An Empirical Study on Diversity Measures and Margin Theory for Ensembles of Classifiers

Marcelo N. Kapp
École de Technologie Supérieure
Montreal, QC, Canada
kapp@livia.etsmtl.ca

Robert Sabourin
École de Technologie Supérieure
Montreal, QC, Canada
robert.sabourin@etsmtl.ca

Patrick Maupin
Defence Research and Development Canada
Valcartier, QC, Canada
patrick.maupin@drdc-rddc.gc.ca

***Abstract*— The main goal of this paper is to investigate the relationship between two theories widely applied to explain the success of classifiers fusion: diversity measures and margin theory. In order to achieve this, we realized an empirical study which evaluates some classical measures related to these two theories with respect to ensembles accuracy. In particular, this study revealed valuable insights on how these two theories can influence each other, and how the application of margin based measures can be useful for the evaluation and selection of ensembles of classifiers with majority voting.**

Keywords: Ensemble of Classifiers, Diversity Measures, Margin Theory, Majority Voting, Classifiers Fusion.

I. Introduction

The fusion of classifier decisions into ensembles has been widely applied to improve the performance of single classifiers. For this reason, over the last years, several works on ensembles of classifiers have been conducted to find measures that could be well correlated with ensembles' accuracy [1]–[12]. However, despite of the efforts, the understanding of the effectiveness of ensembles methods has still intrigued many authors and claimed for new researches on classifiers fusion.

Nevertheless, a consensus in the literature indicates the presence of some diversity between the ensembles members as the main factor for improving the overall performance [4], [7]–[10]. Whereas, as far as we know, even though it is well accepted that diversity is a necessary condition for improving the majority vote accuracy, there is no general agreement on how to quantify or to deal with it.

On the other hand, margin theory has also allured some attention in the literature, since it seems to be able to cast the study of ensembles of classifiers into a large margin classifiers context (as Support Vector Machines [13], for example). In short, the margin theory was firstly applied by Schapire et al. [2] to provide an explanation on how the boosting method works. After that, other authors have used this theory to create new ensemble methods [14], [15]. Indeed, the margin theory seems to be an efficient tool to understand and evaluate the ensemble's learning.

In light of all this, the main goal of this paper is to investigate through of an empirical study what is the possible relationship between the diversity and margin theories, and in particular how they influence the majority vote accuracy. In order to achieve this, we start our study by surveying some classical diversity measures and some measures related to the margin theory. Afterward, an experimental protocol similar to one introduced by Valentini [11] for characterizing ensembles of Support Vector Machines is employed to evaluate the measures and draw some results. In addition, a discussion on the obtained results is also offered, in which we try to answer some questions currently found in the literature, such as: which measure could offer the best guidance to evaluate the classifiers fusion? How are the diversity measures related to each other? Is there a relationship between diversity, margins, and ensemble accuracy? Which are the best measures for observing such relationship? Finally, we conclude this study which surely provides valuable insights on methods for fusion evaluation and selection of ensembles of classifiers.

This paper is organized as follows. Section II surveys classical measures to estimate diversity for classification fusion. Section III introduces the margin theory for ensemble of classifiers and measures related to it. Section IV describes the experimental protocol applied and the obtained results. Finally, we outline the conclusions in Section V.

II. Diversity Measures

Diversity has been quantified in several ways for classification fusion. As a result, different measures have been proposed in the literature. In this work, we apply seven well-known diversity measures which usually are grouped into two types: pairwise and non-pairwise [8]. Moreover, in here each diversity measure name is accompanied with a downward arrow $\downarrow$ or upward arrow $\uparrow$ indicating if the diversity obtained is decreasing or increasing with its value.

A. Pairwise Measures

In pairwise measures, firstly the diversity between all pairs of classifiers is calculated. Thereafter, the overall diversity measure values are computed as the mean of the pairwise values. For instance, given L classifiers, $\frac{L\times(L-1)}{2}$ pairwise diversities are measured, and the final diversity $\bar{d}$ is defined by an average:

$$\bar{d} = \frac{2}{L(L-1)} \sum_{\substack{i,j=1,\ldots,L \\ i\neq j}} d_{i,j} \tag{1}$$

In general for a pairwise measure, N is the total number of samples, N^{11} is the number of times that both classifiers

are correct, N^{00} represents the number of times that both classifiers are incorrect, and N^{10} and N^{01} denote the number of times when just the first or second classifier is correct, respectively. Below, we describe some pairwise measures applied in this work.

