International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 23, No. 4 (2009) 659–686

World Scientific
www.worldscientific.com

COMPOUND DIVERSITY FUNCTIONS FOR ENSEMBLE SELECTION

ALBERT HUNG-REN KO* and ROBERT SABOURIN†

LIVIA, École de Technologie Supérieure, University of Quebec
1100 Notre-Dame West Street, Montreal, Quebec, Canada H3C 1K3
**albert@livia.etsmtl.ca*
†*robert.sabourin@etsmtl.ca*

ALCEU DE SOUZA BRITTO JR.

PPGIA, Pontifical Catholic University of Parana
Rua Imaculada Conceicao, 1155, PR 80215-901, Curitiba, Brazil
alceu@ppgia.pucpr.br

An effective way to improve a classification method's performance is to create ensembles of classifiers. Two elements are believed to be important in constructing an ensemble: (a) the performance of each individual classifier; and (b) diversity among the classifiers. Nevertheless, most works based on diversity suggest that there exists only weak correlation between classifier performance and ensemble accuracy. We propose compound diversity functions which combine the diversities with the performance of each individual classifier, and show that there is a strong correlation between the proposed functions and ensemble accuracy. Calculation of the correlations with different ensemble creation methods, different problems and different classification algorithms on 0.624 million ensembles suggests that most compound diversity functions are better than traditional diversity measures. The population-based Genetic Algorithm was used to search for the best ensembles on a handwritten numerals recognition problem and to evaluate 42.24 million ensembles. The statistical results indicate that compound diversity functions perform better than traditional diversity measures, and are helpful in selecting the best ensembles.

Keywords: Diversity; ensemble of classifiers; pattern recognition; majority voting.

1. Introduction

The purpose of pattern recognition systems is to achieve the best possible classification performance. A number of classifiers are tested in these systems, and the most appropriate one is chosen for the problem at hand. Different classifiers usually make different errors on different samples, which means that, by combining classifiers, we can arrive at an ensemble that makes more accurate decisions.[5,17,21,24,26,35,38] In order to have classifiers with different errors, it is advisable to create diverse classifiers. For this purpose, diverse classifiers are grouped together into what is known

as an Ensemble of Classifiers (EoC). There are several methods for creating diverse classifiers, among them Random Subspaces,[15] Bagging and Boosting.[13,20,29] The Random Subspaces method creates various classifiers by using different subsets of features to train them. Because problems are represented in different subspaces, different classifiers develop different borders for the classification. Bagging generates diverse classifiers by randomly selecting subsets of samples to train classifiers. Intuitively, based on different sample subsets, classifiers would exhibit different behaviors. Boosting uses parts of samples to train classifiers as well, but not randomly; difficult samples have a greater probability of being selected, and easier samples have less chance of being used for training. With this mechanism, most created classifiers will focus on hard samples and can be more effective.

There are two levels of problems in optimizing the performance of an EoC. First, how are classifiers selected, given a pool of different classifiers, to construct the best ensemble? Second, given all the selected classifiers, what is the best rule for combining their outputs? These two problems are fundamentally different, and should be solved separately to reduce the complexity of optimization of EoCs; the former focuses on ensemble selection[3,5,18,21,25,28,34] and the latter on ensemble combination, i.e. the choice of fusion functions.[17,26,28,32,38] For ensemble selection, the problem can be considered in two steps: (a) find a pertinent objective function for selecting the classifiers; and (b) use a pertinent searching algorithm to apply this criterion. Obviously, a correct criterion is one of the most crucial elements in selecting pertinent classifiers.[5,21,25,28] It is considered that, in a good ensemble, each classifier is required to have different errors, so that they will be corrected by the opinions of the whole group.[17,20,21,27,28] This property is regarded as the diversity of an ensemble.

Diversity is important for ensemble selection and cannot be substituted by fusion functions. There are several reasons for this: First, for a large number of classifiers, fusion functions need to take into account all classifier outputs for each evaluation,[3] whereas pairwise diversity measures can be calculated beforehand, and evaluating them is less time-consuming and more effective. Second, classifiers can be created and ensembles can be trained along with diversity.[12,22] Third, we need to optimize fusion functions in order to combine classifiers,[17] since, without knowing the best fusion functions, it would be premature to use them for ensemble selection. Given that different fusion functions need to be evaluated, any preselected fusion function might not be optimal for the ensemble selection. According to the "no free lunch" theorem,[36,37] it is understandable that a search algorithm based on one fusion function might not be better than another search algorithm based on a more common objective function. Based on these arguments, we consider ensemble selection and ensemble combination as two different problems, each of which should be solved separately.

Nevertheless, there is no universal definition of diversity, and therefore a number of different diversity measures have been proposed.[1,9,11,14,15,18,21,25,34] What is more, it has been observed that, even with so many different diversity measures,

clear correlations between ensemble accuracy and diversity measures cannot be found,[5,20,21] leading some researchers to consider diversity measures to be unnecessary for ensemble selection.[28] To sum up, the concept of diversity does help, but both theoretical and experimental approaches showing that strong correlations between diversity measures and ensemble accuracy are lacking. Given the challenge of using diversity for ensemble selection, we argue that the lack of correlation between ensemble accuracy and diversity does not imply that there is no direct relationship between them, but that diversity should be taken into account with the performance of individual classifiers. We suggest that such compound diversity functions can give the best correlation with ensemble accuracy. Here are the key questions that need to be addressed:

(1) Diversity is important, but it has only a weak correlation with the ensemble accuracy. Can we combine the diversity with the classifier accuracies to achieve a higher correlation with the ensemble accuracy?
(2) Is there any effect on such a correlation, e.g. from the number of classes or the number of classifiers?
(3) Can the diversity combined with the classifier accuracy be effective for ensemble selection?

To answer these questions, we derive compound diversity functions by combining diversities and the performances of individual classifiers, and we show that, with such functions, there are strong correlations between the diversity measures and ensemble accuracy. Furthermore, we demonstrate the impact on the correlation between the accuracy and the diversity with different ensemble creation methods, with different number of classifiers and with different number of classes. However, the problem of EoC optimization is very complex. In addition to diversity issues, it is also related to fusion functions for classifier combination and to searching algorithms for ensemble selection. The contribution of this paper constitutes only part of an improved understanding of the use of diversity for ensemble selection.

The paper is organized as follows. In the next section, we investigate the dilemma of the lack of correlation between diversity and ensemble accuracy. In Sec. 3, we provide the reason for why the compound diversity functions might work. In Sec. 4, we discuss how the number of classifiers and the number of classes might influence the correlation between ensemble accuracy and compound diversity functions. Section 6 presents basic diversity measures that would be tested in the experiments. Correlations with ensemble accuracy are measured on 0.624 million ensembles in Sec. 6. In Sec. 7, we use the proposed compound functions as objective functions for ensemble selection among 42.24 million ensembles. A discussion and our conclusion are contained in the final sections.

2. Dilemma of the Ambiguity Towards the Ensemble Accuracy

In this section, we adopt the framework established in Ref. 5 to discuss the impediment to using the ambiguity to estimate ensemble accuracy. For readers not familiar

with the work in Refs. 5 and 19, we present a short introduction here, but the original papers offer far more details. The main point is to decompose the mean square error of an ensemble into an ambiguity part and a non-ambiguity part, and we can find the variance terms in both the ambiguity part and the non-ambiguity part. As a result, when we try to maximize the ambiguity among classifiers, we will also affect the non-ambiguity part. That is the reason that an increase in the diversity will not necessarily guarantee a decrease in the global ensemble error.

To start, we need to introduce the concept of the bias-variance decomposition[5–7,10,16] Briefly speaking, attempts to reduce the bias component will cause an increase in variance, and vice versa.

Suppose that the response variable is binary, i.e. $y \in \{0, 1\}$, the probability of a sample x belonging to a class y can be $P(y|x)$, and the classification task is to estimate this probability $E\{y|x\} = P(y|x)$ based on a sequence of the N observations $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. Assume that we have a classifier f trained with a particular dataset D, the probability of a data point x belonging to a class predicted by the classifier f can be written as $f(x, D)$. To measure the effectiveness of the $f(x, D)$ as a predictor of the $E\{y|x\}$, we can simply calculate its mean square error (MSE).[19]

$$E\{(f(x,D) - E\{y|x\})^2\} = (E\{f(x,D)\} - E\{y|x\})^2 + E\{(f(x,D) - E\{f(x,D)\})^2\} \quad (1)$$

$$\text{or} \quad \text{MSE}\{f\} = \text{bias}(f)^2 + \text{var}(f) \quad (2)$$

where $E\{f(x, D)\}$ is the expectation of the classifier $f(x, D)$ with the respect to the training set D, i.e. the average over the ensemble of the possible D. We can deduct that:

$$\text{bias}(f) = E\{f(x,D)\} - E\{y|x\} \quad (3)$$

$$\text{var}(f) = E\{(f(x,D) - E\{f(x,D)\})^2\} \quad (4)$$

This form can be further decomposed into bias-variance-covariance.[5,34] For an ensemble with L classifiers, the averaged bias of the ensemble members is defined as:

$$\bar{b} = \frac{1}{L}\sum_{i}^{L}(E\{f_i(x,D_i)\} - E\{y|x\}) \quad (5)$$

where D_i is the dataset used to train the classifier f_i. We note that $E\{f_i(x, D_i)\}$ is the average over the ensemble of the possible D, and thus all classifiers will have the same $E\{f(x, D)\}$. We just keep the notation for the clarity and for consistency with Ref. 5. Then, the averaged variance of the ensemble members will be:

$$\bar{v} = \frac{1}{L}\sum_{i}^{L}(E\{(f_i(x,D_i) - E\{f_i(x,D_i)\})^2\}) \quad (6)$$

and the averaged covariance of the ensemble members will be:

$$\bar{c} = \frac{1}{L(L-1)} \sum_{i}^{L} \sum_{j \neq i}^{L} E\{(f_i(x, D_i) - E\{f_i(x, D_i)\})(f_j(x, D_j) - E\{f_j(x, D_j)\})\} \quad (7)$$

