Data Mining in GIS: A Novel Context-Based Fuzzy Geographically Weighted Clustering Algorithm

Le Hoang Son, Pier Luca Lanzi, Bui Cong Cuong, and Hoang Anh Hung

***Abstract*— Geographic Information Systems (GIS) play a very important role to researches and industries. In fact, there have been some studies related to the development of this kind of systems as well as methods of mining attribute GIS data. In this paper, we will explore an aspect of Data Mining in GIS, that is, GIS Clustering for Geo-Demographic Analysis and present a novel context-based fuzzy geographically weighted clustering algorithm to solve such task.**

***Index Terms*—Clustering, data mining, GIS, Geo-Demographic analysis.**

I. Introduction

Applications of Geographic Information Systems (GIS) can be found in many fields of societies. Especially when integrating with the Internet, the importance of GIS is getting more obvious. Indeed, many examples of WebGIS are shown from environment management, land use monitoring to agriculture, etc. [6]. In these researches, WebGIS was used as a tool to display and analyze two dimensional GIS data on a web environment. The results then were either printed or brought back to managers to make efficient policies. Another example can be found from the literature [9] where open-source WebGIS was used for sustainable use and management of natural resources that are getting more and more important to our lives and facing excessive exploitation problems.

From these examples, a question is arisen: *how to extract information or knowledge from a large GIS data?* Obviously, if we have effective Data Mining methods in GIS then more precise policies can be brought to reality. Moreover, they can be the basis to deploy 3D WebGIS systems. In Ref. [7], the authors have pointed out some new trends and prospects on the development of three dimensional WebGIS systems in which spatial analysis and 2D GIS data mining are the main focuses in the following years. Along with another recent study from Ref. [8], it can be concluded that Data Mining in GIS is a new trend nowadays.

Geo-Demographic Analysis is a typical example in GIS mining. It is described as "*the analysis of spatially referenced Geo-Demographic and lifestyle data*" [5] and widely used in the public and private sectors for the planning and provision of products and services. Geo-Demographic Analysis often uses clustering techniques that are used to classify the Geo-Demographic data into groups, making the data more manageable for analysis purposes. Clustering identifies a number of Geo-Demographic groups (clusters), each one with a Geo-Demographic profile. Each geographical area under consideration is then assigned to a group based on its similarity to the group profile [5].

Our contribution in this paper is the introduction of a novel context-based fuzzy geographically weighted clustering algorithm so called CFGWC. It is carefully tested through experiments and proved the advantage in comparison with the state-of-the-art algorithm for this problem.

The rest of the paper is organized as follows. Section 2 introduces some related researches concerning Geo-Demographic Analysis. The main algorithm CFGWC will be presented in Section 3. Section 4 presents some results extracted from experiments. Finally, we make conclusion and future works in the last section.

II. Related Works

In Geo-Demographic Analysis, fuzzy clustering methods are often used. Because they assign a membership value for each area instead of assigning a geographical area to a single group, the issues of ecological fallacy are overcome. The fuzzy clustering algorithm typically used in Geo-Demographic Analysis is the fuzzy c-means clustering algorithm of Bezdek [1] known as FCM.

However, FCM misses geographical factor in its design. For example, consider that residential area Type 27 is the same and behaves in the same way wherever it happens to be located. But what happens if Type 27 areas respond differently depending on their map locations? To overcome this shortcoming, Feng and Flowerdew [4] proposed an extension to the fuzzy clustering technique, which provides for the *ex post facto* adjustment of the cluster membership values based on "neighbourhood effects". The neighbourhood effects incorporate geography into the model. The neighbourhood effects formula adjusts the cluster membership as shown in equation (1),

$$u_j' = \alpha \times u_j + \beta \times \frac{1}{A} \times \sum_{i=1}^{c} w_{ij} \times u_i \quad (1)$$

where u_j' is the new cluster membership of area j and u_j is the old cluster membership of this area. Two parameters α and β are scaling variables to affect the proportion of

Manuscript received April 14, 2012; revised May 30, 2012. This work is sponsored by a research grant of NAFOSTED and Vietnam National University, Hanoi (QGTĐ.11.01).

