Expert Systems With Applications 51 (2016) 143–150

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A new fast fuzzy partitioning algorithm

Rudolf Scitovski[a], Ivan Vidović[b,*], Dražen Bajer[b]

[a] Department of Mathematics, University of Osijek, Trg Lj. Gaja 6, HR, Osijek 31000, Croatia
[b] Faculty of Electrical Engineering, University of Osijek, Cara Hadrijana 10b, HR, Osijek 31000, Croatia

ARTICLE INFO

Keywords:
Fuzzy clustering
Fuzzy *c*-means
Fuzzy locally optimal partition
Fuzzy globally optimal partition
DIRECT
Incremental algorithm

ABSTRACT

In this paper, a new fast incremental fuzzy partitioning algorithm able to find either a fuzzy globally optimal partition or a fuzzy locally optimal partition of the set $\mathcal{A} \subset \mathbb{R}^n$ close to the global one is proposed. This is the main impact of the paper, which could have an important role in applied research. Since fuzzy k-optimal partitions with $k = 2, 3, \ldots, k_{max}$ clusters are determined successively in the algorithm, it is possible to calculate corresponding validity indices for every obtained partition. The number k_{max} is defined in such a way that the objective function value of optimal partition with k_{max} clusters is relatively very close to the objective function value of optimal partition with $(k_{max}-1)$ clusters. Before clustering, the data are normalized and afterwards several validity indices are applied to partitions of the normalized data. Very simple relationships between used validity indices on normalized and original data are given as well. Hence, the proposed algorithm is able to find optimal partitions with the most appropriate number of clusters. The algorithm is tested on numerous synthetic data sets and several real data sets from the UCI data repository.

1. Introduction

Clustering of a data set into conceptually meaningful groups, yielding a partition of that data set, is a problem widely encountered in different research areas. In this paper, we consider the problem of searching for a *Fuzzy Globally Optimal Partition* (FGOP) with the most appropriate number of clusters. Fuzzy clustering serves as an important tool for solving specific problems in different areas of applied research. Let us mention only image and signal processing, medical diagnosis, tomography, neural networks, chemistry, biology, astronomy, speech recognition, environmental sciences, etc. (Bandyopadhyay & Saha, 2013; Bezdek, Keller, Krisnapuram, & Pal, 2005; Theodoridis & Koutroumbas, 2009).

The best known method for solving this problem (in some particular sense) uses the well known fuzzy *c*-means (FCM). It gives a locally optimal partition, whose closeness to the FGOP strongly depends on an initial approximation. The choice of initial centers impacts directly its convergence speed, as well as the quality of partitions found. Stetco, Zeng, and Keane (2015) presented a new approach for choosing the initial centers for FCM, with the aim of increasing the convergence speed. It is based on the technique proposed by Arthur and Vassilvitskii (2007). The first center is picked uniformly at random from the data set, while the others are picked probabilistically, favoring data points more distant from the already selected ones. A drawback of the presented approach is that it introduces an additional user parameter which controls the probabilities assigned to data points.

Direct applications of well-known global optimization methods are not acceptable for searching FGOP (Grbić, Nyarko, & Scitovski, 2013; Paulavičius & Žilinskas, 2014; Pintér, 1996) because of the large number of independent variables in objective function (coordinates of cluster centers) and large number of stationary points of objective function, but different hybrid approaches were proposed that combine global optimization methods with FCM, thus, typically, yielding better performance (Tvrdik & Křivý, 2015). Most often bio-inspired optimization algorithms are used. Silva Filho, Pimentel, Souza, and . Oliveira (2015) presented a hybrid approach that is based on the one proposed by Izakian and Abraham (2011). Both combine particle swarm optimization and FCM, where an initial search performed by particle swarm optimization is followed by FCM in order to improve those partitions. Another hybrid approach for the localization of retinal blood vessels, combining an artificial bee colony optimization algorithm and pattern search, was proposed in Hassanien, Emary, and Hossam (2015). After the cluster centers are obtained by artificial bee colony optimization, the pattern search algorithm is utilized in order to refine them furthermore. Another example can be found in Tang, Zhang, Wang, Wang, and Liu (2014), where the authors proposed a hybrid for the estimation of missing traffic data, which combines FCM and a

* Corresponding author. Tel.: +385 31 495 422; fax: +385 31 224 605.
E-mail addresses: scitowsk@mathos.hr (R. Scitovski), ividovi2@etfos.hr (I. Vidović), dbajer@etfos.hr (D. Bajer).

http://dx.doi.org/10.1016/j.eswa.2015.12.034
0957-4174/

genetic algorithm. The FCM algorithm is used for generating an initial data model which is subsequently refined by the genetic algorithm. The refinement is performed as long as the deviation of the model output on known data is not less than a user defined threshold. A number of hybrid approaches have also been proposed for hard clustering. For example, Tvrdik and Křivý (2015) proposed a hybrid of differential evolution and k-means.

