Clustering Validity Checking Methods: Part II

Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis
Department of Informatics,
Athens University of Economics & Business
Email: {mhalk, yannis, mvazirg}@aueb.gr

ABSTRACT
Clustering results validation is an important topic in the context of pattern recognition. We review approaches and systems in this context. In the first part of this paper we presented clustering validity checking approaches based on internal and external criteria. In the second, current part, we present a review of clustering validity approaches based on relative criteria. Also we discuss the results of an experimental study based on widely known validity indices. Finally the paper illustrates the issues that are under-addressed by the recent approaches and proposes the research directions in the field.

Keywords: clustering validation, pattern discovery, unsupervised learning.

1. INTRODUCTION
Clustering is perceived as an unsupervised process since there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data. As a consequence, the final partitions of a data set require some sort of evaluation in most applications [15]. For instance questions like "how many clusters are there in the data set?", "does the resulting clustering scheme fits our data set?", "is there a better partitioning for our data set?" call for clustering results validation and are the subjects of a number of methods discussed in the literature. They aim at the quantitative evaluation of the results of the clustering algorithms and are known under the general term *cluster validity* methods.

In Part I of the paper we introduced the fundamental concepts of *cluster validity* and we presented a review of clustering validity indices that are based on *external* and *internal* criteria. As mentioned in Part I these approaches are based on statistical tests and their major drawback is their high computational cost. Moreover, the indices related to these approaches aim at measuring the degree to which a data set confirms an a-priori specified scheme.

On the other hand, clustering validity approaches which are based on *relative criteria* aim at finding the best clustering scheme that a clustering algorithm can define under certain assumptions and parameters. Here the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different parameter values.

The remainder of the paper is organized as follows. In the next section we discuss the fundamental concepts and representative indices for validity approaches based on relative criteria. In Section 3 an experimental study based on some of these validity indices is presented, using synthetic and real data sets. We conclude in Section 4 by summarizing and providing the trends in cluster validity.

2. Relative Criteria.
The basis of the above described validation methods is statistical testing. Thus, the major drawback of techniques based on internal or external criteria is their high computational demands. A different validation approach is discussed in this section. It is based on relative criteria and does not involve statistical tests. The fundamental idea of this approach is to choose the best clustering scheme of a set of defined schemes according to a pre-specified criterion. More specifically, the problem can be stated as follows:

" Let P_{alg} be the set of parameters associated with a specific clustering algorithm (e.g. the number of clusters n_c). Among the clustering schemes C_i, i=1,.., n_c, defined by a specific algorithm, for different values of the parameters in P_{alg}, choose the one that best fits the data set."

Then, we can consider the following cases of the problem:

I) ***P_{alg}* does not contain the number of clusters, *n_c*, as a parameter**. In this case, the choice of the optimal parameter values are described as follows: We run the algorithm for a wide range of its parameters' values and we choose the largest range for which n_c remains constant (usually $n_c << N$ (number of tuples)). Then we choose as appropriate values of the P_{alg} parameters the values that correspond to the middle of this range. Also, this procedure identifies the number of clusters that underlie our data set.

II) ***P_{alg}* contains *n_c* as a parameter**. The procedure of identifying the best clustering scheme is based on a

validity index. Selecting a suitable performance index, q, we proceed with the following steps:

- the clustering algorithm is run for all values of n_c between a minimum n_{cmin} and a maximum n_{cmax}. The minimum and maximum values have been defined a-priori by user.
- For each of the values of n_c, the algorithm is run r times, using different set of values for the other parameters of the algorithm (e.g. different initial conditions).
- The best values of the index q obtained by each n_c is plotted as the function of n_c.

Based on this plot we may identify the best clustering scheme. We have to stress that there are two approaches for defining the best clustering depending on the behaviour of q with respect to n_c. Thus, if the validity index does not exhibit an increasing or decreasing trend as n_c increases we seek the maximum (minimum) of the plot. On the other hand, for indices that increase (or decrease) as the number of clusters increase we search for the values of n_c at which a significant local change in value of the index occurs. This change appears as a "knee" in the plot and it is an indication of the number of clusters underlying the dataset. Moreover, the absence of a knee may be an indication that the data set possesses no clustering structure.

In the sequel, some representative validity indices for crisp and fuzzy clustering are presented.

2.1 Crisp clustering.

Crisp clustering, considers non-overlapping partitions meaning that a data point either belongs to a class or not. In this section discusses validity indices suitable for crisp clustering.

2.1.1 The modified Hubert Γ statistic.

The definition of the modified Hubert Γ [18] statistic is given by the equation

$$\Gamma = (1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P(i,j)\cdot Q(i,j) \qquad (1)$$

where N is the number of objects in a dataset, M=N(N-1)/2, P is the proximity matrix of the data set and Q is an N×N matrix whose (i, j) element is equal to the distance between the representative points (v_{ci}, v_{cj}) of the clusters where the objects x_i and x_j belong.

