Theoretical and Applied Informatics
Vol. 29 (2017), no. 1-2, pp. 25–39
DOI: 10.20904/291-2025

Using Diversity for Classifier Ensemble Pruning: An Empirical Investigation

Muhammad Atta Othman Ahmed[1*] Luca Didaci[1†] Bahram Lavi[1‡]
Giorgio Fumera[1§]

[1]Department of Electrical and Electronic Engineering,
University of Cagliari, Piazza D'armi, 09123 Cagliari, Italy

Abstract The concept of 'diversity' has been one of the main open issues in the field of multiple classifier systems. In this paper we address a facet of diversity related to its effectiveness for ensemble construction, namely, explicitly using diversity measures for ensemble construction techniques based on the kind of overproduce and choose strategy known as ensemble pruning. Such a strategy consists of selecting the (hopefully) more accurate subset of classifiers out of an original, larger ensemble. Whereas several existing pruning methods use some combination of individual classifiers' accuracy and diversity, it is still unclear whether such an evaluation function is better than the bare estimate of ensemble accuracy. We empirically investigate this issue by comparing two evaluation functions in the context of ensemble pruning: the estimate of ensemble accuracy, and its linear combination with several well-known diversity measures. This can also be viewed as using diversity as a regularizer, as suggested by some authors. To this aim we use a pruning method based on forward selection, since it allows a direct comparison between different evaluation functions. Experiments on thirty-seven benchmark data sets, four diversity measures and three base classifiers provide evidence that using diversity measures for ensemble pruning can be advantageous over using only ensemble accuracy, and that diversity measures can act as regularizers in this context.

Keywords Multiple classifier systems; Ensemble pruning; Diversity measures

Received 08 Mar 2017 **Revised** 17 Oct 2017 **Accepted** 01 Feb 2017

*E-mail: muhammad.ahmed@diee.unica.it
†E-mail: didaci@diee.unica.it
‡E-mail: lavi.bahram@diee.unica.it
§E-mail: fumera@diee.unica.it

1 Introduction

During twenty years of research in the classifier ensemble field, understanding the notion of *diversity* has been one of the main goals [1, 2]. A general agreement exists on the qualitative definition of diversity and on its role in classifier ensembles; basically, to obtain an effective (accurate) ensemble, its members should be as accurate *and* diverse as possible, where 'diverse' means that they should not make coincident errors [1, 2]. Individual accuracy and diversity are well-known to be contrasting goals, which means that a trade-off between them has to be achieved. On the other hand, formally defining and measuring diversity, as well as explicitly using it for ensemble construction, turned out to be not straightforward.

A number of diversity measures have been proposed over the years [1, 2, 3]. Most measures have been derived intuitively, as attempts to formally characterize the pattern of error of individual classifiers (e.g., the Double-Fault and Disagreement measures [2]). In particular, it has been clearly pointed out that diversity measures alone can not be monotonically related to ensemble accuracy, since the latter depends on a trade-off between diversity and individual classifiers' performance [2, 4]. For instance, searching for a diversity measure strongly related to ensemble performance runs the risk of 'replacing a simple calculation of the ensemble error by a clumsy proxy which we call diversity' [2] (ch. 8). A few other measures have been inspired by *exact* error decompositions derived in the regression field, despite the lack of a direct analogy to classification problems [5]. The Kohavi-Wolpert Variance [3] (and our attempt in [6]) was inspired by the bias-variance-covariance error decomposition of [7]. The measure derived in [8] (which we extended in [6]) was inspired by the *ambiguity* decomposition of [9], and provided useful insights, leading to the concept of 'good' and 'bad' patterns of diversity. Such measures were motivated by the goal of obtaining exact, additive decompositions of the ensemble error into terms accounting for individual classifiers' performance, and terms hopefully interpretable as diversity. Several authors also analyzed, empirically or analytically, the connection between ensemble performance on one side, and the pattern of individual classifiers' performance and existing diversity measures on the other side (e.g., [4, 10]). Such a relationship turned out to be far from clear-cut, and no 'right' diversity measure has emerged so far.

Beside theoretical investigations on defining diversity and using this concept to explain ensemble performance, a considerable research effort has been spent toward the practical goal of *explicitly* using diversity measures for ensemble construction. Among existing methods, almost all follow the *overproduce and choose* approach. It consists of first generating a large ensemble (e.g., using Bagging) and then selecting the most accurate subset of classifiers. The overproduce and choose approach is also known as ensemble *pruning*, *selection* or *thinning*. It is supported by theoretical and empirical evidence showing that a (suitable) subset of the available classifiers could outperform the original ensemble [11, 12, 13].

