

Diversity control for improving the analysis of consensus clustering

Milton Pividori^{a,b,*}, Georgina Stegmayer^a, Diego Milone^a

^a*Research Institute for Signals, Systems and Computational Intelligence, sinc(i), UNL-CONICET, Department of Informatics, FICH, Universidad Nacional del Litoral, Ciudad Universitaria CC 217, Santa Fe 3000, Argentina*

^b*Center of Research and Development of Information Systems Engineering, CIDISI, UTN-CONICET, Department of Information Systems Engineering, FRSF, Universidad Tecnológica Nacional, Lavaisse 610, Santa Fe 3000, Argentina*

Abstract

Consensus clustering has emerged as a powerful technique for obtaining better clustering results, where a set of data partitions (ensemble) are generated, which are then combined to obtain a consolidated solution (consensus partition) that outperforms all of the members of the input set. The diversity of ensemble partitions has been found to be a key aspect for obtaining good results, but the conclusions of previous studies are contradictory. Therefore, ensemble diversity analysis is currently an important issue because there are no methods for smoothly changing the diversity of an ensemble, which makes it very difficult to study the impact of ensemble diversity on consensus results. Indeed, ensembles with similar diversity can have very different properties, thereby producing a consensus function with unpredictable behavior.

*Corresponding author

Email addresses: mpividori@sinc.unl.edu.ar (Milton Pividori),
gstegmayer@sinc.unl.edu.ar (Georgina Stegmayer), dmilone@sinc.unl.edu.ar
(Diego Milone)

URL: www.sinc.unl.edu.ar (Milton Pividori)

In this study, we propose a novel method for increasing and decreasing the diversity of data partitions in a smooth manner by adjusting a single parameter, thereby achieving fine-grained control of ensemble diversity. The results obtained using well-known data sets indicate that the proposed method is effective for controlling the dissimilarity among ensemble members to obtain a consensus function with smooth behavior. This method is important for facilitating the analysis of the impact of ensemble diversity in consensus clustering.

Keywords: Cluster ensembles, consensus clustering, diversity analysis, diversity control, ensemble diversity

1. Introduction

Clustering is fundamental for understanding the structure of data [45] and it has been used in a wide range of areas, including engineering, financial, biological science, and medical applications [29, 25, 34, 44, 40]. However, the correct choice of a clustering algorithm, or even setting its parameters, requires knowledge of the data set and the data distribution assumed by algorithms, since they can strongly affect the final results obtained [22]. Clustering algorithms have been developed to solve a wide range of different problems, but there is no universal method that can be applied to solve all. Thus, different but equally valid solutions can be obtained using various algorithms, which is one reason why clustering is considered to be an ill-posed problem among researchers [24, 50].

In the past decade, consensus clustering (or cluster ensembles) has emerged as a powerful approach for mitigating the issues of conventional cluster

analysis. First, a set of data partitions is generated, which is called an *ensemble*. Next, a *consensus function* combines the ensemble into a consolidated solution or *consensus partition*, which has greater overall accuracy [39, 21, 19, 54, 56, 42, 10, 18]. Given the ill-posed nature of clustering, the accuracy is typically measured by comparing the final solution with a known reference partition, which is generally based on the class labels associated with the data set [20, 50, 30, 41, 46]. Although this reference partition might not be the only valid structure for the data, many studies have tried to determine how ensembles should be built, and which characteristics they should have to obtain high accuracy. In particular, among these characteristics, the level of disagreement between ensemble members, which is called the *ensemble diversity*, has been identified as a key factor in the cluster ensemble problem [7, 20, 17], and various diversity measures have been proposed [15, 14, 2, 33, 55].

Many strategies have been used to explore how diversity affects consensus performance [8, 20], where they usually aim to generate a set of ensembles with different diversity, before observing the performance of the consensus function. One of the most common approaches involves generating the ensemble members by randomly varying a parameter [8, 27, 15, 17], which can be the clustering algorithm itself [36, 28], the number of clusters [12, 49, 57], or its initialization [37, 26]. Instead of changing the clustering algorithm, a common method involves changing the data by randomly selecting subsamples [11, 32, 38, 48], using different features [43, 52, 37, 23, 51], employing random projections [8, 37, 39], or combining several methods together [47, 53]. An alternative to the purely random approach generates the

ensemble that maximizes a given criterion. First, a pool of partitions is created by using the strategies described above and a subset of this pool is then selected, which maximizes the objective function. For example, this greedy approach was used in [20], where a set of criteria were defined to obtain low, medium, and highly diverse ensembles.

These methods have been used widely to explore the dissimilarity within an ensemble, but the results obtained indicate that there is an important problem with current methods for ensemble diversity analysis. Indeed, previous studies provide opposing opinions regarding this issue, where some have suggested that more diverse ensembles are better for obtaining more accurate solutions [8, 20], whereas others have proposed that moderate diversity is the preferred choice [15]. In addition to these contradictory results, high variability has been found not only among data sets but also when different ensemble generation strategies are employed. Moreover, plots of the accuracy as a function of diversity have shown that ensembles with similar diversity can differ greatly in their accuracy [15, 20]. These confusing results show that current approaches can generate diversity but they cannot control it, and this limitation may lead to unpredictable outputs by the consensus method. This is an important issue and it must be addressed before any analysis of the impact of diversity on consensus clustering. This unpredictable behavior occurs because as one diversity measure is being observed, another properties of the ensemble are changing, thereby leading to erratic behavior by the consensus function. In addition, it is difficult to generate ensembles that are uniformly distributed in the diversity range under evaluation, which could lead to a biased analysis. Both of these reasons demonstrate the need to control the

ensemble diversity in order to effectively analyze its impact on the consensus results.

Due to the importance of diversity in consensus clustering, the issues highlighted above motivated us to unveil a new problem in this area and to propose a novel method that allows fine-grained control of the ensemble diversity. To the best of our knowledge, no methods have been proposed previously for controlling disagreement among ensemble partitions. Our method extracts information from the ensemble structure and then uses it to make small changes, which decrease and increase the diversity among ensemble members in a smooth manner. The results that we obtained using six well-known data sets demonstrate that this method is effective for controlling the ensemble diversity, where the consensus function behaves in a smooth manner, thereby providing a novel approach for studying the impact of diversity on consensus clustering.

The remainder of this paper is organized as follows. In Section 2, we explain the problems with current methods and we define the steps in the diversity control method. Section 3 describes the data sets and performance measures used for testing. Section 4 presents the evaluation procedure and the results obtained. In Section 5, we summarize our conclusions as well as suggesting possible improvements and future research.

2. Novel method for controlling diversity

Current methods assume implicitly that ensembles with a particular level of diversity are comparable; thus, equal diversity values should represent similar ensembles, or at least similar inputs for the consensus function, which is

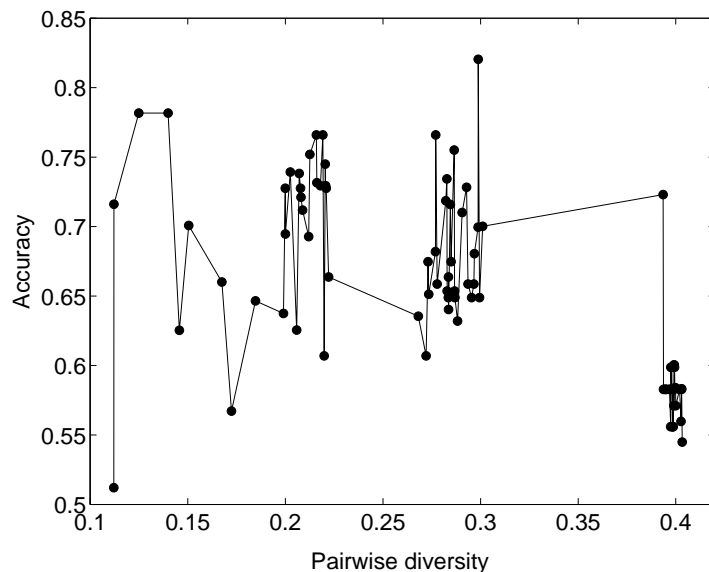


Figure 1: Accuracy of the consensus partition as a function of the pairwise diversity for the Wine data set using the method proposed by [20].

expected to produce similar results. However, this might not be the case in practice. An example of such results is shown in Figure 1, where the accuracy of the consensus partition [20] is plotted as a function of the pairwise ensemble diversity. Similar results can be found in [15]. Two problems are evident based on this plot. The first is the behavior of the output consensus accuracy (y-axis) when the pairwise diversity (x-axis) is around 0.21, where the diversity values are close to each other but many points differ greatly in their accuracy. A similar behavior can be observed around a diversity value of 0.28. Thus, ensembles with similar diversity can represent very different inputs for the consensus function. The second problem is that the diversity range is not always sampled uniformly; for example, there are far less ensembles with diversity values in $[0.10, 0.20]$ and even none in $[0.31, 0.38]$. These

Table 1: Notations used in the proposed method for diversity control.

Symbol	Description
π	Single data partition
Π	Ensemble with groups of partitions
q	Number of groups of partitions in Π
Π_i	Group of partitions (member of Π)
p_i	Size of group Π_i
$\tilde{\pi}_i$	Representative partition for group Π_i
Π^*	Ensemble with controlled diversity
π^*	Final consensus partition
ϕ_i	Clusterer
m	Number of clusterers used
n	Number of times a clusterer is run
Υ	Normalized mutual information

two issues make the study of diversity a fairly difficult problem. As stated earlier, a possible explanation for this behavior is that different properties of the ensemble may change while a single diversity measure is being observed. This makes it very difficult to analyze how diversity affects the consensus function, because it is not possible to make strong conclusions when the target variable (the accuracy in this case) exhibits erratic behavior. Furthermore, the analysis is made even more difficult due to the lack of uniformity in the sampling method for the range of diversity under examination. Thus, it is necessary to control the ensemble diversity in a smooth manner before it can be studied correctly.

