"Good" and "Bad" Diversity in Majority Vote Ensembles

Gavin Brown[1] and Ludmila I. Kuncheva[2]

[1] School of Computer Science, University of Manchester, UK
gavin.brown@cs.manchester.ac.uk
[2] School of Computer Science, Bangor University, UK
l.i.kuncheva@bangor.ac.uk

Abstract. Although diversity in classifier ensembles is desirable, its relationship with the ensemble accuracy is not straightforward. Here we derive a decomposition of the majority vote error into three terms: average individual accuracy, "good" diversity and "bad diversity". The good diversity term is taken out of the individual error whereas the bad diversity term is added to it. We relate the two diversity terms to the majority vote limits defined previously (the patterns of success and failure). A simulation study demonstrates how the proposed decomposition can be used to gain insights about majority vote classifier ensembles.

1 Introduction

The topic of 'diversity' has been a favourite buzzword in the multiple classifier systems community for well over a decade [1, 2]. Numerous diversity measures have been proposed, measured and maximised, all with the goal to increase ensemble performance by balancing "individual accuracy" against "diversity". It is therefore ironic that after so much time and effort, we still have no uniquely agreed definition for "diversity" [3, 4].

In this work we adopt the perspective that a diversity measure should be *naturally* defined as a consequence of two decisions in the design of the ensemble learning problem: the choice of *error function*, and the choice of *combiner function*. When discussing 'diversity', we often overlook these, implicitly adopting the *zero-one loss* (classification error) function, and the *majority vote* combiner. The principle of combining multiple predictions can of course be applied in many learning scenarios, such as regression [5], or unsupervised learning [6]. Depending on the situation, it may be appropriate to adopt other loss functions, such as the cross-entropy, or the squared loss. We might also consider other combiner functions, such as the average or product rule.

These two design decisions turn out to be very interesting for the diversity debate. It turns out that, for particular choices of error/combiner function, a definition of diversity *naturally* emerges. For a real-valued target y, if the

N. El Gayar, J. Kittler, and F. Roli (Eds.): MCS 2010, LNCS 5997, pp. 124–133, 2010.

ensemble is a linear combiner $\bar{f} = \frac{1}{T}\sum_t f_t$, and we assess it with the squared loss function, it is well appreciated that

$$(\bar{f} - y)^2 = \frac{1}{T}\sum_{t=1}^{T}(f_t - y)^2 - \frac{1}{T}\sum_{t=1}^{T}(f_t - \bar{f})^2. \quad (1)$$

The decomposition [7] states that the squared loss of the ensemble is *guaranteed* to be less than or equal to the average squared loss of the individuals. The difference is what we might call a 'diversity' term, measuring the average squared loss of the individuals from the ensemble prediction. The decomposition *only* holds when the ensemble combiner $\bar{f}$ is the (weighted) arithmetic mean of the individual predictions. In classification problems, it is common to minimize the *cross-entropy* loss function, which is derived from a Kullback-Leibler divergence. If we combine our multiple class probability estimates with a normalized geometric mean[1], then we have the decomposition,

$$D_{KL}(y||\bar{f}) = \frac{1}{T}\sum_{t=1}^{T} D_{KL}(y||f_t) - \frac{1}{T}\sum_{t=1}^{T} D_{KL}(\bar{f}||f_t). \quad (2)$$

The KL divergence of the ensemble from a target distribution, is *guaranteed* to be less than or equal to the average divergence of the individual estimates [8]. The difference is again what we might appreciate as diversity, measuring the average divergence of the individuals from the geometric mean.

In this paper we ask the question, *can a similar decomposition hold for classification error and majority vote combiners?* The answer turns out to be less straightforward than the above, and results in *two* diversity terms, which can both help and hinder the ensemble performance. In section 2 we present the main result, and in section 3 relate it to the patterns of 'success' and 'failure' in classifier combining [9]. Section 4 presents a simulation study monitoring the behaviour of the diversity terms in different situations.

