Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 26 (2005) 1275–1291

Pattern Recognition Letters

www.elsevier.com/locate/patrec

A cluster validity index for fuzzy clustering

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

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

Received 26 September 2004
Available online 19 December 2004

Abstract

Cluster validity indexes have been used to evaluate the fitness of partitions produced by clustering algorithms. This paper presents a new validity index for fuzzy clustering called a partition coefficient and exponential separation (PCAES) index. It uses the factors from a normalized partition coefficient and an exponential separation measure for each cluster and then pools these two factors to create the PCAES validity index. Considerations involving the compactness and separation measures for each cluster provide different cluster validity merits. In this paper, we also discuss the problem that the validity indexes face in a noisy environment. The efficiency of the proposed PCAES index is compared with several popular validity indexes. More information about these indexes is acquired in series of numerical comparisons and also three real data sets of Iris, Glass and Vowel. The results of comparative study show that the proposed PCAES index has high ability in producing a good cluster number estimate and in addition, it provides a new point of view for cluster validity in a noisy environment.

Keywords: Fuzzy clustering; Cluster validity; Fuzzy *c*-means; Fuzzy *c*-partitions; Partition coefficient and exponential separation

1. Introduction

Cluster analysis is a method for clustering a data set into groups of similar characteristics. It is an approach to unsupervised learning and also one of the major techniques in pattern recognition. The conventional (hard) clustering methods restrict each point of the data set to exactly one cluster. Since Zadeh (1965) proposed fuzzy sets that produced the idea of allowing to have membership functions to all clusters, fuzzy clustering has been widely studied and applied in a variety of substantive areas (Bezdek, 1981; Höppner et al., 1999; Yang, 1993; Baraldi and Blonda, 1999a,b).

In the fuzzy clustering literature, the fuzzy *c*-means (FCM) clustering algorithm and its variation are the most well-known and used methods (Bezdek, 1981; Höppner et al., 1999; Yang,

[*] Corresponding author. Tel.: +886 3 265 3100; fax: +886 3 265 3199.
E-mail address: msyang@math.cycu.edu.tw (M.-S. Yang).

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.11.022

1993). However, it is necessary to preassume the number c of clusters for these fuzzy clustering algorithms. In general, the number c is unknown. The problem for finding an optimal c is usually called cluster validity (Bezdek, 1974a,b). The issue of cluster validity methods is a broad one. In this paper we focus on the cluster validity of fuzzy partitions. The objective is to find optimal c clusters that can validate the best description of the data structure. Each of these optimal c clusters should be compact and separated from other clusters.

The first proposed fuzzy cluster validity functions associated with FCM are the partition coefficient (PC) and partition entropy (PE) (Bezdek, 1974a,b). Dave (1996) proposed a modified partition coefficient (MPC) index by changing its range to the interval $[0,1]$. The MPC index behaves like the fuzzy set decomposition of Backer and Jain (1981). Above indexes have the disadvantage of lack for the connection to the geometrical structure of data. The separation coefficient proposed by Gunderson (1978) was the first validity index that explicitly takes into account the data geometrical properties. Indexes in this class include the XB index proposed by Xie and Beni (1991), FS index proposed by Fukuyama and Sugeno (1989), SC index proposed by Zahid et al. (1999), the fuzzy hypervolume (FHV) and partition density (PD) indexes proposed by Gath and Geva (1989).

Once the partition is obtained by a clustering method, the validity function can help us to validate whether it accurately presents the data structure or not. We know that there are broadly extended types of FCM in the literature, such as Gustafson and Kessel (1979), Krishnapuram and Kim (1999), Bezdek et al. (1981a,b), Dave (1992), Gath and Geva (1989), Wu and Yang (2002), Yu and Yang (accepted for publication), etc. However, validity indexes are considered to be independent of clustering algorithms. Most clustering algorithms, such as mentioned above, can generate fuzzy partitions and cluster centers for a given data set. Thus, we only consider the standard FCM clustering algorithm for all validity indexes.

In this paper, we present a new validity index for fuzzy clustering called a partition coefficient and exponential separation (PCAES) index. It uses the factors from a normalized partition coefficient and an exponential separation measure for each cluster and then pools these two factors to create the PCAES validity index. We also discuss the case that validity indexes face in a noisy environment. Most validity indexes measure the degree of compactness and separation for the data structure in all of c clusters and then finds an optimal c that each one of these optimal c clusters is compact and separated from other clusters. If the data set contains some noisy points that may be far away from other clustered points, it can be visualized that validity indexes will take the noisy point into a compact and separated cluster. However, the proposed new validity measure can give another point of view in this noisy environment. For each identified cluster, we can measure the potential for the cluster to be a well identified cluster. Under this criterion, a noisy point will not have enough potential to be a cluster. This consideration has a different merit from most cluster validity indexes. The merit of the proposed validity index is not only to find an optimal cluster number estimate and also to provide more information about the data structure in a noisy environment.

The remainder of this paper is organized as follows. In Section 2, we review fuzzy clustering algorithms with several popular validity indexes. A new cluster validity index is then proposed for fuzzy clustering in Section 3. The proposed index is created by considering the compactness and separation measures for each cluster and also the data structure. Section 4 presents some numerical examples with the comparisons. In Section 5, we use three real data sets of Iris, Glass and Vowel to have more comparisons. Many interesting phenomena can be found in these comparison results. Conclusions are drawn in Section 6.

2. Some validity indexes for fuzzy clustering

Since Zadeh (1965) introduced the concept of fuzzy sets, a great deal of research on fuzzy clustering has been conducted. In the literature on fuzzy clustering, the fuzzy c-means (FCM) is the best-known fuzzy clustering method. Let $X = \{x_1, \ldots, x_n\}$ be a data set in an s-dimensional Euclidean space R^s with its ordinary Euclidean norm $\|\cdot\|$ and let c be a positive integer larger than one. A partition of X into c clusters can be presented using mutually disjoint sets $X_1,\ldots,X_c$ such that $X_1 \cup \cdots \cup X_c = X$ or equivalently by the indicator functions $\mu_1,\ldots,\mu_c$ such that $\mu_i(x) = 1$ if $x \in X_i$ and $\mu_i(x) = 0$ if $x \notin X_i$ for all $i = 1,\ldots,c$. The set of indicator functions $\{\mu_1,\ldots,\mu_c\}$ is called a hard c-partition of X into c clusters. Consider an extension to allow $\mu_i(x)$ to be membership functions of fuzzy sets μ_i on X assuming values in the interval $[0,1]$ such that $\sum_{i=1}^{c}\mu_i(x) = 1$ for all x in X. In this case, $\{\mu_1,\ldots,\mu_c\}$ is called a fuzzy c-partition of X. The FCM clustering is an iterative algorithm using the necessary conditions for a minimizer of the objective function J_m with

$$J_m(\mu,a) = \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\|x_j - a_i\|^2, \quad m > 1 \tag{1}$$

where $\{\mu_1,\ldots,\mu_c\}$ with $\mu_{ij} = \mu_i(x_j)$ is a fuzzy c-partition and $\{a_1,\ldots,a_c\}$ is the set of c cluster centers. The weighting exponent m is called the fuzzifier which can have influence on the clustering performance of FCM (Cannon et al., 1986; Pal and Bezdek, 1995; Yu et al., 2004).

The validity function can help us to validate whether it accurately presents the data structure or not, when the partition is obtained by a clustering method. In the fuzzy clustering literature, the FCM algorithm is the best known which has many generalization types. Say for examples, Gustafson and Kessel (1979) and Krishnapuram and Kim (1999) generalized FCM with a fuzzy covariance matrix for improving the ability to detect different cluster shapes (especially for hyperellipsoidal shapes) in the data set. Bezdek et al. (1981a,b) presented a variety of FCM to detect non-hyperellipsoidal shaped substructure with fuzzy c-varieties. Dave (1992) proposed a fuzzy c-shells method for detecting shell cluster shapes, especially for circular and elliptical curve boundaries. Gath and Geva (1989) and Wu and Yang (2002) had used an exponential-type distance to have the clustering results more robust. Recently, Yu and Yang (accepted for publication) had considered a broad type of generalized FCM. However, validity indexes are considered to be independent of clustering algorithms. Most clustering algorithms, such as mentioned above, can generate fuzzy partitions and cluster centers for a given data set. Because we are focusing on the performance of validity indexes, not for clustering algorithms, we only consider the standard FCM clustering algorithm for all validity indexes.

After a fuzzy c-partition $\{\mu_1,\ldots,\mu_c\}$ is provided by a fuzzy clustering algorithm such as FCM, we may ask whether it accurately presents the data structure or not. This is a cluster validity problem. Since most of fuzzy clustering methods need to preassume the number c of clusters, a validity criterion for finding an optimal c which can completely describe the data structure becomes the most studied topic in cluster validity. Several popular validity indexes are reviewed as follows.

