PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT

Pattern Recognition 37 (2004) 487–501

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Validity index for crisp and fuzzy clusters

Malay K. Pakhira[a,*], Sanghamitra Bandyopadhyay[b], Ujjwal Maulik[c]

[a] *Department of Computer Science and Technology, Kalyani Government Engineering College, Kalyani, West Bengal 741 235, India*
[b] *Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700 108, India*
[c] *Department of Computer Science and Technology, Kalyani Government Engineering College, Kalyani 741 235, India*

Received 29 April 2002; received in revised form 11 June 2003; accepted 11 June 2003

Abstract

In this article, a cluster validity index and its fuzzification is described, which can provide a measure of goodness of clustering on different partitions of a data set. The maximum value of this index, called the *PBM-index*, across the hierarchy provides the best partitioning. The index is defined as a product of three factors, maximization of which ensures the formation of a small number of compact clusters with large separation between at least two clusters. We have used both the k-means and the expectation maximization algorithms as underlying crisp clustering techniques. For fuzzy clustering, we have utilized the well-known fuzzy c-means algorithm. Results demonstrating the superiority of the *PBM-index* in appropriately determining the number of clusters, as compared to three other well-known measures, the Davies–Bouldin index, Dunn's index and the Xie–Beni index, are provided for several artificial and real-life data sets.

Keywords: Clustering; Expectation maximization algorithm; Fuzzy c-means algorithm; k-Means algorithm; Unsupervised classification; Validity index

1. Introduction

Clustering [1–5] is an unsupervised classification method when the only data available are unlabelled, and no structural information about it is available. In clustering (also known as exploratory data analysis), a set of patterns, usually vectors in a multi-dimensional space, are organized into coherent and contrasted groups, such that patterns in the same group are similar in some sense and patterns in different groups are dissimilar in the same sense. The purpose of any clustering technique is to evolve a partition matrix $U(X)$ of a given data set X (consisting of, say, n patterns, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$) so as to find a number, say K, of clusters $(C_1, C_2, \ldots, C_K)$. The partition matrix $U(X)$ of size $K \times n$ may be represented as $U = [u_{kj}]$, $k = 1, \ldots, K$ and $j = 1, \ldots, n$, where u_{kj} is the membership of pattern $\mathbf{x}_j$ to clusters C_k. In crisp partitioning of the data, the following condition holds: $u_{kj} = 1$ if $\mathbf{x}_j \in C_k$, otherwise $u_{kj} = 0$. The purpose is to classify data set X such that

$$C_i \neq \emptyset \quad \text{for } i = 1, \ldots, K,$$

$$C_i \cap C_j = \emptyset \quad \text{for } i = 1, \ldots, K, \ \ j = 1, \ldots, K \text{ and } i \neq j$$

and

$$\bigcup_{i=1}^{K} C_i = X.$$

In the case of fuzzy clustering, the purpose is to evolve an appropriate partition matrix $U = [u_{kj}]_{K \times n}$, where $u_{kj} \in [0, 1]$, such that u_{kj} denotes the grade of membership of the jth element to the kth cluster. In fuzzy partitioning of the data, the following conditions hold:

$$0 < \sum_{j=1}^{n} u_{kj} < n \quad \text{for } k = 1, \ldots, K,$$

* Corresponding author. Tel.: +91-33-2582-6680; fax: +91-33-2582-1309.
E-mail addresses: mkp@kucse.wb.nic.in (M.K. Pakhira), sanghami@isical.ac.in (S. Bandyopadhyay), ujjwal_maulik@kucse.wb.nic.in (U. Maulik).

0031-3203/$30.00 © 2003 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2003.06.005

$$\sum_{k=1}^{K} u_{kj} = 1 \quad \text{for } j = 1, \ldots, n$$

and

$$\sum_{k=1}^{K} \sum_{j=1}^{n} u_{kj} = n.$$

The k-means algorithm [5] is one of the very well-known partitional clustering method that produces the minimum-squared-error partitions. When the number of clusters is known a priori, the k-means algorithm optimizes the distance criterion either by minimizing the within cluster spread, or by maximizing the inter cluster separation. The expectation maximization (EM) algorithm [6] is considered to be an appropriate optimization algorithm for constructing proper statistical models of data. It provides a probabilistic clustering where each data element has a certain probability of being a member of any cluster. Unlike the k-means algorithm, it does not depend on any distance measure, and accommodates categorical and continuous data in a superior manner.

The two fundamental questions that need to be addressed in any typical clustering scenario are: (i) how many clusters are actually present in the data, and (ii) how real or good is the clustering itself. That is, whatever may be the clustering technique, one has to determine the number of clusters and also the validity of the clusters formed [7]. The measure of validity of the clusters should be such that it will be able to impose an ordering of the clusters in terms of its goodness. In other words, if $U_1, U_2, \ldots, U_m$ be m partitions of X, and the corresponding values of a validity measure be $V_1, V_2, \ldots, V_m$, then $V_{k1} \geqslant V_{k2} \geqslant \cdots \geqslant V_{km}$, $\forall ki \in \{1, 2, \ldots, m\}$, $i = 1, 2, \ldots, m$ will indicate that $U_{k1} \uparrow U_{k2} \uparrow \ldots \uparrow U_{km}$. Here '$U_i \uparrow U_j$' indicates that partition U_i is a better clustering than U_j. Note that a validity measure may also define a increasing sequence instead of an decreasing sequence of $V_{k1}, \ldots, V_{km}$.

In this paper, we describe an index, called *PBM-index*, which can be used to associate a measure with different partitions of a data set; the maximum value of which indicates the appropriate partitioning. Therefore, if the number of clusters, K, is varied within some range, and an underlying clustering technique is used to partition the data, then the value of K corresponding to the maximum value of *PBM-index* will indicate the correct number of clusters present in the data. The effectiveness of this index, for determining the appropriate number of clusters, is demonstrated for four artificial and two real-life data sets.

Other well-known cluster validity indices, available in the literature, are the Davies–Bouldin (DB) index [8], Dunn's Index [9] (both for hard clusters primarily), and the Xie–Beni (XB) index [10] (for fuzzy clusters). Davies–Bouldin index is a function of ratio of the sum of within cluster scatter to between cluster scatter. Dunn's index is a ratio of within cluster and between cluster separations. The Xie–Beni index is a ratio of the fuzzy within cluster sum of squared distances to the product of the number of elements and the minimum between cluster separation. In order to demonstrate the effectiveness of the *PBM-index*, we compare its performance with the other indices for evolving the proper number of clusters for four artificial and four real-life data sets. For this purpose, both the k-means [5] and EM algorithms [6] have been used as the underlying clustering strategy.

In a part of the investigation, a fuzzified version of the *PBM-index* is proposed. Again, the maximum value of the fuzzy index over different fuzzy partitions of the data indicates the appropriate clustering. For this purpose, the fuzzy c-means (FCM) algorithm [11] is used as the underlying clustering technique. FCM uses the principles of fuzzy sets to partition a data into a fixed number, c, of clusters; thereby providing the appropriate $c \times n$ partition matrix. The performance of the fuzzy *PBM* index is compared with the XB-index in determining the proper number of fuzzy clusters for different data sets.

2. *PBM-index*: a measure for cluster validity

In this section, we first define the *PBM-index*. This is followed by an explanation of the interaction among the different components of the index so that it can approximately indicate the proper partitioning of the data.

2.1. Definition

The *PBM-index* is defined as follows:

$$PBM(K) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K \right)^2, \tag{1}$$

where K is the number of clusters. Here,

$$E_K = \sum_{k=1}^{K} E_k, \tag{2}$$

such that

$$E_k = \sum_{j=1}^{n} u_{kj} \|\mathbf{x}_j - \mathbf{z}_k\| \tag{3}$$

and

$$D_K = \max_{i,j=1}^{K} \|\mathbf{z}_i - \mathbf{z}_j\|. \tag{4}$$

n is the total number of points in the data set, $U(X) = [u_{kj}]_{K \times n}$ is a partition matrix for the data and z_k is the center of the kth cluster. The objective is to maximize this index in order to obtain the actual number of clusters.