Similarly, we can define the normalized Hubert Γ statistic, given by equation

$$\hat{\Gamma} = \frac{[(1/M)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X(i,j)-\mu_X)(Y(i,j)-\mu_Y)]}{\sigma_X\sigma_Y}. \qquad (2)$$

If the $d(v_{ci}, v_{cj})$ is close to $d(x_i, x_j)$ for i, j=1,2,..,N, P and Q will be in close agreement and the values of Γ and $\hat{\Gamma}$ (normalized Γ) will be high. Conversely, a high value of Γ ($\hat{\Gamma}$) indicates the existence of compact clusters. Thus, in the plot of normalized Γ versus n_c, we seek a significant knee that corresponds to a significant increase of normalized Γ. The number of clusters at which the knee occurs is an indication of the number of clusters that occurs in the data. We note, that for $n_c = 1$ and $n_c = N$ the index is not defined.

2.1.2 Dunn family of indices.

A cluster validity index for crisp clustering proposed in [5], aims at the identification of "compact and well separated clusters". The index is defined in equation (3) for a specific number of clusters

$$D_{n_c} = \min_{i=1,\ldots,n_c}\left\{\min_{j=i+1,\ldots,n_c}\left(\frac{d(c_i,c_j)}{\max_{k=1,\ldots,n_c} diam(c_k)}\right)\right\} \qquad (3)$$

where $d(c_i, c_j)$ is the dissimilarity function between two clusters c_i and c_j defined as $d(c_i,c_j) = \min_{x\in C_i, y\in C_j} d(x,y)$, and diam(c) is the diameter of a cluster, which may be considered as a measure of clusters' dispersion. The diameter of a cluster *C* can be defined as follows:

$$diam(C) = \max_{x,y\in C} d(x,y) \qquad (4)$$

If the dataset contains compact and well-separated clusters, the distance between the clusters is expected to be large and the diameter of the clusters is expected to be small. Thus, based on the Dunn's index definition, we may conclude that large values of the index indicate the presence of compact and well-separated clusters.

Index D_{nc} does not exhibit any trend with respect to number of clusters. Thus, the maximum in the plot of D_{nc} versus the number of clusters can be an indication of the number of clusters that fits the data.

The problems of the Dunn index are: i) its considerable time complexity, ii) its sensitivity to the presence of noise in datasets, since these are likely to increase the values of diam(c) (i.e., dominator of equation (3)).

Three indices, are proposed in [14] that are more robust to the presence of noise. They are known as Dunn-like indices since they are based on Dunn index. Moreover, the three indices use for their definition the concepts of Minimum Spanning Tree (MST), the relative neighbourhood graph (RNG) and the Gabriel graph respectively [18].

Consider the index based on MST. Let a cluster c_i and the complete graph G_i whose vertices correspond to the vectors of c_i. The weight, w_e, of an edge, *e,* of this graph equals the distance between its two end points, x, y. Let E_i^{MST} be the set of edges of the MST of the graph G_i and e_i^{MST} the edge in E_i^{MST} with the maximum weight. Then the diameter of C_i is defined as the weight of e_i^{MST}. Dunn-like index based on the concept of the MST is given by equation

$$D_{n_c} = \min_{i=1,\ldots,n_c} \left\{ \min_{j=i+1,\ldots,n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1,\ldots,n_c} diam_k^{MST}} \right) \right\} \quad (5)$$

The number of clusters at which D_m^{MST} takes its maximum value indicates the number of clusters in the underlying data. Based on similar arguments we may define the Dunn-like indices for GG and RGN graphs.

2.1.3 The Davies-Bouldin (DB) index.

A similarity measure R_{ij} between the clusters C_i and C_j is defined based on a measure of dispersion of a cluster C_i , let s_i, and a dissimilarity measure between two clusters d_{ij}. The R_{ij} index is defined to satisfy the following conditions [4]:

1. $R_{ij} \geq 0$
2. $R_{ij} = R_{ji}$
3. if $s_i = 0$ and $s_j = 0$ then $R_{ij} = 0$
4. if $s_j > s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik}$
5. if $s_j = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$.

These conditions state that R_{ij} is non-negative and symmetric.

A simple choice for R_{ij} that satisfies the above conditions is [4]:

$$R_{ij} = (s_i + s_j)/d_{ij}. \quad (6)$$

Then the DB index is defined as

$$DB_{n_c} = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \quad (7)$$

$$R_i = \max_{i=1,\ldots,n_c, i \neq j} R_{ij}, \; i = 1,\ldots,n_c$$

It is clear for the above definition that DB_{nc} is the average similarity between each cluster c_i, i=1, …, n_c and its most similar one. It is desirable for the clusters to have the minimum possible similarity to each other; therefore we seek clusterings that minimize DB. The DB_{nc} index exhibits no trends with respect to the number of clusters and thus we seek the minimum value of DB_{nc} in its plot versus the number of clusters.