1) Q average (↓): This measure is computed for pairs of classifiers i and j as:

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{2}$$

2) Disagreement measure (↑): This measure denotes the ratio between the number of observations where one classifier is correct and the other is incorrect with respect to the total number of observations [16]. For a pair of classifiers i and j, it is computed by:

$$D_{i,j} = \frac{N^{10} + N^{01}}{N} \tag{3}$$

3) Double-fault measure (↓): The double-fault measure estimates the probability of coincident errors for a pair of classifiers. It is defined for a pair of classifiers i and j as [16], [17]:

$$DF_{i,j} = \frac{N^{00}}{N} \tag{4}$$

B. Non-pairwise Measures

Unlike pairwise measures, non-pairwise measures are not calculated by comparing pairs of classifiers, but by comparing all L classifiers as a whole. Below there are some examples of these types of measures:

1) Kohavi-Wolpert (KW) variance (↑): Let $l(x_j)$ be the number of classifiers that correctly recognize x_j. From the formula for the variance [1], the diversity measure becomes:

$$kw = \frac{1}{NL^2}\sum_{j=1}^{N} l(x_j)(L - l(x_j)) \tag{5}$$

2) Generalized diversity (↑): Let Z be a random variable to represent the proportion of classifiers that are incorrect on a randomly drawn sample x, p_i is the probability that $Z = i/L$, and $p(i)$ is the probability that i randomly chosen members will be wrong on a randomly chosen x. The generalized diversity is defined as [18]:

$$p(1) = \sum_{i=1}^{L}\frac{i}{L}p_i, \quad p(2) = \sum_{i=1}^{L}\frac{i(i-1)}{L(L-1)}p_i \tag{6}$$

$$GD = 1 - \frac{p(2)}{p(1)} \tag{7}$$

3) Ambiguity (↑): The ambiguity measure was proposed by Zenobi and Cunningham [5]. Basically, it measures the disagreement among the classifiers predictions $\hat{y}_j$ with respect to the majority prediction y_m, where the factor correctness is not important. The ambiguity measure can be defined as:

$$A = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} \parallel y_{im} \neq \hat{y}_{ij} \parallel \tag{8}$$

4) Difficulty (↓): Unlike the ambiguity measure, the difficulty measure [19] like most of the measures is calculated taking into account the base classifiers' correctness. The goal is to measure the degree of classification difficulty of samples. Basically, this measure is defined to be the variance of a X random variable which denotes the proportion of classifiers that correctly classify a sample x: $\theta = Var(X)$. This way, a small variance indicates that the ensemble is diverse while a high one indicates the opposite.

III. Margin Theory

The margin theory was originally applied to explain the success of Boosting [2] and to develop the Support Vector Machines theory [13]. For ensembles of classifiers, the concept of the margin follows the same idea introduced by Schapire et al. [2]. However, it can be computed according to two different definitions. In general, the margin of a sample x is computed by Equation 9 [12], [20], where v_y is the number of votes for the true class and v_c is the number of votes for any other class.

$$margin(x, y) = v_y - \sum_{\substack{c=1,\ldots,L \\ c\neq y}}^{L} v_c \tag{9}$$

On the other hand, the margin of a sample x can also be obtained as follows:

$$margin(x, y) = v_y - \max_{\substack{c=1,\ldots,L \\ c\neq y}} v_c \tag{10}$$

Naturally, for two-class problems these definitions are quite similar. The main difference is that, while the first definition applies a sum operation, the second definition computes a max operation. Moreover, a major concern needs to be solved in relation to multi-class problems. By the first definition, the margins can represent a lower bound, since they can assume negative values even when the correct label gets the most of votes (when there is a plurality, but not a majority) [20]. Therefore, note that this way the margins can vary more producing a major "diversity" in terms of their values. In this study we employ the first definition in our experiments, since it is the most used in the literature [12], [20].

In addition, further than the notion of margins, there are three main measures related to this theory. The first one is called *minimum margin*. The minimum margin of an ensemble of classifiers on a dataset D is defined as the smallest value of margin obtained to any correct label [20]. Therefore, the minimum margin is governed by:

$$min(margin(D)), \tag{11}$$

given the final decision was correct. The second one is called cumulative margins distributions. Usually, cumulative margins distributions are computed by two simple steps. First, the set of margin values from a dataset is sorted. Next, for each possible value of margin is calculated the percentage of the samples which their margins are lower or equal to the current value. Graphics of cumulative distribution of margins were

firstly introduced by Schapire et al. [2] to demonstrate that Boosting maximizes margins. Finally, the third measure is called *average margin* which is denoted as an average over all margins of a dataset D:

$$Avg.margin(D) = \frac{1}{N}\sum_{i=1}^{N} margin(x_i, y_i) \qquad (12)$$

IV. Experimental Protocol

In order to analyze the relationship between diversity measures and the margin theory previously introduced, an experimental protocol similar to one realized by Valentini [11], [21] for characterizing ensembles of Support Vector Machines (SVM) was carried out.

