

# Investigating the influence of the choice of the ensemble members in accuracy and diversity of selection-based and fusion-based methods for ensembles

Anne M.P. Canuto <sup>\*</sup>, Marjory C.C. Abreu, Lucas de Melo Oliveira, João C. Xavier Jr.,  
Araken de M. Santos

*Informatics and Applied Mathematics Department, Federal University of Rio Grande do Norte (UFRN), Natal, RN 59072-970, Brazil*

Received 7 December 2005; received in revised form 25 August 2006

Available online 2 November 2006

Communicated by M. Kamel

## Abstract

One of the most important steps in the design of a multi-classifier system (MCS), also known as ensemble, is the choice of the components (classifiers). This step is very important to the overall performance of a MCS since the combination of a set of identical classifiers will not outperform the individual members. The ideal situation would be a set of classifiers with uncorrelated errors – they would be combined in such a way as to minimize the effect of these failures. This paper presents an extensive evaluation of how the choice of the components (classifiers) can affect the performance of several combination methods (selection-based and fusion-based methods). An analysis of the diversity of the MCSs when varying their components is also performed. As a result of this analysis, it is aimed to help designers in the choice of the individual classifiers and combination methods of an ensemble.

© 2006 Elsevier B.V. All rights reserved.

**Keywords:** Diversity measures; Classifier ensembles; Selection-based combination methods; Fusion-based combination methods

## 1. Introduction

It is well known that substantial improvements can be obtained in difficult pattern recognition problems by combining or integrating the outputs of multiple classifiers. Classifier combinations (Multi-classifier systems or ensembles) exploit the idea that different classifiers, also referred as to experts or recognition modules – can offer complementary information about patterns to be classified, improving the effectiveness of the overall recognition process

(see, for example, Giacinto and Roli, 2001; Kuncheva, 2004; Sharkey, 1999).

In the literature, the use of multi-classifier system (MCS) has been widely used for several pattern recognition tasks. In the last decade, for instance, a large number of papers (Canuto et al., 2001; Czyz et al., 2004; Lemieux and Parizeau, 2003; Sharkey, 1999; Zhou and Zhang, 2002) have proposed the combination of multiple classifiers for designing high performance classification systems, in areas which include alphanumeric character recognition (Canuto et al., 2001), face recognition (Czyz et al., 2004; Lemieux and Parizeau, 2003; Zhou and Zhang, 2002), among others.

Diversity has been recognized as a very important feature in classifier combination and has been addressed by several authors, as in (Kuncheva and Whitaker, 2003; Shipp and Kuncheva, 2002; Tsymbal et al., 2005; Windeatt et al., 2005). The main reason for this is that if there are

<sup>\*</sup> Corresponding author. Tel.: +55 84 3215 3815; fax: +55 84 3215 3213.

E-mail addresses: [anne@dimap.ufrn.br](mailto:anne@dimap.ufrn.br) (A.M.P. Canuto), [marjory.abreu@gmail.com](mailto:marjory.abreu@gmail.com) (M.C.C. Abreu), [lmoliveira@gmail.com](mailto:lmoliveira@gmail.com) (L. de Melo Oliveira), [jcxavier01@yahoo.com](mailto:jcxavier01@yahoo.com) (J.C. Xavier Jr.), [arakenmedeiros@yahoo.com](mailto:arakenmedeiros@yahoo.com) (A.d.M. Santos).

many different classifiers, it is sensible to expect an increase in the overall performance when combining them. Then, it is intuitively accepted that classifiers to be combined should be diverse, since there is clearly no advantage to be gained from an ensemble that is composed of a set of identical classifiers.

The main aim of this paper is to investigate the importance of the choice of the components of an ensemble (ensemble members) in the diversity and accuracy of different structures (hybrid and non-hybrid) of ensembles. Also, several combination methods will be investigated in order to define which combination methods are more affected by the choice of the ensemble members.

This paper is divided into five sections and organized as follows. Some research works related to the subject of this paper are presented in Section 2. Multi-classifier systems are described in Section 3, focusing on the combination methods as well as some diversity measures that can be used in these systems. Section 4 shows the experimental work using hybrid and non-hybrid multi-classifier systems, applied to four different databases. It shows accuracy and diversity of these systems, illustrating the effect of varying the ensemble members in the combination methods. Finally, in Section 5, it is presented the final remarks of this work.

## 2. Related works

There is a wide variety of researches analysing diversity in ensembles, such as in (Banfield et al., 2005; Duin and Tax, 2000; Kuncheva and Whitaker, 2003; Ruta and Gabrys, 2005; Shipp and Kuncheva, 2002; Tsymbal et al., 2005; Windeatt et al., 2005), among others. In (Kuncheva, 2004; Kuncheva and Whitaker, 2003; Shipp and Kuncheva, 2002), for instance, the authors have compared the correlation between several diversity measures and linear and non-linear fusion-based methods. In contrast, in (Duin and Tax, 2000), for instance, a comparison of several structures of ensemble using twelve different classification methods and 10 linear and non-linear fusion-based combination was performed, but without using any analysis of diversity measure. In the mentioned paper, analysis using different types of classification components and different features for each classification were performed. In addition to that, in (Windeatt et al., 2005), diversity measures have been used in order to analyze the correlation of training patterns, selecting the structure of an ensemble based on information about the training set. Diversity measures have also been used in conjunction with training strategies, such as boosting and bagging (Kuncheva et al., 2002; Shipp and Kuncheva, 2002). Also, diversity measures have been used to select classifiers to compose an ensemble, such as in (Banfield et al., 2005; Giacinto and Roli, 2001; Ruta and Gabrys, 2005) or to perform feature selection (Tsymbal et al., 2005). Finally, in previous works of the authors, such as in (Canuto et al., 2005; Canuto et al., 2005), they evaluate the performance of ensembles using different

ensemble members, but only fusion-based methods and only one ensemble size (five members) were used.

Several works have been published about analysis and comparison of strategies for constructing and combining classifier ensembles, which deal with the issues addressed in this paper. However, most of the works related to diversity and ensemble performance use linear and non-linear non-trainable fusion-based combiners, in which usually both are non-trainable methods. Also, most of them do not analyze the choice of the ensemble members as an important aspect to affect diversity and accuracy of the ensembles. Unlike most of the aforementioned researches, a comparison is performed in this paper, analysing different fusion-based and selection-based methods. In addition to that, the fusion-based methods are linear, non-linear and computational intelligent ones.

At the end of this analysis, it is aimed to have a general picture of which combination methods are more sensitive to the choice of the ensemble members. According to Kuncheva (2004), it is believed that the type of classifier diversity that can be more useful to improve ensemble performance depends also on the particular combination rule used. Based on this, in this paper, an analysis of how the choice of the ensemble members can affect accuracy and diversity of ensembles is performed and how this can affect the accuracy of different combination rules. In this sense, it is aimed to show, that when using one combination method, a more careful choice of the ensemble members has to be done, while for other methods, this choice is not a strong problem.

## 3. Multi-classifier systems

The need to have a computational system that works with pattern recognition on an efficient way has motivated the interest in the study of multi-classifiers systems (MCS) (Czyz et al., 2004; Kuncheva, 2004; Sharkey, 1999), also known as ensembles. The main idea of using ensembles is that the combination of classifiers can lead to an improvement in the performance of a pattern recognition system in terms of better generalisation and/or in terms of increased efficiency and clearer design. In the design of ensembles, three basic steps can be defined, which are: the choice of the organisation of its components, the choice of ensemble members and the choice of the combination methods that will be used.

- A multi-classifier system can be defined according to the organisation of its components as modular and ensemble. In the modular approach, each classifier becomes responsible for a part of the whole system and they are usually linked in a serial way. In the ensemble approach, all classifiers are able to answer to the same task in a parallel or redundant way, in which a combination module is responsible for providing the overall system output (Kuncheva, 2004). In this paper, the ensemble approach will be taken into account.

- In the choice of the ensemble members, the members (classifiers) of an ensemble are chosen and implemented. The correct choice of the set of classifiers is fundamental to the overall performance of an ensemble. The main aim of combining classifiers is to improve their generalisation ability and recognition performance.
- Once a set of classifiers has been created and the strategy for organizing them has been defined, the next step is to choose an effective way of combining their outputs in ensembles. According to their functioning, two strategies of combination methods are discussed in the literature on classifier combination, which are: selection-based and fusion-based methods.

In the MCS context, there is also some work reported in applying learning schemes in an ensemble. Usually, these schemes work on the training sets which are to be presented to the classifiers. The main aim is either to improve the generalisation of the ensemble or to minimize the correlated error of the classifiers within the system (increase the diversity of the classifiers). The main learning methods used in ensembles are Bagging (Breiman, 1996, 1999) and Boosting (Freund, 1995, 2002).

### 3.1. Combination methods

#### 3.1.1. Fusion-based methods

Several fusion-based methods have been proposed in the literature and these can be classified according to their characteristics as Linear, Non-linear, Statistical and Computational Intelligent combiners.

- *Linear combination methods*: Currently, the simplest ways to combine multiple neural networks are the sum and average of the neural networks' outputs (Kuncheva, 2004). Such methods are known as linear combining techniques.
- *Non-linear methods*: This class includes rank-based combiners, such as Borda Count (Sharkey, 1999), as well as majority voting strategies (Kuncheva, 2004).
- *Statistical-based methods*: In this class, statistical methods, such as the Dempster–Shafer technique (Mitchell, 1997) as well as Bayesian combination methods (Mitchell, 1997) are used to combine the output of the classifiers.
- *Computational intelligent methods*: Within the group of methods based on combination via computational intelligence techniques, it can be found fuzzy integral (Canuto et al., 2001), neural networks (Canuto et al., 2001) and genetic algorithms (Kuncheva, 2004).

In this paper, seven different fusion-based methods will be used, which are: majority vote, naïve Bayesian, Sum, Average, Median, MLP and Fuzzy MLP.

#### 3.1.2. Selection-based methods

In selection-based methods, only one classifier is needed to correctly classify the input pattern. The choice of a clas-

sifier to label the input pattern is made during the operation phase. This choice is typically based on the certainty of the current decision. Preference is given to more certain classifiers.

There are also the hybrid methods, in which selection and fusion techniques are used in order to provide the most suitable output to classify the input pattern. Although these methods are considered as hybrid methods, they use the selection procedure as first option and they will be considered as selection-based methods in this paper. Of the methods explained below, the first one is a selection-based method and the last two methods are hybrid ones.

