Diversity in classifier ensembles: fertile concept or dead end?

Luca Didaci, Giorgio Fumera, and Fabio Roli

Dept. of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
email: {luca.didaci, fumera, roli}@diee.unica.it
WWW home page: http://prag.diee.unica.it/en

Abstract. Diversity is deemed a crucial concept in the field of multiple classifier systems, although no exact definition has been found so far. Existing diversity measures exhibit some issues, both from the theoretical viewpoint, and from the practical viewpoint of ensemble construction. We propose to address some of these issues through the derivation of decompositions of classification error, analogue to the well-known bias-variance-covariance and ambiguity decompositions of regression error. We then discuss whether the resulting decompositions can provide a more clear definition of diversity, and whether they can be exploited more effectively for the practical purpose of ensemble construction.

Keywords: Diversity, Bias-variance-covariance decomposition, Ambiguity decomposition

1 Introduction

The concept of "diversity" is deemed among the most important in the field of multiple classifier systems (MCSs), both theoretically, as a way to understand how MCSs work, and as a practical tool for constructing effective classifier ensembles [21, 46]. However, its exact understanding and definition is still a relevant open issue. For instance, quoting from [46] (Sect. 5.1): "It is no doubt that understanding diversity is the holygrail in the field of ensemble learning".

Besides the obvious observation that combining identical classifiers is useless, the concept of diversity has roots in theoretical arguments (e.g., [36, 22, 15, 9]), also inspired by other domains like software engineering [24, 28]. In particular, it has been influenced by the *bias-variance-covariance* (BVC) [37] and the *ambiguity* [18, 4] decompositions of the error of regressor ensembles. Moreover, wide empirical evidence motivates the potential usefulness of combining *non-identical* classifiers. This lead to the widely accepted idea that: (i) there exists a property of MCSs that can be defined as "diversity", can be quantitatively defined and thus measured, and is related to ensemble accuracy (together with the accuracy of individual classifiers); (ii) such a property can be practically exploited to construct an effective ensemble of classifiers.

The concept of diversity has been investigated in the MCS literature from different aspects: several diversity measures have been defined (e.g., [17, 26, 8, 13, 32, 20, 1, 40, 23]); several methods for MCS construction have been proposed, explicitly using diversity measures (e.g., [41, 13, 1, 2, 16, 40, 23]); existing diversity measures have been analysed to understand whether and how much they are "correlated" with ensemble accuracy [32, 20, 35]; several authors analysed the concept of diversity itself [31, 19, 4, 5, 7].

However, although all the existing measures reflect intuitive notions of diversity, none of them has been derived from an exact decomposition of the ensemble error, contrary to the ambiguity decomposition; none of them exhibits a clear trade-off with the average error of ensemble members, in determining the ensemble error [20, 35]; and effective techniques for MCS construction like bagging and boosting do not make explicit use of diversity measures. These issues led some authors even to question the practical usefulness of measuring diversity in MCSs: "[...] the question of the participation of diversity measures in designing classifier ensembles is still open. Directly calculating the accuracy for the chosen combination method makes more sense than calculating the diversity and trying to predict the accuracy. Even if the measure of diversity is easier to calculate than some combination methods, the ambiguous relationship between diversity and accuracy discourages optimising the diversity" [32] (sect. 7); "The quest for defining and using diversity might be heading toward a dead end or might result in powerful new ensemble-building methodologies" [21] (sect. 10.5); "It is not yet known whether diversity is really a driving force, or actually a trap since it might be just another appearance of accuracy" [46] (sect. 5.1).

On the basis of the above premises, in this paper we address the issue of diversity with the following goals: (i) deriving *exact* decompositions of the ensemble error for any combining rule and any number of classes; and exploiting them to understand (ii) whether they can provide a more clear understanding of diversity, and (iii) whether they can be exploited for ensemble construction, more effectively than existing measures. After an overview on BVC and ambiguity decompositions for regression problems in Sect. 2, in Sect. 3 we address issue (i) above by deriving the analogue of these decompositions for the ensemble classification error. In particular, we consider the Kohavi-Wolpert bias-variance decomposition [17] to derive a BVC-like decomposition, while our ambiguity-like decomposition generalises the one of [7]. We then address issues (ii) and (iii) above in Sect. 4. We finally suggest some directions for future work in Sect. 5.