Since ensemble pruning has exponential complexity in the size of the original ensemble, several heuristics have been proposed. In this context, diversity measures have been used in the objective function of pruning methods, to attain a trade-off between individual classifiers' performance and diversity. The effectiveness of using diversity measures to this aim has however been questioned by several authors, based also on empirical evidence [3, 4, 13], and [2] (ch. 8.3). In particular, its actual advantage over directly evaluating ensemble performance (estimated, e.g., from validation

data) is not clear yet. It is also well known that popular and effective ensemble construction techniques like Bagging and Boosting do not use any explicit diversity measure. Nevertheless, despite the questionable effectiveness of heuristic pruning approaches, a theoretically grounded analysis in [14] related to ensembles of binary classifiers combined by majority voting has shown that (a suitable measure of) diversity can have a regularization effect in ensemble pruning.

Based on the above premises, the aim of this work is to compare the effectiveness of explicitly using existing diversity measures in ensemble pruning, against the direct estimation of ensemble performance. This is a follow-up of our preliminary work [15]. In particular, inspired by [14], we evaluate whether several well-known diversity measures can have a regularization effect on the (estimate of) ensemble accuracy. To this aim we consider a pruning method based on the forward selection (FS) algorithm, since it allows a direct comparison between evaluation functions. We then compare the estimated ensemble accuracy against its linear combination with a given diversity measure, using the latter as a regularizer. We carry out experiments on 37 benchmark data sets. We use the popular Bagging as the ensemble construction technique and majority voting as the fusion rule, and evaluate a subset of the ten well-known diversity measures analyzed in [3]. Our results show that using diversity measures for ensemble pruning can be advantageous over using only ensemble accuracy, and that diversity measures can act as regularizers in this context.

2 Previous Work on Using Diversity for Ensemble Design

As pointed out in Sec. 1, diversity measures have been explicitly used so far for ensemble construction only in pruning methods. The only exception is [16], where a diversity measure was used in an ensemble *learning* algorithm.

In [17] ensemble pruning methods have been categorized as follows:

- **Ranking-based**: individual classifiers are first ranked according to some criterion, and then the top-L ones are selected as the final ensemble.
- **Clustering-based**: individual classifiers are first clustered based on the similarity of their predictions; each cluster is then pruned to remove redundant classifiers, and the remaining ones in each cluster are finally combined.
- **Optimization-based**: methods search for a subset of the original ensemble that optimizes a given objective function, which can include a diversity measure. To avoid exhaustive search, three main heuristic search strategies have been proposed: hill climbing, genetic algorithms, and semi-definite programming.

In particular, several optimization-based pruning methods use the forward or backward search (FS/BS) strategy [18, 19, 20, 21, 22, 23, 24].

Given an initial ensemble, FS picks the best individual classifier and iteratively selects among the remaining classifiers the one that maximizes a given objective function. It stops either when a predefined ensemble size is reached, or when all the classifiers from the original ensemble have been selected; in the latter case, FS returns the best ensemble among the ones obtained at each

iteration. The BS algorithm works similarly, iteratively removing from E one classifier at a time. More refined versions of FS/BS have also been proposed, which include a back-fitting step [19].

In the context of optimization-based pruning, three kinds of objective functions have been proposed so far:

- The ensemble accuracy [19, 21], combined with a diversity measure in [14].
- A given diversity measure (disregarding the performance of individual classifiers and of the ensemble) [19, 20, 23].
- Ad hoc measures specifically devised for ensemble pruning, which combine into a single scalar the individual classifiers' performance and the *complementarity* (diversity) between their errors [18, 22, 23, 24].

A different and theoretically grounded view on the role of diversity in ensemble pruning was proposed in [14], in the context of ensembles of binary classifiers combined by majority voting: using a suitable diversity measure it was shown that promoting diversity can be seen as a regularization technique. A pruning method was also proposed based on these results, which exploits a strategy similar to FS: it starts with the most accurate classifier from the original ensemble, then iteratively sorts the remaining classifiers based on their diversity (evaluated using the proposed measure) with the current sub-ensemble, and among the most diverse ones it selects the classifier which leads to the next most accurate sub-ensemble.

Cavalcanti et al. [25] tackle the problem of diversity measures for ensemble pruning using genetic algorithm. Also, in [26] another method for ensemble pruning using margin and diversity based measure is proposed by Guo et al.

It is also worth mentioning two ensemble construction techniques [27, 28] which are not pruning techniques but are related to the pruning criteria considered in this work. They consist of building individual classifiers from different subsets of the available features, analogously to the well known Random Subspace Method [29]. The difference with respect to RSM is that they use a feature selection criterion analogous to the optimization-based pruning criterion mentioned above (including FS in [28]), and evaluate the individual classifiers on the basis of a trade-off between individual classifiers' accuracy and diversity. In particular, in [28] a linear combination of these two quantities was used as the objective function, and five different measures of diversity were considered.

