Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 2353–2363
Pattern Recognition Letters
www.elsevier.com/locate/patrec

New indices for cluster validity assessment

Minho Kim, R.S. Ramakrishna *

Department of Information and Communications, Gwangju Institute of Science and Technology, 1 Oryong-dong, Buk-gu, Gwangju 500-712, Republic of Korea

Received 5 August 2004; received in revised form 15 March 2005
Available online 22 June 2005

Communicated by W. Pedrycz

Abstract

Cluster validation is a technique for finding a set of clusters that best fits natural partitions (of given datasets) without the benefit of any a priori class information. A *cluster validity index* is used to validate the outcome. This paper presents an analysis of design principles implicitly used in defining cluster validity indices and reviews a variety of existing cluster validity indices in the light of these principles. This includes an analysis of their design and performance. Armed with a knowledge of the limitations of existing indices, we proceed to remedy the situation by proposing six new indices. The new indices are evaluated through a series of experiments.

Keywords: Unsupervised learning; Cluster validity index; Clustering algorithm

1. Introduction

Clustering operation attempts to partition a set of objects into several subsets. The idea is that the objects in each subset are indistinguishable under some criterion of similarity. It is one of the most important tasks in data analysis. It finds application in bioinformatics, web data analysis, information retrieval, CRM (customer relationship managements), text mining, and scientific data exploration, to name only a few major areas (Han and Kamber, 2001; Jain et al., 1999; Kim et al., 2001).

Clustering refers to unsupervised learning as opposed to (supervised) classification. That is, it does not make use of class information about each data object during learning. However, most clustering algorithms require that the input parameters be tuned in some way for optimal results. For example, the *K*-means requires *k*, the number of clusters, as an input. As is well known, the quality of the output depends rather strongly on this

* Corresponding author. Tel.: +82 62 970 2217; fax: +82 62 970 2204.
E-mail address: rsr@gist.ac.kr (R.S. Ramakrishna).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.04.007

parameter. In order to suitably adjust (tune) the input parameter, the algorithm has to utilize class information. But this militates against the spirit of pure clustering. In recent times *cluster validity indices* (CVIs) have attracted attention as a means to resolve such contradiction (Akaike, 1974; Bensaid et al., 1996; Bezdek and Pal, 1998; Calinski and Harabasz, 1974; Davies and Bouldin, 1979; Dunn, 1973; Halkidi and Vazirgiannis, 2000, 2001; Kim et al., 2001, 2003; Maulik and Bandyopadhyay, 2002; Schwarz, 1978; Xie and Beni, 1991). CVIs can signal perfect tuning of input parameters for optimal outcome by assuming a minimum (or maximum) value. The quality of the result is embodied in the number of computed clusters and *purity* of each cluster. That is, the final output is optimal if the number of clusters is the same as that which best fits a dataset while at the same time, purity is as high as possible. Here, purity is the sum of data objects in the majority class in each cluster. The number of clusters is related to cluster purity. This can be inferred from the frequently observed phenomenon that cluster purity deteriorates noticeably if the estimated number of clusters is different from that of the given dataset. Therefore, most CVIs focus on finding the optimal number of clusters.

Many conventional CVIs have experimented with new types of intra-cluster and/or inter-cluster distances, and a combination thereof. However, fundamental principles for designing CVIs have rarely been clearly spelt out, if at all. Some CVIs have features, which are quite contrary to the very spirit of CVIs.

In this paper, we present the basic design principles. We also examine existing CVIs under this light. We present some new CVIs that address the drawbacks of existing ones.

For this purpose, we subdivide CVIs into two categories: *summation* type and *ratio* type. The type is determined by the way the intra-cluster distance and the inter-cluster distance are coupled. For many CVIs, averaging is a step in computing intra-cluster distances. It is well known that averaging smears the input. Therefore, it may mask the discriminatory capacity of CVIs.

We propose new CVIs as alternatives to v_{sv} (Kim et al., 2001), SD (Halkidi and Vazirgiannis, 2000), XB (Xie and Beni, 1991), and DB (Davies and Bouldin, 1979). Design principles of conventional CVIs are found wanting in their sweep: clearly they fail to account for all the subtleties of real world applications. We develop a technique to redeem the ratio-type CVIs. We also propose two new CVIs. The improved performance of the proposed CVIs is demonstrated through experiments.

The rest of the paper is organized as follows. Section 2 introduces CVIs. The new CVIs are proposed in Section 3. Experimental studies and conclusions are given in Sections 4 and 5, respectively.

2. Previous works

The performance of many clustering algorithms is critically dependent on the characteristics of the dataset and the input parameters. Improper input parameters may lead to clusters that deviate from those in the dataset. In order to determine the input parameters that lead to clusters that best fit a given dataset, we need reliable guidelines for evaluating the clusters. Popular techniques employ cluster validity indices (CVI).

Most CVIs are usually defined by combining the following pair of evaluation criteria (Berry and Linoff, 1997):

1. *Compactness*: This measures closeness of cluster elements. Typical example is the variance. Of course, variance also indicates how different the members are. However, a low value of variance is an indicator of closeness.
2. *Separability*: This indicates how distinct two clusters are. It computes the 'distance' between two different clusters. The distance between representative objects of two clusters is a good example. This measure has been widely used due to its computational efficiency and effectiveness for hypersphere-shaped clusters.

A good clustering algorithm will have small intra-cluster distances and large inter-cluster distances.

CVIs can be classified into two kinds according to the way these two distances are combined. One of them is the ratio of the intra-cluster distance to

the inter-cluster distance or vice versa. Examples of this kind are: Dunn (Dunn, 1973), DB (Davies and Bouldin, 1979), XB (Xie and Beni, 1991), and I (Maulik and Bandyopadhyay, 2002). The other is the (properly) weighted sum of these two distances. This type is exemplified by: v_{sv} (Kim et al., 2001), SD (Halkidi and Vazirgiannis, 2000), and S_Dbw (Halkidi and Vazirgiannis, 2001).

