Information Sciences 305 (2015) 208–218

Contents lists available at ScienceDirect

Information Sciences

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

An approach to cluster separability in a partition

K. Sabo *, R. Scitovski

Department of Mathematics, University of Osijek, Trg Ljudevita Gaja 6, HR-31000 Osijek, Croatia

ARTICLE INFO

Article history:
Received 23 October 2013
Received in revised form 31 January 2015
Accepted 5 February 2015
Available online 11 February 2015

2000 MSC:
62H30
68T10
90C26
90C27
91C20
47N10

Keywords:
Clustering
Data mining
Cluster separability
Separability balls

ABSTRACT

In this paper, we consider the problem of cluster separability in a minimum distance partition based on the squared Euclidean distance. We give a characterization of a well-separated partition and provide an operational criterion that gives the possibility to measure the quality of cluster separability in a partition. Especially, the analysis of cluster separability in a partition is illustrated by implementation of the *k*-means algorithm.

1. Introduction

Clustering or grouping a set of data points into conceptually meaningful clusters is a well-studied problem in recent literature [2,3,9,10,18,20,22,27], and it has practical importance in a wide variety of applications such as computer vision, signal-image-video analysis, multimedia, networks, biology, medicine, geology, psychology, business, politics and other social sciences.

Let $I = \{1, \ldots, m\}$ and $J = \{1, \ldots, k\}$. A partition of the set $\mathcal{A} = \{a_i \in \mathbb{R}^n : i \in I\}$ into k disjoint subsets $\pi_1, \ldots, \pi_k, 1 \leqslant k \leqslant m$, such that

$$\bigcup_{i=1}^{k} \pi_i = \mathcal{A}, \quad \pi_r \cap \pi_s = \emptyset, \quad r \neq s, \quad |\pi_j| \geqslant 1, \quad \forall r, s, j \in J, \tag{1}$$

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

Any function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+, \ \mathbb{R}_+ := [0, +\infty\rangle$, with the following property

* Corresponding author.
E-mail addresses: ksabo@mathos.hr (K. Sabo), scitowsk@mathos.hr (R. Scitovski).

http://dx.doi.org/10.1016/j.ins.2015.02.011
0020-0255/

$$(\forall (x,y)\in\mathbb{R}^n\times\mathbb{R}^n)\quad d(x,y)\geqslant 0\quad\text{and}\quad d(x,y)=0\Longleftrightarrow x=y,$$

is called a distance-like function (see, e.g., [10,27]). Let $d:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}_+$, be a distance-like function. Then for each cluster $\pi_j\in\Pi$ its center c_j is defined by

$$c_j=c(\pi_j):=\underset{x\in\operatorname{conv}\pi_j}{\arg\min}\sum_{a_i\in\pi_j}d(x,a_i), \tag{2}$$

where $\operatorname{conv}\pi_j$ denotes the convex hull of the cluster π_j. It is said that the partition $\Pi^\star\in\mathcal{P}(\mathcal{A},k)$ is a globally optimal k-partition if

$$\Pi^\star=\underset{\Pi\in\mathcal{P}(\mathcal{A},k)}{\arg\min}\,\mathcal{F}(\Pi),\quad \mathcal{F}(\Pi)=\sum_{j=1}^{k}\sum_{a_i\in\pi_j}d(c_j,a_i), \tag{3}$$

where $\mathcal{F}:\mathcal{P}(\mathcal{A},k)\to\mathbb{R}_+$ is the objective function.

Conversely, for a given set of different points $z_1,\ldots,z_k\in\mathbb{R}^n$, by applying the *minimum distance principle* (see, e.g., [10,24]), one can define the partition $\Pi=\{\pi(z_1),\ldots,\pi(z_k)\}$,

$$\pi(z_j)=\{a\in\mathcal{A}:d(z_j,a)\leqslant d(z_s,a),\ \forall s=1,\ldots,k\},\quad j\in J, \tag{4}$$

where a tie-breaker rule is needed in case of equality.

