Computers & Geosciences 59 (2013) 124–131

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

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

A fast partitioning algorithm and its application to earthquake investigation

Rudolf Scitovski [a,*], Sanja Scitovski [b]

[a] Department of Mathematics, University of Osijek, Trg Lj. Gaja 6, HR – 31 000 Osijek, Croatia
[b] Faculty of Civil Engineering, University of Osijek, Crkvena 21, HR – 31 000 Osijek, Croatia

ARTICLE INFO

Article history:
Received 29 September 2012
Received in revised form
3 June 2013
Accepted 10 June 2013
Available online 26 June 2013

Keywords:
Center-based clustering
Globally optimal partition
Approximate optimization
DIRECT
Earthquake
Seismic activity

ABSTRACT

In this paper a new fast partitioning algorithm able to find either a globally optimal partition or a locally optimal partition of the set $\mathcal{A}\subset\mathbb{R}^n$ close to the global one is proposed. The performance of the algorithm in terms of CPU time shows significant improvement in comparison with other incremental algorithms. Since optimal partitions with 2, 3,… clusters are determined successively in the algorithm, it is possible to calculate corresponding clustering validity indexes for every number of clusters in a partition. In that way the algorithm also proposes the appropriate number of clusters in a partition. The algorithm is illustrated and tested on several synthetic and seismic activity data from a wider area of the Republic of Croatia in order to locate the most intense seismic activity in the observed area.

1. Introduction

Clustering or grouping a data set into conceptually meaningful clusters is a well-studied problem in recent literature. It has practical importance in a wide variety of applications such as earthquake investigation, pattern recognition, facility location problem, text classification, machine learning, business, biology, agriculture, medicine, psychology, etc. (Adelfio et al., 2012; Colombo et al., 1997; Durak, 2011; Iyigun, 2007; Morales-Esteban et al., 2010; Pintér, 1996; Sabo et al., 2011, 2013; Vazler et al., 2012).

Searching for an optimal partition in general is a complex global optimization problem which can have several local and global minima (Grbić et al., in press; Evtushenko, 1985; Pardalos and Coleman, 2009; Pintér, 1996). Hence, numerous methods simplifying the problem are proposed in literature, but they may not lead to a globally optimal partition.

The second problem in cluster analysis that is often considered is determining the appropriate number of clusters in a partition. If the number of clusters is not given in advance, defining an appropriate number of clusters in a partition is a complex problem (see e.g. Gan et al., 2007; Iyigun, 2007; Kogan, 2007; Vendramin et al., 2009).

In our paper, a new incremental algorithm of searching for an optimal partition is proposed. The algorithm represents a generalization of known incremental algorithms (Bagirov and Ugon, 2005; Bagirov, 2008; Bagirov et al., 2011; Likas et al., 2003), and uses the DIRECT algorithm for a global optimization of the Lipschitz continuous function (Gablonsky, 2001; Finkel, 2003; Jones et al., 1993) in order to find a good initial approximation for the k-means algorithm. The algorithm locates either a globally optimal partition or a locally optimal partition close to the global one.

The proposed algorithm is applied in earthquake investigation using the data freely available on the website: http://earthquake.usgs.gov/earthquakes/eqarchives/epic/. Only data that refer to a wider area of the Republic of Croatia have been extracted from the database. Using the aforementioned algorithm, spatial locations of seismic activity centers are detected.

The paper is organized as follows: In Section 2, some basic terms and facts about data clustering are mentioned. In Section 3, a new algorithm of searching for an optimal partition is constructed. A new algorithm is illustrated and compared with other similar algorithms on several synthetic and empirical examples. In Section 4, the mentioned algorithm is applied on the example of detecting spatial locations of seismic activity centers in a wider area of the Republic of Croatia. Conclusions and future work are discussed in Section 5.

2. Data clustering

The given data point set $\mathcal{A}=\{a_i\in\mathbb{R}^n : i=1,\ldots,m\}$, where $n\geq 1$ represents the number of features in the data, should be partitioned into $1\leq k\leq m$ nonempty disjoint subsets (clusters) $\pi_1,\ldots,\pi_k$.