2.2. Explanation

As formulated in Eq. (1), *PBM-index* is a composition of three factors, namely, $1/K$, E_1/E_K and D_K. The first

factor decreases as K increases and this, therefore, reduces the index value. The second factor consists of the ratio of E_1, which is constant for a given data set, and E_K, which decreases with increase in K. Hence, *PBM-index* increases as E_K decreases. This, in turn, indicates that formation of more number of clusters, which are compact in nature, would be encouraged. Finally, the third factor, D_K, measuring the maximum separation between a pair of clusters, increases with the value of K. Note that this value is bounded up by the maximum separation between two points in the data set. Thus, the three factors are found to compete with each other critically. It may appear that consideration of D_K, as defined in Eq. (4), may lead to undesirable clustering where the two maximally separated clusters (say A and B) have been found, while another cluster C (which can be divided into more than one clusters) may lie in between. However, in such situations, considering only D_K, C cannot be divided into its component clusters. We show that in such cases, the other two factors become dominant and are able to provide the requisite clustering in the following paragraphs.

The first factor indicates the divisibility of a K cluster system. This reduces with increase in K. The second factor includes the sum of intra cluster distances for the complete data set taken as a single cluster and that for the K-cluster system. Here, the denominator is decreasing with increase in K and the numerator is fixed. The fixed numerator is used only to eliminate chances that the second factor becomes very small. This factor is a measure of the compactness of a K cluster system, and we want to increase it. The third factor D_K is the maximum inter cluster separation in a K cluster system. This factor signifies between cluster separation, and we want to increase it. So while the first factor is decreasing, the other two are increasing with increase in K. This is justified because we want to keep the number of clusters as small as possible while increase the compactness and separation as much as possible.

The use of D_K, as the measure of separation, requires further elaboration. Instead of using the maximum separation between two clusters, several other alternatives could have been used. For example, if D_K was the sum of pairwise inter cluster distances in a K-cluster structure, then it would increase geometrically with K, and thus we would face a situation when there might be no terminating condition. This might lead to the formation of maximum possible number of clusters equal to the number of elements in the data set. If D_K was the average inter cluster distance then it would decrease at each step with K, instead of being increased. So, this will only leave us with the minimum possible number of clusters. The minimum distance between two clusters may be another choice for D_K. However, this measure would also decrease significantly with increase in the number of clusters. So this would lead to a structure where the loosely connected sub-structures remain as they were, where in fact a separation was expected. Thus maximum separability may not be attained.

In contrast, if we consider the maximum inter cluster separation then we see that this tends to increase significantly until we reach the maximum separation among compact clusters and then it becomes almost constant. The upper bound of this value, which is equal to the maximum separation between any two points, is only attainable when we have two extreme data elements as two single element clusters. But the terminating condition is reached well before this situation. This is the reason that we try to improve the maximum distance between two maximally separate clusters. Since the second and third factors play important role in increasing the index by improving compactness and separation, it seems that revealing the maximum separation attainable at each stage provides sufficient information.

If there is any intermediate divisible cluster(s) between the extreme ones, this fact is taken into account by the number of clusters and the compactness factors. It is seen that when division is possible in the intermediate cluster, the second factor overrides the effect of the first one, and the reverse is true for indivisible cluster. In order to show the above fact analytically, we consider spherically approximated clusters. We assume that each cluster may be approximated by a hyper-sphere having uniform distribution of elements. If a d dimensional data set is considered and $r_1, r_2, \ldots, r_d$ be radii of a cluster along these directions, then $r = (r_1 + r_2 + \cdots + r_d)/d$ is considered to be the radius of the approximated cluster. Compactness of an individual cluster can be measured by the inverse of its sum of within cluster Euclidean distances. Thus, if we have a spherically approximated cluster with radius r and n number of elements, then its compactness is the inverse of $nr/2$.

If such a cluster of radius r having n number of elements be divisible into two equal halves with radius $r/2$ and $n/2$ number of elements in each, then its compactness will be approximately doubled upon division. If we try to divide a compact cluster into two equal halves, then its compactness will be increased by a factor less than two.

Let E_K denote the sum of within cluster distances of a K-cluster configuration, and P_K denote the corresponding *PBM-index* value. Let us assume a K cluster configuration where $K-1$ number of clusters are of radius r having n elements in each, and one intermediate cluster is of radius $2r$ with $2n$ elements. The sum of within cluster Euclidean distances for this configuration is

$$E_K = (K-1)\frac{nr}{2} + 2nr. \tag{5}$$

For this configuration, the larger cluster will be the natural candidate for division at the next step. Let us assume again that division will produce two clusters of equal sizes. Now two cases may arise. The original cluster may have a clear tendency for division or it may be a compact one. In the former case we have

$$E_{K+1} = (K-1)\frac{nr}{2} + nr \tag{6}$$

thus,

$$\frac{E_K}{E_{K+1}} = \frac{K+3}{K+1} \tag{7}$$

and

$$\frac{P_K}{P_{K+1}} = \frac{(K+1)(K+1)}{K(K+3)} = \frac{K^2+2K+1}{K^2+3K}, \tag{8}$$

which is less than 1 for all $K > 1$, i.e., a division is suggested.

In the latter case after division, the radius of each of the resultant clusters will be $(r+(d-1)2r)/d = ((2d-1)r)/d$. So, we have,

$$E_{K+1} = \left[(K-1) + 2\,\frac{(2d-1)}{d}\right]\frac{nr}{2} \tag{9}$$

thus,

$$\frac{E_K}{E_{K+1}} = \frac{K+3}{(K-1)+2\,(2d-1)/d} \tag{10}$$

and

$$\frac{P_K}{P_{K+1}} = \frac{(K+1)((K-1)+2\,(2d-1)/d)}{K(K+3)}$$

$$= \frac{(K^2-1)+2(K+1)(2d-1)/d}{K^2+3K}, \tag{11}$$

which is greater than 1 for all K (and $d > 1$), i.e., division is not suggested. These observation are also verified by our experimental results.

So, it is seen that, at small values of K, the second and third factors play important role in revealing the maximum attainable separation and getting compact clusters. As K grows, the effect of these two factors is overcome by the first factor. If, in any case, the maximum separation is reached before the desired compactness, then the first and the second factors interact to produce the desired compactness.

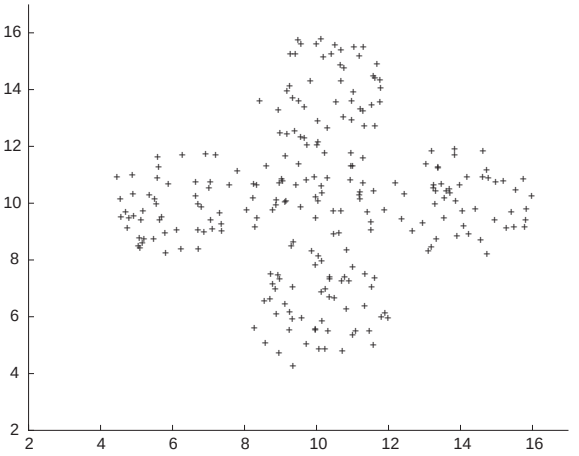

Fig. 1. Circular_5_2.

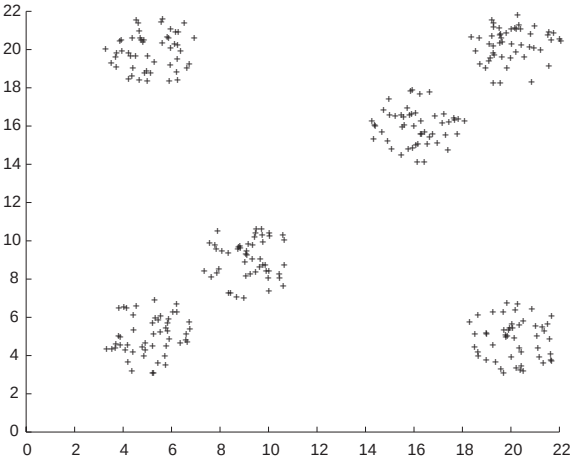

Fig. 2. Circular_6_2.

3. Experimental results

Four artificial and four real-life data sets are considered for experiments. They are described first. This is followed by a demonstration of the variation of *PBM-index* with the number of clusters, when the *k*-means algorithm and the EM algorithm are used as underlying clustering mechanisms. Finally, a comparison of *PBM-index* with the Davies–Bouldin (DB) index , Dunn's index and the Xie–Beni (XB) index is made in terms of the number of clusters and the clustering obtained for the above-mentioned data sets.

