Single and Multi-Objective Genetic Algorithms for the Selection of Ensemble of Classifiers

Eulanda M. Dos Santos, Robert Sabourin and Patrick Maupin

Abstract— **Many recent works have investigated methods to select subsets of classifiers instead of combining all available classifiers. The majority of these works has concluded that the combiner error rate is better than diversity to guide the selection process in order to identify the best performing subset of classifiers. However, the classifier selection process has to take into account three different aspects: complexity, overfitting and performance. These aspects of the selection process have not yet been tackled simultaneously in the literature. The study presented in this paper, deals with these three aspects in a handwritten digit recognition problem. Different search criteria such as diversity, error rate and number of classifiers are applied in single and multi-objective optimization approaches using genetic algorithms. In our experiments, we observed that error rate applied in a single optimization approach was the best objective function to increase performance. The generalized diversity and interrater agreement measures, combined with error rate in pairs of objective functions were the best measures to reduce complexity and keep good performance in a multi-objective optimization approach. Finally, the performance of the solutions found in both, single and multi-objective optimization processes were increased by applying a global validation method to reduce overfitting.**

I. Introduction

Traditionally, the selection of classifiers has focused on finding the most relevant subset of classifiers in order to improve the combined performance. It works as follows: given a pool of classifiers C, one applies search algorithms to select the best performing subset of classifiers L, where $|L| \leq |C|$. It is an overproduce and choose strategy [1]. When dealing with selection of classifiers, two important aspects should be analysed: the search criterion and the search algorithm [2]. Evolutionary algorithms are attractive search algorithms since they appear to fit well with selection of classifiers problems in the context of optimization of ensemble of classifiers (EoCs) [3]. Moreover, Ruta and Gabrys [2] observed that population-based genetic algorithms are good algorithms for classifier selection problems, due to the possibility of dealing with a population of solutions rather than only one solution.

The problem of choosing the most appropriate search criterion received a lot of attention in the recent literature, without much consensus. There is an agreement on the important role played by diversity. Ensembles can be more accurate than individual classifiers only when classifiers members present diversity among them [4]. On the other hand, the relationship between diversity measures and accuracy is not clear. Kuncheva and Whitaker [4] showed that diversity and accuracy do not exhibit a strong relationship and concluded that accuracy estimation can not be substituted for diversity. These results were confirmed by Ruta and Gabrys [2] in the context of classifier subset selection. They used diversity measures to guide the selection of classifiers in order to reduce the generalization error. They concluded that diversity was not a better measure than the combined error rate to find ensembles with good performance.

The combination of classification error rate and diversity as search criteria using multi-objective optimization methods offers a better way to overcome an apparent dilemma by allowing the simultaneous use of both search criteria. It is not surprising that this idea was already investigated in the literature. Zenobi and Cunningham [5] created k Nearest Neighbour (kNN) based EoCs by applying a feature subset selection approach using ambiguity (as defined in [5]) and error rate as objective functions and a hill-climbing search method. They showed that such a combined approach outperformed EoCs generated using the error rate as the only objective function. However, Tremblay et al. [6] applied multi-objective genetic algorithms (MOGA) by maximizing jointly recognition rate and ambiguity (as defined in [5]). They concluded that MOGA guided by such a couple of objective functions did not find better EoCs than Single-Objective Genetic Algorithm (GA) using only recognition rate as objective function. It is important to note that ambiguity was the only diversity measure investigated in both of these works on multi-objective selection of EoCs.

All these previous works have in common one characteristic: performance of the solutions is the only criterion used to determine whether a selection criterion is better than the others. However, to present high reliability, a classifier selection system has to focus on performance, complexity and overfitting [2]. In the present paper we address these three aspects namely, performance in the sense of high recognition rate, complexity in the sense of number of classifiers and finally, overfitting in the sense of a global validation control. These three aspects are analyzed in the context of population based evolutionary algorithms with single and multi-objective functions. Fourteen (14) objective functions are used to guide the optimization process in both, single and multi-objective approaches. These objective functions are: twelve (12) diversity measures, plus the ensemble's combined error rate ($1-$recognition rate) and cardinality (number of classifiers).

Eulanda M. dos Santos, École de technologie supérieure, Montreal, Canada; (email: eulanda@livia.etsmtl.ca).

Robert Sabourin, École de technologie supérieure, Montreal; (email: Robert.Sabourin@etsmtl.ca).

Patrick Maupin, Defence Research and Development Canada, Valcartier (DRDC Valcartier);(email: Patrick.Maupin@drdc-rddc.gc.ca).