*3.1.2.1. Dynamic classifier selection based on local accuracy class (Dcs-LA).* Woods et al. (1997) use local analysis of competence to nominate a classifier to label an input. According to Woods, the main steps to calculate the local class accuracy (LCA) for a test pattern ( $\mathbf{x}$ ) can be defined as follows:

1. Take the class labels provided by all classifiers.
2. For each classifier ( $D_i$ ,  $i = 1, \dots, L$ ) find  $k$  ( $k = 10$  is recommended) points closest to  $\mathbf{x}$  for which  $D_i$  has provided the same label.
3. Calculate the proportion of points in which  $D_i$  has provided the true label and make it be the local class accuracy of this classifier.
4. Choose the classifier with the highest LCA. Three main situations may occur, which are:
  - 4.1. If there is only one winner, let it label  $\mathbf{x}$ .
  - 4.2. If two classifiers are tied, choose a third classifier with the second highest LCA.
  - 4.3. If all classifiers are tied, pick a random class label among the tied labels (Woods et al., 1997).

It is important to emphasize that the selection of the most suitable classifier is made during the test phase and only in case there is a disagreement among the classifiers.

*3.1.2.2. Dynamic classifier selection based on multiple classifier behavior (Dcs-MCS) (Giacinto and Roli, 2001).* There are two main differences of this method and the previous one. Firstly,  $k$  is variable and, secondly, a classifier is selected if and only if the highest LCA is substantially higher than the LCA values of the other classifiers. Otherwise, the test pattern is classified by majority vote technique. The main steps of the DCS-MCS are the following:

1. For each test pattern, select the  $k$  nearest neighbor.
2. Select only the neighbor which has similarity higher than a threshold.
3. Calculate the competence of each classifier to the selected neighbor.
4. If the best classifier is substantially higher than the other, select it.
5. Otherwise, use the majority vote fusion method.

3.1.2.3. *Dynamic classifier selection using also decision templates (Dcs-DT)* (Kuncheva, 2002). As the previous one, this method can switch between combination and fusion techniques (hybrid method). The main steps of this method are described as follows:

- After the training process of all classifiers, the training patterns are grouped into clusters, using a  $k$ -Means clustering procedure (Kuncheva, 2002).
- The classification accuracy of all classifiers is estimated and the classifier with the highest classification is defined for each cluster.
- When a test pattern is clamped into the system, the nearest cluster center is found. If the best classifier of the nearest cluster is significantly better than the others, then let the best classifier label the test pattern. Otherwise, a fusion technique is used with all classifiers.

The main difference of this method and the previous one is that a statistical test is performed in order to define whether the best classifier (highest LCA) is significantly different from the others. In (Kuncheva, 2002), a paired  $t$ -test was used to define the significance of the winner. If the winner is not significantly different from the others, a fusion scheme based on a decision template matrix is performed (Kuncheva, 2002).

### 3.2. Diversity in ensembles

As already mentioned, there is no gain in a multi-classifier system (MCS) that is composed of a set of identical classifiers. The ideal situation, in terms of combining classifiers, would be a set of classifiers that reach an appropriate trade-off between the accuracy of each member of the ensemble and uncorrelation between their errors. Diversity can be reached in three different ways:

- Variations of the parameters of the classifiers (e.g., varying the initial parameters, such as the weights and topology, of a neural network model (Windeatt et al., 2005)).
- Variations of the training dataset of the classifiers (e.g., the use of learning strategies such as Bagging and Boosting (Kuncheva, 2004)).
- Variations of the type of classifier (e.g., the use of different types of classifiers, such neural networks and decision trees, as members of an ensemble – hybrid ensembles).

In this paper, variations of the diversity are captured using different types of classifiers and different parameters. There are different diversity measures available from different fields of research. In (Kuncheva et al., 2002; Kuncheva and Whitaker, 2003), for instance, 10 diversity measures are defined, which can be classified as pairwise (classifiers are considered on a pairwise basis and then average the results) and non-pairwise (the whole group of classifiers is considered). It is important to emphasize that no clear rela-

tionships have been found so far between the different diversity measures proposed in the literature, and the accuracy of classifier ensembles (Kuncheva, 2004; Kuncheva and Whitaker, 2003). In this paper, two measures are used, in which one of them is pairwise and the other one is non-pairwise.

#### 3.2.1. The double fault measure (Giacinto and Roli, 2001)

This measure uses the proportion of the cases that have been misclassified by both classifiers and it is defined as follows:

$$DF_{i,k} = \frac{N^{00}}{N^{11} + N^{10} + N^{01} + N^{00}} \quad (1)$$

where  $N^{00}$  is the number of patterns in which both are wrong;  $N^{11}$  is the number of patterns in which both are correct;  $N^{01}$  is the number of patterns in which the first is wrong and the second is right;  $N^{10}$  is the number of patterns in which the first is right and the second is wrong.

This is a dissimilarity measure in the numerical taxonomy literature (Giacinto and Roli, 2001).

#### 3.2.2. The entropy measure ( $E$ ) (Kuncheva, 2004; Kuncheva and Whitaker, 2003)

It is a non-pairwise measure and it is based on the assumption that the highest diversity among classifiers is manifested by  $[L/2]$  of the votes in  $y_i$  with the same value (0 or 1) and the other  $L - [L/2]$  with the alternative value. If they all were 0's or 1's, there is no disagreement and they are not diverse. One possible diversity measure based on this concept can be defined as (Kuncheva, 2004; Kuncheva and Whitaker, 2003):

$$E = \frac{1}{N} \sum_{m=1}^N \frac{1}{(L - [L/2])} \min\{l(z_m), L - l(z_m)\} \quad (2)$$

$l(z_m)$  is the number of classifiers that correctly recognizes  $z_m$ .  $E$  varies between 0 and 1, where 0 indicates no difference and 1 indicates the highest possible diversity.

## 4. Experimental work

In this experimental work, an empirical comparison of ensembles using several structures and sizes is performed. The base classifiers to be used in this investigation are: MLP (Multi-layer Perceptron), Fuzzy MLP, K-NN (nearest neighbor), RBF (radial basis function), SVM (Support Vector Machine), J48 decision tree and JRIP (Optimized IREP). The choice of the aforementioned classifiers was due to the different learning bias that they use during their functioning. JRIP, for instance, is a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP (Canuto et al., 2005). On the other hand, Fuzzy Multi-layer Perceptron (MLP) is a variation of MLP that clamps desired membership values (calculated incorporating fuzzy

concepts) to the output nodes during the training phase instead of choosing binary values as in a winner-take-all method (Canuto et al., 2001). The errors may be then back-propagated with respect to the desired membership values as the fuzzy desired output. In (Canuto et al., 2001), a method to calculate the fuzzy desired output was described which is suitable for binary images.

Experiments were conducted using four different ensemble sizes, using 3, 5, 7 and 9 base classifiers. The choice of a small number of classifiers is due to the fact that diversity has a strong effect in the performance of ensembles when using less than 10 classifiers (Kuncheva, 2004). In this sense, it is believed that the accuracy of the combination methods used in ensembles with few members is more sensitive to variations in the ensemble members than using larger numbers of members. For each ensemble size, hybrid and non-hybrid structures of ensembles were created. For the non-hybrid structures, the same type of the classification methods was used, varying the initial and training parameters of each method. For instance, in a MLP classifier, different number of hidden neurons, learning rate, momentum and interactions are used to create different versions of the same classifiers.

In order to make the choice of the structures (hybrid) of the ensembles more systematic, ensembles with 1 (NH), 3 (HYB 3), 5 (HYB 5) and 7 (HYB 7) different types of classifiers were taken into consideration (for ensemble sizes 3, 5, 7 and 9). As there are several possibilities for each hybrid structure, this paper presents the average of the accuracy delivered by all possibilities of the corresponding hybrid structure. For instance, when using ensembles with 3 types of classifiers (HYB 3) for ensembles with three base classifiers, there are nine possible structures. In this way, ensemble with 3 types of classifiers (HYB 3) represents the average of these nine possibilities. For all four ensemble sizes (3, 5, 7 and 9), the same procedure is performed. In this way, it becomes easier to make a better analysis of the results.

All the classification methods used in this study, apart from Fuzzy MLP, were obtained from the WEKA machine learning visual package (<http://www.cs.waikato.ac.nz/~ml/weka/>). The selection-based methods as well as most of the combination methods were implemented in Java language.

#### 4.1. Databases

Four different databases are used in this investigation, which are described as follows.

- Database A: It is a breast cancer database, which was developed by the University of Wisconsin and available at Blake et al. (1998). Instances were extracted from images of a fine needle aspirate (FNA) of a breast mass and they describe features of the cell nucleus. A total number of 32 attributes were used, in which 30 of them are real-valued input features. Also, a total number of 682 instances were used, which are equally distributed into malignant and benign examples.

- Database B: an image database, where instances were drawn randomly from a database of seven outdoor images (segmentation dataset from the UCI repository (Blake et al., 1998)). The images were hand-segmented to create a classification for every pixel. Each instance is a  $3 \times 3$  region. Nineteen continuous attributes were extracted from the region and there are seven different classes of regions, which are: brickface, sky, foliage, cement, window, path and grass.
- Database C: It is a protein database which represents a hierarchical classification, manually detailed, and represents known structures of proteins. The main protein classes are all- $\alpha$ , all- $\beta$ ,  $\alpha/\beta$ ,  $\alpha + \beta$  and small. It is an unbalanced database, which has a total of 582 patterns, in which 111 patterns belong to class all- $\alpha$ , 177 patterns to class all- $\beta$ , 203 patterns to  $\alpha/\beta$ , 46 patterns to class  $\alpha + \beta$  and 45 patterns to class small.
- Database D: It is a primate splice-junction gene sequences (DNA) with associated imperfect domain theory, obtained from Blake et al. (1998). A total of 3190 Instances using 60 attributes were used. These attributes describe sequences of DNA used in the process of creation of proteins.

#### 4.2. Cross-validation

In order to evaluate the robustness of the classifier, a common methodology is to perform cross-validation on the classifier. 10-fold cross-validation has been proved to be statistically good enough in evaluating the performance of the classifier (Mitchell, 1997). In 10-fold cross-validation, the training set is equally divided into 10 different subsets. Nine out of 10 of the subsets are used to train the classifier and the tenth subset is used as the test set. The procedure is repeated 10 times, with a different subset being used as the test set.

