Between Two Extremes: Examining Decompositions of the Ensemble Objective Function

Gavin Brown[1], Jeremy Wyatt[2], and Ping Sun[2]

[1] School of Computer Science, University of Manchester,
Kilburn Building, Oxford Road, Manchester, M13 9PL
gavin.brown@manchester.ac.uk
http://www.cs.man.ac.uk/~gbrown/

[2] School of Computer Science, University of Birmingham,
Edgbaston Park Road, Birmingham, B15 2TT
{j.l.wyatt,p.sun}@cs.bham.ac.uk
http://www.cs.bham.ac.uk/~jlw/

Abstract. We study how the error of an ensemble regression estimator can be decomposed into two components: one accounting for the individual errors and the other accounting for the correlations within the ensemble. This is the well known Ambiguity decomposition; we show an alternative way to decompose the error, and show how both decompositions have been exploited in a learning scheme. Using a scaling parameter in the decomposition we can blend the gradient (and therefore the learning process) smoothly between two extremes, from concentrating on individual accuracies and ignoring diversity, up to a full non-linear optimization of all parameters, treating the ensemble as a single learning unit. We demonstrate how this also applies to ensembles using a soft combination of posterior probability estimates, so can be utilised for classifier ensembles.

1 Introduction

It is well recognised that for best performance, an ensemble of estimators should exhibit some kind of disagreement on certain datapoints. When estimators produce class labels and are combined by a majority vote, this is the often cited, but little understood notion of "diversity". When in a regression framework, using estimators combined by a simple averaging operation, the notion of disagreement between estimators is rigorously defined: with a single estimator, we have the well known bias-variance trade-off [5], and with an ensemble of estimators, we have a bias-variance-covariance trade-off [10]. The regression diversity issue can now be understood quite simply: in a single estimator we have a two way trade-off, and in a regression ensemble the optimal "diversity" is that which optimally balances the bias-variance-covariance *three way* trade-off.

The understanding of regression ensembles is therefore quite mature. The understanding of classification ensembles, using a majority vote combiner, is substantially less well developed; see [1] for a recent survey of the field. However, this lack of understanding can be partially side-stepped considering that a classification problem can be reformulated as a regression problem by approximating the class posterior probabilities. In this case the theory is more well developed—Fumera and Roli [4] represent the current state of the art and Kuncheva [8] provides an excellent overview of the area.

Though the bias-variance-covariance decomposition fully quantifies the regression diversity issue, a more well known result is the *Ambiguity decomposition* [7]. This paper will summarise recent work explaining the link between these decompositions, and show an alternative new decomposition. We will see how these decompositions can be exploited in a learning procedure.

The structure of this paper is as follows. in section 2 we show two ways in which the ensemble error can be decomposed: one that has been seen in the literature before, and one new to the literature. In section 3 we illustrate how these can be used in a learning procedure. In section 4 we present a theoretical analysis of the new decomposition and learning procedure. In section 5 we show emprical results and finally conclude with observations on possible future work.

2 Decomposing the Ensemble Objective Function

2.1 The Ambiguity Decomposition

Let us assume that the ensemble combination function is a mean of the M ensemble member outputs, that is $\bar{f} = \frac{1}{M}\sum_i f_i$. For convenience of notation we have omitted dependence on any particular input $\mathbf{x}$; it can be assumed f_i is the output of estimator i for a single arbitary input. The quadratic loss of this estimator from its target d is:

$$e_{ens} = (\bar{f} - d)^2 \tag{1}$$

The Ambiguity Decomposition [7], states that the ensemble error is *guaranteed to be lower than the average individual error at an arbitrary datapoint.* Formally, this is:

$$\begin{aligned} e_{ens} = (\bar{f} - d)^2 &= \frac{1}{M}\sum_i (f_i - d)^2 - \frac{1}{M}\sum_i (f_i - \bar{f})^2 \\ &= \qquad\quad \overline{e} \qquad\quad - \qquad\quad \overline{a} \end{aligned} \tag{2}$$

This illustrates that the correlations between the members is a fundamental part of the quadratic error. If we take the expected value of eq (1) with respect to all possible training sets of fixed size, then we have the *mean squared error*. The *bias-variance-covariance decomposition* [10] shows that this can be broken down into three components:

$$E\{(\bar{f} - d)^2\} = \overline{bias}^2 + \frac{1}{M}\overline{var} + \left(1 - \frac{1}{M}\right)\overline{covar} \tag{3}$$