Some alternative definitions of the dissimilarity between two clusters as well as the dispersion of a cluster, c_i, is defined in [4].

Three variants of the DB_{nc} index are proposed in [14]. They are based on MST, RNG and GG concepts similarly to the cases of the Dunn-like indices.

Other validity indices for crisp clustering have been proposed in [3] and [13]. The implementation of most of these indices is computationally very expensive, especially when the number of clusters and objects in the data set grows very large [19]. In [13], an evaluation study of thirty validity indices proposed in literature is presented. It is based on tiny data sets (about 50 points each) with well-separated clusters. The results of this study [13] place Caliski and Harabasz(1974), Je(2)/Je(1) (1984), C-index (1976), Gamma and Beale among the six best indices. However, it is noted that although the results concerning these methods are encouraging they are likely to be data dependent. Thus, the behaviour of indices may change if different data structures were used. Also, some indices based on a sample of clustering results. A representative example is Je(2)/Je(1) whose computations based only on the information provided by the items involved in the last cluster merge.

2.1.4 RMSSDT, SPR, RS, CD.

This family of validity indices is applicable in the cases that hierarchical algorithms are used to cluster the datasets. Hereafter we refer to the definitions of four validity indices, which have to be used simultaneously to determine the number of clusters existing in the data set. These four indices are applied to each step of a *hierarchical* clustering algorithm and they are known as [16]:

- *Root-mean-square standard deviation (RMSSTD) of the new cluster*
- *Semi-partial R-squared (SPR)*
- *R-squared (RS)*
- *Distance between two clusters (CD).*

Getting into a more detailed description of them we can say that:

RMSSTD of a new clustering scheme defined at a level of a clustering hierarchy is the square root of the variance of all the variables (attributes used in the clustering process). This index measures the homogeneity of the formed clusters at each step of the hierarchical algorithm. Since the objective of cluster analysis is to form homogeneous groups the RMSSTD of a cluster should be as small as possible. In case that the values of RMSSTD are higher than the ones of the previous step, we have an indication that the new clustering scheme is worse.

In the following definitions we shall use the term SS, which means *Sum of Squares* and refers to the equation:

$$SS = \sum_{i=1}^{N} (X_i - \overline{X})^2 \quad (8)$$

Along with this we shall use some additional symbolism like:

i) SS_w referring to the sum of squares within group,
ii) SS_b referring to the sum of squares between groups.
iii) SS_t referring to the total sum of squares, of the whole data set.

SPR for a the new cluster is the difference between SS_w of the new cluster and the sum of the SS_w's values of clusters joined to obtain the new cluster (*loss of homogeneity*), divided by the SS_t for the whole data set. This index measures the loss of homogeneity after merging the two clusters of a single algorithm step. If the index value is zero then the new cluster is obtained by merging two perfectly homogeneous clusters. If its

value is high then the new cluster is obtained by merging two heterogeneous clusters.

RS of the new cluster is the ratio of SS_b over SS_t. SS_b is a measure of difference between groups. Since $SS_t = SS_b + SS_w$, the greater the SS_b the smaller the SS_w and vise versa. As a result, the greater the differences between groups are the more homogenous each group is and vise versa. Thus, RS may be considered as a measure of dissimilarity between clusters. Furthermore, it measures the degree of homogeneity between groups. The values of RS range between 0 and 1. In case that the value of RS is zero (0) indicates that no difference exists among groups. On the other hand, when RS equals 1 there is an indication of significant difference among groups.

The *CD* index measures the distance between the two clusters that are merged in a given step of the hierarchical clustering. This distance depends on the selected representatives for the hierarchical clustering we perform. For instance, in case of *Centroid hierarchical clustering* the representatives of the formed clusters are the centers of each cluster, so CD is the distance between the centers of the clusters. In case that we use *single linkage* CD measures the minimum Euclidean distance between all possible pairs of points. In case of *complete linkage* CD is the maximum Euclidean distance between all pairs of data points, and so on.

Using these four indices we determine the number of clusters that exist in a data set, plotting a graph of all these indices values for a number of different stages of the clustering algorithm. In this graph we search for the steepest knee, or in other words, the greater jump of these indices' values from higher to smaller number of clusters.

