

An Exponential Cluster Validity Index for Fuzzy Clustering with Crisp and Fuzzy Data

M.H. Fazel Zarandi^{1,*}, M.R. Faraji¹ and M. Karbasian²

Abstract. This paper presents a new cluster validity index for finding a suitable number of fuzzy clusters with crisp and fuzzy data. The new index, called the ECAS-index, contains exponential compactness and separation measures. These measures indicate homogeneity within clusters and heterogeneity between clusters, respectively. Moreover, a fuzzy c -mean algorithm is used for fuzzy clustering with crisp data, and a fuzzy k -numbers clustering is used for clustering with fuzzy data. In comparison to other indices, it is evident that the proposed index is more effective and robust under different conditions of data sets, such as noisy environments and large data sets.

Keywords: Fuzzy clustering; Cluster validity index; Fuzzy c -mean algorithm; Fuzzy k -numbers clustering; Fuzzy numbers; Compactness; Separation.

INTRODUCTION

Cluster analysis is the art of partitioning a given data set into similar clusters (groups, subsets, classes), where the partitions should have the following two properties:

- (a) Homogeneity within the clusters,
- (b) Heterogeneity between clusters [1].

There are two scopes for clustering: hard clustering and fuzzy clustering. In hard clustering, each point of a data set is assigned to exactly one cluster, while in fuzzy clustering, each point of the data set belongs to several clusters as a matter of degree in $[0, 1]$. One of the most popular fuzzy clustering methods is Fuzzy c -Means (FCM), proposed by Dunn [2], and then generalized by Bezdek [3]. The fuzzy c -means clustering model was the first model that was computationally efficient and powerful and, therefore, represents the best-known and used clustering approach [4]. Given an unlabeled data set, $X = \{x_1, x_2, \dots, x_n\} \subseteq \mathbb{R}^p$ (n and p are the number and dimension of data, respectively), the FCM

partitions the data set into c clusters by minimizing the evaluation function, $J_m(U, V) = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|x_j - v_i\|^2$, where u_{ij} is the membership degree of data point x_j to the i th cluster, v_i is the cluster centroid of cluster i , and $\|x_j - v_i\|^2$ is the Euclidian distance between x_j and v_i .

Bezdek [3] presented the FCM clustering algorithm as follows.

Algorithm

Fuzzy c -means clustering method (FCM): Given an unlabeled data set, $X = \{x_1, x_2, \dots, x_n\} \subseteq \mathbb{R}^p$ (n and p are the number and dimension of data, respectively), the number of clusters (c), the weighting exponent (m) and the termination criterion (ϵ), this method partitions data set X into c desired clusters with high homogeneity within and heterogeneity between clusters, while minimizing $J_m(U, V)$. It should be noted that the FCM algorithm depends on the initial seed, U_0 (initial membership degree matrix), m , and c . The fuzzy c -mean clustering algorithm [3] is demonstrated in Figure 1.

In the FCM algorithm, first we need to determine the suitable number of clusters. In literature, many studies in dealing with this problem are available and, so, there are many cluster validity indices in this regard. Compactness and separation are two criteria for the clustering evaluation and selection of

1. Department of Industrial Engineering, Amirkabir University of Technology, Tehran, P.O. Box 15875-4413, Iran.

2. School of Economic Science, Ministry of Finance and Economics, Tehran, P.O. Box 15875-111, Iran.

*. Corresponding author. E-mail: zarandi@aut.ac.ir

Received 26 September 2009; received in revised form 7 June 2010; accepted 16 October 2010

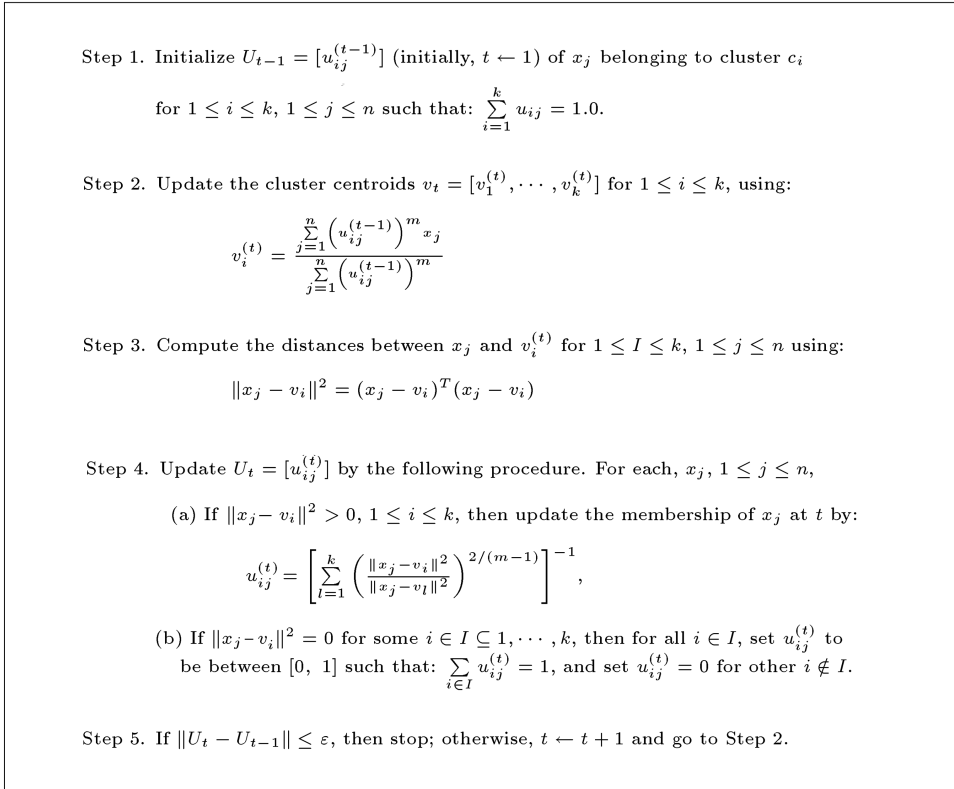


Figure 1. Fuzzy c -means clustering algorithm (FCM).

an optimal clustering scheme [5]. The variation of data within clusters indicates compactness and isolation between clusters indicates separation, respectively.

However, since there is no specified relation to determine these criteria besides which the structure of the data set is usually unknown, then the various indices are proposed. The experimental results on well-known data sets show some of these indices cannot quantify the compactness and separation of data. Also some, regarding huge data, cannot recognize the actual number of clusters correctly, and a number of indices in noisy environments are not robust and have different results [5,6].

In this paper, we present a new cluster validity index called an Exponential Compactness And Separation (ECAS) index. It uses normalized exponential compactness and separation measures. This index was inspired by other indices, especially the proposed index by Wu and Yang [7], and it will be discussed later in detail. Also, the proposed index is modified to be able to validate fuzzy clustering with fuzzy data.

The rest of the paper is organized as follows: the next section reviews several cluster validity functions. The following two sections present the proposed validity index for fuzzy clustering with crisp and fuzzy data, separately. Subsequently, the Experimental Result

Section uses several well-known artificial data sets and image data sets to test and validate the proposed index with crisp and fuzzy data. In the final section, conclusions and future works are presented.

CLUSTER VALIDITY INDICES

In this section, we review some cluster validity indices used for fuzzy clustering. To be more familiar with the indices, we refer to the work of Wang and Zhang [5] and Kim et al. [8], which review many fuzzy cluster validity indices. Table 1 lists a number of popular cluster validation indices, which are evaluated in our study.

Validity indices in a fuzzy environment can be divided into three groups [5]:

1. Indices involving only the membership values,
2. Indices involving the membership values and data set,
3. Other approaches for fuzzy cluster validity.

From Table 1, V_{PC} , V_{PE} and V_{MPC} belong to group 1, and V_{FS} , V_{XB} , V_K , V_{FHV} , V_{PBMF} , V_{PCAES} and V_W belong to group 2. However, some studies such as Falasconi et al. [9] consider another group for indices based on fuzzy hypervolume and density; this classification includes indices like V_{FHV} .

The first group of indices involves only the membership degrees, but not the data set itself. The main advantage of this group is its suitable computational effort. However, some of the drawbacks are:

1. The monotonous dependency on the number of clusters,

2. The sensitivity to fuzzifier m , and, most importantly,
3. The lack of a direct connection to the geometry of the data, since it does not use the data itself [5].

The second group includes indices that use both compactness and separation measurements, but due

Table 1. Ten validity indices for fuzzy clustering.