II. Parameter Settings on Experiments

The experiments were carried out using NIST Special Database 19 (NIST SD19) which is a popular database used to investigate digit recognition algorithms. It is composed of 10 digit classes extracted from eight handwritten sample forms (hsf) series hsf-{0,1,2,3,4,6,7,8}. It is originally divided into 3 sets: hsf-{0123}, hsf-7 and hsf-4. The last two sets are referred herein as data-test1 (60,089 samples) and data-test2 (58,646 samples). Data-test2 is well known to be more difficult to classify than data-test1 [7]. On the basis of the results available in the literature, the representation proposed by Oliveira et al. [8] appears to be well defined and well suited to NIST SD19 database. The features are a combination of concavity, contour and surface of characters. The final feature vector is composed of 132 components: 78 for concavity, 48 for contour and 6 for surface.

Random Subspace-based ensembles of kNN are used in our experiments. Random Subspace is one of the most popular ensemble construction method as well as Bagging and Boosting. It was proposed by Ho in [9] and works as follows: given a data set $S = \{s_1, \ldots, s_N\}$, $s_i \in \Re_n$, different subspaces of the feature space $\Re_n$ are randomly chosen. Each random subspace is used to train one individual classifier. Then, individual classifier outputs are combined. This way, a small number of features are used, reducing the training-time process and the so-called curse of dimensionality. Since Random Subspace presents the advantage to deal with huge feature spaces, kNN appears to be a good candidate as learner in a Random Subspace-based ensemble. Indeed, Ho [10] advocates the fact that by using Random Subspace to generate ensemble of kNN we may reach high generalization rates and avoid the high dimensionality problem which is the main problem with kNN classifiers.

An ensemble of 100 kNN classifiers was generated using the Random Subspace method. EoCs are combined by majority voting. Experimental tests comparison were conducted to set up parameters such as: k value and the number of prototypes to kNN classifiers, the number of subspace dimensions, the size of optimization and validation data sets. Table I (a) summarizes the parameter set used.

TABLE I
Experiments parameters

a) Classifier and ensemble generation	
Number of nearest neighbours (k)	1
Random subspace sample size	32
Training dataset (hsf-{0123})	5,000
Optimization dataset size (hsf-{0123})	10,000
Validation dataset (hsf-{0123})	10,000
Test dataset 1 (hsf-7)	60,089
Test dataset 2 (hsf-4)	58,646
b) Genetic Algorithms parameters	
Population size	128
Number of generations	1000
Probability of crossover	0.8
Probability of mutation	0.01

The population-based evolutionary algorithms used in this work are both, single and multi-objective genetic algorithms. NSGA-II (Elitism Non-Dominated Sorting Algorithm) [11] is the multi-objective algorithm applied. Such algorithm presents two important characteristics: a full elite-preservation strategy and a diversity-preserving mechanism using crowding distance as distance measure. Crowding distance does not need any parameter to be set [11]. The selection of EoCs is applied in the context of genetic algorithms based on binary vectors. Since we use a base line EoC composed of 100 classifiers, each individual is represented by a binary vector of size 100. Each bit determines whether a classifier is active (1) or not (0). Additional experiments were carried out to define the genetic parameters. Table I (b) shows the parameters settings employed. The same parameters were used for both genetic algorithms.

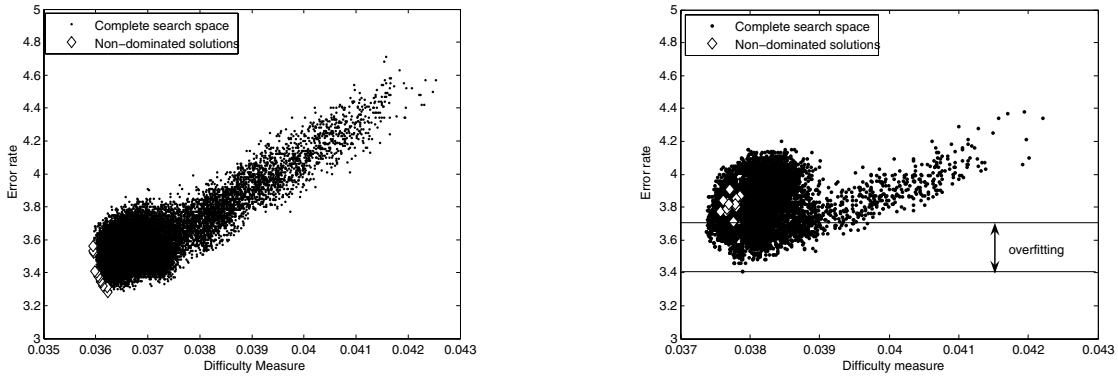
TABLE II
List of diversity measures used in the experiments

Name	**Label**	**Type**
Ambiguity	A	Dissimilarity
Coincident failure diversity	CFD	Dissimilarity
Correlation coefficient	Cor	Similarity
Difficulty measure	Dif	Similarity
Disagreement	Dis	Dissimilarity
Double-fault	DF	Similarity
Entropy	E	Dissimilarity
Fault majority	FM	Dissimilarity
Generalized diversity	GD	Dissimilarity
Interrater agreement	IA	Similarity
Kohavi-Wolpert	KW	Dissimilarity
Q-statistic	Q	Similarity