3.1. Data sets

The four artificial data sets are called *Circular_5_2*, *Circular_6_2*, *Elliptical_10_2* and *Spherical_4_3*. The names imply the structure of the classes, concatenated with the number of clusters actually present in the data and the number of dimensions. For example, for the *Circular_5_2* data the clusters are circular in nature, there are five clusters and the dimension is 2. As can be seen, the number of clusters range from four to ten. The data sets *Circular_5_2*, *Circular_6_2*, *Elliptical_10_2* and *Spherical_4_3* are demonstrated in Figs. 1–4, respectively.

The four real-life data sets are *Iris*, *Crude_oil*, *Cancer* and *Kalazaar*.

Iris data: This data represents different categories of irises having four feature values. The four feature values represent the sepal length, sepal width, petal length and the petal width in centimeters [12]. It has three classes Setosa, Versicolor and Virginica, with 50 samples per class. It is known that two classes Versicolor and Virginica have some amount of overlap while the class Setosa is linearly separable from the other two. Most techniques reported

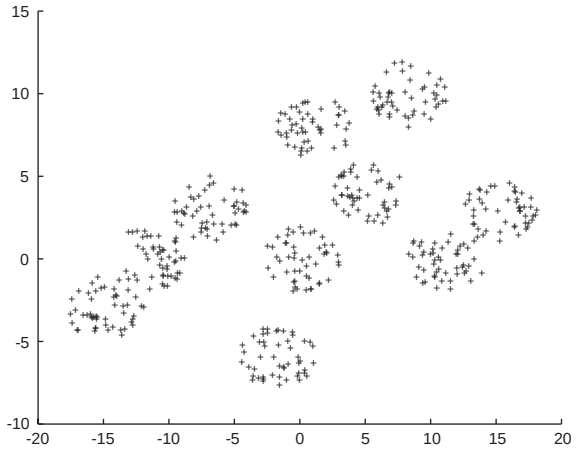

Fig. 3. Elliptical_10_2.

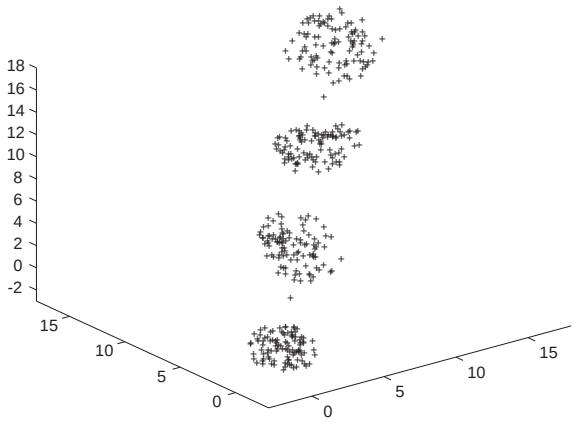

Fig. 4. Spherical_4_3.

in the literature usually provide two clusters for this data [13,14].

Crude Oil data: This overlapping data [15] has 56 data points, five features. The data set is known to have three classes.

Cancer data: The *Cancer* data is the Wisconsin Breast Cancer data set available at [http://www.ics.uci.edu/mlearn/MLRepository.html]. Each pattern has nine features corresponding to *clump thickness*, *cell size uniformity*, *cell shape uniformity*, *marginal adhesion*, *single epithelial cell size*, *bare nuclei*, *bland chromatin*, *normal nucleoli* and *mitoses*. There are two categories in the data: malignant and benign. The two classes are known to be linearly inseparable. There are a total of 683 points in the data set.

Kalazaar data: The *Kalaazar* data [16] consists of 68 patterns in four dimensions. There are two classes: diseased and normal/cured, and four input features/symptoms. These symptoms are the measurements of blood urea (mg%), serum creatinine (mg%), urinary creatinine (mg%) and creatinine clearance (ml/min).

3.2. Results

Several runs of the k-means algorithm are executed for a fixed value of K, and the clustering corresponding to the run that provides the minimum value of the sum of within cluster distances is assumed to be appropriate (for which values of *PBM-index* and other indices are calculated). This is done, since it is known that the k-means algorithm often gets stuck at suboptimal configurations depending on the choice of initial cluster centers. Several re-initialization are therefore necessary to overcome this problem. The value of K is varied in the range [2,40]. Although, the maximum value of K, i.e., K_{max}, is taken to be 40 for the present investigation, it should vary for different data sets. Generally, the maximum number of clusters that can be present in a data set having n elements should not exceed $\sqrt{n}$. This value is considered as a rule of thumb in clustering literature [11]. Note that, as $K \rightarrow n$, the term $E_K \rightarrow 0$; thereby *PBM-index* grows to a very high value. On the basis of this observation, $\sqrt{n}$ can be recommended as a safe measure for K_{max}.

The variation of the *PBM-index* with the number of clusters is shown in Figs. 5–10 for the above-mentioned data sets. As expected, both Figs. 5 and 6 (corresponding to *Circular_5_2* and *Circular_6_2* data sets) show that the value of *PBM-index* peaks at cluster numbers 5 and 6, respectively, indicating that these are the correct number of clusters for the corresponding data sets. Similarly, from Fig. 7, it is found that for *Elliptical_10_2*, the maximum value of *PBM-index* is obtained at $K = 10$, while for *Spherical_4_3* this happens at $K = 4$ (see Fig. 8). It may be verified from Figs. 1–4 that *PBM-index* actually attains its maximum values when the number of clusters is equal to that present in the respective data sets in all the cases.

It is evident from Figs. 9 and 10 that for *Iris* and *Crude_Oil*, clear maximas exist at $K = 3$ (with values 24.886 and 457.781, respectively). Both these data sets are known to contain three classes of patterns [12,15]. From Figs. 11 and 12, it is seen that for the *Cancer* and *Kalazaar* data, maximas exist at $K = 2$ (with value 145.181) and $K = 3$ (with value 910.747), respectively. These data sets are known to have two clusters each. Thus for the *Kalazaar* data *PBM*-index fails to detect the correct number of clusters when k-means algorithm is used as the underlying clustering technique. In Section 3.3, it is shown that the other indices are also unable to detect the correct number of clusters for this data set.

The clustered *Iris* data set, along with the corresponding centers, is shown in Fig. 15 for two features, namely, petal length and sepal width. An interesting observation in this regard is that although we know that the *Iris* data set has three physical classes, very few automatic clustering techniques, reported in the literature, can actually come up

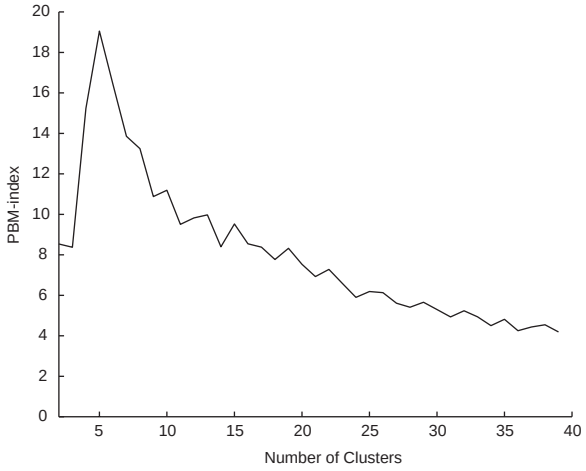

Fig. 5. Variation of the *PBM-index* with the number of clusters for circular_5_2.

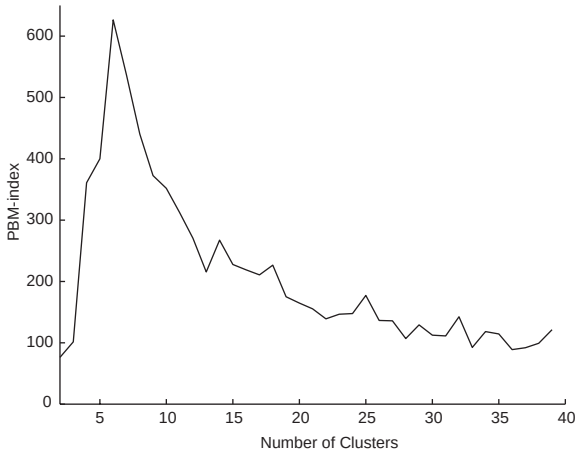

Fig. 6. Variation of the *PBM-index* with the number of clusters for Circular_6_2.

