Machine Learning, 51, 181–207, 2003
© 2003 Kluwer Academic Publishers. Manufactured in The Netherlands.

Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy

LUDMILA I. KUNCHEVA l.i.kuncheva@bangor.ac.uk
CHRISTOPHER J. WHITAKER c.j.whitaker@bangor.ac.uk
School of Informatics, University of Wales, Bangor, Dean Street, Bangor, Gwynedd, LL57 1UT, UK

Editor: Robert E. Schapire

Abstract. Diversity among the members of a team of classifiers is deemed to be a key issue in classifier combination. However, measuring diversity is not straightforward because there is no generally accepted formal definition. We have found and studied ten statistics which can measure diversity among binary classifier outputs (correct or incorrect vote for the class label): four averaged pairwise measures (the Q statistic, the correlation, the disagreement and the double fault) and six non-pairwise measures (the entropy of the votes, the difficulty index, the Kohavi-Wolpert variance, the interrater agreement, the generalized diversity, and the coincident failure diversity). Four experiments have been designed to examine the relationship between the accuracy of the team and the measures of diversity, and among the measures themselves. Although there are proven connections between diversity and accuracy in some special cases, our results raise some doubts about the usefulness of diversity measures in building classifier ensembles in real-life pattern recognition problems.

Keywords: pattern recognition, multiple classifiers ensemble/committee of learners, dependency and diversity, majority vote

1. Introduction

Combining classifiers is now an established research area known under different names in the literature: committees of learners, mixtures of experts, classifier ensembles, multiple classifier systems, consensus theory, etc. Bagging and boosting methods for team members generation, and variants thereof such as arcing and wagging, have been shown to be very successful (Bauer & Kohavi, 1999; Drucker et al., 1994; Schapire, 1999). The key to the success of these algorithms is that, intuitively at least, they build a set of *diverse* classifiers. In general, in both methods the individual classifiers are designed on different subsets of the training data. In the boosting algorithm, the sampling distribution is updated before drawing the training data for the new member of the team. The likelihood for those objects that have been misclassified by the previously generated members of the team is increased, and the classifier team grows progressively "diverse". There is no explicit measure of diversity involved in the process but it is assumed that diversity is a key factor for the success of this algorithm.

Diversity has been recognized as a very important characteristic in classifier combination (Cunningham & Carney, 2000; Krogh & Vedelsby, 1995; Rosen, 1996; Lam, 2000; Littlewood & Miller, 1989). However, there is no strict definition of what is intuitively perceived as diversity, dependence, orthogonality or complementarity of classifiers. Many measures of the connection between two classifier outputs can be derived from the statistical literature (e.g., Sneath & Sokal (1973)). There is less clarity on the subject when three or more classifiers are concerned. There are methods, formulas and ideas aiming at quantifying diversity but, owing to the lack of a definition, little is put on a rigorous or systematic basis.

The general anticipation is that diversity measures will be helpful in designing the individual classifiers and the combination technology. Here are the key questions that need to be addressed:

1. How do we define and measure diversity?
2. How are the various measures of diversity related to each other?
3. How are the measures related to the accuracy of the team?
4. Is there a measure that is best for the purposes of developing committees that minimize error?
5. How can we use the measures in designing the classifier ensemble?

The rest of the paper seeks answers to these questions. To answer question one, the current results from the literature are summarized in Section 2, and ten measures of diversity are introduced in Sections 3 and 4. Section 5 gives a comment on question 3. Section 6 explains the experiments performed in relation to questions 2 and 3. In Section 7, possible answers to questions 4 and 5 are considered.

2. Diversity in classifier ensembles

Let $\mathcal{D} = \{D_1, \ldots, D_L\}$ be a set (pool, committee, mixture, team, ensemble) of classifiers, $\Omega = \{\omega_1, \ldots, \omega_c\}$ be a set of class labels and $\mathbf{x} \in \Re^n$ be a vector with n features to be labeled in Ω. There are three general possibilities for the classifier outputs

1. *A c-element vector* $\mu^i = [d_{i,1}(\mathbf{x}), \ldots, d_{i,c}(\mathbf{x})]^T$ with supports for the classes (a special case of this vector is a probability distribution over Ω estimating the posterior probabilities $P(\omega_s \mid \mathbf{x})$, $s = 1, \ldots, c$). Thus, for each input $\mathbf{x}$ we have L c-dimensional vectors of support.
2. *Class label.* $D_i(\mathbf{x}) \in \Omega,\ i = 1, \ldots, L$.
3. *Correct/incorrect decision (the oracle output).* The output $D_i(\mathbf{x})$ is 1 if $\mathbf{x}$ is recognized correctly by D_i, and 0, otherwise. This is an "oracle" type of output because it assumes that we know the correct label of $\mathbf{x}$, at least for the points in some finite $\mathbf{Z} \subset \Re^n$.

