ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Model-based clustering of high-dimensional data: A review

Charles Bouveyron [a], Camille Brunet-Saumard [b,*]

[a] Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne, France
[b] Laboratoire LAREMA, UMR CNRS 6093, Université d'Angers, France

ARTICLE INFO

Article history:
Received 28 June 2012
Received in revised form 12 December 2012
Accepted 15 December 2012
Available online xxxx

Keywords:
Model-based clustering
High-dimensional data
Dimension reduction
Regularization
Parsimonious models
Subspace clustering
Variable selection
Software
R package

ABSTRACT

Model-based clustering is a popular tool which is renowned for its probabilistic foundations and its flexibility. However, high-dimensional data are nowadays more and more frequent and, unfortunately, classical model-based clustering techniques show a disappointing behavior in high-dimensional spaces. This is mainly due to the fact that model-based clustering methods are dramatically over-parametrized in this case. However, high-dimensional spaces have specific characteristics which are useful for clustering and recent techniques exploit those characteristics. After having recalled the bases of model-based clustering, dimension reduction approaches, regularization-based techniques, parsimonious modeling, subspace clustering methods and clustering methods based on variable selection are reviewed. Existing softwares for model-based clustering of high-dimensional data will be also reviewed and their practical use will be illustrated on real-world data sets.

1. Introduction

Clustering is a data analysis tool which aims to group data into several homogeneous groups. The clustering problem has been studied for years and usually occurs in applications for which a partition of the data is necessary. In particular, more and more scientific fields require to cluster data in the aim to understand or interpret the studied phenomenon. Earliest approaches were based on heuristic or geometric procedures. They relied on dissimilarity measures between pairs of observations. A popular dissimilarity measure is based on the distance between groups, previously introduced by Ward (1963) for hierarchical clustering. In the same way, the *k*-means algorithm (MacQueen, 1967) is perhaps the most popular clustering algorithm among the geometric procedures. Clustering was also defined in a probabilistic framework, allowing to formalize the notion of clusters through their probability distribution. One of the main advantages of this probabilistic approach is in the fact that the obtained partition can be interpreted from a statistical point of view. The first works on finite mixture models were from Wolfe (1963), Scott and Symons (1971) and Duda et al. (2000). Since then, these models have been extensively studied and, thanks to works such as those of McLachlan and Basford (1988), McLachlan and Peel (2000), Banfield and Raftery (1993) or Fraley (1998) and Fraley and Raftery (2002), model-based clustering has become a popular and reference technique.

Nowadays, the measured observations in many scientific domains are frequently high-dimensional and clustering such data is a challenging problem (Tran et al., 2006; von Borries and Wang, 2009; Tritchler et al., 2005), particularly for model-based methods. Indeed, model-based methods show a disappointing behavior in high-dimensional spaces. They suffer from the well-known *curse of dimensionality* (Bellman, 1957) which is mainly due to the fact that model-based

* Correspondence to: Laboratoire LAREMA, Université d'Angers, 2 bd Lavoisier, 49045 Angers Cédex, France.
E-mail address: camille.brunet@gmail.com (C. Brunet-Saumard).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.12.008

Please cite this article in press as: Bouveyron, C., Brunet-Saumard, C., Model-based clustering of high-dimensional data: A review. Computational Statistics and Data Analysis (2013), doi:10.1016/j.csda.2012.12.008

clustering methods are over-parametrized in high-dimensional spaces. Furthermore, in several applications, such as mass spectrometry or genomics, the number of available observations is small compared to the number of variables and such a situation increases the problem difficulty. Since the dimension of observed data is usually higher than their intrinsic dimension, it is theoretically possible to reduce the dimension of the original space without loosing any information. For this reason, dimension reduction methods are frequently used in practice to reduce the dimension of the data before the clustering step. Feature extraction methods, such as principal component analysis (PCA), or feature selection methods are very popular. However, dimension reduction usually does not consider the classification task and provide a sub-optimal data representation for the clustering step. Indeed, dimension reduction methods imply an information loss which could have been discriminative.

To avoid the drawbacks of dimension reduction, several approaches have been proposed to allow model-based methods to efficiently cluster high-dimensional data. This work proposes to review the alternatives to dimension reduction for dealing with high-dimensional data in the context of model-based clustering. Earliest approaches include constrained and parsimonious models or regularization. More recently, subspace clustering techniques and variable selection techniques have been proposed to overcome the limitations of previous approaches. Subspace clustering techniques are based on probabilistic versions of the factor analysis model. This modeling allows to cluster the data in low-dimensional subspaces without reducing the dimension. Conversely, variable selection techniques do reduce the dimension of the data but select the variables to retain regarding the clustering task. Both techniques turn out to be very efficient and their practical use will be discussed as well in this article.

This article is organized as follows. Section 2 briefly recalls the bases of mixture modeling and its inference with the EM algorithm. Section 3 introduces the curse of dimensionality in model-based clustering. Approaches based on dimension reduction, regularization and parsimonious models are reviewed in Section 4. Then, Sections 5 and 6 present respectively the approaches based on subspace clustering and variable selection. Existing softwares for model-based clustering of high-dimensional data are also reviewed in Section 7 and their practical use is discussed in Section 8. Finally, some concluding remarks are made in Section 9.

2. The mixture model and the EM algorithm

This section first recalls the bases of mixture modeling and its inference with the expectation–maximization (EM) algorithm.

2.1. The mixture model

Let us consider a data set of n observations $\{y_1, \ldots, y_n\} \in \mathbb{R}^p$ that one wants to divide into K homogeneous groups. The aim of clustering is to determine, for each observation y_i, the value of its unobserved label z_i such that $z_i = k$ if the observation y_i belongs to the kth cluster. To do so, model-based clustering (Fraley and Raftery, 2002; McLachlan and Peel, 2000) considers the overall population as a mixture of the groups and each component of this mixture is modeled through its conditional probability distribution. In this context, the observations $\{y_1, \ldots, y_n\} \in \mathbb{R}^p$ are assumed to be independent realizations of a random vector $Y \in \mathbb{R}^p$ whereas the unobserved labels $\{z_1, \ldots, z_n\}$ are assumed to be independent realizations of a random variable $Z \in \{1, \ldots, K\}$. The set of pairs $\{(y_i, z_i)\}_{i=1}^n$ is usually referred to as the complete data set. By denoting by g the probabilistic density function of Y, the finite mixture model is:

$$g(y) = \sum_{k=1}^{K} \pi_k f_k(y), \tag{1}$$

where π_k $\left(\text{with the constraint } \sum_{k=1}^{K} \pi_k = 1\right)$ and f_k respectively represent the mixture proportion and the conditional density function of the kth mixture component. Furthermore, the clusters are often modeled by the same parametric density function in which case the finite mixture model is:

$$g(y) = \sum_{k=1}^{K} \pi_k f(y; \theta_k), \tag{2}$$

where θ_k is the parameter vector for the kth mixture component. For a set of observations $y = \{y_1, \ldots, y_n\}$, the log-likelihood of this mixture model is then:

$$\ell(\theta; y) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f(y_i; \theta_k)\right). \tag{3}$$

However, the inference of this model cannot be directly done through the maximization of the likelihood since the group labels $\{z_1, \ldots, z_n\}$ of the observations are unknown. Indeed, due to the exponential number of solutions to explore, the

Please cite this article in press as: Bouveyron, C., Brunet-Saumard, C., Model-based clustering of high-dimensional data: A review. Computational Statistics and Data Analysis (2013), doi:10.1016/j.csda.2012.12.008