The relationship between these two decompositions has been found [2] to be:

$$E\{\frac{1}{M}\sum_i (f_i - d)^2\} = \overline{bias}^2 + \varOmega \tag{4}$$

$$E\{\frac{1}{M}\sum_i (f_i - \bar{f})^2\} = \varOmega - \frac{1}{M}\overline{var} - \left(1 - \frac{1}{M}\right)\overline{covar} \tag{5}$$

where the interaction between the expected average error and the expected Ambiguity is the $\varOmega$ term:

$$\begin{aligned}\varOmega &= \frac{1}{M}\sum_i E\{(f_i - E\{\bar{f}\})^2\} \\ &= \overline{var} + \frac{1}{M}\sum_i (E\{f_i\} - E\{\bar{f}\})^2 \end{aligned} \tag{6}$$

The Ambiguity decomposition is useful for a number of reasons, the primary one being that since the Ambiguity is target-independent, it provides a way to estimate generalisation error of an ensemble from *unlabelled data* [7].

2.2 An Alternative Decomposition

We note that the ensemble objective can also be decomposed as so:

$$\begin{aligned} e_{ens} &= (\bar{f} - d)^2 \\ &= (\frac{1}{M}\sum_i f_i - d)^2 \\ &= \frac{1}{M^2}\sum_i \Big[(f_i - d)\sum_j (f_j - d)\Big] \\ &= \frac{1}{M^2}\sum_i (f_i - d)^2 + \frac{1}{M^2}\sum_i (f_i - d)\sum_{j\neq i}(f_j - d) \end{aligned} \tag{7}$$

Regarding now eq (4), a simple deduction can be made:

$$E\{\frac{1}{M^2}\sum_i (f_i - d)^2\} = \frac{1}{M}\Big[\overline{bias}^2 + \varOmega\Big] \tag{8}$$

and therefore:

$$\begin{aligned} &E\{\frac{1}{M^2}\sum_i (f_i - d)\sum_{j\neq i}(f_j - d)\} \\ &\qquad = \left(1 - \frac{1}{M}\right)\Big[\overline{bias}^2 + \varOmega\Big] - \varOmega + \frac{1}{M}\overline{var} + \left(1 - \frac{1}{M}\right)\overline{covar} \end{aligned} \tag{9}$$

It can be seen that through this new decomposition, part of the $\overline{bias}^2$ and Ω has "moved over" to the second term of the decomposition. This new decompoistion obviously does not share the useful target-independence property of the Ambiguity decomposition—however, the decompositions do share another property which makes them interesting to exploit in a learning scheme. This will be the focus of the next section.

3 Using the Decompositions for Learning

3.1 A Useful Property of the Decompositions

If we were to train a *single* estimator, the error and associated gradient are:

$$e_i = (f_i - d)^2 \qquad \frac{\partial e_i}{\partial f_i} = 2(f_i - d) \tag{10}$$

In the previous section we described two ways to decompose the ensemble error in (1) into *two* additive components. The decompositions share the property that the first component is directly proportional to the error gradient of a single estimator. By this we mean that if we calculate the gradient of the first term on the right hand side of eq (2), we have:

$$\frac{\partial \frac{1}{M}\sum_i (f_i - d)^2}{\partial f_i} = \frac{2}{M}(f_i - d) = \frac{1}{M}\frac{\partial e_i}{\partial f_i} \tag{11}$$

and the same applies to the new decomposition we have proposed, in eq (7), we have:

$$\frac{\partial \frac{1}{M^2}\sum_i (f_i - d)^2}{\partial f_i} = \frac{2}{M^2}(f_i - d) = \frac{1}{M^2}\frac{\partial e_i}{\partial f_i} \tag{12}$$

Imagine now that we take one of the decompositions as our error function to minimise, and place a scaling parameter in front of its second component. So for the Ambiguity decomposition we would have:

$$e_{amb} = \frac{1}{M}\sum_i (f_i - d)^2 - \gamma \frac{1}{M}\sum_i (f_i - \bar{f})^2 \tag{13}$$

where the γ is our scaling parameter. If we set $\gamma = 0$, the search landscape will be exactly equivalent (aside of a constant $\frac{1}{M}$ factor) to that of a single estimator, meaning all the minima will be in the same locations. If we set $\gamma = 1$ such that the two components are balanced, this will be equivalent to training the ensemble as a single unit, albeit a very complicated unit. This allows us to test a simple hypothesis–*"on a given dataset, which is better: a single complex machine, or an ensemble of separately trained simpler machines?"*. With γ, we blend between the extremes—sometimes a simple ensemble will have best performance ($\gamma = 0$), and sometimes a single, complex machine ($\gamma = 1$) will prevail. This could also be interpreted as varying the *fit of the model* from few up to many degrees of freedom.

