ADAPTIVE FUZZY PROBABILISTIC CLUSTERING OF INCOMPLETE DATA

Yevgeniy Bodyanskiy, Alina Shafronenko, Valentyna Volkova

Abstract: *in the paper new recurrent adaptive algorithm for fuzzy clustering of data with missing values is proposed. This algorithm is based on fuzzy probabilistic clustering procedures and self-learning Kohonen's rule using principle "Winner-Takes-More" with Cauchy neighborhood function.*

Using proposed approach it's possible to solve clustering task in on-line mode in situation when the amount of missing values in data is too big.

Keywords*: fuzzy clustering, Kohonen self-organizing network, learning rule, incomplete data with missing values.*

ACM Classification Keywords*: 1.2.6 [Artificial Intelligence]: Learning – Connectionism and neural nets; 1.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search – Control theory; 1.5.1 [Pattern Recognition]: Clustering – Algorithms.*

Introduction

The problem of data sets described by vector-images clustering often occurs in many applications associated with Data Mining, but recently the focus on Fuzzy Clustering [Bezdek, 1981; Hoeppner, 1999; Xu, 2009], when processed vector-image with different levels of probabilities, possibilities or memberships, can belong to more than one class.

However, there are situations when the data sets contain missing values, the information that is lost. In this situation more effective is to use mathematical apparatus of Computational Intelligence [Rutkowski, 2008] and, first of all artificial neural networks [Marwala, 2009], that solve task of restoring the lost observations and modifications of the popular method of fuzzy c-means [Hathaway, 2001], which solve the problem of clustering without recovery of data.

Existing approaches for data processing with missing values [Zagoruyko, 1979; Zagoruyko, 1999], are efficient in cases when the massive of the original observations is given in batch form and does not change during the processing. At the same time, there is a wide class of problems in which the data that arrive to the processing, have the form of sequence that is feed in real time as it occurs in the training of Kohonen self-organizing maps [Kohonen, 1995] or their modifications [Gorshkov, 2009]. In this regard we introduced [Bodyanskiy, 2012] the adaptive neuro-fuzzy Kohonen network to solve the problem of clustering data with gaps based on the strategy of partial distances (PDS FCM). However, in situations where the number of such missing values is too big, the strategy of partial distances may be not effective, and therefore it may be necessary, along with the solution of fuzzy clustering simultaneously estimate the missing observations. In this situation, a more efficient is approach that is based on the optimal expansion strategy (OCS FCM) [Hathaway, 2001]. This work is devoted to the task of on-line data clustering using the optimal expansion strategy, adapted to the case when information is processed in a sequential mode, and its volume is not determined in advance.

Adaptive probabilistic fuzzy clustering of data with missing values based on the optimal expansion strategy

Baseline information for solving the task of clustering in a batch mode is the sample of observations, formed from N n-dimensional feature vectors $X = \{x_1, x_2, ..., x_N\} \subset R^n, x_k \in X, k = 1, 2, ..., N$. The result of clustering is the partition of original data set into m classes $(1 < m < N)$ with some level of membership $U_q(k)$ of k-th feature vector to the q-th cluster $(1 \le q \le m)$. Incoming data previously are centered and standardized by all features, so that all observations belong to the hypercube $[-1,1]^n$. Therefore, the data for clustering form array $\tilde{X} = \{\tilde{x}_1, ..., \tilde{x}_k, ..., \tilde{x}_N\} \subset R^n$, $\tilde{x}_k = (\tilde{x}_{k1}, ..., \tilde{x}_{ki}, ..., \tilde{x}_{kn})^T$, $-1 \le \tilde{x}_{ki} \le 1$, $1 < m < N$, $1 \le q \le m$, $1 \le i \le n$, $1 \le k \le N$ that is, all observations $\tilde{x}_{ki}$ are available for processing.

