A Comparative Analysis of Genetic Algorithm and Ant Colony Optimization to Select Attributes for an Heterogeneous Ensemble of Classifiers

Laura E A Santana, Ligia Silva Anne M P Canuto, Fernando Pintro and Karliane O Vale

***Abstract*— In the context of ensemble systems, feature selection methods can be used to provide different subsets of attributes for the individual classifiers, aiming to reduce redundancy among the attributes of a pattern and to increase the diversity in such systems. Among the several techniques that have been proposed in the literature, optimization methods have been used to find the optimal subset of attributes for an ensemble system. In this paper, an investigation of two optimization techniques, genetic algorithm and ant colony optimization, will be used to guide the distribution of the features among the classifiers. This analysis will be conducted in the context of heterogeneous ensembles and using different ensemble sizes.**

I. Introduction

One of the possible alternatives to enhance the efficiency of pattern recognition systems is to combine several classification systems within one structure, forming a classifier combination system. The main example of this type of system is ensemble (also known as committees). Ensembles are based on the idea that a pool of different classifiers can offer complementary information about the patterns to be classified. In such a context, the combination of these classifiers aims to improve the accuracy of the whole classification process. As a consequence of this, in the last decade, ensembles of classifiers have been widely used for several pattern recognition tasks. These systems are composed of a set of individual (base) classifiers, organized in a parallel way, that receive the input patterns and send their output to a combination method which is responsible for providing the final output of the system.

However, the use of ensembles can lead to an increasing in the time processing of a system, since they are more complex than single classifiers. In this case, the use of ensembles has to be well justified in order to overcome the increased complexities of these systems. In addition, there is clearly no accuracy gain in a system that is composed of a set of identical individual classifiers [8, 12]. In fact, the ideal situation is a set of individual classifiers with uncorrelated errors — they are combined in such a way as to minimize the effect of these failures. In other words, the individual classifiers should be diverse among themselves.

The use of feature selection or data distribution methods in ensemble systems usually increases the diversity of the members of an ensemble. This is because the individual classifiers will perform the same task (classification of the same input patters) but they were built using different subsets of features. In addition, these methods can reduce the dimensionality of the individual classifiers, reducing the overall complexity of the ensemble systems. In this case, the use of feature selection methods can help to reduce the complexity of ensemble as well as to increase the diversity of the individual classifiers of these systems.

There are some papers in the literature which address feature selection methods in ensembles, such as in [3, 5, 6, 10, 11, 13, 14, 15, 17, 19, 23]. However, the majority of these papers address the homogenous structures of ensemble. Unlike these works, this paper will investigate the use of feature selection methods in the context of heterogeneous ensembles (individual classifiers are generated by applying different learning algorithms).

Of the techniques used to select attributes for the individual classifiers, the optimization techniques have also been successfully used, such as in [13, 14, 17, 19, 23]. Of the optimization techniques, genetic algorithms [3, 7, 14, 17] and ant colony optimization [13, 19, 23] are among the techniques more used. Although these techniques have been widely used, there is no comparative analysis of more than one optimization technique to select the attributes in the context of heterogeneous ensembles and with different ensemble sizes. In this paper, an investigation of the use of genetic algorithm and ant colony optimization for finding the optimal subset of attributes will be done. The main aim of this analysis is to evaluate the performance of both optimization techniques in the choice of the attributes for an

Manuscript received February 05, 2010. This work was supported in part by the CNPq, under processes number 556424/2008-5 and 200755/2009-9.

Laura E A Santana is with Informatics and Applied Mathematics Department – Federal University of Rio Grande do Norte (UFRN), Natal, RN - BRAZIL, 59072-970 Tel.: +55-84-3215-3815, lauraemmanuella@yahoo.com.br

Ligia Silva is with Informatics and Applied Mathematics Department – Federal University of Rio Grande do Norte (UFRN), Natal, RN - BRAZIL, 59072-970 Tel.: +55-84-3215-3815 ligia.cefet@gmail.com

Anne M P Canuto is with Informatics and Applied Mathematics Department – Federal University of Rio Grande do Norte (UFRN), Natal, RN - BRAZIL, 59072-970 Telephone:+55-84-3215-3815, anne@dimap.ufrn.br.

Fernando Pintro is with Informatics and Applied Mathematics Department – Federal University of Rio Grande do Norte (UFRN), Natal, RN - BRAZIL, 59072-970 Tel.: +55-84-3215-3815 fernoxp@hotmail.com

Karliane O Vale is with Informatics and Applied Mathematics Department – Federal University of Rio Grande do Norte (UFRN), Natal, RN - BRAZIL, 59072-970 Tel.: +55-84-3215-3815 karlianev@gmail.com

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

ensemble system. In order to do this analysis, heterogeneous structures will be used in ensembles composed of 3, 6 and 12 individual classifiers. These systems will use five different combination methods.

