Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 1, Number 8; October, 2014 pp. 87-90

http://www.krishisanskriti.org/jbaer.html

Fuzzy C-Means in Finding Available Structure

R. Devi[1], S.R. Kannan[2], S. Ramathilagam[3]

[1,2]*Dept. of Mathematics, Pondicherry University, Pondicherry*
[3]*Dept. of Mathematics, Periyar Govt. Arts College, Cuddalore, Tamil Nadu*

Abstract: **Finding available structures in high dimensional database is a difficult task, due to vagueness spread in the actual value of objects of database. Researchers have invented numbers of clustering techniques in selecting the relevant objects of available structures of high dimensional database, but the techniques have failed to provide proper outcome results with less error. Therefore this paper aims to present suitable clustering techniques which can capable in finding perfect structures. This paper introduces the effective clustering techniques in the combination of fuzzy c-means, and typicality of possibilistic c-means. The performance of proposed method is evaluated through the experimental work on benchmarks database.**

Keywords: **Fuzzy C-Means, Kernel Distances, Uncertain Objects, Cancer Databases.**

1. INTRODUCTION

Due to large attributes of object in high dimensional database, the sparseness difference occurred between the data objects. Further the uncertainty objects of high dimensional database contain incomplete information. Therefore, the clustering techniques have failed to find the dissimilarity between the uncertain objects in high dimensional real world database for analyzing available information in the database [2, 9]. . In order to provide proper results with high dimensional uncertain objects of database, recently fuzzy set based fuzzy clustering [1, 3, 6, 7, 8, 10, 11, 17] has been implemented effectively. The membership of fuzzy set can work well with the uncertain database [25]. The fuzzy set based fuzzy clustering allows gradual memberships to the data points to place an object in all clusters.

This gives the flexibility to express that data points belong to more than one cluster at the same time. The memberships offer a much finer degree of detail of the data model to cluster it into several groups and memberships can also express how ambiguously or definitely a data point should belong to a cluster [4, 5]. Even though there are lots of benefits using fuzzy c-means algorithms, it has considerable drawbacks such as the result of clustering process deteriorates while uncertainty exists in the high dimensional database [12, 13, 14, 15, 19, 24]. This paper attempts to provide more suitable clustering techniques using fuzzy clustering methods for analyzing high–dimensionality database. In order to find an effective membership for the points which have equidistant for two clusters this paper obtained the possiblistic c-means based objective function of fuzzy c-means. The performance of obtaining membership to the noisy object is improved by relaxing the membership constraints using the typicality of the possibilistic c-means [16, 18, 20, 21, 22, 23, 26].

The rest of this paper is organized as follows. Section 2 contains the proposed algorithm. The experimental results on Checkerboard Dataset are reported in Section 3. Section 4 provides conclusion of this paper.

2. 2. PROPOSED ALGORITHM

In this subsection we introduce effective fuzzy clustering technique to find the similar patterns or subtypes of cancers in high – dimensional cancer database which is corrupted by similar intensities between objects, missing values and other noises by scanning process of gene expression. This paper incorporates fuzziness weighting exponent, the expression of possibilistic typical weighting exponent (τ) and tangent kernel induced distance with the objective of proposed fuzzy c-means to capture the meaningful information from cancer database. The proposed objective function of Tangent Fuzzy Possibilistic C-Means is given by

$$J_{\text{TFPCM}}(U,V)=2\sum_{k=1}^{n}\sum_{i=1}^{c}\left(u_{ik}^{m}+\tau_{ik}^{\eta}\right)\left(1-T_B(x_k,v_i)\right) \qquad (1)$$

Where $T_B(x_k,v_i)=1-\tanh\left(\frac{-B(x_k,v_i)}{\sigma^2}\right)$, $B(x_k,v_i)=\frac{|x_k-v_i|^2}{|x_k+v_i|^2}$, and the

T_B represents tangent bary curtis kernel induced distance. m & η in (5) are weighting exponents. The weighting exponents compute the amount of fuzziness in the resulting classification in order to obtain proper center of cluster from the database which has similar gene expression. By minimizing the equation (1) we have obtained the degrees of membership, typicality and the cluster centers. To minimize the equation (1) subject to the conditions, the Lagrangian multiplier rule is used.