In this section, we propose a diversity control method. Table 1 presents the notations used and the general process is depicted in Figure 2. The first step generates an ensemble Π from the data. This ensemble is then divided

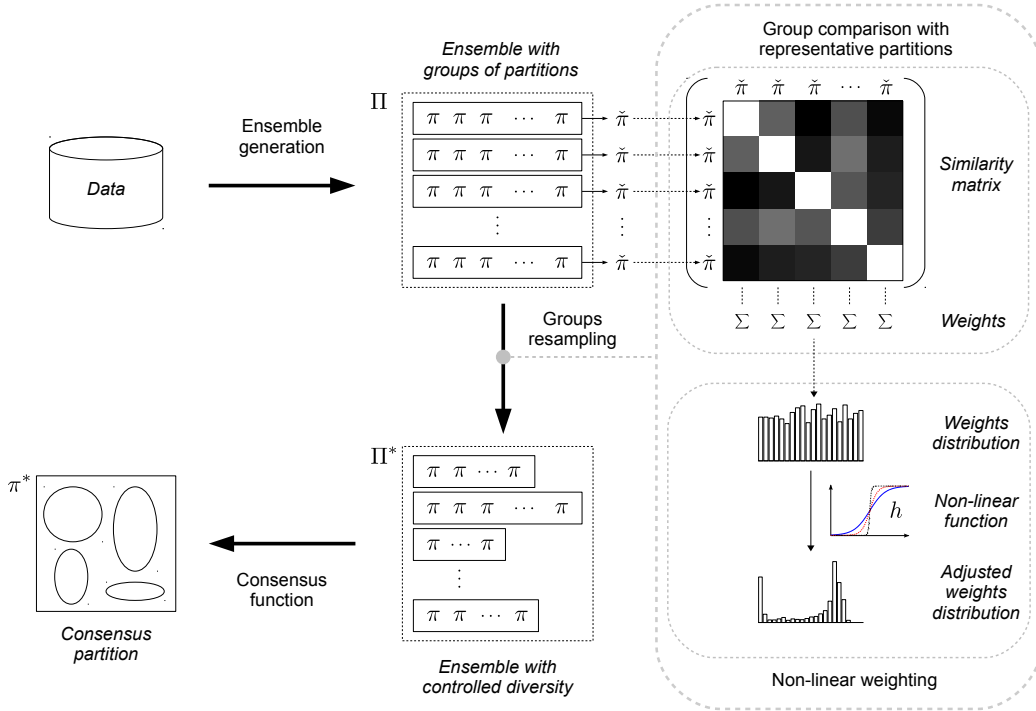


Figure 2: The diversity control method: steps and intermediate outcomes.

into *groups of partitions*, which are created by clustering over the ensemble members. The ensemble with controlled diversity Π^* , which is produced by resampling from the original groups, is finally combined to obtain the consensus partition. The resampling step is shown in detail at the right of the figure. First, information about the ensemble structure is extracted, which is obtained by comparing each group with the others. Instead of a direct group comparison, *representative partitions* $\tilde{\pi}$ are defined for each group. A weight is calculated for each group based on the similarity matrix between the representatives. Next, the distribution of these weights is adjusted using a nonlinear function and they are employed to resample from the groups of partitions in Π . Based on the adjusted distribution, the resampling process

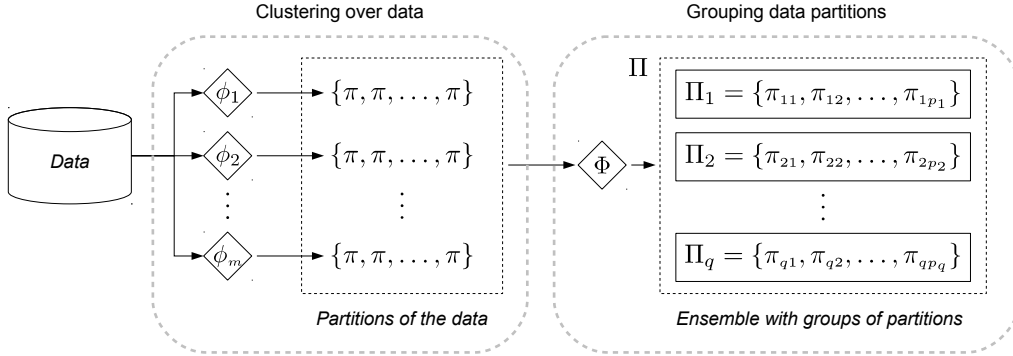


Figure 3: Ensemble generation based on groups of partitions. The ensemble generation process comprises two phases: 1) clustering over data and 2) grouping data partitions.

modifies the group sizes to change the ensemble diversity in a smooth manner. Thus, a new ensemble Π^* is created and its diversity is controlled by the distribution of the weights. Finally, a consensus function is used to derive the final partition π^* from Π^* . The following subsections explain these steps in detail.

2.1. Ensemble generation based on groups of partitions

Ensemble generation, which is the first step of the diversity control method, receives an input data set and produces an ensemble organized into groups of partitions. This step is depicted in detail in Figure 3 and it comprises two phases: 1) *clustering over data* where several partitions are obtained and 2) *grouping data partitions*, which identify different groups of partitions.

Phase 1: Clustering over data. Several methods can be used to create an ensemble. A common approach involves producing different partitions of the data using a single clustering algorithm and randomly varying some of its parameters [26, 20]. In particular, *k*-means is a frequent choice and one of the following schemes for selecting the number of clusters is commonly

employed: k is fixed and the cluster centers are randomly initialized [15, 57]; or k is chosen randomly within an interval $[k_{min}, k_{max}]$ [13, 26, 18]. Our proposed approach uses a combination of both. A diagram illustrating this clustering process is shown in the first part of Figure 3. A clusterer ϕ_i is defined for each k in the chosen interval, where ϕ_i is run n times with random initializations, thereby producing a set of $m \times n$ partitions. This process aims to obtain as many different partitions as possible. The ensembles generated in this manner have a known structure with sets of partitions, where each set is generated by the same clusterer only by varying the initialization.

Phase 2: Grouping data partitions. After producing the initial ensemble, the second phase of the ensemble generation step involves gathering information about its structure. This phase basically applies a clustering algorithm Φ to the previously generated ensemble members, which obtains groups of partitions denoted as Π_i , with size p_i . Therefore, this phase does not modify the initial ensemble and it simply regroups its members. Three grouping algorithms are proposed for this purpose, each of which produces groups of partitions with different properties. The first algorithm is the simplest and it groups the ensemble members according to their number of clusters. Thus, each group of partitions produced using this approach has the same k and this method is called ensemble grouping by k (Gk). The second method, called full ensemble grouping (FG), involves applying a hierarchical agglomerative clustering algorithm to the ensemble members. This method uses the distance $1 - \Upsilon(\pi_i, \pi_j)$, where $\Upsilon(\pi_i, \pi_j)$ is the normalized mutual information (NMI) [36] between partitions π_i and π_j . In contrast to Gk , FG allows to specify the number of groups q to be produced. The third

algorithm, called *FGk*, is a combination of both, where it takes the groups formed by *FG* and then splits them according to *Gk*.

2.2. Group comparison with representative partitions

The previous step of the method produces an ensemble Π with groups of partitions Π_i . The next step involves comparing those groups to obtain information about the structure of Π , which can be achieved simply by performing a pairwise comparison between groups. However, given the structure of Π , it is possible to define a more efficient method for comparing groups. It should be noted that each group member shares some properties with the others, and thus each group of partitions can be considered as a cluster. Therefore, instead of comparing all of the group members, the similarity between two groups can be estimated by comparing their corresponding representative partitions. The representative partition $\tilde{\pi}_i$ (see Figure 2) is the single partition that best represents its group Π_i . A convenient definition of the representative that maximizes the information shared with group members is:

$$\tilde{\pi}_i = \arg \max_{\tilde{\pi}} \sum_{\forall \pi \in \Pi_i} \Upsilon(\pi, \tilde{\pi}). \quad (1)$$

A consensus function can be used as a method for obtaining a partition $\tilde{\pi}$. Group members could certainly have different numbers of clusters (particularly when *FG* is employed), so it is necessary to decide how many clusters the representative partition should contain. This is achieved by obtaining a consensus partition for each k in Π_i , where that with the maximum agreement among group members is finally selected as the representative. For instance,

	$\tilde{\pi}_1$	$\tilde{\pi}_2$	$\tilde{\pi}_3$	$\tilde{\pi}_4$	$\tilde{\pi}_5$
$\tilde{\pi}_1$		0.71	0.80	0.75	0.69
$\tilde{\pi}_2$	0.71		0.79	0.75	0.63
$\tilde{\pi}_3$	0.80	0.79		0.79	0.70
$\tilde{\pi}_4$	0.75	0.75	0.79		0.68
$\tilde{\pi}_5$	0.69	0.63	0.70	0.68	
\bar{r}_i	0.74	0.72	0.77	0.74	0.68

Figure 4: Similarity matrix for the representative partitions generated for the Wine data set. The last row represents the column average (discarding the main diagonal).

if group Π_i contains many data partitions with $k = 3, 5$ and 7 , then three consensus partitions are derived from Π_i , with $3, 5$ and 7 clusters. Thus, the consensus partition that maximizes (1) is selected as the representative for group Π_i .

After obtaining all of the representative partitions, they are compared with each other to produce a similarity matrix

$$r_{ij} = \Upsilon(\tilde{\pi}_i, \tilde{\pi}_j). \quad (2)$$

This matrix contains information about the ensemble structure. The similarity of a group with respect to the other groups is estimated by $\bar{r}_i = \frac{1}{q-1} \sum_{j \neq i} r_{ij}$. An example of a similarity matrix is shown in Figure 4, which compares five representatives. According to this similarity matrix, group Π_3 is most similar to the rest of the ensemble and Π_5 is most different. This information can be used to modify Π and obtain a new ensemble Π^* with

controlled diversity.

2.3. Nonlinear group weighting and resampling

After gathering the information about the ensemble structure, the next stages produce a new ensemble Π^* (with controlled diversity) by adjusting the distribution of \bar{r}_i to obtain a weight w_i for each group and then resampling from the groups in ensemble Π . If small increases in diversity are desired, it is intuitive to increase the size of the most different groups slightly, whereas the proportion of the most similar groups is reduced. By contrast, the opposite operation should decrease the diversity. Both mechanisms provide a way of controlling the diversity.

