Combining Diversity and Classification Accuracy for Ensemble Selection in Random Subspaces

Albert Hung-Ren Ko, Robert Sabourin, and Alceu de Souza Britto, Jr.

Abstract— **An ensemble of classifiers has been shown to be effective in improving classifier performance. Two elements are believed to be viable in constructing an ensemble: a) the classification accuracy of each individual classifier; and b) diversity among the classifiers. Nevertheless, most works based on diversity suggest that there exists only weak correlation between diversity and ensemble accuracy. We propose compound diversity functions which combine the diversities with the classification accuracy of each individual classifier, and show that with Random subspaces ensemble creation method, there is a strong correlation between the proposed functions and ensemble accuracy. The statistical result indicates that compound diversity functions perform better than traditional diversity measures.**

I. 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 [1, 9, 14, 19, 20]. 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 [2], Bagging and Boosting [13, 22, 23]. 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.

Albert Hung-Ren Ko is with LIVIA, École de Technologie Supérieure, University of Quebec, Montreal, Quebec, H3C 1K3, Canada (email: albert@livia.etsmtl.ca).

Robert Sabourin is with LIVIA, École de Technologie Supérieure, University of Quebec, Montreal, Quebec, H3C 1K3, Canada (email: robert.sabourin@etsmtl.ca).

Alceu de Souza Britto, Jr. is with PPGIA, Pontifical Catholic University of Parana, PR 80215-901, Curitiba, Brazil (email: alceu@ppgia.pucpr.br).

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 [1, 5, 9, 10, 15, 25] and the latter on ensemble combination, i.e. the choice of fusion functions [14, 19, 20, 25, 27]. 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 [1, 9, 25]. 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 [1, 13, 14, 21, 25]. 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 [15], 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 [6, 26]. Third, we need to optimize fusion functions in order to combine classifiers [14], since, without knowing the best fusion functions, it would be premature to use them for ensemble selection. 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, 2, 4, 5, 10]. 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 [1, 9, 13], leading some researchers to consider diversity measures to be unnecessary for ensemble selection [25]. 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 classification accuracy 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) Is there a correlation between the diversity and ensemble accuracy?
2) Which is the best diversity measure for observing such a correlation?
3) Is there any effect on such a correlation, e.g. from the number of classifiers?
4) Can diversity be effective for ensemble selection?

To answer these questions, we derive compound diversity functions by combining diversities and the classification accuracies 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 Random Subspaces as the ensemble creation method, 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.

II. DIVERSITY MEASURES

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 $d_{12}, d_{13}, ..., d_{(L-1)L}$ will be calculated, and the final diversity $\bar{d}$ will be its average [1]:

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

   This type of diversity includes: Q-statistics, the correlation coefficient, the disagreement measure and the double fault [1, 2, 15].

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, Kohavi-Wolpert variance, the measurement of interrater agreement, the measure of difficulty, generalized diversity and coincident failure diversity [1, 5, 15].

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 [1]. We discuss this problem and its theoretical obstacle in the next section.

III. DILEMMA OF THE AMBIGUITY TOWARDS THE ENSEMBLE ACCURACY

The theoretical problem concerning the ensemble accuracy is very complicated, in general, while the behavior of a zero-one loss error is an important topic, there is simply no clear analog of the bias-variance-covariance decomposition when we have a zero-one loss function [9, 18]. To have more insights into the problem, we focus on continuous loss functions instead of zero-one loss function. Because the concept of ensemble diversity can be understood via bias-variance decomposition [3, 9, 11, 17, 18] in continuous loss functions, we could have more leverage on the problem. Here, we adopt the framework established in [9] to discuss the impediment to using the ambiguity to estimate ensemble accuracy.