If we decompose the mean square error for this ensemble of L classifiers, we get:

$$\mathrm{MSE}(L) = E\left\{\left(\left(\frac{1}{L}\sum_{i}^{L} f_i(x, D_i)\right) - E\{y|x\}\right)^2\right\} \quad (8)$$

$$= \bar{b}^2 + \frac{1}{L}\bar{v} + \frac{L-1}{L}\bar{c} \quad (9)$$

To determine the link between $\mathrm{MSE}(L)$ and the ambiguity, which measures the amount of variability among classifier outputs in ensembles, we need to apply ambiguity decomposition. It has been proved[19] that, at a single data point, the quadratic error of the ensemble f_{ens} is guaranteed to be less than or equal to average quadratic error of the individual classifiers[19]:

$$(f_{\mathrm{ens}} - E\{y|x\})^2 = \sum_{i}^{L} w_i (f_i(x, D_i) - E\{y|x\})^2 - \sum_{i}^{L} w_i (f_i(x, D_i) - f_{\mathrm{ens}})^2 \quad (10)$$

where w_i is the weight of classifier $f_i(x, D_i)$ in the ensemble, and $0 \leq w_i \leq 1$. If every classifier $f_i(x, D_i)$ has the same output, then the second term is 0, and f_{ens} would be equal to the average quadratic error of the individual classifiers. Note that the ensemble function is a convex combination ($\sum_i^L w_i = 1$):

$$f_{\mathrm{ens}} = \sum_{i}^{L} w_i f_i(x, D_i) \quad (11)$$

For the $\mathrm{MSE}(L)$ of this ensemble of classifiers, suppose that every classifier has the same weight, i.e. $\forall i, w_i = \frac{1}{L}$, so f_{ens} is merely the average function of all individual classifiers $f_{\mathrm{ens}} = \bar{f}$. Consequently, the ambiguity decomposition can be written as:

$$(\bar{f} - E\{y|x\})^2 = \frac{1}{L}\sum_{i}^{L} (f_i(x, D_i) - E\{y|x\})^2 - \frac{1}{L}\sum_{i}^{L} (f_i(x, D_i) - \bar{f})^2 \quad (12)$$

Note that its expectation is exactly Eqs. (8) and (9):

$$E\left\{\frac{1}{L}\sum_{i}^{L} (f_i(x, D_i) - E\{y|x\})^2 - \frac{1}{L}\sum_{i}^{L} (f_i(x, D_i) - \bar{f})^2\right\}$$

$$= \bar{b}^2 + \frac{1}{L}\bar{v} + \frac{L-1}{L}\bar{c} \quad (13)$$

The ambiguity is the second term on the left-hand side in Eq. (13), and it can be written as[19]:

$$E\left\{\left(\frac{1}{L}\sum_{i}^{L}(f_i(x,D_i)-\bar{f})^2\right)\right\}$$

$$=\frac{1}{L}\sum_{i}^{L}E\{(f_i(x,D_i)-E\{f_i(x,D_i)\})^2\}-E\{(\bar{f}-E(\bar{f}))^2\} \tag{14}$$

$$=\bar{v}-\mathrm{var}(\bar{f})=\bar{v}-\frac{1}{L}\bar{v}-\frac{L-1}{L}\bar{c} \tag{15}$$

The first term on the left-side in Eq. (13) is the sum of averaged bias and averaged variance of classifiers:

$$E\left\{\frac{1}{L}\sum_{i}^{L}(f_i(x,D_i)-E\{y|x\})^2\right\}=\bar{b}^2+\bar{v} \tag{16}$$

As stated in Ref. 5, the term $\bar{v}$, the average variance, exists in both the ambiguity part and the non-ambiguity part of MSE(L). This means that we cannot simply maximize the ambiguity without affecting the bias component of MSE(L). When we try to maximize the ambiguity among classifiers, we actually maximize the difference between its variance $\bar{v}$ and its covariance $\bar{c}$. If the term $\bar{v}$ increases, the non-ambiguity part of MSE(L) will increase too. This is why, in general, an increase in the diversity measure will not necessarily guarantee a decrease in the global ensemble error. We need to mention that the above discussion is with respect to a single data point, but the results can generalize to the full space.[5]

3. Proposed Compound Diversity Functions

The above section shows that MSE(L) can be decomposed into an ambiguity part and a non-ambiguity part, and because the variance terms exist in both parts, there is no easy solution to minimize MSE(L) by simply maximizing the ambiguity. In this section, however, we will show that in some certain circumstances, MSE(L) can have another form of the decomposition. Based on this decomposition, we propose an indirect approximation of MSE(L) with only the average errors of individual classifiers and the diversities of classifier-pairs. The proposed approximation might thus help reduce MSE(L) for the ensemble selection. First, suppose that we have an ensemble with only 2 classifiers $f_i(D_i), f_j(D_j)$, and that classifiers $f_i(D_i)$ and $f_j(D_j)$ have the recognition rates a_i and a_j on a data set X, respectively, and the average error of classifier $f_i(D_i)$ is $(1-a_i)$, and the average error of classifier $f_j(D_j)$ is $(1-a_j)$ and the diversity d_{ij} is measured between them. With only two classifiers, we get $L=2$ in Eqs. (6) and (7). As a result, at any data point $x \in X$, the ambiguity between $f_i(x,D_i)$ and $f_j(x,D_j)$ is exactly half of the difference between

their variance and covariance in Eq. (15):

$$\begin{aligned}\text{amb}_{ij} &= \frac{1}{2}(\bar{v} - \bar{c}) \\ &= \frac{1}{4}(E\{(f_i(x, D_i) - E\{f_i(x, D_i)\})^2\} + E\{(f_j(x, D_j) - E\{f_j(x, D_j)\})^2\} \\ &\quad - 2 \cdot E\{(f_i(x, D_i) - E\{f_i(x, D_i)\}) \cdot (f_j(x, D_j) - E\{f_j(x, D_j)\})\}) \quad (17)\end{aligned}$$

If we use $L = 2$ in Eq. (9) and replace $\frac{1}{2}(\bar{v} - \bar{c})$ by amb_{ij}, we can write MSE(2) as:

$$\text{MSE}(2) = \bar{b}^2 + \frac{1}{2}(\bar{v} + \bar{c}) = \text{amb}_{ij} + \bar{b}^2 + \bar{c} \quad (18)$$

As a result of this decomposition, there are basically two MSE(2) terms, the first being the ambiguity of the ensemble, and the second being the sum of the averaged covariance and the averaged bias of individual classifiers. Using Eq. (17), we can write the above equation as:

$$\text{MSE}(2) = \bar{b}^2 + \bar{v} - \frac{1}{2}(\bar{v} - \bar{c}) = \bar{b}^2 + \bar{v} - \text{amb}_{ij} \quad (19)$$

where $\text{amb}_{ij} = \frac{1}{2}(\bar{v} - \bar{c})$. The point is that we have the term $\bar{b}^2 + \bar{v}$ instead of $\bar{b}^2 + \bar{c}$, and one way to approximate the $\bar{b}^2 + \bar{v}$ of the ensemble is through the $\text{var}(f) + \text{bias}(f)^2$ of each individual classifier f, which is exactly the MSE of each individual classifier. Despite this, we do not have its exact value of the $\text{var}(f) + \text{bias}(f)^2$ of the classifier f at each data point. However, we have the average of its zero-one loss error[7] on the whole data set X, i.e. $(1 - a_i)$. The behavior of a zero-one loss error is much more complicated, and up to now there has simply been no clear analog of the bias-variance-covariance decomposition when we have a zero-one loss function.[5,7] Nevertheless, it is still reasonable to assume that the larger the MSE of a classifier at each data point x, the larger its average zero-one loss error on the whole data set X should be. We need to draw some assumptions to get the reasonable approximation here. First, we want to approximate the value of $\bar{b}^2 + \bar{v}$ in Eq. (19), but what we know is the average error rate $(1 - a_i)$ of any given classifier f_i. So suppose that:

(1) For any classifier f_i, $(1 - a_i) \approx \alpha_i(\text{var}(f_i) + \text{bias}(f_i)^2)$.
(2) All classifiers in the ensemble have similar MSE(f).

The first assumption gives that $(1-a_i) \approx \alpha_i(\text{var}(f_i) + \text{bias}(f_i)^2)$ for f_i and $(1-a_j) \approx \alpha_j(\text{var}(f_j) + \text{bias}(f_j)^2)$ for f_j. Still, owing to the lack of exact values for α_i and α_j, there is no easy solution to the approximation of the sum of averaged bias and averaged variance. But, if the second assumption stands, i.e. these individual classifiers have a similar MSE(f), and one could obtain a reasonable approximation of $(\bar{b}^2 + \bar{v})$ by calculating the geometric mean of individual classifier's $(\text{var}(f) +$

$\text{bias}(f)^2$). As a result, the term $\bar{b}^2 + \bar{v}$ might be approximated by the error rates of individual classifiers based on the above assumptions:

$$(\bar{b}^2 + \bar{v}) \approx \gamma((1 - a_i) \cdot (1 - a_j))^{\frac{1}{2}} \tag{20}$$

Now, we want to approximate the value of the ambiguity amb_{ij} in Eq. (19) with the diversity measures. Again, we need to suppose that:

- The diversity measures represent approximations of the ambiguity among classifiers, i.e. $d_{ij} \propto \text{amb}_{ij}$, $0 \leq d_{ij} \leq 1$.

Using the assumption, the term d_{ij} has a high correlation with $\text{amb}_{ij} = \frac{1}{2}(\bar{v} - \bar{c})$, and the approximation of $\frac{1}{2}(\bar{v} - \bar{c})$ can be written as:

$$\text{amb}_{ij} \approx \delta \cdot d_{ij} \tag{21}$$

For an approximation to MSE(2), i.e. $\bar{b}^2 + \bar{v} - \text{amb}_{ij}$, given the approximation $(\bar{b}^2 + \bar{v})$ as $\gamma \cdot ((1 - a_i) \cdot (1 - a_j))^{\frac{1}{2}}$, and the approximation of their diversity $(\bar{v} - \bar{c})$ as $\delta \cdot d_{ij}$, we could not achieve any exact solution due to the lack of values γ and δ. Again, we need to make some assumptions to go further:

- The ambiguity term and the non-ambiguity term have similar weights in MSE(2).

