Evolving Ensemble of Classifiers in Random Subspace

Albert Hung-Ren Ko
LIVIA
École de Technologie Supérieure
University of Quebec
1100 Notre-Dame West Street
Montreal, QC, H3C 1K3
Canada
albert@livia.etsmtl.ca

Robert Sabourin
LIVIA
École de Technologie Supérieure
University of Quebec
1100 Notre-Dame West Street
Montreal, QC, H3C 1K3
Canada
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

ABSTRACT

Various methods for ensemble selection and classifier combination have been designed to optimize the results of ensembles of classifiers. Genetic algorithm (GA) which uses the diversity for the ensemble selection could be very time consuming. We propose compound diversity functions as objective functions for a faster and more effective GA searching. Classifiers selected by GA are combined by a proposed pairwise confusion matrix transformation, which offer strong performance boost for EoCs.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*classifier design and evaluation*

General Terms

Pairwise Confusion Matrix Transformation Algorithm

Keywords

Fusion Function, Combining Classifiers, Diversity, Confusion Matrix, Pattern Recognition, Majority Voting, Ensemble of Learning Machines.

1. INTRODUCTION

Different classifiers usually make different errors on different samples, which means that we can arrive at an ensemble that makes more accurate decisions by combining classifiers [1, 9, 14, 16, 18, 27]. For this purpose, diverse classifiers are grouped together into what is known as an Ensemble of Classifiers (EoC). There are two problems in optimizing the performance of an EoC: first, how classifiers are selected, given a pool of different classifiers, to construct the best ensemble; and second, given all the selected classifiers, choosing the best rule to combine their outputs. These problems are fundamentally different, and should be solved separately to reduce the complexity involved in optimizing EoCs; the former focuses on ensemble selection [1, 9, 10, 15, 25] and the latter on ensemble combination, i.e. the choice of fusion functions [8, 14, 25].

Several important factors must be considered for an EoC: (a) find a pertinent objective function for selecting the classifiers; (b) use a pertinent searching algorithm to apply this criterion; and (c) use a adequate fusion function to combine classifier outputs. Diversity measures are designed as objective functions for ensemble selection [1, 2, 4, 6, 10], but their performances are not convincing. Moreover, when genetic algorithm (GA) is used as a searching algorithm for ensemble selection, the evaluation of non-pairwise diversity measures may be very time consuming. On the other hand, some different fusion functions have been suggested for combining classifiers [8, 9, 10, 11, 12, 14, 16, 25], but they are either based on strong assumptions [17, 20, 22], such as simple fusion functions, or required a large data set, such as trained fusion functions [9, 10, 11, 12]. Given insufficient training samples, simple fusion functions may outperform some trained fusion functions [22]. Here are the key questions that need to be addressed:

1. Can GA searching which uses the diversity be fast and effective for ensemble selection?
2. Can we take both the diversity and classifier accuracy into account in selecting classifiers?
3. Can a trained fusion function be effective without large training samples?
4. Can we take the interaction among classifiers into account in combining classifiers?

To answer these questions, we propose a method for selecting and combining classifiers (Fig. 1). Compound diversity functions (CDF) combine diversity measures with classification accuracy of individual classifiers in a pairwise manner, and thus allow fast and effective GA searching for ensemble selection. With the same fashion, pairwise confusion matrix (PCM) transforms an EoC into an ensemble of classifier pairs (Fig. 2). With the prospect of using classifier pairs, CDFs can more precisely maximize the diversity between classifier pairs, and given those diverse classifier pairs, PCM can obtain useful probabilities for classifier combination, transform the crisp class label outputs into class probability

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.

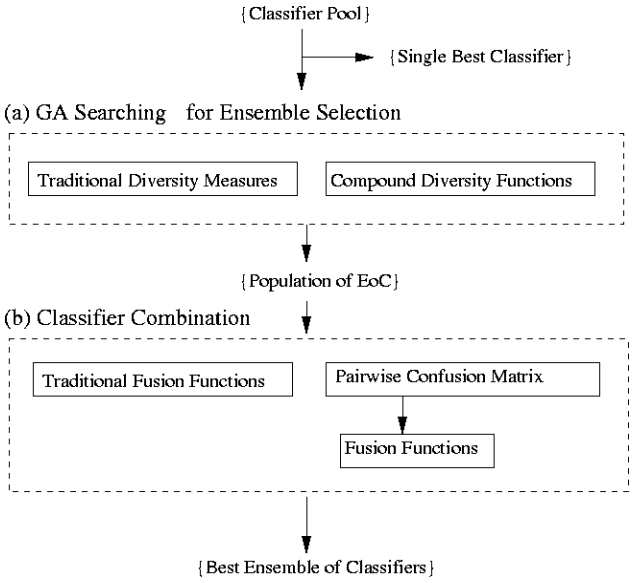

Figure 1: The EoC optimization approach includes ensemble selection and classifier combination. The ensemble selection is carried out by GA searching with various objective functions.

outputs and reduce the number of samples needed for ensemble training. With GA searching, experimental results suggest that the performance of a PCM can be a notch above that of the simple majority voting rule, and the performance of CDFs is apparently better than traditional diversity measures.