Le Hoang Son and Hoang Anh Hung are with the Center for High Performance Computing, VNU University of Science, Vietnam (e-mail: sonlh@vnu.edu.vn).

Pier Luca Lanzi is with the Department of Electronics and Information, Politecnico di Milano, Milan, Italy (e-mail: lanzi@elet.polimi.it).

Bui Cong Cuong is with the Institute of Mathematics, Vietnamese Academy of Science, Vietnam (e-mail: ccuong@inbox.com).

the original membership which satisfy the condition,

$$\alpha + \beta = 1 \tag{2}$$

A is a factor to scale the "sum" term to the range 0 to 1. The weighted membership is calculated as follows,

$$w_{ij} = p_{ij}^{b} / d_{ij}^{a} \tag{3}$$

where p is the length of the common boundary between areas i and j. The number d_{ij} is the distance between areas i and j. Two numbers a and b are user definable parameters.

However, Feng and Flowerdew's neighbourhood effects have some limitations. *Firstly*, they ignore the effects of areas which have no common boundaries. *Secondly*, they exclude the effects of population - a key Geo-Demographic consideration. To overcome these limitations, a modified version of the cluster membership adjustment was proposed that incorporates a spatial interaction effect model [5]. Originated from the principles of geographical spatial interaction [2] by incorporating a basic spatial interaction model into the weighting of the memberships as well as included the neighbourhood effects in fuzzy clustering algorithm, this makes cluster centers "geographically aware".

In the literature [5], Mason and Jacobson proposed a model to calculate the influence of one area upon another as the product of the populations of the areas. A distance decay effect is implemented in the divisor. This effect is implemented through the weighting factor as described in equation (4),

$$w_{ij} = \left(m_i \times m_j\right)^{b} / d_{ij}^{a} \tag{4}$$

where m_i, m_j are the population of areas i and j, respectively. The number d_{ij} is the distance between areas i and j. Two numbers a and b are user definable parameters. A is a factor to scale the "sum" term, and is calculated across all clusters, ensuring that the sum of the memberships for a given area for all clusters is equal to one.

The algorithm FGWC has been being the state-of-the-art in Geo-Demographic Analysis.

III. Context Based Fuzzy Geographically Weighted Clustering Algorithm

A. Basic Ideas

Mason and Jacobson's algorithm [5] has a limitation: its velocity. Because a cluster membership modification process has to be done in each step, this definitely increases the computational time. Moreover, the standard FCM is a direction-free construction, which means it is regardless if a dimension is an input or an output variable. This may lead to a not very reasonable distribution of prototypes or centres of groups in that the algorithm sweeps over even unrelated areas in data space.

For the given problem, an idea of the context-based algorithm is invoked. Originated from the result of [10] and especially the new one [3], we try to attach "context" to the above algorithm. Indeed, we define a context variable in order to narrow the origin dataset under some conditions of certain dimensions. Because only a subset of origin dataset which has considerable meaning to the context is invoked, the velocity and efficiency of clustering can be improved considerably and the result focuses on the area that really has many relevant points. For example, if we want to look for a shopping area then a new context "shopping" will be put to the algorithm to reduce the search space. In a specific case, the context-sensitive algorithm allows us to concentrate the classification into a subspace due to conditions of some dimensions showed in defined context. What is more, the speed of CFGWC is relatively faster than FGWC in case of little context variables.

