UDK 519.7:004.8

Y. Bodyanskiy, A. Shafronenko

ON-LINE ROBUST FUZZY CLUSTERING BASED ON SIMILARITY MEASURE

Abstract. The problem of fuzzy adaptive on-line clustering of data distorted by outliers sequentially supplied to the processing when the original sample volume and the number of distorted observations are unknown is considered. The probabilistic and possibilistic clustering algorithms for such data, that are based on the similarity measure of a special kind that weakens or overwhelming outliers are proposed.

Index Terms. Computational and artificial intelligence, fuzzy neural networks, machine learning, multi-layer neural network.

Introduction

The problem of clustering data sets are often occurs in many practical problems, and its solution has been successfully used mathematical apparatus of computational intelligence [1] and first of all, artificial neural networks and soft computing methods (in the case of overlapping classes) is usually assumed that original array is specified a priori and processing is made in batch mode. Here as one of the most effective approach based on using FCM [2], which is reduced as a result to minimize the objective function with constraints of special form.

Real data often contains abnormal outliers of different nature, for example, measurement errors or distributions with "heavy tails". In this situation classic FCM is not effective because the objective function based on the Euclidean metric, only reinforces the impact of outliers. In such conditions it is advisable to use robust objective functions of special form [3], the overwhelming influence of outliers. For information processing in a sequential mode, in [4-6] adaptive procedures on-line fuzzy clustering have been proposed, which are in fact on-line modifications of FCM, where instead of the Euclidean metric robust objective functions, weaken the influence of outliers where used.

Statement of the Problem

The problem of robust on-line clustering based on objective functions is proposed using similarity measure of special form, which allows to synthesize efficient and numerically simple algorithms is proposed.

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 are previously 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.

We have developed numerically simple on-line procedure for partitioning sequentially fed to the data processing $\tilde{x}_k$ on m perhaps overlapping classes, while it is not known in advance whether $\tilde{x}_k$ undistorted or contains missing values and outliers. Furthermore, it is assumed that the amount of information under processing is not known in advance and is increased with time.

Adaptive Fuzzy Robust Data Clustering Based on Similarity Measure

As already mentioned, to solve the problem of fuzzy clustering of data containing outliers the special objective functions of the form [3-6] can be used, by some means these anomalies overwhelming, and the problem itself is associated with the minimization of these functions. From a practical point of view it is more convenient to use instead of the objective functions, based on the metrics, the so-called measures of similarity (SM) [7], which are subject to more lenient conditions than metrics:

$$\begin{cases} S(\tilde{x}_k,\tilde{x}_p)\ge 0, \\ S(\tilde{x}_k,\tilde{x}_p)=S(\tilde{x}_p,\tilde{x}_k), \\ S(\tilde{x}_k,\tilde{x}_k)=1\ge S(\tilde{x}_k,\tilde{x}_p). \end{cases}$$

(no triangle inequality), and clustering problem can be "tied" to maximize these measures.

If the data are transformed so that $-1\le\tilde{x}_{ki}\le 1$ the measure of similarity can be structured so as to suppress unwanted data lying at the edges of interval $[-1,1]$.

Figure 1 illustrates the use of similarity measure as the Cauchy function with different parameters width $\sigma^2<1$