Non-hierarchical clustering. In the case of nonhierarchical clustering (e.g. K-Means) we may also use some of these indices in order to evaluate the resulting clustering. The indices that are more meaningful to use in this case are RMSSTD and RS. The idea, here, is to run the algorithm a number of times for different number of clusters each time. Then the respective graphs of the validity indices is plotted for these clusterings and as the previous example shows, we search for the significant "knee" in these graphs. The number of clusters at which the "knee" is observed indicates the optimal clustering for our data set. In this case the validity indices described before take the following form:

$$RMSSTD = \left[\frac{\sum_{\substack{i=1\ldots nc \\ j=1\ldots v}} \sum_{k=1}^{n_{ij}} (x_k - \overline{x}_k)^2}{\sum_{\substack{i=1\ldots nc \\ j=1\ldots v}} (n_{ij} - 1)} \right] \quad (9)$$

and

$$RS = \frac{SS_b}{SS_t} = \frac{SS_t - SS_w}{SS_t} \Rightarrow$$

$$RS = \frac{\left\{ \sum_{j=1\ldots v} \left[\sum_{k=1}^{n_j} (x_k - \overline{x}_k)^2 \right] \right\} - \left\{ \sum_{\substack{i=1\ldots c \\ j=1\ldots v}} \left[\sum_{k=1}^{n_{ij}} (x_k - \overline{x}_k \right. \right.}{\sum_{j=1\ldots v} \left[\sum_{k=1}^{n_j} (x_k - \overline{x}_k)^2 \right]} \quad (10)$$

where n_c is the number of clusters, d the number of variables(data dimensionality), n_j is the number of data values of j dimension while n_{ij} corresponds to the number of data values of j dimension that belong to cluster i. Also $\overline{x_j}$ is the mean of data values of j dimension.

2.1.5 The SD validity index.

Another clustering validity approach is proposed in [8]. The SD validity index definition is based on the concepts of a*verage scattering for clusters* and *total separation between clusters*. In the sequel, we give the fundamental definition for this index.

Average scattering for clusters. The average scattering for clusters is defined as

$$Scatt(n_c) = \frac{1}{n_c} \sum_{i=1}^{n_c} \|\sigma(v_i)\| / \|\sigma(X)\| \quad (11)$$

Total separation between clusters. The definition of total scattering (separation) between clusters is given by equation (12)

$$Dis(n_c) = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^{n_c} \left(\sum_{z=1}^{n_c} \|v_k - v_z\| \right)^{-1} \quad (12)$$

where $D_{max} = \max(||v_i - v_j||)\ \forall i, j \in \{1, 2,3,\ldots, n_c\}$ is the maximum distance between cluster centers. The $D_{min} = \min(||v_i - v_j||)\ \forall i, j \in \{1, 2,\ldots, n_c\}$ is the minimum distance between cluster centers.

Now, we can define a validity index based on equations (11) and (12) as follows

$$SD(n_c) = a \cdot Scat(n_c) + Dis(n_c) \quad (13)$$

where *a* is a weighting factor equal to $Dis(c_{max})$ where c_{max} is the maximum number of input clusters.

The first term (i.e., $Scat(n_c)$ is defined in equation (11) indicates the average compactness of clusters (i.e., intra-cluster distance). A small value for this term indicates compact clusters and as the scattering within clusters increases (i.e., they become less compact) the value of $Scat(n_c)$ also increases. The second term $Dis(n_c)$ indicates the total separation between the n_c clusters (i.e., an indication of inter-cluster distance). Contrary to the first term the second one, $Dis(n_c)$, is influenced by the geometry of the clusters and increase with the number of clusters. The two terms of *SD* are of the different range, thus a weighting factor is needed in order to incorporate both terms in a balanced way. The number

of clusters, n_c, that minimizes the above index is an optimal value. Also, the influence of the maximum number of clusters c_{max}, related to the weighting factor, in the selection of the optimal clustering scheme is discussed in [8]. It is proved that *SD* proposes an optimal number of clusters almost irrespectively of the c_{max} value.

2.1.6 The S_Dbw validity index.

A recent validity index is proposed in [10]. It exploits the inherent features of clusters to assess the validity of results and select the optimal partitioning for the data under concern. Similarly to the SD index, its definition is based on compactness and separation taking also into account density.

In the following, S_Dbw is formalized based on: i. clusters' compactness (in terms of intra-cluster variance), and ii. density between clusters (in terms of inter-cluster density).

Let $D=\{v_i | i=1,\ldots, n_c\}$ be a partitioning of a data set S into c convex clusters where v_i is the center of each cluster as it results from applying a clustering algorithm to S.

Let *stdev* be the average standard deviation of clusters defined as: $stdev = \frac{1}{c}\sqrt{\sum_{i=1}^{n_c} \|\sigma(v_i)\|}$.

Further the term $||x||$ is defined as : $||x|| = (x^T x)^{1/2}$, where x is a vector.