Current findings and results about diversity can be summarized as follows:

– Assume that classifier outputs are estimates of the *posterior probabilities*, $\hat{P}_i(\omega_s \mid \mathbf{x})$, $s = 1, \ldots, c$, so that the estimate $\hat{P}_i(\omega_s \mid \mathbf{x})$ satisfies

$$\hat{P}_i(\omega_s \mid \mathbf{x}) = P(\omega_s \mid \mathbf{x}) + \eta_s^i(\mathbf{x}), \tag{1}$$

where $\eta_s^i(\mathbf{x})$ is the error for class ω_s made by classifier D_i.

The outputs for each class are combined by averaging, or by an order statistic such as minimum, maximum or median. Tumer and Ghosh (1996a, 1996b, 1999) derive an expression about the added classification error (i.e., the error above the Bayes error) of the team under a set of assumptions

$$E_{add}^{ave} = E_{add}\left(\frac{1+\delta(L-1)}{L}\right), \tag{2}$$

where E_{add} is the added error of the individual classifiers (all have the same error), and δ is a correlation coefficient.[1]

One of the assumptions is that a classifier produces independent estimates of the posterior probabilities, $\hat{P}_i(\omega_s \mid \mathbf{x}), s = 1, \ldots, c$. This is not the case, as by design $\sum_s \hat{P}_i(\omega_s \mid \mathbf{x}) = 1$. The derivation also assumes independence between the estimates for different classes from two different classifiers, i.e., $\hat{P}_i(\omega_s \mid \mathbf{x})$ and $\hat{P}_j(\omega_k \mid \mathbf{x})$, $s, k = 1, \ldots c$, $s \neq k$, $i, j = 1, \ldots, L$, $i \neq j$. There is no information available about whether violation of these assumptions will have a substantial effect on the derived relationship.

Ignoring the problems listed above, the main result of Tumer and Ghosh is that positively correlated classifiers only slightly reduce the added error, uncorrelated classifiers reduce the added error by a factor of $1/L$, and negatively correlated classifiers reduce the error even further. But note that there is a limit on the largest absolute value of a negative pairwise correlation among L classifiers. Tumer and Ghosh (1996b, 1999) do not mention the case of negative correlation although it clearly supports their thesis that the smaller the correlation, the better the ensemble. A negative correlation between the continuous-valued outputs has been sought, predominantly by altering the available training set or parameters of the classifier (Dietterich, 2000a; Hashem, 1999; Krogh & Vedelsby, 1995; Liu & Yao, 1999; Opitz & Shavlik, 1999; Parmanto et al., 1996; Giacinto & Roli, 2001; Rosen, 1996; Sharkey & Sharkey, 1997; Skalak, 1996; Tumer & Ghosh, 1999).

- When classifiers output class labels, the classification error can be decomposed into bias and variance terms (also called 'spread') (Bauer & Kohavi, 1999; Breiman, 1999; Kohavi & Wolpert, 1996) or into bias and spread terms. In both cases the second term can be taken as the diversity of the ensemble. These results have been used to study the behavior of classifier ensembles in terms of the bias-variance trade-off.

Despite the theory that is available for continuous-valued outputs, many authors discuss the concept of diversity in terms of correct/incorrect (oracle) outputs. The rest of this paper concerns only the oracle outputs.

3. Pairwise diversity measures

3.1. The Q statistics

Let $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ be a labeled data set, $\mathbf{z}_j \in \Re^n$ coming from the classification problem in question. We can represent the output of a classifier D_i as an N-dimensional binary vector

Table 1. A 2 × 2 table of the relationship between a pair of classifiers.

	D_k correct (1)	D_k wrong (0)
D_i correct (1)	N^{11}	N^{10}
D_i wrong (0)	N^{01}	N^{00}

Total, $N = N^{00} + N^{01} + N^{10} + N^{11}$.

$\mathbf{y}_i = [y_{1,i}, \ldots, y_{N,i}]^T$, such that $y_{j,i} = 1$, if D_i recognizes correctly $\mathbf{z}_j$, and 0, otherwise, $i = 1, \ldots, L$.

