Pattern Recognition Letters 29 (2008) 2067–2077

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

DIVFRP: An automatic divisive hierarchical clustering method based on the furthest reference points

Caiming Zhong [a,b,*], Duoqian Miao [a,c], Ruizhi Wang [a,c], Xinmin Zhou [a]

[a] School of Electronics and Information Engineering, Tongji University, Shanghai 201804, PR China
[b] College of Science and Technology, Ningbo University, Ningbo 315211, PR China
[c] Tongji Branch, National Engineering and Technology Center of High Performance Computer, Shanghai 201804, PR China

ARTICLE INFO

Article history:
Received 25 September 2007
Received in revised form 21 May 2008
Available online 22 July 2008

Communicated by L. Heutte

Keywords:
Divisive clustering
Automatic clustering
Furthest reference point
Dissimilarity measure
Peak
Spurious cluster

ABSTRACT

Although many clustering methods have been presented in the literature, most of them suffer from some drawbacks such as the requirement of user-specified parameters and being sensitive to outliers. For general divisive hierarchical clustering methods, an obstacle to practical use is the expensive computation. In this paper, we propose an automatic divisive hierarchical clustering method (DIVFRP). Its basic idea is to bipartition clusters repeatedly with a novel dissimilarity measure based on furthest reference points. A sliding average of sum-of-error is employed to estimate the cluster number preliminarily, and the optimum number of clusters is achieved after spurious clusters identified. The method does not require any user-specified parameter, even any cluster validity index. Furthermore it is robust to outliers, and the computational cost of its partition process is lower than that of general divisive clustering methods. Numerical experimental results on both synthetic and real data sets show the performances of DIVFRP.

1. Introduction

Clustering is an unsupervised classification technique in pattern analysis (Jain et al., 1999). It is defined to divide a data set into clusters without any prior knowledge. Objects in a same cluster are more similar to each other than those in different clusters. Many clustering methods have been proposed in the literature (Xu and Wunsch, 2005; Jain et al., 1999). These methods can be roughly classified into following categories: hierarchical, partitional, density-based, grid-based and model-based methods. However, the first two methods are the most significant algorithms in clustering communities. The hierarchical clustering methods can be further classified into agglomerative methods and divisive methods. Agglomerative methods start with each object as a cluster, recursively take two clusters with the most similarity and merge them into one cluster. Divisive methods, proceeding in the opposite way, start with all objects as one cluster, at each step select a cluster with a certain criterion (Savaresi et al., 2002) and bipartition the cluster with a dissimilarity measure.

In general, partitional clustering methods work efficiently, but the clustering qualities are not as good as those of hierarchical methods. The *K*-means (MacQueen, 1967) clustering algorithm is one of well-known partitional approaches. Its time complexity is $O(NKId)$, where N is the number of objects, K is the number of clusters, I is the number of iterations required for convergence, and d is the dimensionality of the input space. In practice, K and d are usually far less than N, it runs in linear time on low-dimensional data. Even though it is computationally efficient and conceptually simple, *K*-means has some drawbacks, such as no guarantee of convergence to the global minimum, the requirement of the number of clusters as an input parameter provided by users, and sensitivity to outliers and noise. To remedy these drawbacks, some variants of *K*-means have been proposed: PAM (Kaufman and Rousseeuw, 1990), CLARA (Kaufman and Rousseeuw, 1990), and CLARANS (Ng and Han, 1994).

To the contrary, hierarchical clustering methods can achieve good clustering results, but only at the cost of intensive computation. Algorithm single-linkage is a classical agglomerative method with time complexity of $O(N^2 \log N)$. Although algorithm CURE (Guha et al., 1998), one improved variant of single-linkage, can produce good clustering quality, the worst-case time complexity of CURE is $O(N^2 \log_2 N)$. Compared to agglomerative methods, divisive methods are more computationally intensive. For bipartitioning a cluster C_i with n_i objects, a divisive method will produce a global optimal result if all possible $2^{n_i-1}-1$ bipartitions are considered. But clearly, the computational cost of the complete enumeration is prohibitive. This is the very reason why divisive methods are seldom applied in practice. Some improved divisive

* Corresponding author. Address: School of Electronics and Information Engineering, Tongji University, Shanghai 201804, PR China. Tel.: +86 21 69589867; fax: +86 21 69589359.
E-mail addresses: zhongcaiming@nbu.edu.cn, charman_zhong@hotmail.com (C. Zhong).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.07.002

methods do not consider unreasonable bipartitions identified by a pre-defined criterion in order to reduce the computational cost (Gowda and Ravi, 1995). Chavent et al. (2007) in a monothetic divisive algorithm use a monothetic approach to reduce the number of admissible bipartitions.

Most traditional clustering methods, such as *K*-means, DBScan (Ester et al., 1996), require some user-specified parameters. Generally, however, the required parameters are unknown to users. Therefore, automatic clustering methods are expected in practical applications. Some clustering methods of this kind have been presented in the literature (Wang et al., 2007; Tseng and Kao, 2005; Garai and Chaudhuri, 2004; Bandyopadhyay and Maulik, 2001; Tseng and Yang, 2001). Roughly these methods can be categorized into two groups: clustering validity index-based methods (Wang et al., 2007; Tseng and Kao, 2005) and genetic scheme-based methods (Garai and Chaudhuri, 2004; Bandyopadhyay and Maulik, 2001; Tseng and Yang, 2001). Wang et al. (2007) iteratively apply the local shrinking-based clustering method with different cluster number *K*s. In the light of CH index and Silhouette index, the qualities of all clustering results are measured. The optimal clustering result with the best cluster quality is selected. Tseng and Kao (2005) use Hubert's Γ index to measure a cluster strength after each adding (or removing) of objects to (or from) the cluster. For genetic scheme-based clustering methods, it is crucial to define a reasonable fitness function. Bandyopadhyay and Maulik (2001) take some validity indices as fitness functions directly. In the methods of Garai and Chaudhuri (2004) and Tseng and Yang (2001), although validity indices are not used directly, the fitness functions are very close to validity indices essentially. So genetic scheme-based methods, in different extents, are dependent on the clustering validity indices. However, clustering validity indices are not a panacea since an index that can deal with different shapes and densities is not available.

