IEMS Vol. 8, No. 2, pp. 101-108, June 2009.

Validation Measures of Bicluster Solutions

Youngrok Lee
Department of Industrial and Management Engineering
Pohang University of Science and Technology, Pohang 790-784, KOREA
E-mail: lyr1004@postech.ac.kr

Jeong-Hwa Lee
Department of Industrial and Management Engineering
Pohang University of Science and Technology, Pohang 790-784, KOREA
E-mail: bls83@postech.ac.kr

Chi-Hyuck Jun†
Department of Industrial and Management Engineering
Pohang University of Science and Technology, Pohang 790-784, KOREA
E-mail: chjun@postech.ac.kr

Received Date, May 4, 2009; Accepted Date, May 28, 2009

Abstract. Biclustering is a method to extract subsets of objects and features from a dataset which are characterized in some way. In contrast to traditional clustering algorithms which group objects similar in a whole feature set, biclustering methods find groups of objects which have similar values or patterns in some features. Both in clustering and biclustering, validating how much the result is informative or reliable is a very important task. Whereas validation methods of cluster solutions have been studied actively, there are only few measures to validate bicluster solutions. Furthermore, the existing validation methods of bicluster solutions have some critical problems to be used in general cases. In this paper, we review several well-known validation measures for cluster and bicluster solutions and discuss their limitations. Then, we propose several improved validation indices as modified versions of existing ones.

Keywords: Biclustering, Clustering, Feature, Object, Validation Index

1. INTRODUCTION

In machine learning, *clustering* is an unsupervised learning method to discover groups of similar objects from a dataset (Bishop 2006). By cluster analysis, we find useful groups which may not be known previously. For example, through clustering, we can find a customer group which shows similar purchasing patterns. Cluster analysis is helpful to investigate relation between objects or to organize a different action to each cluster.

Biclustering is a method of detecting homogeneous and uniquely characterized subsets of objects and features (or attributes) from an original dataset (Cheng and Church 2000). Whereas traditional clustering algorithms group objects similar in a whole feature set in a cluster, biclustering methods find groups of objects which have similar values or patterns in some features. For example, a customer group which shows similar purchasing patterns for pet food can be found as a bicluster even though the group does not show similarity for other features. Hence, in many cases, biclustering finds useful groups that may not be obtained by clustering.

The difference between the results of clustering and biclustering is represented by Figure 1. Whereas clustering partitions a dataset in one direction as in (a) or (b), biclustering discovers groups by extracting a subset of both objects and features as in (c). Various types of bicluster solutions are introduced in Madeira and Oliveira (2004).

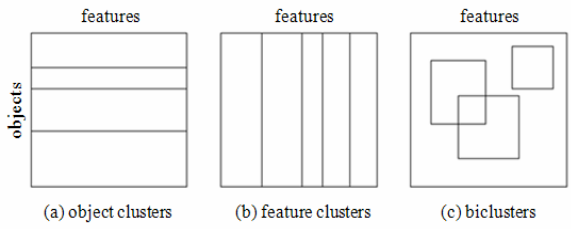

Figure 1. Cluster solutions and bicluster solutions.

† : Corresponding Author

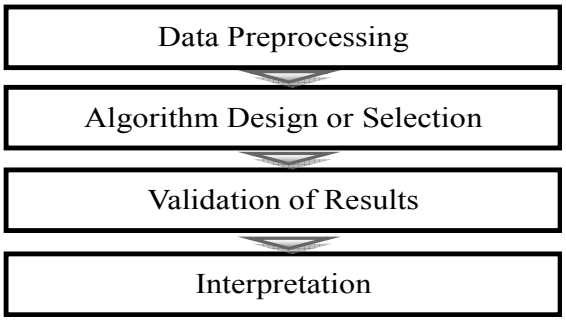

Figure 2. Procedure of Clustering and Biclustering.

Halkidi *et al*. (2001) and Xu and Wunsch(2005) propose four-step procedure of cluster analysis like Figure 2. Bicluster analysis also follows this procedure. The third step, *validation*, is a process of evaluating the results of cluster or bicluster analysis quantitatively. This process is very important for reducing the cost of interpretation. If a non-informative solution is not filtered at the validation step, we might make unnecessary efforts.