This paper is divided into seven sections and it is organized as follows. Section 2 describes the research works related to the subject of this paper. In Section 3, a brief description of ensemble systems is presented, focusing on the use of feature selection methods in ensemble systems. Section 4 presents a brief description of genetic algorithm and ant colony optimization, while the description and results provided by the empirical analysis are shown in Sections 5 and 6. Finally, Section 7 presents the final remarks of this paper.

II. RELATED WORKS

Feature selection methods try to reduce the dimensionality of the attributes of a dataset, spotting the best ones. The attribute subset selection can be defined as the process that chooses the best attributes subset according to a certain criterion, excluding the irrelevant or redundant attributes. In using feature selection methods, it is aimed to improve the quality of the obtained results. In the context of ensembles, for instance, feature selection methods provide different subsets of attributes for the individual classifiers, aiming to reduce redundancy among the attributes of a pattern and to increase the diversity in such systems. In this paper, feature selection methods will be used in ensemble systems. Hence, hereafter, the term feature selection will be used in the context of ensembles.

Recently, several authors have investigated the use of feature selection methods in ensembles, such as in [3, 5, 6, 10, 11, 13, 14, 15, 17, 19, 23]. These methods can be broadly divided into two main approaches, which are:

- Filter: In this approach, as it can be found in [16, 18], there is no need for a classification method to be used during the feature selection process. In other words, the feature selection process is independent from the classification method. In [18], for instance, the authors used different criteria to rank the attributes and to distribute them among the classifiers of different structures of ensembles;
- Wrapper: In this approach, as it can be found in [3, 6, 15], the feature selection process is dependent from the classification method. The feature subset is chosen based on the classification method used. Two different classification methods lead to different feature subset chosen.

In traditional feature selection, the wrapper approach involves the computational overhead of evaluating candidate feature subsets by executing a given learning algorithm on the dataset using each feature subset under consideration. The filter approach is generally computationally more efficient than the wrapper approach. However, the major drawback of the filter approach is that an optimal selection of features may not be independent of the inductive and representational biases of the learning algorithm that is used to construct the classifier.

In the context of ensembles, the major drawback of the wrapper approach is emphasized, since the fitness function usually has to take into consideration the accuracy of the whole ensemble system, increasing even further the complexity of this function. In some works, as in [17], the fitness function is based on the accuracy of individual classifiers and only after the choice of a reduced set of attributes is done, these attributes are distributed over the individual classifiers. However, these works do not consider the accuracy of an ensemble and, most of the time, the used fitness functions do not reflect the real situation of the ensemble systems. Alternatively, the work in [25] has used both filter and wrapper approaches, in which the filter approach was used in the first phase, while a genetic search is employed in the second phase. It does smooth out the problem, but it still has a reasonable complexity involving in its processing.

Additionally, as the ensemble system used a two-step decision making process (the individual classifier level and the combination method level), the dependency of the chosen subset with the classification methods can be smoothed out, since the classifier accuracy is not the only parameter to define the accuracy of the ensemble system. Furthermore, when using heterogeneous ensembles, different types of classification methods are used and a poor performance of one individual classifier (bad choice of the subset of attributes) can be overcome by a good performance of a different classifier. Because of this, the use of the filter approach to select different subsets of attributes for the individual classifiers in an ensemble has become an interesting (and efficient) choice.

Independently of the used approach, optimization techniques can also be used to automate the search for the optimum attribute subsets. Recently, several authors have investigated genetic algorithms (GA) to design ensemble of classifiers [4, 7, 15, 16]. In [7], for instance, authors suggest two simple ways to use genetic algorithm to design an ensemble of classifiers. They present two versions of their algorithm. The former uses just disjoint feature subsets while the latter considers (possibly) overlapping feature subsets. The fitness function employed is the accuracy of the ensemble.

On the other hand, works using ant colony optimization (ACO) for selecting attributes to an ensemble system can be found in [13, 19, 23]. In [13], the authors used a ACO to select attributes for an ensemble system. In this work, a priori information about the attributes and the size of the attribute set were used as heuristic information during the ACO processing. The authors applied their system in face recognition and they showed that the use of ACO was more positive than those based on genetic algorithms.

Thus, unlike most of the previous works reviewed, this paper uses feature selection methods in the context of heterogeneous ensembles. In addition, it analyzes comparatively the performance of two optimization

techniques in order to select the optimum subset of attributes. However, this analysis is conducted in heterogeneous structures and applied in ensembles of different sizes.

