Ambiguity-Guided Dynamic Selection of Ensemble of Classifiers

Eulanda M. dos Santos
Automated Production Engineering
École de technologie supérieure
Montreal, Canada
Email: eulanda@livia.etsmtl.ca

Robert Sabourin
Automated Production Engineering
École de technologie supérieure
Montreal, Canada
Email: Robert.Sabourin@etsmtl.ca

Patrick Maupin
Defence Research and Development Canada
DRDC Valcartier
Valcartier, Canada
Email: Patrick.Maupin@drdc-rddc.gc.ca

***Abstract*— Dynamic classifier selection has traditionally focused on selecting the most accurate classifier to predict the class of a particular test pattern. In this paper we propose a new dynamic selection method to select, from a population of ensembles, the most confident ensemble of classifiers to label the test sample. Such a level of confidence is measured by calculating the ambiguity of the ensemble on each test sample. We show theoretically and experimentally that choosing the ensemble of classifiers, from a population of high accurate ensembles, with lowest ambiguity among its members leads to increase the level of confidence of classification, consequently, increasing the generalization performance. Experimental results conducted to compare the proposed method to static selection and DCS-LA, demonstrate that our method outperforms both DCS-LA and static selection strategies when a population of high accurate ensembles is available.**

Keywords: ensemble of classifiers, dynamic selection, diversity measures, genetic algorithms.

I. Introduction

The application of an ensemble creation method, such as bagging [1], boosting [2] and random subspace [3], generates a set of classifiers C, where $C = \{C_1, C_2, \ldots, C_n\}$. Given such a pool of classifiers, the selection of classifiers has focused on finding the most relevant subset of classifiers L, rather than combining all available n classifiers, where $|L| \leq |C|$. Indeed, the selection of classifiers relies on the idea that either each classifier member is an expert in some local regions of the feature space [4] or component classifiers are redundant [5]. When an ensemble of classifiers (EoC) is selected, such a selection process includes fusion since the selected EoC must be combined.

Classifier selection techniques may be divided into two categories: static and dynamic. In the first case, the best subset of classifiers found during the training phase is used for the classification phase [5], [6]. In the second case, the selection is made during the classification phase taking into account the characteristics of the sample to be classified [4], [7]. The static selection of classifiers task may suffer from a problem: a fixed EoC defined during the training phase may not be well adapted for the whole test set. This problem is similar to searching for a universal best individual classifier, i.e., due to differences among samples, there is no an individual classifier perfectly adapted for every test sample. In addition, as stated by the "No Free Lunch" theorems [8], no algorithm may be assumed to be better than any other algorithm when averaged over all possible class of problems.

Another problem on static selection of EoCs occurs when population-based approaches, especially Pareto-based algorithms, are used as search algorithms during the selection process. Pareto-based algorithms, for instance, NSGA-II - elitist non-dominated sorting genetic algorithm [11], use Pareto dominance to solve multi-objective optimization problems (MOOP) such as the selection of EoCs guided by diversity and classification accuracy as a pair of objective functions [5]. A Pareto front is a set of nondominated solutions representing different tradeoffs between the multiple objective functions. When one solution is better according to one objective, it is often worse according to the other objective functions. Due to the different tradeoffs over the Pareto front, it is difficult to choose the best EoC to classify the test data set unless only one objective function is taken into account to point out the best solution. This is a persistent problem in MOOP applications. In [5], for example, the solution with highest recognition rate was chosen to classify the test samples even though the optimization process was guided by both diversity and recognition rate measures.

In dynamic classifier selection, rather than using a fixed EoC, selection is done based directly on the test patterns. The goal is to select systematically at least one classifier that can be able to correctly classify each test sample. The dynamic classifier selection process is traditionally divided into two phases: a partition generation procedure and the selection of the best classifier for each partition. In [9], local regions are determined according to the features values from an independent data set. The performance of each classifier from the pool is measured for each partition. Then, during the test phase, the most accurate classifier into the regions sharing the same features values with the test sample is selected to label it.

Woods et al. [4] and Giacinto and Roli [7] proposed dynamic classifier selection approaches based on estimating local competence into selection regions defined by k nearest neighbors (kNN). In [4] the DCS-LA (Dynamic Classifier Selection with Local Accuracy) algorithm is presented. Feature subset selection was applied to create different training sets

TABLE I

COMPILATION OF SOME OF THE METHODS REPORTED IN THE DYNAMIC CLASSIFIER SELECTION LITERATURE.