Robustness to outliers is an important property for clustering algorithms. Clustering algorithms that are vulnerable to outliers (Patan and Russo, 2002) may use some outlier detection mechanisms (Aggarwal and Yu, 2001; Ramaswamy et al., 2000; Breunig et al., 2000; Knorr and Ng, 1998) to eliminate the outliers in data sets before clustering proceeds. However, since this is an extra task, users prefer to clustering algorithms robust to outliers.

In this paper, we propose an efficient divisive hierarchical clustering algorithm with a novel dissimilarity measure (DIVFRP). Based on the furthest reference points, the dissimilarity measure makes the partition process robust to outliers and reduces the computational cost of partitioning a cluster C_i to $O(n_i \log n_i)$. After a data set being partitioned completely, the algorithm employs a sliding average of differences between neighboring pairs of sum-of-errors to detect potential peaks and determine the candidates of the cluster number. Finally, spurious clusters are removed and the optimal cluster number *K* is achieved. Our experiments demonstrate these performances. The remaining sections are organized as follows: algorithm DIVFRP is presented in Section 2. Section 3 presents experimental results. The performances are studied in Section 4. Section 5 concludes the paper.

2. The clustering algorithm

We begin our discussion of the clustering algorithm DIVFRP by considering the concept of general clustering algorithm.

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_N\}$ be a data set, where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{id})^T \in \Re^d$ is a feature vector, and x_{ij} is a feature. A general clustering algorithm attempts to partition the data set $\boldsymbol{X}$ into *K* clusters: $C_0, C_1, \ldots, C_{K-1}$ and one outlier C_{outlier} set according to the similarity or dissimilarity measure of objects. Generally, $C_i \neq \emptyset$, $C_i \cap C_j = \emptyset$, $\boldsymbol{X} = C_0 \cup C_1 \cup \ldots \cup C_{K-1} \cup C_{\text{outlier}}$, where $i = 0, 1, \ldots, K-1$, $j = 0, 1, \ldots, K-1$, $i \neq j$.

The algorithm DIVFRP comprises three phases:

1. Partitioning a data set.
2. Detecting the peaks of differences of sum-of-errors.
3. Eliminating spurious clusters.

2.1. Partitioning a data set

2.1.1. The dissimilarity measure based on the furthest reference points

Similarity or dissimilarity measures are essential to a clustering scheme, because the measures determine how to partition a data set. In a divisive clustering method, let C_i be the cluster to be bipartitioned at a step of the partitioning process, $g(C_x, C_y)$ be a dissimilarity function. If the divisive method bipartitions C_i into C_{i1} and C_{i2}, the pair (C_{i1}, C_{i2}) will maximize the dissimilarity function *g* (Theodoridis and Koutroumbas, 2006). According to this definition of dissimilarity, we design our dissimilarity measure as follows.

For a data set consisting of two spherical clusters, our dissimilarity measure is on the basis of the observation: the distances between points in a same cluster and a certain reference point are approximative. We call the distances a representative. For the two clusters, two representatives exist with respect to a same reference point. Assume that there exits a point on the line that passes through the two cluster mean points, and both clusters are on the same side of the point. Taking the point as the reference point, one will get the maximum value of the difference between the two representatives. On the contrary, if the reference point is on the perpendicular bisector of the line segment that ends at the two cluster mean points, one will get the minimum value. However, it is difficult to get the ideal reference point since the cluster structure is unknown. We settle for the furthest point from the centroid of the whole data set instead, because it never lies between the two cluster mean points and two clusters must be on the same side of it. Fig. 1 illustrates the dissimilarity measure based the furthest point and how the cluster being split.

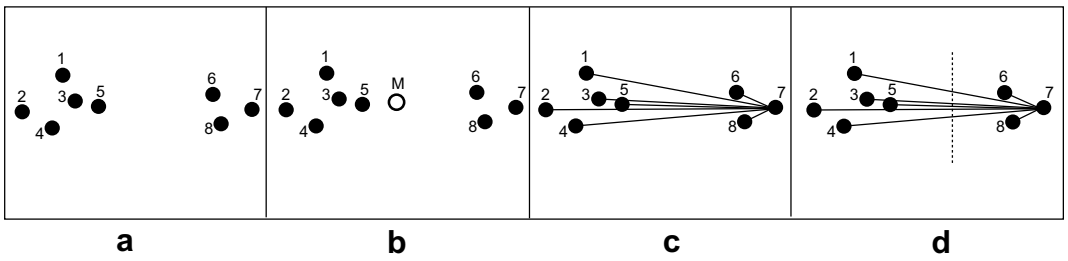

Fig. 1. Illustration of the dissimilarity measure and a split. In (a), a data set with two spherical clusters is shown. In (b), the hollow point *M* is the mean point of the data set; point 7 is the furthest point to the mean and selected as the reference point. In (c), distances from all points including the reference point to the reference are computed. In (d), the neighboring pair $< d_{r6}, d_{r5} >$ with maximum difference between its two elements is selected as the boundary, with which the cluster is split.