HEURISTIC POSSIBILISTIC CLUSTERING AND SWITCHING REGRESSION

D.A. VIATTCHENIN
National Academy of Sciences of Belarus
Minsk, Belarus
e-mail: viattchenin@mail.ru

Abstract

The paper deals in a preliminary way with the problem of forming of the switching regression models. The proposed approach is based on the clustering of the switching regression data set using a heuristic D-AFC-TC-algorithm of possibilistic clustering. An idea of the D-AFC-TC-algorithm is presented and the method of switching regression models forming is outlined. A numerical example is given and the result of application of the D-AFC-TC-algorithm to the two-dimensional data set is considered. Preliminary conclusions are stated.

1 Preliminaries

Regression analysis is a method of modeling the relation between explanatory and response variables. Usually, a single regression model is used for fitting the data set. However, it is necessary to have more than one regression model in some cases, for example, c number of regression models, for fitting a data set. The kind of models called switching regressions.

Let us assume that $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is the data set where each explanatory observation $x_i = (x_{i1}, \ldots, x_{im}) \in \Re^m$ has corresponding response observation y_i. The switching regression is employed to find c linear regressions which take the form

$$y_{il} = b_{l0} + b_{l1}x_{i1} + \ldots + b_{lm}x_{im}, \quad l = 1, \ldots, c \tag{1}$$

that best fit the data structure.

Some fuzzy clustering techniques were proposed for the switching regression problems solving. In the first place, Hathaway and Bezdek [1] combined switching regressions with fuzzy c-means clustering method and referred to them as fuzzy c-regression method. The $FCRM$-algorithm yields simultaneous estimates of the parameters of c regression models, together with a fuzzy partitioning of the data. From other hand, Wu, Yang and Hsieh [6] propose a mountain c-regression method based on the modified mountain clustering method [7]. The MCR-algorithm can form well-estimated c regression models for switching regression data sets. However, both considered methods depend on the number c of clusters in the original data set and the number c is unknown very often. So, a problem of estimation of the unknown number c regression models must be solved in the first place. The goal of the paper is a consideration of possibilities of an application of the heuristic $D-AFC-TC$-algorithm of possibilistic clustering to forming of unknown c regression models. For this purpose, a consideration

of an idea of the $D-AFC-TC$-algorithm is presented. A methodology for the data preprocessing is described. Results of application of the $D-AFC-TC$-algorithm of possibilistic clustering to two data sets and obtained switching regression models are considered. Preliminary conclusions are formulated.

2 An Introduction to the D-AFC-TC-algorithm

An outline for the heuristic approach to possibilistic clustering was outlined in [3], where a basic version of direct clustering algorithm was described. The basic version of the algorithm, which is described in [3], can be called the $D-AFC(c)$-algorithm and the possibilistic interpretation of the approach was given in [5]. The concepts of a fuzzy α-clusters and an allotment among fuzzy α-clusters are basic concepts of the approach. Detection of the allotment $R^*(X)$ among given number c of particularly separate fuzzy α-clusters $A^l_{(\alpha)}$, $l \in \{1,\ldots,c\}$ is the aim of the classification in the case of the data processing by the $D-AFC(c)$-algorithm.

From other hand, the $D-AFC-TC$-algorithm [4] is a modification of the $D-AFC(c)$-algorithm based on the transitive closure [2] of the matrix of fuzzy tolerance relation. Moreover, detection of a unique allotment among unknown number c of fuzzy α-clusters is the aim of classification in the case of the $D-AFC-TC$-algorithm. For the purpose, a heuristic of the leap of the tolerance threshold values α, $\alpha \in (0,1]$ was used in [4]. The allotment $R^\alpha_z(X) = \{A^l_{(\alpha)} | l = \overline{1,c}, \alpha \in (0,1]\}$ among the unknown number c of fully separate fuzzy α-clusters, the corresponding value of tolerance threshold α, typical points τ^l, $l = 1,\ldots,c$ of fuzzy α-clusters and their prototypes $\bar{\tau}^l$, $l = 1,\ldots,c$ coordinates are the results of classification in this case.

