Applied Soft Computing 22 (2014) 566–584

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Enhancing clustering quality of geo-demographic analysis using context fuzzy clustering type-2 and particle swarm optimization

Le Hoang Son*

VNU University of Science, Vietnam National University, Viet Nam

ARTICLE INFO

Article history:
Received in revised form 14 February 2014
Available online 2 May 2014

Keywords:
Context clustering
Fuzzy clustering type-2
Geo-demographic analysis
Heuristic algorithms
Particle swarm optimization

ABSTRACT

Geo-Demographic Analysis, which is one of the most interesting inter-disciplinary research topics between Geographic Information Systems and Data Mining, plays a very important role in policies decision, population migration and services distribution. Among some soft computing methods used for this problem, clustering is the most popular one because it has many advantages in comparison with the rests such as the fast processing time, the quality of results and the used memory space. Nonetheless, the state-of-the-art clustering algorithm namely FGWC has low clustering quality since it was constructed on the basis of traditional fuzzy sets. In this paper, we will present a novel interval type-2 fuzzy clustering algorithm deployed in an extension of the traditional fuzzy sets namely Interval Type-2 Fuzzy Sets to enhance the clustering quality of FGWC. Some additional techniques such as the interval context variable, Particle Swarm Optimization and the parallel computing are attached to speed up the algorithm. The experimental evaluation through various case studies shows that the proposed method obtains better clustering quality than some best-known ones.

Introduction

Geo-Demographic Analysis (GDA), which was defined as "*the analysis of spatially referenced geo-demographic and lifestyle data*"[33], is one of the most interesting inter-disciplinary research topics between Geographic Information Systems and Data Mining, and is widely used in the public and private sectors for the planning and provision of products and services. There are various examples showing the needs of GDA in practical applications. Shelton et al. [34] performed a geo-demographic classification for mortality patterns in Britain and found the main causes of deaths in England and Wales from 1981 to 2000 associated with geographical locations in a map so that they could assist decision makers in better understanding the distribution of major causes. Michael [23] conducted a GDA analysis to gather community attitudes on the future growth of Werri Beach and Gerringong, NSW (Nelson), Australia focusing primarily on what actions Council should take to manage population growth within existing neighborhoods. Páez et al. [29] presented a geo-demographic framework using data from Montreal, Canada to identify potential commercial partnerships that could exploit the characteristics of smart cards. Campbell et al. [8] provided a detailed GDA of over 37,000 gifted and talented students admitted to the National Academy for Gifted and Talented Youth in England in 2003/2005 and showed that National Academy had nonetheless reached significant numbers of students in the poorest areas, something over 3000 students, and 8% of students identified as gifted and talented at this stage. Day et al. [11] took a survey that determined clusters of nations grouped by health outcomes by comparing life expectancy and a range of health system indicators within and between each cluster in order to provide sensible groupings for international comparisons. Some other typical applications of GDA such as the spatial and socio-economic determinants of tuberculosis, urban green space accessibility for different ethnic and religious groups, children disorders investigation, etc. could be referenced in the articles [1,6,9,32,36,37].

In order to perform GDA, some soft computing methods are often used such as *Principal Component Analysis* (PCA), *Self-Organizing Maps* (SOM) and clustering. Walford [41] described a method using PCA to study the spatial distribution of the 1991 census data scores. However, results of PCA depend on the scaling of the variables, and its applicability is limited by certain assumptions made in the derivation. Loureiro et al. [21] introduced the use of SOM as an adequate tool for GDA. Based on the variations in edge length in a path between two units on the SOM, the authors presented a new way of calculating fuzzy memberships of fuzzy clustering method. However, it requires a lot of memory spaces to store all neurons and weights; what is more the speed of training

* Correspondence to: 334 Nguyen Trai, Thanh Xuan, Hanoi 010000, Viet Nam. Tel.: +84 904171284; fax: +84 0438623938.
E-mail addresses: sonlh@vnu.edu.vn, chinhson2002@gmail.com

http://dx.doi.org/10.1016/j.asoc.2014.04.025
1568-4946/

phase is quite slow. Because of some limitations in those methods, clustering is often used instead because it has many advantages in comparison with the rests such as the fast processing time, the quality of results and the used memory space. Our previous work in [36] made an overview about some clustering methods for GDA such as Fuzzy C-Mean (FCM) [3], the agglomerative hierarchical clustering [11], Neighborhood Effects (NE) [13], K-Means clustering [20] and Fuzzy Geographically Weighted Clustering (FGWC) [24]. Among them, FGWC was considered the most favorite algorithm and was used in most of research articles about GDA applications.

$$u'_k = \alpha \times u_k + \beta \times \frac{1}{A} \times \sum_{j=1}^{c} w_{kj} \times u_j \quad (1)$$

$$\alpha + \beta = 1 \quad (2)$$

$$w_{kj} = \frac{(pop_k \times pop_j)^b}{d_{kj}^a} \quad (3)$$

FGWC calculates the influence of one area upon another by Eqs. (1)–(3) where u_k' (u_k) is the new (old) cluster membership of the area k. Two parameters α and β are the scaling variables. pop_k, pop_j are the populations of areas k and j, respectively. The number d_{kj} is the distance between k 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.