III. Ensemble systems

As previously mentioned, the goal of using ensembles is to improve the performance of a pattern recognition system in terms of better generalization and/or in terms of increased efficiency and clearer design [12]. There are two main issues in the design of an ensemble: the ensemble components and the combination methods that will be used. In relation to the first issue, the members of an ensemble are chosen and implemented. The correct choice of the set of individual classifiers is fundamental to the overall performance of an ensemble. The ideal situation would be a set of individual classifiers with uncorrelated errors - they would be combined in such a way as to minimize the effect of these failures. In other words, the individual classifiers should be diverse among themselves. According to its structure, an ensemble can be divided in two main approaches: heterogeneous and homogeneous. The first approach combines different types of classification algorithms as individual classifiers. On the other hand, the second approach combines classification algorithms of the same type.

Once a set of classifiers has been created and selected, the next step is to choose an effective way of combining their outputs. The choice of the best combination method for an ensemble implies in the performance of exhaustive testing. Indeed, the choice of the combination method of an ensemble is very important and difficult to achieve. There are a great number of combination methods reported in the literature [12]. According to their functioning, there are three main strategies of combination methods: fusion-based, selection-based, and hybrid methods.

- Fusion based Methods: In this class, it is assumed that all classifiers are equally experienced in the whole feature space and the decisions of all of classifiers are taken into account for any input pattern. There are a vast number of fusion-based methods reported in the literature. Examples of this class are: sum, majority voting, naïve bayesian, neural networks, fuzzy neural networks, fuzzy connectives among others;
- Selection-based Methods: In this class, only one classifier is needed to correctly classify the input pattern in selection-based methods. In order to do so, it is important to define a process to choose a member of the ensemble to make the decision, which is usually based on the input pattern to be classified. The choice is typically based on the certainty of the current decision. Preference is given to more accurate classifiers. One of the main methods in classifier selection is dynamic classifier selection (DCS), proposed in [24];
- Hybrid Methods: They are the ones in which selection and fusion techniques are used in order to provide the most suitable output to classify the input pattern. Usually, there is a criterion process to decide whether to use the selection or combination method. The main idea is to use selection only and if only the best classifier is really good to classify the testing pattern. Otherwise, a combination method is used. Two main examples of hybrid methods are: dynamic classifier selection based on multiple classifier behavior (Dcs-MCS) and dynamic classifier selection using also decision templates (Dcs-DT) [12].

In this paper, five different combination methods will be used, in which one of them is a hybrid-based method (Dcs-MCB) and the remaining four methods are fusion-based ones (2 non-trainable –sum and voting – and 2 trainable – naïve bayesian and a MLP neural networks).

A. Diversity in Ensembles

As already mentioned, there is no gain in ensembles that are composed of a set of identical classifiers. The ideal situation, in terms of combining classifiers, would be a set of classifiers that present uncorrelated errors. In other words, the ensemble must show diversity among the members in order to improve the performance of the individual classifiers. Diversity in ensemble systems can be reached when the individual classifiers are built under different circumstances, such as in the following ways:

- Different parameter settings of the classifiers: In this approach, diversity can be reached through the use of different initial parameters setting of the classification methods. In a neural network, for instance, this would mean varying the weights and topology of a neural network model.
- Different classifier training datasets: In this approach, diversity can be reached through the use of learning strategies such as Bagging and Boosting or the use of feature distribution methods.
- Different classifier types: In this approach, diversity can be reached through the use of different types of classifiers. For instance, usually an ensemble which is composed of neural network and decision tree is more diverse than ensembles composed of only neural networks or decision trees.

In this paper, variations of the diversity are captured when using different types of classifiers and different training datasets (feature selection methods).

IV. Optimization techniques

Optimization can be defined as the search for the optimal solution for a given problem. The idea is to find the optimal values to solve a problem and the techniques used in this search are called optimization techniques. Several techniques have been proposed to solve these problems. Intuitively, it is possible to state that these techniques try to optimizing the value of some objective function, subject to any resource and/or other constraints such as legal, input, environmental, and behavioral restrictions.

However, there are some problems that cannot be tackled by classical methods, since it either needs a high computing demanding or it is unfeasible to solve. In this case, this drawback generates demand for other types of algorithms, such as heuristic optimization approaches. In the next two subsections, two heuristic optimization techniques (genetic algorithm and ant colony optimization) will be briefly described.

A. Genetic Algorithm

Algorithms (GAs) were first developed by Holland in 1960 [9]. They are considered as stochastic global optimization methods inspired by biological mechanisms such as evolution and hereditary. Lately, they have been widely used in different tasks, such as numerical optimization [28], optimization of neural networks [10], in multiagent systems [16], among others.

When using genetic algorithms, the search space is used to build the chromosomes, in which each possible solution for a problem is coded as a chromosome (individual). The set of these individuals is called population. The initial population (population of the first interaction) of a genetic algorithm can be chosen in several ways, being the most common way the random choice.