(a) The first validity index associated with FCM was the partition coefficient (Bezdek, 1981; Trauwaert, 1988) defined by

$$\mathrm{PC}(c) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{2} \tag{2}$$

where $\frac{1}{c} \leqslant \mathrm{PC}(c) \leqslant 1$. In general, we find an optimal cluster number c^* by solving $\max_{2 \leqslant c \leqslant n-1} \mathrm{PC}(c)$ to produce a best clustering performance for the data set X.

(b) The partition entropy (Bezdek, 1974a,b) was defined by

$$\mathrm{PE}(c) = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}\log_2\mu_{ij} \tag{3}$$

where $0 \leqslant \mathrm{PE}(c) \leqslant \log_2 c$. In general, we find an optimal c^* by solving $\min_{2 \leqslant c \leqslant n-1}\mathrm{PE}(c)$ to produce a best clustering performance for the data set X.

(c) Both PC and PE possess monotonic evolution tendency with c. Modification of the PC index proposed by Dave (1996) can reduce the monotonic tendency and was defined by

$$\mathrm{MPC}(c) = 1 - \frac{c}{c-1}(1 - \mathrm{PC}(c)) \tag{4}$$

where $0 \leqslant \mathrm{MPC}(c) \leqslant 1$. Note that the MPC index is equivalent to the non-fuzziness index (NFI) (Robubens, 1978). In general, an optimal cluster number c^* is found by solving $\max_{2 \leqslant c \leqslant n-1}\mathrm{MPC}(c)$ to produce a best clustering performance for the data set X.

Above indexes use only fuzzy memberships. This may be lack for the connection to the geometrical structure of data. The following indexes simultaneously take into account fuzzy memberships and the data structure. Some existing indexes in this class are briefly reviewed as follows.

(d) A validity function proposed by Fukuyama and Sugeno (1989) was defined by

$$\begin{aligned}\mathrm{FS}(c) &= \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\|x_j - a_i\|^2 \\ &\quad - \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\|a_i - \bar{a}\|^2 \\ &= J_m(\mu, a) + K_m(\mu, a) \qquad (5)\end{aligned}$$

where $\bar{a} = \sum_{i=1}^{c} a_i/c$. $J_m(\mu,a)$ is the FCM objective function which measures the compactness and $K_m(\mu,a)$ measures the separation. In general, an optimal c^* is found by solving $\max_{2\leqslant c\leqslant n-1}\mathrm{FS}(c)$ to produce a best clustering performance for the data set X.

(e) A validity function proposed by Xie and Beni (1991) with $m = 2$ and modified by Pal and Bezdek (1995) was defined by

$$\mathrm{XB}(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\|x_j - a_i\|^2}{n\min_{i,j}\|a_i - a_j\|^2} = \frac{J_m(\mu,a)/n}{\mathrm{Sep}(a)} \qquad (6)$$

$J_m(\mu,a)$ is a compactness measure and $\mathrm{Sep}(a)$ is a separation measure. In general, an optimal c^* is found by solving $\max_{2\leqslant c\leqslant n-1}\mathrm{XB}(c)$ to produce a best clustering performance for the data set X.

(f) A validity function proposed by Zahid et al. (1999) was defined by

$$\mathrm{SC}(c) = \mathrm{SC}_1(c) - \mathrm{SC}_2(c) \qquad (7)$$

where

$$\mathrm{SC}_1(c) = \frac{\sum_{i=1}^{c}\|a_i - \bar{a}_i\|^2/c}{\sum_{i=1}^{c}\left(\sum_{j=1}^{n}(\mu_{ij}^{m})\|x_j - a_i\|^2/\sum_{j=1}^{n}\mu_{ij}\right)} \qquad (8)$$

and

$$\mathrm{SC}_2(c) = \frac{\sum_{i=1}^{c-1}\sum_{l=i+1}^{c}\left(\sum_{j=1}^{n}\left(\min\left(\mu_{ij},\mu_{lj}\right)\right)^2\right)/\sum_{j=1}^{n}\min\left(\mu_{ij},\mu_{lj}\right)}{\sum_{j=1}^{n}\left(\max_{1\leqslant i\leqslant c}\mu_{ij}\right)^2/\sum_{j=1}^{n}\max_{1\leqslant i\leqslant c}\mu_{ij}} \qquad (9)$$

Both SC_1 and SC_2 measure the ratio of separation and compactness. SC_1 considers the geometrical properties of the data structure and membership functions. SC_2 considers only the fuzzy memberships. In general, we find an optimal c^* by solving $\min_{2\leqslant c\leqslant n-1}\mathrm{SC}(c)$ to produce a best clustering performance for the data set X.

(g) The fuzzy hypervolume (FHV) validity function proposed by Gath and Geva (1989) was defined by

$$\mathrm{FHV}(c) = \sum_{i=1}^{c}[\det(F_i)]^{1/2} \qquad (10)$$

where

$$F_i = \frac{\sum_{j=1}^{n}\left(\mu_{ij}\right)^m\left(x_j - a_j\right)\left(x_j - a_i\right)^{\mathrm{T}}}{\sum_{j=1}^{n}\left(\mu_{ij}\right)^m} \qquad (11)$$

The matrix F_i denotes the fuzzy covariance matrix of cluster i. A fuzzy partition can be expected to have a low $\mathrm{FHV}(c)$ value if the partition is tight. An extremum for this index would ideally indicate a good partition. Thus, we find an optimal c^* by solving $\min_{2\leqslant c\leqslant n-1}\mathrm{FHV}(c)$ to produce a best clustering performance for the data set X.

These indexes are either the most cited or newly proposed validity indexes for fuzzy clustering. These indexes have a common objective for finding a good estimate of a cluster number c so that each one of c clusters is compact and separated from other clusters. If the data set contains some noisy points, it can be visualized that validity indexes will take each noisy point into a compact and separated cluster. In next section, we propose a new validity index to have another point of view from these indexes. For each identified cluster, we measure the potential to see whether the identified cluster has ability to be a good cluster or not. Under this criterion, a noisy point will not have enough potential to be a cluster. Our new consideration can give an impressive result in a noisy environment. The efficiency of the proposed index will be compared with these validity indexes in series of numerical and real data sets in Sections 4 and 5.

3. The proposed validity index for fuzzy clustering

We now propose a new validity index for fuzzy clustering. Let $X = \{x_1, \ldots, x_n\}$ be a data set in R^s.

Assume that $\mu = \{\mu_1, \ldots, \mu_c\}$ is a fuzzy c-partition based on a fuzzy clustering (e.g. FCM) algorithm. We consider two factors with a normalized partition coefficient and an exponential separation measure to validate each cluster. We then pool these two terms to create a new validity index, called a partition coefficient and exponential separation (PCAES) index. We first define the PCAES index for cluster i as

$$\mathrm{PCAES}_i = \sum_{j=1}^{n} \mu_{ij}^2/\mu_M - \exp\left(-\min_{k\neq i}\{\|a_i - a_k\|^2\}/\beta_\mathrm{T}\right) \tag{12}$$

where

$$\mu_M = \min_{1\leqslant i\leqslant c}\left\{\sum_{j=1}^{n}\mu_{ij}^2\right\} \quad \text{and} \quad \beta_\mathrm{T} = \frac{\sum_{l=1}^{c}\|a_l - \bar{a}\|^2}{c} \tag{13}$$

We use the term of a normalized partition coefficient (NPC) with

$$\sum_{j=1}^{n}\mu_{ij}^2/\mu_M \tag{14}$$

to measure the compactness for the cluster i relative to the most compact cluster which has the compactness measure μ_M. This term is similar to the compactness measure for cluster i used in the PC index, where the measure is taken as an average, not as a relative value. The compactness value in (14) will belong to the interval $(0,1]$.

The exponential-type separation measure for cluster i with

$$\exp\left(-\min_{k\neq i}\{\|a_i - a_k\|^2\}/\beta_\mathrm{T}\right) \tag{15}$$

takes advantage of exponential function that measures the distance between cluster i and its closest cluster. This measure is similar to an exponential function of the separation measure $\mathrm{Sep}(a)$ in (6) defined by the XB index. Moreover, we consider it relative to β_T of the total average distance measure for all clusters. The total average distance measure of all clusters is similar to the separation measure $K_m(\mu, a)$ in (5) defined by the FS index. We take the exponential function to make the separation measure in the interval $(0,1]$ and also make the compactness (14) and separation (15) to have the same range (or degree) of measurement. Another motivation for taking the exponential function is that an exponential operation is highly useful in dealing with the classical Shannon entropy (Pal and Pal, 1991, 1992) and cluster analysis (Gath and Geva, 1989; Wu and Yang, 2002). Especially, Wu and Yang (2002) had claimed that an exponential-type distance gives robust property based on the influence function analysis.