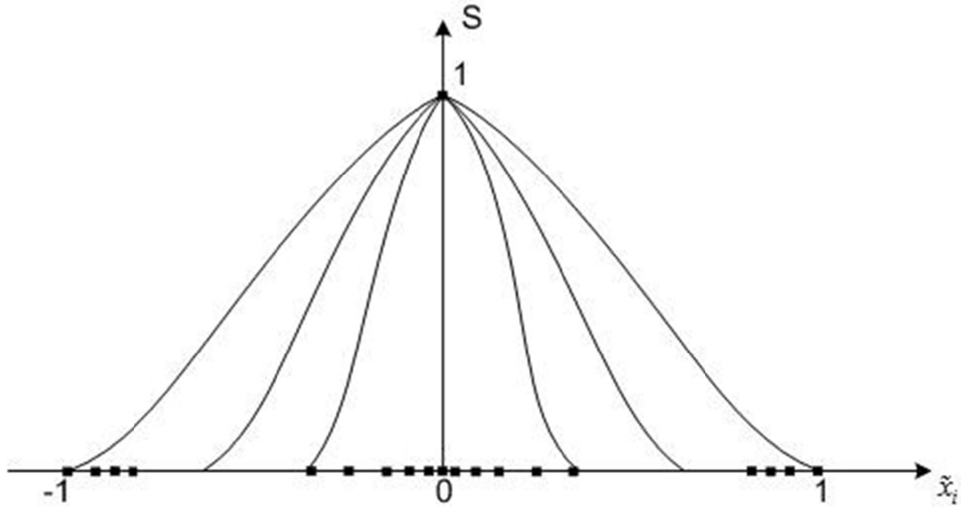

Fig. 1 Similarity measure based on the Cauchy

Picking up the width parameter σ^2 functions

$$S(\tilde{x}_k, w_q) = \frac{1}{1 + \frac{\left\| \tilde{x}_k - w_q \right\|^2}{\sigma^2}} = \frac{\sigma^2}{\sigma^2 + \left\| \tilde{x}_k - w_q \right\|^2} = \frac{\sigma^2}{\sigma^2 + D^2(\tilde{x}_k, w_q)}, \quad (1)$$

possible exclude the effect outliers, which in principle can not be done using the Euclidean metric

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

Further, by introducing the objective function based on similarity measure (1)

$$E_S(U_q(k), w_q) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^{\beta}(k) S(\tilde{x}_k, w_q) = \sum_{k=1}^{N} \sum_{q=1}^{m} \frac{U_q^{\beta}(k)\sigma^2}{\sigma^2 + \left\| \tilde{x}_k - w_q \right\|^2},$$

probabilistic constraints

$$\sum_{q=1}^{m} U_q(k) = 1,$$

Lagrange function

$$L_S(U_q(k), w_q, \lambda(k)) = \sum_{k=1}^{N}\sum_{q=1}^{m} \frac{U_q^{\beta}(k)\sigma^2}{\sigma^2 + \left\|\tilde{x}_k - w_q\right\|^2} + \\ + \sum_{k=1}^{N} \lambda(k)(\sum_{q=1}^{m} U_q(k) - 1) \quad (3)$$

(here $\lambda(k)$ - indefinite Lagrange multipliers) and solving the system of Karush-Kuhn-Tucker equations, we arrive at the solution

$$\begin{cases} U_q(k) = \dfrac{(S(\tilde{x}_k, w_q))^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(S(\tilde{x}_k, w_l))^{\frac{1}{1-\beta}}}, \\ \lambda(k) = -(\sum\limits_{l=1}^{m}(\beta S(\tilde{x}_k, w_l))^{\frac{1}{1-\beta}})^{1-\beta}, \\ \nabla_{w_q} L_S(U_q(k), w_q, \lambda(k)) = \sum\limits_{k=1}^{N} U_q^{\beta}(k) * \\ \quad * \dfrac{\tilde{x}_k - w_q}{(\sigma^2 + \left\|\tilde{x}_k - w_q\right\|^2)^2} = \vec{0}. \end{cases} \quad (4)$$

The last equation (4) has no analytic solution, so to find a saddle point of the Lagrangian (3) we can use the procedure of Arrow-Hurwitz-Uzawa, as a result of which we obtain the algorithm

$$\begin{cases} U_q(k+1) = \dfrac{(S(\tilde{x}_{k+1}, w_q))^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(S(\tilde{x}_{k+1}, w_l))^{\frac{1}{1-\beta}}}, \\ w_q(k+1) = w_q(k) + \eta(k+1) U_q^{\beta}(k+1) *, \\ \quad * \dfrac{\tilde{x}_{k+1} - w_q}{(\sigma^2 + \left\|\tilde{x}_{k+1} - w_q\right\|^2)^2} = \\ = w_q(k) + \eta(k+1)\varphi_q(k+1)(\tilde{x}_{k+1} - w_q) \end{cases} \quad (5)$$

where

 ISSN 1562-9945

$$\varphi_q(k+1) = \frac{\tilde{x}_{k+1} - w_q}{(\sigma^2 + \|\tilde{x}_{k+1} - w_q\|^2)^2}$$

neighbourhood robust functions of WTM-self-learning rule.

