An Empirical Investigation on the Use of Diversity for Creation of Classifier Ensembles

Muhammad A.O. Ahmed, Luca Didaci, Giorgio Fumera(✉), and Fabio Roli

Department of Electrical and Electronic Engineering, University of Cagliari,
Piazza d'Armi, 09123 Cagliari, Italy
{muhammad.ahmed,didaci,fumera,roli}@diee.unica.it
http://pralab.diee.unica.it

Abstract. We address one of the main open issues about the use of diversity in multiple classifier systems: the effectiveness of the explicit use of diversity measures for creation of classifier ensembles. So far, diversity measures have been mostly used for ensemble pruning, namely, for selecting a subset of classifiers out of an original, larger ensemble. Here we focus on pruning techniques based on forward/backward selection, since they allow a direct comparison with the simple estimation of accuracy of classifier ensemble. We empirically carry out this comparison for several diversity measures and benchmark data sets, using bagging as the ensemble construction technique, and majority voting as the fusion rule. Our results provide further and more direct evidence to previous observations against the effectiveness of the use of diversity measures for ensemble pruning, but also show that, combined with ensemble accuracy estimated on a validation set, diversity can have a regularization effect when the validation set size is small.

Keywords: Diversity · Ensemble pruning · Forward/backward selection · Ensemble construction

1 Introduction

After about twenty years of active research in the classifier ensemble field, understanding the notion of diversity remains one of the main open problems [11,25]. On the one hand, there is a general agreement on the qualitative definition of diversity and on its role, e.g.: "it is desired that the individual learners should be *accurate and diverse*" [25]; "Common sense suggests that the classifiers in the ensemble should be as accurate as possible and should not make coincident errors" [11] (Chap. 8). On the other hand, measuring diversity and explicitly using it for ensemble construction exhibits several open issues.

A number of **diversity measures** have been proposed over the years [9,11,25]. Most measures have been derived intuitively, as attempts to formally characterize the pattern of individual classifiers' errors (e.g., the Double-Fault and Disagreement measures [11]). In particular, it has been clearly pointed out that diversity measures alone can not be monotonically related to ensemble accuracy, since the

F. Schwenker et al. (Eds.): MCS 2015, LNCS 9132, pp. 206–219, 2015.
DOI: 10.1007/978-3-319-20248-8_18

latter depends instead on a trade-off between diversity and individual classifiers' performance [11,19]; quoting from [11] (Chap. 8), looking 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." A few other measures have been inspired by *exact* error decompositions derived in the regression field, despite the lack of a direct analogy with regression problems was pointed out in [2]: the Kohavi-Wolpert Variance [9] (and our attempt in [6]) was inspired by the bias-variance-covariance error decomposition [21], and the measure derived in [3] (which we extended in [6]) by the ambiguity decomposition [8]. The rationale of such measures is to look for exact, additive decompositions of the ensemble error into terms accounting for individual classifiers' performance, and terms hopefully interpretable as diversity; the results of [3] provided useful insights, leading to the concept of "good" and "bad" 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., [10,19]). Such a relationship turned out to be far from clear-cut, and no "right" diversity measure has emerged so far.

Almost all the existing methods that **explicitly use diversity for ensemble construction** follow the overproduce and choose approach (except for [24], where a diversity measure is used in an ensemble *learning* algorithm). It consists of first generating a large ensemble (e.g., using bagging) and then selecting the most accurate subset of classifiers (usually with a predefined size). This is known as ensemble *pruning*, *selection* or *thinning*. Since this problem 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 look for a trade-off between individual classifiers' performance and diversity. The effectiveness of such an approach has however been questioned by several authors, based also on empirical evidences [11,19] (Chap. 8.3). In particular, its actual advantage over directly evaluating ensemble performance (estimated, e.g., from validation data) is not clear yet. On the other hand, it is well known that popular and effective ensemble construction techniques like bagging and boosting do not use any explicit diversity measure.