To guide the optimization process, the following measures were applied as objective functions:

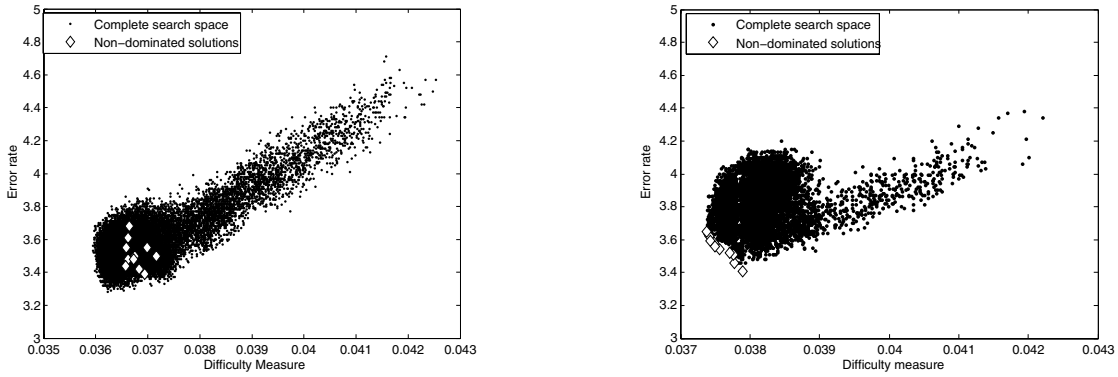
- **Error rate** - It is the most evident objective function. By applying search on minimizing error rate we may accomplish the main objective in pattern recognition, i.e., find predictors with high performance.
- **Cardinality** - Inspired by feature subset selection methods, where it is possible to increase recognition rates while reducing the number of features, the minimization of the number of classifiers appears to be a good objective function. The hope is to increase the recognition rate while minimizing the number of classifiers in order to accomplish, both performance and complexity requirements.
- **Diversity** - We use in this work 12 diversity measures (Table II): 10 measures grouped by Kuncheva and Whitaker [4] - correlation coefficient, coincident failure diversity, disagreement, double-fault, difficulty measure, entropy, generalized diversity, interrater agreement, Kohavi-Wolpert, Q-statistic, plus fault majority (as defined in [2]) and ambiguity [5]. Dissimilarity measures are maximized while similarity measures are minimized during the optimization process.

Once all parameters have been defined, we shall turn now to the three aspects we have proposed to deal with, i.e., overfitting, performance and complexity. Next sections detail our analysis.



(a) Optimization

(b) Validation



(c) Optimization

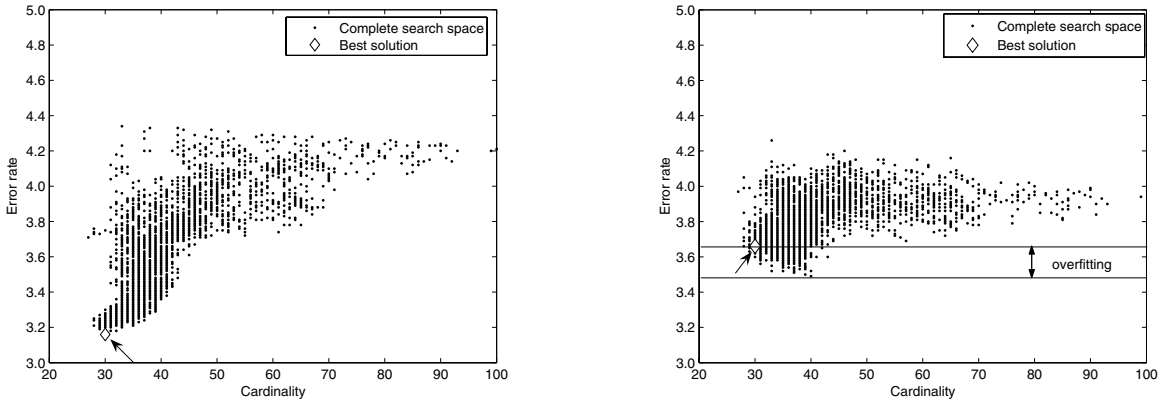
(d) Validation

Fig. 1. Optimization using NSGA-II and the couple of objective functions: difficulty measure and error rate. Objective space and Pareto-front projected on the optimization 1(a) and on the validation 1(b) data sets. 1(b) Overfitting between the most accurate solution on the validation data set and the most accurate solution from the Pareto-front. 1(c) Pareto-front from the auxiliary archive projected on the optimization and on the validation data sets 1(d).