In our previous work [15] we carried out a preliminary comparison between using the ensemble accuracy as the evaluation measure and using existing, ad hoc measures proposes for pruning methods, that combine the individual (not the ensemble's) classifiers' performance and the complementarity between their errors. In this work we carry out a direct comparison of ensemble accuracy against its combination with well-known diversity measures that do not include individual classifiers' performance, and are not specifically devised for ensemble pruning.

3 Aim of this work

As mentioned in Sec. 1, many existing ensemble pruning methods use heuristic evaluation functions that combine the performance of individual classifiers and some measure of their diversity.

It is then interesting to understand whether and under what conditions such evaluation functions are more effective (in terms of the performance of the resulting ensemble) than directly evaluating the performance of the considered ensembles (estimated, e.g., from validation data) during the pruning procedure. Quite surprisingly, so far such a comparison has been carried out by only a few authors [14, 19, 22, 23, 24], and only with a limited scope. In particular, it was often limited to the proposed evaluation measure, and using different and incomparable experimental set-up (i.e., different data sets, base classifiers, ensemble construction methods, etc.). We also point out that, among these works, only in [14, 24] the use of the proposed evaluation functions provided a statistically significant improvement over a direct estimation of ensemble performance.

To sum up, so far no clear evidence has been provided about the effectiveness of using diversity measures for ensemble pruning. A notable exception is the work of [14], where an original view of the role of diversity as a regularizer in ensemble design was proposed and theoretically investigated, in the case of binary classifiers combined by majority voting, and with a specific diversity measure. Their theoretical results showed that promoting diversity during ensemble design can actually have a regularization effect. Based on these results, a specific ensemble pruning method was then proposed in [14].

Based on the above premises, and inspired by [14], the aim of this work is to investigate whether also existing diversity measures can have a regularization effect in ensemble pruning, with respect to the (estimate of) ensemble accuracy. More precisely, we consider two evaluation functions: ensemble accuracy A alone, and its linear combination with a given diversity measure D, given by $A + \lambda D$ (with $\lambda > 0$), which is the usual form of regularization terms.

To carry out a direct comparison between such evaluation functions we consider a pruning method based on the forward selection (FS) algorithm. We first build an ensemble of N classifiers using a given ensemble construction technique, then we use FS to obtain a subset of $L < N$ classifiers, for a given L. We consider the basic version of FS: it starts with the best (estimated) individual classifier of the original ensemble, then it iteratively selects from the remaining classifiers the one that provides the best evaluation function (either A or $A+\lambda D$) on the new candidate ensemble. The pseudo code is shown in Alg. 1.

Require: an ensemble $\mathbf{E}$ of N classifiers; a desired ensemble size $L < N$; a validation set $\mathbf{V}$; an objective function f_{obj} (to be computed on $\mathbf{V}$).
$C \leftarrow$ The most accurate individual classifier from $\mathbf{E}$
$\mathbf{S} \leftarrow \{C\}$
For counter $= 2, \ldots, L$ **do**
  $C_t \leftarrow \text{argmax}_{C \in \mathbf{E}/\mathbf{S}} f_{obj}(\mathbf{S} \cup \{C\})$
  $\mathbf{S} \leftarrow (\mathbf{S} \cup \{C_t\})$
End for
return S

Algorithm 1 Forward Selection algorithm for ensemble pruning.

	ρ	Dis	DF	KW	κ	E	θ	GD	CFD
Q	0.9945	−0.9840	0.5578	−0.9840	−0.9840	0.9943	0.9352	−0.8210	−0.8396
ρ		−0.9710	0.5491	−0.9710	−0.9710	0.9998	0.9546	−0.8256	−0.8463
Dis			−0.5648	1.0000	1.0000	−0.9713	−0.8619	0.7978	0.8258
DF				−0.5648	−0.5648	0.5490	0.4922	−0.8879	−0.8951
KW					1.0000	−0.9713	−0.8619	0.7978	0.8258
κ						−0.9713	−0.8619	0.7978	0.8258
E							0.9548	−0.8257	−0.8462
θ								−0.7970	−0.8002
GD									0.9927

Table 1 Correlation coefficient between each pair of the diversity measures considered in [3].