Dunn is based on the inter-cluster distance (used when a single linkage algorithm chooses two clusters to be merged) and the diameter of a cluster hypersphere. In recent times, several variants of Dunn have been proposed (Bezdek and Pal, 1998). DB is defined by the average of cluster evaluation measures of all the clusters. XB adopted the minimum distance between any pair of clusters and the global average of distances between each data object and clusters as inter- and intra-cluster distances, respectively. Unlike XB, the index I made use of the maximum inter-cluster distance and summed the distances (instead of averaging) multiplied by the number of clusters. v_{sv} chose the minimum inter-cluster distance and the average of cluster variances as its components and they are min–max normalized for the purpose of weighting. SD is defined by a total separation measure and the average of normalized cluster variances. The total separation measured at $nc = nc_{max}$ is multiplied by intra-cluster distances. S_Dbw replaced the total separation with the density of data objects in the middle of two clusters and omitted the weighting factor.

Another classification of CVIs is based on whether they are applied to hard (crisp) or fuzzy clustering evaluation. In hard clustering, each data object is allowed to have non-overlapping inclusion in a cluster, while in fuzzy clustering, it can have overlapping inclusion. The grade of inclusion of a data object x_k in a cluster C_j is denoted by u_{kj}, the degree of membership. XB and I are used in fuzzy clustering. More fuzzy CVIs can be found in (Bezdek et al., 1999; Kim et al., 2003).

3. Analysis and new CVIs

In order to understand how CVIs spot the best clustering result, we need to understand their design principles. The two kinds of CVI described in Section 2 are based on different design approaches. First, consider the summation type CVIs. Fig. 1(a) illustrates the basic design principles. It shows two graphs obtained by plotting intra-cluster distance (dW) and inter-cluster distance (dB) against the number of clusters (nc). These graphs assume ideal conditions. That is, it is assumed that dW increases sharply as nc decreases from $nc_{optimal}$ to $nc_{optimal} - 1$; and dB decreases sharply as nc decreases from $nc_{optimal} + 1$ to $nc_{optimal}$. Due to these assumptions, the graph of the CVI formed by summing these two graphs has a minimum at $nc = nc_{optimal}$; and hence, we can locate the optimal number of clusters of a given dataset. Fig. 1(a) has another hidden assumption involving the weights on $dW$and dB. Improper weights may lead to incorrect decision on the optimal number of clusters. The problem of improper weights is apparent in the experimental results (Section 4). Similar graphs can be found in (Kim et al., 2001).

In order for the CVIs to succeed in distinguishing the optimal clustering result, all the above assumptions should be valid. However, some of the conventional CVIs do not appear to take this validity factor into consideration. This is emphatically so with regard to dW. Let us examine the problem further (Eq. (1)). As explained in Section 2, dW represents compactness of clusters as a whole; and a measure of dW, viz., variance, decreases as compactness increases. The compactness of a cluster changes slowly when nc decreases in the range $nc > nc_{optimal}$ because data objects which are originally in a single natural cluster are divided into several clusters over that range and merging is imperative for some of them. On the contrary, if nc decreases over the range $nc < nc_{optimal}$, sharp decrease in compactness (compared with its value before merging) is expected due to unnecessary merging. In other words, merging is likely to cause steep increase in values of dW. This phenomenon is particularly conspicuous in hierarchical clustering. Fig. 2 illustrates this.

However, dW in v_{sv}, i.e., v_u (in Eq. (1)) is the average of *decompactnesses* (mean absolute deviation in this case) of each cluster. This approach tends to hide the effect of a cluster obtained by unnecessary merging (i.e., steep increase in values

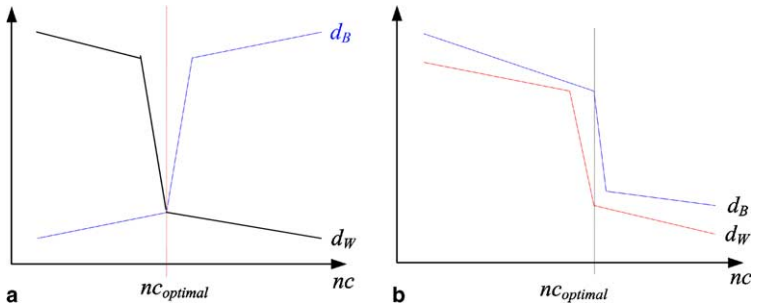

Fig. 1. Design principles for a summation- and a ratio-type CVI: (a) summation-type and (b) ratio-type.

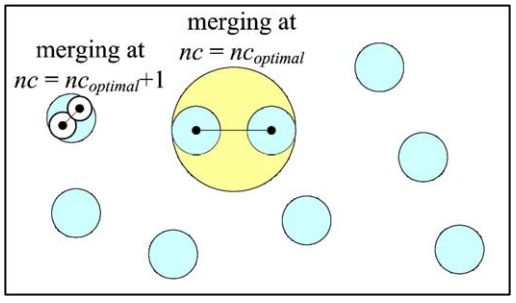

Fig. 2. Illustration of changes in compactness when merging happens at $nc = nc_{\text{optimal}} - 1$ and $nc = nc_{\text{optimal}}$.

