Two Levels of Extensions of Validity Function Based Fuzzy Clustering

Minyar Sassi
Ecole Nationale d’Ingénieurs de Tunis, Tunisia
minyar.sassi@enit.rnu.tn

Amel Grissa Touzi
Ecole Nationale d’Ingénieurs de Tunis, Tunisia
amel.touzi@enit.rnu.tn

Habib Ounelli
Faculté des Sciences de Tunis, Tunisia
habib.ounelli@fst.rnu.tn

ABSTRACT
Since clustering is an unsupervised method and there is no a-priori indication for the actual number of clusters presented in a data set, there is a need of some kind of clustering results validation. In this paper, we propose a new cluster validity index for the fuzzy clustering algorithms. This validation includes two levels. The first during the clustering process for identifying the worst cluster to delete it. The second includes the validity function for evaluating the set of the resulting partitions.

Key Words: Fuzzy C-Means Algorithm, Validity Function, Modified Xie-Beni validity index, Clustering Process.

1. Introduction
Cluster analysis has been playing an important role in solving many problems in medicine, psychology, biology, sociology, pattern recognition and image processing. Grouping the data objects into clusters is the main step. Clustering algorithms attempt to assess the interaction among patterns by organizing patterns into clusters such that patterns within a cluster are more similar to each other than are patterns belonging to different clusters. Extensive and good overview of clustering algorithms can be found in the literature [1,2].
The data objects are grouped into clusters based on a number of different approaches. Partition-based clustering and hierarchical clustering are two of the main techniques. Partition-based clustering often starts from an initial partition and optimizes (usually locally) a clustering criterion. Hard clustering, in which each data object belongs to a single cluster, and fuzzy clustering, in which each object is assigned a degree of membership in ever cluster, are the two main partition-based techniques. We mention some algorithm like K-Means [3], K-Medoids [4], and the Fuzzy C-Means (FCM) [5]. Hierarchical clustering techniques generate a nested series of partitions based on a criterion, which measures the similarity between clusters or the separability of a cluster, for merging or splitting clusters. We can mention, BIRCH (Balanced Iterative Reducing and Clustering using Hierarches) algorithm [6], CURE (Clustering Using REpresentatives) algorithm [7]. Other techniques include density-based clustering like DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [8], DENCLUE (Density clustering) [9] and grid-based clustering like WaveCluster (Clustering using Wavelet Trans-formation) [10].
It has been demonstrated that fuzzy
clustering, using algorithms such as the FCM, has clear advantages over crisp and probabilistic clustering methods.

One of the problems of FCM clustering algorithm is the difficulty in determining the optimal number of clusters. The proposed solutions to this problem usually rely on validity measures [11,12], such as the Dunn’s separation index [13,14], the Bezdek’s partition coefficient [15], the Xie-Beni’s separation index [16], and the Gath-Geva’s index [11], etc.

For the time being, there is no universal solution to this problem, since clusters may differ in shapes and sizes. Therefore in order to obtain optimal number of fuzzy cluster, two conditions are necessary:

- A good validity function for the evaluation of the goodness of clusters for varying number of clusters,
- A good cluster algorithm that can automatically produce and optimize the number of clusters.

Once these two requirements are met, the strategy of getting the optimal number of fuzzy clusters easy: produce the optimal solution for each potential number of clusters, and use the validity function to choose the best one, thus automatically deciding on the number of clusters.

To satisfy the first condition, we propose a new validity function based on Xie-Beni validity index to evaluate the compactness and separation of different clustering results for a given data set. Experiments show that, unlike the situation when Xie-Beni validity index is used, even when the number of clusters is very large, our validity function still works well.

For the second condition, we are particularly interested in the FCM algorithm proposed by Bezdek [17]. This algorithm requires the user to pre-define the number of clusters C; however, it is not always possible to know C in advance. Because the fuzzy partitions obtained using the FCM algorithm depend on the choice of C, it is necessary to validate each of the C-partitions once they are found [15]. This validation is carried out by the new validity function, which evaluates each of the fuzzy C-partitions and determines the optimal partition or the optimal number of clusters C from them.