In [6] we discussed the above issues, focusing on the derivation of exact decompositions of the ensemble error, and outlined several research directions. One of them, which we start addressing in this work, consists of comparing the effectiveness of explicitly using diversity measures in ensemble pruning, with the simple estimation of ensemble performance. Although many pruning methods have been proposed so far, the above comparison has been carried out by only a few authors, and with a limited scope. In this work we focus on pruning methods based on forward/backward selection (FS/BS) algorithms, which are the easiest ones on which such a comparison can be made, and carry out an empirical investigation on 23 benchmark data sets, using the popular bagging as the ensemble construction technique, and majority voting as the fusion rule. We evaluate ten well known diversity measures analyzed in [9], and five measures specifically defined for ensemble pruning. We also evaluate the effect of the validation set size on ensemble pruning effectiveness.

Algorithm 1. Forward Selection algorithm for ensemble pruning

Input: an ensemble E of N classifiers; a desired ensemble size $L < N$; a validation set V; an objective function m (to be computed on V)
Output: a subset of L classifiers from E

$C \leftarrow$ the most accurate individual classifier from E
$S \leftarrow \{C\}$
for $k = 2, \ldots, L$ **do**
  $C^* \leftarrow \arg\max_{C \in E \setminus S} m(S \bigcup \{C\})$
  $S \leftarrow S \bigcup \{C^*\}$
end for
return S

2 Ensemble Pruning with Forward/Backward Selection

Ensemble pruning methods can be categorized as follows [20]:

- **Ranking-based:** individual classifiers are first ranked according to some criterion, and then the top-L are selected to form 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 (often implemented as FS or BS), genetic algorithms, and semi-definite programming.

We focus on optimization-based methods in which FS/BS is used, since they allow a direct comparison between the simple estimation of ensemble accuracy and objective functions involving diversity. Several pruning methods based on FS/BS, together with specific objective functions, have been proposed so far, including [1,4,13–17]. Given an initial ensemble E of size N, FS constructs a pruned ensemble S of size $L < N$ by starting from the best individual classifier from E, and iteratively adding a classifier to S by maximizing a given objective function (see Algorithm 1).[1] 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 [13].

Three kinds of objective functions have been proposed so far:

- The ensemble performance, [13] (reduce-error pruning technique), [4,12].
- Diversity measures alone, disregarding the performance of individual classifiers and of the ensemble, [13] (Kullback-Leibler Divergence pruning), [17] and [1] (kappa-thinning).

[1] If no predefined size is given, FS stops when all the classifiers from E have been added, and returns the best ensemble among the N ones obtained at every iteration.

– Measures specifically defined for ensemble pruning. They combine into a single scalar the individual classifiers' performance and the *complementarity* between their errors [14–16] and [1] (AID thinning and Concurrency thinning). We will refer to them as *pruning measures*.

Among the existing pruning measures, we focus on the following ones. Let $(\mathbf{x}, y)$ denote a sample with its class label, V the validation set, E and S the original and the current pruned ensemble, C^* the candidate classifier to be added to (or removed from) S, and $S(\mathbf{x})$ the label assigned to $\mathbf{x}$ by S.

– A measure aimed at minimizing the number of coincident errors between ensemble members, when majority voting is used, to be used in the FS algorithm [16] (Sect. 5.2). It selects the classifier C^* that correctly labels the highest number of validation samples, among the ones misclassified by the majority of classifiers in the current ensemble S:

$$C^* = \arg\min_{C \in E \setminus S} \sum_{(\mathbf{x},y) \in V} \begin{array}{r} I\left[C(\mathbf{x}) \neq y \wedge S(\mathbf{x}) \neq y\right] \\ - I\left[C(\mathbf{x}) = y \wedge S(\mathbf{x}) \neq y\right], \end{array} \tag{1}$$

  where $I[A] = 1$ if $A =$True, and $I[A] = 0$ otherwise.
– Two measures proposed in [14] to be used in the FS algorithm, with the majority voting rule: Complementariness (the sum of validation samples which are wrongly classified by the current ensemble, but not by the candidate classifier) and Margin Distance. The former is a variant of Eq. (1). They are respectively defined as:

$$C^* = \arg \max_{C \in E \setminus S} \sum_{(\mathbf{x},y) \in V} I\left[C(\mathbf{x}) = y \wedge S(\mathbf{x}) \neq y\right], \tag{2}$$

$$C^* = \arg \min_{C \in E \setminus S} \left\| \mathbf{o} - \frac{1}{|E|} \left(\mathbf{c}_C + \sum_{C' \in S} \mathbf{c}_{C'} \right) \right\|_2^2, \tag{3}$$

  where $\mathbf{c}_{C'}$ is a $|V|$-dimensional vector whose i-th element is defined as:

$$2I[C'(\mathbf{x}_i) = y_i] - 1 \in \{-1, +1\},$$

  and $\mathbf{o}$ is defined as a constant vector whose components are all identical to some value p, with $0 < p < 1$.
– A measure proposed in the context of the Concurrency thinning technique in [1], based on BS. It chooses the classifier to be removed from S with the aim of penalizing the agreement on correctly classified samples (this is a variant of Eq. (1) as well):

$$C^* = \arg\min_{C \in S} \sum_{(\mathbf{x},y) \in V} \begin{array}{r} I\left[C(\mathbf{x}) = y \wedge S(\mathbf{x}) = y\right] \\ + 2I\left[C(\mathbf{x}) = y \wedge S(\mathbf{x}) \neq y\right] \\ - 2I\left[C(\mathbf{x}) \neq y \wedge S(\mathbf{x}) \neq y\right]. \end{array} \tag{4}$$

– The Uncertainty Weighted Accuracy (UWA), to be used in the FS algorithm; it was proposed in [15] as a variant of the Concurrency measure of Eq. (4):

$$C^* = \arg\max_{C \in E \setminus S} \sum_{(\mathbf{x},y) \in V} \begin{array}{l} NF(\mathbf{x}) \times I\left[C(\mathbf{x}) = y \wedge S(\mathbf{x}) = y\right] \\ + NT(\mathbf{x}) \times I\left[C(\mathbf{x}) = y \wedge S(\mathbf{x}) \neq y\right] \\ - NF(\mathbf{x}) \times I\left[C(\mathbf{x}) \neq y \wedge S(\mathbf{x}) = y\right] \\ - NT(\mathbf{x}) \times I\left[C(\mathbf{x}) \neq y \wedge S(\mathbf{x}) \neq y\right], \end{array} \tag{5}$$

where $NT(\mathbf{x})$ and $NF(\mathbf{x})$ are the number of classifiers in S that classify $\mathbf{x}$ respectively correctly and wrongly.

3 Aim of This Work

A comparison between the effectiveness of directly using ensemble performance as the objective function, and using measures involving diversity, has been carried out by a few authors [1,12–15], often limited to the specific evaluation measure they were proposing, and using different and incomparable experimental setups (different data sets, base classifiers, ensemble construction methods, etc.). We also point out that only in [12,15] the use of pruning measures provided a statistically significant improvement over the use of ensemble performance.

Our aim is thus to carry out an extensive experimental investigation of FS/BS-based ensemble pruning methods, focused on the comparison between the use of ensemble performance as the objective function, and the use of measures involving diversity. To this aim, we focus on the basic FS/BS algorithm without back-fitting, and consider three kinds of objective functions:

1. Ensemble accuracy.
2. A generic diversity measure, focusing on the ones analyzed in [9]. Although diversity alone is deemed to be not effective for ensemble pruning [11,19], we consider also this option to provide a more direct evidence to these findings.
3. Pruning measures, which combine individual classifiers' performance and complementarity: we consider the ones described in Sect. 2, Eqs. (1)–(5).

We also consider another way to combine ensemble performance and diversity. Since diversity measures are not homogeneous to classification accuracy, to avoid combining them with individual classifiers' accuracy in an arbitrary way (e.g., by a linear combination), we use a two-stage FS/BS: first we select $M < N$ classifiers using either ensemble accuracy or diversity; then we further select $L < M$ classifiers using the other measure. Algorithm 2 shows the version in which ensemble accuracy is used at the first stage. In our experiments we considered both versions.

