Pattern Recognition 60 (2016) 25–36

Contents lists available at ScienceDirect

Pattern Recognition

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

Towards parameter-independent data clustering and image segmentation

Jian Hou [a,*], Weixue Liu [b], Xu E [b], Hongxia Cui [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 21 October 2015
Received in revised form
17 March 2016
Accepted 28 April 2016
Available online 17 May 2016

Keywords:
Dominant sets
Clustering
Image segmentation
Similarity matrix
Similarity measure

ABSTRACT

While there are a large amount of clustering algorithms proposed in the literature, the clustering results of existing algorithms usually depend on user-specified parameters heavily, and it is usually difficult to determine the optimal parameters. With the pairwise data similarity matrix as the input, dominant sets clustering has been shown to be an effective data clustering and image segmentation approach, partly due to its ability to find out the underlying data structure and determine the number of clusters automatically. However, we find that the original dominant sets algorithm is sensitive to the similarity measures used in building the similarity matrix. This means that parameter tuning is required to generate satisfactory clustering results, and dominant sets clustering results are also parameter dependent. In order to remove the dependence on the user-specified parameter, we study how the similarity measures influence the dominant sets clustering results. As a result, we propose to transform similarity matrices by histogram equalization before clustering. While this transformation is shown to remove the sensitiveness to similarity measures effectively, it also results in over-segmentation. Therefore in the next step we present a cluster extension method to overcome the over-segmentation effect and generate more reasonable clustering results. We test the enhanced clustering algorithm in both data clustering and image segmentation experiments, and comparisons with the state-of-the-art algorithms validate the effectiveness of our algorithm.

1. Introduction

Clustering is widely used in various fields, including pattern recognition and image processing, etc. There have been a large amount of clustering algorithms in the literature, and some traditional clustering algorithms include *k*-means, BIRCH, DBSCAN [1], EM (Expectation Maximization) and CLIQUE [2]. In recent years, spectral clustering, e.g., normalized cuts (NCuts) [3], receives a lot of attention. Spectral clustering groups data into clusters based on the eigen-structure of the pairwise data similarity matrix. Another popular clustering algorithm is the so-called affinity propagation [4], which finds out the cluster members as the result of affinity messages passing among input data. In [5] the authors present a density peak based method, which uses the local density and minimum distance to higher density of each data to identify cluster centers and then assign other data to clusters. Some other important progresses in clustering and its application in image segmentation include [6–13].

Although there are a lot of clustering algorithms proposed from different perspectives and designed for different purposes, existing algorithms usually require user-specified parameters as input, and their clustering performance depends heavily on these parameters. One typical example of such parameters is the number of clusters, which is required as input by many clustering algorithms, e.g., k-means and NCuts. Some other algorithms, e.g., DBSCAN and affinity propagation, are able to determine the number of clusters by themselves. However, they require other parameters as input. In fact, DBSCAN needs to be given a neighborhood radius and the minimum number of data in this neighborhood, and affinity propagation requires the preference values of all data to be specified. While the density peak based algorithm proposed in [5] generates impressive clustering results on some datasets, its clustering performance depends on the cutoff distance and some other parameters evidently.

Some methods have been proposed to determine the above-mentioned parameters. For example, the authors of [14–17] present some methods to determine the appropriate number of clusters, and the authors of [4] published a method to calculate the

[*] Corresponding author.
E-mail addresses: dr.houjian@gmail.com (J. Hou), bhulwx@qq.com (W. Liu), exu21@163.com (X. E), cuihongxia@bhu.edu.cn (H. Cui).

http://dx.doi.org/10.1016/j.patcog.2016.04.015
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

range of preference values of data. In addition, correlation clustering [18,19] has been proposed to be a parameter-independent clustering algorithm. However, the automatic determination of parameters in clustering algorithms is still an open problem in general. In existing works these parameters are often determined empirically.

Noticing that it is usually not a trivial task to determine the appropriate parameters for existing clustering algorithms, in this paper we aim to explore a parameter-independent clustering approach. Our work is partly based on the study of the dominant sets (DSets) clustering [20,21] algorithm. As a graph-theoretic concept of a cluster, a dominant set is defined as a subset of data with high internal similarity and low external one. This definition enables a dominant set to be regarded as a cluster. DSets clustering accomplishes the clustering process by extracting dominant sets sequentially and determines the number of clusters automatically. Since its proposal, DSets clustering has been successfully applied in various applications, including soft clustering [22], image segmentation [21], human activity analysis [23], bioinformatics [24], object detection [25] and image classification [26], etc.

However, we have found that the original DSets clustering algorithm is sensitive to the similarity measures used in building the similarity matrix as input. Usually when the data to be clustered are represented as points in a vector space, we need to build the similarity matrix with each entry in the form of $s(i,j)=\exp(-d(i,j)/\sigma)$, where $d(i,j)$ is the distance between two data i and j, and σ is a regulation parameter. From this similarity representation we see that given a set of data to be clustered, different σ results in different similarity matrices, and this in turn leads DSets clustering to generate different clustering results. This sensitiveness to similarity measures means that we need an appropriate σ to generate satisfactory clustering results. In other words, although DSets clustering does not explicitly involve any parameter, it is also parameter dependent implicitly. Although it is possible to build non-parametric similarity, e.g., cosine or histogram intersection, it is shown in [27] that these measures are much inferior to distance based measures in DSets clustering potentially. Therefore in this paper we build the similarity matrix based on $s(i,j)=\exp(-d(i,j)/\sigma)$.