Based on this assumption, the value MSE(2) can be approximated as the product of the error rates of each classifier and their pairwise diversity. Given $0 \leq d_{ij} \leq 1$, we have $0 \leq 1 - d_{ij} \leq 1$, and we define an index for the approximation of MSE(2) as:

$$\widetilde{\text{MSE}_{ij}} \equiv (1 - d_{ij}) \cdot ((1 - a_i) \cdot (1 - a_j))^{\frac{1}{2}} \tag{22}$$

For multiple classifiers, the direct approximation of MSE(L) is much more complex and its term of covariance cannot easily be substituted. Still, we can regard multiple classifiers as a network of classifier-pairs, and we might use the average error of each individual classifier and the diversity between each classifier-pair for an indirect approximation of MSE(L). Given the number of selected classifiers $L \geq 2$, and we have $\widetilde{\text{MSE}(L)} \approx (\prod_{i=1}^{L}(1 - a_i))^{\frac{1}{L}}(\prod_{i,j=1,i\neq j}^{L}(1 - d_{i,j}))^{\frac{1}{L\times(L-1)}}$. By calculating their product, we can get an approximation of ensemble accuracy without any consideration for the type of fusion functions. It is important to note that different diversity measures are supposed to have different sorts of relationships with ensemble accuracy. Some diversity measures calculate the ambiguity among classifiers, where positive correlation with ensemble accuracy is expected; others actually measure the similarity among classifiers, where there would be a negative correlation between them and ensemble accuracy. In the case where the diversity measures represent the ambiguity, we combine the diversity measures with the error rates of each individual classifier:

$$\widehat{\text{div}_{\text{amb}}} = \left(\prod_{i=1}^{L}(1 - a_i)\right)^{\frac{1}{L}} \left(\prod_{i,j=1,i\neq j}^{L}(1 - d_{i,j})\right)^{\frac{1}{L\times(L-1)}} \tag{23}$$

where a_i is the correct classification rate of classifier f_i, and $d_{i,j}$ is the measured diversity between classifier f_i and classifier f_j. Apparently we have $\frac{L\times(L-1)}{2}$

diversity measures on different classifier-pairs. Here, $1 - a_i$ is the error rate of classifier f_i, and $(1 - d_{i,j})$ can be interpreted as the similarity between classifier f_i and classifier f_j. Thus, $\widehat{\text{div}_{\text{amb}}}$ is, in fact, an estimation of the likelihood of a common error being made by all classifiers. In other words, we expect $\widehat{\text{div}_{\text{amb}}}$ to have positive correlation with ensemble accuracy. However, if the diversity measures represent the similarity, the proposed compound diversity function should be:

$$\widehat{\text{div}_{\text{sim}}} = \left(\prod_{i=1}^{L}(1 - a_i)\right)^{\frac{1}{L}} \left(\prod_{i,j=1,i\neq j}^{L} d_{i,j}\right)^{\frac{1}{L\times(L-1)}} \tag{24}$$

where $d_{i,j}$ should be interpreted as the similarity between f_i and f_j in this case. So, $\widehat{\text{div}_{\text{sim}}}$ ought to mean the likelihood of a common error being made by all the classifiers. We expect negative correlation between the $\widehat{\text{div}_{\text{sim}}}$ and ensemble accuracy. While it is true that these approximations lead to strong correlations with MSE(L) for a fixed number of classifiers L, the bottom line is that the ensemble selection will result in the minimization of L for the proposed compound diversity function, if L is set as a free parameter. This is substantiated below:

Suppose that there are a total of M classifiers in the pool, and we intend to select a subset of L classifiers, $L \leq M$, which can construct an EoC with the best accuracy by a simple majority voting rule.[27,28,30] For the pairwise diversity measures, suppose that for all classifiers $f_1 \sim f_M$, we measure the diversity d_{ij} on $\frac{M(M-1)}{2}$ classifier-pairs $c_{ij}, 1 \leq i, j \leq M, i \neq j$. Intuitively, there exists at least one classifier-pair $\widehat{c_{ij}}$ with the maximum pairwise diversity $\widehat{d_{ij}}$ that is larger than or equal to any pairwise diversity of other classifier-pairs d_{ij}, for $1 \leq i, j \leq M, i \neq j$. As a consequence, the maximum pairwise diversity $\widehat{d_{ij}}$ of classifier-pair $\widehat{c_{ij}}$ is larger than the diversities of any other selected L classifiers, given that $2 \leq L \leq M$:

$$\forall L, \quad \widehat{d_{ij}} \geq E\{d_{ij}\} = d_L \tag{25}$$

where $E\{d_{ij}\}$ is the mean of the pairwise diversities of L selected classifiers. This means that if we use pairwise diversity as an objective function for ensemble selection, and if the number of classifiers is set as a free parameter, it is quite possible that we will get only one classifier-pair. The proposed compound functions are based on diversity measured in a pairwise manner, even taking into account the individual classifiers' error rates, ensembles with fewer classifiers are more likely to be favored in the ensemble selection. With regard to this effect, functions with various number of classifiers shall be rescaled by [a]:

$$\widehat{\text{div}_{\text{amb}}} = \frac{L}{L-1}\left(\prod_{i=1}^{L}(1 - a_i)\right)^{\frac{1}{L}} \left(\prod_{i,j=1,i\neq j}^{L}(1 - d_{i,j})\right)^{\frac{1}{L\times(L-1)}} \tag{26}$$

[a]In practice, when L is large, it is possible that we need to multiply a coefficient η on the compound diversity functions, so that the lower bound of evaluated compound diversity values will not exceed machine capacity and precision.

$$\widehat{\mathrm{div}_{\mathrm{sim}}} = \frac{L}{L-1}\left(\prod_{i=1}^{L}(1-a_i)\right)^{\frac{1}{L}}\left(\prod_{i,j=1,i\neq j}^{L} d_{i,j}\right)^{\frac{1}{L\times(L-1)}} \tag{27}$$

Thus, to combine a specific diversity measure with the error rates of each individual classifiers, one must select between both compound diversity functions. $\widehat{\mathrm{div}_{\mathrm{amb}}}$ must be used when diversity measure is based on the ambiguity among classifiers, where positive correlation with ensemble accuracy is expected, while $\widehat{\mathrm{div}_{\mathrm{sim}}}$ must be used when the diversity is based on the similarity among classifiers, where there would be a negative correlation between them and ensemble accuracy.

4. Concern about the Number of Classes and the Number of Classifiers

The measures $\widehat{\mathrm{div}_{\mathrm{sim}}}$ and $\widehat{\mathrm{div}_{\mathrm{amb}}}$ are supposed to have a strong correlation with the MSE of the ensemble, but this MSE never reaches 100% correlation with ensemble error, for several reasons: First, the ensemble error is a zero-one loss error, while the MSE of the ensemble is based on bias, variance and covariance terms. Second, ensemble error is influenced by the way classifiers are combined, i.e. by the choice of fusion functions, while the MSE of the ensemble does not take fusion functions into consideration when combining ensembles. Third, ensemble error is involved in more complicated situations and is related to other concerns, such as the number of classes and the number of classifiers (see the following discussion). For these reasons then, it is not hard to see why $\widehat{\mathrm{div}_{\mathrm{sim}}}$ and $\widehat{\mathrm{div}_{\mathrm{amb}}}$ will not be perfectly correlated with the ensemble error. However, we need to know more about what its limitations are.

Given the complexity of the problem of ensemble selection, and the various *ad hoc* methods for combining classifiers, it is impossible at this stage to create a flawless and complete framework for understanding the limitations of the estimation of ensemble accuracy with compound diversity functions. With this in mind, we set up some preconditions for a special case study as the first step towards gaining these understandings. We suppose that each classifier produces labels of samples as outputs, and we need to fix a fusion function for combining classifiers in an ensemble in our case study. A number of different fusion functions can be used,[17] but for, simplicity and effectiveness,[28] suppose that a simple majority voting rule[27,28,30] constitutes the fusion function of ensemble outputs. Based on these conditions, we wish to know whether or not:

(1) Given an ensemble of classifiers, is it possible that some classifiers make more (or less) error without changing the ensemble outputs?
(2) Given an ensemble of classifiers, is it possible that some classifier-pairs have greater (or less) diversity without changing the ensemble outputs?
(3) If the above two concerns are true, how different can they be while maintaining the same ensemble outputs?

It is not hard to answer the first two questions. When a simple majority voting rule is used, a correct ensemble output depends on the proportion of classifiers correctly classifying this sample. For a sample x in a T-class problem, suppose that the correct class is $i, 1 \leq i \leq T$. The ensemble will give correct output only under the condition $\forall j,\ c(i)_T > c(j)_T$, for $1 \leq i, j \leq T, i \neq j$, where $c(i)_T$ is the number of classifiers making a decision on class i, and $c(j)_T$ is the number of classifiers making a wrong decision on another class j, in a T-class problem. Under the condition $\forall j, c(i)_T > c(j)_T$, the $c(i)_T$ can decrease, and the $c(j)_T$ can increase, and the ensemble can still give the correct output.

A similar reasoning can apply to diversities, because the change in the error rates of each individual classifier will eventually affect the diversities among them. It is apparent that the different error rates of individual classifiers and the different diversities among them can achieve the same ensemble outputs by a simple majority voting rule. We know that there is an unavoidable systematic estimation bias on the correlation measurement with ensemble accuracy for this fusion function. In fact, since this problem results from classifiers combining by a simple majority voting rule, and not from a particular ensemble selection criterion, the effect will occur for any objective functions on ensemble selection.

