17. R. C. O'Reilly and Y. Munakata, Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain. Cambridge, MA: MIT Press, 2000. 18. A. Lazar, G. Pipa, and J. Triesch, Fading memory and time series prediction in recurrent networks with different forms of plasticity. Neural Networks, vol. 20, no. 3, pp. 312–322, Apr. 2007. 19. A. Cichocki and R. Unbehauen, Neural Networks for Optimization and Signal Processing. Chichester: Wiley, 1993. 20. J. Wang, "Analysis and design of a k-winners-take-all model with a single state variable and the Heaviside step activation function", IEEE Trans. Neural Networks, vol. 21, no. 9, pp. 1496–1506, Sept. 2010. 21. P. V. Tymoshchuk, "A dynamic K-winners take all analog neural circuit," in Proc. IVth IEEE Int. Conf. Perspective technologies and methods in MEMS design, L'viv, 2008, pp. 13–18. 22. Q. Liu and J. Wang, "Two k-winners-take-all networks with discontinuous activation functions", Neural Networks, vol. 21, pp. 406–413, Mar. – Apr. 2008. 23. Тимощук П. Аналогова структурно-функціональна нейронна схема визначення максимальних сигналів // Комп'ютерні науки та інформаційні технології. – 2012. – № 744. – С. 10–17. (Вісн. Нац. ун-ту "Львівська політехніка"). 24. Тимощук П. Математична модель нейронної схеми типу "K-Winners-Take-All" обробки дискретизованих сигналів // Вісн. Нац. ун-ту "Львівська політехніка" "Комп'ютерні системи проектування. Теорія і практика". – 2010. – № 685. – С. 45–50 (). 25. B. D. Calvert and C. A. Marinov, "Another K-winners-take-all analog neural network", IEEE Trans. Neural Networks, vol. 4, no. 1, pp. 829–838, Jul. 2000.

UDC 004.032.26

I. Perova, Ye. Bodyanskiy
Kharkiv National University of Radio Electronics

ADAPTIVE FUZZY CLUSTERING BASED ON MANHATTAN METRICS IN MEDICAL AND BIOLOGICAL APPLICATIONS

Розглянуто алгоритм нечіткої кластеризації даних за наявності аномальних спостережень. Запропонований рекурсивний алгоритм нечіткої кластеризації даних ґрунтується на використанні манхеттенської метрики, що забезпечує високу швидкість обробки інформації та просту обчислювальну реалізацію. Результат апробації на даних медико-біологічних досліджень підтверджує ефективність запропонованого підходу.

Ключові слова: алгоритм нечіткої кластеризації, манхеттенська метрика, функція Лагранжа.

The problem of fuzzy clustering on the basis of the probabilistic fuzzy approach under the presence of outliers in data is considered. Recursive fuzzy clustering algorithm is proposed, which optimizes the objective function based on Manhattan metrics provides high speed of information processing and simple computational realization. The results of real data clustering confirm the effectiveness of proposed approach in medical data mining tasks.

Key words: fuzzy clustering algorithm, Manhattan metrics, Lagrange function

Introduction

Clustering and classification of datasets of different nature are now key problems of data mining, and effective solving of this tasks is important for knowledge acquisition by analysis of observations.

Generally, cluster analysis is algorithmic basis of data classification by means of separation of the available data into a number of classes (clusters) without a priori defined membership of any observation sample to one of the class (unsupervised learning). In the traditional (crisp) approach it is assumed that every observation belongs to only one class. The k-means algorithm [1] and the nearest-neighbor rule [2] are most popular examples of this approach. It is much more natural to assume that every observation may

belong to several clusters at same time with certain degrees of membership. This assumption is the basis of fuzzy cluster analysis [3, 4]. At present time many fuzzy clustering approaches are widely used, such as Bezdek's fuzzy c-means [3], the Gustafson-Kessel algorithm [5], fuzzy k-nearest neighbors [6], fuzzy shell cluster analysis by Klawonn-Kruse-Timm [7], mountain clustering by Yager and Filev [8] e.a. The approaches mentioned above are capable of efficient data clustering when the clusters are overlapping, but only with the assumption that the clusters are compact, i.e. they do not have abrupt (anomalous) outliers. Whereas real datasets usually contain up to 20% of outliers [9-11], the assumption of clusters compactness may sometimes become inadequate.