Once an initial population is created, the individuals of this population are assessed trough a fitness function, which informs the goodness of a chromosome in the solution of the optimization task. In other words, it indicates how close a chromosome is from the optimal solution. In addition, fitness is the function to be optimized (minimization or maximization) in a genetic algorithm. Based on this fitness function, chromosomes are selected and some genetic operators (mutation, crossover and so on) are applied in the selected chromosomes, forming new ones. The idea is that these chromosomes evolve, always creating better individuals until it reaches the global optimum [9,10].

B. Ant-Colony Optimization

In the early 1990s, ant colony optimization (ACO) was introduced by M. Dorigo as a novel nature-inspired metaheuristic for the solution of hard combinatorial optimization (CO) problems [22]. ACO belongs to the class of metaheuristics, which are approximate algorithms used to obtain good enough solutions to hard CO problems in a reasonable amount of computation time.

Some studies of the ethologists tried to understand how almost blind animals like ants could manage to establish shortest route paths from their colony to feeding sources and back. As a result of these studies, it was found that the ants used pheromone trails to communicate information among individuals regarding paths, and was used to decide where to go. A moving ant lays some pheromone (in varying quantities) on the ground, thus marking the path by a trail of this substance. While an isolated ant moves essentially at random, an ant encountering a previously laid trail can detect it and decide with high probability to follow it, thus reinforcing the trail with its own pheromone. The collective behaviour that emerges is a form of autocatalytic behaviour where the more the ants following a trail, the more attractive that trail becomes for being followed. The process is thus characterized by a positive feedback loop, where the probability with which an ant chooses a path increases with the number of ants that previously chose the same path.

Based on these studies, the ant-colony optimization (ACO) technique was proposed. ACO is a population-based metaheuristic which distributes the search activities over so-called "*ants*". In other words, the activities are divided among agents with very simple basic capabilities which, to some extent, mimic the behaviour of real ants in the search for food.

It is important to emphasize that ACO has not been created as a simulation of ant colonies, but to use the metaphor of artificial ant colonies and apply them as an optimization tool.

The idea of ACO is to model the problem as an environment, in which it is possible to create a set of artificial ants which move around this environment. In the beginning of the processing, as there is no information about the path to go from one point to the other, the choice of the ants about which way to is completely random. During the processing, the idea is that if at a given point an ant has to choose among different paths, those which were heavily chosen by preceding ants (that is, those with a high trail level) are chosen with higher probability. Furthermore high trail levels are synonymous with short paths.

In general perspective, the ACO approach attempts to solve an optimization problem by repeating the following two steps:

- The candidate solutions are constructed using a pheromone model, that is, a parametrized probability distribution over the solution space;
- The candidate solutions are used to modify the pheromone values in a way that is deemed to bias future sampling toward high quality solutions.

V. Experimental Setting Up

In order to analyze the performance of the optimization techniques in the choice of the attributes for ensemble systems, an empirical analysis is performed in the context of heterogeneous ensembles. In this analysis, the performance of the ensemble systems without any feature selection (called Ensembles with no distribution, or simply, No Dist) will be compared with ensembles using attributes chosen by genetic algorithm (GA-based ensembles) and ant colony optimization (ACO-based ensembles). The idea is to evaluate the effect of using these heuristic optimization techniques in the choice of attributes for ensembles.

As already mentioned, three different ensemble sizes will be used in this investigation, using 3, 6 and 12 individual classifiers. For each system size, several different configurations will be used. In all these configurations, different base components (heterogeneous systems) will be used. Three types of individual classifiers are used to

compose the individual classifiers of the ensembles, k-NN (k Nearest Neighbor), DT (Decision Tree) and a neural network (MLP - Multilayer Perceptron). As there are several possibilities of combination of the individual classifiers, this paper presents the average of the accuracy delivered by all possibilities of the heterogeneous structures.

For both optimization techniques, the fitness of the possible solutions is analyzed in terms of a correlation measure. Pearson's Product Moment Correlation Coefficient (PMCC) is a non-parametric measure of correlation. In other words, it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables [20]. This measure can be used, for instance, to define the correlation (dependency) of the features of an input pattern. This measure will be used in one parameter called intra-classifier correlation.