2 Decomposition of the Majority Vote Error

Considers a two-class problem with class labels in the set $\{-1, +1\}$. Let $\Phi = \{\phi_1, \phi_2, \ldots, \phi_T\}$ be a set of classifiers, where T is odd. Denote by $h_t(\mathbf{x}) \in \{-1, +1\}$ the output of classifier ϕ_t for input $\mathbf{x}$. Let $y(\mathbf{x}) \in \{-1, +1\}$ be the true label of $\mathbf{x}$. The zero-one loss of ϕ_t for $\mathbf{x}$ is

$$e_t(\mathbf{x}) = \begin{cases} 0, y(\mathbf{x}) = h_t(\mathbf{x}) \\ 1, y(\mathbf{x}) \neq h_t(\mathbf{x}) \end{cases} = \frac{1}{2}\left(1 - y(\mathbf{x})\, h_t(\mathbf{x})\right). \quad (3)$$

If Φ is taken to be a classifier ensemble, the majority vote output for input $\mathbf{x}$ is

$$H(\mathbf{x}) = \text{sign}\left(\frac{1}{T}\sum_{t=1}^{T} h_t(\mathbf{x})\right), \quad (4)$$

[1] Equivalent to the product rule.

where $H(\mathbf{x}) \in \{-1, +1\}$. The zero-one loss of the ensemble for input $\mathbf{x}$ is

$$e_{\text{maj}}(\mathbf{x}) = \frac{1}{2}\ (1 - y(\mathbf{x})\ H(\mathbf{x}))\,. \tag{5}$$

Define also the *disagreement* between classifier ϕ_t and the ensemble as

$$\delta_t(\mathbf{x}) = \frac{1}{2}\ (1 - h_t(\mathbf{x})\ H(\mathbf{x}))\,. \tag{6}$$

Take the difference between the ensemble loss and the average individual loss,

$$\begin{aligned}
\Delta &= e_{\text{maj}}(\mathbf{x}) - e_{\text{ind}}(\mathbf{x}) \\
&= \frac{1}{2}\ (1 - y(\mathbf{x})\ H(\mathbf{x})) - \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\ (1 - y(\mathbf{x})\ h_t(\mathbf{x})) && (7)\\
&= \frac{1}{2} - \frac{1}{2}y(\mathbf{x})\ H(\mathbf{x}) - \frac{1}{2} + \frac{1}{2T}\sum_{t=1}^{T} y(\mathbf{x})\ h_t(\mathbf{x}) && (8)\\
&= -y(\mathbf{x})\ H(\mathbf{x})\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\left(1 - \frac{h_t(\mathbf{x})}{H(\mathbf{x})}\right). && (9)
\end{aligned}$$

Since $H(\mathbf{x}) \in \{-1, +1\}$, we can write $\frac{h_t(\mathbf{x})}{H(\mathbf{x})} = h_t(\mathbf{x})H(\mathbf{x})$, so we have:

$$\begin{aligned}
\Delta &= -y(\mathbf{x})\ H(\mathbf{x})\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\,(1 - h_t(\mathbf{x})\ H(\mathbf{x})) && (10)\\
&= -y(\mathbf{x})\ H(\mathbf{x})\frac{1}{T}\sum_{t=1}^{T}\delta_t(\mathbf{x}). && (11)
\end{aligned}$$

This demonstrates that *the difference between the majority voting loss and the average individual loss can be directly expressed in terms of the average classifier disagreement.* In summary:

$$e_{\text{maj}} = e_{\text{ind}} - y(\mathbf{x})\ H(\mathbf{x})\frac{1}{T}\sum_{t=1}^{T}\delta_t(\mathbf{x}) \tag{12}$$