Even though various biclustering algorithms have appeared in the last decade, there are only a few studies for the quantitative evaluation of biclustering solutions. Furthermore, existing valuation methods of biclustering are quite incomplete as compared with those of clustering. For this reason, bicluster analysis has not been applied widely although it can give us valuable information.

Jain and Dubes (1988) and Halkidi *et al.* (2001) categorize validation indices into three types; internal, external and relative indices. More simply, Handl *et al.* (2005) classify validation measures into two types; internal and external measures. External indices are distinguished from internal indices by the presence of prior information of known categories. We also categorize validation indices into two types like Handl *et al.* (2005).

The purpose of this paper is to review the existing validation indices of bicluster solutions and to propose improved measures. The rest of this paper is organized as follows: in Section 2, validation indices of partitioning cluster solutions are reviewed; in Section 3, existing validation indices of bicluster solutions are reviewed; in Section 4, drawbacks of the existing indices are investigated and improved validation indices are proposed; finally, in Section 5, conclusions are made.

2. VALIDATION OF CLUSTER SOLUTIONS

Cluster analysis is mainly divided into hierarchical clustering and partitioning (Jain and Dubes, 1988). Some validation measures are applicable only to hierarchical clustering because they use the result of each step of agglomerative or divisive process. On the other hand, validation indices of partitioning solutions evaluate only the final result of partitioning. In this research, we concern only with validation measures which are applicable to partitioning.

Partitioning methods again can be categorized into two types; crisp partitioning and fuzzy partitioning (Halkidi *et al*., 2001). In the crisp clustering, a data point is assigned to exactly one cluster. On the contrary, in the fuzzy clustering, an object can correspond to many clusters with membership values in the range of (0, 1). In this research, we concentrate only on the crisp clustering.

2.1 Notations

Let $\mathbf{X} = (x_{ij})$ be the input dataset with n objects and m features. Also, let $\mathbf{x}_i = (x_{i1}, \cdots, x_{im})$ be the i th object of $\mathbf{X}$. Then, a crisp cluster solution C with K clusters can be defined as

$$C = \{C_1, C_2, \cdots, C_K\} \tag{1}$$

where C_k represents the k th cluster including n_k objects and $C_i \cap C_j = \phi$ for all $i \neq j$. So,

$$\sum_{k=1}^{K} n_k = n \tag{2}$$

The center of cluster C_k is denoted by

$$\mathbf{z}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i \tag{3}$$

and the center of the whole objects is represented as

$$\mathbf{z} = \frac{1}{n} \sum_{k=1}^{K} n_k \mathbf{z}_k \tag{4}$$

2.2 Internal Indices

Internal indices are validation measures which evaluate clustering results using only information intrinsic to the input dataset. Internal measures evaluate cluster solutions mainly in three view points; compactness, connectedness and separateness (Handl *et al*., 2005).

Both compactness and connectedness evaluate coherence or homogeneity of objects within a cluster. Compactness represents similarity of objects within a cluster. If the center or centroid of a cluster represents all objects in the cluster well, the compactness of the cluster is highly evaluated. On the contrary, connectedness represents density of a cluster. Connectedness of a cluster is high when all objects are connected with each other directly or indirectly by other objects in the cluster.

On the other hand, separateness evaluates heterogeneity between clusters. Separation quantifies how a cluster is apart from other clusters. It means that a cluster should have particular characteristics so that it should be discriminated from other clusters.

When we evaluate cluster solutions, we should con-

sider both homogeneity within a cluster and heterogeneity between clusters. Therefore, widely used validation indices combine compactness or connectedness and separateness into a single index.

The index proposed by Dunn (1973) combines compactness and separateness. Let $d(\mathbf{x}, \mathbf{y})$ be a metric of distance between two objects $\mathbf{x}$ and $\mathbf{y}$. Then, the Dunn's index can be formulated as

$$DI = \frac{\min_{\mathbf{x}\in C_i,\, \mathbf{y}\in C_j,\, 1\le i\ne j\le K} d(\mathbf{x}, \mathbf{y})}{\max_{\mathbf{x},\, \mathbf{y}\in C_i,\, 1\le i\le K} d(\mathbf{x}, \mathbf{y})} \tag{5}$$

Since the numerator increases as clusters are separated from each other and the denominator decreases as the homogeneity of each cluster increases, maximization of the Dunn's index is desired.