of dW) because of averaging. This problem becomes worse when nc_{optimal} is large since the increase of dW caused by unnecessary merging is relatively insignificant for large nc and its effect can easily be masked. For instance, let us assume that the decompactness of cluster C_i, say, $dW(C_i)$, is increased from 5 to 10 due to unnecessary merging when $nc = nc_{\text{optimal}}$. If $nc = 10$, the contribution (effect) of $dW(C_i)$ to dW is just $dW(C_i)/nc = 10/10 = 1.0$. That is, the contribution is reduced from 5 (=10 − 5) to 1. Moreover, if $nc = 20$, then the contribution is further reduced to 10/20 = 0.5. In order to deal with this problem, we propose v_{u}^* as given in Eq. (2). Also, we propose v_{sv}^*, a variant of v_{sv} utilizing v_{u}^*:

$$\begin{aligned} v_{\text{u}}(nc) &= \frac{1}{nc}\sum_{i=1}^{nc}\left(\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i)\right),\\ v_{\text{o}}(nc) &= \frac{nc}{d_{\min}}, \quad d_{\min} = \min_{i\neq j} d(c_i,c_j),\\ v_{\text{sv}}(nc) &= v_{\text{uN}}(nc) + v_{\text{oN}}(nc), \end{aligned} \tag{1}$$

where $v_{\text{uN}}(nc)$ and $v_{\text{oN}}(nc)$ are min–max *normalized versions* of $v_{\text{u}}(nc)$ and $v_{\text{o}}(nc)$, respectively,

$$\begin{aligned} v_{\text{u}}^*(nc) &= \max_{i=1,\ldots,nc}\{v_{\text{u},i}(nc)\}\\ &= \max_{i=1,\ldots,nc}\left\{\frac{1}{n_i}\sum_{x\in C_i} d(x,c_i)\right\},\\ v_{\text{sv}}^*(nc) &= v_{\text{uN}}^*(nc) + v_{\text{oN}}(nc). \end{aligned} \tag{2}$$

As 'Scat' in SD (Eq. (3), where $D_{\max} = \max_{i\neq j} d(c_i,c_j)$, $D_{\min} = \min_{i\neq j} d(c_i,c_j)$) resorts to averaging, the same problem is expected. Hence, we define 'Scat*' as an alternative (Eq. (4)). SD* is also redefined (we did not consider S_Dbw, a variant of SD, because of its high computational cost).

$$\begin{aligned} \text{Scat}(nc) &= \frac{1}{nc}\sum_{i=1}^{nc}\frac{\|\sigma(c_i)\|}{\|\sigma(X)\|},\\ \text{Dis}(nc) &= \frac{D_{\max}}{D_{\min}}\sum_{i=1}^{nc}\left(\sum_{j=1}^{nc} d(c_i,c_j)\right)^{-1},\\ \text{SD}(nc) &= a\cdot\text{Scat}(nc) + \text{Dis}(nc),\\ a &= \text{Dis}(nc_{\max}); \end{aligned} \tag{3}$$

$$\begin{aligned} \text{Scat}^*(nc) &= \max_{i=1,\ldots,nc}\left\{\frac{\|\sigma(c_i)\|}{\|\sigma(X)\|}\right\},\\ \text{SD}^*(nc) &= a\cdot\text{Scat}^*(nc) + \text{Dis}(nc),\\ a &= \text{Dis}(nc_{\max}). \end{aligned} \tag{4}$$

Now let us examine the design principles of ratio type CVIs (please see Fig. 1(b)). We observe the following phenomena:

(i) There is a steep increase in dW when nc decreases from nc_{optimal} to $nc_{\text{optimal}} - 1$.

(ii) There is a sharp increase in dB when nc decreases from $nc_{\text{optimal}} + 1$ to nc_{optimal}.

We investigate how these assumptions influence the CVIs. The value of a ratio type CVI is dependent on the relative values of dW and dB. Let the relative values of dW and dB be $dW'(nc) = dW(nc)/dW(nc_0)$ and $dB'(nc) = dB(nc)/dB(nc_0)$, respectively. Let CVI be the ratio of dW to dB, i.e., $\text{CVI}(nc) = dW(nc)/dB(nc)$. We have,

(i) $dW'(nc) > dB'(nc) \Rightarrow \text{CVI}(nc) > \text{CVI}(nc_0)$,
(ii) $dW'(nc) < dB'(nc) \Rightarrow \text{CVI}(nc) < \text{CVI}(nc_0)$.

As shown in Fig. 1(b), dW has a relatively larger increase at $nc = nc_{\text{optimal}}$ compared with values at $nc > nc_{\text{optimal}}$ than has dB. That is, $dW'(nc_{\text{optimal}}) \ll dB'(nc_{\text{optimal}})$. However, relative changes of values of dW and dB are similar at $nc \neq nc_{\text{optimal}}$. Therefore, CVI has the minimum value at $nc = nc_{\text{optimal}}$. If CVI is defined by dB/dW, it has the maximum value at $nc = nc_{\text{optimal}}$.

Ratio type CVIs also face the same problem as faced by summation type CVIs if they utilize the average (de)compactness as do v_{sv} and SD. We can observe the problem in XB (Eq. (5)). This can be addressed in the same way. Hence, XB is redefined as XB^* (Eq. (6)):

$$\text{XB}(nc) = \frac{\sum_{k=1}^{nc}\sum_{j=1}^{N} u_{kj}^2 d(x_j, c_k)^2}{N \cdot \min_{i,j\neq i} d(c_i, c_j)^2}, \tag{5}$$

$$\text{XB}^*(nc) = \frac{\max_{k=1,\ldots,nc}\left\{\frac{\sum_{j=1}^{N} u_{kj}^2 \|x_j - c_k\|^2}{n_k}\right\}}{\min_{i,l\neq i}\|c_i - c_l\|^2}. \tag{6}$$

Fig. 1(b) has another hidden assumption. It refers to the proviso that the indicated pattern of dW and dB around $nc = nc_{\text{optimal}}$ occurs only once. However, this is not guaranteed in real-world applications. For example, if a heap of clusters with similar dW are separated by similar distances, the pattern can occur several times at $nc < nc_{\text{optimal}}$, and moreover, they may have even shaper changes. That is, even at $nc < nc_{\text{optimal}}$, there may be situations where $dW'(nc) < dB'(nc)$, i.e., $\text{CVI}(nc) < \text{CVI}(nc_{\text{optimal}})$ when we set $dW'(nc)$ and $dB'(nc)$ as $dW(nc)/dW(nc_{\text{optimal}})$ and $dB(nc)/dB(nc_{\text{optimal}})$, respectively.

