VNU Journal of Science: Comp. Science & Com. Eng., Vol. 32, No. 3 (2016) 32-38

Some Improvements of Fuzzy Clustering Algorithms Using Picture Fuzzy Sets and Applications for Geographic Data Clustering

Nguyen Dinh Hoa[1,*], Le Hoang Son[2] , Pham Huy Thong[2]

[1]*VNU Information Technology Institute, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*
[2]*VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam*

Abstract

This paper summarizes the major findings of the research project under the code name QG.14.60. The research aims to enhancement of some fuzzy clustering methods by the mean of more generalized fuzzy sets. The main results are: (1) Improve a distributed fuzzy clustering method for big data using picture fuzzy sets; design a novel method called DPFCM to reduce communication cost using the facilitator model (instead of the peer-to-peer model) and the picture fuzzy sets. The experimental evaluations show that the clustering quality of DPFCM is better than the original algorithm while ensuring reasonable computational time. (2) Apply picture fuzzy clustering for weather nowcasting problems in a novel method called PFS-STAR that integrates the STAR technique and picture fuzzy clustering to enhance the forecast accuracy. Experimental results on the satellite image sequences show that the proposed method is better than the related works, especially in rain predicting. (3) Develop a GIS plug-in software that implemented some improved fuzzy clustering algorithms. The tool supports access to spatial databases and visualization of clustering results in thematic map layers.

Received 20 June 2016, Revised 04 October 2016, Accepted 18 October 2016

Keywords: Spatial clustering, fuzzy clustering, distributed clustering, picture fuzzy set, weather nowcasting, spatio-temporal regression.

1. Introduction

Geographic data clustering problems work with spatial data. These problems have many important applications in the economic development and social activities, from the geo-economic analysis, marketing analysis, environmental resources management to processing the satellite remote sensing images, weather forecasting, pollution predictions, diseases preventions, etc ... However, mining geographic data to extract information from the database of a geographic information system (GIS) has many challenges. The database of GIS contains large amounts of data, which increases day by day; the data volume to be processed is often large, even very large [3]. Attribute data fields are often multi-dimensional and correlated. Clustering multi-dimensional data, especially in the case of large data sets is a difficult problem.

Attribute data in GIS are varied, may be collected from various sources and have different forms and representations; Data can be quantitative or qualitative (classified in categories), multimedia data (meteorological images, remote sensing images). Classification

* Corresponding author. E-mail.: hoand@vnu.edu.vn

in categories is inherently fuzzy. We want to classify, by example, a region as "flat", "moderate slope," or "very steep". The interpretation of remote sensing images based on the different colors is another example of the fuzzy nature of clustering geographic data.

It is difficult in general to get the consistent clustering geographic data and the unique interpretation of results. Fuzzy approach aims to overcome some disadvantages of clear (hard) clustering for better quality. Using fuzzy set we can make suitable modifications to traditional clear clustering methods and apply to processing geographical data.

Recently, many researches focus on fuzzy clustering to handle geographic data (see the review in [5, 11, 13]). Several research groups in Vietnam and particularly in VNU Hanoi have published the works on data clustering, in which there are some researches in the direction of clustering geographical data. The promising results on fuzzy clustering of geographic data had been published by the research team at the Center for High Performance Computing, University of Science, VNU [7,8,9]. The authors have improved fuzzy clustering algorithm through the expansion of the fuzzy set concept. Instead of the classic fuzzy set, the process of clustering uses the new fuzzy concept such as the intuitionistic fuzzy set [1.16] and more recently the picture fuzzy set [4].

Research project "Development of advanced data clustering algorithms for geographic information systems and applications" under the code name QG.14.60 aims to continue the researches in this direction. The application of expanded fuzzy concept as intuitionistic fuzzy sets, picture fuzzy sets will allow to enhance the quality of clustering. On the other hand, to handle large data sets in clustering geographic data for the real life applications, it is necessary to improve performance of the algorithms, to increase the speed of convergence in the distributed clustering scenario in particular. The development of a tool for data clustering and integrating it into the geographic information systems as a utility to assist users is also a task to be completed by the project team.