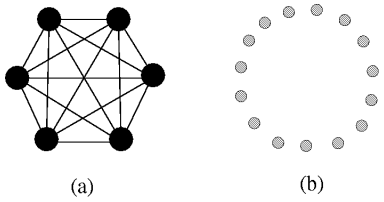

Figure 2: An example of pairwise confusion matrices transformation in a 6-classifier ensemble. (a) The original ensemble with 6 classifiers; and (b) the transformation yields to 15 classifier pairs, each classifier pair is equal to the link between two classifiers in (a).

The paper is organized as follows. The proposed ensemble selection method is presented in section 2. In section 3, the proposed pairwise confusion matrix is presented. Experimental results of both ensemble selection and classifier combination are compared in section 4. Discussion and conclusion are presented in the remaining sections.

2. OBJECTIVE FUNCTIONS FOR ENSEMBLE SELECTION

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, 3, 13, 14, 25]. This property is regarded as the diversity of an ensemble.

2.1 Traditional 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]:

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

   This type of diversity includes: Q-statistics (Q), the correlation coefficient (COR), the disagreement measure (DM) and the double fault (DF) [1, 2, 4, 15].

2. Non-Pairwise diversity measures
   There are other 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 (EN), Kohavi-Wolpert variance (KW), the measurement of interrater agreement (INT), the measure of difficulty (DIFF), generalized diversity (GD) and coincident failure diversity (CFD) [1, 6, 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]. For this reason, we propose the compound diversity functions (CDF) in the next section.

2.2 Compound Diversity Functions (CDF)

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:

$$\widehat{div_{amb}} = (\prod_{i=1}^{L}(1-a_i))^{\frac{1}{L}}(\prod_{i,j=1,i\neq j}^{L}(1-d_{i,j}))^{\frac{1}{L\times(L-1)}} \quad (2)$$

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 words, we expect $\widehat{div_{amb}}$ to be negatively correlated with ensemble accuracy, such as for DM, KW, EN, GD and CFD.

However, if the diversity measures represent the similarity, the proposed compound diversity function should be:

$$\widehat{div_{sim}} = (\prod_{i=1}^{L}(1-a_i))^{\frac{1}{L}}(\prod_{i,j=1,i\neq j}^{L} d_{i,j})^{\frac{1}{L\times(L-1)}} \quad (3)$$

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 made by all the classifiers. We expect negative correlation between the $\widehat{div_{sim}}$ and ensemble accuracy, such as for DF, INT, DIFF, Q and COR.

CDFs are based on diversity measured in a pairwise manner, even taking into account the individual classifiers error rates, and ensembles with fewer classifiers are more likely to be favored in ensemble selection. With regard to this effect, functions with various numbers of classifiers shall be rescaled by:

$$\widehat{div_{amb}} = \frac{L}{L-1}(\prod_{i=1}^{L}(1-a_i))^{\frac{1}{L}}(\prod_{i,j=1,i\neq j}^{L}(1-d_{i,j}))^{\frac{1}{L\times(L-1)}} \quad (4)$$

$$\widehat{div_{sim}} = \frac{L}{L-1}(\prod_{i=1}^{L}(1-a_i))^{\frac{1}{L}}(\prod_{i,j=1,i\neq j}^{L} d_{i,j})^{\frac{1}{L\times(L-1)}} \quad (5)$$

Using these CDFs, we can select ensembles by taking into account both diversity between classifiers and their classification accuracy. Moreover, since CDF applies a pairwise manner to measure the diversity, all diversity and all individual classifier accuracy can be measured beforehand, and the evaluation of GA calculates simply their product. It also assures that the use of PCM, which is also executed in a pairwise way, can enjoys the diversity between classifiers and explores the most available information.

3. FUSION FUNCTIONS FOR CLASSIFIER COMBINATION

3.1 Traditional Fusion Functions

Several simple fusion functions for combining classifiers have been proposed, such as Maximum Rule (MAX), Minimum Rule (MIN), Sum Rule (SUM), Product Rule (PRO) and simple Majority Voting Rule (MAJ) [14, 16, 17, 20, 26]. These directly compare the outputs from all individual classifiers in an ensemble, and do not require any further training. Some related theoretical studies are presented in [14, 16, 20]. These simple fusion functions are straightforward. Since they are relatively simple and do not explore the relationships between classifiers or those between classes, they are suboptimal [17], and, as stated in [5, 20], these fusion functions rely on the very restrictive assumption of the independence of estimates. To address this shortcoming, other, more sophisticated strategies have been proposed which use more available information in combining classifiers [9, 10, 11, 12], such as Naive Bayes (NB) [9, 16], Decision Templates (DT) [12], Behavior-Knowledge Space (BKS) and Wernecke's method (WER) [11].

