

Evaluating the validity of clustering results based on density criteria and multi-representatives

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

Abstract- Although the goal of clustering is intuitively compelling and its notion arises in many fields, it has been difficult to define a unified approach to address the clustering problem and thus diverse clustering approaches abound in the research community. These approaches are based on different clustering principles and assumptions and they often lead to qualitatively different results. As a consequence the results of clustering algorithms (i.e. data set partitionings) need to be evaluated as regards their validity based on widely accepted criteria.

In this paper a cluster validity index, $CDbw$, is introduced which assesses compactness and separation of the partitions generated by a clustering algorithm. The cluster validity index, given a data set and a set of clustering algorithms, enables: i) the selection of the input parameter values that lead an algorithm to the best possible partitioning of the data set, and ii) the selection of the algorithm that provides the optimal partitioning of the data set. $CDbw$ handles efficiently arbitrarily shaped clusters by representing each cluster with a number of points rather than by a single representative point. The properties of the validity index are theoretically justified. A full implementation and experimental results confirm the reliability of the validity index showing also that its performance compares favorably to that of several others.

1. INTRODUCTION

Cluster analysis aims at providing useful information by organizing data into groups (clusters) such that data in a cluster are more similar to each other than are data belonging to different clusters. The study of clustering is only unified at this very general level of description while at the level of methods and algorithms one can encounter a multitude of clustering techniques (agglomerative, spectral, centroid etc). These techniques are based on diverse underlying principles and assumptions and they often lead to different results [20]. Since clustering is unsupervised and there is no a priori knowledge for data distribution in the underlying set, the significance of the clusters defined for a data set needs to be validated.

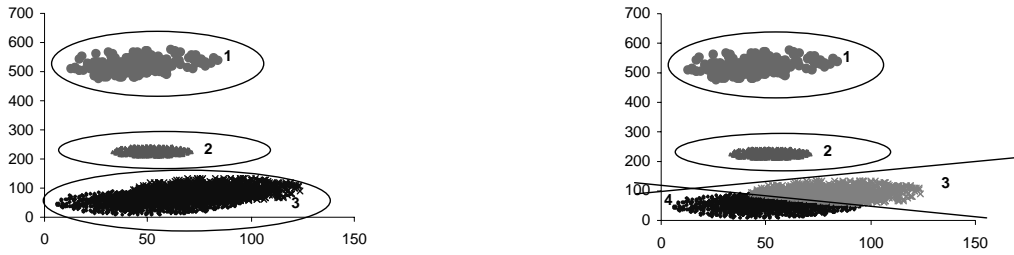


Figure 1: The different partitionings defined by K-Means when it runs with different input parameter values.

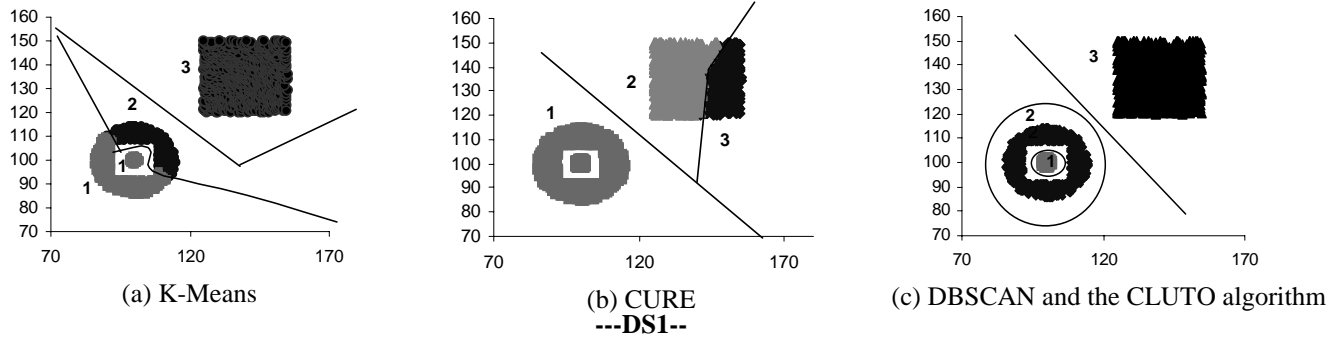


Figure 2: Partitioning of DS1 into three clusters as defined by different clustering algorithms.

The *cluster validity* is a broad issue and subject of endless arguments since the notion of “good” clustering is strictly related to the application domain and its specific requirements. Nevertheless it is generally accepted that the answer to the validity of the clustering results has to be sought in measures of separation among the clusters and cohesion within clusters. Then to define these measures and evaluate clusters we have to take into account specific aspects of their definition. In the context of this paper we address the cluster validity problem based on the density notion of clusters, which is widely accepted in the clustering field. Since the majority of clustering algorithms partition a data set considering that each data point belongs to one and only one cluster, the proposed approach is defined in the context of crisp clustering. However, its extension to fuzzy clustering domain is considered to be a further work issue.

Below we define the terms and concepts that will be used through out the paper.

Given a multi-dimensional data set, the data space is not usually uniformly occupied. This implies that one can identify sparse and crowded (dense) places in the data space. We claim that such a data set (e.g. the data sets in Figure 1 and Figure 2) presents *clustering tendency* or it possesses a *clustering structure* [29].

Assuming that S is a data set presenting clustering tendency and there is a partitioning C of S that represents its dense areas as distinct partitions (i.e. the underlying clusters in S), we call C *actual*

partitioning. Then, the results of a clustering algorithm A applied to S comprise a partitioning of S into a set of clusters that is called *clustering scheme*. If for each cluster C_i there is a partition of the actual partitioning P_j such that $C_i = P_j$ (i.e. contain the same data objects) then we claim that the algorithm discovered the *real* clusters or the *actual partitioning*.

Of course there are cases that an algorithm A applied to S with different input parameter values (for brevity ipvs), results in different clustering schemes none of which resembles the actual partitioning. Among these clustering schemes, the one that is most similar to (approximates) the actual partitioning is further called optimal partitioning¹ of S by A. In other words, the optimal partitioning refers to the best possible partitioning of S among those defined by the clustering approaches. As it will be further discussed, it is important to discover the ipvs for A applied to S that result in the optimal partitioning. In general terms, the validity of clustering results relies on i) the inherent features of the data set under concern (such as geometry and density distribution of clusters), and ii) the clustering criteria and assumptions adopted by the algorithm.

1.1 Motivation

As we have already mentioned *data clustering* is a complex problem and its interpretation varies in different application domains. Thus a multitude of clustering methods has been developed and is available in the literature [20]. However, given a data set and a clustering algorithm running on it with different ipvs, we obtain different partitionings of the data set into clusters. Then we need to decide which the partitioning is that best fits the data set under concern among the defined ones. This, the *cluster validity* problem, is generally accepted as a cornerstone issue of the clustering process.

There are two aspects in checking the validity of clustering results with regard to a data set: i) the choice of the appropriate ipvs for a clustering algorithm, and ii) the choice of the algorithm resulting in the optimal partitioning (as defined above).

In the sequel we motivate these aspects, using examples. As Figure 1 depicts, different ipvs lead to different clustering results (here, the K-Means [2] algorithm is used). The data set is falsely partitioned in most of the cases. Only one set of ipvs (i.e. number of clusters = 3) lead to the actual partitioning of the data set. If there is no prior knowledge about the data structure, it is difficult to find the optimal ipvs for the algorithm under concern.

¹ In the context of this paper the terms “partitioning” and “clustering scheme” are interchangeable.

Cross-algorithm comparison takes place in the example of Figure 2 where different clustering algorithms (K-Means [2], CURE [10], DBSCAN [6], a clustering algorithm provided by the CLUTO toolkit [19]) are used to partition the DS1 data set. The algorithms ran under specific set of ipvs to define a partitioning of DS1 into three clusters. It is clear from Figure 2a and Figure 2b that K-Means and CURE (with $a=0.3$, $r=10$) partition DS1 wrongly into three clusters. On the other hand, DBSCAN and the CLUTO algorithm (see Figure 2c) with suitable ipvs give better results since it partitioned the data set discovering its real three clusters. Hereafter, we use the term “*correct number of clusters*” to refer to the number of clusters in the actual partitioning of a data set while the number of clusters in case of optimal partitioning is further called “*optimal number of clusters*”.

The visualization curse. As the above examples show, visual perception of the clusters structure enables a profound assessment of the partitioning validity. In almost all cases most the experimental evaluation of clustering algorithms [1, 6, 10, 11, 12, 22] is based on 2D-data sets. Thus, the reader is able to visually verify the validity of the results. Hence visualization of the data set is perceived as a useful verification of the clustering results. However, in case of large multi-dimensional data sets (e.g. more than three dimensions), effective visualization can be cumbersome. Moreover the perception of clusters based on visualization is a difficult task for humans not accustomed to higher dimensional spaces. What is needed is a visual-aids-free assessment of some objective criterion, indicating the validity of the results of a clustering algorithm. This should be applicable to a potentially high dimensional data set and handle efficiently arbitrarily shaped clusters (i.e. clusters of non-spherical geometry).

In this paper we define and evaluate a cluster validity index, *CDbw* (Composed Density between and within clusters) and a methodology that, given a data set S , is able to discover, (a) assuming a clustering algorithm and its results with different ipvs, the values of the algorithm’s input parameters that result in the optimal partitioning of S (i.e. the best possible partitioning of S among those defined by the algorithm), assuming a set of algorithms and taking into account the results of (a), the algorithm that returns the optimal partitioning of S .

In some cases a clustering algorithm finds the correct number of clusters but partitions the data set in a wrong way (i.e. it does not find the real clusters). Assuming different clustering algorithms’ sessions²

² *Clustering algorithm session*: the application of the clustering algorithm to a data set using specific values for its input parameters

on the data set under concern, resulting in different partitionings but all containing the correct number of clusters, *CDbw* enables finding the optimal partitioning of a data set among the aforementioned ones. Moreover, *CDbw* adjusts well to non-spherical cluster geometries, contrary to the validity indices proposed in the literature (an overview is presented in [29, 13]). It achieves this by considering multiple representative points per cluster. The validity index is theoretically proved to take its maximum value for the partitioning that best fits the underlying data (herein referred to as the optimal partitioning) among those defined for the data set under concern. The cluster validity index is fully implemented and experiments prove its efficiency for various data sets and clustering algorithms.

The rest of the paper is organized as follows. Section 2 reviews cluster validity related concepts and some cluster validity criteria related to our work. We motivate and define our cluster validity index in Section 3. Section 4 follows presenting a theoretical study of *CDbw*. In Section 5 we describe an experimental study of our approach while we present its comparison to other cluster validity indices. Finally, we conclude in Section 6 by briefly presenting our contributions and indicate directions for further research.

2. RELATED WORK

The fundamental clustering problem is to partition a data set into groups (i.e. clusters), such that the data points in a cluster are more similar to each other than points in different clusters [10]. In the clustering process, there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data [2]. This is what distinguishes clustering from classification [7].

There is a multitude of clustering methods available in the literature, which can be broadly classified into the following categories [11, 20]: i) *Partitional clustering*, ii) *Hierarchical clustering*, iii) *Density-based clustering*, iv) *Grid-based clustering*.

For each of these types there exists a wealth of categories and different algorithms [1, 11, 12] for finding the clusters. In general terms, clustering algorithms are based on a criterion for judging the validity of a given partitioning. Moreover, they define a partitioning of a data set based on certain assumptions and *not* the optimal one that fits the data set.

Since clustering algorithms discover clusters, which are not known a-priori, the final partitioning of a data set requires some sort of evaluation in most applications [24]. Requirements for the evaluation of

clustering results are well known in the research community and a number of efforts have been made especially in the area of pattern recognition. However, the issue of cluster validity is rather under-addressed in the area of databases and data mining applications, even though recognized as important. There are three approaches to investigate cluster validity [29]. The first is based on *external criteria*. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set. The second approach is based on *internal criteria*, meaning that the results of a clustering algorithm are evaluated in terms of quantities that involve the vectors of the data set themselves (e.g. proximity matrix). The third approach is based on *relative criteria*. Here the basic idea is to choose the best clustering scheme of a set of defined schemes according a pre-specified criterion. A number of validity indices appear in the literature for each of the above approaches [29]. A cluster validity index for crisp clustering proposed in [4], attempts to identify “compact and well-separated clusters”. Other validity indices for crisp clustering have been proposed in [3] and [30]. The implementation of most of these indices is computationally expensive, especially when the number of clusters and number of objects in the data set grows a lot [31]. In [21] an evaluation study of thirty validity indices proposed in the literature is presented. The results of this study rate the indices 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, i.e. the characteristics of data can affect their performance in an unpredictable way. Thus there is no guarantee that they will be optimal for real data set. For fuzzy clustering [29], Bezdek proposed the partition coefficient (1974) and the classification entropy (1984). The limitations of these indices are [3]: i) their monotonous dependency on the number of clusters, and ii) the lack of direct connection to the geometry of the data. Other fuzzy validity indices are proposed in [9, 31]. We should mention that the evaluation of proposed indices and the analysis of their reliability are limited.