* Corresponding author. Tel.: +385 31 224 800; fax: +385 31 224 801.
E-mail addresses: scitowsk@mathos.hr (R. Scitovski), sscitov@unios.hr (S. Scitovski).

0098-3004/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2013.06.010

Such partition will be denoted by Π, and the set of all partitions of the set $\mathcal{A}$ consisting of k clusters will be denoted by $\mathcal{P}(\mathcal{A};m,k)$.

Suppose also that a weight $w_i>0$ is associated to each data point. 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), which has at least positive definiteness property, then to each cluster $\pi_j\in\Pi$ we can associate its center c_j defined by

$$c_j=c(\pi_j):=\arg\min_{x\in\mathrm{conv}(\pi_j)}\sum_{a_i\in\pi_j}w_i d(x,a_i), \tag{1}$$

where $\mathrm{conv}(\pi_j)$ is a convex hull of the set π_j. After that, by introducing the objective function $\mathcal{F}:\mathcal{P}(\mathcal{A};m,k)\to\mathbb{R}_+$ we can define the quality of a partition and search for the *globally optimal k-partition* by solving the following optimization problem:

$$\arg\min_{\Pi\in\mathcal{P}(\mathcal{A};m,k)}\mathcal{F}(\Pi),\quad \mathcal{F}(\Pi)=\sum_{j=1}^{k}\sum_{a_i\in\pi_j}w_i d(c_j,a_i). \tag{2}$$

Conversely, for a given set of centers $c_1,\ldots,c_k\in\mathbb{R}^n$, by applying the minimal distance principle, we can define the partition $\Pi=\{\pi(c_1),\ldots,\pi(c_k)\}$ of the set $\mathcal{A}$ which consists of the clusters

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

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

$$\arg\min_{c_1,\ldots,c_k\in\mathbb{R}^n}F(c_1,\ldots,c_k),\quad F(c_1,\ldots,c_k)=\sum_{i=1}^{m}w_i\min_{1\le s\le k}d(c_s,a_i). \tag{3}$$

The solution of (2) and (3) coincides. Namely, it is easy to verify the following equalities:

$$\begin{aligned}F(c_1^\star,\ldots,c_k^\star)&=\sum_{i=1}^{m}w_i\min_{1\le s\le k}d(c_s^\star,a_i)=\sum_{j=1}^{k}\sum_{a_i\in\pi(c_j^\star)}w_i\min_{1\le s\le k}d(c_s^\star,a_i)\\&=\sum_{j=1}^{k}\sum_{a_i\in\pi(c_j^\star)}w_i d(c_j^\star,a_i)=\mathcal{F}(\Pi^\star),\end{aligned} \tag{4}$$

where $\Pi^\star=\{\pi(c_1^\star),\ldots,\pi(c_k^\star)\}$. Thereby, the objective function F is a symmetric function which can have a large number of independent variables, it does not have to be either convex or differentiable, and generally it may have at least $k!$ local and global minima (Grbić et al., in press). Therefore, this becomes a complex GOP.

2.1. Choice of a distance-like function

Among many well-known distance-like functions (Durak, 2011; Kogan, 2007; Teboulle, 2007) we will mention only two which will be used in numerical experiments in Section 3.2 and in application to seismic activity in Section 4. In some concrete applications, the choice of the corresponding distance-like function is very important.

The most popular distance-like function is *the Least Squares* (*LS*) *distance-like function* $d_{LS}:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}_+$, $d_{LS}(x,y)=\|x-y\|_2^2$. In this case, the cluster center is called the centroid and it can be simply obtained as a weighted arithmetic mean

$$c_j=\arg\min_{x\in\mathrm{conv}(\pi_j)}\sum_{a_i\in\pi_j}d_{LS}(x,a_i)=\frac{1}{W_j}\sum_{a_i\in\pi_j}w_i a_i,\quad W_j=\sum_{a_i\in\pi_j}w_i. \tag{5}$$

Centroid c_j has the property that the weighted sum of squares of Euclidean distances of points from the cluster π_j to its center c_j is minimal. From the physical point of view, the centroid c_j can be understood as a center of gravity of the set $\mathcal{A}$ with weights $w_i>0$ of its points.