The rest of this paper is organized as follows. Section 2 describes the distributed fuzzy clustering method for big data using picture fuzzy sets called DPFCM. An application of picture fuzzy clustering for weather nowcasting problems in a novel method called PFS-STAR is presented in section 3. Section 4 introduces the GIS plug-in tool *SpatialClust* that implements some improved fuzzy clustering algorithms. Summary and conclusion follows in section 5.

2. Distributed Clustering Method Using Picture Fuzzy Sets - DPFCM

2.1. Fuzzy clustering with picture fuzzy sets

The concept of picture fuzzy sets [4] is suggested in the case of opinion polls. The voter opinions on the decision in question can be one of four types: yes, no, abstain, and refusal to answer. A picture fuzzy set is then defined as a collection of elements x, each associated with three measures $\mu_S(x)$, $\eta_S(x)$, $v_S(x)$ as follows:

$S = \{(x, \mu_S(x), \eta_S(x), \xi_S(x))\}$;

These measures subject to the constraints:

$\mu_S(x) \in [0,1]$, $\eta_S(x) \in [0,1]$, $\xi_S(x) \in [0,1]$.

$\mu_S(x) + \eta_S(x) + \xi_S(x) \in [0,1]$.

$\mu_S(x)$ is called the positive degree of membership of x, $\eta_S(x)$ is the neutral degree and $\xi_S(x)$ is the negative degree. The refusal degree of an element is calculated as $\gamma_S(x) = 1 - (\mu_S(x) + \eta_S(x) + \xi_S(x))$.

In [15] the authors have proposed a picture fuzzy clustering algorithm, using the concept of picture fuzzy sets instead of the classical fuzzy set. The algorithm bases on the well-known fuzzy clustering algorithm FCM [2], but besides

the positive factors u_{kj}, the negative and neutral factors also included in each steps to calculate the membership degree of the data point j to the cluster k. The objective function to minimize is the following:

$$J = \sum_{k=1}^{N}\sum_{j=1}^{C}\left(u_{kj}\left(2-\xi_{kj}\right)\right)^m \left\|X_k - V_j\right\|^2 + \sum_{k=1}^{N}\sum_{j=1}^{C}\eta_{kj}\left(\log\eta_{kj} + \xi_{kj}\right) \to \min \quad (1)$$

The variables $u_{kj}, \eta_{kj}, \xi_{kj}$ subject to the constraints:

$$u_{kj}, \eta_{kj}, \xi_{kj} \in [0,1], \quad (2)$$

$$u_{kj} + \eta_{kj} + \xi_{kj} \le 1, \quad (3)$$

$$\sum_{j=1}^{C}\left(u_{kj}\left(2-\xi_{kj}\right)\right) = 1, \quad (4)$$

$$\sum_{j=1}^{C}\left(\eta_{kj} + \frac{\xi_{kj}}{C}\right) = 1,\ k = \overline{1,N},\ j = \overline{1,C} \quad (5)$$

The steps of algorithm are as follows:

- *Initial step:* $t = 0$; randomly initialize the variables $u_{kj}^{(t)}, \eta_{kj}^{(t)}, \xi_{kj}^{(t)}$ ($k = \overline{1,N}$, $j = \overline{1,C}$) so that the conditions (2-3) are satisfied;

- **Step 1:** *t= t+1*; calculate the cluster centers V_j using the formula below

$$V_j = \frac{\sum_{k=1}^{N}\left(u_{kj}\left(2-\xi_{kj}\right)\right)^m X_k}{\sum_{k=1}^{N}\left(u_{kj}\left(2-\xi_{kj}\right)\right)^m},\ j = \overline{1,C}, \quad (6)$$