III. The Overfitting Aspect

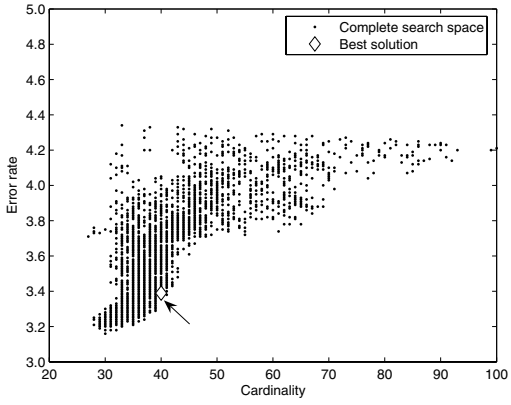
In general, the selection of EoCs relies on the idea of overproduce and choose strategy [1]. "Overproduce" due to the process of optimization of EoCs using search algorithms (GA and NSGA-II, in this paper), and "choose" due to the selection of the best EoC to classify the test samples. In this context of selecting the best EoC, we may categorize three different strategies: 1 - selection without validation; 2 - selection with partial validation and 3 - selection with global validation. The first and simplest procedure relies on selecting the best EoC on the same data set used during the search process. There is no independent validation data set. In this case, optimization process is performed using a data set, for a fixed number of generations. A population of solutions (Pareto-front for NSGA-II and the last generation for GA) is found and analyzed. Then, the same optimization data set is used to identify the best performing EoC. This procedure was applied in [12] for ensemble feature selection. However, it is well accepted in the literature that an independent data set (validation data set) must be used to validate selection methods in order to reduce overfitting and increase the generalization ability.

Following this idea, in the second selection procedure when the optimization process is finished, each EoC from the last population of solutions is used to classify an independent validation data set. The best solution on such validation set is then picked up to classify the test samples. Tremblay et al. [6] have applied the second procedure in a problem of selecting classifiers. Non-dominated Sorting Algorithm (NSGA) was used as search algorithm. Ruta and Gabrys [2] have applied the second procedure in their experiments with several different search algorithms, including GA. However, according to Radtke et al. [13], despite such apparent success, this strategy fails to address the overfitting phenomenon in the context of multi-objective optimization for classification systems. Although using an independent data set to validate the Pareto front solutions, they stressed that overfitting phenomenon may still be present.

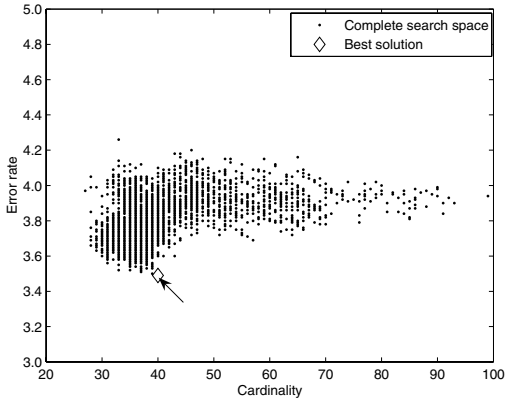


(a) Optimization

(b) Validation



(c) Optimization



(d) Validation

Fig. 2. Optimization using GA with error rate as objective function. Complete search space and best solution from last generation projected on the optimization 2(a) and on the validation 2(b) data sets. 2(b) overfitting between the most accurate solution found on the validation set and the most accurate solution from population of last generation. The complete search space and the global best solution from the auxiliary archive projected on the optimization 2(c) and on the validation 2(d) data sets.

The overfitting phenomenon is illustrated in Figure 1. NSGA-II was employed using the pair of objective functions: minimize jointly error rate and difficulty measure. Figure 1(a) depicts an example of the evolution of the optimization process after 1000 generations. Each point on the plot corresponds to an EoC and diamonds represent the Pareto front. When the objective space and Pareto-front are projected on validation data set (Figure 1(b)), the results are somewhat unexpected. The Pareto-front solutions found during the optimization process are far from being the best solutions found with the validation set. As the figure shows, the best solutions found on the validation data set are discarded during the optimization process. Therefore, the overfitting problem is still detected. In [13] and in our example (Figure 1) the overfitting phenomenon is shown on multi-objective optimization problems. However, we attempt to show experimentally in this paper that overfitting problem may also be detected in a single-objective optimization scenario. Figure 2 shows the objective space after 1000 generations using GA as search algorithm and the minimization of error rate as objective function. Diamonds represent the best solution. Even though we employed error rate as objective function, we show in Figure 2 plots of error rate versus number of classifiers to better illustrate the problem. The objective space and the best solution are projected on the optimization (Figure 2(a)) and on the validation data sets (Figure 2(b)). The overfitting problem is also detected for GA when the minimization of error rate is used as objective function.

