering analysis is to classify elements into categories based on their similarity. Clustering by fast search and find of density peaks (CFSFDP) has been proven to be an effective and novel algorithm, which identifies the centers of clusters with density maxima. However, the performance of CFSFDP is quite sensitive to the estimation of densities, that is exactly the selection of the cutoff distance (d_c). In a conventional way, the selection of d_c is based on subjective experience. It meets difficulties in finding an appropriate d_c, especially for detecting nonspherical clusters, because CFSFDP cannot perform well when there are more than one density peak for one cluster. Besides, another barrier of applying CFSFDP is that manual interaction is always required for making an effective selection of cluster centers. In this paper, a new density-based clustering algorithm, clustering by defining and merging candidates of cluster centers via independence and affinity (CDMC-IA), is proposed. With its strategy, an appropriate value of cutoff distance d_c can be well suggested and the robustness of the method itself is enhanced. Moreover, CDMC-IA introduces a new quantity independence to sort and select cluster centers, instead of human based selection from decision graph. Another quantity affinity is also introduced, which well handles multiple density peaks existing in one cluster and is able to assign each data point to the its targeted cluster. The performance of applying conventional clustering methods to benchmark datasets will be compared with the proposed method in this paper.

Keywords:
Clustering, Kernel density estimation, Cutoff distance, Independence, Affinity.

1. Introduction

Clustering analysis plays a critical role in data mining, financial data analysis, engineering signal processing, and data mining [1, 2]. Its applications range from pattern recognition [3-5], image processing [6-9], recommendation [10], and etc. Clustering is to classify data into different categories, or clusters, based on a measure of similarity, while data that are dissimilar are grouped into different clusters [11]. Several different clustering methods categorized into density based, hierarchical, partitioning, model based and grid based have been proposed [12]. Some popular clustering algorithms are K-means, Fuzzy C-means (FCM), Hierarchical clustering, and etc. Density based clustering methods can easily detect the arbitrary shape of clusters in large spatial databases [13, 14]. The density based spatial clustering of applications with noise (DBSCAN) [15], is one of the most popular clustering algorithms. DBSCAN is robust against noise, however not fully deterministic for border points, and the cluster shapes depend upon the parameters input [13]. Recently, an alternative algorithm implementing clustering by fast search and find of density peaks was proposed [14]. CFSFDP is mainly based on these two assumptions: 1) the density of central point of a cluster is higher than its neighbors; 2) the cluster centers are at a relatively large distance from these points having a higher density. Thus, CFSFDP is to firstly find density peaks. Then after, CFSFDP provides a decision graph to the analyzer for the selection of cluster centers. Finally, data points will be assigned to the targeted cluster based on the Euclidean distance.

Despite some clustering analysis platforms, like ClustEval, shows CFSFDP is one of the top-performing methods [16], there are several problems remained.

(i) The first limitation of CFSFDP is that the performance of CFSFDP is highly depending upon the selection of cutoff distance d_c, especially when the dataset is in arbitrary shape. An inappropriate d_c will result in assigning the data point to a wrong cluster. The conventional way suggests that with the desired d_c, the number of neighbors of each data point in a dataset should be 1-2% of entire dataset. Even though, the clustering results with a suggested d_c may still be unsatisfying.