The third question depends on the nature of the pattern recognition problems and cannot be easily estimated. It is impossible to say in what way this estimation bias will affect the correlation between compound diversity functions and ensemble accuracy. But among those problems are two elements resulting in this estimation bias on correlation measurements between $\widehat{\text{div}_{\text{sim}}}/\widehat{\text{div}_{\text{amb}}}$ and ensemble accuracy:

(1) the number of classes of the problem,
(2) the number of classifiers selected from the pool to construct the ensemble.

As we mentioned before, an ensemble can maintain the same outputs under the condition that $\forall j, c(i)_T \geq c(j)_T$. For a given sample in a T-class problem, suppose that the ensemble output remains the same. We define a margin $m(T), m(T) \geq 0$ to be the number of correct classifiers exceeding the threshold of being majority[13,24,29]:

$$m(T) = c(i)_T - \rho(T) \tag{28}$$

where $\rho(T)$ is the threshold of the majority voting in a T-class problem. Usually $\rho(T)$ represents the second most popular vote[13]:

$$\rho(T) = \max c(j)_T, 1 \leq j \leq T, j \neq i \tag{29}$$

Intuitively, given that the output of the ensemble remains unchanged, we still have:

$$c(i)_T \geq \rho(T), 1 \leq i \leq T \tag{30}$$

Given that all classifiers have choices on T classes, we can expect both $c(i)_T$ and $\rho(T)$ to decrease when T increases. The larger the number of classes is, the fewer votes are obtained for each class.

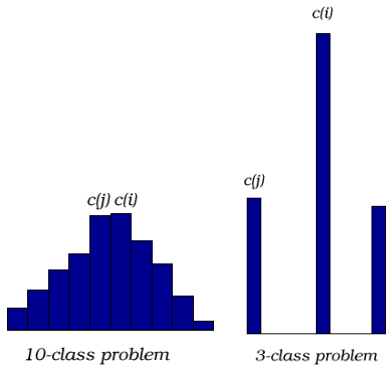

Fig. 1. Distribution of 100 votes in ensembles: (a) 10-class problem; (b) 3-class problem.

As we can see in Fig. 1, for a 10-class problem, class i received the majority vote, but the margin $m(10)$ with the second most popular voted class j is very small. This means that the ensemble can change its decision with several different votes, therefore the measured error rates and diversities are more accurate in estimating ensemble accuracy. By contrast, for a 3-class problem, the margin $m(3)$ between $c(i)_3$ and $c(j)_3$ is huge, which means that more classifiers are allowed to change their individual outputs while the ensemble can still maintain the same outputs. In this case, the estimation will be much worse and the correlation with ensemble accuracy will have deteriorated. The margin $m(T)$ is proportional to this estimation bias. For the problem with T classes, given L classifiers, then we can define the margin $m(T)$ as:

$$m(T) = L \cdot (P(c(i)_T|t(x) = i) - P(c(j)_T|t(x) = i)) \tag{31}$$

Thus, we note that it is also proportional to the number of classifiers of ensemble L. This indicates that the estimation bias in the correlation measurement between ensemble accuracy and $\widehat{\text{div}_{\text{sim}}}/\widehat{\text{div}_{\text{amb}}}$ will become larger when more classifiers are used. This estimation bias results directly from the nature of a zero-one loss error, and from the simple majority voting rule for combining classifiers. No matter which objective function for ensemble selection is used, we will encounter a loss of correlation with ensemble accuracy. The influence of the number of classes affects not only the margin of the majority voting, but also the sensitivity of the whole voting network as well, especially in the measure of diversity. Figure 2(a) shows that, on an ensemble of 7-classifiers, there are two groups of classifiers with different opinions in a 2-class problem ($C1 \sim C4$, and $C5 \sim C7$), and the majority voting rule needs at least four votes from classifiers for a decision to be made. By contrast, in a 6-class problem, the majority could be represented with only two votes [Fig. 2(b)], we have six groups with different outputs ($C1$ agrees with $C2$, but $C3$, $C4$, $C5$ and $C6$ all differ from one another). Note that we have the same margin of one vote in both cases. If we consider the majority class shifting into another class, six pairwise diversities have to be modified in 2-class problems (i.e. if $C4$ agrees with

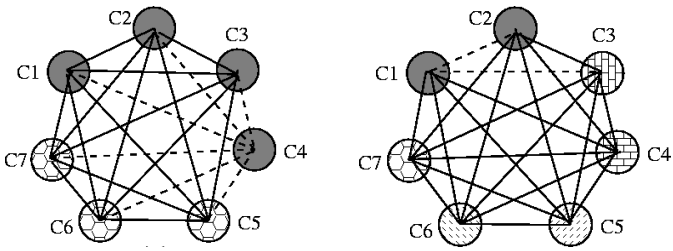

Fig. 2. An ensemble of seven classifiers (C1 $\sim$ C7); the shadowed circles represent the classifiers needed to achieve the majority, the solid lines the pairwise diversities among classifiers, and the dashed lines the required modified-pairwise-diversities if the majority is shifted into another class: (a) at least four votes needed in 2-class problems; six modified pairwise-diversities needed for majority-shifting; (b) at least two votes needed in 6-class problems; two modified pairwise-diversities needed for majority-shifting.

$C5$, $C6$, $C7$, diversities must change between $C4$ and all other classifiers); and only two pairwise diversities need to be modified in 6-class problems (i.e. if $C1$ agrees with $C3$, diversities change only between $C1$ and $C3$, $C1$ and $C2$). This indicates that a large number of diversity changes in low-class problems may not affect the final output, but in high-class problems a slight change in diversity may lead to another final decision. Thus, the measure $\widehat{\text{div}_{\text{sim}}}/\widehat{\text{div}_{\text{amb}}}$ is much more sensitive to ensemble behavior in high-class problems than it is in low-class problems.

This suggests that the implementation of proposed compound diversity functions should be much more effective dealing with high-class problems. Moreover, the fewer classifiers are selected in an ensemble, the more accurate the correlation between ensemble accuracy and compound diversity functions shall be.

5. Diversity Measures

Before we carried out the correlation measurements, we need to introduce some diversity measures that would be evaluated in our experiments. The traditional concept of diversity is composed of the terms of correct/incorrect classifier outputs. By comparing these correct/incorrect outputs among classifiers, their respective diversity can be calculated. In general, there are two kinds of diversity measures:

(1) Pairwise diversity measures
Diversity is measured between two classifiers. In the case of multiple classifiers, diversity is measured on all possible classifier-pairs, and global diversity is calculated as the average of the diversities on all classifier-pairs. That is, given L classifiers, $\frac{L\times(L-1)}{2}$ pairwise diversities $d_{12}, d_{13}, \ldots, d_{(L-1)L}$ will be calculated, and the final diversity $\bar{d}$ will be its average[21]:

$$\bar{d} = 2 \times \frac{\sum_{ij} d_{ij}}{L \times (L-1)}, i \leq j \tag{32}$$

This type of diversity includes: Q-statistics,[1,3] the correlation coefficient,[21] the disagreement measure[15] and the double fault.[11]

(2) Non-Pairwise diversity measures
There are others diversities that are not pairwise, i.e. they are not calculated by comparing classifier-pairs, but by comparing all classifiers directly. This type of diversity includes: the Entropy measure,[21] Kohavi–Wolpert variance,[18] the measurement of interrater agreement,[3,9] the measure of difficulty,[14] generalized diversity[25] and coincident failure diversity.[25]

Most research suggests that neither type of diversity is capable of achieving a high degree of correlation with ensemble accuracy, as only very weak correlation can be observed.[21] As we see in Sec. 3, the proposed compound diversity functions might represent better correlations with the ensemble accuracy. To verify its usefulness, we carried out the experiments of the correlation measurements in the next section.

6. Correlations between Diversity and Ensemble Accuracy

To make sure that the normalized compound diversity function is valid for the estimation of ensemble accuracy, we tested it on problems extracted from UCI machine learning repository. There are several requirements for the selection of pattern recognition problems. First, we should test three types of ensemble creation method: Random Subspaces, Bagging and Boosting. Thus, the databases must have a large feature dimension for Random Subspaces. Second, to avoid the dimensional curse during training, each database must have sufficient samples of its feature dimension. Third, to avoid identical samples being trained in Random Subspaces, only databases without symbolic features are used. Fourth, to simplify the problem, we do not use databases with missing features. In accordance with the requirements listed above, we carried out our experiments on four databases selected from the UCI data repository (see Table 1).

For each of four databases, for each of three ensemble creation methods (Random Subspaces, Bagging and Boosting), and for each of three classification algorithms, 18 classifiers were generated as the pool for base classifiers. Classifiers were then selected from this pool to construct ensembles. The three different classification

Table 1. UCI data for ensembles of classifiers.

Database	Classes	Train	Test	Features	Random Subspace	Bagging	Boosting
Wisconsin Breast-Cancer	2	284	284	30	5	66%	66%
Satellite	6	4435	2000	36	4	66%	66%
Image Segmentation	7	210	2100	19	4	66%	66%
Letter Recognition	26	10007	9993	16	12	66%	66%

algorithms used in our experiments are Naive Bayesian Classifiers (NBC), Quadratic Discriminant Classifiers (QDC), and 5-Layer Neural Network Classifiers (NNC) with Back-Propagation.[8] To better understand the influence of the number of classifiers on the correlation between diversity and ensemble accuracy, ensembles were composed from $3 \sim 15$ classifiers. In total, we evaluated 13 different numbers of classifiers for ensembles. All correlations are measured for ensembles with the same number of classifiers, then the mean values of correlations from different numbers of classifiers are calculated. To obtain the most accurate measure, 50 ensembles were constructed with the same number of selected classifiers for each database, for each classification algorithm, for each ensemble method and for each different number of classifiers. We repeated this process 30 times to obtain a reliable evaluation. The simple majority voting rule is used as the fusion function for the evaluation of the global performances of related EoC. A total of $3 \times 3 \times 4 \times 13 \times 50 \times 30 = 0.702$ million ensembles should be evaluated. But, due to the dimensional curse, NNC did not have sufficient samples for training on the Image Segmentation problem or on the Satellite problem for Bagging or for Boosting. This occurred on $1 \times 2 \times 2 \times 13 \times 50 \times 30 = 0.078$ million ensembles, so in total $0.702 - 0.078 = 0.624$ million ensembles were evaluated in the experiment.