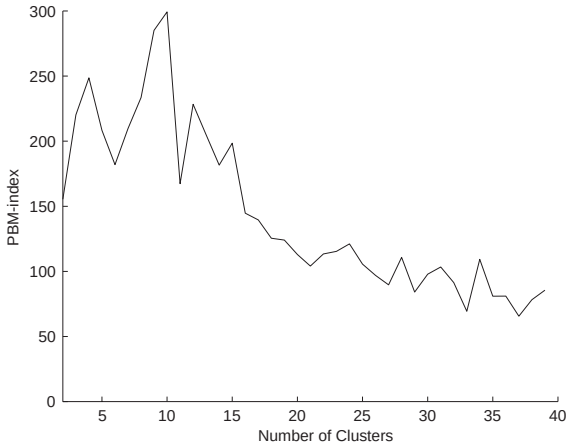

Fig. 7. Variation of the *PBM-index* with the number of clusters for Elliptical_10_2.

with three clusters for this data. In most of the situations, the number of clusters found for this data is equal to two [13,14].

3.3. Comparison with other indices

The Davies–Bouldin index: This index is a function of the ratio of the sum of *within-cluster scatter* to *between-cluster separation*. The scatter within the ith cluster is computed as

$$S_{i,q} = \left(\frac{1}{|C_i|} \sum_{x \in C_i} \{ \|x - z_i\|_2^q \} \right)^{1/q} \tag{12}$$

and the distance between cluster C_i and C_j is defined as

$$d_{ij,t} = \left\{ \sum_{s=1}^{p} |z_{is} - z_{js}|^t \right\}^{1/t} = \|z_i - z_j\|_t. \tag{13}$$

$S_{i,q}$ is the qth root of the qth moment of the points in cluster i with respect to their mean, and is a measure of the dispersion of the points in cluster i. Specifically, $S_{i,1}$, used in this article, is the average Euclidean distance of the vectors in class i to the centroid of class i. $d_{ij,t}$ is the Minkowski distance of order t between the centroids that characterize clusters i and j. Subsequently we compute

$$R_{i,qt} = \max_{j, j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\}. \tag{14}$$

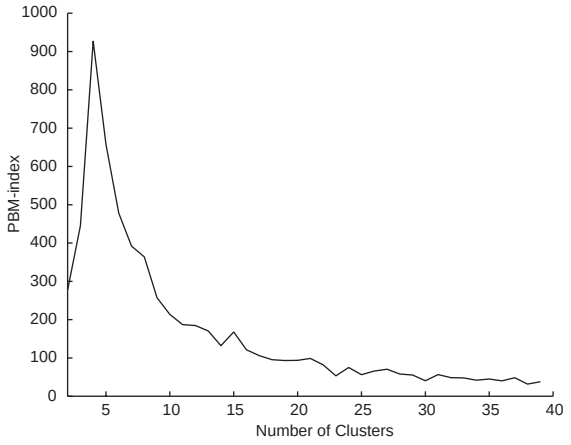

Fig. 8. Variation of the *PBM-index* with the number of clusters for Spherical_4_3.

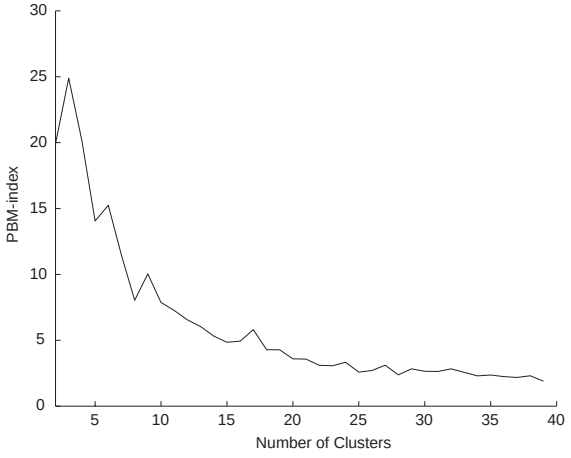

Fig. 9. Variation of the *PBM-index* with the number of clusters for *Iris*.

The Davies–Bouldin (DB) index is then defined as

$$DB = \frac{1}{K}\sum_{i=1}^{K} R_{i,qt}. \tag{15}$$

The objective is to minimize the DB index for achieving proper clustering.

Dunn's index: Let S and T be two nonempty subsets in R^N. Then the diameter Δ of S and set distance δ between S and T are

$$\Delta(S) = \max_{x,y\in S}\{d(x,y)\} \tag{16}$$

and

$$\delta(S,T) = \min_{x\in S, y\in T}\{d(x,y)\}, \tag{17}$$

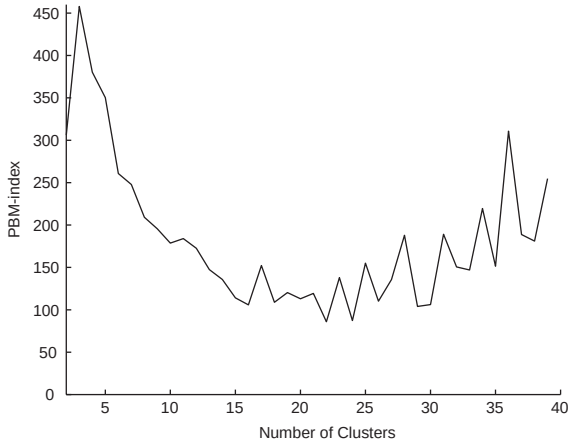

Fig. 10. Variation of the *PBM-index* with the number of clusters for *Crude_Oil*.

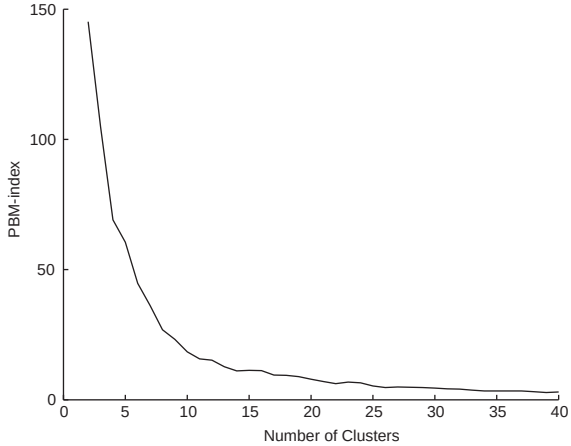

Fig. 11. Variation of the *PBM-index* with the number of clusters for *Cancer*.

where $d(x,y)$ is the distance between points x and y. For any partition Dunn defined the following index:

$$v_D = \min_{1\leqslant i\leqslant K}\left\{\min_{1\leqslant j\leqslant K, j\neq i}\left\{\frac{\delta(C_i,C_j)}{\max_{1\leqslant k\leqslant K}\Delta(C_k)}\right\}\right\}. \tag{18}$$

Larger values of v_D correspond to good clusters, and the number of clusters that maximizes v_D is taken as the optimal number of clusters.

In Ref. [17] generalized Dunn's index is presented. This general form is

$$v_{D_{ij}} = \min_{1\leqslant s\leqslant K}\left\{\min_{1\leqslant t\leqslant K, t\neq s}\left\{\frac{\delta_i(C_s,C_t)}{\max_{1\leqslant k\leqslant K}\Delta_j(C_k)}\right\}\right\}. \tag{19}$$

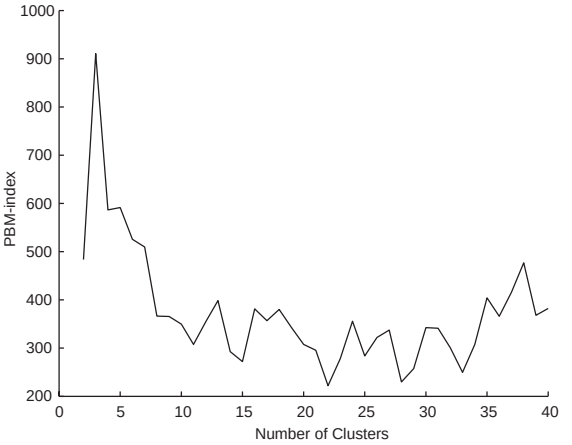

Fig. 12. Variation of the *PBM-index* with the number of clusters for *Kalazaar*.