Optimizing the equation (1), we have obtained a generalized membership equations u_{ik} and typicality τ_{ik} forthe iterative solution of an objective function.

$$\Rightarrow u_{ik} = \frac{\left(\dfrac{1}{1-T_B(x_k, v_i)}\right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left(\dfrac{1}{1-T_B(x_k, v_j)}\right)^{\frac{1}{m-1}}} \qquad (2)$$

The typicality τ_{ik} is as:

$$\Rightarrow \tau_{ik} = \frac{\left(\dfrac{1}{(1-T_B(x_k, v_i))}\right)^{\frac{1}{\eta-1}}}{\sum_{l=1}^{n}\left(\dfrac{1}{(1-T_B(x_l, v_i))}\right)^{\frac{1}{\eta-1}}} \qquad (3)$$

Optimizing the equation (1), this paper obtains the equations for updating the cluster center or prototypes of TFPCM.

$$v_i^t = \frac{\sum_{K=1}^{n}(\mu_{ik}^m + \tau_{ik}^\eta)[T_B(x_k, v_i^{t-1})]T_B'(x_k, v_i^{t-1})B_d'(x_k, v_i^{t-1})x_k}{\sum_{K=1}^{n}(\mu_{ik}^m + \tau_{ik}^\eta)[T_B(x_k, v_i^{t-1})]T_B'(x_k, v_i^{t-1})B_d(x_k, v_i^{t-1})} \qquad (4)$$

where t represents the iteration count,

$$T_B(x_k, v_i) = 1 - \tanh\left(\frac{-B(x_k, v_i)}{\sigma^2}\right),$$

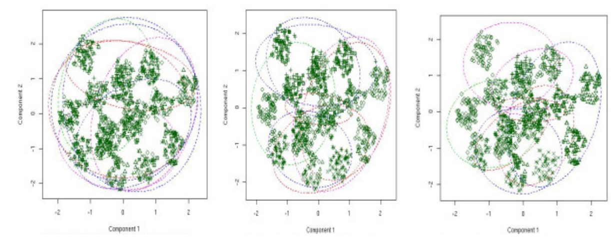
$$B(x_k, v_i) = \frac{|x_k - v_i|^2}{|x_k + v_i|^2},$$

and

$$B_d(x_k, v_i) = \frac{\left[|x_k + v_i|^2 - |x_k - v_i|^2\right]}{\left(|x_k + v_i|^2\right)^2}, \quad B_d'(x_k, v_i) = \frac{\left[|x_k + v_i|^2 + |x_k - v_i|^2\right]}{\left(|x_k + v_i|^2\right)^2}$$

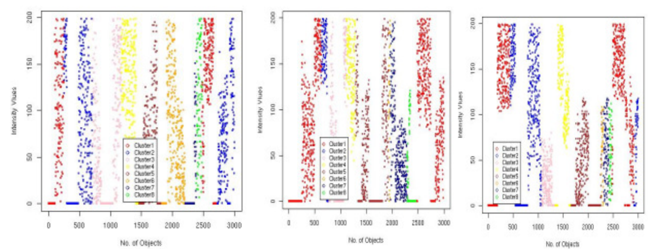
3. EXPERIMENTAL RESULTS WITH CHECKERBOARD DATASET

This subsection presents the results implementing the algorithms FPCM, KFCM, KPFCM, TFPCM, on Checkerboard database [40] for showing the clustering performance of proposed method with large amount of database. The first experiment starts with FPCM, KFCM, and KPFCM on real Checkerboard dataset with 3 attributes includes eight classes. Figs. 1 (a-c) show the level lines of the membership functions of obtained eight clusters on Checkerboard dataset by FPCM, KFCM and KPFCM. The level lines are obtained based on the resulted memberships of each object in the database. The reallocated 1000 Checkerboard dataset into eight clusters using the experimental results of FPCM, KFCM and KPFCM are given in Figs. 2 (a-c) for getting the difference in the actual eight classes in an original checkerboard dataset. Further from Figs. (1-2) it has been observed that the FPCM, KFCM and KPFCM algorithms have failed to cluster the data objects into well separated clusters, there are several overlapping clusters obtained in capturing eight classes.