Before presenting alternatives to alleviate this problem, we make a few observations. As for $nc < nc_{\text{optimal}}$, an increase in the value of dW leads to an increase in the value of CVI. This is desirable in order for CVI to have the minimum at $nc = nc_{\text{optimal}}$. Let us define two terms as follows: $\text{diff}_{dW} = dW(nc) - dW(nc+1)$ and $\text{maxDiff}(nc) = \max_{n_{\max},\ldots,nc}\text{diff}_{dW} \cdot \text{maxDiff}(nc)$ has the following properties:

(i) Relatively small values at $nc \geqslant nc_{\text{optimal}}$,
(ii) Large values only at $nc < nc_{\text{optimal}}$.

These are desirable properties of dW. Furthermore, maxDiff(nc) can augment dW without any side effect due to these properties. Thus, we propose a new CVI, XB^{**}, augmented by maxDiff(nc) from XB^* as

$$\text{XB}^{**}(nc) = \frac{\max_{k=1,\ldots,nc}\left\{\frac{\sum_{j=1}^{N} u_{kj}^2 \|x_j - c_k\|^2}{n_k}\right\} + \text{maxDiff}(nc)}{\min_{i,j\neq i}\|c_i - c_j\|^2}. \tag{7}$$

Among ratio type CVIs, DB (in Eq. (8)) contains another assumption besides those explained above. As defined in Eq. (8), DB is the average of the maximum of R_{ij} of each cluster. R_{ij} has the maximum in the following situations:

(i) when d_{ij} dominates,
(ii) when $(S_i + S_j)$ dominates,
(iii) by a combination of d_{ij} and $(S_i + S_j)$.

Here 'dominate' means that it is a decisive factor in determining $\max\{R_{ij}\}$. In case of (i), d_{ij} has a relatively small value compared with $(S_i + S_j)$ and is usually $\min\{d_{ij}\}$. This indicates a situation where two clusters are located very close to each other and they need to be merged. In case of (ii), $(S_i + S_j)$ has relatively very large value, usually when $(S_i + S_j) = \max\{(S_i + S_j)\}$. This means that an unnecessary merging has taken place. Finally (case (iii)), R_{ij} has the maximum value when neither $\min\{d_{ij}\}$ nor $\max\{(S_i + S_j)\}$ occurs, i.e., by some combination of d_{ij} and $(S_i + S_j)$. Summing up, we observe that DB can effectively have the optimal

value, that is, the minimum value when d_{ij} dominates at $nc > nc_{\text{optimal}}$ and $(S_i + S_j)$ dominates at $nc < nc_{\text{optimal}}$. From the assumption that in an ideal situation $1/\min\{d_{ij}\}$ and $\max\{(S_i + S_j)\}$ have relatively large values when $nc > nc_{\text{optimal}}$ and $nc < nc_{\text{optimal}}$, respectively, as shown in Fig. 1(b), we can redefine DB as DB* (Eq. (9)):

$$\mathrm{DB}(nc) = \frac{1}{nc}\sum_{i=1}^{nc}\left(\max_{j=1,\ldots,nc,j\neq i} R_{ij}\right), \tag{8}$$

where $R_{ij} = \frac{S_i+S_j}{d_{ij}}$, $S_i = \frac{1}{n_i}\sum_{x\in C_i} d(x, c_i)$, and $d_{ij} = d(c_i, c_j)$;

$$\mathrm{DB}^*(nc) = \frac{1}{nc}\sum_{i=1}^{nc}\left(\frac{\max_{k=1,\ldots,nc,k\neq i}\{S_i + S_k\}}{\min_{l=1,\ldots,nc,l\neq i}\{d_{i,l}\}}\right). \tag{9}$$

We can augment DB* by exploiting the intuition about maxDiff(nc) about XB** for a similar reason. The new definition is

$$\begin{aligned}
&\mathrm{DB}^{**}(nc)\\
&\quad = \frac{1}{nc}\sum_{i=1}^{nc}\left(\frac{\max_{k=1,\ldots,nc,k\neq i}\{S_i + S_k\} + \mathrm{maxDiff}_i(nc)}{\min_{l=1,\ldots,nc,l\neq i}\{d_{i,l}\}}\right),\\
&\mathrm{maxDiff}_i(nc) = \max_{nc_{\max},\ldots,nc} \mathrm{diff}_i(nc),\\
&\mathrm{diff}_i(nc) = \max_{k=1,\ldots,nc,k\neq i}\{S_i(nc) + S_k(nc)\}\\
&\qquad\qquad - \max_{k=1,\ldots,nc+1,k\neq i}\{S_i(nc+1) + S_k(nc+1)\}.
\end{aligned} \tag{10}$$

Performance comparisons of new CVIs proposed in this section as well as conventional CVIs are undertaken in Section 4 through comprehensive experiments.