4 Experimental Setting

We chose 23 benchmark data sets from the UCI Machine Learning Repository Database,[2] with at least 350 samples, only numerical attributes, and without

[2] http://www.ics.uci.edu/~mlearn/MLRepository.html.

Algorithm 2. Two-stage Forward Selection algorithm for ensemble pruning

Input: a classifier ensemble E of size N; a desired ensemble size $L < N$; an intermediate ensemble size M, with $L < M < N$; a validation set V; a diversity measure d
Output: a subset of L classifiers from E

step 1 (accuracy-based pruning): select from E an ensemble E' of size M using Algorithm 1, and using classification accuracy as the objective function m
step 2 (diversity-based pruning): select from E' an ensemble S of size L using Algorithm 1, and using d as the objective function m
return S

missing values (see Table 1). We used bagging to construct the original ensemble, majority voting as the combining rule, and two different base classifiers: multi-layer perceptron neural networks (MLP-NN) with one hidden layer containing ten units, and decision trees (DT). For MLP-NN we used the standard Matlab implementation[3], learning rate $\eta = 0.05$, and maximum number of training epochs equal to 300. For DTs we used the code of [11] (par. 2.A.2.1), with the Gini impurity criterion, χ^2 stopping criterion, and the default threshold equal to 1 for the pre-pruning stopping criterion. We set the size of the original ensemble to $N = 100$, and considered four different sizes of the pruned ensembles: $L = 5, 15, 25$ and 35.

We used only FS-based pruning. In the two-stage Algorithm 2 we set the size M of the first-stage pruned ensemble to $M = L + \lfloor (N - L)/2 \rfloor$. Since FS-based pruning starts from the best individual classifier, to better appreciate its effectiveness we chose the training set size of each data set in preliminary experiments, by maximizing the difference between the accuracy of an ensemble of 100 classifiers (constructed by bagging) and of the best individual classifier (see the right-most column of Table 1). We then set the size of the validation as one third of the training set, and used the remaining samples as a testing set. We also used only half of the validation set (one sixth of the training set) to evaluate the effect of validation set size on the performance of ensemble pruning. We considered the ten diversity measures analyzed in [9] (the ones in the top rows of Table 2), as well as measures in Eqs. (1)–(5), which combine into a single scalar the individual classifiers' performance and the complementarity between their errors (the ones in the bottom five rows of Table 2).

We carried out 20 runs of the experiments. At each run we selected the training, validation and testing sets by stratified random sampling (no data set was originally subdivided into a training and a testing set). We applied bagging to the training set, to construct the original ensemble of $N = 100$ classifiers. We then run Algorithm 1 separately using as the objective function the ensemble accuracy, each diversity measure, and the pruning measures in Eqs. (1)–(5). We also run the two-stage Algorithm 2 in both versions (using accuracy either at the first or at the second stage), for each diversity measure. We finally computed, separately for each data set, pruning method, base classifier,

[3] http://it.mathworks.com/help/nnet/ref/patternnet.html.

Table 1. Characteristics of the data sets. The two rightmost columns report the size of the training set for the two base classifiers, as a fraction of the whole data set.

Dataset	Samples	Classes	Features	Tr. set size	
				MLP-NN	DT
Australian	690	2	14	0.42	0.42
Balance scale	625	3	4	0.18	0.42
Blood transfusion	748	2	4	0.48	0.60
Breast cancer	699	2	9	0.30	0.12
Bupa	345	2	6	0.54	0.06
Checker board	1000	2	2	0.36	0.30
Coil 2000	9822	2	85	0.06	0.18
Cone tours	2000	3	2	0.06	0.24
Contraceptive	1473	3	9	0.36	0.60
ILPD	583	2	9	0.50	0.06
Laryngeal 2	692	2	16	0.06	0.48
Monk2	432	2	6	0.48	0.06
Page blocks	5473	5	10	0.06	0.42
Phoneme	5404	2	5	0.36	0.30
Pima Indians	768	2	8	0.54	0.30
Pop failures	540	2	20	0.42	0.30
Ring	7400	2	20	0.42	0.30
SaHeart	462	2	4	0.54	0.18
Sata log image seg	2310	7	19	0.44	0.30
Landsat Satellite	6435	7	36	0.60	0.48
Spam base	4601	2	57	0.42	0.30
Townorm	7400	2	20	0.12	0.30
Wine quality	4898	7	11	0.18	0.30