Reference number	**Classifier Members**	**Ensemble Construction**	**Partition Method**
[4]	Heterogeneous	Feature subset selection	kNN
[7]	Heterogeneous	Feature subset selection	kNN
[9]	Decision Trees	Different blocks of samples	subset of features values

for each classifier member (5 heterogeneous classifiers) and the individual opinions were combined using majority voting. During the test phase, the algorithm defines local regions as the set of kNN in the training data set surrounding each test sample. Thus, an estimation of local accuracy of each classifier is performed and the most locally accurate classifier is then selected to estimate the true class of the test pattern. The algorithm proposed in [7] is very similar to Woods' method. The difference relies on the definition of the local regions. Once the kNN of the test sample are identified, the local region used to estimate the individual performances of the classifiers is defined as the nearest neighbors sharing similar multiple classifier behavior with the test sample to be classified. Such a multiple classifier behavior is computed as a vector of the classes assigned by each individual classifier to a particular test sample.

However, since only one classifier is used to label a test sample, these methods may face a drawback: when the local expert does not correctly classify the test pattern, such a misclassification can not be avoided [12]. Therefore, the selection of an EoC instead of only one classifier could decrease misclassification. In addition, other disadvantages of methods based on selection regions defined by kNN, such as DCS-LA, can be pointed out: the parameter k and the distance function, which must be predetermined and the presence of the noise in the neighborhood of each sample can affect the performance of the method. Moreover, they are very time consuming processes. Table IV summarizes methods reported in the literature dealing with the dynamic selection of classifiers. It is important to see that all of these methods are conducted using a partition generation procedure and the selection of the best classifier for each partition

In this paper, we propose a new method for dynamic selection of EoCs. The proposed method relies on choosing dynamically from a population of EoCs the solution with largest extent of consensus, i.e. lowest ambiguity among the outputs of the classifiers, to predict the true class of the test pattern. The objective is to overcome both static and dynamic selection drawbacks mentioned before, namely the use of a fixed EoC to classify the whole test data set in static selection and the use of only one local expert to classify each test sample in dynamic selection. We attempt to prove theoretically and experimentally that choosing the EoC, from a population of high accurate EoCs, with lowest ambiguity among its members leads to increase the "degree of certainty" of the classification [14], consequently, increasing the generalization performance.

The random subspace method is applied to generate an initial pool of 100 Decision Trees (DT)-based classifiers. Thus, the populations of EoCs are generated by population-based genetic algorithms in optimization processes using both single- and multi-objective functions. Five objective functions are used to generate the populations of EoCs. These objective functions comprise four diversity measures and the ensemble's combined error rate ($1-$recognition rate). For multi-objective problems we avoid choosing only one solution over the Pareto front. Instead, a best solution is dynamically selected according to the ambiguity presented by each EoC. In addition, the DCS-LA method is adapted to the context of populations of EoCs to be compared to our method by experiments on three different databases. These experiments show that our method is superior to DCS-LA, besides to avoid the DCS-LA's disadvantages mentioned in this introduction.

The following section depicts our proposed new dynamic selection method. Then, section III describes how DCS-LA has been adapted to EoC selection problems. The experimental protocol employed is described in section IV. Finally, the experiments and the results are presented in section V. Conclusions and suggestions for future work are discussed in section VI.

II. AMBIGUITY-GUIDED DYNAMIC SELECTION OF EOC

In this section we describe the task of selecting an EoC from a population of ensembles and the method we propose to solve it.

A. Problem definition

Given a set of classifiers $C = \{C_1, C_2, \ldots, C_n\}$ which have been already trained, search algorithms are used to search for the best subset of classifiers. When dealing with such an optimization problem, two important aspects should be analyzed: the search criterion and the search algorithm [6]. The combined error rate, diversity measures and ensemble size are search criteria traditionally employed [5], [6]. Evolutionary algorithms are attractive search algorithms since they appear to fit well with the selection of classifier problems in the context of EoC optimization. Moreover, Ruta and Gabrys [6] observe that population-based GAs are good for classifier selection problems because of the possibility of dealing with a set of solutions rather than only one, which can be important in performing a post-processing phase. The dynamic selection method proposed in this paper is such a post-processing strategy.

In single-objective optimization problems, the best solution according to the objective function used to guide the search is often picked up to classify the test set. However, the selection task is more difficult when a Pareto front is involved. When the optimization of EoCs is guided by multi-objective functions, for instance the classification error rate and diversity measures as a pair of objective functions, Pareto-based evolutionary algorithms are solutions to such a kind of problem. These algorithms use Pareto dominance to perform the reproduction of the individuals.

A Pareto front is a set of nondominated solutions representing different tradeoffs between the multiple objective functions. A solution $p^{(1)}$ is said to dominate solution $p^{(2)}$, $p^{(1)} \preceq p^{(2)}$, if $p^{(1)}$ is no worse than $p^{(2)}$ on all the objective functions and $p^{(1)}$ is better than $p^{(2)}$ in at least one objective function. Figure 1 shows the Pareto front produced using the following pair of objective functions: minimize the error rate and maximize diversity (ambiguity as defined in [13]). Each circle on the plot corresponds to an EoC.