Validity Index	Function Description	Optimal Cluster Number
Partition coefficient [3,10]	$V_{PC} = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2$	$\max\{V_{PC}(U, c_i, m)\}$
Partition entropy [11,12]	$V_{PE} = -\frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n u_{ij} \log u_{ij}$	$\min\{V_{PE}(U, c_i, m)\}$
Modification of partition coefficient [13]	$V_{MPC} = 1 - \frac{c}{c-1}(1 - V_{PC})$	$\max\{V_{MPC}(U, c_i, m)\}$
Fukuyama and Sugeno [14]	$V_{FS} = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \ x_j - v_i\ ^2 - \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \ v_i - \bar{v}\ ^2$	$\min\{V_{FS}(U, c_i, m)\}$
Xie and Beni [15]	$V_{XB} = \frac{J_m(U, V)/n}{Sep(V)} = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \ x_j - v_i\ ^2}{n \min_{i \neq j} \ v_i - v_j\ ^2}$	$\min\{V_{XB}(U, c_i, m)\}$
Kwon [16]	$V_K = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 \ x_j - v_i\ ^2 + \frac{1}{c} \sum_{i=1}^c \ v_i - \bar{v}\ ^2}{\min_{i \neq j} \ v_i - v_j\ ^2}$	$\min\{V_K(U, c_i, m)\}$
Fuzzy hypervolume validity [17]	$V_{FHV} = \sum_{i=1}^c [\det(F_i)]^{1/2},$ $F_i = \frac{\sum_{j=1}^n u_{ij}^m (x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^n u_{ij}^m}$	$\min\{V_{FHV}(U, c_i, m)\}$
Pakhira-Bandyopadhyay-Maulik [18]	$V_{PBMF} = \left(\frac{1}{c} \times \frac{\sum_{j=1}^n \ x_j - v_i\ }{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \ x_j - v_i\ ^2} \times \max_{i,j=1}^c \ v_i - v_j\ \right)^2$	$\min\{V_{PBMF}(U, c_i, m)\}$
Partition coefficient and exponential separation [7]	$V_{PCAES} = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 / u_M$ $- \sum_{i=1}^c \exp(-\min_{k \neq i} \{\ v_i - v_k\ ^2 / \beta_T\})$ $u_M = \min_{1 \leq i \leq c} \left\{ \sum_{j=1}^n u_{ij}^2 \right\},$ $\beta_T = \frac{\sum_{l=1}^c \ v_l - \bar{v}\ ^2}{c}$	$\max\{V_{PCAES}(U, c_i, m)\}$
W-index [6]	$V_W(U, V) = \frac{Var^N(U, V)}{Sep^N(c, V)} = \left(\frac{Var(U, V)}{Var_{\max}} \right) / \left(\frac{Sep(c, U)}{Sep_{\max}} \right)^t$ where: $Var(U, V) = \left[\sum_{i=1}^c \sum_{j=1}^n u_{ij} \left(1 - \exp\left(-\frac{\ x_j - v_i\ ^2}{\beta}\right) \right) / n(i) \right] \times \left(\frac{c+1}{c-1} \right)^{1/2}$ $\beta = \left(\frac{\sum_{j=1}^n \ x_j - \bar{v}\ ^2}{n} \right)^{-1}$ $Sep(c, U) = 1 \max_{i \neq j} [\max_{x_k \in X} \min(u_{ik}, u_{jk})]$ $Var_{\max} = \max_c Var(U, V),$ $Sep_{\max} = \max_c Sep(c, V)$	$\min\{V_W(U, c_i, m)\}$

x_j is the j th data point, v_i are cluster centers, c_i is the number of clusters, \bar{v} is the grand mean of all data, x_j and u_{ij} is the membership value of data x_j of class c_i .

to the variety of the used measurements, they have different results. With respect to [7], V_{FS} does not use a good separation measurement ($\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|v_i - \bar{v}\|^2$) for the data structure, because $\|v_i - \bar{v}\|^2$ is not a good separation measure for cluster i . On the other hand, in V_{XB} , the extension of $V_{XB}(V_K)$ and V_{PBMF} indices, and the used separation, measures, $\min_{i \neq j} \|v_i - v_j\|^2$ and $\max_{i,j=1}^c \|v_i - v_j\|^2$, are considered for all clusters, not for each cluster [7].

Now, we focus our attention on V_{PCAES} and V_W . The partition coefficient and exponential separation index ($PCAES$) proposed by Wu and Yang [7] contains two factors to validate compactness and separation for each cluster. The $PCAES$ index for each cluster i is defined as follows:

$$PCAES_i = \sum_{j=1}^n u_{ij}^2 / u_M - \exp \left(- \min_{k \neq i} \{ \|v_i - v_k\|^2 \} / \beta_T \right), \quad (1)$$

where:

$$u_M = \max_{1 \leq i \leq c} \left\{ \sum_{j=1}^n u_{ij}^2 \right\},$$

and:

$$\beta_T = \frac{\sum_{l=1}^c \|v_l - \bar{v}\|^2}{c}.$$

$\sum_{j=1}^n u_{ij}^2 / u_M$ is a Normalized Partition Coefficient (NPC) and is used to measure the compactness of cluster i relative to the most compact cluster that has the compactness measure u_M . Also, $\exp(-\min_{k \neq i} \{ \|v_i - v_k\|^2 \} / \beta_T)$ is an exponential separation measure for cluster i and is used to measure the distance between cluster i and its closest cluster. The large $PCAES_i$ value means that cluster i is compact inside and separated from other $(c-1)$ clusters.

Finally, Wu and Yang defined the $PCAES$ validity index as follows [7]:

$$V_{PCAES} = PCAES(c) = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 / u_M - \sum_{i=1}^c \exp \left(- \min_{k \neq i} \{ \|v_i - v_k\|^2 \} / \beta_T \right). \quad (2)$$

The $PCAES$ validity index takes advantage of the exponential function in the separation measure. An optimal c^* can be found by solving $\max_{2 \leq c \leq n-1} PCAES(c)$

to produce a best clustering performance for data set X .

Also, in the literature of fuzzy cluster validity, Zhang et al. [6] presented a new index called w index or V_W . This index is defined as follows:

$$V_W(U, V) = \frac{Var^N(U, V)}{Sep^N(c, V)} = \frac{\left(\frac{Var(U, V)}{Var_{\max}(U, V)} \right)}{\left(\frac{Sep(c, V)}{Sep_{\max}(c, V)} \right)}, \quad (3)$$

where:

$$Var(U, V) = \left[\frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij} \left(1 - \exp \left(- \frac{\|x_j - v_i\|^2}{\beta} \right) \right)}{n(i)} \right] \times \left(\frac{c+1}{c-1} \right)^{1/2}, \quad (4)$$

$$Sep(c, U) = 1 - \max_{i \neq j} \left[\max_{x_k \in X} \min(u_{ik}, u_{jk}) \right], \quad (5)$$

$$Var_{\max} = \max_c Var(U, V),$$

$$Sep_{\max} = \max_c Sep(c, V),$$

$$\beta = \left(\frac{\sum_{j=1}^n \|x_j - \bar{v}\|^2}{n} \right)^{-1}.$$

In V_W , $\left(\frac{c+1}{c-1} \right)^{1/2}$ is used only to adjust the value of the compactness measure or variation measure. V_W takes advantage of the exponential function to validate the compactness measure ($Var(U, V)$), and also it involves the distance between the data point and cluster centers, which is necessary for determining the dissimilarity of the data points. The experimental results in Wang and Zhang [5], and Zhang et al. [6] on the well-known artificial data sets have shown that V_W has one of the best performances compared to other indices. On the other hand, in the experimental results of Wang and Zhang [5], and Zhang et al. [6], it can be observed that V_{PCAES} incorrectly recognized the suitable cluster numbers in Dataset_5_2, Dataset_6_2 and Dataset_10_2. These three data sets are simulated and used in this paper and, therefore, are shown in Figure 2.

By focusing our attention on these data sets, it can be observed that Dataset_6_2 and Dataset_10_2, visually, have well-separated clusters, but some of these clusters are located near to each other and,

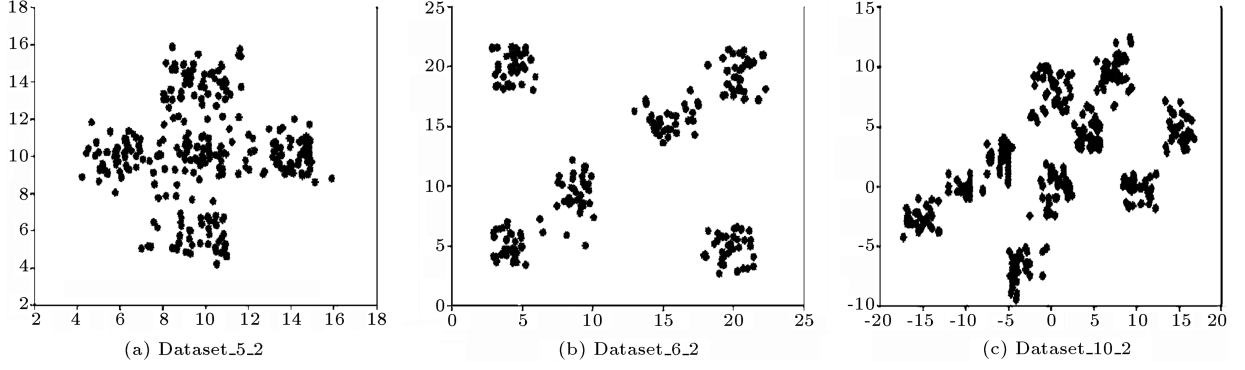


Figure 2. The PCAES index incorrectly recognized the suitable number of clusters for these data sets.

therefore, their cluster centers are close together. So, these factors, with respect to definitions of the compactness and separation measure, emphasize likely V_{PCAES} practiced weak in compactness measure. In fact, this vision is almost true (almost, because separation and compactness measures to validate clustering should be considered concurrently), because V_{PCAES} only uses membership values to validate the compactness measure, and does not consider the structure of data, i.e. the relative distance between the data set and cluster centers are not taken into account.

On the other hand, the advantages of taking the exponential function have been understood in earlier work on cluster analysis [6]. In this research, we define a new compactness measure instead of the compactness measure in V_{PCAES} , in an exponential type, which can be led to fortify V_{PCAES} . We also reformed the separation measure in V_{PCAES} that led to a new cluster validity index.