Since there is no guarantee on controlling overfitting using the second method, a more appropriate method is necessary. An approach to reduce the overfitting problem on the context of multi-objective optimization is proposed in [13]. They advocate that instead validating only the solutions from the Pareto-front, validation has to be done at each new

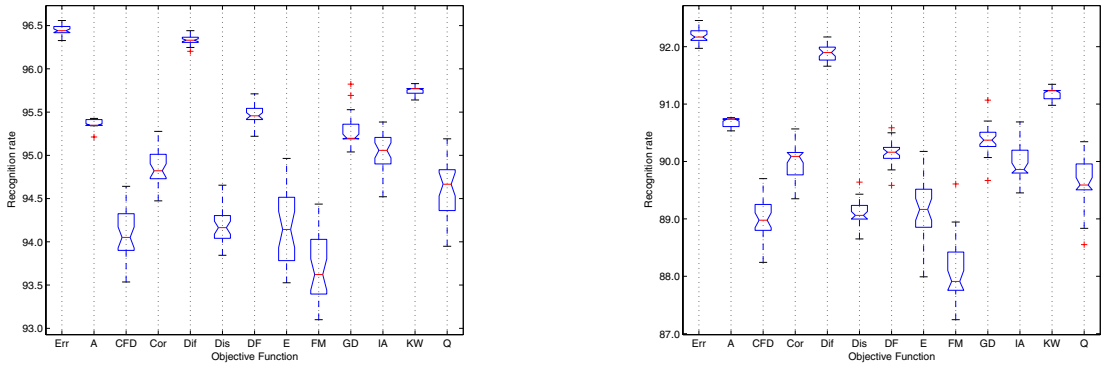


Fig. 3. Results of 30 replications using GA and 13 different objective functions. The performances were calculated on the Data-test1 (left) and on the Data-test2 (right).

generation using an auxiliary archive. They call this method "selection with global validation", the third selection strategy aforementioned. It works as follows: at each new generation during the optimization process, all solutions are validated, thus the non-dominated solutions on such validation set are stored in an auxiliary archive. When the optimization process is finished, two sets of non-dominated solutions are available: 1 - the traditional Pareto-front found on optimization data set and 2 - the set of non-dominated solutions obtained on validation data set. Figure 1(d) depicts an example of this third method where the solutions stored in the auxiliary archive (diamonds) are projected on the objective space using the validation data set while in Figure 1(c) they are projected on the objective space using the optimization data set. The solutions stored in the archive are very different to the Pareto-front solutions found on the optimization data set.

We show in this paper that the same approach may also be applied in the context of single objective optimization. To accomplish this, we propose to change the method in order to adapt it to single objective optimization problems. Since there is no Pareto-front when using GA, our interest is towards the global best solution. In this case, it is sufficient to validate all solutions at each new generation and keep stored, in the auxiliary archive, the global best solution found on the validation data set. When the optimization process is finished, the last population of solutions found during the optimization process and the global best solution found on the validation data set are available. When error rate is the objective function, the global best solution is calculated in terms of error rate values. Figure 2(d) shows the global best solution stored in the archive projected on validation data set while in Figure 2(c) it is projected on optimization data set.

These preliminary studies suggest that when there is overfitting on single and multi-objective optimization of EoCs, the procedure applying global validation is useful in order to reduce such overfitting phenomenon. These experiments lead us to an important question: can overfitting phenomenon be detected when diversity measures are used to guide the single-objective optimization process? In order to answer this question, Section VI summarizes the results of our experiments comparing the three selection strategies mentioned above. Thirteen different objective functions, including 12 diversity measures were applied. When the optimization process is guided by diversity, the global best solution is determined in terms of diversity values.

IV. Performance Analysis

In order to define the best objective function and the best search algorithm to our problem we carried out an experimental investigation focused on performance (recognition rate) and complexity (number of classifiers). Based on the results presented in Section III, both NSGA-II and GA were applied taking into account the auxiliary archive (global validation strategy) to reduce overfitting. The objective functions described in Section II were employed to guide the optimization process.

The first question to answer is: which measure is the best objective function to find high performing EoCs? Among measures featured in Section II, error rate and diversity measures are the most evident candidates. The direct way to compare these measures is to apply a single objective optimization approach. This direct comparison allows us to verify the possibility of using diversity instead recognition rate to find high performing EoCs.

A. Experiments with GA

GA-based experiments were conducted to compare 13 different objectives functions: error rate and the 12 diversity measures mentioned in Section II. More details about these diversity measures can be found in [4], [2] and [5]. Each experiment was replicated 30 times in order to better compare the results. Therefore, each of the 13 objective functions employed generated 30 optimized EoCs. Genetic parameters are shown in Table I (b). Figure 3 shows the comparison results of the 30 replications on Data-test1 (left) and on Data-test2 (right).