3.2 An Existing Training Scheme that Exploits the Property

Negative Correlation (NC) Learning [9] adds penalty terms of a particular form to the error function of an individual estimator. In the original heuristic formulation of NC, the penalty term existed in several different forms:

$$e_i^{(1)} = \frac{1}{2}(f_i - d)^2 + \gamma(f_i - \bar{f})\sum_{j \neq i}(f_j - \bar{f}) \tag{14}$$

$$e_i^{(2)} = \frac{1}{2}(f_i - d)^2 + \gamma(f_i - d)\sum_{j \neq i}(f_j - d) \tag{15}$$

$$e_i^{(3)} = \frac{1}{2}(f_i - d)^2 + \gamma(f_i - 0.5)\sum_{j \neq i}(f_j - 0.5) \tag{16}$$

The training scheme is implemented as follows:

1. Let M be the final number of predictors required.
2. Take a training set $z = \{(\mathbf{x}_1, d_1), ..., (\mathbf{x}_N, d_N)\}$.
3. For each training pattern in z from $n = 1$ to N do :
   (a) Calculate $\bar{f} = \frac{1}{M}\sum_i f_i(\mathbf{x}_n)$
   (b) For each estimator from $i = 1$ to M,
   perform a *single* update for each weight w in estimator i according to one of error functions (14), (15), or (16).
4. Repeat from step 3 for a desired number of epochs.

◇

After training, for any new testing point the output of the ensemble is given by the simple average combination. The γ parameter controls a trade-off between the objective and penalty terms. With $\gamma = 0$ we would have an ensemble with each estimator training with plain gradient descent exactly equivalent to training a set of estimators independently of one another. If γ is increased, more and more emphasis would be placed on the correlations by minimising the penalty.

The first form in (14) has been thoroughly investigated in a regresssion setting [2]. This can be summarised by noting that $(f_i - \bar{f}) = -\sum_{j \neq i}(f_j - \bar{f})$, and therefore:

$$\begin{aligned} e_i^{(1)} &= \frac{1}{2}(f_i - d)^2 + \gamma(f_i - \bar{f})\sum_{j \neq i}(f_j - \bar{f}) \\ &= \frac{1}{2}(f_i - d)^2 - \gamma(f_i - \bar{f})^2 \end{aligned} \tag{17}$$

The similarity to the Ambiguity decomposition can be seen immediately. This link to the Ambiguity and Bias-Variance-Covariance decompositions allowed a solid grounding, as well as a proven upper bound on the γ penalty coefficient entirely *independent* of all parameters except M, the size of the ensemble [2]. In benchmarks against other ensemble techniques such as Mixtures of Experts, Adaboost.R1 and Bagging, it was found to be a competitive technique. The remaining two forms were not as well understood. However, noting our new decomposition, a slight rearrangement shows:

$$e_{new} = \frac{1}{M^2} \sum_i \left[(f_i - d)^2 + (f_i - d) \sum_{j \neq i} (f_j - d) \right] \tag{18}$$

An immediate similarity to (15) can be seen, though the exact relationship is not yet clear. This form of NC was found [9] to sometimes be more successful on classification problems, than the original penalty. In the next section we proceed to analyze the decomposition, in an attempt to identify why this penalty may have outperformed the original.

4 Gradient Analysis of the New Decomposition

We have seen that the quadratic error of the simple average ensemble estimator can be decomposed in two ways: the Ambiguity decomposition [7] and our new decomposition in eq (7). Now that we know the ensemble error can be viewed in a composite form (1) and two decomposed forms, we can calculate the gradients either way. The composite form gives a simple result of:

$$\frac{\partial e_{ens}}{\partial f_i} = (\bar{f} - d) \cdot \frac{1}{M} \tag{19}$$

Performing this instead starting from our new decomposed form shows more interesting results. Before we begin the calculation, we first break it into two components, where the first term concerns estimator i, and the second concerns all the other estimators $k \neq i$:

$$\begin{aligned} e_{ens} = &\frac{1}{M^2} \left[\frac{1}{2}(f_i - d)^2 - \frac{1}{2}(f_i - d) \sum_{j \neq i} (f_j - d) \right] \\ &+ \frac{1}{M^2} \sum_{k \neq i} \left[\frac{1}{2}(f_k - d)^2 - \frac{1}{2}(f_k - d) \sum_{j \neq k} (f_j - d) \right] \end{aligned}$$