2 Background: decompositions of regression error

In regression problems, an unknown function has to be estimated using a set d of n samples of its input-output pairs, $(\mathbf{x}, y) \in \mathbb{R}^m \times \mathbb{R}$.[1] Assume that a learning algorithm is used, which produces the estimator $f(\mathbf{x}; d)$ when trained on d. To simplify notation, in the following we will write f in place of $f(\mathbf{x}; d)$.

[1] Throughout the paper we will use uppercase letters to denote random variables, and the corresponding lowercase letters to denote specific values.

The expectation of the mean squared error (MSE) of f on a given input $\mathbf{x}$, taken over random training sets D of size n and over $P[Y|\mathbf{x}]$, can be written in terms of the well-known bias-variance (BV) decomposition [12]:

$$E_{D,Y|\mathbf{x}}\left[(f-Y)^2\right] = \mathit{bias}_f^2 + \mathit{var}_f + \mathit{noise}\ , \tag{1}$$

where *noise* equals the variance of Y given $\mathbf{x}$, and is independent on D, while

$$\mathit{bias}_f^2 = \left(E_D[f] - E[Y|\mathbf{x}]\right)^2,\quad \mathit{var}_f = E_D\left[(f - E_D[f])^2\right]\ . \tag{2}$$

It is known that, usually, bias can be reduced only at the expense of a higher variance, and vice versa, and that an effective variance reduction technique consists of linearly combining an ensemble of N different regressors:

$$f_{\text{ens}}(\mathbf{x}; d) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x}; d)\ . \tag{3}$$

The BV decomposition of f_{ens} can be rewritten in the form of a bias-variance-covariance (BVC) decomposition [37, 4]. Let us define

$$\begin{aligned} &\overline{\mathit{bias}} = \tfrac{1}{N}\textstyle\sum_j \mathit{bias}_{f_j}\ ,\quad \overline{\mathit{var}} = \tfrac{1}{N}\textstyle\sum_j \mathit{var}_{f_j}\ ,\\ &\overline{\mathit{cov}} = \tfrac{1}{N(N-1)}\textstyle\sum_{i,j\neq i} E_D\left[(f_i - E_D[f_i])(f_j - E_D[f_j])\right]\ . \end{aligned} \tag{4}$$

It then follows that:

$$E_D[(f_{\text{ens}} - E[Y|\mathbf{x}])^2] = \overline{\mathit{bias}}^2 + \frac{1}{N}\overline{\mathit{var}} + (1-\frac{1}{N})\overline{\mathit{cov}}\ . \tag{5}$$

This highlights that the variance reduction effect strongly depends on the amount of correlation between the outputs of individual regressors: the lower the correlation (i.e., the lower the term $\overline{\mathit{cov}}$, which can also be negative), the higher the reduction of variance.

The MSE of f_{ens} can also be written equivalently in terms of the *ambiguity* decomposition, which, for a given $(\mathbf{x}, y)$ and d, is given by [18, 4]:

$$(y - f_{\text{ens}})^2 = \frac{1}{N}\sum_{i=1}^{N}(y - f_i)^2 - \frac{1}{N}\sum_{i=1}^{N}(f_i - f_{\text{ens}})^2\ . \tag{6}$$

Differently from the BVC decomposition, the ambiguity decomposition highlights a trade-off between the average accuracy of individual regressors, and their deviation from the ensemble output. The latter term was called "ambiguity" (hence the name of the decomposition),[2] and can be easily interpreted in terms of diversity between individual regressors. Therefore, this provides a clear, formal definition of "diversity" for regression problems [4].

[2] The ambiguity term is related to the correlation among individual regressors. The beneficial effects of negative correlation had already been pointed out in [29].

From the practical viewpoint of ensemble construction, the ambiguity decomposition was successfully exploited by the Negative Correlation Learning (NCL) method [25].[3] NCL is a *parallel*, gradient-descent learning algorithm, whose objective function is given by the linear combination of the MSE of individual regressors (the first term of the ambiguity decomposition), minus a term proportional to the corresponding ambiguity. Instead of independently training the individual regressors first, and then computing the coefficients of their linear combination, NCL pursues both goals simultaneously. This may allow it to attain a better trade-off between accuracy and diversity.

In principle, the ambiguity decomposition could also be exploited in the context of an overproduce and choose strategy, for selecting the best subset of regressors s' out of a given, larger set s. The members of s can be first independently trained (by minimising their individual MSE); then, the subset s' exhibiting the highest ambiguity should be selected. However, all works on regressor ensemble selection we are aware of did not use this approach, but relied on the direct estimate of the ensemble MSE [43, 30, 14, 39, 27, 34, 38], with the only exception of [11]. However, in [11] diversity measures inspired by the ones defined for classification problems were used, instead of the clean ambiguity term. The above overproduce and choose strategy, implemented by maximising some diversity measure, is used by several classifier ensemble construction techniques, instead. We will further discuss this point in Sect. 4.

