Contents lists available at ScienceDirect

Pattern Recognition

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

Robust cluster validity indexes

Kuo-Lung Wu[a,*], Miin-Shen Yang[b], June-Nan Hsieh[b]

[a]Department of Information Management, Kun Shan University, Yung-Kang, Tainan 71023, Taiwan
[b]Department of Applied Mathematics, Chung Yung Christian University, Chung-Li 32023, Taiwan

ARTICLE INFO

Article history:
Received 23 October 2008
Received in revised form 19 January 2009
Accepted 23 February 2009

Keywords:
Cluster validity index
Clustering algorithms
Fuzzy c-means
Partition membership
Mean
Median
Robust
Noise
Outlier

ABSTRACT

Cluster validity indexes can be used to evaluate the fitness of data partitions produced by a clustering algorithm. Validity indexes are usually independent of clustering algorithms. However, the values of validity indexes may be heavily influenced by noise and outliers. These noise and outliers may not influence the results from clustering algorithms, but they may affect the values of validity indexes. In the literature, there is little discussion about the robustness of cluster validity indexes. In this paper, we analyze the robustness of a validity index using the φ function of M-estimate and then propose several robust-type validity indexes. Firstly, we discuss the validity measure on a single data point and focus on those validity indexes that can be categorized as the mean type of validity indexes. We then propose median-type validity indexes that are robust to noise and outliers. Comparative examples with numerical and real data sets show that the proposed median-type validity indexes work better than the mean-type validity indexes.

1. Introduction

After clustering results from a clustering algorithm are obtained, it is important to validate if it accurately presents the actual structure of data. As Pal and Bezdek [1] pointed out, once clusters are found, it is necessary to validate them. Most clustering algorithms can generate partition memberships and prototypes for a given data set. However, they must presume the cluster number c in which this number is generally unknown. In this case, validity indexes can be used to find an optimal c where they are supposed to be independent of clustering algorithms. Thus, a cluster validity index should be suitable for most clustering algorithms. The problem of finding an optimal cluster number c is usually called cluster validity. Cluster validity indexes have been used to evaluate the fitness of partition memberships produced by a clustering algorithm. Many cluster validity indexes for clustering algorithms had been proposed in the literature. The objective of those validity indexes is to find an optimal cluster number c that can validate the best description of the data structure. In general, it is assumed that these optimal c clusters should be compact within the cluster and well separated from other clusters.

In cluster analysis, fuzzy c-means (FCM) algorithm is best-known. The first proposed cluster validity index associated with FCM was the partition coefficient (PC) [2–4]. Subsequently, partition entropy (PE) [3], proportion exponent [5], normalization of PC and PE [6–9], a performance measure [10], minimum and mean hard tendencies [11] were proposed. However, most of these indexes only considered partition memberships, not considered the geometrical structure of data. The separation coefficient proposed by Gunderson [12] seems to be the first validity index that explicitly takes into account the data geometrical properties. Indexes in this class include another parts of cluster validity indexes for clustering. These include FS (Fukuyama and Sugeno [13]), the fuzzy hypervolume (FHV) and partition density (PD) (Gath and Geva [14]), XB (Xie and Beni [15]), compose within and between scattering (CWB) (Rezaee et al. [16]), ratio fuzzy separation/fuzzy compactness (SC) (Zahid et al. [17]), a B_{crit} index based on the compactness of clusters and spatial separation (Boudraa [18]), a so-called SCF index based on the fuzzy compactness and separation with union and intersection of the clusters (Fadili et al. [19], overlap and separation (OS) Kim et al. [20]), ratio between compactness and separation (RCS) (Tsekouras and Sarimveis [21]), PBM-index (Pakhira et al. [22]), and PC and exponential separation (PCAES) (Wu and Yang [23]).

In this paper, we will discuss the robustness of validity indexes. We know that both PC and PE have monotonic tendencies with cluster number c. Robust versions of PC and PE with respect to the cluster number c were proposed in [7–9,26]. However, there has been less discussion about the influences of noise and outliers on cluster validity indexes. In general, a validity index should be independent of clustering algorithms. However, the noise and outliers may

[*] Corresponding author.
E-mail address: klwu@mail.ksu.edu.tw (K.-L. Wu).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.02.010

change the values of a validity index, even though the results from clustering algorithms are not influenced by these noise and outliers. To consider this robustness problem, we propose robust cluster validity indexes based on the median in this paper. The rest of this paper is organized as follows. Section 2 discusses the validity measure on a single data point. We then show that most validity indexes for clustering can be reformulated as mean-type validity indexes. The robustness of a validity index is analyzed by the φ function of M-estimate. Section 3 presents the proposed median-type validity indexes which are M-estimators with a constant φ function. We demonstrate the robustness to noise and outliers for these median types. Numerical examples and several real data sets are presented in Section 4. Finally, conclusions are stated in Section 5.