Above all, we observe that most trained fusion functions tend to explore more information from the training set. For this reason, most classifier combination strategies need to take the interaction between classifiers and between classes into consideration. If these elements are ignored, as with NB, then the performance cannot be satisfactory. If these elements are fully explored, as with BKS or WER, given the complicated behavior of classifiers in an ensemble, especially in a high class dimension and with a large number of classifiers, the number of samples can scarcely be sufficient, and the probabilities obtained will usually be unreliable.

Herein lies the problem with training ensembles for combining classifiers. The fact that an ensemble acts in an extremely large space means that we need to use a method which is both effective and accurate. Given this dilemma, we propose a method which considers an ensemble of classifier pairs rather than an EoC. This will increase the number of members in ensembles, and thus is more stable by SUM rule. It also generates ensemble members with lower estimation error because of better information provided by classifier pairs, and thus offers improvement for MAJ rule. The proposed pairwise confusion matrix (PCM) transformation is a practical solution for both lowering the estimation error and the variance of the average estimation error, which involves the pairwise interaction between classifiers over their class label outputs. We detail this transformation and its use in combining classifiers in the next section.

3.2 Pairwise Confusion Matrix Transformation (PCM)

The dilemma of EoCs is that, given a limited number of samples, we need to take into account the interaction among classifiers. PCM makes use of pairwise estimation to solve this problem. If we only take classifier pairs into account, we need only calculate the probability $P(l|c(i),c(j),x)$, where $c(i)$ and $c(j)$ are the decisions of classifier $f(i)$ and classifier $f(j)$ over a sample x respectively. For $P(l|c(i),c(j),x)$, given T classes there are only $T\times T^2 = T^3$ different situations, and if the number of samples N is large enough, i.e. $N \gg T^3$, we can obtain a reliable estimation of this probability. This probability can be approximated by calculating PCM:

$$P(l|c(i),c(j)) = N(x\in l, c(i), c(j))/N(c(i),c(j)) \quad (6)$$

where $N(c(i),c(j))$ is the total number of samples on which classifier $f(i)$ gives crisp output $c(i)$, and classifier $f(j)$ gives crisp output $c(j)$, while $N(x\in l, c(i),c(j))$ is the number of samples the real class label of which is l, $1\leq l\leq T$. The probability $P(l|c(i),c(j),x)$ is, in fact, the concept of a 3-dimensional confusion matrix, where the decision of classifier $c(i)$, the decision of classifier $c(j)$ and the real class label of samples represent each dimension. For any sample x with a class label k, PCM provides a pairwise matrix of classifier $f(i)$ and classifier $f(j)$, with the probability of how likely it will be classified as class $c(i)$ by $f(i)$ and as class $c(j)$ by $f(j)$. For any sample x classified as class l by classifier $f(i)$, PCM provides a partial confusion matrix between classifier $f(j)$ and the real class labels of samples. All the confusion matrices of classifier $f(j)$ can be derived quickly from any pairwise confusion matrices concerning $f(j)$:

$$P(l|c(j),x) = \sum_{i=1}^{T} P(l|c(i),c(j),x) \quad (7)$$

where $c(i)$ constitutes the class label outputs of classifier $f(i)$. In other words, it is a cube of T^3 cells with N samples filled in; since L classifiers mean $\frac{L\times(L-1)}{2}$ classifier pairs, we can obtain $\frac{L\times(L-1)}{2}$ pairwise confusion matrices (PCM).

The probabilities from these pairwise confusion matrices offer several advantages over the traditional ensemble combination strategies: (a) they do not require the class probability outputs of each sample but only the class label outputs of each sample from individual classifiers; (b) they transform the simple class label outputs into the class probability outputs; and (c) they take into account of the interaction between classifiers. Note that the use of pairwise confusion matrices is a transformation, not an actual classifier com-

Table 1: UCI data for ensembles of classifiers. Tr = Training Samples; Ts = Test Samples; RS-Card. = Random Subspace Cardinality.

Database	Classes	Tr	Ts	Features	RS-Card.
Ionosphere	2	175	175	34	20
Liver-Disorders	2	172	172	6	4
Pima-Diabetes	2	384	384	8	4
Breast-Cancer	2	284	284	30	5
Wine	3	88	88	13	6
Image Segmentation	7	210	2100	19	4
Letter Recogntion	26	10000	10000	16	12

bination scheme. Based on these pairwise class probabilities, we can apply other different classifier combination rules. We give two examples of the application of PCMs in general fusion functions:

1. Pairwise Confusion Matrix using Sum Rule (PCM-SUM)
   Assign x $\rightarrow$ k if

$$\frac{2}{L \times (L-1)} \sum_{i,j=1, i>j}^{L} P(k|c(i), c(j), x) = \max_{l=1}^{T} \frac{2}{L \times (L-1)} \sum_{i,j=1, i>j}^{L} P(l|c(i), c(j), x) \quad (8)$$

2. Pairwise Confusion Matrix using Majority Voting Rule (PCM-MAJ)
   This rule is similar to the simple MAJ rule, but uses the pairwise probability $P(l|c(i), c(j), x)$ from the classifier pair $f(i)$ and $f(j)$ instead of the simple probability $P_i(l|x)$ from a single classifier $f(i)$ considering class l.