It may have 18 different forms depending on the functional forms used to select δ_i and Δ_j. These indices are denoted by $v_{D_{ij}}$ for $1 \leqslant i \leqslant 6$ and $1 \leqslant j \leqslant 3$. In our experiments we used the index form $v_{D_{53}}$. The functional forms for different δ_i s and Δ_j s may be found in Ref. [17].

The Xie–Beni index: This is a fuzzy clustering index. We mention it here briefly. The generalized version of this index [11] is given by

$$S = \frac{J_m}{n * (d_{min})^2}, \tag{20}$$

where J_m is the sum of squared errors objective function for fuzzy clustering [11] and is given by

$$J_m(U,Z) = \sum_{j=1}^{n}\sum_{k=1}^{K} (u_{kj})^m \|\mathbf{x}_j - \mathbf{z}_k\|^2, \tag{21}$$

where $1 \leqslant m < \infty$. Here, U is a partition matrix, $U = [u_{kj}] \in R^{Kn}$. u_{kj} is interpreted to be the grade of membership of $\mathbf{x}_j$ in the kth cluster. Z is the set of cluster centers, i.e., $Z = \{\mathbf{z}_k\} \in R^n$. The relations used for computing U and Z are same as those used in Ref. [11].

d_{min} is the minimum inter cluster distance. The minimum value of S in the hierarchy corresponds to the number of clusters present in the data set.

Table 1 provides the actual number of clusters present in the above-mentioned data sets, and the number of clusters indicated by the indices, namely, the DB index, Dunn's index, the XB-index and *PBM-index* (which attains its maximum value) using the k-means algorithm. As can be seen from the table, while *PBM-index* is able to provide the appropriate number of clusters for all the data sets except the *Kalazaar* data, the DB-index fails for *Elliptical_0_2*, *Iris*, *Crude_Oil* and *Kalazaar* data sets, Dunn's index fails for *Elliptical_10_2*, *Iris* and *Kalazaar* data sets and the XB-index fails for all other data sets except *Circular_5_2*, *Spherical_4_3* and *Cancer*. In Ref. [17] it is shown that for the *Iris* data, Dunn's index is able to find either two or three cluster solutions. But out of total 18 possible forms of the generalized Dunn's index, only two can detect three cluster solution and the remaining 16 (including $v_{D_{53}}$) can detect 2 cluster solution. In Ref. [14] a scale space-based method is described, where also it is shown that for the *Iris* data the two and three cluster solutions are sustained for comparable scale intervals with two cluster being for a marginally longer scale interval.

Table 2 provides a comparison of the DB-index, Dunn's index, the XB-index and PBM-index, when the EM algorithm is used for clustering. From this table, it can be seen that the DB-index, Dunn's index and the XB-index fail for *Elliptical_10_2*, *Iris* and *Crude_Oil* data sets. In addition the Dunn's index also fails for the *Kalazaar* data set. It is worthwhile to mention here that for the *Kalazaar* data, the PBM-index, DB-index and XB-index are able to detect the correct number of clusters. All these indices failed to detect this by the k-means algorithm. This result is expected due to more perfect clustering by the EM algorithm.

Tables 3–6 present the variation of the DB-index, Dunn's index, the XB-index and *PBM-index* with the number of clusters in the range [2,15] for all the data sets when the k-means algorithm is used for clustering. The optimum values of the indices are presented in boldface in the tables. Note that, from Table 6, it is seen that for the *Kalazaar* data, the minimum value of the DB-index (0.838) is obtained for $K = 15$ and the second lower value (0.913) for $K = 13$. However, as mentioned in Ref. [11], the maximum possible number of clusters that one should consider for a data set having n elements is $\sqrt{n}$ which is between 8 and 9 for this data set. Therefore, the index value (0.918) for $K = 4$ is highlighted. Table 7 shows the variation of the PBM-index only with the number of clusters for the data sets when the EM method is used as the underlying clustering method. Since *Iris* is a widely used data set, we compare the cluster centers obtained for this data when the k-means method and the EM method are used for clustering in conjunction with the *PBM-index* in Table 8. As can be seen, the obtained cluster centers are quite close to the actual ones.

The clustered *Elliptical_10_2* data, for which *PBM-index* is maximized (when K was equal to 10) is shown in Fig. 13. Similar is the situation when the EM algorithm is used to cluster the *Elliptical_10_2* data. A better cluster structure compared to that obtained with the k-means method is observed. These clusters are shown in Fig. 14. Fig. 15 shows the almost correctly clustered *Iris* data (in two dimensions) for which *PBM-index* attained the maximum value. Likewise if the EM algorithm is used as the underlying clustering technique, *PBM-index* attains its maximum value for a three cluster solution; but in this case the clustering is more perfect. These are shown in Fig. 16. It is seen that

Table 1
Comparison of the number of clusters identified by the DB index, Dunn's index, the XB-index and *PBM-index* using the *k*-means algorithm

Data set	Actual number of clusters	Number of clusters obtained			
		DB-index	Dunn's index	XB-index	PBM-index
Circular_5_2	5	5	5	5	5
Circular_6_2	6	6	6	4	6
Spherical_4_3	4	4	4	4	4
Elliptical_10_2	10	8	2	4	10
Iris	3	2	2	2	3
Crude_Oil	3	2	3	2	3
Cancer	2	2	2	2	2
Kalazaar	2	4	3	4	3

Table 2
Comparison of the number of clusters identified by the DB-index, Dunn's index, the XB-index and *PBM-index* using the EM algorithm

Data set	Actual number of clusters	Number of clusters obtained			
		DB-index	Dunn's index	XB-index	PBM-index
Circular_5_2	5	5	5	5	5
Circular_6_2	6	6	6	6	6
Spherical_4_3	4	4	4	4	4
Elliptical_10_2	10	8	2	2	10
Iris	3	2	2	2	3
Crude_Oil	3	2	2	2	3
Cancer	2	2	2	2	2
Kalazaar	2	2	4	2	2

Table 3
Values of the DB-index, Dunn's index, the XB-index and *PBM-index* in the range of $K = 2, \ldots, 15$ for different data sets using the *k*-means algorithm (Entries in bold face indicate the optimal values for respective indices)

Number of clusters	Data set							
	Circular_5_2				*Circular_6_2*			
	DB	DUNN	XB	PBM	DB	DUNN	XB	PBM
2	1.099	0.949	0.415	8.535	1.006	1.105	0.286	73.922
3	0.784	1.007	0.233	9.428	0.633	1.038	0.152	96.798
4	0.679	1.364	0.442	15.461	0.361	2.158	**0.043**	360.856
5	**0.671**	**1.409**	**0.143**	**19.143**	0.373	0.951	0.140	412.415
6	0.778	0.913	0.388	15.408	**0.355**	**2.242**	0.055	**626.419**
7	0.831	0.949	0.319	14.408	0.544	0.749	0.555	508.040
8	0.862	0.893	0.366	14.266	0.725	0.751	0.542	403.234
9	0.917	0.811	0.316	10.886	0.678	0.669	0.516	321.442
10	0.930	0.808	0.375	11.197	0.775	0.721	0.533	343.217
11	0.883	0.811	0.336	9.310	0.629	0.661	1.247	296.727
12	0.938	0.803	0.357	10.410	0.862	0.535	0.437	261.651
13	0.816	0.902	0.461	10.311	0.983	0.479	0.494	268.898
14	0.815	0.692	0.556	9.499	0.807	0.566	0.524	234.571
15	0.766	0.698	0.572	9.463	0.820	0.443	0.459	200.894

Table 4
Values of the DB-index, Dunn's index, the XB-index and *PBM-index* in the range of $K = 2, \ldots, 15$ for different data sets using the *k*-means algorithm (Entries in bold face indicate the optimal values for respective indices)