Assuming the value fazzifier as $\beta = 2$, we arrive at a robust variant of FCM:

$$\begin{cases} U_q(k+1) = \dfrac{(S(\tilde{x}_{k+1}, w_q))}{\sum\limits_{l=1}^{m} (S(\tilde{x}_{k+1}, w_l))}, \\ w_q(k+1) = w_q(k) + \eta(k+1) \dfrac{U_q^2(k+1)}{(\sigma^2 + \|\tilde{x}_{k+1} - w_q\|^2)^2}. \end{cases}$$

Further, using the concept of accelerated time, it's possible to introduce robust adaptive probabilistic fuzzy clustering procedure in the form

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{(S(\tilde{x}_k, w_q^{(\tau)}(k)))^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m} (S(\tilde{x}_k, w_l^{(\tau)}))^{\frac{1}{1-\beta}}}, \\ w_q^{(Q)}(k) = w_q^{(0)}(k+1), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1) * \\ \quad * \dfrac{(U_q^{(Q)}(k))^{\beta}}{(\sigma^2 + \|\tilde{x}_{k+1} - w_q^{(\tau)}(k+1)\|^2)^2} * \\ \quad * (\tilde{x}_{k+1} - w_q^{(\tau)}(k+1)). \end{cases} \qquad (6)$$

Similarly, it's possible to synthesize a robust adaptive algorithm for possibilistic [8] fuzzy clustering using criterion

$$E_S(U_q(k), w_q, \mu_q) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^{\beta}(k) S(\tilde{x}_k, w_q) + \\ + \sum_{q=1}^{m} \mu_q \sum_{k=1}^{N} (1 - U_q(k))^{\beta}.$$

Solving the problem of optimization, we obtain the solution:

$$\begin{cases} U_q(k+1) = \left(1 + \left(\dfrac{S(\tilde{x}_{k+1}, w_q(k))}{\mu_q(k)}\right)\right)^{-1}, \\ w_q(k+1) = w_q(k) + \eta(k+1) U_q^{\beta}(k+1) * \\ \quad * \dfrac{\tilde{x}_{k+1} - w_q(k)}{(\sigma^2 + \left\| \tilde{x}_{k+1} - w_q(k) \right\|^2)^2}, \\ \mu_q(k+1) = \dfrac{\sum_{p=1}^{k+1} U_q^{\beta}(p) S(\tilde{x}_p, w_q(k+1))}{\sum_{p=1}^{k+1} U_q^{\beta}(p)}, \end{cases} \quad (7)$$

receiving at $\beta = 2$ the form

$$\begin{cases} U_q(k+1) = \dfrac{1}{1 + \dfrac{S(\tilde{x}_{k+1}, w_q(k))}{\mu_q(k)}}, \\ w_q(k+1) = w_q(k) + \eta(k+1) \dfrac{U_q^2(k+1)}{(\sigma^2 + \left\| \tilde{x}_{k+1} - w_q(k) \right\|^2)^2} * \\ \quad * (\tilde{x}_{k+1} - w_q(k)), \\ \mu_q(k+1) = \dfrac{\sum_{p=1}^{k+1} U_q^2(p) S(\tilde{x}_p, w_q(k+1))}{\sum_{p=1}^{k+1} U_q^2(p)}. \end{cases}$$

And, finally, introducing the accelerated time we obtain the procedure

$$U_q^{(\tau+1)}(k) = \frac{1}{1 + \left(\dfrac{S(\tilde{x}_k, w_q^{(\tau)}(k))}{\mu_q^{(\tau)}(k)}\right)^{\frac{1}{\beta-1}}}, \quad (8a)$$

$$w_q^{(Q)}(k) = w_q^{(0)}(k+1), \quad (8b)$$

$$w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1) \frac{(U_q^{(Q)}(k))^{\beta}}{(\sigma^2 + \left\| \tilde{x}_{k+1} - w_q^{(\tau)}(k+1) \right\|^2)^2} (\tilde{x}_{k+1} - w_q^{(\tau)}(k+1)), \quad (8c)$$