- **Step 2:** Update the u_{kj} , η_{kj}, ξ_{kj} by the formula (7-9)

$$u_{kj} = \frac{1}{\sum_{i=1}^{C}\left(2-\xi_{kj}\right)\left(\frac{\left\|X_k - V_j\right\|}{\left\|X_k - V_i\right\|}\right)^{\frac{2}{m-1}}}, \quad (7)$$

$k = \overline{1,N}$, $j = \overline{1,C}$,

$$\eta_{kj} = \frac{e^{-\xi_{kj}}}{\sum_{i=1}^{C} e^{-\xi_{ki}}}\left(1 - \frac{1}{C}\sum_{i=1}^{C}\xi_{ki}\right), \quad (8)$$

$k = \overline{1,N}$, $j = \overline{1,C}$,

$$\xi_{kj} = 1 - \left(u_{kj} + \eta_{kj}\right) - \left(1 - \left(u_{kj} + \eta_{kj}\right)^{\alpha}\right)^{\frac{1}{\alpha}}, \quad (9)$$

$k = \overline{1,N}$, $j = \overline{1,C}$.

- **Step 3:** Stop the loop if the total changes of variables in updating step less than the predefined threshold:

$$\left\|u^{(t)} - u^{(t-1)}\right\| + \left\|\eta^{(t)} - \eta^{(t-1)}\right\| + \left\|\xi^{(t)} - \xi^{(t-1)}\right\| \le \varepsilon$$

or the step counter greater than *maxSteps*; otherwise, return to Step **1**.

2.2. DPFCM - Distributed fuzzy clustering using picture fuzzy sets

In [17] the authors have proposed a fuzzy clustering algorithm CDFCM for distributed computing environments with the peer-to-peer communicational model (P2P). In this algorithm, the cluster centers and the fuzzy membership factors of data points are calculated at every peer site and then updated in each iteration using only the results of the peer neighbors. This process is repeated until a stopping criterion is satisfied. CDFCM is considered as one of the most effective fuzzy clustering algorithms for distributed computing environments.

By analysis in details we realize that communication costs for each iteration of the algorithm CDFCM is high, approximately $p.n_{loc}$, where p is the number of peers and n_{loc} is the average number of neighbors of one peer. Also, because the algorithm only use the nearby local results to update in each iterations, so the final clustering result may not be of highest quality.

Our idea of improving the algorithm CDFCM is that we can reduce communication costs and improve the quality of clustering results through using the picture fuzzy clustering and the facilitator model instead of the peer-to-peer communicational model. The proposed method is called DPFCM (distributed fuzzy picture clustering method).

- At the local level, each peer site performs picture fuzzy clustering in each iteration;

- At the global level, all the peer sites transfer the results to the unique master site which plays the role of a facilitator in the communication process. Thus, in one updating step at the global level, the cost to complete the communication process is of order of *p*. Moreover, the global information allows to improve the quality of clustering.

The experimental evaluation was conducted upon the *benchmark datasets* from UCI Machine Learning Repository, namely: *IRIS, GLASS, IONOSPHERE, HABERMAN and HEART*. The speed of convergence and the cluster validity measurements are evaluated. The average number of iterations AIN is obviously better if smaller, where as the average classification rate ACR and the average normalized mutual information ANMI [6] are the bigger the better.

The table below compares the quality of our clustering algorithm DPFCM with some other algorithms.

