A Diversity-Accuracy Measure for Homogenous Ensemble Selection

S. Taleb Zouggar[1]*, A. Adla[2]

[1] Department of Economics, University of Oran 2 (Algeria)
[2] Department of Computer Science, University of Oran 1 (Algeria)

Received 25 December 2017 | Accepted 30 May 2018 | Published 29 June 2018

Abstract

Several selection methods in the literature are essentially based on an evaluation function that determines whether a model M contributes positively to boost the performances of the whole ensemble. In this paper, we propose a method called DIversity and ACcuracy for Ensemble Selection (DIACES) using an evaluation function based on both diversity and accuracy. The method is applied on homogenous ensembles composed of C4.5 decision trees and based on a hill climbing strategy. This allows selecting ensembles with the best compromise between maximum diversity and minimum error rate. Comparative studies show that in most cases the proposed method generates reduced size ensembles with better performances than usual ensemble simplification methods.

Keywords

Machine Learning, Classification, Decision Trees, Ensemble Methods, Bagging, Ensemble Pruning, Hill Climbing.

DOI: 10.9781/ijimai.2018.06.005

I. Introduction

Decision trees are classification methods for generating mutually exclusive decision rules structured in trees that must be simple with maximum performances on the learning sample.

These methods have the advantage to generate intelligible rules but are less efficient than the competing methods because of a significant variance due to the trees instability [5] that increases their generalization error rate [20].

To alleviate this weakness and find a compromise between the complexity of a model and its generalization reliability, stopping criteria as well as post-pruning algorithms have been proposed. These techniques reduce the variance but deteriorate the performances (increase the value of the bias). Another solution is to improve an unstable learning algorithm by using it several times to construct a set of different models [6], [40], [15], [18]. The generated models are then aggregated in order to combine their predictions. In the case of learning by regression, the aggregation is based on individual models predictions, whereas in the case of learning by classification, it is done by a majority vote among the classes predicted by the different models.

The effectiveness of these methods is related to the construction of a collection of classifiers that are both sufficiently precise (performance) and sufficiently diverse (diversity). Diversity offers the opportunity to benefit from the complementarity of the individual models that make up the ensemble.

If each classifier is precise and does not commit errors on the same individuals as others, then the uncorrelated errors of the different classifiers are removed using the voting process.

The decision trees ensemble methods [43], [8], [9], [28], [25], [26], [27], [1] are homogeneous ensemble methods for which the basic model used is a tree induction algorithm.

The ensemble methods are noise-resistant; do not suffer from over-learning and give good performances [41], but have the disadvantage of relying on a large number of models, which can have as consequences increased learning time, storage resources [31] and the prediction time related to the interrogation of all models in the set.

The aim of ensemble pruning methods is to improve both the efficiency (prediction time) and the prediction performances [35] because a large number of models increase the computational complexity but guarantees a great diversity within the ensemble. This diversity is represented by models with good or bad predicting performances.

Models with poor performance negatively affect the overall performance of the ensemble. Eliminating them while maintaining a large diversity of the remaining elements in the ensemble, improves performance while reducing prediction time.

We propose a new method to simplify homogeneous ensembles composed of C4.5 decision trees [39]. This method is based on a DHCEP (Directed Hill Climbing Ensemble Pruning) strategy with a multi-objective function to evaluate the relevance of an ensemble of trees. The function, used in a Hill Climbing process in Forward Selection (FS), allows selection of ensembles with the best compromise between maximum diversity and minimum error rate. The motivation behind the joint use of the two criteria is that there is a correlation between the individual performance of classifiers and their diversity. The more accurate the classifiers, the less they disagree. The use of one of the two properties is not sufficient to find the best performing ensemble.

* Corresponding author.
E-mail address: souad.taleb@gmail.com

The proposed new multi-objective function is based on this compromise between individual performance of trees and their diversity. A comparison with UWA methods [36], Complementariness (Comp) [32], and Margin Distance Minimization (MARGIN) [33] shows, in most cases, that the proposed method allows generating ensembles that are both smaller in size and more efficient, than those of the methods cited above. This reduced number of trees allows a gain in memory space and computing time which can be very significant for large samples.

The paper is organized as follows: In Section 2 we present a state-of-the-art on ensembles selection methods with more details on UWA [36], Complementariness [32], and Margin Distance minimization [33].

These methods allow simplifying ensembles and will serve as a basis for comparison (The source code for these different methods is available athttp://mlkd.csd.auth.gr/ensemblepruning.html [36]). In Section 3, we present the DIACES method, detailing the proposed new function as well as the path strategy. Section 4 contains all the experiments and analysis of the obtained results, whereas in the last section we conclude and propose some insights on future work.

II. Bagging, Aggregation, and Hill Climbing

In this section, we highlight the basic elements used in this paper, namely, the bagging method used for the generation of the initial ensemble, aggregation by unweighted and weighted vote.

A. Diversification by Bagging

Bagging Bootstrap Aggregating is a resampling method introduced by Breiman in 1996 [6]. Given a learning sample Ω_L and a learning method which generates a predictor $\hat{h}(.,\Omega_L)$ using Ω_L. The principle of bagging is to draw several bootstrap samples ($\Omega_L^{\theta 1}$, …,$\Omega_L^{\theta q}$) and generate for each one a collection of predictors ($\hat{h}(.,\Omega_L^{\theta 1})$, …, $\hat{h}(.,\Omega_L^{\theta q})$) using the base learning method for finally aggregating them.