ISSN 1562-9945

$$\mu_q^{(\tau+1)}(k) = \frac{\sum_{p=1}^{k} (U_q^{(\tau+1)}(p))^{\beta} S(\tilde{x}_p, w_q^{(\tau+1)}(k))}{\sum_{p=1}^{k} (U_q^{(\tau+1)}(p))^{\beta}}. \quad (8d)$$

Experimental Research

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

Iris. This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. Predicted attribute: class of iris plant. This is an exceedingly simple domain. Number of Instances: 150 (50 in each of three classes). Number of Attributes: 4 numeric, predictive attributes (sepal length in cm, sepal width in cm, petal length in cm, petal width in cm) and the 3 classes: Iris Setosa, Versicolour, Virginica.

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).

Partition Coefficient (PC): measures the amount of "overlapping" between clusters.

Classification Entropy (CE): it measures the fuzzyness of the cluster partition only, which is similar to the Partition Coefficient.

Partition Index (SC): is the ratio of the sum of compactness and separation of the clusters. It is a sum of individual cluster validity measures normalized through division by the fuzzy cardinality of each cluster. SC is useful when comparing different partitions having equal number of clusters. A lower value of SC indicates a better partition.

Separation Index (S): on the contrary of partition index (SC), the separation index uses a minimum-distance separation for partition validity.

Xie and Beni's Index (XB): it aims to quantify the ratio of the total variation within clusters and the separation of clusters. The optimal number of clusters should minimize the value of the index.

Dunn's Index (DI): this index is originally proposed to use at the identification of "compact and well separated clusters". So the result of the clustering has to be recalculated as it was a hard partition algorithm.

The experimental results are presented in the table 1.

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

Table I

Results of Experiments

Algorithms	Iris UCI repository					
	PC	CE	SC	S	XB	DI
Adaptive fuzzy robust clustering data based on similarity measure	0,0199	0,0122	-2439807	0,0022	0,0015	1
FCM	0,8011	0,3410	0,2567	0,0030	7,1965	0,0080

Fig. 2 shows the work of data clustering algorithm that is not corrupted by abnormal outliers.

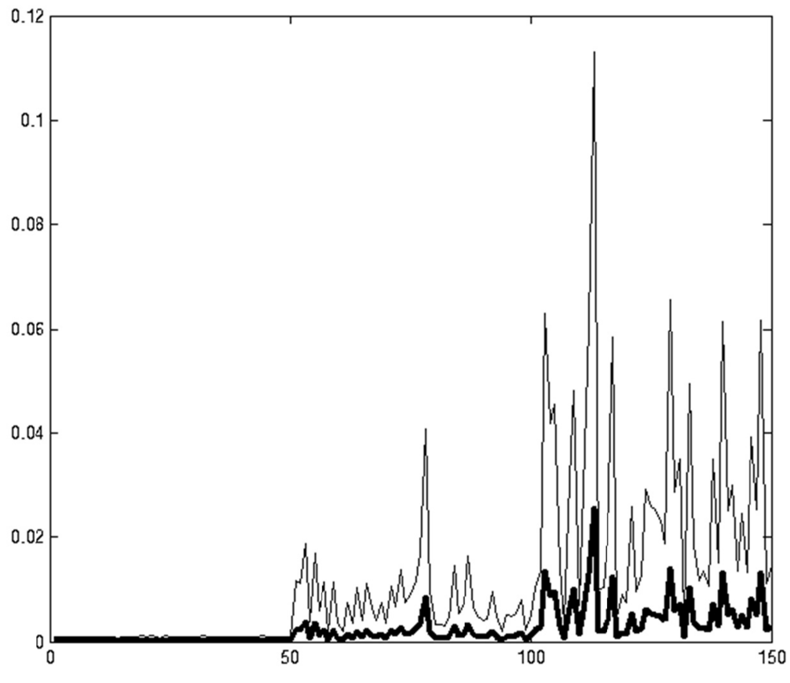

Fig. 2. Data clustering that is not corrupted by abnormal outliers, where solid line is membership level; bold line is function of similarity measure.