Although FGWC is the most popular clustering algorithm for GDA, it still contains some limitations such as the speed of computing and the clustering quality. One of our previous works in [35] presented a method so-called CFGWC to accelerate the speed of computing of FGWC by attaching the context variable terms. Other works in [36,37] have showed some preliminary results in improving the clustering quality of FGWC through intuitionistic fuzzy sets and geographical spatial effects. Thus, *our focus* in this work is to continue with the clustering quality problem of FGWC. Based upon the observation that FGWC was constructed on the basis of the traditional fuzzy sets, which contain some limitations in membership degrees as pointed out by Mendel [25], this fosters us to improve FGWC in an extension of the traditional fuzzy sets to enhance the clustering quality of the algorithm. Now, let us explain why clustering algorithms on the traditional fuzzy sets have low clustering quality.

According to Mendel [25], the traditional fuzzy sets cannot process some exceptional cases where the membership degrees are not the crisp values but the fuzzy ones instead. For example, the possibility to get tuberculosis disease of a patient concluded by a doctor is from 60 to 80 percents after examining all symptoms. Even if some modern medical machines are provided, the doctor cannot give an exact number of that possibility. This shows the fact that crisp membership values cannot model some situations in the real world and should be replaced with the fuzzy ones. Rhee [30] stated that using the traditional fuzzy sets often results in bad clustering quality because their uncertainties such as distance measure, fuzzifier, centers, prototype and initialization of prototype parameters can create imperfect representations of the pattern sets. For example, in case of pattern sets that contain clusters of different volume or density, it is possible that patterns staying on the left side of a cluster may contribute more for the other rather than this cluster so that choosing suitable value for the fuzzifier is difficult. Bad selection can yield undesirable clustering results for pattern sets that include noise. Because of those limitations, some preliminary results of deploying fuzzy clustering methods in an extension of the traditional fuzzy sets so-called *Interval Type-2 Fuzzy Sets* (IT2FS) have been introduced. Mendel [25] described the definition of IT2FS as follows.

$$\tilde{A} = \left\{ (x, u, \mu_{\tilde{A}}(x, u) = 1) | \forall x \in A, \forall u \in J_X \subseteq [0, 1] \right\}. \quad (4)$$

From Eq. (4), we recognize that IT2FS is a generalization of the traditional fuzzy sets since IT2FS will return to the traditional fuzzy sets when there is no uncertainty in the third dimension. Based upon this definition, some authors introduced several interval type-2 fuzzy clustering algorithms such as in the works of Hwang and Rhee [15] and Rhee [30]. Specifically, Hwang and Rhee [15] presented a type-2 fuzzy clustering algorithm to solve the problem of choosing distance measures in FCM algorithm, taking the difference of each type-2 membership function area with the corresponding type-1 membership value. Rhee [30] presented an improvement of this algorithm using two different values of fuzzifiers to solve the uncertainty of fuzzifier in FCM. Some other variants of the interval type-2 fuzzy clustering algorithms could be referenced in [2,10,12,14,17,19,22,26,27,31,42].

Motivated by those results, in this article, we will present a novel interval type-2 fuzzy clustering algorithm so-called *Context Fuzzy Geographically Weighted Clustering on IT2FS* or in short CFGWC2 to enhance the clustering quality of FGWC. The difference of CFGWC2 with those interval type-2 fuzzy clustering algorithms above is two fold: *Firstly*, CFGWC2 is specially designed for the GDA problem that requires the modification of geographical spatial effects to the algorithm itself; *secondly*, it is equipped with some additional techniques to speed up the whole algorithm, namely:

- *An interval context variable*, which is an extension of the single context variable of Pedrycz [28], is proposed and used to clarify the clustering results and accelerate the computing speed.
- In order to avoid bad initialization, which may occur in other interval type-2 fuzzy clustering algorithms, and to converge quickly to the (sub-) optima solutions, a meta-heuristic optimization method namely *Particle Swarm Optimization* – PSO [18] is used to determine good initial centers for CFGWC2.
- Since context values in the interval context variable can be simultaneously processed in CFGWC2, *parallel computing technique* is adapted to CFGWC2 to reduce the computational costs.

What have been listed in those bullets are *our contributions* in this paper. The proposed algorithm will be implemented and compared with some relevant methods in term of clustering quality to verify its efficiency.

The rests of this paper are organized as follows. Section "The proposed methodology" elaborates the proposed method in details including those additional techniques one-after-another. The numerical experiments through various case studies and discussions are given in Section "Results". Finally, Section "Conclusions" gives the conclusions and outlines future works of this article.

The proposed methodology

In the previous section, we have known that CFGWC2 is an interval type-2 fuzzy clustering algorithm equipped with some additional techniques such as the interval context variable, PSO and the parallel computing for the GDA problem. Since those techniques are necessary for the description of CFGWC2, they are firstly presented in Sections "Using PSO for the determination of initial centers" and "The interval context". The CFGWC2 algorithm accompanied with the parallel computing mechanism will be described in Section "Evaluation by various case studies".