International Journal of Approximate Reasoning 94 (2018) 60–75

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Robust, fuzzy, and parsimonious clustering, based on mixtures of Factor Analyzers [☆]

Luis Angel García-Escudero [a], Francesca Greselin [b,*], Agustin Mayo Iscar [a]

[a] Department of Statistics and Operational Research and IMUVA, University of Valladolid, Spain
[b] Department of Statistics and Quantitative Methods, Milano-Bicocca University, Italy

ARTICLE INFO

Article history:
Received 11 January 2017
Received in revised form 20 January 2018
Accepted 21 January 2018

Keywords:
Fuzzy clustering
Robust clustering
Unsupervised learning
Factor analysis
Hard contrast
Dimension reduction
Outliers identification

ABSTRACT

A clustering algorithm that combines the advantages of fuzzy clustering and robust statistical estimators is presented. It is based on mixtures of Factor Analyzers, endowed by the joint usage of impartial trimming and constrained estimation of scatter matrices, in a modified maximum likelihood approach. The algorithm generates a set of membership values, that are used to fuzzy partition the data set and to contribute to the robust estimates of the mixture parameters. The adoption of clusters modeled by Gaussian Factor Analysis allows for dimension reduction and for discovering local linear structures in the data. The new methodology has been shown to be resistant to different types of contamination, by applying it on artificial data. A brief discussion on the tuning parameters, such as the trimming level, the fuzzifier parameter, the number of clusters and the value of the scatter matrices constraint, has been developed, also with the help of some heuristic tools for their choice. Finally, a real data set has been analyzed, to show how intermediate membership values are estimated for observations lying at cluster overlap, while cluster cores are composed by observations that are assigned to a cluster in a crisp way.

1. Introduction

Fuzzy clustering is a method of data analysis and pattern recognition which allocates a set of observations to clusters in a "fuzzy" way, more formally, constructs a "membership" matrix whose (i, g)-th element represents "the degree of belonging" of the i-th observation to the g-th cluster.

In this sense, the clusters are "fuzzy sets" as defined in [45]. In our case, the fuzzy sets will be based on mixtures of Gaussian factor analyzers, hence they have a strong probabilistic meaning and our rules of operation are not those proposed by Zadeh but those that come naturally from probability theory.

Factor analysis is a widely employed method to be used when, as it happens in many phenomena, several observed variables could be explained by a few unobserved ones, exploiting their correlations. It is a powerful method, whose scope of application is unfortunately limited by the assumption of global linearity for representing data. To widen its applicability, [21] and [26] introduced the idea of combining one of the basic form of dimensionality reduction – factor analysis – with a basic method for clustering – the Gaussian mixture model, thus arriving at the definition of mixtures of factor analyzers. At the same time, [40,41] and [3] considered the related model of mixtures of principal component analyzers, for a similar

☆ This paper is part of the Virtual special issue on Soft methods in probability and statistics, Edited by Barbara Vantaggi, Maria Brigida Ferraro, Paolo Giordani.

* Corresponding author.
E-mail addresses: lagarcia@eio.uva.es (L.A. García-Escudero), francesca.greselin@unimib.it (F. Greselin), agustin@med.uva.es (A.M. Iscar).

https://doi.org/10.1016/j.ijar.2018.01.001

purpose. Further references may be found in [33] (chapter 8). Factor analysis is related to principal component analysis (PCA) [41]; however, these two methods have many conceptual and algorithmic differences, with the most significant being that the former aims at modelling *correlation* between variables and searches for underlying linear structures, the second focuses on their *variances* and identifies a reduced set of linear transformations of variables, maximizing their variance.

With reference to other techniques for data reduction, we want in particular to mention [5], where it has been shown that the principal components corresponding to the larger eigenvalues do not necessarily contain information about group structure. Therefore, data reduction and clustering *separately* may not be a good idea. Data reduction can ameliorate clustering and classification results, but combining variable selection *and* clustering can give improved results. Mixtures of factor analysers are designed exactly for simultaneously performing clustering and dimension reduction. Using this approach, we aim at finding local linear models in clusters, that could be particularly useful for datasets with a high number of observed variables. Beyond its parsimony, this method often provides a better insight into the structure of the original data.