4 Diversity Measures

In this section we describe the diversity measures used in this work. We started from the ten measures analyzed in [3]: Q-statistic (Q), Correlation coefficient (ρ), Disagreement (Dis), Double-fault (DF), Kohavi-Wolpert variance (KW), Interrater agreement (κ), Entropy (E), Difficulty (θ), Generalised diversity (GD) and Coincident failure diversity (CFD). They include pairwise and non-pairwise measures (i.e., measures that are defined on two classifiers, or on a classifier ensemble of any size), respectively Q, ρ, Dis, DF, and E, KW, κ, θ, GD, CFD; and measures that require the true label of the samples on which they are computed (all except E and Dis), and measures that do not (E and Dis). For pairwise measures, the diversity of an ensemble of more than two classifiers is computed as their average value over all distinct pairs of ensemble members.

In [3] it was observed that some of the considered measures are strongly correlated (positively or negatively). We therefore decided to select only a subset of the least correlated measures. To this aim we estimated the correlation between all pairs of such measures by simulating the outputs of two binary classifiers on 1,500 input instances. For both classifiers we randomly and independently generated 1,500 binary values (0 and 1) from a uniform distribution, which represent either incorrect (0) and correct (1) decisions, in the case of diversity measures defined in terms of classification outcomes (correct/incorrect, which requires the true class label to be known), or the predicted labels of a two-class problem (which does not require the true class labels), in the case of diversity measures defined in terms of classifier decisions (namely, Entropy and Disagreement). We repeated the above procedure for twenty times, and evaluated the correlation coefficient between every distinct pair of diversity measures.

These values are reported in Tab. 1. It is worth noting that our results qualitatively agree with the ones reported in [3], although they have been obtained using different data.

Based on these results, we first selected the two least correlated measures, i.e., θ and DF (their correlation is 0.4492, see Tab. 1). All the other measures exhibit a quite high correlation with either θ or DF. Among them, we selected two further measures exhibiting the lowest maximum correlation with θ and DF, which turn out to be Dis and GD. Note that the four selected measures include pairwise and non-pairwise measures, as well as measures defined in terms of classification outcomes and in terms of classifier decisions. We report their definition for

completeness (see [3] for the definition of the other measures).

Considering two classifiers C_1 and C_2 and assuming m the number of instances on which these measures are computed, a the number of instances correctly classified by both C_1 and C_2, b the number of instances correctly classified only by C_1, c the number of instances correctly classified only by C_2, d the number of instances incorrectly classified by both C_1 and C_2, and p_i the accuracy of C_i $(i = 1, 2)$ estimated on the same set of instances.
DF is a pairwise measure proposed in [30]:

$$DF = \frac{d}{m} \ . \tag{1}$$

GD is a non-pairwise measure proposed in [31]:

$$GD = 1 - \frac{p_2}{p_1} \ . \tag{2}$$

Dis is a pairwise measure proposed in [32]:

$$Dis = \frac{b + c}{m} \ . \tag{3}$$

Finally, Difficulty (θ) is a non-pairwise measure proposed in [33], which is defined as the variance of the pairwise Dis measure computed for all distinct pairs of classifiers:

$$\theta = Var(Dis) \ . \tag{4}$$

5 Experimental Setting

As explained in Sec. 3, the aim of our experiments is to compare two ensemble evaluation functions for ensemble pruning, using the basic FS pruning strategy described in Alg. 1: the ensemble performance, evaluated as the classification accuracy A estimated from validation data, and its linear combination with a given diversity measure D evaluated on the same validation set, $A+\lambda D$, with $\lambda > 0$.

To this aim we create an initial ensemble E composed of $N = 100$ classifiers, and prune it to an ensemble of L classifiers, with $L = 5, 15, 25, 35$, using the FS algorithm. We used Bagging to obtain E, as it is a well-known ensemble creation technique, and has already been used to this aim for ensemble pruning, e.g. [11, 34]. We used majority voting as the combining rule, since it is the standard choice for Bagging [35].

In our experiments we used three different base classifiers: Multi-Layer Perceptron Neural Networks (NN), Decision Trees (DT) and K-Nearest Neighbors (K-NN). We used their standard Matlab implementation (Neural Networks and Statistics and Machine Learning Toolboxes). In particular, for NNs we used the `patternnet` function with a learning rate $\eta = 0.05$, gradient descent with momentum as the learning algorithm, and a maximum of 1000 epochs as a stop criterion. For DTs we used the Gini impurity criterion, the χ^2 stopping criterion, and the default threshold equal to 1 for the pre-pruning stopping criterion. For K-NN we used $K = 1$.

In the evaluation function $A + \lambda D$ we used several values of λ: $0.2, 0.5$, and 0.7. We also considered the four diversity measures chosen in Sec. 4: DF, θ, Dis and GD.