Then the overall inter-cluster density is defined as:

Definition 1. *Inter-cluster Density* (ID) - It evaluates the average density in the region among clusters in relation to the density of the clusters. The goal is the density among clusters be low in comparison with the density in the clusters. Then, inter-cluster density is defined as follows:

$$Dens\_bw(n_c) = \frac{1}{n_c \cdot (n_c - 1)} \sum_{i=1}^{n_c} \left(\sum_{\substack{j=1 \\ i \neq j}}^{n_c} \frac{density(u_{ij})}{\max\{density(v_i), density(v_j)\}} \right), \quad (14)$$

where v_i, v_j are the centers of clusters c_i, c_j, respectively and u_{ij} the middle point of the line segment defined by the clusters' centers v_i, v_j . The term density(u) defined in equation (15):

$$density(u) = \sum_{l=1}^{n_{ij}} f(x_l, u), \quad (15)$$

where n_{ij} is the number of tuples that belong to the cluster c_i and c_j, i.e., $x_l \in c_i$, $c_j \subseteq S$. It represents the number of points in the neighbourhood of u. In our work, the neighbourhood of a data point, u, is defined to be a hyper-sphere with center u and radius the average standard deviation of the clusters, *stdev*. More specifically, function *f(x,u)* is defined as:

$$f(x,u) = \begin{cases} 0, & \text{if } d(x,u) > stdev \\ 1, & \text{otherwise} \end{cases} \quad (16)$$

A point belongs to the neighbourhood of u if its distance from u is smaller than the average standard deviation of clusters. Here we assume that data ranges are normalized across all dimensions of finding the neighbours of a multidimensional point [1].

Definition 2. *Intra-cluster variance.* It measures the average scattering of clusters. Its definition is given by equation 11.

Then the validity index S_Dbw is defined as:

$$S\_Dbw(n_c) = Scat(n_c) + Dens\_bw(n_c) \quad (17)$$

The definition of S_Dbw indicates that both criteria of "good" clustering (i.e., compactness and separation) are properly combined, enabling reliable evaluation of clustering results. Also, the density variations among clusters are taken into account to achieve in more reliable results. The number of clusters, n_c, that minimizes the above index an optimal value indicating the number of clusters present in the data set.

Moreover, an approach based on the S_Dbw index is proposed in [9]. It evaluates the clustering schemes of a data set as defined by different clustering algorithms and selects the algorithm resulting in optimal partitioning of the data.

In general terms, S_Dbw enables the selection both of the algorithm and its parameters values for which the optimal partitioning of a data set is defined (assuming that the data set presents clustering tendency). However, the index cannot handle properly arbitrarily shaped clusters. The same applies to all the aforementioned indices.

2.2 Fuzzy Clustering.

In this section, we present validity indices suitable for fuzzy clustering. The objective is to seek clustering schemes where most of the vectors of the dataset exhibit high degree of membership in one cluster. Fuzzy clustering is defined by a matrix $U=[u_{ij}]$, where u_{ij} denotes the degree of membership of the vector x_i in cluster j. Also, a set of cluster representatives is defined. Similarly to crisp clustering case a validity index, q, is defined and we search for the minimum or maximum in the plot of q versus n_c. Also, in case that q exhibits a trend with respect to the number of clusters, we seek a significant knee of decrease (or increase) in the plot of q.

In the sequel two categories of fuzzy validity indices are discussed. The first category uses only the memberships values, u_{ij}, of a fuzzy partition of data. The second involves both the U matrix and the dataset itself.

2.2.1 Validity Indices involving only the membership values.

Bezdek proposed in [2] the *partition coefficient*, which is defined as

$$PC = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{n_c} u_{ij}^2 \quad (18)$$

The PC index values range in $[1/n_c, 1]$, where n_c is the number of clusters. The closer to unity the index the "crisper" the clustering is. In case that all membership values to a fuzzy partition are equal, that is, $u_{ij}=1/n_c$, the PC obtains its lower value. Thus, the closer the value of PC is to $1/n_c$, the fuzzier the clustering is. Furthermore, a value close to $1/n_c$ 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 follows

$$PE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_c} u_{ij}\cdot \log_a\left(u_{ij}\right) \qquad (19)$$

where a is the base of the logarithm. The index is computed for values of n_c greater than 1 and its values ranges in $[0, \log_a n_c]$. The closer the value of PE to 0, the crisper the clustering is. As in the previous case, index values close to the upper bound (i.e., $\log_a n_c$), 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 PC) or decrease (for PE) in the plots of the indices versus the number of clusters.,
ii) their sensitivity to the fuzzifier, m. The fuzzifier is a parameter of the fuzzy clustering algorithm and indicates the fuzziness of clustering results. Then, as $m\rightarrow 1$ the indices give the same values for all values of n_c. On the other hand when $m\rightarrow \infty$, both PC and PE exhibit significant knee at $n_c = 2$.
iii) the lack of direct connection to the geometry of the data [3], since they do not use the data itself.