A bootstrap sample Ω_a^l is obtained by randomly drawing n observations in the starting sample Ω_L. Each observation has the probability of $1/n$ of being shot; $|\Omega_L|=n$, the random variable Θ_l represents the random drawing.

Initially, Bagging was introduced with a decision tree as basic rule. But the schema is general and can be apply to other basic rules. In Fig. 1 is presented the principle of Bagging.

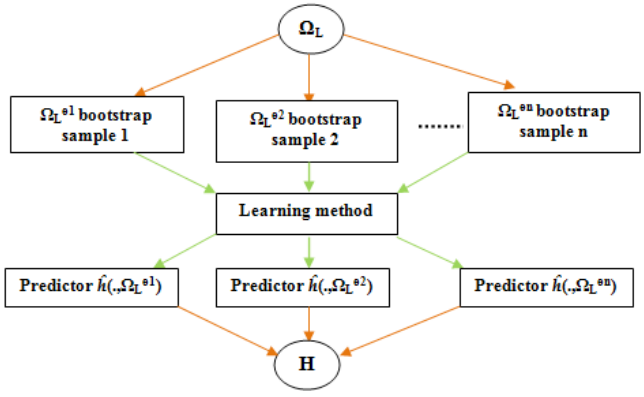

Fig. 1. Representative diagram of Bagging.

B. Aggregation (Unweighted Vote, Weighted Vote)

The unweighted and weighted voting are the most used methods for combining (aggregating) whether homogenous or heterogeneous models. In ensemble methods each model, for an instance, gives a class value, a probability, and the class with most votes, highest average probability is assigned to the instance by the ensemble.

In weighted vote, the classification models are associated with weights assigned relatively to their classification accuracy. Formally this can be written [36]:

Let x be an instance and $m_{i,\ i\ =\ 1..k}$ a set of models that output a probability distribution $m_i(x,c_j)$ for each class $c_{j\ ,\ j\ =\ 1\ .\ .\ .\ n}$. The output of the (weighted) voting method y(x) for instance x is given by the following mathematical expression:

$$y(x) = argmax_{c_j} \sum_{i=1}^{k} w_i m_i(x, c_j) \quad (1)$$

C. Hill Climbing

Hill climbing is an optimization technique belonging to the family of local search. The algorithm starts with any solution to a problem, then tries iteratively to find a better solution by changing one element of the solution. If the change produces a better solution (maximize or minimize the evaluation function used for the course), an incremental change is made to the new solution. The process is repeated until no improvements can be found (the function reached the maximum or the minimum).

Hill climbing attempts to maximize (or minimize) a target function f(X) where X is a vector of continuous and/or discrete values. Each iteration, hill climbing will adjust a single element in X and determine if the change improves the value of f(X). Any change improving the function f(X) is accepted, the process continues until no amelioration of the function can be found.

For ensemble selection, DHCEP (Directed Hill Climbing Ensemble Pruning) is used, in this case the vector X is composed of classifiers or predictors.

The course can be realized either in backward elimination or in forward selection, in the first case the whole ensemble is considered as a solution and then repeatedly elements not improving the evaluation function are eliminated one by one, in the second case we initialize with an element randomly and we add the elements that improve the evaluation function one by one. The elements to be added or removed are part of the neighborhood of the current solution. In Fig. 2, a hill climbing diagram for an ensemble composed of four models is presented.

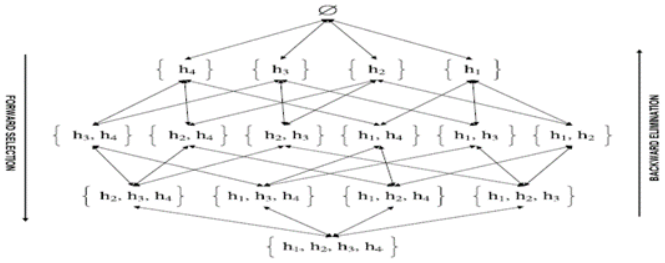

Fig. 2. Hill climbing search for selection in an ensemble composed of four classifiers [36].

III. Related Work

Several methods and heuristics have been proposed to reduce the size of a set of classifiers. They are categorized in four main classes:1) ordering-based methods [31], [47], [33], [34], 2) clusters based methods [16], [21], [29], [19], 3) optimization based methods [49], [50], and 4) the others [31], [17], [10], [32], [3].

Among these methods, a large number for ensemble pruning based on a hill climbing research process have recently been proposed [31], [17], [10], [32], [3]. The methods differ from each other by the adopted

research directions, the different evaluation measures or the evaluation ensembles. It will be noted that some methods use the learning sample for evaluation while others promote the use of a separate validation set. The latter depends mainly on the availability of the data.

A first type of approaches uses performance measures. Fan et al. [17] propose a profit-based evaluation function and propose dynamic scheduling to accelerate the prediction process. For the reduction of the ensemble size, the total benefit is used as selection criterion in conjunction with a greedy search algorithm with and without back fitting. The path begins with the model with the greatest benefit. A set of instances x is considered, each instance x can be positive or negative, B(x) denotes the benefit of predicting x as positive and the total benefit BT=$\sum_x B(x)$, the authors choose the sub ensembles which maximize the total benefit.

Empirical evaluations of several data sets have revealed that the profit-driven greedy approach with or without back fitting eliminates 90% of the size of an ensemble maintaining or sometimes exceeding the total benefit on the test sample of the original ensemble. The authors have also studied the possibility of combining diversity and total benefit, but the experimental results have shown that the total benefit is a good criterion by itself.