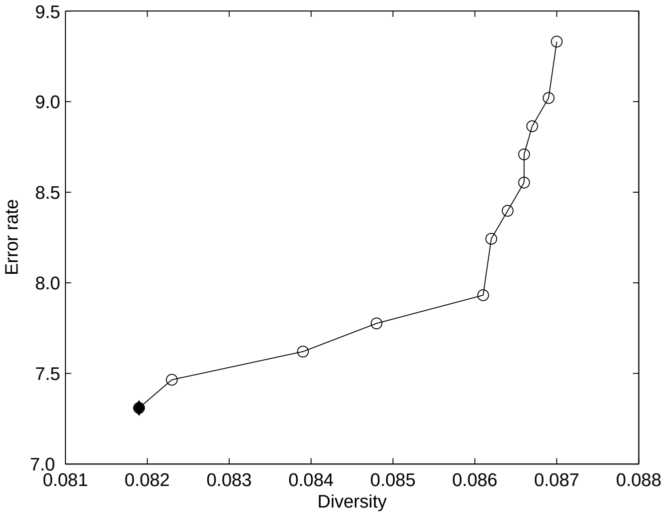

Fig. 1. The Pareto front produced by NSGA-II

Although the solutions over the Pareto front are equally important, the EoC with lowest error rate (black circle in Figure 1) is usually chosen to classify the test samples in a static selection approach. We propose in this paper an ambiguity-guided dynamic selection method to dynamically choose the best solution to classify individually each test sample, instead selecting the solution with lowest error rate. Such ambiguity measure is not related to the objective functions used to generate the population of solutions. Ambiguity used during the optimization process is a global diversity measure because it is calculated for the whole data set. Ambiguity used to dynamically select the best EoC to classify the test samples is a local ambiguity since it is calculated for each test pattern individually. Therefore, our local ambiguity measure is a selection criteria employed whatever objective functions used to guide the optimization, as described in next section.

B. Proposed method

The reject mechanism for an EoC is based on the extent of consensus among its members. The decision to reject a pattern is related to the confidence level of the EoC. Indeed, it is better to reject the pattern where the EoC presents low confidence level to take a decision. Such confidence level is related to the *a posteriori* probability $P(z_i|x)$ that a particular pattern x comes from class z_i. Hansen et al [14] have observed that $P(z_i|x)$ on the context of EoC, may be calculated based on the extent of consensus as follows: given an EoC C of n classifiers, where $C = \{C_1, C_2, \ldots, C_n\}$, let $v(i|x)$ be the number of votes for class z_i given x, $P(z_i|x)$ can be calculated as:

$$P(z_i|x) = \frac{v(I(x)|x)}{n} \tag{1}$$

where,

$$I(x) = \arg_{i=1}^{n} \max v(i|x) \tag{2}$$

It is important to note that Equation 1 is the complement of the ambiguity diversity measure calculated for a particular pattern. The classification ambiguity, proposed by Zenobi and Cunningham [13] is defined as:

$$a_i(x) = \begin{cases} 0 & \text{if } \text{Class}V_i(x) = \text{Class}\bar{V}(x) \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

where a_i is the ambiguity of the i^{th} classifier on the observation x, Class$V_i(x)$ is the class determined by the classifier and Class$\bar{V}(x)$ is the ensemble output. The ambiguity of the ensemble is:

$$\bar{A} = \frac{1}{|X|.|C|} \sum_{i \in C} \sum_{x \in X} a_i(x) \tag{4}$$

where C denotes an EoC and X the data set, and $|C|$ and $|X|$ their respective cardinalities.

If we define $|X| = 1$, i.e., ambiguity calculated for each pattern isolated, ambiguity $\bar{A}$ becomes the complement of the extent of consensus in Equation 1, thus $\bar{A} + P(z_i|x) = 1$. Therefore, when such a local ambiguity calculated for each sample is minimized, the extent of consensus is maximized. Taking into account the majority voting as the combination function, ambiguity measures the number of classifiers in disagreement with the majority voting whereas the extent of consensus measures the number of classifiers in agreement with the majority voting. Consequently, the certainty of correct classification of the EoC is maximized. Another important point to observe is that Equation 4 can be calculated in dynamic selection of EoCs because no information on the correctness of the output is needed.

When a correct classification takes place, a minimum ambiguity value leads to maximizing the margins, which is related to the certainty of the EoC's classification. Even though ambiguity measures the variance of ensemble around the majority voting without take into account the label of the sample. Let $\mu_i(x), i = 1, \ldots, z$ be the degree of support for