In short, the experimental setup has been organized into two steps. First, we have selected a complex synthetic problem denoted P2 and two other real-world problems: Satimage and Letter. P2 [21] is a classification problem that consists of two classes (I and II), where each decision region for each class is delimited by one, two, or even more than four equations and without overlapping between the distributions, see Figure 1. The four equations are simple polynomial and trigonometric functions, as follows:

$$\begin{aligned} Eq_1(x) &= 2 \times sin(x) + 5 \\ Eq_2(x) &= (x-2)^2 + 1 \\ Eq_3(x) &= -0.1x^2 + 0.6 \times sin(4x) + 8 \\ Eq_4(x) &= \frac{(x-10)^2}{2} + 7.902 \end{aligned} \qquad (13)$$

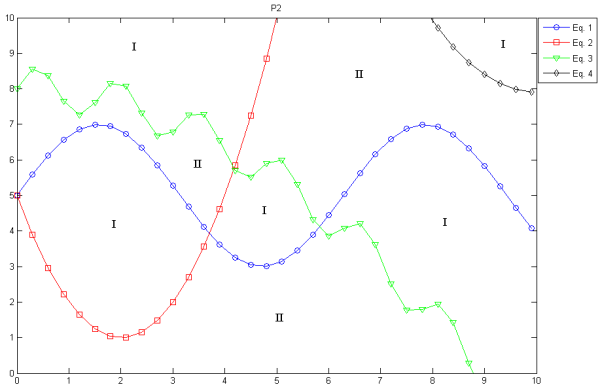


Fig. 1. Illustration of the P2 Problem.

For P2 problem, a large dataset was generated and split into a small training set and a large testing set composed of 100 and 10,000 samples respectively. On the other hand, Satimage and Letter are multi-classes problems from Satlog collection [22]. For these problems the same original distributions of samples for training and validation sets were used. Table I lists the main information about the databases of these three problems.

After that, ensembles of SVMs with RBF-kernel varying the C and γ parameters were built based on the Bagging method [23]. Therefore, ensemble members were created by taking random samples with replacement from a given original training set D, and by building them on D_i bootstrapped subsets, where $i = 1, \ldots, L$. L was set to 50 for all problems. Finally, for a test sample x, the final classification decision was made by taking the majority vote over the class labels produced by each ensemble member. In order to adapt SVM for the multi-class problems, an one-against-one strategy was employed. Moreover, a RBF kernel was used because it nonlinearly maps samples into a higher dimensional space. Furthermore, this kernel has also obtained superior power of generalization and lower complexity than the Polynomial kernel [21], for example. Specifically, the variations of the C and γ parameters were done based on these values:

$$\begin{cases} \gamma \in & \{10000, 2500, 100, 25, 4, 1, 0.25, 0.04, 0.01, 25e03, \\ & 4e04, 1e04, 25e05, 11e05, 6e06, 4e06, 1e06\} \\ C \in & \{0.01, 0.1, 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\} \end{cases}$$

Above all, 17x12=204 combinations of models were trained and evaluated on each D_i subset of data totalizing more than 30,600 different RBF-SVMs for all databases. After that, the measures introduced in Sections II and III were evaluated over the ensembles. In addition, we have also computed the average loss of predictions between base classifiers outputs $\hat{y_j}$ and a true class y_i^*. In fact, the average loss represents the mean error rate between the ensemble members as is defined in Equation 14. Furthermore, the generalization error is computed according to Equation 15, where y_m denotes the majority vote from an ensemble.

$$A.Loss = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} \parallel \hat{y}_{ij} \neq y_i^* \parallel \qquad (14)$$

$$G.Error = \frac{1}{N}\sum_{i=1}^{N} \parallel y_{im} \neq y_i^* \parallel \qquad (15)$$

The obtained results are shown in Tables II, III, and IV, and depicted in Figures 4 (a)-(f). Of course, note that although it is theorically possible that ensembles composed of members less performing than 50% reach good performances combined, in here as it occurs in most of the studies on ensembles of classifiers, the relevance of the results is better visualized on ensembles with the average loss lower than 50%. Taking into account this, the results in the Tables II-IV where picked out from ensembles with the average loss (mean error rate) lower or equal to 40%.

TABLE I
Information on the databases.

Database	# Classes	# Features	Training	Test
P2	2	2	100	10,000
Satimage	6	36	3,104	1,331
Letter	26	16	10,500	4,500

TABLE II

RESULTS OBTAINED ON THE P2 DATABASE.

Measures	C	γ	Value	A.Loss(%)	G.Error(%)
Avg. Loss	2	100	0.1719	17.19	12.78
Gen. Error	1	100	0.1274	17.59	12.74
Diff.(↓)	0.1	25	0	35.06	28.91
Amb.(↑)	0.1	25	0.2563	35.06	28.91
DF(↓)	2	100	0.1006	17.19	12.78
Dis.(↑)	0.1	25	0.3508	35.06	28.91
K-W(↑)	0.1	25	0.1719	35.06	28.91
Gen. Div.(↑)	0.1	25	0.5003	35.06	28.91
Q Avg.(↓)	0.1	25	0.3100	35.06	28.91
Min. Marg.(↑)	0.1	25	0	35.06	28.91
Avg. Margin(↑)	2	100	0.6561	17.19	12.78
CI(↓)	2	100	0.6734	17.19	12.78

TABLE III

RESULTS OBTAINED ON THE SATIMAGE DATABASE.