In real datasets, noise and outliers contaminate the data collection, and any assumed model is only an approximation to reality. Unfortunately, one single outlier – if located sufficiently far away – can completely ruin the results of many (hard and fuzzy) clustering methods, in the sense that at least one of the component parameters estimate can be arbitrarily large. Remarkably, the fuzzy community has been traditionally interested in robustness issues in Cluster Analysis (see, e.g., the large amount of references in the two review papers [10,1]). In fuzzy clustering, the aim is to obtain a collection of membership values $u_{ig} \in [0,1]$ for all $i = 1 \ldots n$ and $g = 1 \ldots G$, where increasing degrees of membership are meant by higher values of the u_{ig}. We may understand why robustness in so critical in fuzzy clustering when we observe that for an outlying observation $\mathbf{x}_i$, generally lying far from the G clusters, we could obtain $u_{ig} \sim 1/G$, while we would expect $u_{ig} \sim 0$ for $g = 1, \ldots, G$, to convey a very low plausibility to belong to any cluster. In addition, if one outlying observation $\mathbf{x}_i$ is placed in a very distant position but closer to cluster g than the others, then typically $u_{ig} \sim 1$ and $\mathbf{x}_i$ would influence heavily the parameters' estimation of cluster g.

Hence we put forward a robust methodology based on trimming, to identify outliers and to assign them zero membership values. In our approach, we will indicate observation $\mathbf{x}_i$ as fully trimmed if $u_{ig} = 0$ for all $g = 1, \ldots, G$ and, thus, this observation has no membership contribution to any cluster. The proposed idea trace back to the seminal paper [8], and is called *impartial trimming*, or trimming self-determined by the data. We look for a method having a reasonably good efficiency (accuracy) at the assumed model; for which small deviations from the model assumptions should impair the performance only by a small amount; and for which larger deviations from the model assumptions should not cause a catastrophe [27].

Among the several methodologies for robust fuzzy clustering, some well-known approaches are the "noise clustering" in [9], the use of more robust discrepancy measures [43,32] and the "possibilistic" clustering in [31]. References concerning robustness in hard (crisp) clustering can be found in [17,12,34].

An interesting robust proposal have been introduced in [6], where the Student's t-mixture (instead of the Gaussian), is exploited for modeling factors and errors, allowing for outlier downweighting during the model fitting procedure. Alternatively, outliers can be accommodated in the model by considering additional mixture components. It turns out that the two alternatives, while adding stability in the presence of outliers of moderate size, do not possess a substantially better breakdown behavior than estimation based on normal mixtures [25]. In other words, one only observation could completely breakdown the estimation. Hence a model-based alternative with good breakdown properties is still missing in the literature.

On the other hand, a further issue inherent to the choice of Gaussian Mixtures (GM) is originated by the unboundedness of the likelihood, turning the maximization of the objective function into an ill-posed problem. Therefore, following a wide literature stream, based on [24,28,16,22], we will adopt a constrained estimation of the component covariances, to avoid singularities and to reduce the appearance of spurious solutions. In particular, the joint use of trimming and constrained estimation for mixtures of factor analyzers (MFA) have previously been studied for non fuzzy clustering in [19]. The advantage of MFA with respect to GM is due to the reduction of the number of free parameters to be estimated. In each component, a reduced number of latent variables, called *factors* and lying in a lower-dimensional subspace, explain the observed variables, through a linear submodel. Therefore, there is scope for a robust fuzzy clustering approach, along the lines of [14], and [11] for fuzzy robust clusterwise regression.

The original contribution of the present paper is to combine: i) robust estimation techniques; ii) fuzzy clustering; iii) dimension reduction through factor analysis; iv) the flexibility of having hard and soft assignments for units (the "hard contrast" property introduced in [35]). The interplay between these four features provides a novel powerful model with a parsimonious parametrization, specially useful in higher dimensional fuzzy clustering problems.

The outline of the paper is as follows. In Section 2 we introduce a fuzzy version of robust mixtures of Gaussian MFA, considering how to implement fuzziness, trimming and constrained estimation. Section 3 presents an efficient algorithm for its estimation. Afterwards, in Section 4, we present results on synthetic data to show the effectiveness of the proposal. A brief discussion on the role of the parameters involved in the fuzzy robust procedure, such as the trimming level, the fuzzifier parameter, the number of clusters and the value of the scatter matrices constraint, has been developed, also with the help of some heuristic tools for their tuning. An application to real data is provided in Section 5, in comparison with competing methods. Conclusions and further work are delineated in Section 6.