This situation often happens when processing medical and biological data sets because human subjective factor plays important role in these tasks.

The source information for all the mentioned algorithms is the data set of N n-dimentional feature vectors $X=\{x(1),x(2),...,x(N)\}$, $x(k)\in R^n$, $k=1,2,...,N$. The output of the algorithm is the separation of the original data into m clusters with some degree of membership $m_q(x(k))$ of the k-th vector to the q-th cluster.

In this paper, we make an attempt to derive an adaptive computationally simple stable fuzzy clustering algorithm for recursive data processing in online mode as more and more data become available, using Manhattan metrics.

Stable probabilistic fuzzy clustering algorithm

Probabilistic fuzzy-clustering approach belong to a class of objective function based algorithm [3] that are designed to solve fuzzy clustering problem via the optimization of a certain predetermined clustering criterion, and are the best-grounded from the mathematical point of view.

For pre-standardized feature vector (the standartization is performed component-wise so that all the feature vectors belong to the hypercube $[-1,1]^n$), the objective function is

$$E^k=\sum_{k=1}^{N}\sum_{q=1}^{m}m_q^b\left(x(k)\right)d\left(x(k),c_q\right) \tag{1}$$

subject to constraints

$$\sum_{q=1}^{m}m_q\left(x(k)\right)=1,\ k=1,...,N, \tag{2}$$

$$0<\sum_{k=1}^{N}m_q\left(x(k)\right)<N,\ q=1,...,m \tag{3}$$

Here $m_q\left(x(k)\right)\in[0,1]$ is the degree of membership of the vector $x(k)$ to the q-th cluster, c_q is the prototype (center) of the q-th cluster, β is a non-negative parameter, referred to "fuzzifier" (usually β=2), $d\left(x(k),c_q\right)$ is the distance between $x(k)$ and c_q in the adopted metrics. The result of clustering is assumed to be $N\times m$ matrix $W=\left\{m_q\left(x(k)\right)\right\}$, referred to as "fuzzy partition matrix".

Note that since the elements of the matrix W can be regarded as the probabilities of the hypotheses of data vector membership to certain clusters, the procedures generated from (1) subject to constraints (2), (3) are referred to as the "probabilistic clustering algorithms".

The distance function $d\left(x(k),c_q\right)$ is usually assumed to be Minkowski L^p metrics [17]

$$d\left(x(k),c_q\right)=\left(\sum_{i=1}^{n}\left|x_i(k)-c_{qi}\right|^p\right)^{\frac{1}{p}},\ p\geq 1 \tag{4}$$

where $x_i(k)$, c_q are i-th components of $(n\times 1)$-vectors $x(k)$, c_q respectively. Assuming β=p=2 leads to the most popular, simple and quite effective Bezdek's fuzzy c-means algorithm [3]

$$m_q\left(x(k)\right)=\frac{\left\|x(k)-c_q\right\|^{-2}}{\sum\limits_{l=1}^{N}\left\|x(k)-c_l\right\|^{-2}}, \tag{5}$$

$$c_q=\frac{\sum\limits_{k=1}^{N}m_q^2\left(x(k)\right)x(k)}{\sum\limits_{k=1}^{N}m_q^2\left(x(k)\right)}. \tag{6}$$

Simplicity of (5) and (6) is determined by using of Euclidean (quadratic) metrics, those derivates on the estimated parameters are linear forms. It allows to obtain a solution in simple analytic form.