PROPOSED CLUSTER VALIDITY INDICES

So far, a variety of validity indices have been proposed for fuzzy clustering by different scholars. These variations have two main reasons:

- In spite of the existence of two characteristics to obtain the actual number of clusters, compactness and separation, a specified relation or equation does not exist to determine the value of these characteristics.
- The nature of data (whose dimensions are often more than 2, and so whose structures are unknown).

In this section, a new cluster validity index is proposed, called an Exponential Compactness And Separation (ECAS) index. This index contains two terms quantifying compactness and separation, respectively. In the next subsection, we explain the proposed index and its related behavior.

Definitions

A compactness measure must consider the variation or scattering of data within a cluster. On the other hand, as already mentioned, the advantages of taking the exponential function has been known in cluster analysis. Especially, Zhang et al. [6] presented an exponential compactness measure (i.e., $Var(U, V)$ in Equation 4) that has been discussed in the previous section. With this background, and almost similar to components of $Var(U, V)$ (i.e. β , $n(i)$ and adjusted value), we propose an exponential compactness measure called EC_{comp} (Exponential Compactness), as follows:

$$EC_{comp}(c) = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \exp \left(- \left(\frac{\|x_i - v_j\|^2}{\beta_{comp}} + \frac{1}{c+1} \right) \right), \quad (6)$$

where β_{comp} is defined as the sample covariance for cluster i , i.e.:

$$\beta_{comp} = \frac{\sum_{k=1}^n \|x_k - \bar{v}\|^2}{n(i)},$$

with:

$$\bar{v} = \frac{\sum_{j=1}^n x_j}{n},$$

where $n(i)$ is the number of data in cluster i , which can be determined by related methods (e.g. using any t-conorm in fuzzy logic between membership degrees of a datum in c clusters). The power of EC_{comp} contains two terms:

$$\|x_i - v_j\|^2 / \beta_{comp},$$

and:

$$\frac{1}{c+1}.$$

The first term is related to the compactness within clusters and the last term is only used to adjust the value of the compactness measure, which will be clarified in the next subsection. Also, β_{comp} plays the role of factor β of the compactness measure used in V_W [6] as a positive constant, but β_{comp} is a relative positive value that adjusts with the amount of data in cluster i (i.e., $n(i)$).

The separation measure must consider the isolation distance between fuzzy clusters. In the previous section, we focused on V_{PCAES} and discussed its advantages and shortcomings in compactness and separation measures. It is found that in V_{PCAES} , the separation measure does not work well in some cases. Therefore, in this research, a new separation measure called ES_{sep} (Exponential Separation) is defined as follows:

$$ES_{sep}(c) = \sum_{i=1}^c \exp \left(- \min_{i \neq k} \left\{ \frac{(c-1) \|v_i - v_k\|^2}{\beta_{sep}} \right\} \right), \quad (7)$$

where β_{sep} is defined as the total average distance measure for all clusters, i.e.:

$$\beta_{sep} = \frac{\sum_{l=1}^c \|v_l - \bar{v}\|^2}{c}.$$

It is easy to verify that ES_{sep} is a modified version of the separation measure used in V_{PCAES} ; in fact, we only added a multiplier $(c-1)$ in the numerator. The separation measure of V_{PCAES} only uses the minimum distance between cluster center i and other $(c-1)$ cluster centers to quantify the isolation between clusters, but the added multiplier in ES_{sep} leads to the use of this minimum distance instead of all $(c-1)$ other distances.

Finally, we normalize and then compose these two terms to create the $ECAS$ -index as follows:

$$V_{ECAS} = ECAS(c) = \frac{EC_{comp}(c)}{\max_c(EC_{comp}(c))} - \frac{ES_{sep}(c)}{\max_c(ES_{sep}(c))}. \quad (8)$$

A large value for the compactness measure over c indicates a compact partition, and a small value for the separation measure over c indicates well-separated clusters. Eventually, an optimal cluster number is obtained by maximizing $ECAS(c)$ over $2 \leq c \leq c_{max}$.

Explanation

More explanations of V_{ECAS} are as follows:

- (a) Since $0 < \exp(-z) \leq 1$, ($\forall z \in \mathbb{R}^+$), then $0 \leq EC_{comp}(c) \leq \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m$ and $0 \leq ES_{sep}(c) \leq$

c . Therefore, in a huge data set, the value of $EC_{comp}(c)$ is very large and the value of $ES_{sep}(c)$ is low (because $\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m$ is very large and c is small). Hence these different scales need to be normalized, so in the V_{ECAS} index, $EC_{comp}(c)$ and $ES_{sep}(c)$ should be divided to $\max_c(EC_{comp}(c))$ and $\max_c(ES_{sep}(c))$ for $c = 2, 3, \dots, c_{max}$ for normalization, respectively.

- (b) When c approaches the number of data samples, n , then $\|x_i - v_j\|^2 \rightarrow 0$, $n(i) \rightarrow 1$ and $\frac{1}{c+1}$ decreases. Therefore, the power of $EC_{comp}(c)$ approaches 0 and so $EC_{comp}(c)$ increases. On the other hand, when $c \rightarrow n$, then each cluster center is equivalent to one data, i.e. $x_j \rightarrow v_j$ and so the power of $ES_{sep}(c)$ increases. Therefore, $ES_{sep}(c)$ increases and restrains an increase in the V_{ECAS} value in total.
- (c) From Equation 6, we can rewrite $EC_{comp}(c)$ as a multiplication of two terms as follows:

$$EC_{comp}(c) = \left[\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \exp \left(- \frac{\|x_i - v_j\|^2}{\beta_{comp}} \right) \right] \times \left[\exp \left(- \frac{1}{c+1} \right) \right]. \quad (9)$$

In Equation 9, $\exp \left(- \frac{1}{c+1} \right)$ is only used to adjust the value of the compactness measure. This adjustment can be cleared with a numerical example, as shown in Figure 3.

Assume the first term of Equation 9, $\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \exp \left(- \frac{\|x_i - v_j\|^2}{\beta_{comp}} \right)$, is denoted by $func_{comp}(c)$ and the second term is denoted by $adj_{comp}(c)$. We cluster the above numerical example using the FCM algorithm over $2 \leq c \leq 100$, and calculate $adj_{comp}(c)$, $func_{comp}(c)$ and $EC_{comp}(c)$. Figure 4 shows their values, and

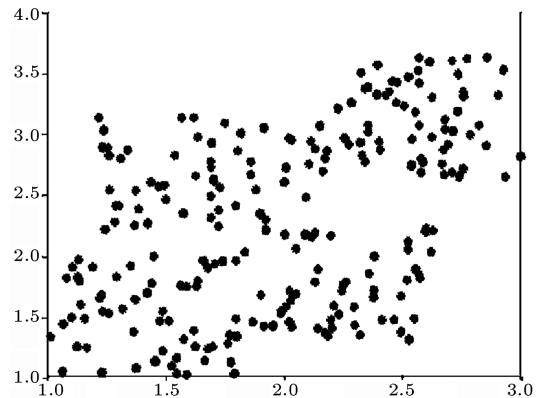


Figure 3. A data set with 4 clusters and 2 dimensions (DataSet_4_2).

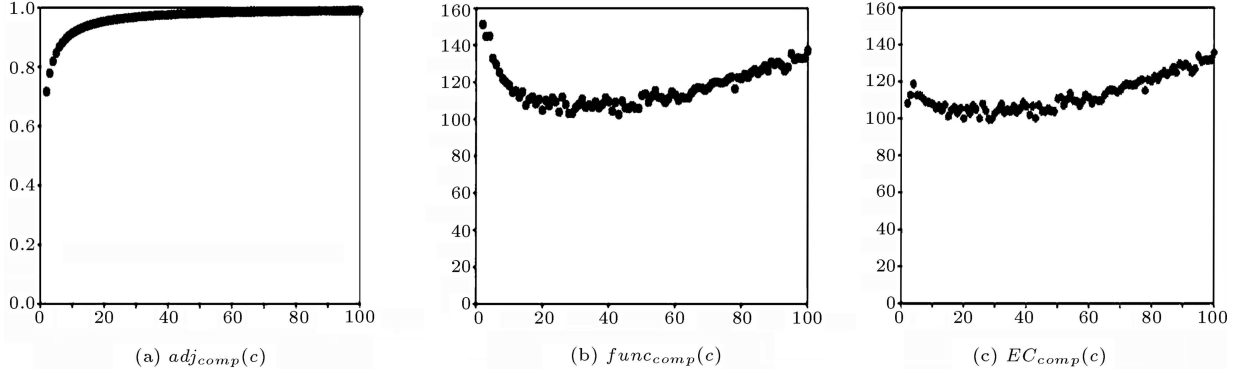


Figure 4. Compactness measure and its terms for Dataset 4.2 over $2 \leq c \leq 100$. (a) Adjust value; (b) compactness function; and (c) compactness measure.

it can be observed that the $EC_{comp}(c)$ curve is the adjusted curve of $func_{comp}(c)$. For example, $func_{comp}(2) = 151.1929$ is greater than $func_{comp}(4) = 144.8754$, but after the adjustment, $EC_{comp}(2) = 108.3345$ is less than $EC_{comp}(4) = 118.6139$. In fact, 4 clusters is the optimal number of clusters for this case.