Therefore, the problem of finding an optimal partition of the set $\mathcal{A}$ can be reduced to the following optimization problem:

$$\underset{z_1,\ldots,z_k\in\mathbb{R}^n}{\arg\min}\,F(z_1,\ldots,z_k),\quad F(z_1,\ldots,z_k)=\sum_{i=1}^{m}\min_{1\leqslant j\leqslant k}d(z_j,a_i). \tag{5}$$

Optimization problems (3) and (5) are equivalent [24]. Global optimization problem (5) can also be found in the literature as a *center-based clustering problem* [9,12,27]. If the *squared Euclidean distance* $d:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}_+, d(x,y)=\|x-y\|^2$ is used, the function F from (5) becomes a standard *k-means* objective function. The objective function $F:\mathbb{R}^{kn}\to\mathbb{R}_+$ defined by (5) can have a large number of independent variables (the number of clusters in the partition multiplied by the dimension of data points: $k\cdot n$), it does not have to be either convex or differentiable and usually it has several local minima. Hence, this becomes a complex global optimization problem.

Furthermore, suppose that $\mathcal{A}\subset\mathbb{R}^n=\{(x_1,\ldots,x_n):x_i\in\mathbb{R}\}$ is a given set. By using the *squared Euclidean distance* $d:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}_+$, $d(x,y)=\|x-y\|^2=\langle x-y,\,x-y\rangle$, where $\langle\cdot,\cdot\rangle$ is the standard inner product, we analyze internal separability of some partition Π of the set of data points $\mathcal{A}$, i.e., we consider the following problem:

> Let $\mathcal{A}\subset\mathbb{R}^n$ be a set, d the squared Euclidean distance and $z_1,\ldots,z_k\in\mathbb{R}^n$ a set of mutually different points (*assignment points*) that determine the partition $\Pi=\{\pi(z_1),\ldots,\pi(z_k)\}$, where $\pi(z_j)$ are given by (4). The question is: *How can the assignment points be changed such that the partition* Π *remains unchanged?*

Especially, an open ball $B(\delta)=\{u\in\mathbb{R}^n:\|u\|<\delta\}$ of radius $\delta>0$ is searched for, such that for an arbitrary set of assignment points $\{\zeta_1,\ldots,\zeta_k\in\mathbb{R}^n:\zeta_j\in z_j+B(\delta)\}$ the clusters $\pi(\zeta_j)$ and $\pi(z_j)$ are equal for all $j\in J$. The ball $B(\delta)$ is said to be a *separability ball of the partition* Π and the corresponding balls

$$z_j+B(\delta):=\{z_j+u:u\in B(\delta)\},\quad j\in J,$$

will be called *separability balls associated with assignment points* $z_1,\ldots,z_k$.

Note that in this way separability balls for all clusters have the same radius δ. The problem could also be formulated such that separability balls are searched for each cluster separately.

There is a rich literature considering similar problems. Some of them will be discussed in detail in the next section, after the term cluster separability in a partition is defined and a characterization of a well-separated partition is given. The problem is first considered for the one-dimensional case, and then in detail for the n-dimensional case. In Section 3, cluster separability in a partition is illustrated by the implementation of the k-means algorithm. Finally, some conclusions are given in Section 4.

2. Cluster separability in a partition

Let $1\leqslant k\leqslant m,\ I=\{1,\ldots,m\},\ J=\{1,\ldots,k\}$, and let $\mathcal{A}=\{a_i\in\mathbb{R}^n:i\in I\}$ be a given data set in $\mathbb{R}^n$. By using the squared Euclidean distance, for a given set of assignment points $z_1,\ldots,z_k\in\mathbb{R}^n$, according to the minimum distance principle, there is a partition $\Pi=\{\pi(z_1),\ldots,\pi(z_k)\}$ made up of clusters

$$\pi(z_j)=\{a\in\mathcal{A}:\|z_j-a\|\leqslant\|z_s-a\|,\ s\in J\},\quad j\in J. \tag{6}$$

Note that each cluster $\pi(z_j)$ depends on the neighboring clusters, and notation $\pi(z_j)$ implies that cluster $\pi(z_j)$ is associated to the center z_j. It is well-known (see, e.g., [10]) that it may happen that some of the clusters are empty sets or that some elements $a\in\mathcal{A}$ appear on the border of two or more clusters $\pi(z_1),\ldots,\pi(z_k)$ determined by assignment points $z_1,\ldots,z_k$