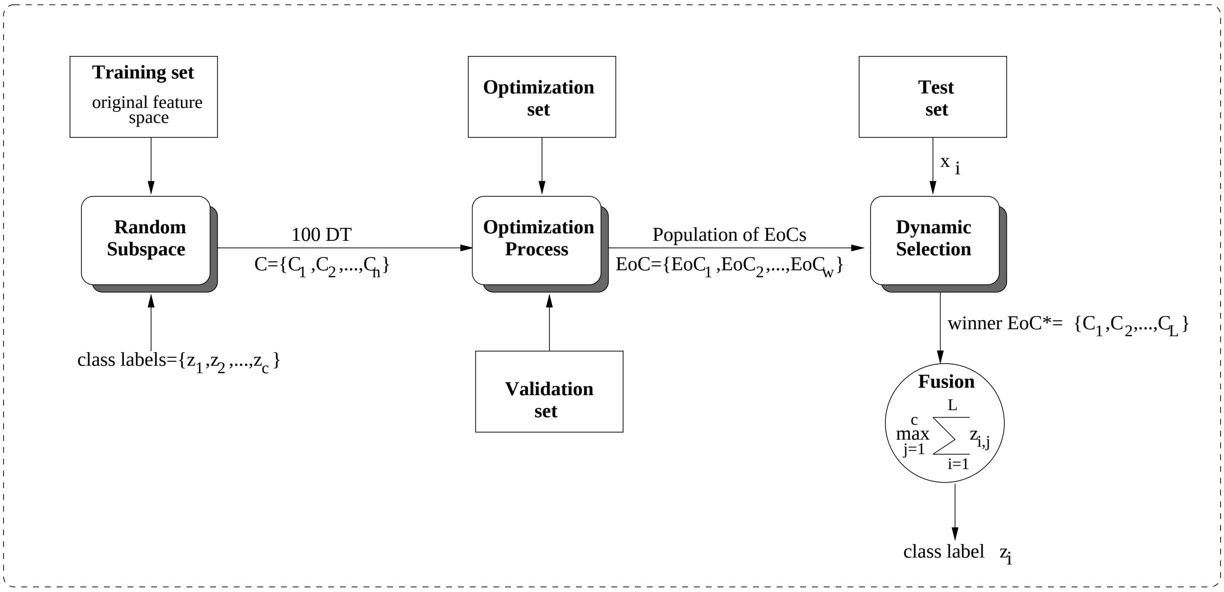

Fig. 2. Overview of the selection process

each class, the margin of sample x is calculated for a z classes' problem as follows:

$$m(x) = \mu_k(x) - \sum_{i \neq k} \mu_i(x) \qquad (5)$$

where k is the real class label of x. Hence, the maximization of the margins leads to increase the certainty of the classification.

Although there is an agreement in the literature related to the important role played by diversity, based on the standpoints presented in this section, we propose a dynamic selection method to select the EoC which presents the smallest ambiguity value to classify the given test pattern. The assumption is that an EoC with low ambiguity presents a high confidence level of classification.

We may now summarize the ambiguity-guided dynamic selection algorithm as follows:

1) Perform the design of a pool of classifiers $C = \{C_1, C_2, \ldots, C_n\}$.
2) Perform an optimization of EoCs task using a search algorithm in order to generate a population of w solutions $EoC = \{EoC_1, EoC_2, \ldots, EoC_w\}$ from the initial pool of classifiers C.
3) For each test sample x_i do:
   - If all w EoCs agree on the label, then classify x_i assigning it the consensus label.
   - If the population of w disagrees on the label, calculate ambiguity according to Equation 4 with $|X| = 1$ for each EoC from the population;
   - Select the EoC with lowest ambiguity value $EoC*$ to classify x_i.

The tie-break procedure is the same as employed in [4] for DCS-LA. It is carried out in the following way:

1) The majority voting among all EoCs with equal competence (same ambiguity value) is calculated and assigned to x_i.
2) If a winner vote cannot be pointed out, the EoC with the second lowest ambiguity value is identified to break the tie.
3) If the tie cannot be broken using the second best EoC, randomly select an EoC to classify x_i.

Figure 2 shows an overview of the algorithm. First, the training set is used by the random subspace method to generate a pool of classifiers C. In this paper, as mentioned in section IV, it is generated a DT-based pool with 100 classifiers. Second, the pool C of classifiers is the input to an optimization process, in this paper it is performed by GAs guided by both single- and multi-objective functions. As explained in IV, an optimization and a validation data sets are used during the optimization process. The population of EoCs is the output of the optimization process. Finally, the dynamic selection process takes place by selecting for each test sample the $EoC*$ with lowest ambiguity. Then, $EoC*$ is combined by majority voting. Majority voting of an EoC with L classifiers, is calculated for a problem with z classes as:

$$z_i = \max_{j=1}^{z} \sum_{i=1}^{L} z_{i,j} \qquad (6)$$

Our dynamic selection method is illustrated in Figure 3. In this example, the output of the optimization process is a Pareto front. $\bar{A}$ is calculated for each solution over the Pareto front. Thus, the solution $EoC*$ with lowest $\bar{A}$ is selected and combined to assign the majority voting class z_i to the test sample x_i.

