Choice of the number of component clusters in mixture models by information criteria

Christian Olivier
SIC-IRCOM, UMR CNRS 6615
Université de Poitiers, France
Olivier@sic.univ-poitiers.fr

Frédéric Jouzel
laboratoire PSI
Université de Rouen, France
Frederic.Jouzel@univ-rouen.fr

Abdelaziz El Matouat
ENS de Fès, Maroc
Abdelaziz.El-Matouat@
univ-rouen.fr

Abstract

This paper considers the problem of choosing the number of component clusters within the context of the standard mixture of multivariate normal distributions. The problem to choose the number of clusters in a clustering procedure has already been dealt with, but still remains opened. We propose to use information criteria to solve this problem in the Gaussian mixture-model approach, which is nowadays a standard approach in clustering. The different criteria are presented and then compared with other well-known criteria on synthetic data sets. Often, the number of clusters k is unknown and needs to be estimated. A two-stage iterative maximum-likelihood procedure is used as a clustering technique to estimate the parameters of the mixture-model. A new criterion φ_β is derived and proposed as a criterion for choosing the number of clusters in the mixture-model context. For comparative purposes, Akaike's information criterion AIC *(1973) and Rissanen's 1978* MDL *criterion are also introduced in the mixture-model context. Numerical examples are shown on simulated normal data sets with a known number of mixture clusters to illustrate the significance of φ_β in choosing the number of clusters and the best fitting model. We demonstrate its efficiency and robustness through experimental results for synthetic mixture data sets.*

1 Introduction and statement of the problem

A general common problem in all clustering techniques is the difficulty of deciding of the number of clusters present in a given data set, cluster validity, and the identification of the appropriate number of clusters. This paper considers the problem of choosing the number of clusters within the context of unsupervised clustering in the framework of pattern recognition.

Most of the existing clustering procedures require prior knowledge of the number of clusters which is often unavailable and has to be estimated [17]. In the mixture-model cluster analysis, often the number of component clusters k is not known and needs to be estimated from the available observed data. This problem is known as the *cluster validation* problem [8]. Nevertheless, this aspect is often omitted in mixture type approaches [5, 16]. Despite the increased number of books appearing on finite mixture distributions such as [5], a relatively little work has been done concerning the choice of the number of component mixture clusters.

In this paper, the clustering viewpoint consists in identifying and describing the class distribution using a sample drawn from the mixture-model, and estimating the number of mixture clusters k in a non-subjective manner. To achieve this, we will use a two stage iterative maximum likelihhod (ML) procedure (EM algorithm) to estimate the parameters in the mixture-model, and develop a new information criterion φ_β as a new index for cluster validation.

The paper is organized as follows. In the next section, we introduce the ML procedure to estimate the parameters of the model. In section 3 we introduce the φ_β criterion to perform order estimate (choice of the number of clusters). We explain in part 3.2 the analytical expression of the criteria in the mixture-model case. Finally, we give in section 4 some results on synthetic datasets.

2 The standard mixture-model cluster analysis

There are a lot of clustering algorithms to perform unsupervised classification, for instance the k-means algorithm or hierarchical classification. A standard and powerful approach is available through a multivariate Gaussian mixture hypothesis. Data are assumed to