Equation (12) is the zero-one loss of a majority vote on a single datapoint $\mathbf{x}$. To calculate the majority vote *classification error*, E_{maj}, we need to integrate with respect to the probability density function $p(\mathbf{x})$. In the following we also take advantage of the fact that $y(\mathbf{x})H(\mathbf{x}) = +1$ on datapoints where the ensemble is correct, and $y(\mathbf{x})H(\mathbf{x}) = -1$ where it is incorrect.

$$\begin{aligned}
E_{\text{maj}} &= \int_{\mathbf{x}} e_{\text{ind}}(\mathbf{x}) - \int_{\mathbf{x}} y(\mathbf{x})\ H(\mathbf{x})\frac{1}{T}\sum_{t=1}^{T}\delta_t(\mathbf{x}) && (13)\\
&= \int_{\mathbf{x}} e_{\text{ind}}(\mathbf{x}) - \underbrace{\int_{\mathbf{x}+}\frac{1}{T}\sum_{t=1}^{T}\delta_t(\mathbf{x})}_{\text{good diversity}} + \underbrace{\int_{\mathbf{x}-}\frac{1}{T}\sum_{t=1}^{T}\delta_t(\mathbf{x})}_{\text{bad diversity}} && (14)
\end{aligned}$$

Here the integral has been separated for two subspaces of the data: $\mathbf{x}+$ where the ensemble is correct, and $\mathbf{x}-$ where it is incorrect.

Equation (14) prompts the following interpretation. The majority vote error has a direct relationship with two components of diversity, measured by the disagreement between the classifier decision $h_t(\mathbf{x})$ and the ensemble decision $H(\mathbf{x})$. We label the two diversity components "good" and "bad" diversity. The good diversity measures the disagreement on datapoints where the ensemble is *correct*—and due to the negative sign, any disagreement on these points *increases* the gain relative to the average individual error. The bad diversity measures the disagreement on datapoints where the ensemble is *incorrect*—here, the diversity term has a positive sign, so any disagreement *reduces* the gain relative to individual error.

Another way to think of this is as the number of votes "wasted" by the ensemble members. For an arbitrary datapoint $\mathbf{x}$, assume the ensemble is already correct. If there is little disagreement, then several votes have been 'wasted', since the same correct decision would have been taken had some classifiers changed their votes. This is the "good" diversity term, measuring disagreement where the ensemble is already correct—more disagreement equates to fewer wasted votes. The "bad" diversity term measures the opposite: disagreement where the ensemble is incorrect—any disagreement equates to a wasted vote, as the individual did not have any effect on the ensemble decision. Thus, increasing good diversity and decreasing bad diversity is equivalent to reducing the number of 'wasted' votes.

We now provide alternative formulations of the good/bad diversity terms, that will facilitate a link to the patterns of "success" and "failure" previously used to study the limits of majority voting [9]. Let $\mathbf{v}$ be a T-dimensional binary random variable such that $v_t = 1$ means that classifier ϕ_t labels $\mathbf{x}$ correctly, and $v_t = 0$ means that the assigned label is incorrect. Construct a T-element vector $\mathbf{1} = [1, 1, \ldots, 1]^T$. Then the scalar product $\mathbf{v}^T\mathbf{1}$ will be the number of correct votes in $\mathbf{v}$. The ensemble is correct when $\mathbf{v}^T\mathbf{1} \geq \frac{T+1}{2}$. The "good diversity" in this case is the number of incorrect votes, i.e., $T - \mathbf{v}^T\mathbf{1}$. When the ensemble is incorrect, the "bad diversity" is the number of correct votes, i.e., $\mathbf{v}^T\mathbf{1}$. The integral across the feature space can be replaced by a summation across all possible values of $\mathbf{v}$ leading to

$$E_{\text{maj}} = E_{\text{ind}} - \underbrace{\frac{1}{T} \sum_{\mathbf{v}^T\mathbf{1} \geq \frac{T+1}{2}} (T - \mathbf{v}^T\mathbf{1})\, p(\mathbf{v})}_{\text{good diversity}} + \underbrace{\frac{1}{T} \sum_{\mathbf{v}^T\mathbf{1} < \frac{T+1}{2}} \mathbf{v}^T\mathbf{1}\, p(\mathbf{v})}_{\text{bad diversity}}. \quad (15)$$

In the following section we show these terms can be related to the patterns of 'success' and 'failure' [9] for voting ensembles.