We have adapted the DCS-LA method to the context of populations of EoCs to be able to compare the ambiguity-guided dynamic selection method proposed in this paper to DCS-LA. The DCS-LA procedure for populations of EoCs is

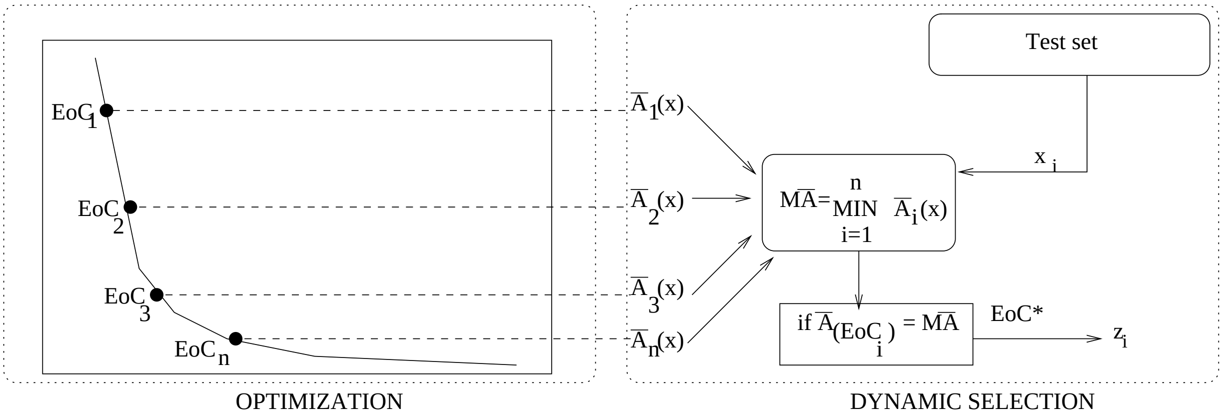

Fig. 3. Ambuiguity-guided dynamic selection using a Pareto front as input.

described in next section.

III. Dynamic ensemble selection with local accuracy

The DCS-LA method was proposed by Woods et al. [4] to dynamically selects the most accurate classifier from a population of classifiers, to predict the label of the test sample. The local accuracy was measured as the kNN of the test sample. They proposed two strategies to measure the local accuracy: 1 - overall local accuracy and 2 - local class accuracy. Since they concluded that the second strategy achieved the best results we use the local class accuracy in this paper.

Given the set of classifiers $C = \{C_1, C_2, \ldots, C_n\}$, let x be the test sample and z the class assigned by classifier C_i to this particular sample, then $C_i(x) = z$, the local class accuracy estimation is computed as:

$$LA_{C_i}(x) = \frac{K_z}{\sum_{i=1}^{n} K_{iz}} \tag{7}$$

where K_z is the number of neighbors of x from the training or validation set for which classifier C_i has correctly assigned class z and $\sum_{i=1}^{n} K_{iz}$ is the total number of neighbors labeled for C_i as class z.

However, DCS-LA cannot be directly employed in problems involving populations of EoCs because it has been originally proposed to deal with populations of classifiers. Therefore, we propose to slightly change the DCS-LA with local class accuracy estimation to such a different problem in the following way: given a population of w EoCs $EoC = \{EoC_1, EoC_2, \ldots, EoC_w\}$ which have been already optimized, x the unknown test sample, z the class assigned by $EoC_i = \{C_1, C_2, \ldots, C_L\}$ to sample x, $|EoC_i| = L$ is the cardinality of the ensemble, local EoC class accuracy is estimated as:

$$LA_{EoC_i}(x) = \frac{\sum_{i=1}^{L} LA_{C_i}(x)}{L} \tag{8}$$

Summarizing, in this paper, the local EoC class accuracy for pattern x is calculated as sum of the proportion of neighbors from the validation data set, surrounding x for each EoC's member divided by the cardinality of the ensemble.

TABLE II
Specifications of the databases used in the experiments.

Data set	Number of samples	Number of features	Features RSS k=1	Pool size
dna	3186	180	45	100
satimage	6435	36	18	100
texture	5500	40	20	100

IV. Experimental Protocol

To carry out the experiments, relatively high-dimensional feature space is necessary due to random subspace ensemble creation method. This aspect was taken into account when selecting the databases used in our experiments. We use the following three databases described in Table II: dna, texture and satimage.

In order to avoid overfitting during the optimization process we employ the same global validation strategy described in [5]. According to this strategy, to conduct the selection of EoCs performed as an optimization process by a search algorithm, one needs at least four different data sets: 1 - a training data set (TRAIN) to train the base classifiers in order to generate the initial pool of classifiers C; 2 - an optimization data set (OPT) used by the search algorithm to calculate fitness; 3 - a validation data set (VAL) to validate the solutions found during the optimization process and 4 - a test data set (TEST) to measure the generalization performance.