These experiments show that:

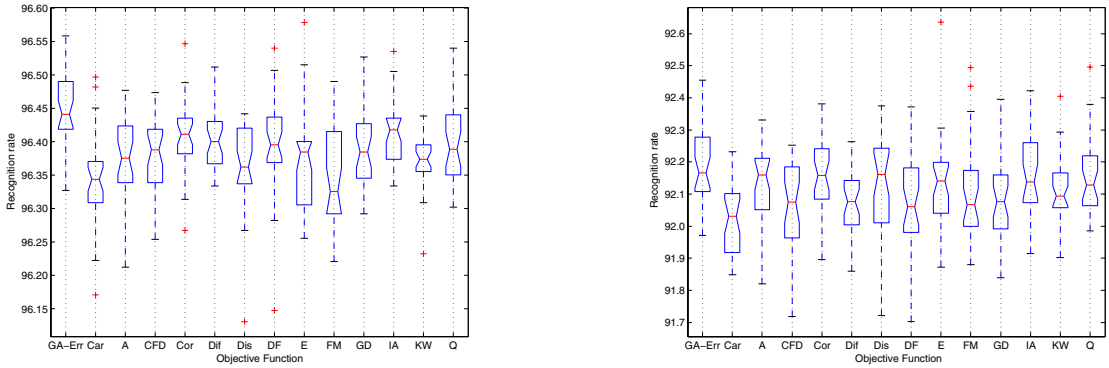


Fig. 4. Results of 30 replications using NSGA-II and 13 different couples of objective functions. The performances were calculated on the Data-test1 (left) and on the Data-test2 (right). The first values represent GA with error rate as objective function.

1) Diversity measures are not better than error rate as objective function to generate high performing EoCs.
2) The most successful diversity measure was Difficulty measure. However, the performance of the EoCs found using this measure was, on average, 0.12% (Data-test1) and 0.31% (Data-test2) worse than the EoCs found using directly error rate.
3) Fault-majority was the worst objective function. It is different from the results presented by Ruta and Gabrys [2]. They have observed that measures with better correlation with majority voting error, i.e., fault-majority and double-fault, are better objective functions to generate high performing EoCs than the other ones. Double-fault was the third better diversity measure and fault-majority was the worst objective function in our experiments.

The results achieved using GA were already expected, except for fault-majority and double-fault. We confirmed the results from previous works, e.g. [2] and [4], that diversity alone can not substitute error rate as objective function to find the highest performing EoCs. Since diversity alone is not better than error rate, can we found better performing EoCs by including both objective functions in the optimization process? We try to answer this question using a multi-objective optimization approach in next section.

B. Experiments with NSGA-II

We pursue our experimental study using now NSGA-II as search algorithm. The preliminary study with GA suggested that diversity alone is not better than error rate to generate the best performing EoCs. This observation motivated the use of both error rate and diversity jointly to guide the optimization process with NSGA-II, since we have the possibility to combine different objective functions. The hope is that higher diversity between base classifiers lead to the selection of high performing EoCs.

Each diversity measure mentioned in Section II was combined with error rate to compose pairs of objective functions to guide the optimization process. Again, the optimization process using each pair of objective functions was replicated 30 times. Figure 4 shows the results of 30 replications on Data-test1 (left) and on Data-test2 (right). It is important to mention that the first value corresponds to the results using GA as search algorithm and error rate as objective function. So we can compare the single and the multi-objective results. The second value corresponds to the results using NSGA-II guided by cardinality and error rate as objective functions (discussed in next section).

From these results some observations can be made:

1) By including both diversity and error rate in a multi-objective optimization process we may find more performing EoCs than using diversity alone, however, the performance of these EoCs are still worse than the performance of the EoCs found using error rate in the single optimization process.
2) The best diversity measures are difficulty, interrater agreement, correlation coefficient, and double-fault on Data-test1. On Data-test2, almost all diversity measures have similar results. It is important to note that the difference among EoCs found using diversity (multi-objective optimization) and EoCs found using only error rate (single objective optimization) was reduced. The EoCs found using the four best diversity measures were on average 0.05% worse than the EoCs found using only error rate on Data-test1 and 0.13% on Data-test2.
3) The two measures pointed out by Ruta and Gabrys [2] as the best diversity measures (double-fault and fault-majority), found better EoCs on multi-objective than in single-objective, as happened for all measures. However, especially on Data-test1, Fault-majority is the worst measure again.

As indicated in Section I, besides performance, we have to take into account complexity and overfitting when selecting EoCs. In next two sections these two aspects are analysed.

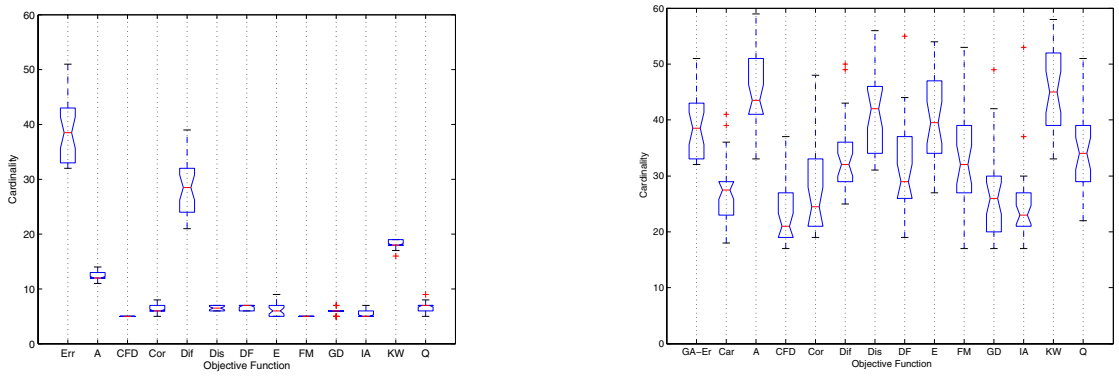


Fig. 5. Cardinality of the EoCs found using GA (left) and NSGA-II (right). Each optimization process was performed 30 times.