There are various statistics to assess the similarity of two classifier outputs (Afifi & Azen, 1979). Yule's Q statistic (1900) for two classifiers, D_i and D_k, is

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \tag{3}$$

where N^{ab} is the number of elements $\mathbf{z}_j$ of $\mathbf{Z}$ for which $y_{j,i} = a$ and $y_{j,k} = b$ (see Table 1).

For statistically *independent* classifiers, the expectation of $Q_{i,k}$ is 0. Q varies between -1 and 1. Classifiers that tend to recognize *the same* objects correctly will have positive values of Q, and those which commit errors on different objects will render Q negative. For a team $\mathcal{D}$ of L classifiers, the averaged Q statistics over all pairs of classifiers is,

$$Q_{av} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k} \tag{4}$$

3.2. *The correlation coefficient* ρ

The correlation between two binary classifier outputs (correct/incorrect), $\mathbf{y}_i$ and $\mathbf{y}_k$, is

$$\rho_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11} + N^{10})(N^{01} + N^{00})(N^{11} + N^{01})(N^{10} + N^{00})}} \tag{5}$$

For any two classifiers, Q and ρ have the same sign, and it can be proved that $|\rho| \leq |Q|$. In Kuncheva et al. (2000) we chose Q to measure the dependency because it has been designed for contingency tables such as Table 1, and is simpler than ρ to calculate from the table entries.

3.3. *The disagreement measure*

This measure was used by Skalak (1996) to characterize the diversity between a base classifier and a complementary classifier, and then by Ho (1998) for measuring diversity in decision forests. It is the ratio between the number of observations on which one classifier is correct and the other is incorrect to the total number of observations. In our notation,

$$Dis_{i,k} = \frac{N^{01} + N^{10}}{N^{11} + N^{10} + N^{01} + N^{00}} \tag{6}$$

3.4. The double-fault measure

This measure was used by Giacinto and Roli (2001) to form a pairwise diversity matrix for a classifier pool and subsequently to select classifiers that are least related. It is defined as the proportion of the cases that have been misclassified by both classifiers, i.e.,

$$DF_{i,k} = \frac{N^{00}}{N^{11} + N^{10} + N^{01} + N^{00}} \tag{7}$$

For all pairwise measures we used the averaged values over the diversity matrix, calculated similarly to (4). We note that all these pairwise measures have been proposed as measures of (dis)similarity in the numerical taxonomy literature (e.g., Sneath & Sokal (1973)).

4. Non-pairwise diversity measures

4.1. The entropy measure E

The highest diversity among classifiers for a particular $\mathbf{z}_j \in \mathbf{Z}$ is manifested by $\lfloor L/2 \rfloor$ of the votes in $\mathbf{Z}_j$ with the same value (0 or 1) and the other $L - \lfloor L/2 \rfloor$ with the alternative value. If they all were 0's or all were 1's, there is no disagreement, and the classifiers cannot be deemed diverse. Denote by $l(\mathbf{z}_j)$ the number of classifiers from $\mathcal{D}$ that correctly recognize $\mathbf{z}_j$, i.e, $l(\mathbf{z}_j) = \sum_{i=1}^{L} y_{j,i}$.

One possible measure of diversity based on this concept is

$$E = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{(L - \lceil L/2 \rceil)} \min\{l(\mathbf{z}_j), L - l(\mathbf{z}_j)\}. \tag{8}$$

E varies between 0 and 1, where 0 indicates no difference and 1 indicates the highest possible diversity.

4.2. Kohavi-Wolpert variance

Kohavi and Wolpert (1996) derived a decomposition formula for the error rate of a classifier. They give an expression of the variability of the predicted class label y for $\mathbf{x}$, across training sets, for a specific classifier model

$$\text{variance}_x = \frac{1}{2} \left(1 - \sum_{i=1}^{c} P(y = \omega_i \mid \mathbf{x})^2 \right). \tag{9}$$

We use this general idea in the following way. We look at the variability of the predicted class label for $\mathbf{x}$ for the given training set using the classifier models $D_1, \ldots, D_L$. Instead of Ω, in this paper we consider two possible classifier outputs: correct and incorrect. The authors point out that $P(y = \omega_i \mid \mathbf{x})$ is estimated as an average over different data sets. In our case, $P(y = 1 \mid \mathbf{x})$ and $P(y = 0 \mid \mathbf{x})$ will be obtained as an average over $\mathcal{D}$, i.e.,

$$\hat{P}(y = 1 \mid \mathbf{x}) = \frac{l(\mathbf{x})}{L} \quad \text{and} \quad \hat{P}(y = 0 \mid \mathbf{x}) = \frac{L - l(\mathbf{x})}{L} \tag{10}$$