2. The M-estimator based cluster validity indexes using the sample mean

Clustering is a technique used to partition a data set $X = \{x_1, \ldots, x_n\} \subset R^s$ into c subsets that can well represent the structure of the data set X. In general, the cluster number c is in $\{2, 3, \ldots, n-1\}$. The data partitions of c clusters can be described by a $c \times n$ partition matrix $U = [\mu_{ij}]_{c\times n}$ where each element μ_{ij} of U represents the membership of x_j belonging to the ith cluster. In general, there are three kinds of partition matrices that are used for clustering: (1) the hard partitions U_H with $\mu_{ij} \in \{0,1\}$ and $\sum_{i=1}^{c} \mu_{ij} = 1$ for each i; (2) the fuzzy partitions U_F with $\mu_{ij} \in [0,1]$ and $\sum_{i=1}^{c} \mu_{ij} = 1$ for each i; (3) the possibilistic partitions U_P with $\mu_{ij} \in [0,1]$ and $\sum_{i=1}^{c} \mu_{ij} > 0$ for each i. For a simple case with $c = 2$, Fig. 1(a) shows the elements of U (i.e. μ_{1j} and μ_{2j}). The possible values of μ_{ij} in U_H, U_F and U_P will occur in the solid circle points, on the line, and inside the rectangle, respectively. For the case of $c = 3$, Fig. 2 also shows the constraints of these different kind of partitions.

2.1. Validity measure on each single data point

The quality of a data partition matrix $U = [\mu_{ij}]_{c\times n}$ indicates how closely the data points are associated with the cluster centers. If the partition membership of one data point belonging to a particular cluster is significantly larger than the memberships to the other clusters, then this data point could be well identified with the particular cluster. For a simple case of $c = 2$, a well clustered data point should have the membership belonging to the shadow area as shown in Fig. 1(b). These areas contain a large μ_{1j} but a small μ_{2j}, or oppositely, a small μ_{1j} but a large μ_{2j}. Since each data point has c memberships, it is desirable to summarize the information contained in the memberships by a single number $v_j = v(x_j)$ which can indicate how well the data point x_j is clustered by a clustering algorithm. A simple definition of v_j can be

$$v_j = \sum_{i=1}^{c} \mu_{ij}^2 \tag{1}$$

The v_j can be seen as a validity measure on a single data point. If we consider a measure with the mean of $v_1, \ldots, v_n$, then the measure will become the PC validity index. Fig. 2(a) shows the values using Eq. (1) for the case of $c = 2$. A large value of v_j will occur in the data points having large μ_{1j} or μ_{2j}. The smallest value $1/c = 0.5$ will occur in the case of $(\mu_{1j}, \mu_{2j}) = (0.5, 0.5)$. Note that Eq. (1) is suitable for use only in fuzzy c-partitions, as shown in Fig. 2(a). Since both μ_{1j} and μ_{2j} can be close to 1 in a possibilistic environment, a completely possibilistic

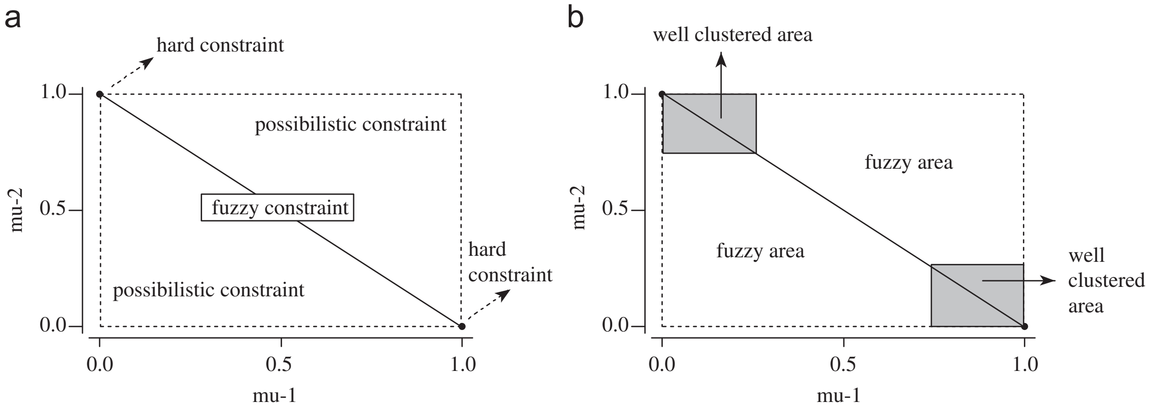

Fig. 1. (a) Possible values of μ_{ij} with $c = 2$. (b) Areas of well clustered data points with $c = 2$.

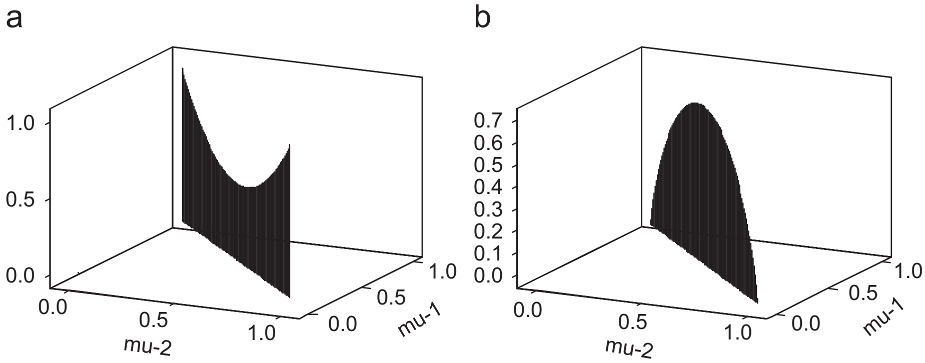

Fig. 2. The validity measure v_j (a) $v_j = PC_j$. (b) $v_j = PE_j$.