Since the compactness and separation for each cluster are restricted on

$$0 < \sum_{j=1}^{n}\mu_{ij}^2/\mu_M \leqslant 1 \tag{16}$$

and

$$0 < \exp\left(-\min_{k\neq i}\left\{\|a_i - a_k\|^2\right\}/\beta_\mathrm{T}\right) \leqslant 1 \tag{17}$$

we then have the boundary for PCAES_i with

$$-1 < \mathrm{PCAES}_i < 1 \quad \text{for all } i = 1, \ldots, c \tag{18}$$

We see that the proposed validity criterion PCAES_i could detect each cluster with two measures from a normalized partition coefficient and an exponential separation. The large PCAES_i value means that the cluster i is compact inside and separated from the other $(c-1)$ clusters. The small or negative value of PCAES_i indicates that cluster i is not a well-identified cluster. Finally, the PCAES validity index is then defined as

$$\begin{aligned}\mathrm{PCAES}(c) &= \sum_{i=1}^{c}\mathrm{PCAES}_i \\ &= \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2/\mu_M \\ &\quad - \sum_{i=1}^{c}\exp\left(-\min_{k\neq i}\left\{\|a_i - a_k\|^2\right\}/\beta_\mathrm{T}\right)\end{aligned} \tag{19}$$

Obviously,

$$-c < \mathrm{PCAES}(c) < c \tag{20}$$

In our validity index, we used PCAES_i first to measure the compactness and separation for each cluster and then summed all PCAES_i as $\mathrm{PCAES}(c)$

to measure the compactness and separation for the data structure. Thus, the total compactness of the data set is measured by the term

$$\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{2}/\mu_{M} \tag{21}$$

which is the normalized PC index and the total separation of the data set is measured by the term

$$\sum_{i=1}^{n}\exp\left(-\min_{k\neq i}\left\{\|a_i-a_k\|^2\right\}/\beta_{\mathrm{T}}\right),\quad k=1,\ldots,c \tag{22}$$

The large PCAES(c) value means that each of these c clusters is compact and separated from other clusters. The small PCAES(c) value means that some of these c clusters are not compact or separated from other clusters. The maximum of PCAES(c), with respect to c, could be used to detect the data structure with a compact partition and well-separated clusters. Thus, an optimal c^* can be found by solving $\min_{2\leqslant c\leqslant n}\mathrm{PCAES}(c)$ to produce a best clustering performance for the data set X.

The consideration of normalizing the partition coefficient can give us a small PCAES$_i$, value when cluster i contains only a few points and the index PCAES will be then relatively small. This gives us an alarm whether noisy points are taken into compact and separated clusters or not. This situation often occurs in real applications. Other indexes do not own this property. Comparisons will be made in Sections 4 and 5. Thus, using the proposed validity index not only gives us an optimal cluster number estimate, but also presents more information about the data structure.

Before we present numerical examples and comparisons in the next section, some comments for these validity indexes are made as follows.

(1) Both PC and PE indexes consider only the compactness measurement for each cluster and for the data structure. They are obviously lack of the connection to the geometrical structure of data.
(2) In FS, because $\|a_i-\bar{a}\|^2$ is not a good separation measure for cluster i, the term $\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\|a_i-\bar{a}\|^2$ is not either good separation measure for the data structure. This is why FS may have some unexpected results in our numerical examples.
(3) Although XB uses the FCM objective function to measure the compactness for each cluster and for the data structure, the used separation measure $\min_{i,j}\|a_i-a_j\|^2$ is considered for all clusters, not for each cluster.
(4) The SC index is the most complicated index of all validity indexes mentioned in this paper. We see that SC always defines the compactness and separation measures for the data structure, not for each cluster.

4. Numerical examples

In this section, some numerical examples are presented to compare the proposed PCAES index with the other seven indexes PC, PE, MPC, FS, XB, SC and FHV. We implemented the FCM clustering algorithm on each data set with the cluster number $c=2,\ldots,c_{\max}\approx\sqrt{n}$.

Example 1. In this four-clusters data set, we show the view inside the PCAES operation. The results of the FCM algorithm for different cluster numbers with c = 2, 3, 4 and 5 are shown in Fig. 1. The PCAES$_i$ for each cluster can be found in Fig. 1. In Fig. 1(a), PCAES(c) = PCAES$_1$ + PCAES$_2$ = 1.9. In Fig. 1(b), PCAES(c) = 2.21. In Fig. 1(c), four clusters are well-clustered and have the maximum PCAES value. In Fig. 1(d), the fifth cluster is generated by splitting the second and fourth clusters and hence the PCAES$_4$ and PCAES$_5$ have negative values −0.0518 and −0.3660 respectively. This result indicates that the fourth and fifth clusters are neither compact inside nor separated from other clusters.

The results from other indexes for this data set are shown in Table 1. By optimizing the validity functions, most of indexes indicate that c^* = 4 is an optimal cluster number estimate for this data set except PC, PE, FS and FHV. However, PC, FS and FHV also consider that c^* = 4 may be a good cluster number estimate. Note that a validity function is a tool to discover the data structure and gives a good cluster number estimate.

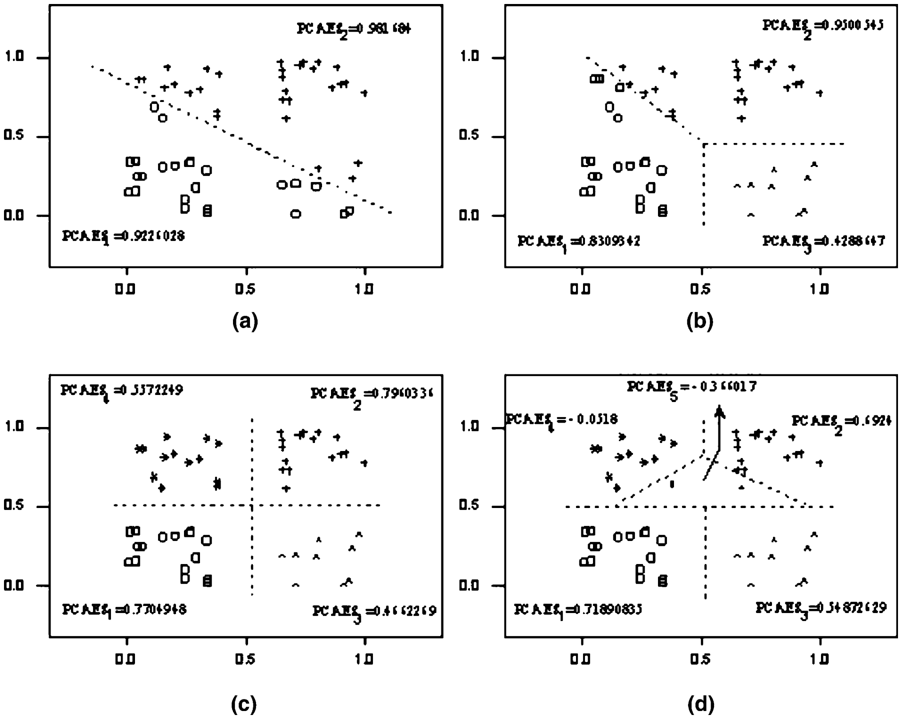


Fig. 1. Clustering results and PCAES values with different cluster number c. (a) $c = 2$, PCAES(c) = 1.9042872, (b) $c = 3$, PCAES(c) = 2.2098534, (c) $c = 4$, PCAES(c) = 2.5899803, (d) $c = 5$, PCAES(c) = 1.5421018.

Table 1
Values of validity indexes in Fig. 1

c	PC	PE	MPC	FS	XB	SC	FHV	PCAES
2	0.752	0.386	0.504	0.206	0.197	0.239	0.1024	1.904
3	0.715	0.522	0.572	−3.759	0.107	0.601	0.0799	2.210
4	0.749	0.529	0.665	−6.560	0.067	1.831	0.0550	2.590
5	0.667	0.702	0.584	−5.995	0.234	1.129	0.0542	1.542
6	0.627	0.800	0.552	−6.038	0.305	0.889	0.0547	−0.166
7	0.637	0.790	0.576	−6.993	0.229	1.413	0.0515	−0.398
8	0.622	0.841	0.568	−6.804	0.178	1.168	0.0513	−0.305

Therefore, graphing the validity function may give us more information. In Fig. 2, the charts are used to describe the validity indexes for this data set. Overall, they give us that $c^* = 4$ is a good cluster number estimate which actually matches the structure of the data set shown in Fig. 1.