Measures	C	γ	Value	A.Loss(%)	G.Error(%)
Avg. Loss	5	1	0.1091	10.91	9.92
Gen. Error	20	1	0.0969	11.06	9.69
Diff.(↓)	0.1	0.25	0.1586	15.37	15.10
Amb.(↑)	200	$6e06$	0.0800	32.94	29.30
DF(↓)	50	1	0.0816	11.08	9.77
Dis.(↑)	1000	0.25	0.0787	12.37	10.59
K-W(↑)	1000	0.25	0.0386	12.37	10.59
Gen. Div.(↑)	1000	0.25	0.3181	12.37	10.59
Q Avg.(↓)	1000	0.25	0.9568	12.37	10.59
Min. Marg.(↑)	50	$1e04$	0.2000	26.59	26.52
Avg. Margin(↑)	5	1	0.7818	10.91	9.92
CI(↓)	50	1	0.4622	11.08	9.77

TABLE IV

RESULTS OBTAINED ON THE LETTER DATABASE.

Measures	C	γ	Value	A.Loss(%)	G.Error(%)
Avg. Loss	10	1	0.0456	4.56	3.44
Gen. Error	20	1	0.0336	4.58	3.36
Diff.(↓)	5	1	0.1298	4.71	3.80
Amb.(↑)	1	25	0.1685	29.94	24.93
DF(↓)	20	1	0.0275	4.58	3.36
Dis.(↑)	1	25	0.1247	29.94	24.93
K-W(↑)	1	25	0.0611	29.94	24.93
Gen. Div.(↑)	500	0.25	0.4063	6.02	4.31
Q Avg.(↓)	2	25	0.9526	28.79	23.78
Min. Marg.(↑)	10	4	-0.2000	4.77	3.47
Avg. Margin(↑)	10	1	0.9088	4.56	3.44
CI(↓)	20	1	0.1247	4.58	3.36

A. Discussion

From the obtained results we could observe some quite interesting aspects related to the relationship between diversity measures, margin theory, and majority vote accuracy. In order to well present this discussion, we firstly examine the obtained results for each theory with respect to the majority vote accuracy, and next we discuss and give details on the relationship between them.

1) Diversity results: Definitely, the obtained results have shown that diversity is very important for majority vote accuracy. For instance, ensembles with the lowest average loss of predictions between their members have not reached the lowest generalization error, see in Tables II-IV and Figures 4 (a)-(f). It means that individual performances of members are one factor that contributes to the overall ensemble performances, but they are not sufficient. Indeed, some diversity is requested to get the highest majority vote performances.

However, as we have outlined in the introduction, the relationship between diversity and ensemble accuracy has been considered sometimes incoherent or confuse in the literature [7], [8]. In particular, through our experiments we could also see that the results for some diversity measures can be quite ambiguous. For example, for all problems we could note that there were diversity measures that have assumed ambiguous values for several kinds of ensembles, even if they have had different mean error rates (average loss) or majority vote error (generalization error). It has occurred mainly to the diversity measures which are focused on the increasing of the variance between the base classifiers outputs, like the Qaverage, Disagreement, Ambiguity, and KW variance measures.

To sum up, we could also note some differences between the diversity measures studied. Such differences allow us to categorize the measures into two groups according to their results in relation to the members or ensembles accuracies. In the first one are the diversity measures that were "weakly" related, such as: Qaverage, Disagreement, Ambiguity, and KW variance measures. On the other hand, Generalized Diversity, Difficulty, and Double-Fault measures belong to the second group denoted as "strongly" related. Yet concerning this last group, we can outline that Double-Fault measure was more related to the ensemble accuracy, followed by the Difficulty and Generalize Diversity measures which were a little bit more ambiguous. Similar observations about the behaviors of the Double-Fault and Difficulty measures have been also done in [7].

2) Margin results: The obtained results for margins based measures have allowed us to explore this theory in a deeper way. In particular, we have evaluated the main measures provided by the margin theory: minimum margin, cumulative margins distributions, and average margin, on test sets in order to obtain some insights on this theory and the majority vote accuracy. In fact, in the literature it is common to find the statement that maximizing the margins on some data decreases the generalization error on future test sets. Taking this into account, in a first moment one would expect that maximizing the minimum margin measure for ensembles should be

accompanied with the minimum generalization error.

Unfortunately, what we could note is that the minimum margin measure can be really instable. For instance, as it can be seen in Figure 4 (c), the tracking of the maximum minimum margin can be quite instable, or even ambiguous around the "best ensemble". It could be one of the factors responsible for in some situations the greed increasing of the minimum margin can be not satisfactory.

Furthermore, after the evaluation of the minimum margin measure, we have also employed cumulative margins distributions in order to analyze the margin values of some ensembles, see in Figure 2. Definitely, the ensembles with best performances have reached larger margins than ensembles with poor performances.

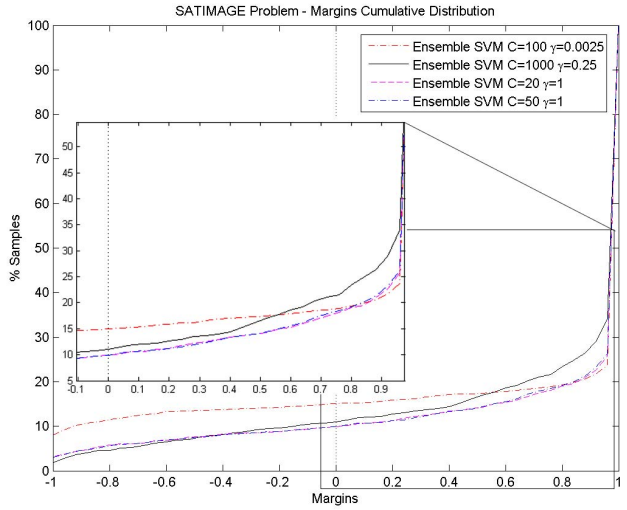


Fig. 2. Some cumulative margins distributions computed on the Satimage problem.