There are other validation indices that combine compactness and separateness using center points of clusters. The CH index by Caliński and Harabasz (1974) is given by

$$CH = \frac{1}{K-1}\sum_{k=1}^{K} n_k d(\mathbf{z}_k, \mathbf{z}) \Big/ \frac{1}{n-K}\sum_{k=1}^{K}\sum_{\mathbf{x}\in C_k} d(\mathbf{x}, \mathbf{z}_k) \tag{6}$$

In the CH index, numerator increases as clusters are separated and denominator decreases as each cluster becomes homogeneous. Therefore, a cluster solution which shows a large CH index is preferred.

Also, Davies and Bouldin (1979) introduce the DB index which is defined as

$$DB = \frac{1}{K}\sum_{k=1}^{K} \max_{l\ne k} \frac{\frac{1}{n_k}\sum_{\mathbf{x}\in C_k} d(\mathbf{x}, \mathbf{z}_k) + \frac{1}{n_l}\sum_{\mathbf{y}\in C_l} d(\mathbf{y}, \mathbf{z}_l)}{d(\mathbf{z}_k, \mathbf{z}_l)} \tag{7}$$

In this case, minimization of the DB index is advisable.

On the other hand, Brock *et al*. (2008) propose an index considering connectivity. Let $N_i(j)$ be the j th nearest neighbor of the object $\mathbf{x}_i$. Also, let $v_{i,N_i(j)}$ be a variable that

$$v_{i,N_i(j)} = \begin{cases} 0 & \text{if } \mathbf{x}_i \text{ and } N_i(j) \text{ are in the same cluster} \\ 1/j & \text{otherwise} \end{cases} \tag{8}$$

Then, the Brock's index can be formulated as

$$BI = \sum_{i=1}^{n}\sum_{j=1}^{L} v_{i,N_i(j)} \tag{9}$$

where L is a decision variable representing the number of neighbors employed to measure connectivity. Then, as each cluster becomes dense and separated from other clusters, the connectivity index becomes small.

2.3 External Indices

In contrast to internal indices, external indices use prior knowledge about groups of objects as extrinsic information. By comparing the obtained cluster solution with the prior knowledge, we can quantify quality of the cluster solution. Let P be the prior s-partition set which is represented as

$$P = \{P_1, P_2, \cdots, P_s\} \tag{10}$$

where P_l represents the l th partition of objects and $P_i \cap P_j = \phi$ for all $i \ne j$. Then, P is considered as a cluster solution which is known in advance.

Rand (1971) proposes a comparison measure of two cluster solutions. Let $S_C(\mathbf{x}, \mathbf{y})$ be 1 if two objects $\mathbf{x}$ and $\mathbf{y}$ belong to the same cluster in the cluster solution C, while it is 0 otherwise. Also, let a, b, c and d be the numbers of pairs of objects $(\mathbf{x}, \mathbf{y})$ which are defined as follows:

$$\begin{aligned} a &= \left| \{(\mathbf{x}, \mathbf{y}) \mid S_C(\mathbf{x}, \mathbf{y}) = 1,\, S_P(\mathbf{x}, \mathbf{y}) = 1\} \right| \\ b &= \left| \{(\mathbf{x}, \mathbf{y}) \mid S_C(\mathbf{x}, \mathbf{y}) = 1,\, S_P(\mathbf{x}, \mathbf{y}) = 0\} \right| \\ c &= \left| \{(\mathbf{x}, \mathbf{y}) \mid S_C(\mathbf{x}, \mathbf{y}) = 0,\, S_P(\mathbf{x}, \mathbf{y}) = 1\} \right| \\ d &= \left| \{(\mathbf{x}, \mathbf{y}) \mid S_C(\mathbf{x}, \mathbf{y}) = 0,\, S_P(\mathbf{x}, \mathbf{y}) = 0\} \right| \end{aligned} \tag{11}$$

Then, the Rand index is defined by

$$RI = \frac{a+d}{a+b+c+d} \tag{12}$$

Hubert and Arabie (1985) modify the Rand index as

$$RI_{adj} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \tag{13}$$

Also, Fowlkes and Mallows (1983) introduce another external index which is expressed by

$$FM = a \Big/ \sqrt{(a+b)(a+c)} \tag{14}$$

Intuitively, above external indices are 1 if the cluster solution C is equivalent to the prior partition set P, and they decreases as C becomes discordant with P. Therefore, a cluster solution whose external index is close to 1 is desired.