Caruana et al. [10] use several performance metrics and a hill climbing strategy for building ensembles of models from libraries of thousands of models. Model libraries are generated using different learning algorithms. The Forward Stepwise selection consists to add models that maximize its performance.

The second type of approaches uses diversity-based measures. Martínez-Muñoz et al. [32] use diversity with a hill climbing forward selection process. The diversity measure called Complementariness is defined as the complementarity of a model h_k with respect to the current sub set and a set of instances of the evaluation sample Eval = $\{(x_i, y_i)\}$ such that |Eval| = n, the measure COM is calculated as follows:

$$COM_{Eval}(h_k, Sub) = \sum_{i=1}^{n} I(y_i = h_k(x_i) and y_i \neq Sub(x_i)) \quad (2)$$

Where I (True) = 1, I (False) = 0, Sub(x_i) is the classification of the instance x_i by the sub-set Sub. The measurement principle is to add to Sub, the model which allows classifying correctly the examples misclassified by the subset.

Martínez-Muñoz et al. [33] propose the minimization of the marginal distance which allows calculating the diversity by associating to each classifier h_t a vector c_t whose dimension is equal to the number of individuals of the evaluation sample. An element $c_t(i)$ takes the value 1 if h_t properly classifies the individual i and -1 otherwise. An average vector C_{Sub}, associated with a subset Sub is calculated as $C_{Sub}= \frac{1}{|Sub|}\sum_{t=1}^{|Sub|} c_t$. The objective is to reduce the Euclidean distance d(o, C_{Sub}) where o is a predefined vector. The measure that represents the margin is written:

$$MAR_{Eval}(h_k, Sub) = d\,(o, \frac{1}{|Sub|+1}(c_k + C_{Sub})) \quad (3)$$

Partalas et al. [35], [36] propose the Uncertainty Weighted Accuracy measure (UWA) in (4) that considers four cases when adding a model to a sub ensemble and using justified weights to distinguish favor cases from others.

$$UWA_{Eval}(h_k, Sub) = \sum_{i=1}^{|Eval|} (\alpha * I(y_i = h_k(x_i) ET y_i \neq Sub(x_i)) - \beta * I(y_i \neq h_k(x_i) ET y_i = Sub(x_i)) + \beta * I(y_i = h_k(x_i) ET y_i = Sub(x_i)) - \alpha * I(y_i \neq h_k(x_i) ET y_i \neq Sub(x_i))) \quad (4)$$

Where the parameters α, β represent respectively the number of models in the sub-set *Sub* correctly classifying the instance (x_i,y_i) and the number of models incorrectly classifying the same instance.

More recent related work [30] theoretically deals with the effect of diversity on voting generalization performance using Probably Approximately Correct (PAC) learning. It is revealed that diversity is closely related to the space complexity hypothesis, and strengthening it can be achieved by applying regularization to ensemble methods. Based on this analysis, the authors apply an explicit regularization of the diversity for the selection of ensembles.

Dai [12] proposes an improvement of the ensemble selection method of the same authors. This method uses backtracking in depth, which is perfectly adapted to systematically seek solutions to combinatorial problems of great magnitude. This improvement concerns the response time of this method, which has been considerably improved in this study.

Zhou et al. [51] propose a new algorithm based on frequent item learning that links data and the simplified ensemble to a transactional database whose transactions are instances and items are classifiers. A Boolean classification matrix is used for each model of the pruned ensemble. Using this matrix, several candidate ensembles are obtained by iterative and incremental extraction of basic classifiers with the best performances.

Bhatnagar et al. [4] perform ensemble selection using a performance-based and diversity-based function that considers the individual performance of classifiers as well as the diversity between pairs of classifiers. A bottom-up search is performed to generate the sub ensembles by adding various pairs of classifiers with high performance.

To simplify a set of classifiers usually involves reducing the number of trees while maximizing performance. Qian et al. [38] adopt the Pareto diagram to solve this two-goal problem using an evolutionary optimization method of Pareto combined with a local search operator. The method is applied in the field of mobile human activity recognition.

Based on the approximate ensembles, Guo et al. [22] propose a new framework for ensemble selection. In this context, the relationship between attributes in an approximate space is considered a priori as well as their degree of maximum dependence. This effectively reduces the search space and increases the diversity of selected sub-ensembles. Finally, to choose the appropriate sub-ensemble, an evaluation function that balances diversity and precision is used. The proposed method allows repetitively changing the search space of the relevant sub-ensembles and selecting the next sub-ensemble from a new search space.

Cavalcanti et al. [11] combine in pairs different matrices of diversity using a genetic algorithm. The combined diversity matrix is then used to group similar (not very diverse) models; they must not belong to the same ensemble. To generate candidate ensembles, the combined diversity matrix is transformed into one or more graphs and then a graph coloring technique is applied.

Guo et al. [23] propose a new metric using the margin (instances) and the diversity (of classifiers) to explicitly evaluate the importance of individual classifiers. By adding the models to the ensemble in decreasing order of the metric, the user can choose the first T models to form a sub-ensemble.

Dai et al. [13] emphasize the utility of optimizing predictive performance together with diversity, which are two indispensable and inseparable parameters for ensemble selection. There have been three measures proposed to simplify ensembles using a greedy algorithm: 1) The first measure simultaneously considers the difference (diversity) between the current subset and the candidate classifier and the performance of each one; 2) The second allows evaluating the diversity within the ensemble and; 3) the last measure reinforces the concern about the accuracy of the resulting sub-ensemble. Experimental results confirm the interest of the three measures which is illustrated by the improvement of performances.

