Performance Evaluation of the Silhouette Index

Artur Starczewski[1(✉)] and Adam Krzyżak[2]

[1] Institute of Computational Intelligence, Częstochowa University of Technology,
Al. Armii Krajowej 36, 42-200 Częstochowa, Poland
artur.starczewski@iisi.pcz.pl
[2] Department of Computer Science and Software Engineering,
Concordia University, Montreal, Canada
and Department of Electrical Engineering,
Westpomeranian University of Technology, 70-313 Szczecin, Poland
krzyzak@cs.concordia.ca

Abstract. This article provides the performance evaluation of the *Silhouette* index, which is based on the so called *silhouette width*. However, the index can be calculated in two ways, and so, the first approach uses the mean of the mean *silhouettes* through all the clusters. On the other hand, the second one is realized by averaging the *silhouettes* over the whole data set. These various approaches of the index have significant influence on indicating the proper number of clusters in a data set. To study the performance of the index, as the underlying clustering algorithms, two popular hierarchical methods were applied, that is, the *complete-linkage* and the *single-linkage* algorithm. These methods have been used for artificial and real-life data sets, and the results confirm very good performances of the index and they also allow to choose the best approach.

Keywords: Clustering · Validity index · Unsupervised classification

1 Introduction

The data clustering is important technique used to split the data elements into the homogeneous subsets (called clusters), inside which elements are more similar to each other, while they are more different in other groups. The clustering algorithms can be classified into some categories, for example; partitional, hierarchical or density-based clustering. Note that the partitional algorithms form, the so called, one-level partitioning of the data, whereas the hierarchical algorithms create multi-level ones. There are a lot of clustering algorithms described in the literature, for example, *k-means* is the popular and well-known partitional algorithm, which has many variations [5,10,15]. On the other hand, among hierarchical methods one can mention such as: *single-linkage*, *complete-linkage* or *average-linkage* [14,18,23]. Moreover, for the density-based methods, clusters are

A. Krzyżak carried out this research during his sabbatical leave from Concordia University.

L. Rutkowski et al. (Eds.): ICAISC 2015, Part II, LNAI 9120, pp. 49–58, 2015.
DOI: 10.1007/978-3-319-19369-4_5

defined as dense regions separated by low density ones and they are capable of finding arbitrary shaped clusters. A very important question that one should consider is how many clusters there are in a given data set. The right answer to this question has a significant influence on the optimal partitioning of data. For most algorithms, this parameter defining a number of clusters must be given *a priori*. Then, the cluster validity indices are used to indicate the perfect partitions of data. For this purpose, lots of various validity indices with clustering algorithms are employed. For example, the popular indices for the crisp clustering include *Dunn* [9], *Davies-Bouldin (DB)* [8] or *Silhouette* index [28]. Note that there are other indices used for the fuzzy clustering, e.g., *partition coefficient (PC)* [2], *Xie and Beni (XB)* [30] or *Fukuyama and Sugeno (FS)* [11] index.

In this paper the performance evaluation of two versions of the *Silhouette* index is described. The first is called $SILv1$-index, and the second $SILv2$-index. The paper is organized as follows. Section 2 describes these two versions of the *Silhouette* index. Sections 3 presents experimental results using artificial and real-life data sets. Finally, there are conclusions in Section 4.

2 Description of the *Silhouette* index

The *Silhouette* index is described in [28]. Let us denote a partition of a data set X by $C = \{C_1, C, ..., C_K\}$, where C_k indicates k^{th} cluster in the data set, and $k = 1, .., K$. This index is based on the so called *silhouette width*, which can be expressed as follows:

$$S(\mathbf{x}) = \frac{b(\mathbf{x}) - a(\mathbf{x})}{max\left(a(\mathbf{x}), b(\mathbf{x})\right)} \tag{1}$$

where $a(\mathbf{x})$ is the within-cluster mean distance defined as the average distance between $\mathbf{x}$ which belongs to C_k and the rest of patterns $\mathbf{x}_k$ belonging to the same cluster, that is

$$a(\mathbf{x}) = \frac{1}{n_k - 1} \sum_{\mathbf{x}_k \in C_k} d\left(\mathbf{x}, \mathbf{x}_k\right) \tag{2}$$

and n_k is a number of patterns in C_k. On the other hand, $b(\mathbf{x})$ is the smallest of the mean distances of $\mathbf{x}$ to the patterns $\mathbf{x}_\iota$ belonging to the other clusters C_ι, where $\iota = 1, ..., K$ and $\iota \neq k$. Thus, the smallest distance can be defined as:

$$b(\mathbf{x}) = \min_{\substack{\iota=1 \\ \iota \neq k}}^{K} \delta(\mathbf{x}, \mathbf{x}_\iota) \tag{3}$$

where the mean distance for C_ι can be written as:

$$\delta(\mathbf{x}, \mathbf{x}_\iota) = \frac{1}{n_\iota} \sum_{\mathbf{x}_\iota \in C_\iota} d\left(\mathbf{x}, \mathbf{x}_\iota\right) \tag{4}$$

and n_ι is a number of patterns in C_ι. Consequently, the *silhouette width* for the given cluster C_k can be expressed as:

$$S(C_k) = \frac{1}{n_k} \sum_{\mathbf{x} \in C_k} S(\mathbf{x}) \tag{5}$$