4. Experimental results

In this section, we evaluate performance of not only the conventional CVIs but also the proposed CVIs. As for the clustering algorithm, we adopt a variant of hierarchical clustering with refinement by K-means at $nc = \sqrt{N}$ and $nc = nc^*$, where nc^* is the number of clusters suggested by a CVI. Since we focus on a crisp (hard) clustering and its validation, the membership degree, u_{kj}, of index I (given in Eq. (11)), XB, and XB's variants in Eqs. (5)–(7) is either 0 or 1. The examination of fuzzy clustering is beyond the scope of this paper:

$$\begin{aligned}
&\mathrm{I}(nc) = \left(\frac{1}{nc}\times\frac{E_1}{E_{nc}}\times D_{nc}\right)^p, \quad p = 2,\\
&E_{nc} = \sum_{k=1}^{nc}\sum_{j=1}^{N} u_{kj}\|x_j - c_k\|, \quad D_{nc} = \max_{i\neq j}\|c_i - c_j\|.
\end{aligned} \tag{11}$$

We make use of eleven datasets: seven of them are synthetic datasets (named Dataset x). The others are X30, Bensaid, Starfield', Iris, and NE. Datasets x's are generated as a two-dimensional dataset to visually check whether CVIs succeed in finding the optimal number of clusters. Three of them are given in Fig. 3. Datasets x are somewhat elliptical in shape. Dataset 3 is a variant of Dataset 2 obtained by controlling distances among clusters as well as width and height of each ellipse. Dataset 5 is generated from Dataset 4. X30 and Bensaid were first introduced to demonstrate their work in (Bezdek and Pal, 1998; Bensaid et al., 1996). Starfield' is a superset of the Starfield dataset (Kim et al., 2003). Each data in the Starfield dataset corresponds to the spatial position of bright stars near Polaris. Iris dataset is four-dimensional and is observed from 3 physical classes. However, it is well known that two classes among them overlap in their numerical representation. Here, clustering is basically based on mathematically defined distance, and the number of primary classes is probably 2. In other words, the optimal nc for the this dataset is debatable (Bezdek and Pal, 1998). NE is a spatial dataset, which consists of postal addresses of three metropolitan cities in US (New York, Philadelphia, and Boston) (Theodoridis, 1996). The number of clusters of each dataset is summarized in Table 1.

In the sequel, three categories of results of performance evaluation are presented. First, the results of conventional CVIs using averaging and those of proposed CVIs remedying the problem caused by the averaging are compared. Second, the performance evaluations of improved ratio-type CVIs are given. Finally, we provide summarized performance evaluation results for various CVIs including the proposed ones in respect of all the test datasets in which the suggested number of clusters of various cluster validity indices and their correctness are exhibited.

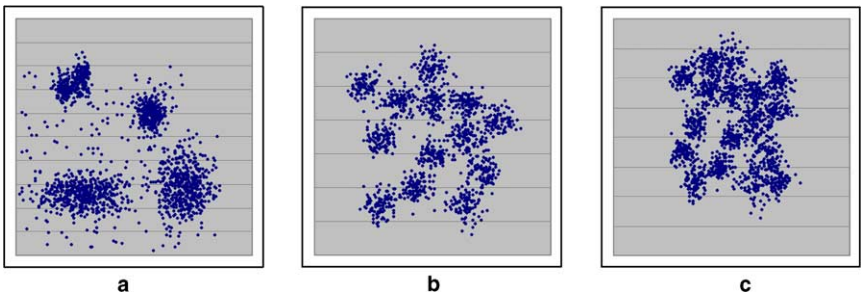

Fig. 3. Synthetic datasets: (a) Dataset 1, (b) Dataset 2 and (c) Dataset 4.

Table 1
The number of clusters of each dataset

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	X30	Bensaid	Starfield	Iris	NE
# clusters	4	13	13	17	17	3	3	9	2 or 3	3

4.1. Effect of averaging in intra-cluster distances and its remedy

First, v_{sv} and v^*_{sv} are compared. Fig. 4 shows experimental results for Dataset 3. The x axis represents the number of clusters and the y axis the parameter values (v_{uN}, v_{oN}, etc.). Note that the graphs in Fig. 4 have 0 value at $nc = 1$. This is due to the assumption that $nc_{optimal} > 1$. Thus, all the values at $nc = 1$ will be ignored. When nc decreases from $nc_{optimal}$ to $nc_{optimal} - 1$, dW should undergo a steep increase as discussed in Section 3. However, such a phenomenon does not occur in the dW graph in Fig. 4(a), i.e., v_{uN} graph when nc decreases from 13 ($nc_{optimal}$) to 12. Even though dB, i.e., v_{oN} shows a property complying with basic design principles of CVIs at $nc = 13$ (steep decrease in v_{oN} graph), v_{sv} fails to pinpoint $nc_{optimal}$ *because of the inherent problems connected with* v_{uN}. That is, v_{sv} has the minimum value at $nc = 11$ instead of at $nc = 13$ ($nc_{optimal}$). On the contrary, v^*_{uN} in Fig. 4(b) rises steeply at $nc = 13$ and hence, v^*_{sv} locates $nc_{optimal}$ exactly. This means that the proposed design approach satisfactorily solves the problem arising due to averaging.

Second, we compare SD with SD*. Note that we did not conduct experiments on S_Dbw since its computational cost is very high. Results on Dataset 2 are given in Fig. 5. We can observe identical effects in the proposed method.

In the third experiment, the performance of XB and XB* is discussed. Comparison of XB and XB* are given in Fig. 6 (Dataset 3). For the purpose of putting three graphs together in one figure, the

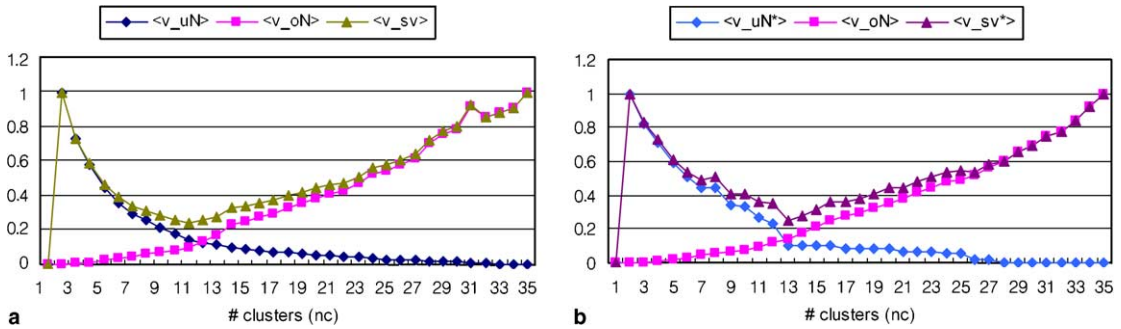

Fig. 4. Comparison of (a) v_{sv} and (b) v^*_{sv} (Dataset 3).