#### 4.3. T-test

The comparison of two supervised learning methods is, often, accomplished by analyzing the statistical significance of the difference between the mean of the classification correct rate, on independent test sets, from the methods evaluated. In order to evaluate the mean of the correct rate, several (distinct) data sets are needed. However, the number of data sets available is often limited. One way to overcome this problem is to divide the data sets into training and test sets by the use of a  $k$ -fold cross-validation procedure (Mitchell, 1997), which has been used in this investigation.

Applying the distinct algorithms to the same folds with  $k$  at least equal to 10, the statistical significance of the differences between the methods can be measured, based on the mean of the correct rate from the test sets. The  $p$ -value provided by the  $t$ -test means the degree of confidence of the result. For instance, when using a confidence level of

95%, one sample is statistically different of the other only if the  $p$ -value is lower than 0.05.

#### 4.4. Individual classifiers

Before starting the investigation of the ensembles, it is important to analyze the performance (accuracy) of the individual classifiers. As already mentioned, variations of the same classifiers were obtained using different setting parameters. A special attention was made for the classifiers to have a similar level of accuracy, since a classifier with a poor accuracy can have a negative influence in the accuracy of the ensemble. As nine variations of each classifier were used in this investigation, for simplicity reasons, only the average accuracies of classifiers are shown in Table 1.

According to the average accuracy provided by the classifiers, it can be seen that all classifiers have delivered a sim-

ilar pattern of accuracy for all four databases. The Fuzzy MLP classifier has provided the highest accuracy for databases A, B and D, while SVM has delivered the highest accuracy for database C.

#### 4.5. Ensembles with three base classifiers

Table 2 shows accuracy and standard deviation of ensembles with three base classifiers applied to all four databases. In this table, 10 different combination methods were analyzed, which are: Voting, Naïve Bayes (NB), Sum, Average, Median, MLP (multi-layer perceptron), Fuzzy MLP and the remaining three are selection-based (Dcs, Dcs-MCS and Dcs-DT). As it can be seen, three of the chosen combination methods (Naïve Bayes, MLP and Fuzzy MLP) are trainable methods. In order to define the training set of these combination methods, approximately 10% of the instances of a database were taken out to create a validation set. For instance, for database D, 290 instances were taken from the dataset to compose the validation set. In this sense, the training set (to be used in the 10-fold cross-validation) will be composed of 2900 instances and the validation set will be composed of 290 instances. These combination methods were then trained using the validation set.

For each combination method, their performance using one type of classifier (non-hybrid structure – NH) and three types of classifiers (hybrid structure – HYB 3) will be investigated. As already mentioned, values presented in Table 2 represents the average accuracy and standard deviation of all possibilities. The bold numbers in this table represents the highest accuracy for each combination method. The

Table 1  
Classifier accuracy (CA) and standard deviation (SD) of the Individual Classifiers

	Database A		Database B		Database C		Database D	
	CA	SD	CA	SD	CA	SD	CA	SD
knn	73.95	3.58	76.39	2.76	69.45	6.19	75.85	2.81
svm	80.05	4.59	84.75	2.66	<b>83.90</b>	5.55	81.89	3.84
mlp	84.61	3.11	84.88	2.47	78.71	3.45	84.69	4.11
fmlp	<b>85.43</b>	3.61	<b>86.05</b>	2.68	83.67	3.46	<b>85.13</b>	2.68
rbf	82.69	3.61	82.39	3.25	81.86	3.63	81.73	2.65
tree	78.58	3.93	81.69	3.51	78.17	4.52	79.99	3.64
jrip	81.11	3.55	82.75	3.99	74.51	5.21	81.88	3.67

Table 2  
Accuracy and standard deviation of hybrid and non-hybrid ensembles with three base classifiers

	Ensembles with three base classifiers					
	Database A (breast cancer)			Database B (image)		
	NH	HYB 3	Dif	NH	HYB 3	Dif
Voting	72.02 ± 7.01	<b>75.58</b> ± 2.92	3.56	80.97 ± 6.06	<b>82.56</b> ± 6.51	1.59
N Bayes	<b>78.76</b> ± 4.71	78.58 ± 2.35	<u>0.18</u>	<b>85.56</b> ± 6.58	85.11 ± 7.14	0.45
Sum	77.54 ± 6.03	<b>78.21</b> ± 2.32	0.67	82.51 ± 5.55	<b>83.02</b> ± 5.26	0.51
Average	77.78 ± 4.69	<b>78.67</b> ± 3.3	0.89	83.96 ± 6.4	<b>84.36</b> ± 7.32	0.4
Median	73.6 ± 3.71	<b>76.94</b> ± 4.23	3.34	81.57 ± 6.62	<b>82.72</b> ± 7.06	1.15
MLP	91.18 ± 3.82	<b>92.37</b> ± 2.17	1.19	<b>95.25</b> ± 2.97	94.49 ± 2.45	0.76
FuzzyMLP	91.95 ± 3.23	<b>94.75</b> ± 2.12	2.8	95.95 ± 2.4	<b>96.35</b> ± 2.32	<u>0.4</u>
DCS-LA	90.8 ± 3.3	<b>92.21</b> ± 1.7	1.41	93.07 ± 2.95	<b>94.79</b> ± 2.52	<u>1.72</u>
DCS-DT	90.79 ± 3.28	<b>94.75</b> ± 2.29	<u>3.96</u>	94.8 ± 3.04	<b>95.63</b> ± 2.71	0.83
DCS-MCB	92.41 ± 2.77	<b>94.45</b> ± 1.71	2.04	95.6 ± 2.04	<b>95.72</b> ± 2.11	0.12
	Database C (proteins)			Database D (splice)		
	NH	HYB 3	Dif	NH	HYB 3	Dif
	NH	HYB 3	Dif	NH	HYB 3	Dif
Voting	69.21 ± 11.03	<b>71.71</b> ± 4.41	<u>2.5</u>	76.15 ± 4.68	<b>78.3</b> ± 4.56	2.15
N Bayes	73.07 ± 9.56	<b>76.51</b> ± 3.59	3.44	81.14 ± 4.46	<b>82.93</b> ± 3.85	<u>1.79</u>
Sum	70.71 ± 10.25	<b>76.02</b> ± 3.7	5.31	78.65 ± 3.48	<b>80.67</b> ± 3.75	2.02
Average	72.31 ± 10.28	<b>76</b> ± 3.83	3.69	79.39 ± 4.4	<b>81.85</b> ± 3.85	2.46
Median	70.05 ± 11.46	<b>74.6</b> ± 4.82	4.55	77.07 ± 5.36	<b>79.59</b> ± 4.75	2.52
MLP	84.33 ± 7.31	<b>90.25</b> ± 2.7	5.92	89.92 ± 2.12	<b>92.73</b> ± 2.74	<u>2.81</u>
FuzzyMLP	89.69 ± 6.14	<b>93.16</b> ± 3.48	3.47	93.02 ± 3.01	<b>95.34</b> ± 2.25	2.32
DCS-LA	81.04 ± 6.64	<b>89.43</b> ± 2.39	<u>8.39</u>	90.35 ± 2.77	<b>92.77</b> ± 1.98	2.42
DCS-DT	86.01 ± 6.91	<b>91.9</b> ± 2.79	5.89	91.69 ± 2.51	<b>94.15</b> ± 2.69	2.46
DCS-MCB	88.39 ± 6.26	<b>93.55</b> ± 2.74	5.16	93.88 ± 2.8	<b>95.87</b> ± 2.78	1.99

fourth and seventh columns of Table 2 show the difference in performance (*Dif*) delivered by the combination method when varying the ensemble members. It is calculated by the difference between the highest and the lowest accuracies. This value aims to define the variation in performance when using different ensemble members. In this sense, the methods which have high values of *Dif* have a strong variation when using different ensemble members. As a consequence, it can be stated that they are strongly affected by the choice of the ensemble members.

The accuracies of the ensembles were, on average, higher than the corresponding individual classifiers. However, some fusion-based methods had delivered a lower accuracy than the best individual classifiers. In analyzing the accuracy of the combination methods, ensembles with Fuzzy MLP had delivered the highest accuracy for databases A, B and C, while Dcs-MCB had provided the highest accuracy for database D. Of the fusion-based methods, the highest average accuracy was reached by ensembles combined by Fuzzy MLP, while Dcs-MCB had delivered the highest average accuracy of the selection-based methods. It is important to emphasize that the highest accuracies (bold numbers in Table 2) delivered by all combination methods were reached when using a hybrid structure (HYB 3), in most of the cases.

#### 4.5.1. Difference in performance

When analyzing the difference in performance delivered by the combination methods, in a general perspective, the lowest variations (*Dif*) of the combination methods were provided by database B, while the highest variations were

provided by database C. This is an expected result since database C is an unbalanced database, while database B is an equally distributed (balanced) database. Another important fact to be observed is that, apart from database D, the highest difference in performance was always reached by a selection-based method (Dcs-LA for databases B and C and Dcs-DT for database A). In contrast, the lowest difference was always reached by a fusion-based method (Naïve Bayes for databases A and D, Fuzzy MLP for database B and Voting for database C). Finally, the lowest average difference in performance (all four databases) delivered by the combination methods is reached by Naïve Bayes (1.46), followed by Average (1.86), Sum (2.13), Fuzzy MLP (2.25), Dcs-MCB (2.33), Voting (2.45), MLP (2.67), Median (2.89), Dcs-LA (3.29) and Dcs-DT (3.49). It is important to analyze that the combination methods that provided the top two highest average differences were selection-based methods.

In order to evaluate whether the difference in performance delivered by the combination methods is significant, the hypothesis tests (*t*-test) comparing the highest and lowest accuracy ensembles, using a confidence level of 95%, is performed. It is important to emphasize that the use of a confidence level of 95% means that two differences will be statistically significant if the *p*-value of this comparison is less than 0.05. If the difference in performance of a combination method is statistically significant, this means that the variation reached when changing the ensemble members is strong enough to be detected by a statistical test. In this sense, it can be stated that this combination method is sensitive to changes in the ensemble members.