3. VALIDATION OF BICLUSTER SOLUTIONS

Similarly as in clustering, validation is an important issue in bicluster analysis. We can categorize validation indices of bicluster solutions into internal indices and external indices. In this section, we review several existing indices.

3.1 Notations

Bicluster solution set M can be defined as a set of biclusters as follows:

$$M = \{B_1, B_2, \cdots, B_K\} \tag{15}$$

where B_k denotes the k th bicluster, $k = 1, \cdots, K$. A bicluster is a combination of two subsets; one is a subset of objects and the other is a subset of features. Therefore, a bicluster B_k can be represented as

$$B_k = (O_k, F_k) = \{(\mathbf{x}_i, \mathbf{y}_j) \mid \mathbf{x}_i \in O_k,\ \mathbf{y}_j \in F_k\} \tag{16}$$

where O_k and F_k are subsets of objects and features, respectively, and $\mathbf{x}_i$ and $\mathbf{y}_j$ denote the i th row and j th column of $\mathbf{X}$, respectively.

Size of a bicluster B_k is defined as

$$|B_k| = |O_k| \times |F_k| \tag{17}$$

where $|O_k|$ and $|F_k|$ are the number of objects and features corresponding to B_k, respectively.

3.2 Internal Indices

Internal indices of bicluster solutions also use information only intrinsic to the dataset and the bicluster solution.

3.2.1 Average Residue

Cheng and Church (2000) define *residue* of an observed value x_{ij} in the bicluster B_k as

$$r_{ij}^{(k)} = x_{ij} - \frac{1}{|O_k|}\sum_{\mathbf{x}_i \in O_k} x_{ij} - \frac{1}{|F_k|}\sum_{\mathbf{y}_j \in F_k} x_{ij} + \frac{1}{|B_k|}\sum_{(\mathbf{x}_i, \mathbf{y}_j) \in B_k} x_{ij} \tag{18}$$

Then, they evaluate each bicluster by the *mean squared residue* which is defined as

$$MSR(B_k) = \frac{1}{|B_k|}\sum_{(\mathbf{x}_i, \mathbf{y}_j) \in B_k} r_{ij}^{(k)} \tag{19}$$

Yang *et al.* (2002) introduce the *average residue* to evaluate the total bicluster solution.

$$ASR = \frac{1}{K}\sum_{k=1}^{K} MSR(B_k) \tag{20}$$

Also, Madeira and Oliveira (2004) introduce the residue of overlapping biclusters with the general additive model or the general multiplicative model to evaluate the bicluster solutions. As the average residue becomes close to 0, the bicluster solution is highly evaluated.

3.2.2 $\bar{\Gamma}$ Index

Santamaría *et al.* (2007) propose an index by imitating the normalized Hubert's statistic (Jain and Dubes, 1988). Let $P = (P_{ij})$ be the proximity matrix of objects so that p_{ij} denotes the distance between two objects $\mathbf{x}_i$ and $\mathbf{x}_j$. Also, let $\mathbf{C} = (c_{ij})$ be the membership matrix that

$$c_{ij} = \frac{1}{1 + k_{ij}} \tag{21}$$

where k_{ij} is the number of biclusters which two objects $\mathbf{x}_i$ and $\mathbf{x}_j$ simultaneously belong to. Then, they define the statistic of objects as

$$\bar{\Gamma}_O = \frac{2}{n(n-1)} \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(p_{ij} - \mu_p\right)\left(c_{ij} - \mu_c\right)}{\sigma_p \sigma_c} \tag{22}$$

where $\mu_p(\mu_c)$ and $\sigma_p(\sigma_c)$ are the mean and the standard deviation of $\mathbf{P}(\mathbf{C})$, respectively.

In the same way, the statistic of features $\bar{\Gamma}_F$ can be formulated. Then, they define the $\bar{\Gamma}$ index by combining the two statistics as follows:

$$\bar{\Gamma} = \frac{n\bar{\Gamma}_O + m\bar{\Gamma}_F}{n + m} \tag{23}$$

Since the numerator increases as similar objects or features are grouped together, a bicluster solution with large $\bar{\Gamma}$ is preferred in the range of [-1, 1].

3.3 External Indices

External indices of biclustering are used to compare two bicluster solutions. If we have prior grouping information, we can evaluate a bicluster solution by comparing with the known information.