(a) by FPCM (b) by KFCM (c) by KPFCM

Fig.1 Obtained size of clusters on checkerboard dataset



(a) by FPCM (b) by KFCM (c) by KPFCM

Fig2 .Separated eight Clusters of 1000 checkerboard dataset

Subsequently, this subsection is introduced the proposed TFPCM algorithm on 1000 checkerboard database for finding eight clusters. The obtained size of clusters using proposed algorithms is given in Figs. 3. The size or level lines of each cluster is identified from the database based on the resulted memberships of data object. The reallocated data into eight clusters based on partitioned results of TFPCM is given in Fig. 4. As shown in Figs. (3-4) the proposed algorithm is tried to correctly identified the eight numbers of clusters from heavily overlapping objects among the objects in dataset.

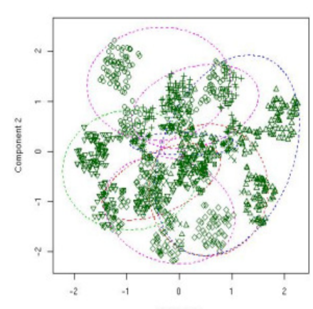


Fig.3 Obtained size of clusters checkerboard

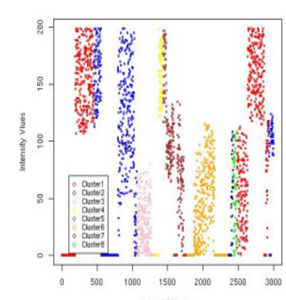


Fig. 4 Allocated 1000 by proposed dataset by proposed algorithm

But from Figs. 3&4 it can be shown that the data elements are almost found within the boundary of each clusters by proposed algorithms and the boundary of clusters have not been affected by outliers. Comparison results in terms of number of iterations for completion of clustering the datasets, clustering accuracy, and running time during experimental works using the TFPCM algorithms on Checkerboard dataset are given in Table

Table 1. Comparison of Iteration Count, Running Time and clustering accuracy

Checkerboard Dataset	FPCM	KFCM	KPFCM	TFPCM
No. of Iterations	19	17	18	5
Clustering Accuracy	51 %	65%	78%	99%
Running Time	1minute	1minute	45 seconds	4 seconds

From Table 1, the best clustering validity, running time, and number of iterations was obtained for proposed methods during the experiment on checkerboard data with eight clusters.

4. CONCLUSION

This paper is introduced novel kernel fuzzy possibilistic c-means based on the membership function of fuzzy c-means, the typicality of possibilistic c-means approaches, kernel functions, for finding available information in high dimensional databases. In order to establish the effectiveness of the proposed methods, this paper demonstrated experimental works on Checkerboard dataset. This paper has reported the superiority of the proposed method through cluster validation, running time, number of iterations and well separated clusters.

5. ACKNOWLEDGEMENTS

This work was financially supported by DST India and NSC Taiwan.

REFERENCES

[1] Berks et al., Fuzzy clustering – a versatile mean to explore medical database, ESIT2000, Aachen, Germany.

[2] Carlos Alzate, Johan A.K. Suykens, Sparse kernel spectral clustering models for large-scale data analysis, Neurocomputing, Volume 74, Issue 9, April 2011, Pages 1382-1390

[3] Congalton, R.G. and Green, K. (1992) Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis Publishers, USA.

[4] Feng Chu and L.P. Wang, "Applications of support vector machines to cancer classification with microarray data," International Journal of Neural Systems, vol.15, no.6, pp.475-484, 2005.

[5] Feng Chu, Wei Xie, and L.P. Wang, "Gene selection and cancer classification using a fuzzy neural network", Proceedings of the North-American Fuzzy Information Processing Conference (NAFIPS 2004), vol.2, pp.555-559, 2004.