In the present work here, a new cluster validity index for fuzzy clustering is proposed. The proposed index consists of two properties: a compactness measure and a separation measure forming the modified Xie-Beni validity index for evaluation of the quality of cluster in two levels (a) for each cluster during the clustering process and (b) for the set of cluster result.

The organization of the rest of the paper is as follows. In section 2, we provide background information of fuzzy clustering and discuss previous work in cluster validity indices. In section 3, the proposed extensions of the validity function are discussed. Section 4 presents the experimental results. Finally, section 5 concludes the paper.

2. Cluster Validity

In this section, we present the FCM clustering algorithm and clustering result validity assessment.

2.1 The Fuzzy C-Means Clustering Algorithm

FCM clustering algorithm [5] employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.

FCM is an iterative algorithm. The aim of FCM is to find cluster centers (centroids) that minimize a dissimilarity function. The basic FCM algorithm can be formulated as follows:

$$\text{Minimize}_d(U,V) = \sum_{j=1}^{N} \sum_{i=1}^{C} u_{ij}^m \| x_j - v_i \|^2$$  \hspace{1cm} (1)

Where the N is the total number of data vectors in a given data set and c is the number of clusters; \( X = \{ x_1, x_2, ..., x_N \} \subset \mathbb{R}^S \) are the feature data and cluster centers; and \( U = \{ u_{ij} \}_{NxC} \) is a fuzzy partition matrix composed of the membership of each feature vector \( x_j \) in each cluster \( i \). \( u_{ij} \) should satisfy
\[ \sum_{i=1}^{C} u_{ij} = 1 \text{ for } j = 1, \ldots, N \text{ and } u_{ij} \geq 0 \text{ for all } i = 1, 2, \ldots, C \text{ and } j = 1, 2, \ldots, N. \] The exponent \( m \) is a parameter, usually called a fuzzifier parameter. To minimize \( J_m(U, V) \), the cluster centers \( v_i \) and the membership matrix \( U \) need to be computed according to the following iterative formula:

To reach a minimum of dissimilarity function there are two conditions. These are given in Eq.\((2)\) and Eq.\((3)\).

\[
\forall i j \quad v_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m} \quad (2)
\]

\[
\forall i j \quad u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( d_{ij} - d_{kj} \right)^{2/(m-1)}} \quad (3)
\]

This algorithm determines the following steps:

**Step 1:** Input the number of clusters \( C \), the fuzzifier \( m \) and the distance function.

**Step 2:** Initialize the cluster centers \( v_i^0 (i = 1, 2, \ldots, C) \)

**Step 3:** Calculate \( u_{ij} (i = 1, 2, \ldots, N; j = 1, 2, \ldots, C) \) using Eq.\((3)\).

**Step 4:** Calculate \( v_i^1 (j = 1, 2, \ldots, C) \) using Eq.\((2)\).

**Step 5:** If \( \max_{1 \leq i \leq C} \left| v_i^0 - v_i^1 \right| \leq \varepsilon \) then go to Step 6; else let \( v_i^0 = v_i^1 (i = 1, 2, \ldots, C) \) and go to Step 3.

**Step 6:** Output the clustering results: cluster centers \( v_i^1 (i = 1, 2, \ldots, C) \), membership matrix \( U \) and the elements of each cluster \( i \), i.e., all the \( x_j \) such that \( u_{ji} > u_{jk} \) for all \( k \neq i \).

### 2.2 Clustering Result Validity Assessment

#### 2.2.1 Problem specification

The objective of the clustering methods is to discover significant groups present in a data set. In general, they should search for clusters whose members are close to each other (in other words have a high degree of similarity) and well separated. A problem we face in clustering is to decide the optimal number of clusters that fits a data set.

#### 2.2.2 Validity Indices

There are two categories of fuzzy validity indices. The first category uses only the memberships values, \( u_{ij} \), of a fuzzy partition of data. On the other hand the latter one involves both the \( U \) matrix and the dataset itself.

**a) Validity Indices involving only the membership values**

Bezdek proposed in [18] the partition coefficient, which is defined as:

\[
\frac{\sum_{j=1}^{N} \sum_{i=1}^{C} u_{ij}^2}{N} \quad (4)
\]

The \( v_{PC} \) index values range in \([1/C, 1]\), where \( C \) is the number of clusters. The closer to the unity the index the “crisper” the clustering is. In case that all membership values to a fuzzy partition are equal, that is, \( u_{ij} = C \), the \( v_{PC} \) obtains its lower value. Thus, the closer the value of \( v_{PC} \) to \( 1/C \), the fuzzier the clustering is. Furthermore, a value close to \( 1/C \) indicates that there is no clustering tendency in the considered dataset or the clustering algorithm failed to reveal it.