V. Complexity Analysis

Feature subset selection approaches rely on an idea that by selecting the most discriminant features we may reduce the number of features and increase the recognition rate since redundant features are discarded. We may establish an analogy between feature subset selection and ensemble of classifiers selection. Following this analogy, including cardinality and error rate as a pair of objective functions to guide the optimization process, could address both complexity and performance aspects. The hope is to discard redundant classifiers in order to decrease error rate. Thus, in this experiment, search with NSGA-II was guided by minimizing jointly error rate and cardinality. In Figure 5 (right) we can see the size of the EoCs found using this couple of objective functions. The cardinality of the EoCs found using the other combinations of objective functions, i.e., diversity and error rate, are also shown to compare the results. Figure 5 (left) shows the cardinality of the EoCs found with GA (Section IV-A). Based on these results we observe that:

1) The analogy between feature subset selection and ensemble selection may be established. The performance of our baseline system, i.e., the pool of 100 kNN (96.28% on Data-test1), is 0.05% worse than the averaged result using cardinality and error rate as objective functions (average of 96.35% on Data-test1), while the averaged cardinality is 27 classifiers. However, better performing EoCs can be found using GA and error rate. Moreover, the combination of cardinality and error rate as objective function did not establish the best trade-off between these two measures. Interrater agreement and generalized combined with error rate generated smaller (24 classifiers on average) and more performing EoCs (96.41% on average on Data-test1).
2) Ambiguity combined with error rate and Kohavi-Wolpert combined with error rate found the biggest EoCs (45 classifiers on average). What is more interesting is that, although we have found the best performing EoCs using GA as search algorithm and error rate as objective function, such single objective function did not find the biggest EoCs.

VI. Overfitting Analysis

We also have carried out a comparative study using the three selection procedures described in Section III: 1- selection without validation; 2 - selection with partial validation and 3 - selection with global validation, to verify whether overfitting is detected and the advantages of using the global validation method. We applied both GA and NSGA-II guided by all objective functions discussed before. The optimization process was replicated 30 times up to 1000 generations. For each replication, a population of solutions was thus found and the three selection procedures were applied to identify the best EoC. Table III summarizes the average error rates of the 30 replications from the experiments using GA and Table IV shows the average error rates of the experiments using NSGA-II. These experiments show that:

1) The overfitting phenomenon was detected in all of multi-objective optimization results. The global validation procedure allowed us to find EoCs with higher power of generalization whatever pair of objective functions used to guide the NSGA-II search.
2) Except for fault-majority, the global validation procedure helped to find EoCs with higher recognition rates in all of single objective optimization problems studied in this paper. Therefore, overfitting was detected even when diversity was used as objective function.
3) Since the selection of the global best solution in terms of diversity leads to increase the generalization performance, there is a relationship, even not clear, between diversity and performance.
4) Data-test2 appears to be more sensitive to the overfitting problem. The reduction on generalization error rate was higher on such data set.

VII. Conclusions

This work presented the experimental results of a study using single and multi-objective selection of ensemble of

TABLE III

GA RESULTS ON COMPARING THE SELECTION PROCEDURES.

	Datatest1			Datatest2		
Objective Function	No Valid	Partial Valid	Global Valid	No Valid	Partial Valid	Global Valid
Err	3.60	3.60	3.55	7.89	7.91	7.80
A	4.70	4.70	4.63	9.42	9.42	9.31
CFD	6.36	6.35	5.92	11.79	11.79	11.02
Cor	5.34	5.61	5.15	10.42	10.64	10.01
Dif	3.76	3.76	3.67	8.43	8.44	8.11
Dis	6.32	6.32	5.81	11.70	11.70	10.92
DF	4.80	4.80	4.53	10.45	10.45	9.85
E	6.11	6.12	5.86	11.30	11.33	10.82
FM	6.00	6.00	6.32	11.39	11.39	11.92
GD	4.76	4.84	4.72	10.29	9.98	9.64
IA	5.17	5.16	4.96	10.24	10.21	10.01
KW	4.28	4.28	4.25	8.89	8.89	8.81
Q	5.73	5.73	5.43	10.97	10.97	10.37

TABLE IV

NSGA-II RESULTS ON COMPARING THE SELECTION PROCEDURES.