Other fusion functions, such as DT or NB, will require further training, but are applicable as well. To prove that PCMs are applicable, we carry out the experiments on classifier combination without ensemble selection in the next section.

3.3 Preliminary Experiments on Fusion Functions

To ensure that the PCM is useful for combining classifiers, we tested it on problems extracted from a UCI machine learning repository. There are several requirements for the selection of pattern recognition problems. First, the databases must have a large feature dimension for the Random Subspace method. 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 Subspace, 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 7 databases selected from the UCI data repository (see Table 1). Among available samples, in general, 50% are used as a training data set, and 50% are used as a test data set, except for the Image Segmentation dataset, whose training data set and test data set have been defined on UCI data repository. Of the training data set, 70% are used for classifier training and 30% are used for validation. Ensemble-training (including BKS, NB and PCM) used the entire available training data set. The cardinality of Random Subspace is set under the condition that all classifiers have recognition rates more than 50%.

The three different classification algorithms used in our experiments are K-Nearest Neighbors Classifiers (KNN), Parzen Windows Classifiers (PWC) and Quadratic Discriminant Classifiers (QDC) [24]. For each of 7 databases and for each of 3 classification algorithms, 10 classifiers were generated as the pool of classifiers. Among these, each classifier has a 50% chance of being selected from this pool to construct ensembles, ensembles were thus constructed by different numbers of classifiers, and at least three classifiers are required for an ensemble. As a result, all ensembles were constructed from $3 \sim 8$ classifiers. 30 ensembles had been generated for each database, for each ensemble generation method and for each classification algorithm. Note that each ensemble can have different number of classifiers. In total, we evaluated $30 \times 7 \times 3 = 630$ ensembles. We then combined these ensembles with 5 different fusion functions (Table 2).

Table 2: Comparison of recognition rates of different fusion functions with Random Subspace on UCI machine learning problems. F.F. = Fusion Functions for classifier combination. Liv.= Liver-Disorder Data; Iono. = Ionosphere Data; Imag. = Image Segmentation; Diab. = Pima Diabete; Canc. = Wisconsin Breast Cancer Data. Lett. = Letter Recogntion. All numbers are in percents ($\%$), the standard variances are indicated in parenthesis. Note that 3 classification algorithms were used and only average values are shown here.

F.F. $\rightarrow$	MAJ	NB	BKS	PCM -MAJ	PCM -SUM
Iono.	81.39 (0.09)	81.47 (0.06)	**90.75 (-)**	83.10 (0.06)	81.09 (0.07)
Liv.	63.90 (0.11)	56.53 (0.24)	**81.01 (0.04)**	65.28 (0.08)	64.96 (0.08)
Diab.	78.94 (0.16)	60.23 (0.60)	**83.68 (0.03)**	80.34 (0.06)	78.30 (0.05)
Canc.	93.54 (0.05)	93.68 (0.48)	92.14 (0.04)	**94.17 (0.03)**	93.54 (0.03)
Wine	84.42 (0.15)	89.96 (0.23)	**94.76 (0.13)**	90.30 (0.24)	88.82 (0.18)
Imag.	75.91 (0.51)	64.78 (2.88)	-	**85.31 (0.19)**	82.98 (0.17)
Lett.	84.24 (0.04)	90.72 (0.04)	-	**91.08 (0.09)**	85.56 (0.09)

In previous studies, BKS has been shown to be comparatively accurate when an ensemble of 3 classifiers is involved [23], but the BKS could be outperformed by most of the other fusion functions when more classifiers are involved [12]. In our study, the BKS apparently performs very well in 2- and 3-class problems. But when the class dimension is larger than 6, due to huge data size and limited computer memory we could not construct the BKS table.

We also observe that PCM-MAJ offers quite stable performance, in general better than that offered by the simple MAJ rule. The t-statistic test shows that the significance level is at 2.78%, so there is little chance for simple MAJ to perform as well as PCM-MAJ. Interestingly, we note that the difference in performance between PCM-MAJ and simple MAJ is somehow related to classifier diversity. In general, the greater the diversity of classifiers, the better PCM-MAJ can outperform simple MAJ. It is not difficult to understand that this property is in some way influenced by the types of classifiers used in experiments, because different classification algorithms lead to different levels of diversity among classifiers. Nevertheless, the ensembles tested were constructed by randomly selected classifiers without any ensemble selection procedure. To better understand the effect of fusion functions on real problems, we must test this rule on a high-class problem with a large data set. It is also advisable that we test different objective functions for ensemble selection. To affirm the use of both CDF and PCM, we need to carry out more experiments, we detail the further experiments in the next section.