The similarity of a group with the rest of the ensemble, \bar{r}_i , is used to calculate a weight w_i . The mechanism employed to obtain fine-grained control over ensemble diversity involves applying a nonlinear function to \bar{r}_i values. This function is used to adjust the distribution of \bar{r}_i in a smooth manner, thereby allowing small increases or decreases in diversity. In the proposed method, we employ a sigmoid function and the computation of the group weights is given by

$$w_i(h) = \frac{p_i}{1 + e^{-h(\bar{r} - \bar{r}_i)}}, \quad (3)$$

where p_i is the size of group Π_i , \bar{r} is the mean of \bar{r}_i , and h controls the change in diversity. A new ensemble Π^* can be created for each h (with $\Pi^* = \Pi$ for $h = 0$). Thus, when $h > 0$, the method makes Π^* more diverse than Π because the most different groups in the ensemble receive higher weights. By contrast, when $h < 0$, Π^* has lower diversity than Π because the most similar

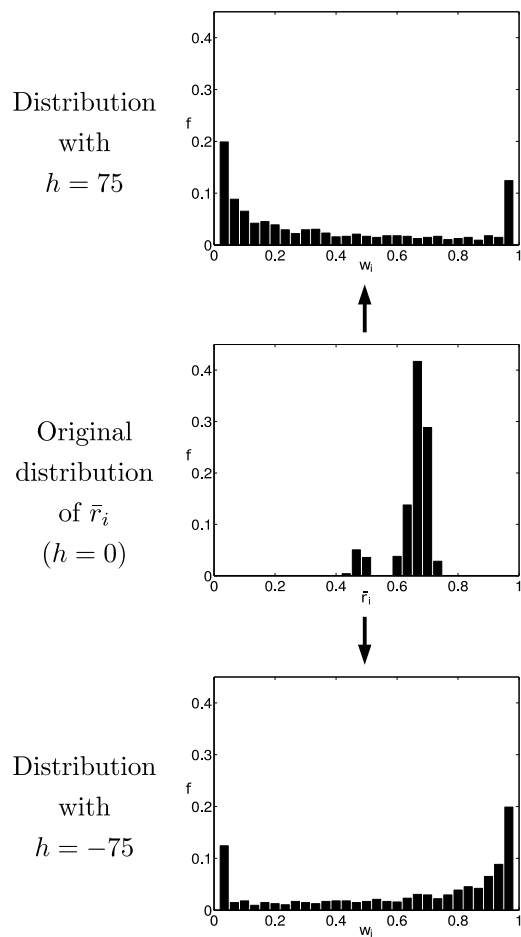


Figure 5: Adjustment of the distribution of \bar{r}_i using a sigmoid function with two different values for parameter h .

groups are favored. It is possible to obtain smooth increases (decreases) in the ensemble diversity by gradually increasing (decreasing) h .

An example of the application of this method is shown in Figure 5, which presents three histograms. The distribution of \bar{r}_i (middle) is modified by using two different values for h (top and bottom). The histogram in the middle shows that \bar{r}_i is greater than the mean for the majority of the groups.

This explains why most of the groups are favored when the method tries to decrease the diversity (see the histogram at the bottom with a sigmoid using $h = -75$). The opposite effect is obtained when the method attempts to increase the ensemble diversity ($h = 75$).

After obtaining the weights w_i , the new groups in ensemble Π^* are produced by resampling the groups in Π with probability $P_i = w_i / \sum_i w_i$. For example, if we suppose that there are three groups, which obtain weights of $w_1 = 0.8$, $w_2 = 0.3$ and $w_3 = 0.4$ for a particular value of h , and the size of ensemble Π^* is 100, then according to P_1 , P_2 and P_3 , 53 of its members will be drawn uniformly from group Π_1 , 20 from Π_2 and 27 from Π_3 . When a slightly different value is used for h to calculate w_i , the amount of partitions drawn from each group will also change slightly, thereby obtaining smooth changes in diversity.

The final step in the overall method involves deriving a consensus partition from ensemble Π^* . In previous studies, several consensus functions have been introduced and different approaches have been used to combine a set of partitions. Among these consensus methods, some of the most popular are based on graph representations [36, 9, 5, 31], and co-association matrices or evidence accumulation [12, 13]. Any consensus approach can be employed during this step in the proposed method.

3. Materials and performance measures

In this section, we describe the data sets and performance measures used to test the proposed method. These measures were used to quantify the performance of the diversity control method in different steps.

Table 2: Descriptions of the artificial and real data sets used in the experiments. N is the number of data points, D the number of dimensions, and k the number of clusters.

Data set	N	D	k
Difficult Doughnut	500	12	2
Four Gaussian	100	12	4
Iris	150	4	3
Ionosphere	351	34	2
Glass	214	9	6
Wine	178	13	3

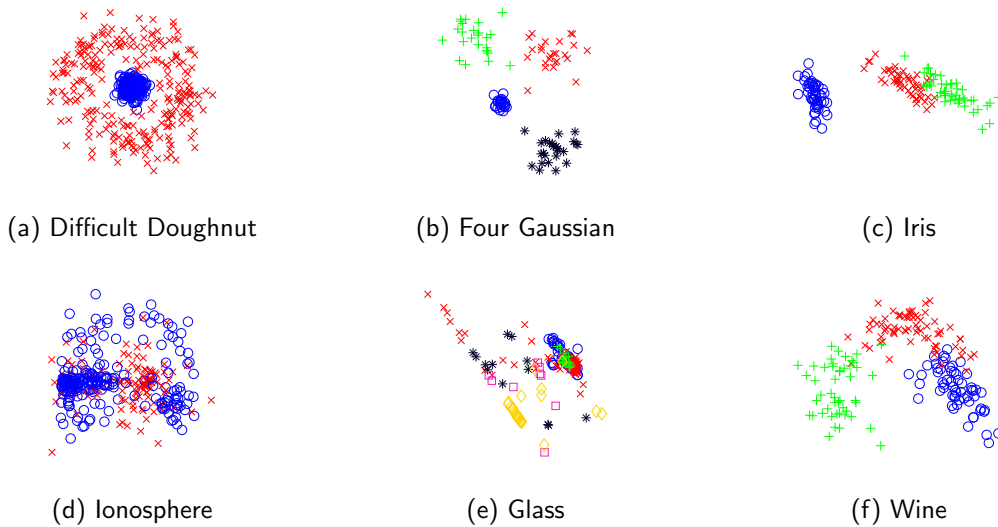


Figure 6: Cluster shapes for the data sets employed. For artificial data sets (6a and 6b) the two noiseless dimensions were used to obtain the plots. For real data sets, principal components analysis projection was employed.

3.1. Data sets

Six well-known artificial and real data sets were employed (Table 2). Difficult Doughnut and Four Gaussian [26] were created artificially by generating

two dimensions of data and then adding 10 dimensions of uniform noise. These two data sets had very different cluster shapes (see Figures 6a and 6b), thereby representing different levels of difficulty for the clustering algorithm. The real data sets were obtained from the UCI repository [1]. As shown in Table 2 and Figure 6, they contained different number of objects, dimensions, and classes, and their clusters had distinct shapes and degrees of overlapping.

3.2. Performance measures

The diversity control method has different steps, so it can be evaluated at the end of each. We assessed its performance using different types of measures, which quantified three distinct aspects of the overall process. The first type comprised *ensemble measures* for quantifying the diversity of the ensembles generated. The next type comprised *clustering measures*, which we used to observe the impact of the generated ensembles on the consensus partition from different perspectives. Finally, a smoothness measure was used to evaluate the fine-grained control over ensemble diversity, which assessed how the ensemble and clustering measures evolved with diversity. A higher level of smoothness was associated with better control over the ensemble diversity.

3.2.1. Ensemble measures

Pairwise diversity. Two pairwise measures were used to quantify the ensemble diversity. Given an ensemble Π^* of size P , the first measures the dissimilarity among the ensemble members [8], which can be defined as the mean of the ensemble members dissimilarities

$$D_{p1} = \frac{2}{P(P-1)} \sum_{i=1}^{P-1} \sum_{j=i+1}^P (1 - \Upsilon(\pi_i, \pi_j)). \quad (4)$$

The second measure, D_{p2} , quantifies the spread of the diversity [15], which is calculated as the standard deviation of the dissimilarities among the ensemble members.

Non-pairwise diversity. The average NMI (ANMI) [36] indicates the average amount of information shared between the consensus partition π^* and the members of its corresponding ensemble Π^* ,

$$\bar{\Upsilon} = \frac{1}{P} \sum_{\forall \pi_i \in \Pi^*} \Upsilon(\pi^*, \pi_i). \quad (5)$$

ANMI can be considered as a level of representativeness of π^* . The complement of (5) can be used as an ensemble diversity measure [15, 27], where D_{np1} is defined as the average of the differences between π^* and each member of Π^* . The standard deviation of these differences represents another measure, which is denoted as D_{np2} .

3.2.2. Clustering measures

These measures comprise external and internal criteria for validating consensus partitions [45, 16]. The external criteria compare the consensus partition with a reference partition of the data, which is drawn independently. By contrast, internal criteria validate π^* using only the inherent information related to the data.

Accuracy: refers to the degree of agreement with a reference or true partition. The accuracy is calculated as

$$A = \Upsilon(\pi^*, \tilde{\pi}), \quad (6)$$

where $\tilde{\pi}$ is the reference partition of the data, which is often based on the class labels (which were available for all of the data sets used in this study).

Compactness: measures the cluster homogeneity in a partition by calculating the intra-cluster variance. The global compactness for partition π^* is given by

$$C = \frac{1}{k} \sum_i^k C_i = \frac{1}{k} \sum_i^k \left(\frac{1}{|\Omega_i|} \sum_{\forall \mathbf{x}_j \in \Omega_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|_2 \right), \quad (7)$$

where k is the number of clusters Ω_i in π^* , \mathbf{x}_j is a data vector, and $\boldsymbol{\mu}_i$ is the centroid of the cluster. Values of C closer to 0 indicate more compact clusters.

Separation: quantifies the degree of separation between clusters in a partition. This measure is given by

$$S = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^k \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2. \quad (8)$$

Higher values of S indicate greater separation of the clusters in partition π^* .

Davies-Bouldin index: this measure combines compactness and separation [4], and it is defined as

$$B = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left(\frac{C_i + C_j}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2} \right). \quad (9)$$

This index is a function of the ratio of the sum of the within-cluster scatter relative to the between-cluster separation. Therefore, B values close to 0 indicate that a partition π^* has compact and separated clusters.

Dunn index: this index combines the dissimilarity between clusters and their diameters [6], and it is given by

$$D = \frac{\min_{0 < m \neq n \leq k} \left\{ \min_{\substack{\forall \mathbf{x}_i \in \Omega_m \\ \forall \mathbf{x}_j \in \Omega_n}} \{ \|\mathbf{x}_i - \mathbf{x}_j\|_2 \} \right\}}{\max_{0 < m \leq k} \max_{\forall \mathbf{x}_i, \mathbf{x}_j \in \Omega_m} \{ \|\mathbf{x}_i - \mathbf{x}_j\|_2 \}}. \quad (10)$$

This index measures the relationship between the minimum inter-cluster distance and the maximum intra-cluster distance. If the partition contains well-separated clusters, the minimum distance among them is usually large and their maximum diameter is expected to be small. Therefore, larger values of D indicate better partitions.

3.2.3. Smoothness measure

Control over the diversity of an ensemble can be assessed by studying the evolution of ensemble and clustering measures, which involves observing the behavior of the inputs of the consensus function (ensembles Π^*) and their corresponding outputs (consensus partitions π^*). On the one hand, if a single diversity measure is being controlled, it is expected that other diversity measures will change in a smooth manner. On the other hand, current approaches assume that ensembles with similar diversity values are similar inputs for the consensus function, and thus they are expected to produce comparable consensus partitions. To verify whether this assumption actually holds, clustering measures can be calculated over the consensus partitions. If they change in a smooth manner, this indicates that ensembles with similar diversity are actually similar inputs for the consensus function.