In order to make the DSets clustering results independent of the parameter σ, we investigate the differences in similarity matrices generated from different σ's. By regarding a similarity matrix as an intensity image, we found that similarity matrices corresponding to different σ's can be regarded as images with different intensity contrasts. In image enhancement, images of different intensity contrasts can be transformed to be of approximately the same contrasts and appearances. This observation indicates that if we transform the similarity matrices from different σ's with some image enhancement technique, it is possible to remove the differences in these similarity matrices, and then remove the differences in clustering results. In implementation we adopt histogram equalization to accomplish the transformation, and show that this transformation can be used to make similarity matrices invariant to σ's. However, we also find that this transformation results in over-segmentation in clustering results. Therefore in the next step we do cluster extension to overcome the effect of over-segmentation. Experiments on both data clustering and image segmentation validate the effectiveness of the enhanced DSets clustering algorithm. A preliminary version of some works in this paper appeared in [28].

The remainder of this paper is organized as follows. In Section 2 we introduce the definition of dominant set and discuss its properties and problems. Section 3 details our approach to transform similarity matrices by histogram equalization, and then Section 4 presents a cluster extension method to solve the over-segmentation problem. The data clustering and image segmentation results with our algorithm and some others are reported in Section 5. Finally, Section 6 concludes this paper.

2. Dominant sets clustering

Traditional clustering algorithms usually accomplish the clustering process by partitioning the given set of data, and each part is regarded as a cluster. In other words, the clusters are obtained as the outcome of a partitioning process simultaneously. In contrast, with the DSets clustering algorithm the clusters are extracted from the input data in a sequential manner. In this part we introduce the definition of dominant set and the DSets clustering algorithm briefly, and refer the interested reader to [20,21,29] for more details.

Firstly, the n data items to be clustered are represented as an undirected, edge-weighted graph $G=(V,E,w)$ without self-loops. Here V denotes the set of nodes and corresponds to the n data items, E reflects the pairwise adjacency relationship among data items, and w is the weight function measuring the similarity between data. We then represent the graph G by its corresponding pairwise $n\times n$ similarity matrix $A=(a_{ij})$, where $a_{ij}=w(i,j)$ if $(i,j)\in E$ and $a_{i,j}=0$ otherwise. As no self-loops exist in the graph, all entries on the main diagonal of A are zero.

Informally, a dominant set can be viewed as a maximal subset with internal coherency. More specifically, a dominant set is such a subset of data that the inside data are similar to each other, and they are dissimilar to the outside data. This property satisfies the basic similarity constraint imposed on a cluster, and enables us to regard a dominant set as a cluster. By extracting dominant sets sequentially, we are able to accomplish the clustering where each cluster corresponds to a dominant set.

In the following we present the formal definition of dominant set. With a non-empty subset $D\subseteq V$, $i\in D$ and $j\notin D$, we define

$$aw_D(i)=\frac{1}{|D|}\sum_{k\in D}a_{ik}. \tag{1}$$

where $|D|$ denotes the size of D

$$\phi_D(i,j)=a_{ij}-aw_D(i). \tag{2}$$

The weight of data i with respect to D is then defined as

$$w_D(i)=\begin{cases}1, & \text{if } |D|=1,\\ \sum_{j\in D\setminus\{i\}}\phi_{D\setminus\{i\}}(j,i)w_{D\setminus\{i\}}(j), & \text{otherwise.}\end{cases} \tag{3}$$

Since $w_D(i)$ is defined in a recursive form and its meaning is not evident, we explain what it represents approximately as follows. From Eq. (2) we see that the parameter $\phi_D(i,j)$ reflects the comparison of two similarities, i.e., the similarity between j and i, and the average similarity between i and all the data in D. Evidently $\phi_D(i,j)>0$ means that i is more closely related with j than with its neighbors in D. Although $w_D(i)$ is defined in a recursive form in Eq. (3), we note that it is calculated totally based on $\phi_{D\setminus\{i\}}(j,i)$, and here j refers to every data in $D\setminus\{i\}$. In other words, we need to calculate the two similarities a_{ji} and $aw_{D\setminus\{i\}}(j)$ for each j in $D\setminus\{i\}$. Since Eq. (3) shows that $w_D(i)$ is calculated in the form of a weighted sum of these $\phi_{D\setminus\{i\}}(j,i)$, we can *approximately* regard $w_D(i)$ as a comparison of the sums of the two similarities, i.e.,

$$\tilde{\delta}(i,D\setminus\{i\})=\sum_{j\in D\setminus\{i\}}a_{ji} \tag{4}$$

and

$$\tilde{\delta}(D\setminus\{i\})=\sum_{j\in D\setminus\{i\}}aw_{D\setminus\{i\}}(j)=\frac{1}{|D\setminus\{i\}|}\sum_{j\in D\setminus\{i\}}\ \sum_{k\in D\setminus\{i\}}a_{jk} \tag{5}$$

If we denote $D\setminus\{i\}$ by S, and divide Eqs. (4) and (5) by $|S|$, we