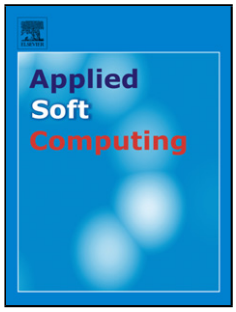

Title: A novel internal validity index based on the cluster centre and the nearest neighbour cluster

Authors: Shibing Zhou, Zhenyuan Xu

PII: S1568-4946(18)30366-1
DOI: https://doi.org/10.1016/j.asoc.2018.06.033
Reference: ASOC 4948

To appear in: *Applied Soft Computing*

Received date: 17-12-2017
Revised date: 20-6-2018
Accepted date: 23-6-2018

Please cite this article as: Zhou S, Xu Z, A novel internal validity index based on the cluster centre and the nearest neighbour cluster, *Applied Soft Computing Journal* (2018), https://doi.org/10.1016/j.asoc.2018.06.033

A novel internal validity index based on the cluster centre and the nearest neighbour cluster

Shibing Zhou [a, *], Zhenyuan Xu [b]

[a] *Department of Computer Science and Technology, Jiangnan University, Wuxi 214122, China*
[b] *School of Science, Jiangnan University, Wuxi 214122, China*

[*] Corresponding author at: Department of Computer Science and Technology, Jiangnan University, Wuxi 214122, China.
E-mail addresses: worldguard@sina.com (S. Zhou), xuzhenyuan1946@hotmail.com (Z. Xu).

Highlights

- We propose a new index to analyse the validity of clustering results.
- We propose a new method to determine the optimal number of clusters.
- The proposed index is an improvement of Silhouette.

Abstract

It is crucial to evaluate the clustering quality in cluster analysis. In this paper, a new internal cluster validity index based on the cluster centre and the nearest neighbour cluster is designed according to the geometric distribution of objects. Moreover, a method for determining the optimal number of clusters is proposed. The new methodology can evaluate the clustering results produced by a certain clustering algorithm and determine the optimal number of clusters for a given dataset. Theoretical research and experimental results indicate the validity and good performance of the proposed index and method.

Keywords

Affinity propagation; cluster validity index; hierarchical clustering; *K*-means; number of clusters.

1. Introduction

Clustering is a primary method of data analysis. Usually, based on similarity or dissimilarity measures, clustering divides objects into many clusters so that the objects in the same cluster are similar and the objects in different clusters are dissimilar. Clustering is widely used in many fields, such as pattern recognition, machine learning, data mining and bioinformatics. Researchers have developed many clustering algorithms. These clustering algorithms can be roughly classified into partitional clustering and hierarchical clustering. Partitional clustering algorithms include *K*-means [1], *K*-medians [2], *K*-medoids [3], fuzzy *C*-means (FCM) [4] and affinity propagation (AP) [5]. Hierarchical algorithms include agglomerative hierarchical clustering (AHC) and divisive hierarchical clustering (DHC) [6–8]. Different clustering algorithms or different configurations of the same algorithm produce different partitions. Moreover, some clustering algorithms must initially be supplied with the number of clusters, known as the *k* parameter. Since the number of clusters is rarely previously known, the usual approach is to run the clustering algorithm several times with a different *k* value for each run. The process of evaluating the partitions produced by clustering algorithms is known as cluster validation, an important subject in cluster analysis. The common approach for this evaluation is to use validity indices. Validity indices are typically classified by researchers into two groups, i.e., internal or external. The main difference between internal and external indices is whether external information is used. Since external validity indices need to know the true cluster labels in advance, they are mainly