3 Decompositions of the Ensemble Classification Error

Several BV decompositions of classification error have been proposed, e.g., [3, 17, 10], and have been used to empirically investigate the variance (and sometimes bias) reduction effect of classifier combination techniques. However, no decomposition analogue to BVC (i.e., explicitly including the outputs of individual classifiers) has been derived yet. This is not straightforward, for instance because the concept of covariance is undefined for categorical outputs (class labels), as pointed out in [4]. Similarly, no ambiguity-like decomposition (i.e., including the average error of individual classifiers) has been derived for MCSs, with the only exception of the one of [7] for two-class problems. Accordingly, existing diversity measures have not been derived from exact decompositions of classification error. However, we point out that they have been empirically and theoretically analysed by investigating whether and how their trade-off with the average error of individual classifiers is related to the ensemble error. In other words, they have been (often implicitly) considered as the equivalent of the ambiguity term in the corresponding decomposition (6).

In the following we show how an analogue of the BVC decomposition can be derived, as well as the analogue of the ambiguity decomposition, which generalises the one of [7] to any number of classes.

[3] NCL was actually defined “heuristically” in [25], with no reference to the ambiguity decomposition. The strong relationship between NCL and the ambiguity decomposition was pointed out and thoroughly analysed in [5].

3.1 A Bias-Variance-Covariance Decomposition for Classifier Ensembles

We consider the Kohavi-Wolpert BV decomposition of classification error (0/1-loss), for a L-class problem [17]. We denote class labels by $y_1, \ldots, y_L$. To further simplify the notation, we define: $P[y_i] = P[Y = y_i|\mathbf{x}]$, and $\hat{P}[y_i] = P_D[f(\mathbf{x}; D) = y_i]$. The loss of a classifier $f(\mathbf{x}; d)$ on a given sample $(\mathbf{x}, y)$, which we denote by $e(\mathbf{x}, y; d)$, equals $I[f(\mathbf{x}; d) \neq y]$, where $I[a] = 1$ (0), if $a =$ True (False). Bias and variance are defined in [17] as follows:

$$bias_f = \frac{1}{2} \sum_{y_i} \left(P[y_i] - \hat{P}[y_i] \right)^2 , \quad var_f = \frac{1}{2} \left(1 - \sum_{y_i} \hat{P}[y_i]^2 \right) . \tag{7}$$

It follows that [17]:

$$E_{D,Y|\mathbf{x}}[e(\mathbf{x}, Y; D)] = bias_f + var_f + noise \ , \tag{8}$$

where $noise = \frac{1}{2}\left(1 - \sum_{y_i} P[y_i]^2\right)$.

We now rewrite the above bias and variance terms for a MCS $\{f_1, \ldots, f_N\}$. We denote by f_{ens} the ensemble output, with no restriction on the combining rule. Adding and subtracting to the expressions of $bias_{f_{\text{ens}}}$ and $var_{f_{\text{ens}}}$ the two terms indicated below, after some manipulations we obtain:

$$\begin{aligned} bias_{f_{\text{ens}}} &= \tfrac{1}{2} \textstyle\sum_{y_i} \left(P[y_i] - \tfrac{1}{\sqrt{N}} \sum_j \hat{P}_j[y_i] + \tfrac{1}{\sqrt{N}} \sum_j \hat{P}_j[y_i] - \hat{P}_{\text{ens}}[y_i] \right)^2 \\ &= \overline{bias} + b, \\ var_{f_{\text{ens}}} &= \tfrac{1}{2} \left(1 - \tfrac{1}{N^2} \textstyle\sum_{j,y_i} \hat{P}_j^2[y_i] + \tfrac{1}{N^2} \sum_{j,y_i} \hat{P}_j^2[y_i] - \sum_{y_i} \hat{P}_{\text{ens}}[y_i]^2 \right) \\ &= \tfrac{1}{N}\overline{var} + v \ , \end{aligned} \tag{9}$$

where $\overline{bias} = \frac{1}{N} \sum_j bias_{f_j}$, $\overline{var} = \frac{1}{N} \sum_j var_{f_j}$, and the terms b and v are given in the online appendix of this paper[4] (they are not reported here due to lack of space). This easily leads us to the analogue of the BVC decomposition (5):

$$E_{D,Y|\mathbf{x}} e_{\text{ens}}(\mathbf{x}, Y; D) = \overline{bias} + \frac{1}{N}\overline{var} + b + v + noise \ . \tag{10}$$

The term $b + v$ corresponds to the covariance term of (5), and its interpretation is under analysis at the time of submitting the camera-ready of this paper.