IV. The DIACES Proposed Measure

Our goal is to construct a distribution of the number of errors associated with each case and to calculate the diversity of this distribution. Our goal is to minimize diversity while maintaining good performance for each classifier.

The set of data Ω is divided into two sub samples Ω_L (generally 80% of Ω) for learning and pruning and Ω_T (generally 20% of Ω) for testing. A bagging ensemble BE of *t* C4.5 trees is constructed, BE = $\{T_1,\ldots,T_i,\ldots,T_t\}$, using Ω_L with $|\Omega_L| = n$. Each tree T_i is represented by a vector $(x_{1i}, x_{2i},\ldots, x_{ji},\ldots, x_{ni})^T$. We have the following notations:

- x_{ij}: Result of classification of the individual i by the tree j, $x_{ij} = 1$ if the individual *i* is misclassified by the tree T_j and $x_{ij} = 0$ otherwise,
- x_{i+}:The total number of errors committed for the individual *I*: $x_{i+} = \sum_{j=1}^{t} x_{ij}$
- X: The total number of errors committed by the set: $X = \sum_{i=1}^{n}\sum_{j=1}^{t} x_{ij}$
- (Θ_i, x_{i+}): The relative distribution of the error frequencies associated with the different cases: $\theta_i = \frac{x_{i+}}{X}, i = 1, n$
- x_{+j}: The number of errors committed by the classifier T_j over all the individuals: $x_{+j} = \sum_{i=1}^{n} x_{ij}$
- e_j: The error rate associated with the tree T_j: $e_j = \frac{x_{+j}}{n}$

The evaluation function to optimize noted S connects diversity Θ_i and the error rate e_j: $S = \sum_{i=1,n} \theta_i^2 + \sum_{j=1,t} e_j^2$.

The component C= $\sum_{i=1}^{n} \theta_i^2$ is a concentration index of error distribution which is derived from the quadratic entropy (or Gini index). The smallest value of C is 1/n, where all x_{i+} have the same value. This situation is best conditioned by the value of X. The minimization of C+E, using two metrics error rates and concentration index of errors distribution, allows having a good compromise between the diversity of the trees and their average performance.

C* represents the normalized coefficient =>C*= $\frac{nxC - x}{nk - x}$

E* represents the normalized coefficient =>E* = $\frac{kn^2E - x^2}{xkn - x}$

Finally the function to minimize S=C*+αE* (α is a parameter determined empirically by the user).

Algorithm 1 presents the proposed method DIACES in a pseudo code:

Algorithm1 DIACES;
Input
  BE = $\{T_1,\ldots,T_t\}$;
  Ω_L: selection set;
  Neighborhood(Ψ_j): Function that returns the subsets of models obtained from Ψ_j by adding a classifier (tree);
Output
  Sub ensemble Ψ_0 of BE;
Begin
Initialize(Ψ_0);
1. Calculate $S(\Psi_0,\Omega_L)$;
if $\exists\ \Psi_j$ such as $S(\Psi_j,\Omega_L) < S(\Psi_0, \Omega_L)$ where $\Psi_j \in$ Neighborhood (Ψ_0)Then Ψ_0=argmin$_{\Psi j}(S(\Psi_j,\Omega_L))$;
Goto 1;
End.

The algorithm complexity consists in calculating the hill climbing path method complexity which is $O(k^2)$ where k is the number of classifiers of the ensemble. The function S is computed from a matrix composed of n rows (number of individuals in the validation set), its complexity is $O(n)$, the calculation of the function is repeated k' times where k' is the number of ensembles traveled in the hill climbing scheme. The complexity of the proposed method is $O(n*k'*k^2)$.

A. Initialization and Path

The path strategy used by our method is a hill climbing strategy whose principle is simple [44]. It consists in reducing the number of ensembles generated in the case of an exhaustive path in which 2^k of ensembles are explored (k is the number of classifiers).

The hill climbing allows obtaining a sub-optimal solution by going through $\frac{k(k+1)}{2}$ subsets, considering a set of states and selecting the next state to be visited from the neighborhood of the current state. In this case, the states are the different ensembles of models and the neighborhood of a subset Sub of BE (set of all hypotheses) is composed of ensembles constructed by adding (forward selection) or deleting (backward elimination) a model of Sub. The method goes across the search space (all subsets of models) from one end to the other; one of the two ends is composed of the empty set and the other of the set of all the models. The complexity of hill climbing is $O(k^2)$.

In our case, we go across the set in forward selection and for the initialization we choose a tree that is not very good and not very bad. The fact of not choosing the most precise tree, as is the case for many methods that adopt a hill climbing method for selection in a set, is due to the fact that in some cases the most accurate tree can be perfect (does not make any error) on the evaluation sample, consequently a subset is produced with a single tree (the initialization tree) which can be very bad in generalization.

The proposed solution consists in choosing a tree with "average" performances on the evaluation sample. First the k trees of the initial set BE are ordered using performance on the evaluation sample in ascending or decreasing way, the ordered set B is then decomposed in 3 subsets BE1, BE2, and BE3, finally the initialization tree is randomly selected from the subset BE2.

According to experimental studies that we have carried out, the proposed multi-objective function directly affects the error rate in generalization (on the test sample Ω_t). In Fig. 3, the curve shows the correlation between the function and the error rate for the Ionosphere dataset.