4. EXPERIMENTS WITH GENETIC ALGORITHM (GA) SEARCHING IN RANDOM SUBSPACE

4.1 Experimental Protocol

We carried out experiments on a 10-class handwritten numeral problem. The data were extracted from $NIST\ SD19$, essentially as in [7], based on the ensembles of KNNs generated by the Random Subspace 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 ($hsf\_\{0-3\}$) to create 100 KNN in Random Subspace, we use relatively small size of data set to better observe the impact of EoC. The optimization set containing 10000 samples ($hsf\_\{0-3\}$) was used for genetic algorithm (GA) searching for ensemble selection. To avoid overfitting during GA searching, the selection set containing 10000 samples ($hsf\_\{0-3\}$) was used to select the best solution from the current population according to the objective function defined, and then to store it in a separate archive after each generation. The same selection set was also used for training fusion functions, including PCM transformation and the NB fusion function. Note that with 100 classifiers and 10 classes, BKS and WER would require constructing a table with 10^{101} cells, which is impossible to realize. Using the best solution from this archive, the test set containing 60089 samples ($hsf\_\{7\}$) was used to evaluate the EoC accuracies.

For the ensemble selection, we tested 3 kinds of different objective functions in this section. The majority voting error (MVE) was tested because of its reputation as one of the best objective functions in selecting classifiers for ensembles [25], it evaluates directly the global EoC performance by MAJ rule. In addition, we also tested 10 different traditional diversity measures and 10 different compound diversity measures which combine the pairwise diversity measures and individual classifier performance to estimate ensemble accuracy. Comparison of these two kinds of objective functions can also allow us to evaluate whether or not the direct use of ensemble accuracy for ensemble selection is adequate for further optimization on combining classifiers.

We tested 21 different objective functions, including Majority Voting Error (MVE) and 10 traditional diversity measures and respective compound diversity functions (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, 4, 6]) (Tables 4, 5 and 6). For fusion functions, because only crisp class outputs were obtained by KNN, MAJ, NB, PCM-MAJ and PCM-SUM were applied.

These objective functions are evaluated by GA searching. We used GA because the complexity of population-based searching algorithms can be flexibly adjusted depending on the size of the population and the number of generations with which to proceed. Moreover, because the algorithm returns a population of the best combinations, it can potentially be exploited to prevent generalization problems [25]. Parameters used in simple GA were set as follows. 100 genes were used and each gene indicates the inclusion or exclusion of each KNN classifier. Crossover probability was set to $p_c = 50\%$, and the mutation probability is set to $\frac{1}{L}$, where L is the number of length of the mutated binary string [19]. The maximum number of generations in GA was set to $m_g = 500$, and the size of the population was set to $s_p = 128$, which means that 64000 ensembles were evaluated in each experiment. With 30 replications,

Table 3: Table of Abbreviations.

Abbreviation	Indication
CDF	Compound Diversity Functions
EoC	Ensemble of Classifiers
GA	Genetic Algorithm
PCM	Pairwise Confusion Matrix
TDM	Traditional Diversity Measures
Abbreviation	**Classification Algorithms**
KNN	K-Nearest Neighbors Classifier
PWC	Parzen Windows Classifiers
QDC	Quadratic Discriminant Classifiers
Abbreviation	**Objective Functions**
COR	Correlation Coefficient
CFD	Coincident Failure Diversity
DF	Double-Fault
DIFF	Difficulty Measure
DM	Disagreement Measure
EN	Entropy Measure
GD	Generalized Diversity
INT	Interrater Agreement
KW	Kohavi-Wolpert Variance
MVE	Majority Voting Error
Q	Q-statistics
CDF-COR	Compound Diversity Function using Correlation Coefficient
CDF-CFD	Compound Diversity Function using Coincident Failure Diversity
CDF-DF	Compound Diversity Function using Double-Fault
CDF-DIFF	Compound Diversity Function using Difficulty Measure
CDF-DM	Compound Diversity Function using Disagreement Measure
CDF-EN	Compound Diversity Function using Entropy Measure
CDF-GD	Compound Diversity Function using Generalized Diversity
CDF-INT	Compound Diversity Function using Interrater Agreement
CDF-KW	Compound Diversity Function using Kohavi-Wolpert Variance
CDF-Q	Compound Diversity Function using Q-statistics
Abbreviation	**Fusion Functions**
MAJ	Simple Majority Voting Rule
MIN	Minimum Rule
MAX	Maximum Rule
PRO	Product Rule
SUM	Sum Rule
BKS	Behavior-Knowledge Space
DT	Decision Templates
NB	Naive Bayes
WER	Wernecke's Method
PCM-MAJ	Pairwise Confusion Matrix using Majority Voting Rule
PCM-SUM	Pairwise Confusion Matrix using Sum Rule

Table 4: Mean recognition rates of ensembles selected by MVE and combined with various fusion functions. All standard variation is smaller than 0.01%.