Substituting (10) in (9),

$$\text{variance}_x = \frac{1}{2}(1 - \hat{P}(y = 1 \mid \mathbf{x})^2 - \hat{P}(y = 0 \mid \mathbf{x})^2), \tag{11}$$

and averaging over the whole of $\mathbf{Z}$, we set the KW measure of diversity to be

$$KW = \frac{1}{NL^2} \sum_{j=1}^{N} l(\mathbf{z}_j)(L - l(\mathbf{z}_j)) \tag{12}$$

Interestingly, KW differs from the averaged disagreement measure Dis_{av} by a coefficient, i.e.,

$$KW = \frac{L-1}{2L} Dis_{av}. \tag{13}$$

(The proof of the equivalence is given in the Appendix.) Therefore, the relationship between the majority vote accuracy (P_{maj}) and Dis_{av} will match exactly the relationship between P_{maj} and KW.

4.3. *Measurement of interrater agreement κ*

A statistic developed as a measure of interrater reliability, called κ, can be used when different raters (here classifiers) assess subjects (here $\mathbf{z}_j$) to measure the level of agreement while correcting for chance (Fleiss, 1981). It has links to the intraclass correlation coefficient and the significance test of Looney (1988).

If we denote $\bar{p}$ to be the average individual classification accuracy, i.e.,

$$\bar{p} = \frac{1}{NL} \sum_{j=1}^{N} \sum_{i=1}^{L} y_{j,i}, \tag{14}$$

then

$$\kappa = 1 - \frac{\frac{1}{L} \sum_{j=1}^{N} l(\mathbf{z}_j)(L - l(\mathbf{z}_j))}{N(L-1)\bar{p}(1-\bar{p})} \tag{15}$$

and so κ is related to KW and Dis_{av} as follows

$$\kappa = 1 - \frac{L}{(L-1)\bar{p}(1-\bar{p})} KW = 1 - \frac{1}{2\bar{p}(1-\bar{p})} Dis_{av}. \tag{16}$$

Fleiss (1981) defines the pairwise κ_p as

$$\kappa_p = \frac{2(N^{11}N^{00} - N^{01}N^{10})}{(N^{11} + N^{10})(N^{01} + N^{00}) + (N^{11} + N^{01})(N^{10} + N^{00})}. \tag{17}$$

However, it can be shown that the (non-pairwise) κ (15) is not obtained by averaging κ_p.

Dietterich (2000b) uses the κ statistic as a measure of diversity between two classifiers. Class label outputs were considered and κ was calculated for each pair of classifiers from their coincidence matrix. Scatterplots called "κ-error diagrams" were given, where kappa was plotted against mean accuracy of the classifier pair.

4.4. The measure of "difficulty" θ

The idea for this measure came from a study by Hansen and Salamon (1990). We define a discrete random variable X taking values in $\{\frac{0}{L}, \frac{1}{L}, \ldots, 1\}$ and denoting the proportion of classifiers in $\mathcal{D}$ that correctly classify an input $\mathbf{x}$ drawn randomly from the distribution of the problem. To estimate the probability mass function of X, the L classifiers in $\mathcal{D}$ are run on the data set $\mathbf{Z}$.

Figure 1 shows three possible histograms of X for $L = 7$ and $N = 100$ data points. We assumed that all classifiers have individual accuracy $p = 0.6$. The leftmost plot shows the histogram if the seven classifiers were independent. In this case $X \times L$ has a Binomial distribution ($p = 0.6, n = L$). The middle plot shows seven identical classifiers. They all recognize correctly *the same* 60 points and misclassify the remaining 40 points in $\mathbf{Z}$. The rightmost plot corresponds to the case of negatively dependent classifiers. They recognize different subsets of $\mathbf{Z}$. The figures in the histogram are calculated so that the sum of all correct votes is $L \times p \times N = 7 \times 0.6 \times 100 = 420$. That is, if m_i denotes the number of data points for $X = \frac{i}{L}$, in all three histograms, $\sum_{i=1}^{L} i\ m_i = 420$.