ensemble size L and validation set size, the average accuracy and its standard deviation on testing samples, over the 20 runs. Due to space limits, we make these results available only from our web site,[4] and only report the results of the statistical significance test. We compared the accuracy of pruned ensembles attained by Algorithm 1 using ensemble accuracy as the objective function, and using each of the other measures (both by Algorithms 1 and 2). To this aim we used the Wilcoxon signed-rank test, which is recommended in [5] for comparing two algorithms over multiple data sets. Our goal was to assess whether the difference was significant, and, if so, whether using ensemble accuracy as the objective function was the best or the worst option. Accordingly, we made two

[4] http://pralab.diee.unica.it/en/MCS2015Appendix1.

Table 2. Diversity measures (top ten rows, from [9]) and pruning measures (in the other rows, defined in Eqs. (1)–(5)) used in the experiments.

Diversity/pruning measure	Abbreviation
Entropy	E
Kohavi-Wolpert	KW
Coincidence failure diversity	CFD
Generalized diversity	GD
Interrater agreement	Kappa
Difficulty	Theta
Q Statistic	Q
Correlation	Rho
Disagreement	D
Double fault	DF
Uncertainty weighted accuracy	UWA
Partridge and yates' measure	PYM
Complementariness	Cs
Margin distance	MD
Concurrency	Cy

one-sided tests (at the $\alpha = 0.05$ level), evaluating the null hypotheses that FS-based pruning using ensemble accuracy (or a measure involving diversity) is not better than using a given measure involving diversity (or ensemble accuracy). Only if *both* null hypotheses are rejected, it can be concluded that there is no statistically significant difference between the two options.

5 Experimental Results

For each pruned ensemble size L, base classifier, and validation set size, Tables 3, 4, 5, 6, 7 and 8 report the comparison between FS-based pruning (Algorithm 1) using ensemble accuracy, and FS-based pruning implemented by Algorithm 1 using either a diversity or a pruning measure, and by Algorithm 2 combining ensemble accuracy and diversity.

Tables 3 and 4 clearly show that using ensemble accuracy often provides a better or comparable pruned ensemble than using any diversity measure alone, or a pruning measure. The only exceptions are GD (with $L = 15$) and UWA (with $L = 35$), using DT as the base classifier and a small validation set (see Table 3).

Interestingly, most of the cases when using diversity attained comparable results occur for three only measures: Entropy, Generalized Diversity and Kappa.

Tables 5, 6, 7 and 8, which refer to the two-stage FS algorithm combining ensemble performance and diversity, show a different pattern, instead. When a larger validation set is used, ensemble accuracy still produces often a better or

Table 3. Comparison of FS-based pruning (Algorithm 1) using ensemble accuracy vs. using each diversity or pruning measure, for different ensemble sizes L and validation set sizes. Base classifier: DT. 'A': using accuracy is statistically significantly better than using the corresponding diversity/other measures, over the 23 data sets; 'D': using the corresponding diversity/other measures is better than ensemble accuracy; '-': there is no statistically significant difference between the two measures.

Diversity	Ensemble size L							
	Val. size: 1/3 Tr. size				Val. size: 1/6 Tr. size			
measure	5	15	25	35	5	15	25	35
E	-	-	-	-	-	-	-	-
KW	A	A	A	A	A	A	A	A
CFD	A	A	A	A	A	A	A	A
GD	-	-	-	-	-	D	-	-
Kappa	-	-	-	-	-	-	-	-
Theta	-	A	-	A	A	-	-	-
Q	-	A	-	A	A	-	-	-
Rho	A	A	A	A	-	A	A	A
D	A	A	A	A	A	A	-	-
DF	A	A	A	A	A	A	A	A
UWA	A	A	A	A	-	-	-	D
PYM	-	-	-	-	-	-	-	-
Cs	A	A	A	A	A	A	A	A
MD	-	-	-	-	-	-	-	-
Cy	-	-	-	-	-	-	-	-