2.2.2 *Indices involving the membership values and the dataset.*

The *Xie-Beni index* [19], XB, also called the compactness and separation validity function, is a representative index of this category.

Consider a fuzzy partition of the data set $X=\{x_j; j=1,.., n\}$ with $v_i (i=1,\ldots, n_c\}$ the centers of each cluster and u_{ij} the membership of data point j with regards to cluster i.

The fuzzy deviation of x_j form cluster i, d_{ij}, is defined as the distance between x_j and the center of cluster weighted by the fuzzy membership of data point j belonging to cluster i.

$$d_{ij}=u_{ij}\,||x_j-v_i||\qquad (20)$$

Also, for a cluster i, the sum of the squares of fuzzy deviation of the data point in X, denoted σ_i, is called variation of cluster i.

The sum of the variations of all clusters, σ, is called *total variation* of the data set.

The term $\pi=(\sigma_i/n_i)$, is called compactness of cluster i. Since n_i is the number of point in cluster belonging to cluster i, π, is the average variation in cluster i.

Also, the separation of the fuzzy partitions is defined as the minimum distance between cluster centers, that is

$$D_{min} = \min||v_i - v_j||$$

Then *XB index* is defined as

$$XB=\pi/N\cdot dmin$$

where N is the number of points in the data set.

It is clear that small values of XB are expected for compact and well-separated clusters. We note, however, that XB is monotonically decreasing when the number of clusters n_c gets very large and close to n. One way to eliminate this decreasing tendency of the index is to determine a starting point, c_{max}, of the monotony behaviour and to search for the minimum value of XB in the range $[2, c_{max}]$. Moreover, the values of the index XB depend on the fuzzifier values, so as if $m\rightarrow\infty$ then $XB\rightarrow\infty$.

Another index of this category is the *Fukuyama-Sugeno index*, which is defined as

$$FS_m = \sum_{i=1}^{N}\sum_{j=1}^{n_c} u_{ij}^m\left(\left\|\mathbf{x_i}-v_\mathbf{j}\right\|_A^2 - \left\|v_j - v\right\|_A^2\right) \qquad (21)$$

where v is the mean vector of X and A is an lxl positive definite, symmetric matrix. When A=I, the above distance becomes the Euclidean distance. It is clear that for compact and well-separated clusters we expect small values for FS_m. The first term in brackets measures the compactness of the clusters while the second one measures the distances of the clusters representatives.

Other fuzzy validity indices are proposed in [6], which are based on the concepts of hypervolume and density. Let Σj the fuzzy covariance matrix of the j-th cluster defined as

$$\Sigma_j = \frac{\sum_{i=1}^{N} u_{ij}^m\left(x_i - v_j\right)\left(x_i - v_j\right)^T}{\sum_{i=1}^{N} u_{ij}^m} \qquad (22)$$

The *fuzzy hyper volume* of j-th cluster is given by equation:

$$V_j = |\Sigma_j|^{1/2}$$

where $|\Sigma_j|$ is the determinant of Σ_j and is a measure of cluster compactness.

Then the *total fuzzy hyper volume* is given by the equation

$$FH = \sum_{j=1}^{n_c} V_j \qquad (23)$$

Small values of FH indicate the existence of compact clusters.

The *average partition density* is also an index of this category. It can be defined as

$$PA = \frac{1}{n_c}\sum_{j=1}^{n_c}\frac{S_j}{V_j} \qquad (24)$$

Then $S_j = \sum_{x\in X_j} u_{ij}$, where X_j is the set of data points that are within a pre-specified region around v_j (i.e., the center of cluster j), is called the sum of the central members of the cluster j.

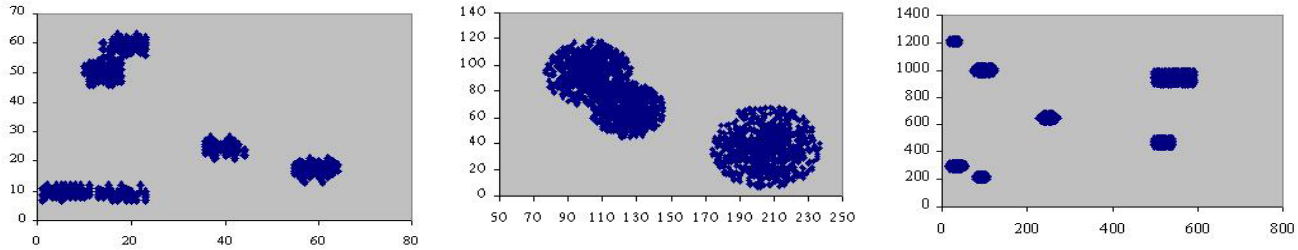

Figure 1.Datasets

Table 1. Optimal number of clusters proposed by validity indices

	DataSet1	DataSet2	DataSet3	Nd_Set
	Optimal number of clusters			
RS, RMSSTD	3	2	5	3
DB	6	3	7	3
SD	4	3	6	3
S_Dbw	4	2	7	3

A different measure is the *partition density index* t defined as

$$PD=S/FH \tag{25}$$

where $S=\sum_{j=1}^{n_c} S_j$.

