Applied Soft Computing 64 (2018) 94–108

Contents lists available at ScienceDirect

Applied Soft Computing

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

Integrating cluster validity indices based on data envelopment analysis

Boseop Kim [a], Hakyeon Lee [b], Pilsung Kang [a,*]

[a] School of Industrial Management Engineering, Korea University, Seoul, South Korea
[b] Department of Industrial and Systems Engineering, Seoul National University of Science and Technology, Seoul, South Korea

ARTICLE INFO

Article history:
Received 6 July 2017
Received in revised form
20 November 2017
Accepted 23 November 2017
Available online 12 December 2017

Keywords:
Clustering validity
Data envelopment analysis
Linear programming
Internal measure

ABSTRACT

Because clustering is an unsupervised learning task, a number of different validity indices have been proposed to measure the quality of the clustering results. However, there is no single best validity measure for all types of clustering tasks because individual clustering validity indices have both advantages and shortcomings. Because each validity index has demonstrated its effectiveness in particular cases, it is reasonable to expect that a more generalized clustering validity index can be developed, if individually effective cluster validity indices are appropriately integrated. In this paper, we propose a new cluster validity index, named Charnes, Cooper & Rhodes – cluster validity (CCR-CV), by integrating eight internal clustering efficiency measures based on data envelopment analysis (DEA). The proposed CCR-CV can be used for purposes that are more general because it extends the coverage of a single validity index by adaptively adjusting the combining weights of different validity indices for different datasets. Based on the experimental results on 12 artificial and 30 real datasets, the proposed clustering validity index demonstrates superior ability to determine the optimal and plausible cluster structures compared to benchmark individual validity indices.

1. Introduction

Clustering is not only one of the most actively studied multivariate data analysis algorithms in the pattern recognition and machine learning fields, but it is also one of the most widely applied algorithms to solve real world problems such as customer segmentation in marketing and new product or service development in retail business [1–3]. The purpose of clustering is to determine a number of groups (clusters) and associated cluster memberships for all records such that records in the same clusters are homogeneous (similar to each other) whereas the records in different clusters are heterogeneous (different from each other). Clustering is regarded as unsupervised learning because there are no explicit answers for the following two questions: (1) what is the optimal number of clusters for a given dataset? and (2) what are the best cluster membership assignments for all records in the dataset?

Because there are no explicit answers for the above questions, it becomes difficult to evaluate the quality of clustering results, which has led to the development of a significant number of different clustering validity indices [4–10]. Although all validity indices agree that an effective clustering result must satisfy the two qualitative principles, i.e., homogeneity within clusters and heterogeneity between clusters, they employ different formulas to quantify these principles. It is accepted that none of the currently exiting cluster validity measures can guarantee the best results for all clustering tasks [11–13]. Hence, in practice, many clustering algorithms are employed to determine the different number of clusters and associated cluster memberships. Then, these clustering results are evaluated by multiple validity measures to allow data analysts or domain experts to determine the most practically plausible clustering result based on their domain knowledge [7].

Clustering validity indices can be grouped into two major categories: external and internal [11]. External indices evaluate the clustering results by comparing the cluster memberships assigned by a clustering algorithm with the previously known knowledge such as externally supplied class labels [14,15]; internal indices evaluate the goodness of the cluster structure by focusing on the intrinsic information of the data itself [12]. Because external indices allow a more objective comparison between clustering algorithms with different parameters, e.g., the number of clusters, they have been adopted to validate any newly proposed clustering algorithm by comparing it with the existing algorithms in the academic

* Corresponding author at: 801A Innovation Hall, Korea University, 145 Anam ro, Seongbuk Gu, Seoul 02841, South Korea.
E-mail addresses: svie89@korea.ac.kr (B. Kim), hylee@snut.ac.kr (H. Lee), pilsung_kang@korea.ac.kr (P. Kang).