We measured ensemble accuracy correlation on ten traditional diversity measures, including the disagreement measure (DM),[15] the double-fault (DF),[11] Kohavi–Wolpert variance (KW),[18] the interrater agreement (INT),[9] the entropy measure (EN),[21] the difficulty measure (DIFF),[14] generalized diversity (GD),[25] coincident failure diversity (CFD),[25] Q-statistics (Q),[1] and the correlation coefficient (COR),[21] as well as on ten respective proposed compound diversity functions [Eqs. (26) and (27)]. They are also compared with the Mean Classifier Error (ME) of individual classifiers. On all training databases, the proportion of selected samples in Bagging and Boosting is 66%. For Random Subspaces, the sizes of subsets of features are decided under the condition that each classifier created must have recognition rates more than 50%.

6.1. *Random subspaces*

In Table 2, we show the correlations between original diversity measures and ensemble accuracy, and the correlation between compound diversity functions and ensemble accuracy. NBC, QDC, and NNC are applied on all databases, and we show their average correlations.

First, we observe that in most cases the ME has an apparent correlation with ensemble accuracy. Furthermore, it shows that, in general, compound diversity functions give better results than the original diversity measures; it can also be perceived that, even though the correlation between ME and ensemble accuracy is weak, compound diversity functions still work well and present stronger correlations with ensemble accuracy than ME. Of all the diversity measures, Q, COR, INT and DIFF are not stable. By contrast, DM, DF, KW, EN, GD and CFD are

Table 2. Correlation for the Random Subspaces method between ensemble accuracy and: (a) Mean Classifier Error; (b) the average of pure diversity measures; (c) the proposed compound diversity functions. The arrows indicate the expected direction of the correlations: ↓ for −1 and ↑ for 1.

	Breast Cancer	Satellite	Image Segmentation	Letter Recognition
ME (Mean Classifier Error) (↓)	−0.4447	−0.5820	−0.6147	−0.4680
Original Diversity Measures				
Disagreement measure (DM) (↑)	−0.0170	0.0779	−0.1860	−0.0577
Double fault (DF) (↓)	−0.3916	−0.1204	−0.4725	−0.3758
Kohavi-Wolpert variance (KW) (↑)	−0.0170	0.0779	−0.1860	−0.0577
Interrater agreement (INT) (↓)	−0.3605	−0.0791	−0.0038	−0.0283
Entropy measure (EN) (↑)	−0.0170	0.0779	−0.1860	−0.0577
Measure of difficulty (DIFF) (↓)	0.2440	−0.1263	0.5518	0.1364
Generalized diversity (GD) (↑)	0.2893	0.0819	0.3547	0.1413
Coincident failure diversity (CFD) (↑)	0.2990	0.0807	0.3603	0.1526
Q-statistics (Q) (↓)	−0.1705	−0.0811	0.1140	0.0460
Correlation coefficient (COR) (↓)	−0.3552	−0.0792	0.0120	−0.0266
Proposed Compound Diversity Functions				
Disagreement measure (DM) (↓)	−0.6379	−0.4563	−0.4310	−0.4449
Double fault (DF) (↓)	−0.4924	−0.4731	−0.5058	−0.4916
Kohavi-Wolpert variance (KW) (↓)	−0.5407	−0.5337	−0.7616	−0.5014
Interrater agreement (INT) (↓)	−0.2416	−0.0462	−0.1010	−0.1496
Entropy measure (EN) (↓)	−0.6379	−0.4563	−0.4310	−0.4449
Measure of difficulty (DIFF) (↓)	−0.3292	−0.2877	0.0708	−0.1200
Generalized diversity (GD) (↓)	−0.4551	−0.4978	−0.5951	−0.4851
Coincident failure diversity (CFD) (↓)	−0.4264	−0.4561	−0.5292	−0.4490
Q-statistics (Q) (↓)	−0.3362	−0.2355	−0.1224	−0.4410
Correlation coefficient (COR) (↓)	−0.2488	−0.0468	−0.0998	−0.1498

quite reliable, as they always offer 43%–76% of correlation with compound diversity functions. Note that in some cases (e.g. Wisconsin breast cancer), their correlation with ensemble accuracy is better than the correlation between ME and ensemble accuracy.

6.2. *Bagging*

The ensembles for the second experiment were created by Bagging. NBC and QDC are used on all the databases. But NNC is implemented on all of them except the Image Segmentation data and the Satellite data, given insufficient samples, because their high feature dimension caused the dimensional curse.

In Table 3, there is a clear correlation between ME and ensemble accuracy, and it is quite strong. Of all the diversities, Q, COR, INT, and DIFF did not perform as well as the others. The GD and CFD results are unstable; sometimes giving good correlation but sometimes not. DM, KW and EN are stable, though a little bit weaker than those in Random Subspaces. Since the selected databases have high feature dimension for the implementation of Random Subspaces, as a result,

Table 3. Correlation for Bagging between ensemble accuracy and: (a) Mean Classifier Error; (b) the average of pure diversity measures; (c) the proposed compound diversity functions. The arrows indicate the expected direction of the correlations: ↓ for −1 and ↑ for 1.

	Breast Cancer	Satellite	Image Segmentation	Letter Recognition
ME (Mean Classifier Error) (↓)	−0.5516	−0.5151	−0.8113	−0.5906
Original Diversity Measures				
Disagreement measure (DM) (↑)	−0.2902	0.1309	−0.2306	0.1771
Double fault (DF) (↓)	−0.0409	−0.2131	−0.3520	−0.2603
Kohavi-Wolpert variance (KW) (↑)	−0.2902	0.1309	−0.2306	0.1771
Interrater agreement (INT) (↓)	−0.0219	−0.1356	0.2298	−0.1340
Entropy measure (EN) (↑)	−0.2902	0.1309	−0.2306	0.1771
Measure of difficulty (DIFF) (↓)	0.4925	−0.2024	−0.3516	0.0224
Generalized diversity (GD) (↑)	−0.1122	0.1313	−0.2273	0.2149
Coincident failure diversity (CFD) (↑)	−0.1178	0.1314	−0.2321	0.2150
Q-statistics (Q) (↓)	0.1068	−0.1283	−0.1692	0.0570
Correlation coefficient (COR) (↓)	−0.0058	−0.1386	−0.1686	−0.1309
Proposed Compound Diversity Functions				
Disagreement measure (DM) (↓)	−0.5269	−0.3689	−0.3700	−0.5656
Double fault (DF) (↓)	−0.3370	−0.4798	−0.6645	−0.5663
Kohavi-Wolpert variance (KW) (↓)	−0.5431	−0.4384	−0.8329	−0.6005
Interrater agreement (INT) (↓)	−0.2086	−0.1798	−0.0050	−0.1443
Entropy measure (EN) (↓)	−0.5269	−0.3689	−0.3700	−0.5656
Measure of difficulty (DIFF) (↓)	−0.2359	−0.3978	−0.3873	−0.3256
Generalized diversity (GD) (↓)	−0.3331	−0.3962	−0.6721	−0.4922
Coincident failure diversity (CFD) (↓)	−0.2864	−0.3672	−0.3683	−0.4702
Q-statistics (Q) (↓)	−0.5094	−0.4559	−0.1190	−0.4109
Correlation coefficient (COR) (↓)	−0.2014	−0.1867	−0.0846	−0.1450

the effect of the dimensional curse might occur for Bagging and for Boosting. KW always performed at $43\% \sim 83\%$ on our compound diversity function.

We note that, in general, the correlations between the diversities and ensemble accuracy for Bagging are weaker than those for Random Subspaces. But, on high-dimension-class problems, (e.g. letter recognition data, image segmentation), the implementation of compound diversity functions is just as good for Bagging as for Random Subspaces. The advantage of compound diversity functions over the original diversity measures can be perceived in this case.

6.3. *Boosting*

The ensembles were created for the third experiment by Boosting, NBC and QDC are used on all databases, but NNC is used on all except the Image Segmentation data and the Satellite data, because, given insufficient samples, their high feature dimension caused the dimensional curse.

On most of the databases, there is a strong correlation between ME and ensemble accuracy (Table 4). Interestingly, it is in Boosting that we see how the implementation of diversity really matters: the correlation by the proposed compound

Table 4. Correlation for Boosting between ensemble accuracy and: (a) Mean Classifier Error; (b) the average of pure diversity measures; (c) the proposed compound diversity functions.The arrows indicate the expected direction of the correlations: ↓ for −1 and ↑ for 1.