- (d) As already mentioned, the proposed separation measure is a modified version of the separation measure defined in V_{PCAES} [7], with this difference; its numerator is multiplied by $(c - 1)$. In fact, V_{PCAES} , for separation of cluster i , only uses the minimum of Euclidean distances between cluster i and other clusters, while V_{ECAS} uses the summation of $c-1$ distances by using the minimum distance instead of all distances.
- (e) Both $EC_{comp}(c)$ and $ES_{sep}(c)$ take advantage of the exponential function to measure distances. The motivations behind and advantages of taking the exponential function is that an exponential operation is highly useful in dealing with classical Shannon entropy [19,20] and cluster analysis [6,7,21]. Especially, Wu and Yang [21] had claimed that an exponential-type distance gives a robust property, based on the influence function analysis.

MODIFIED CLUSTER VALIDITY FOR CLUSTERING WITH FUZZY DATA

In this section, we modify V_{ECAS} for fuzzy clustering with LR -type fuzzy data. This type of fuzzy data can be collected in a matrix called the LR fuzzy data matrix:

$$\tilde{X} \equiv \{\tilde{x}_{ij} = (m_{ij}, \alpha_{ij}, \beta_{ij})_{LR} : i = 1, I, j = 1, J\},$$

where i and j denote the units and fuzzy variables, respectively; $\tilde{x}_{ij} = (m_{ij}, \alpha_{ij}, \beta_{ij})_{LR}$ represents LR fuzzy variable j observed on the i th unit where m_{ij} denotes the mean, and α_{ij} and β_{ij} indicate the left and right spread, with the following membership function:

$$\mu(\tilde{u}_{ij}) = \begin{cases} L\left(\frac{m_{ij} - \tilde{u}_{ij}}{\alpha_{ij}}\right) \tilde{u}_{ij} \leq m_{ij} & (\alpha_{ij} > 0) \\ R\left(\frac{\tilde{u}_{ij} - m_{ij}}{\beta_{ij}}\right) \tilde{u}_{ij} \geq m_{ij} & (\beta_{ij} > 0) \end{cases} \quad (10)$$

where $L(w_{ij})$ (and $R(w_{ij})$) is a decreasing “shape” function \mathbb{R}^+ to $[0, 1]$ with $L(0) = 1$; $L(w_{ij}) < 1$ for all $w_{ij} > 0$, $\forall i, j$; $L(w_{ij}) > 0$ for all $w_{ij} < 1$, $\forall i, j$; $L(1) = 0$ (or $L(w_{ij}) > 0$ for all w_{ij} , $\forall i, j$, and $L(+\infty) = 0$) [4,22]. In LR -type fuzzy data, the Triangular Fuzzy Numbers (TFNs) are most commonly used. For a LR -type fuzzy data, $\tilde{x} = (m_{\tilde{x}}, \alpha_{\tilde{x}}, \beta_{\tilde{x}})$, if $L(x) = R(x) = 1 - x$, then \tilde{x} is called a TFN, i.e.:

$$\mu(x) = \begin{cases} 1 - \frac{m_{\tilde{x}} - x}{\alpha_{\tilde{x}}} & \text{for } x \leq m_{\tilde{x}} (\alpha_{\tilde{x}} > 0) \\ 1 - \frac{m_{\tilde{x}} - x}{\beta_{\tilde{x}}} & \text{for } x \geq m_{\tilde{x}} (\beta_{\tilde{x}} > 0) \end{cases} \quad (11)$$

In the literature, there exist several research papers in fuzzy clustering with fuzzy data. Yang and Ko [23] presented a FKNC or FCNC model (fuzzy k -numbers clustering model) that uses a square distance between each pair of fuzzy numbers. Yang and Liu [24] extended the FKNC model and proposed a FCMCCFV model for high-dimensional fuzzy data (a fuzzy c -means clustering model for a conical fuzzy vector). Hung and Yang [25] proposed an AFKNC model (alternative fuzzy k -numbers clustering model) that it is a modified version of the FKNC model that uses an exponential type of FKNC distance. Also, the FWCNC model (weighted fuzzy c -means clustering model) proposed by D’Urso and Giordani [26] considers fuzzy data with a symmetric LR membership function. They introduced a weighted square distance measure between fuzzy data.

However, we use the square distance and FKNC model presented by Yang and Ko [23], which is a popular model in the literature of fuzzy clustering with fuzzy data [4]. Yang and Ko defined a new type of distance, d_{LR} or d_{YC} for any \tilde{X} and \tilde{Y} with

$\tilde{X} = (m_{\tilde{x}}, \alpha_{\tilde{x}}, \beta_{\tilde{x}})$ and $\tilde{Y} = (m_{\tilde{y}}, \alpha_{\tilde{y}}, \beta_{\tilde{y}})$ as follows [23]:

$$\begin{aligned} d_{LR}^2(\tilde{X}, \tilde{Y}) = & (m_{\tilde{x}} - m_{\tilde{y}})^2 + ((m_{\tilde{x}} - l\alpha_{\tilde{x}}) \\ & - (m_{\tilde{y}} - l\alpha_{\tilde{y}}))^2 + ((m_{\tilde{x}} + r\beta_{\tilde{x}}) \\ & - (m_{\tilde{y}} + r\beta_{\tilde{y}}))^2, \end{aligned} \quad (12)$$

where $l = \int_0^1 L^{-1}(\omega)d(\omega)$ and $r = \int_0^1 R^{-1}(\omega)d(\omega)$. Here, l and r are parameters that summarize the shape of the left and right tails of a membership function. Then each value of l and r gives a particular membership function; especially when $l = r = \frac{1}{2}$, we have a TFN [4].

Therefore, the FKNC model for LR -type Fuzzy data is characterized as [23]:

min:

$$\begin{aligned} \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m d_{LR}^2(\tilde{x}_j, \tilde{v}_i) = & \sum_{i=1}^c \sum_{j=1}^n u_{ik}^m [(m_j - m_{ci})^2 \\ & + [(m_j - l\alpha_j) - (m_{ci} - l\alpha_{ci})]^2 \\ & + [(m_j + r\beta_j) - (m_{ci} + r\beta_{ci})]^2], \end{aligned} \quad (13)$$

where \tilde{v}_i indicates the i th cluster center and $(m_{ci}, \alpha_{ci}, \beta_{ci})$ the mean and left and right spreads of \tilde{v}_i . By solving the optimization of the above objective function by means of the Lagrangian multiplier method, the following iterative solutions are obtained:

$$\begin{aligned} u_{ij} = & \frac{1}{\sum_{k=1}^c \left[\frac{LR d_{ij}^2}{LR d_{ik}^2} \right]^{2/m-1}}, \\ m_{ci} = & \frac{\sum_{k=1}^n u_{ki}^m [3m_k - l(\alpha_k - \alpha_{ci}) + r(\beta_k - \beta_{ci})]}{3 \sum_{k=1}^n u_{ik}^m}, \end{aligned}$$

and:

$$\begin{aligned} \alpha_{ci} = & \frac{\sum_{k=1}^n u_{ik}^m [m_{ci} + l\alpha_k - m_k]}{l \sum_{k=1}^n u_{ik}^m}, \\ \beta_{ci} = & \frac{\sum_{k=1}^n u_{ik}^m [m_k + r\beta_k - m_{ci}]}{r \sum_{k=1}^n u_{ik}^m}. \end{aligned}$$

With respect to these iterative solutions, the FKNC model must start with an initial fuzzy partition (for more details see [23]).

With this background, the compactness and separation measures of V_{ECAS} for fuzzy clustering with fuzzy data are modified and presented as follows:

The compactness measure can be modified as:

$$\begin{aligned} EC'_{comp}(c) = & \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \\ & \exp \left(- \left(\frac{d_{LR}^2(\tilde{x}_j, \tilde{v}_i)}{\beta'_{comp}} + \frac{1}{c+1} \right) \right), \end{aligned} \quad (14)$$

where:

$$\beta'_{comp} = \frac{\sum_{k=1}^n d_{LR}^2(\tilde{x}_k, \tilde{v})}{n(i)},$$

and $\tilde{v} = (m_{\tilde{v}}, \alpha_{\tilde{v}}, \beta_{\tilde{v}})$ is a sample mean of all data and $m_{\tilde{v}} = \frac{\sum_{k=1}^n m_k}{n}$, $\alpha_{\tilde{v}} = \frac{\sum_{k=1}^n \alpha_k}{n}$ and $\beta_{\tilde{v}} = \frac{\sum_{k=1}^n \beta_k}{n}$ [25].

Also, the separation measure is modified as follows:

$$ES'_{sep}(c) = \sum_{i=1}^c \exp \left(- \min_{i^1k} \left\{ \frac{(c-1)d_{LR}^2(\tilde{v}_i, \tilde{v}_k)}{\beta'_{sep}} \right\} \right), \quad (15)$$

where:

$$\beta'_{sep} = \frac{\sum_{t=1}^c d_{LR}^2(\tilde{v}_t, \tilde{v})}{c}.$$

Finally, the modified ECAS index for fuzzy clustering with fuzzy data is defined as follows:

$$\begin{aligned} V'_{ECAS} = & ECAS'(c) \\ = & \frac{EC'_{comp}}{\max_c(EC'_{comp})} - \frac{ES'_{sep}}{\max_c(ES'_{sep})}. \end{aligned} \quad (16)$$

Similar to the crisp data, a suitable number of clusters for fuzzy data is obtained by maximizing V'_{ECAS} over c . Since we modified V_{ECAS} with a change in its distance and used the fuzzy data instead of the crisp one, it is expected that V'_{ECAS} inherits the properties of V_{ECAS} . It can be seen, in our numerical examples, that V'_{ECAS} is robust to noise points as well as V_{ECAS} .

In the next section, to test and demonstrate the effectiveness of the proposed index for recognizing the suitable number of clusters, we use a number of well-known artificial data sets, image data sets and fuzzy data sets that exist in the literature.