Fusion Functions → / Objective Functions ↓	MAJ	NB	PCM -MAJ	PCM -SUM
MVE	96.45 %	96.27 %	**96.94 %**	96.43 %

Table 5: Mean recognition rates of ensembles selected by *traditional diversity measures* and combined with various fusion functions. O.F. = Objective Functions for ensemble selection; F.F. = Fusion Functions for classifier combination. All standard variation is smaller than 0.01%.

O.F. → / F.F. ↓	CFD	COR	DM	DF	DIFF
simple MAJ	93.66 %	92.42 %	91.56 %	94.10 %	96.24 %
NB	93.93 %	93.52 %	92.86 %	94.31 %	96.12 %
PCM-MAJ	93.22 %	92.47 %	91.84 %	93.58 %	**96.63 %**
PCM-SUM	93.70 %	92.99 %	92.42 %	94.29 %	95.78 %
O.F. → / F.F. ↓	EN	GD	INT	KW	Q
simple MAJ	90.04 %	93.26 %	93.04 %	95.72 %	91.96 %
NB	91.81 %	94.15 %	94.15 %	96.07 %	93.12 %
PCM-MAJ	91.12 %	93.46 %	93.46 %	96.53 %	91.91 %
PCM-SUM	90.85 %	93.72 %	93.72 %	95.73 %	92.68 %

40.32 million ensembles were searched and evaluated. A threshold of 3 classifiers was applied as the minimum number of classifiers for an EoC during the whole searching process. The use of archive stored the best individual found in selection set.

We need to address the fact that the classifiers used were generated with feature subsets having only 32 features out of a total of 132. The weak classifiers can help us better observe the effects of EoCs. If a classifier uses all available features and all training samples, a much better performance can be observed [21]. But, since this is not the objective of this paper, we focus on the improvement of EoCs by optimizing fusion functions on combining classifiers.

4.2 Experimental Results

The benchmark KNN classifier uses all 132 features, and so, with $K = 1$ we can have 93.34% recognition rates. The combination of all 100 KNN by simple MAJ gives 96.28% classification accuracy, and gives 96.96% by PCM-MAJ. The possible upper limit of classification accuracy (the oracle) is defined as the ratio of samples which are classified correctly by at least one classifier in a pool to all samples. The oracle is 99.95% for KNN.

First we used MVE as the objective function for the ensemble selection. We applied different fusion functions and observed that PCM-MAJ performed better than MAJ and achieved 96.94% recognition rate (Table 4).

To verify that PCM-MAJ can offer comparable performance with other objective functions, we used traditional diversity measures for the ensemble selection. While some traditional diversity measures did show some improvement by using PCM, others deteriorated (Table 5). It is important to note that, except for KW, which always finds 17 classifiers for an EoC, and DIFF with 21 classifiers, all diversity measures selected ensembles made up of 3 classifiers. Since 3 classifiers were transformed into 3 classifier pairs by PCM, we can foresee that PCM cannot be of much more advantage for an ensemble with only 3 classifiers. Among all traditional diversity measures, DIFF gives the best performance with 96.63% recognition rate.

Table 6: Mean recognition rates of ensembles selected by *compound diversity functions (CDFs)* and combined with various fusion functions. ; F.F. = Fusion Functions for classifier combination. All standard variation is smaller than 0.01%.

O.F. → / F.F. ↓	CDF-CFD	CDF-COR	CDF-DM	CDF-DF	CDF-DIFF
simple MAJ	96.22 %	96.29 %	96.19 %	96.20 %	96.23 %
NB	95.78 %	95.77 %	95.79 %	95.76 %	95.80 %
PCM-MAJ	**96.88 %**	**96.88 %**	96.84 %	96.82 %	96.87 %
PCM-SUM	96.21 %	96.21 %	96.17 %	96.17 %	96.21 %
O.F. → / F.F. ↓	CDF-EN	CDF-GD	CDF-INT	CDF-KW	CDF-Q
simple MAJ	96.18 %	96.19 %	96.22 %	96.20 %	96.20 %
NB	95.75 %	95.75 %	95.81 %	95.74 %	95.79 %
PCM-MAJ	96.85 %	96.86 %	96.87 %	96.82 %	96.86 %
PCM-SUM	96.19 %	96.21 %	96.22 %	96.16 %	96.21 %

By contrast, the compound diversity functions (CDFs) are much more stable as objective functions (Table 6). Compare Table 5 and Table 6, we can see that the ensembles selected by CDFs are more stable than those selected by traditional diversity measures. However, 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. Among all CDFs, CDF-CFD and CDF-COR give the best performance with 96.88% recognition rate.