The autocorrelation coefficient at lag 1 can be used as a smoothness

measure. It is known that this coefficient can be employed to detect non-randomness [3], and this conforms with the intuitive concept of smoothness. From this viewpoint, the randomness in a sequence refers to the independence or unpredictability of one value relative to another. For example, for accuracy, this coefficient can be calculated as

$$\rho_1 = \frac{\sum_d (A_d - \bar{A})(A_{d+1} - \bar{A})}{\sum_d (A_d - \bar{A})^2}, \quad (11)$$

where A is a sequence of accuracies sorted by diversity d and \bar{A} is their mean. Any of the ensemble or clustering measures defined previously can be used instead of the accuracy. Thus, values of ρ_1 closer to 1 indicate the smooth behavior of a sequence of values, so control over the diversity of the ensemble is better. By contrast, ρ_1 near 0 indicates rough changes, thereby demonstrating almost no control of the diversity of an ensemble.

4. Results and discussion

We evaluated the proposed method in several different ways, as described in this section. First, we specified the experimental setup and determined the ability of the method to control the ensemble diversity for all of the data sets described in Section 3.1. After that, the evolution of the clustering measures was used to indicate the effects of the method on the performance of the consensus function. Finally, we calculated the autocorrelation at lag 1 (ρ_1) for the ensemble and clustering measures.

4.1. Experimental setup

For phase 1 in the ensemble generation step (Section 2.1), the base clustering algorithm employed was k -means, with $k_{min} = 2$ and $k_{max} = 20$ ($m = 19$),

and 20 random initializations ($n = 20$) for each k . The number of groups of partitions produced in phase 2 was 19 for both Gk and FG ($q = 19$). The representative of each group and the consensus partition π^* obtained at the final step were derived using a supra-consensus function [36]. This supra-consensus function employed 10 different consensus methods and it selected the method that maximized the ANMI. These internal methods comprised three graph-based functions: Cluster-based similarity partitioning algorithm (CSPA), HyperGraph partitioning algorithm (HGPA) and Meta-CLustering Algorithm (MCLA) [36], as well as seven methods based on evidence accumulation [13]. The latter methods accumulated the results found by all of the ensemble members in a co-association matrix, before applying a hierarchical agglomerative clustering algorithm to derive a consensus partition. Seven types of distances were used as linkage criteria: nearest (single), furthest (complete), unweighted average, weighted average, unweighted center of mass, weighted center of mass, and inner squared (minimum variance). The entire experiment was repeated 50 times.

For further details and testing with more experimental setups, the full source code¹ is freely available for download. In addition, the diversity control algorithm can be tested rapidly using a *web-demo*², which is an accessible web interface developed with [35].

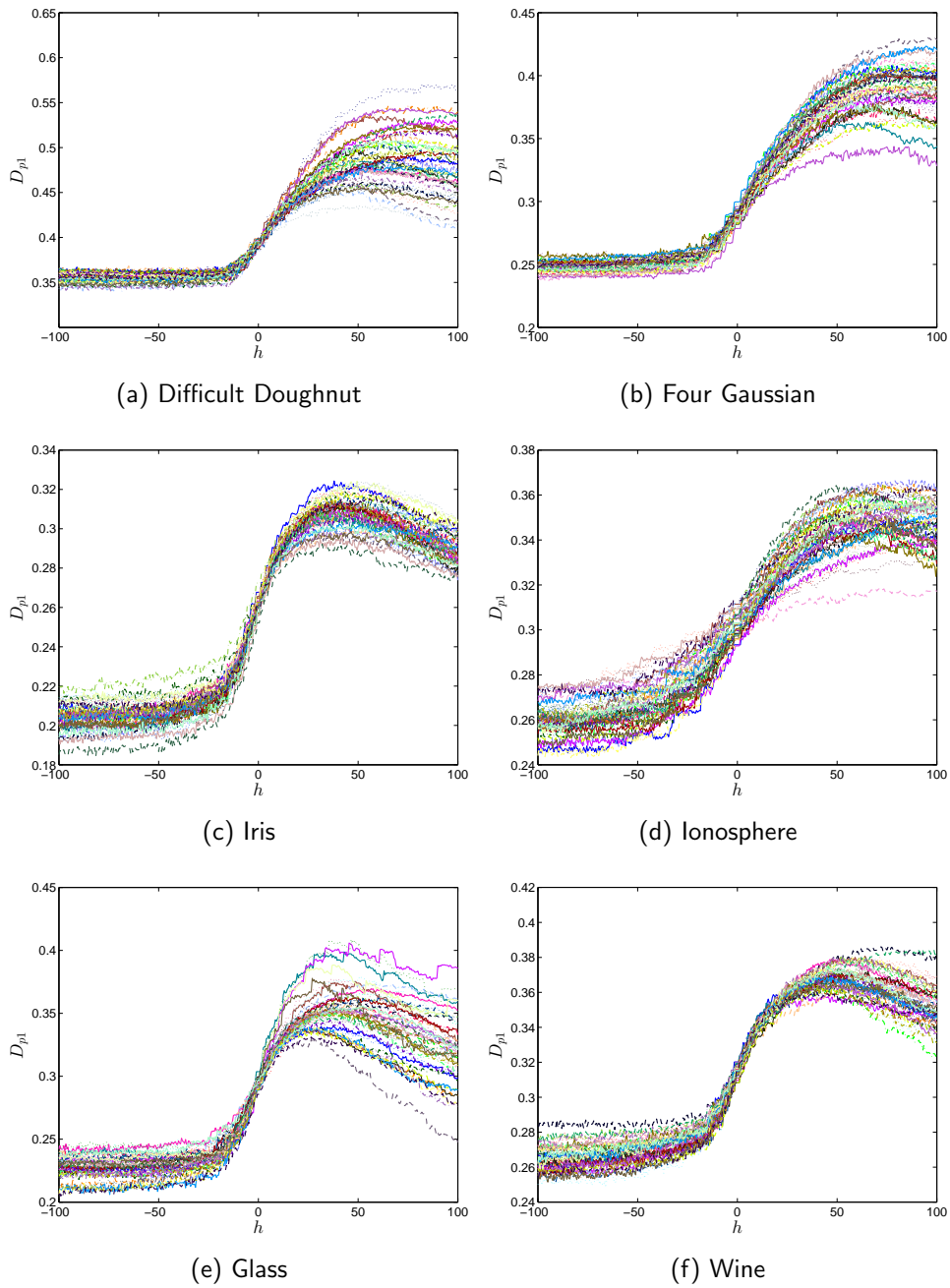


Figure 7: Evolution of the control of diversity for six data sets.

4.2. Control of ensemble diversity

The diversity control results for all of the data sets are shown in Figure 7, in which the pairwise ensemble diversity (D_{p1}) is plotted as a function of the sigmoid parameter h . Each repetition is indicated by a different line style and color. A single ensemble Π^* was created for each h value by considering a wide interval $[-100, 100]$ in order to observe how the method behaved with intermediate cases, but also with the extreme cases. The size of Π^* was 100 and the grouping algorithm was FG.

The results show that although different behaviors were observed for each data set, in all cases the method could increase and decrease the pairwise ensemble diversity by adjusting h . In addition, the method effectively produces a smooth change in D_{p1} , where the full diversity range was uniformly sampled. When the method was used to decrease or increase D_{p1} , each curve finally converged to a certain level, although each arrived at different lower and upper bounds. Saturation occurred at both sides of the curve because when higher values were used for h , minor changes were observed in the sigmoid function, and thus there was almost no difference in the size of the groups among the newly created ensembles. In this extreme situation, the only source of change was the random selection of members from each group of partitions. Therefore, no changes were expected in the ensemble diversity at the highest values of h . Another interesting behavior was observed as the diversity increased ($h > 0$), where a point of maximum diversity was reached and it then decreased to a certain level. This change occurred because high

¹<https://sourceforge.net/projects/sourcesinc/files/divcontrol/0.10/>

²<http://fich.unl.edu.ar/sinc/web-demo/divcontrol/>

values of h made some $w_i \approx 0$, thereby producing empty groups, whereas others w_i were sufficiently large to introduce repeated partitions.

The results show clearly that the proposed method could decrease or increase the ensemble diversity in an effective manner. A wide range was employed for h , but in practice, it would be necessary to focus on a more useful range for h values in the vicinity of $h = 0$, where the diversity exhibits large changes. h can take continuous values, so it is easy to reduce the step size and explore this range, which we tested in the following experiments. The diversity could be controlled independently of the data characteristics by controlling the diversity for data sets with compact and well-separated clusters (e.g., Four Gaussian), as well as those with far more complex structures (such as Ionosphere and Glass). In addition, it should be noted that the full diversity range was uniformly sampled. This results in a fine-grained method to control the level of disagreement among ensemble members in a smooth manner.

4.3. Performance evaluation

Next, we calculated the classical performance measures described in Section 3.2 for the final consensus partition, which allowed us to observe how these quantities evolved for ensembles with controlled diversity. The results are shown in Figure 8 for the six measures calculated based on the Wine data set³. The average of all the repetitions is plotted for each h as well as the corresponding confidence intervals in grey ($\alpha = 0.05$).

In addition to the control of ensemble diversity, as demonstrated ear-

³Full results for all of the data sets are available in the supplementary material.