It should be of course noted that we have multiplied through by a constant $\frac{1}{2}$, as is usual in gradient descent training of MLPs, which were the estimator used in this paper, but the analysis applies in general to any estimator. The partial derivative of this result with respect to f_i is:

$$\begin{aligned} \frac{\partial e_{ens}}{\partial f_i} = &\frac{1}{M^2} \left[(f_i - d) + \frac{1}{2} \sum_{j \neq i} (f_j - d) \right] \\ &+ \frac{1}{M^2} \sum_{k \neq i} \left[\frac{1}{2}(f_k - d) \right] \end{aligned}$$

Or rearranged:

$$\frac{\partial e_{ens}}{\partial f_i} = \frac{1}{M^2}\left[(f_i - d) + \frac{1}{2}\sum_{j\neq i}(f_j - d) + \frac{1}{2}\sum_{k\neq i}(f_k - d)\right] \tag{20}$$

We have omitted details of this calculation for space considerations, though it is fairly involved and recommended that the reader attempt it to give assurance of the result and implications described in the remainder of the section. If we examine the final derivative, ignoring the constant scaling factor of $\frac{1}{M^2}$, we see the ensemble gradient has been broken into three components, which we will now label and make use of. Gradient component "A" is:

$$A = (f_i - d) \tag{21}$$

Noting that the second and third components are identical (apart from the indices j and k we have chosen), we have gradient components "B" and "C":

$$B = C = \frac{1}{2}\sum_{j\neq i}(f_j - d) \tag{22}$$

Though they are identical, it can be seen from the breakdown we have given that "B" is contributed by the ith estimator, whereas the "C" component is contributed by the sum of the other estimators. Remembering the NC error, using the second penalty term, eq (15), we have:

$$\begin{aligned}\frac{\partial e_i^{(2)}}{\partial f_i} &= (f_i - d) + \gamma\sum_{j\neq i}(f_j - d)\\ &= \quad A \quad + \ \gamma(B + C)\end{aligned}$$

Reintroducing the $\frac{1}{M^2}$ scaling factor, we note that when $\gamma = 1$ we have:

$$\frac{\partial e_{ens}}{\partial f_i} = \frac{1}{M^2}\frac{\partial e_i^{(2)}}{\partial f_i} \tag{23}$$

or rearranged:

$$\frac{\partial e_i^{(2)}}{\partial f_i} = M^2 \cdot \frac{\partial e_{ens}}{\partial f_i} \tag{24}$$

The gradient of the ith estimator's error function when using NC and setting $\gamma = 1.0$ is proportional to, but M^2 times steeper than the ensemble error gradient. This means that the minima will all be in the same locations in the search space, but indicates that a much faster convergence down the landscape should be observed, with the obvious consequence of possibly overshooting the minimum. At $\gamma = 0.5$, it exactly models the individual estimator's contribution to the ensemble error. However, since the estimators' outputs are combined, the errors cannot be assumed to be independent of one another; with $\gamma = 1.0$, it "simulates" the gradients of the remaining estimators, only possible because $B = C$.

5 Empirical Results

In the original experiments on NC [9], the penalty term in (15) was found to be more successful on classification problems. With our new knowledge on where the penalty is derived from, and noting that training was performed over a fixed number of iterations, we hypothesize that in fact the steepness of the landscape simply allowed faster convergence. We therefore engage in a short empirical test to verify this, using a dataset that NC is known to perform well on, the *Phoneme* data [6]. A full empirical benchmarking of the new penalty is outside the scope of this paper.

We use an ensemble of 20 MLPs each with 6 hidden nodes. We perform a five-fold cross validation, using 1 fold for training, 1 fold as validation data for early stopping, and 3 folds as testing data. Training is stopped by monitoring validation data for a rise over a 500 epoch moving window, at this point weights are reset to the best point within the training period so far. Results are in figure 1 and 2, also indicating 95% confidence intervals. Results show statistically significant improvements in comparison to a simple ensemble, in both cases when γ is set optimally. The correlation between the validation and testing curves indicate it is possible to use validation error to select this γ for use on testing data. At optimal γ there is no significant difference between the penalties. However, if we regard figure 3, we can see the number of iterations required to converge to this minimum error: here we see that the new penalty does indeed converge much faster, in almost half the number of training iterations.