EXPERIMENTAL RESULTS

This section presents the effectiveness of V_{ECAS} and V'_{ECAS} by applying some widely used data sets, and by making an extensive comparison with a number of

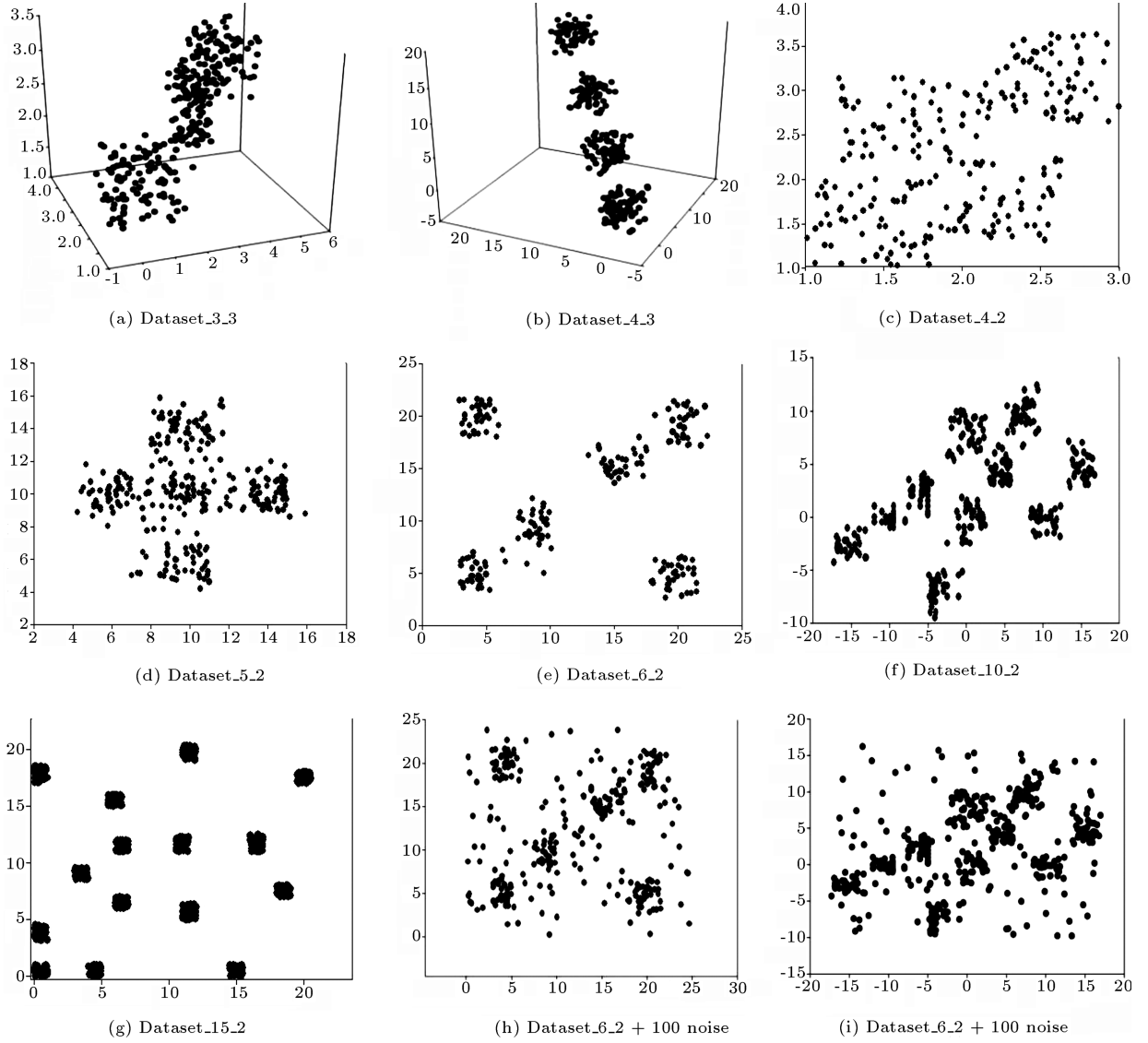


Figure 5. The nine artificial data sets.

predefined cluster validity indices. We used the FCM algorithm (initialed with U_0) where the initializing values are fuzzifier $m = 2$, termination criterion $\varepsilon = 10^{-5}$, and distance function $\|\cdot\|$ is the Euclidean distance. Moreover, we choose $c_{\min} = 2$ and $c_{\max} = \sqrt{n}$ based on Bezdek's suggestion [27].

Crisp Data Sets

In this subsection, to test V_{ECAS} and then to compare it with the aforementioned cluster validity indices, we used nine artificial and eight well-known data sets. The nine artificial data sets are called DataSet_3_3, DataSet_4_3, DataSet_4_2, DataSet_5_2, DataSet_6_2, DataSet_10_2, DataSet_15_2, Dataset_6_2 + 100 noise and Dataset_10_2+100 noise, some of which are similar to data sets used by Wang and Zhang, and Zhang et al. [5,6]. The names of the first seven data sets imply

the number of clusters actually present in the data and the number of dimensions, respectively. For example, for the DataSet_15_2 data, there are fifteen clusters and the dimension is 2. Also, 100 noise points are added to the Dataset_6_2 and Dataset_10_2, which are called Dataset_6_2+100 noise and Dataset_10_2+100 noise. The aim of this work is to demonstrate the robustness of the ECAS-index. These data sets are demonstrated in Figure 5, respectively.

Eight well-known data sets with Iris, Wisconsin Breast Cancer (WBCD), Wisconsin Diagnostic Breast Cancer (WDBC), Wine, Bupa Liver Disorder, Butterfly, Example_1 and Example_2 are used as the test data sets for cluster validity index comparisons. Table 2 shows these data sets and their characteristics. The first five data sets are real life data sets and are freely available at: <http://www.ics.uci.edu/~mllearn/databases.html>.

Table 2. The eight well-known data sets with their characteristics.

Data Sets	Number of Samples	Dimension	Best Partitioning
Iris	150	4	2 or 3
WBCD	683	9	2
WDBC	569	30	2
Wine	178	13	3
Liver Disorder	345	6	2
Butterfly	15	2	2
Example_1	16	2	3
Example_2	16	2	4

The last three data sets are similar to data sets used by Wang and Zhang [5]. The data sets are normalized in experiments. Also, Butterfly, Example_1 and Example_2 data are shown in Figure 6, respectively.

Experimental Results for Crisp Data

This subsection presents the experiment results of the abovementioned numerical examples to compare the proposed V_{ECAS} with the other ten indices, V_{PC} , V_{PE} , V_{MPC} , V_{FS} , V_{XB} , V_K , V_{PBMF} and V_{FHV} , V_{PCAES} and V_W . The optimal number of clusters is found in the maximum value of V_{ECAS} and these optimums for the aforementioned data sets are demonstrated in Figure 7. Table 3 shows the optimal number of clusters in the seventeen data sets, and the results found by the ten indices. Column 2 of Table 3 indicates the optimal number of clusters, i.e., C^* , and the other columns show the results obtained using each index. In this table, the last column shows the results of V_{ECAS} . As observed, V_{ECAS} correctly recognizes the optimal number of clusters for all data sets.

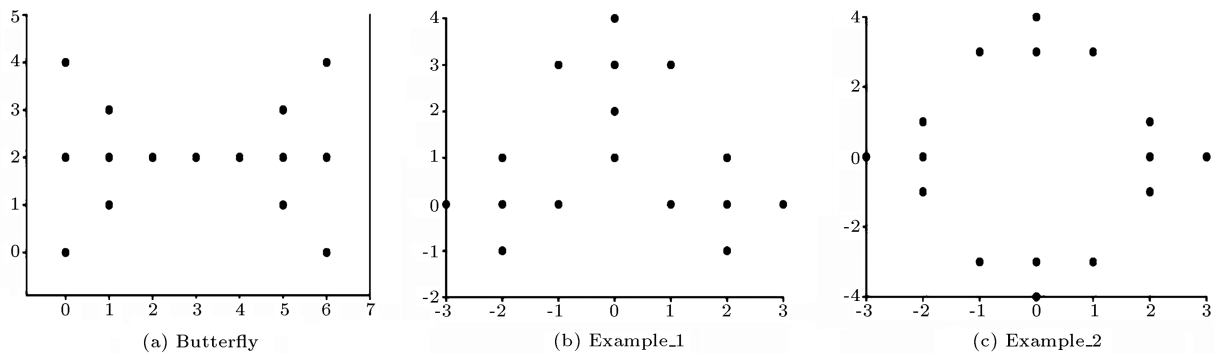
The experimental results show that V_{PE} correctly recognizes the optimal number of clusters for 8 data sets, V_{PC} , V_{FS} and V_{XB} for 11 data sets, V_K and

V_{PCAES} for 12 data sets, V_{FHV} for 15 data sets, V_{MPC} for 16 data sets, and V_{PBMF} , V_W and V_{ECAS} for all data sets. It should be noted that among these indices, V_{ECAS} , V_W , V_{MPC} , V_{FS} , V_{FHV} and V_{PBMF} are robust to noise. Hence the proposed index, V_{ECAS} , is one of the most effective indices considered.