3.2 Ambiguity-like Decomposition for Classifier Ensembles

The only decomposition of the classification error (0/1-loss) of an ensemble, analogue to the ambiguity decomposition, has been derived in [7], for two-class problems. Denoting the class labels by the values $\{-1, +1\}$, the loss of a classifier

[4] http://prag.diee.unica.it/pra/bib/didaciMCS2013

on a sample $(\mathbf{x}, y)$ can be expressed as $e_f(\mathbf{x}, y; d) = \frac{1}{2}(1 - y \times f)$. Denoting by $\overline{e}(\mathbf{x}, y; d)$ the average loss of an ensemble of N classifiers, it follows that [7]:

$$e_{\text{ens}}(\mathbf{x}, y; d) = \overline{e}(\mathbf{x}, y; d) - y \times f_{\text{ens}} \times \frac{1}{N} \sum_{j=1}^{N} \delta_j(\mathbf{x}, y; d) \ , \tag{11}$$

where $\delta_j(\mathbf{x}, y; d) = \frac{1}{2}(1 - f_j \times f_{\text{ens}})$. This term is a measure of the disagreement between classifier f_j and the ensemble. The decomposition (11) appears thus very similar to the ambiguity decomposition (6). However, the second term in the right-hand side (RHS) of (11) also includes the true class label y, contrary to the ambiguity term in (6). The interpretation of decomposition (11) is very clear: it shows that a lower average accuracy of individual classifiers can be compensated by a higher disagreement with the ensemble, as far as the ensemble remains correct. This latter condition is due to the presence of the y term in the RHS of (11). We point out that decomposition (11) is valid for any combining rule, although in [7] only majority voting was considered.

Here we show that a more general decomposition can be obtained, for any number of classes. To this aim, we can exploit the BVC-like decomposition (10). We denote the expected average misclassification probability of individual classifiers on a point $\mathbf{x}$, $E_{D,Y|\mathbf{x}}\left[\frac{1}{N}\sum_j e_j(\mathbf{x}, y; d)\right]$, by $\overline{e}(\mathbf{x})$. It is easy to see that $\overline{e}(\mathbf{x}) = \overline{bias} + \overline{var} + noise$. Rewriting (10) by adding and subtracting the term $\frac{N-1}{N}\overline{var}$, after some manipulations we obtain:

$$\begin{aligned} e_{\text{ens}}(\mathbf{x}) &= \overline{e}(\mathbf{x}) - \textstyle\sum_{y_i} P[y_i] \frac{1}{N} \sum_j \left(\hat{P}_{\text{ens}}[y_i] - \hat{P}_j[y_i]\right) \\ &= \overline{e}(\mathbf{x}) - \frac{1}{N} \textstyle\sum_j \left(P_{D,Y|\mathbf{x}}[f_{\text{ens}} = Y|\mathbf{x}] - P_{D,Y|\mathbf{x}}[f_j = Y|\mathbf{x}]\right) \ . \end{aligned} \tag{12}$$

The same result can also be obtained by directly computing $E_{D,Y|\mathbf{x}}[e_{ens}(\mathbf{X}, Y; D) - \overline{e}(\mathbf{X}, Y; D)]$, which is the approach followed in [7]. Obviously, for $L = 2$, the expectation of (11) with respect to $D, Y|\mathbf{x}$ equals (12).