However, in general only the cumulative margins distribution graphics may not be enough to evaluate ensembles, since they do not represent a single value which should be more useful. Then, in face of this we extend our analysis on the average margin measure. In fact, we have observed that this measure is very stable. Particulary, we can assert that is more relevant to compute the average margin than the minimum one. In contrast, although the average margin over test instances represents an estimate of expected margin for a classification problem [24], after an analysis of the results, it is clear that it is strictly related to the average loss (mean error rate) of the base classifiers. Indeed, one can see that the maximum values of average margin correspond to the minimum values of average loss in Tables II-IV and Figures 4(a)-(f).

Therefore, maximizing the average margin points out the ensembles composed of the strongest individual members in a given pool. In general, this fact is not much interesting since there is a great tendency that in a "limit" of the highest possible individual performances, the base classifiers will be very similar, with so low diversity, and hence their team may be not reach the maximum majority vote accuracy. Unquestionably, despite of the fact that in most of the times the maximum values of average margins accompany the minimum values of generalization error for some combinations of the C and γ parameters, usually the ensembles with the maximum average margin and minimum generalization error (majority vote error) in the extreme diverge, since their final solutions and results were different (see Tables II-IV). Regarding this capability of only pointing out the ensembles with the strongest members, we have started to question the usefulness of the margin theory for ensemble methods like Bagging or Random Subspaces [25], which unlike Boosting have a fixed size and do not multiple rounds on training data for changing their parameters.

In an effort to find another margin based measure able to predict a "better boundary" related to the majority vote error, we have examined more carefully the relationship between the expected majority vote error rate and the margins of the ensembles with the largest average margin and lowest generalization error. Particulary, we have compared their histograms formed only by frequencies of the "correct" margins with respect to the total of test samples. The histograms for the Satimage and Letter problems are depicted in Figures 3 (a)-(d).

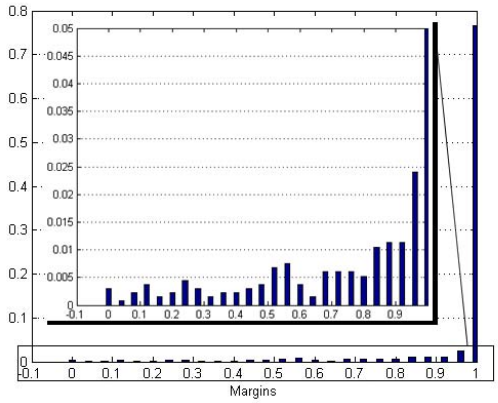
Based on the results it is possible to observe that ensembles with the lowest generalization error (Figures 3 (b) and (d)) have obtained margins with more "plurality" of values between the margins with a correct decision than ensembles with the largest average margin (Figures 3 (a) and (c)). It is clearly noticeable that while ensembles with the largest average margin reach high values of margins, ensembles with the lowest generalization error obtain margins relatively high, but tends also to produce values of margins more varied for the correct classified samples. These results have demonstrated how important is a balance between the increasing of the margins accompanied of some control to decrease or keep low the correlation between the ensemble members.

Thus, inspired by this idea, we have evaluated another measure denoted here as CI-measure, since it is based on the Chebishev's inequality. Above all, despite of the fact that we have considered the margin definitions in a slightly different way, the CI-measure represents widely the same idea of the upper bound defined by Breiman for the Random Forest method [15].

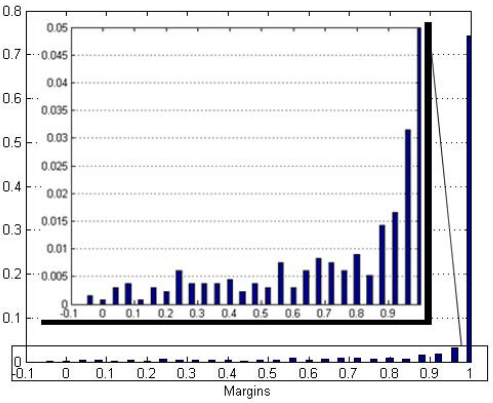
More specifically, Breiman has applied the Chebishev's inequality in order to establish a relation between the strength of the base classifiers (average margin) and the dependence between them (correlation) for predicting the generalization error. Thus, as demonstrated by Breiman, assuming an *average margin* ≥ 0, the Chebyshev's inequality provides:

$$PE^* \leq \frac{var(margins(D))}{(avg.margin(D))^2} \tag{16}$$

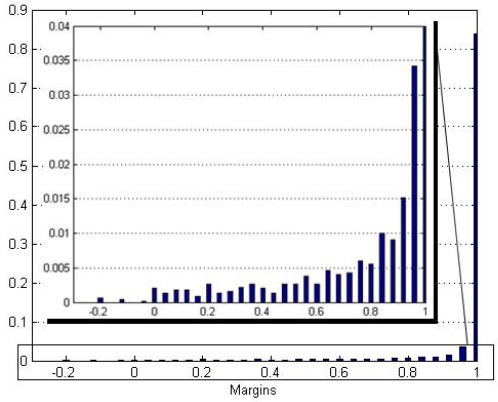
where PE^* is the probability of error. Using definitions, he has proved that the variance of the margins is lower or equal to the average of the correlation coefficients of pairs of classifiers times an average of variance between them. Then, it gives also a measure for the generalization error