At the same time in medical tasks it is more naturally to use Manhattan metrics (p=1 in (4)), i.e.

$$d\left(x(k),c_q\right)=\sum_{i=1}^{n}\left|x_i(k)-c_{qi}\right|=\left|x(k)-c_q\right| \tag{7}$$

whose gradient respectively c_q has the form

$$\nabla_{c_q}d\left(x(k),c_q\right)=-sign\left(x(k)-c_q\right) \tag{8}$$

where

$$sign\left(x(k)-c_q\right)=\left(sign\left(x_1(k)-c_{q1}\right),...,sign\left(x_i(k)-c_{qi}\right),...,sign\left(x_n(k)-c_{qn}\right)\right)^T.$$

By introducing the goal function of probabilistic fuzzy clustering

$$E\left(m_q(k),c_q\right)=\sum_{k=1}^{N}\sum_{q=1}^{m}m_q^b\left(x(k)\right)\sum_{i=1}^{n}\left|x_i(k)-c_{qi}\right|=\sum_{k=1}^{N}\sum_{q=1}^{m}m_q^b\left(x(k)\right)\left|x(k)-c_q\right|=$$

$$=\sum_{k=1}^{N}\sum_{q=1}^{m}m_q^b\left(x(k)\right)d\left(x(k),c_q\right) \tag{9}$$

and taking into consideration the constraints (2) we can write the Lagrange function

$$L\left(m_q(k),c_q,l(k)\right)=\sum_{k=1}^{N}\sum_{q=1}^{m}m_q^b\left(x(k)\right)\sum_{i=1}^{n}\left|x_i(k)-c_{qi}\right|+\sum_{k=1}^{N}l(k)\left(\sum_{q=1}^{m}m_q(k)-1\right), \tag{10}$$

where $\lambda(k)$ is an undetermined Lagrange multiplier that guarantees the fulfillment of the constraints (2), (3). The saddle point of the Lagrange function (10) could be found solving the following system of Karush-Kuhn-Tucker equations

$$\begin{cases}\dfrac{\partial L\left(m_q(k),c_q,l(k)\right)}{\partial m_q(k)}=0,\\ \dfrac{\partial L\left(m_q(k),c_q,l(k)\right)}{\partial l(k)}=0,\\ \nabla_{c_q}L\left(m_q(k),c_q,l(k)\right)=\mathbf{0}.\end{cases} \tag{11}$$

Solving the first and the second equations of the system (11) leads to well-known result

$$\begin{cases}m_q\left(x(k)\right)\dfrac{d\left(\mathrm{x}(k),c_q\right)^{\frac{1}{1-b}}}{\sum\limits_{l=1}^{m}\left(d\left(x(k),c_l\right)\right)^{\frac{1}{1-b}}}=0,\\ l(k)=-\left(\sum\limits_{l=1}^{m}\left(b\,d\left(x(k),c_l\right)\right)^{\frac{1}{1-b}}\right)^{1-b}.\end{cases} \tag{12}$$

but the third one

$$\nabla_{c_q} L\left(m\left(x(k)\right), c_q, l\left(k\right)\right) = \sum_{k=1}^{N} m_q^b\left(x(k)\right) \nabla_{c_q} d\left(x(k), c_q\right) = \mathbf{r} 0 \tag{13}$$

obviously has no analytical solution. The solution of (13) could be computed with use of a local modification of Lagrange function [12] and the recursive fuzzy clustering algorithms [22]. Furthermore, searching the saddle point of the local Lagrange function

$$L\left(m_q\left(x(k)\right), c_q, l\left(k\right)\right) = \sum_{q=1}^{m} m_q^b\left(x(k)\right) d\left(x(k), c_q\right) + l\left(k\right)\left(\sum_{q=1}^{m} m_q\left(x(k)\right) - 1\right) \tag{14}$$

using the Arrow-Hurwitz-Uzawa procedure gives the following algorithm:

$$\begin{cases} m_q\left(x(k)\right) \dfrac{d\left(\mathrm{x}(k), c_q^{(k)}\right)^{\frac{1}{1-b}}}{\sum_{l=1}^{m}\left(d\left(x(k), c_l^{(k)}\right)\right)^{\frac{1}{1-b}}}, \\ c_{qi}\left(k+1\right) = c_{qi}\left(k\right) - h\left(k\right) \dfrac{\partial L_k\left(m_q\left(x(k)\right), c_q, l\left(k\right)\right)}{\partial c_{qi}} = \end{cases} \tag{15}$$