Also in testing V_{ECAS} for large data sets, we apply it to image segmentation. Image segmentation refers to the process of partitioning a digital image into multiple regions (sets of pixels), typically used to locate objects and boundaries. One method that has been developed for image segmentation is the clustering method. Each pixel of a color image in HSV color space corresponds to hue, saturation and value (these three features in HSL color space are hue, saturation and lightness). So a color image of size $m \times n$ pixels corresponds to an array of size $m \times n$ and three dimensions. Figure 8 shows the six color images in image processing that are called Penguin, Water Lily, Im_1, Im_2, Im_3 and Im_4. The data sets corresponding to these images are shown in Figure 9. The Penguin, Im_1, Im_2, and Im_3 images are of size 128×128 in pixels, and the Water Lily and Im_4 images are of size 81×123 in pixels. Also it is necessary to state that the first two color images are also used by Zhang et al. [6].

Because \sqrt{n} is very large and, practically, in image data, the optimal number of clusters is found in $c^* \ll \sqrt{n}$, then we selected, randomly, $c_{\max} = 50$, which is equivalent to earlier works such as Zhang et al. [6].

Table 4 summarizes the results of partitioning the six abovementioned image data sets, which were obtained using the aforementioned cluster validity indices. Among these data sets, Im_2 is remarkable (Figures 10c and 10d). In Im_3, clearly the three clusters are the optimal number of clusters, as shown in Figure 10c, but its data set (Figure 9c) cannot demonstrate this issue. The results of the image segmentation show that V_{ECAS} correctly recognize the optimal number of clusters for five image data sets, V_{PBMF} and V_{FS} for 3 data sets, and the other indices do not have considerable output. Thus in huge data

**Figure 6.** The three well-known data sets.

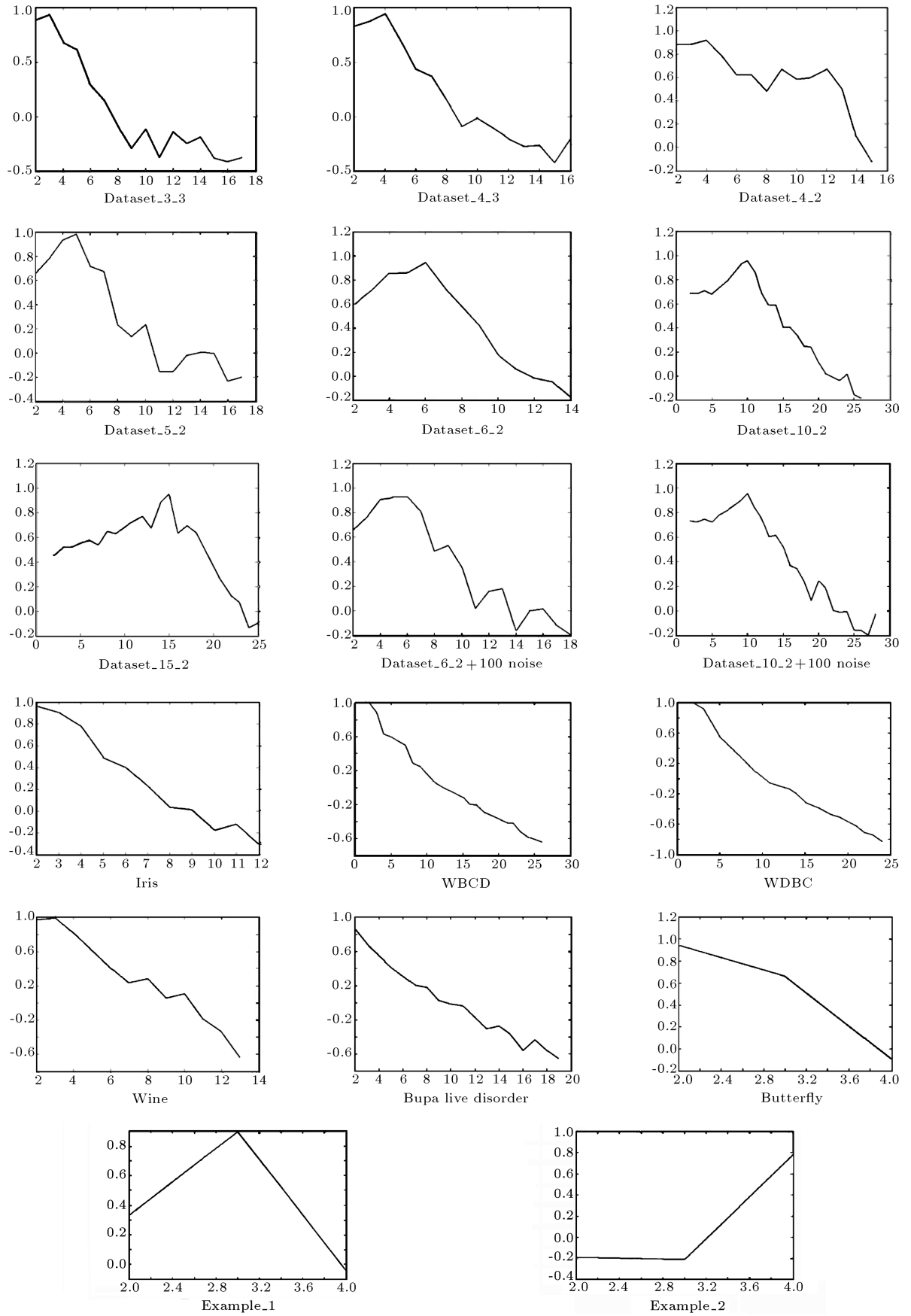


Figure 7. The maximum value of index corresponds to the optimal number of clusters.

Table 3. Values of c obtained by each cluster validity indices for seventeen data sets.

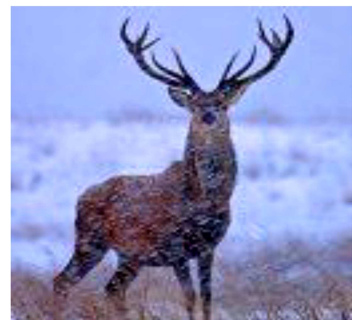
Dataset	C^*	V_{PC}	V_{PE}	V_{MPC}	V_{FS}	V_{XB}	V_K	V_{FHV}	V_{PBMF}	V_{PCAES}	V_W	V_{ECAS}
Dataset_3_3	3	2	2	3	3	2	2	3	3	3	3	3
Dataset_4_3	4	4	2	4	4	4	4	4	4	4	4	4
Dataset_4_2	3 or 4	2	2	2	11	13	4	14	3	4	4	4
Dataset_5_2	5	4	2	5	5	4	4	5	5	4	5	5
Dataset_6_2	6	6	6	6	6	4	4	6	6	4	6	6
Dataset_10_2	10	2	2	10	10	10	10	10	10	4	10	10
Dataset_15_2	15	15	17	15	15	15	15	15	15	15	15	15
Dataset_6_2 +100 noise	6	6	2	6	6	4	4	6	6	4	6	6
Dataset_10_2 +100 noise	10	2	2	10	10	7	7	10	10	4	10	10
Iris	2 or 3	2	2	2	5	2	2	3	3	2	3	2
WBCD	2	2	2	2	12	2	2	2	2	2	2	2
WDBC	2	2	2	2	4	2	2	2	2	2	2	2
Wine	3	2	2	3	13	3	3	3	3	3	3	3
Live disorder	2	2	2	2	4	2	2	18	2	2	2	2
Butterfly	2	2	2	2	2	2	2	2	2	2	2	2
Example_1	3	3	3	3	3	3	3	3	3	3	3	3
Example_2	4	4	4	4	4	4	4	4	4	4	4	4



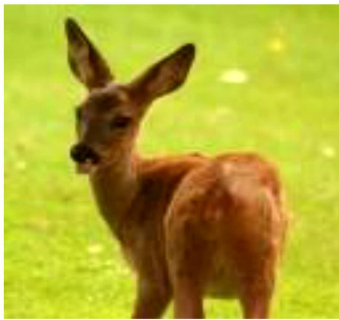
(a) Penguin



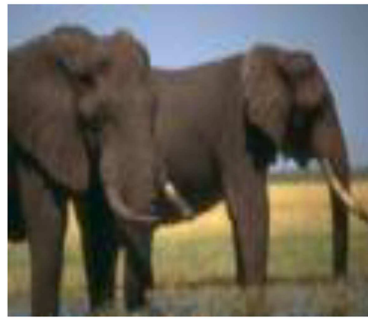
(b) Im_1



(c) Im_2



(d) Im_3



(e) Im_4



(f) Water Lily

Figure 8. Original color images to image segmentation.