[6] Gordon et al., "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma," Cancer Research, vol. 62, pp. 4963-4967, 2002.

[7] Hu Yang, Nicolino J. Pizzi,"Biomedical Data Classification Using Hierarchical Clustering" , Proc IEEE Canadian Conf Elect Comput Eng, Niagara Falls, 2004.

[8] Jaya Rama Krishnaiah et al., Data Analysis of Bio-Medical Data Mining using Enhanced Hierarchical Agglomerative Clustering , International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, Pp. 43-49, September 2012, ISSN: 2277-3754.

[9] Jezewski M., An application of modified fuzzy clustering to medical data classification, Journal of Medical Informatics and Technologies, 2011, Vol. 17, pp. 51-57.

[10] José A. Castellanos-Garzón, Carlos Armando García, Paulo Novais, Fernando Díaz: A visual analytics framework for cluster analysis of DNA microarray data. Expert Syst. Appl. 40(2): 758-774 (2013)

[11] Kenneth Revett et. al., An Analysis of a Lymphoma/Leukaemia Dataset Using Rough Sets and Neural Networks, M.S. Szczuka et al. (Eds.): ICHIT 2006, LNAI 4413, pp. 229–239, 2007, Springer-Verlag Berlin Heidelberg 2007

[12] Liang Bai et al., An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data, Knowledge-Based Systems 24 (2011) 785–795

[13] Liu et al., Performance research of Gaussian function weighted fuzzy C-means algorithm. In S. J. Maybank, M. Ding, F. Wahl, Y. Zhu (Eds.), Pattern recognition and computer vision. Proceedings of SPIE (Vol. 6788, pp. 67881Q-1–67881Q-7).

[14] Lunetta, R.S. et al. (1991) Remote sensing and geographic information system data integration: error sources and research issues. Photogramm. Eng. Remote Sens., 57,677–687.

[15] S. Mitra and Y. Hayashi, "Bioinformatics with soft computing," IEEE Transactions on Systems, Man and Cybernetics, Part C, vol.36, pp.616 -635, 2006.

[16] Omnia Ossama, Hoda M.O. Mokhtar , Mohamed E. El-Sharkawi, An extended k-means technique for clustering moving objects, Egyptian Informatics Journal (2011) 12, 45–51

[17] Rousseeuw PJ (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

[18] Shixiong Xia, Qiang Liu, Yong Zhou, Bing Liu, 2012 International Conference on Electrical and Computer Engineering, Advances in Biomedical Engineering, Vol.11, pg 227-232

[19] Smet, Frank De, Mathys, Janick, Marchal, Kathleen, Thijs, Gert, Moor, Bart De and Moreau, Yves, Adaptive quality-based clustering of gene expression profiles. Bioinformatics, 18:735–746, 2002.

[20] Tamayo, P., et al.: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic

differentiation. Proceedings of the National Academy of Sciences of the United States of America 96(6), 2907 (1999)

[21] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M.: Systematic determination of genetic network architecture. Nature Genetics 22, 281–285 (1999).

[22] Tseng, V.S., Kao, C.-P.: Efficiently Mining Gene Expression Data via a Novel Parameterless Clustering Method. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2 (2005) 355-365.

[23] Troyanskaya, O.,Cantor M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman R. Missing value estimation methods for dna microarrays, Bioinformatics. 2001 Jun;17(6):520-5

[24] D.Vanisri, C.Loganathan, An Efficient Fuzzy Possibilistic C-Means with Penalized and Compensated Constraints, Global Journal of Computer Science and Technology, Volume 11 Issue Version 1.0 March 2011.

[25] Yang, M.-S., & Tsai, H.-S. (2008). A Gaussian kernel-based fuzzy c-means algorithm with a spatial bias correction. Pattern Recognition Letters, 29, 1713–1725.

[26] Zadeh, L.A., Fuzzy sets. Information Control 8, 338–353, 1965.