Let M_1 and M_2 be bicluster solutions which consist of K_1 and K_2 biclusters, respectively. We consider that one of them is the obtained solution and the other is the prior solution. Then, we can denote each bicluster solution as

$$M_j = \{B_1^{(j)}, B_2^{(j)}, \cdots, B_{K_j}^{(j)}\},\quad j = 1,\ 2 \tag{24}$$

where $B_k^{(j)} = (O_k^{(j)},\ F_k^{(j)})$.

Prelić *et al.*, (2006) propose the external index based on the Jaccard index (Downton and Brennan, 1980). The Prelić index compare two solutions based on categorization of objects as follows:

$$I_{Prelic}(M_1,\ M_2) = \frac{1}{K_1}\sum_{i=1}^{K_1} \max_j J\left(O_i^{(1)},\ O_j^{(2)}\right) \tag{25}$$

where $J(A, B)$ is the Jaccard index for two sets A and B:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{26}$$

Liu and Wang (2007) propose another external index which compares two solutions considering both objects and features. Their index (the LW index) can be formulated as

$$I_{LW}(M_1, M_2) = \frac{1}{K_1}\sum_{i=1}^{K_1}\max_j \frac{|O_i^{(1)} \cap O_j^{(2)}| + |F_i^{(1)} \cap F_j^{(2)}|}{|O_i^{(1)} \cup O_j^{(2)}| + |F_i^{(1)} \cup F_j^{(2)}|} \tag{27}$$

Whereas above two indices are based on the Jaccard index, Santamaría *et al*. (2007) propose an external index based on the Dice index (Dice, 1945) which is called the F_1 measure by Turner *et al*. (2005) in biclustering cases. The Santamaría index computes the overall relevance of two bicluster solutions as follows:

$$I_{Santamaria}(M_1, M_2) = \frac{1}{K_1}\sum_{i=1}^{K_1}\max_j D\left(B_i^{(1)}, B_j^{(2)}\right) \tag{28}$$

Where $D(A, B)$ is the Dice index given by

$$D(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{29}$$

Above three indices lie in the range of [0, 1]. The indices which are close to 1 mean that the two bicluster solutions are similar to each other.

While the Prelić index compares only object sets and the LW index compares object sets and feature sets independently, the Santamaría index compares two solutions using pairs of objects and features. Therefore, the Santamaría index is the most conservative index among above three indices.

4. NEW MEASURES FOR BICLUSTER SOLUTIONS

By comparing validation indices of biclustering with those of clustering, we can find some defects of the indices reviewed in Section 3. In this section, we reveal problems of the existing indices by raising some issues in validating bicluster solutions and propose new measures.

4.1 New Internal Index based on Average Residue

The average residue reviewed in Section 3.2.1 can be compared to compactness or connectedness in clustering. It considers only the homogeneity within a bicluster and never concerns about the heterogeneity between biclusters or significance of extracted biclusters. Therefore, the average residue prefers small biclusters. For example, if a bicluster consists of only one object and one feature, the mean squared residue of the bicluster is zero regardless of the distribution of input dataset.

Also, the average residue is significantly affected by the scale of the input dataset. For example, if the whole input values are multiplied by 0.1, the average residue will be reduced by 99% without change of the bicluster solution. Therefore, the average residue is inappropriate to compare bicluster solutions with various preprocessing of the input dataset.

To resolve these problems, combining the concept of separateness to the average residue is required. For example, we can evaluate the distance between two biclusters by computing the mean squared residue of the super bicluster including the two biclusters. Let $B_i \oplus B_j$ be the super bicluster of the two biclusters so that