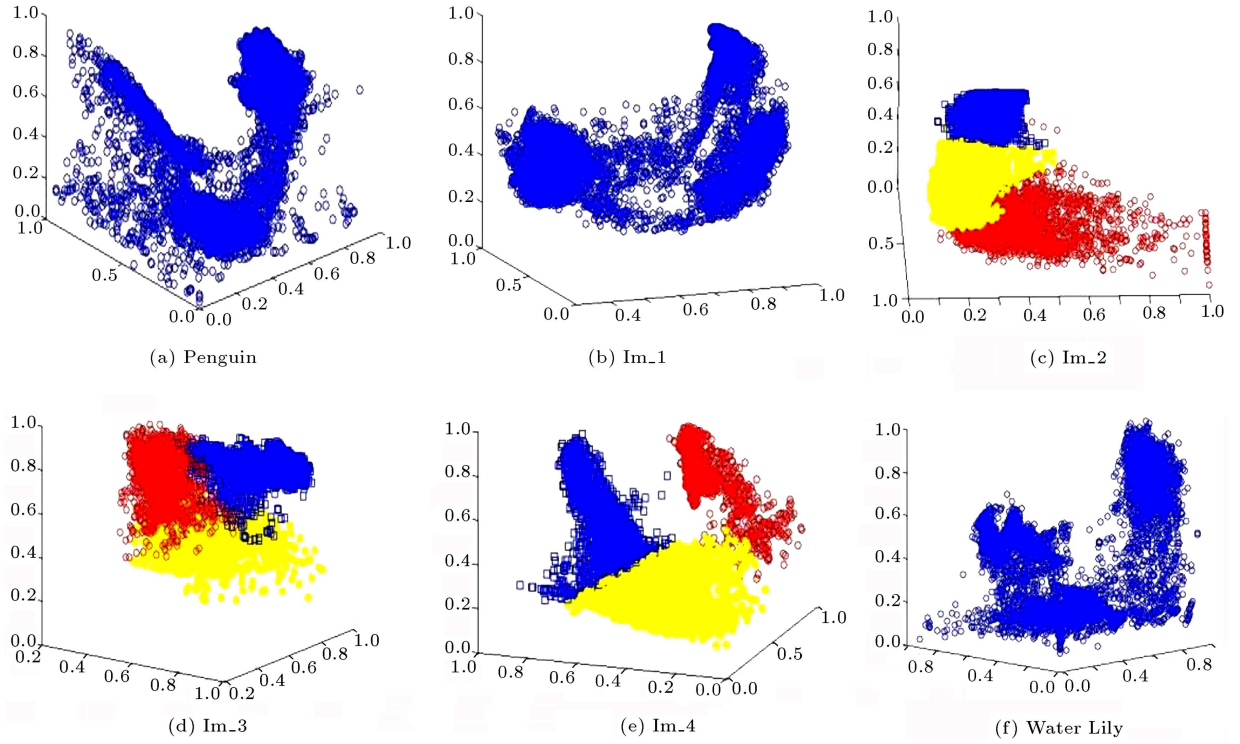


Figure 9. The six image data sets correspond to images in Figure 8.

Table 4. Values of c obtained by each cluster validity indices for image data sets.

Dataset	C^*	V_{PC}	V_{PE}	V_{MPC}	V_{FS}	V_{XB}	V_K	V_{FHV}	V_{PBMF}	V_{PCAES}	V_W	V_{ECAS}
Penguin	3	2	2	2	3	2	2	6	2	2	3	2
Water Lily	3	2	2	2	5	2	2	9	3	3	2	3
Im_1	3	2	2	2	6	3	3	7	3	2	2	3
Im_2	3	2	2	2	3	2	2	2	4	2	2	3
Im_3	3	2	2	2	3	2	2	2	3	2	2	3
Im_4	3	2	2	2	11	2	2	11	4	2	3	3

sets, V_{ECAS} is obviously a superior index in comparison with others.

Experimental Results for Fuzzy Data

In this subsection, to test V'_{ECAS} , we use three fuzzy data sets that are available in literature regarding fuzzy clustering with fuzzy data: Taiwanese tea data set, Example.1 and Example.2. The Taiwanese tea data set is used by Hung and Yang [25] and De Oliveira and Pedrycz [4], and Example.1 and Example.2 data sets are used by Hung and Yang [25]. These data sets are demonstrated in Figure 11. The Taiwanese tea data set is drawn by Hung and Yang [25], regarding the evaluation of 70 kinds of Taiwanese tea, and with respect to four attributes: appearance, tincture, liquid

color and aroma. Taiwanese tea evaluation comes under the subjective judgment of experts at five different quality levels: perfect, good, medium, poor and bad. Finally, Hung and Yang [25] generated 70 triangular fuzzy numbers with five different clusters. Therefore, a cluster validity index must recognize five clusters for the Taiwanese tea data set. On the other hand, in Example.1, there are 20 triangular fuzzy numbers, which are shown in Figure 11b, and intuitively $c = 2$ is a suitable number of clusters for them. Moreover, Example.2 consists of the same 20 fuzzy numbers with an added point that can be regarded as a noise or an outlier [25]. We implemented the FKNC model and V'_{ECAS} on these three data sets in order to test V'_{ECAS} . The optimal number of clusters is found in the maximum value of V'_{ECAS} . The results of the clustering

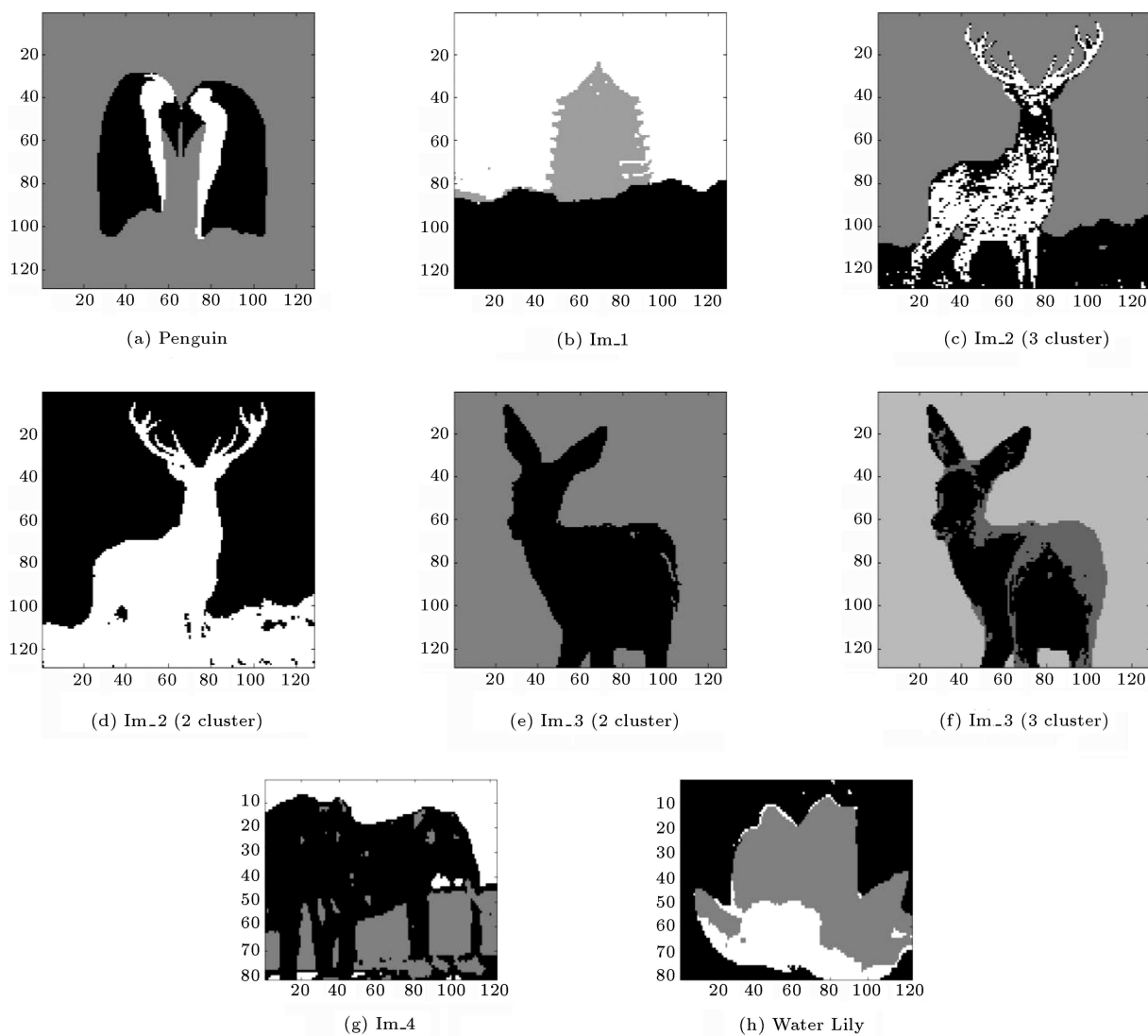


Figure 10. Segmentation results of the fuzzy c -means clustering algorithms.

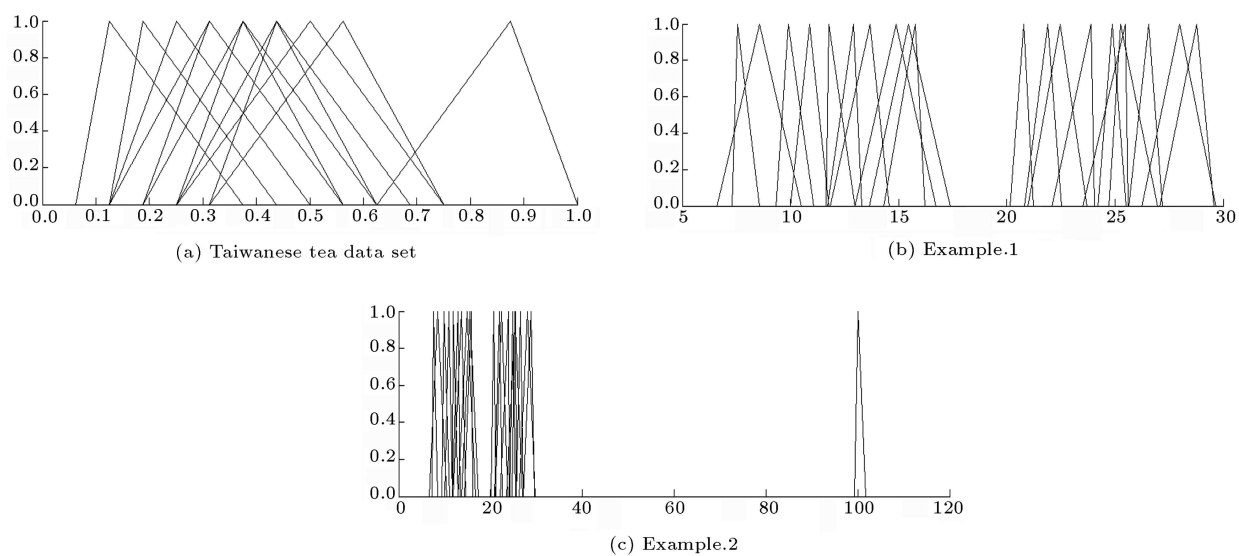


Figure 11. Three triangular fuzzy data sets for fuzzy clustering with fuzzy data.