- Intra-classifier correlation (or simply intra correlation or intra): it is a criterion that defines the correlation within one classifier. In other words, it describes the correlation that might exist among the attributes of one classifier. The correlation of each classifier is calculated and it is then averaged to provide the intra-class correlation. The main aim of this criterion is to choose attributes for one classifier which are as uncorrelated as possible. It is envisaged to focus on the diversity of the classifiers separately;

In order to obtain a better estimation of the accuracy rates, a 10-fold cross validation method is applied to all ensembles (as well as individual classifiers). Thus, all accuracy results presented in paper refer to the mean over 10 different test sets. Furthermore, some of the combination methods are trainable methods. In this sense, the datasets are divided into 11 folds (subsets) of equal size (keeping the distributions of the classes in each fold). From these sets, 10 of them are used in the 10-fold cross validation procedure and the remaining one works as the validation set to train the combination methods. In addition, as the optimization techniques used are a non-deterministic, 10 runs of each technique were performed.

In order to compare the accuracy of the ensemble systems, a statistical test will be applied, which is called hypothesis test (t-test) [21]. It is a test which involves testing two learned hypotheses on identical test sets. In order to perform the test, a set of samples (ranking values) from both algorithms should be used. Based on the information provided, along with the number of samples, the significance of the difference between the two sets of samples, based on a degree of freedom (α), is defined. In this paper, the confidence level is 95% ($\alpha = 0.05$).

A. Datasets

Two different datasets are used in this investigation, which are described as follows:

- Protein dataset: It is a protein dataset which represents a hierarchical classification, manually detailed, and represents known structures of proteins. They are organized according to their evolutionary and structural relationship. The main protein classes are all-α, all-β, α/β, α+β and small. A total of 126 attributes is used in this dataset. It is an unbalanced dataset, which has a total of 582 patterns, in which 111 patterns belong to class all-α, 177 patterns to class all-β, 203 patterns to α/β, 46 patterns to class α+β and 45 patterns to class small.
- Outdoor Images (or simply image) dataset. This dataset was taken from the UCI repository (segmentation dataset) [1]. The 2.310 instances were drawn randomly from a dataset of 7 outdoor images. The images were hand-segmented to create a classification for every instance, where each instance is a 3x3 region. Eighteen attributes were extracted from each region.

B. Genetic Algorithm

In order to generate the initial population for the genetic algorithm, an initial pool with a pre-defined number of classifiers is used. In this paper, a population of 50 chromosomes is used. A binary chromosome is used to represent a possible solution for the problem. The size of the chromosome is L *X N,* where *L* represents the number of classifiers of the ensemble and *N* represents the number of attributes of the dataset. In this chromosome, the first N bits will represent the feature subset for classifier C_1, followed by the N bits for classifier C_2, and so on. Figure 1 represents a typical example of an individual (ensemble) that represents an ensemble composed of three classifiers (C_0, C_1 and C_2) and the dataset has 10 attributes. In this example, the attributes 1, 2, 5, 7 are assigned to classifier C_0, attributes 1, 7 and 8 are assigned to classifier C_1 and attributes 7 and 10 are assigned to classifier C_2.

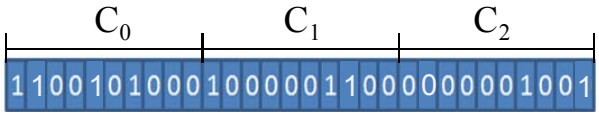

Figure 1. Chromosome representation of the genetic algorithm

The genetic algorithm applies crossover and mutation genetic operator. It also keeps the best individual to the next generation (elitism). Its ending condition is the maximum number of iterations, which were different for both datasets (500 for image and 2500 for protein dataset).

C. Ant-Colony Optimization

The use of ant colony optimization to find the optimal subsets of features for the individual classifiers of an ensemble requires its representation as a graph, where the vertices (node) are the attributes of the dataset and the edges represent the trails to the next attributes. In other words, when there is a trail which links two attributes *i* and *j,* this means that attribute *j* was chosen right after the choice of

attribute *i* in the solution of the problem. In this case, the search for the optimal subset of attributes can be represented by the trail of an ant in the graph, where a certain number of X nodes (attributes) have been visited and a termination conditions is met. In this paper, a initial analysis was carried a value of N/2 was found to be the best value for X, where N is the number of attributes of the dataset, was found to reach the best results.

In this paper, the original ACO algorithm was applied, such as in [22]. In the ACO algorithm, the probability of an ant *k* choose a trail (i,j) is given by equation (1),

$$P_{ij}^{k}(t) = \frac{\tau_{ij}^{\alpha}(t) \cdot n_{ij}^{\beta}}{\sum_{x \in N^k} \tau_{ix}^{\alpha}(t) \cdot n_{ix}^{\beta}}, \quad if\ i \in N^k \qquad (1)$$

Where:

N^k is the subset of attributes which has not been yet chosen by ant *k*;

$\tau_{ij}(t)$ is the amount of pheromone in the trail which links attributes *i* and *j*;

n_{ij} is the heuristic function of the trail which links attributes *i* and *j* and it is given by $1/CR_{ij}$, where CR_{ij} is the intra-correlation between attributes *i* and *j* (described in Section V);

α and β are weigts assigned to the pheromone and the heuristic information, respectively.