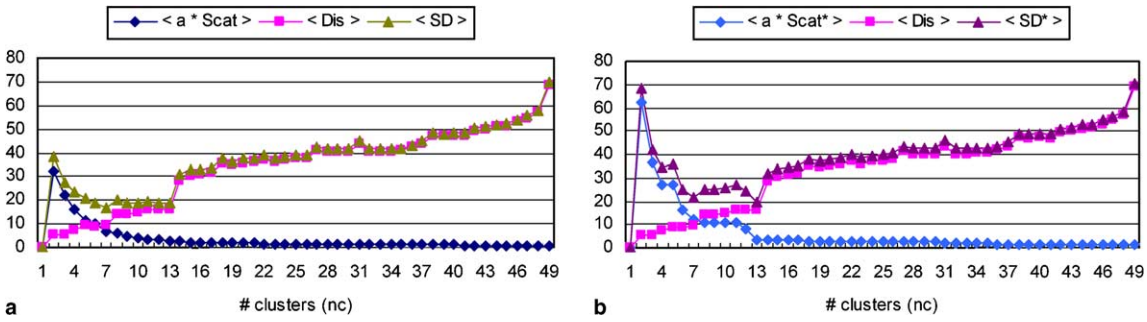

Fig. 5. Comparison of (a) SD and (b) SD* (Dataset 2).

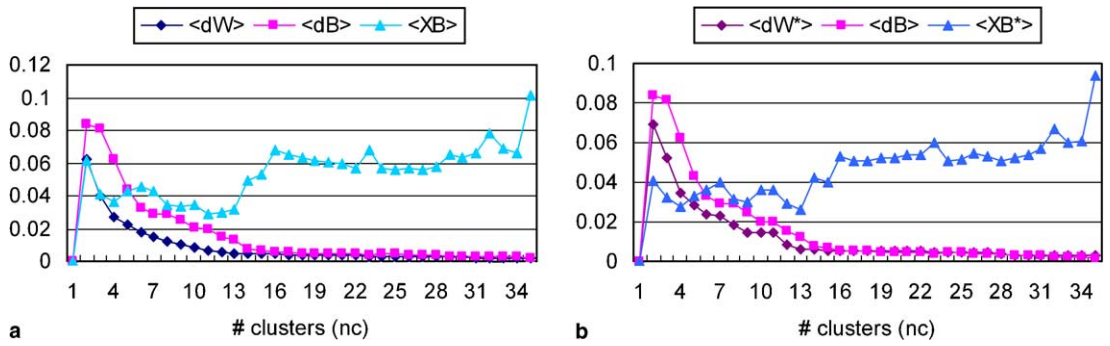

Fig. 6. Comparison of (a) XB and (b) XB* (Dataset 3).

values of some of the measures (dW, dB, and CVI) in the figure have been artificially scaled. Since real values are used in calculating the CVIs, this adjustment has no effect on performance comparisons. This experiment also illustrates how XB* deals with the averaging problem. At first sight, one may not perceive the difference between the dW and the dW^* graphs. However, while dW of XB has 12.8% increase, dW^* of XB* has 34.2% increase when nc changes from 13 ($nc_{optimal}$) to 12. Conversely this means that the proposed approach leads to dW^* having relatively small values at $nc = 13$ compared with other ncs and hence, leads to $dW^{*\prime}(13) < dB'(13)$; that is, XB*(13) is the minimum as discussed in Section 3.

4.2. Performance evaluation of improved ratio-type CVIs

To begin with, we compare the performance of XB* and XB** in Fig. 7. Results in Fig. 7 were obtained from Dataset 4. In Fig. 7(a), XB* fails to have the minimum value at $nc = 17$ ($nc_{optimal}$) although dW^* and dB have (relatively) steep transitions around that nc. The reason is that those patterns occur several times and that one (at $nc = 4$) of them leads to the smaller value of XB* (than the pattern at $nc = 13$). In other words, the ideal pattern (Fig. 1(b)) may not be found in real world applications. Instead, the situation observed in this experiment may be encountered more often. Such phenomenon is also observed in many other experiments. XB** is an augmented version of XB* (through maxDiff(nc)), which has large values only over the range $nc < nc_{optimal}$ as described in Section 3. Therefore, we now have $dW'(nc_{optimal}) > dB'(nc_{optimal})$ (i.e., XB**($nc_{optimal}$) has the minimum). Clearly, the problem has been addressed adequately.

Next, DB, DB*, and DB** are discussed. Fig. 8 provides results on DB and DB* for Dataset 3, which indicates that DB* finds the exact number

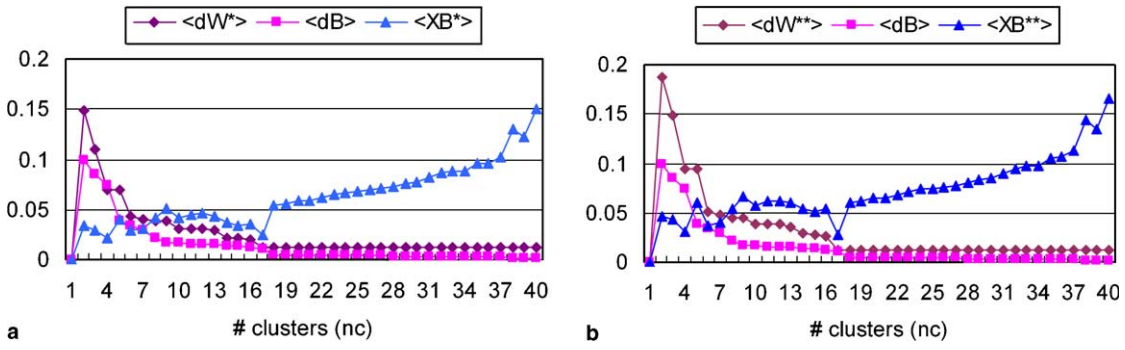

Fig. 7. Comparison of (a) XB* and (b) XB** (Dataset 4).