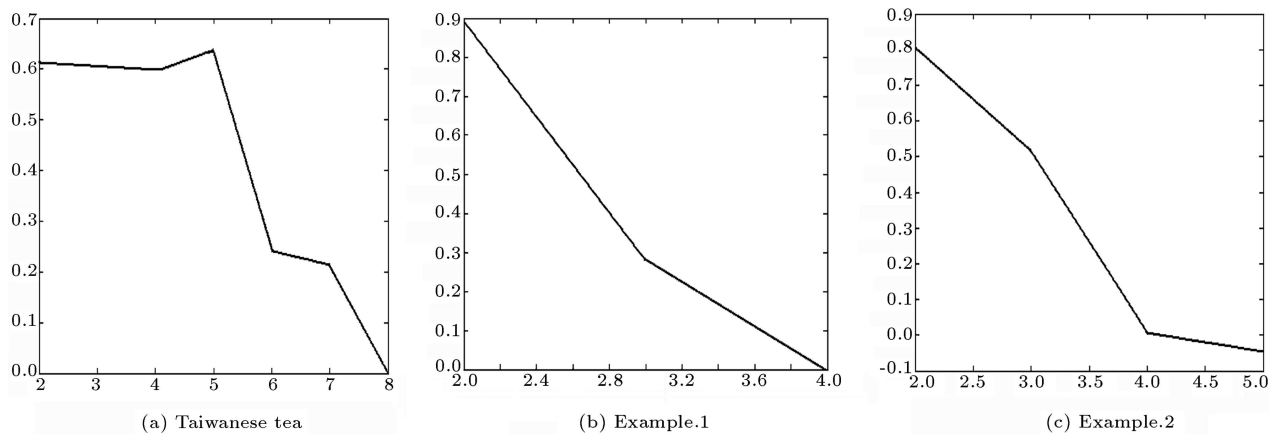


Figure 12. The maximum value of index corresponds to the optimal number of clusters.

of three fuzzy data sets are demonstrated in Figure 12. In fact, V'_{ECAS} in three data sets correctly recognized the optimal number of clusters.

CONCLUSIONS AND FUTURE WORK

In this paper, an exponential cluster validity index, based on compactness and separation measures, has been proposed. Compactness indicates variation of the data within clusters and separation indicates isolation between clusters. The numerical performance of the proposed index was compared with some well known cluster validity indices, using different types of numerical and real data sets. Moreover, it was applied to color image segmentation to show the performance of its validity. Next, the proposed cluster validity index was extended to be capable of validating clustering with fuzzy data. In both proposed indices, it is shown that they have good performance in determining the suitable number of clusters, and are robust in noisy environments.

This paper has some potential future work. In this research, the value of weighting exponent m is assumed to be predefined. Developing a new strategy to select an appropriate m , based on the behavior of the index, needs further work. Moreover, in this research, the norm of the proposed index is Euclidian. Generating a heuristic to select the suitable norm, based on the situation of the data set, needs more investigation.

REFERENCES

1. Höppner, F. et al., *Fuzzy Cluster Analysis: Methods for Classification Data Analysis and Image Recognition*, Chapter 1, Wiley, New York (1999).
2. Dunn, J.C. "A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters", *J. Cybernet.*, **3**, pp. 32-57 (1974).
3. Bezdek, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York (1981).
4. De Oliveira, J.V. and Pedrycz, W., *Advances in Fuzzy Clustering and Its Applications*, Chapter 8, Wiley, New York (2007).
5. Wang, W. and Zhang, Y. "On fuzzy cluster validity indices", *Fuzzy Sets and Systems*, **158**, pp. 2095-2117 (2007).
6. Zhang, Y. et al. "A cluster validity index for fuzzy clustering", *Information Sciences*, **178**, pp. 1205-1218 (2008).
7. Wu, K.L. and Yang, M.S. "A cluster validity index for fuzzy clustering", *Pattern Recognition Lett.*, **26**, pp. 1275-1291 (2005).
8. Kim, D.W. et al. "On cluster validity index for estimation of the optimal number of fuzzy clusters", *Pattern Recognition*, **37**, pp. 2009-2025 (2004).
9. Falasconi, M. et al. "A stability based validity method for fuzzy clustering", *Pattern Recognition*, **43**, pp. 1292-1305 (2010).
10. Trauwaert, E. "On the meaning of Dunn's partition coefficient for fuzzy clusters", *Fuzzy Sets and Systems*, **25**, pp. 217-242 (1988).
11. Bezdek, J.C. "Cluster validity with fuzzy sets", *J. Cybernet.*, **3**, pp. 58-73 (1974).
12. Bezdek, J.C. "Numerical taxonomy with fuzzy sets", *J. Math. Biol.*, **1**, pp. 57-71 (1974).
13. Dave, R.N. "Validating fuzzy partition obtained through c -shells clustering", *Pattern Recognition Lett.*, **17**, pp. 613-623 (1996).
14. Fukuyama, Y. and Sugeno, M. "A new method of choosing the number of clusters for the fuzzy c -means method", in *Proceedings of Fifth Fuzzy Systems Symposium*, pp. 247-250 (1989).
15. Xie, X.L. and Beni, G. "A validity measure for fuzzy clustering", *IEEE Trans. Pattern Anal. Mach. Intell.*, **13**(8), pp. 841-847 (1991).
16. Kwon, S.H. "Cluster validity index for fuzzy clustering", *Electron Lett.*, **34**(22), pp. 2176-2178 (1998).

17. Gath, I. and Geva, A.B. "Unsupervised optimal fuzzy clustering", *IEEE Trans. Pattern Anal. Mach. Intell.*, **11**(7), pp. 773-781 (1989).
18. Pakhira, M.K. et al. "Validity index for crisp and fuzzy clusters", *Pattern Recognition*, **37**, pp. 487-501 (2004).
19. Pal, N.R. and Pal, S.K. "Entropy: a new definition and its applications", *IEEE Trans. Systems Man Cybernet.*, **21**, pp. 1260-1270 (1991).
20. Pal, N.R. and Pal, S.K. "Some properties of the exponential entropy", *Information Sciences*, **66**, pp. 119-137 (1992).
21. Wu, K.L. and Yang, M.S. "Alternative *c*-means clustering algorithms", *Pattern Recognition*, **35**, pp. 2267-2278 (2002).
22. Zimmermann, H.J., *Fuzzy Set Theory and Its Applications*, 4th Ed., Kluwer Academic press, Dordrecht (2001).
23. Yang, M.S. and Ko, C.H. "On a class of fuzzy *c*-numbers clustering procedures for fuzzy data", *Fuzzy Sets and Systems*, **84**, pp. 49-60 (1996).
24. Yang, M.S. and Liu, H.H. "Fuzzy clustering procedures for conical fuzzy vector data", *Fuzzy Sets and Systems*, **106**, pp. 189-200 (1999).
25. Hung, W.L. and Yang, M.S. "Fuzzy clustering on LR-type fuzzy numbers with an application in Taiwanese tea evaluation", *Fuzzy Sets and Systems*, **150**, pp. 561-577 (2005).
26. D'Urso, P. and Giordani, P. "A weighted fuzzy *c*-means clustering model for fuzzy data", *Computational Statistics & Data Analysis*, **50**, pp. 1496-1523 (2006).
27. Bezdek, J.C., *Pattern Recognition in Handbook of Fuzzy Computation*, Chapter F6, IOP Publishing Ltd., Boston, NY (1998).

BIOGRAPHIES

Mohammad Hossein Fazel Zarandi is a Professor at the Department of Industrial Engineering of Amirkabir University of Technology, Tehran, Iran, and a member of the Knowledge-Information Systems Laboratory at the University of Toronto, Ontario, Canada. His main research interests focus on: Intelligent Information Systems, Soft Computing, Computational Intelligence, Fuzzy Sets and Systems, Multi-agent Systems, Networks, Meta-heuristics and Optimization. Professor Fazel Zarandi is the author of numerous books, scientific papers and technical reports in the above areas most of which are accessible through the Web. He has taught several courses in a number of universities throughout Iran and North America in the areas of: Fuzzy Systems Engineering, Decision Support Systems, Management Information Systems, Artificial Intelligence and Expert Systems, Systems Analysis and Design, Scheduling, Neural Networks, Simulations and Production Planning and Control.

Mohammad Reza Faraji received his B.S. in Applied Mathematics from Zanjan University in 2004. He obtained his M.S. degree in the field of sociology-economy systems engineering from the Industrial Engineering Department of Amirkabir University of Technology in 2009.

Masoud Karbasian has a Ph.D. degree in Management. His main research areas include: Data Analysis, Organizational Management and Globalization, the area in which he has also authored several books. His recent research focus is on Data Mining in Uncertain Environments.