Information Sciences 405 (2017) 1–17

Contents lists available at ScienceDirect

Information Sciences

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

Parameter independent clustering based on dominant sets and cluster merging

Jian Hou[a,*], Weixue Liu[b]

[a] College of Engineering, Bohai University, Jinzhou 121013, China
[b] College of Information Science and Technology, Bohai University, Jinzhou 121013, China

ARTICLE INFO

Article history:
Received 11 April 2016
Revised 1 April 2017
Accepted 5 April 2017
Available online 5 April 2017

Keywords:
Clustering
Dominant sets
Cluster merging
Parameter independent

ABSTRACT

Clustering is an important unsupervised learning approach with wide application in data mining, pattern recognition and intelligent information processing. However, existing clustering algorithms usually involve one or more user-specified parameters as input and their clustering results depend heavily on these parameters. In order to solve this problem, we present a parameter independent clustering algorithm based on the dominant sets algorithm and cluster merging. In the first step histogram equalization transformation is applied to solve the parameter dependence problem of the dominant sets algorithm. We provide the theoretic foundation of this method and discuss the implementation details. The clustering result is then refined with a cluster merging method, which is based on a new clustering quality evaluation criterion. We use extensive experiments on several datasets to validate each step and the whole procedures of our algorithm. It is shown that our parameter independent algorithm performs comparably to some existing clustering algorithms which benefit from user-specified parameters.

1. Introduction

Data clustering refers to the task of grouping objects into clusters so that the data in the same cluster are similar and those in different clusters are dissimilar. The popular clustering algorithms include k-means, DBSCAN [7], Fuzzy C-means [34], BIRCH, EM and CLIQUE [1], and some recent works on clustering include [5,24,25,28,38]. As an important unsupervised learning approach, clustering is widely used in pattern recognition, data mining and intelligent information processing and is potentially useful in other fields [35,36].

In recent decades, graph based clustering has attracted much attention due to its great potential demonstrated in practice. By representing the data relationship with an edge-weighted graph and the corresponding pairwise similarity matrix, graph based clustering algorithms intend to partition the graph to obtain clusters. By making use of the rich data distribution information captured in the pairwise similarity matrix, graph based algorithms have been shown to generate superior results in many applications. As one of the most well-known graph based clustering algorithms, the normalized cuts (NCuts) algorithm [30] has been widely used as a benchmark in data clustering and image segmentation. Spectral clustering utilizes the eigen-structure of the pairwise similarity matrix to perform dimension reduction, and then accomplishes the clustering with a simple algorithm, e.g., k-means, in the new data space of reduced dimension. The affinity propagation (AP) algorithm

* Corresponding author.
E-mail address: dr.houjian@gmail.com (J. Hou).

http://dx.doi.org/10.1016/j.ins.2017.04.006
0020-0255/

[3] passes the affinity messages among input data iteratively and finds out the cluster centers and members gradually. The AP algorithm has obtained successful application in human face clustering and gene detection, etc. Another graph based algorithm worth mentioning is the dominant sets (DSets) algorithm [27]. The DSets algorithm defines a dominant set as a graph-theoretic concept of a cluster and extracts the clusters (dominant sets) in a sequential manner. The DSets algorithm has been shown to be effective in various tasks including image segmentation [12,14], object detection [33], human activity analysis [10] and object classification [13,15], etc.