We can further rewrite decomposition (12) in the case of a fixed training set d, i.e., by taking the expectation of $e_{\text{ens}}(\mathbf{x}, y; d)$ with respect of $P[Y|\mathbf{x}]$ only:

$$e_{\text{ens}}(\mathbf{x}; d) = \overline{e}(\mathbf{x}; d) - \frac{1}{N} \sum_j \left(P_{Y|\mathbf{x}}[f_{\text{ens}} = Y|\mathbf{x}] - P_{Y|\mathbf{x}}[f_j = Y|\mathbf{x}]\right) \ . \tag{13}$$

In particular, for a single sample $(\mathbf{x}, y)$ and a single training set d, we obtain the generalisation of (11) for $L > 2$:

$$e_{f_{\text{ens}}}(\mathbf{x}, y; d) = \overline{e}(\mathbf{x}, y; d) - \frac{1}{N} \sum_j \left(I[f_{\text{ens}} = y] - I[f_j = y]\right) \ . \tag{14}$$

Expressions (12)–(14) are thus three different versions of a general ambiguity-like decomposition of the ensemble error, that is valid for any number of classes and any combining rule.[5] Comparing (12) to the BVC-like decomposition (10), to understand the correspondence between their terms (as done in [4] for regression error), is the subject of our ongoing work.

[5] This decomposition can also be easily extended to any loss function.

4 Discussion

Let us recall the second and third issues mentioned in the introduction. Can the second term of decomposition (12)–(14) be interpreted as a diversity measure? Can it be practically exploited for ensemble construction? In particular, is it more effective than existing diversity measures, and than the direct estimate of the ensemble error, e.g., in terms of estimation reliability, computational complexity, or the possibility of estimating it using unlabelled samples only? We address these issues in the following.

4.1 Interpretation of the ambiguity-like decomposition

For two-class problems, the second term in the RHS of decomposition (11) can be interpreted as a diversity measure, in terms of the disagreement between the individual classifiers and the ensemble [7], similarly to the ambiguity term in (6). In the general case when $L > 2$, it is easy to see that a similar interpretation can be given for the second term in the RHS of decomposition (14). However, for $L > 2$ the disagreement is not expressed in terms of the class labels, but in terms of the *correctness* of such choices (they coincide only when $L = 2$).

In [7] the decomposition (11) was further analysed by considering the case of zero Bayes error (i.e., when y is a deterministic function of $\mathbf{x}$), and by taking the expectation of (11) over $P[\mathbf{X}]$, which gives the error probability of the ensemble.[6] Taking into account that $y \times f_{\text{ens}} = +1$ (-1) when the ensemble is correct (wrong), and denoting by $\mathbf{x}+$ and $\mathbf{x}-$ the corresponding regions in feature space, one obtains [7]:

$$e_{\text{ens}}(d) = \overline{e}(d) - \int_{\mathbf{x}+} \frac{1}{N} \sum_{j=1}^{N} \delta_j(\mathbf{x}; d) \mathrm{d}\mathbf{x} + \int_{\mathbf{x}-} \frac{1}{N} \sum_{j=1}^{N} \delta_j(\mathbf{x}; d) \mathrm{d}\mathbf{x} \ . \tag{15}$$

This highlights that increasing the disagreement is beneficial on samples where the ensemble is correct, while it is detrimental on samples where the ensemble is wrong. Accordingly, the corresponding diversity components were named respectively “good” and “bad” diversity in [7].

It is now easy to see that the same interpretation can be given when $L > 2$, from decomposition (14), provided that “disagreement” is intended as explained above. On samples where the ensemble is correct, increasing the disagreement is beneficial, i.e., the highest number of individual classifiers should *misclassify* such samples, *independently* on the specific class labels they choose. Increasing the disagreement is detrimental on samples misclassified by the ensemble, instead: this means as well that the highest number of individual classifiers should *misclassify* such samples. Accordingly, the concept of good and bad diversity can be extended to $L > 2$, by considering the above definition of disagreement.

[6] This analysis can be easily extended to the case of non-zero Bayes error.

4.2 Practical exploitation of diversity measures

Here we discuss the practical usefulness of diversity measures, including the diversity term of the ambiguity-like decomposition (14), for ensemble construction.

We first point out that the diversity term in (14) depends on the specific combining rule. This is a consequence of the fact that the error of a given ensemble depends on the combining rule, and that the first term in the RHS of (14) is the average error of ensemble members. However, existing diversity measures do not depend on the combining rule. Actually, although they are not explicitly tailored to a specific rule, most of them seem related to majority voting [35]. Even interesting measures recently proposed, using information theory, do not take into account the combining rule [6, 44]. The pros and cons of using a single diversity measure for all combining rules, and of using different measures tailored to specific rules, have been discussed in [21] (sect. 10.5): "The problem is that the 'clean' diversity measure might be of little use due to its weak relationship with the ensemble accuracy [...]. On the other hand, the more we involve the ensemble performance into defining diversity, the more we are running onto the risk of trying to replace a simple calculation of the ensemble error by a clumsy estimate that we call diversity.'