A few other indices are proposed and discussed in [11, 13].

2.3 Other approaches for cluster validity

Another approach for finding the best number of cluster of a data set was proposed in [17]. It introduces a practical clustering algorithm based on Monte Carlo cross-validation. More specifically, the algorithm consists of M cross-validation runs over M chosen train/test partitions of a data set, D. For each partition u, the EM algorithm is used to define n_c clusters to the training data, while n_c is varied from 1 to c_{max}. Then, the log-likelihood $L_c^u(D)$ is calculated for each model with n_c clusters. It is defined using the probability density function of the data as

$$Lk(D)=\sum_{i=1}^{N}\log f_k(x_i/\Phi_k) \tag{26}$$

where f_k is the probability density function for the data and Φ_k denotes parameters that have been estimated from data. This is repeated M times and the M cross-validated estimates are averaged for each n_c. Based on these estimates we may define the posterior probabilities for each value of the number of clusters n_c, $p(n_c/D)$. If one of $p(n_c/D)$ is near 1, there is strong evidence that the particular number of clusters is the best for our data set.

The evaluation approach proposed in [17] is based on density functions considered for the data set. Thus, it is based on concepts related to probabilistic models in order to estimate the number of clusters, better fitting a data set, and it does not use concepts directly related to the data, (i.e., inter-cluster and intra-clusters distances).

3. AN EXPERIMENTAL STUDY

In this section we present a comparative experimental evaluation of the important validity measures, aiming at illustrating their advantages and disadvantages.

We consider relative validity indices proposed in the literature, such as RS-RMSSTD [16], DB [18] and the most recent ones SD [8], and S_Dbw[10]. The definitions of these validity indices can be found in Section2.1. For our study, we used four synthetic two-dimensional data sets further referred to as DataSet1, DataSet2, DataSet3 (see Figure 1a-c) and a six-dimensional dataset Nd_Set containing three clusters.

Table 1 summarizes the results of the validity indices (RS, RMSSDT, DB, SD and S_Dbw), for different clustering schemes of the above-mentioned data sets as resulting from a clustering algorithm. For our study, we use the results of the algorithms K-Means[12] and CURE[7] with their input value (number of clusters) ranging between 2 and 8. Indices RS, RMSSTD propose the partitioning of DataSet1 into three clusters while DB selects six clusters as the best partitioning. On the other hand, SD and S_Dbw select four clusters as the best partitioning for DataSet1, which is also the correct number of clusters fitting the underlying data. Moreover, the indices S_Dbw and DB select the correct number of clusters(i.e., seven) as the optimal partitioning for DataSet3 while RS, RMSSTD and SD select the clustering scheme of five and six clusters respectively. Also, all indices propose three clusters as the best partitioning for Nd_Set. In the case of DataSet2, DB and SD select three clusters as the optimal scheme, while RS-RMSSDT and S_Dbw select

two clusters (i.e., the correct number of clusters fitting the data set).

Here, we have to mention that a validity index *is not a clustering algorithm itself* but a measure to evaluate the results of clustering algorithms and gives an indication of a partitioning that best fits a data set. The semantics of clustering is not a totally resolved issue and depending on the application domain we may consider different aspects as more significant. For instance, for a specific application it may be important to have well separated clusters while for another to consider more the compactness of the clusters. In the case of S_Dbw, the relative importance of the two terms on which the index definition is based can be adjusted. Having an indication of a good partitioning as proposed by the index, the domain experts may analyse further the validation procedure results. Thus, they could select some of the partitioning schemes proposed by S_Dbw, and select the one better fitting their demands for crisp or overlapping clusters. For instance DataSet2 can be considered as having three clusters with two of them slightly overlapping or having two well-separated clusters. In this case we observe that S_Dbw values for two and three clusters are not significantly different (0.311, 0.324 respectively). This is an indication that we may select either of the two partitioning schemes depending on the clustering interpretation. Then, we compare the values of *Scat* and *Dens_bw* terms for the cases of two and three clusters. We observe that the two clusters scheme corresponds to well-separated clusters (Dens_bw(2)= 0.0976 < Dens_bw(3)= 0.2154) while the three-clusters scheme contains more compact clusters (Scat(2)= 0.21409 >Scat(3)= 0.1089).