3 Patterns of Success and Failure

The *pattern of success* and the *pattern of failure* were introduced as special cases illustrating the limits of the majority vote [9]. Given a set of classifiers of the same individual accuracy p, the pattern of *success* is the most favorable distribution of the correct votes, leading to the largest improvement on p. To achieve this, we define a probability distribution over all possible combinations of correct/incorrect votes. Each combination where exactly $\frac{T+1}{2}$ votes are correct, appears with probability α. The only other combination of votes with non-zero probability is when all votes are incorrect. In this pattern, there are no wasted votes, as each vote is needed to ensure the smallest majority of correct votes.

For example, consider three classifiers, $\Phi = \{\phi_1, \phi_2, \phi_3\}$, each of accuracy $p = 0.6$. Denote again a correct vote by 1, and an incorrect vote by 0. The pattern of success is constructed by assigning probability α to each of the three combinations of votes 011, 101, 110, and probability $1 - 3\alpha$ to 000. Since each of the classifiers will be accurate in two of these combinations, to make up the individual accuracy of 0.6, α must be 0.3. The majority vote error for this example is $1 - 3\alpha = 0.1$ [9]. This calculation can now be easily demonstrated using equation (15).

$$\begin{aligned} E_{\text{maj}} &= E_{\text{ind}} - \frac{1}{3}(3 \times (3-2) \times \alpha) + \frac{1}{3}(0 \times (1-3\alpha)) \\ &= 0.4 - \frac{1}{3} \times 3 \times 0.3 = 0.4 - 0.3 = 0.1. \end{aligned}$$

In the pattern of failure, the correct votes are distributed in such a way that they are one shy of majority, so the largest quantity of correct votes are wasted. In the example above, the pattern of failure is constructed by assigning probabilities β to combination of votes 001, 010 and 100, and probability $1-3\beta$ to combination 111. Each of the three classifiers is accurate only in combination 111 (probability $1 - 3\beta$) and one of the combinations with one correct vote (probability β). To ensure that the individual accuracy is $p = 0.6$, we have $1 - 2\beta = 0.6$, hence $\beta = 0.2$. Using equation (15) again,

$$\begin{aligned} E_{\text{maj}} &= E_{\text{ind}} - \frac{1}{3}((3-3) \times (1-3\beta)) + \frac{1}{3}(3 \times 1 \times \beta)) \\ &= 0.4 + \frac{1}{3} \times 3 \times 0.2 = 0.4 + 0.2 = 0.6. \end{aligned}$$

In the general definition of the pattern of success, each of the possible $\binom{T}{\frac{T+1}{2}}$ vote combinations where the ensemble is correct appears with probability α, and the combination where $\mathbf{v}^T\mathbf{1} = 0$ appears with probability $1 - \binom{T}{\frac{T+1}{2}}\alpha$. All other vote combinations have zero probability. Then, substituting these values into (15), we have

$$E_{\text{maj}} = (1-p) - \frac{1}{T}\left(\binom{T}{\frac{T+1}{2}} \times \frac{T-1}{2} \times \alpha\right) \tag{16}$$

To ensure that the individual accuracy of all classifiers is p, α must satisfy

$$\alpha = \frac{p}{\binom{T-1}{\frac{T-1}{2}}}. \tag{17}$$

Substituting (17) in (16) and applying simple algebraic manipulation, we recover exactly the expression for the upper bound on majority vote accuracy, defined in [9]

$$E_{\text{maj}} = \max\left\{0,\ \ E_{\text{ind}} - \frac{T-1}{T+1}(1 - E_{\text{ind}})\right\}. \tag{18}$$

In a similar way, substituting the appropriate figures for the pattern of failure, (15) leads to the general expression for the lower bound on majority vote error:

$$E_{\text{maj}} = E_{\text{ind}}\left(1 + \frac{T-1}{T+1}\right). \tag{19}$$

Equation (15) gives a direct relationship between the majority vote error and the patterns of success/failure, where the bad and the good diversity partly "neutralise" one another.