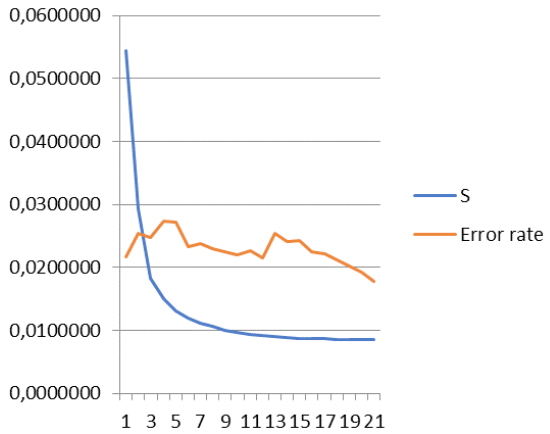

Fig. 3. Correlation between the function Sand the error rate on ΩT for Ionosphere dataset.

V. Experiments

Experiments consist in building homogeneous sets by sampling the starting sample and using the C4.5 decision tree generation algorithm as a basic rule.

A. Materials and Methods

The Weka platform [46] is used as a source for the C4.5 learning algorithm and validation. For this purpose, we consider 24 Benchmarks of the UCI Repository [2] which are described in Table I, the table is composed of five columns:

ID: Identifier of the data set, NI: Number of Instances, ND: Number of Descriptors, CM: Class Modalities, α: the value of the parameter α for each data set.

TABLE I. Description of the Data Sets used for the Comparison

	id	NI	ND	CM	α
Audiology	b1	226	69	24	181
Breast-Cancer	b2	286	9	2	229
Breast w	b3	699	9	2	559
CMC	b4	1473	9	3	1178
Diabetes	b5	768	8	2	614
Credit-a	b6	690	15	2	552
Credit-g	b7	1000	20	2	800
Heart-Statlog	b8	270	13	2	216
Anneal	b9	798	38	6	638
Balance-Scale	b10	625	4	3	500
Colic	b11	368	22	2	294
Haberman	b12	306	3	2	245
Titanic	b13	390	13	2	312
Primarytumor	b14	339	17	2	271
Sonar	b15	195	60	2	156
Soybean	b16	683	35	19	546
Vehicle	b17	946	18	4	757
Vote	b18	435	16	2	348
Vowel	b19	990	13	11	792
Autos	b20	205	25	6	164
Glass	b21	214	9	6	171
Hearth-h	b22	294	13	5	235
Ionosphere	b23	351	34	2	281
Lymph	b24	148	18	4	118

The initial sample decomposition is based on the results presented in [37]. It concludes that for the case of simplification of homogeneous set, pruning on the same learning sample allows obtaining better results than using separate samples, which is also an advantage in the case of unavailability of large amounts of data.

The initial sample is subdivided into two sub-samples, one of which is composed of 80% of the individuals and will be used for learning (model generation) and evaluation (pruning), while the remaining 20% will be used for testing.

The method we propose, DIACES, is compared to a set of ensemble pruning methods based on diversity: UWA, COM, and MARGIN, detailed the state of art. For all these methods, the unweighted majority vote is used for the combination of models and the performance calculation.

The methods use a forward selection strategy in a hill climbing scheme. The stop criterion for methods in literature is the performance on the evaluation sample which generates subsets of reduced sizes compared with the usual stop criteria defined as a fixed number of models [35]. In our case we use the same function for both the path and the stop.

B. Results and Analysis

A first criterion to compare the different pruning methods is the performance of the subsets obtained. For each method and each dataset, 10 different draws are made for which the averages of the success rates obtained for each draw are calculated.

For 24 benchmarks, DIACES shows better performances in 15 cases followed by the COM method with 6 victories, 3 victories for UWA and finally MARGIN with two victories. For the average success rate on all datasets, DIACES is ranked first with an average rate of 80.48%, exceeding the other methods with a rate of at least 0.9%.

Compared to UWA, DIACES improves performance by about 3% on the Audiology, Breast cancer, Vehicle, Primary Tumor, and Titanic datasets. Moreover, DIACES improves the performances obtained by the COM and MARGIN methods by at least 0.2%, 4% on Audiology data set.

To ensure whether the performance differences between DIACES and the other methods are signifi cant; we will use the statistical test of the sign. The test allows ensuring the effectiveness of the proposed method for any data collection. The test considers the method achieving the best success rate and compares it to DIACES. Based on the results presented in Table II, we compare DIACES to COM which gives an average success rate of 79.87% following the next steps:

1. Calculate the performance difference D between DIACES and COM methods.
2. Calculate the p-value, using the sign test:

Success number	7
p-value	0.04

The p-value < 0.05 (0.05 being the risk value) implies that the performance difference between DIACES and COM is significant, and consequently DIACES is better than all the other methods.

Depending on the average success rates values, a rank is assigned to each method for each benchmark. In this way the different methods are compared according to their average ranks which is an appropriate criterion of comparison [14], the ranks and average ranks of the methods are presented in Table III. AVR represents the average rank of the generated sub ensembles based on success rates.