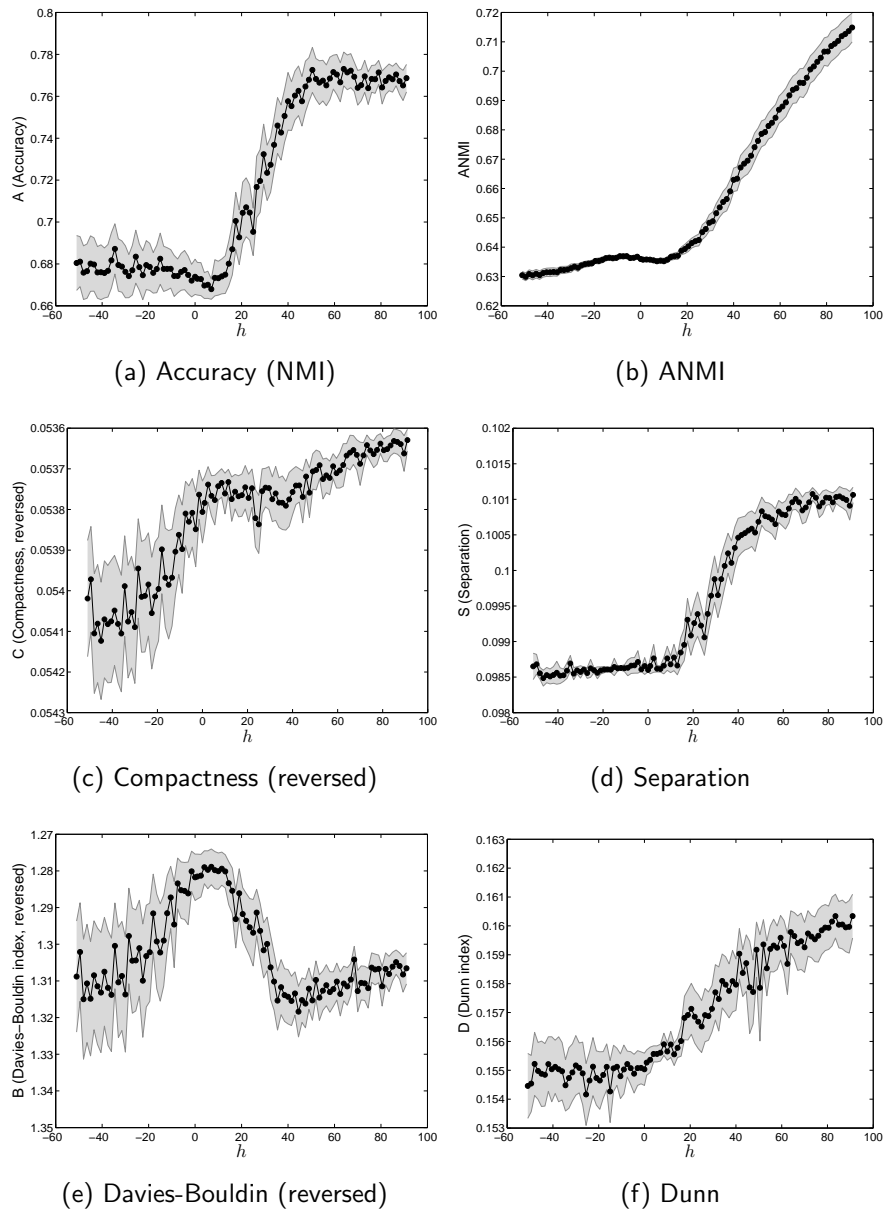


Figure 8: Six performance measures calculated for the consensus partitions obtained from the Wine data set. The compactness and Davies-Bouldin index are presented in reverse, so the values at the top are better for all of the measures.

lier, the results showed that the proposed method could also induce smooth changes in the performance measures. In most cases, the quantities tended to improve with higher values of h (more diverse ensembles). Most of them did not appear to be affected greatly, but the accuracy and ANMI exhibited relatively high increases. Thus, the partitions π^* appeared to have almost equally compact and separated clusters for all values of h , but they definitely represented different partitions of the data. Therefore, the proposed method could change the final partition properties in a smooth manner by controlling the diversity. The ability to produce smooth changes in the performance measures is a very important feature of the proposed method and future studies could explore the influence of the ensemble on the consensus quality.

4.4. Smoothness of the performance measures

The smoothness (ρ_1) was determined for the ensemble and clustering measures using different approaches for ensemble generation. Two existing methods were compared with our proposed approach: a random generation method (RN) described by [15, 8, 27, 17], which randomly creates a set of ensembles with different diversities; and the method proposed by [20], which we refer to as category-based (CB), where this method generates a set of ensembles that belong to “low,” “medium,” and “high” diversity categories. The proposed method for diversity control used three alternative algorithms to generate groups: Gk , FG and FGk . Only a useful range for h was employed in these tests, i.e., a range where the diversity exhibited large changes. Sequences of ensembles were generated for these five methods and sorted by ascending diversity. The average smoothness of the ensemble and clustering measures were determined for all of the repetitions.

Table 3: Smoothness of the ensemble measures.

ρ_1	Difficult Doughnut					Four Gaussian				
	RN	CB	<i>Gk</i>	FG	<i>FGk</i>	RN	CB	<i>Gk</i>	FG	<i>FGk</i>
D_{p2}	0.69	0.95	<u>0.98</u>	<u>0.98</u>	<u>0.98</u>	0.75	0.95	<u>0.98</u>	<u>0.98</u>	<u>0.98</u>
D_{np1}	0.66	0.71	<u>0.96</u>	<u>0.97</u>	<u>0.98</u>	0.15	0.74	<u>0.95</u>	<u>0.97</u>	<u>0.98</u>
D_{np2}	0.48	<u>0.95</u>	0.86	<u>0.97</u>	<u>0.92</u>	0.03	0.57	0.64	0.80	<u>0.89</u>

ρ_1	Iris					Ionosphere				
	RN	CB	<i>Gk</i>	FG	<i>FGk</i>	RN	CB	<i>Gk</i>	FG	<i>FGk</i>
D_{p2}	0.61	0.95	<u>0.98</u>	<u>0.98</u>	<u>0.98</u>	0.17	0.90	<u>0.97</u>	<u>0.98</u>	<u>0.98</u>
D_{np1}	0.07	0.70	<u>0.97</u>	<u>0.97</u>	<u>0.98</u>	0.65	0.62	<u>0.98</u>	<u>0.98</u>	<u>0.98</u>
D_{np2}	0.10	0.84	<u>0.96</u>	<u>0.97</u>	<u>0.97</u>	0.50	0.72	<u>0.97</u>	<u>0.97</u>	<u>0.98</u>

ρ_1	Glass					Wine				
	RN	CB	<i>Gk</i>	FG	<i>FGk</i>	RN	CB	<i>Gk</i>	FG	<i>FGk</i>
D_{p2}	0.51	0.95	0.97	0.97	<u>0.98</u>	0.56	0.93	<u>0.98</u>	<u>0.98</u>	<u>0.98</u>
D_{np1}	0.46	0.92	<u>0.94</u>	<u>0.96</u>	<u>0.97</u>	0.02	0.64	<u>0.96</u>	<u>0.96</u>	<u>0.96</u>
D_{np2}	0.02	0.86	<u>0.92</u>	0.88	<u>0.94</u>	0.17	0.73	<u>0.93</u>	<u>0.94</u>	<u>0.93</u>

The smoothness results for the ensemble measures are presented in Table 3, where the average ρ_1 was calculated for three diversity measures: D_{p2} , D_{np1} and D_{np2} . Bold numbers indicate the maximum ρ_1 obtained for each data set and ensemble measure. The differences between underlined and non-underlined values are statistically significant ($\alpha = 0.05$). For all of the data sets, we found that RN always obtained the lowest scores for ρ_1 . Thus, the ensemble diversity values for D_{p2} , D_{np1} and D_{np2} changed roughly compared with D_{p1} . By contrast, CB obtained higher ρ_1 values, which indicates that it could generate ensembles that changed more smoothly than the random approach. This may be explained by the “greedy” nature of CB when creating the diversity categories because although it contains a random com-

Table 4: Smoothness of the clustering measures.

ρ_1	Difficult Doughnut					Four Gaussian				
	RN	CB	Gk	FG	FGk	RN	CB	Gk	FG	FGk
<i>A</i>	0.03	<u>0.91</u>	0.77	0.95	0.82	0.03	0.64	0.60	<u>0.76</u>	0.79
<i>C</i>	0.02	<u>0.88</u>	0.68	0.93	0.77	0.03	0.47	0.55	0.66	0.75
<i>S</i>	0.02	<u>0.92</u>	0.77	0.95	0.82	0.04	0.47	<u>0.60</u>	<u>0.72</u>	0.77
<i>B</i>	0.02	0.89	0.80	0.95	0.82	0.03	0.17	<u>0.60</u>	<u>0.71</u>	0.78
<i>D</i>	0.28	0.35	0.51	0.73	0.57	0.02	0.57	0.65	0.74	0.80

ρ_1	Iris					Ionosphere				
	RN	CB	Gk	FG	FGk	RN	CB	Gk	FG	FGk
<i>A</i>	0.04	0.39	0.61	0.81	0.71	0.06	0.58	0.32	0.43	0.41
<i>C</i>	0.01	<u>0.45</u>	0.29	<u>0.59</u>	0.70	0.19	0.18	0.34	0.50	0.31
<i>S</i>	0.01	0.41	0.48	0.78	<u>0.75</u>	0.15	0.61	0.57	0.46	0.55
<i>B</i>	0.07	0.16	0.52	0.76	<u>0.70</u>	0.23	0.50	0.56	0.45	0.52
<i>D</i>	0.07	0.44	0.67	0.85	0.71	0.05	0.27	0.45	0.33	0.27

ρ_1	Glass					Wine				
	RN	CB	Gk	FG	FGk	RN	CB	Gk	FG	FGk
<i>A</i>	0.04	0.32	<u>0.74</u>	0.83	<u>0.78</u>	0.05	0.40	0.56	0.71	0.62
<i>C</i>	0.03	0.28	<u>0.60</u>	<u>0.60</u>	0.70	0.03	0.46	0.28	0.37	0.39
<i>S</i>	0.11	0.57	<u>0.87</u>	<u>0.88</u>	0.90	0.03	0.55	0.72	0.81	0.72
<i>B</i>	0.11	0.42	<u>0.87</u>	<u>0.88</u>	0.89	0.03	0.35	0.28	0.44	0.38
<i>D</i>	0.04	0.57	<u>0.77</u>	<u>0.77</u>	0.82	0.04	0.33	0.37	0.50	0.39

ponent (the first partition of the ensemble is selected randomly), this method selects the next partition incrementally according to the target diversity category. These diversity categories generally produce a set of similar ensembles, thereby obtaining smooth differences among them. However, this behavior has a side effect because the diversity categories frequently differ greatly from each other, so CB does not sample uniformly from the diversity range.

The diversity control method (with all three alternatives) obtained the highest ρ_1 values for all of the data sets, with statistically significant differences compared with the other methods. This is because the ability to control diversity resulted in small changes among the ensembles, and thus *all* of the ensembles differed from each other in a smooth manner, whereas CB created cohesive but separate ensemble categories. Overall, all of the grouping algorithms obtained similar performance, but FG*k* produced the best results followed closely by FG. The higher performance of these algorithms compared with G*k* was due to the cohesion between the generated groups of partitions. Indeed, G*k* only groups by *k*, which might leave very different partitions within the same group.