comparable pruned ensemble; however, for ensembles of DTs it never outperforms the combination of ensemble performance and diversity; moreover, it almost always performs worse with respect to the Double Fault (DF) measure. When a smaller validation set is used, together with DT classifiers, instead, combining ensemble accuracy and diversity is often better, or at least not worse, than using only ensemble accuracy (four right-most columns of Tables 5 and 7, vs the same columns of Table 3). Remarkably, this happens for most diversity measures.

These results seem to suggest that estimating the ensemble performance is the best option for FS-based pruning, provided that a sufficiently large validation set is available. Otherwise, a combination of ensemble performance and diversity can be advantageous, at least for some types of base classifiers. One possible explanation is that diversity measures have a *regularization* effect capable of preventing over-fitting, to some extent, as already argued in [12]. This is an interesting and non-straightforward property, which is worth investigating more throughly.

Table 4. Comparison of FS-based pruning (Algorithm 1) using ensemble accuracy vs. using each diversity or pruning measure, for a validation set size equal to 1/3 and 1/6 of the training set size. Base classifier: MLP-NN. See caption of Table 3 for the meaning of table entries.

	Ensemble size L							
Diversity	Val. size: 1/3 Tr. size				Val. size: 1/6 Tr. size			
measure	5	15	25	35	5	15	25	35
E	-	-	-	-	A	-	-	-
KW	A	A	A	A	-	-	-	-
CFD	A	A	A	A	-	-	-	-
GD	-	-	-	-	-	-	-	-
Kappa	-	-	-	-	-	-	-	-
Theta	A	A	A	A	-	-	-	-
Q	A	A	A	A	-	-	-	-
Rho	A	A	A	A	-	-	-	-
D	A	A	A	A	-	-	-	-
DF	A	A	A	A	-	-	-	-
UWA	-	-	-	-	-	-	-	-
PYM	-	-	-	-	-	-	-	-
Cs	A	A	A	A	A	A	A	A
MD	-	-	-	-	-	-	-	-
Cy	-	-	-	-	-	-	-	-

Table 5. Comparison of FS-based pruning (Algorithm 1) using ensemble accuracy vs Algorithm 2 using ensemble accuracy at the first stage and each diversity measure at the second stage. Base classifier: DT. See caption of Table 3 for the meaning of table entries.

	Ensemble size L							
Diversity	Val. size: 1/3 Tr. size				Val. size: 1/6 Tr. size			
measure	5	15	25	35	5	15	25	35
E	-	-	-	-	D	D	D	D
KW	-	-	-	-	D	D	D	D
CFD	-	-	-	D	D	D	D	D
GD	-	-	-	D	D	D	D	-
Kappa	-	-	-	-	D	D	D	-
Theta	-	-	-	-	D	D	D	-
Q	-	-	-	-	-	D	D	-
Rho	-	-	-	-	D	D	D	-
D	-	-	-	-	D	D	D	-
DF	-	D	D	D	D	D	D	D

Table 6. Comparison of FS-based pruning (Algorithm 1) using ensemble accuracy vs Algorithm 2 using ensemble accuracy at the first step and each diversity measure at the second stage, for a validation set size equal to 1/3 and 1/6 of the training set size. Base classifier: MLP-NN. See caption of Table 3 for the meaning of table entries.

Diversity measure	Ensemble size L							
	Val. size: 1/3 Tr. size				Val. size: 1/6 Tr. size			
	5	15	25	35	5	15	25	35
E	A	A	A	A	A	A	A	A
KW	A	A	A	A	A	A	A	A
CFD	-	-	-	-	A	-	D	-
GD	-	-	-	-	A	-	-	-
Kappa	A	A	-	-	A	A	A	A
Theta	A	-	-	-	A	-	-	-
Q	A	-	-	A	A	A	A	A
Rho	A	A	A	-	A	A	A	A
D	A	A	A	A	A	A	A	A
DF	D	D	D	D	-	-	-	-

Table 7. Comparison of FS-based pruning (Algorithm 1) using ensemble accuracy vs Algorithm 2 using each diversity measure at the first stage and ensemble accuracy at the second stage. Base classifier: DT. See caption of Table 3 for the meaning of table entries.