Number of clusters	Data set							
	Spherical_4_3				*Elliptical_10_2*			
	DB	DUNN	XB	PBM	DB	DUNN	XB	PBM
2	0.534	1.852	0.756	274.903	0.753	**1.795**	0.160	155.421
3	0.511	1.042	0.081	395.571	0.882	1.128	0.217	220.218
4	**0.441**	**1.939**	**0.052**	**941.876**	0.709	1.202	**0.148**	248.635
5	0.755	0.589	0.678	667.766	0.660	0.684	0.426	208.356
6	0.787	0.484	1.191	485.160	0.650	0.812	0.242	181.965
7	0.824	0.482	0.781	368.502	0.736	0.741	0.221	209.766
8	1.032	0.473	1.045	314.586	**0.565**	0.741	0.159	233.508
9	1.042	0.476	0.895	240.224	0.589	0.673	0.206	285.089
10	0.936	0.421	0.914	227.246	0.666	0.691	0.728	**299.288**
11	0.916	0.433	1.132	193.554	0.676	0.494	0.815	167.370
12	0.934	0.394	1.027	155.672	0.630	0.863	0.509	228.336
13	0.974	0.364	0.935	149.530	0.693	0.638	0.499	204.845
14	0.870	0.346	1.876	141.732	0.685	0.686	0.724	181.617
15	1.077	0.457	1.382	140.119	0.706	0.205	1.091	198.430

Table 5
Values of the DB-index, Dunn's index, the XB-index and *PBM-index* in the range of $K = 2, \ldots, 15$ for different data sets using the *k*-means algorithm (Entries in bold face indicate the optimal values for respective indices)

Number of clusters	Data set							
	Iris				*Crude_Oil*			
	DB	DUNN	XB	PBM	DB	DUNN	XB	PBM
2	**0.404**	**1.902**	**0.066**	19.923	**0.647**	1.303	**0.122**	306.006
3	0.666	1.282	0.164	**24.886**	0.706	**1.319**	0.149	**457.781**
4	0.776	0.946	0.260	20.139	0.794	0.888	0.351	380.317
5	0.902	0.582	0.430	14.060	0.859	0.678	0.440	350.305
6	0.878	0.545	0.509	15.246	0.964	0.755	0.598	260.788
7	1.017	0.638	0.476	11.449	0.935	0.429	0.522	247.797
8	0.981	0.451	0.749	8.041	0.940	0.566	0.778	209.125
9	1.002	0.427	0.492	10.039	0.852	0.345	0.524	209.125
10	1.032	0.510	0.908	7.877	0.978	0.456	0.515	195.587
11	0.964	0.486	0.791	7.262	0.909	0.350	1.493	178.687
12	0.982	0.469	0.815	6.552	0.839	0.446	1.872	184.027
13	1.018	0.356	0.440	6.041	0.914	0.318	1.160	172.520
14	1.128	0.377	0.975	5.317	0.875	0.304	1.354	135.771
15	1.054	0.449	0.988	4.852	0.938	0.251	1.268	114.075

the percentages of correct clustering, for the *Iris* data, using the *k*-means and the EM algorithm are 93.33 and 98.67, respectively.

It is observed that for clustering *Iris* using the EM algorithm, the number of clusters detected by both Dunn's index and the XB-index is 2. The values of the Dunn's index for $K = 2$ is 1.985 and for $K = 3$ is 1.097, and that of the XB-index for $K = 2$ is 0.065 and for $K = 3$ is 0.221.

3.4. Fuzzification of the index

In this section, we propose a fuzzy version of the *PBM*-index denoted by *PBMF*. The fuzzy index is obtained by incorporating fuzzy distances. It is defined as follows:

$$PBMF = \left(\frac{1}{K} \times \frac{E_1}{J_m} \times D_K \right)^2. \tag{22}$$

Table 6
Values of the DB-index, Dunn's index, the XB-index and *PBM-index* in the range of $K = 2, \ldots, 15$ for different data sets using the *k*-means algorithm (Entries in bold face indicate the optimal values for respective indices)

Number of clusters	Data set							
	Cancer				*Kalazaar*			
	DB	DUNN	XB	PBM	DB	DUNN	XB	PBM
2	**0.757**	**0.930**	**0.149**	**145.181**	1.029	0.450	0.310	486.354
3	1.535	0.675	0.432	104.962	1.101	**0.519**	0.471	**910.747**
4	1.604	0.679	0.451	73.630	**0.918**	0.458	**0.292**	777.546
5	1.655	0.580	2.042	60.531	1.106	0.316	0.758	577.664
6	1.627	0.258	2.215	44.857	0.973	0.449	0.465	602.602
7	1.615	0.256	2.214	33.841	0.922	0.445	0.389	416.579
8	1.773	0.229	2.212	27.910	1.038	0.320	0.685	425.636
9	1.595	0.250	1.924	21.414	0.984	0.322	0.682	438.088
10	1.575	0.247	1.963	18.022	1.066	0.297	0.578	325.102
11	1.675	0.224	4.561	15.809	0.924	0.362	0.457	381.356
12	1.607	0.224	1.862	13.462	0.919	0.354	0.380	373.390
13	1.557	0.181	4.630	11.802	0.913	0.256	0.514	343.570
14	1.631	0.208	1.683	10.993	1.003	0.354	0.616	337.034
15	1.556	0.195	3.874	9.633	0.838	0.253	0.557	311.357

Table 7
Values of *PBM-index* in the range of $K = 2, \ldots, 15$ for different data sets using the EM algorithm (Entries in bold face indicate the optimal values for *PBM-index*)

Number of clusters	Data set							
	Circular_5_2	*Circular_6_2*	*Spherical_4_3*	*Elliptical_10_2*	*Iris*	*Crude_Oil*	*Cancer*	*Kalazaar*
2	6.152	64.220	274.903	154.143	20.193	291.611	**142.127**	**822.548**
3	6.729	77.354	379.794	196.928	**22.365**	**410.278**	74.269	395.127
4	15.001	158.456	**941.876**	224.964	14.311	369.958	62.428	522.027
5	**18.700**	412.415	671.521	179.121	12.754	351.747	41.402	480.707
6	15.931	**528.128**	478.087	221.211	11.229	276.520	32.593	411.033
7	14.288	503.542	376.419	210.464	12.502	238.090	16.856	397.755
8	15.201	405.217	286.345	237.906	8.226	187.752	20.557	359.754
9	10.788	353.571	265.228	282.613	9.316	221.240	15.785	337.086
10	11.682	301.500	225.182	**299.402**	7.710	143.237	15.856	322.372
11	10.479	285.368	200.978	220.768	6.273	178.103	14.090	327.805
12	10.765	256.163	175.113	178.983	5.885	195.534	9.076	280.291
13	9.365	266.871	151.374	199.532	5.533	160.509	10.480	251.697
14	9.229	226.105	142.167	189.317	4.729	137.451	10.087	195.523
15	9.064	184.089	139.431	185.032	4.161	120.741	8.121	181.819

Table 8
Comparison of the cluster centers obtained for the *Iris* data using the *k*-means and the EM methods with the original centers

Cluster number	Centers for *Iris* data											
	Actual cluster centers				*Centers with k-means*				*Centers with EM*			
1	6.588	2.974	5.552	2.026	6.854	3.077	5.715	2.054	6.673	3.002	5.531	1.988
2	5.006	3.428	1.464	0.246	5.006	3.428	1.464	0.246	5.006	3.428	1.462	0.246
3	5.936	2.770	4.260	1.326	5.884	2.741	4.389	1.434	5.817	2.731	4.229	1.338

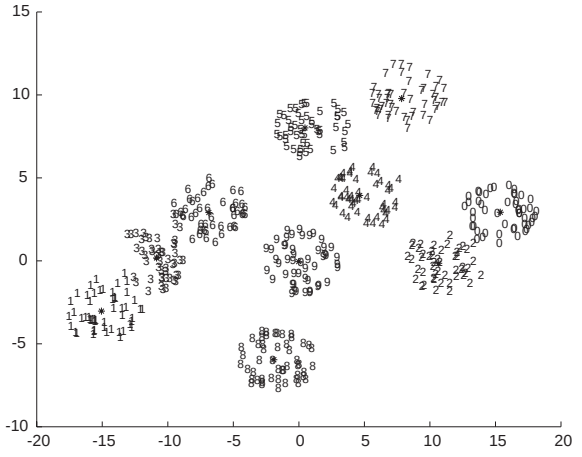

Fig. 13. Clustered *Elliptical_10_2* data and the corresponding Centers (*) for Ten Clusters (indicated by '0' to '9') obtained by the *k*-means algorithm, corresponding to the optimal value of *PBM-index* (=299.288).