Table 3

*p*-Values of the statistical test performed between the highest and lowest accuracies delivered by the combination methods for databases A–D

	Ensembles with three base classifiers				Ensembles with five base classifiers			
	A	B	C	D	A	B	C	D
Voting	2.14E–06	<b>0.0536</b>	0.0183	0.0014	2.88E–11	<b>0.0736</b>	9.78E–06	0.0096
N Bayes	<b>0.366</b>	<b>0.3341</b>	0.0004	0.0026	2.75E–05	<b>0.1435</b>	0.0010	0.0182
Sum	<b>0.145</b>	<b>0.2705</b>	9.72E–07	0.0002	2.43E–09	5.65E–06	1.22E–06	0.0018
Average	<b>0.0674</b>	<b>0.3540</b>	0.0004	6.57E–05	3.17E–11	<b>0.1757</b>	0.0007	0.0449
Median	1.02E–07	<b>0.1393</b>	0.0001	0.0006	7.40E–08	<b>0.2149</b>	0.0001	<b>0.0793</b>
MLP	0.00416	0.0334	5.97E–13	6.04E–12	0.0068	0.0008	8.94E–16	0.0001
FuzzyMLP	3.20E–11	<b>0.1305</b>	1.59E–06	9.98E–09	5.03E–08	<b>0.1771</b>	3.78E–09	0.0001
Dcs-LA	0.0001	2.41E–05	4.45E–25	8.87E–11	1.64E–09	3.52E–05	3.62E–15	7.06E–06
Dcs-DT	6.96E–18	0.0296	1.09E–13	3.06E–09	2.59E–10	0.0023	1.46E–12	0.0051
Dcs-MCB	2.41E–09	<b>0.3561</b>	1.28E–12	2.70E–06	0.0001	<b>0.2153</b>	8.91E–11	0.0681
	Ensembles with seven base classifiers				Ensembles with nine base classifiers			
	A	B	C	D	A	B	C	D
Voting	4.01E–27	<b>0.1498</b>	4.57E–07	1.28E–05	3.12E–28	3.15E–06	1.92E–42	9.95E–11
N Bayes	3.51E–12	0.0311	3.58E–08	2.81E–06	1.61E–21	1.47E–05	1.21E–31	3.27E–20
Sum	9.77E–11	<b>0.0510</b>	1.35E–06	7.95E–07	4.51E–27	3.65E–13	1.71E–50	5.02E–13
Average	1.73E–05	<b>0.2091</b>	1.02E–05	0.0037	1.15E–10	0.0006	6.81E–16	6.79E–08
Median	3.54E–07	<b>0.2055</b>	5.83E–06	0.0062	9.13E–11	0.0021	3.78E–20	4.62E–06
MLP	2.09E–05	0.0157	7.61E–10	<b>0.0831</b>	3.32E–19	0.0002	1.28E–33	1.72E–28
FuzzyMLP	0.0119	0.0133	0.0020	0.0009	3.90E–10	0.0247	4.87E–24	4.35E–08
Dcs-LA	2.70E–06	0.0044	2.39E–05	2.23E–05	5.79E–23	3.79E–05	1.51E–22	2.87E–31
Dcs-DT	2.64E–06	0.0163	4.44E–05	0.0001	1.52E–18	0.0035	3.52E–14	7.15E–10
Dcs-MCB	0.0215	0.0073	0.0006	0.0009	1.73E–07	0.0003	4.08E–11	0.0044

Table 3 shows the  $p$ -values for all differences in performance, including all four databases. In this table, the bold numbers represent differences in performance which are not statistically significant. It was observed from the top left part of Table 3 that ensembles with MLP, Dcs-LA and Dcs-DT combination methods have differences in performance which are statistically significant in all four databases. Ensembles that have differences in performances which are statistically significant in three databases (apart from database B) are Voting, Median, Fuzzy MLP and Dcs-MCB. Finally, ensembles that have differences in performances which are statistically significant for two databases are NB, Sum and Average. Based on this result, it could be concluded that ensembles with MLP, Dcs-LA and Dcs-DT are the top three methods affected by variations in the ensemble members (in all four databases). In contrast, ensembles with NB, Sum and average are the least three methods affected by variations in the ensemble members, since the difference in performance is statistically significant in only two databases.

#### 4.5.2. Diversity

In order to analyze the level of diversity of the ensembles, Table 4 shows two diversity measures applied for all ensemble structures and sizes. The diversity measures were described in Section 3.2. As already mentioned, double fault is a pairwise measure while entropy is a non-pairwise one. The first and second lines of Table 4 show the diversity measures provided by ensembles with three base classifiers. As it can be observed from Table 4, in most of the cases, variations in the ensembles member were reflected in the diversity of the ensembles in an expected way.

In analyzing the behavior of the diversity measures when varying from non-hybrid (NH) to the hybrid (HYB 3) structures of the ensembles, it could be seen that the

double fault measure has decreased, while entropy has increased in three databases (apart from database B). For both diversity measures, this means that hybrid ensembles had provided higher level of diversity than the non-hybrid ones. This is also reflected in the accuracy of the ensembles since the highest accuracies were reached by the hybrid structures (HYB 3), in almost all cases. In this sense, based on the experiments of this paper, there might be a relation between accuracy and diversity emerged by the choice of the ensemble members. However, it is important to emphasize that this is a data-dependent result, in which this may not be true when using other databases.

#### 4.6. Ensembles with five base classifiers

Table 5 shows the accuracy and standard deviation of ensembles with five base classifiers applied to all four databases. In this table, 10 different combination methods were analyzed, which are: Voting, Naïve Bayes (NB), Sum, Average, Median, MLP (multi-layer Perceptron), Fuzzy MLP and the remaining three are selection-based (Dcs, Dcs-MCS and Dcs-DT). For each combination method, their performance using one type of classifier (NH), three types of classifiers (HYB 3) and five types of classifiers (HYB 5) will be investigated. As in the previous section, the fourth and seventh columns of Table 5 show the difference in performance ( $Dif$ ) delivered by the combination method when varying the ensemble members.

The accuracies of the ensembles were, on average, higher than the corresponding ensembles with three base classifiers. One interesting fact to be noticed is that all three selection-based methods did not improve the accuracy in comparison with ensembles with three base classifiers for database C, which is an unbalanced database. In contrast, this fact was not observed for the fusion-based methods. In

Table 4  
Two diversity measures applied to all ensemble structures and sizes for all four databases

	A		B		C		D	
	Double fault	Entropy	Double fault	Entropy	Double fault	Entropy	Double fault	Entropy
<i>Ensembles with three base classifiers</i>								
NH	0.307	0.800	0.279	0.837	0.314	0.824	0.250	0.826
HYB 3	0.282	0.823	0.267	0.826	0.290	0.826	0.216	0.854
<i>Ensembles with five base classifiers</i>								
NH	0.290	0.824	0.227	0.859	0.287	0.836	0.224	0.789
HYB 3	0.257	0.815	0.220	0.867	0.314	0.839	0.345	0.791
HYB 5	0.253	0.825	0.219	0.870	0.233	0.865	0.215	0.820
<i>Ensembles with seven base classifiers</i>								
NH	0.280	0.816	0.249	0.854	0.267	0.847	0.240	0.751
HYB 3	0.239	0.849	0.257	0.859	0.242	0.860	0.239	0.782
HYB 5	0.236	0.871	0.239	0.824	0.210	0.873	0.224	0.796
HYB 7	0.200	0.920	0.260	0.830	0.190	0.900	0.280	0.810
<i>Ensembles with nine base classifiers</i>								
NH	0.339	0.816	0.261	0.814	0.234	0.817	0.281	0.816
HYB 3	0.333	0.832	0.254	0.788	0.244	0.812	0.246	0.828
HYB 5	0.279	0.849	0.291	0.849	0.257	0.839	0.239	0.857
HYB 7	0.277	0.846	0.280	0.843	0.276	0.853	0.277	0.859

Table 5

Accuracy and standard deviation of hybrid and non-hybrid ensembles with five base classifiers