Diversity easure	Ensemble size L							
	Val. size: 1/3 Tr. size				Val. size: 1/6 Tr. size			
	5	15	25	35	5	15	25	35
E	-	-	-	-	D	D	D	-
KW	-	-	-	-	D	D	-	-
CFD	-	-	-	-	D	D	-	D
GD	-	-	D	D	-	D	D	D
Kappa	-	-	-	-	-	D	D	-
Theta	-	-	-	-	-	D	D	-
Q	-	-	-	-	-	D	-	-
Rho	-	-	-	-	D	D	D	-
D	-	-	-	-	D	D	-	-
DF	-	-	D	D	-	D	D	D

Table 8. Comparison of FS-based pruning (Algorithm 1) using ensemble accuracy vs Algorithm 2 using each diversity measure at the first stage and ensemble accuracy at the second stage, for a validation set size equal to 1/3 and 1/6 of the training set size. Base classifier: MLP-NN. See caption of Table 3 for the meaning of table entries.

	Ensemble size L							
Diversity	Val. size: 1/3 Tr. size				Val. size: 1/6 Tr. size			
Measure	5	15	25	35	5	15	25	35
E	-	A	A	A	-	-	-	-
KW	-	A	A	A	A	A	-	A
CFD	-	-	-	-	A	D	D	-
GD	-	D	-	-	-	D	D	-
Kappa	-	A	A	-	A	A	-	-
Theta	-	A	A	-	A	A	-	-
Q	A	A	A	A	A	-	A	A
Rho	-	A	A	A	A	A	-	A
D	A	A	A	A	A	-	-	A
DF	-	-	-	-	-	-	-	-

6 Discussion

We empirically investigated the effectiveness of explicitly using diversity measures for FS-based ensemble pruning, vs the simple estimation of ensemble accuracy. On the one hand, our results provide a more direct evidence in support of previous findings that using diversity measures alone is not effective for ensemble pruning [11,19], and in particular are in agreement with the well-established fact that diversity is not monotonically related to ensemble accuracy [11]. On the other hand, they suggest that, combined with the performance of individual classifiers, diversity can be useful to FS-based pruning when a small validation set is available. It seems therefore that diversity has a regularization effect. This possible effect has already been argued through the derivation of generalization bounds in [22], in the context of constructing ensembles of support vector machines, as well as in [12], in the context of FS-based ensemble pruning. However, in [12] the effect of different validation set sizes was not assessed, and only one diversity and two pruning measures were considered for comparison (Table 6).

To sum up, what our results provide is not a sharp conclusion either in favor or against the effectiveness of explicitly using diversity measures for ensemble pruning. Instead, and perhaps more interestingly, they provide some hints on the conditions under which diversity can be useful, and clearly suggest as a future research direction a more thorough investigation of the effect of validation set size. Our analysis can also be extended to other pruning methods categorized in [20] as optimization-based, which use genetic algorithms [7,23] or a kind of

best-first search [18], where ensemble accuracy can also be used as the objective function. Finally, this investigation can be extended to regression problems, in which the exact Ambiguity decomposition includes a diversity term which does *not* depend on ground truth, contrary to most diversity measures for classification problems, including all the ones in [9] considered in this work, and the one in [3] derived from an *exact* Ambiguity-like decomposition; this allows it to be computed also on a set of *unlabeled* samples, thus potentially reducing the effect of over-fitting when a small set of (labelled) validation samples is available.

Acknowledgments. This work has been partly supported by the project CRP-59872 funded by Regione Autonoma della Sardegna, L.R. 7/2007, Bando 2012.