We carried out our experiments on 37 benchmark data sets from the UCI Machine Learning Repository Database,[1] containing only numerical attributes and no missing values (see Tab. 2). They represent a remarkable range of classification problems: the number of patterns ranges from 160 to 10992, the number of classes from 2 to 10, and feature set size from 2 to 85. We randomly subdivided each data set, using stratified sampling, into a training set, a validation set and a test set. The size of the training set is defined as explained in Sec. 5.1. The size of the validation set was chosen as 1/3 of the training set, and the remaining instances were used as the testing set. We repeated this procedure for 20 runs, and evaluated the resulting average accuracy on testing samples.

5.1 Choice of the training set size

For each data set we chose the training set size that maximizes the (estimated) difference between the highest and lowest accuracy attained by different ensembles of a given size L. The rationale is that, if all ensembles of L classifiers obtained from the initial ensemble E exhibit a similar accuracy, it becomes difficult to evaluate the difference (if any) between different pruning methods (in our case, different evaluation functions used in the same pruning method). Fig. (1) illustrates the idea.

To this aim we carried out preliminary experiments, considering training sets sizes ranging from 1% to 70% of the whole data set. For NNs, we also considered different numbers of hidden units, between 3 and 20. Since considering different ensemble sizes L is computationally costly, and obviously considering all possible subsets of size L of a given ensemble is infeasible, we only considered ensembles of size $L = \frac{N}{2} = 50$, and estimated the performance of the best and worst such ensembles with the ones of ensembles made up of the L best and by the L worst individual classifiers.

The resulting training set sizes used in the rest of our experiments are shown in Tab. 3. For NNs the number of hidden units is also shown.

5.2 Statistical test

To compare the two considered ensemble pruning evaluation functions we carried out a test of statistical significance between the corresponding average test set accuracy over the different runs of our experiments. To this aim we chose the Wilcoxon signed-rank test, as it is recommended in [36] for comparing two algorithms over multiple data sets, which is the setting considered in our experiments. This is a non-parametric statistical hypothesis test that can be used to determine whether two dependent samples were drawn from populations having the same distribution. This test is used to evaluate the statistical significance of the obtained results, i.e., whether it is possible to reject the null hypothesis that the observed values – in our case, the accuracies obtained by different ensembles – are different only by chance. We used a p-value of 0.05.

[1] `http://archive.ics.uci.edu/ml/index.php`

Dataset	**Classes**	**Instances**	**Features**
Bank Note	2	1372	4
Banana	2	5300	2
Blood Transfusion	2	748	4
Cardiotocography	3	2126	22
Pop Failures	2	540	20
SatLogLandSetSat	6	6435	36
SataLogImageSeg	7	2310	19
Spam Base	2	4601	57
Thyroid	3	7200	21
Wine Quality	7	4898	11
Australian	2	690	14
Balance Scale	3	625	4
Bands	2	365	19
Breast Cancer	2	699	9
Bupa	2	345	6
Checker Board	2	1000	2
Cleveland	5	297	13
Coil2000	2	1286	85
Contours	3	2000	2
Contraceptive	3	1473	9
Dermatology	6	358	34
Hayes Roth	3	160	4
ILPD	2	583	9
Laryngeal 2	2	692	16
Marketing	9	6876	13
Monk 2	2	432	6
Page Plocks	5	5473	10
Pen based	10	10992	16
Phoneme	2	3186	5
Pima	2	768	8
Ring	2	7400	20
Saheart	2	462	4
Segment	7	2310	19
Spectfheart	2	267	44
Vehicle	4	846	18
WDBC	2	569	30
Yeast	10	1484	8

Table 2 Characteristics of the data sets.

5.3 Experimental Results

For each pruned ensemble size L, base classifier, diversity measure and value of λ, Tab. 4 shows the results of our experiments in terms of the statistical significance of the difference in test set accuracy of the FS pruning method implemented using the two considered evaluation functions. More precisely, the null hypothesis is that there is no difference between these evaluation functions. In Tab. 4 entries marked with 'A' mean that for the corresponding pruned ensemble size,