The outliers were added manually. Next figure (fig. 3) shows sensitivity to outliers in the data set Iris.

 ISSN 1562-9945

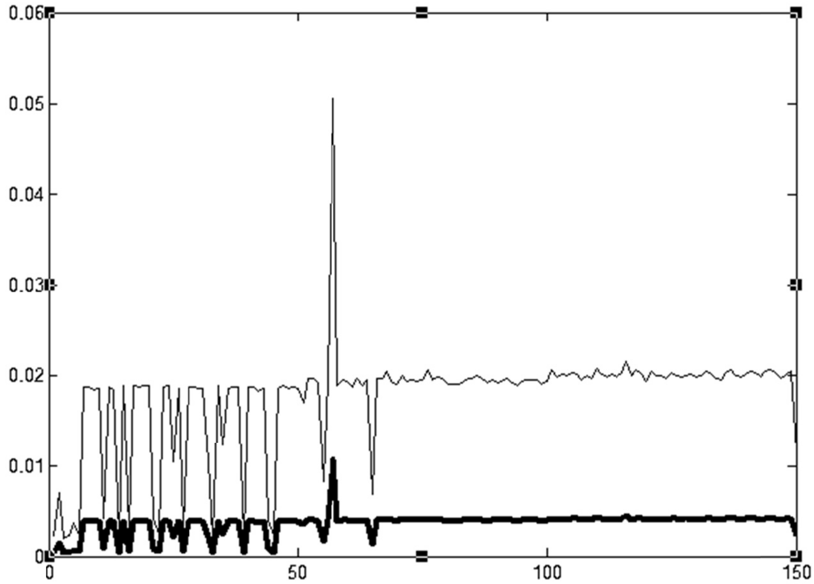

Fig. 3. Sensitivity to outliers in the data set Iris, where solid line is membership level; bold line is function of similarity measure.

Conclusions

There is a group of robust adaptive fuzzy clustering algorithms introduced, allowing in on-line mode to process distorted data containing outliers. The basis of the proposed algorithms is using of classical procedures as fuzzy c-means of J. Bezdek, T. Kohonen self-learning, as well as specially introduced similarity measure allowing to work with distorted information. The algorithms are simple in numerical implementation, being essentially gradient optimization procedures for objective functions of special form.

REFERENCES

1. L.Rutkowski. Computational Intelligence. Methods and Techniques. Berlin-Heidelberg: Springer-Verlag, 2008.-514p.
2. J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. – N.Y.: Plenum, 1981. – 272p.
3. R.N. Dave, R. Krishnapuram. Robust cluatering methods: A unified view// IEEE Trans. on Fuzzy Systems. – 1997. – 5. - №2. – P.270-293.
4. BodyanskiyYe., Gorshkov Ye., Kokshenev I., Kolodyazhniy V. Robust recursive fuzzy clustering algorithms//Proc. East West Fuzzy Coll. – Zittau/Goerlitz: HS, 2005. – P. 301-308.
5. BodyanskiyYe. Computational Intelligence Techniques for Data Analisis. – Vol. P. – 72. – Bonn: GI, 2005. – P. 15-36.

6. Kokshenev I., BodyanskiyYe., Gorshkov Ye., Kolodyazhniy V. Outlier resistant recursive fuzzy clustering algorithm / In "Computational Intelligence: Theory and Application". – Ed. by B.Reusch-Advances in Soft Computing, Vol. 38. – Berlin-Heidelberg: Springer-Verlag, 2006. – P. 647-652.
7. Sepkovski J.J. Quantified coefficients of association and measurement of similarity // J. Int. Assoc. Math. – 1974. – 6. – №2. – P. 135-152.
8. L. Keller, R. Krishnapuram, N.R. Pal. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. – N.Y.:Springer Science + Business Media, Inc., 2005. – 776p.

 ISSN 1562-9945