References

1. Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: EnsembleUWA diversity measures and their application to thinning. Inf. Fusion **6**(1), 49–62 (2005)
2. Brown, G., Wyatt, J.L., Harris, R., Yao, X.: Diversity creation methods: a survey and categorisation. Inf. Fusion **6**(1), 5–20 (2005)
3. Brown, G., Kuncheva, L.I.: "Good" and "Bad" diversity in majority vote ensembles. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 124–133. Springer, Heidelberg (2010)
4. Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A.: Ensemble selection from libraries of models. In: 21st International Conference on Machine Learning, p. 18. ACM (2004)
5. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2005)
6. Didaci, L., Fumera, G., Roli, F.: Diversity in classifier ensembles: fertile concept or dead end? In: Zhou, Z.-H., Roli, F., Kittler, J. (eds.) MCS 2013. LNCS, vol. 7872, pp. 37–48. Springer, Heidelberg (2013)
7. Ko, A.H.-R., Sabourin, R., de Souza Britto Jr., A.: Compound diversity functions for ensemble selection. Int. J. Patt. Rec. Artif. Int. **23**(4), 659–686 (2009)
8. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In: Tesauro, G., Touretzky, D.S., Leen, T.K. (eds.) Advances in Neural Information Processing Systems 7, pp. 231–238. MIT Press, Cambridge (1995)
9. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach. Learn. **51**(2), 181–207 (2003)
10. Kuncheva, L.I.: A bound on kappa-error diagrams for analysis of classifier ensembles. IEEE Trans. Knowl. Data Eng. **25**(3), 494–501 (2013)
11. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms, 2nd edn. Wiley, Hoboken (2014)
12. Li, N., Yu, Y., Zhou, Z.-H.: Diversity regularized ensemble pruning. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012, Part I. LNCS, vol. 7523, pp. 330–345. Springer, Heidelberg (2012)
13. Margineantu, D.D., Dietterich, T.G.: Pruning adaptive boosting. In: 14th International Conference Machine Learning, pp. 378–387. Morgan Kaufmann (1997)
14. Martinez-Munoz, G., Suarez, A.: Aggregation ordering in bagging. In: International Conference on Artificial Intelligence and Applications, pp. 258–263 (2004)

15. Partalas, I., Tsoumakas, G., Vlahavas, I.P.: An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. Mach. Learn. **81**, 257–282 (2010)
16. Partridge, D., Yates, W.B.: Engineering multiversion neural-net systems. Neural Comput. **8**(4), 869–893 (1996)
17. Prodromidis, A., Stolfo, S.J.: Pruning meta-classifiers in a distributed data mining system. In: Proceedings of the 1st National Conference on New Information Technologies, pp. 151–160 (1998)
18. Rokach, L.: Collective-agreement-based pruning of ensembles. Comp. Stat. Data Anal. **53**(4), 1015–1026 (2009)
19. Tang, E.K., Suganthan, P.N., Yao, X.: An analysis of diversity measures. Mach. Learn. **65**, 247–271 (2006)
20. Tsoumakas, G., Partalas, I., Vlahavas, I.: An ensemble pruning primer. In: Okun, Oleg, Valentini, Giorgio (eds.) Applications of Supervised and Unsupervised Ensemble Methods. SCI, vol. 245, pp. 1–13. Springer, Heidelberg (2009)
21. Ueda, N., Nakano, R.: Generalization error of ensemble estimators. In: International Conference on Neural Networks, pp. 90–95 (1996)
22. Yu, Y., Li, Y.-F., Zhou, Z.-H.: Diversity regularized machine. In: 22nd International Joint Conference on Artificial Intelligence, pp. 1603–1608 (2011)
23. Zhou, Z.-H., Wu, J., Tang, W.: Ensembling neural networks: many could be better than all. Artif. Intell. **137**(1–2), 239–263 (2002)
24. Yu, Y., Li, Y.-F., Zhou, Z.-H.: Diversity regularized machine. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, pp. 1603–1608 (2011)
25. Zhou, Z.-H.: Ensemble Methods: Foundations and Algorithms. CRC Press, USA (2012)