$$= c_{qi}\left(k\right) + h\left(k\right) m_q^b\left(x(k)\right) sign\left(x_i(k) - c_{qi}\left(k\right)\right)$$

where $h\left(k\right)$ is the learning rate parameter, $c_{qi}\left(k\right)$ is the *i*-th component of the *q*-th prototype vector calculated at the *k*-th step, or the same in vector form

$$\begin{cases} m_q\left(x(k)\right) \dfrac{d\left(\mathrm{x}(k), c_q\left(k\right)\right)^{\frac{1}{1-b}}}{\sum_{l=1}^{m}\left(d\left(x(k), c_l\left(k\right)\right)\right)^{\frac{1}{1-b}}}, \\ c_q\left(k+1\right) = c_q\left(k\right) + h\left(k\right) m_q^b\left(x(k)\right) sign\left(x(k) - c_q\left(k\right)\right), \end{cases} \tag{16}$$

that from computational point of view is essentially simpler that the robust fuzzy clustering algorithm, proposed in [23].

Especially simple form this algorithm obtains when β=2

$$\begin{cases} m_q\left(x(k)\right) \dfrac{\left|\mathrm{x}(k) - c_q\left(k\right)\right|^{-1}}{\sum_{l=1}^{m}\left|\mathrm{x}(k) - c_l\left(k\right)\right|^{-1}}, \\ c_q\left(k+1\right) = c_q\left(k\right) + h\left(k\right) m_q^2\left(x(k)\right) sign\left(x(k) - c_q\left(k\right)\right). \end{cases} \tag{17}$$

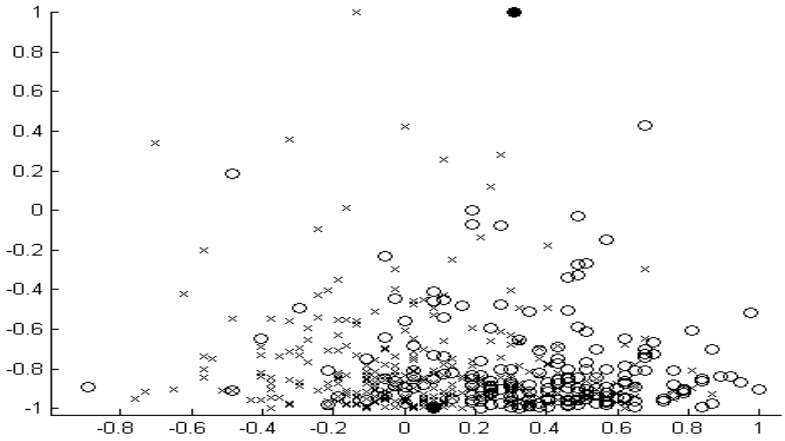

Fig. 1

Experiments

We have used the proposed algorithm in the problem of data clustering of data set from the UCI machine learning database: "heart-disease" [24]. This data set contains 2 clusters. On fig.1 the results of clustering labeled as ‘o’ and ’x’ are shown using two-dimensional projection from nine-dimensional of 1 and 3 properties. Centers of clusters was labeled as ‘•’. Using adaptive fuzzy clustering based on Manhattan metrics provides satisfactory quality of clustering that is better than quality of clustering based on standard fuzzy-c-means algorithm.

Conclusion

In the paper stable adaptive probabilistic fuzzy clustering algorithm based on the objective function of a special form (Manhattan metrics), suitable for heavy-tailed data distribution with outliers, is proposed. The algorithm could be used in a wide range of applications, such as medical data mining, fault detection, pattern recognition in self-organizing mode when the size of the data set is not known a priori, and the data must be processed in sequential mode, those is typical for medical and biological research.