Table 4 presents the changes in the outputs of the consensus function, where five clustering measures are shown: accuracy (*A*), compactness (*C*), separation (*S*), the Davies-Bouldin index (*B*), and the Dunn index (*D*). The trends in these results were similar to those obtained for the ensemble measures. The ensembles generated by RN allowed the consensus function to derive unpredictable partitions. Using this method, the smoothness scores for clustering measures were always close to 0, which indicates that the ensembles changed in a highly irregular manner according to the measures based on the output of the consensus function. Although CB was far superior to RN, the diversity control method obtained the best performance. In addition to producing ensembles that changed in a smooth manner from the viewpoint of diversity (Table 3), the diversity control method could also induce smooth changes in the consensus partitions.

For the diversity control method, a joint analysis of the smoothness for

the ensemble and clustering measures showed that given a slight change in the input, the consensus function could produce an acceptable slight change in the output. This was the case for the Four Gaussian, Iris, and Glass data sets, but especially for Difficult Doughnut, where the proposed method appeared to even control the quality of the consensus partition.

These results suggest that current methods can generate ensembles with differences in diversity, but their internal structures might actually differ in an unpredictable manner. Thus, these methods are less useful and reliable in cases where it is necessary to analyze how diversity affects the consensus performance. By contrast, our proposed method can control a particular measure of ensemble diversity (as shown in Section 4.2), but it also guarantees that all aspects of the ensemble change in a smooth manner. This was demonstrated by the high smoothness scores obtained from the perspectives of the diversity and consensus function. Therefore, it can be stated that the proposed method is more suitable for generating ensembles that differ according to the fine-grained control of their diversity.

5. Conclusions

In this study, we identified and addressed a problem in existing methods for generating diverse ensembles. Thus, when these methods are used to create sets of clustering solutions with several levels of disagreement, ensembles with similar diversity may actually have very different properties, which leads to unpredictable behavior by the consensus function. As a consequence, studying ensemble diversity is a difficult problem. By contrast, our proposed method can increase and decrease the dissimilarity among ensemble members

in a smooth manner, thereby providing a novel approach for ensemble diversity analysis. Our method analyzes the structure of the ensemble and makes small changes to it, which allows the fine-grained control of diversity. The performance of our method was demonstrated experimentally by evaluating different aspects of the overall method. The proposed method could produce ensembles that changed not only in a smooth manner from the viewpoint of diversity, but also according to different quality measures based on the output of the consensus function.

The proposed diversity control method is an important step in the development of a more robust approach for effectively exploring how ensemble diversity affects the final quality of clusters. In future research, we will perform more in-depth investigations using different data sets, base clustering algorithms, consensus functions, and validation measures.

6. Acknowledgements

This study was supported by the National Scientific and Technical Research Council [PIP 2013 117], National University of Litoral [CAI+D 2011 548], and National Agency for Science and Technology Promotion (ANPCyT) [PICT 2014 2627].

References

- [1] D.J. Newman A. Asuncion. UCI machine learning repository, 2007.
- [2] Javad Azimi and Xiaoli Fern. Adaptive cluster ensemble selection. In *Proceedings of the 21st International Joint Conference on Artificial Intelli-*

- gence, IJCAI'09, pages 992–997, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.
- [3] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9781118619063.
- [4] David L. Davies and Donald W. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1(2):224–227, April 1979. ISSN 0162-8828. doi: 10.1109/TPAMI.1979.4766909.
- [5] Carlotta Domeniconi and Muna Al-Razgan. Weighted cluster ensembles: Methods and analysis. *ACM Trans. Knowl. Discov. Data*, 2(4):17:1–17:40, 2009. ISSN 1556-4681. doi: 10.1145/1460797.1460800.
- [6] J. C. Dunn. Well separated clusters and optimal fuzzy-partitions. *Journal of Cybernetics*, 4(1):95–104, 1974.
- [7] Xiaoli Z. Fern and Wei Lin. Cluster ensemble selection. *Stat. Anal. Data Min.*, 1(3):128–141, November 2008. ISSN 1932-1864. doi: 10.1002/sam.v1:3.
- [8] Xiaoli Zhang Fern and Carla E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *ICML-2003*, pages 186–193, 2003.
- [9] Xiaoli Zhang Fern and Carla E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04*, pages

- 36–, New York, NY, USA, 2004. ACM. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015414.
- [10] Alessandro Fiori, Andrea Mignone, and Giuseppe Rospo. Decoclu: Density consensus clustering approach for public transport data. *Information Sciences*, 328:378–388, 2016. ISSN 0020-0255. doi: 10.1016/j.ins.2015.08.054.
- [11] B. Fischer and J.M. Buhmann. Bagging for path-based clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(11): 1411–1415, November 2003. ISSN 0162-8828. doi: 10.1109/TPAMI.2003.1240115.
- [12] A. L N Fred and A.K. Jain. Data clustering using evidence accumulation. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 276–2804, 2002. doi: 10.1109/ICPR.2002.1047450.
- [13] AL.N. Fred and AK. Jain. Combining multiple clusterings using evidence accumulation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(6):835–850, June 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.113.
- [14] Francesco Gullo, Andrea Tagarelli, and Sergio Greco. *Diversity-based Weighting Schemes for Clustering Ensembles*, chapter 37, pages 437–448. 2009. doi: 10.1137/1.9781611972795.38.
- [15] Stefan T. Hadjitodorov, Ludmila I. Kuncheva, and Ludmila P. Todorova. Moderate diversity for better cluster ensembles. *Information Fusion*, 7(3):264–275, 2006. ISSN 1566-2535. doi: 10.1016/j.inffus.2005.01.008.

- [16] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001. ISSN 0925-9902. doi: 10.1023/A:1012801612483.
- [17] Jie Hu, Tianrui Li, Hongjun Wang, and Hamido Fujita. Hierarchical cluster ensemble model based on knowledge granulation. *Knowledge-Based Systems*, 91:179–188, 2016. ISSN 0950-7051. doi: 10.1016/j.knosys.2015.10.006.
- [18] Dong Huang, Jianhuang Lai, and Chang-Dong Wang. Ensemble clustering using factor graph. *Pattern Recognition*, 50:131–142, 2016. ISSN 0031-3203. doi: 10.1016/j.patcog.2015.08.015.
- [19] Xiaodi Huang, Xiaodong Zheng, Wei Yuan, Fei Wang, and Shanfeng Zhu. Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization. *Information Sciences*, 181(11):2293–2302, 2011. ISSN 0020-0255. doi: 10.1016/j.ins.2011.01.029.
- [20] N. Iam-On, T. Boongoen, S. Garrett, and C. Price. A link-based approach to the cluster ensemble problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(12):2396–2409, 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.84.
- [21] N. Iam-On, T. Boongeon, S. Garrett, and C. Price. A link-based cluster ensemble approach for categorical data clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 24(3):413–425, March 2012. ISSN 1041-4347. doi: 10.1109/TKDE.2010.268.

- [22] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.*, 31(8):651–666, June 2010. ISSN 0167-8655. doi: 10.1016/j.patrec.2009.09.011.
- [23] Liping Jing, Kuang Tian, and Joshua Z. Huang. Stratified feature sampling method for ensemble clustering of high dimensional data. *Pattern Recognition*, 48(11):3688–3702, 2015. ISSN 0031-3203. doi: 10.1016/j.patcog.2015.05.006.
- [24] Jon Kleinberg. An impossibility theorem for clustering. In *Neural Information Processing Systems*, pages 446–453. MIT Press, 2002.
- [25] Gang Kou, Yi Peng, and Guoxun Wang. Evaluation of clustering algorithms for financial risk analysis using {MCDM} methods. *Information Sciences*, 275:1–12, 2014. ISSN 0020-0255. doi: 10.1016/j.ins.2014.02.137.
- [26] L. I. Kuncheva and D. P. Vetrov. Evaluation of stability of k-Means cluster ensembles with respect to random initialization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(11):1798–1808, November 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.226.
- [27] L.I. Kuncheva and S.T. Hadjitodorov. Using diversity in cluster ensembles. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 2, pages 1214–12192, October 2004. doi: 10.1109/ICSMC.2004.1399790.
- [28] M.H.C. Law, A.P. Topchy, and A.K. Jain. Multiobjective data clustering. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004*.

- Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II-424-II-430 Vol.2, June 2004. doi: 10.1109/CVPR.2004.1315194.
- [29] J. Lee and Kuo-Hsun Hsu. Gea: A goal-driven approach to discovering early aspects. *Software Engineering, IEEE Transactions on*, 40(6):584–602, June 2014. ISSN 0098-5589. doi: 10.1109/TSE.2014.2322368.
- [30] Liu Limin and Fan Xiaoping. A new selective clustering ensemble algorithm. In *e-Business Engineering (ICEBE), 2012 IEEE Ninth International Conference on*, pages 45–49, 2012. doi: 10.1109/ICEBE.2012.17.
- [31] Selim Mimaroglu and Ertunc Erdil. Combining multiple clusterings using similarity graph. *Pattern Recognition*, 44(3):694–703, 2011. ISSN 0031-3203. doi: 10.1016/j.patcog.2010.09.008.
- [32] Behrouz Minaei-Bidgoli, Alexander P. Topchy, and William F. Punch. A comparison of resampling methods for clustering ensembles. In *IC-AI*, pages 939–945, 2004.
- [33] M.C. Naldi, A.C.P.L.F. Carvalho, and R.J.G.B. Campello. Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery*, 27(2):259–289, 2013. ISSN 1384-5810. doi: 10.1007/s10618-012-0290-x.
- [34] G. Stegmayer, D. H. Milone, L. Kamenetzky, M. G. López, and F. Carrari. A biologically-inspired validity measure for comparison of clustering methods over metabolic datasets. *IEEE ACM Transactions in Computational Biology and Bioinformatics*, 9(3):706–716, May 2012.

- [35] Georgina Stegmayer, Milton Pividori, and Diego H. Milone. A very simple and fast way to access and validate algorithms in reproducible research. *Briefings in Bioinformatics*, 2015. doi: 10.1093/bib/bbv054. URL <http://bib.oxfordjournals.org/content/early/2015/07/27/bib.bbv054.abstract>.
- [36] Alexander Strehl, Joydeep Ghosh, and Claire Cardie. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.
- [37] A. Topchy, A.K. Jain, and W. Punch. Combining multiple weak clusterings. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 331–338, November 2003. doi: 10.1109/ICDM.2003.1250937.
- [38] A. Topchy, B. Minaei-Bidgoli, A.K. Jain, and W.F. Punch. Adaptive clustering ensembles. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 272–2751, August 2004. doi: 10.1109/ICPR.2004.1334105.
- [39] A. Topchy, A.K. Jain, and W. Punch. Clustering ensembles: models of consensus and weak partitions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12):1866–1881, 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.237.
- [40] R. Vidal. Subspace clustering. *Signal Processing Magazine, IEEE*, 28(2): 52–68, March 2011. ISSN 1053-5888. doi: 10.1109/MSP.2010.939739.