	Ensembles with five base classifiers							
	Database A (breast cancer)				Database B (image)			
	NH	HYB 3	HYB 5	<i>Dif</i>	NH	HYB 3	HYB 5	<i>Dif</i>
Voting	73.85 ± 6.26	80.75 ± 6.11	<b>82.61</b> ± 5.72	<u>8.76</u>	<b>85.26</b> ± 3.48	84.43 ± 3.73	85 ± 2.65	0.83
N Bayes	78.44 ± 5.12	<b>81.58</b> ± 4.66	81.23 ± 5.57	3.14	<b>89.74</b> ± 4.68	88.97 ± 4.5	89.35 ± 4.61	0.77
Sum	77.56 ± 5.9	81.12 ± 3.78	<b>84.24</b> ± 3.99	6.68	86.24 ± 4.09	85.45 ± 3.07	<b>87.86</b> ± 2.05	<u>2.41</u>
Average	75.98 ± 4.9	81.63 ± 3.88	<b>83.03</b> ± 4.89	7.05	<b>88.15</b> ± 4.86	87.4 ± 5.25	88.01 ± 6.58	0.75
Median	76.02 ± 5.61	80.72 ± 5.45	<b>83.5</b> ± 8.31	7.48	<b>85.73</b> ± 5.3	84.84 ± 6.62	85.82 ± 6.57	0.89
MLP	93.47 ± 4.85	94.42 ± 2.41	<b>95.45</b> ± 1.42	<u>1.98</u>	95.49 ± 3	94.59 ± 2.21	<b>95.83</b> ± 1.64	1.24
FuzzyMLP	94.27 ± 2.85	96.43 ± 2.01	<b>97.27</b> ± 2.26	3	96.86 ± 2.08	96.58 ± 1.9	<b>96.92</b> ± 1.92	0.34
DCS-LA	91.2 ± 3.58	93.99 ± 2.22	<b>95.53</b> ± 3.02	4.33	94.51 ± 2.78	95.51 ± 2.3	<b>96.45</b> ± 1.33	1.94
DCS-DT	91.83 ± 2.99	94.33 ± 2.91	<b>95.47</b> ± 2.05	3.64	96.1 ± 1.94	<b>97.06</b> ± 2.27	96.58 ± 1.54	0.96
DCS-MCB	94.13 ± 3.34	95.32 ± 1.97	<b>96.37</b> ± 2.29	2.24	96.54 ± 1.87	96.67 ± 2.12	<b>96.82</b> ± 1.68	<u>0.28</u>
	Database C (proteins)				Database D (splice)			
	NH	HYB 3	HYB 5	<i>Dif</i>	NH	HYB 3	HYB 5	<i>Dif</i>
	NH	HYB 3	HYB 5	<i>Dif</i>	NH	HYB 3	HYB 5	<i>Dif</i>
Voting	72.84 ± 6.79	75.95 ± 4.18	<b>78.45</b> ± 5.43	5.61	78.86 ± 4.95	79.56 ± 4.62	<b>81.31</b> ± 5.62	<u>2.45</u>
N Bayes	77.88 ± 8.22	79.39 ± 3.97	<b>82.39</b> ± 4.77	<u>4.51</u>	82.16 ± 3.57	<b>83.42</b> ± 4.02	82.57 ± 4.92	1.26
Sum	74.58 ± 6.99	78.47 ± 4.02	<b>80.81</b> ± 4.87	6.23	80.58 ± 2.8	81.49 ± 3.4	<b>82.5</b> ± 3.97	1.92
Average	76.31 ± 8.47	78.93 ± 5.06	<b>81.15</b> ± 4.95	4.84	81.48 ± 3.47	<b>82.57</b> ± 4.49	82.47 ± 4.15	1.09
Median	73.65 ± 7.8	77.34 ± 5.47	<b>79.6</b> ± 6.81	5.95	79.86 ± 5.55	80.62 ± 5.71	<b>81.52</b> ± 6.49	1.66
MLP	84.47 ± 4.88	89.19 ± 2.56	<b>92.28</b> ± 2.75	7.81	92.99 ± 2.11	93.45 ± 2.35	<b>94.65</b> ± 2.44	1.66
FuzzyMLP	89.27 ± 4.48	93.61 ± 3.9	<b>94.38</b> ± 3.37	5.11	94.97 ± 2.23	95.33 ± 2.19	<b>96.64</b> ± 2.2	1.67
DCS-LA	80.41 ± 7.05	86.65 ± 3.1	<b>90.8</b> ± 2.32	<u>10.39</u>	93.16 ± 2.44	94.38 ± 1.98	<b>95.34</b> ± 2.37	2.18
DCS-DT	85.33 ± 5.68	89.44 ± 2.07	<b>92.88</b> ± 2.77	7.55	93.76 ± 3	94.93 ± 3.03	<b>95.29</b> ± 2.85	1.53
DCS-MCB	87.94 ± 5.91	92.05 ± 1.87	<b>94.89</b> ± 2.6	6.95	95.38 ± 2.17	95.57 ± 2.05	<b>96.06</b> ± 2.4	<u>0.68</u>

analyzing the accuracy of the combination methods, as in the previous section, the highest accuracy was always reached by ensembles combined by Fuzzy MLP (databases B and C) and Dcs-MCB (databases A and D). As it was expected, of the fusion-based methods, the highest average accuracy was reached by ensembles combined by Fuzzy MLP, while Dcs-MCB had delivered the highest average accuracy of the selection-based methods. As in the previous section, the highest accuracies (bold numbers in Table 5) delivered by all combination methods were always reached when using the totally hybrid structure (HYB 5), followed by the partial hybrid structure (HYB 3) and by the non-hybrid structure (NH), for all four databases.

#### 4.6.1. Difference in performance

When analyzing the difference in performance delivered by the combination methods (*Dif*), in a general perspective, there was an improvement in the difference in performance, when compared with ensembles using three base classifiers. The magnitude of the increase is higher for the non-trainable fusion-based methods, such as Voting, Sum, Average and Median. In comparing ensembles combined by Fuzzy MLP and by Dcs-MCB, the magnitude of the increase was basically the same, being the two least affected methods when increasing the number of classifiers in an ensemble from 3 to 5. Unlike the previous section, it was observed that, apart from database C, the highest difference in performance was always reached by a fusion-based method (Voting for databases A and D, Sum for database B and Naïve Bayes for database C). In contrast, the lowest difference was reached by a fusion-based method in two

databases (MLP for database A and Naïve for database C) and by a selection-based method in two databases (Dcs-MCB for databases B and D).

As in the previous section, the lowest average difference in performance (all four databases) delivered by the combination methods is reached by Naïve Bayes (2.42), followed by Fuzzy MLP (2.53), Dcs-MCB (2.58), MLP (3.17), Dcs-DT (3.42), Average (3.43), Median (4.00), Sum (4.31), Voting (4.41) and Dcs-LA (4.71). It is important to emphasize the improvement reached by the selection-based methods in the list of average difference, when compared with ensembles using three base classifiers. Dcs-DT, for instance, was the ninth method in the list of average difference for ensembles with three base classifiers and, now, it is in fourth place. In addition to this, ensemble with Dcs-DT is the only method in which the average difference decreased, in comparison with ensembles using three base classifiers.

In order to evaluate whether the difference in performance delivered by the combination methods is significant, the hypothesis tests (*t*-test) comparing the highest and lowest accuracy ensembles, using a confidence level of 95%, is performed. Table 3 shows the *p*-values for all differences in performance, including all four databases. It was observed from the top right part of Table 3 that ensembles with Sum, MLP, Dcs-LA and Dcs-DT combination methods have differences in performance which are statistically significant in all four databases. Ensembles that have differences in performances which are statistically significant in three databases (apart from database B) are Voting, NB, Average, Fuzzy MLP and Dcs-MCB. Finally, the only ensemble in which the difference in performance is statistically significant

cant for two databases is Median. Based on this result, it could be concluded that ensembles with Sum, MLP, Dcs-LA and Dcs-DT are the top four methods affected by variations in the ensemble members (in all four databases).

#### 4.6.2. Diversity

In order to analyze the level of diversity of the ensembles, Table 3 shows two diversity measures applied for all ensemble structures and sizes. The fourth, fifth and sixth lines of Table 3 show the diversity measures provided by ensembles with five base classifiers. An interesting fact to be noticed from Table 3 is that, unlike the previous section, the double fault measure did not reflect the expected decrease in the diversity of the ensembles. For databases C and D, the double fault measure did not decrease when varying from non-hybrid (NH) to hybrid (HYB3 and HYB 5). It is important to emphasize that both databases are unbalanced ones. In contrast, the entropy measure has increased when changing from non-hybrid to hybrid structures for three databases (apart from database A). This means that, according to the entropy measure, hybrid ensembles had provided higher level of diversity than the non-hybrid ones, which is also reflected in the accuracy of the ensembles. In this sense, there might be a relation between accuracy and diversity emerged by the choice of the ensemble members, at least for the entropy measure.

#### 4.7. Ensemble with seven base classifiers

Table 6 shows the accuracy and standard deviation of ensembles with seven base classifiers applied to all four data-

bases. For each combination method, their performance using one type of classifier (NH), three types of classifiers (HYB 3), five types of classifiers (HYB 5) and seven types of classifiers (HYB 7) will be investigated. As in the previous sections, the fourth and seventh columns of Table 6 show the difference in performance (*Dif*) delivered by the combination method when varying the ensemble members.

The accuracies of the ensembles were, on average, higher than the corresponding ensembles with five base classifiers. However, for database C, the opposite fact was observed, in which the accuracies delivered by ensembles with seven base classifiers are slightly lower than ensembles with five base classifiers.

In analyzing the accuracy of the combination methods, unlike the previous section, the highest accuracy was always reached by ensembles combined by Dcs-MCB, followed closely by ensembles with Fuzzy MLP. As it was expected, of the fusion-based methods, the highest average accuracy was reached by ensembles combined by Fuzzy MLP, while Dcs-MCB had delivered the highest average accuracy of the selection-based methods. As in the previous section, the highest accuracies (bold numbers in Table 6) delivered by all combination methods were always reached when using a totally hybrid structure (HYB 7), followed by the two partially hybrid structures (HYB 5 and HYB 3) and by the non-hybrid structure (NH), for all four databases.

##### 4.7.1. Difference in performance

When analyzing the difference in performance delivered by the combination methods, in a general perspective, there

Table 6  
Accuracy and standard deviation of hybrid and non-hybrid ensembles with seven base classifiers