4 Simulation Experiment

At this stage it is difficult to recommend a way to use the decomposition straightforwardly in a classifier ensemble algorithm. The simulation study illustrates how the decomposition can help to gain insight into the internal workings of majority vote classifier ensembles.

4.1 Data

The 2-class, 2-d data set is shown in Figure 1. The data is generated uniformly in the unit square and the classes are labelled according to a rotated checker board[2]. The problem has a Bayes error of 0.

4.2 Experimental Protocol

The ensemble was sampled from a pool of linear classifiers. Each classifier was constructed by drawing a random line through a point in the unit square. The two sides of the line were labelled in the two classes in the way that gave accuracy greater than 0.5. Thus the classifiers were only slightly better than chance. The following steps were carried out

[2] Matlab code for N data points, uniform distribution, checkerboard with side a, rotated by θ:

```
function[d,labd]=gendatcb(N,a,theta)
d=rand(N,2);
d_transformed=[d(:,1)*cos(theta)-d(:,2)*sin(theta),d(:,1)*sin(theta)+d(:,2)*cos(theta)];
s=ceil(d_transformed(:,1)/a)+floor(d_transformed(:,2)/a);labd=2-mod(s,2);
```

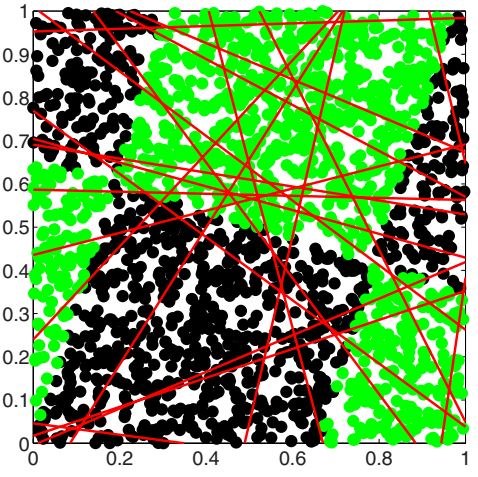

Fig. 1. Rotated checker board data and the boundaries of 20 random linear classifiers. The checker board was generated with side $a = 0.63$ and rotation angle $\theta = 0.3$.

1. Generate an initial pool Φ consisting of 2000 random linear classifiers.
2. The ensemble size T was varied as

$$T \in \{1, 3, 5, 9, 13, 19, 31, 51, 71, 101, 201, 501, 1001\}.$$

3. Generate and test fifty ensembles for each value of T :
   (a) T classifiers were sampled without replacement from the pool Φ.
   (b) A testing data set of 1000 2-d points was generated.
   (c) The testing accuracy of the ensemble, the average individual accuracy, and the two diversity terms were estimated and stored. The ensemble labelled the data points through the majority vote.
4. The stored values were averaged across the 50 runs.

To evaluate the effect of the individual accuracy on the ensemble accuracy and the diversity terms, the above protocol was repeated for a new "selected" pool of classifiers, engineered so as to have individual accuracy significantly better than random. The procedure was to first generate 8000 random classifiers, then select the 2000 with highest accuracy.

4.3 Results

To evaluate the success of the ensemble with increasing T, the majority vote errors for the random and the selected ensembles are plotted in Figure 2. As expected, the error drops with T. The "Random" ensembles show better improvement on the individual error rate but the ensemble error does not reach the one when sampling the ensemble from the selected set.