The partition entropy coefficient is another index of this category. It is defined as:

\[
\frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{C} \left[ u_{ji} \log_a(u_{ji}) \right] \quad (5)
\]

Where \( a \) is base of the algorithm. The index is computed for values of \( C \) greater than 1 and its values ranges in \([0, \log_a C]\). The closer the value of \( v_{PE} \) to 0, the harder the clustering is. As in the previous case, the values of index close to the upper bound (i.e., \( \log_a C \),
indicate absence of any clustering structure in the dataset or inability of the algorithm to extract it. The drawbacks of these indices are:

i) Their monotonous dependency on the number of clusters. Thus, we seek significant knees of increase (for \( v_{PC} \)) or decrease (for \( v_{PE} \)) in plot of the indices versus the number of clusters.

ii) Their sensitivity to the fuzzifier, \( m \). More specifically, as \( m \to 1 \) the indices give the same values for all values of \( C \).

On the other hand when \( m \to \infty \), both \( v_{PC} \) and \( v_{PE} \) exhibit significant knee at \( C = 2 \).

b) Indices involving the membership values and the dataset.

The Xie-Beni index, \( v_{XB} \), also called the compactness and separation validity function, is a representative index of this category. Then the XB index is defined as:

\[
v_{XB} = \frac{\sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^2 \| x_j - v_i \|^2}{N \left( \min_{i \neq k} \| v_i - v_k \|^2 \right)}
\]  

(6)

Where \( N \) is the number of points in the data set. It is clear that small values of \( v_{XB} \) are expected for compact and well-separated clusters. We note, however, that \( v_{XB} \) is monotonically decreasing when the number of clusters \( C \) gets very large and close to \( N \). One way to eliminate this decreasing tendency of the index is by calculating the variance of each cluster by summing over only the members of each cluster rather than over all \( N \) for each cluster, which contrasts with the original Xie-Beni validity measure. Hence, the most desirable partition is obtained by maximizing \( v_{XB} \). We define a Modified Xie-Beni validity index as:

\[
v_{MXB} = \frac{D_{\min}^2}{\sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^2 \| v_j - v_i \|^2}
\]  

(8)

\( D_{\min} \) is the minimum distance between the cluster centers.

\[
D_{\min} = \min_{i \neq k} \| v_i - v_k \|^2
\]

(9)

Evaluating the quality of the clustering results is another important issue. Clustering is an unsupervised procedure. There is no objective criterion for evaluating the clustering results; they are assessed using a cluster validity index. In general, geometric properties, including the separation between clusters and compactness within a cluster, are often used to measure the quality. The cluster validity index also plays an important role in determining the number of clusters. It is

3. Validity Functions Extensions

In this section, we present our fuzzy clustering approach.

3.1 Motivations

While basing on Eq. (6), we notice that \( v_{XB} \) is monotonically decreasing when the number of clusters \( C \) gets very large and close to \( N \). One way to eliminate this decreasing tendency of the index is by calculating the variance of each cluster by summing over only the members of each cluster rather than over all \( N \) for each cluster, which contrasts with the original Xie-Beni validity measure. Hence, the most desirable partition is obtained by maximizing \( v_{MXB} \).
expected that the optimal value of the cluster validity index should be obtained at the true number of clusters. A general approach for determining the number of clusters is to select the optimal value of a certain cluster validity index. Whether a cluster validity index yields the true number of clusters is a criterion for the validity index. Most existing criteria give good results for data sets with well separated clusters, but usually fail for complex data sets.

3.2 The Proposed Validity Index and Validation Algorithm

The Modified Xie-Beni validity index is defined by Eq. (8). As we can see, this new validity index uses the minimum distance between the cluster centers. We extend this validity index for calculating the measure of separation and compactness to evaluate individual clusters.