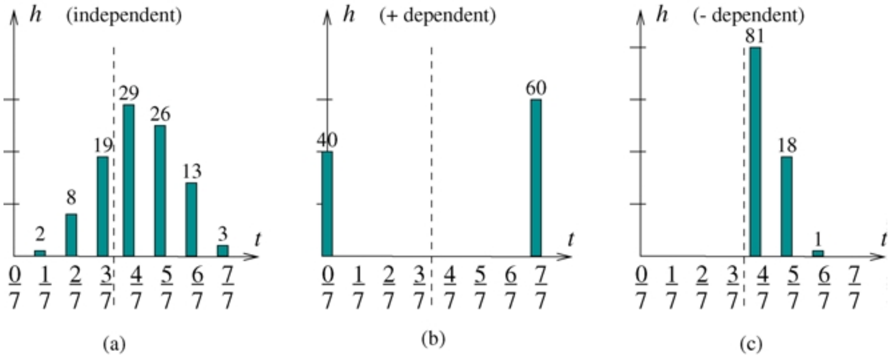

Figure 1. Patterns of "difficulty" for three classifier teams with $L = 7$, $p = 0.6$, and $N = 100$. The dashed line is the majority vote border. The histograms show the number of points (out of 100) which are correctly labeled by i of the L classifiers. The x-axis is "proportion correct", i.e., i/L.

Hansen and Salamon (1990) talk about a pattern of "difficulty" of the points in the feature space, which in our case is represented by the histogram over **Z**. If the same points have been *difficult* for all classifiers, and the other points have been *easy* for all classifiers, we obtain a plot similar to the middle one (no diversity in the team). If the points that were difficult for some classifiers were easy for other classifiers, the distribution of X is as the one on the right. Finally, if each point is equally difficult for all classifiers, the distribution on the left is the most likely one.

The target is a measure of diversity based on the *distribution of difficulty*. We can capture the distribution shape by using the variance of X. Diverse teams $\mathcal{D}$ will have smaller variance of X (right plot). Teams of similar classifiers will have high variance, as the pattern in the middle plot. Let the three variables X in figure 1(a), 1(b) and 1(c) be X_a, X_b, and X_c, respectively. The three variances are

$$\theta_a = \mathit{Var}(X_a) = 0.034, \quad \theta_b = \mathit{Var}(X_b) = 0.240, \quad \theta_c = \mathit{Var}(X_c) = 0.004.$$

Based on this, we define the *difficulty* θ to be $\mathit{Var}(X)$. For convenience we can scale θ linearly into [0, 1], taking $p(1-p)$ as the highest possible value. The higher the value of θ, the worse the classifier team. Ideally, $\theta = 0$, but this is an unrealistic scenario. More often, real classifiers are positively dependent and will exhibit patterns similar to figure 1(b).

4.5. *Generalized diversity*

This measure has been proposed by Partridge and Krzanowski (1997).

Let Y be a random variable expressing the proportion of classifiers (out of L) that are incorrect (or *fail*) on a randomly drawn object $\mathbf{x} \in \Re^n$. Denote by p_i the probability that $Y = \frac{i}{L}$. (Note that $Y = 1 - X$, where X was introduced for θ). Denote by $p(i)$ the probability that i randomly chosen classifiers will fail on a randomly chosen $\mathbf{x}$. Suppose that two classifiers are randomly picked from $\mathcal{D}$. Partridge and Krzanowski argue that maximum diversity occurs when failure of one of these classifiers is accompanied by correct labeling by the other classifier. In this case the probability of both classifiers failing is $p(2) = 0$. Minimum diversity occurs when failure of one is always accompanied by failure of the other. Then the probability of both classifiers failing is the same as the probability of one randomly picked classifier failing, i.e., $p(1)$. Using

$$p(1) = \sum_{i=1}^{L} \frac{i}{L} p_i, \quad \text{and} \quad p(2) = \sum_{i=1}^{L} \frac{i}{L} \frac{(i-1)}{(L-1)} p_i, \tag{18}$$

the generalization diversity measure GD is defined as

$$GD = 1 - \frac{p(2)}{p(1)}. \tag{19}$$

GD varies between 0 (minimum diversity when $p(2) = p(1)$) and 1 (maximum diversity when $p(2) = 0$).