Table 1. Clustering quality of algorithms [10]

Dataset	DPFCM	FCM	WEFCM	PFCM	Soft-DKM	CDFCM
ACR (%)						
IRIS	**96.04**	89.33	96.66	89.33	87.38	95.90
GLASS	**53.33**	42.08	54.39	42.08	40.50	52.96
IONOSPHERE	**75.26**	70.94	76.58	70.94	67.77	75.26
HABERMAN	**76.50**	51.96	77.12	51.96	51.42	74.68
HEART	**71.89**	51.31	72.88	51.31	50.24	71.95
ANMI						
	DPFCM	FCM	WEFCM	PFCM	Soft-DKM	CDFCM
IRIS	**0.8785**	0.7433	0.8801	0.7433	0.7294	0.8705
GLASS	**0.4175**	0.2974	0.4263	0.2974	0.2848	0.4170
IONOSPHERE	**0.1961**	0.1299	0.2026	0.1299	0.1028	0.1961
HABERMAN	**0.0826**	0.0024	0.0992	0.0024	0.0018	0.0610
HEART	**0.0395**	0.0052	0.0445	0.0052	0.0028	0.0408

The results presented in the table show that the clustering quality of DPFCM is mostly better than those of three distributed clustering algorithms, namely CDFCM, Soft-DKM and PFCM. It is also better than the traditional centralized clustering algorithm FCM, and is a little worse than the centralized weighted clustering WEFCM. There are some cases, for example, of the IONOSPHERE and the HEART dataset, DPFCM results in clustering quality of the same order or a little worse than CDFCM.

For the speed of convergence, the comparison of AIN of DPFCM with the others shows the disadvantage of DPFCM as expected, but the differences of AINs are not much.

The above results were published in the international scientific journal "Expert Systems with Applications" [10].

3. Application of picture fuzzy clustering in analysis of meteorological images for weather nowcasting

One of the methods of predicting the weather, called weather nowcasting, is on the basis of analysis of the satellite images sequence by combining the spatio-temporal autoregressive (STAR) model with fuzzy clustering. There are publications in this research domain. Recently Shukla and colleagues [14] have proposed a number of technical improvements to raise the accuracy.

However, because using classical fuzzy sets, the image areas of ambiguous interpretation or lack of clarity have the negative impacts to the prediction result. Picture fuzzy clustering [15] using more advanced fuzzy concept has been shown that is better than the traditional fuzzy clustering. Our idea is advancing the research of Shukla et al, through combining the primary STAR techniques with picture fuzzy clustering to create a new weather prediction method, called Picture Fuzzy Clustering - Spatiotemporal autoregressive (PFC-STAR). We hope that the combination can improve the quality of the prediction results. The proposed PFC-STAR method involves three steps:

- The pixels of satellite images (training samples) are divided into groups by using picture fuzzy clustering algorithm proposed in [15].

- All the elements of these clusters in training samples are then labeled and filtered using the Discrete Fourier Transform to clarify non-predictable scale to increase the time range of predictability.

- Finally, the next sequence of images are predicted through spatio-temporal auto-regression method, which allows the weather forecast for the chosen geographic area in a short time ahead.

- The experimental evaluation of the proposed method was conducted on the personal computer of 2 GB RAM, 2.13 GHz core 2 Duo, upon the data sets, which is the sequence of satellite images of the Southeast Asia region. Each data set includes 5 satellite images taken over a time period from 9:30 to 13:30, of 100 x 100 pixels in size. Comparison of the results showed that the method proposed here is better than the relevant methods of weather nowcasting, especially with higher precision of the rain-rate regression.

The above results have been presented and published in the Proceedings of the International Symposium on Geo-informatics for Spatial Infrastructure Development in Earth and Allied Sciences (GIS-IDEAS)" [12].

Table 2. Comparison of RMSE and computational time of PFC-STAR and the method of Shukla et al [12]

Data	RMSE (%)		Computational time (sec)	
	PFC-STAR	Shukla et al. (2014)'s method	PFC-STAR	Shukla et al. (2014)'s method
Malaysia	26.77	27.11	362.745	359.88
Luzon – Philippines	33.61	33.45	345.672	343.43
Jakarta – Indonesia	30.12	32.04	342.76	339.97