Given a dataset of N attributes $X = \{X_1,..,X_N\}$. Supposed that missing data have been processed, our purpose is to classify them into C clusters. We will work in r - dimension space ($X \in R^r$) with X_k is the k^{th} point and V_i is the center of i^{th} cluster. Then, we define a context variable in $Y \in X$ whose definition is stated through the map,

$$\begin{aligned} A &: Y \to [0,1] \\ y_k &\mapsto f_k = A(y_k). \end{aligned} \tag{5}$$

The value f_k can be understood as the representation for the level of relation of the k^{th} point to the supposed context Y. These are some ways to define the relation between f_k and the membership of k^{th} point to the i^{th} cluster, for instance, using the sum operator (6) or maximum operator (7).

$$\sum_{j=1}^{C} u_{kj} = f_k \,;\; k = \overline{1,N}, \tag{6}$$

$$\max_{j=1}^{C} u_{kj} = f_k \,;\; k = \overline{1,N}. \tag{7}$$

The basic objective function is,

$$J = \sum_{k=1}^{N}\sum_{j=1}^{C} u_{kj}^{m} \left\| X_k - V_j \right\|^2 \to \min, \tag{8}$$

where m is a coefficient of fuzziness and u_{kj} is an element of partition matrix U defined as follow,

$$\begin{aligned} U(f) = \Bigg\{ & u_{kj} \in [0,1] : \sum_{j=1}^{c} u_{kj} = f_k, \forall k = 1,..,N; \\ & \sum_{k=1}^{N} u_{kj} < N, \forall j = 1,..,C \Bigg\}. \end{aligned} \tag{9}$$

B. The Proposed Algorithm

The Context Fuzzy Geographically Weighted Clustering algorithm (CFGWC) has five steps:

1. *Initiate the matrix* $U(t)$ *at* $t = 0$.
2. *Re-calculate centers of each clusters according to*

$$V_j = \frac{\sum_{k=1}^{N} u_{kj}^{m} \times X_k}{\sum_{k=1}^{N} u_{kj}^{m}} \,;\; j = \overline{1,C}. \tag{10}$$

3. *Re-calculate matrix* $U(t+1)$ as follows,

$$u_{kj} = \frac{f_k}{\sum_{i=1}^{C}\left(\frac{\|X_k - V_j\|}{\|X_k - V_i\|}\right)^{\frac{2}{m-1}}}; \quad k = \overline{1,N};\ j = \overline{1,C}. \tag{11}$$

4. *Adjust the partition matrix following by geographical characteristics,*

$$u_{kj}' = \alpha \times u_{kj} + \beta \times \frac{1}{A} \times \sum_{i=1}^{c} w_{ij} \times u_{ki}; \quad k = \overline{1,N};\ j = \overline{1,C}. \tag{12}$$

All parameters in equation (12) were previously defined through equations (2) and (4). However, the parameter A is a factor to scale the "sum" term, and is calculated across all clusters, ensuring that the sum of the memberships for a given area for all clusters is equal to f_k.

5. *If the error of the partition matrix* $\|U(t+1) - U(t)\|$, *defined through some analysis normal, is less than given threshold* δ *then the algorithm stops, else return Step 2.*

We arrive at the formula in Step 3 by transforming the condition $U \in U(f)$ to a standard unconstrained optimization by making use of Lagrange multipliers and determining a critical point of the resulting function. That means we only need to change the total membership of each point to all the groups. That sum is not necessary equal to 1, but it can vary from 0 to 1. It is obvious from those formulae that, if a point has no meanings in a certain context, its contextual value f_k will be equal to 0, and it plays no role in re-manipulating the positions of centres and the membership measures. The target function of the algorithm remains unchanged.

IV. Experimental Results

In this section, we have implemented the above two algorithms (Neighbourhood Effects- NE, FGWC) in addition to the proposed (CFGWC) in C programming language and executed them on a Linux Cluster 1350 with eight computing nodes of 51.2GFlops. Each node contains two Intel Xeon dual core 3.2GHz, 2GB Ram. These algorithms was run against a test dataset of socio-economic demographic variables from United Nation Organization in 2005, using $a = b = 1$, fuzzy exponent is 2 and β values from 0.1 to 0.5. Table 1 shows the running time of three algorithms.