Given the features of a certain sample $x \in X$, assume that we have a classifier f trained with a particular dataset X, the expectation of the output of the classifier can be written as $E(f(x))$. For convenience, we denote $E\{f\}$ instead of $E(f(x))$. If the correct value of the output is r, then we can write the bias of the classifier f as:

$$bias(f) = E\{f\} - r \tag{2}$$

and the variance of the classifier f can be written as:

$$var(f) = E\{(f - E\{f\})^2\} \tag{3}$$

Now, the mean square error (MSE) of this classifier f can be exactly represented by its variance and bias:

$$E\{(f-r)^2\} = (E\{f\} - r)^2 + E\{(f - E\{f\})^2\} \tag{4}$$

$$or\ MSE\{f\} = bias(f)^2 + var(f) \tag{5}$$

This form can be further decomposed into bias-variance-covariance [9, 10]. For L classifiers, the averaged bias of the ensemble members is defined as:

$$\bar{b} = \frac{1}{L}\sum_{i}^{L}(E\{f_i\} - r) \tag{6}$$

Then, the averaged variance of the ensemble members will be:

$$\bar{v} = \frac{1}{L}\sum_{i}^{L}(E\{(f_i - E\{f_i\})^2\}) \tag{7}$$

and the averaged covariance of the ensemble members will be:

$$\bar{c} = \frac{1}{L(L-1)}\sum_{i}^{L}\sum_{j\neq i}^{L} E((f_i - E\{f_i\})(f_j - E\{f_j\})) \tag{8}$$

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

$$MSE(L) = E\{((\frac{1}{L}\sum_{i}^{L} f_i) - r)^2\} \quad (9)$$

$$= \bar{b}^2 + \frac{1}{L}\bar{v} + \frac{L-1}{L}\bar{c} \quad (10)$$

To determine the link between $MSE(L)$ and the ambiguity, which measures the amount of variability among classifier outputs in ensembles, we need to apply ambiguity decomposition. It has been proved [12] 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 [12]:

$$(f_{ens} - r)^2 = \sum_{i}^{L} w_i(f_i - r)^2 - \sum_{i}^{L} w_i(f_i - f_{ens})^2 \quad (11)$$

where w_i is the weight of classifier f_i in the ensemble, and $0 \leq w_i \leq 1$. If every classifier f_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 ($\sum_i^L w_i = 1$):

$$f_{ens} = \sum_{i}^{L} w_i f_i \quad (12)$$

For the $MSE(L)$ 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 $f_{ens} = \bar{f}$. Consequently the ambiguity decomposition can be written as:

$$(\bar{f} - r)^2 = \frac{1}{L}\sum_{i}^{L}(f_i - r)^2 - \frac{1}{L}\sum_{i}^{L}(f_i - \bar{f})^2 \quad (13)$$

Note that its expectation is exactly eq.9 and eq.10:

$$E\{\frac{1}{L}\sum_{i}^{L}(f_i - r)^2 - \frac{1}{L}\sum_{i}^{L}(f_i - \bar{f})^2\} = \bar{b}^2 + \frac{1}{L}\bar{v} + \frac{L-1}{L}\bar{c} \quad (14)$$

The ambiguity is the second term on the left-hand side in eq.14, and it can be written as [12]:

$$E\{(\frac{1}{L}\sum_{i}^{L}(f_i - \bar{f})^2)\} \quad (15)$$

$$= \frac{1}{L}\sum_{i}^{L} E\{(f_i - E\{f_i\})^2\} - E\{(\bar{f} - E(\bar{f}))^2\} \quad (16)$$

$$= \bar{v} - var(\bar{f}) = \bar{v} - \frac{1}{L}\bar{v} - \frac{L-1}{L}\bar{c} \quad (17)$$

The first term of the left-side in eq.14 is the sum of averaged bias and averaged variance of classifiers:

$$E\{\frac{1}{L}\sum_{i}^{L}(f_i - r)^2\} = \bar{b}^2 + \bar{v} \quad (18)$$

As stated in [9], 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.

