Expert Systems with Applications 38 (2011) 11829–11836

Contents lists available at ScienceDirect

Expert Systems with Applications

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

The extended fuzzy C-means algorithm for hotspots in spatio-temporal GIS

Ferdinando Di Martino, Salvatore Sessa *

Università degli Studi di Napoli Federico II, Dipartimento di Costruzioni e Metodi Matematici in Architettura, Via Monteoliveto 3, 80134 Napoli, Italy

A R T I C L E I N F O

Keywords:
EFCM
FCM
GIS
Hotspot

A B S T R A C T

In spatial analysis buffer impact areas are called hotspots and are determined by means of density clustering methods. In a previous work, we found these hotspots in the context of a Geographic Information System (GIS) by using the extended fuzzy C-means (EFCM). Here we show how the spatial distribution of the hotspots can evolve temporally and like applicational example, we present the spatial-temporal evolution in the period 2000–2006 of the fire point-events data of the Santa Fè district (NM) (downloaded from URL: www.fs.fed.us/r3/gis/sfe_gis.shtml).

1. Clustering algorithms and GIS

It is well known that the primitive operations in a Geographic Information System (GIS) concern points, lines and polygons. Thus the geographic location, for instance, of a criminal event is represented from a point, which is center of a circle considered as buffer influence area of that event and its radius is usually called the buffer distance. The buffer area is called hotspot (Chainey, Reid, & Stuart, 2002; Grubesic & Murray, 2001). Another typical example of hotspot is the circle having center in the epicenter of a earthquake.

When the user faces with many point-events data, the buffer distance can also be a constant value for all these points (cf. Fig. 1) and the related circles are merged for creating new buffer areas (cf. Fig. 2). If one must manipulate a huge number of point-events, the classical density clustering methods are not adequate in the determination of hotspots.

Based fuzzy C-means algorithms (FCM) (Bezdek, 1981) seem more adequate: indeed it has been used in the determination of hotspots with high number of crimes (e.g., Chainey et al., 2002; Harries, 1999; McGuire & Williamson, 1999; Murray, McGuffog, Western, & Mullins, 2001). As in our previous papers (Di Martino, Loia, & Sessa, 2007; Di Martino & Sessa, 2009) we continue to propose the usage of the extended fuzzy C-means (EFCM) (Kaymak, Babuska, Setnes, Verbruggen, & van Nauta Lemke, 1997; Kaymak & Setnes, 2002) algorithm for three advantages with respect to the FCM's: robustness to noise and outliers, linear computational complexity and automatic determination of the optimal number of clusters. After the determination of the hotspots, we analyze their temporal evolution between two successive years. In GIS literature the spatio-temporal evolution of some topics is already known (cf., e.g., Liu, Zhang, Cai, & Tong, 2010). In Section 2 we give an overview of the EFCM algorithm. In Section 3 we give an application of the EFCM algorithm in the specific problem of fire prevention of a forest area, located in New Mexico: indeed we construct hotspots which represent dangerous areas of fire point-events. In Section 4 we study their spatio-temporal evolution in the period 2000–2006, by making a comparison of the hotspots for every pair of successive years. Section 5 concludes the paper.

2. EFCM algorithm: an overview

For sake of completeness, we recall the main steps of the EFCM (Kaymak & Setnes, 2002). Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset R^n$, $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T \in R^n$, be the dataset, where x_{ij} is the jth component (feature) of the vector $\mathbf{x}_i$, $j = 1, \ldots, N$. In the classical FCM algorithm the objective function to be minimized is the following:

$$J(\mathbf{X}, \mathbf{U}, \mathbf{V}) = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2 \tag{1}$$

where C is the initial (fixed a priori) number of the clusters, u_{ij} is the membership degree of $\mathbf{x}_j$ to the ith cluster ($i = 1, \ldots, C$) forming the generic entry of the matrix $\mathbf{U}$, $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_C\} \subset R^n$ is the set of the centers of the C clusters (prototypes), m is the fuzzifier parameter and d_{ij} is the distance between the center $v_i = (v_{1i}, v_{2i}, \ldots, v_{nj})^T$ of the ith cluster and the jth vector x_j, calculated as

$$d_{ij} = \sqrt{(\mathbf{x}_j - \mathbf{v}_i)^T S (\mathbf{x}_j - \mathbf{v}_i)} \tag{2}$$

being S is a symmetric norm matrix. The center of each cluster prototype is defined as

$$\mathbf{v}_i = \frac{\sum_{j=1}^{N} u_{ij}^m \mathbf{x}_j}{\sum_{j=1}^{N} u_{ij}^m} \tag{3}$$

* Corresponding author. Tel.: +39 0812538907; fax: +39 0812538909.
E-mail addresses: fdimarti@unina.it (F. Di Martino), sessa@unina.it (S. Sessa).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.03.071