The results in Table 1 clearly show that the proposed algorithm CFGWC is faster than FGWC and not too much slower than NE in most cases. When the parameter β increases from 0.1 to 0.5, more geographic modification have to be done then the computation time is higher as a result. However, the running time of CFGWC is faster than NE when this parameter is small. Therefore, the first notice derived from this test is that CFGWC is better than FGWC and even NE in some cases.

In Fig. 1, we study the change in membership of four algorithms. Using the dataset of UNO above, we split them into two groups: "High" and "Low". The figure below shows membership degrees of second group of 30 typical countries. Obviously, the membership degree of NE is near to the one of FCM because NE only modifies geo-location effects after running FCM algorithm. Due to the geo-modification in each step, the line in FGWC is quite far from the ones of FCM and NE. As being mentioned before, the CFGWC algorithm concentrates on some specific contexts. Indeed, its line is near to FGWC's in some cases. This may help us focus on some main relevant results in all spaces.

TABLE I: The Comparison of 3 Algorithms

Number of elements	β	The running times of algorithms (sec)		
		NE	FGWC	CFGWC
100	0.1	0.439102	0.659341	0.384615
	0.2	0.439127	0.824176	0.439560
	0.3	0.494455	0.879121	0.494505
	0.4	0.659341	0.934066	0.769231
	0.5	0.660345	1.052133	0.879121
200	0.1	0.659341	0.879121	0.661231
	0.2	0.714286	0.912616	0.725632
	0.3	0.724387	0.914202	0.819618
	0.4	0.964272	1.137693	1.021739
	0.5	1.043956	1.463481	1.243875

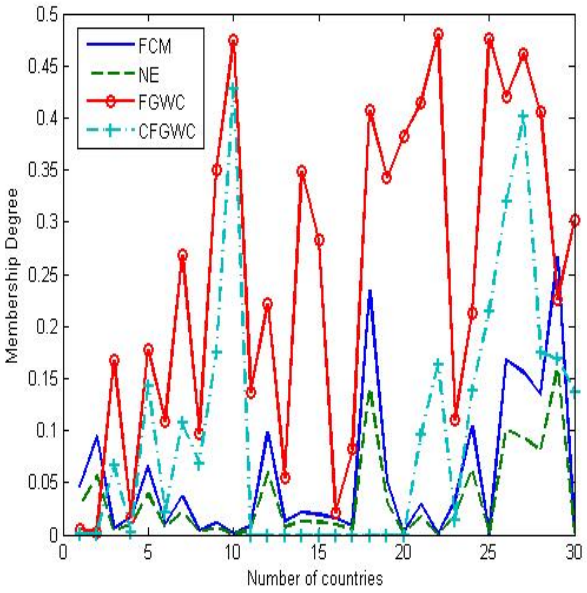

Fig. 1. The change of membership degrees in the dataset

V. Conclusion

This paper aims to emphasize a new promising trend in Data Mining namely as GIS Mining by pointing out some recent researches as well as motivations behind. Throughout a brief summary about Geo-Demographic Analysis, we strongly believe that GIS mining will be intensively studied in nearly future. Besides, we also presented a new algorithm for Geo-Demographic Analysis problem. This algorithm was carefully verified and compared with some best known ones.

The results in experiment showed the efficiency of our solution.

In the future, we will concentrate on the parallel version of CFGWC algorithm as well as integrate it to real GIS applications.

REFERENCES

[1] Bezdek, J.C., R. Ehrlich, et al., "FCM: the fuzzy c-means clustering algorithm," *Computers & GeoSciences*, vol. 10, pp. 191-203, 1984.

[2] Birkin, M and G. P. Clarke, "Spatial Interaction in Geography," *Geography Review*, vol. 4, no. 5, pp. 16-24, 1991.

[3] Bui Cong Cuong, Le Hoang Son and Hoang Thi Minh Chau, "Some Context Fuzzy Clustering Methods for Classification Problems," In *Proceedings of the 2010 Symposium on Information and Communication Technology* (SoICT '10), Hanoi, Viet Nam, August 27 - 28, 2010, pp. 34 - 40.