Another approach for finding the optimal number of clusters in a data set is proposed in [27]. It introduces a practical clustering algorithm based on Monte Carlo cross-validation. This approach differs significantly from the one we propose. While we evaluate clustering schemes based on widely recognized validity criteria of clustering, the evaluation approach proposed in [27] is based on density functions considered for the data set. Thus, it uses concepts related to probabilistic models in order to

estimate the number of clusters, better fitting a data set, while we use concepts directly related to the data.

3. A CLUSTER VALIDITY INDEX BASED ON DENSITY

Following up the examples discussed in the Introduction section, each clustering algorithm provides a partitioning of a data set but does not deal with the validity of the clustering results. For instance the algorithm DBSCAN [6] defines clusters based on density variations, considering values for the cardinality and radius of an object's neighbourhood. On the other hand, K-Means partitions a data set into a pre-specified number of clusters based on an objective function that attempts to minimize the distance of every data object from the center of cluster to which it belongs. Though the aforementioned algorithms attempt to find the best possible partitions for the given input parameter values, there is no indication that the resulting partitions are the optimal or even the ones presented in the data set.

The fundamental criteria for clustering algorithms include *compactness* and *separation* of clusters. However, the clustering algorithms aim at satisfying these criteria based on initial assumptions (e.g. initial locations of the cluster centers) or input parameter values (e.g. the number of clusters, minimum diameter or number of points in a cluster). What is missing is an approach that satisfies a global optimization of the clustering criteria, comparing the different clustering schemes defined for a data set.

An important aspect of such an approach is to define the measures of the clustering criteria, i.e. to define the measures based on which we will evaluate the compactness and separation of clusters. A commonly used measure of clusters compactness is the variance while the average distance between clusters (e.g. distance between clusters centers) is considered to be a standard measure of clusters' separation. However, there are aspects of clusters (e.g. density variations, data scattering) that are not taking into account, though they are strictly related to data. For instance, assume the data set in Figure 2 and its respective partitionings presented in Figure 2b and Figure 2c. If we take into account the variance of clusters as a measure of their compactness, the clusters in Figure 2b will be considered to be more compact than those of Figure 2c. This can be justified if we consider that the cluster variance only measures the distance of points belonging to a cluster from the cluster center while it does not take into account either the distribution of data points in the cluster or any changes observed in the data distribution (that is, dense areas followed by low-density areas and vice versa) within the considered

clusters. With reference to our example, though the cluster 1, depicted in Figure 2b, contains areas of low density and the distribution of data points present significant changes through the cluster, its variance is estimated to be almost the same even not less than the average variance of the clusters 1 and 2, as Figure 2c depicts.

Another issue of concern is the *geometry* of the clusters that has been treated in several algorithms recently [6, 25]. The problem is that when a cluster's geometry is deviating from the hyper-spherical shape algorithms generally have problems detecting them and even in cases that an algorithm achieves to handle arbitrarily shaped clusters, it is based on specific assumptions.

Based on the above observations we propose a cluster validity index, *CDbw*, taking into account a) *density distribution between and within clusters* to assess the compactness and separation of the defined clusters, b) *changes of the density distribution* within clusters to assess the clusters' cohesion, and c) requirements for handling *awkward cluster geometries*.

The cluster's geometry issue is tackled in *CDbw* by considering multiple representative points for each cluster defined by an algorithm. This approach improves geometry-related efficiency compared to other related ones (a survey of cluster validity approaches is presented in [13]) that consider a single representative point per cluster.

In this section, we formalize a cluster validity index putting emphasis on the geometry of clusters. It is a relative validity index since it compares a set of different clustering schemes defined for a data set and selects the one that best fits it. Before we proceed with the definition of the cluster validity index, in Section 3.4, some concepts that are fundamental for *CDbw* are introduced.

3.1 Cluster representative points definition

Let $D = \{V_1, \dots, V_c\}$ be a partitioning of a data set S into c clusters where V_i is the set of representative points of the cluster C_i , such that $V_i = \{v_{i1}, \dots, v_{ir} \mid r = \text{number of representatives per cluster}\}$ and v_{ij} is the j th representative of the cluster C_i . Each cluster is represented by a set of r points that are generated by selecting well-scattered points within the cluster. These r points achieve to capture the geometry of the respective cluster.

The contribution of the proposed approach is the use of multi-representatives in cluster validity so as to capture the shape of the clusters in the clustering evaluation process and not a method to define representative points. Then we select to use one of the widely used approaches, which is based on

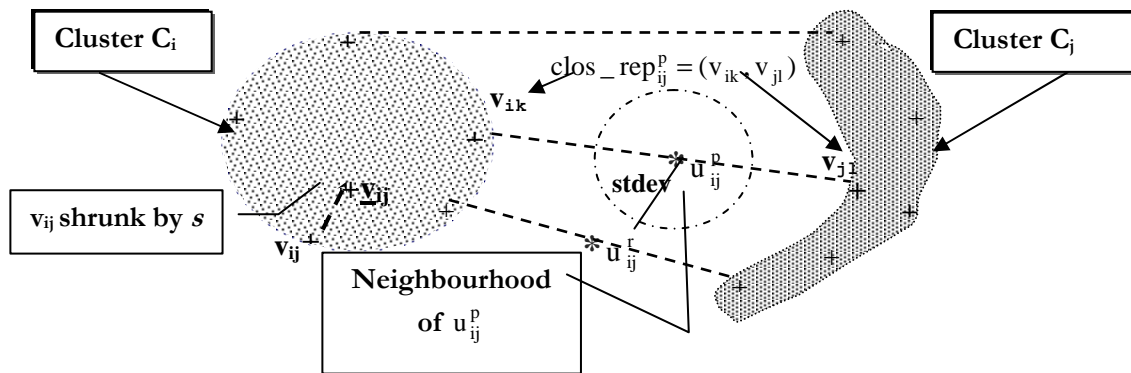


Figure 3. Inter-cluster density definition

“furthest-first” technique [14], to define the representative points of the clusters. The main idea of this procedure is briefly presented below. However, other approaches for finding the clusters’ representative can be used as well.

DEFINITION 1. *Multiple representatives.* Assume a data set S and the set of representative points V_i of cluster C_i . Let S_i be the subset assigned to C_i and $C_i.center$ be the center (centroid) of a cluster (i.e. the mean value of S_i). Then the set of r representative points is defined as follows:

```

Points function Rep_points (Points  $S_i$ , integer  $r$ ) {
  Points  $V_i := \emptyset$ 
  for  $i := 1$  to  $r$  do {
     $max\_dist := 0$ 
    for each point  $p \in S_i$  do {
      if  $i = 1$ 
         $min\_dist := dist(p, C_i.center)$ 
      else
         $min\_dist := \min\{dist(p, q) : q \in V_i\}$ 
        if ( $min\_dist \geq max\_dist$ ) {
           $max\_dist := min\_dist$ 
           $far\_point := p$ 
        }
       $V_i := V_i \cup \{far\_point\}$ 
    }
  }
  Return  $V_i$ 
} //end

```

Algorithm explanation. The procedure for defining the representatives of a cluster C_i is iterative. In the first iteration, the point farthest from the center of the cluster under consideration is chosen as the first representative point. In each subsequent iteration, a point from the cluster is chosen that is farthest from the previously chosen representative points. Thus the function results in a set of points that represent the geometry of the cluster periphery.

The extended analysis on the procedure for selecting the representatives is out of the scope of this paper. However, we carried out a study to examine the influence of the number of representative points, r , to the cluster validity results and we discuss its results in Section 6.

DEFINITION 2.1. *Closest Representative points.* Let V_i and V_j be the set of representatives of the clusters C_i and C_j respectively. A representative point of C_i , let v_{ik} , is considered to be the *closest representative* in C_i of the representative v_{jl} of the cluster C_j , further referred to as $closest_rep^i(v_{jl})$, if v_{ik} is the representative point of C_i with the minimum distance from v_{jl} , i.e. $d(v_{jl}, v_{ik}) = \min_{v_{ix} \in V_i} \{d(v_{jl}, v_{ix})\}$,

where d is the Euclidean distance. The set of closest representatives of C_j with respect to C_i is defined as follows:

$$CR_j^i = \{(v_{ik}, v_{jl}) \mid v_{jl} = \text{closest_rep}^j(v_{ik})\}$$

DEFINITION 2.2. *Respective Closest Representative points.* The set of respective representative points of the clusters C_i and C_j is defined as the set of mutual closest representatives of the clusters under concern, i.e. $R_{CR_{ij}} = \{(v_{ik}, v_{jl}) \mid v_{ik} = \text{closest_rep}^i(v_{jl}) \text{ and } v_{jl} = \text{closest_rep}^j(v_{ik})\}$

In other words, the $R_{CR_{ij}}$ set is defined as the intersection of the closest representative of C_i with respect to C_j and the closest representative of C_j with respect to C_i , i.e. $R_{CR_{ij}} = CR_i^j \cap CR_j^i$.

3.2 Clusters' Separation in terms of density

The inter-cluster density evaluates the average density in the area between clusters. Here, the term “area between clusters” implies the area between the respective closest representatives of the clusters. Considering that representative points efficiently capture the shape and extent of the clusters, the density in the area between closest points of clusters is an indication of how close are the clusters.

DEFINITION 3. *Density between clusters* – Let $\text{clos_rep}_{ij}^p = (v_{ik}, v_{jl})$ be the p th pair of respective closest representative points of clusters C_i and C_j , i.e. $\text{clos_rep}_{ij}^p \in R_{CR_{ij}}$, and u_{ij}^p the middle point of the line segment defined by the p th pair clos_rep_{ij}^p (see Figure 3). The density between the clusters C_i and C_j is defined as follows:

$$\text{Dens}(C_i, C_j) = \frac{1}{|R_{CR_{ij}}|} \sum_{p=1}^{|R_{CR_{ij}}|} \left(\frac{d(\text{clos_rep}_{ij}^p)}{2 \cdot \text{stdev}} \cdot \text{density}(u_{ij}^p) \right) \quad \text{Eq. 1}$$

where $d(\text{clos_rep}_{ij}^p)$ is the Euclidean distance between the pair of points defined by $\text{clos_rep}_{ij}^p \in R_{CR_{ij}}$, and $|R_{CR_{ij}}|$ presents the cardinality of the set $R_{CR_{ij}}$. The term stdev is the average standard deviation of the considered clusters and is given by the equation:

$$\text{stdev} = \frac{1}{c} \sum_{i=1}^c \|\text{stdev}_i\| \quad \text{Eq. 2}$$

where c is the number of clusters and $\text{stdev}_i = (\text{stdev}_i^1, \dots, \text{stdev}_i^d)$ (d denotes the dimension of the data under concern) measures the deviation of the points belonging to the cluster C_i from the cluster center, while the term $\|\text{stdev}_i\|$ is defined as: $\|\text{stdev}_i\| = (\text{stdev}_i^T \text{stdev}_i)^{1/2}$.

The term $\text{density}(u_{ij}^p)$ is defined in Eq. 3:

$$\text{density}(u_{ij}^p) = \frac{\sum_{l=1}^{n_i+n_j} f(x_l, u_{ij}^p)}{n_i + n_j} \quad \text{Eq. 3}$$

where x_l corresponds to the data points of the clusters under concern (i.e. $x_l \in C_i \cup C_j$), while n_i and n_j are the number of points that belong to clusters C_i and C_j respectively. It represents the percentage of points in C_i and C_j that belong to the neighbourhood of u_{ij}^p . To define the neighbourhood of a data point, the scattering of data points on each dimension is considered to be an important factor. For example, if the scattering of data is large and the neighbourhood of points is small then there might be no points included in the neighbourhood of any of the data points. On the other hand, if the scattering is small and the neighbourhood is large, then the entire data set might be in the neighbourhood of all the data points. Selecting different neighbourhoods for different data sets can reasonably solve this problem. The standard deviation can be used to approximately represent the scatter of data points. Thus the neighbourhood of a point can be considered as a function of standard deviation on each dimension. In the context of this paper, the neighbourhood of a data point, u_{ij}^p , is defined as the hyper-sphere centered at u_{ij}^p (see Figure 3) with radius the average standard deviation of the considered clusters, $stdev$. Then the function $f(x, u_{ij})$ is defined as:

$$f(x, u_{ij}) = \begin{cases} 1, & \text{if } d(x, u_{ij}) < stdev \text{ and } x \neq u_{ij} \\ 0, & \text{otherwise} \end{cases} \quad \text{Eq. 4}$$

where $stdev$ is the standard deviation of clusters under concern as defined in Eq. 2.