Example 2. In Fig. 3(a), we give a data set made up of four clusters with some bridge points. Intuitively, $c = 4$ is suitable for the data set. We then move the four clusters close together from Fig. 3(a) to form Fig. 3(b). The results from the validity indexes for these two data sets are shown in Table 2. The estimated cluster numbers by the validity indexes are shown on the top of figures in Fig. 3. For the data set in Fig. 3(a), all indexes produce the expected results with $c^* = 4$ except the PE index. However, the PE index indicates that $c^* = 4$ may be another good cluster number estimate from Table 2. When the bridge points are closed to be connected as shown in Fig. 3(b), XB and SC indexes take the bridge

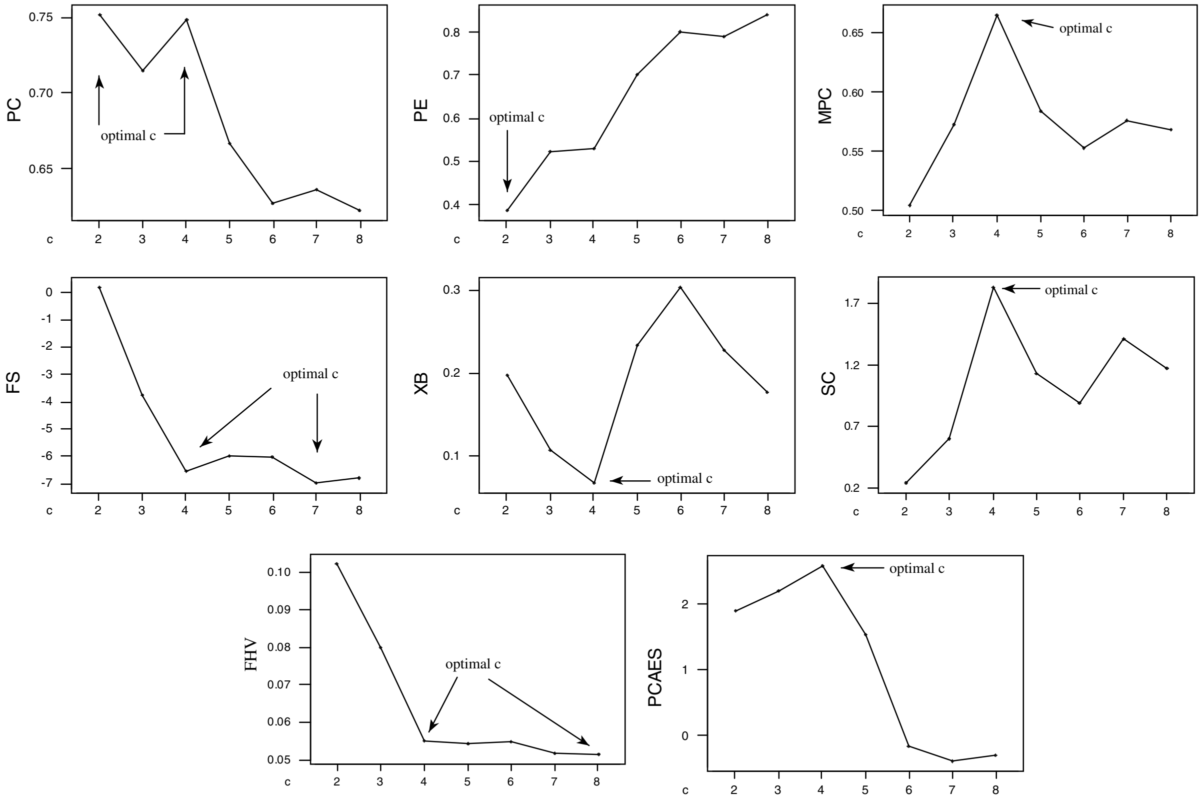


Fig. 2. Graphic chart for Table 1.

index	PC	PE	MPC	FS	XB	SC	FHV	PCAES
optimal c	4	2	4	4	4	4	4	4

index	PC	PE	MPC	FS	XB	SC	FHV	PCAES
optimal c	2	2	4	7	5	5	8	4

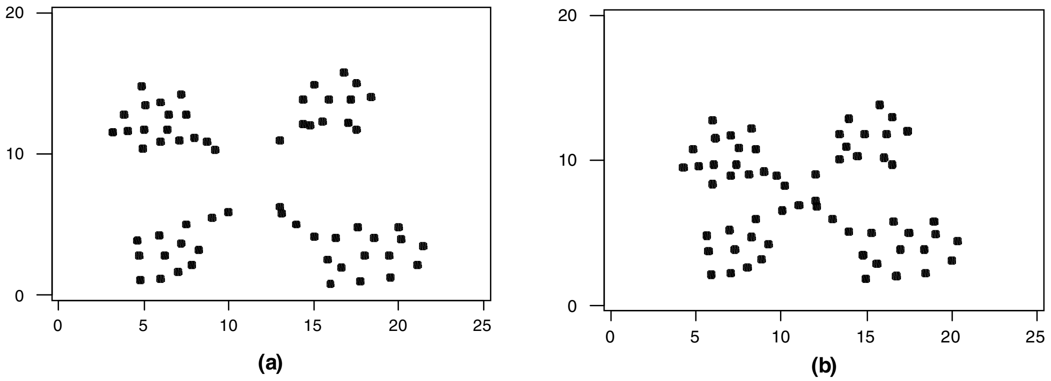


Fig. 3. The data set with a bridge structure.

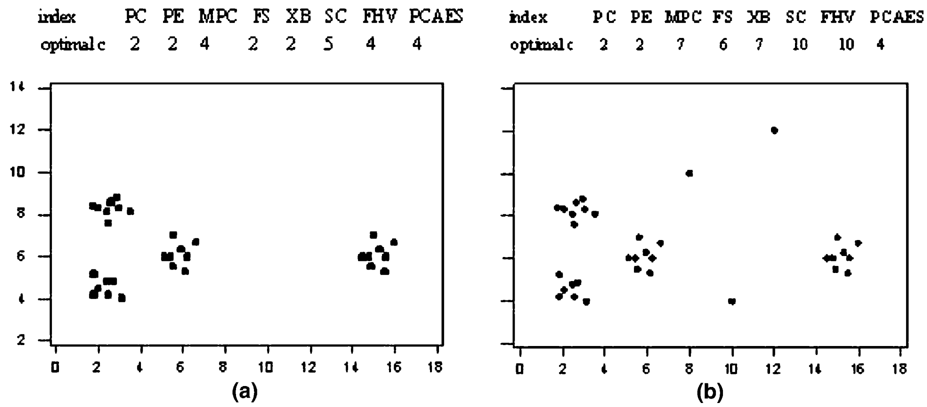
points into a compact and separated cluster with an optimal cluster number estimate $c^* = 5$. PC and PE indicate that $c^* = 2$ is the best cluster number estimate. FS and FHV indexes give the optimal

Table 2
Values of validity indexes in Fig. 3(a) and (b)

c	PC	PE	MPC	FS	XB	SC	FHV	PCAES
Fig. 3(a)								
2	0.733	0.414	0.466	133	0.204	0.221	28.387	1.834
3	0.708	0.533	0.562	−983	0.157	0.457	18.961	2.306
4	0.792	0.450	0.722	−2235	0.055	2.353	10.252	2.803
5	0.743	0.561	0.678	−2211	0.102	2.135	10.593	1.788
6	0.692	0.669	0.631	−2151	0.283	1.823	10.943	0.258
7	0.670	0.730	0.616	−2184	0.244	1.759	10.910	−0.764
8	0.633	0.799	0.581	−1975	0.219	1.734	10.340	−1.041
Fig. 3(b)								
2	0.741	0.412	0.481	−1	0.185	0.354	16.683	1.757
3	0.645	0.629	0.468	−427	0.255	0.207	15.739	2.231
4	0.707	0.599	0.609	−1108	0.103	1.249	9.964	2.598
5	0.671	0.700	0.588	−1121	0.091	1.337	9.719	2.570
6	0.623	0.817	0.547	−1086	0.317	0.819	10.369	1.209
7	0.601	0.877	0.535	−1130	0.246	0.898	9.935	0.074
8	0.580	0.932	0.520	−1088	0.266	0.759	9.510	0.169

cluster number $c^* = 7$ and 8, respectively. We find that both MPC and PCAES indexes give an optimal cluster number $c^* = 4$. According to the index values shown in Table 2, PC, XB and SC give the information that $c^* = 4$ is also a good cluster number estimate for the data set in Fig. 3(b). The cluster generated by the bridge points does not have a large PCAES$_i$ value and it does not have enough potential to be a well-identified cluster. Thus, PCAES indicates that $c^* = 4$ is a good cluster number estimate which is coincident to the MPC index.

Example 3. In this example, we will show a special property of the proposed PCAES index that can avoid taking a single noisy point into a compact and well-separated cluster. Fig. 4(a) presents a data set that one cluster is far away from other three clusters. In Fig. 4(b), three noisy points are added to the data set of Fig. 4(a). The results from the indexes are shown in Table 3. In Fig. 4(a), by optimizing the validity functions, PC, PE, FS and XB produce the same result with the optimal cluster number $c^* = 2$ where the three clusters on the left are synthesized into a cluster and the right



index	PC	PE	MPC	FS	XB	SC	FHV	PCAES
optimal c	2	2	4	2	2	5	4	4

index	PC	PE	MPC	FS	XB	SC	FHV	PCAES
optimal c	2	2	7	6	7	10	10	4

Fig. 4. (a) The data set without noisy points, (b) the same data set with three noisy points.