Thus, to be able to conduct experiments using the three small databases described in Table II using the global validation strategy, all experiments in this paper were carried out using 10-fold cross-validation. The original data set is separated into 10 folds, one fold for TEST, a second fold for VAL, a third one for OPT and the remaining 7 folds are combined to compose TRAIN. Then, TRAIN is used to generate the initial pool of classifiers C, OPT is used by the search algorithms to the optimization process and VAL is used to perform the global validation strategy. This process is repeated for each fold and the final error rates are reported as the average over 10. It is important to mention that each optimization process was replicated 30 times and the result of each replication was computed. The results reported in Table V were obtained as the mean of the error rates from each fold over 10 and the final result is the mean of the error rates over 30 replications.

An ensemble of 100 DT classifiers trained on each TRAIN fold was generated using the random subspace method. The C4.5 algorithm [17] (Release 8) was used to construct the trees in this paper. The EoCs were combined by majority voting. The size of the subsets of features used by random subspace is shown in Table II.

The error rate and diversity measures were applied to guide the optimization process. We employ 4 of the diversity measures proposed in the literature: coincident failure diversity (CFD), difficulty measure (Dif) and double-fault (DF) [15], plus ambiguity (A) (as defined in [13]). It is worth noting that *dissimilarity* measures such as A and CFD, must be maximized, while *similarity* measures such as Dif and DF, must be minimized when used as objective functions during the optimization process.

The population-based evolutionary algorithms used in this work are both single- and multi-objective genetic algorithm (MOGA). Since, single-objective genetic algorithm (GA) can find few solutions in its last population, we fixed a size of the population to be the input for our ambiguity-guided dynamic selection method. In this way, instead picking up only the best solution, we find the N best solutions obtained on validation data set to compose the population of EoCs for the dynamic selection step. We set $N = 21$ in our experiments. When MOGA is used, we pick up all solutions over the Pareto front to be the input of the dynamic selection process.

Among several Pareto-based evolutionary algorithms proposed in the literature, NSGA-II [11] appears to be interesting due two important characteristics: a full elite-preservation strategy and a diversity-preserving mechanism using the crowding distance as the distance measure. The crowding distance does not need any parameter to be set [11]. The elitism concept is used to provide means to keep the good solutions among generations and the diversity-preserving mechanism is used to allow a better spread among the solutions over the Pareto front. Based on these characteristics, in our experiments MOGA used is NSGA-II.

The selection of EoCs is applied in the context of GAs based on binary vectors. Since we use baselines EoC composed of 100 classifiers, each individual is represented by a binary vector with a size of 100. Each bit determines whether a classifier is active (1) or not (0). We defined the same genetic parameters employed in [5]. Table III shows the parameter settings employed. The same parameters were used for both single- and multi-objective GAs.

TABLE III

GENETIC ALGORITHMS PARAMETERS

Population size	128
Number of generations	1000
Probability of crossover	0.8
Probability of mutation	0.01

V. EXPERIMENTS AND DISCUSSION

In our experiments, GAs are employed to generate populations of EoCs to be the input to the dynamic selection process. The error rate is applied individually as single-objective function using GA and the diversity measures are used in pairs of objective functions combined with the error rate for NSGA-II. The objective is to point out the best selection strategy, for instance, static or dynamic, as well as to compare our ambiguity-guided dynamic selection method to DCS-LA. We set $k = 10$ for experiments with DCS-LA as employed in [4]. The static selection results were obtained by picking up the best solution obtained on validation data set by GA and calculating the solution's error rate on test data set. For NSGA-II, the EoC presenting the lowest error rate over the Pareto front was chosen as the static solution used to classify the whole test data set.

Table IV presents some of the results reported in the literature dealing with the selection of classifiers, except for reference [3], on the databases used in this paper. In this way it is possible to have an overview about the results obtained in the literature, even though the partition of the data was not the same as conducted in this paper.

Table V summarizes the mean and the standard deviation of the error rates obtained on three databases comparing static selection, DCS-LA and ambiguity-guided dynamic selection. Values are shown in bold for the lowest error rates obtained. We also report the error rates obtained on combining the initial pool of classifiers by majority voting and the results from an oracle. Such an oracle classifies correctly the test sample if any of the classifiers members predicts the correct label to the sample.

The overall best results on dna database were obtained on the populations of EoCs found by NSGA-II using Dif and the error rate as pair of objective functions. These results indicate an order relation between the three methods tested. Static selection was worse than both dynamic selection methods and DCS-LA was worse than ambiguity-guided dynamic selection. However, static selection was better than both dynamic approaches in three problems. Applying our method on the populations of EoCs generated by GA guided by the error rate and NSGA-II guided by CFD-error and DF-error produced better results than DCS-LA. Only on the population of EoCs generated by NSGA-II guided by A combined with the error rate, DCS-LA outperforms our method. The results obtained using our method on populations of EoCs found by NSGA-II guided by Dif and the error rate were better than the error rate of our baseline EoC composed of 100 DT.

The results obtained on the remaining databases do not show the same order relation as in satimage results. For satimage, our method was better than the static selection using populations of EoCs generated by NSGA-II guided by A and

TABLE IV

COMPILATION OF SOME OF THE RESULTS REPORTED IN THE EOC SELECTION LITERATURE (RSS: RANDOM SUBSPACE; DT: DECISION TREES; HET: HETEROGENEOUS CLASSIFIERS).