Introducing the objective function of clustering [Bezdek, 1981]

$$E(U_q(k), w_q) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^{\beta}(k) D^2(\tilde{x}_k, w_q)$$

with constraints $\sum_{q=1}^{m} U_q(k) = 1,\ 0 < \sum_{k=1}^{N} U_q(k) < N$ and solving standard nonlinear programming problem, we get the probabilistic fuzzy clustering algorithm [Hoeppner, 1999;Xu, 2009]

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{(D^2(\tilde{x}_k, w_q^{(\tau)}))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (D^2(\tilde{x}_k, w_l^{(\tau)}))^{\frac{1}{1-\beta}}}, \\ w_q^{(\tau+1)} = \dfrac{\sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta} \tilde{x}_k}{\sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta}}, \end{cases} \quad (1)$$

where w_q - prototype (centroid) of q-th cluster, $\beta > 1$ - parameter that is called fuzzyfier and defines "vagueness" of boundaries between classes, $D^2(\tilde{x}_k, w_q)$ - the distance between $\tilde{x}_k$ and w_q in adopted metric, $\tau = 0, 1, 2, ...$ - index of epoch of information processing which is organized as a sequence of $w_q^{(0)} \to U_q^{(1)} \to w_q^{(1)} \to U_q^{(2)} \to ...$. The calculation process continues until satisfy the condition

$$\left\| w_q^{(\tau+1)} - w_q^{(\tau)} \right\| \le \varepsilon \ \forall\ 1 \le q \le m,$$

(here ε - defines threshold of accuracy) or until the specified maximum number of epochs Q ($\tau = 0, 1, 2, ..., Q$).

Note also that when $\beta = 2$ and

$$D^2(\tilde{x}_k, w_q) = \left\| \tilde{x}_k - w_q \right\|^2,$$

we get a popular algorithm of Bezdek's fuzzy c-means (FCM)[Bezdek, 1981].

The process of fuzzy clustering can be organized in on-line mode as sequentially processing. At this situation batch algorithm (1) can be rewritten in recurrent form [Bodyanskiy, 2005]

$$\begin{cases} U_q(k+1) = \dfrac{(D^2(\tilde{x}_{k+1}, w_q(k)))^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(D^2(\tilde{x}_{k+1}, w_l(k)))^{\frac{1}{1-\beta}}}, \\ w_q(k+1) = w_q(k) + \eta(k+1)U_q^{\beta}(k+1)(\tilde{x}_{k+1} - w_q(k)), \end{cases} \quad (2)$$

where $\eta(k+1)$ - learning rate parameter, $U_q^{\beta}(k+1)$ - bell-shaped neighborhood function of neuro-fuzzy Kohonen network (Cauchy function), designed to solve the problems of fuzzy clustering [Gorshkov, 2009; Bodyanskiy, 2012], based on the principle "Winner Takes More» (WTM) [Kohonen, 1995].

In the presence of an unknown number of missing values in vector images $\tilde{x}_k$, that form array $\tilde{X}$, following [Hathaway, 2001], we introduce the sub-arrays:

$X_F = \{\tilde{x}_k \in \tilde{X} \mid \tilde{x}_k - vector\ containing\ all\ components\}$;

$X_P = \{\tilde{x}_{ki}, 1 \le i \le n, 1 \le k \le N \mid values\ \tilde{x}_k,\ available\ in\ \tilde{X}\}$;

$X_G = \{\tilde{x}_{ki} = ?, 1 \le i \le n, 1 \le k \le N \mid values\ \tilde{x}_k, absent\ in\ \tilde{X}\}$.

The optimal expansion strategy consists in the fact that the elements of sub-array X_G are considered as additional variables, which are estimated by minimization of objective function E. Thus, in parallel with clustering (optimization E by $U_q(k)$ and w_q) estimation of missing observations is made (optimization E by $\tilde{x}_{ki} \in X_G$). In this case, the algorithm of fuzzy c-means based on the optimal expansion strategy can be written as the following sequence of steps [Hathaway, 2001]:

1. Setting the initial conditions for the algorithm: $\beta > 0$; $1 < m < N$; $\varepsilon > 0$; $w_q^{(0)}$; $1 \le q \le m$; $\tau = 0,1,2,\ldots,Q$; $X_G^{(0)} = \{-1 \le \hat{x}_{ki}^{(0)} \le 1\}$, where $X_G^{(0)} - N_G (1 \le N_G \le (n-1)N)$ arbitrary initial estimates $\hat{x}_{ki}^{(0)}$ of missing values $\tilde{x}_{ki} \in X_G$;
2. Calculation of membership levels by solving the optimization problem:
$$U_q^{(\tau+1)}(k) = \underset{U_q(k)}{\arg\min} E(U_q(k), w_q^{(\tau)}, X_G^{(\tau)}) = \frac{(D^2(\hat{x}_k^{(\tau)}, w_q^{(\tau)}))^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(D^2(\hat{x}_k^{(\tau)}, w_l^{(\tau)}))^{\frac{1}{1-\beta}}} = \frac{(\left\|\hat{x}_k^{(\tau)} - w_q^{(\tau)}\right\|^2)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(\left\|\hat{x}_k^{(\tau)} - w_l^{(\tau)}\right\|^2)^{\frac{1}{1-\beta}}}$$
(here vector $\hat{x}_k^{(\tau)}$ differs from $\tilde{x}_k$ by replacing missing values $\tilde{x}_{ki} \in X_G$ by estimates $\hat{x}_{ki}^{(\tau)}$ that are calculated for the τ-th epoch of data processing);
3. Calculation the centroids of clusters:
$$w_q^{(\tau+1)} = \underset{w_q}{\arg\min} E(U_q^{(\tau+1)}(k), w_q, X_G^{(\tau)}) = \frac{\sum\limits_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}\hat{x}_k^{(\tau)}}{\sum\limits_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}};$$
4. Checking the stop conditions:
if $\left\|w_q^{(\tau+1)} - w_q^{(\tau)}\right\| < \varepsilon\ \forall\ 1 \le q \le m$ or $\tau = Q$, then the algorithm terminates, otherwise go to step 5;
5. Estimation of missing observations by solving the optimization problem:
$$X_G^{(\tau+1)} = \underset{X_G}{\arg\min} E(U_q^{(\tau+1)}(k), w_q^{(\tau+1)}, X_G)$$

or, equivalently

$$\frac{\partial E(U_q^{(\tau+1)}(k), w_q^{(\tau+1)}, X_G)}{\partial \hat{x}_{ki}} = 0,$$

That leads to

$$\hat{x}_{ki}^{(\tau+1)} = \frac{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k))^{\beta} w_{qi}^{(\tau+1)}}{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k))^{\beta}}.$$

Information processing with this algorithm is organized as a sequence

$$w_q^{(0)} \to U_q^{(1)} \to \hat{x}_{ki}^{(1)} \to w_q^{(1)} \to U_q^{(2)} \to \ldots \to w_q^{(\tau)} \to U_q^{(\tau+1)} \to \hat{x}_{ki}^{(\tau+1)} \to w_q^{(\tau+1)} \to \ldots \to w_q^{(Q)}$$

thus it is possible to organize on-line clustering by type of procedure (2). For this purpose we introduce two time scales: real time $k = 1, 2, \ldots, N, \ldots$, and accelerated computing time $\tau = 0, 1, 2, \ldots, Q$. Here we assume that between two instants of real time k and $k+1$ implemented Q iterations of accelerated time.

Then we can write procedure

$$\begin{cases} U_q^{(\tau+1)}(k+1) = \dfrac{(\left\| \hat{x}_{k+1}^{(\tau)} - w_q(k) \right\|^2)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (\left\| \hat{x}_{k+1}^{(\tau)} - w_l(k) \right\|^2)^{\frac{1}{1-\beta}}}, \\ \hat{x}_{k+1,i}^{(\tau+1)} = \dfrac{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta} w_{qi}(k)}{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta}}, \\ w_q(k+1) = w_q(k) + \eta(k+1)(U_q^{(Q)}(k+1))^{\beta} * \\ \qquad\qquad * (\hat{x}_{k+1}^{(Q)} - w_q(k)), \end{cases} \tag{3}$$

which shows that the memberships and missing observations are calculated in accelerated time, and centroids - in real time by WTM selflearning rule.

Of course centroids can be recalculated in accelerated time too:

$$\begin{cases} U_q^{(\tau+1)}(k+1) = \dfrac{(\left\| \hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1) \right\|^2)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (\left\| \hat{x}_{k+1}^{(\tau)} - w_l^{(\tau)}(k) \right\|^2)^{\frac{1}{1-\beta}}}, \\ w_q^{(0)}(k+1) = w_q^{(Q)}(k), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1)(U_q^{(\tau+1)}(k+1))^{\beta}(\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)), \\ \hat{x}_{k+1,i}^{(\tau+1)} = \dfrac{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta} w_{qi}^{(\tau+1)}(k+1)}{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta}}, \end{cases} \tag{4}$$

in this case anyway, both in (3) and (4) operation of summation about k is absent, that for large N can involve a lot of memory.

