Pareto Analysis for the Selection of Classifier Ensembles

Eulanda M. Dos Santos
Ecole de technologie superieure
1100 rue Notre-Dame ouest
Montreal, Canada
eulanda@livia.etsmtl.ca

Robert Sabourin
Ecole de technologie superieure
1100 rue Notre-Dame ouest
Montreal, Canada
Robert.Sabourin@etsmtl.ca

Patrick Maupin
Defence Research and Development Canada
2459 Pie XI Blvd North
Val-Bélair, Canada
Patrick.Maupin@drdc-rddc.gc.ca

ABSTRACT

The overproduce-and-choose strategy involves the generation of an initial large pool of candidate classifiers and it is intended to test different candidate ensembles in order to select the best performing solution. The ensemble's error rate, ensemble size and diversity measures are the most frequent search criteria employed to guide this selection. By applying the error rate, we may accomplish the main objective in Pattern Recognition and Machine Learning, which is to find high-performance predictors. In terms of ensemble size, the hope is to increase the recognition rate while minimizing the number of classifiers in order to meet both the performance and low ensemble size requirements. Finally, ensembles can be more accurate than individual classifiers only when classifier members present diversity among themselves. In this paper we apply two Pareto front spread quality measures to analyze the relationship between the three main search criteria used in the overproduce-and-choose strategy. Experimental results conducted demonstrate that the combination of ensemble size and diversity does not produce conflicting multi-objective optimization problems. Moreover, we cannot decrease the generalization error rate by combining this pair of search criteria. However, when the error rate is combined with diversity or the ensemble size, we found that these measures are conflicting objective functions and that the performances of the solutions are much higher.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology

General Terms

Experimentation

Keywords

Classifier ensembles, ensemble selection, Pareto analysis, diversity measures

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.

1. INTRODUCTION

Diversity is considered the key issue for employing classifier ensembles successfully [6]. It is intuitively accepted that ensemble members must be different from each other, exhibiting especially diverse errors [3]. Although the concept of diversity is still considered an ill-defined concept [3], there are several measures of diversity reported in the literature [7]. Moreover, the most widely used ensemble creation techniques, bagging, boosting and the random subspace method are focused on incorporating the concept of diversity into the construction of effective ensembles.

Another approach to explicitly enforce diversity during the generation of ensembles is the so-called *overproduce-and-choose strategy* (OCS) [9]. Methods based on OCS are divided into two phases: (1) *overproduction*; and (2) *selection*. An initial *large* pool of classifiers $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ is constructed at the first phase using the training dataset $\mathcal{T}$. The second phase is devoted to generate and test different combinations of the initial classifiers c_i in order to identify the best subset of classifiers C_j^*. The selected ensemble C_j^* is then combined to estimate the class labels ω_k of the samples contained in the test dataset $\mathcal{G}$. Figure 1 illustrates the OCS phases. Hence, the objective of OCS is to find the most relevant subset of classifiers based on the assumption that classifiers in $\mathcal{C}$ are redundant [6]. It is also interesting to note that the selection phase required by OCS can be easily formulated as an optimization problem in which a search algorithm operates by minimizing/maximizing one objective function or a set of objective functions.

Taking into account that highly accurate and reliable classification is required in Machine Learning and Pattern Recognition practical applications, ideally, ensemble classifier members must be accurate and different from each other to ensure performance improvement. Therefore, the key challenge for classifier ensemble research is to understand and measure diversity in order to establish the perfect trade-off between diversity and accuracy [6]. The literature has shown that OCS allows the selection of accurate and diverse classifier members [9] by using the combination of the error rate and diversity as search criteria. Zenobi and Cunningham [14] used ambiguity (as defined in [14]) and the error rate to guide a hill-climbing search method. Tremblay et al. [11] employed a *multi-objective genetic algorithm* MOGA (a modified version of Non-dominated Sorting GA - NSGA [4]) guided by pairs of objective functions composed of the error rate with the following four diversity measures: ambiguity [14], fault majority [10], entropy and Q-statistic [7].

However, since OCS relies on the idea that component

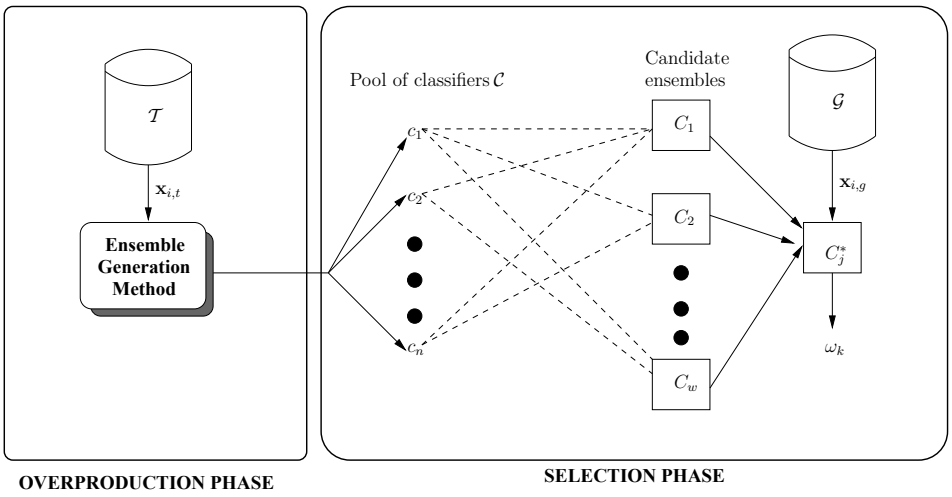


Figure 1: Overview of the OCS process. OCS is divided into the overproduction and the selection phases. The overproduction phase creates a large pool of classifiers, while the selection phase focus on finding the most relevant subset of classifiers.