4. Developing data clustering tool as a plug-in for GIS

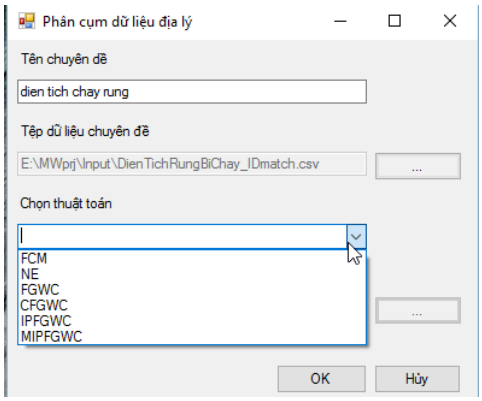
For the convenience of users in mining geographical data, a data clustering engine should be developed and integrated into GIS to support direct access of spatial database for reading input data and displaying the results on the map layers.

MapWindow is an open source GIS software that Windows users are familiar with and it is currently being developed and the latest version released continuously. MapWindow support plug-ins in the form of dynamic link libraries (.dll *), and the development environment such as Visual Studio Community Edition is available for free download. This tool supports using the language C# and dot.NET frame. Our implementation of the proposed algorithms to run experimental evaluation is conducted using C / C ++, therefore the Visual Studio development environment in the most suitable choice to put our source code into.

The plug-in named *SpatialClust* is a clustering tool module for geographical data, which deployed several fuzzy clustering algorithms with improvements that our team has proposed as presented above. Restrictions on computational resources of a plug-in does not allow to implement the distributed algorithms or to process large data sets. Hence,

only some appropriate algorithms are included in the tool, namely: FCM, NE, FGWC, CFGWC, IPFGWC, MIPFGWC. The plug-in supports direct access of spatial database for reading attribute values and displaying the resulting clusters in different colors on the map.

Input: data file format is *.csv (coma separated values). All the GIS software have to support importing and exporting data in the *.shp format of one map layer to the *.csv format.



Picture 1. Dialog box for choosing input data and algorithm.

Output: there are two types:

1. Output as text file (*.txt or plain text) to provide enough detail for the purposes of analysis and evaluation of algorithms or for the subsequent treatment, if any.

2. Displaying visually on the map: in parallel with printing the results to a text file, the tool allows updated cluster labels directly to the cluster column of database beneath and by setting GIS functionalities users can show visualization of clusters on maps. For this purpose, the properties table of map layer must have the last column named CLUSTER.

5. Summary and conclusions

The research we carried out in the research project has contributed to improve fuzzy clustering algorithms, distributed fuzzy clustering to process large data sets in order to apply for geographical data clustering. The results contribute to better address real-world problems we meet in many application areas.

The distributed fuzzy clustering algorithm to handle large data sets using picture fuzzy sets called DPFCM has improved overall clustering quality in comparison with the algorithm of Chen and colleagues [17]. Clustering quality of DPFCM is better than some clustering algorithms of the same type, but the computational time does not add much. The new weather nowcasting method PFC-STAR using picture fuzzy sets instead of classical fuzzy sets has allowed raising the quality of predictions in comparison with the method of Shukla et al [14], especially in predicting rain-rate. We can conclude that the use of picture fuzzy clustering actually had a positive impact on the quality of the clustering results for the problems related to the inherently fuzzy concepts.

The software tool for data clustering integrated into MapWindow as a plug-in that performs typical fuzzy clustering algorithms and the improvements proposed in our researches will help to promote practical applications of geographic data mining in various domains.

Acknowledgements

The authors would like to thank the colleagues for comments through discussions in the scientific seminars which help to correct the errors and to complete the results achieved. We also express our sincere thanks to VNU Hanoi for funding the research project under the code name QG.14.60 and for other supports to conduct the research.

References

[1] Atanassov, K. T. (1986). *Intuitionistic fuzzy sets*. Fuzzy Sets and Systems, 20, 87-96.