IV. PROPOSED COMPOUND DIVERSITY FUNCTIONS

Even though the ambiguity among classifiers is not a guarantee of ensemble accuracy, it has been shown that this ambiguity is a necessary condition for an ensemble to achieve a high degree of accuracy [1, 19, 20]. To compensate for the ambiguity dilemma, and to use its intrinsic property to reduce $MSE(L)$, we propose compound diversity functions using the diversity measures in a pairwise fashion to estimate ensemble accuracy. First, suppose that we have an ensemble with 2 classifiers f_i, f_j, and that classifiers f_i and f_j have the recognition rates a_i and a_j respectively, the error of classifier f_i is $(1 - a_i)$, the error of classifier f_j is $(1 - a_j)$ and the diversity d_{ij} is measured between them.

With only two classifiers, we get $L = 2$ in eq.7 and eq.8. As a result, the ambiguity between f_i and f_j is exactly half of the difference between their variance and covariance in eq.17:

$$amb_{ij} = \frac{1}{2}(\bar{v} - \bar{c}) \quad (19)$$

$$= \frac{1}{4}(E\{(f_i - E\{f_i\})^2\} + E\{(f_j - E\{f_j\})^2\} - 2 \cdot E\{(f_i - E\{f_i\}) \cdot (f_j - E\{f_j\})\})$$

If we use $L = 2$ in eq.10 and replace $\frac{1}{2}(\bar{v} - \bar{c})$ by amb_{ij}, we can write $MSE(2)$ as:

$$MSE(2) = \bar{b}^2 + \frac{1}{2}(\bar{v} + \bar{c}) = amb_{ij} + \bar{b}^2 + \bar{c} \quad (20)$$

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. The dilemma for the approximation is that, even if we can measure the ambiguity as a diversity measure, the sum of the averaged covariance and the averaged bias of individual classifiers is not easy to measure. Still, using the eq.19, we can also write the above equation as:

$$MSE(2) = \bar{b}^2 + \bar{v} - \frac{1}{2}(\bar{v} - \bar{c}) = \bar{b}^2 + \bar{v} - amb_{ij} \quad (21)$$

where $amb_{ij} = \frac{1}{2}(\bar{v} - \bar{c})$. 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 $var(f) + bias(f)^2$ of each individual classifier f, which is exactly the MSE of each individual classifier. Despite this, we do not have its exact value. We obtain instead its zero-one loss error [18], i.e. $(1 - a_i)$. As we mentioned before, up to now there has simply been no clear analog of the bias-variance-covariance decomposition for zero-one loss functions [9, 18]. Nevertheless, it is still reasonable to assume that the larger the MSE of each individual classifier, the larger the zero-one

loss error should be. This means that we can suppose that $(1-a_i) \approx \alpha_i(var(f_i)+bias(f_i)^2)$ for f_i and $(1-a_j) \approx \alpha_j(var(f_j)+bias(f_j)^2)$ for f_j.

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 suppose that the individual classifiers have a similar $MSE(f)$, i.e. if $var(f_j)+bias(f_j)^2$ is close to $var(f_i)+bias(f_i)^2$, one could obtain an proportional approximation of $(\bar{b}^2+\bar{v})$ by calculating $((var(f_i)+bias(f_i)^2)*(var(f_j)+bias(f_j)^2))^{\frac{1}{2}}$. As a result, the term $\bar{b}^2+\bar{v}$ could be proportionally approximated by the error rates of individual classifiers:

$$(\bar{b}^2+\bar{v}) \approx \gamma((1-a_i)\cdot(1-a_j))^{\frac{1}{2}} \tag{22}$$

However, given that diversity measures represent approximations of the ambiguity among classifiers, suppose that $d_{ij} \propto amb_{ij}$, given that $0 \leq d_{ij} \leq 1$. The term d_{ij} will have a high correlation with $amb_{ij} = \frac{1}{2}(\bar{v}-\bar{c})$, and the approximation of $\frac{1}{2}(\bar{v}-\bar{c})$ can be written as:

$$(\bar{v}-\bar{c}) \approx \delta \cdot d_{ij} \tag{23}$$

For a proportional approximation to $MSE(2)$, i.e. $\bar{b}^2+\bar{v}-amb_{ij}$, given proportional approximation $(\bar{b}^2+\bar{v})$ as $\gamma\cdot((1-a_i)\cdot(1-a_j))^{\frac{1}{2}}$, and the proportional 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. If $\bar{b}^2+\bar{v} \gg \frac{1}{2}(\bar{v}-\bar{c})$, then we only need the Mean Classifier Error $\gamma\cdot((1-a_i)\cdot(1-a_j))^{\frac{1}{2}}$ for the approximation of $MSE(2)$. If $\bar{b}^2+\bar{v} \ll \frac{1}{2}(\bar{v}-\bar{c})$, then we just need diversity $\delta \cdot d_{ij}$ as the approximation. But so far, both are considered insufficient for ensemble selection. Nevertheless, if we assume that classifier-error and classifiers-diversity have similar weights in $MSE(2)$:

$$|\bar{b}^2+\bar{v}-\frac{1}{2}(\bar{v}-\bar{c})| \leq \epsilon(\bar{b}^2+\bar{v}+\frac{1}{2}(\bar{v}-\bar{c})), \epsilon \ll 1 \tag{24}$$

then 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 proportional approximation of $MSE(2)$ as:

$$\widetilde{MSE_{ij}} \equiv (1-d_{ij})\cdot((1-ai)\cdot(1-aj))^{\frac{1}{2}} \tag{25}$$

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 (Fig.1). Consequently, the approximation of $MSE(L)$ will depend on the individual classifier errors and the respective diversity between classifiers. Given the number of selected classifiers $L \geq 2$, and we have $\widetilde{MSE(L)} \sim (\prod_{i=1}^{L}(1-a_i))^{\frac{1}{L}}(\prod_{i,j=1,i\neq j}^{L}(1-d_{i,j}))^{\frac{1}{L\times(L-1)}}$. Fig.1 shows the mechanism of such an approximation. Each circle represents the error rate of each individual classifier, and each line represents the diversity among them. By calculating their product, we can get an approximation of ensemble accuracy without any consideration for the type of fusion functions.

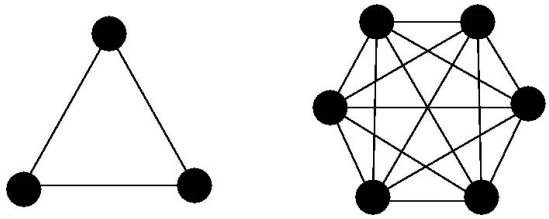

Fig. 1. The relationship between the error rate of individual classifiers (circles) and the diversity between them (lines).

It is important to note that different diversity measures are supposed to have different sorts of relationships with ensemble accuracy. Some diversity measures measure 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:

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

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

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

where $d_{i,j}$ should be interpreted as the similarity between f_i and f_j in this case. So, $\widehat{div_{sim}}$ ought to mean the likelihood of a common error being by all the classifiers. We expect negative correlation between the $\widehat{div_{sim}}$ 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 [8, 21, 25]. 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 $\widehat{c_{ij}}$ of classifier-pair $\widehat{c_{ij}}$ is larger than the diversities of any other selected L classifiers, given that $2 \leq L \leq M$:

$$\forall L, \quad \widehat{d_{ij}} \geq E\{d_{ij}\} = d_L \tag{28}$$

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's 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:

$$\widehat{div_{amb}} = \frac{L}{L-1}\left(\prod_{i=1}^{L}(1-a_i)\right)^{\frac{1}{L}}\left(\prod_{i,j=1,i\neq j}^{L}(1-d_{i,j})\right)^{\frac{1}{L\times(L-1)}} \tag{29}$$

$$\widehat{div_{sim}} = \frac{L}{L-1}\left(\prod_{i=1}^{L}(1-a_i)\right)^{\frac{1}{L}}\left(\prod_{i,j=1,i\neq j}^{L}d_{i,j}\right)^{\frac{1}{L\times(L-1)}} \tag{30}$$