classifiers are redundant, an analogy can be established between feature subset selection and OCS. Feature subset selection (FSS) approaches work by selecting the most discriminant features in order to reduce the number of features and to increase the recognition rate. Following this analogy, the selection phase of OCS could focus on discarding redundant classifiers in order to increase performance and reduce ensemble size. In [11], it has been shown that it is possible to reduce the number of classifiers while at the same time increasing performance by using both the error rate and ensemble size to guide the search phase.

Therefore, diversity measures, performance and ensemble size appear to be the most relevant measures to be employed in order to identify the best classifier ensemble C_j^*. In this paper we present an experimental study combining these three main measures to make up pairs of objective functions to guide the selection phase formulated as a multi-objective optimization problem. *Multi-objective genetic algorithms* (MOGAs) are often solutions to optimization processes guided by multi-objective functions. Among several MOGAs proposed in the literature, we use in this paper NSGA-II (elitist non-dominated sorting GA) [4]. This algorithm presents 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 require any parameter to be set [4]. Elitism is used to provide a means to keep good solutions among generations, and the diversity preservation mechanism is used to allow a better spread among the solutions over the Pareto front. Solutions more widely spread are better since a larger spread may indicate a better coverage of the true Pareto front [2].

In despite of the diversity preservation mechanism, objective functions in multi-objective optimization problems must be conflicting objective functions in order to provide spread over the Pareto front. In this paper we use two Pareto quality metrics based on calculating the overall Pareto spread and k^{th} objective Pareto spread introduced by Wu and Azarm [13] to show whether or not the pairs of objective functions used in our experiments are conflicting objective functions. Our objective is to verify how these Pareto quality measures may help the analysis of the relationship between diversity measures, performance and ensemble size. As a consequence, we study how these classifier ensembles measures are related to the performance of the solutions found at the end of the optimization process.

The paper is organized as follows. The Pareto quality metrics are described in section 2. Then, the experiments and the results are presented in section 3. Conclusions and suggestions for future work are discussed in section 4.

2. PARETO ANALYSIS

Objective functions in multi-objective optimization problems are often conflicting. Since different tradeoffs are established over the Pareto front, when one solution is better according to one objective, it is often worse according to the remaining objective functions. Indeed, Deb [5] points out that in a two conflicting objective problem, for instance, if a ranking of non-dominated solutions is carried out in an ascending order according to one objective function, a ranking in a descending order is obtained according to the other objective function.

In terms of OCS, we mentioned in the introduction that the error rate, ensemble size and diversity measures are the most frequent objective functions employed in the literature. Various approaches defining diversity have been proposed in the literature. We employ 12 diversity measures (Table 1) in the optimization processes conducted in this paper. Ten measures were grouped by Kuncheva and Whitaker [7]: correlation coefficient, coincident failure diversity, disagreement, double-fault, difficulty measure, entropy, generalized diversity, interrater agreement, Kohavi-Wolpert, Q-statistic. Fault majority was proposed by Ruta and Gabrys [10] and ambiguity was defined by Zenobi and Cunningham [14]. It is important to mention that *dissimilarity* measures must be maximized, while *similarity* measures must be minimized when used as objective functions during the optimization process. The pairwise measures are calculated for each pair

Table 1: List of search criteria used in the optimization processes. The optimization specifies whether the search criterion must be minimized (similarity) or maximized (dissimilarity) and the type indicates whether the diversity measure is must be calculated for each pair of classifiers or on the whole candidate ensemble.

Name	Label	Optimization	Type
Error rate	ϵ	-	-
Ensemble size	ζ	-	-
Ambiguity	γ [14]	Dissimilarity	Non pairwise
Coincident failure diversity	σ [7]	Dissimilarity	Non pairwise
Correlation coefficient	ρ [7]	Similarity	Pairwise
Difficulty measure	θ [7]	Similarity	Non pairwise
Disagreement	η [7]	Dissimilarity	Pairwise
Double-fault	δ [7]	Similarity	Pairwise
Entropy	ξ [7]	Dissimilarity	Non pairwise
Fault majority	λ [10]	Dissimilarity	Pairwise
Generalized diversity	τ [7]	Dissimilarity	Non pairwise
Interrater agreement	κ [7]	Similarity	Non pairwise
Kohavi-Wolpert	ψ [7]	Dissimilarity	Non pairwise
Q-statistic	Φ [7]	Similarity	Pairwise

of classifiers, while the non-pairwise measures are calculated on the whole ensemble C_j. Summarizing, as shown in Table 1, our objective functions comprise twelve diversity measures, the ensemble's error rate (ϵ) and ensemble size (ζ).