$$B_i \oplus B_j = (O_i \cup O_j,\ F_i \cup F_j) \tag{30}$$

Then, the distance between B_i and B_j can be formulated as

$$d\left(B_i, B_j\right) = MSR\left(B_i \oplus B_j\right) \tag{31}$$

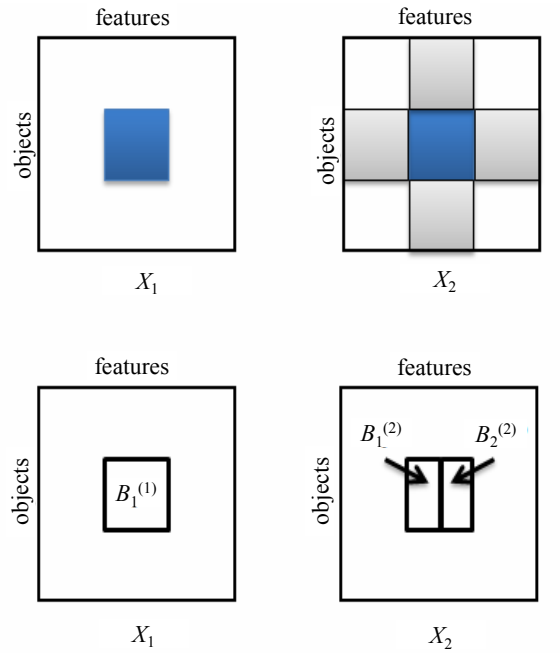

Figure 3. Example datasets and bicluster solutions.

By dividing the residue within biclusters by the distance between biclusters, we can construct a combined internal index like indices in Section 2.2. We define the new validation measure as

$$NSR = \frac{\sum_{k=1}^{K} MSR(B_k)}{\sum_{k=1}^{K} \min_{j=1, \cdots, K} d(B_k, B_j)} \tag{32}$$

where '*NSR*' is the abbreviation of the *normalized squared residue*.

A small value of NSR in Eq. (32) is preferred because the value decreases as coherence within biclusters increases and the separateness between biclusters increases. Generally, NSR is less than 1 because the mean squared residue of each bicluster is less than that of the super bicluster. Therefore, if NSR is close to 1 or larger, some biclusters in the solution should be merged.

Suppose, for example, in Figure 3, that M_1 and M_2 are two possible bicluster solutions of the dataset X_1. Elements of the dark area of X_1 were generated from the Gaussian distribution with mean 3 and standard deviation 0.3. Then, the average residue of M_1 is 0.093 and the average residue of M_2 is 0.094, so we cannot be sure which one is better on the basis of the average residue. However, since the value of NSR for the bicluster solution M_2 is 0.99, which is close to 1, we conclude that the two biclusters of M_2 should be merged to one bicluster like in M_1.

Since the normalized squared residue not only considers the separateness of biclusters but also normalizes the scale of the input dataset, it might be more appropriate to evaluate a bicluster solution than the average residue.

4.2 Modified $\bar{\Gamma}$ Index

The $\bar{\Gamma}$ index resolves problems of the average residue. Since the $\bar{\Gamma}$ index tends to group similar objects and features in the same bicluster as many as possible, the problem of overestimation of a small bicluster is reduced. Also, the $\bar{\Gamma}$ index is relatively free from the scaling problem because of the denominator σ_p.

However, since it is just combination of two validation indices of one-way clustering, it may be inappropriate in general for bicluster solutions. A bicluster solution which is not much relevant to a one-way cluster solution might be underestimated by the $\bar{\Gamma}$ index, even though it is informative.

If we obtain the normalized Hubert's statistics for each bicluster in both the object and the feature directions, this problem can be reduced. Let O and F be the whole object set and feature set of the input dataset X. Also, for a bicluster $B_k = (O_k, F_k)$, let X_k be the subset of the input dataset which consists of the object set O and the feature subset F_k. Then, the proximity matrix $\mathbf{P}$ can be constructed with the distance metric p_{ij} in the feature space F_k, not F. The distance measure might be defined differently according to the type of each bicluster. Also, let $\mathbf{C} = (c_{ij})$ be the membership matrix that

$$c_{ij} = \begin{cases} 0 & \text{if } \mathbf{x}_i, \mathbf{x}_j \in O_k \\ 1 & \text{otherwise} \end{cases} \tag{33}$$

Then the statistic of the bicluster B_k in the object direction can be defined as Eq. (22). Let Γ_{O_k} represent the statistic. Then we can define the statistic of objects as

$$\Gamma_O = \frac{1}{K}\sum_{k=1}^{K}\Gamma_{O_k} \tag{34}$$

In this manner, we also compute the statistic Γ_F. By averaging the two statistics, a single index validating a bicluster solution can be obtained as follows:

$$\Gamma_{proposed} = \left(\Gamma_O + \Gamma_F\right)/2 \tag{35}$$

As for the $\bar{\Gamma}$ index, a bicluster solution with a large value of the $\Gamma_{proposed}$ index is preferred in the range of [-1, 1].