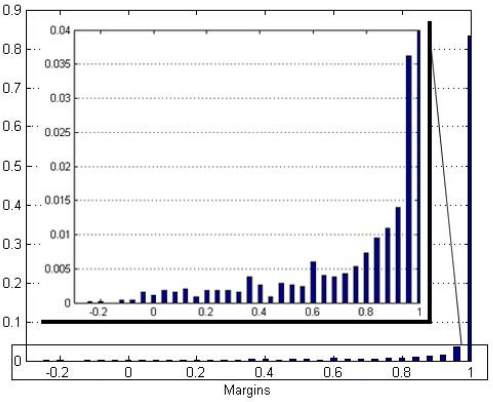
(a) Satimage ($C = 5, \gamma = 1$)



(b) Satimage ($C = 20, \gamma = 1$)



(c) Letter ($C = 10, \gamma = 1$)



(d) Letter ($C = 20, \gamma = 1$)

Fig. 3. Histograms of the 'correct" margins frequencies from ensembles with the largest average margin ((a) and (c)), and with the lowest generalization error ((b) and (d)) from Tables III and IV for the Satimage and Letter problems, respectively.

which although loose, it has produced interesting results, see in Tables II-IV and Figures 4 (a)-(f). By employing the CI-measure with the margin definition used here, it means that the ensembles must be sufficiently confident on their decisions with a certain majority of the votes, at least 50% in average (to obtain positive average margins). Whereas, it fulfills the same suggestive function for ensembles with majority voting in general as VC-type bounds do for other types of classifiers [15].

Now after the discussion on the obtained results for diversity measures and margin measures and their complex relations to the majority vote error, we finally discuss the relationship between these two theories.

3) Diversity x Margin results: Recently, Tang et al. [12] have demonstrated that enlarging of the margins can be the same that the enlarging of diversity. Nonetheless, based on the obtained results we can introduce more details on this relationship. Most importantly, we have observed that the maximization of the margins of ensembles with fixed size means to decrease the mean error rate of the team and also in parallel to increase the diversity. In fact, the diversity is only appropriately increased if it is represented by diversity measures "strongly" related to the average loss and hence to the average margin, such as: Generalized Diversity, Difficulty, and Double-Fault measures, since the other measures are more related just to the variance of the outputs and seem do not regard the individual members performances.

Furthermore, analyzing the results we could observe that the diversity measure Double-Fault, and the margin based measure CI-measure, were the two measures more related to the generalization error for all the problems. Maybe Double-Fault has produced a stable behavior because if strong classifiers are available (high average margin), this measure seeks to decrease the probability of identical errors (correlation between the members).

Therefore, in this point of view the relationship between the diversity and margin theories becomes strong, since as we have previously described, the CI-measure works according to this same balance. However, the CI-measure takes the advantage that such mechanism of balance is explicit at its formulation, the average margin is related to the strength of the base classifiers, and the variance of the margins can be

seen as diversity represented by the correlation between the base classifiers. Therefore, the boundary provided by the CI-measure seems to be a good evaluation measure for ensembles.

Overall, we can assert that if the relationship between some diversity measures and accuracy is not so strong, as a consequence it will not be regarding the margin theory too. Thus, if one regards the Accuracy-Diversity dilemma which states that highly accurate classifiers cannot be very diverse [7]. In fact, it is true since it is benefitial that the base classifiers be very strong (with large average margins), but also with the lowest possible correlation between them.

V. Conclusion

Several studies have provided evidences that diversity and large margins are very important to reach high performance in the classifiers fusion. On one hand Kuncheva et al. [7] based on the diversity theory have observed that boosted ensembles can produce higher diversity and generally higher accuracy. On the other hand, some authors have asserted that the increasing of margins is also responsible for decreasing the generalization error with boosted ensembles [2]. What could these two theories have in common? Motivated by these facts, in this paper we have investigated whether there is a useful relationship between the diversity and margin theories, and naturally the ensemble accuracy.

In short, even though it is not straightforward to characterize these relationships, we have drawn some conclusions. Firstly, diversity measures are indeed inadequate to evaluate or track the improvements in the overall performance. Mainly those measures more related to the variance between the ensembles members. This fact could explain why seeking diversity explicitly is ineffective to point out ensembles with optimal generalization performance. Moreover, only the increasing of the margins on a test dataset may not be a good option for ensembles of classifiers. Besides, the minimum margin measure seems not to be stable, and average margins indicated just ensembles composed of the strongest individual classifiers, but not with the best fusions.