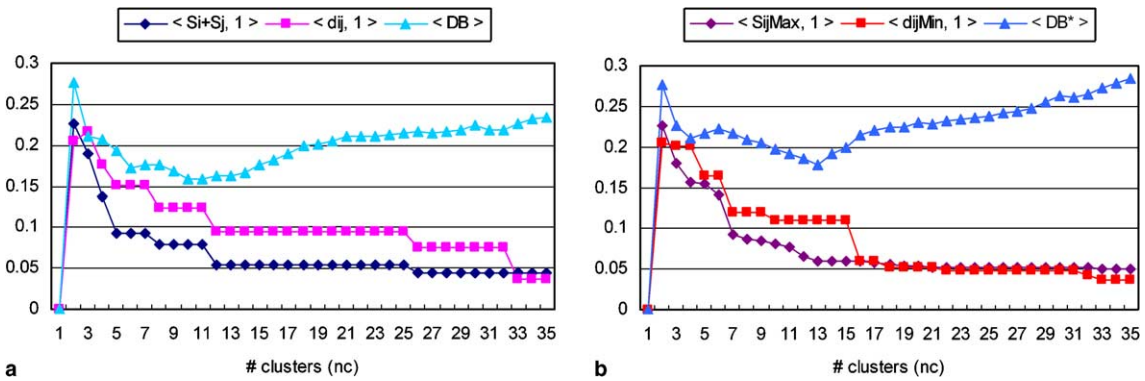

Fig. 8. Comparison of (a) DB and (b) DB* (Dataset 3).

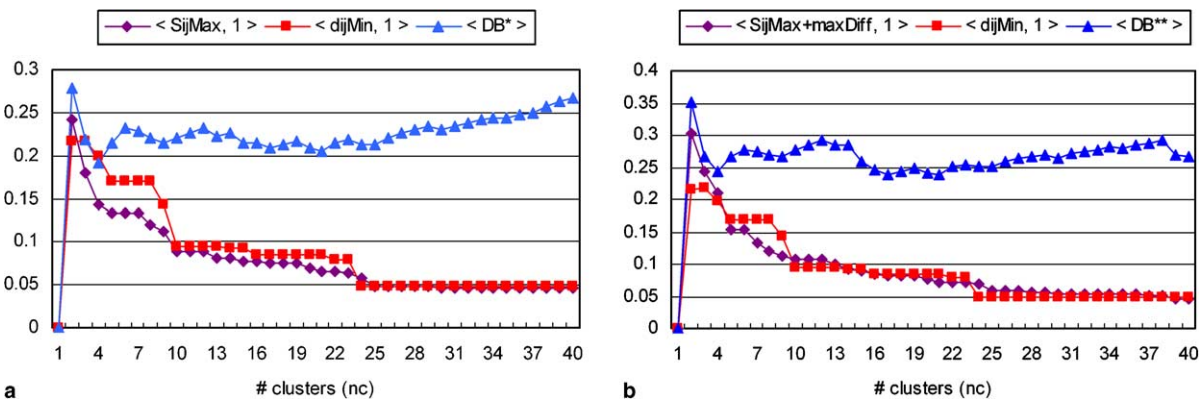

Fig. 9. Comparison of (a) DB* and (b) DB** (Dataset 5).

of natural clusters ($nc = 13 = nc_{\text{optimal}}$) while DB estimates the optimal value to be 11. In Fig. 9, DB* is compared with DB** augmented by max-$\text{Diff}_i(nc)$ from DB*. Results are for Dataset 5.

From the results, we can see that the problem caused by multiple patterns of Fig. 1(b) over $nc < nc_{optimal}$ has been alleviated.

4.3. Summary of performance evaluations

Table 2 summarizes experimental results on CVIs for various datasets. The table shows that the CVIs may be ordered as follows: $v_{sv} < SD < I < XB < DB$. One of the reasons why CVIs of the DB class have the best results is that the averaging used in DB and its variants is different from that in other CVIs. While the other CVIs make use of averaging with respect to dW of each cluster, CVIs of the DB class perform averaging with respect to the results produced by combining information in dW and dB of each class. That is, CVIs of the DB class integrate information of each class, each of which preserves pattern information (of Fig. 1(b)). Additionally, they can lessen the effect of outliers.

In view of the larger scope, the ratio type CVIs show better results than do summation type CVIs. This is due to the weighting problem discussed in Section 3. This problem is shown in Fig. 10, which shows the result in respect of v^*_{sv} for Dataset 1. Although v^*_{uN} and v_{oN} generate proper patterns at $nc = 4$ ($nc_{optimal}$), v^*_{sv} fails to have the minimum value at that point. This happens because v^*_{uN} has relatively larger value than v_{oN} at $nc = 4$, which implies that weighting for v^*_u and v_o (i.e., min–max normalization) is inadequate. Summation type CVIs always have the drawback of adequate weighting. This is another parameter tuning problem.

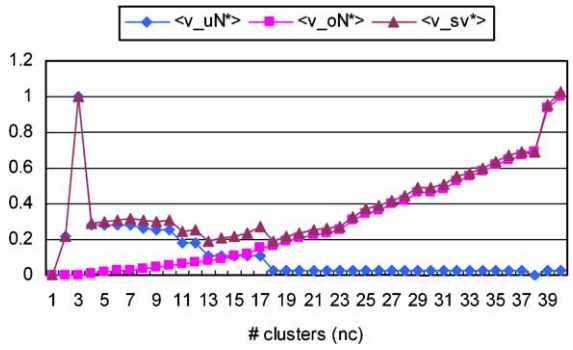

Fig. 10. Weakness found in v^*_{sv} (Dataset 1).

