Computational Statistics and Data Analysis 61 (2013) 124–136

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

A fast algorithm for robust constrained clustering

Heinrich Fritz [a], Luis A. García-Escudero [b,*], Agustín Mayo-Iscar [b]

[a] Department of Statistics and Probability Theory, Vienna University of Technology, Austria

[b] Department of Statistics and Operations Research and IMUVA, University of Valladolid, Spain

ARTICLE INFO

Article history:
Received 13 March 2012
Received in revised form 20 November 2012
Accepted 23 November 2012
Available online 29 November 2012

Keywords:
Cluster analysis
Robustness
Impartial trimming
Classification EM algorithm
TCLUST

ABSTRACT

The application of "concentration" steps is the main principle behind Forgy's *k*-means algorithm and the fast-MCD algorithm. Despite this coincidence, it is not completely straightforward to combine both algorithms for developing a clustering method which is not severely affected by few outlying observations and being able to cope with non spherical clusters. A sensible way of combining them relies on controlling the relative cluster scatters through constrained concentration steps. With this idea in mind, a new algorithm for the TCLUST robust clustering procedure is proposed which implements such constrained concentration steps in a computationally efficient fashion.

1. Introduction

It is easy to realize that there are clear relations between Forgy's *k*-means algorithm (Forgy, 1965) and the fast-MCD algorithm (Rousseeuw and van Driessen, 1999). These two widely applied algorithms play a clear key role in cluster analysis and in robust statistics, respectively. The connection between them mainly refers to the application of the so-called "concentration" steps. Roughly speaking, in these concentration steps, the closest observations to a given center are considered in order to update this center estimate, such that the algorithm searches for regions with a high concentration of observations.

A great drawback when using the *k*-means method is that it ideally searches for spherically scattered clusters with similar sizes. Further, the presence of a certain fraction of outlying observations could negatively affect its performance (see, e.g., García-Escudero et al., 2010).

Under the previous premises, it seems quite logical to try to combine the clustering ability of *k*-means with the ability to robustly estimate covariance structures provided by the fast-MCD algorithm.

The trimmed *k*-means algorithm (García-Escudero et al., 2003) can be seen as a simple combination of *k*-means and fast-MCD algorithms, where spherical clusters are still assumed. In each concentration step, the proportion α of the most remote observations (considering Euclidean distances) to the previous k centers are discarded. Subsequently, k new centers are obtained by using the group means of the non-discarded observations. Note that the approach simplifies to the well-known Forgy's *k*-means algorithm when the trimming level α is set to 0. More information on the trimmed *k*-means approach can be found in Cuesta-Albertos et al. (1997) and García-Escudero and Gordaliza (1999).

* Correspondence to: Departamento de Estadística e I.O., Facultad de Ciencias, Paseo de Belén 7, 47011 Valladolid, Spain. Tel.: +34 983 185878; fax: +34 983 185861.

E-mail addresses: heinrich@fritz.cc (H. Fritz), lagarcia@eio.uva.es (L.A. García-Escudero), agustinm@eio.uva.es (A. Mayo-Iscar).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.11.018

It is also a logical step to think about the trimmed k-means algorithm but considering Mahalanobis distances $(\boldsymbol{x}_i - \boldsymbol{m}_j)'\boldsymbol{S}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{m}_j)$ (as the fast-MCD algorithm does) instead of Euclidean distances. In this case, the centers $\boldsymbol{m}_j$ and scatter matrices $\boldsymbol{S}_j$ for $j = 1, \ldots, k$ would be updated by computing sample means and sample covariance matrices of the non-discarded observations assigned to each group. Unfortunately, this "naive" combination of algorithms does not provide sensible clustering results, since large clusters tend to "eat" smaller ones. This problem was already noticed in Maronna and Jacovkis (1974) in the untrimmed case ($\alpha = 0$).

For avoiding this drawback, additional constraints are introduced, which limit the difference between the cluster scatters. In fact, many well-known clustering methods implement (implicitly and explicitly) such constraints. For example, the k-means method assumes the same spherical scatter for all the clusters.

Hathaway (1985), in a pioneering work on the mixture fitting framework, proposed constraining the relative differences between cluster scatters through a constant c that controls the strength of the constraints. With this idea in mind, García-Escudero et al. (2008) introduces the TCLUST method which is based on controlling the relative sizes of the eigenvalues of the cluster scatter matrices.