	Breast Cancer	Satellite	Image Segmentation	Letter Recognition
ME (Mean Classifier Error) (↓)	−0.4828	−0.5173	−0.3405	−0.6148
Original Diversity Measures				
Disagreement measure (DM) (↑)	−0.1392	−0.2849	−0.2370	0.4086
Double fault (DF) (↓)	−0.0047	0.3131	0.2549	−0.3408
Kohavi-Wolpert variance (KW) (↑)	−0.1392	−0.2849	−0.2370	0.4086
Interrater agreement (INT) (↓)	−0.0538	0.1283	−0.1497	−0.3926
Entropy measure (EN) (↑)	−0.1392	−0.2849	−0.2370	0.4086
Measure of difficulty (DIFF) (↓)	0.3652	0.3505	0.2647	−0.1940
Generalized diversity (GD) (↑)	−0.0576	−0.2949	−0.2410	0.4092
Coincident failure diversity (CFD) (↑)	−0.0558	−0.3115	−0.2436	0.4109
Q-statistics (Q) (↓)	0.0873	0.1923	0.0471	−0.2980
Correlation coefficient (COR) (↓)	−0.0638	0.1293	−0.1498	−0.3912
Proposed Compound Diversity Measures				
Disagreement measure (DM) (↓)	−0.5599	−0.1080	−0.0219	−0.5410
Double fault (DF) (↓)	−0.3878	−0.0462	0.0364	−0.5351
Kohavi-Wolpert variance (KW) (↓)	−0.5487	−0.4489	−0.3708	−0.5681
Interrater agreement (INT) (↓)	−0.1807	0.0607	−0.0275	−0.3129
Entropy measure (EN) (↓)	−0.5599	−0.1080	−0.0219	−0.5410
Measure of difficulty (DIFF) (↓)	−0.2825	0.0729	0.0854	−0.4388
Generalized diversity (GD) (↓)	−0.3459	−0.2538	−0.1226	−0.5226
Coincident failure diversity (CFD) (↓)	−0.3182	−0.0660	−0.0008	−0.4693
Q-statistics (Q) (↓)	−0.5448	−0.1134	−0.0299	−0.3180
Correlation coefficient (COR) (↓)	−0.1980	0.0611	−0.0272	−0.3130

diversity function could be equivalent to or better than that of ME, which means that, for Boosting, the notion of diversity does help to obtain a strong correlation with ensemble accuracy. Nevertheless, we also perceive that the correlations between the diversities and ensemble accuracy are weaker for Boosting than those for Bagging and for Random Subspaces for low-dimension-class problems. But, when the number of classes is large (e.g. letter recognition data), the correlation on Boosting can be as good as that on Bagging, and the notion of diversity is quite well with compound diversity functions. In high-class-problems, the useful diversity measures appear to be DM, DF, KW, EN, DIFF, GD and CFD. They offer correlations between 46%–56%.

6.4. *Discussion on the correlation between diversity and ensemble accuracy*

In all three ensemble creation methods, we first note that the proposed compound diversity functions correlate much stronger with the ensemble accuracy than the traditional diversity measures. Second, comparison of the various ensemble creation methods suggests that, in Random Subspaces, the proposed compound diversity functions generally have the strongest correlations with ensemble accuracy, better

than in Bagging or in Boosting. Nevertheless, considering the correlation with ensemble accuracy, compound diversity functions could perform better than ME in Boosting. This suggests that the issue of ensemble diversity is crucial in Boosting.

It is certain that the number of classifiers has an impact on the correlation between compound diversity functions and ensemble accuracy. We found the strongest correlation with ensemble accuracy on the minimum number of classifiers, i.e. when ensembles were constructed with only three classifiers. But this correlation could decrease to nearly 0 when the number of classifiers is close to the total number of classifiers available in the pool, as we explained in Sec. 5. A typical example is shown in Fig. 3, and this tendency is observed on all our experimental problems. This is the reason why the measured average correlation is not too significant compared with the ME.

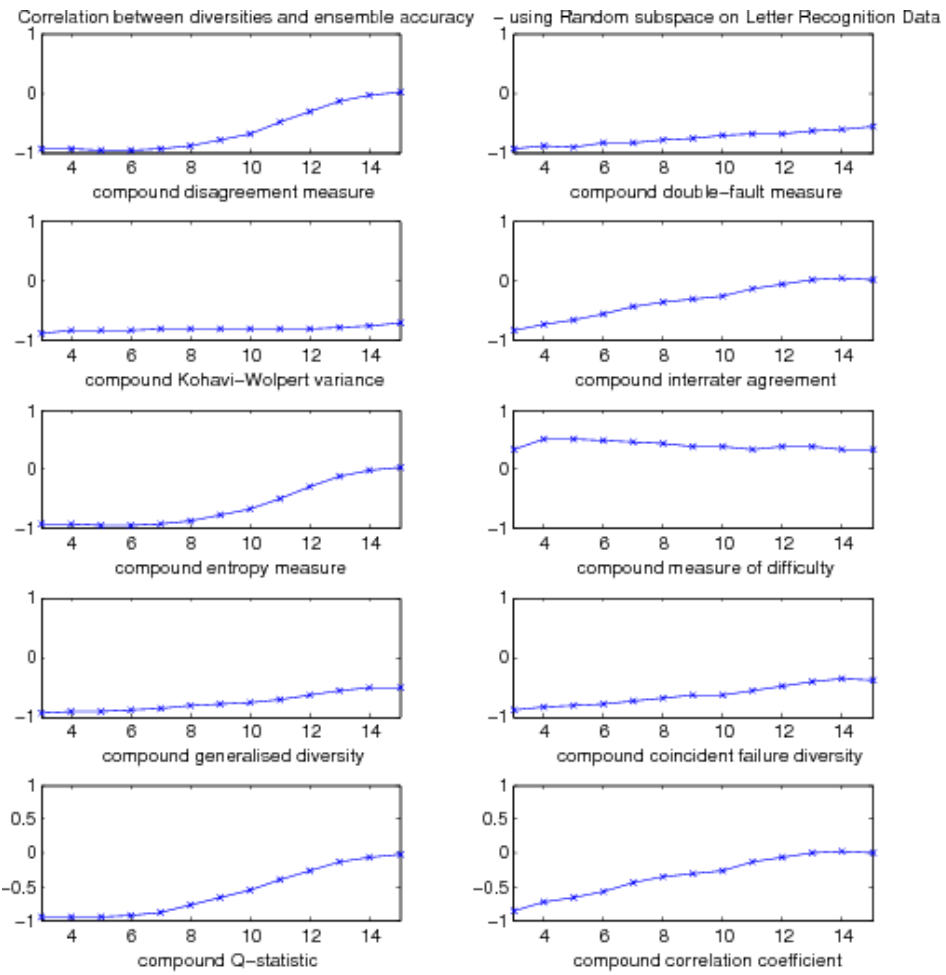

Fig. 3. The correlations between the CDFs and the accuracy on the letter recognition problem extracted from the UCI machine learning database with the Random subspaces as the ensemble creation method. We observe that the larger the ensemble size, the lower the correlation.

7. Ensemble Selection and Diversity as Objective Function

Even though the experiment shows that the compound diversity functions are strongly correlated with ensemble accuracy, it is important to show that such functions can be used as objective functions for ensemble selection. Thus, we carried out a number of experiments using different diversities as objective functions for ensemble selection. These objective functions are evaluated by genetic algorithm (GA) searching. We used a GA because the complexity of population based searching algorithms can be flexibly adjusted depending on the size of the population and the number of the generations to proceed. Moreover, because the algorithm returns population of the best combination, it can be potentially exploited to prevent generalization problems.[28] We tested 20 different diversities, including ten compound diversity functions and ten original diversity measures. Besides these 20 different objective functions, we also used the Mean Classifier Error (ME) and the error of ensembles applying the majority voting (MVE). We then compared their effectiveness as objective functions for the creation of the EoC.

7.1. *Experimental protocol for ensemble selection*

We carried out experiments on a ten-class handwritten-numeral problem. The data was extracted from NIST SD19, essentially as in Ref. 33, based on the ensembles of KNNs generated by the Random Subspaces method. The NIST SD19 contains more than 400,000 isolated handwritten digits organized into seven partitions or series of images, denoted by: $hsf\_\{0-3\}$, $hsf\_\{4\}$, $hsf\_\{6\}$ and $hsf\_\{7\}$. We used nearest neighbor classifiers ($K = 1$), each NN classifier having a different feature subset of 32 features extracted from the total of 132 features. Four databases were used: the training set with 5000 samples (from partitions $hsf\_\{0-3\}$) was used to create 100 NN in Random Subspaces, and the optimization set containing 10000 samples ($hsf\_\{0-3\}$) was used for GA searching. To avoid overfitting during GA searching, the validation set containing 10000 samples ($hsf\_\{0-3\}$) was used to select the best solution from the current population according to the defined objective function, and then to store it in a separate archive after each generation. Using the best solution from this archive, the test set containing 60089 samples ($hsf\_\{7\}$) was used to evaluate the accuracies of EoC. We used GA as the searching algorithm, with 128 individuals in the population and with 500 generations, which means 64,000 ensembles were evaluated in each experiment. The mutation probability was set to 0.01, and the crossover probability to 50%. With 22 different objective functions (Mean Classifier Error (ME), Majority Voting Error (MVE), ten original diversity measures, and ten compound diversity functions) and 30 replications, 42.24 million ensembles were searched and evaluated. A threshold of three classifiers was applied as the minimum number of classifiers for EoC during the whole searching process. Experimental results are reported in Table 5.

First, we see that the use of traditional diversity measures does not always give satisfying performance. The results show that the selected ensembles perform

Table 5. The recognition rates of the ensembles selected by different objective functions, including traditional diversity measures and compound diversity functions (CDF), on NIST SD19 handwritten numerals.