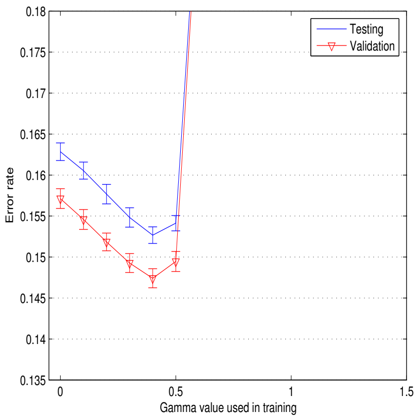

Fig. 1. Error using original penalty ($\bar{f}$)

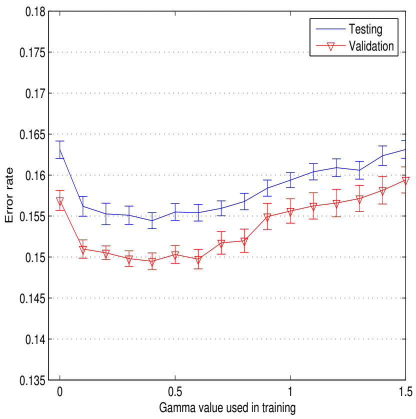

Fig. 2. Error using new penalty (d)

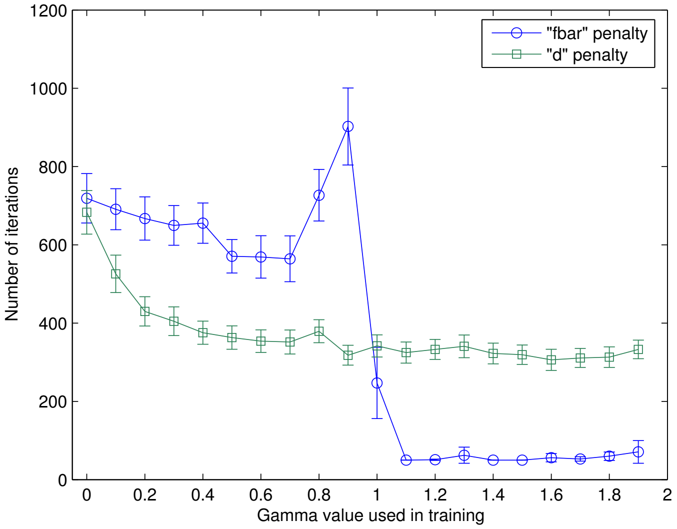

Fig. 3. Optimum number of training iterations

6 Conclusions

We examined what happens when we decompose the objective function of the ensemble into two components, one accounting for individual accuracy, and one accounting for the effect of correlations between the estimators. It should be noted *the decompositons in this paper applies only to ensembles using the simple average combination method, and a quadratic loss function*. We presented the Ambiguity decomposition [7], and a *new* alternative decomposition of the ensemble error. We showed how these decompositions have been exploited in a learning scheme, and used analysis of the error gradients to explain why one may outperform the other in certain situations.

In this work, we stated two assumptions which characterise the ensemble: a simple average combination function, and a quadratic loss error function. What happens when we use different assumptions? Are there decompositions when using other combiners, like the median and mode rules? Or other loss functions than quadratic? In classification, we are usually not interested in the quadratic loss from the true posterior probabilities - but more so in the *classification error rate*. This uses a *zero-one* loss function - are there analytic decompositions when this is the case? Current evidence [3] shows that an additive decomposition *does not exist*.

From this perspective, it seems obvious that any formulation of classification diversity will be intrinsically tied to 1) the loss function, and 2) the combination function, of the ensemble. Therefore any study citing the utility of “diversity” has a duty to present observations in this context, and not simply use “diversity” as if it were a mysterious panacea.

References

1. Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: A survey and categorisation. *Journal of Information Fusion*, 6(1):5–20, March 2005.
2. Gavin Brown, Jeremy Wyatt, and Peter Tino. Managing diversity in regression ensembles. *Journal of Machine Learning Research*, 6, 2005.
3. J.H. Friedman. Bias, variance, 0-1 loss and the curse of dimensionality. Technical report, Stanford University, 1996.
4. Giorgio Fumera and Fabio Roli. Linear combiners for classifier fusion: Some theoretical and experimental results. In *Proc. Int. Workshop on Multiple Classifier Systems (LNCS 2709)*, pages 74–83, Guildford, Surrey, June 2003. Springer.
5. S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.
6. University College London Neural Network Group. The Elena Project. http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA/elena.htm.
7. Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *NIPS*, 7:231–238, 1995.
8. Ludmila Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley Press, 2004. ISBN 0-471-21078-1.
9. Yong Liu. *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.
10. N. Ueda and R. Nakano. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks*, pages 90–95, 1996.