4. CONCLUSIONS AND TRENDS IN CLUSTERING VALIDITY

Cluster validity checking is one of the most important issues in cluster analysis related to the inherent features of the data set under concern. It aims at the evaluation of clustering results and the selection of the scheme that best fits the underlying data.

The majority of algorithms are based on certain criteria in order to define the clusters in which a data set can be partitioned. 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. We presented a survey of the most known validity criteria available in literature, classified in three categories: external, internal, and relative. Moreover, we discussed some representative validity indices of these criteria along with a sample experimental evaluation.

As we discussed earlier the validity assessment approaches works better when the clusters are mostly compact. However, there are a number of applications where we have to handle arbitrarily shaped clusters (e.g. spatial data, medicine, biology). In this case the traditional validity criteria (variance, density and its continuity, separation) are not any more sufficient.

There is a need for developing quality measures that assess the quality of the partitioning taking into account i. intra-cluster quality, ii. inter-cluster separation and iii. geometry of the clusters, using sets of representative points, or even multidimensional curves rather than a single representative point.

Another challenge is the definition of an integrated quality assessment model for data mining results. The fundamental concepts and criteria for a global data mining validity checking process have to be introduced and integrated to define a quality model. This will contribute to more efficient usage of data mining techniques for the extraction of valid, interesting, and exploitable patterns of knowledge.

ACKNOWLEDGEMENTS

This work was supported by the General Secretariat for Research and Technology through the PENED ("99EΔ 85") project. We thank C. Amanatidis for his suggestions and his help in the experimental study. Also, we are grateful to C. Rodopoulos for the implementation of CURE algorithm as well as to Dr Eui-Hong (Sam) Han for providing information and the source code for CURE algorithm.

REFERENCES

[1] Michael J. A. Berry, Gordon Linoff . *Data Mining Techniques For marketing, Sales and Customer Support*. John Willey & Sons, Inc, 1996.

[2] Bezdeck, J.C, Ehrlich, R., Full, W.. "FCM:Fuzzy C-Means Algorithm", *Computers and Geoscience,* 1984.

[3] Dave, R. N. . "Validating fuzzy partitions obtained through c-shells clustering*", Pattern Recognition Letters*, Vol .17, pp613-623, 1996.

[4] Davies, DL, Bouldin, D.W. "A cluster separation measure". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 1, No2, 1979.

[5] Dunn, J. C. . "Well separated clusters and optimal fuzzy partitions", *J. Cybern.* Vol.4, pp. 95-104, 1974.

[6] Gath I., Geva A.B. "Unsupervised optimal fuzzy clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 11(7), 1989.

[7] Guha, S., Rastogi, R., Shim K. (1998). "CURE: An Efficient Clustering Algorithm for Large Databases", *Published in the Proceedings of the ACM SIGMOD Conference*.

[8] Halkidi, M., Vazirgiannis, M., Batistakis, I.. "Quality scheme assessment in the clustering process", *Proceedings of PKDD*, Lyon, France, 2000.

[9] Halkidi M, Vazirgiannis M., "A data set oriented approach for clustering algorithm selection", *Proceedings of PKDD*, Freiburg, Germany, 2001

[10] M. Halkidi, M. Vazirgiannis, "Clustering Validity Assessment: Finding the optimal partitioning of a

data set", to appear in the *Proceedings of ICDM*, California, USA, November 2001.

[11] Krishnapuram, R., Frigui, H., Nasraoui. O. "Quadratic shell clustering algorithms and the detection of second-degree curves", *Pattern Recognition Letters,* Vol. 14(7), 1993

[12] MacQueen, J.B (1967). "Some Methods for Classification and Analysis of Multivariate Observations*", In Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability*, Volume I: Statistics, pp281-297.

[13] Milligan, G.W. and Cooper, M.C.. "An Examination of Procedures for Determining the Number of Clusters in a Data Set", *Psychometrika*, Vol.50, pp 159-179, 1985.

[14] Pal, N.R., Biswas, J.. "Cluster Validation using graph theoretic concepts". *Pattern Recognition,* Vol. 30(6), 1997.

[15] Rezaee, R, Lelieveldt, B.P.F., Reiber, J.H.C. "A new cluster validity index for the fuzzy c-mean", *Pattern Recognition Letters*, 19, pp. 237-246, 1998.

[16] Sharma, S.C.. *Applied Multivariate Techniques*. John Willwy & Sons, 1996.

[17] Smyth, P. "Clustering using Monte Carlo Cross-Validation". *Proceedings of KDD Conference*, 1996.

[18] Theodoridis, S., Koutroubas, K.. *Pattern recognition,* Academic Press, 1999.

[19] Xie, X. L, Beni, G.. "A Validity measure for Fuzzy Clustering*", IEEE Transactions on Pattern Analysis and machine Intelligence,* Vol.13, No4, 1991.