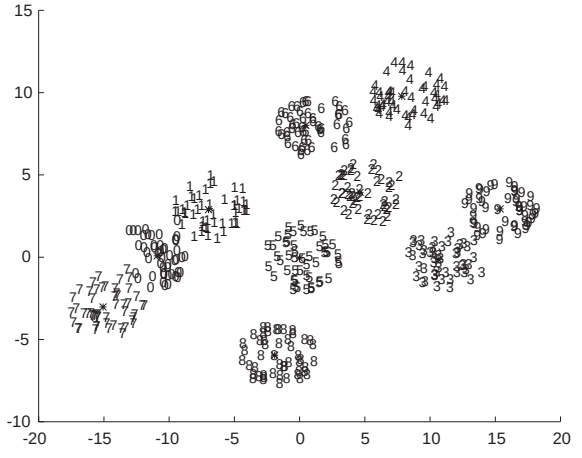

Fig. 14. Clustered *Elliptical_10_2* data and the corresponding centers (*) for ten clusters (indicated by '0' to '9') obtained by the EM algorithm, corresponding to the optimal value of *PBM-index* (=299.402).

Here, J_m is considered to be

$$J_m(U,Z) = \sum_{j=1}^{n}\sum_{k=1}^{K}(u_{kj})^m \|\mathbf{x}_j - \mathbf{z}_k\|. \tag{23}$$

We have taken $m = 1.5$, and have used the fuzzy *c*-means algorithm(FCM) [11] for clustering.

The number of clusters determined by the fuzzy versions of the XB-index and *PBM*-index are shown in Table 9. As can be seen from the tables, the *PBMF*-index is again able

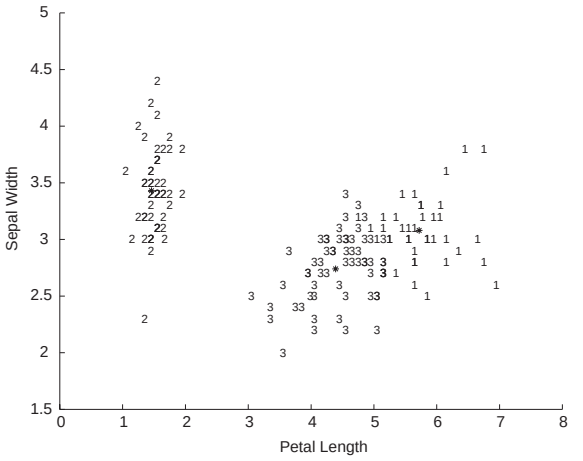

Fig. 15. Clustered *Iris* data and the corresponding centers (*) for three clusters (indicated by '1' to '3') obtained by the *k*-means algorithm, corresponding to the optimal value of *PBM-index* (=24.886).

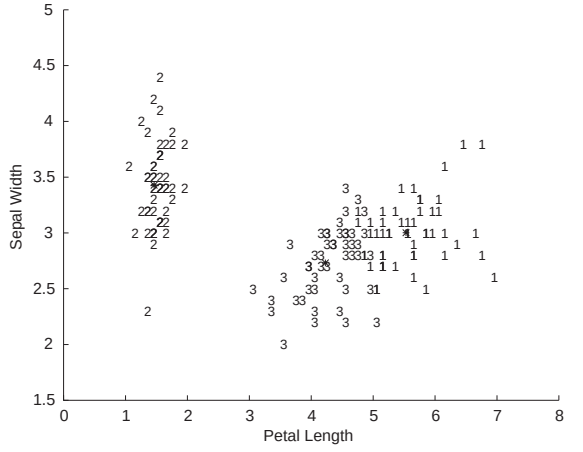

Fig. 16. Clustered *Iris* data and the corresponding centers (*) for three clusters (indicated by '1' to '3') obtained by the EM algorithm, corresponding to the optimal value of *PBM-index* (=22.365).

to correctly identify the number of clusters for all data sets except the *Kalazaar* data for which the XB-index identified five clusters instead of 2. The variations of the fuzzy XB-index and fuzzy *PBM*-index values for $K = 2$–15 are shown in Tables 10 and 11, respectively.

4. Discussion and conclusions

A cluster validity index is described in this article. It is found to attain its maximum value when the data is properly clustered. Therefore, this new index may be used for evolving the appropriate number of clusters in a given data

Table 9
Comparison of the number of clusters identified by the fuzzy XB-index and fuzzy *PBM-index* using fuzzy *c*-means algorithm when $m = 1.5$

Data set	Actual number of clusters	Number of clusters obtained	
		XB-index	PBMF-index
Circular_5_2	5	4	5
Circular_6_2	6	4	6
Spherical_4_3	4	4	4
Elliptical_10_2	10	9	10
Iris	3	2	3
Crude_Oil	3	2	3
Cancer	2	2	2
Kalazaar	2	5	3

set. Moreover, proper partitioning of the data set may also be achieved using the *PBM-index*. The performance of this index for providing the correct number of clusters is compared with those of the well known the DB-index, Dunn's index and the XB-index where the new index is found to significantly outperform all three of them. The fuzzy version of this index is also found to perform well for all the data sets considered.

Here, the expression for the index is raised to a power of 2 in order to improve the contrast between the index values for consecutive K-values. Note that in this article, for crisp clustering, we have used the k-means and EM algorithms as underlying clustering methodologies. Since the number of clusters needs to be specified a priori for both these algorithms, they have to be executed several times for different

Table 10
Values of fuzzy *XB-index* in the range of $K = 2, \ldots, 15$ (when $m = 1.5$) for different data sets using the fuzzy *c*-means algorithm (Entries in bold face indicate the optimal values for *XB-index*)

Number of clusters	Data set							
	Circular_5_2	*Circular_6_2*	*Spherical_4_3*	*Elliptical_10_2*	*Iris*	*Crude_Oil*	*Cancer*	*Kalazaar*
2	0.334	0.308	0.074	0.144	**0.062**	**0.108**	**0.135**	0.283
3	0.201	0.136	0.084	0.219	0.156	0.131	0.512	0.435
4	**0.115**	**0.043**	**0.052**	0.135	0.220	0.258	0.717	0.313
5	0.132	0.136	0.603	0.196	0.277	0.225	1.907	**0.272**
6	0.279	0.054	1.175	0.188	0.836	0.249	1.736	0.370
7	0.266	0.560	0.512	0.189	2.411	0.383	2.239	0.426
8	0.396	0.554	0.958	0.149	0.551	0.360	1.511	0.465
9	0.336	0.499	1.164	**0.123**	0.599	0.312	1.451	0.517
10	0.308	0.450	0.691	0.461	1.265	0.334	1.381	0.290
11	0.273	0.538	1.044	0.296	1.254	0.284	3.871	0.369
12	0.246	0.402	0.598	0.471	0.638	0.459	3.697	0.324
13	0.212	0.392	0.710	0.433	1.069	0.399	3.531	0.289
14	0.212	0.329	0.708	0.371	0.970	0.430	3.355	0.284
15	0.271	0.366	0.603	0.362	0.923	0.362	4.306	0.287

Table 11
Values of fuzzy *PBM-index* in the range of $K = 2, \ldots, 15$ (when $m = 1.5$) for different data sets using the fuzzy *c*-means algorithm (Entries in bold face indicate the optimal values for *PBM-index*)

