Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 279 (2015) 112–129

FUZZY sets and systems

www.elsevier.com/locate/fss

On robust fuzzy c-regression models

Jacek M. Leski [a,b,*], Marian Kotas [a]

[a] Institute of Electronics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

[b] Department of Computer Medical Systems, Institute of Medical Technology and Equipment, Roosevelt St. 118, 41-800 Zabrze, Poland

Received 24 April 2013; received in revised form 14 May 2014; accepted 6 December 2014

Available online 10 December 2014

Abstract

One of the most popular clustering methods based on minimization of a criterion function is the fuzzy c-means one. Its generalization by application of hyperplane shaped prototypes of the clusters is known as the **F**uzzy **C**-**R**egression **M**odels (FCRM) method. Although with this generalization many new applications of clustering emerged, it appeared to be rather sensitive to poor initialization and to the presence of noise and outliers in data. In this paper we introduce a new objective function, using the Huber's M-estimators and the Yager's OWA operators to overcome the disadvantages of the approach considered. We derive and describe an algorithm for minimization of the objective function defined. We have called it the **F**uzzy **C**-**O**rdered-**R**egression **M**odels (FCORM) clustering algorithm. The algorithm is compared to a few other important reference ones. To this end experiments on synthetic data with various types of noise and different numbers of outliers are carried out. We investigate the methods performance in the conditions that can be encountered in signal analysis. Large-scale simulations demonstrate the competitiveness and usefulness of the method proposed.

Keywords: Fuzzy clustering; Fuzzy c-means; Fuzzy c-regression models; Ordered weighted averaging; Robust methods; Signal analysis

1. Introduction

The unsupervised classification of data into groups is called clustering. The method plays an important role in many engineering fields, such as pattern recognition, computer vision, machine learning, image analysis, communication, knowledge discovery, data mining and so on [1,10,19]. In the traditional so-called hard clustering, the groups (clusters) are disjoint. Each data item belongs to one cluster only. In [43] Zadeh introduced the notion of a fuzzy membership function. It allowed to associate with each data item and each cluster a real number in the interval [0,1] representing the "grade of membership" of this item in the cluster considered. This way Zadeh formed the basis for the development of fuzzy (or soft) clustering. However, the idea itself has been introduced by Ruspini [35] and Dunn [9], and generalized by Bezdek who developed an approach based on criterion function minimization [1].

* Corresponding author at: Institute of Electronics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland. Tel.: +48 32 2372004; fax: +48 32 2372225.

E-mail address: jacek.leski@polsl.pl (J.M. Leski).

http://dx.doi.org/10.1016/j.fss.2014.12.004

One of the most popular clustering methods based on criterion function minimization is the **F**uzzy **C**-**M**eans (FCM) method which has successfully been applied to a wide variety of problems [1,5,6,21,24,25,33]. Many modifications of the FCM method have been proposed in the literature. Most of them rely on an inclusion of additional information into the clustering process. In an important group of FCM modifications, the information about the shapes of clusters prototypes is exploited. This information is relayed to the algorithms as the constraints on these shapes: in [1] and [26] the prototypes were constrained to linear varieties or linear elliptotypes in a feature space, in [12] to hyperellipsoidal ones and in [11] to hyperspheres in a feature space. This work deals with a modification of the **F**uzzy **C**-**R**egression **M**odels (FCRM) method, which is also called as the method of **F**uzzy **S**witching **R**egression **M**odels (FSRM)[1] [13], where the prototypes are constrained to functions (usually but not necessarily the linear ones). Many modifications of the above algorithm have recently appeared in the literature. The method proposed by Menard introduces additionally a so called noise cluster and, besides, it takes into account uncertainty with regard to data membership in linear regression models, near the models intersection [29]. In [23] ε-insensitive loss function was applied to the determination of data distances from the regression lines. A combination of clustering methods and the method of support vector regression was presented in [3]. The concept of α-cut implementation of the fuzzy c-regression models was introduced by Yang et al. [39]. To obtain robustness against bad initialization, the mountain clustering method was used [38] for the determination of linear regression parameters. In [40] support vector machines method was combined with clustering in the kernel space to improve robustness in classification problems.

In another group of fuzzy clustering modifications, the information about the data non-Gaussian distribution and the presence of noise and outliers is taken into account. This group includes: possibilistic clustering [8,19], fuzzy noise clustering [7], L_p $(0 < p < 1)$ norm clustering [14], L_1 and L_∞ norm clustering [2,16,17], fuzzy c-ordered-means clustering [27], time-domain-constrained clustering [25], ε-insensitive fuzzy clustering [20] and ε-insensitive fuzzy c-regression models clustering [23].

Two different approaches to clustering algorithms robustifying against outliers are worth emphasizing. They are: application of reweighting scenario based Huber's M-estimators [15] and the use of the ordered weighted averaging (OWA) operation [41]. In [27] both approaches were combined for the first time (the aim was to robustify a modification of the FCM method). The goal of this paper is to show that the Huber's M-estimators and the Yager's OWA operators can be used together to improve robustness of the method of the fuzzy c-regression models significantly. The second goal of this work is to investigate the performance of the proposed method when applied to data in the presence of noise and outliers. We will first present the proposed fuzzy c-ordered-regression models method, which exploits both approaches to be more robust, and then we will investigate its performance.

The investigations are aimed not only to present the proposed method competitiveness with respect to the reference methods but also to show its high practical value. To this end its application to the Q wave onset determination in the ECG signals can be considered. The Q wave begins the depolarization of the heart ventricles. Depolarization is followed by repolarization which ends at the offset of the T wave. Thus the both phases of the heart contraction are covered by a so called QT interval, presented in Fig. 1. Precise determination of the interval borders is of high clinical importance, yet it is rather difficult when the analyzed signals are noisy [18]. Fitting two linear models to the signal segments preceding and following the Q wave, we can determine the wave onset at the point where the lines intersect. Thus the switching regression models can help us to solve this important problem from the field of biomedical signal processing.

The paper is organized as follows: A detailed description of the Fuzzy C-Ordered Regression Models (FCORM) method is presented in Section 2. The experiments on different synthetic data, simulating the real life signals, are presented and discussed in Section 3. Finally, conclusions are drawn in Section 4.

2. Fuzzy c-ordered regression models

Suppose we have a set, $\mathrm{Tr}^{(N)} = \{(\mathbf{x}_k, y_k)\}_{k=1}^{N}$, where each independent datum $\mathbf{x}_k \in \mathbb{R}^t$ has a corresponding dependent datum $y_k \in \mathbb{R}$, and N is dataset cardinality. Although the data pairs $(\mathbf{x}_k, y_k)$ are unlabeled, we assume that they are drawn from the switching regression models, which consist of c models in the following linear form: $y_k = w_0^{(i)} + \widetilde{\mathbf{w}}^{(i)\top}\mathbf{x}_k + e_k^{(i)}$ for $k = 1, 2, \cdots, N$, where $e_k^{(i)}$ represents uncertainty (with zero mean) for the kth pair

[1] In the further part of this work the method will be denoted as FCRM(FSRM).