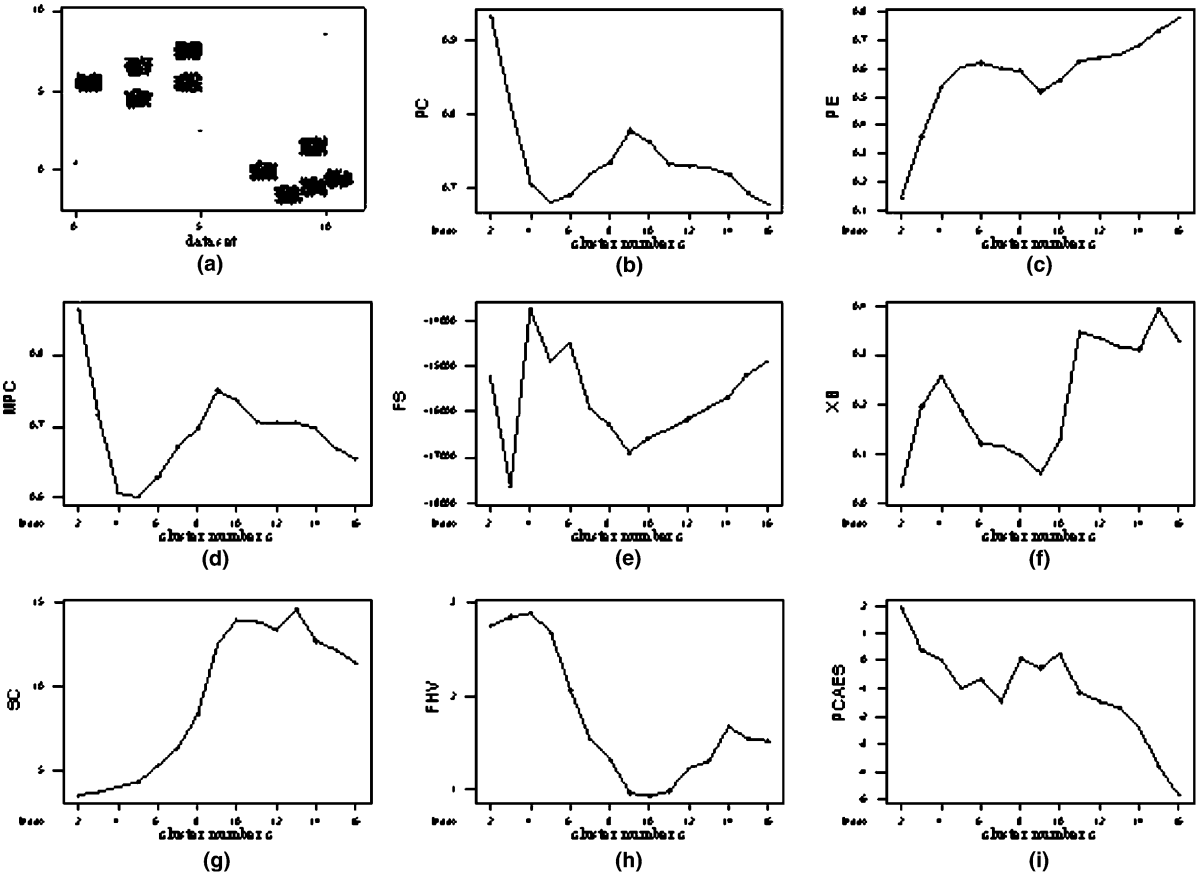
Table 3
Values of validity indexes in Fig. 4(a) and (b)

c	PC	PE	MPC	FS	XB	SC	FHV	PCAES
Fig. 4(a)								
2	0.941	0.128	0.882	−8.57	0.0288	5.50	3.097	1.341
3	0.810	0.330	0.714	−698	0.1625	6.30	1.995	1.400
4	0.915	0.204	0.887	−853	0.0292	17.66	0.859	1.736
5	0.845	0.320	0.806	−806	0.4169	22.47	0.890	−0.301
6	0.780	0.438	0.736	−745	0.3381	19.85	0.935	−1.659
Fig. 4(b)								
2	0.902	0.183	0.804	−722	0.047	2.60	5.814	1.377
3	0.780	0.386	0.670	−634	0.254	2.72	4.657	1.354
4	0.855	0.310	0.807	−806	0.103	4.11	2.970	1.687
5	0.866	0.310	0.833	−899	0.056	5.36	2.743	1.508
6	0.870	0.320	0.844	−938	0.040	6.76	1.785	0.768
7	0.878	0.314	0.858	−936	0.025	13.58	1.060	0.359
8	0.825	0.398	0.800	−932	0.370	15.96	0.995	−1.228
9	0.780	0.477	0.752	−861	0.311	16.30	0.988	−2.738
10	0.762	0.518	0.735	−916	0.287	17.43	0.986	−3.785

hand cluster remains an isolated cluster. They also indicate that $c^* = 4$ may be another good cluster number estimate. The SC index gives the optimal cluster number $c^* = 5$. MPC, FHV and PCAES indexes indicate that $c^* = 4$ is optimal for the data set in Fig. 4(a).

In Fig. 4(b), most validity indexes are affected by these noisy points. We find that MFC and XB with the optimal cluster number $c^* = 7$ take each noisy point into a compact and well-separated cluster. SC and FHV present a monotonic tendency of the cluster number c where they give a largest optimal cluster number estimate $c^* = 10$ in all indexes. PC and PE give a least optimal cluster number $c^* = 2$ in all indexes. Our proposed PCASE index gives an optimal cluster number estimate $c^* = 4$ that presents a property without being affected by noisy points. Overall, $c^* = 7$ from MFC and XB and $c^* = 4$ from PCAES seem to be better matching the structure of the data set shown in Fig. 4(b) where both present two different points of view for cluster validity. MPC and XB take each noisy point into a compact and well-separated cluster. PCAES gives another viewpoint that, although each noisy point is separated from the other clusters, each one of them has not enough potential to be a well-identified cluster. The reason is that the compactness measure for each cluster is relative to the most compact cluster and hence the compactness measure for each noisy point is small. We mention that our objective in this example is not to use validity indexes for detecting noisy points, but for demonstrating the property of considerations involving the compactness and separation measures for each cluster to provide a different merit of cluster validity.

Example 4. This is a 10-clusters data set that each cluster contains 100 random sample uniformly generated on a rectangle grid as shown in Fig. 5(a). Five clusters are separated on the top of the left and three clusters are connected on the bottom of the right. We add three noisy points in this data set. The results of each validity index are shown in Fig. 5(b)–(i). Note that a reasonable cluster number estimate for this data set could be $c^* = 2$ (i.e. five clusters are clustered as a cluster) or $c^* = 8$ (i.e. three connected clusters on the bottom of the right are considered as a cluster) or $c^* = 10$ (i.e. original real cluster number) or $c^* = 13$ (i.e. each noisy point is considered as a cluster). PC, PE and MPC present the same results that 2 and 9 are good cluster number estimates according to the local extremes of the index curve. FS shows



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 5. (a) The data set with three noisy points, (b)–(i) results of eight validity indexes.

that 3, 5 and 9 are good cluster number estimates. XB presents that 2, 9 and 14 are suitable estimates. The SC index considers each noisy point as a cluster with $c^* = 13$. The FHV index shows that 2 and 10 are good cluster number estimates which quite match the structure of data shown in Fig. 5(a). The PCASE index gives the information that 2, 6, 8 and 10 are good cluster number estimates. The results from PCAES present a good coincidence to the data structure shown in Fig. 5(a).

Example 5. We test the efficiency of the validity indexes for a data set of size 10,000. Fig. 6(a) shows the 100×100 pixels image data set with 500 uniformly noisy points. The image has two clusters where one cluster is black and another one is grey and in addition, the image uniformly includes 500 noisy points as shown in Fig. 6(a). The results of validity indexes are shown in Fig. 6(b)–(i). A suitable cluster number in segmenting this image data set is 2. The PC, PE, MPC and XB indexes present the same result with the cluster number estimate $c^* = 3$. The FS and SC indexes show the unexpected results for the image. The FHV presents a flat curve after $c = 3$. The proposed PCASE curve shows that $c^* = 2$ is an optimal cluster number estimate, and also with a local maximum $c = 5$ that has a similar behavior as the XB index. In this case, the PCASE index presents a property without being affected by noisy points for this large data set.