V. 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 with Random Subspaces ensemble creation method. There are several requirements for the selection of pattern recognition problems. First, 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 4 databases, for each of 3 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 algorithms used in our experiments are Naive Bayesian Classifiers (NBC), Quadratic Discriminant Classifiers (QDC), and 5-Layer Neural Network Classifiers (NNC) with Back-Propagation [24]. 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

TABLE I

UCI DATA FOR ENSEMBLES OF CLASSIFIERS. C: NUMBER OF CLASSES; TR: NUMBER OF TRAINING SAMPLES; TS: NUMBER OF TEST SAMPLES; FEAT: NUMBER OF TOTAL FEATURES. RS: CARDINALITY OF FEATURES FIXED IN RANDOM SUBSPACES.

Database	C	Tr	Ts	Feat	RS
Wisconsin Breast-Cancer	2	284	284	30	5
Satellite	6	4435	2000	36	4
Image Segmentation	7	210	2100	19	4
Letter Recognition	26	10007	9993	16	12

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. A total of $3\times4\times13\times50\times30 = 0.234$ million ensembles should be evaluated.

We measured ensemble accuracy correlation on 10 traditional diversity measures, including the disagreement measure (DM), the double-fault (DF), Kohavi-Wolpert variance (KW), the interrater agreement (INT), the entropy measure (EN), the difficulty measure (DIFF), generalized diversity (GD), coincident failure diversity (CFD), Q-statistics (Q) and the correlation coefficient (COR) [1, 2, 5, 15], as well as on 10 respective proposed compound diversity functions (eq.26 & eq.27). They are also compared with the Mean Classifier Error (ME) of individual classifiers. 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% .

A. Measured Correlations

In the 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.

First, we observe that ME has an apparent correlation with ensemble accuracy (Table 2). Furthermore, it shows that, in general, compound diversity functions give better results than the original diversity measures. Of all the diversity measures, Q, COR, INT and DIFF are not stable. By contrast, DM, DF, KW, EN, GD and CFD are quite reliable. 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. The advantage of compound diversity functions over the original diversity measures can be perceived in this case.

It is certain that the number of classifiers has an impact on the correlation between compound diversity functions and ensemble accuracy. We found that the strongest correlation with ensemble accuracy on the minimum number of classifiers, i.e. when ensembles were constructed with only 3

TABLE II

CORRELATION 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 SIGNS OF CORRELATION WITH ENSEMBLE ACCURACY [1, 13].

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

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. This is the reason why the measured average correlation is not too significant.

VI. 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 [25]. We tested 20 different diversities, including 10 compound diversity functions and 10 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.

A. Experimental Protocol for Ensemble Selection

We carried out experiments on a 10-class handwritten-numeral problem. The data was extracted from $NISTSD19$, essentially as in [7], based on the ensembles of KNNs generated by the Random Subspaces method. We used nearest neighbor classifiers ($K = 1$) for KNN, each KNN 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 ($hsf\_\{0-3\}$) was used to create 100 KNN 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 is 0.01. With 22 different objective functions (Mean Classifier Error (ME), Majority Voting Error (MVE), 10 original diversity measures, and 10 compound diversity functions) and 30 replications, 42.24 million ensembles were searched and evaluated. A threshold of 3 classifiers was applied as the minimum number of classifiers for EoC during the whole searching process.

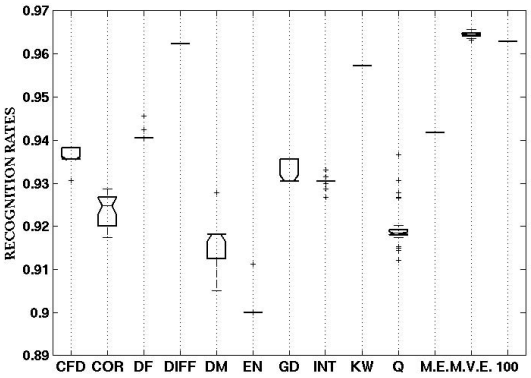

Fig. 2. The recognition rates achieved by EoCs selected by original diversity measures, compared with the Mean Classifier Error (ME), Majority Voting Error (MVE), and the ensemble of all (100) KNN classifiers