Database	**Reference number**	**Method**	**Classifiers**	**Data Partition**	**Error (%)**
Dna	[3]	RSS	DT	Holdout	9.19
Satimage	[16]	DT	Het	Cross valid	10.82
Texture	[4]	DT	Het	Holdout	0.75

TABLE V

MEAN VALUES ON 30 REPLICATIONS AND STANDARD DEVIATION VALUES

Dna 100DT = 5.05 Oracle = 0.03					
	NSGA-II				GA
Method	**A**	**CFD**	**Dif**	**DF**	**Error**
Static	5.93 ± 0.29	6.36 ± 0.42	4.77 ± 0.17	5.67 ± 0.33	4.97 ± 0.16
DCS-LA	7.05 ± 0.21	7.54 ± 0.28	4.63 ± 0.15	7.23 ± 0.23	5.14 ± 0.14
$\bar{A}$ dynamic	7.94 ± 0.23	6.50 ± 0.29	$\mathbf{4.59} \pm 0.17$	6.27 ± 0.23	4.95 ± 0.18
Satimage 100DT = 8.64 Oracle=0.22					
Static	8.77 ± 0.11	9.19 ± 0.18	8.76 ± 0.18	9.03 ± 0.17	8.82 ± 0.12
DCS-LA	9.30 ± 0.09	10.66 ± 0.19	10.12 ± 0.13	10.12 ± 0.15	8.96 ± 0.09
$\bar{A}$ dynamic	$\mathbf{8.73} \pm 0.10$	9.22 ± 0.19	9.17 ± 0.10	9.17 ± 0.15	8.78 ± 0.12
Texture 100DT = 2.56 Oracle=0.02					
Static	2.93 ± 0.15	3.20 ± 0.21	2.41 ± 0.09	2.91 ± 0.18	2.51 ± 0.09
DCS-LA	3.42 ± 0.17	3.85 ± 0.14	2.47 ± 0.07	3.51 ± 0.16	3.20 ± 0.09
$\bar{A}$ dynamic	2.98 ± 0.08	3.02 ± 0.14	$\mathbf{2.35} \pm 0.06$	2.97 ± 0.14	2.41 ± 0.05

the error rate and populations generated by GA guided only by the error rate. Static selection was better than both dynamic methods again in three problems. However, our method was better than DCS-LA in all five problems. Besides, the best overall results on the database were obtained employing our method on populations of EoCs found by NSGA-II guided by A and the error rate as pair of objective functions. Even though, selection did not outperform combination because the error rate of our baseline EoC (100 DT) was better than all results obtained on selection.

For texture database, static selection was better than our dynamic method in only two problem. In addition, our ambiguity-guided dynamic selection approach applied on populations of EoCs generated by NSGA-II guided by Dif combined with the error rate, achieved the overall best results on texture database and its results were better than the results of the baseline EoC.

The results indicate that our ambiguity-guided dynamic selection method outperforms DCS-LA. However, the quality of the population of EoC critically affect the performance of both dynamic methods. The results obtained on populations of EoCs found by NSGA-II guided by Dif combined with the error rate show this problem. Since such a combination of search algorithm and objective functions generates a population of high accurate classifiers, our method outperforms static selection. It is important to mention that the size of the initial pool of classifiers may play an important role. For satimage database, for example, since the selection was not better than combination, we can conclude that the results could be different if the initial pool of classifiers had more than 100 classifiers. Indeed, in majority of the literature on dynamic classifier selection [4], [7] either the differences between combination and selection are very small or combination is better than selection.

However, the results obtained using oracle indicate that the dynamic selection results can still be improved. The huge difference between oracle and combination results indicate that the selection of classifiers can be an effective task. Moreover, it is also important to mention that the static selection results were obtained using a validation control method to reduce overfitting in the optimization process. This strategy surely increases the generalization performance of the static solutions.

VI. CONCLUSIONS

In this paper we have proposed a new method to dynamic selection of ensembles of classifiers guided by ambiguity. It relies on choosing the EoC with lowest ambiguity among its members to predict the class of the test pattern. This method was applied in populations of EoCs generated by single- and multi-objective genetic algorithms. The error rate and diversity measures were employed as objective functions to guide the search process. In addition, DCS-LA was adapted to deal with populations of EoCs instead of populations of classifiers.

The experiments on three different databases demonstrated that the proposed ambiguity-guided method outperforms DCS-

LA. However, it was not better than static selection in the majority of the problems investigated. The excellent results from an oracle indicates that the dynamic selection may be improved to be able to outperform static selection. For future works we plan to investigate and propose strategies to improve the dynamic selection results. Moreover, other ensemble creation methods such as bagging and different base classifier will also be investigated.