https://doi.org/10.1016/j.asoc.2017.11.052
1568-4946/

Table 1
Examples of internal clustering validity indices

Index	Abb.	Definition	Optimal Value
Root-mean-square std dev	*RMSSTD*	$\sum_i \sum_{x \in C_i} \|x - c_i\|^2 / \lceil P \sum_i (n_i - 1)^{\frac{1}{2}} \rceil\}$	Elbow
R-squared	*RS*	$(\sum_{x \in D} \|x - c\|^2 - \sum_i \sum_{x \in C_i} \|x - c_i\|^2) / \sum_{x \in D} \|x - c\|^2$	Elbow
Modified Hubert τ statistic	τ	$\frac{2}{n(n-1)} \sum_{x \in D} \sum_{y \in D} d(x, y) d_{x \in C_i, y \in C_j}(c_i, c_j)$	Elbow
Calinski-Harabasz index	*CH*	$\frac{\sum_i n_i d^2(c_i, c) / (NC - 1)}{\sum_i \sum_{x \in C_i} d^2(x, c_i) / (n - NC)}$	Max
I index	*I*	$(\frac{1}{NC} \cdot \frac{\sum_{x \in D} d(x,c)}{\sum_i \sum_{x \in C_i} d(x,c_i)} \cdot \max_{i,j} d(c_i, c_j))^2$	Max
Dunn's indices	*D*	$\min_i \left\{ \min_j \left(\frac{\max_{x \in C_i, y \in C_j} d(x, y)}{\max_k \left\{ \max_{x,y \in C_k} d(x, y) \right\}} \right) \right\}$	Max
Silhouette index	*S*	$\frac{1}{NC} \sum_i \{ \frac{1}{n_i} \sum_{x \in C_i} \frac{b(x) - a(x)}{max[b(x), a(x)]} \}$	Max
Davies-Bouldin index	*DB*	$\frac{1}{NC} \sum_i \max_{j, j \neq i} \{ [\frac{1}{n_i} \sum_{x \in C_i} d(x, c_i) + \frac{1}{n_j} \sum_{x \in C_j} d(x, c_j)] / d(c_i, c_j) \}$	Min
Xie-Beni index	*XB*	$[\sum_i \sum_{x \in C_i} d^2(x, c_i)] / [n \cdot \min_{i,j \neq i} d^2(c_i, c_j)]$	Min
SD validity index	*SD*	$\mathrm{Dis}(NC_{max}) Scat(NC) + Dis(NC)\ \mathrm{Scat}(NC) = \frac{1}{NC} \sum_i \|\sigma(C_i)\| / \|\sigma(D)\|\ \mathrm{Dis}(NC) = \frac{\max_{i,j} d(c_i, c_j)}{\min_{i,j} d(c_i, c_j)} \sum_i (\sum_j d(c_i, c_j))^{-1}$	Min
S_Dbw validity index	S_Dbw	$\mathrm{Scat}(NC) + Dens\_bw(NC)\ \mathrm{Dens}_{\mathrm{bw}(NC)} = \frac{NC}{NC(NC-1)} \sum_i [\sum_{j, j \neq i} \frac{\sum_{x \in C_i \cup C_j} f(x, u_{i,j})}{max\{\sum_{x \in C_i} f(x, c_i), \sum_{x \in C_j} f(x, c_j)\}}]$	Min

Reprinted from Ref. [6].

research [16,17]. However, they cannot be applied to real world problems because such external information is not readily available. Hence, internal validation indices are more commonly used in practice. The main advantage of internal validation measures is that they do not require any prior knowledge on the clustering structure of a given dataset [7]. They evaluate the compactness within clusters and separation between clusters based on their own formulas. Different formulas are a result of considering the impact of various factors such as noise, density, sub-clusters, skewed distributions, and monotonicity of index [6].

Because datasets have their own intrinsic characteristics, there is no single unique internal validity measure that is best fitted to all data structures [11,12]. In supervised learning, it is also known that there is no single algorithm that outperforms the other algorithms for all datasets [18]. However, if multiple algorithms are properly combined, the predictive performance of this combination, known as an ensemble, is typically superior to single algorithms [19–21]. Similarly, in unsupervised clustering, it is a reasonable expectation that the effectiveness of clustering validity measures can be improved if they can be collectively used with an appropriate integration technique. To this end, some studies attempted to form an ensemble of multiple clustering validity measures to resolve the limitations of individual validity measures. For example, Jaskowiak et al. [22] constructed an ensemble validity measure based on 28 different measures with nine different selection strategies. However, none of them has a sound theoretical basis for integration, but the integration is done empirically. Kou et al. [23] employed three multiple-criteria decision-making (MCDM) methods for evaluating clustering results for financial risk analysis. Although their idea of using MCDM as a tool for integrating validity measures is interesting, the experiments have some limitations; they only considered financial datasets and the results were inconsistent with the known properties of the adopted methods.

In this paper, we propose an integrated clustering validity measure named Charnes, Cooper & Rhodes – cluster validity (CCR-CV), which combines eight internal validity indices based on data envelopment analysis (DEA). There exist two main difficulties of cluster validity integration. First, some validity indices are designed to be minimized with the optimal cluster structure whereas others are designed to be maximized [6]. Moreover, the combining weights must not be fixed constants; rather, they must vary according to the intrinsic characteristics of the dataset. DEA was originally developed to evaluate the efficiency of a system by measuring the ratio of the weighted sum of the output components to the weighted sum of the input components [24]. The weights are not fixed; rather they are determined by solving an optimization problem, considering not only the features of the system itself but also its competitors. Therefore, we employ four validity indices pursuing maximization as the output component and four validity indices pursuing minimization as the input component to define the efficiency of DEA. To determine the appropriate combining weights of the validity indices for a certain clustering algorithm with its associated parameters, we formulate the optimization problem using all candidate algorithm-parameter pairs. Hence, we expect that the coverage of the proposed clustering validity index can be extended with superior performance compared to the individual indices.

The remainder of this paper is organized as follows. In Section 2, we briefly review the internal validity measures used in this study. In Section 3, we introduce DEA and the formulation of the optimization problem used in DEA. Then, we demonstrate the proposed DEA-based integrated clustering validity measures. In Section 4, experimental settings including dataset description, clus-