In order to illustrate whether or not these objective functions produce conflicting objective problems, we will use examples obtained in one replication using the NIST-digits database (section 3). An initial pool of *k Nearest Neighbors* kNN classifiers was generated at the overproduction phase using the random subspace method. The maximum number of generations was fixed at $1,000$ and the size of the population of individuals is 128.

In Figure 2(a) the search was guided by the minimization of ϵ and the difficulty measure θ. The Pareto front solutions are shown in an ascending order of ϵ, consequently, in a descending order of the θ. Figure 2(b) shows the Pareto front found using the following pair of objective functions: jointly minimize ϵ and ζ. Once again, the solutions were ordered according to ϵ, and in a descending order of ζ. In Figure 2(c) the pair of objective functions employed was the minimization of ζ and θ. The Pareto solutions are shown in an ascending order of θ (descending order of ζ). However, the same behavior cannot be detected in Figure 2(d) where it is shown an example of the evolution of the optimization process, which was guided by the following pair of objective functions: jointly minimize ζ and the interrater agreement κ. It can be seen that only one solution was found over this Pareto front. The first two figures show that both θ or ζ combined with ϵ are conflicting objectives. In Figure 2(c) it is shown that θ and ζ are also conflicting objectives. In contrast, κ and ζ are not conflicting objectives.

In order to show whether or not the pairs of objective functions composed by the measures summarized in Table 1, are conflicting objective functions we apply here two quality metrics based on calculating the overall Pareto spread and the k^{th} objective Pareto spread introduced by Wu and Azarm [13]. Considering a multi-objective problem with m objective functions $f_1, f_2, \ldots, f_m$, we show in this section how to calculate the two measures of spread taking into account, without loss of generality, all objective functions to be minimized and equally important.

Given w be the possible worst solution and b be the possible best solution, the objective space should be scaled by the following equation for any feasible point x_k:

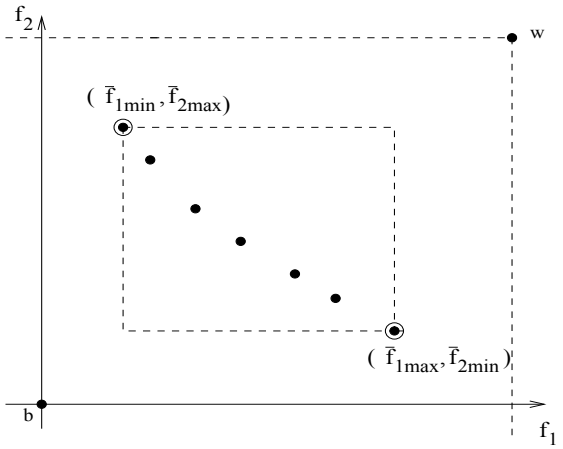


Figure 3: Scaled objective space of a two-objective problem used to calculate the overall Pareto spread and the k^{th} objective Pareto spread.

$$\bar{f}_j(x_k) = \frac{f_j(x_k) - f_j^b}{f_j^w - f_j^b} \quad (1)$$

In a two-objective problem, the scaled objective space is a hyper-rectangle, as shown in Figure 3 defined by the scaled worst $(1,1)$ and best $(0,0)$ solutions.

2.1 Overall Pareto Spread (OPS)

OPS measures the width of the Pareto front over the objective space considering all objective functions together. This measure is defined as:

$$OPS = \prod_{i=1}^{m} |max_{k=1}^{\bar{np}}[\bar{f}_i(x_k)] - min_{k=1}^{\bar{np}}[\bar{f}_i(x_k)] \quad (2)$$

where $\bar{np}$ is the total number of Pareto solutions.

When the measures used to guide the optimization process are no conflicting objective functions, OPS is equal to