Among the CVIs in the Table 2, v_{sv} and SD has the worst result and DB** the best. Augmentation by $\max Diff_i(nc)$ in addition to advantages discussed so far enables DB** to show the best performance.

5. Conclusions

In this paper, we have provided two kinds of basic design principles for summation- and ratio-type CVIs and critically scrutinized various conventional CVIs from that vantage point. We have also proposed new CVIs that do not suffer from the usual drawbacks of the traditional CVIs. Analyses and experimental studies show that many of the CVIs that compute intra-cluster distance (dW) by averaging effectively filter out the steep

Table 2
Suggested number of clusters by various cluster validity indices and their correctness for various datasets

	v_{sv}	v^*_{sv}	SD	SD*	I	XB	XB*	XB**	DB	DB*	DB**
Dataset 1	15 (×)	18 (×)	4 (○)	4 (○)	4 (○)	4 (○)	4 (○)	4 (○)	4 (○)	4 (○)	4 (○)
Dataset 2	13 (○)	13 (○)	7 (×)	13 (○)	13 (○)	13 (○)	13 (○)	13 (○)	13 (○)	13 (○)	13 (○)
Dataset 3	11 (×)	13 (○)	6 (×)	11 (×)	11 (×)	11 (×)	13 (○)	13 (○)	11 (×)	13 (○)	13 (○)
Dataset 4	15 (×)	16 (×)	7 (×)	6 (×)	6 (×)	4 (×)	4 (×)	17 (○)	8 (×)	17 (○)	17 (○)
Dataset 5	16 (×)	18 (×)	7 (×)	6 (×)	4 (×)	4 (×)	4 (×)	4 (×)	17 (○)	4 (×)	17 (○)
X30	3 (○)	3 (○)	2 (×)	2 (×)	4 (×)	3 (○)	3 (○)	3 (○)	3 (○)	3 (○)	3 (○)
Bensaid	3 (○)	3 (○)	3 (○)	3 (○)	5 (×)	3 (○)	3 (○)	3 (○)	3 (○)	3 (○)	3 (○)
Starfield	2 (×)	6 (×)	3 (×)	3 (×)	5 (×)	9 (○)	9 (○)	9 (○)	9 (○)	9 (○)	9 (○)
Iris	7 (×)	4 (×)	7 (×)	5 (×)	3 (○)	2 (○)	2 (○)	2 (○)	2 (○)	2 (○)	3 (○)
NE	6 (×)	6 (×)	2 (○)	2 (○)	3 (○)	2 (○)	2 (○)	2 (○)	2 (○)	2 (○)	3 (○)
# correct	3	4	3	4	4	7	8	9	8	9	10

Correctness is marked by ○ or × in the parenthesis.

transition of intra-cluster distance of a cluster, caused by unnecessary merging. That is, it turns out that averaging violates the basic design guidelines. We have proposed a new approach that defines dW as the maximum of the intra-cluster distances of each cluster. New CVIs, i.e., v_{sv}^{*}, SD*, and XB* have also been proposed as a result. The proposed CVIs show improved performance. We have also suggested that maxDiff can be beneficial to ratio type CVIs even when the datasets are quite complex. It was applied to XB* and DB* to obtain XB** and DB**. The ratio type CVIs perform better than do summation types. v_{sv} and SD has returned the worst results and DB** the best results, which is consistent with their type. The CVIs proposed in this paper are expected to facilitate wider acceptance of clustering algorithms.

Acknowledgements

This work was supported by the Gwangju Institute of Science and Technology (GIST) through the Campus Internationalization project and the Ministry of Education (MOE) through the Brain Korea 21 (BK21) project.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723.

Bensaid, A.M. et al., 1996. Validity-guided (re)clustering with applications to image segmentation. IEEE Trans. Fuzzy Syst. 4 (2), 112–123.

Berry, M.J.A., Linoff, G., 1997. Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons, Berlin.

Bezdek, J.C., Pal, N.R., 1998. Some new indexes of cluster validity. IEEE Trans. Syst. Man, Cyber.—Part B 28 (3), 301–315.

Bezdek, J.C. et al., 1999. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers, Dordrecht.

Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Comm. Statist. 3, 1–27.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intel. (PAMI) 1 (2), 224–227.

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cyber. 3, 32–57.

Halkidi, M., Vazirgiannis, M., 2000. Quality scheme assessment in the clustering process. In: Proc. PKDD (Principles and Practice of Knowledge Discovery in Databases), Lyo, France. Lecture Notes in Artificial Intelligence. Springer-Verlag GmbH, vol. 1910, pp. 265–276.

Halkidi, M., Vazirgiannis, M., 2001. Clustering validity assessment: finding the optimal partitioning of a dataset. Proc. Internat. Conf. Data Mining (ICDM), California, USA. pp. 187–194.

Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann, Los Altos, CA.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: A review. ACM Comput. Surveys 31 (3), 264–323.

Kim, D.-J., Park, Y.-W., Park, D.-J., 2001. A novel validity index for determination of the optimal number of clusters. IEICE Trans. Inform. Syst. E84-D (2), 281–285.

Kim, D.-W., Lee, K.H., Lee, D., 2003. Fuzzy cluster validation index based on inter-cluster proximity. Pattern Recognition Lett. 24, 2561–2574.

Maulik, U., Bandyopadhyay, S., 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Anal. Mach. Intel. (PAMI) 24 (12), 1650–1654.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.

Theodoridis, Y., 1996. Spatial datasets—An unofficial collection. Available from: <http://www.dias.cti.gr/~ytheod/research/datasets/spatial.html>.

Xie, X.L., Beni, G.A., 1991. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intel. (PAMI) 3 (8), 841–846.