- [41] Fei Wang, Changshui Zhang, and Naijiang Lu. Boosting GMm and its two applications. In *Multiple Classifier Systems*, volume 3541 of *Lecture Notes in Computer Science*, pages 12–21. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-26306-7. doi: 10.1007/11494683_2.
- [42] Fei Wang, Xin Wang, and Tao Li. Generalized cluster aggregation. In *IJCAI*, pages 1279–1284, 2009.
- [43] Junjie Wu, Hongfu Liu, Hui Xiong, Jie Cao, and Jian Chen. K-means-based consensus clustering: A unified view. *Knowledge and Data Engineering, IEEE Transactions on*, 27(1):155–169, January 2015. ISSN 1041-4347. doi: 10.1109/TKDE.2014.2316512.
- [44] Rui Xu and D.C. Wunsch. Clustering algorithms in biomedical research: A review. *Biomedical Engineering, IEEE Reviews in*, 3:120–154, 2010. ISSN 1937-3333. doi: 10.1109/RBME.2010.2083647.
- [45] Rui Xu and Don Wunsch. *Clustering*. Wiley-IEEE Press, 2009. ISBN 9780470276808.
- [46] Yu-Meng Xu, Chang-Dong Wang, and Jian-Huang Lai. Weighted multi-view clustering with feature selection. *Pattern Recognition*, 53:25–35, 2016. ISSN 0031-3203. doi: 10.1016/j.patcog.2015.12.007.
- [47] Fan Yang, Xuan Li, Qianmu Li, and Tao Li. Exploring the diversity in cluster ensemble generation: Random sampling and random projection. *Expert Systems with Applications*, 41(10):4844–4866, 2014. ISSN 0957-4174. doi: 10.1016/j.eswa.2014.01.028.

- [48] Y. Yang and J. Jiang. Hybrid sampling-based clustering ensemble with global and local constitutions. *IEEE Transactions on Neural Networks and Learning Systems*, 2015. ISSN 2162-237X. doi: 10.1109/TNNLS.2015.2430821. in press.
- [49] Yun Yang and Ke Chen. Temporal data clustering via weighted clustering ensemble with different representations. *Knowledge and Data Engineering, IEEE Transactions on*, 23(2):307–320, February 2011. ISSN 1041-4347. doi: 10.1109/TKDE.2010.112.
- [50] Jinfeng Yi, Tianbao Yang, Rong Jin, A.K. Jain, and M. Mahdavi. Robust ensemble clustering by matrix completion. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1176–1181, 2012. doi: 10.1109/ICDM.2012.123.
- [51] Z. Yu, P. Luo, J. You, H. S. Wong, H. Leung, S. Wu, J. Zhang, and G. Han. Incremental semi-supervised clustering ensemble for high dimensional data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):701–714, 2016. ISSN 1041-4347. doi: 10.1109/TKDE.2015.2499200.
- [52] Zhiwen Yu, Hau-San Wong, and Hongqiang Wang. Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics*, 23(21):2888–2896, 2007. doi: 10.1093/bioinformatics/btm463.
- [53] Zhiwen Yu, Hau-San Wong, Jane You, Guoxian Yu, and Guoqiang Han. Hybrid cluster ensemble framework based on the random combination

- of data transformation operators. *Pattern Recognition*, 45(5):1826–1837, 2012. ISSN 0031-3203.
- [54] Zhiwen Yu, Jane You, Hau-San Wong, and Guoqiang Han. From cluster ensemble to structure ensemble. *Information Sciences*, 198:81–99, 2012. ISSN 0020-0255. doi: 10.1016/j.ins.2012.02.019.
- [55] Zhiwen Yu, Le Li, Yunjun Gao, Jane You, Jiming Liu, Hau-San Wong, and Guoqiang Han. Hybrid clustering solution selection strategy. *Pattern Recognition*, 47(10):3362–3375, 2014. ISSN 0031-3203. doi: 10.1016/j.patcog.2014.04.005.
- [56] Zhiwen Yu, Le Li, Hau-San Wong, Jane You, Guoqiang Han, Yunjun Gao, and Guoxian Yu. Probabilistic cluster structure ensemble. *Information Sciences*, 267:16–34, 2014. ISSN 0020-0255. doi: 10.1016/j.ins.2014.01.030.
- [57] Caiming Zhong, Xiaodong Yue, Zehua Zhang, and Jingsheng Lei. A clustering ensemble: Two-level-refined co-association matrix with path-based transformation. *Pattern Recognition*, 48(8):2699–2709, 2015. ISSN 0031-3203. doi: 10.1016/j.patcog.2015.02.014.

Diversity control for improving the analysis of consensus clustering

Supplementary material

Milton Pividori

Georgina Stegmayer

Diego Milone

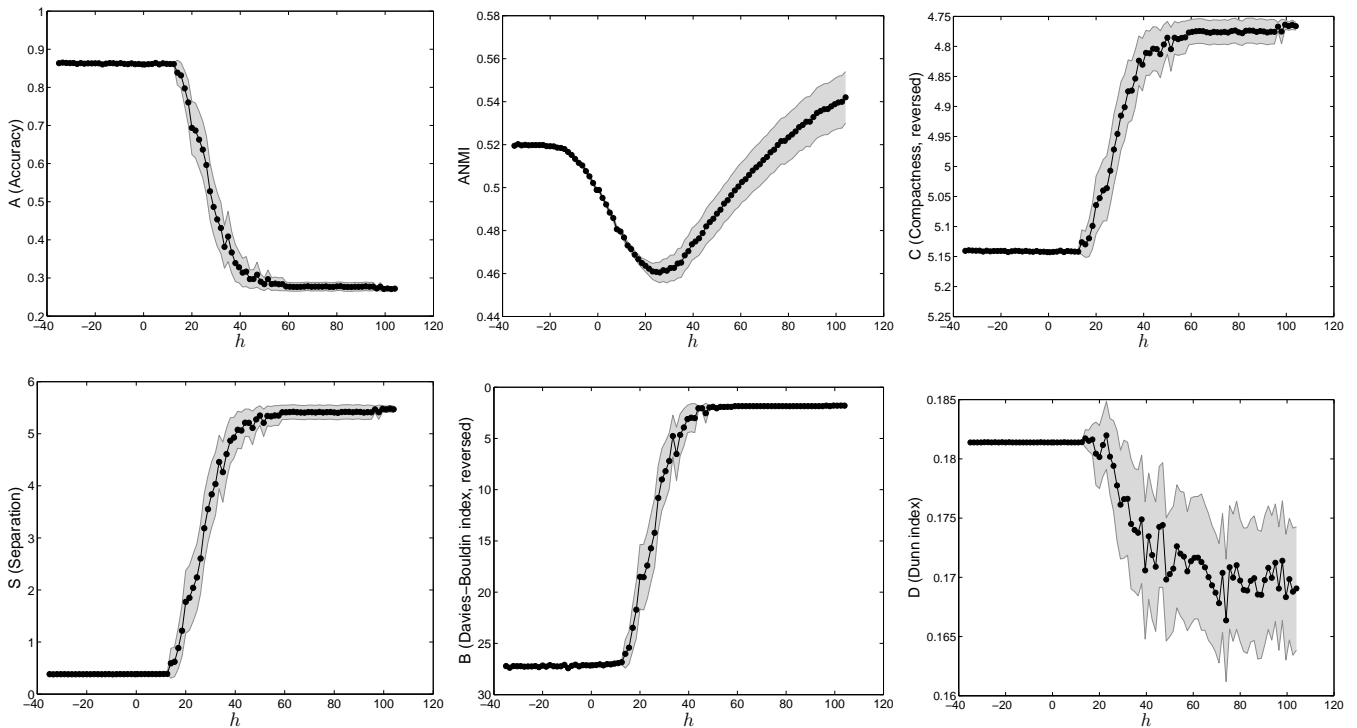


Figure 1: Six performance measures calculated for the consensus partitions obtained from the Difficult Doughnut data set. The compactness and Davies-Bouldin index are presented in reverse, so the values at the top are better for all of the measures.

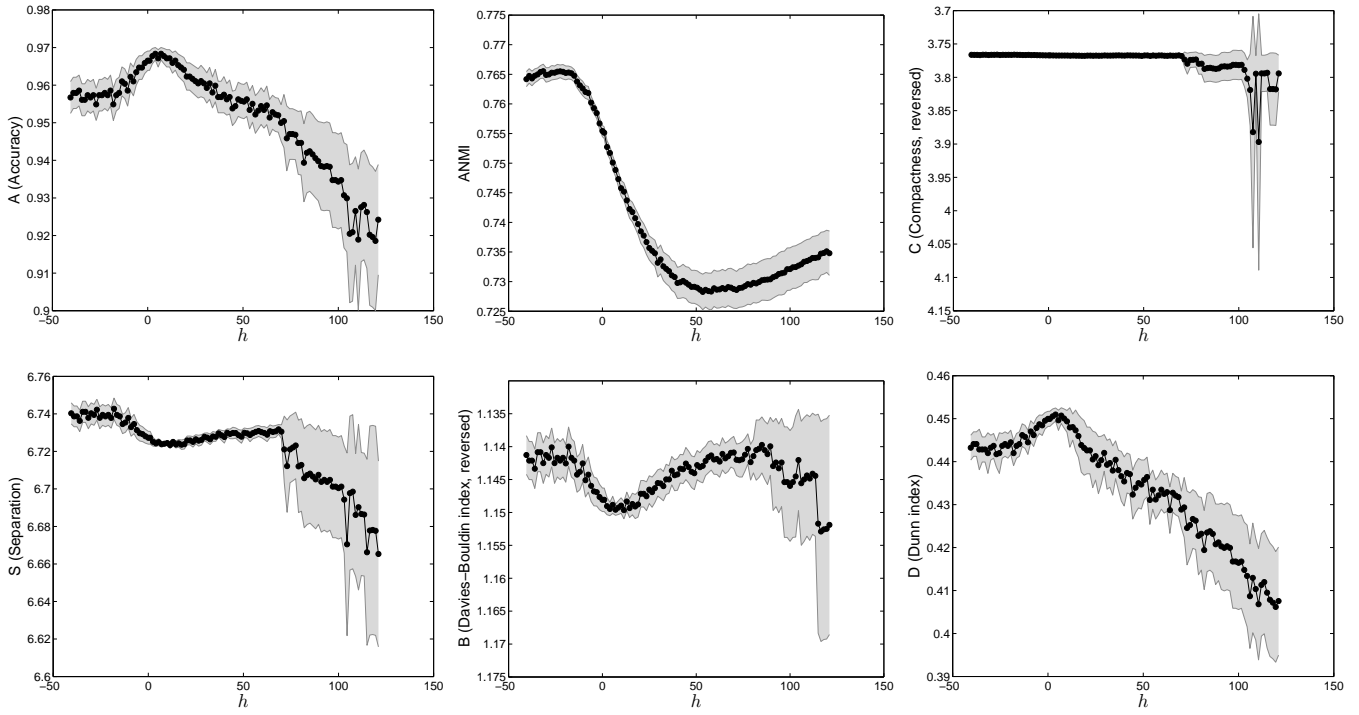


Figure 2: Six performance measures calculated for the consensus partitions obtained from the Four Gaussian data set. The compactness and Davies-Bouldin index are presented in reverse, so the values at the top are better for all of the measures.