A point belongs to the neighbourhood of u_{ij}^p if its distance from u_{ij}^p is smaller than the average standard deviation of the clusters (i.e. $d(x, u_{ij}^p) < stdev$). On the other hand, the actual area between clusters whose density we are interested in estimating is defined to be the area between the respective closest representative points (see Figure 3) and its size is defined to be $d(\text{clos_rep}_{ij}^p)$. Since the term $\text{density}(u_{ij}^p)$ represents the density in the area whose size is defined by the standard deviation of the considered clusters (i.e. the hyper-sphere with diameter $2 \cdot stdev$), the actual density between the clusters will correspond to the percentage $\frac{d(\text{clos_rep}_{ij}^p)}{2 \cdot stdev}$ of points belonging to the neighbourhood of u_{ij}^p (i.e. $\text{density}(u_{ij}^p)$). The above justifies the definition of density between the p th pair of the respective representatives of clusters C_i and C_j as $\frac{d(\text{clos_rep}_{ij}^p)}{2 \cdot stdev} \cdot \text{density}(u_{ij}^p)$. \square

DEFINITION 4. *Inter-cluster Density* - Let $\mathbf{C} = \{C_i \mid i=1, \dots, c\}$ be a partitioning of a data set into c clusters, $c > 1$. The Inter-cluster density measures for each cluster $C_i \in \mathbf{C}$, the maximum density between

C_i and the other clusters in \mathbf{C} . More specifically, it is defined by Eq. 5:

$$\text{Inter_dens}(\mathbf{C}) = \frac{1}{c} \cdot \sum_{i=1}^c \max_{\substack{j=1, \dots, c \\ j \neq i}} \{\text{Dens}(C_i, C_j)\} \quad \text{Eq. 5}$$

where $c > 1$, $c \neq n$. \square

DEFINITION 5. *Clusters' separation (Sep).* It evaluates the separation of clusters taking into account the Inter-cluster density with respect to the distance between clusters. A good partitioning is characterized by long distances between clusters' representatives and low density between them (i.e. well-separated clusters). Then, the clusters' separation is defined by the equation (Eq. 6):

$$\text{Sep}(\mathbf{C}) = \frac{\frac{1}{c} \cdot \sum_{\substack{i=1 \\ i \neq j}}^c \min_{\substack{j=1, \dots, c \\ i \neq j}} \{\text{Dist}(C_i, C_j)\}}{1 + \text{Inter_dens}(c)}, c > 1, c \neq n \quad \text{Eq. 6}$$

where $\text{Dist}(C_i, C_j) = \frac{1}{|\text{RCR}_{ij}|} \sum_{p=1}^{|\text{RCR}_{ij}|} d(\text{clos_rep}_{ij}^p)$ and $|\text{RCR}_{ij}|$ is the cardinality of the set RCR_{ij} as defined earlier.

Note: According to the definitions above, *Inter_dens* assesses the maximum number of points distributed in the area between the considered clusters. This is an indication of how close are the clusters. In general terms, it is expected the maximum density to be detected between a cluster and its closest one, since the closest are the clusters the more probable is to find area of high density between them. Also the area between clusters is measured in terms of the distance between their respective closest representatives. Then, *Sep(C)* can be perceived to measure the respective number of data points per unit of space between the closest clusters, i.e. the relative density between clusters.

3.3 Clusters' compactness in terms of density

Cluster's compactness is a measure of cluster's inherent quality, which increases when the clusters are characterized by high internal density. In the context of this paper we exploit the concept of multiple representative points for a cluster (as defined earlier). The average internal density of clusters is defined as the percentage of the cluster points that belong to the neighbourhood of representative points within the considered clusters. Therefore, the density within clusters is reaching higher values as the partitioning improves (i.e. approximates actual partitioning).

We previously introduced the concept of multiple representative points. They are initially generated by selecting well-scattered points in the cluster that represent well the geometric features of the cluster.

We exploit these points to assess the separation of the clusters as it is described above. Besides cluster's separation, we also take into account cluster's compactness and cohesion. This implies that clusters should not only be well separated but also dense. Without loss of generality the center of a cluster (the mean of the data points in the cluster) can be perceived as a good approximation of the cluster space core³. Then we gradually shift the representative points towards the clusters' center in order to take instances of the initial representative points at different places of cluster space. Measuring the density in the neighbourhood of these representatives, the density distribution within a cluster can be estimated.

Let \underline{v}_{ij} , further called *shrunk representative*, correspond to the j th representative point of the cluster C_i , v_{ij} , shrunk (shifted) towards the center of the cluster by a shrinking factor $s \in [0, 1]$ (see Figure 3). Thus the k th dimension of \underline{v}_{ij} can be defined as $\underline{v}_{ij}^k = v_{ij}^k + s \cdot (C_i^k \text{.center} - v_{ij}^k)$, where $C_i \text{.center}$ is the center of cluster C_i . The shrinking factor, s , is user-defined to control the compactness of clusters in the validity checking process according to the application needs. A high value of s shrinks the representatives closer to the cluster center and thus it favors more compact clusters. On the other hand, a small value of s shrinks more slowly the representatives and the validity checking process favors elongated clusters.

To eliminate the influence of s to the cluster validity results, the density within clusters is estimated for different values of s . More specifically, the value of s is increasing so that the representative points are gradually shrunk and the respective values of density are calculated at these shrunk points. The average value of a cluster's intra-cluster density, as calculated for the different values of s , is considered to be the density within the cluster under concern. It is evident that we are able to get a better view of density distribution within the considered cluster, calculating the density at different areas of the cluster.

DEFINITION 6. *Relative intra-cluster density* measures the relative density within clusters with respect to (wrt.) a shrinking factor s . This implies the number of points that belong to the neighbourhood of the

³ Here the "cluster core" is used to denote the most central point of cluster space. It is not necessarily a point within the cluster but it is perceived as a point around which the data points belonging to the cluster are distributed.

representative points of the defined clusters shrunk by s , let \underline{v}_{ij} , (i.e. points belong to the hyper-sphere centered at \underline{v}_{ij} with $stdev$ radius). Then the relative *Intra-cluster density* with respect to the factor s is defined as follows:

$$\text{Intra_dens}(\mathbf{C}, s) = \frac{\text{Dens_cl}(\mathbf{C}, s)}{c \cdot stdev}, \quad c > 1, \quad \text{where } \text{Dens_cl}(\mathbf{C}, s) = \frac{1}{r} \sum_{i=1}^c \sum_{j=1}^r \text{density}(\underline{v}_{ij}) \quad \text{Eq. 7}$$

The term $\text{density}(\underline{v}_{ij})$ is defined in Eq. 8:

$$\text{density}(\underline{v}_{ij}) = \sum_{l=1}^{n_i} f(x_l, \underline{v}_{ij}) / n_i, \quad \text{Eq. 8}$$

where n_i is the number of the points, x_l , that belong to the cluster C_i , i.e. $x_l \in C_i \subseteq S$. It represents the proportion of points in cluster C_i that belong to the neighbourhood of a representative $\underline{v}_{ij} \forall j$ (i.e. the representatives of C_i shrunk by a factor s). The neighbourhood of a data point, \underline{v}_{ij} , is defined to be a hyper-sphere centered at \underline{v}_{ij} with radius the average standard deviation of the considered clusters, $stdev$. The function $f(x, \underline{v}_{ij})$ in Eq. 8 is defined as in Eq. 4. \square

DEFINITION 7. The *compactness* of a clustering scheme \mathbf{C} in terms of density is defined by the equation:

$$\text{Compactness}(\mathbf{C}) = \sum_s \text{Intra_dens}(\mathbf{C}, s) / n_s \quad \text{Eq. 9}$$

where n_s denotes the number of different values considered for the factor, s , based on which the density at different areas within clusters is calculated. Usually, we consider that the values of the shrinking factor, s , is gradually increasing in $[0.1, 0.8]$ (the cases that $s = 0$, and $s \geq 0.9$ refer to the trivial case that the representative points correspond to the boundaries and center of cluster respectively).

Then considering that the representative points are shrunk by a factor $0.1 \leq s \leq 0.8$, and $s_i = s_{i-1} + 0.1$, we get from Eq. 9:

$$\text{Compactness}(\mathbf{C}) = \sum_{s \in [0.1, 0.8]} \text{Intra_dens}(\mathbf{C}, s) / 8 \quad \text{Eq. 9a}$$

On other words, in the context of this paper the term $\text{Compactness}(\mathbf{C})$ corresponds to the average density within a set of clusters, \mathbf{C} , defined for a data set.

3.4 Assessing the quality of a clustering scheme

In the previous sections (Section 3.2 and Section 3.3) we introduce some measures based on which the *compactness* and *separation* of clusters are evaluated. However, none of these measures could lead to a

reliable evaluation of clusters' validity if they are taken into account separately. Thus the requirement for a global measure that assesses the quality of a clustering scheme in terms of its validity arises.

The above motivate us to proceed with the introduction of the terms i) *Clusters' cohesion* and ii) *Separation wrt. Compactness*. They rely on requirements of "good clustering" and aim at giving a global assessment of clusters' quality.

3.4.1 Clusters' Cohesion

Besides the compactness of clusters, another requirement of clusters' quality is that the changes of density distribution within clusters to be significantly small. This implies that not only the average density within clusters (as measured by *Compactness*) has to be high but also the density changes as we move within the clusters have to be small. The above requirements are strictly related with the evaluation of clusters' cohesion, i.e. the density-connectivity of objects belonging in the same clusters.

DEFINITION 8. *Intra-density changes*, measures the changes of density within clusters. It is given by the equation:

$$\text{Intra_change}(\mathbf{C}) = \frac{\sum_{i=1, \dots, n_s} |(\text{Intra_dens}(\mathbf{C}, s_i) - \text{Intra_dens}(\mathbf{C}, s_{i-1}))|}{(n_s - 1)} \quad \text{Eq. 10}$$

where n_s is the number of different values that the factor s takes. Significant changes to the intra-cluster density indicate that there are areas of high density that followed by areas of low density and vice versa. □

DEFINITION 9. *Cohesion* measures the density within clusters with respect to the density changes observed within them. It is defined as follows:

$$\text{Cohesion}(\mathbf{C}) = \frac{\text{Compactness}(\mathbf{C})}{1 + \text{Intra_change}(\mathbf{C})} \quad \text{Eq. 11}$$

3.4.2 Separation wrt Compactness

The optimal partitioning (see Introduction for a definition of the term) requires maximal compactness (i.e. intra-cluster density) in such a way that the clusters are well separated and vice versa. This implies that *compactness and separation* are closely related measures of clusters' quality. Furthermore there are cases in which the clusters' separation tends to be meaningless with regard to the clusters' quality, if it is considered independently of clusters' compactness. For instance, assume the data set depicted in Figure 4 and its partitionings presented in Figure 4a and Figure 4d. Though both of these partitionings contain well-separated clusters, the partitioning of Figure 4a contains more compact clusters than the

one of Figure 4d and therefore it is selected as the optimal partitioning. Then, it is evident that we need to evaluate a measure, which evaluates the separation of clusters in conjunction with their compactness.

DEFINITION 10. *SC (Separation wrt Compactness)* evaluates the clusters' separation with respect to their compactness:

$$SC(\mathbf{C}) = \text{Sep}(\mathbf{C}) \cdot \text{Compactness}(\mathbf{C}) \quad \text{Eq. 12}$$

In other words, considering a data set and its clustering scheme \mathbf{C} , SC is defined as the product of the density between clusters ($\text{Sep}(\mathbf{C})$) and the density within the clusters defined in \mathbf{C} ($\text{Compactness}(\mathbf{C})$).