From the review above we see that numerous clustering algorithms have been proposed, and some of them show impressive performance in clustering tasks. However, all of these algorithms require one or more input parameters explicitly or implicitly, and their clustering results usually depend heavily on the parameters. The k-means algorithm must be fed with the number of clusters, which is not easy to determine in many cases. As spectral clustering algorithms usually adopt k-means as a step, these algorithms also require the number of clusters to be determined beforehand. While DBSCAN and AP are able to determine the number of clusters by themselves, DBSCAN requires as input a neighborhood radius and the minimum number of data in the neighborhood, and AP requires the preference values of the data to be specified. In both cases the variance of input parameters has a significant influence on the clustering results. Although the DSets algorithm uses only the pairwise similarity matrix as input and no parameters are involved explicitly, parameters may be introduced in the case that the data for clustering are represented as feature vectors. Specifically, a commonly used similarity measure of two data items x and y is $s(x, y) = exp(-d(x, y)/\sigma)$, where $d(x, y)$ is the Euclidean distance and σ is a regularization parameter. With the same set of data, different σ's lead to different similarity matrices, which are found to result in different DSets clustering results. With these parameter dependent algorithms, we need a careful parameter tuning process in order to obtain satisfactory clustering results. This makes these algorithms less attractive in practical applications.

While the majority of existing clustering algorithms are parameter dependent, the efforts to achieve parameter independence can be traced back to [19]. In this paper we present a parameter independent clustering algorithm on the basis of the DSets algorithm and cluster merging. In applying the DSets algorithm to cluster the data represented as feature vectors, the parameter σ influences the pairwise similarity matrix directly, and then the clustering results indirectly. In [11] the authors propose to transform the similarity matrix with histogram equalization, so that the new similarity matrix is no longer influenced by σ. With this transformation, the DSets algorithm generates almost identical clustering results with different σ's. However, this transformation is also found to result in over-small clusters. This problem is solved in [11] by expanding the clusters, where the expansion method involves user-specified parameters. In this paper we solve the small-cluster problem with a different approach which is independent of parameters. Specifically, we merge the over-small clusters to increase cluster size, and the cluster merging method is based on the relationship between intra-cluster and inter-cluster similarities. By making use of the nice properties of the DSets algorithm, this parameter independent method is shown to solve the small-cluster problem effectively and improve the clustering quality evidently.

A preliminary version of some works in this paper appear in [17]. Compared with [17], the contributions of this paper are as follows. First, we discuss in theory why histogram equalization transformation of similarity matrices is able to eliminate the influence of σ's on DSets clustering results. Second, we show that in our algorithm, the similarity matrices obtained by histogram equalization transformation perform no worse than those obtained from fixed σ's. Third, we present an internal criterion to evaluate the clustering quality, which is shown to perform better than the existing Davies–Bouldin index and the Dunn index. In addition, we validate both the major steps and the whole procedure of our algorithm with extensive experiments, which make our conclusions more convincing.

The rest of this paper is organized as follows. The concept and properties of the DSets algorithm are presented briefly in Section 2. Then we analyze the problems of the DSets algorithm and present our cluster merging method in Section 3. Section 4 reports the experimental results of the proposed algorithm and finally Section 6 concludes this paper.

2. Dominant sets

Considering that our algorithm is based in part on the DSets algorithm, in this section we introduce the concept of dominant set and the properties of this algorithm briefly. More details of this algorithm can be found in [26,27].

As dominant set is a graph-theoretic concept of a cluster, we represent the n data to be clustered with an undirected edge-weighted graph $G = (V, E, w)$ without self-loops, where V is the vertex set containing all the data, E denotes the edge set consisting of the adjacency relationship among the data, and w is the weight function measuring how closely the data are related. If we represent the pairwise $n \times n$ similarity matrix as $A = (a_{ij})$, we see that $a_{ij} = w(i, j)$ if $(i, j) \in E$ and $a_{i,j} = 0$ otherwise.

The definition of dominant set is presented as follows. With a non-empty subset $D \subseteq V$ and two data $p \in D$, $q \notin D$, we define

$$\phi_D(p, q) = a_{pq} - \frac{1}{n}\sum_{k\in D} a_{pk}, \tag{1}$$

and then

$$w_D(p) = \begin{cases} 1, & \text{if} \quad |D| = 1, \\ \sum\limits_{k\in D\setminus\{p\}} \phi_{D\setminus\{p\}}(k, p) w_{D\setminus\{p\}}(k), & \text{otherwise.} \end{cases} \tag{2}$$