All the approaches mentioned so far have one significant drawback in common. Namely, all assume that the number of clusters inherent to a data set is known a priori. Thus, their applicability is limited to such cases. Nonetheless, different approaches attempting to solve that problem can be found in the literature. Kashan, Rezaee, and Karimiyan (2013) proposed an approach for automatic fuzzy clustering adapting the grouping evolution strategies, which was originally proposed for hard clustering. Another approach for automatic fuzzy clustering has been proposed by Peng, Wang, Shi, Riscos-Nunez, and Perez-Jimenez (2015). It is based on a tissue-like P system and uses, with regard to the previously mentioned approach, fixed size objects for representing cluster centers.

In our paper, a new fast incremental fuzzy partitioning algorithm able to find either a FGOP or a fuzzy locally optimal partition close to the global one is proposed. The algorithm successively searches for optimal partitions with $k = 2, 3, \ldots$ clusters until the relative difference in the objective function values becomes smaller than some small $\epsilon > 0$. In each iteration the FCM and DIRECT algorithm for global optimization are combined. The drawback of the proposed algorithm lies in the fact that we cannot be certain if a FGOP or only a partition close to a FGOP was found. Numerous experiments conducted on synthetic data sets and on real data sets from the UCI data repository show that the algorithm gives a FGOP or a partition very close to a FGOP. Furthermore, the algorithm is suitable for applying different validity indices in order to determine the most appropriate number of clusters in a partition.

The paper is organized as follows. In the next section, the problem statement is given. In Section 3, the problem for searching for a fuzzy locally and globally optimal partition is described and a new efficient incremental algorithm for searching for a globally optimal partition is proposed. In Section 4, the proposed algorithm is tested on several real data sets. Finally, some conclusions are given in Section 5.

2. Problem statement

Given is a data points set $\mathcal{A} = \{a^i = (a^i_1, \ldots, a^i_n) : i = 1, \ldots, m\} \subset [\alpha, \beta] \subset \mathbb{R}^n$, where $\alpha = (\alpha_1, \ldots, \alpha_n)^T$, $\beta = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n$ and $[\alpha, \beta] = \{x \in \mathbb{R}^n : \alpha_i \le x_i \le \beta_i\}$.

If components $a^i_1, \ldots, a^i_n$ of the data point $a^i \in \mathcal{A}$ are not of equal range, i.e. if components of the vector α, resp. vector β, are mutually significantly different, they should be normalized first. This means that set $\mathcal{A}$ should be transformed into set $\mathcal{B} = \{\mathcal{T}(a^i) : a^i \in \mathcal{A}\} \subset [0, 1]^n$ by the mapping $\mathcal{T} : [\alpha, \beta] \to [0, 1]^n$, where

$$T(x) = D(x - \alpha), \quad D = \operatorname{diag}\left(\tfrac{1}{\beta_1 - \alpha_1}, \ldots, \tfrac{1}{\beta_n - \alpha_n}\right). \tag{1}$$

After clustering set $\mathcal{B}$, the obtained results will be transformed back into interval $[\alpha, \beta]$ by the mapping $\mathcal{T}^{-1} : [0, 1]^n \to [\alpha, \beta]$, where

$$T^{-1}(x) = D^{-1}x + \alpha. \tag{2}$$

Therefore, it is furthermore assumed that the whole data set $\mathcal{A}$ is contained in the hypercube $[0, 1]^n$.