Number of clusters	Data set							
	Circular_5_2	*Circular_6_2*	*Spherical_4_3*	*Elliptical_10_2*	*Iris*	*Crude_Oil*	*Cancer*	*Kalazaar*
2	9.666	90.821	290.259	187.060	21.828	352.688	**170.668**	595.979
3	12.236	117.701	598.341	282.274	**28.217**	**535.784**	146.827	**1238.021**
4	20.693	372.336	**960.814**	298.668	24.741	509.315	122.966	1013.571
5	**22.935**	417.548	808.379	313.113	23.194	525.968	106.170	980.565
6	20.467	**638.371**	590.393	297.770	21.557	329.858	94.615	912.366
7	20.178	527.483	468.896	287.642	18.464	394.300	85.470	764.219
8	19.651	489.693	476.680	284.901	15.944	367.489	75.899	718.408
9	17.131	417.426	440.664	338.523	14.410	327.031	67.109	704.994
10	18.199	402.123	365.957	**345.923**	12.953	327.574	59.707	695.149
11	16.581	361.228	359.292	315.631	12.208	336.502	59.044	742.347
12	16.695	291.795	239.754	291.795	10.976	311.628	56.596	697.123
13	15.650	319.158	325.261	277.956	9.847	270.208	52.258	706.695
14	14.761	340.550	250.304	248.598	9.885	326.127	45.387	630.308
15	14.138	256.352	251.064	286.219	6.235	316.342	42.667	671.984

values of K, varying in the range $[2, K_{max}]$. Similarly, for the fuzzy index, the FCM algorithm is run several times. The *PBM-index* is computed for the partitions thus obtained with different K and the number of clusters corresponding to the maximum value of this index is taken to provide the actual number of clusters. In this regard, a technique may be developed, for example by using evolutionary computation [18], which can evolve the appropriate number of clusters automatically using the *PBM-index*. Moreover, extensive experimentation with the *PBMF-index* is required to firmly establish its effectiveness. The authors are currently working in that direction.

5. Summary

Clustering is an unsupervised classification scheme where no a priori knowledge of the data set is available. In clustering (also known as exploratory data analysis), a set of patterns, usually vectors in a multidimensional space, are organized into a number of coherent and contrasted groups, such that patterns in the same group are similar in some sense and patterns in different groups are dissimilar in the same sense. Clustering can be performed either in crisp or fuzzy mode. In crisp mode, the clusters are disjoint, i.e., one pattern can belong to one and only one cluster. In fuzzy clustering, each pattern can be a member of all the clusters with a certain grade of membership. When only two hard grades, 0 and 1, are available, fuzzy clustering reduces to crisp clustering.

In clustering, the role of a validity index is very important. It helps to determine the appropriate number of clusters present in a data set. In the literature of clustering, a large number of cluster validity indices are there. Among them the Davies–Bouldin (DB) index and Dunn's index are highly used for crisp clustering, whereas, the Xie–Beni (XB) index is widely used for fuzzy clustering. Besides these indices, there are many other indices. But none of the indices perform satisfactorily for wide range of data sets. Owing to this reason it is necessary to develop a suitable index which can be applied to various data sets.

In this article, we propose a cluster validity index which can work for both crisp and fuzzy clustering. We have provided a detailed mathematical analysis of the index in support of the work-ability of the proposed index.

Like other indices, the proposed index is also an optimizing index, i.e., it can be used in association with an optimization algorithm in order to search for better clustering. The index can provide a measure of goodness of clustering on different partitions of a data set. The maximum value of this index, called the *PBM-index*, across the hierarchy provides the best partitioning. The index is defined as a product of three factors, maximization of which ensures the formation of a small number of compact clusters with large separation between at least two clusters.

We have used both the k-means and the EM algorithms as underlying crisp clustering techniques. For fuzzy clustering, we have utilized the well-known fuzzy c-means algorithm.

We have tested the superiority of the proposed *PBM-index* in appropriately determining the number of clusters, as compared to above-mentioned Davies–Bouldin index, Dunn's index and the Xie–Beni index, by using it for several artificial and real-life data sets. The data sets include four artificial and four real-life data. The artificial data sets include two- and three-dimensional data where the number of clusters vary from four to ten. The real-life data sets have dimensions in the range of four to nine and number of clusters vary from two to four. It is found that the proposed index outperforms the other indices for all the data sets.

References

[1] M.R. Anderberg, Cluster Analysis for Application, Academic Press, New York, 1973.
[2] J.A. Hartigan, Clustering Algorithms, Wiley, New York, 1975.
[3] P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, London, 1982.
[4] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[5] J.T. Tou, R.C. Gonzalez, Pattern Recognition Principles, Addison-Wesley, Reading, 1974.
[6] P.S. Bradley, U. Fayyad, C. Reina, Scaling EM (expectation maximization) clustering to large databases, Technical Report, Microsoft Research Centre, Redmond, USA, 1998.
[7] R.C. Dubes, A.K. Jain, Clustering techniques: the user's dilemma, Pattern Recognition 8 (1976) 247–260.
[8] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (1979) 224–227.
[9] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (1973) 32–57.
[10] X.L. Xie, G.A. Beni, Validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 3 (3) (1991) 841–846.
[11] N.R. Pal, J.C. Bezdek, On cluster validity for the fuzzy c-means model, IEEE Trans. Fuzzy Systems 3 (3) (1995) 370–379.
[12] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals Eugenics 3 (1936) 179–188.
[13] J.C. Bezdek, N.R. Pal, Some new indexes of cluster validity, IEEE Trans. Syst. Man Cybern. 28 (1998) 301–315.
[14] R. Kothari, D. Pitts, On finding the number of clusters, Pattern Recognition Lett. 20 (1999) 405–416.
[15] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall, London, 1982.
[16] S. Mitra, S.K. Pal, Fuzzy multi-layer perceptron, inferencing and rule generation, IEEE Trans. Neural Networks 6 (1995) 51–63.
[17] J. Bezdek, N.R. Pal, Some new indexes of cluster validity, IEEE Trans. Syst. Man Cybern. B 28 (1998) 301–315.
[18] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, New York, 1989.

About the Author—MALAY K. PAKHIRA received his B.Sc. degree in Physics and B.Tech. degree in Computer Science and Technology from the University of Calcutta, Calcutta, India in 1987 and 1990, respectively. He received his Masters in Computer Science and Engineering from Jadavpur University, Calcutta, India in 1992. Currently he is a lecturer in Computer Science and Technology at Kalyani Government Engineering College, West Bengal, India. At present he is doing his Ph.D. research work on Unsupervised Pattern Classification. His research interests include Image Processing, Pattern Recognition, Evolutionary Algorithms, Soft Computing and Data Mining.

About the Author—SANGHAMITRA BANDYOPADHYAY did her Bachelors in Physics and Computer Science in 1988 and 1991, respectively, from University of Calcutta, Calcutta, India. Subsequently, she did her Masters in Computer Science from Indian Institute of Technology, Kharagpur, India in 1993 and Ph.D. in Computer Science from Indian Statistical Institute, Calcutta, India in 1998. Currently she is an Assistant Professor at Indian Statistical Institute, Calcutta, India. She is the first recipient of *Dr. Shanker Dayal Sharma Gold Medal* and *Institute Silver Medal* for being adjudged the best all round postgraduate performer in 1994. She has worked in Los Alamos National Laboratory in 1997 as a graduate research assistant and in the University of New South Wales, Sydney, Australia. as a post doctoral fellow. Dr. Bandyopadhyay received the Indian National Science Academy (INSA) and the Indian Science Congress Association (ISCA) *Young Scientist Awards* in 2000, and Indian National Academy of Engineers (INAE) *Young Scientist Awards* in 2002. Her research interests include Evolutionary and Soft Computing, Pattern Recognition, Data Mining, Parallel Processing and Distributed Computing.

About the Author—UJJWAL MAULIK did his Bachelors in Physics and Computer Science in 1986 and 1989 respectively from University of Calcutta, Calcutta, India. Subsequently, he did his Masters and Ph.D. in Computer Science in 1991 and 1997, respectively, from Jadavpur University, India. Dr. Maulik has worked as a scientist in Center for Adaptive Systems Application, Los Alamos, and Los Alamos National Laboratory, New Mexico, USA in 1997. In 1999, he went on a postdoctoral assignment to University of New South Wales, Sydney, Australia. He is the recipient of the Government of India BOYSCAST fellowship for doing research in University of Texas at Arlington, USA in 2001. Dr. Maulik has been elected Fellow of the Institute of Electronics and Telecommunication Engineers (IETE). He is currently an assistant professor in the Department of Computer Science, Kalyani Government Engineering College, India. His research interests include Parallel and Distributed Processing, Natural Language Processing, Evolutionary Computation, Pattern Recognition and Data Mining.