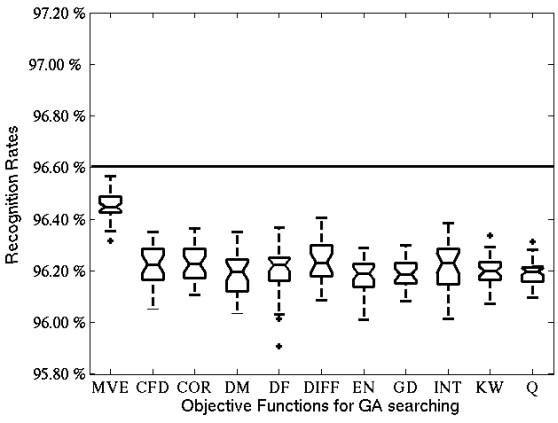

Figure 3: The recognition rates achieved by EoCs selected by 10 compound diversity functions (CDFs) and Majority Voting Error (MVE), using the *simple MAJ* as fusion function.

Comparing PCM-MAJ and the simple MAJ as fusion functions, we can see that in general PCM-MAJ offers a better performance than the simple MAJ (Fig. 3 and 4). For the ensembles selected by CDFs, a 95% confidence interval indicates the improvement of $0.6\% \sim 0.67\%$ in the recognition rates using PCM-MAJ.

Comparing different objective functions for the ensemble selection, the proposed CDFs do improve the performance of EoCs, and always perform better than the respective original diversity measures, their performances being much close to those ensembles ob-

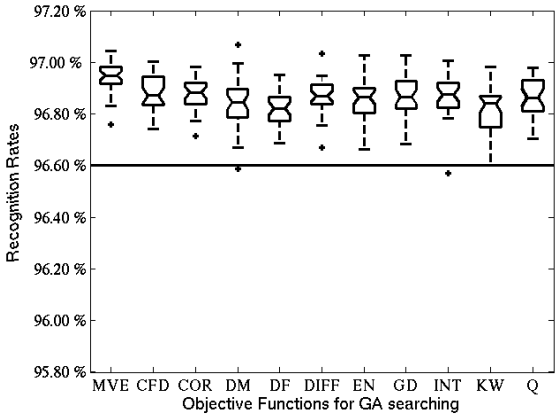

Figure 4: The recognition rates achieved by EoCs selected by 10 compound diversity functions (CDFs) and Majority Voting Error (MVE), using *PCM-MAJ* as fusion function.

Table 7: 95% Confidence Intervals indicate the differences of the recognition rates between MVE and other objective functions. TDM = Traditional Diversity Measures; CDF = Compound Diversity Functions.

Compared → Objective Functions	MVE outperforms TDM by	MVE outperforms CDF by
with simple MAJ	$3.59\% \sim 5.86\%$	$0.21\% \sim 0.28\%$
with PCM-MAJ	$3.39\% \sim 3.6\%$	$0.05\% \sim 0.12\%$

tained with the MVE objective function. On Table 7, we measured the difference of the performances in terms of 95% confidence intervals, and we see that with the simple MAJ, MVE outperforms traditional diversity measures by $3.59 \sim 5.86\%$, but MVE outperforms CDFs only by $0.21\% \sim 0.28\%$. When PCM-MAJ is used, then the difference between MVE and CDFs is even smaller, MVE is better than CDFs only by $0.05\% \sim 0.12\%$. 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, given that these CDFs do not take into account of any fusion functions, the ensemble outputs can be further optimized using various classifier-combining methods [14, 25]. As we can see that with PCM-MAJ ensembles selected by CDFs enjoy much more performance improvement than those selected by MVE.

Considering different ensemble selection schemes, and comparing the boost of PCM over simple MAJ, MVE so far enjoys a 0.49% boost on KNN classifiers and is still the best (Table 4). But CDFs provide better improvement, up to 0.67% (EN) (Table 6), even though their final results are not quite as good as those of MVE as objective functions.

Until recently, there have been few other fusion functions that perform better than simple MAJ for crisp class output classifiers. But, when PCM transformation is carried out, and those classifier pairs from ensembles are evaluated by PCM-MAJ, we observe a boost in the recognition rates of EoCs, the results achieved by PCM-MAJ being a notch above those of simple MAJ. This affirms the improvement brought about by PCM (See Figs. 3 and 4).

5. DISCUSSION

For EoCs, the ideal is to obtain the probability $P(l|c(1), \cdots, c(i), \cdots, c(L), x)$ for the whole data set X, where l is the possible class label, and $c(1), \cdots, c(i), \cdots, c(L)$ are decisions of individual classifiers $f(1), \cdots, f(i), \cdots, f(L)$ respectively. But, in reality, this approach might not work owing to the limitation with respect to the number of samples. Instead of estimating $P(l|c(1), \cdots, c(i), \cdots, c(L), x)$, the proposed PCM deals with the probability $P(l|c(i), c(j), x)$ from pairwise confusion matrices on an evaluated class l, and thus is much more applicable, while at the same time taking into account classifier interaction.