Overall, we could also conclude that in a certain fashion, the relationship between the diversity and margin theories can be quite interdependent. Since the generalization error can be well estimated by the combination of strong precision of the base classifiers and a relative diversity between them. However, no sacrificing of accuracy of the base classifier is needed for diversity. In light of this, the use of measures based on the margin theory, like the CI-measure, can be a good direction for future researches on evaluation and selection of ensemble of classifiers.

References

[1] Ron Kohavi and David H. Wolpert. Bias plus variance decomposition for zero-one loss functions. In Lorenza Saitta, editor, *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 275–283. Morgan Kaufmann, 1996.

[2] Robert E. Schapire, Yoav Freund, Peter Barlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 322–330, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[3] Pedro Domingos. A unified bias-variance decomposition and its applications. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 231–238, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[4] Padraig Cunningham and John Carney. Diversity versus quality in classification ensembles based on feature selection. In *ECML '00: Proceedings of the 11th European Conference on Machine Learning*, pages 109–116, London, UK, 2000. Springer-Verlag.

[5] Gabriele Zenobi and Padraig Cunningham. Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pages 576–587, London, UK, 2001. Springer-Verlag.

[6] Dymitr Ruta and Bogdan Gabrys. A theoretical analysis of the limits of majority voting errors for multiple classifier systems. *Pattern Anal. Appl.*, 5(4):333–350, 2002.

[7] Ludmila Kuncheva, Marina Skurichina, and Robert P. W. Duin. An experimental study on diversity for bagging and boosting with linear classifiers. *Information Fusion*, 3(4):245–258, 2002.

[8] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, 51(2):181–207, 2003.

[9] Terry Windeatt. Vote counting measures for ensemble classifiers. *Pattern Recognition*, 36(12):2743–2756, 2003.

[10] Gavin Brown, Jeremy L. Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.

[11] Giorgio Valentini. An experimental bias-variance analysis of svm ensembles based on resampling techniques. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(6):1252–1271, 2005.

[12] E. K. Tang, P. N. Suganthan, and X. Yao. An analysis of diversity measures. *Mach. Learn.*, 65(1):247–271, 2006.

[13] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer Verlag., 1995.

[14] L. Breiman. Arcing the edge, 1997.

[15] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[16] D. B. Skalak. The Sources of Increased Accuracy for Two Proposed Boosting Algorithms. In P. Chan, editor, *Working Notes of the AAAI Workshop on Integrating Multiple Learned Models*, pages 120–125, 1996.

[17] Giorgio Giacinto and Fabio Roli. An approach to the automatic design of multiple classifier systems. *Pattern Recogn. Lett.*, 22(1):25–33, 2001.

[18] D. Partridge and W. Krzanowski. Design of effective neural network ensembles for image classification purposes. *Software diversity: practical statistics for its measurement and exploitation*, 39(10):707–717, 1997.

[19] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(10):993–1001, 1990.

[20] Adam J. Grove and Dale Schuurmans. Boosting in the limit: maximizing the margin of learned ensembles. In *AAAI '98/IAAI '98: Procs. of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 692–699.

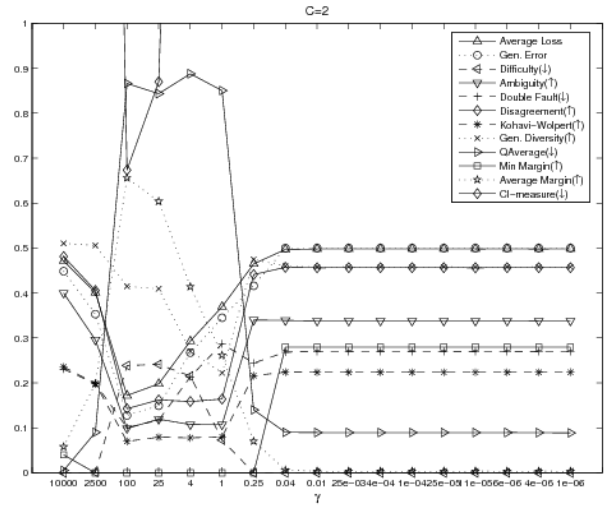
[21] G. Valentini. *Ensemble methods based on bias-variance analysis*. PhD thesis, University of Genova, Genova, Italy, 2003.

[22] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. Machine learning, 1994.

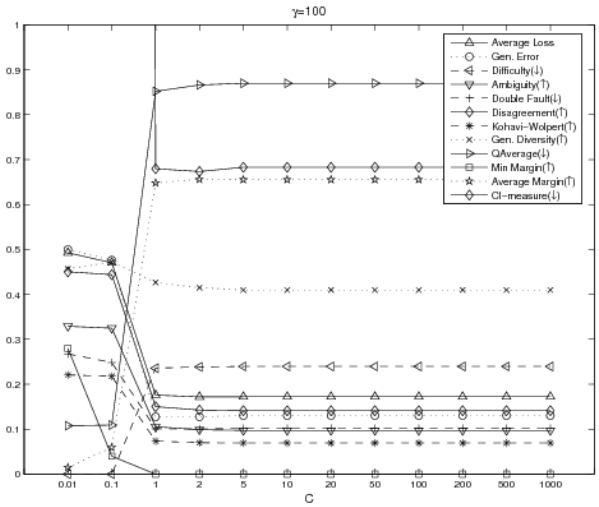
[23] Leo Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

[24] Alexey Tsymbal, Mykola Pechenizkiy, and Padraig Cunningham. Dynamic integration with random forests. In *ECML*, pages 801–808, 2006.