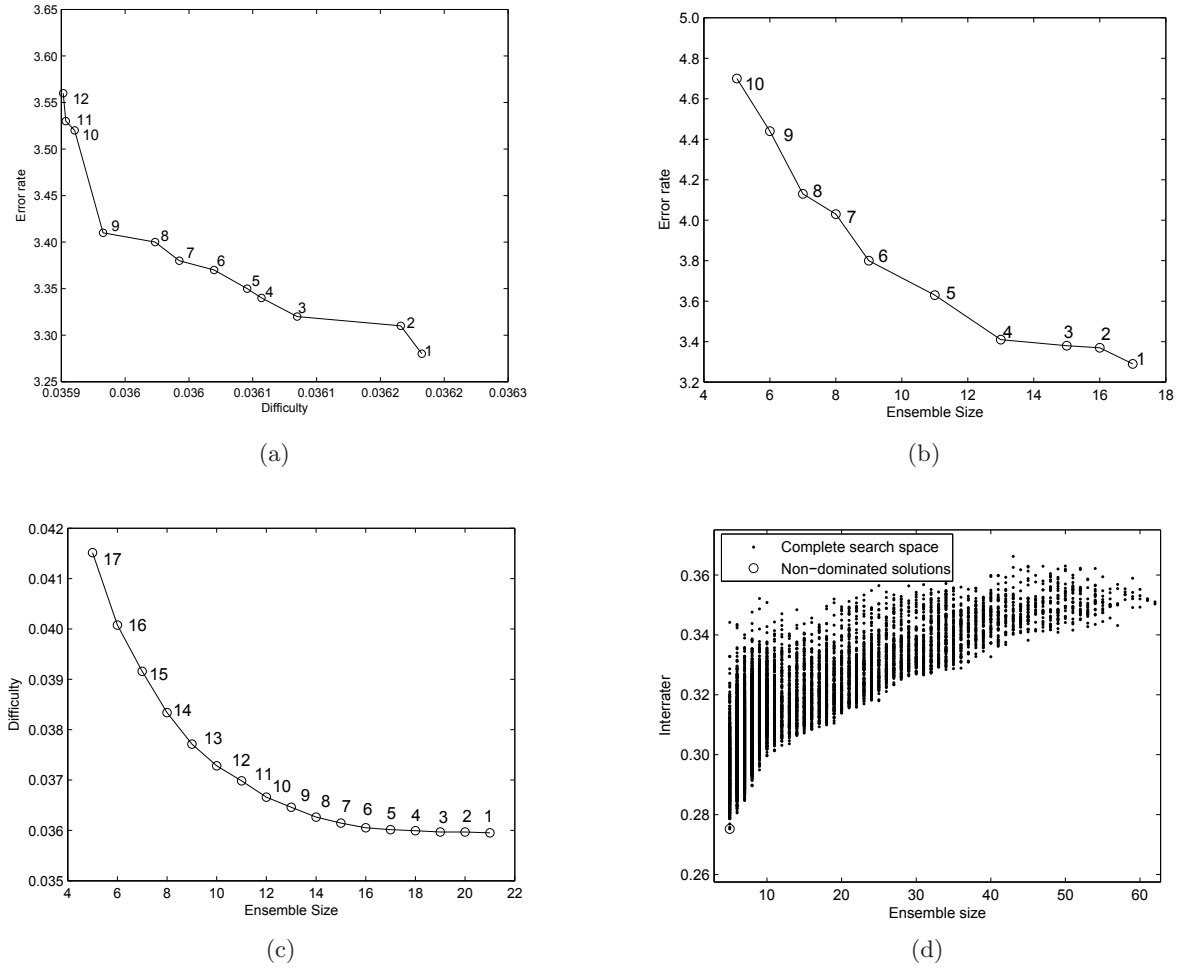


Figure 2: Set of non-dominated solutions after 1,000 generations found using NSGA-II and the pairs of objective functions: jointly minimize the error rate and the difficulty measure (a), jointly minimize the error rate and ensemble size (b), jointly minimize ensemble size and the difficulty measure (c) and jointly minimize ensemble size and the interrater agreement (d).

Table 2: Worst and best points, minima and maxima points, IPS and OPS for each set of non-dominated solutions shown in Figure 2.

Objective Functions	Original values				Scaled values					
	w	b	min	max	w	b	min	max	IPS	OPS
Difficulty θ (f_1)	1	0	0.0359	0.0362	1	0	0.0359	0.0362	0.0003	
Error rate ϵ (f_2)	100	0	3.2800	3.5600	1	0	0.0328	0.0356	0.0028	0.1 x 10^{-5}
Ensemble size ζ (f_1)	100	0	5	17	1	0	0.0500	0.1700	0.1200	
Error rate ϵ (f_2)	100	0	3.2900	4.7000	1	0	0.0329	0.0470	0.0141	0.17 x 10^{-2}
Ensemble size ζ (f_1)	100	0	5	21	1	0	0.0500	0.2100	0.1600	
Difficulty θ (f_2)	1	0	0.0415	0.0356	1	0	0.0356	0.0415	0.0059	0.9 x 10^{-3}
Ensemble size ζ (f_1)	100	0	5	5	1	0	0.0500	0.0500	0.0000	
Interrater κ (f_2)	1	0	0.2752	0.2752	1	0	0.2752	0.2752	0.0000	0.0000

0. Thus, a wider spread leads to more diversity over the Pareto front, which is a desired Pareto property. However, there may be a drawback to this quality measure since it does not take into account the individual objective spread. For instance, considering the example shown in Figure 3, if instead of 7 solutions the Pareto front was composed of the following two solutions $(\bar{f}_{1min}, \bar{f}_{2max})$ and $(\bar{f}_{1max}, \bar{f}_{2min})$, the same OPS would be obtained. However, the diversity observed over a Pareto front composed of only two solutions is much smaller than the diversity over a Pareto front composed of 7 solutions. The k^{th} objective Pareto spread [13] focus on overcoming this problem.