In the pheromone update for each trail, the following equation is used.

$$\tau_{ij}^{(t)} = (1-\rho)\tau_{ij}^{(t-1)} + \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \qquad (2)$$

Where:

ρ represents the evaporation rate of trail until time t;

$\tau_{ij}^{(t-1)}$ is the intensity of pheromone on the trail (i,j) at time t-1;

$\Delta\tau_{ij}^{k}$ is the amount of pheromone on trail (i,j) on the current iteraction. This value can be calculated using the following equation.

$$\Delta\tau_{ij}^{k} = \begin{cases} Q/L_k, if\ the\ k-th\ ant\ uses\ i\ in\ your\ solution \\ 0, otherwise \end{cases} \qquad (3)$$

Where:

Q is a constant;

L_k is the intra-correlation of the final solution built by ant k.

In this paper, a colony is composed of 30 ants. Each ant is responsible for building a possible solution at each iteration of the algorithm. After a defined number of iterations (termination condition), the solution which contains the highest intra-correlation value is chosen as being the final solution.

VI. Results and Discussion

In this section, the results of the empirical analysis are illustrated. Table I shows the results (accuracy and standard deviation) of the ensemble systems with three individual classifiers, using feature selection (ACO and GA) and without (No Dist). These systems were applied to the datasets described in Section V.A.

TABLE I: ACCURACY AND STANDARD DEVIATION OF THE ENSEMBLE SYSTEMS COMPOSED OF THREE CLASSIFIERS

	IMAGE DATASET		
	No Dist	AG	ACO
Vote	96.80±1.64	94.14±2.75	96.94±1.83
Sum	97.82±1.86	95.07±3.03	97.42±2.03
NN	97.33±1.77	93.36±4.01	94.50±3.88
Naive	97.95±1.92	95.61±2.64	97.16±2.07
DCS	97.95±1.63	94.04±3.02	96.93±2.06
	SCOOP DATASET		
	No Dist	AG	ACO
Vote	79.56±3.30	79.17±3.29	80.39±4.02
Sum	74.71±5.58	73.72±4.27	75.36±4.93
NN	81.19±4.15	78.72±3.26	80.01±4.16
Naive	80.69±3.05	79.24±3.92	80.11±4.1
DCS	79.68±3.38	78.52±3.35	79.84±4.1

In a general perspective, it can be noticed from Table I that the ensemble systems using both feature selections had accuracy level lower than the ensembles without feature selection. It is believed that the small number of individual classifiers is the main reasons for this decrease in the accuracy level. It is expected that the use of feature selection causes a decrease in the performance of the individual classifiers. In using a small number of individual classifiers, the combination of weaker classifiers was not sufficient to avoid a decrease in the performance of the ensemble systems.

However, it could be observed that, when using ACO, the accuracy of the ensemble systems had improvements in all analyzed cases, when compared with the ensembles using GA. The highest difference of accuracy level (ACO versus GA) was reached when using DCS combination method for Image and using Sum method for protein dataset.

When applying the statistical test to analyze the the feature selection methods (ACO versus GA), a t-test was applied comparing the accuracy of ACO-based ensembles with GA-based ones. As a result of this test, it was found that the number of statistically significant improvements was 9 (out of 10). The only case in which there is no evidence to state that the accuracy of ACO-based ensembles is statistically higher than GA-based ensembles is when using Naïve combination method for protein dataset.

Still in the analysis from a perspective of the statistical *t*-test, the increase in the accuracy level of the ensemble systems without feature selection was compared with the best feature-based ensembles (ACO). The test has

shown that the improvements were statistically significant in only 3 cases (out of 10). The use of neural network as a combination method was not positive for ensembles using ACO, since the decrease of accuracy was statistically significant for both datasets (p-value = 3.01E-10 for Image and p-value = 0.045 for protein). In addition, the use of DCS combination for image dataset caused a statistically significant decrease in the accuracy level of ACO-based ensembles (p-value = 0.001).

In summarizing, the use of ACO was positive since ACO-based ensembles had statistically significant improvements, when compared with GA-based ensembles. In addition, they had performance similar with ensembles without feature selections (only 3 statistically significant decreases).

A. Ensembles with 6 individual Classifiers

Table II presents the accuracy level and standard deviation of ensemble systems with six individual classifiers. It can be seen from Table II that the increase in the number of individual classifiers caused an increase in the accuracy level of the feature selection-based ensembles (ACO and GA). In Table II, it is possible to see that the highest accuracy level was reached by either ACO or GA, in some cases (for example, using Sum method for both datasets). In relation to the optimization technique, as it occurred in Table I, the use ACO led to an improvement in the accuracy level in most of the analyzed cases. However, the accuracy levels of GA-based ensembles are much closer to the ACO-based ensembles, when compared with ensembles with three individual classifiers.