Dataset	**hidden units**	**NN**	**DT**	***K*-NN**
Bank Note	12	0.1	0.6	0.6
Banana	3	0.7	0.7	0.7
Blood Transfusion	3	0.5	0.4	0.1
Cardiotocography	7	0.1	0.6	0.2
Pop Failures	3	0.5	0.6	0.6
SatLogLandSetSat	12	0.6	0.5	0.1
SataLogImageSeg	20	0.6	0.5	0.1
Spam Base	3	0.4	0.4	0.4
Thyroid	3	0.1	0.3	0.3
Wine Quality	7	0.4	0.6	0.5
Australian	12	0.4	0.5	0.5
Balance Scale	12	0.5	0.6	0.2
Bands	3	0.1	0.6	0.3
Breast Cancer	20	0.4	0.6	0.2
Bupa	12	0.4	0.5	0.6
Checker Board	12	0.6	0.6	0.1
Cleveland	7	0.6	0.5	0.6
Coil2000	3	0.1	0.6	0.6
Contours	20	0.5	0.6	0.3
Contraceptive	3	0.6	0.6	0.6
Dermatology	7	0.4	0.3	0.3
Hayes Roth	12	0.6	0.4	0.6
ILPD	3	0.1	0.5	0.1
aryngeal 2	3	0.2	0.5	0.1
Marketing	7	0.6	0.6	0.6
Monk 2	12	0.6	0.5	0.2
Page Plocks	7	0.4	0.5	0.6
Pen based	8	0.7	0.3	0.7
Phoneme	7	0.1	0.6	0.6
Pima	12	0.6	0.6	0.1
Ring	20	0.5	0.5	0.6
Saheart	12	0.4	0.4	0.6
Segment	20	0.5	0.6	0.3
Spectfheart	20	0.2	0.4	0.6
Vehicle	12	0.6	0.5	0.6
WDBC	3	0.6	0.3	0.1
Yeast	7	0.3	0.6	0.1

Table 3 For each data set, the number of hidden units for the NN base classifiers (second column) and the training set size for the three base classifiers (NNs, DTs and k-NNs) is shown.

base classifier, diversity measure and value of λ, using only ensemble accuracy (estimated from validation data) as the evaluation function is significantly better (according to Wilcoxon signed-rank test) than using its linear combination with the diversity measure. Entries marked with 'D' mean the opposite (the latter evaluation function is significantly better than the former). We

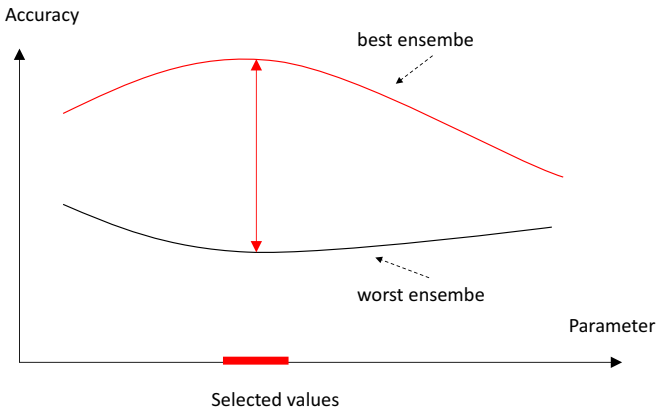

Figure 1 Qualitative illustration of the criterion used for choosing the training set size and the number of hidden units in NN classifiers (X axis): maximizing the accuracy gap between the best and the worst ensemble of a given size (see text for the details).

point out that the null hypothesis has always been rejected; therefore, every entry of Tab. 4 is marked with either 'A' or 'D'.

These results provide a quite strong evidence that a linear combination of ensemble accuracy and of a diversity measure between ensemble members outperforms the use of ensemble accuracy alone as the pruning evaluation function, to a statistically significant extent.

The table clearly shows that using $A + \lambda D$ as the evaluation function in the FS algorithm provides a statistically significantly better pruned ensembles than using accuracy alone, in almost all the considered cases. The only exceptions can be observed for the largest considered ensembles ($L = 35$) of DT classifiers, when DF and θ were used as diversity measures, and the λ coefficient was 0.2 and 0.5; and for ensembles of various sizes of NN classifiers, when the other two diversity measures (Dis and GD) were used and the λ coefficient was 0.5 and 0.7. It is also worth noting that the $A + \lambda D$ evaluation function always outperformed its counterpart A for ensembles of K-NN classifiers, and with the only exception of the largest ensembles ($L = 35$) for the DT classifier. With regard to the diversity measures, using DF, θ and GD in the $A + \lambda D$ evaluation function turned out to be worse than using A alone only for 2 out of the 108 combinations of pruned ensemble size, base classifier and value of λ (3 diversity measures, 4 ensemble sizes, 3 base classifiers and 3 values of λ); using Dis, this happened for 4 out of the 36 combinations. Due to the lack of space, we have not included the detailed results in the paper. Detailed results are available on the Pralab website[2].

As far as our experiments are concerned, we can conclude that well-known, 'generic' ensemble diversity measures (i.e., not specifically devised for ensemble pruning) seem to be useful when used together with ensemble accuracy as the pruning evaluation function. In particular, such diversity measures seem to act as regularizers of the estimated ensemble accuracy, which is in agreement with the more specific results of [14].