3.4.3 *CDbw* definition

A reliable cluster validity index has to correspond to all the requirements of “good” clustering. This implies that it has to evaluate the cohesion of clusters as well as the separation of clusters in conjunction with their compactness. These requirements motivate the definition of the *validity index CDbw*. It is based on the terms defined in the equations Eq. 11 and Eq. 12 and is given by the following equation:

$$\text{CDbw}(\mathbf{C}) = \text{Cohesion}(\mathbf{C}) \cdot SC(\mathbf{C}), c > 1 \quad \text{Eq. 13}$$

The above definitions refer to the case that a data set possesses clustering tendency, i.e. the data vectors can be grouped into at least two clusters. The validity index is not defined for $c = 1$. A detailed discussion on *CDbw* and its properties is presented in Section 4.

3.5 Further discussion on *CDbw* definition

The definition of *CDbw* (Eq. 13) indicates that all the criteria of “good” clustering (i.e. cohesion of clusters, compactness and separation) are taken into account, enabling reliable evaluation of clustering results. A clustering scheme with compact and well-separated clusters with few variation of the density distribution within clusters results in high values for both *CDbw* terms (i.e. Cohesion and Separation wrt. Compactness). Therefore it converges to a maximum value when the optimal partitioning is achieved (a proof is provided in the next section). Moreover, *CDbw* exhibits no monotonous trends with respect to the number of clusters and thus in the plot of *CDbw* versus the number of clusters we seek the maximum value of *CDbw*. The absence of a clear local maximum in the plot is an indication that the data set under consideration possesses no clustering structure.

In the trivial case that each point is considered as a separate cluster, i.e. $c = n$, the standard deviation of the clusters is 0. Then Eq. 1 and Eq. 7 cannot be defined when $c = n$. However, this is not a serious problem. In real-world cases, if the data can be organized into compact and well-separated clusters (i.e. the data set possesses a clustering structure), its optimal partitioning will correspond to a set of clusters whose number ranges between 2 and $n-1$.

Nevertheless, considering the semantics of the terms *Intra_dens* and *Inter_dens*, which in the trivial case ($c=n$), cannot be defined based on Eq. 1 and Eq. 7, we proceed with the following statement:

In the trivial case that each point is a separate cluster, i.e. $c = n$, the standard deviation of clusters is 0. Then:

- According to Eq. 3 the term $\text{density}(u_{ij}^p)$ is zero for any pair of the defined clusters. This implies that the density between clusters is also zero, i.e.

$$\text{Dens}(C_i, C_j) = 0 \quad \forall i, j \in [1, n] \Rightarrow \text{Inter_dens}(\mathbf{C}) = 0, \text{ where } \mathbf{C} = \{C_i \mid i=1, \dots, n\}$$

- The intra-cluster density measures the average density in the neighbourhood of the clusters' shrunken representatives. In case that $c = n$, there is only one point that belong to a cluster which is also considered as representative. According to Eq. 4: $\forall x, x \neq \underline{v}_{ij}$ and $d(x, \underline{v}_{ij}) = \text{stdev}_i = 0 \Rightarrow f(x, \underline{v}_{ij}) = 0$. Therefore the density within clusters is :

$$\text{Dens_cl}(\mathbf{C}, s) = \text{Dens_cl}(\mathbf{C})/n = 0, \forall s \text{ and } \mathbf{C} = \{C_i \mid i=1, \dots, n\}$$

Then, without loss of generality, we claim that $\text{Intra_dens}(\mathbf{C}, s) = 0 \quad \forall s$, when $c=n$. As a consequence, we get from Eq. 9 that $\text{Compactness}(\mathbf{C}) = 0$, when $c = n$.

Hence, based on Eq. 13, $\text{CDBw}(\mathbf{C}) = 0$ when $c = n$.

In the sequel we proceed with a theoretical study on *CDBw* which justifies its ability to select the optimal partitioning of a data set S given that S possesses clustering structure. A study of the *CDBw* mathematical properties is presented in Appendix I. It proves that it is bounded and presents no monotonous trends with respect to the number of clusters.

4. DISCUSSION ON CDBW PROPERTIES

In this section we study the effectiveness of *CDBw* in selecting the optimal partitioning (as defined in the Introduction) among those defined by different clustering algorithms' sessions. We prove that *CDBw* converges to a maximum value when the optimal partitioning of a data set is considered.

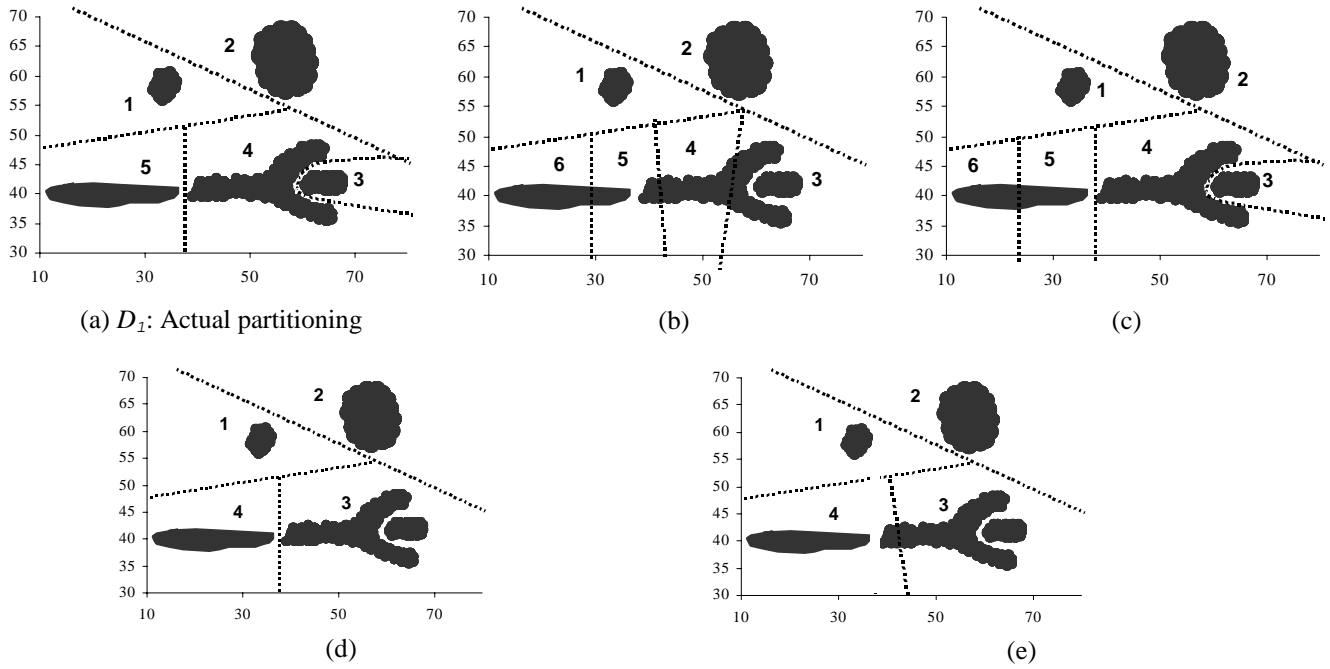


Figure 4. A data set S partitioned into different number of clusters.

Assuming a data set S , the behavior of the validity index is studied in the following cases:

- i) the clustering schemes, D_i , defined for S consist of different number of clusters.
- ii) different clustering schemes defined for S all of them with the same number of clusters.

In the sequel we discuss in further detail the above cases of clustering results. We assume that the considered data sets presents clustering tendency (i.e. their data can be organized into clusters), and that there is no overlap among clusters.

Case 1: Assume a data set S and a set of different partitionings D_i defined for S each of which corresponds to a different number of clusters. The value of $CDbw$ is maximized when the optimal partitioning is found.

Let n be the correct number of clusters of the data set S corresponding to the partitioning D_1 (actual partitioning of S): $D_1(n, S) = \{D_{1.c_i}\}, i=1, \dots, n$ and

let m be the number of clusters of another partitioning D_r of S : $D_r(m, S) = \{D_{r.c_j}\}, j = 1, \dots, m$.

Let $CDbw_{D_1}$ and $CDbw_{D_r}$ be the respective values of the cluster validity index for the clustering schemes. Then, we consider the following sub-cases:

- i) Assume the number of clusters in D_r to be more than the real clusters (i.e. $m > n$), and parts of the real clusters (corresponding to D_1) grouped into clusters of D_r (e.g. cluster 5 in Figure 4b).

Let $fC_{D_1} = \{f_{C_{D_1p}} \mid p=1, \dots, n, f_{C_{D_1p}} \subseteq D_{1.c_i}, i=1, \dots, n\}$ be a set of fractions of clusters in D_1 .

Similarly, we define $fc_{D_r} = \{fc_{D_{rk}} \mid k = 1, \dots, nf_r, fc_{D_{rk}} \subseteq D_r.c_j, j=1, \dots, m\}$. Then:

a. There is at least one cluster of D_r that is formed by a union of some of D_I 's cluster fractions (e.g. cluster 3 in Figure 4b consists of the cluster 3 and a fraction of the cluster 4 in the partitioning of Figure 4a). This can be formalized as follows:

$\exists D_r.c_i: D_r.c_i = \cup fc_{D_{1p}},$ where $p \in \{1, \dots, nf_1\}$, nf_1 is the number of considered fractions of clusters in D_I ,

b. There is at least one cluster of D_I that is formed by a union of some of D_r 's cluster fractions (e.g. cluster 4 in D_I is formed by the union of cluster fractions of the clusters 3, 4 and 5 in Figure 4b), i.e. $\exists D_1.c_i: D_1.c_i = \cup fc_{D_{rk}},$ where $k \in \{1, \dots, nf_r\}$, where nf_r is the number of considered fractions of clusters in D_r ,

In this case, some of the clusters in D_r include regions of low density while significant changes to the density distribution are observed within the clusters of D_r (for instance cluster 3 in Figure 4b). Thus, the value of the first term of $CDbw$ (Cohesion of D_r) is smaller than the cohesion of D_I (i.e. $Cohesion_{D_r} < Cohesion_{D_I}$). On the other hand, the second term (Separation wrt. Compactness of clusters) is also decreasing compared to the corresponding term in D_I (i.e. $SC_{D_r} < SC_{D_I}$). This is because some of the real clusters are split in case of D_r and therefore there are areas between clusters that are of high density (e.g. clusters 3 and 4 in Figure 4b). Then, since both $CDbw_{D_r}$ terms decrease in comparison to $CDbw_{D_I}$ we conclude that $CDbw_{D_I} > CDbw_{D_r}$.

ii) Let D_r be a partitioning where more clusters than in D_I are formed (i.e. $m > n$). We assume that at least one of the D_I clusters is split to more than one partitions each of which form a separate cluster in D_r (e.g. cluster 5 in D_I is split into two clusters (i.e. cluster 5 and cluster 6) in case of the partitioning presented in Figure 4c) while none of D_I clusters are subsets of D_r clusters (as Figure 4c depicts). That is, $\exists D_1.c_i: D_1.c_i = \cup c_{D_{rj}}, j \in \{1, \dots, m\}$, and $\forall D_r.c_j: D_r.c_j \neq \cup D_1.c_i, i \in \{1, \dots, n, n = \text{number of clusters in } D_I\}$.