ACKNOWLEDGMENT

This research is supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), Brazil, and Defence Research and Development Canada, DRDC-Valcartier under the contract W7701-2-4425.

REFERENCES

[1] L. Breiman, "Machine Learning," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[2] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *XIII International Conference on Machine Learning*, 1996, pp. 148–156.

[3] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, vol. 20, no. 8, pp. 832–844, 1998.

[4] K. Woods and W.P. Kegelmeyer and K. Bowyer, "Combination of Multiple Classifiers Using Local Accuracy Estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, vol. 19, no. 4, pp. 405–410, 1997.

[5] E.M. Dos Santos and R. Sabourin and P. Maupin, "Single and Multi-Objective Genetic Algorithms for the Selection of Ensemble of Classifiers," in *International Joint Conference on Neural Networks*, Vancouver-Canada, 2006, pp.5377–5384.

[6] D. Ruta and B. Gabrys, "Classifier Selection for Majority Voting," *Information Fusion*, vol. 6, no. 1, pp. 163–168, 2005.

[7] G. Giacinto and F. Roli, "Dynamic classifier selection based on multiple classifier behaviour," *Pattern Recognition*, vol. 34, pp. 1879–1881, 2001.

[8] D. Corne and J. Knowles, "No free lunch and free leftovers theorems for multiobjective optimization problems," in *Evolutionary Multi-Criterion Optimization*, 2003, pp. 327–341.

[9] X. Zhu and X. Wu and Y. Yang, "Dynamic Selection for Effective Mining from Noisy Data Streams," in *Fourth IEEE International Conference on Data Mining*, 2004, pp. 305–312.

[10] L.I. Kuncheva, "Cluster-and-Selection Model for Classifier Combination," in *International Conference on Knowledge Based Intelligent Engineering Systems and Allied Technologies*, 2000, pp. 185–188.

[11] K. Deb,*Multi-Objective Optimization using Evolutionary Algorithms*, England:John Wiley & Sons, LTD, 2001.

[12] H.W. Shin and S.Y. Sohn, "Selected tree classifier combination based on both accuracy and error diversity," *Pattern Recognition*, vol. 38, pp.191–197, 2005.

[13] G. Zenobi and P. Cunningham, "Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error," in *European Conference on Machine Learning*, 2001, pp. 576–587.

[14] L.K. Hansen and C. Liisber and P. Salomon,"The Error-Reject Tradeoff," *Open Systems & Information Dynamics*, vol. 4, pp. 159–184, 1997.

[15] 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.

[16] L. Didaci, G. Giacinto, F. Roli and G.L. Marcialis, "A study on the performances of the dynamic classifier selection based on local accuracy estimation," *Pattern Recognition*, vol. 28, pp.2188–2191, 2005.

[17] J.R. Quinlan,*Programs for Machine Learning*, Morgan Kaufmann, 1993.

APPENDIX

Four diversity measures were employed in this work. Since ambiguity was defined in equation 4, we present here the three other diversity measures studied: coincident failure diversity, difficulty measure and double-fault.

In order to simplify the description of the measures, we use a notation provided by Kuncheva and Whitaker [15]. Let L be the set of available classifiers and $Z = z_1, \ldots, z_N$ the data set, N^{ab} denotes the number of examples classified in Z, where a, b may assume the value of 1 when the classifier is correct and 0 otherwise. $l(z_j)$ denotes the number of classifiers that correctly classify (z_j).

It is important to observe that coincident failure diversity and difficulty measure are calculated on the whole ensemble of L classifiers. Double-fault is a pairwise measure, which means that it is calculated for each pair of classifiers and then, it is averaged over all pairs of L classifiers. Moreover, *dissimilarity* measures must be maximized, while *similarity* measures must be minimized when used as objective functions during the optimization process

A. Difficulty Measure

Given X calculated from $\{\frac{0}{L}, \frac{1}{L}, \ldots, 1\}$, which represents the number of classifiers in L that correctly classify a pattern z_i, this similarity measure may be calculated as:

$$\theta = Var(X)$$

B. Coincident Failure Diversity

This measure is the result of a modification to Generalized Diversity [15], which is based on the same distribution proposed for the difficulty measure. Here, however, $Y = \frac{i}{L}$ denotes the proportion of classifiers that do not correctly classify a randomly chosen sample x. Letting $p(i)$, be the probability that i classifiers fail when classifying a sample x, this dissimilarity measure is defined as follows:

$$CFD = \begin{cases} 0, & p_0 = 1.0 \\ \frac{1}{(1-p_0)} \sum_{i=1}^{L} \frac{i}{L} \frac{(i-1)}{(L-1)} p_i & p_0 < 0 \end{cases}$$

C. Double-fault

It is a pairwise similarity measure defined as:

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

The average of over all pairs of classifiers is:

$$DF = \frac{2}{L(L-2)} \sum_{i=1}^{L-1} \sum_{k=1+1}^{L} DF_{i,k}$$