Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 286–298
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

CLUES: A non-parametric clustering method based on local shrinking

Xiaogang Wang[a,*], Weiliang Qiu[b], Ruben H. Zamar[c]

[a]*Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ont., Canada M3J 1P3*
[b]*The Channing Laboratory, Harvard Medical School, USA*
[c]*Department of Statistics, University of British Columbia, Canada*

Available online 17 December 2006

Abstract

A novel non-parametric clustering method based on non-parametric local shrinking is proposed. Each data point is transformed in such a way that it moves a specific distance toward a cluster center. The direction and the associated size of each movement are determined by the median of its *K*-nearest neighbors. This process is repeated until a pre-defined convergence criterion is satisfied. The optimal value of the number of neighbors is determined by optimizing some commonly used index functions that measure the strengths of clusters generated by the algorithm. The number of clusters and the final partition are determined automatically without any input parameter except the stopping rule for convergence. Experiments on simulated and real data sets suggest that the proposed algorithm achieves relatively high accuracies when compared with classical clustering algorithms.

Keywords: Automatic clustering; *K*-nearest neighbors; Local shrinking; Number of clusters

1. Introduction

Clustering is the process of partitioning a set of objects into subsets based on some measure of similarity (or dissimilarity) between pairs of objects. Cluster analysis has many applications in data mining where large data sets, such as marketing data, need to be partitioned into much smaller and homogeneous groups. Cluster analysis is also widely used to analyze biological data. For example, given a set of gene expression data, a cluster of genes could suggest either these genes have a similar function in the cell or that they are regulated by the same transcription factor. For many clustering algorithms, such as *K*-means (MacQueen, 1967; Hartigan and Wong, 1979) and PAM (Kaufman and Rousseeuw, 1990), the number of clusters or sub-populations needs to be specified by the user. The determination of the number of clusters is one of the most difficult problems in cluster analysis.

Most of the methods for estimating the number of clusters or sub-populations can be classified into several categories. The first category is to select the number of clusters by optimizing a certain measure of strength of the clusters (Tibshirani et al., 2000). This category embraces various methods of estimating the number of components of mixture of distributions (Fraley and Raftery, 2002). The second category of methods first partitions the data into many small clusters, and then merges these small clusters until no clusters can be merged (Frigui and Krishnapuram, 1999). Another strategy is to

* Corresponding author. Tel.: +1 416 736 2100x33938.
E-mail address: stevenw@mathstat.yorku.ca (X. Wang).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.12.016

extract one cluster at a time (Zhung et al., 1996). Moreover, mode detection or bump hunting methods (Cheng and Hall, 1998; Hall and Heckman, 2000) can also be used to determine the number of clusters.

Recently, a new category for estimating the number of clusters has emerged. The main idea is to first iteratively move data points toward cluster centers and use the number of convergence points as the number of clusters. The key issue here is how to move data points toward their cluster centers. One approach, gravitational clustering (Wright, 1977; Kundu, 1999; Sato, 2000; Wang and Rau, 2001) can be interpreted from the point of view of field theory in physics: each data point is considered as a particle of unit mass with zero velocity which is gradually moving toward the cluster center due to gravitation. Another approach, mean-shift clustering (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 1999, 2000, 2001, 2002) originates from an idea in kernel density estimation: data points are transformed toward denser regions by using kernel functions.

In this paper, we propose an automatic clustering algorithm that determines the number of clusters and the partition without any input parameter except a convergence criterion. Our algorithm also shrinks data points towards cluster centers as in the mean-shift algorithm. This process is repeated until a specified criterion is satisfied. The final partition can then be obtained as if all the data points had converged to the cluster centers. In our algorithm, however, the shrinking process is determined by the K-nearest neighbor approach (Mack and Rosenblatt, 1979) instead of kernel functions. Since our method is based on the K-nearest neighbor approach, it is very adaptive to the local geometrical structure of the sample space and well suited for dealing with high-dimensional sparse samples. This property leads to better clustering especially when the clusters are of irregular shapes.

Since the shrinking in our algorithm is dictated by the K-nearest neighbor approach and the final partition is a function of K, one would question how this value should be determined. To resolve this critical issue of choosing the value for K, our algorithm searches for the value of K that maximizes an index function such as the CH index (Calinski and Harabasz, 1974) or the Silhouette index (Kaufman and Rousseeuw, 1990). To be more specific, our algorithm starts from a small K and gradually increases the size of K until a measure of the strength of clusters, such as the CH index or the Silhouette index, is optimized. The estimation of the number of clusters and the ultimate partition are then obtained simultaneously based on the value of the optimal K. Our algorithm bears similarities to the second class of methods to determine the number of clusters as it starts from many small clusters first and then merge them together. However, the number of clusters is not determined by the user. Instead, it is obtained automatically through optimizing a measure of strength of clusters.

The rest of the paper is organized as follows. In Section 2, we present our clustering algorithm, CLUES. In Section 3, we describe the data sets used to study the performance of our method and discuss the results. In Section 4, we provide the conclusion and briefly discuss some future works.

2. The algorithm

CLUES (CLUstEring based on local Shrinking) algorithm consists of three major elements:

(1) shrinking procedure;
(2) partition procedure;
(3) determination of the optimal K.

We will describe the shrinking procedure in the next subsection. We will then discuss the partition procedure in the following subsection. The measure of strength of clusters such as the CH index and Silhouette index are discussed in Section 2.3. The determination of K is discussed in Section 2.4.

2.1. Shrinking procedure

The key idea of shrinking procedure resembles the gravitational clustering and the data sharpening procedure. As in the gravitational clustering, each data point can be viewed as a particle in a gravitational field with unit mass and zero velocity at the beginning. The local gravitational field would pull each data point into a denser region according to some gravitational laws. This can also be called sharpening effect as the boundary of each cluster should be clearer after this step. Therefore, the minimum distance among clusters will increase as the procedure goes on. In the sparse section of the sample space, the magnitude of the movement can be relatively large. In the dense region, however,