In this case, the clusters of both D_I and D_r present no significant changes to their density distribution. Thus the cohesion of clusters in D_r , $Cohesion_{D_r}$, is slightly smaller or vaguely the same as compared to the cohesion of D_I , $Cohesion_{D_I}$. As a consequence, we get $Cohesion_{D_r} \approx Cohesion_{D_I}$. On the other hand, the term of $CDbw$, Separation wrt. Compactness, is decreasing as some of the clusters in D_I (corresponding to the real clusters) are split in case of D_r . Therefore

there are areas between clusters that are of high density (for instance cluster 3 in Figure 4c) and then $SC_{D_r} \ll SC_{D_1}$. Based on the above discussion and taking into account that the decrease of SC (separation wrt. compactness) is significantly higher than the decrease of Cohesion we conclude that $CDbw_{D_1} > CDbw_{D_r}$.

iii) Let D_r be a partitioning with less clusters than in D_1 ($m < n$) and two or more of D_1 clusters are grouped in a cluster of D_r (as Figure 4d depicts). Then, $\exists D_{r,c_j}: D_{r,c_j} = \cup D_{1,c_i}$, where $i \in \{1, \dots, n\}$. In this case, the value of the first term of $CDbw_{D_r}$, $Cohesion_{D_r}$, significantly decreases as compared to $Cohesion_{D_1}$ (i.e. the value of the clusters' cohesion in D_1) since we can observe significant changes to the density distribution within the clusters of D_r . As a consequence, $Cohesion_{D_r} \ll Cohesion_{D_1}$. On the other hand, the separation wrt. the compactness of the clusters in D_r is slightly decreasing or remains vaguely the same as compared to this of D_1 . This is because similarly to D_1 clusters (Figure 5a), there are no dense areas between the D_r clusters. However, the clusters of D_1 are more compact than these of D_r . Then, we claim that $SC_{D_1} \approx SC_{D_r}$. Based on the above discussion and considering the decrease in D_r 's cluster cohesion and compactness in conjunction with the fact that there are no significant changes to the separation of clusters, we claim that $CDbw_{D_1} > CDbw_{D_r}$.

iv) Assume the number of clusters in D_r to be less than the real clusters (i.e. $m < n$), and parts of the real clusters (corresponding to D_1) are grouped into clusters of D_r (e.g. Figure 4e). This is almost similar to the sub-case (i). More specifically, let fc_{D_1} and fc_{D_r} be the set of cluster fractions in D_1 and D_r respectively (as defined above in the sub-case (i)). Then:

a. There is at least one cluster of D_r that is formed by a union of D_1 's cluster fractions (e.g. cluster 4 in Figure 4e consists of cluster 5 and a fraction of cluster 4 in D_1). This can be formalized as follows:

$\exists D_{r,c_i}: D_{r,c_i} = \cup fc_{D_1 p}$, where $p \in \{1, \dots, nf_1\}$, nf_1 is the number of considered fractions of clusters in D_1 , and

b. There is at least one cluster of D_1 that is formed by a union of D_r 's cluster fractions (e.g. cluster 4 in D_1 is formed by the union of fractions of the clusters 3 and 4 in the partitioning depicted in Figure 4e), i.e. $\exists D_{1,c_i}: D_{1,c_i} = \cup fc_{D_r k}$, where $k \in \{1, \dots, nf_r\}$, and nf_r is the number of considered fractions of clusters in D_r .

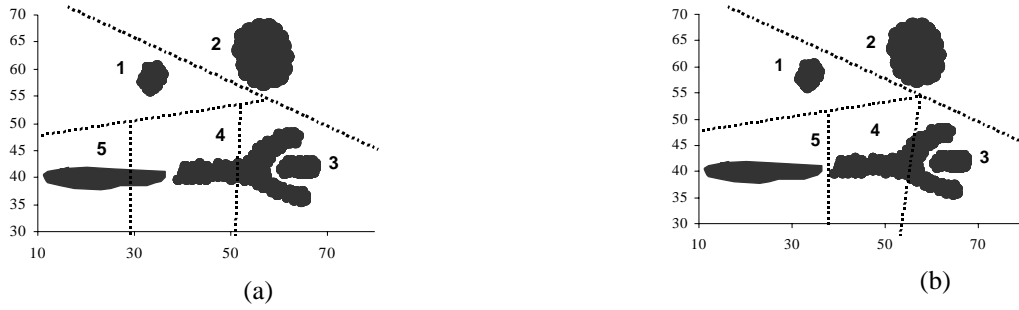


Figure 5. A data set S partitioned wrongly in five clusters

In this case, some of the clusters in D_r include regions of low density while significant changes to the density distribution are observed within the clusters of D_r (for instance cluster 3 in Figure 4e). As a consequence the value of D_r 's cluster cohesion is smaller than the cohesion of D_I (i.e. $\text{Cohesion}_{D_r} < \text{Cohesion}_{D_I}$). Moreover, some of the real clusters are split in D_r and therefore there are areas between clusters that are of high density (e.g. clusters 3 and 4 in Figure 4e). Therefore, we claim that the term of separation wrt. compactness of clusters in D_r is decreasing compared to the corresponding term in D_I (i.e. $\text{SC}_{D_r} < \text{SC}_{D_I}$). Then, since both CDBw_{D_r} terms decrease in comparison to CDBw_{D_I} we get that $\text{CDBw}_{D_I} > \text{CDBw}_{D_r}$.

These are all the sub-cases of partitionings with different number of clusters that can be defined for a given data set. Based on the above discussion, we conclude that in each cases, CDBw selects D_I (i.e. actual partitioning) as the optimal partitioning.

Case 2: Consider a data set S and the different partitionings D_i defined for S . Assuming that each D_i consists of a number of clusters equal to the correct one, the value CDBw is maximized when the optimal partitions are found for the correct number of clusters.

Let D_r be a partitioning with the same number of clusters as the correct one (i.e. the number of clusters presented in the actual partitioning), D_I (i.e. $m = n$, see Figure 4a). Furthermore, we assume that one or more of the real clusters (as defined earlier) corresponding to D_I are split and their parts are grouped into different clusters in D_r (as in Figure 5a,b). This implies that assuming $\text{fc}_{D_I} = \{\text{fc}_{D_I p} \mid p \in \{1, \dots, \text{nf}_1\}\}$, $\text{fc}_{D_I p} \subseteq D_{I.c_i}$, $i=1, \dots, n$ to be a set of cluster fractions in D_I , there is at least one cluster in D_r that consists of a subset of D_I 's cluster fractions, i.e. $\exists D_{r.c_j}: D_{r.c_j} = \cup_{p \in \{1, \dots, \text{nf}_1\}} \text{fc}_{D_I p}$. In this case, the D_r clusters contain areas of low-density, thus significant density variations within clusters are observed (e.g. cluster 3 and cluster 4 in Figure 5a or cluster 3 in Figure 5b). As a consequence, in case of D_r the cohesion of clusters, Cohesion_{D_r} , decreases as compared to the respective term of D_I , i.e.

$Cohesion_{D_r} < Cohesion_{D_1}$. On the other hand, some of the clusters in D_r are split and therefore there are areas between clusters that are of high density (for instance areas between clusters 3 and 4, and clusters 4 and 5 in Figure 5a or between clusters 3 and 4 in Figure 5b). Therefore, the separation wrt, the compactness of clusters in D_r is less than in D_1 , i.e. $SC_{D_r} < SC_{D_1}$. Then it is evident that $CDbw_{D_1} > CDbw_{D_r}$.

Conclusion: To summarize, CDbw achieves to select in each case the optimal partitioning of a data set among those defined by clustering algorithms applied to a data set. In other words, if a data set, S, possesses a clustering structure and an algorithm achieves to find its actual partitioning, C_{opt} , then the value of CDbw corresponding to C_{opt} will be the maximum among the respective values for other partitionings of S.

5. TIME COMPLEXITY

The complexity of the cluster validity index $CDbw$, is based on the complexity of the terms *Cohesion* and *Separation wrt. Compactness* as defined in the equations Eq. 11 and Eq. 12 respectively. Let d be the number of attributes (data set dimension); c be the number of clusters; n be the number of the data points; r be the number of a cluster's representatives. Then the complexity of selecting the closest representative points of c clusters is $O(dc^2r^2)$. Based on their definitions, the computational complexity of SC depends on the complexity of clusters' compactness (*Compactness*) and separation (*Sep*) that is $O(ncrd)$ and $O(ndc^2)$ respectively. Then the complexity of SC is $O(ndr^2c^2)$. Furthermore, based on Eq. 11, the computational complexity of clusters' cohesion (*Cohesion*) is $O(ncrd)$. Then, we conclude that $CDbw$ complexity is $O(ndr^2c^2)$. Usually, $c, d, r \ll n$, therefore the complexity of the validity index for a specific clustering scheme is $O(n)$. The complexity of the whole cluster validity procedure to find the optimal partitioning of a data set, S, among a set of k different partitionings defined by a clustering algorithm will be $O(kn)$. In the context of this paper, the clustering process is not considered as part of the cluster validity process. Given that k is significantly smaller than n (number of data points), the complexity of the clustering validity process will be $O(n)$.

To quantify the above we experimented with different data sets and measured the time complexity of the proposed cluster validity approach with respect to the size of the data set and the number of clusters. The data sets used for the experiments are synthetic generating according to the normal distribution. Figure 6a shows that our method scales linearly as the size of the data set increases from 10^3 to 10^6 points. Moreover, the running time of the cluster validity approach for data sets with 8 to 40

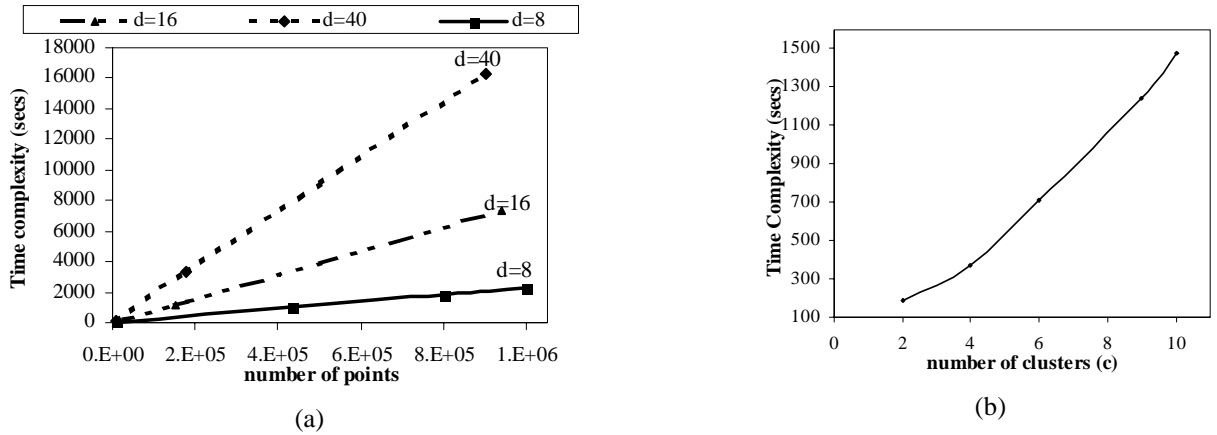


Figure 6: Execution time in seconds as function of (a) the number of points (b) the number of clusters

dimensions data is measured. As Figure 6a depicts the $CDbw$ execution time increases linearly to the number of points independently of the data dimensionality. On the other hand, Figure 6b plots the running time of the cluster validity approach versus the number of clusters. We used a 6-dimensional data set with 147.200 points and considering its clustering schemes with 2 to 10 clusters we compute the respective values of $CDbw$. Figure 6b shows, as it is expected based on theoretical study of the time complexity, that the cluster validity approach scales nearly quadratic as the number of clusters increases but since c is usually a small integer, it creates no problem.

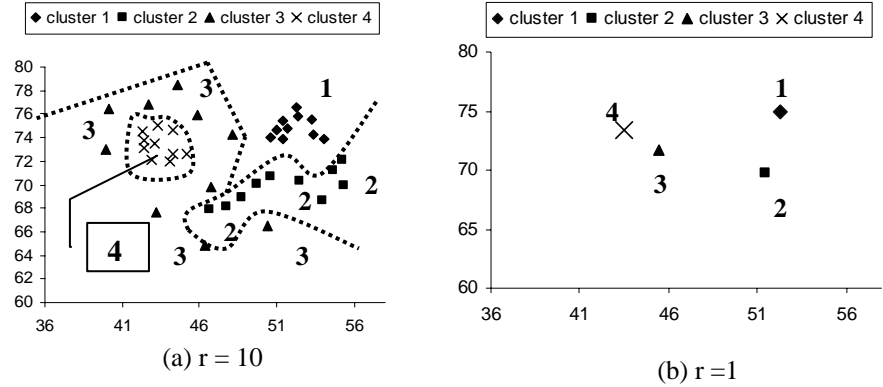
6. EXPERIMENTAL EVALUATION

Based on the definition of the validity index $CDbw$ described in Section 3 we implemented a cluster validity checking system in JAVA. The data sets are stored in Microsoft Access 2000 relational tables while the connection to the database is implemented using the JDBC API [34].

We carried out experiments with representative clustering algorithms of different categories, i.e. partitional, hierarchical and density-based ones using real-world and synthetic data sets. In all cases our approach performs favourably selecting the optimal partitioning for a data set among those defined by a clustering algorithm under different assumptions. Additionally, experiments using results of different clustering algorithms were carried out. This study showed that $CDbw$ can be used to evaluate comparatively results of different clustering algorithms and therefore selects the algorithm that discovers the optimal partitioning for a data set. The experimental section concludes with a comparison of $CDbw$ to some of the most important validity indices proposed in the literature.