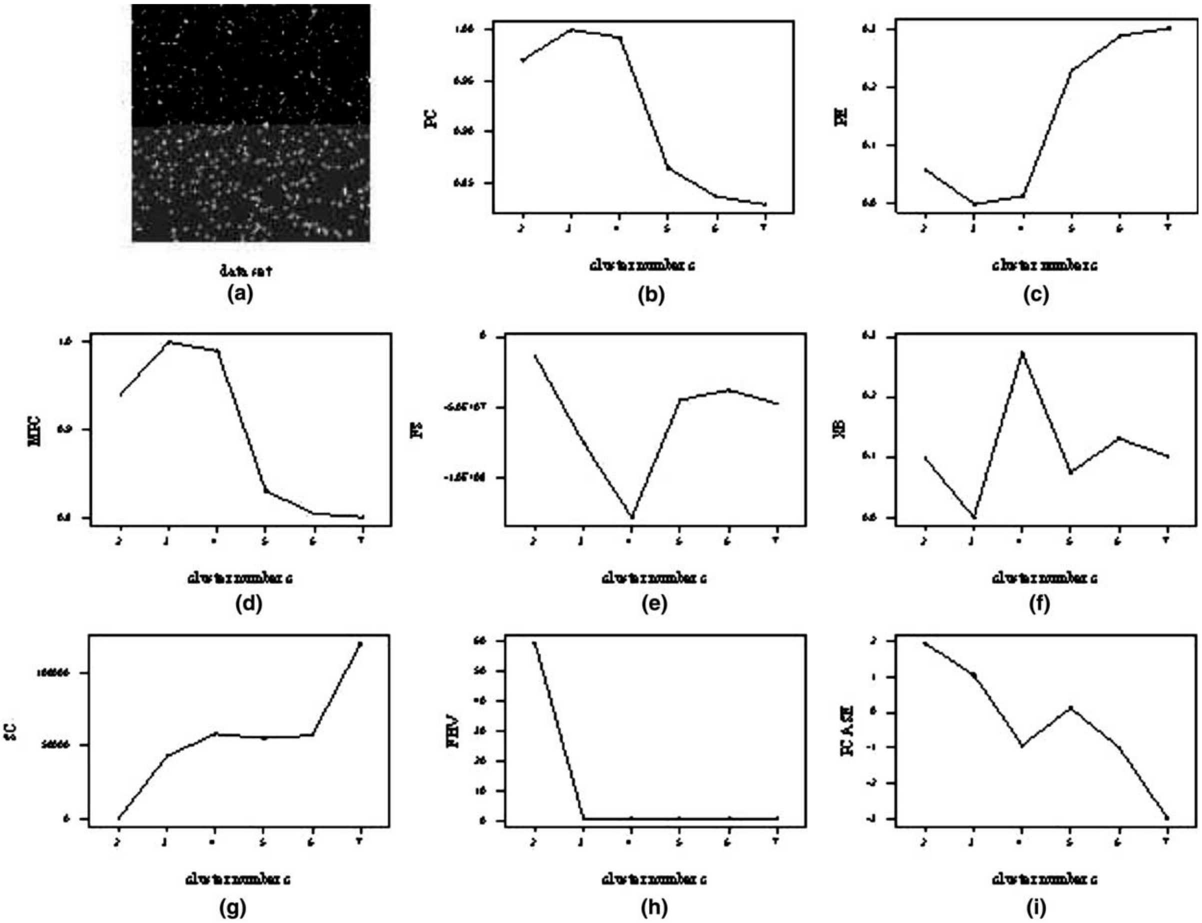


Fig. 6. Results of validity indexes for the two-clusters synthetic image data set with 100 ∗ 100, pixels and 500 noisy points.

Example 6. In this example, we consider a data set drawn from a normal mixture of 10 clusters with the same standard deviation 0.2 where each cluster has 1000 data points. We then uniformly add 1000 noisy points in this data set. The histogram of data is shown in Fig. 7(a). Since the range of the 1000 noisy points are large relative to these 10,000 normal mixture data, the influence of noisy points on the validity indexes will be significant. Most indexes indicate that $c = 2$ and 12 are suitable cluster number estimates. We find that the PCASE index gives the information that $c^* = 10$ is a suitable cluster number estimate. This result from the PCASE index is coincident to the original data structure. To have validity indexes in real applications, we will implement three real data sets of Iris, Glass and Vowel in the next section.

5. Real data sets with Iris, Glass and Vowel

In this section, three real data sets with Iris (Anderson, 1935; Bezdek et al., 1999), Glass and Vowel (Blake and Merz, 1998) are used as the test sets for validity index comparisons. Most clustering problems are solved by minimizing the constructed dispersion measures. In general, each dimension presents one characteristic of data in an s-dimensional data set where each characteristic has different dispersion. Thus, the results from

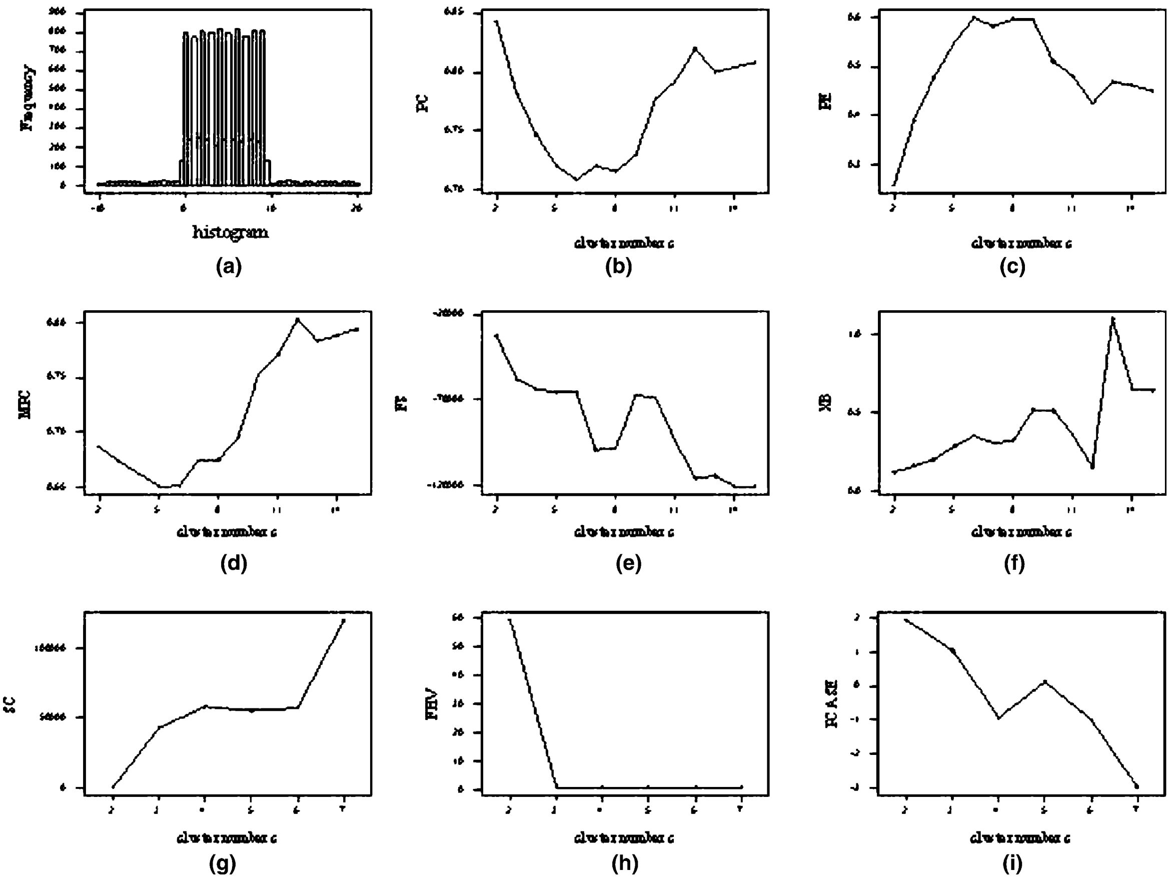


Fig. 7. Results of validity indexes for the 10-clusters normal mixture data set where each cluster contains 1000 data points, and there are 1000 noisy points uniformly added in the data set.

minimizing the total dispersion measure may discard the effects of some characteristics, especially, for those that have small dispersion values. This situation frequently occurs in high dimensional data sets. To use sufficiently all the information of characteristics, we shall normalize the data set. Suppose that we have a data set $X = \{x_1, \ldots, x_n\}$ in an s-dimensional space with each $x_j = (x_{j1}, \ldots, x_{js})$. We normalize the data by replacing x_{jk} with x'_{jk} as

$$x'_{jk} = \frac{x_{jk} - \sum_{l=1}^{n} \frac{x_{lk}}{n}}{\sqrt{\sum_{l=1}^{n}\left(x_{lk} - \sum_{l=1}^{n} \frac{x_{lk}}{n}\right)^2/(n-1)}},$$

$$k = 1, \ldots, s, \quad j = 1, \ldots, n \tag{23}$$

After normalization, each characteristic in the data set will have a common sample mean and dispersion. Thus, we normalize the data sets Iris, Glass and Vowel before we analyze them.