On the other hand, as pointed out in Sect. 3, existing measures are usually considered as the equivalent of the ambiguity term in regression. Indeed, they have often been analysed by investigating whether and how their trade-off with the average error of individual classifiers is related to the ensemble error, but no clear correlation has been found [32, 20, 35]. This raised some doubts about the usefulness of existing measures for ensemble construction. Some authors even argued that a direct estimation of ensemble accuracy can be more effective. For instance, see the quote from [32], reported in Sect. 1; and: "In our opinion, the existing diversity measures are [...] not [sufficient] for [selecting base classifiers]" [35] (sect. 4). Such doubts are strengthened by the following fact: overproduce and choose methods for ensemble construction, that make explicit use of diversity measures, did not provide evidence that such an approach is more effective than directly estimating ensemble accuracy [41, 13, 1, 2, 16, 40]. In particular, besides [41, 40], where such a comparison has not been made, in [13, 1, 2] the use of diversity measures did not provide any significant accuracy improvement, and in [16] the direct estimation of classifier accuracy turned out to significantly outperform the use of diversity measures.

Consider now the *exact* decomposition of the ensemble error (14), or the equivalent (for two-class problems) decomposition (11). Can their diversity terms be exploited more effectively in the context of overproduce and choose methods? At least at a first glance, the answer seems negative. The reason is that to compute these diversity terms (on a given set of samples, e.g., a validation set) one needs to know both the ensemble output and the correct class label of each sample. However, this also allows one to directly estimate the ensemble accuracy. We point out that a similar issue arises about the use of the ambiguity decomposition for regression problems as well, as mentioned in Sect. 2. However, even though computing the ambiguity term in (6) is not computationally cheaper than directly computing the ensemble MSE, the former does not involve the correct output y, which allows one to compute it using also *unlabelled* samples. It would be interesting to investigate whether this can be actually advantageous. It is worth noting that the use of unlabelled samples to promote diversity in MCSs has been suggested in [42].

Consider now the use of diversity measures for ensemble construction strategies analogue to NCL, i.e., for directly constructing a MCS without overproducing first, and then selecting a subset of classifiers. In this context, it is pertinent to note that well-known MCS construction techniques like bagging, random forests, random subspace, and AdaBoost, are effective even though they do not explicitly use any diversity measure (see, e.g., [21], chapter 10). By the way, they are all tailored to majority voting (or weighted voting, in the case of AdaBoost) [35]. On the one hand, it is commonly believed that such techniques "can be interpreted as building diverse base classifiers implicitly" [35]. This fact has also inspired the idea of investigating what objective function, and thus, what diversity measure, is implicitly optimised by such techniques.[7] On the other hand, the above discussion about existing measures and about the diversity terms in (14) and (11), strengthens the doubt that they are not more useful in practice than directly estimating ensemble accuracy. Indeed, they seem only "descriptive", i.e., they formalise the intuition that (at least for the majority voting rule) an effective ensemble is made up of classifiers that are accurate "enough" on different regions of the feature space, such that (ideally) a majority of them correctly classifies each sample. This is exactly the goal that techniques like bagging pursue, using different strategies, without explicitly relying on diversity measures. To our knowledge, the only method analogue to NCL proposed so far (besides the direct use of NCL with base classifiers like neural networks) is the one of [45]. It simultaneously trains a set of two-class linear classifiers, and computes the weights of their linear combination, using a SVM-like learning algorithm. The objective function aims at jointly maximising individual accuracy and diversity. Diversity is measured as the average pairwise disagreement between individual classifiers. This method exhibited comparative performance with bagging and AdaBoost. On the other hand, the considered diversity measure does not coincide with the ambiguity term (11). A further investigation of this method is thus interesting.

To sum up, existing diversity measures are at most an approximation of the "real" diversity term, in the context of exact decompositions of the ensemble error like (14) and (11), in which the first term is given by the average error of ensemble members. On the other hand, the practical usefulness of diversity measures, even exact ones, remains questionable. In the next section we will indicate possible research directions to address this issue.

[7] Zhi-Hua Zhou, MCS 2010 panel discussion: `http://www.diee.unica.it/mcs/mcs2010/paneldiscussion.html`

5 Suggestions for future research

On the basis of the above results and discussion, we conclude this paper by suggesting some research directions, aimed at better understanding whether the explicit use of diversity measures can be useful in practice, for ensemble construction.