The compactness of ith cluster \( C_i \) denoted as \( cp_i \) is given by:

\[
cp_i = \frac{\sum_{x_j \in C_i} u_i(x_j)^2}{\sum_{x_j \in C_i} u_i(x_j)^2 \| x_j - v_i \|^2}
\]  

Where \( u_i(x_j) \) is the membership value of \( x_j \) belonging to ith cluster. \( v_i \) is the center of ith cluster \( C_i \).

The separation of \( C_i \) denoted as \( sp_i \), is given by:

\[
sp_i = \min_{1 \leq j \leq C, i \neq j} \| v_i - v_j \|^2
\]

Where \( v_i \) is the center of ith cluster \( C_i \), \( v_j \) is the center of jth cluster \( C_j \).

The separation-compactness of \( C_i \) denoted as \( sc_i \) is given by:

\[
sc_i = sp_i \times cp_i
\]

A small value of \( sc_i \) indicates the worst cluster to be deleted. Therefore, the optimal fuzzy c-partition or the optimum value of \( c \) is obtained by minimizing \( \nu_{MXB} \) over \( C = C_{\text{max}}, ..., 2 \).

We now describe the procedural steps for the validation of the FCM using the proposed validity index \( \nu_{MXB} \).

**Step 1:** initialize the parameters related to the FCM and the validity index \( C = C_{\text{max}}, C_{\text{max}} = 2 \), \( C_{\text{opt}} = \sqrt{N} \) and initialize \( C \) cluster centers.

**Step 2:** Apply the basic FCM clustering algorithm.

**Step 3:** Compute and store the Modified Xie-Beni validity index \( \nu_{MXB} \) for the fuzzy partition obtained in step 2.

**Step 4:**

Repeat

Calculate \( sc_i \) for \( i = 1,...,C \).

Delete the worst cluster.

Recalculate the new centers.

Decrease \( C \rightarrow C - 1 \); perform the basis FCM algorithm based on parameter \( C \) to find the optimal cluster centers.

Calculate validity function \( \nu_{MXB} \) for new clusters using Eq. (8), denote it as \( \nu'_{MXB} \); if

\[
\nu'_{MXB} > \nu_{MXB} \quad \nu_{MXB} = \nu'_{MXB},
\]

\( C_{\text{opt}} = C \).

**Until** \( c \leq 2 \).

**Step 5:** Output \( V = \{v_1, v_2, ..., v_{\text{opt}}\} \) as an optimal cluster centers.

There is no general agreement on what value to use for \( C_{\text{max}} \). The value of \( C_{\text{max}} \) can be chosen according to the user’s knowledge of the data set; however, as this is not always possible, a rule of thumb that many investigators use is \( C_{\text{max}} \leq \sqrt{N} \).

4. Experimental Results

To demonstrate the effectiveness with which the proposed index determines the optimal partition, we conducted extensive comparisons with other indices on a number of widely used data sets. The
proposed index $v_{MXB}$ was compared with four fuzzy cluster validity indices mentioned in Section 2.2: Bezdek’s $v_{PC}$ [18] and $v_{PE}$ [18], Xie and Beni’s $v_{XB}$ [16], Fukuyama and Sugeno’s $v_{FS}$ [19].

First, we tested the cluster validity indices for three data sets. The performance of a fuzzy cluster validity index depends on the outcome of a fuzzy clustering algorithm. A validity index is not able to provide desirable evaluation when the used clustering algorithm is not appropriate to the partitioning of a given data set. For a fair comparison, we choose three data sets that the standard FCM can easily discriminate, and choose four validity indices that is known to provide good evaluation results for the FCM. For each data set, we performed the FCM algorithm at each value of $C = C_{\text{max}}, \ldots, 2$.

The parameters of the FCM were set to a termination criterion $\varepsilon = 0.001$, weighting exponent $m = 2$ and $\|\ast\|^2$ was the square of the Euclidean norm. Initial centroids were randomly selected.

Four data sets were used to evaluate the validation performance of each index: the X30 [20], BENSAID [21] and IRIS [22] data sets.

The first data set contains $N = 30$ points. The data set has $C = 3$ compact, well-separated clusters with 10 points per cluster.