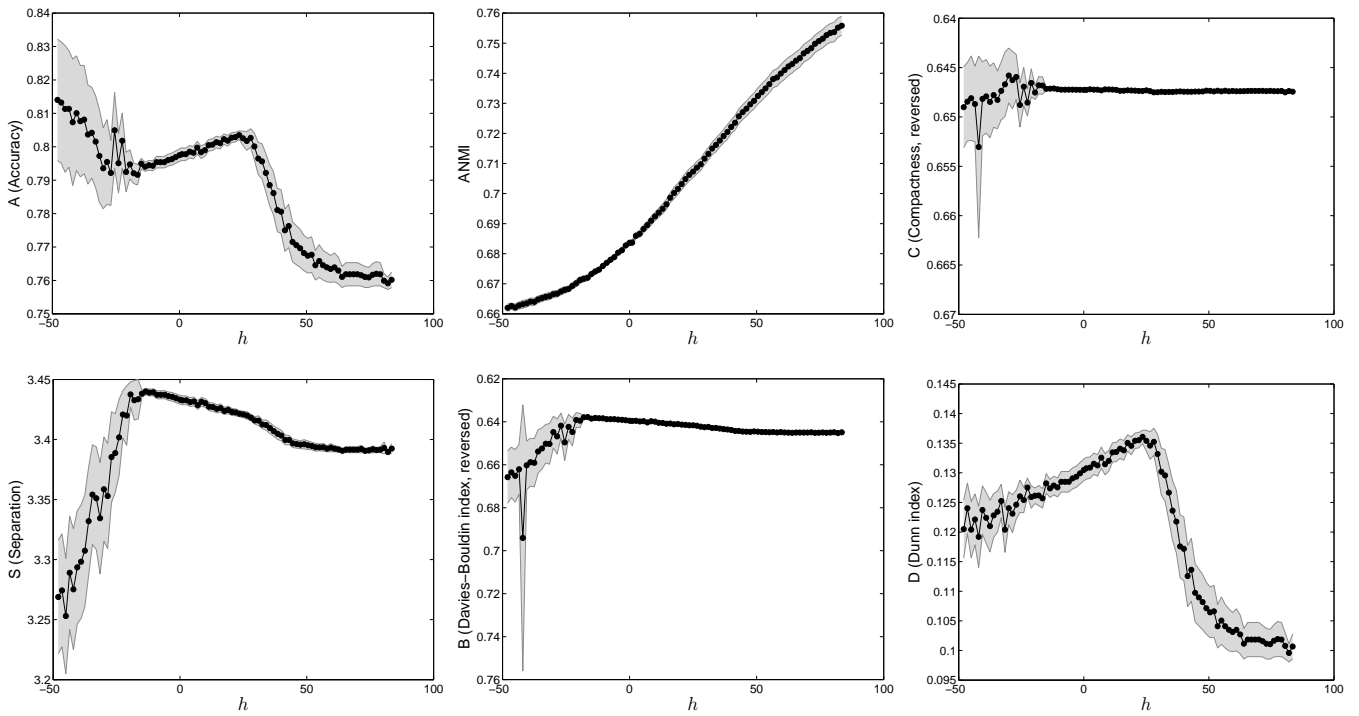


Figure 3: Six performance measures calculated for the consensus partitions obtained from the Iris data set. The compactness and Davies-Bouldin index are presented in reverse, so the values at the top are better for all of the measures.

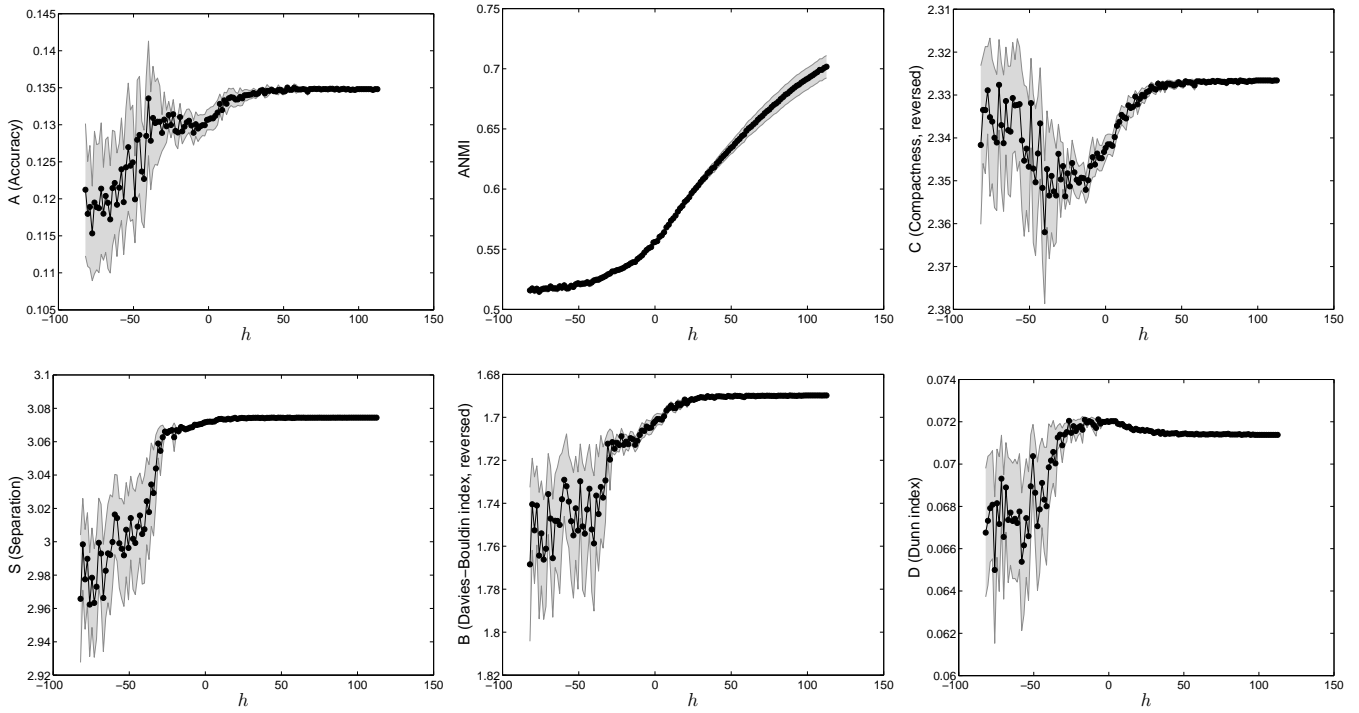


Figure 4: Six performance measures calculated for the consensus partitions obtained from the Ionosphere data set. The compactness and Davies-Bouldin index are presented in reverse, so the values at the top are better for all of the measures.

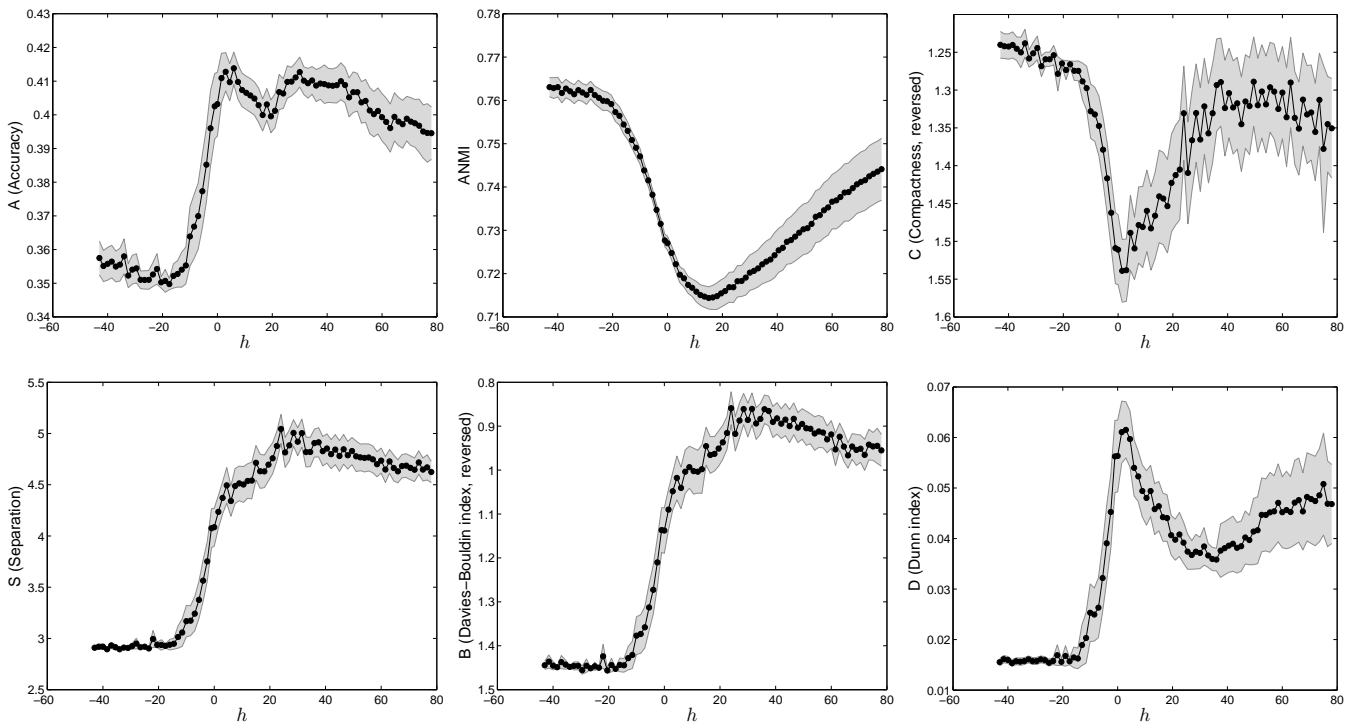


Figure 5: Six performance measures calculated for the consensus partitions obtained from the Glass data set. The compactness and Davies-Bouldin index are presented in reverse, so the values at the top are better for all of the measures.