Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems ••• (••••) •••–•••

FUZZY sets and systems

www.elsevier.com/locate/fss

Clustering of fuzzy data and simultaneous feature selection: A model selection approach

Arkajyoti Saha [a], Swagatam Das [b,*]

[a] *Stat-Math Unit, Indian Statistical Institute, 203 B.T.Road, Kolkata-700108, W.B., India*
[b] *Electronics and Communication Sciences Unit, Indian Statistical Institute, 203 B.T.Road, Kolkata-700108, W.B., India*

Received 1 August 2016; received in revised form 12 July 2017; accepted 28 November 2017

Abstract

Fuzzy data occurs frequently in the fields of decision making, social sciences, and control theory. We consider the problem of clustering fuzzy data along with automatic component number detection and feature selection. A model selection criterion called minimum message length is used to address the problem of component number selection. The Bayesian framework can be adopted here, by applying an explicit prior distribution over the parameter values. We discuss both uninformative and informative priors. For the latter, a gradient descent algorithm for automatic optimization of the prior hyper-parameters is presented. The problem of simultaneous feature selection involves ordering the discriminative features according to their relative importance, and at the same time eliminating non-discriminative features. The feature selection problem is also formulated as a parameter estimation problem by extending the concept of feature saliency. Then the estimation can be computed simultaneously with the clustering steps. By combining the clustering, the cluster number detection and the feature selection into one estimation problem, we modified the fuzzy Expectation–Maximization (EM) algorithm to perform all of the estimation. Evaluation criteria are proposed and empirical study results are reported to showcase the efficacy of our proposals.

Keywords: Fuzzy data; Component number selection; Model selection; Minimum Message Length (MML); Bayesian; Jeffreys prior; DNW prior

1. Introduction

Often, the classes of objects encountered in the physical world suffer from non-random imprecision. Especially in the fields of artificial intelligence, pattern recognition, communication of information, decision making, social sciences, industrial engineering (automobile, control engineering, medicine, logistics, and vehicular communications), we frequently encounter cases where a precise value of a variable is either not available, cannot be measured, is economically infeasible, or is not required. It might also be the case that the magnitudes are ill defined or only linguistic values for the magnitudes are available. Fuzzy data helps us to mathematically model the vagueness with

[*] Corresponding author.
E-mail addresses: arkajyotisaha93@gmail.com (A. Saha), swagatam.das@isical.ac.in (S. Das).

https://doi.org/10.1016/j.fss.2017.11.015
0165-0114/© 2017 Elsevier B.V. All rights reserved.

the least possible loss of information, while retaining the inherent simplicity of the data. The statistical formalization of the idea of imprecision in data can be performed in the following two ways.

1. **Physical interpretation of fuzzy data:** In this interpretation of fuzzy data, the independent existence of fuzzy datum is assumed, i.e. it is not associated with any underlying precise variable [1]. This assumes the data under consideration is intrinsically fuzzy and follows mathematical formalism of fuzzy random variables, which are defined as mappings from a probability space to a fuzzy subset [2] with certain measurability properties. Some recent literature follows this interpretation of fuzzy data for estimation and hypothesis testing purposes [3,4].
2. **Epistemic interpretation of fuzzy data:** In this approach, the fuzzy numbers are assumed to "imperfectly specify a value that is existing and precise, but not measurable with exactitude under the given observation conditions" [1]. Unlike the previous approach, here, a fuzzy datum is associated with an existing and precise random variable. It is interpreted as a "possibility distribution associated to a precise realization of a random variable that has only been partially observed" [5]. We adopt this approach towards the fuzzy data in this article.

Clustering is the method of partitioning the data based on some (dis)similarity measures, such that the data within a cluster is as similar as possible, and data from different clusters are as dissimilar as possible. Under the assumption that the data are produced (following some probability distribution) from one of a number (unknown) of alternative sources of generation, finite mixture models are rich enough to be extended to an involved statistical model by which issues like selection of an optimal number of clusters, feature saliency and the validity of a given model can be addressed in a formal and structured way [6].

Clustering of fuzzy data is a topic of interest to the modern research on imprecise data analytics [7–14]. Though the mixture model has been extensively used in clustering of crisp data it was first introduced in the fuzzy setup in [5,15]. The fuzzy EM algorithm performs the mixture model based clustering of fuzzy data. The major limitations with this algorithm are summarized below:

1. It does not take into consideration the number of components, which is one of the most important things to be determined in the clustering of a dataset and is unknown in general.
2. It does not consider selecting the most useful features, and hence suffers from poor performance in the presence of noise variables and higher computational load.
3. The standard fuzzy EM algorithm, which is used to fit the finite mixture model, suffers from problems [6] associated with the basic EM algorithm, like sensitivity to initialization and convergence to boundary of the parameter set [5].

In [15], the Monte-Carlo estimation was introduced to take care of the non-Gaussian scenario, which is developed as a generalization of [5]. The issue with the convergence to local maximum and dependence on initialization was taken care of with Bayesian approaches. In this article, we extend the Bayesian approaches in the perspective of feature selection. For the sake of computational simplicity, we have restricted our attention to only Gaussian mixture models. Similar to [15], the developed algorithm can also be extended in the non-Gaussian scenario with Monte-Carlo estimation.

In this paper, we introduce an automated model selection approach for the clustering of fuzzy data. We summarize the main contributions of the paper as follows:

1. In order to automatically determine the number of components, we incorporate a classical Bayesian model-selection technique, namely the Minimum Message Length (MML) criterion [16], along with different variations and choices of priors, in the conventional Fuzzy EM setup, which solves the problem of high sensitivity to initialization and convergence to the boundary of parameter set.
2. We formulate and incorporate the feature selection problem as an estimation problem by extending the concept of feature saliency [17] in the fuzzy data setup and develop a modified fuzzy EM algorithm to address the problem of feature saliency determination. To avoid the dependency on "good initialization", here we introduce the MML criterion as the model selection criterion with two different choices of priors (uninformative and informative). In the corresponding analysis, this guarantees that feature saliency of irrelevant or noisy features are driven towards zero, resulting in automatic feature selection.

Please cite this article in press as: A. Saha, S. Das, Clustering of fuzzy data and simultaneous feature selection: A model selection approach, Fuzzy Sets Syst. (2017), https://doi.org/10.1016/j.fss.2017.11.015