	Ensembles with seven base classifiers									
	Database A (breast cancer)					Database B (image)				
	NH	HYB 3	HYB 5	HYB 7	<i>Dif</i>	NH	HYB 3	HYB 5	HYB 7	<i>Dif</i>
Voting	76.51 ± 6.41	<b>80.99</b> ± 3.04	74.87 ± 3.06	76.26 ± 2.69	6.12	84.86 ± 3.94	84.54 ± 3.06	84.71 ± 4.59	<b>85.56</b> ± 1.15	<u>1.02</u>
N Bayes	79.39 ± 4.32	84.28 ± 2.18	87.09 ± 2.73	<b>91.06</b> ± 3.94	6.78	87.89 ± 6.05	89.17 ± 3.63	90.06 ± 5.05	<b>91.56</b> ± 2.29	<u>3.67</u>
Sum	77.1 ± 6.15	<b>81.97</b> ± 3.14	78.25 ± 3.18	81.26 ± 3.69	4.87	86.13 ± 3.16	86.26 ± 3.25	85.73 ± 3.62	<b>87.67</b> ± 1.98	1.94
Average	78.17 ± 5.1	83.05 ± 4.64	83.05 ± 5.31	<b>85.69</b> ± 4.69	7.52	87.63 ± 5.58	87.51 ± 5.35	88.28 ± 5.34	<b>89</b> ± 7.33	1.49
Median	76.03 ± 7.2	<b>81.36</b> ± 6.2	77.86 ± 7	81.26 ± 5.98	5.33	85.47 ± 6.57	85.2 ± 5.96	84.73 ± 6.4	<b>86.58</b> ± 7.99	1.85
MLP	89.73 ± 5.82	95.03 ± 2.11	95.52 ± 1.35	<b>97.81</b> ± 1.84	<u>8.08</u>	96.49 ± 1.85	96.01 ± 2.64	96.32 ± 2.59	<b>97.89</b> ± 1.99	1.88
FuzzyMLP	93.75 ± 3.41	95.95 ± 1.36	95.78 ± 1.18	<b>96.33</b> ± 2.39	2.58	97.17 ± 1.8	97.15 ± 2.12	<b>97.42</b> ± 2.47	95.56 ± 2.09	1.86
DCS-LA	92.5 ± 3.1	95.15 ± 1.42	95.66 ± 1.24	<b>97.36</b> ± 1.13	4.86	95.8 ± 2.02	95.95 ± 1.79	96.56 ± 1.9	<b>97.56</b> ± 0.97	1.76
DCS-DT	92.33 ± 3.06	95.23 ± 2.04	96.02 ± 1.72	<b>97.23</b> ± 2.14	4.9	96.36 ± 1.69	96.73 ± 1.89	97.01 ± 2.0	<b>97.56</b> ± 1.07	1.2
DCS-MCB	94.51 ± 2.6	96.27 ± 2.11	94.89 ± 1.85	<b>96.28</b> ± 1.99	<u>1.77</u>	96.45 ± 1.78	97.09 ± 1.89	96.92 ± 2.49	<b>97.98</b> ± 2	1.53
	Database C (proteins)					Database D (splice)				
	NH	HYB 3	HYB 5	HYB 7	<i>Dif</i>	NH	HYB 3	HYB 5	HYB 7	<i>Dif</i>
	NH	HYB 3	HYB 5	HYB 7	<i>Dif</i>	NH	HYB 3	HYB 5	HYB 7	<i>Dif</i>
Voting	71.84 ± 6.05	73.26 ± 4.97	78.5 ± 5.27	<b>82.59</b> ± 5.24	10.75	77.42 ± 5.66	79.15 ± 4.96	80.87 ± 4.28	<b>85.56</b> ± 2.28	<u>8.14</u>
N Bayes	75.14 ± 5.72	78.65 ± 3.76	82.76 ± 6.19	<b>86.59</b> ± 5.48	<u>11.45</u>	84.99 ± 4.27	84.03 ± 4.62	85.59 ± 3.61	<b>91.11</b> ± 2.19	7.08
Sum	74.31 ± 5.87	77.03 ± 3.58	81.86 ± 6.09	<b>84.33</b> ± 5.71	10.02	82.86 ± 4.55	82.37 ± 3.92	84.08 ± 4.22	<b>88.78</b> ± 2.04	6.41
Average	74.88 ± 7	78.28 ± 5.03	82.08 ± 6.94	<b>85.33</b> ± 5.14	10.45	84.08 ± 4.55	83.6 ± 5.07	84.6 ± 5.39	<b>88.47</b> ± 7.94	4.87
Median	73.23 ± 6.89	74.55 ± 5.91	79.49 ± 6.16	<b>83.97</b> ± 5.89	10.74	80.98 ± 6.61	80.89 ± 6.49	81.82 ± 6.57	<b>86.48</b> ± 7.98	5.59
MLP	84.64 ± 5.98	86.55 ± 2.59	88.94 ± 2.66	<b>92.64</b> ± 4.3	8	92.64 ± 2.85	94.04 ± 2.53	94.98 ± 2.5	<b>95.19</b> ± 1.97	2.55
FuzzyMLP	90.25 ± 4.16	90.48 ± 2.33	91.34 ± 3.17	<b>94.36</b> ± 3.59	<u>4.11</u>	94.46 ± 2.95	95.3 ± 2.59	95.94 ± 1.97	<b>97.66</b> ± 2.68	3.2
DCS-LA	82.42 ± 6.69	84.14 ± 3.72	88.26 ± 2.64	<b>91.76</b> ± 3.67	9.34	92.72 ± 3.16	94.29 ± 2.29	96.05 ± 2.18	<b>97.09</b> ± 0.79	4.37
DCS-DT	85.52 ± 5.81	87.04 ± 2.89	91.26 ± 2.17	<b>93.19</b> ± 1.28	7.67	94.55 ± 2.54	95.38 ± 2.6	<b>96.31</b> ± 2.66	95.7 ± 1.18	1.76
DCS-MCB	88.39 ± 6.45	91.29 ± 2.06	93.22 ± 1.91	<b>95.37</b> ± 238	6.98	95.71 ± 2.84	95.34 ± 2.68	95.54 ± 2.65	<b>96.99</b> ± 1.88	<u>1.65</u>

was an improvement in the difference in performance, when compared with ensembles with five base classifiers. In comparing ensembles with Fuzzy MLP and with Dcs-MCB, the magnitude of the increase was basically the same, being the two least affected methods when increasing the number of classifiers in an ensemble from 5 to 7. As in the previous section, it was observed that the highest difference in performance was always reached by a fusion-based method. In contrast, the lowest difference was reached by a fusion-based method in two databases (Fuzzy MLP for database C and Voting for database B) and by a selection-based method in two databases (Dcs-MCB for databases A and D). The lowest average difference (all four databases) delivered by the combination methods is reached by Fuzzy MLP (2.94), followed by Dcs-MCB (2.98), Dcs-DT (3.88), Dcs-LA (5.08), MLP (5.13), Sum (5.81), Median (5.88), Average (6.08), Voting (6.51), and Naïve Bayes (7.25). It is important to emphasize the improvement of the ensembles combined by Dcs-LA, which was the last in the list of average difference for ensembles with five base classifiers and, now, it is in fourth place. Another interesting fact is that the selection-based methods are in the top four places in the list of average differences, which is the opposite of ensembles with three base classifiers.

In order to evaluate whether the difference in performance delivered by the combination methods is significant, the hypothesis tests ( $t$ -test) comparing the highest and lowest accuracy ensembles, using a confidence level of 95%, is performed. Table 3 shows the  $p$ -values for all differences in performance, including all four databases. It was observed from the bottom left part of Table 3 that ensembles with NB, Fuzzy MLP, Dcs-LA, Dcs-MCB and Dcs-DT combination methods have differences in performance which are statistically significant in all four databases. All other Ensembles have differences in performances which are statistically significant three databases (Voting, Sum, Average, Median and MLP). It is important to emphasize that although ensembles with Fuzzy MLP and Dcs-MCB have the top two lowest average differences, these differences are statistically significant because the standard deviation of these combination methods are low. On the other hand, ensembles combined by Average or Median, for instance, have high standard deviations for database B (around 5.0), making the statistical test concludes that this difference is not significant.

#### 4.7.2. Diversity

In order to analyze the level of diversity of the ensembles, Table 3 shows two diversity measures applied for all ensemble structures and sizes. Lines 8–11 of Table 3 show the diversity measures provided by ensembles with seven base classifiers. An interesting fact to be noticed from Table 3 is that, as in the previous section, the double fault measure did not reflect the expected decrease in the diversity of the ensembles. For databases B and D, the double fault measure did not decrease when varying from non-

hybrid to hybrid. In contrast, the entropy measure has increased when changing from non-hybrid to hybrid structures for three databases (apart from database B). This means that, according to the entropy measure, hybrid ensembles had provided higher level of diversity than the non-hybrid ones. This is also reflected in the accuracy of the ensembles. In this sense, based on the experiments of this paper, there might be a relation between accuracy and diversity emerged by the choice of the ensemble members, at least for the entropy measure.

#### 4.8. Ensembles with nine base classifiers

Table 7 shows the accuracy and standard deviation of ensembles with nine base classifiers applied to all four databases. For each combination method, their performance using one type of classifier (NH), three types of classifiers (HYB 3), five types of classifiers (HYB 5) and seven types of classifiers (HYB 7) will be investigated. As in the previous sections, the fourth and seventh columns of Table 7 show the difference in performance ( $Dif$ ) delivered by the combination method when varying the ensemble members.

The accuracies of the ensembles were, on average, higher than the corresponding ensembles with seven base classifiers. In analyzing the accuracy of the combination methods, as in the previous section, the highest accuracy was always reached by ensembles combined by a selection-based method. Also, the highest accuracies were reached when using a totally hybrid structure (HYB 7), followed by the two partially hybrid structures (HYB 5 and HYB 3) and by the non-hybrid structure (NH), for all four databases.

As it was expected, of the fusion-based methods, the highest average accuracy was reached by ensembles combined by Fuzzy MLP, while Dcs-MCB had delivered the highest average accuracy of the selection-based methods. It is important to emphasize the increase in performance reached by the Dcs-DT and Dcs-LA (Dcs-MCB was always between the best or second best combiner). For database C, for instance, the accuracies of Dcs-LA and Dcs-DT were even higher than Fuzzy MLP, which was not the case for ensembles with fewer base classifiers.

##### 4.8.1. Difference in performance

When analyzing the difference in performance delivered by the combination methods, in a general perspective, there was an improvement in the difference in performance, when compared with ensembles with seven base classifiers. As in the previous section, it was observed that the highest difference in performance was always reached by a fusion-based method. In contrast, the lowest difference was reached by a fusion-based method in only one database (Fuzzy MLP for database B) and by a selection-based method in three databases (Dcs-MCB for databases A and D and Dcs-DT for database C).

The lowest average difference in performance (all four databases) delivered by the combination methods is reached by Dcs-MCB (1.94), followed by Fuzzy MLP

Table 7

Accuracy and standard deviation of hybrid and non-hybrid ensembles with nine base classifiers