2.2 K^{th} Objective Pareto Spread (IPS)

According to Wu and Azarm[13], IPS is an additional measure to OPS, since it takes into account the range of diversity of each objective function, as follows:

$$IPS_j = |max_{i=1}^{n_p}(\bar{f}_j(x_i)) - min_{i=1}^{n_p}(\bar{f}_j(x_i))| \quad (3)$$

In the two-objective problem illustrated in Figure 3, IPS and OPS are respectively:

$$IPS_{f_1} = |\bar{f}_{1max} - \bar{f}_{1min}| \quad (4)$$

$$IPS_{f_2} = |\bar{f}_{2max} - \bar{f}_{2min}| \quad (5)$$

$$OPS = IPS_{f_1} IPS_{f_2} \quad (6)$$

The IPS is important because it allows us to measure the diversity over the Pareto front and to show whether or not spread is wider for one objective function than for the others. As a consequence, we may verify when the objective functions are equally diverse.

Table 2 presents IPS and OPS calculated for each Pareto front shown in Figure 2. The difficulty measure θ and the error rate ϵ are conflicting objective functions, but there is much more variation among ϵ values than among θ values. Ensemble size ζ and ϵ are conflicting objective functions, and there is more variation in ζ values than in ϵ values. In addition, θ and ζ are conflicting objective functions, but once again, there is less variation among θ values. It is important to note that there is more variation among θ values when this diversity measure is combined with ζ than when it is combined with ϵ. Finally, the interrater agreement κ and ζ are not conflicting objective functions. In next section we present experimental results employing these Pareto quality measures in our problem of selecting classifier ensembles. All objective functions shown in Table 1 are investigated.

3. EXPERIMENTS

A series of experiments has been carried out to calculate the two Pareto quality measures presented in last section. We analyze the Pareto fronts obtained by combining diversity measures, the error rate and ensemble size in pairs of objective functions used to guide NSGA-II at the selection phase of OCS. The objective of this analysis is to verify how these quality measures are related to the performance of the best classifier ensembles obtained at the end of the OCS process. In section 3.1 we present the results of the Pareto analysis and in section 3.2, our analysis in terms of performance of solutions.

Table 3: Specifications of the NIST-digits dataset used in the experiments (RSS: random subspace).

Number of classes	10
Number of features	132
Train Set $\mathcal{T}$	5,000
Optimization Set $\mathcal{O}$	10,000
Validation Set $\mathcal{V}$	10,000
Test Set $\mathcal{G}$	60,089
Number of features for RSS	32
Initial pool size	100

Table 4: NSGA-II parameters

Population size	128
Number of generations	1000
Probability of crossover	0.8
Probability of mutation	0.01
One-point crossover and bit-flip mutation	

We used the NIST digits Special Database 19 (NIST SD19) in our experiments, called NIST-digits. The original dataset was partitioned into four independent datasets: $\mathcal{T}$, $\mathcal{O}$, $\mathcal{V}$ and $\mathcal{G}$, using the classical holdout validation strategy. This is due to the fact that OCS requires at least four datasets. The ensemble creation method is employed using a training dataset ($\mathcal{T}$) to generate the initial pool of classifiers, $\mathcal{C}$. Thus, the search algorithm calculates fitness on $\mathcal{O}$ by testing different candidate ensembles. The best candidate ensemble (C_j^*) is identified in $\mathcal{V}$ to prevent overfitting. Finally, the generalization performance of C_j^* is measured using a test dataset ($\mathcal{G}$). We employ the representation proposed by Oliveira et al. [8], which is composed of 132 features. Table 3 lists important information about the database and the partitions used to compose the four separate sets.

We chose kNN as the base classifier in our experiments. We used $k = 1$ without fine-tuning this parameter in order to avoid additional experiments. The random subspace method was used to generate one pool of 100 homogeneous classifiers. The size of the subsets of features used by the random subspace method is shown in Table 3. The optimization processes were conducted by NSGA-II using binary vectors. Since we used initial pools of classifiers composed of 100 members, each individual was represented by a binary vector with a size of 100. Each bit determines whether a classifier is active (1) or not (0). The genetic parameters are summarized in Table 4.

3.1 Pareto Analysis Results

We have initially calculated IPS and OPS in order to show which measures, summarized in Table 1, are conflicting objective functions. The diversity measures were used by NSGA-II in pairs of objective functions combined with the error rate ϵ. Moreover, NSGA-II employed pairs of objective functions combining either the diversity measures or ϵ with ensemble size. In this way, we have the following three main groups of pairs of objective functions: (1) diversity measures combined with ϵ; (2) ensemble size ζ combined with ϵ; and (3) diversity measures combined with ζ.

Table 5 shows the results obtained for each pair of ob-

jective functions. It is important to mention that the optimization processes were replicated 30 times owing to the use of a stochastic search algorithm. Thus, the values reported in this table were obtained as the mean of the 30 replication results. Table 5 shows the values of IPS and OPS, as well as the difference between the k^{th} objective Pareto spreads. This difference indicates the variation among objective functions values. Moreover, it is implicitly assumed that the non-dominated set obtained at each run of NSGA-II is a good approximation of the true Pareto-optimal.