TABLE II: ACCURACY AND STANDARD DEVIATION OF THE ENSEMBLE SYSTEMS COMPOSED OF SIX CLASSIFIERS

	IMAGE DATASET		
	No Dist	AG	ACO
Vote	96.12±1.71	92.98±4.39	91.68±4.77
Sum	96.82±1.83	97.00±2.45	97.36±2.27
NN	97.82±1.66	96.79±2.48	97.13±2.43
Naive	97.90±1.68	97.75±1.83	97.50±2.04
DCS	97.79±1.67	96.6±2.29	97.16±2.13
	SCOOP DATASET		
	No Dist	AG	ACO
Vote	79.69±4.19	80.30±3.53	80.72±3.99
Sum	74.77±5.09	75.70±4.56	76.61±4.59
NN	80.56±3.41	81.66±3.56	82.01±3.46
Naive	80.81±2.50	81.51±3.84	81.64±3.90
DCS	81.25±2.88	81.29±3.50	81.88±3.66

When analyzing the results of the feature selection methods from a perspective of the statistical *t*-test, the number of statistically significant improvements reached by ACO was 1 (out of 10), when compared with GA-based ensembles. It is a massive reduction in relation to the results in Table I. It shows that the improvement in the accuracy level happened in both feature selection methods, making them have similar performance.

In comparing ACO-based ensembles with ensembles without feature selection from the perspective of the statistical *t*-test, it is observed an inversion in the behavior of the systems. Instead of having statistically significant decreases, ACO-based ensembles have now statistically significant increases in 4 cases (Vote, Sum, NN and DCS for protein dataset). It proves the improvement in the accuracy of the ACO-based ensembles when increasing the number of individual classifiers.

B. Ensembles with 12 individual Classifiers

Table III shows the results of the ensemble systems composed of twelve individual classifiers. As it happened in Table II, the accuracy level of the feature selection-based methods increased and surpassed the accuracy of the ensembles without feature selection in most of the cases. In addition, the performance of GA-based ensembles improved, when compared with ACO-based ensembles.

TABLE III: ACCURACY AND STANDARD DEVIATION OF THE ENSEMBLE SYSTEMS COMPOSED OF TWELVE CLASSIFIERS

	IMAGE DATASET		
	No Dist	AG	ACO
Vote	96.82±1.78	90.12±6.27	88.33±5.42
Sum	97.88±1.82	98.29±1.72	97.17±2.37
NN	97.90±1.60	98.31±1.72	97.05±2.29
Naive	98.03±1.59	98.87±1.22	97.79±1.75
DCS	97.90±1.37	98.27±1.64	97.09±2.21
	SCOOP DATASET		
	No Dist	AG	ACO
Vote	80.18±5.37	79.50±4.37	80.96±3.72
Sum	77.73±5.11	77.58±4.66	77.25±4.59
NN	80.94±2.78	82.71±3.75	82.94±3.37
Naive	80.94±2.19	82.80±4.1	82.67±3.77
DCS	81.25±2.76	82.62±3.83	82.65±3.74

When applying the statistical test in the feature selection-based methods, the GA-based ensembles had statistically significant improvements in 5 cases, all of them in the Image dataset. In analyzing the opposite way (ACO-versus GA), it was observed that there is no statistical evidence to state that the accuracies of the ensembles using ACO are different from the results with ensembles with GA for all cases.

In the statistical analysis of the systems with and without feature selection, the GA-based and ACO-based ensembles had statistically significant improvements in 3 cases (NN, Naïve and DCS for protein dataset). In the remaining 7 cases, there is no statistical evidence to state that the accuracies of the ensembles using feature selection are different from the results with ensembles without feature selection.

VII. FINAL REMARKS

In this paper, an investigation of two optimization techniques to feature selection in ensembles was performed. This analysis was done in the context of heterogeneous ensembles and three ensemble sizes were used, which were: 3, 6 and 12 individual classifiers. These ensemble systems were applied to two different datasets.

Through this analysis, it could be observed that as the number of individual classifiers increased, the performance of the feature selection based ensembles increased and surpassed the performance of the ensembles without feature selections. When using ensembles with 12 individual classifiers, for instance, the accuracy level of the feature selection based ensembles was higher than the ones without feature selection in most of the analyzed cases.

In relation to the performance of the individual optimization techniques, it could be noted that ACO has affected more positively the accuracy of the ensembles with fewer individual classifiers (3 and 6). As the number of individual classifiers increased, the performance of the GA-based ensembles increased and they become the ensembles with the highest accuracy for most of cases. In other words, based on this empirical analysis, it is possible to conclude that when using small ensembles (small number of individual classifiers), the best option is ACO. However, for large ensembles, the best choice is GA.