Example 7 (*Iris data set*). The Iris data set (Anderson, 1935; Bezdek et al., 1999) has $n = 150$ points in an $s = 4$ dimensional space. It consists of three clusters. Two clusters have substantial overlapping. Thus, one can argue $c = 2$ or $c = 3$ for the Iris data set. The validity indexes of the Iris data set are shown in Fig. 8. The PC, PE, MPC and XB indexes show that $c^* = 2$ is an optimal cluster number estimate. The result from the FS index is unexpected. If we only consider the values of c between 2 and 6 in Fig. 8(g), the FHV

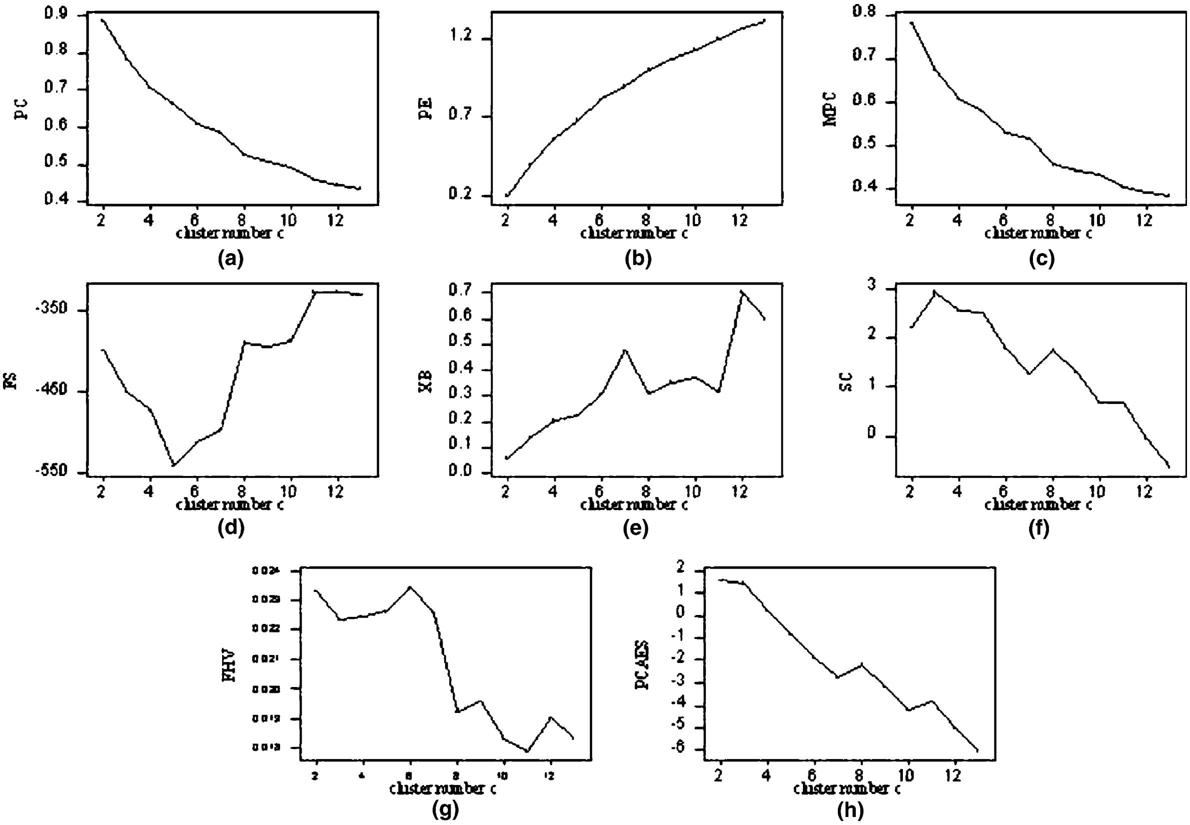


Fig. 8. Results of validity indexes for the Iris data set.

index indicates that the optimal cluster number estimate for the Iris data set is $c^* = 3$. This result is exactly the same as FHV shown in (Gath and Geva, 1989). The FHV index also shows that 8 and 11 may be another good estimates according to the local minimums of the curve. The SC index shows that $c^* = 3$ is a good estimate which is also coincident to the result shown in (Zahid et al., 1999). In our simulations, the SC index is more sensitive to initial values of clustering algorithms than other indexes. Different initial values always leads to different results. The proposed PCASE index gives that $c^* = 2$ is an optimal cluster number estimate, but it also shows that $c^* = 3$ may be another good cluster number estimate. Overall, most validity indexes give the optimal cluster number estimate $c^* = 2$ or 3 for the Iris data set.

Example 8 (*Glass data set*). The Glass data set (Blake and Merz, 1998) has $n = 214$ points in an $s = 9$ dimensional space. It consists of six clusters. Since the clusters of this data set are heavily overlapped, the validity indexes are difficult to produce a good cluster number estimate. Fig. 9 shows the cluster validity results for the normalized Glass data set. The MPC index shows that $c^* = 3$, 4, 5 and 7 are all good cluster number estimates. The optimal cluster number estimates for this data set obtained by the XB and FHV indexes are 5 and 4, respectively. Others present the monotonic tendency of the cluster number c. Note that, if the data set has many undistinguishable clusters, the validity indexes may present a monotonic tendency of c. In this situation, the extreme value in validity index curves may not be a suitable cluster number estimate. This can also be found in the

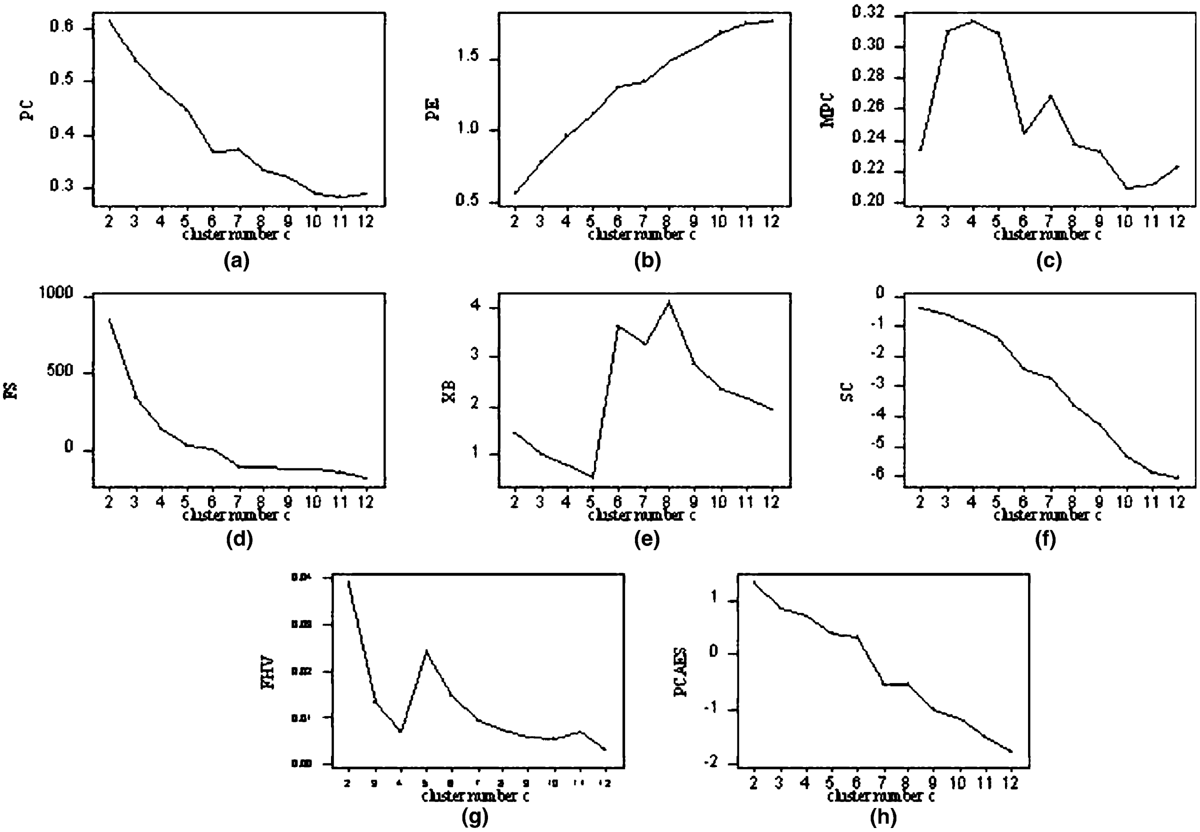


Fig. 9. Results of validity indexes for the normalized Glass data set.

next example. Although the index curve may present a monotonic tendency, the information about the data structure can also be found by checking for the trend of the index curve. Note that, the PCAES value is positive when $c^* = 6$ and negative when $c^* = 7$ and there is a large decreasing in PCAES values when $c^* = 6$–7. Therefore, the PCAES index offers the information that $c^* = 6$ is a good cluster number estimate for the Glass data set.

Example 9 (*Vowel data set*). The Vowel data set (Blake and Merz, 1998) has $n = 990$ points in an $s = 10$ dimensional space that has 11 clusters. Fig. 10 shows the validity results for the normalized Vowel data set. Most indexes present a monotonic tendency of the cluster number c except XB and PCAES. Although most validity indexes cannot validate a good cluster number estimate in the Vowel data set, the proposed PCAES index indicates that $c^* = 3$ or 11 may be a good cluster number estimate according to the local maximums of the index curve. In this data set, the PCASE index actually offers more information about the structure of the Vowel data set than other validity indexes.