A partition of the set $\mathcal{A} = \{a^i \in [0, 1]^n : i = 1, \ldots, m\} \subset \mathbb{R}^n$ into k disjoint subsets $\pi_1, \ldots, \pi_k$, $1 \le k \le m$, such that

$$\bigcup_{j=1}^{k} \pi_j = \mathcal{A}, \quad \pi_r \cap \pi_s = \emptyset, \quad r \ne s, \quad |\pi_j| \ge 1, \quad j = 1, \ldots, k, \tag{3}$$

will be denoted by $\Pi(\mathcal{A}) = \{\pi_1, \ldots, \pi_k\}$ and the set of all such partitions will be denoted by $\mathcal{P}(\mathcal{A}; k)$. The elements $\pi_1, \ldots, \pi_k$ of the partition Π are called *clusters*.

If $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$, $\mathbb{R}_+ = [0, +\infty\rangle$ is some distance-like function (see e.g. Kogan (2007); Teboulle (2007)), then to each cluster $\pi_j \in \Pi$ we can associate its center c_j defined by

$$c_j := \operatorname*{argmin}_{x \in [0,1]^n} \sum_{a^i \in \pi_j} d(x, a^i). \tag{4}$$

After that, by introducing the objective function $\mathcal{F} : \mathcal{P}(\mathcal{A}; k) \to \mathbb{R}_+$, the quality of a partition can be defined, and searching for the *globally optimal k-partition* comes down to solving the following optimization problem

$$\operatorname*{argmin}_{\Pi \in \mathcal{P}(\mathcal{A};k)} \mathcal{F}(\Pi), \quad \mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a^i \in \pi_j} d(c_j, a^i), \quad c = (c_1, \ldots, c_k). \tag{5}$$

Conversely, for a given set of centers $c_1, \ldots, c_k \in [0, 1]^n$, by applying the minimal distance principle, the partition $\Pi = \{\pi(c_1), \ldots, \pi(c_k)\}$ of set $\mathcal{A}$ consisting of clusters:

$$\pi(c_j) = \{a \in \mathcal{A} : d(c_j, a) \le d(c_s, a),\ \forall s = 1, \ldots, k\}, \quad j = 1, \ldots, k,$$

can be defined. Thereby, one has to take into account that every element of set $\mathcal{A}$ occurs in one and only one cluster. Therefore, the problem of finding an optimal partition of set $\mathcal{A}$ can be reduced to the following *global optimization problem* (see e.g. Späth (1983); Teboulle (2007))

$$\operatorname*{argmin}_{c \in [0,1]^{n \times k}} F(c), \quad F(c) = \sum_{i=1}^{m} \min_{1 \le j \le k} d(c_j, a^i). \tag{6}$$

The solutions of (5) and (6) coincide (Scitovski & Scitovski, 2013; Späth, 1983).

Furthermore, let $U \in \{0, 1\}^{m \times k}$ be a matrix such that

$$u_{ij} = \begin{cases} 1, & \text{if } a^i \in \pi_j \\ 0, & \text{if } a^i \notin \pi_j \end{cases}, \quad i = 1, \ldots m, \quad j = 1, \ldots, k, \tag{7}$$

$$\sum_{j=1}^{k} u_{ij} = 1, \quad i = 1, \ldots, m. \tag{8}$$

Then (5) can be rewritten as (Bezdek, Ehrlich, & Full, 1984; Bezdek et al., 2005; Theodoridis & Koutroumbas, 2009)

$$\operatorname*{argmin}_{c \in [0,1]^{n \times k},\ u_{ij} \in \{0,1\}} F(c, U), \quad F(c, U) = \sum_{i=1}^{m} \sum_{j=1}^{k} u_{ij} d(c_j, a^i). \tag{9}$$

In order to ensure all conditions from (3), the following should be added to conditions (7) and (8):

$$\sum_{i=1}^{m} u_{ij} \ge 1, \quad j = 1, \ldots, k. \tag{10}$$

Assuming that elements $a^i \in \mathcal{A}$ can partially belong to different clusters, then, due to (8), it must be $u_{ij} \in [0, 1]$ (Scitovski & Sabo, 2014). According to (Bezdek et al., 1984; Bezdek et al., 2005; Theodoridis & Koutroumbas, 2009), the membership grade of a^i in cluster π_j is determined by u^q_{ij}, where parameter $q > 1$ is called the *fuzzifier*, and the objective function becomes

$$F(c, U) = \sum_{i=1}^{m} \sum_{j=1}^{k} u^q_{ij}(c)\, d(c_j, a^i). \tag{11}$$