	Ensembles with nine base classifiers									
	Database A (breast cancer)					Database B (image)				
	NH	HYB 3	HYB 5	HYB 7	Dif	NH	HYB 3	HYB 5	HYB 7	Dif
Voting	75.32 ± 5.87	78.57 ± 2.98	84.31 ± 3.26	<b>85.69</b> ± 2.16	<u>10.37</u>	85.21 ± 4.6	82.86 ± 4.63	<b>86.07</b> ± 4.11	85.69 ± 2.16	3.21
N Bayes	81.21 ± 6.12	83.98 ± 2.51	88.78 ± 3.67	<b>90.51</b> ± 3.22	9.3	88.89 ± 6.66	87.9 ± 3.87	<b>90.7</b> ± 4.6	90.51 ± 3.22	2.8
Sum	79.23 ± 5.56	82.98 ± 3.35	87.2 ± 3.51	<b>88.84</b> ± 2.24	9.61	85.01 ± 3.38	85.36 ± 3.42	87.31 ± 4.07	<b>88.84</b> ± 2.24	<u>3.83</u>
Average	80.11 ± 7.03	83.23 ± 4.97	87.08 ± 6.42	<b>87.95</b> ± 6.5	7.84	88.41 ± 6.03	86.91 ± 5.21	<b>89.75</b> ± 5.99	87.95 ± 6.5	2.84
Median	78.25 ± 7.95	81.86 ± 6.39	85.36 ± 6.64	<b>86.88</b> ± 6.82	8.63	84.8 ± 4.7	84.13 ± 5.48	86.39 ± 4.65	<b>86.88</b> ± 6.82	2.75
MLP	91.68 ± 3.61	93.46 ± 2.11	95.48 ± 2.61	<b>96.72</b> ± 1.9	5.04	95.49 ± 1.81	95.43 ± 2.58	95.7 ± 2.4	<b>96.72</b> ± 1.9	1.29
FuzzyMLP	93.96 ± 2.88	95.86 ± 2.37	96.71 ± 2.12	<b>96.73</b> ± 2.0	2.77	96.84 ± 2.61	96.44 ± 2.51	<b>97.2</b> ± 2.4	96.73 ± 2	<u>0.76</u>
DCS-LA	93.09 ± 2.41	96.23 ± 1.69	96.63 ± 1.13	<b>96.89</b> ± 1.2	3.8	95.95 ± 2.17	95.55 ± 1.98	96.18 ± 1.83	<b>96.62</b> ± 1.2	1.07
DCS-DT	93.34 ± 2.77	95.81 ± 1.49	<b>97.04</b> ± 1.34	96.86 ± 1.35	3.7	96.4 ± 2.38	96.43 ± 2.15	<b>97.39</b> ± 1.87	97.18 ± 1.41	0.99
DCS-MCB	95.09 ± 3.1	95.52 ± 1.81	97.1 ± 2.15	<b>97.38</b> ± 1.78	<u>2.29</u>	96.54 ± 1.8	96.47 ± 2.62	<b>97.73</b> ± 1.87	97.38 ± 1.78	1.19
	Database C (proteins)					Database D (splice)				
	NH	HYB 3	HYB 5	HYB 7	Dif	NH	HYB 3	HYB 5	HYB 7	Dif
	NH	HYB 3	HYB 5	HYB 7	Dif	NH	HYB 3	HYB 5	HYB 7	Dif
Voting	74.8 ± 3.93	78.36 ± 3	82.54 ± 4.84	<b>85.43</b> ± 2.19	10.63	77.45 ± 3.88	79.45 ± 4.57	81.61 ± 4.72	<b>82.56</b> ± 4.87	5.11
N Bayes	81.5 ± 2.64	82.96 ± 2.17	86.92 ± 6.4	<b>90.09</b> ± 3.91	8.59	82.42 ± 3.47	83.6 ± 3.93	85.08 ± 3.97	<b>88.8</b> ± 3.57	<u>6.38</u>
Sum	75.92 ± 3.65	79.64 ± 2.25	85.45 ± 5.34	<b>88.17</b> ± 2.37	<u>12.25</u>	80.38 ± 3.03	82.06 ± 4.01	83.01 ± 3.48	<b>85.9</b> ± 5.04	5.52
Average	79.94 ± 4.08	81.79 ± 3.63	86.49 ± 6.53	<b>87.81</b> ± 6.05	7.87	81.27 ± 4.86	82.26 ± 5.03	84.56 ± 5.38	<b>86.46</b> ± 6.12	5.19
Median	75.06 ± 6.22	79.09 ± 5.48	83.76 ± 7.49	<b>86.49</b> ± 6.4	11.43	79.44 ± 6.67	80.15 ± 5.94	81.94 ± 5.55	<b>84.71</b> ± 6.86	5.27
MLP	89.76 ± 2.84	91.83 ± 1.47	93.03 ± 2.39	<b>96.19</b> ± 1.78	6.43	91.23 ± 2.38	93 ± 3.59	93.33 ± 2.53	<b>96.2</b> ± 1.78	4.97
FuzzyMLP	91.71 ± 2.92	92.59 ± 2.56	94.6 ± 2.34	<b>96.78</b> ± 1.88	5.07	94.42 ± 1.97	95.34 ± 2.35	<b>96.5</b> ± 2.36	96.38 ± 1.88	2.08
DCS-LA	91.97 ± 2.7	92.83 ± 2.31	95.1 ± 2.81	<b>96.3</b> ± 1.55	4.33	91.09 ± 2.74	92.89 ± 2.68	95.53 ± 2.61	<b>96.46</b> ± 1.16	5.37
DCS-DT	93.59 ± 3.02	95.15 ± 2.67	96.58 ± 1.83	<b>96.93</b> ± 1.46	<u>3.34</u>	94.07 ± 3.08	95.38 ± 3.15	96.63 ± 3.32	<b>96.77</b> ± 1.62	2.7
DCS-MCB	93.8 ± 3.38	95.87 ± 2.28	96.29 ± 2.63	<b>96.99</b> ± 1.7	3.19	95.58 ± 2.82	95.96 ± 2.61	96.43 ± 2.84	<b>96.67</b> ± 1.96	<u>1.09</u>

(2.67), Dcs-DT (2.68), Dcs-LA (3.64), MLP (4.43), Average (5.94), Naïve Bayes (6.77), Median (7.02), Voting (7.33), and Sum (7.8). It is important to emphasize that the average difference in performance of the ensembles with Fuzzy MLP was not higher than Dcs-MCB for the first time, since for all other ensemble sizes Fuzzy MLP had lower average difference. Another interesting fact is that, as in the previous section, the selection-based methods are in the top four places in the list of average differences, which is the opposite of ensembles with three base classifiers. Also, for the top five ensembles, the average difference decreased, when compared with ensembles with seven base classifiers. This shows that these methods are not affected when increasing the number of classifiers in an ensemble from 7 to 9.

In order to evaluate whether the difference in performance delivered by the combination methods is significant, the hypothesis tests (*t*-test) comparing the highest and lowest accuracy ensembles, using a confidence level of 95%, is performed. Table 3 shows the *p*-values for all differences in performance, including all four databases. It was observed from the bottom right part of Table 3 that all ensembles have differences in performance which are statistically significant in all databases. It could be concluded that all ensembles are equally affected by variations in the ensemble members when using ensembles with nine base classifiers, since their differences in performance is statistically significant in all four databases.

#### 4.8.2. Diversity

In order to analyze the level of diversity of the ensembles, Table 3 shows two diversity measures applied for all ensemble

structures and sizes. Lines 13–16 of Table 3 show the diversity measures provided by ensembles with nine base classifiers. As in the previous sections, the double fault measure did not reflect the expected decrease in the diversity of the ensembles, for all four databases, showing no relation between accuracy and diversity. In contrast, the entropy has always increased, when varying from non-hybrid to hybrid structures. This means that, based on the experiments of this paper, there might be a relation between accuracy and diversity emerged by the choice of the ensemble members, at least for the entropy measure.

#### 4.9. Discussion of the results

The main aim of this investigation was to analyze the effect of the choice of the ensemble members in diversity and accuracy of the ensembles. In order to do that, several structures and sizes of ensembles were analyzed. As a result of this investigation, it can be concluded that, as expected, the pattern of performance is data-dependent. However, some general statements can be drawn from this analysis, which can be described as follows.

- The highest accuracies were always reached by the hybrid structures, for all four ensemble sizes analyzed. The general pattern of performance was that the highest accuracies were reached by totally hybrid ensembles (HYB 7), followed by partially hybrid (HYB 5 and HYB 3) and by non-hybrid structures (NH).
- The variation in performance when using different ensemble members (*Dif* values) usually increased when

increasing the number of classifiers in an ensemble. However, when increasing from 7 to 9 base classifiers, the difference in performance decreased for some combination methods.

- The selection-based methods had higher *Dif* values than the fusion-methods for ensembles with few base classifiers (for instance, three base classifiers). This means that these methods are more sensitive to variations in the ensemble members than the fusion-based methods, when using ensembles with few base classifiers. However, as the number of base classifiers increases, this situation changes. For instance, for ensembles with nine base classifiers, all three selection-based methods were in the top four methods with low average *Dif* values. This means that the more base classifiers an ensemble has, the least sensitive to variations in the ensemble members the selection-based methods are.
- However, as already mentioned, the *Dif* values tend to increase when increasing the number of base classifiers for all combination methods. As a consequence of this, the differences in performance tend to be statistically significant for all combination methods. For instance, when using three base classifiers, only three (out of 10) combination methods have differences in performance statistically significant for all four databases. In contrast, when using nine base classifiers, all 10 combination methods have differences in performance statistically significant for all four databases.
- In relation to the diversity measures used in this investigation, the double fault measure only reflected the expected increase of diversity when using ensembles with three base classifiers. For all other ensemble sizes, the double fault measure had an inconstant behavior. On the other hand, entropy, which is a non-pairwise measure, did reflect the expected increase of diversity with all ensemble sizes. In this sense, there might be a relation between accuracy and diversity emerged by the choice of the ensemble members when using three base classifiers, illustrated by both diversity measures. For the other three ensemble sizes, the relation between accuracy and diversity was illustrated only by the entropy measure. It is important to emphasize that this result was not an expected one since that no clear relationships have been found so far between the two diversity measures (double fault and entropy) and the accuracy of classifier ensembles in the literature. This might be an indication that it is a data-dependent result, in which this may not be true when using other databases.

#### 4.10. Analyzing the functioning of selection-based methods

As it can be noticed from this investigation, the selection-based methods are more sensitive to variations in the ensemble members than the fusion-based methods when using ensembles with few base classifiers. In order to under-

stand more the effects of the choice of the ensemble members in the functioning of the DCS method, an analysis of classifier distribution was performed. In this analysis, it is shown the number of classifiers that were activated during the test procedure, for each class. After that, an average of the number of activated classifiers by the number of classes is calculated. Classifiers that were activated less than 10% of the test patterns were discarded. The ideal is that few classifiers should be activated to recognize the test patterns of a class (low average number of activated classifiers). When having few activated classifiers, it could be said that there is a high predominance of the classifiers over the classes, usually activating the same classifiers for test patterns of the same class.

As a result of this analysis, it could be observed that the non-hybrid structure (NH) uses more activated classifiers to recognize the test patterns of a class than the hybrid ones (HYB 3, HYB 5 or HYB 7). This fact is more evident when using ensembles with three base classifiers, in which the average number of activated classifiers for all hybrid ensembles is, on average, 1.25, while it was 2.1 for the non-hybrid structure (NH). This shows that the predominance of the classifiers for the non-hybrid ensembles is lower than for the hybrid ones. It is believed that it is caused by the similar behavior of the components of the non-hybrid structures. Also, it could be noticed that, for all other ensemble sizes, the number of activated classifiers is always lower for the hybrid structures than for the non-hybrid one.