For example, in Figure 3, suppose that M_1 is a bicluster solution of the dataset X_2. Elements of the dark area are same as X_1. Elements of the white and gray

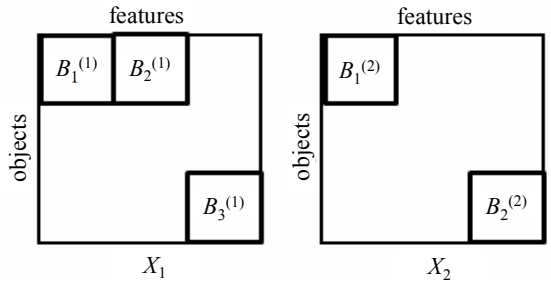

Figure 4. A pair of bicluster solutions whose external index is asymmetric.

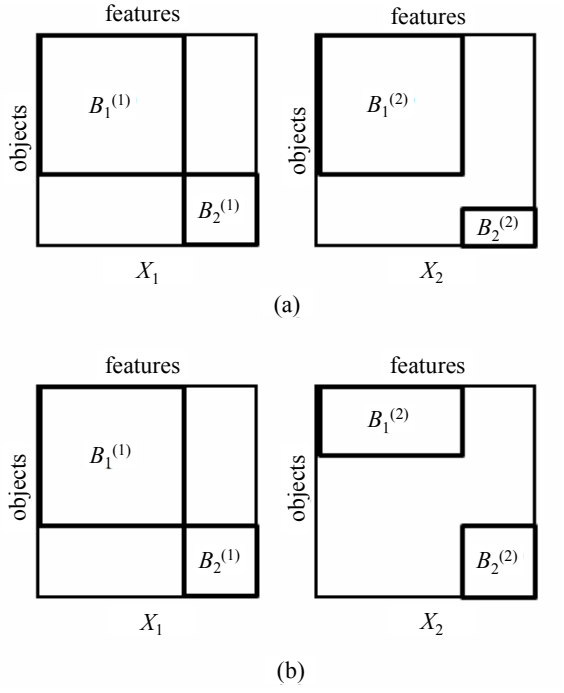

Figure 5. Two pair of bicluster solutions whose existing external indices are same.

areas were generated from the Gaussian distribution with mean 0 and standard deviation 1 and 2, respectively. Using the Euclidean distance metric, the $\bar{\Gamma}$ index of M_1 is -0.015, which shows that the solution M_1 is totally non-informative. This result is caused by the gray areas which grow distance between objects or features within the bicluster. However, when applying the new measure, the value of Eq. (35) is 0.79, which concludes that the solution M_1 seems to be good. In such case in general, the $\Gamma_{proposed}$ index might be more appropriate to validate bicluster solutions than the $\bar{\Gamma}$ index.

4.3 New External Indices

External indices in Section 3.2 have two common problems. One of them is that the indices are asymmetric. In contrast to external indices of clustering, in general,

$$I(M_1, M_2) \neq I(M_2, M_1) \tag{36}$$

where I represents an external index of biclustering. Figure 4 shows an example of two bicluster solutions. In this case, $I(M_2, M_1)$ is 1 because every bicluster in M_2 completely reappears in M_1. However, $I(M_1, M_2)$ is less than 1 because $B_2^{(1)}$ does not reappear in M_2.

This problem is caused by the 'max' operation in computing the indices. One simple way to resolve this asymmetry is averaging two indices as follows:

$$I_{sym}(M_1, M_2) = \{I(M_1, M_2) + I(M_2, M_1)\}/2 \tag{37}$$

Another problem of existing external indices is that they do not consider the size of biclusters in comparing two bicluster solutions. Figure 5 shows two pairs of two bicluster solutions. In Figure 5(a), objects of the small biclusters are matched 50%. On the other hand, in Figure 5(b), objects of the large biclusters are matched 50%. Even though the matching area in Figure 5(a) is larger, the existing external indices of solutions in Figure 5(a) and 5(b) are same.

To add the effect of the bicluster size to the external indices, weighted average can be used. Following indices give more weights to a large bicluster.

$$I_{Proposed1}(M_1, M_2) = \frac{\sum_{i=1}^{K_1} \left|B_i^{(1)}\right| \times \max_j J\left(B_i^{(1)}, B_j^{(2)}\right)}{\sum_{i=1}^{K_1} \left|B_i^{(1)}\right|} \tag{38}$$

$$I_{Proposed2}(M_1, M_2) = \frac{\sum_{i=1}^{K_1} \left|B_i^{(1)}\right| \times \max_j D\left(B_i^{(1)}, B_j^{(2)}\right)}{\sum_{i=1}^{K_1} \left|B_i^{(1)}\right|} \tag{39}$$

The proposed external index (called Proposed 1) in Eq. (38) is a weighted average of the Jaccard index whereas the proposed index (called Proposed 2) in Eq. (39) is a weighted average of the Dice index.