TABLE II. Comparing Success Rates for Different Methods

	DIACES	UWA	COM	MARGIN
b1	**0.85432**	0.81773	0.80661	0.81108
b2	0.72618	**0.72977**	0.72631	0.72205
b3	**0.97459**	0.95032	0.95248	0.9496
b4	0.52106	0.52617	0.52515	**0.52651**
b5	**0.77519**	0.76469	0.76727	0.76524
b6	0.85329	0.86085	**0.86882**	0.86156
b7	**0.742**	0.73	**0.7365**	0.729
b8	0.81857	0.81848	**0.82034**	0.81478
b9	**0.99651**	0.9916	0.99048	0.99104
b10	0.8145	0.8144	**0.8264**	0.8208
b11	0.81598	0.84245	**0.83423**	0.83012
b12	**0.73913**	0.73118	0.7131	0.72131
b13	**0.75509**	0.72176	0.73075	0.73331
b14	**0.45732**	0.41787	0.41587	0.42685
b15	**0.79043**	0.78048	0.77803	0.77071
b16	**0.9537**	0.93676	0.94705	0.93897
b17	**0.78626**	0.74612	0.74198	0.74553
b18	0.95571	0.954	0.95171	**0.9586**
b19	0.89749	**0.92016**	0.91058	0.91866
b20	0.81362	**0.81703**	0.80483	0.80972
b21	**0.789**	0.7619	0.77284	0.77047
b22	**0.80951**	0.79652	0.80858	0.80687
b23	**0.9235**	0.90856	0.90142	0.91998
b24	**0.84**	0.81376	0.83789	0.82411
AVSR	0.8084	0.7980	0.7987	0.7986

TABLE III. Ranking Methods Based on Success Rate

	DIACES	UWA	COM	MARGIN
b1	1	2	4	3
b2	3	1	2	4
b3	1	3	2	4
b4	4	2	3	1
b5	1	4	2	3
b6	4	3	1	2
b7	1	3	2	4
b8	2	3	1	4
b9	1	2	4	3
b10	3	4	1	2
b11	4	1	2	3
b12	1	2	4	3
b13	1	4	3	2
b14	1	3	4	2
b15	1	2	3	4
b16	1	4	2	3
b17	1	2	4	3
b18	2	3	4	1
b19	4	1	3	2
b20	2	1	4	3
b21	1	4	2	3
b22	1	4	2	3
b23	1	3	4	2
b24	1	4	2	3
AVR	**1.81**	2.81	2.71	2.67

We note that the proposed method has the best average rank with 1.81, followed by MARGIN method with 2.67, COM comes in the third position with 2.67, and finally the UWA method with 2.81.

A second comparison criterion is the size of the ensemble obtained. For each benchmark and on 10 draws, the average sizes of the subsets obtained is calculated.

Table IV shows the average sizes (over 10 iterations) of the sub sets obtained for each method. The number of models selected is reduced for all methods compared to the original size of the ensemble.

TABLE IV. Sub Ensembles Sizes Generated Using the Different Methods

	DIACES	UWA	COM	MARGIN
b1	**10.2**	14.6	14	21.7
b2	**11.2**	11.4	11.3	15.5
b3	**9.8**	11.7	13.1	14.3
b4	**18**	48.7	38.2	42.1
b5	**14.2**	20.4	26.3	36.3
b6	**13.3**	16.1	18.7	18
b7	**16.9**	23.2	25.7	33.2
b8	**12.7**	14.7	17.9	18.7
b9	**2.6**	3.4	2.9	13.8
b10	**14.7**	26.9	22.4	21.8
b11	12.2	8.2	**4.7**	7
b12	12.5	13.5	15.6	24.1
b13	**15.2**	22.4	24	34.1
b14	**13**	53	37.2	33.6
b15	10.7	**7**	8.8	15
b16	**12.2**	**12.2**	19.3	23.5
b17	**13.4**	31.6	33.8	37.7
b18	10.8	**4.6**	9.4	11
b19	**12.4**	18.8	13.1	22.7
b20	**11.00**	17	14.5	20
b21	**13**	15.9	20.6	16.2
b22	**13.2**	15.7	17.2	19.7
b23	10.6	**6.7**	7.9	12.7
b24	10.9	**10.6**	15.3	21.5
AVSZ	**12.28**	17.84	17.99	22.26

The reductions for all methods compared to the initial ensemble of all models vary between 93% and 97%.

The new method allows obtaining subsets of reduced sizes compared to the other methods for 17 data sets, 66% of cases; UWA comes in second place with 6 victories, and finally the COM method which counts one victory.

The study of algorithmic complexity makes it possible to calculate the quantity of resources (time or space) necessary for its execution. The complexities in time of the proposed method are calculated on the 24 data sets (For each one of them it is a question of calculating the average of time and space on 10 iterations) on a machine treating 109 instructions/seconds (1 Gigahertz) and a memory of 3 Gigabytes. The Hill Climbing search method is fast because it does not cover all the possible case combinations, the maximum time is calculated for cmc and credit-g data sets for which the search lasts 5.12 ms, the minimum times are 0.15 ms for the anneal data set; For other data sets the times vary between 3.7 ms and 0.5 ms. The 24 data sets have a total time of 0.82 seconds.

VI. Conclusion and Future Works

The ensemble methods improve the performance of an unstable classifier but have the disadvantage of the loss of readability of the model provided; Composed of a large number of distinct trees and therefore more difficult to synthesize by humans.

This is why other methods have also been proposed to synthesize not the results but the structure of a tree in the form of a "consensus" from a set of classifiers of the same type [42] [45] nevertheless with a deterioration in the quality of prediction.

In this paper we presented a new evaluation function combining performance and diversity for selection in a homogeneous ensemble used in a process of climbing hill path. The method was evaluated on several benchmarks and compared to pruning homogeneous ensembles in literature.

The results show that the proposed method obtains ensembles with performances exceeding the ensembles obtained by the compared methods. In addition the interest of the function is that it can be used equally with homogeneous sets as well as heterogeneous sets (the basic rules are not the same).