Before we proceed with the discussion on the results of our approach experimental evaluation, we briefly present the results of a study we carried out to examine the influence of the number of

representatives to the validity results. Based on this we also select the number of representatives that we use for the experiments discussed below .



Tuning the number of representatives r . We considered data sets containing clusters of different shapes and varied the number of representatives, r , in the range from 1 (equivalent to the case where only the cluster center is used for representation) to 30. We observed that *CDbw* did not always select the optimal partitioning when r took values less than 5.

Especially, in the case that $r = 1$, (i.e. cluster center is considered as the representative), *CDbw* failed to select the correct number of clusters for the data sets with arbitrarily shaped clusters. On the other hand, for values of r greater than 5, *CDbw* always found the optimal partitioning for the data set even in case of data sets containing arbitrarily shaped clusters. In order to illustrate the effectiveness of the representative points to capture the geometry of clusters, we plot in Figure 7 the representatives of clusters in DS4 (see Figure 14) for $r = 1$ and $r = 10$. As Figure 7b depicts, a single representative point cannot efficiently represent the shape of clusters in DS4, contrary to the case where more than one representatives ($r = 10$) are used (see Figure 7a). This also justifies the advantage of our cluster validity approach in comparison to traditional ones. Furthermore, we observed that there were no significant changes to the efficiency of *CDbw* when r took values greater than 10. In general, we conclude that a number of representatives around 10 ($r \geq 10$) achieves to capture to a satisfactory degree the geometry of clusters and then *CDbw* gives reliable results with regard to the selection of the optimal partitioning of a data set. For the experiments discussed below, ten representatives have been used to represent the clusters under evaluation.

Data sets. We experiment with different synthetic data set in order to evaluate the performance of the proposed cluster validity approach in various cases of the data sets' structure. We used a synthetic data generator driven by a set of parameters, such as data dimension, the number of clusters, the number of points per cluster, the clusters' geometry (i.e. cycle, rectangle). Moreover the specific characteristics of

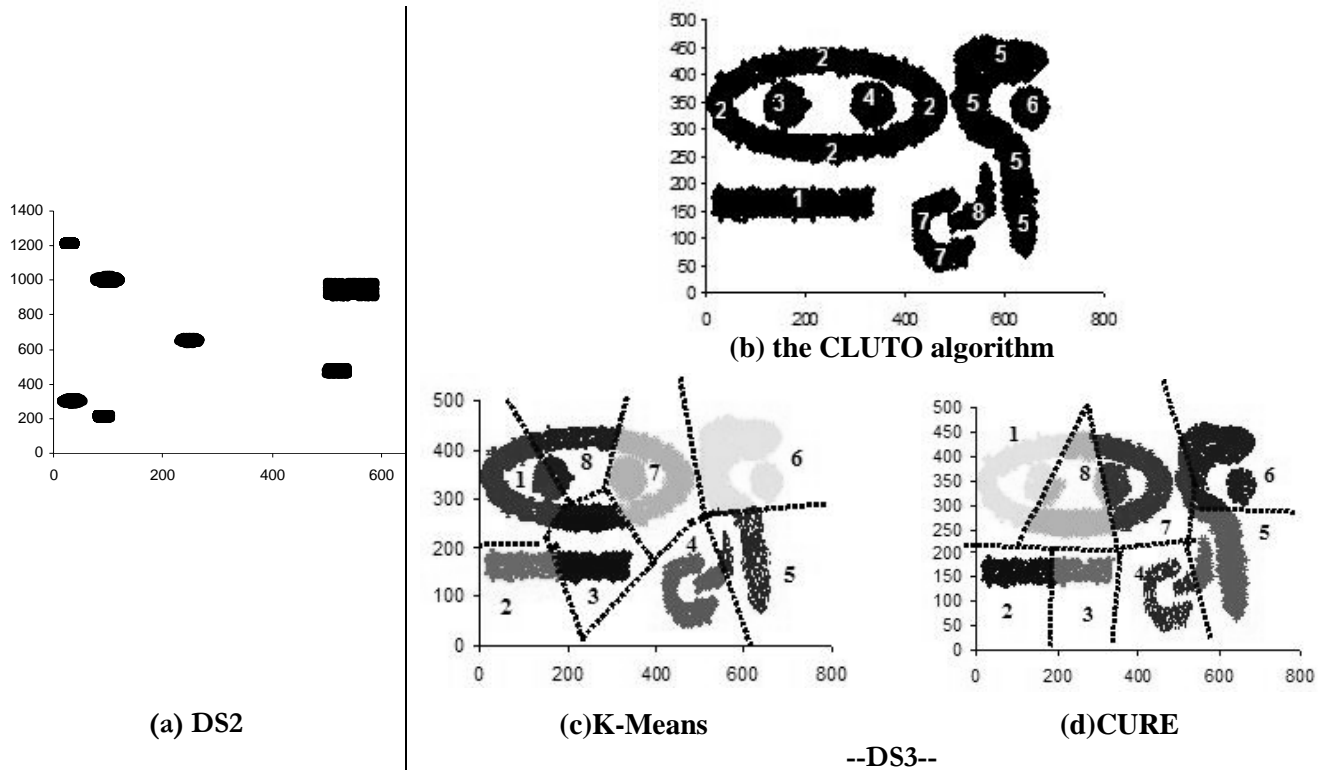
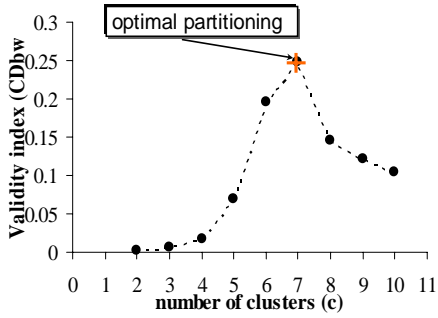


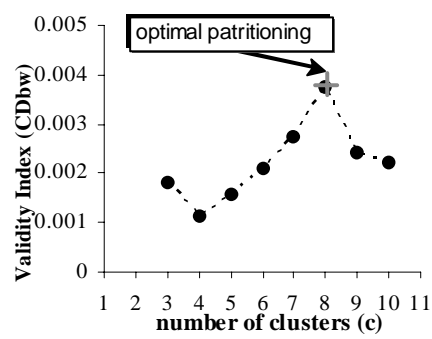
Figure 8: Synthetic data sets.: a)Visualization of DS2 and Partitioning of DS3 using b) CLUTO, c) K-Means, d) CURE

each cluster have to be determined, i.e. the radius and center for circular clusters, the coordinates of the lower vertex and the length of edge for rectangular clusters. Then each cluster is defined as a hyper-rectangle or a hyper-cycle and the points in the interior of the cluster are uniformly distributed. The clusters of the considered data sets can have non-spherical shapes and different average densities. Finally, we use two real-world data sets from the spatial and biomedical domain to evaluate the performance of our approach in real world applications.

Below we discuss the *CDbw* validation approach in cases of two-dimensional and multi-dimensional data sets as well as in case of non-spherical clusters (i.e. arbitrarily shaped clusters). We experiment with various data sets containing different numbers of clusters of various shapes. Due to space constraints we report only results for data sets containing less than 10 clusters. As regards the data dimensionality, it ranges between 2 and 120 dimensions. The rest of our results with data sets of higher dimensions and data sets containing a larger number of clusters are qualitatively similar to those presented herein, thus they are omitted for brevity. Also, for clarity purposes and to avoid cases that an algorithm fails to define clusters in a data set because of the nature of data (e.g. no clustering tendency, noise etc), we chose to describe experiments on data that possess a clear clustering structure. We have ignored the presence of noise in data so that the experimental study is independent of the efficiency of



(a)



(b)

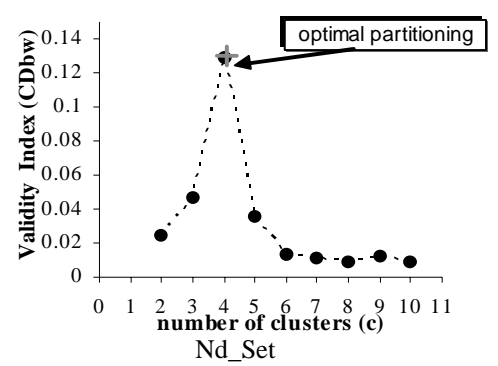


Figure 9: $CDBw$ as a function of number of clusters for a) DS2 (CURE), b) DS3 (CLUTO)

Figure 10. $CDBw$ vs the number of clusters for a 120-dimensional data set

the clustering algorithms to handle noise.

6.1 Selecting the optimal partitioning defined by a clustering algorithm

The goal of this experiment is to evaluate the effectiveness of $CDBw$ with regard to the selection of the optimal clustering scheme among those defined by a clustering algorithm. The initial set of experiments refers to a 2-dimensional data set, DS2, containing 7 clusters (see Figure 8a). We ran CURE [10] with $a=0.3$, r (number of representatives) = 10 and the number of clusters (denoted by c) ranging between 2 and 10. In the context of this paper different values of c correspond to different clustering schemes of a data set. For each of the partitionings obtained we computed the respective value of $CDBw$. As Figure 9a depicts the clustering scheme of seven clusters, which corresponds to the real clusters, is proposed as the optimal partitioning of DS2. This is also the number of partitions that corresponds to the real clusters of DS2 (see Figure 8a). Another 2-dimensional data set used in the context of this experimental study was DS3. It is a synthetic data set generated based on the data set used in [17] and it contains 8 clusters (see Figure 8b). We used a clustering algorithm provided by the CLUTO toolkit⁴ to discover the clustering scheme of DS3 with the number of clusters ranging in [3, 10]. For each of the partitionings obtained the $CDBw$ value is computed and the respective graph of the validity index values vs the number of clusters is created (see Figure 9b). Based on this graph we observe that $CDBw$ reaches its maximum at the partitioning of the eight clusters. Then it is proposed as the optimal partitioning of DS3. We note that the selected set of partitions also corresponds to the real clusters of DS3 (see Figure 8b).

⁴ The results of this experiment was obtained by running the 'vcluster' program with parameters $clmethod=graph$, $sim=dist$, $agglofrom=30$ and the number of clusters ranging between 3 and 10.

Multi-dimensional data sets. The validity of clustering results (i.e. that the set has been well partitioned) can be visually verified only in 2D or 3D cases. In higher dimensions it is difficult to verify the resulting clusters. The proposed validity index, $CDbw$, tackles this problem giving an indication of the optimal clustering scheme without visualization of the data set. We have experimented with various data sets but due to lack of space, herein, we select to present the behaviour of $CDbw$ when it is used to evaluate a set of different partitionings that have been defined for a data set with 120 attributes. This data set, further referred to as Nd_Set, contains four clusters and this is also verified by the $CDbw$ approach. We ran the CURE algorithm (with $a=0.3$, $r=10$) on the data repeatedly, with the number of clusters c in the range 2 to 10. For each value of c , we obtained a partitioning of the data and calculated the respective value of $CDbw$. The plot of $CDbw$ vs. the number of clusters (corresponding to the different partitionings of the data set) is depicted in Figure 10 and as we can observe $CDbw$ takes its maximum value when $c = 4$. Thus the partitioning of four clusters is proposed as the optimal partitioning of the data set under concern as defined by CURE.

Real-world data sets. One of the data sets we used in the current study contains parts of Greek roads network [28]. The roads are represented by their MBR approximations' vertices and the data set is further called R_D1. The goal of applying cluster analysis to this data set is to detect significant groups in the considered part of the Greek road map. It is evident that given the road map we are not able to know a priori which are the 'inherent' groups and thereby a visual-free validation approach is required. Then, different clustering schemes were discovered by CURE and they are evaluated based on the $CDbw$ approach. Figure 12 depicts the behaviour of $CDbw$ with respect to the number of clusters and we observe that $CDbw$ is maximized for four clusters. According to this procedure the partitioning of four is selected as the optimal partitioning of R_D1. To justify our claim we consider the visualization of the clustering scheme that is maximized for four clusters. According to this procedure the partitioning of four is selected as the optimal partitioning of R_D1. To justify our claim we consider the visualization of the clustering scheme that is selected as the optimal partitioning of R_D1. It is depicted in Figure 11. We observe that the selected clustering scheme of four clusters is a good approximation of groups into which the data of R_D1 can be organized.