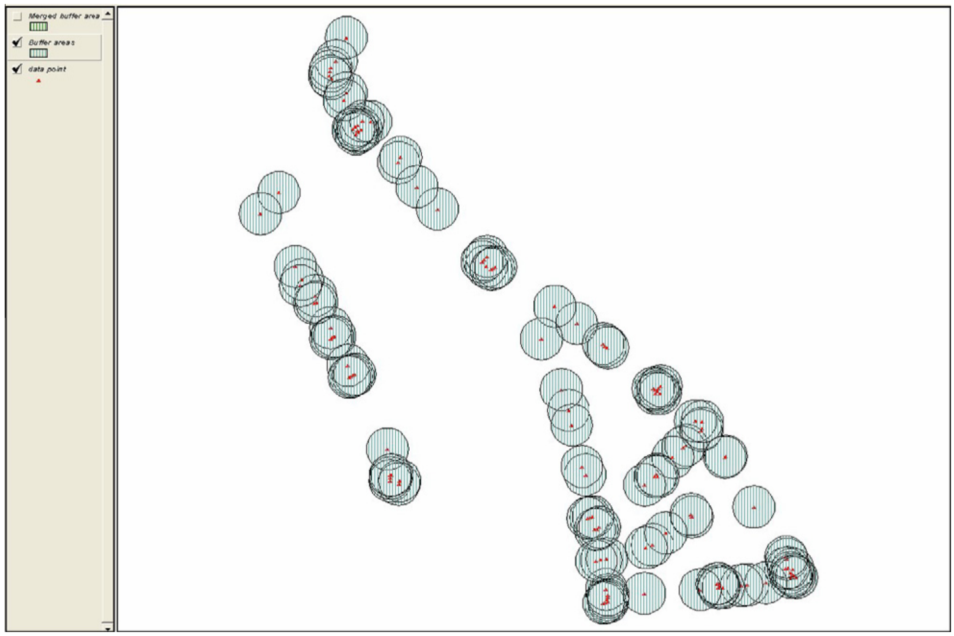

Fig. 1. Example of circular buffer areas.

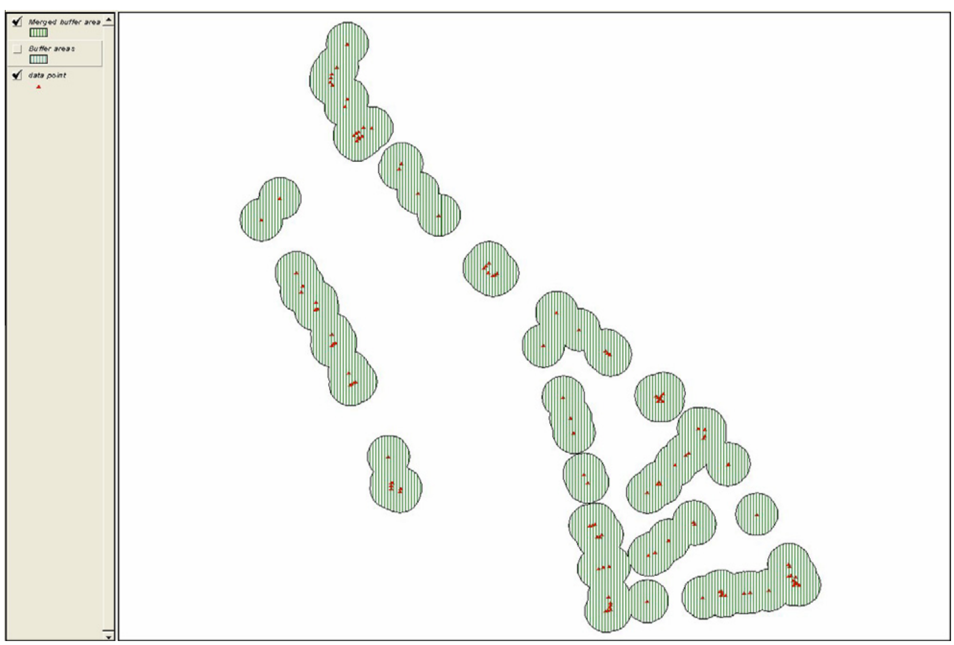

Fig. 2. Example of merged buffer areas.

for $i = 1,\ldots,C$ and the membership degrees u_{ij} are given by

$$u_{ij} = \frac{1}{\left(\sum_{k=1}^{C} \frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{2}{m-1}}} \tag{4}$$

subjected to the constraints:

$$\sum_{i=1}^{C} u_{ij} = 1 \quad \forall j \in \{1,\ldots,N\} \tag{5}$$

$$0 < \sum_{j=1}^{N} u_{ij} < N \quad \forall i \in \{1,\ldots,C\} \tag{6}$$

Initially the u_{ij}'s and the $\mathbf{v}_i$ are assigned randomly and updated in each iteration. If $\mathbf{U}^{(l)} = \left(u_{ij}^{(l)}\right)$ is the matrix $\mathbf{U}$ calculated at the lth step, the iterative process stops when

$$\left\|\mathbf{U}^{(l)} - \mathbf{U}^{(l-1)}\right\| = \max_{i,j}\left|u_{ij}^{(l)} - u_{ij}^{(l-1)}\right| < \varepsilon \tag{7}$$

where $\varepsilon > 0$ is a prefixed parameter. Generally speaking, in the FCM the distribution of the points-data is sensitive to the initialization phase. Indeed there are two noteworthy shortcomings: the (a priori) choice of the parameter C could influence the final number of clusters and moreover the number of the zones with low density of points data could be enormous, so influencing the final distribution of the points in the various clusters.

In order to overcome these shortcomings, the authors (Kaymak et al., 1997; Kaymak & Setnes, 2002) propose the EFCM algorithm, in which the cluster prototypes are hyperspheres in the case of the Euclidean metric. Indeed, if r_i is the radius of V_i, we say that x_j