These results show that all diversity measures are conflicting objective functions when combined with ϵ. Except for Kohavi-Wolpert ψ, the difficulty measure θ and ambiguity γ, there is more variation among diversity values than among ϵ values. Ensemble size ζ and ϵ are also conflicting objective functions and there is more variation among ζ values. However, there are only three diversity measures which are conflicting objective functions when combined with ζ: the same ψ, θ and γ.

The literature has shown that pairwise diversity measures and ζ are not conflicting objective functions. Aksela and Laaksonen [1] observe that pairwise measures always try to find the two most diverse classifiers, which leads to the minimization of the number of classifiers during the optimization process. Our results confirm this observation since all pairwise measures (see Table 1) employed in our experiments are not conflicting objective functions when combined with ζ. However, our results show that most of the non-pairwise measures lead also to the minimization of ζ. Only γ, θ and ψ are conflicting objective functions when combined with ζ. Consequently, these three non-pairwise measures do not minimize ζ. Figure 4 confirms these observations. Figure 4(a) presents a graph containing the size of the classifier ensembles found in 30 replications generated by each pair of objective functions composed by ζ and the 12 diversity measures. We also show in Figure 4(b) the ensemble size of the classifier ensembles found by NSGA-II combining ϵ with diversity and ϵ with ζ. It is important to mention that the minimum number of classifiers allowed by the search algorithms was 5. This fixed minimum ensemble size was defined to avoid generating too small ensembles.

According to Whitaker and Kuncheva [12] non-pairwise measures are calculated by using either entropy or correlation between individual outputs and the ensemble's output or distribution of "difficulty" of the data samples. Among the three diversity measures, which are conflicting objective functions when combined with ζ, two measures are based on the variance over the dataset (θ and ψ), and γ is based on the variance among ensemble's members. Thus, our observation is that, besides the pairwise diversity measures pointed out by Aksela and Laaksonen [1], all non-pairwise diversity measures based on entropy or correlation between individual outputs are also not able to find the best ensemble's size, i.e. they minimize ζ during the optimization process.

Moreover, there is no guarantee of finding the best ensemble's size using one of the three diversity measures based on variance, specially ambiguity, which generates the less diverse Pareto front (Table 5 and Figure 5(a)) and the smallest classifier ensembles (Figure 4(a)). However, these three measures have such a property that makes them able to generate larger classifier ensembles, which is useful in OCS methods since we deal with a large initial pool of classifiers. The two diversity measures based on the variance over the dataset found more diverse Pareto front and larger classifier ensembles than ambiguity (based on the variance of the classifiers' output). The difficulty measure θ was better than Kohavi-Wolpert ψ in these two aspects, as it can seen in Figure 5(b) for ψ and in Figure 2(c) for θ, as well as in Table 5. In next section we analyze how these observations may be related to the performance of the ensembles found at the end of the optimization process.

3.2 Performance Analysis

We continue our experimental study by analyzing the recognition rates achieved by the classifier ensembles found in the selection process of our OCS. Figure 6 shows the performances of the selected classifier ensembles found in 30 replications of the optimization processes. Figure 6(a) shows the performance of the classifier ensembles found by NSGA-II when combining ζ with diversity measures, while Figure 6(b) shows the performances of the ensembles found when combining ϵ with diversity measures and with ζ.

These results show that the three diversity measures γ, θ and ψ, which are conflicting objective functions when combined with ζ, are the most successful diversity measures in terms of performance when NSGA-II was guided by pairs of objective functions made up of ensemble size ζ and diversity. The best diversity measure was θ. However, the performance of the ensembles obtained using ζ and ϵ are better (Figure 6(b)). In fact, our results indicate that diversity combined with ζ in pairs of objective functions does not find high-performance classifier ensembles. Taking into account all the results obtained in the series of experiments, the performances of the ensembles found using these pairs of objective functions showed the worst performances. It is expected that ensembles which are too small will perform considerably less well than other approaches such as combining the initial pools of classifiers or selecting the best subset of classifiers using NSGA-II guided by ϵ combined either with diversity or with ζ. However, the analogy between FSS and OCS may be established. The performance of our baseline system, i.e. the pool of 100 kNN (96.28%), is 0.07% worse than the average result using ζ and ϵ as the objective functions (average of 96.35% in Figure 4), while the averaged ensemble size is 27 classifiers (Figure 4(b)).

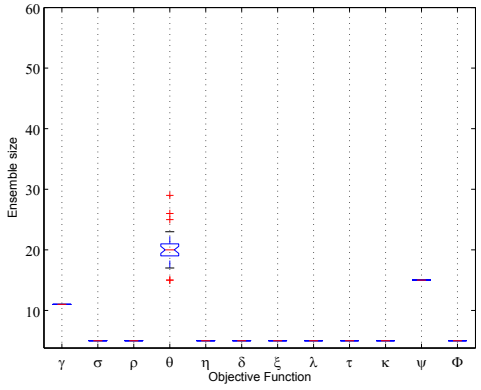
Therefore, we observe that since ensemble size and diversity are not conflicting objective functions, we cannot decrease the generalization error rate by combining this pair of objective functions. However, by including both diversity and ϵ in a multi-objective optimization process, we may find the most high-performance classifier ensembles. The best diversity measures in this main group of pairs of objective functions are difficulty θ, interrater agreement κ, correlation coefficient ρ and double-fault δ.