[2] `http://pralab.diee.unica.it/en/TAAI2018Appendix1`

		L=5			L=15			L=25			L=35		
Base	**Diversity**	λ			λ			λ			λ		
classifier		0.2	0.5	0.7	0.2	0.5	0.7	0.2	0.5	0.7	0.2	0.5	0.7
DT	DF	D	D	D	D	D	D	D	D	D	**A**	**A**	D
	Theta	D	D	D	D	D	D	D	D	D	**A**	**A**	D
	DIS	D	D	D	D	D	D	D	D	D	D	D	D
	GD	D	D	D	D	D	D	D	D	D	D	D	D
KNN	DF	D	D	D	D	D	D	D	D	D	D	D	D
	Theta	D	D	D	D	D	D	D	D	D	D	D	D
	DIS	D	D	D	D	D	D	D	D	D	D	D	D
	GD	D	D	D	D	D	D	D	D	D	D	D	D
NN	DF	D	D	D	D	D	D	D	D	D	D	D	D
	Theta	D	D	D	D	D	D	D	D	D	D	D	D
	DIS	D	**A**	**A**	D	D	**A**	D	D	**A**	D	D	D
	GD	D	**A**	**A**	D	D	D	D	D	D	D	D	D

Table 4 Outcome of the statistical significance test for the comparison between the use of the evaluation functions A and $A + \lambda D$ (see text) for ensemble pruning, for several ensemble sizes L, values of λ, base classifiers and diversity measures. 'A' means that the evaluation function A is statistically significantly better than $A + \lambda D$, 'D' means the opposite (see text for the details).

6 Conclusions

Whereas the usefulness of diversity measures for ensemble construction has been questioned by some authors, their specific role as regularizers has been recently pointed out in [14] based on theoretical results as well as on empirical evidence in the context of ensemble pruning, although in a specific setting (binary classifiers, and an ad hoc diversity measure).

As a follow-up of our preliminary work [15], in this paper we investigated the effectiveness of well-known, generic diversity measures in ensemble pruning. In particular, we considered their use in the ensemble evaluation function of pruning methods based on the forward search strategy, by linearly combining them with ensemble accuracy (estimated from validation data). This can be viewed as using diversity measures as regularizers, in the spirit of [14].

As far as our experiments are concerned, our empirical results provided evidence that also generic ensemble diversity measures can be useful when used together with ensemble accuracy as the pruning evaluation function. This is in agreement with the results we obtained in [15], related to ad hoc evaluation functions proposed by other authors for ensemble pruning, that combine individual classifiers' (not ensemble) accuracy and diversity (more precisely, complementarity between their errors). Our results also show that also generic diversity measures can have a regularization effect on the estimated ensemble accuracy, in the context of ensemble pruning. This provides some evidence that the results of [14], related to a specific diversity measure, could be extended to generic diversity measures.

References

[1] Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. CRC press, 1st edition, 2012.

[2] L. I. Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014. DOI: 10.1002/9781118914564.index.

[3] L. I. Kuncheva and Ch. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003. DOI: 10.1023/A:1022859003006.

[4] E. K. Tang, P. N. Suganthan, and X. Yao. An analysis of diversity measures. *Machine Learning*, 65(1):247–271, 2006. DOI: 10.1007/s10994-006-9449-2.

[5] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, mar 2005. DOI: 10.1016/j.inffus.2004.04.004.

[6] L. Didaci, G. Fumera, and F. Roli. Diversity in classifier ensembles: Fertile concept or dead end? In *Multiple Classifier Systems*, pages 37–48. Springer Berlin Heidelberg, 2013. DOI: 10.1007/978-3-642-38067-9_4.

[7] N. Ueda and R. Nakano. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks (ICNN'96)*. IEEE. DOI: 10.1109/icnn.1996.548872.

[8] G. Brown and L. I. Kuncheva. "Good" and "bad" diversity in Majority Vote Ensembles. In *Multiple Classifier Systems*, pages 124–133. Springer Berlin Heidelberg, 2010. DOI: 10.1007/978-3-642-12127-2_13.

[9] P. Sollich and A. Krogh. Learning with ensembles: How over-fitting can be useful. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS'95, pages 190–196. MIT Press, 1995.

[10] L. I. Kuncheva. A bound on kappa-error diagrams for analysis of classifier ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 25(3):494–501, 2013. DOI: 10.1109/tkde.2011.234.

[11] G. Martínez-Muñoz and A. Suárez. Pruning in ordered bagging ensembles. In *Proceedings of the 23rd international conference on Machine learning – ICML'06*. ACM Press, 2006. DOI: 10.1145/1143844.1143921.