Mahalanobis distance-like function $d_M:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}_+$, $d_M(x,y)=(x-y)S(x-y)^T$, ($S>0$ symmetric positive definite matrix), takes into consideration the correlations within a data set (Durak, 2011). The matrix S is a symmetric positive definite covariance matrix. It can be easily seen that the cluster center is the same when using the LS-distance-like function. Note also that the Mahalanobis distance-like function becomes an LS-distance-like function if S is the identity matrix and both of these distance-like functions have a symmetry property, but they do not satisfy the triangle inequality.

3. Searching for a globally optimal partition

Given is a data points set $\mathcal{A}\subset[\alpha,\beta]\subset\mathbb{R}^n$, where $\alpha=(\alpha_1,\ldots,\alpha_n)$, $\beta=(\beta_1,\ldots,\beta_n)\in\mathbb{R}^n$ and $[\alpha,\beta]=\{x\in\mathbb{R}^n:\alpha_i\le x_i\le\beta_i\}$, thereby to each data point $a^i\in\mathcal{A}$ a weight $w_i>0$ is associated. The goal is to determine a partition $\Pi^\star=\{\pi_1^\star,\ldots,\pi_k^\star\}$ with centers $c_1^\star,\ldots,c_k^\star$ as a solution of GOP (2), or equivalently (3).

Since our objective function (3) is a Lipschitz continuous function (Pintér, 1996; Sabo et al., 2013), there are numerous methods for solving this GOP (Evtushenko, 1985; Floudas and Gounaris, 2009; Neumaier, 2004; Pintér, 1996). One of the most popular algorithms for solving a GOP for the Lipschitz continuous function is the DIRECT (DIviding RECTangles) algorithm (Finkel, 2003; Gablonsky, 2001; Jones et al., 1993). Because of the symmetry property of the function F there are at least $k!$ solutions of this problem. That was a motive for developing a very efficient special version of the DIRECT algorithm for symmetric functions in Grbić et al. (in press). Complexity of this problem is specially emphasized if the number of features n or the number of data points m is large.

Instead of searching for the GOP, various simplifications are often proposed in the literature that would find a partition for which we usually do not know how close it is to the globally optimal one. The most popular algorithm of searching for a locally optimal partition is a well-known k-means algorithm (see e.g. Kogan, 2007; Rizman-Žalik, 2008; Späth, 1983; Teboulle, 2007). If we have a good initial approximation, this algorithm can provide an acceptable solution (Volkovich et al., 2007). In case we do not have a good initial approximation, the algorithm should be restarted with various random initializations, as proposed by Leisch (2006).

3.1. A new algorithm

Our paper proposes a new efficient algorithm of searching for an optimal partition as a natural generalization of different incremental algorithms (Likas et al., 2003; Bagirov, 2008; Bagirov et al., 2011). For that purpose we define the sequence of objective functions

$$F_k:\underbrace{\mathbb{R}^n\times\cdots\times\mathbb{R}^n}_{k}\to\mathbb{R}_+,$$

$$F_k(c_1,\ldots,c_k)=\sum_{i=1}^{m}w_i\min\{d(c_1,a^i),\ldots,d(c_k,a^i)\}. \tag{6}$$

For $k=1$, the function F_1 attains its global minimum at the point $c_1^\star\in[\alpha,\beta]$ given by (1).

For $k>1$, we determine an optimal k-partition with centers $c_1^\star,\ldots,c_k^\star$ by the following incremental algorithm.

Algorithm 1. (Searching for an optimal k-partition)

Step 1: Let $\hat{c}_1,\ldots,\hat{c}_{k-1}$ be the centers obtained in the previous step as an approximation of a global minimizer of the function F_{k-1} and let

$$F_{k-1}(\hat{c}_1,\ldots,\hat{c}_{k-1})=\sum_{i=1}^{m}w_i\delta_{k-1}^i,$$

$$\delta_{k-1}^i=\min\{d(\hat{c}_1,a^i),\ldots,d(\hat{c}_{k-1},a^i)\}, \tag{7}$$