When no class probability outputs are provided, most simple fusion functions, such as MAX, MIN, SUM and PRO, cannot be applied. The only available simple fusion function is the simple MAJ. For trained fusion functions, DT requires the class probability outputs from classifiers, and to deal with a problem involving crisp class label outputs, only NB or BKS, WER are applicable. However, for high-class dimension problems and large-size ensembles, there is no way to use BKS or WER, e.g. a 10-class problem with 100 classifiers requires the construction of a table with 10^{101} cells. On all selected UCI machine learning problems, PCM-MAJ almost always outperforms simple MAJ as a fusion function for combining classifiers. Moreover, the difference in performance between PCM-MAJ and simple MAJ is to some extent correlated with the diversity of ensembles, especially when KNN is used in Random Subspace.

Considering objective functions for ensemble selection, CDFs do give stable and better performance than traditional diversity measures, and they enjoy a strong boost when PCM is applied. Another advantage of CDFs 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.

6. CONCLUSION

In this paper, we use compound diversity functions for ensemble selection, which take into account the ensemble diversity and individual classifier classification accuracy in a pairwise manner, and pairwise confusion matrix for classifier combination, which transforms crisp class label outputs into class probability outputs and thus takes into account the interaction of classifiers in a pairwise manner. To conclude, the proposed method has some significant advantages:

1. In general, CDFs allows a fast and effective GA searching for ensemble selection.

2. In general, PCM offers a strong performance boosting for ensembles selected by CDFs.

3. Because of its pairwise nature, it does not need too many samples for training compared with BKS or WER.

The experiment reveals that the performance of PCM is promising. Intuitively, PCM can also be used for other trained fusion functions, such as NB or DT. This will require another training data set, but we are interested in investigating the potential use of PCM in improving the performance of trained fusion functions.

The key element that makes an ensemble of classifier pairs outperform an EoC is that the use of PCM takes the interaction into consideration. The pairwise manner may still be sub-optimal, but,

if the class dimension is low and we have few classifiers and a large number of samples, PCM can be upgraded to the third degree, i.e. we can obtain the probabilities of any class label l by calculating $P(l|c(i), c(j), c(h), x)$ based on three classifier outputs $c(i), c(j), c(h)$. This would require the construction of 4-dimensional confusion matrices and allow us to interpret the interaction of three classifiers at the same time. Another possible improvement scheme would be the use of PCM-MAJ as an objective function for ensemble selection. In the same way that simple MAJ is used for ensemble selection (i.e. MVE) and for classifier combination, one can also apply PCM-MAJ for both ensemble selection and classifier combination.

Given that this exploratory work has been accomplished 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 classifier combination as well as ensemble selection, with more pattern recognition problems and more classification methods. We carried out experiments only in Random Subspace as ensemble creation method, and it will be of great interest to measure the impact of the proposed PCM and CDF on boosting and bagging as well.

Acknowledgment

This work was supported in part by grant OGP0106456 to Robert Sabourin from the NSERC of Canada. We would like to thank E. dos Santos and L. S. Oliveira for contributing data.

7. 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] 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

[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] K. Turner and J. Ghosh, "Error Correlation and Error Reduction in Ensemble Classifiers," *Connection Science*, vol. 8, no. 3-4, pp. 385-404, 1996

[6] J. L. Fleiss, B. Levin, and M. C. Paik, "Statistical Methods for Rates and Proportions," Second Edition, New York: John Wiley & Sons, 2003

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

[8] 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

[9] 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

[10] Y. S. Huang and C.Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 90-93, 1995

[11] K. D. Wernecke, "A coupling procedure for discrimination of mixed data," *Biometrics*, vol. 48, pp. 97-506, 1992

[12] L.I. Kuncheva, J.C. Bezdek and R.P.W. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern Recognition*, vol. 34, no. 2, pp. 299-314, 2001

[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] L. Xu, A. Krzyzak and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418-435, 1992

[17] R. P. W. Duin, "The Combining Classifier: To Train or Not to Train?" *16th International Conference on Pattern Recognition (ICPR)*, vol. 2, pp. 20765, 2002

[18] L. K. Hansen, C. Liisberg and P. Salamon, "The error-reject tradeoff," *Open Systems and Information Dynamics*, vol. 4, pp. 159-184, 1997

[19] A. E. Eiben, R. Hinterding, and Z. Michalewicz, "Parameter control in evolutionary algorithms", *In IEEE Transactions on Evolutionary Computation*, vol.3, no. 2, pp. 124-141, 1998

[20] L. I. Kuncheva, "A Theoretical Study on Six Classifier Fusion Strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281-286, 2002

[21] J. Milgram, M. Cheriet and R. Sabourin, "Estimating Accurate Multi-class Probabilities with Support Vector Machines," *International Joint Conference on Neural Networks 2005 (IJCNN 2005)*, pp. 1906-1911, 2005.

[22] S. Raudys, "Experts' boasting in trainable fusion rules," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1178 - 1182, 2003

[23] 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

[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] J. Kittler and F. M. Alkoot, "Relationship of Sum and Vote Fusion Strategies," *Multiple Classifier Systems (MCS)*, pp. 339-348, 2001

[27] 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," *Multiple Classifier Systems (MCS)*, pp. 273-282, 2004