We plan, in future work, to use this new pruning measure and apply it for selection in a random forest ensemble, knowing that a random forest ensemble improves the performance of a bagging [7].We also propose to study the possibility of using another path strategy for the selection. The third contribution consists in using the method in the field of predicting student performance in education institutions and comparing the results with those obtained in [24]. In their work [48] use various classification methods (Neural networks, Kppv) separately for the diagnosis of breast cancer. We propose to use a heterogeneous ensemble composed of several classification methods for the same application. The ensemble will then be simplified by using the proposed measure.

The last point consists in finding a value for the parameter α for which we have noticed during empirical research that an appropriate value, determined in a non-empirical way, could significantly improve the results as observed in our experiments, for this we will use the Pareto principle.

References

[1] Y. Amit, D. Geman, "Shape quantization and recognition with randomized trees," Neural Computation, (9), 1997, pp. 1545–1588.

[2] D.N.A. Asuncion. (2007). UCI machine learning repository. Available at http://www.ics.uci.edu/~mlearn/MLRepository.

[3] R.E. Banfield, L.O. Hal, K.W. Bowyer, W.P. Kegelmeyer, "Ensemble diversity measures and their application to thinning," Information Fusion,6 (1), 2005, pp. 49–62.

[4] V. Bhatnagar, M. Bhardwaj,S. Sharma, S. Haroon, "Accuracy-diversity based pruning of classifier ensembles,"Progress in Artificial Intelligence, 2(2-3), 2014, pp. 97–111.

[5] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman and Hall, New York, 1984.

[6] L. Breiman, "Bagging Predictors," Machine Learning, 26, No. 2, 1996, 123-140.

[7] L. Breiman, "Random Forests," Machine Learning, 45(1), 2001, pp. 5-32.

[8] B.G. Buchanan, E.H.Shortliffe, "Rule Based Expert Systems," Addison-Wesley, Reading, Massachusetts, 1984, pp. 288-291.

[9] P.L. Bogier, "Shafer-Dempster reasoning with applications to multisensor target identification," 1EEE Trans. Sys. Man. Cyb. SMC-17, 1984, pp 968-977.

[10] R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes, "Ensemble selection from libraries of models," In Proceedings of the 21st international conference on machine learning, 2004, pp. 18.

[11] G.D.C. Cavalcanti, L.S. Oliveira, T.J.M. Moura, G.V. Carvalho, "Combining diversity measures for ensemble pruning,"In Pattern Recognition Letters, Volume 74, 2016, pp. 38-45.

[12] Q. Dai, "An efficient ensemble pruning algorithm using One-Path and Two-Trips searching approach," In Knowledge-Based Systems, Volume 51, 2013, pp. 85-92.

[13] Q. Dai, R. Ye, Z. Liu., "Considering diversity and accuracy simultaneously for ensemble pruning," In Applied Soft Computing, Volume 58, 2017, pp. 75-91.

[14] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, 7, 2006, pp 1–30.

[15] T. Dietterich, "Ensemble Methods in Machine Learning," Lecture Notes in Computer Science, 1857, 2000, pp.1-15.

[16] P. Domingos, "Knowledge acquisition from examples via multiple models," In: Proc. 14th International Conference on Machine Learning, Morgan Kaufmann, 1997, pp. 98–106.

[17] W. Fan, F. Chu, H. Wang, P.S. Yu, "Pruning and dynamic scheduling of cost- sensitive ensembles," In Eighteenth national conference on artificial intelligence, American association for artificial intelligence, 2002, pp. 146–151.

[18] Y. Freund, R.E. Schapire, "A Short Introduction to Boosting," J. Japanese Soc, Artificial Intelligence Res., 14(5), 1999, pp. 771-780 (1999).

[19] Q. Fu, S.X. Hu, S.Y. Zhao, "Clusterin-based selective neural network ensemble," Journal of Zhejiang University SCIENCE6A(5), 2005, pp. 387-392.

[20] P. Geurts, "Contributions to decision tree induction: bias/variance tradeoff and time series classification," Phd. Thesis, Department of Electronical Engineering and Computer Science, Liège Univ of Liège, Belgium, 2002.

[21] G. Giacinto, F. Roli, G. Fumera, "Design of effective multiple classifier systems by clustering of classifiers," In: 15th International Conference on Pattern Recognition, ICPR 2000, 2000, pp. 160-163.

[22] Y. Guo, L. Jiao, S. Wang., F. Liu, K. Rong, T. Xiong, "A novel dynamic rough subspace based selective ensemble," In Pattern Recognition, Volume 48, Issue 5,2015, pp. 1638-1652.

[23] H. Guo, H. Liu, R. Li, C. Wu, Y. Guo, M. Xu, "Margin and diversity based ordering ensemble pruning," In Neurocomputing, 2017; ISSN 0925-2312.

[24] A.K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," International Journal of Interactive Multimedia and Artificial Intelligence, vol. In Press, issue In Press, no. In Press, pp. 1-6, 02/2018.

[25] T.K. Ho, J.J. Hull, S.N. Srihari, "Decision Combination in Multiple Classifier Systems," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 1, 1994, pp. 66-75.

[26] T.K. Ho, "Random Decision Forests," Proc. Third Int'l Conf, Document Analysis and Recognition, 1995, pp. 278-282.

[27] T.K. Ho, "The random subspace method for constructing decision forests," IEEE Trans, on Pattern Analysis and Machine Intelligence, 20(8), 1998, pp. 832-844.

[28] S. Kwok, C. Carter, "Multiple decision trees," Uncertainty in Artificial Intelligence 4, ed. Shachter, R., Levitt, T., Kanal, L., and Lemmer, J., North-Holland, 1990, pp. 327-335.