1. MacQueen, J. On convergence of k-means and partitions with minimum average variance. – Ann. Math. Statist., 36, 1965 . – 1084 p. 2. Cover, T.M. Estimates by the nearest-neighbor rule. IEEE Trans. on Information Theory, 14, 1968. – 50–55 p. 3. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. – Plenum Press, New York, 1981. 4. Höppner, F., Klawonn, F., Kruse, R.: Fuzzy-Clusteranalyse. Verfahren für die Bilderkennung, Klassifikation und Datenanalyse. – Vieweg, Braunschweig, 1996. 5. Gustafson, E.E., Kessel, W.C.: Fuzzy clustering with a fuzzy covariance matrix. – Proc. IEEE CDC, San Diego, California, 1979. – 761–766 p. 6. Keller, J.M., Gray, M.R., Givens, J.A., Jr.: A fuzzy k-nearest neighbor algorithm. – IEEE Trans. on Syst., Man and Cybern. 3, 1985. – 32–57 p. 7. Klawonn, F., Kruse, R., Timm, H.: Fuzzy Shell Cluster Analysis. In: Della Riccia, G., Lenz, H.J., Kruse, R. (eds.): Learning, Networks and Statistics. – Springer-Verlag, Wien, 1997. – 105–120 p. 8. Yager, R.R., Filev, D.P.: Approximate clustering via the mountain method. – IEEE Trans. on Syst., Man and Cybern. 24, 1994. – 1279–1284 p. 9. Barnett, V., Lewis, T.: Outliers in Statistical Data. – John Wiley & Sons, Chichester-New York-Brisbane-Toronto, 1978. 10. Rey, W.J.J.: Robust Statistical Methods. Lecture Notes in Mathematics, Vol. 690. – Springer-Verlag, Berlin-Heidelberg-New York , 1978. 11. Huber, P.J. Robust Statistics. – John Wiley & Sons, New York, 1981. 12. Looney, C.G. A fuzzy clustering and fuzzy merging algorithm. – Technical Report, CS-UNR-101-1999. – URL: http://sherry.ifi.unizh.ch/looney99fuzzy.html 13. Looney, C.G. A fuzzy classifier with ellipsoidal Epanechnikovs. – Technical Report, Computer Science Department, University of Nevada, Reno, NV, 2001. – URL: http://sherry.ifi.unizh.ch/looney01fuzzy.html 14. Tsuda, K., Senda, S., Minoh, M., Ikeda, K.: Sequential fuzzy cluster extraction and its robustness against noise. – Systems and Computers in Japan, 28, 1997. – 10-17p. 15. Höppner, F., Klawonn, F.: Fuzzy clustering of sampled functions. Proc. 19th Int. Conf. of the North American Fuzzy Information Processing Society (NAFIPS), Atlanta, USA, 2000. – 251–255 p. 16. Georgieva, O., Klawonn, F.: A clustering algorithm for identification of single clusters in large data sets. – Proc. 11th East-West Fuzzy Colloquium. – HS Zittau-Görlitz, 2004. – 118–125 p. 17. Pau, L.F.: Failure Diagnosis and Performance Monitoring. – Marcel Dekker Inc., NY, 1981. 18. Tsypkin, Ya.Z.: Foundations of Information Theory of Identification. – Nauka, Moscow, 1984. 19. Holland, P.W., Welsh, R.E.: Robust regression using iteratively re-weighted least squares. – Comm. Statist. Theory and Methods A6, 1977. – 813-827p. 20. Welsh, R.E. Nonlinear statistical data analysis. – Proc. Comp. Sci. and Statist. Tenth Ann. Symp. Interface. Nat’l Bur. Stds. Gaithersburg, MD, 1977. – 77–86 p. 21. Chichocki, A., Unbehauen, R.: Neural Networks for Optimization and Signal Processing. Teubner, Stuttgart (1993). 22. Bodyanskiy, Ye., Kolodyazhniy, V., Stephan, A.: Recursive fuzzy clustering algorithms. Proc. 10th East-West Fuzzy Colloquium. HS Zittau-Görlitz, 2002. – 276–283 p. 23. Bodyanskiy Ye., Gorshkov Ye., Kokshenev I., Kolodyazhniy V. Outlier resitant recursive fuzzy clustering algorithms. – Proc. 12-th East-West Fuzzy Colloquium. HS Zittan-Görlitz, 2005. – 301–308 p. 24. David W. Aha UCI Repository of machine learning databases. – URL: https://archive.ics.uci.edu/ml/datasets/Heart+Disease CA: University of California, Department of Information and Computer Science, 1988.