First, we see that the use of traditional diversity measures does not always give satisfying performance. The results show that the selected ensembles perform poorly, most of them are even worse than those chosen by ME. Apparently there are many outliers indicated in the box plot (Fig. 2), 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

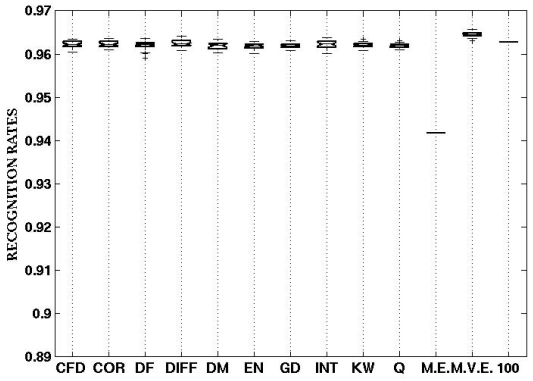

Fig. 3. The recognition rates achieved by EoCs selected by compound diversity functions, compared with the Mean Classifier Error (ME), Majority Voting Error (MVE), and the ensemble of all (100) KNN classifiers

lead to great instability. This phenomenon is understandable, in light of the fact that the 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., 3 classifiers in this experiment. Moreover, some non-pairwise diversity measures will lead to 3 classifiers, since it will not be easy to find an ensemble with greater diversity than the ensemble composed of the 3 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 [8]. 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. 3). Most EoCs selected by them are constructed by $35 \sim 60$ classifiers, which is about half the total of 100 classifiers. Compared with the EoCs found by MVE with $19 \sim 35$ classifiers, the sizes of EoCs selected by the compound diversity functions are larger, but the performances are quite stable and comparable. 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.

VII. 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 Random Subspaces ensemble creation method. 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 [13], 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. 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 KNN classifiers and reduce the number of classifiers by half. The only two original diversity measures that have better performances than ME are DIFF and KW, the others appearing to be inefficient. 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. 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 [14, 21, 25]. This is a huge 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.

VIII. CONCLUSION

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, which is much faster than other objective functions. Given L evaluated classifiers on N samples, traditional non-pairwise diversity measures have the complexity of $O(LN)$ for each evaluation, while compound diversity functions only have the complexity as $O(L+\frac{L(L-1)}{2})$ to calculate the product of all classification accuracies and pairwise diversities. Given $L \ll N$, compound diversity functions offer a significant speed-up. For more flexible ensemble selection methods, such as adaptive ensemble selections and dynamic ensemble selections [26], this will be a great advantage.

Given that this exploratory work has been accomplished with 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. We carried out experiments only in Random Subspaces as ensemble creation method, and it will be of great interest to measure the impact of the proposed CDF on boosting and bagging as well. The problems associated with optimizing ensembles include not only diversity, but also searching algorithms [25] and fusion functions [14]. 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. Moreover, it will be essential to address the task of combining classifiers - the choice of fusion function - in the near future.

ACKNOWLEDGMENT

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

REFERENCES

[1] L. I. Kuncheva and C. J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181-207, 2003

[2] T.K. Ho, "The random space method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998

[3] G. Brown, J. Wyatt and P. Sun, "Between Two Extremes: Examining Decompositions of the Ensemble Objective Function," *International Workshop on Multiple Classifier Systems (MCS 2005)*, pp. 296-305, 2005

[4] G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image and Vision Computing*, vol. 19, no. 9-10, pp. 699-707, 2001

[5] R. Kohavi, and D.H. Wolpert, "Bias Plus Variance Decomposition for Zero-One Loss Functions," *In Proceedings of the International Machine Learning Conference (ICML 1996)*, pp. 275-283, 1996

[6] P. Melville and R. J. Mooney, "Creating Diversity in Ensembles Using Artificial Data," *Information Fusion*, vol. 6, no. 1, pp. 99-111, 2005