The matrix of coefficients of pair wise dissimilarity between objects $I = [\mu_I(x_i,x_j)]$, $i,j = 1,\ldots,n$ can be obtained after application of some distance to the matrix of the normalized data $X_{n\times m} = [\mu_{x_i}(x^t)]$, $i = 1,\ldots,n$, $t = 1,\ldots,m$. The most widely used distance for fuzzy sets x_i, x_j, $i,j = 1,\ldots,n$ in $X = \{x_1,\ldots,x_n\}$ is the squared normalized Euclidean distance [2]:

$$\varepsilon(x_i,x_j) = \frac{1}{m}\sum_{t=1}^{m}\left(\mu_{x_i}(x^t) - \mu_{x_j}(x^t)\right)^2. \quad i,j = \overline{1,n}. \tag{2}$$

3 An Outline of the Proposed Method

The basic idea of the proposed method is to constructing of regression models in the each fuzzy cluster of the obtained allotment. We know that any two points can form a line. In particular, any two points in a two-dimensional data set can have a regression equation $y = b_0 + b_1 x$ where the parameters of the equation is denoted by $\vec{b} = (b_0, b_1)$. So, for data points located around a regression line, the parameters $\vec{b}_l = (b_{l0}, b_{l1})$, $2 \le l \le c$ for these points should be very close, where c is the unknown number of regression models. That is why, the unknown number c of fuzzy clusters must be detected and the regression model (1) must be constructed in the each fuzzy cluster.

Thus, the $(m+1)$-dimensional data set can be processed by the $D-AFC-TC$-algorithm of possibilistic clustering and the equation (1) can be constructed in all fuzzy clusters of the obtained allotment $R^*(X) = \{A^1_{(\alpha)}, \quad , A^c_{(\alpha)}\}$ The typical point τ^l and the prototype $\bar{\tau}^l$ of each fuzzy cluster $A^l_{(\alpha)}$, $l \in \{1, \ldots, c\}$ can be selected as the points for forming of a regression line in each fuzzy cluster of the obtained allotment $R^*(X)$ among unknown number c of fuzzy clusters So, the regression model for each fuzzy cluster can be constructed immediately For the purpose, corresponding linear equations must be solved and parameters $\vec{b}_l = (b_{l0}, b_{l1}, \quad , b_{lm})$, $2 \leq l \leq c$ must be estimated

4 An Illustrative Example

Let us consider an application of the $D-AFC-TC$-algorithm to the switching regression problem for a simple illustrative example. For the purpose, the two-dimensionality data set $X = \{x_1, \ldots, x_{20}\}$ was generated. Let x be the explanatory variable and y be the response variable. So, the matrix of the object data $X_{20\times 2} = [x^t_i]$, $i = 1, \ldots, 20$, $t = 1, 2$ where $x = x^1$ and $y = x^2$ was processed by the $D-AFC-TC$-algorithm using the squared normalized Euclidean distance (2) The data set and obtained results are presented in Figure 1.