Also, for the selection-based methods (DCS-DT and DCS-MCS), an analysis of the fusion frequency during the test phase was performed (number of times that the fusion method was used). Fig. 1 shows the fusion frequency of DCS-DT and DCS-MCS for the non-hybrid ensembles (NH), ensembles with three types of classifiers (HYB 3), ensembles with five types of classifiers (HYB 5) and ensembles with seven types of classifiers (HYB 7). These structures were analyzed with all four ensemble sizes (ES 3, ES 5, ES 7 and ES 9), for all four databases. As it can be observed from Fig. 1, the fusion frequency is always higher for the DCS-DT method. This is because this method uses a statistical test in order to decide whether to use the selection or fusion method. On the other hand, the DCS-MCS uses a threshold to decide between selection and fusion, which is not a statistical test. In addition to that, the highest fusion frequencies are reached by the non-hybrid structures for almost all cases, decreasing as the ensembles become more hybrid (HYB 5 and HYB 7). In addition, ensembles with nine base classifiers (ES 9) did not show a decrease in the fusion frequency, since all ensembles structures had similar values. The results in Fig. 1 only confirm the lack of predominance of classifiers when using non-hybrid ensembles. The selection-based methods (Dcs-DT and Dcs-MCB) decrease the competence of the classifiers in the classes to be recognized. In this sense, the fusion method is more activated. This is reflected in the variation of performance delivered by the selection-based methods

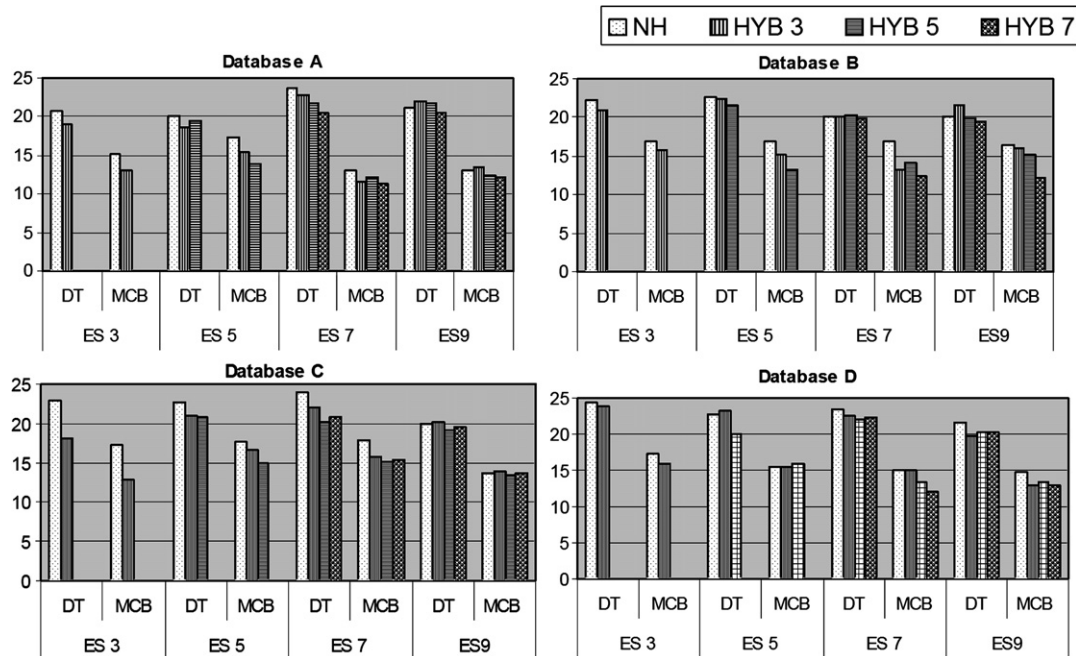


Fig. 1. The fusion frequency used in the DCS-DT and DCS-MCS methods.

when varying ensemble members, mainly in ensembles with few base classifiers.

## 5. Final remarks

In this paper, an investigation of different ensemble structures and sizes was performed. The main aim of this paper was to investigate how hybrid (different models of classifiers) and non-hybrid (same model of classifiers) ensemble structures behave, in terms of accuracy and diversity, when varying the ensemble members. Also, it aimed to investigate the effect of varying the ensemble members in the performance of some combination methods. This investigation was done using four different databases.

Through this analysis, it could be observed that the highest accuracies were almost always reached by the hybrid structures, for all four ensemble sizes analyzed. Of the fusion-based methods, the highest average accuracy was reached by ensembles combined by Fuzzy MLP, while Dcs-MCB had delivered the highest average accuracy of the selection-based methods. In addition to this, the selection-based methods are more sensitive to variations in the ensemble members (higher *Dif* values) than the fusion-based methods, when using ensembles with few base classifiers. When increasing the number of base classifiers, the selection-based become less sensitive to variations in the ensemble members. However, the *Dif* values tend to increase when increasing the number of base classifiers for all combination methods. As a consequence of this, the differences in performance tend to be statistically significant for all combination methods. In other words, all the combination methods tend to be more sensitive to changes

in the ensemble members when using a high number of base classifiers.

Finally, when analyzing the diversity of the ensembles, it could be noticed that the entropy measure illustrated the expected increase in diversity reached by the choice of the ensemble members. However, the double fault measure only reflected the relation with the expected diversity for ensembles with three base classifiers. It is important to emphasize that entropy is a non-pairwise measure, while double fault is a pairwise measure.

The results obtained in this paper show that the choice of the ensemble members is important to the accuracy and diversity of ensembles. Of course, it is not the only factor to be taken into account. Other parameters, such as: training data, feature input and/or initial parameters of the classifiers have also to be taken into account.

## References

- Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P., 2005. Ensemble diversity measures and their application to thinning. *Information Fusion* 6 (1), 49–62 (Special issue on Diversity in Multiple Classifier System).
- Blake, C.L., Merz, C.J., 1998. UCI Repository of machine learning databases. University of California, Department of Information and Computer Science, Irvine, CA. <<http://www.ics.uci.edu/~mllearn/MLRepository.html>>.
- Breiman, L., 1996. Bagging predictors. *Machine Learning* 24, 123–140.
- Breiman, L., 1999. Using adaptive bagging to Debias regressions, Technical report 547, Dept. Statistics, University of California, Berkeley.
- Canuto, A., 2001. Combining neural networks and fuzzy logic for applications in character recognition. Ph.D. thesis, University of Kent.
- Canuto, A.M.P., Souto, M.C.P., Santos, A., Silva, S.M., Bezerra, V.S.A., 2005. Comparative analysis of the performance of hybrid and

- non-hybrid multi-classifier systems. In: Internat. Joint Conf. on Neural Networks (IJCNN), 2005, Montreal, Proc. of IJCNN 2005.
- Canuto, A.M.P., Oliveira, L., Xavier Junior, J., Santos, A., Abreu, M., 2005. Performance and diversity evaluation in hybrid and non-hybrid structures of ensembles. In: 5th Internat. Conf. on Hybrid Intelligent Systems, pp. 285–290.
- Czyz, J., Sadeghi, M., Kittler, J., Vandendorpe, L., 2004. Decision fusion for face authentication. In: Proc. First Internat. Conf. on Biometric Authentication, pp. 686–693.
- Duin, R.P.W., Tax, D.M.J., 2000. Experiments with classifier combining rules. In: Proc. First Internat. Workshop on Multiple Classifier Systems. Lecture Notes in Computer Science, vol. 1857, pp. 16–29.
- Freund, W., 1995. Boosting a weak learning algorithm by majority. Inform. Comput.
- Giacinto, G., Roli, F., 2001. Design of effective neural network ensembles for image classification. Image Vis. Comput. J. 19 (9–10), 697–705.
- Giacinto, G., Roli, F., 2001. Dynamic classifier selection based on multiple classifier behaviour. Pattern Recognition 34, 1879–1881.
- Kuncheva, L., 2002. Switching between selection and fusion in combining classifiers: An experiment. IEEE Trans. Systems Man Cybernet. – Part B 32 (2), 146–155.
- Kuncheva, L.I., 2004. Combining Pattern Classifiers. Methods and Algorithms. Wiley.
- Kuncheva, L., Whitaker, C., 2003. Measures of diversity in classifier ensembles. Machine Learning 51, 181–207.
- Kuncheva, L.I., Skurichina, M., Duin, R., 2002. An experimental study on diversity for bagging and boosting with linear classifier. Information Fusion 3 (2), 245–258.
- Kuncheva, L.I., Skurichina, M., Duin, R.P.W., 2002. An experimental study on diversity for bagging and boosting with linear classifiers. Information Fusion 3 (2), 245–258.
- Lemieux, A., Parizeau, M., 2003. Flexible multi-classifier architecture for face recognition systems. In: The 16th Internat. Conf. on Vision Interface.
- Mitchell, T., 1997. Machine Learning. McGraw-Hill.
- Ruta, D., Gabrys, B., 2005. Classifier selection for majority voting. Information Fusion 6 (1), 49–62 (Special issue on Diversity in Multiple Classifier System).
- Sharkey, A.J.C., 1999. Multi-net system. In: Sharkey, A.J.C. (Ed.), Combining Artificial Neural Nets: Ensemble and Modular Multi-net Systems. Springer-Verlag, pp. 1–30.
- Shipp, C.A., Kuncheva, L.I., 2002. Relationships between combination methods and measures of diversity in combining classifiers. Information Fusion 3 (2), 135–148.
- Shipp, C.A., Kuncheva, L.I., 2002. An investigation into how adaboost affects classifier diversity. In: Proc. IPMU, pp. 203–208.
- Tsymbol, A., Pechenizkiy, M., Cunningham, P., 2005. Diversity in search strategies for ensemble feature selection. Information Fusion 6 (1), 83–98 (Special issue on Diversity in Multiple Classifier System).
- Windeatt, T., 2005. Diversity measures for multiple classifier system analysis and design. Information Fusion 6 (1), 21–36 (Special issue on Diversity in Multiple Classifier System).
- Woods, K., Kegelmeyer, W., Bowyer, K., 1997. Combination of multiple classifiers using local accuracy estimates. IEEE Trans. Pattern Anal. Machine Intell. 19 (4), 405–410.
- Zhou, D., Zhang, J., 2002. Face recognition by combining several algorithms. Pattern Recognition 3 (3), 497–500.