[2] Bezdek, J.C., R. Ehrlich, et al (1984), *FCM: the fuzzy c-means clustering algorithm*, Computers and Geosciences, 10, pp.191-203

[3] Brinkoff, T., Kriegel, H.-P. (1994), *The Impact of Global Clustering on Spatial Database Systems*, Proceedings of the 2th VLDB Conference, Santiago, Chile, pp. 168-179.

[4] Bui Cong Cuong, Vladik Kreinovich, *Picture Fuzzy Sets - a new concept for computational intelligence problems*, Proceeding of 2013 Third World Congress on Information and Communication Technologies (WICT 2013), 1-6.

[5] Deepti Joshi, *Polygonal Spatial Clustering*, Ph.D. Dissertation, University of Nebraska, 2011.

[6] Huang, H. C., Chuang, Y. Y., & Chen, C. S. (2012), *Multiple kernel fuzzy clustering*, IEEE Transactions on Fuzzy Systems, 20(1), 120-134.

[7] Le Hoang Son, Bui Cong Cuong, Pier Luca Lanzi, Hoang Anh Hung (2011) *Data Mining in GIS: A Novel Context-Based Fuzzy Geographically Weighted Clustering Algorithm*. International Journal of Machine Learning and Computing.

[8] Le Hoang Son (2011), Nguyen Dinh Hoa, Pier Luca Lanzi, and Bui Thi Huong Lan, A Combination of Clustering Techniques and Fuzzy Control in 2D Polygon Determination for the Terrain Splitting and Mapping Problem, International Journal of Computer and Electrical Engineering 3(5), pp. 682 – 689.

[9] Le Hoang Son, Bui Cong Cuong, Pier Luca Lanzi, Nguyen Tho Thong (2012), *A Novel Intuitionistic Fuzzy Clustering Method for Geo-Demographic Analysis*, Expert Systems with Applications.

[10] Le Hoang Son (2015), "*DPFCM: A novel distributed picture fuzzy clustering method on picture fuzzy sets*", Expert Systems with Applications, 42 (2015) pp. 51-66.

[11] Neethu C V, Subu Surendran, *Review of Spatial Clustering Methods*, International Journal of Information Technology Infrastructure, Volume 2, No.3, May - June 2013.

[12] Nguyen Dinh Hoa, Pham Huy Thong, Le Hoang Son, "*Weather Nowcasting from Satellite Image Sequences Using Picture Fuzzy Clustering and Spatial-temporal Regression"*, International Symposium on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences (GIS-IDEAS), Danang, Vietnam, December, 7th-9th , 2014, pp. 137-142

[13] M. Perumal, B. Velumani, A. Sadhasivam, and K. Ramaswamy, (2015), *Spatial Data Mining Approches for GIS - A Brief Review*, Conference paper, January 2015, © Springer International Publishing Switzerland.

[14] Shukla, B. P., Kishtawal, C. M., & Pal, P. K. (2014),*Prediction of Satellite Image Sequence for Weather Nowcasting Using Cluster-Based Spatiotemporal Regression*, IEEE Transactions on Geoscience and Remote Sensing, 52(7), 4155 - 4160.

[15] Thong, P.H., Son, L.H. (2014). *A new approach to multi-variables fuzzy forecasting using picture fuzzy clustering and picture fuzzy rules interpolation method*, Proceeding of 6th International Conference on Knowledge and Systems Engineering (KSE 2014), October 9-11, 2014, Hanoi, Vietnam, 679 - 690.

[16] Visalakshi, N. K., Thangavel, K., & Parvathi, R. (2010). *An intuitionistic fuzzy approach to distributed fuzzy clustering*, International Journal of Computer Theory and Engineering, 2 (2), 1793–8201.

[17] Zhou, J., Chen, C., Chen, L., & Li, H. (2013). *A collaborative fuzzy clustering algorithm in distributed network environments*, IEEE Transactions on Fuzzy Systems. http://dx.doi.org/10.1109/TFUZZ.2013.2294205.