[29] A. Lazarevic, Z. Obradovic, "The effective pruning of neural network classifiers," In: 2001 IEEE/INNS International Conference on Neural Networks, IJCNN 2001, pp. 796-801.

[30] N. Li, Y. Yu, H. Zhou, "Diversity Regularized Ensemble Pruning," In Proceedings of the 23rdEuropean Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I (ECMLPKDD'12), Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 2012, pp. 330-345.

[31] D.D. Margineantu, T.G.Dietterich, "Pruning adaptive boosting," In: Proc Of the 14th International Conference on Machine Learning, 1997, pp. 211-218.

[32] G. Martínez-Muñoz, A. Suarez, "Aggregation ordering in bagging," In International Conference on Artificial Intelligence and Applications (IASTED), 2004, pp. 258–263.

[33] G. Martínez-Muñoz, A. Suarez, "Pruning in ordered bagging ensembles," In 23rd international conference in machine learning (ICML-2006), 2006, pp. 609–616.

[34] G. Martínez-Muñoz, A. Suarez, "Using boosting to prune bagging ensembles," Pattern Recognition Letters 28 (1), 2007, pp. 156–165.

[35] I. Partalas, G. Tsoumakas, I. Vlahavas, " Focused ensemble selection: A diversity- based method for greedy ensemble selection," In: M. Ghallab, C.D. Spyropoulos, N. Fakotakis, N.M. Avouris (eds.) ECAI 2008 - 18th European Conference on Artificial Intelligence, Patras, Greece, Proceedings, Frontiers in Artificial Intelligence and Applications, vol. 178, 2008, pp. 117-121.

[36] I. Partalas, G. Tsoumakas, I. Vlahavas, "An ensemble uncertainty aware measure for directed hill climbing ensemble pruning," Machine Learning, 81, 2010, pp. 257–282.

[37] I. Partalas, G. Tsoumakas, I. Vlahavas, "A Study on Greedy Algorithms for Ensemble Pruning," Technical Report TR-LPIS-360-12, LPIS, Dept. of Informatics, Aristotle University of Thessaloniki, Greece, 2012.

[38] C. Qian, Y. Yu, Z.H. Zhou, "Pareto ensemble pruning," In: AAAI Conference on Artificial Intelligence, 2015.

[39] J.R. Quinlan, "C4.5: Programs For Machine Learning," Morgan Kaufmann Publishers, San Mateo, California, 1993.

[40] R.E. Schapire, Y. Freund, P. Bartlett, "Lee W.S., Boosting the margin: a new explanation for the effectiveness of voting methods," In Douglas H. Fisher, editor, Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97), 1997, pp. 322–330, Morgan Kaufmann.

[41] Settouti, N., M. E. A. Bechar, and M. A. Chikh, "Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classification Task, " International Journal of Interactive Multimedia and Artificial Intelligence, vol. 4, issue Special Issue on Artificial Intelligence Underpinning, no. 1, pp. 46-51, 09/2016.

[42] W.D. Shannon, D. Banks, "Combining classification trees using MLE," Statist. Med., 18(6), 1999, pp. 727-740.41

[43] S. Shlien, "Non parametric classification using matched binary decision trees," Pattern Recognition Letters, 13(2), 1992, pp. 83-88.

[44] G. Tsoumakas, I. Partalas, I. Vlahavas, "An Ensemble Pruning Primer," Applications of Supervised and Unsupervised Ensemble Methods (Eds.) Okun and Valentino, 2009, pp. 1-13, Springer-Verlag.

[45] J. T. L. Wang, K. Zhang, "Finding similar consensus between trees: an algorithm and a distance hierarchy," Pattern Recognition, 2001, 34:127.137.

[46] I. H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, 2005.

[47] Y. Yang, K. Korb, K. Ting, "Webb G., Ensemble selection for superparent-one-dependence estimators," In: AI 2005: Advances in Artificial Intelligence, 2005, pp. 102-112.

[48] Youh, H., and G. Rumbe, "Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 1, issue A Direct Path to Intelligent Tools, no. 3, pp. 5-12, 12/2010.

[49] H. Zhou, J. Wu, W. Tang, "Ensembling neural networks: Many could better than all," Artificial intelligence 137 (1-2), 2002, pp. 239-263.

[50] H. Zhou, W. Tang, "Selective ensemble of decision trees," In: 9thInternational Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 2003, pp. 476-483.

[51] H. Zhou, X. Zhao , X. Wang, "An effective ensemble pruning algorithm based on frequent patterns," In Knowledge-Based Systems, Volume 56, 2014, pp. 79-85.

Souad Taleb Zouggar

Souad Taleb Zouggar is Assistant Professor in Computer Science at University of Oran 2, Algeria. She received a PhD in Computer Science from University of Oran in 2014. She has published papers on Machine Learning, Data mining. Her research interests focus on Knowledge Discovery in Databases, Machine Learning, Optimization, and Decision Support systems.

Abdelkader Adla

Abdelkader Adla is full Professor in Computer Science at University of Oran 1, Algeria. He received a PhD in Computer Science, Artificial Intelligence from Paul Sabatier University Toulouse III, France. He received also a State Doctorate in Computer-Aided Design and Simulation from University of Oran in 2007. He has published papers on collaborative decision making, Decision Support Systems (DSS), distributed group DSS and multi-agents DSS. His research interests focus on Group DSS, Group work facilitation, Cooperative and collaborative systems, Knowledge and Multi-agent decision support systems.