4. CONCLUSIONS

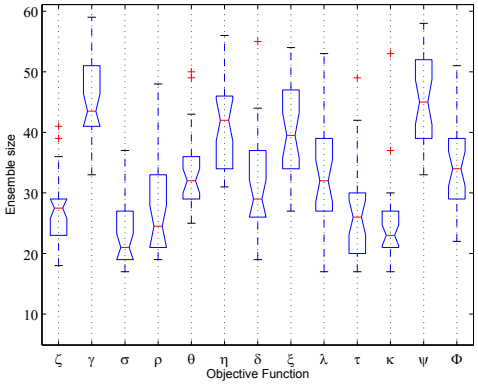
In this paper we have applied two Pareto spread quality measures to analyze the relationship among the three main search criteria used to select classifier ensembles, which are the ensemble error rate, ensemble size and diversity measures. The experiments were conducted using NSGA-II and the three search criteria were combined in pairs of objective functions in three main groups: (1) diversity measures combined with the error rate; (2) ensemble size combined with the error rate; and diversity measures combined with ensemble size.

Table 5: The average k^{th} objective Pareto spread IPS values and the average overall Pareto spread OPS values for each pair of objective function (div: diversity).

	Error rate ϵ				**Ensemble size ζ**			
Diversity	IPS_ϵ	IPS_{div}	$IPS_{div} - IPS_\epsilon$	OPS	IPS_ζ	IPS_{div}	$IPS_\zeta - IPS_{div}$	OPS
γ	0.0187	0.0254	0.0067	0.5 x10^{-3}	0.0503	0.0030	-0.0474	0,0001
σ	0.0134	0.0368	0.0234	0.5 x10^{-3}	0.0000	0.0000	0.0000	0.0000
ρ	0.0213	0.0442	0.0229	0.9 x 10^{-3}	0.0000	0.0000	0.0000	0.0000
θ	0.0025	0.0006	-0.0019	0.2 x 10^{-5}	0.1733	0.0058	-0.1675	0.0010
η	0.0330	0.0428	0.0098	0.14 x 10^{-2}	0.0000	0.0000	0.0000	0.0000
δ	0.0159	0.0068	-0.0091	0.1 x 10^{-3}	0.0000	0.0000	0.0000	0.0000
ξ	0.0249	0.0659	0.0410	0.17 x 10^{-2}	0.0000	0.0000	0.0000	0.0000
λ	0.0192	0.0460	0.0268	0.9 x 10^{-3}	0.0000	0.0000	0.0000	0.0000
τ	0.0150	0.0357	0.0207	0.5 x 10^{-3}	0.0000	0.0000	0.0000	0.0000
κ	0.0159	0.0447	0.0288	0.7 x 10^{-3}	0.0000	0.0000	0.0000	0.0000
ψ	0.0119	0.0113	-0.0006	0.1 x 10^{-3}	0.1000	0.0079	-0.0921	0.0008
Φ	0.0265	0.0927	0.0662	0.25 x 10^{-2}	0.0000	0.0000	0.0000	0.0000
	Ensemble size ζ and Error rate ϵ							
	IPS_ϵ		IPS_ζ		$IPS_\zeta - IPS_\epsilon$		OPS	
	0.0119		0.2163		0.2045		0.0025	

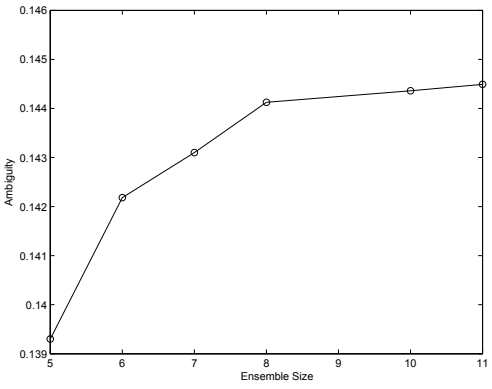


(a)

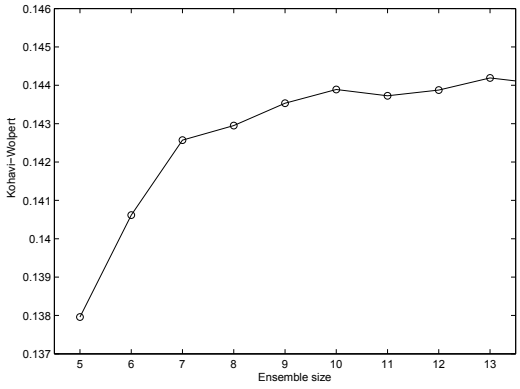


(b)

Figure 4: Size of the ensembles found using 12 diversity measures combined with ensemble size ζ (a) and the error rate ϵ (b). Results from combining ζ with ϵ are also shown in (b). Plus sign indicates outliers.