1. The BVC-like decomposition (10), and in particular the term corresponding to covariance, deserves further analysis. A comparison with the ambiguity-like decomposition (12) is interesting, to understand the correspondence between their terms, as done in [4] for the BVC and ambiguity decompositions of regression error.
2. The ambiguity-like decomposition (12) should be extended to loss functions different than 0/1. It should also be further analysed with respect to specific combining rules, different from majority voting (which has been considered in [7]). In particular, it would be interesting to investigate whether the average error of individual classifiers is the most suitable as the first term of such a decomposition, for any combining rule.
3. The effectiveness of explicitly using diversity measures in ensemble construction methods with the overproduce and choose strategy, should be thoroughly compared with the direct estimation of ensemble accuracy. This should be done also for regression problems, where the ambiguity term seems in principle more advantageous than the corresponding one for classification problems.
4. It is also interesting to compare the diversity terms derived from exact decompositions of the ensemble error like (14) and (11), with existing diversity measures. This can help understanding which of these measures is a better approximation to the "real" diversity, also with respect to a specific combining rule. This could even suggest new diversity measures that better approximate the "real" one, and are also of practical use.

Acknowledgments. The authors thank the reviewers for their constructive comments and criticisms to the first version of this paper. This work has been partly supported by the project CRP-18293 funded by Regione Autonoma della Sardegna, L.R. 7/2007, Bando 2009.

References

1. Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: A new ensemble diversity measure applied to thinning ensembles. In: 4th Int. Workshop on Multiple Classifier Systems, pp. 306–316, Springer, 2003.
2. Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: Ensemble diversity measures and their application to thinning. Information Fusion 6, 49–62 (2005)
3. Breiman, L.: Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California, Berkeley, CA (1996)

4. Brown, G., Wyatt, J.L., Harris, R., Yao, X.: Diversity creation methods: a survey and categorisation. Information Fusion 6, 5–20 (2005)
5. Brown, G., Wyatt, J.L., Tino, P.: Managing diversity in regression ensembles. Journal of Machine Learning Research 6, 1621–1650 (2005)
6. Brown, G.: An information theoretic perspective on multiple classifier systems. In: 8th Int. Workshop on Multiple Classifier Systems, pp. 344–353, Springer (2009)
7. Brown, G., Kuncheva, L.I.: GOOD and BAD diversity in majority vote ensembles. In: 9th Int. Workshop on Multiple Classifier Systems, pp. 124–133, Springer (2010)
8. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning 40, 139–157 (2000)
9. Dietterich, T.G.: Ensemble methods in machine learning. In: 1st Int. Workshop on Multiple Classifier Systems, pp. 1–15, Springer (2000)
10. Domingos, P.: A unified bias-variance decomposition for zero-one and squared loss. In: 7th Int. Conf. on Artificial Intelligence, pp. 564–569 (2000)
11. Dutta, H.: Measuring Diversity in Regression Ensembles. In: 4th Indian Int. Conf. on Artificial Intelligence, pp. 2220–2236 (2009)
12. Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural computation 4, 1–58 (1992)
13. Giacinto, G., Roli, F.: Design of effective neural network ensembles for image classification purposes. Image and Vision Computing 19, 699–707 (2001)
14. Hernandez-Lobato, D., Martinez-Munoz, G., Suarez, A.: Pruning in ordered regression bagging ensembles. In: Int. Joint Conf. Neural Net., pp. 1266–1273 (2006)
15. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 226–239 (1998)
16. Ko, A.H.-R., Sabourin, R., DeSouza Britto Jr., A.: Compound diversity functions for ensemble selection, Int. J. Patt. Rec. Artificial Intelligence 23, 659–686 (2009)
17. Kohavi, R., Wolpert, D.H.: Bias plus variance decomposition for zero-one loss functions. In: 13th Int. Conf. Mac. Learn., pp. 275–283, Morgan Kaufmann (1996)
18. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In: Adv. in Neural Inf. Proc. Systems 7, pp. 231–238, MIT Press (1995)
19. Kuncheva, L.I.: That elusive diversity in classifier ensembles. In: 1st Iberian Conf. on Patt. Rec. and Image Analysis, pp. 1126–1138, Springer (2003)
20. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mac. Learn. 51, 181–207 (2003)
21. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, Hoboken, NJ (2004)
22. Lam, L., Suen, C.Y.: Application of majority voting to pattern recognition: an anal- ysis of its behavior and performance. IEEE Transactions on Systems, Man, and Cybernetics - Part C 27, 553–568 (1997)
23. Li, N., Yu, Y., Zhou, Z.-H.: Diversity Regularized Ensemble Pruning. In: Europ. Conf. Mach. Learn. Knowledge Discovery in Databases (part I), pp. 330–345, Springer (2012)
24. Littlewood, B., Miller, D.R.: Conceptual modeling of coincident failures in multiversion software. IEEE Transactions on Software Engineering 15, 1596–1614 (1989)
25. Liu, Y.: Negative Correlation Learning and Evolutionary Neural Network Ensembles. PhD thesis, University College, The University of New South Wales, Australian Defence Force Academy, Canberra, Australia (1998)
26. Margineantu, D.D., Dietterich, T.G.: Pruning adaptive boosting. In: 14th Int. Conf. on Machine Learning, pp. 211–218 (1997)