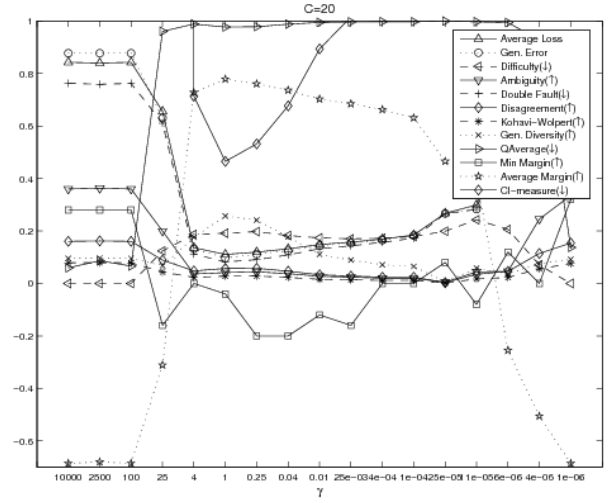
[25] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844, 1998.



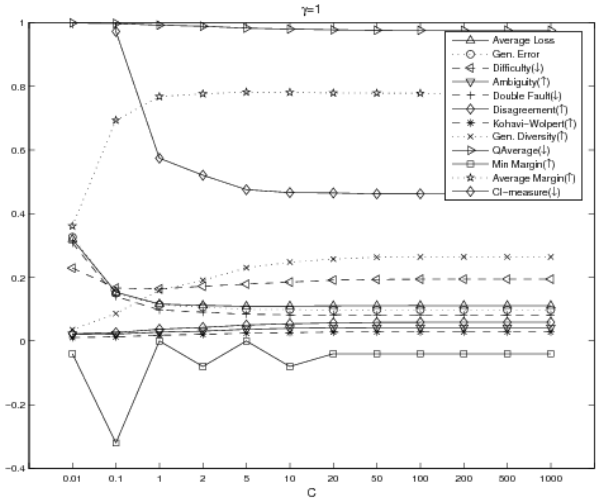
(a) P2 - fixed C



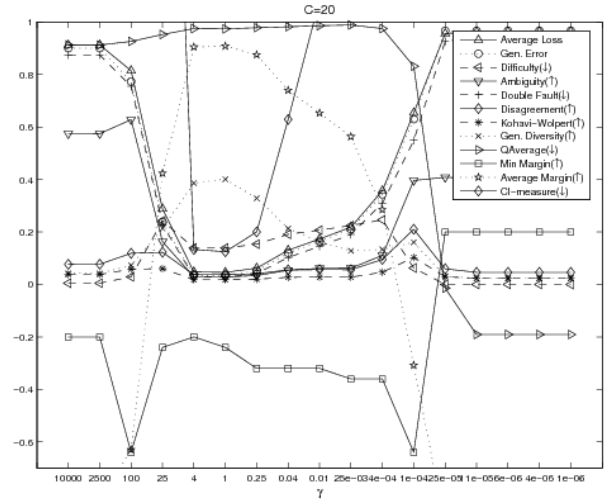
(b) P2 - fixed γ



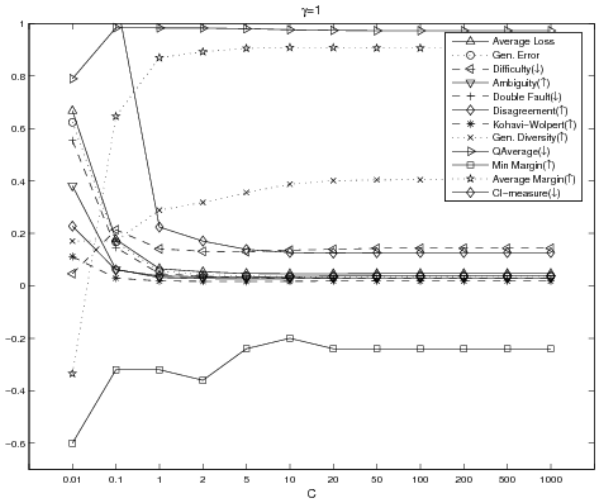
(c) Satimage - fixed C



(d) Satimage - fixed γ



(e) Letter - fixed C



(f) Letter - fixed γ

Fig. 4. Some results obtained for ensembles with the best combinations of C and γ parameters on two different perspectives for all the databases. In the first place, Figures (a), (c), and (e) depict results for ensembles with the best combination by fixing and varying the C and γ parameters, respectively. In the second place, Figures (b), (d), (f) depict the results obtained by fixing the best γ parameter found and varying C.