Table 1 lists the results of validity indices for $C = C_{\text{max}}, \ldots, 2 \approx 5$. For each $C$, index values were computed for each of the five validity indices considered. Seven indices including the proposed $v_{MXB}$ correctly identify the optimal $C$. The optimal $c$’s of $v_{FS}$ is at $C = 4$ indicating that it had failed to recognize the correct number $C$.

The second data set contains 49 data points in two dimensional measurement spaces, and consists of three clusters.

Table 2 shows the results obtained using the various validity indices with respect to $C = C_{\text{max}}, \ldots, 2$, where $C_{\text{max}} = \sqrt{N} \approx 7$. The optimal number of clusters in the BENSAID’s set, $C = 3$, was correctly recognized by five cluster validity indices including the proposed index $v_{MXB}$. Among the unsuccessful algorithm, $v_{PE}$ indicate the presence of two clusters. $v_{FS}$ indicate optimal values of $C \approx 7$.

Table 3 list the validation results of each index for the IRIS data set for $C = C_{\text{max}}, \ldots, 2$ where $C_{\text{max}} = \sqrt{N} \approx 12$. IRIS has $N = 150$ data points in a four-dimensional measurement space that represent three physical clusters. In their numerical representation two of the three clusters are hardly discernable, while the third is well separated from the other two. On can argue in favor of both $C = 2$ and 3 for IRIS because of the substantial overlap of two of the clusters. We also take $C = 2$ as the optimal choice in view of the
geometric structure of IRIS as mentioned by Pal and Bezdek [2]. The optimal $C = 2$ is identified by $v_{PC}$, $v_{PE}$, $v_{XB}$ and $v_{MXB}$. In contrast, $v_{FS}$ yield optimal values at $C = 3$.

Table 3: Cluster Validity Values for the IRIS data set

<table>
<thead>
<tr>
<th>$c$</th>
<th>$v_{PC}$</th>
<th>$v_{PE}$</th>
<th>$v_{XB}$</th>
<th>$v_{FS}$</th>
<th>$v_{MXB}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0.39</td>
<td>0.62</td>
<td>1.45</td>
<td>-292.54</td>
<td>0.29</td>
</tr>
<tr>
<td>11</td>
<td>0.41</td>
<td>0.61</td>
<td>1.33</td>
<td>-290.04</td>
<td>0.30</td>
</tr>
<tr>
<td>10</td>
<td>0.44</td>
<td>0.52</td>
<td>1.30</td>
<td>-325.74</td>
<td>0.37</td>
</tr>
<tr>
<td>9</td>
<td>0.49</td>
<td>0.44</td>
<td>1.12</td>
<td>-331.92</td>
<td>0.43</td>
</tr>
<tr>
<td>8</td>
<td>0.55</td>
<td>0.41</td>
<td>0.90</td>
<td>-334.98</td>
<td>0.45</td>
</tr>
<tr>
<td>7</td>
<td>0.59</td>
<td>0.38</td>
<td>0.78</td>
<td>-375.90</td>
<td>0.59</td>
</tr>
<tr>
<td>6</td>
<td>0.60</td>
<td>0.30</td>
<td>0.30</td>
<td>-391.48</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>0.62</td>
<td>0.29</td>
<td>0.40</td>
<td>-371.28</td>
<td>0.66</td>
</tr>
<tr>
<td>4</td>
<td>0.69</td>
<td>0.20</td>
<td>0.40</td>
<td>-406.39</td>
<td>0.79</td>
</tr>
<tr>
<td>3</td>
<td>0.75</td>
<td>0.14</td>
<td>0.09</td>
<td>-449.11</td>
<td>0.84</td>
</tr>
<tr>
<td>2</td>
<td>0.81</td>
<td>0.11</td>
<td>0.07</td>
<td>-400.94</td>
<td>0.89</td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, we have proposed a new cluster validity index for the fuzzy clustering algorithms. The proposed index introduced a modification to the Xie-Beni validity measure. Thus, the optimal fuzzy partition is obtained by maximizing $v_{MXB}$ with respect to $c$. The proposed index works well even when the number of clusters is very large. The performance of the proposed index on various data sets demonstrated its effectiveness.

As future works, more tests and comparisons will be done on both the modified Xie-Beni validity function and other validity indices.

References:


Large Database (VLDB98), (NY USA), pp. 428-439, 1998.


