Information Sciences 317 (2015) 202–223

Contents lists available at ScienceDirect

Information Sciences

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

A novel kernel fuzzy clustering algorithm for Geo-Demographic Analysis

Le Hoang Son *

VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan, Hanoi, Viet Nam

ARTICLE INFO

Article history:
Received 3 January 2014
Received in revised form 19 April 2015
Accepted 25 April 2015
Available online 6 May 2015

Keywords:
Fuzzy clustering
Geo-Demographic Analysis
Intuitionistic possibilistic fuzzy clustering
Kernel-based clustering
Spatial Interaction – Modification Model

ABSTRACT

Geo-Demographic Analysis (GDA) is a major concentration of various interdisciplinary researches and has been used in many decision-making processes regarding the provision and distribution of products and services in society. Machine learning methods namely Principal Component Analysis, Self-Organizing Map, K-Means, fuzzy clustering and fuzzy geographically weighted clustering were proposed to enhance the quality of GDA. Among them, the state-of-the-art method – Modified Intuitionistic Possibilistic Fuzzy Geographically Weighted Clustering (MIPFGWC) has some drawbacks such as: (i) using the Euclidean similarity measure often results in high error rate and sensitivity to noises and outliers; (ii) updating the membership matrix by the Spatial Interaction – Modification Model (SIM^2) model leads to new centers not being "geographically aware". In this paper, we present a novel fuzzy clustering algorithm named as Kernel Fuzzy Geographically Clustering (KFGC) that utilizes both the kernel similarity function and the new update mechanism of the SIM^2 model to remedy the disadvantages of MIPFGWC. Some supported properties and theorems of KFGC are also examined in the paper. Specifically, the differences between solutions of KFGC and those of MIPFGWC and of some variants of KFGC are theoretically validated. Lastly, experimental analysis is performed to compare the performance of KFGC with those of the relevant algorithms in terms of clustering quality.

1. Introduction

Prior to the definition of Geo-Demographic Analysis (GDA) problem, let us consider an example to demonstrate the roles of GDA to practical applications.

Example 1. A hot-spot analysis for the number of viral hemorrhagic fever cases in Vietnam in 2011 is examined in Fig. 1. The results are expressed in a map showing various groups determined by intervals of cases such as [0,47] and [48,233]. From this fact, decision makers could observe the most dangerous places and issue appropriate medical measures to prevent such the situation in the future. The distribution can be expressed by linguistic labels such as "High cases of viral hemorrhagic fever" and "Low cases of viral hemorrhagic fever" to eliminate the limitations of boundary points in the intervals. Such hot-spot analysis in this example is a kind of Geo-Demographic Analysis regarding the classification of a geographical area according to a given subject, e.g. viral hemorrhagic fever cases. The classification could be done for spatial data both in point

* Tel.: +84 904171284; fax: +84 0438623938.
E-mail addresses: sonlh@vnu.edu.vn, chinhson2002@gmail.com

http://dx.doi.org/10.1016/j.ins.2015.04.050
0020-0255/

Lists of abbreviation

Terms	Explanation
GIS	Geographical Information Systems
GDA	Geo-Demographic Analysis
SIM	Spatial Interaction Model
SIM-PF	Spatial Interaction Model with Population Factor
SIM^2	Spatial Interaction – Modification Model
SOM	Self-Organizing Maps
PCA	Principal Component Analysis
K-Means	a hard partition based clustering algorithm
FCM	Fuzzy C-Means
NE	Neighborhood Effects
FGWC	Fuzzy Geographically Weighted Clustering
IPFGWC	Intuitionistic Possibilistic Fuzzy Geographically Weighted Clustering
MIPFGWC	Modified Intuitionistic Possibilistic Fuzzy Geographically Weighted Clustering
KFGC	Kernel Fuzzy Geographically Clustering
IFV	a spatial clustering quality validity index
UNO	United Nations Organization

and region standards of Geographical Information Systems (GIS), and a number of points/regions that both share common characteristics in the spatial and attribute data forming a group marked by a unique symbol and color[1] in the map. Intuitively, GDA can be regarded as the spatial clustering in GIS.

Definition 1. Given a geo-demographic dataset X consisting of N data points where each data point is equivalent to a point/region of spatial data in GIS. This data point is characterized by many geo-demographic attributes where each one could be considered as a subject for clustering. The objective of GDA is to classify X into C clusters so that,

$$J = \sum_{k=1}^{N}\sum_{j=1}^{C} u_{kj} \times \|X_k - V_j\| \to \min, \tag{1}$$

$$\begin{cases} u_{kj} \in [0,1], \\ \sum_{j=1}^{C} u_{kj} = 1, \\ u_{kj} = u_{kj}(w_j), \\ V_j = V_j(w_j), \\ k = \overline{1,N}; \quad j = \overline{1,C}, \end{cases} \tag{2}$$

where u_{kj} is the membership value of data point X_k ($k = \overline{1,N}$) to cluster jth ($j = \overline{1,C}$). V_j is the center of cluster jth ($j = \overline{1,C}$). w_j is the weight of cluster jth ($j = \overline{1,C}$) showing the influence of spatial relationships in a map. It is often calculated through a spatial model such as *Spatial Interaction Model* (SIM) [6], *Spatial Interaction Model with Population Factor* (SIM-PF) [14,25,27] or *Spatial Interaction – Modification Model* (SIM^2) [26].

GDA is widely used in the public and private sectors for planning and provision of products and services. In GDA, geo-demographic attributes are used to characterize essential information of population at a certain geographical area and a specific point of time. Some common geo-demographics can be named but a few such as gender, age, ethnicity, knowledge of languages, disabilities, mobility, home ownership, and employment status. One of the most useful functions of GDA is the capability to visualize the geo-demographic trends by locations and time stamps such as the study of the average age of a population over time and the investigation of the migration trends of local people in a town. As illustrated in Example 1, the results of GDA are depicted on a map that demonstrates the distribution of several distinct groups. Various distribution maps can be combined or overlapped into a single one so that users could observe the tendency of a certain group over time for the analyses of geo-demographic trends. Both distributions and trends of values within a geo-demographic variable are of interest in GDA. By providing essential information about geo-demographic distributions and trends, GDA assists effectively for many decision-making processes involving the provision and distribution of products and services in society, the determination of common population's characteristics and the study of population variation. This clearly demonstrates the important role of GDA to practical applications nowadays.

[1] For interpretation of color in Figs. 1 and 3, the reader is referred to the web version of this article.