The next data set we considered comes from the biomedical domain. The data are collected from Greek

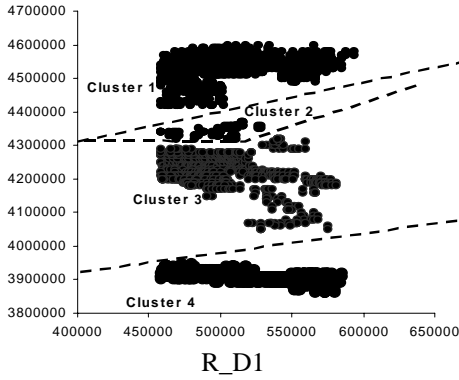


Figure 11. A Real Data Set representing a part of the Greek roads network

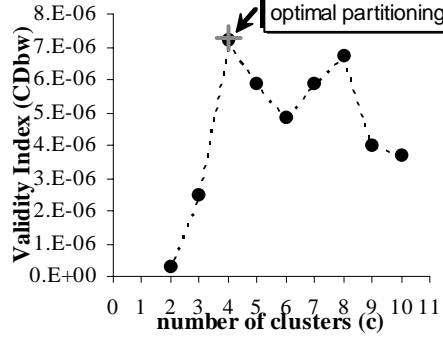


Figure 12. $CDBw$ as a function of number of clusters for real data representing a part of Greek road network.

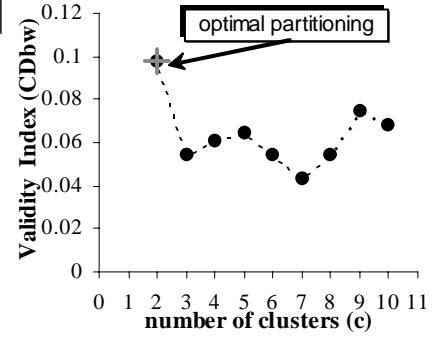


Figure 13. $CDBw$ vs the numbers of clusters for the OXA_VAN data set

hospitals based on the daily isolations of the Microbiology laboratory [23, 33]. They are measurements of the isolated organisms’ resistances to specific antibiotics. In the context of this experiment we focus on discovering the overall distribution of patterns in the set of *Sau* organism resistances to two antibiotics (OXA, VAN) selected by experts, further referred to as OXA_VAN. Then the “Average Linkage” algorithm (an hierarchical algorithm) is used to define the clusters in OXA_VAN. The clustering result was a dendrogram each level of which corresponds to a different partitioning of the data set. Considering the results of clustering algorithm for 2 to 10 clusters, nine different partitionings are defined. Then the values of $CDBw$ are computed and the plot of its values versus the number of clusters is created (see Figure 13). In this plot we seek the maximum value of $CDBw$. The partitioning of two clusters is proposed as the optimal one since at this value of c (i.e. clustering scheme) the validity index reaches its maximum. The effectiveness of the cluster validity result is verified by the experts who claimed that the proposed clustering scheme corresponds to the real clusters into which the analysed set of organisms can be classified based on their resistances to the antibiotics OXA and VAN. More specifically, the first group refers to *Sau* organisms that are resistant to VAN and susceptible to OXA while the second one is the subclass of *Sau* organisms that present resistance to both OXA and VAN. Then, the $CDBw$ approach assists with discovering significant patterns presented in the underlying data.

6.2 Optimal algorithm selection

In previous sections, we performed experiments with different $ipvs$ for clustering algorithms. In the sequel we show that $CDBw$, assuming a data set and a set of clustering algorithms, enables the selection of the optimal partitioning among those defined by different clustering algorithms. Thus, the

Table 1: Optimal partitioning found by *CDbw* for different clustering algorithms

No clusters	K-Means		DBSCAN		CURE r =10, a=0.3		CLUTO (clmethod=graph -sim=dist - agglofrom=20)	
	Input	CDbw Value	Input	CDbw Value	Input	CDbw Value	Input	CDbw Value
6	C=6	0.01036	-	-	C=6	0.01149	6	0.0237
5	C=5	0.02257	-	-	C=5	0.02768	5	0.0263
4	C=4	0.02009	-	-	C=4	0.02432	4	0.0264
3	C=3	0.01993	Eps=2,MinPts=4	0.032	C=3	0.02004	3	0.032
2	C=2	0.02743	Eps=10,MinPts=4	0.02743	C=2	0.02743	2	-

(a) DS1

No clusters	K-Means		DBSCAN		CURE r =10, a=0.3		CLUTO (clmethod=graph - sim=dist -agglofrom=30)	
	Input	CDbw Value	Input	CDbw Value	Input	CDbw Value	Input	CDbw Value
9	9	7.809E-4	-	-	9	8.491E-4	9	0.0024
8	8	7.661E-4	Eps=13,MinPts=35	0.00366	8	9.849E-4	8	0.0037
7	7	7.495E-4	Eps=12,MinPts=17	0.0027	7	0.00113	7	0.0028
6	6	7.844E-4	Eps=13,MinPts=15	0.0023	6	4.7152E-4	6	0.00209
5	5	6.115E-4	Eps=14,MinPts=15	0.0016	5	4.378E-4	5	0.00155
4	4	7.242E-4	Eps=15,MinPts=15	0.0011	4	4.8914E-4	4	0.00113
3	3	0.00103	Eps=16,MinPts=15	0.0018	3	6.5291E-4	3	0.00179
2	2	8.078E-4	Eps=22,MinPts=15	0.0012	2	1.1606E-4	2	-

(b) DS3

No clusters	K-Means		DBSCAN		CURE r =10, a=0.3		CLUTO (clmethod=graph, sim=dist, agglofrom=30)	
	Input	CDbw Value	Input	CDbw Value	Input	CDbw Value	Input	CDbw Value
6	C=6	0.0542	-	-	C=6	0.01234	6	0.06737
5	C=5	0.0440	-	-	C=5	0.0616	5	0.08407
4	C=4	0.0307	Eps=1,MinPts=4	0.1057	C=4	0.0272	4	0.1026
3	C=3	0.0175	Eps=2,MinPts=15	2.89E-06	C=3	0.0229	3	0.0518
2	C=2	0.0494	Eps=2,MinPts=10	0.0749	C=2	0.0408	2	0.0749

(c) DS4

clustering algorithm that finds the optimal partitioning can be selected. Let S be a data set on which we ran different clustering algorithms $A=\{A_i\}$ using for each of them the optimal ipvs, P_{opt} , as found by *CDbw*. Let $\{C_k(A(P_{opt}))\}$ be the clustering schemes resulting from the execution of the aforementioned algorithms. It is noteworthy that the values of *CDbw* are comparable for different clustering algorithms since the definition of the validity index only depends on the partitioning and not on the algorithm itself.

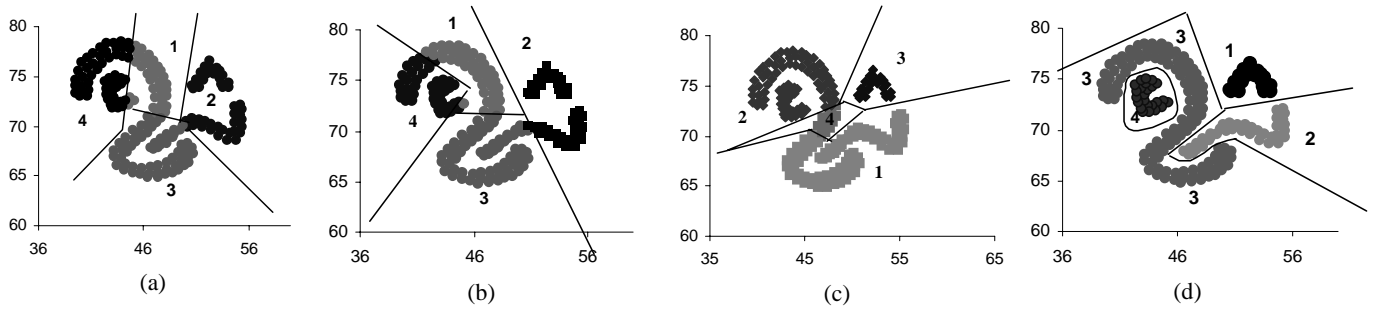
Then, we find the algorithm that results in the optimal partitioning for the data set under concern by solving $\max_{A_i \in A} \left\{ \max_{\{C_k(A_i(P_j))\}} \{C_{Dbw}(C_k(A_i(P_j)))\} \right\}$. In other words, for each of the clustering algorithms A_i the optimal partitioning, C_{opti} , is selected. Then $C_{opt} = \{C_{opti}(A_i(P_{opt})), A_i \in A\}$ is perceived as the set of optimal partitionings defined by the algorithms under concern. The partitioning with the highest *CDbw* value in C_{opt} is selected as the overall optimal partitioning and the respective algorithm running with

the Popt set of ipvs is considered to be the most appropriate algorithm for the data set under concern. In the sequel, we present the results of an experimental study we carried out using four widely used algorithms, one from each of the popular clustering algorithm categories: K-Means (partitional), DBSCAN (density-based), CURE (hierarchical) and an algorithm provided by the CLUTO toolkit which is based on the two-phase clustering approach of the CHAMELEON [KH+99] algorithm.

The goal of the experimental results discussed below is not to evaluate the algorithms themselves and make inferences about their performance. On the other hand, we aim to show that the proposed cluster validity approach is independent of the clustering approach used to define the partitioning of a data set. It is applied to the results of any clustering algorithm and achieves to select the partitioning that best fits the data under concern. Then the clustering algorithms are only considered to be the tools that assist with the definition of the results to which the cluster validity approach will be applied.

Table 1a presents the $CDbw$ values for the clustering schemes of the synthetic data set DS1 as defined by K-Means, DBSCAN, CURE and the CLUTO algorithm respectively. In the context of this experimental study we evaluate the algorithm results for the set of the specific ipvs that is used. As Figure 2 depicts, the real clusters in DS1 are three. However, the majority of clustering algorithms fail to partition it in a right way. Figure 2(a) and Figure 2 (b) present the proposed partitioning of DS1 into three clusters as defined by K-Means and CURE respectively while Figure 2 (c) presents the set of three clusters as defined by DBSCAN and the CLUTO clustering algorithm. We observe that DBSCAN and the CLUTO algorithm are the only algorithms that achieve to discover the real clusters of DS1. This is also verified by Table1a, which presents the values of $CDbw$ for the clustering schemes defined by the considered clustering algorithms. $CDbw$ takes its maximum value for the clustering scheme of three clusters as defined by DBSCAN and CLUTO, which also corresponds to the actual partitioning of DS1. It is also interesting that the values of $CDbw$ corresponding to the optimal partitioning defined by DBSCAN and CLUTO are the same (0.32). This justifies our claim that $CDbw$ values do not depend on the algorithms.

The next data set containing clusters with non-standard geometries that we used was DS3 (see Figure 8). More specifically, we consider the clustering schemes defined by the algorithms mentioned above while their input parameters' values are depicted in Table 1b. In case of DS3, $CDbw$ takes its maximum value for the partitioning of eight clusters as defined by the CLUTO algorithm. We note that



--DS4--

Figure 14: Partitioning of DS4 into four clusters as defined by (a) K-Means, (b) CURE, (c) the CLUTO algorithm and (d) DBSCAN

these are the real eight clusters presented in the DS3 as Figure 8 also depicts. On the other hand, DBSCAN running with the set of ipvs, Eps=13, MinPts=35, discovers a set of eight clusters that approximates the real ones but it considers a part of the cluster 8 (in Figure 8b) as noise. This observation justifies the slight decrease of the *CDbw* value in relation with its values when the real eight clusters are defined. Also the partitionings of DS3 into eight clusters as defined by K-Means and CURE are depicted in Figure 8c and Figure 8d respectively. It is obvious that all algorithms except the CLUTO algorithm fail to partition it properly even in case that the correct number of clusters (i.e. $c=8$) is considered.

Another example that an algorithm clusters a data set finding the correct number of clusters but the wrong partitions is presented in Figure 14. *CDbw* is able to evaluate the results of different clustering algorithms and select the optimal partitioning among those defined by the aforementioned algorithms, i.e. to select the optimal algorithm for a data set. This is verified by the experiments. According to Table 1c, in case of DS4 (see Figure 14), *CDbw* takes its maximum value for the partitioning of four clusters as defined by DBSCAN. Figure 14c presents the partitioning of DS4 into four clusters as defined by DBSCAN while the clustering result of K-Means, CURE and the CLUTO algorithm into four clusters is presented in Figure 14a and Figure 14c respectively. It is obvious that K-Means, CURE (with $a=0.3$, $r=10$) and the CLUTO algorithm (clmethod=graph, sim=dist, agglfrom=30) failed to partition DS4 properly, even in case that the correct number of clusters (i.e. $c=4$) is considered.