100 NN	ME	MVE
96.28 ± 0.00 %	94.18 ± 0.00 %	96.45 ± 0.05 %

DM	KW	EN	GD	CFD
91.56 ± 0.46 %	95.72 ± 0.00 %	90.04 ± 0.21 %	93.26 ± 0.25 %	93.66 ± 0.18 %
INT	DIFF	DF	Q.	COR
93.04 ± 0.11 %	96.24 ± 0.00 %	94.10 ± 0.13 %	91.96 ± 0.52 %	92.44 ± 0.37 %

CDF-DM	CDF-KW	CDF-EN	CDF-GD	CDF-CFD
96.19 ± 0.09 %	96.20 ± 0.06 %	96.18 ± 0.08 %	96.19 ± 0.05 %	96.22 ± 0.08 %
CDF-INT	CDF-DIFF	CDF-DF	CDF-Q.	CDF-COR
96.22 ± 0.09 %	96.23 ± 0.08 %	96.20 ± 0.10 %	96.20 ± 0.05 %	96.23 ± 0.07 %

poorly, most of them are even worse than those chosen by ME. Apparently there are many outliers indicated in the box plot (Fig. 4), which are values exceeding the distance of 1.5 interquartile range ($Q_U - Q_L$) from either end of the box, which means that searching by the traditional diversity measures could lead to great instability. This phenomenon is understandable, in light of the fact that the

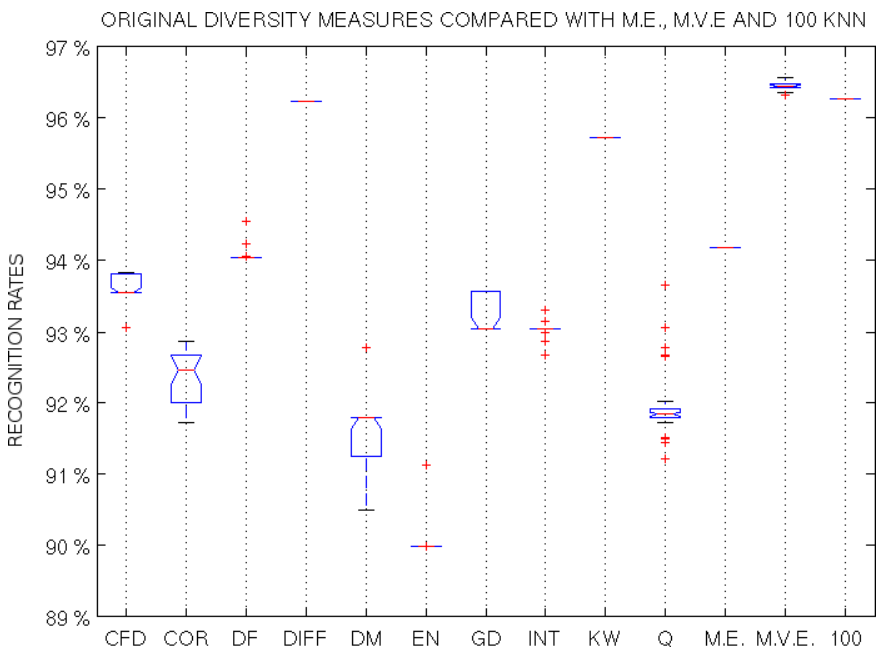

Fig. 4. The recognition rates achieved by EoCs selected by original diversity measures, compared with the Mean Classifier Error (ME), Majority Voting Error (MVE), and the ensemble of all (100) NN classifiers.

original diversity measures were designed to optimize diversity among classifiers, and they do not target ensemble accuracy directly. The result also confirms the lack of correlation between most diversity measures and ensemble accuracy.

As we predicted, all pairwise diversity measures will lead to the minimum number of classifiers, i.e. three classifiers in this experiment. Moreover, some non-pairwise diversity measures will lead to three classifiers, since it will not be easy to find an ensemble with greater diversity than the ensemble composed of the three most diverse classifiers. The only two diversity measures that can resist the minimum-converging tendency are KW, which always finds 17 classifiers for EoC, and DIFF with 21 classifiers. DIFF performs relatively well in this case, as had been shown in Ref. 30. It seems that DIFF, the minimization of the variance of the proportion of correct classifiers on all samples, encourages fairly distributed difficulty, instead of selecting the most diverse classifiers. To arrive at a fair distribution of difficulty, a number of classifiers would be required. Even DIFF did not have strong correlation with ensemble accuracy in our previous correlation measurement; it does guarantee a comparable performance in this case.

By contrast, the proposed compound diversity functions are much more stable (Fig. 5). Most EoCs selected by them are constructed by 35–60 classifiers, which is about half the total of 100 classifiers. Compared with the EoCs found by MVE with 19–35 classifiers, the sizes of EoCs selected by the compound diversity functions

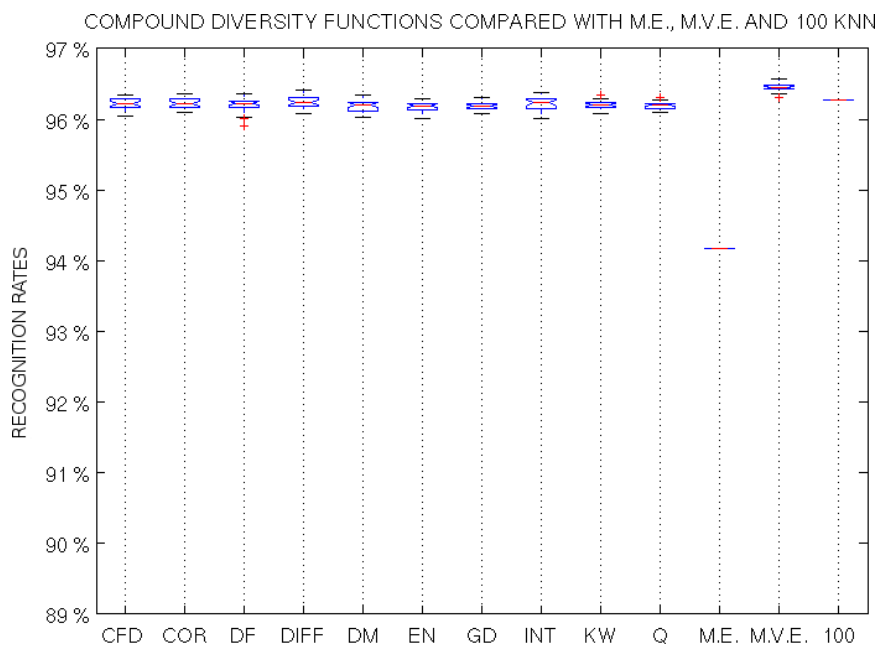

Fig. 5. The recognition rates achieved by EoCs selected by compound diversity functions, compared with the Mean Classifier Error (ME), Majority Voting Error (MVE), and the ensemble of all (100) NN classifiers.

are larger, but the performances are quite stable. Though MVE is still clearly better with the significance $p < 0.01$, the differences in recognition rates with EoCs selected by MVE are usually less than 0.3%. This indicates that the EoCs selected by the proposed compound functions are quite generalized and fit different fusion functions.

Finally, we point out that, among all diversity measures, the compound diversity functions always perform better than the original diversity measures. While most of the original diversity measures perform worse than ME, the use of the compound diversity functions gives much better results than ME. Furthermore, all compound diversity functions achieve similar performances; which should result from the strong correlations among most of them.

8. Discussion

Previous published studies suggested that diversity is not unequivocally related to ensemble accuracy, and it is our objective to demonstrate that the implementation of diversity can help in ensemble selection. As we can see in these experiments, there are correlations between the proposed compound diversity functions and ensemble accuracy. The result also suggests that DM, KW, EN, GD and CFD are stable for all ensemble creation methods. Performance depends strongly on the accuracy of individual classifiers, but, in general, an equivalent or stronger correlation could be achieved with compound diversity functions, especially with KW.

In contrast to the use of the original diversity measures, which show no strong intercorrelation,[20] these compound diversity functions do have strong intercorrelations, except for COR, DIFF, INT and Q. This means that most diversities have similar indication, and so the creation of new diversity measures might not be a priority, but rather consideration of how to use diversities for ensemble selection. With the Random Subspaces method, this correlation is stronger than it is in either Bagging or Boosting. In general, a decrease in correlation is observed when the number of selected classifiers increases, but this was not the case for high-class problems, as we predicted.

Based on GA searching, we see that the compound diversity functions apparently outperform the original diversity measures and the Mean Classifier Error as objective functions for ensemble selection, and even exceed the performance of the ensemble of all 100 NN classifiers and reduce the number of classifiers by half. The proposed compound diversity functions do improve the performance of EoCs, and always perform better than the respective original diversity measures, their performances being much close to those ensembles obtained with the MVE objective function.

Recall that MVE is used both for ensemble selection and for classifier combination, and thus, it is understandable that MVE will have the best performance as the objective function. But, it is possible that when different fusion functions are used, MVE will not be the best choice as an objective function. An ensemble combined

with Decision Template (DT), for example, might not have the best performances when it is evaluated by MVE. The "no free lunch" theorem[36,37] has also supported the idea that no search algorithm will be optimal in all situations.

Given that these compound diversity functions do not take into account of any fusion functions, the ensemble outputs can be further optimized using various classifier-combining methods.[17,27,28] This is an advantage for modular approaches to further optimize searching algorithms and fusion functions. All the compound diversity functions worked well for ensemble selection in our experiment, even some that had previously been measured and found to have weaker correlation with ensemble accuracy. This indicates a strong similarity among most of the compound diversity functions in the pattern recognition problems evaluated.

The result encourages further exploration of the implementation of compound diversity functions, and the pertinence of these functions for use with different searching algorithms. Moreover, it suggests that the problem resides in finding ways to amalgamate diversities and individual classifier errors, rather than allowing diversity measures to select EoCs single-handedly. Another advantage of compound diversity functions is that they can be calculated beforehand, since diversities are measured in a pairwise manner, and error rates are measured on each classifier; thus, for time-consuming searching methods, such as GA or exhaustive searching, ensemble accuracy can be estimated quickly by simply calculating the products of the diversity measures and individual classifier errors. Given L classifiers and N samples on a C-class problem, the complexity of the CDFs is $O(L + \frac{L(L-1)}{2})$, the complexity of non-pairwise traditional diversity measures is $O(LN)$, and the complexity of the MVE is $O(LNC)$. The CDFs thus has the lower cost for the ensemble selection.

9. Conclusion

Diversity used to be regarded as useful, but not unequivocally related to ensemble accuracy. In this exploratory work on diversity, we show that, with the proper compound diversity functions, there are strong correlations between the diversities and ensemble accuracy. Moreover, using population-based GA searching, the compound diversity functions do improve the recognition rates of the ensembles. We have drawn up some conclusion based on our experiments:

(1) Diversities and the performances of individual classifiers should be taken into account together.
(2) Compound diversity functions have stronger correlations with the ensemble accuracy than the traditional diversity measures.
(3) Compared with MVE, compound diversity functions have lower cost for the ensemble selection.
(4) In general, ensembles selected by different compound diversity functions have so far been found to have similar performances for GA searching, with the significance $p \geq 0.1$.