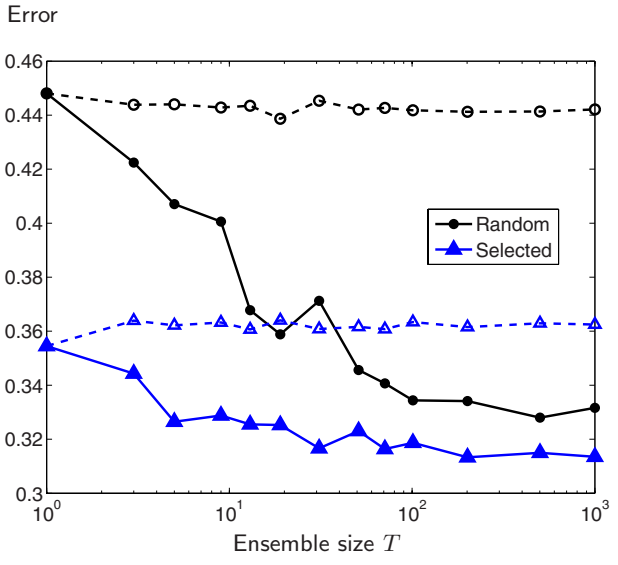

Fig. 2. Majority vote error (solid lines) and average individual error (dotted lines) versus the ensemble size T. "Random" corresponds to the original classifier pool, where the individual classifiers are only slightly better than chance. "Selected" corresponds to the selected pool, engineered to have higher accuracy.

Figure 3 shows the scatter plot of the Good diversity versus Bad diversity terms. Three values of T were chosen for the illustration, 3, 31, and 1001, shown with different markers. Each point represents one ensemble, thus there are 50 points for each marker, for each of the 50 trials. The diagonal line corresponds to when good = bad diversity, in which case they cancel each other out, and ensemble error is just equal to the individual error. Points *above* the diagonal line show ensembles where good diversity is larger than bad diversity, therefore the ensemble improves on the average individual error. The improvement of the ensemble with respect to the average individual error can be read off the plot as the vertical distance from the point to the diagonal line.

The ensembles in subplot (a) are more dispersed than those in plot (b). This reveals that both good and bad diversities reach larger values for the random pool than for the selected pool of classifiers for the same ensemble size T. This can be explained with the fact that higher individual error rate allows for a "weak" majority, where the ensemble is correct but there are many incorrect votes as well. This effect is especially visible for $T = 3$. While the "random" scenario offers a range of very good and very bad ensembles (scattered far above or below the diagonal line), the "selected" pool has a modest range of ensembles. For larger T, both ensemble pools produce compact clusters of points suggesting that the good-bad diversity ratio stabilises with T. Even though the clusters for $T = 1001$ are compact, they are positioned differently with respect to the diagonal line.