Conclusion: Based on above experimental study we conclude that *CDbw* does not only select the optimal partitioning among the results of a specific clustering algorithm but also assists to find the optimal partitioning among the results of different algorithms. Moreover, *CDbw* handles efficiently arbitrarily shaped clusters considering multi-representative points to describe the structure of clusters.

6.3 Comparison to other cluster validity indices

In this Section we compare $CDbw$ to four of the most important validity indices proposed in the literature⁵, such as RS - $RMSSTD$ [26], DB [29], SD [15] and S_Dbw [16]. The definition of the above indices is presented in [13].

$RMSSTD$ and RS are representative examples of statistical validity indices and are jointly taken into account indicating the optimal number of clusters. The optimal partitioning of a data set corresponds to the number of clusters for which a significant local change in values of RS and $RMSSTD$ occurs. As regards DB , SD and S_Dbw the clustering scheme for which the validity indices take its minimum value is selected as the optimal partitioning. We carried out an evaluation study comparing $CDbw$ to the aforementioned validity indices using the 2-dimensional data sets DS1 (see Figure 2), DS3, DS4, (see Figure 8a and Figure 8b respectively) and DS4 (see Figure 14). We also considered the real data set R_D1 and 120-dimensional data set, Nd_Set discussed above.

Table 2 summarizes the results of the validity indices (RS , $RMSSTD$, DB , SD , S_Dbw and $CDbw$), for different clustering schemes of the aforementioned data sets as defined by a clustering algorithm (K-Means, CURE or DBSCAN). The comparison of validity indices refers to the same clustering results for each of the considered data sets. Also, we consider that there is at least a partitioning of the data sets among the evaluated ones that corresponds to their real clusters. In case of DS2, R_D1 and Nd_Set we use the results of the algorithms K-Means and CURE. Indices $CDbw$, SD_bw and DB select the correct number of clusters (i.e. seven) as optimal partitioning for DS2 whereas SD and $RSDDT&RS$ select six and five clusters respectively. Considering the different partitionings of R_D1 as defined by CURE the validity indices $RMSSTD$ and RS propose the partitioning of three clusters. Also the cluster validity indices DB , SD select the clustering scheme of two, five and eight clusters respectively whereas S_Dbw and $CDbw$ consider the partitioning of four clusters as the optimal one. We note that among the partitionings considered for R_D1 the one of four clusters best fits the real clusters in the underlying data. As regards Nd_Set, SD proposes three clusters as its optimal partitioning, while S_Dbw selects the partitioning of eight clusters. On the other hand, $CDbw$, $RMSSTD&RS$ and DB propose four clusters, which corresponds to the number of clusters in the actual partitioning of Nd_Set.

⁵ Though Calinski and Harabasz (1974) is proposed as one of the best indices in [21] we exclude it from this study since it is predecessor of the four considered validity indices (RS , $RMSSTD$, DB , SD and S_Dbw). Moreover its definition is based on terms that are similar to these used by the most recent indices $RMSSTD$ and RS (see [26]).

In case of DS1, DS3 and DS4 (containing arbitrarily shaped clusters), we consider the results of DBSCAN and the CLUTO algorithm since they achieve to handle efficiently arbitrarily shaped clusters. *CDbw* finds the

Table 2: Optimal partitioning proposed by validity indices compared with *CDbw**

	DS1	DS2	DS3	DS4	R_D1	Nd_Set
Actual partitioning	3	7	8	4	4	4
RS, RMSSTD	2	5	3	3	3	4
DB	2	7	4	3	2	4
SD	2	6	4	2	5	2
S_Dbw	2	7	10	3	4	8
CDbw	3	7	8	4	4	4

* In this table the numbers in the cells represent the clustering scheme containing the respective number of clusters.

real four clusters as the optimal partitioning for DS4 (see Figure 8c), on the contrary to *RS&RMSSTD*, *S_Dbw* and *DB*, which propose three clusters as the optimal partitioning and *SD* that proposes the partitioning of two clusters. In case of DS1, only *CDbw* proposes the partitioning into three clusters while all the others select the partitioning of two clusters as the optimal one. Similarly, *CDbw* selects the partitioning that contains the real eight clusters as the optimal one for DS3 whereas *RMSSTD&RS*, *DB*, *SD* and *S_Dbw* fail, selecting the clustering schemes of three, four and ten clusters respectively. Based on the above observations we conclude that *CDbw* achieves to find the clustering scheme that best fits a data set, while other validity indices fail in some cases, especially when the data sets contain arbitrarily shaped clusters.

7. CONCLUSIONS

The main concern in the clustering process is to reveal the *optimal partitioning* of the data set into clusters. In most of the cases the users visually verify the clustering results. However, in case of voluminous and/or multidimensional data sets where efficient visualization is difficult or even impossible, it becomes tedious to know if the results of clustering are valid or not.

In this paper, we defined a new validity index, *CDbw*, and a methodology that given a data set, S, and a set of algorithms $A=\{algi\}$ enables: i) finding the set of ipvs that lead each algi to the best possible clustering results and ii) taking into account results of (i), finding algi that returns the optimal partitioning of S among those defined by the algorithms under concern.

CDbw adjusts well to non-spherical and skewed cluster geometries, contrary to the validity indices proposed in the literature. It achieves this by considering multi-representative points per cluster. The reliability of the validity index is theoretically studied and it is proved that *CDbw* reaches its maximum when the optimal partitioning is achieved. The proposed cluster validity index is fully implemented

and experiments prove its efficiency for various data sets and algorithms. The current experimentation uses the results of at least four algorithms, covering all clustering algorithm's categories.

The proposed approach is defined in the context of crisp clustering corresponding to the results of the majority of clustering algorithms. However, the issue of handling uncertainty is very important with regard to the quality of clustering results. Then one of our future work direction will be the definition of the cluster validity index so that both geometrical and fuzzy aspects of data distribution to be taken in account. Furthermore, we plan an extension of this effort to be directed towards an integrated algorithm for cluster analysis putting emphasis on the geometric features of clusters, using sets of representative points, or even multidimensional curves.

ACKNOWLEDGEMENTS

We wish to thank C. Rodopoulos and C. Amanatidis for the implementation of CURE algorithm. We are indebted to our colleagues at the Medical School of National and Kapodistrian University of Athens Associate Prof. A. Vatopoulos and Dr. J. Papaparaskevas, who provided all the necessary information for the experiments on epidemiological data. Also we are grateful to Drs Joerg Sander and Eui-Hong (Sam) Han for providing information and the source code for DBSCAN and CURE algorithms respectively. This work was partially funded by the Information Society Technologies program of the European Commission, Future and Emerging Technologies under the IST-2001-33058 PANDA project (2001-2004).

REFERENCES

- [1] R. Agrawal, J. Gehrke, D. Gunopulos, & P. Raghavan, (1998), Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, In *Proceedings ACM SIGMOD International Conference on Management of Data*, June 2-4, Seattle, Washington, USA
- [2] M. Berry, & G. Linoff, (1996). *Data Mining Techniques For marketing, Sales and Customer Support*. John Willey & Sons, Inc.
- [3] R. N. Dave, (1996). Validating fuzzy partitions obtained through c-shells clustering, *Pattern Recognition Letters*, Vol .17 (pp 613-623).
- [4] J. C. Dunn, (1974). Well-separated clusters and optimal fuzzy partitions, *J. Cybern.* Vol.4 (pp. 95-104),
- [5] M. Ester, H-P. Kriegel, J. Sander, M. Wimmer, & X. Xu, (1998). Incremental Clustering for Mining in a Data Warehousing Environment, In *Proceedings of 24th VLDB Conference*, New York, USA.
- [6] M. Ester, H-P Kriegel, J. Sander, & X. Xu, (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, In *Proceedings of 2nd Int. Conf. On Knowledge Discovery and Data Mining*, Portland, OR (pp. 226-231).
- [7] U. Fayyad, G. Piatetsky-Shapiro, P. Smuth, & R. Uthurusamy, (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press.
- [8] U. Fayyad, & R. Uthurusamy, (1996). Data Mining and Knowledge Discovery in Databases, *Communications of the ACM*. Vol.39, No11, November.
- [9] I. Gath, & B. Geva (1989). Unsupervised Optimal Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 11, No7, July.
- [10] S. Guha, R. Rastogi, & K. Shim, (1998). CURE: An Efficient Clustering Algorithm for Large Databases, In *Proceedings ACM SIGMOD International Conference on Management of Data*, June 2-4, Seattle, Washington, USA.

- [11] S. Guha, R. Rastogi, & K. Shim, (1999). ROCK: A Robust Clustering Algorithm for Categorical Attributes, In *Published in the Proceedings of the IEEE Conference on Data Engineering*, Data, June 1-3, 1999, Philadelphia, Pennsylvania, USA.
- [12] A. Hinneburg, & D. Keim, (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New York City, August 27-31.
- [13] M. Halkidi, Y. Batistakis, & M. Vazirgiannis, (2001). On Clustering Validation Techniques”, *Journal of Intelligent Information Systems Journal*, 17:2/3, (107-145).
- [14] D. Hochbaum and D. B. Shmoys (1985). "A best possible heuristic for the k-center problem", *Mathematics of Operations Research*, Vol 10:180--184.
- [15] M. Halkidi, M. Vazirgiannis, & Y. Batistakis, (2000). Quality scheme assessment in the clustering process, In *Proceedings of PKDD Conference*, Lyon, France.
- [16] M. Halkidi, & M. Vazirgiannis, (2001). Clustering Validity Assessment: Finding the optimal partitioning of a data set, In *Proceedings of ICDM Conference*, California, USA, November.
- [17] George Karypis, Eui-Hong (Sam) Han, and Vipin Kumar (1999). "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling", *IEEE Computer*.
- [18] G. Karypis. *CLUTO: A clustering Toolkit*. Release 2.1.1. <http://www-users.cs.umn.edu/~karypis/cluto/>
- [19] Kleinberg. "An impossibility theorem for clustering". In Proc. of the 16th conference on Neural Information Processing Systems, 2002.
- [20] A.K Jain, M.N. Murty, &P.J. Flynn, (1999). Data Clustering: A Review, *ACM Computing Surveys*, Vol. 31, No3.
- [21] G.W. Milligan, & M.C. Cooper, (1985). "An Examination of Procedures for Determining the Number of Clusters in a Data Set", *Psychometrika*, 50 (159-179).
- [22] R. Ng, J. Han, (1994). "Efficient and Effective Clustering Methods for Spatial Data Mining". *Proceeding of the 20th VLDB Conference*, Santiago, Chile.
- [23] T.F. O'Brien, & J.M. Stelling, (1995). WHONET: an information system for monitoring antimicrobial resistance. *Emerging Infectious Diseases*, 1995; 1: 66-__.
- [24] R. Rezaee, B. Lelieveldt, & J. Reiber, (1998). "A new cluster validity index for the fuzzy c-mean", *Pattern Recognition Letters*, 19 (237-246).
- [25] C. Sheikholeslami, S. Chatterjee, & A. Zhang, (1998). WaveCluster: A-MultiResolution Clustering Approach for Very Large Spatial Database. In *Proceedings of 24th VLDB Conference*, New York, USA.
- [26] S.C Sharma, (1996). *Applied Multivariate Techniques*. John Willwy & Sons.
- [27] P. Smyth, (1996). Clustering using Monte Carlo Cross-Validation. In *Proceedings of KDD Conference* (126-133).
- [28] Y. Theodoridis. *Spatial Datasets: an "unofficial" collection*. <http://dias.cti.gr/~ytheod/research/datasets/spatial.html>
- [29] S. Theodoridis, & K. Koutroubas, (1999). *Pattern recognition*, Academic Press.
- [30] G. W. Milligan, S.C. Soon, & L. M. Sokol, (1983). "The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5 (40-47).
- [31] X. Xie, & G. Beni, (1991). A Validity measure for Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and machine Intelligence*, Vol.13, No4, August.
- [32] T. Zhang, R. Ramakrishnan, & M. Linvy, (1996). BIRCH: An Efficient Method for Very Large Databases, In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, June 4-6, Montreal, Canada.
- [33] PENED: Research & Development for Knowledge Discovery in Medical Data. <http://www.db-net.aueb.gr/projects/pened/>
- [34] JDBC Technology. <http://java.sun.com/products/jdbc/>