Table 1 shows the calculated external indices of two pairs of bicluster solutions in Figure 5. Differently from the existing three external indices, the proposed indices conclude that the pair of bicluster solutions in Figure 5(a) is much similar to each other than those in Figure 5(b).

Table 1. The external indices of the cases in the Figure 5.

Index	(a)	(b)
Prelić	0.75	0.75
Liu and Wang	0.88	0.88
Santamaría	0.83	0.83
Proposed 1	0.90	0.60
Proposed 2	0.93	0.73

5. CONCLUSIONS

Both clustering and biclustering are very useful analysis to find unknown informative groups from a dataset. In many cases, a bicluster solution is more informative than a standard clustering result. However, the research for the validation of bicluster solutions is still demanding. The bicluster solution should be carefully validated because poor validation leads useless or wrong information or high costs in interpreting the solutions.

In this paper, we reviewed existing validation indices of cluster and bicluster solutions by categorizing into internal and external indices. Then, we raised several issues in using the existing indices. Also, we proposed approaches to resolving those problems. We may need more extensive simulation study to demonstrate the performance of the proposed measures in a future.

ACKNOWLEDGMENT

This work was supported by the KOSEF through the National Core Research Center for System Bio-Dynamics at POSTECH.

REFERENCES

Aguilar-Ruiz, J. (2005), Shifting and scaling patterns from gene expression data. *Bioinformatics*, **21**, 3840-3845.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer, NY.

Brock, G., Pihur, V., Datta, S., and Datta, S. (2008), clValid, an R package for cluster validation. *Journal of Statistical Software*, **25**, 1-22.

Caliński, T., and Harabasz, J. (1974), A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, **3**, 1-27.

Cheng, Y. and Church, G. (2000), Biclustering of expression data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 93-103.

Davies, D. and Bouldin, D. (1979), A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 224-227.

Dice, L. (1945), Measures of the amount of ecologic association between species. *Ecology*, **26**, 297-302.

Downton, M. and Brennan, T. (1980), Comparing classifications: an evaluation of several coefficients of partition agreement. *Class. Soc. Bull*, **4**, 53-54.

Dunn, J. (1973), A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybernetics and Systems*, **3**, 32-57.

Fowlkes, E. and Mallows, C. (1983), A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**, 553-569.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001), On clustering validation techniques, *Journal of Intelligent Information Systems*, **17**, 107-145.

Handl, J., Knowles, J., and Kell, D. (2005), Computational cluster validation in post-genomic data analysis, *Bioinformatics*, **21**, 3201-3212.

Hubert, L. and Arabie, P. (1985), Comparing partitions. *Journal of Classification*, **2**, 193-218.

Jain, A. and Dubes, R. (1988), *Algorithms for clustering data*, Prentice-Hall, Englewood Cliff, NJ.

Liu, X. and Wang, L. (2007), Computing the maximum similarity biclusters of gene expression data. *Bioinformatics*, **23**, 50-56.

Madeira, S. and Oliveira, A. (2004), Biclustering algorithms for biological data analysis: a survey. *IEEE /ACM Transactions on Computational Biology and Bioinformatics*, **1**, 24-45.

Prelić, A., Bleuler, S., Zimmermann, P.,Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006), A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122-1129.

Rand, W. (1971), Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846-850.

Santamaría, R., Quintales, L., and Therón, R. (2007), Methods to bicluster validation and comparison in microarray data. *Lecture Notes in Computer Science: Proceedings of IDEAL'07*, 780-789.

Turner, H., Bailey, T., and Krzanowski, W. (2005), Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics and Data Analysis*, **48**, 235-254.

Xu, R. and Wunsch, D., II (2005), Survey of clustering algorithms, *IEEE Transactions on Neural Networks*, **16**, 645-678.

Yang, Y., Wang, W., Wang, H., and Yu, P. (2002), δ-clusters: capturing subspace correlation in a large data set, *Proceedings. 18th International Conference on Data Engineering*, 517-528.