The TCLUST method has good robustness behavior and nice theoretical properties (the existence of solutions for both sample and population problems, together with the consistency of sample solutions to population ones). Unfortunately, from a computational viewpoint, solving the TCLUST problem is not an easy task. Although an algorithm for solving this problem was given in García-Escudero et al. (2008), the most critical issue there was how to enforce the eigenvalue ratio constraints. This is clearly its computational bottle-neck because a complex optimization problem must be solved in each concentration step. To be more precise, a maximization of a $(k \times p)$-variate function with $\binom{k\times p}{2}$ constraints needs to be solved (k stands for the number of clusters and p for the data dimension). This makes the algorithm computationally unfeasible even for moderate values of k and/or p.

In this work, we present an algorithm for implementing the constrained concentration steps, which clearly speeds up the previous TCLUST algorithm and makes it computationally feasible for practical applications. This algorithm only requires the evaluation of a not very complex function $2pk + 1$ times in each concentration step.

The proposed algorithm can be seen as a Classification EM algorithm (Schroeder, 1976; Celeux and Govaert, 1992) and, more generally, as a generalized k-means algorithm (Bock, 2007). Note that the proposed algorithm allows to exactly solve the (constrained) maximization step, which forces the trimmed likelihood target function to increase monotonically through the iterations.

An implementation of the algorithm described in this work is available through the R package `tclust` available at http://CRAN.R-project.org/package=tclust. A description of how this R package can be used in practical applications can be found in Fritz et al. (2012). In this work, we detail the algorithms internally applied by this package.

The methodology behind the discussed approach is explained in Section 2, while the algorithm is presented in Section 3. Section 4 contains a brief simulation study, investigating the performance of the algorithm and it is compared to other closely related ones in Section 5. Section 6 explains how this algorithm allows the practical application of exploratory tools which help us to decide on the number of clusters and the trimming level. Section 7 finally presents concluding thoughts.

2. Constrained robust clustering and TCLUST

Given a sample of observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in $\mathbb{R}^p$ and $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, the probability density function of a p-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, we consider the following general *robust constrained clustering problem* for a fixed trimming level α:

> Search for a partition $R_0, R_1, \ldots, R_k$ of the indices $\{1, \ldots, n\}$ with $\#R_0 = \lceil n\alpha \rceil$, centers $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k$ in $\mathbb{R}^p$, symmetric positive semidefinite $p \times p$ scatter matrices $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_k$ and weights $p_1, \ldots, p_k$ with $p_j \in [0, 1]$ and $\sum_{j=1}^k p_j = 1$, which maximizes

$$\sum_{j=1}^{k} \sum_{i \in R_j} \log\big(p_j \phi(\boldsymbol{x}_i; \boldsymbol{m}_j, \boldsymbol{S}_j)\big). \tag{2.1}$$

Depending on the constraints imposed on the weights p_j and on the scatter matrices $\boldsymbol{S}_j$, the maximization of (2.1) for $\alpha = 0$ leads to well established clustering procedures. For instance, assuming equal weights $p_1 = \cdots = p_k$ and scatter matrices $\boldsymbol{S}_1 = \cdots = \boldsymbol{S}_k = \sigma^2 \boldsymbol{I}$ with $\boldsymbol{I}$ being the identity matrix and $\sigma > 0$ yields the k-means method. The determinantal criterion introduced by Friedman and Rubin (1967) is obtained when assuming $p_1 = \cdots = p_k$ and $\boldsymbol{S}_1 = \cdots = \boldsymbol{S}_k = \boldsymbol{S}$ with $\boldsymbol{S}$ being a positive definite matrix. In general, the "likelihood" in (2.1) when $\alpha = 0$ and $p_1 = \cdots = p_k$ is often referred to as the Classification-Likelihood (see, e.g., Scott and Symons, 1971). The use of (2.1) assuming different weights p_j goes back to Symons (1981) and Bryant (1991) and is also known as the penalized Classification-Likelihood criterion.

Trimmed alternatives to the previously commented approaches can be constructed by introducing a trimming level $\alpha > 0$ to (2.1), which yields "trimmed likelihoods". This way, for instance, the trimmed k-means method in Cuesta-Albertos et al. (1997) extends k-means and the trimmed determinantal criterion in Gallegos and Ritter (2005) extends the