[7] G. Tremblay, R. Sabourin, and P. Maupin, "Optimizing Nearest Neighbour in Random Subspaces using a Multi-Objective Genetic Algorithm," *In Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)*, pp 208-211, 2004

[8] C. A. Shipp and L. I. Kuncheva, "Relationships Between Combination Methods and Measures of Diversity in Combining Classifiers," *International Journal of Information Fusion*, vol. 3, no. 2, pp. 135 - 148, 2002

[9] G. Brown, J. Wyatt, R. Harris and X. Yao, "Diversity Creation Methods: A Survey and Categorisation," *International Journal of Information Fusion*, vol. 6, no. 1, pp. 5-20, 2005

[10] N. Ueda and R. Nakano, "Generalization error of ensemble estimators," *In Proceedings of International Conference on Neural Networks (ICNN 1996)*, pp. 90-95, 1996

[11] B. E. Geman, S. and R. Dorsat, "Neural Networks and the Bias / Variance Dilemma," *Neural Computation*, no. 4, pp. 1-58, 1992

[12] A. Krogh and J. Vedelsby, "Neural Network Ensembles, Cross Validation, and Active Learning," *Advances in Neural Information Processing Systems*, vol. 7, pp. 231–238, 1995

[13] L. I. Kuncheva, M. Skurichina, and R. P. W. Duin, "An Experimental Study on Diversity for Bagging and Boosting with Linear Classifiers," *International Journal of Information Fusion*, vol. 3, no. 2, pp. 245-258, 2002

[14] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, no. 3, pp. 226–239, 1998

[15] R. E. Banfield, L. O. Hall, K. W. Bowyer and W. P. Kegelmeyer, "A New Ensemble Diversity Measure Applied to Thinning Ensembles," International Workshop on Multiple Classifier Systems (MCS 2003), pp. 306 - 316, 2003

[16] 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," *International Workshop on Multiple Classifier Systems (MCS 2002)*, pp. 62-71, 2002

[17] G. James, "Variance and Bias for General Loss Functions," *Machine Learning*, vol. 51, no. 2, pp. 115-135, 2003.

[18] P. Domingos, "A Unified Bias-Variance Decomposition and its Applications," *International Conference on Machine Learning (ICML 2000)*, pp. 231-238, 2000

[19] 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," *International Workshop on Multiple Classifier Systems (MCS 2004)*, pp. 273-282, 2004

[20] E. Pekalska, M. Skurichina and R. P. W. Duin, "Combining Dissimilarity-Based One-Class Classifiers," *International Workshop on Multiple Classifier Systems (MCS 2004)*, pp. 122-133, 2004

[21] D. Ruta and B. Gabrys, "Analysis of the Correlation between Majority Voting Error and the Diversity Measures in Multiple Classifier Systems," *In Proceedings of the 4th International Symposium on Soft Computing*, 2001

[22] R. E. Schapire, Y. Freund, P. Bartlett and W. S. Lee, "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998

[23] A. Grove and D. Schuurmans, "Boosting in the limit: Maximizing the Margin of Learned Ensembles," *In Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pp. 692-699, 1998

[24] R.P.W. Duin, "Pattern Recognition Toolbox for Matlab 5.0+," available free at: ftp://ftp.ph.tn.tudelft.nl/pub/bob/prtools

[25] D. Ruta and B. Gabrys, "Classifier Selection for Majority Voting," *International Journal of Information Fusion*, pp. 63-81, 2005

[26] G. Giacinto and F. Roli, "Dynamic Classifier Selection Based on Multiple Classifier Behaviour," *Pattern Recognition*, vol. 34, no. 9, pp. 179-181, 2001

[27] D. M. J. Tax, M. Van Breukelen, R. P. W. Duin, and J. Kittler, "Combining Multiple Classifiers by Averaging or by Multiplying," *Pattern Recognition*, vol. 33, no. 9, pp.1475-1485, 2000