Of course that the results obtained in this paper is still initial, since it used only two datasets. However, a wider investigation using more datasets and ensemble sizes and structures is the subject of an ongoing research.

REFERENCES

[1] Blake, C., & Merz, C. (1998). UCI Machine Learning Repository. Retrieved 07 01, 2008, from http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2003). A new ensemble diversity measure applied to thinning ensembles. 4th International Workshop on Multiple Classifier Sysems, (pp. 306-316).

[3] Chandra, A. (2004). Evolutionary framework for the creation of diverse hybrid ensembles for better generalisation. Master thesis, University of Birmingham, School of Computer Science.

[4] Chandra, A., & Yao, X. (2005). Evolutionary Framework for the Construction of Diverse Hybrid Ensembles. 13th European Symposium on Artificial Neural Networks, (pp. 253-258). Bruges (Bélgica).

[5] M. Bacauskiene, A. Verikas, A. Gelzinisa and D.Valinciusa (2009). A feature selection technique for generation of classification committees and its application to categorization of laryngeal images. Pattern Recognition 42(5), 645 – 654, 2009.

[6] J Derrac, S García and F Herrera (2009). A First Study on the Use of Coevolutionary Algorithms for Instance and Feature Selection. 4th International Conference of Hybrid Artificial Intelligence Systems (HAIS), LNCS 5572, 557-564, 2009.

[7] Kuncheva, L., Jain, L.C.: Designing classifier fusion systems by genetic algorithms. IEEE Trans. Evol. Comput. **4**(4), 327–336 (2000)

[8] Hansen, L., & Salamon, P. (1990). Neural network ensembles. IEEE Trans Pattern Analysis and Machine Intelligence , 12, pp. 993-1001.

[9] Holland, J. H. (1992). Adaptation in Nature and Artificial System. Cambridge: MIT Press.

[10] J Hua, W D. Tembe and E R. Dougherty (2009). Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition 42(3), 409-424, 2009

[11] R Jensen and Q Shen (2009). New Approaches to Fuzzy-Rough Feature Selection. IEEE Transactions on Fuzzy Systems, 17(4), 824-838.

[12] Kuncheva, L. I. (2004). Combing Pattern Classifiers. New Jersey: Wiley.

[13] H Kanan and K Faez. An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system. Applied Mathematicsand Computation, pp. 716-725, 2008.

[14] Lee, M. et al. A two-step approach for feature selection and classifier ensemble construction in computer-aided diagnosis. IEEE International Symposium on Computer-Based Medical System, pp. 548-553,. Albuquerque: IEEE Computer Society, 2008.

[15] Liu, Y., Yao, X., & Higuchi, T. (2000). Evolutionary ensembles with negative correlation learning. IEEE Transactions on Evolutionary Computation , 4 (4), p. 380.

[16] Oliveira, D. F., Canuto, A., & Campos, A. (2006). GNeurAge: An Evolutionary Agent-Based System for Classification Tasks. Sixth International Conference on Hybrid Intelligent Systems, (pp. 11-11). Rio de Janeiro.

[17] L Oliveira, M Morita and R Sabourin. Feature selection for ensembles applied to handwriting recognition. International Journal of Document Analysis , 262-279, 2006.

[18] Sylvester, J., & Chawla, N. V. (2006). Evolutionary Ensemble Creation and Thinning.

[19] K Robbins, W Zhang and J Bertrand. The ant colony algorithm for feature selection in high-dimension gene expression data for disease classiication. Mathematical Medicine and Biology (pp. 413-426). Oxford University Press, 2007.

[20] P Chen and P Popovich, Correlation: Parametric and Nonparametric Measures. Sage Publications, 1st edition, 2002

[21] J Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research 7: 1–30, 2006.

[22] M. Dorigo, Optimization, learning and natural algorithms (in Italian), Ph.D. Thesis, Dipartimento di Elettronica, Politecnico di Milano, Italy, 1992.

[23] Y He, D. Chen and W Zhao. Ensemble classifier system based on ant colony algorithm and its application in chemical pattern classification. Chemometrics and Intelligent Laboratory Systems, pp. 39-49, 2006.

[24] Woods, K., Kegelmeyer, W. P., & Bowyer, K. (1997). Combination of multiple classiers using local accuracy estimates. IEEE Transactions on Pattern Analysis and Machine Intelligence , 19, pp. 405-410.

[25] M. Bacauskiene, A. Verikas, A. Gelzinisa and D.Valinciusa (2009). A feature selection technique for generation of classification committees and its application to categorization of laryngeal images. Pattern Recognition 42(5), 645 – 654, 2009.