(a)



(b)

Figure 5: Set of non-dominated solutions found by NSGA-II and the pairs of objective functions: minimize ensemble size and maximize ambiguity (a) and minimize ensemble size and maximize Kohavi-Wolpert (b).

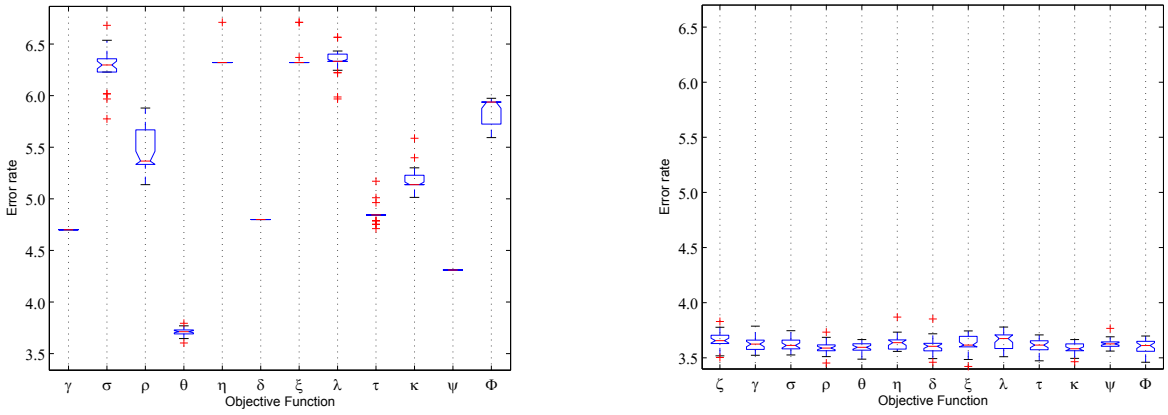


Figure 6: Performance of the classifier ensembles found using NSGA-II with pairs of objective functions made up of 12 diversity measures and ensemble size ζ (a), and the error rate ϵ (b). The results obtained by combining ζ with ϵ are also shown in (b). Plus sign indicates outliers, which are points beyond the ends of the whiskers.

The experiments demonstrated that all diversity measures are conflicting objective functions when combined with the error rate. In addition, ensemble size and the error rate are conflicting objective functions. However, there are only three diversity measures which are conflicting objective functions when combined with ensemble size. As a consequence, since the minimum number of classifiers is often achieved when using this combination of objective functions, we cannot decrease the generalization error rate by combining them. In contrast, by combining the error rate with either diversity or the ensemble size, we may find much more high-performance classifiers. The best performances were obtained by combining error rate with diversity.

For future works we plan to investigate how Pareto spread measures may be used to help dynamic classifier ensemble selection processes.

5. ACKNOWLEDGMENTS

This research was 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.

6. REFERENCES

[1] M. Aksela and J. Laaksonen. Using diversity of errors for selecting members of a committee classifier. *Pattern Recognition*, 39(4):608–623, 2006.

[2] S. Ando and E. Suzuki. Distributed multi-objective ga for generating comprehensive pareto front in deceptive optimization problems. In *Proceedings of the IEEE CEC*, pages 1569–1576, 2006.

[3] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.

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

[5] K. Deb. Unveiling innovative design principles by means of multiple conflicting objectives. *Engineering Optimization*, 35(5):445–470, 2003.

[6] B. Gabrys and D. Ruta. Genetic algorithms in classifier fusion. *Applied Soft Computing*, 6(4):337–347, 2006.

[7] L. Kuncheva and C. Whitaker. Measures of diversity in classifier ensembles an their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[8] L. Oliveira, R. Sabourin, F. Bortolozzi, and C. Suen. Automatic recognition of handwritten numerical strings: A recognition and verification strategy. *IEEE TPAMI*, 24(11):1438–1454, 2002.

[9] F. Roli, G. Giacinto, and G. Vernazza. Methods for designing multiple classifier systems. In *Proceedings of MCS*, pages 78–87, 2001.

[10] D. Ruta and B. Gabrys. Classifier selection for majority voting. *Information Fusion*, 6(1):163–168, 2005.

[11] G. Tremblay, R. Sabourin, and P. Maupin. Optimizing nearest neighbour in random subspaces using a multi-objective genetic algorithm. In *Proceedings of ICPR*, pages 208–211, Cambridge, UK, 2004.

[12] C. Whitaker and L. Kuncheva. Examining the relationship between majority vote accuracy and diversity in bagging and boosting. Technical report, School of Informatics, Unversity of Wales, 2003.

[13] J. Wu and S. Azarm. Metrics for quality assessment of a multiobjective design optimization solution set. *Transactions ASME, Journal of Mechanical Design*, 123:18–25, 2001.

[14] G. Zenobi and P. Cunningham. Using diversity in preparing ensembles of classifiers based on differente feature subsets to minimize generalization error. In *Proceedings of XII European Conference on Machine Learning*, pages 576–587, Freiburg, Germany, 2001.