Given that this exploratory work has been accomplished with different ensemble creation methods, considering different numbers of classifiers of ensembles, evaluating millions of ensembles, but with a restricted number of classification algorithms, and in a limited number of problems, it will be advisable to carry out more experiments on ensemble selection, with more pattern recognition problems and more classification methods. The problems associated with optimizing ensembles include not only diversity, but also searching algorithms[28] and fusion functions.[17] The next step will be to test different searching algorithms with the proposed compound diversity functions, for the purpose of optimizing the ensemble selection process.

Acknowledgment

This work was supported in part by grant OGP0106456 to Robert Sabourin from the NSERC of Canada.

References

1. A. A. Afifi and S. P. Azen, *Statistical Analysis: A Computer Oriented Approach*, 2nd edn. (Academic Press, New York, 1979).
2. Y. Amit and D. Geman, Shape quantization and recognition with randomized trees, *Neural Comput.* **9** (1997) 545–1588.
3. R. E. Banfield, L. O. Hall, K. W. Bowyer and W. P. Kegelmeyer, A new ensemble diversity measure applied to thinning ensembles, *Int. Workshop on Multiple Classifier Systems* (*MCS 2003*) (2003), pp. 306–316.
4. L. Breiman, Random forests, *Mach. Learn.* **45** (2001) 5–32.
5. G. Brown, J. Wyatt, R. Harris and X. Yao, Diversity creation methods: a categorisation, *Int. J. Inform. Fus.* **6**(1) (2005) 5–20.
6. G. Brown, J. Wyatt and P. Sun, Between two extremes: examining decompositions of the ensemble objective function, *Int. Workshop on Multiple Classifier Systems (MCS 2005)* (2005), pp. 296–305.
7. P. Domingos, A unified bias-variance decomposition and its applications, *Int. Conf. Mach. Learn. (ICML 2000)* (2000), pp. 231–238.
8. R. P. W. Duin, Pattern Recognition Toolbox for Matlab 5.0+, available free at: ftp://ftp.ph.tn.tudelft.nl/pub/bob/prtools
9. J. L. Fleiss, B. Levin and M. C. Paik, *Statistical Methods for Rates and Proportions*, 2nd edn. (John Wiley & Sons, New York, 2003).
10. B. E. Geman and S. R. Dorsat, Neural networks and the bias/variance dilemma, *Neural Comput.* **4** (1992) 1–58.
11. G. Giacinto and F. Roli, Design of effective neural network ensembles for image classification purposes, *Image Vis. Comput.* **19**(9–10) (2001) 699–707.
12. G. Giacinto and F. Roli, Dynamic classifier selection based on multiple classifier behaviour, *Patt. Recogn.* **34**(9) (2001) 179–181.
13. A. Grove and D. Schuurmans, Boosting in the limit: maximizing the margin of learned ensembles, *Proc. Fifteenth Nat. Conf. Artif. Intell.* (1998), pp. 692–699.
14. L. K. Hansen and P. Salamon, Neural network ensembles, *IEEE Trans. Patt. Anal. Mach. Intell.* **12** (1990) 993–1001.
15. T. K. Ho, The random space method for constructing decision forests, *IEEE Trans. Patt. Anal. Mach. Intell.* **20** (8) (1998) 832–844.

16. G. James, Variance and bias for general loss functions, *Mach. Learn.* **51**(2) (2003) 115–135.
17. J. Kittler, M. Hatef, R. Duin and J. Matas, On combining classifiers, *IEEE Trans. Patt. Anal. Mach. Intell.* **20**(3) (1998) 226–239.
18. R. Kohavi and D. H. Wolpert, Bias plus variance decomposition for zero-one loss functions, *Proc. Int. Mach. Learn. Conf.* (*ICML 1996*) (1996), pp. 275–283.
19. A. Krogh and J. Vedelsby, Neural network ensembles, cross validation, and active learning, *Advances in Neural Inform. Process. Syst.* **7** (1995) 231–238.
20. L. I. Kuncheva, M. Skurichina and R. P. W. Duin, An experimental study on diversity for bagging and boosting with linear classifiers, *Int. J. Inform. Fus.* **3**(2) (2002) 245–258.
21. L. I. Kuncheva and C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Mach. Learn.* **51**(2) (2003) 181–207.
22. P. Melville and R. J. Mooney, Creating diversity in ensembles using artificial data, *Inform. Fus.* **6**(1) (2005) 99–111.
23. L. Z. Oliveira, M. Morita, R. Sabourin and F. Bortolozzi, Multi-objective genetic algorithms to create ensemble of classifiers, *Proc. 3rd Int. Conf. Evolutionary Multi-Criterion Optimization (EMO 2005)* (2005), pp. 592–606.
24. D. Opitz and R. Maclin, Popular ensemble methods: an empirical study, *J. Artif. Intell. Res.* **11** (1999) 169–198.
25. D. Partridge and W. Krzanowski, Software diversity: practical statistics for its measurement and exploitation, *Inform. Softw. Technol.* **39** (1997) 707–717.
26. E. Pekalska, M. Skurichina and R. P. W. Duin, Combining dissimilarity-based one-class classifiers, *Int. Workshop on Multiple Classifier Systems (MCS 2004)* (2004), pp. 122–133.
27. D. Ruta and B. Gabrys, Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems, *Proc. 4th Int. Symp. Soft Computing*, 2001.
28. D. Ruta and B. Gabrys, Classifier selection for majority voting, *Int. J. Inform. Fus.* (2005), pp. 63–81.
29. R. E. Schapire, Y. Freund, P. Bartlett and W. S. Lee, Boosting the margin: a new explanation for the effectiveness of voting methods, *Ann. Statist.* **26**(5) (1998) 1651–1686.
30. C. A. Shipp and L. I. Kuncheva, Relationships between combination methods and measures of diversity in combining classifiers, *Int. J. Inform. Fus.* **3**(2) (2002) 135–148.
31. M. Skurichina, L. I. Kuncheva and R. P. W. Duin, Bagging and boosting for the nearest mean classifier: effects of sample size on diversity and accuracy, *Int. Workshop on Multiple Classifier Systems (MCS 2002)* (2002), pp. 62–71.
32. D. M. J. Tax, M. Van Breukelen, R. P. W. Duin and J. Kittler, Combining multiple classifiers by averaging or by multiplying, *Patt. Recogn.* **33**(9) (2000) 1475–1485.
33. G. Tremblay, R. Sabourin and P. Maupin, Optimizing nearest neighbour in random subspaces using a multi-objective genetic algorithm, *Proc. 17th Int. Conf. Patt. Recogn. (ICPR 2004)* (2004), pp. 208–211.
34. N. Ueda and R. Nakano, Generalization error of ensemble estimators, *Proc. Int. Conf. Neural Networks (ICNN 1996)*, (1996), pp. 90–95.
35. G. I. Webb and Z. Zheng, Multistrategy ensemble learning: reducing error by combining ensemble learning techniques, *IEEE Trans. Knowl. Data Engin.* **16**(8) (2004) 980–991.
36. D. Whitley, Functions as permutations: regarding no free lunch, Walsh analysis and summary statistics, *Parallel Problem Solving from Nature (PPSN 2000)* (2000), pp. 169–178.

37. D. H. Wolpert and W. G. Macready, No free lunch theorems for search, *IEEE Trans. Evolut. Comput.* **1** (1997) 67–82.
38. H. Zouari, L. Heutte, Y. Lecourtier and A. Alimi, Building diverse classifier outputs to evaluate the behavior of combination methods: the case of two classifiers, *Int. Workshop on Multiple Classifier Systems (MCS 2004)* (2004), pp. 273–282.

Albert Hung-Ren Ko received M.Sc.A, degree in artificial intelligence and pattern recognition from the Université Pierre et Marie Curie in 2002, and his Ph.D degree in pattern recognition from the École de Technologie Supérieure, Université du Quebec in 2007.

His research interests are in ensemble classification methods, small world structure and neural networks.

R. Sabourin joined in 1977 the Physics Department of the Montreal University where he was responsible for the design, experimentation and development of scientific instrumentation for the Mont Mégantic Astronomical Observatory. His main contribution was in the design and the implementation of a microprocessor-based fine tracking system combined with a low-light level CCD detector. In 1983, he joined the staff of the École de Technologie Supérieure, Université du Québec, in Montréal where he co-founded the Dept. of Automated Manufacturing Engineering where he is currently Full Professor and teaches pattern recognition, evolutionary algorithms, neural networks and fuzzy systems. In 1992, he joined also the Computer Science Department of the Pontifícia Universidade Católica do Paraná (Curitiba, Brazil) where he was co-responsible for the implementation in 1995 of a master program and in 1998 a PhD program in applied computer science. Since 1996, he is a senior member of the Centre for Pattern Recognition and Machine Intelligence (CENPARMI, Concordia University).

Dr Sabourin is the author (and co-author) of more than 250 scientific publications including journals and conference proceeding. He was co-chair of the program committee of CIFED'98 (Conférence Internationale Francophone sur l'Écrit et le Document, Québec, Canada) and IWFHR'04 (9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan). He was nominated as Conference co-chair of the next ICDAR'07 (9th International Conference on Document Analysis and Recognition) that has been held in Curitiba, Brazil in 2007.

His research interests are in the areas of handwriting recognition, signature verification, watermarking algorithms and biocryptography.

Alceu de Souza Britto Jr. received M.Sc. degree in industrial informatics from the Federal Center for Technological Education of Parana (Brazil) in 1996, and Ph.D. degree in computer science from Pontifical Catholic University of Parana (PUCPR, Brazil). In 1989 he joined the Computer Science Department of the Ponta Grossa University (Brazil). In 1995, he also joined the Computer Science Department of the PUCPR. From July 1998 to July 2000 he worked on handwriting recognition area in Montreal (Canada) in the laboratories of the Centre for Pattern Recognition and Machine Intelligence (CENPARMI-Concordia University) and the École de Technologie Supérieure (ÉTS-Université du Quebec). From 2004 to 2005, he was the director of the Computer Science Undergraduate Course at PUCPR. Since 2001, he joined the Post Graduate Program in Applied Informatics of the PUCPR, where his research interests are in the areas of pattern recognition, computer vision, document analysis and handwriting recognition.