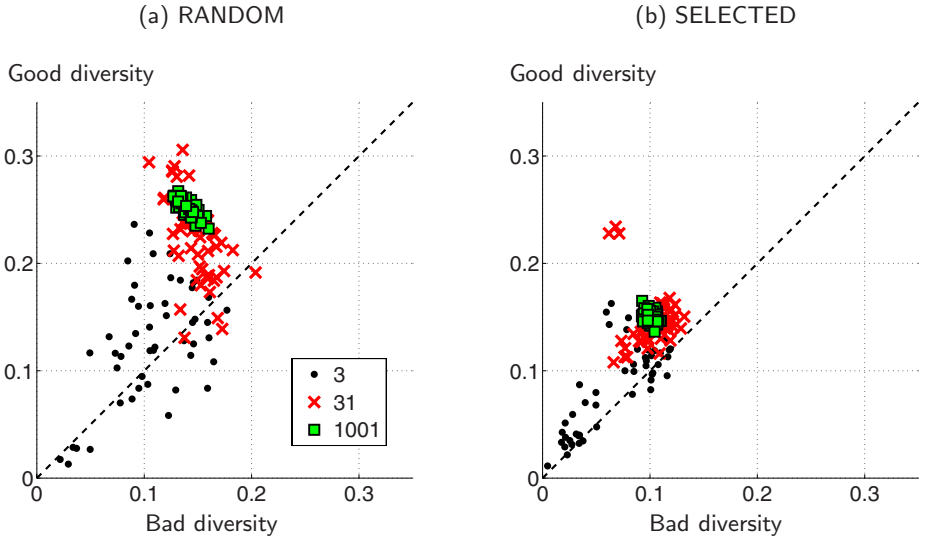

Fig. 3. Scatterplots of good diversity versus bad diversity terms for the "random" and the "selected" pools of classifiers for 3 values of T

The ensembles sampled from the "random" pool are further up compared to these sampled from the "selected" pool, which is mirrored in the larger versus smaller improvement on the individual errors in the rightmost points in Figure 2. While the improvement on the individual error can be gauged from Figure 2, the spread of the values of the good and bad diversity cannot.

Interestingly, the points for $T = 3$ and $T = 31$ are not shaped as coherent clusters but are rather split into sub-clusters (more visible in subplot (b)). The group of crosses far above the main cluster in subplot (b) corresponds to ensembles with dramatic improvement on the individual error rates (close to the pattern of success). This may give a lead towards creating ensembles that magnify good diversity while keeping the bad diversity at bay.

5 Conclusions

In this paper we adopted the perspective that a diversity measure should be naturally derived as a direct consequence of two factors: the loss function of interest, and the combiner function. We presented a decomposition of the classification error, using the majority vote combiner, into three terms: individual accuracy, 'good' diversity, and 'bad' diversity. A larger value of the good diversity reduces the majority vote error, whereas a larger value of bad diversity *increases* the error. We showed a direct relation of these concepts to the upper/lower limits defined on majority voting error [9]. A simulation study illustrated that the diversity terms tend to exhibit a large variance in smaller ensembles, and stabilize with very large ensembles.

The decomposition lends direct support not only to the existing theory of majority voting [9], but also to existing algorithms. The DECORATE algorithm [10] uses artificially constructed data examples to induce a diversity in majority voting by making the individuals disagree wherever possible with the ensemble. The results of this paper suggests it may be possible to construct more targeted algorithms, which directly magnify the "good" diversity while suppressing the "bad" diversity.

References

1. Brown, G.: Ensemble Learning. In: Encyclopedia of Machine Learning. Springer Press, Heidelberg (2010)
2. Kuncheva, L.: Combining pattern classifiers: methods and algorithms. Wiley-Interscience, Hoboken (2004)
3. Kuncheva, L.: That elusive diversity in classifier ensembles. In: Perales, F.J., Campilho, A.C., Pérez, N., Sanfeliu, A. (eds.) IbPRIA 2003. LNCS, vol. 2652, pp. 1126–1138. Springer, Heidelberg (2003)
4. Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity creation methods: a survey and categorisation. Information Fusion 6(1), 5–20 (2005)
5. Brown, G., Wyatt, J., Tiňo, P.: Managing diversity in regression ensembles. Journal of Machine Learning Research 6, 1650 (2005)
6. Strehl, A., Ghosh, J.: Cluster ensembles: a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research 3, 583–617 (2003)
7. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In: Advances in neural information processing systems, pp. 231–238 (1995)
8. Heskes, T.: Bias/variance decompositions for likelihood-based estimators. Neural Computation 10(6), 1425–1433 (1998)
9. Kuncheva, L., Whitaker, C., Shipp, C., Duin, R.: Limits on the majority vote accuracy in classifier fusion. Pattern Analysis & Applications 6(1), 22–31 (2003)
10. Melville, P., Mooney, R.: Creating diversity in ensembles using artificial data. Information Fusion 6(1), 99–111 (2005)