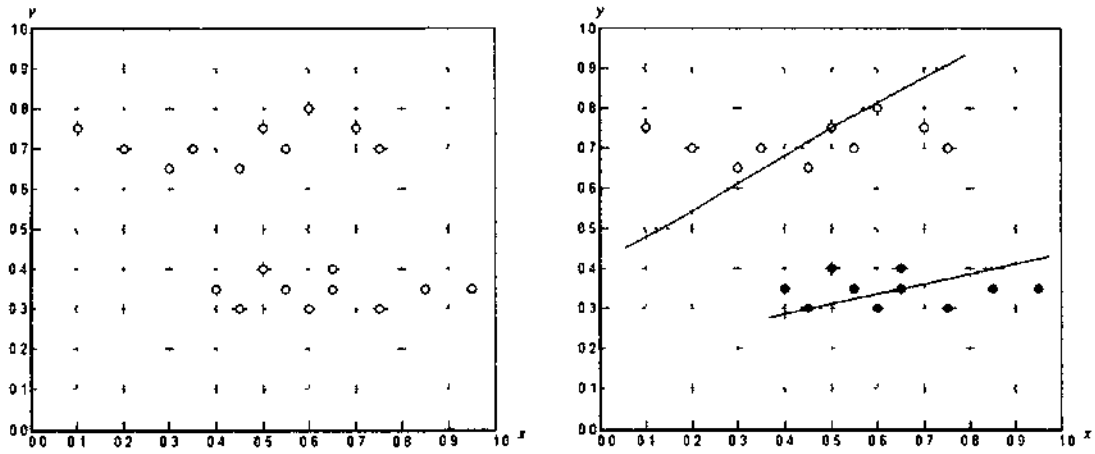

Figure 1 The two-cluster switching regression data set and the D-AFC-TC-algorithm results for the data set

By executing the $D-AFC-TC$-algorithm to the data set, the allotment $R^*(X)$ among two well-separated fuzzy clusters which corresponds to the result, is received for the tolerance threthold $\alpha = 0\ 9925$ and a support of each fuzzy cluster $A^l_{(\alpha)}$, $l = 1, 2$ is formed by ten elements Fuzzy clusters prototypes are represented in Figure 1 by $\cdot$, elements of the first class are represented by $\circ$, and elements of the second class are represented by $\cdot$. Parameters of the regression line of the first class are $\vec{b}_1 = (b_{1\,0} = 0.4, b_{1\,1} = 0.7)$ and parameters of the regression line of the second class are

$\vec{b}_2 = (b_{2\,0} = 0.09, b_{2\,1} = 0.4)$. In other words, $y_1 = 0.4 + 0.7x_1$ is the regression equation for the first class and $y_2 = 0.09 + 0.4x_1$ is the regression equation for the second class.

5 Concluding Remarks

The result of the numerical experiment seems to be satisfactory. The result of application of the $D - AFC - TC$-algorithm to the data set shows that the $D - AFC - TC$-algorithm is an effective clustering procedure for classification and obtained classification results can be useful for forming of unknown c regression models.

The extracted unknown c prototypes of fuzzy clusters can be applied for solving the initial-value problem for the $FCRM$-algorithm.

Acknowledgements

I am grateful to Dr. Jan W. Owsinski and Prof. Janusz Kacprzyk for their interest in the investigation and support. I also thank to Mr. Aliaksandr Damaratski and Dr. Eduard Snezhko for elaborating experimental software.

References

[1] Hathaway R.J., Bezdek J.C. (1993). Switching Regression Models and Fuzzy Clustering. *IEEE Transactions on Fuzzy Systems.* Vol. **1**, pp. 195-204.

[2] Kaufmann A. (1975). *Introduction to the Theory of Fuzzy Subsets.* Academic Press, New York.

[3] Viattchenin D.A. (2004). A New Heuristic Algorithm of Fuzzy Clustering. *Control & Cybernetics.* Vol. **33**, pp. 323-340.

[4] Viattchenin D.A. (2007). Direct Algorithms of Fuzzy Clustering Based on the Transitive Closure Operation and Their Application to Outliers Detection. *Artificial Intelligence.* No. **3**, pp. 205-216. (in Russian)

[5] Viattchenin D.A. (2008). On Possibilistic Interpretation of Membership Values in Fuzzy Clustering Method Based on the Allotment Concept. *Proceedings of the Institute of Modern Knowledge.* No. **3**, pp. 85-90. (in Russian)

[6] Wu K.-L., Yang M.-S., Hsieh J.-N. (2010). Mountain c-Regressions Method. *Pattern Recognition.* Vol. **43**, pp. 86-98.

[7] Yang M.-S., Wu K.-L. (2005). A Modified Mountain Clustering Algorithm. *Pattern Analysis and Applications.* Vol. **8**, pp. 125-138.