	Datatest1			Datatest2		
Objective Function	No Valid	Partial Valid	Global Valid	No Valid	Partial Valid	Global Valid
Card	3.66	3.67	3.65	8.08	8.08	7.98
A	3.66	3.70	3.63	7.90	7.97	7.87
CFD	3.67	3.68	3.62	8.09	8.11	7.94
Cor	3.60	3.59	3.59	7.91	7.90	7.84
Dif	3.64	3.63	3.60	8.12	8.11	7.93
Dis	3.66	3.69	3.64	7.99	8.01	7.88
DF	3.63	3.64	3.60	8.12	8.14	7.93
E	3.66	3.67	3.63	7.96	7.98	7.87
FM	3.69	3.70	3.65	8.02	8.01	7.90
GD	3.65	3.63	3.61	8.08	8.00	7.92
IA	3.62	3.60	3.59	7.96	7.92	7.84
KW	3.67	3.67	3.63	7.96	7.97	7.89
Q	3.65	3.63	3.60	7.98	7.94	7.84

classifiers taking into account three factors: overfitting, performance and complexity. The objective was to determine the best objective function and ultimately, to select between single and multi-objective approach. An ensemble of 100 kNN classifiers generated using Random Subspace was used as initial pool of classifiers. Fourteen different objective functions were applied: 12 diversity measures, error rate and cardinality (number of classifiers). NSGA-II was the multi-objective genetic algorithm used. The overfitting aspect was dealt with by keeping an auxiliary archive to store validated solutions. This method, called "global validation" approach, relies on the idea that instead of validating only the last generation solutions (GA) or Pareto-front solutions (NSGA-II), it is necessary to validate solutions in each generation during the optimization process.

Our results confirmed the observation from previous works that diversity alone can not be better than error rate to find the most accurate EoCs. When both, error rate and diversity are combined in a multi-objective approach, the performance of the solutions found using diversity are much higher than the performance of the solutions found using diversity in a single approach. However, the performance of the EoCs found using diversity measures combined with error rate to guide the selection were still worse than the performance of the EoCs found using only error rate, although the difference was reduced. The combined minimization of cardinality and error rate did not accomplish the objective to reduce complexity while increasing performance. In fact, generalized diversity and interrater agreement measures combined with error rate established the best trade-off between performance and cardinality. We conclude that the advantage of using a multi-objective optimization combining diversity and error rate is to find small EoCs with reasonable performance. Finally, the global validation selection procedure helped to increase the generalization performance of the EoCs found in both single and multi-objective experiments, even when diversity was used as objective function. Although not as clearly as one might have hoped, our results outlined a relationship between performance and diversity as criteria for the selection of EoCs with high performance.

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] G. Giacinto and F. Roli, "An Approach to the Automatic Design of Multiple Classifier Systems," *Pattern Recognition Letters*, vol. 22, pp. 25–33, 2001.

[2] D. Ruta and B. Gabrys, "Classifier Selection for Majority Voting," *Information Fusion*, vol. 6, pp. 163–168, 2005.

[3] K. Sirlantzis and M.C. Fairhurst and R.M. Guest, "An evolutionary algorithm for classifier and combination rule selection in multiple classifier system," *Proc. ICPR*, 2002, pp. 771–774.

[4] L.I. Kuncheva and C.J. Whitaker, "Measures of diversity in classifier ensembles an their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, pp. 181–207, 2002.

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

[6] G. Tremblay and R. Sabourin and P. Maupin, "Optimizing Nearest Neighbour in Random Subspaces using a Multi-Objective Genetic Algorithm," *Proc. ICPR*, 2004.

[7] P.J. Gother, *NIST Special Database 19 - Handprited forms and characters database,*, National Institute of Standard and Technology - NIST : database CD documentation, 1995.

[8] L.S. Oliveira and R. Sabourin and F. Bortolozzi and C.Y. Suen, "Automatic Recognition of Handwritten Numerical Strings: A Recognition and Verification Strategy," *IEEE TPAMI*, vol. 24, pp. 1438–1454, 2002.

[9] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE TPAMI*, vol. 20, pp. 832–844, 1998.

[10] T.K. Ho, "Nearest Neighbors in Random Subspaces," *Proc. International Workshop on Statistical Techniques in Pattern Recognition*, 1998, pp. 640–648.

[11] K.Deb, *Multi-Objective Optimization using Evolutionay Algorithms*, England: Jhon Wiley Sons, 2002, Second Edition.

[12] A. Tsymbal and M. Pechenizkiy and P. Cunningham, "Diversity in Search Strategies for Ensemble Feature Selection," *Information Fusion*, vol. 6, pp. 83–98, 2005.

[13] P.V.W. Radtke and R. Sabourin and T. Wong, "Impact of Solution Over-fit on Evolutionary Multi-Objective Optimization for Classification Systems," *Submitted to IEEE Transactions on Evolutionary Computation*, 2006.