27. Partalas, I., Tsoumakas, G., Hatzikos, E.V., Vlahavas, I.P.,: Greedy regression ensemble selection: Theory and an application to water quality prediction. Information Sciences 178, 3867–3879 (2008)
28. Partridge, D., Krzanowski, W.J.: Software diversity: practical statistics for its measurement and exploitation. Information & Software Technology 39, 707–717 (1997)
29. Perrone, M.P., Cooper, L.N.: When networks disagree: Ensemble methods for neural networks. In: Mammone, R.J. (ed.) Artificial Neural Networks for Spech and Vision, pp. 126–142. Chapman & Hall, New York (1993)
30. Rooney, N., Patterson, D., Nugent, C.: Reduced ensemble size stacking. In: 16th Int. Conf. on Tools with Artificial Intelligence, pp. 266–271 (2004)
31. Sharkey. A.J.C., Sharkey, N.E., Combining diverse neural nets. The Knowledge Engineering Review 12, 231–247 (1997)
32. Shipp, C.A., Kuncheva, L.I.: Relationships between combination methods and measures of diversity in combining classifiers. Information Fusion 3, 135–148 (2002)
33. Sirlantzis, K., Hoque, S., Fairhurst, M.C.: Diversity in multiple classifier ensembles based on binary feature quantisation with application to face recognition. Applied Soft Computing 8, 437-445 (2008)
34. Sun, Q., Pfahringer, B.: Bagging Ensemble Selection for Regression. In: Australasian Conference on Artificial Intelligence, pp. 695-706, Springer (2012)
35. Tang, E.K., Suganthan, P.N., Yao, X.: An analysis of diversity measures. Machine Learning 65, 247–271 (2006)
36. Tumer, K., Ghosh, J.: Analysis of decision boundaries in linearly combined neural classifiers. Pattern Recognition 29, 341–348 (1996)
37. Ueda, N., Nakano. R.: Generalization error of ensemble estimators. In: Int. Conf. Neural Networks, pp. 90–95 (1996)
38. Wang, D., Alhamdoosh, M.: Evolutionary extreme learning machine ensembles with size control. Neurocomputing 102, 98–110 (2013)
39. Yu, Y., Zhou, Z.-H., Ting, K.M.: Ting: Cocktail Ensemble for Regression. In: 7th Int. Conf. Data Mining, pp. 721–726, IEEE Computer Society (2007)
40. Yu, Y., Li, Y.-F., Zhou, Z.-H.: Diversity regularized machine. In: 22nd Int. Joint Conf. on Artificial Intelligence, pp. 1603–1608 (2011)
41. Zenobi G., Cunningham P.: Using Diversity in Preparing Ensembles of Classifiers Bases on Different Feature Subsets to Minimize Generalization Error. In: European Conf. on Machine Learning (2001)
42. Zhang, M.-L., Zhou, Z.-H.: Exploiting unlabeled data to enhance ensemble diversity. Data Min. Knowl. Disc. 26, 98–129 (2013)
43. Zhou, Z.-H., Wu, J., Tang, W.: Ensembling neural networks: many could be better than all. Artificial Intelligence 137, 239–263 (2002)
44. Zhou, Z.-H., Li, N.: Multi-Information Ensemble Diversity. In: 9th Int. Workshop on Multiple Classifier Systems, pp. 134–144, Springer (2010)
45. Yu, Y., Li, Y.-F., Zhou, Z.-H.: Diversity regularized machine. In: Proc. 22nd Int. Joint Conf. on Artificial Intelligence, pp. 1603–1608 (2011)
46. Zhou, Z.-H.: Introduction to Ensemble Methods. CRC Press, 2012.