Experiments

Experimental research conducted on two samples of data such as Wine and Iris of UCI repository.

To estimate the quality of the algorithm we used quality criteria partitioning into clusters such as: Partition Coefficient (PC), Classification Entropy (CE), Partition Index (SC), Separation Index (S), Xie and Beni's Index (XB), Dunn's Index (DI) [Xu, 2009].

We also compared the results of our proposed algorithm with other more well-known such as Fuzzy C-means (FCM) clustering algorithm and Gustafson-Kessel clustering algorithm.

The proposed algorithm shown better results than the FCM and Gustafson-Kessel clustering algorithm.

Conclusion

The problem of probabilistic fuzzy adaptive clustering, containing a priori unknown number of gaps, based on the optimal expansion of data strategy is considered. The proposed algorithms are based on the recurrent optimization of a special type of goal functions. Missing observations are replaced by their estimates also obtained in the solution of optimization problem. Centroids of recovered clusters are tuned using a procedure close to the T.Kohonen WTM-rule with the function of the neighborhood (membership), having the Cauchian form.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Bezdek, 1981] J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981.

[Bodyanskiy, 2005] Ye. Bodyanskiy. Computational intelligence techniques for data analysis. Lecture Notes in Informatics. Bonn: GI, 2005, V. P-72, P. 15-36.

[Bodyanskiy, 2012] BodyanskiyYe., Shafronenko A., Volkova V. Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network. In “Artificial Intelligence Methods and Techniques for Business and Engineering Applications” – Rzeszow-Sofia: ITHEA, 2012. – P. 287-296.

[Gorshkov, 2009] Ye. Gorshkov, V. Kolodyazhniy, Ye. Bodyanskiy. New recursive learning algorithms for fuzzy Kohonen clustering network. Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronc Systems. (Rapperswil, Switzerland, June 21-24, 2009) Rapperswil, Switzerland, 2009, P. 58-61.

[Hathaway, 2001] R.J. Hathaway, J.C Bezdek. Fuzzy c-means clustering of incomplete data. IEEE Trans. on Systems, Man, and Cybernetics, №5, 31, 2001, P. 735-744.

[Hoeppner, 1999] F Hoeppner, F. Klawonn, R. Kruse, T. Runker. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition. Chichester, John Wiley &Sons, 1999.

[Kohonen, 1995] T. Kohonen. Self-Organizing Maps. Berlin: Springer-Verlag, 1995.

[Krishnapuram, 1993] R. Krishnapuram, J.M. Keller. A possibilistic approach to clustering. Fuzzy Systems, 1993, 1, №2, P.98-110.

[Marwala, 2009] T Marwala. Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques. Hershey-New York, Information Science Reference, 2009.

[Rutkowski, 2008] L.Rutkowski. Computational Intelligence.Methods and Techniques. Berlin-Heidelberg: Springer-Verlag, 2008.

[Xu, 2009] R. Xu, D.C. Wunsch. Clustering. Hoboken, N.J. John Wiley & Sons, Inc., 2009.

[Zagoruyko, 1979] N.G. Zagoruyko. Empirical predictions. Novosibirsk, Nauka, 1979 (in Russian).

[Zagoruyko, 1999] N.G. Zagoruyko. Applied Data Analysis and Knowledge. Novosibirsk, 1999 (in Russian).

Authors' Information

Yevgeniy Bodyanskiy – Professor, Dr. – Ing. habil., Scientific Head of Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, 14 Lenin Ave., Office 511, 61166 Kharkiv, Ukraine; e-mail: bodya@kture.kharkov.ua

Major Fields of Scientific Research: Artificial neural networks, Fuzzy systems, Hybrid systems of computational intelligence

Alina Shafronenko – Ph.D student in Artificial Intelligence dept., Kharkiv National University of Radioelectronics Lenin Ave., 14, Kharkiv, 61166, Ukraine; e-mail: alinashafronenko@gmail.com

Major Fields of Scientific Research: neural networks, neural network processing of data with gaps, fuzzy clustering, clustering of data

Valentyna Volkova *- Candidate of Technical Science (Ph.D.), Senior lecturer in Artificial Intelligence dept., Kharkiv National University of Radioelectronics Lenin Ave., 14, Kharkiv, 61166, Ukraine; e-mail: volkova@kture.kharkov.ua*

Major Fields of Scientific Research: neural networks, fuzzy clustering, clustering of data