[12] Z.-H. Zhou and W. Tang. Selective ensemble of decision trees. In *Lecture Notes in Computer Science*, pages 476–483. Springer Berlin Heidelberg. DOI: 10.1007/3-540-39205-x_81.

[13] Z.-H. Zhou, J. Wu, and W. Tang. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1-2):239–263, 2002. DOI: 10.1016/s0004-3702(02)00190-x.

[14] N. Li, Y. Yu, and Z.-H. Zhou. Diversity regularized ensemble pruning. In *Machine Learning and Knowledge Discovery in Databases*, pages 330–345. Springer Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-33460-3_27.

[15] M. A. O. Ahmed, L. Didaci, G. Fumera, and F. Roli. An empirical investigation on the use of diversity for creation of classifier ensembles. In *Multiple Classifier Systems*, pages 206–219. Springer International Publishing, 2015. DOI: 10.1007/978-3-319-20248-8_18.

[16] Y. Yu, Y.-F. Li, and Z.-H. Zhou. Diversity regularized machine. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence – Volume Volume Two*, IJCAI'11, pages 1603–1608. AAAI Press, 2011. DOI: 10.5591/978-1-57735-516-8/IJCAI11-269.

[17] G. Tsoumakas, I. Partalas, and I. Vlahavas. An ensemble pruning primer. In *Studies in Computational Intelligence*, pages 1–13. Springer Berlin Heidelberg, 2009. DOI: 10.1007/978-3-642-03999-7_1.

[18] D. Partridge and W. B. Yates. Engineering multiversion neural-net systems. *Neural Computation*, 8(4):869–893, 1996. DOI: 10.1162/neco.1996.8.4.869.

[19] D. D. Margineantu and T. G. Dietterich. Pruning adaptive boosting. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 211–218. Morgan Kaufmann Publishers Inc., 1997.

[20] A. L. Prodromidis and S. J. Stolfo. Pruning meta-classifiers in a distributed data mining system. In *In Proc of the First National Conference on New Information Technologies*, pages 151–160, 1998.

[21] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes. Ensemble selection from libraries of models. In *Twenty-first international conference on Machine learning – ICML '04*. ACM Press, 2004. DOI: 10.1145/1015330.1015432.

[22] G. Martınez-Munoz and A. Suárez. Aggregation ordering in bagging. In *Proc. of the IASTED International Conference on Artificial Intelligence and Applications*, pages 258–263. Citeseer, 2004.

[23] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. Ensemble diversity measures and their application to thinning. *Information Fusion*, 6(1):49–62, 2005. DOI: 10.1016/j.inffus.2004.04.005.

[24] I. Partalas, G. Tsoumakas, and I. Vlahavas. An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Machine Learning*, 81(3):257–282, 2010. DOI: 10.1007/s10994-010-5172-0.

[25] G. D. C. Cavalcanti, L. S. Oliveira, T. J.M. Moura, and G. V. Carvalho. Combining diversity measures for ensemble pruning. *Pattern Recognition Letters*, 74:38–45, 2016. DOI: 10.1016/j.patrec.2016.01.029.

[26] H. Guo, H. Liu, R. Li, Ch. Wu, Y. Guo, and M. Xu. Margin & diversity based ordering ensemble pruning. *Neurocomputing*, 275:237–246, 2018. DOI: 10.1016/j.neucom.2017.06.052.

[27] G. Zenobi and P. Cunningham. Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. In *Machine Learning: ECML 2001*, pages 576–587. Springer Berlin Heidelberg, 2001. DOI: 10.1007/3-540-44795-4_49.

[28] A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Diversity in search strategies for ensemble feature selection. *Information Fusion*, 6(1):83–98, 2005. DOI: 10.1016/j.inffus.2004.04.003.

[29] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. DOI: 10.1109/34.709601.

[30] G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification purposes. *Image and Vision Computing*, 19(9–10):699–707, 2001. DOI: 10.1016/s0262-8856(01)00045-2.

[31] D. Partridge and W. Krzanowski. Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology*, 39(10):707–717, 1997. DOI: 10.1016/s0950-5849(97)00023-2.

[32] D. B. Skalak. The sources of increased accuracy for two proposed boosting algorithms. In *Proc. American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*, pages 120–125, 1996.

[33] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990. DOI: 10.1109/34.58871.

[34] G. Martinez-Muoz, D. Hernandez-Lobato, and A. Suarez. An analysis of ensemble pruning techniques based on ordered aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):245–259, 2009. DOI: 10.1109/tpami.2008.78.

[35] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. DOI: 10.1007/bf00058655.

[36] J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.