[4] Feng, Z. and R. Flowerdew, *Fuzzy Geodemographics: a contribution from fuzzy clustering methods. Innovations in GIS 5*, London: Taylor & Francis, 1998.

[5] G. A. Mason and R. D. Jacobson, "Fuzzy Geographically Weighted Clustering," In *Proceedings of the 9th International Conference on GeoComputation*, Maynooth, Eire, Ireland, 3-5 September 2007, (electronic proceedings on CD-ROM).

[6] Le Hoang Son, "A WebGIS application in agricultural land management," *VNU Journal of Science, Natural Sciences and Technology*, vol. 25, no. 4, pp. 234 – 240, 2009.

[7] Le Hoang Son, "On the Development of Three Dimensional WebGIS Systems: Some New Trends and Prospects," In *Proceedings of the 2010 3rd IEEE International Conference on Computer Science and Information Technology* (IEEE ICCSIT 2010), Chengdu, China, July 9 - 11, 2010, vol. 1, pp. 182 – 186.

[8] Liu Jia, Liu Lin, "Research on GIS Data Mining Method," In *Proceedings of the 2010 3rd IEEE International Conference on Computer Science and Information Technology* (IEEE ICCSIT 2010), Chengdu, China, July 9 - 11, 2010, vol. 8, pp. 568-571.

[9] Le Hoang Son, Nguyen Quoc Huy, Nguyen Tho Thong and Tran Thi Kim Dung, "An effective solution for sustainable use and management of natural resources through WebGIS Open Source and Decision-Making Support Tools," In *Proceeding of the 5th International Conference on GeoInformatics for Spatial-Infrastructure Development in Earth and Allied Sciences* (GIS-IDEAS 2010), Hanoi, Vietnam, December 9-11, 2010, pp. 87 – 92.

[10] W. Pedrycz, "Conditional fuzzy C-mean," *Pattern Recognition Lett*; vol. 17, pp. 625-632, 1996.

Le Hoang Son is a researcher at the Center for High Performance Computing, VNU University of Science. His major field includes Data Mining, Geographic Information Systems and Parallel Computing. He is a member of IACSIT and also an associate editor of the International Journal of Engineering and Technology (IJET). He also served as a reviewer for PACIS 2010, ICMET 2011, ICCTD 2011, International Journal of Computer and Electrical Engineering (IJCEE), Imaging Science Journal (IMS). Email: sonlh@vnu.edu.vn. Tel.: +84-904-171-284.

Pier Luca Lanzi was born in Turin, Italy, in 1967. He received the Ph.D. degree in Computer and Automation Engineering from the Politecnico di Milano in 1999. He is associate professor at the Politecnico di Milano, Dept. of Electronics and Information. His research areas include evolutionary computation, reinforcement learning, machine learning. He is interested in applications to data mining and computer games. He is member of the editorial board of the Evolutionary Computation Journal, the IEEE Transaction on Computational Intelligence and AI in Games, and Evolutionary Intelligence. He is also the founding editor in chief of the SIGEVOlution, the newsletter of the ACM Special Interest Group on Genetic and Evolutionary Computation. He served as the general chair for the 2009 IEEE Conference on Computational Intelligence in Games (CIG-2009) and for the 2011 Genetic and Evolutionary Computation Conference (GECCO-2011). Email: lanzi@elet.polimi.it

Bui Cong Cuong is an associate professor at Institute of Mathematics, Vietnam Academy of Science and Technology (VAST). His major fields include Soft Computing, Intelligence Computational, Fuzzy Clustering and Neuro Computing. He published more than 40 papers at national conferences and journals as well as international ones. Email: ccuong@inbox.com.

Hoang Anh Hung is a researcher at the Center for High Performance Computing, VNU University of Science. His major fields include Software Engineering and Knowledge Discovery. He has a good experience in developing some tools for these fields. Email: anhhungnxh@gmail.com.