6. Conclusions

In this paper, we reviewed several validity indexes and then proposed a new validity index, called a PCAES index. In the proposed PCAES index, we give each identified cluster a normalized partition coefficient and exponential separation index to measure the potential whether the identified cluster has the ability to be a well-identified

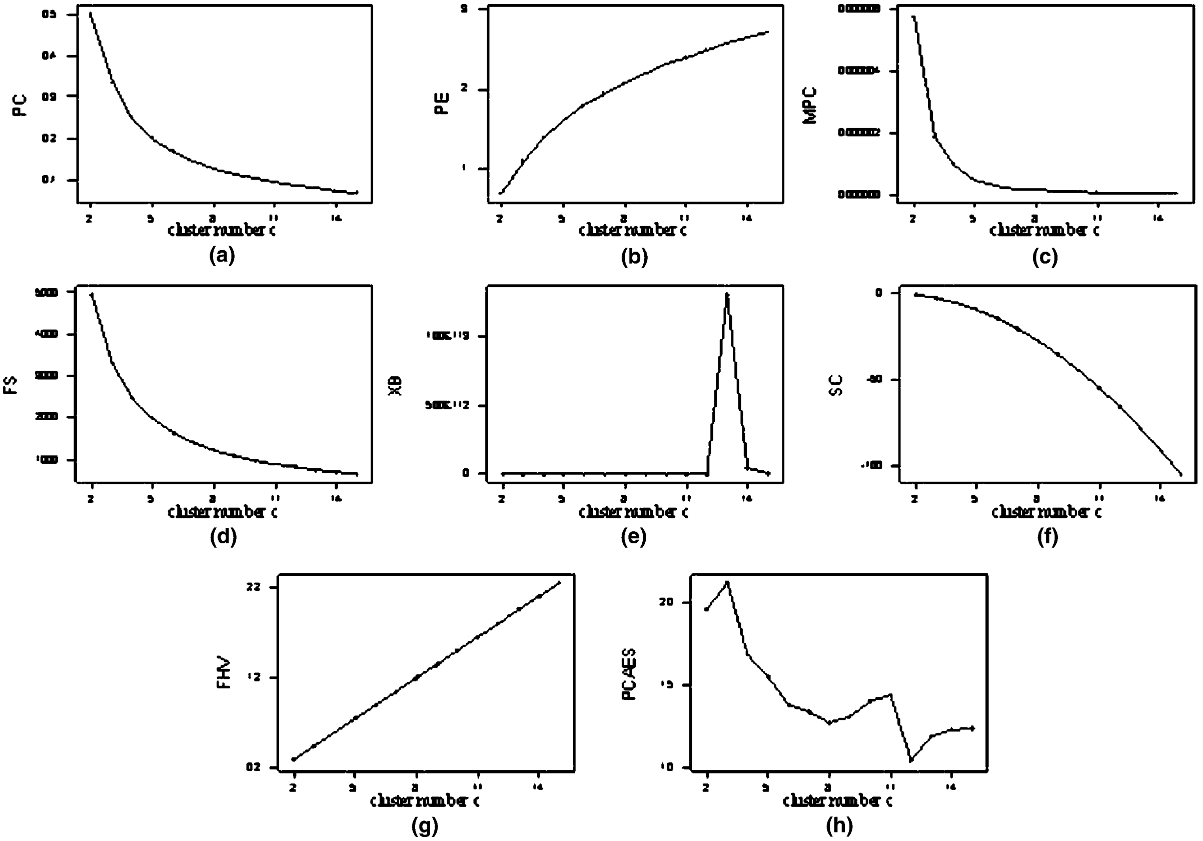


Fig. 10. Results of validity indexes for the normalized Vowel data set.

cluster or not. Most validity indexes will take a noisy point into a compact and well-separated cluster. In our PCAES index, each cluster has compactness and separation measures. If one of the *c* clusters has not enough potential to be a well-identified (valid) cluster, these *c* clusters will not validly describe the data structure. Under the proposed PCAES consideration, a noisy point will not have enough potential to be a cluster so that it can give impressive results in a noisy environment. In simulations, several different types of numerical and real data sets were given to compare the PCAES index with the PC, PE, MFC, FS, XB, SC and FHV indexes. The proposed PCAES index has an important merit for validating the goodness of fit in partitions from the FCM clustering algorithm. It also provides a new point of view for cluster validity in a noisy environment.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported in part by the National Science Council of Taiwan under Grant NSC-89-2118-M-033-001.

References

Anderson, E., 1935. The IRISes of the Gaspe peninsula. Bull. Amer. IRIS Soc. 59, 2–5.

Backer, E., Jain, A.K., 1981. A clustering performance measure based on fuzzy set decomposition. IEEE Trans. Pattern Anal. Machine Intell. 3, 66–74.

Baraldi, A., Blonda, P., 1999a. A survey of fuzzy clustering algorithms for pattern recognition—Part I. IEEE Trans. System Man Cybernet.—Part B 29, 778–785.

Baraldi, A., Blonda, P., 1999b. A survey of fuzzy clustering algorithms for pattern recognition—Part II. IEEE Trans. System Man Cybernet.—Part B 29, 786–801.

Bezdek, J.C., 1974a. Cluster validity with fuzzy sets. J. Cybernet. 3, 58–73.

Bezdek, J.C., 1974b. Numerical taxonomy with fuzzy sets. J. Math. Biol. 1, 57–71.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Bezdek, J.C., Coray, C., Gunderson, R., Watson, J., 1981a. Detection and characterization of cluster substructure: I. Linear structure: Fuzzy *c*-lines. SIAM J. Appl. Math. 40 (2), 339–357.

Bezdek, J.C., Coray, C., Gunderson, R., Watson, J., 1981b. Detection and characterization of cluster substructure: II. Fuzzy *c*-varieties and convex combinations thereof. SIAM J. Appl. Math. 40 (2), 358–372.

Bezdek, J.C., Keller, J.M., Krishnapuram, R., Kuncheva, L.I., Pal, N.R., 1999. Will the Real Iris data please stand up? IEEE Trans. Fuzzy Systems 7, 368–369.

Blake, C.L., Merz, C.J., 1998. UCI repository of machine learning databases, a huge collection of artificial and real-world data sets. Available from: <http://www.ics.uci.edu/~mlearn/MLRepository.html>.

Cannon, R.L., Dave, J.V., Bezdek, J.C., 1986. Efficient implementation of the fuzzy *c*-means clustering algorithm. IEEE Trans. Pattern Anal. Machine Intell. 8, 248–255.

Dave, R.N., 1992. Generalized fuzzy *c*-shells clustering and detection of circular and elliptical boundaries. Pattern Recognition 25 (7), 713–721.

Dave, R.N., 1996. Validating fuzzy partition obtained through *c*-shells clustering. Pattern Recognition Lett. 17, 613–623.

Fukuyama, Y., Sugeno, M., 1989. A new method of choosing the number of clusters for the fuzzy *c*-means method. Proceeding of fifth Fuzzy Syst. Symp., pp. 247–250.

Gath, I., Geva, A.B., 1989. Unsupervised optimal fuzzy clustering. IEEE Trans. Pattern Anal. Machine Intell. 11, 773–781.

Gunderson, M., 1978. Application of fuzzy ISODATA algorithms to star tracker pointing systems. Proc. 7th Triennial World IFCA Cong., Helsinki, Filind, pp. 1319–1323.

Gustafson, D.E., Kessel, W., 1979. Fuzzy clustering with a fuzzy covariance matrix. Proc. IEEE Conf. Decision Contr., San Diego, CA, pp. 761–766.

Höppner, F., Klawonn, F., Kruse, R., Runkler, T., 1999. Fuzzy Cluster Analysis: Methods for Classification Data Analysis and Image Recognition. Wiley, New York.

Krishnapuram, R., Kim, J., 1999. A note on the Gustafson–Kessel and adaptive fuzzy clustering algorithms. IEEE Trans. Fuzzy Systems 7, 453–461.

Pal, N.R., Bezdek, J.C., 1995. On cluster validity for fuzzy *c*-means model. IEEE Trans. Fuzzy Systems 3, 370–379.

Pal, N.R., Pal, S.K., 1991. Entropy, a new definition and its applications. IEEE Trans. System Man Cybernet. 21, 1260–1270.

Pal, N.R., Pal, S.K., 1992. Some properties of the exponential entropy. Inform. Sci. 66, 119–137.

Robubens, M., 1978. Pattern classification problems and fuzzy sets. Fuzzy Sets Systems 1, 239–253.

Trauwaert, E., 1988. On the meaning of Dunn's partition coefficient for fuzzy clusters. Fuzzy Sets Systems 25, 217–242.

Wu, K.L., Yang, M.S., 2002. Alternative *c*-means clustering algorithms. Pattern Recognition 35, 2267–2278.

Xie, X.L., Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Machine Intell. 13, 841–847.

Yang, M.S., 1993. A survey of fuzzy clustering. Math. Comput. Modelling 18, 1–16.

Yu, J., Cheng, Q., Huang, H., 2004. Analysis of the weighting exponent in the FCM. IEEE Trans. Systems Man Cybernet.—Part B 34, 634–639.

Yu, J., Yang, M.S., accepted for publication. Optimality test for a generalized FCM and its application to parameter selection. IEEE Trans. Fuzzy Systems.

Zadeh, L.A., 1965. Fuzzy sets. Inform. Control 8, 338–353.

Zahid, N., Limouri, M., Essaid, A., 1999. A new cluster-validity for fuzzy clustering. Pattern Recognition 32, 1089–1097.