4.6. *Coincident failure diversity*

Coincident failure diversity is a modification of *GD* also proposed by Partridge and Krzanowski (1997).

$$CFD = \begin{cases} 0, & p_0 = 1.0; \\ \dfrac{1}{1-p_0} \displaystyle\sum_{i=1}^{L} \frac{L-i}{L-1} p_i, & p_0 < 1 \end{cases} \tag{20}$$

This measure is designed so that it has a minimum value of 0 when all classifiers are always correct or when all classifiers are simultaneously either correct or wrong. Its maximum value 1 is achieved when all misclassifications are unique, i.e., when at most one classifier will fail on any randomly chosen object.

5. Relationship between diversity and accuracy of the ensemble

Table 2 shows a summary of the 10 measures of diversity including their types, and literature sources.

We have to be cautious when designing diversity measures based on oracle outputs. The measures should not be a replacement for the estimate of the accuracy of the team but should stem from the intuitive concept of diversity. For example, the double fault measure (*DF*) is clearly related to the team accuracy as small values of the measure (higher diversity) will favor more accurate teams. On the other hand, *E* is not designed to be directly related to the accuracy of the team. We can show by an example that the relationship between *E* and the accuracy of the majority vote by the team is not predefined by design. Assume that we have 3 classifiers, two of which are the same (D_1 and D_2), and always give the correct

Table 2. Summary of the 10 measures of diversity.

Name		↑ / ↓	P	S	Reference
Q-statistic	Q	(↓)	Y	Y	(Yule, 1900)
Correlation coefficient	ρ	(↓)	Y	Y	(Sneath & Sokal, 1973)
Disagreement measure	D	(↑)	Y	Y	(Ho, 1998; Skalak, 1996)
Double-fault measure	DF	(↓)	Y	N	(Giacinto & Roli, 2001)
Kohavi-Wolpert variance	kw	(↑)	N	Y	(Kohavi & Wolpert, 1996)
Interrater agreement	κ	(↓)	N	Y	(Dietterich, 2000b; Fleiss, 1981)
Entropy measure	Ent	(↑)	N	Y	(Cunningham & Carney, 2000)
Measure of difficulty	θ	(↓)	N	N	(Hansen & Salamon, 1990)
Generalised diversity	GD	(↑)	N	N	(Partridge & Krzanowski, 1997)
Coincident failure diversity	CFD	(↑)	N	N	(Partridge & Krzanowski, 1997)

Note: The arrow specifies whether diversity is greater if the measure is lower (↓) or greater (↑). ‘P’ stands for ‘Pairwise’ and ‘S’ stands for ‘Symmetrical’.

class label, and the third one (D_3) is always wrong (i.e., $\mathbf{y}_1$ and $\mathbf{y}_2$ contain only 1's, and $\mathbf{y}_3$ contains only zeros). Then $E = 1$ (highest diversity), and $P_{maj} = 1$. Assume now that D_2 is replaced by a classifier identical to D_3, so that now $P_{maj} = 0$. But E is still 1, indicating the highest possible diversity.

Ruta and Gabrys (2001) looked into measures of diversity and their relationship with the majority vote accuracy. They raised an interesting point of *symmetry* of the diversity measures. If we ignore the context of the oracle outputs and treat them as two different symbols, the measures of diversity should be symmetrical with respect to swapping the 0 and the 1. Table 2 shows which of the 10 diversity measures are symmetric and which are not. It is perhaps not surprising that later we show that the non-symmetrical measures exhibit a slightly higher correlation with the team accuracy and tend to be grouped together.

It is difficult to judge which measure best expresses the concept of diversity, which leaves question 4, as stated in the introduction, open for further debate. The next section describes a set of experiments to examine the relationship between the measures themselves and between the measures and the team accuracy.

6. Experiments

6.1. *The experimental plan*

There is no standard experimental set up to build a variety of classifier teams. Thus, *the probability* that a team $\mathcal{D}$ occurs in a real-life experiment is a vacuous notion. Then how do we create classifier teams for the purpose of finding out whether there is a relationship between diversity and accuracy of the team?

Section 6.2 gives the results of an experiment where a measure of diversity is chosen in advance (here Q_{av}). We do not have a limit for the theoretical range of the values of Q_{av}. Assuming equal individual accuracy p, which is above a certain threshold depending on L, the range for Q_{av} is from -1 to 1. However, with more than two classifiers, a large negative value for Q_{av} is difficult to achieve in practice. Classifier *outputs* were generated in the form of correct/incorrect votes (the oracle outputs) with predefined pairwise Q. We sample to ensure that all values of Q_{av} are approximately equally represented.

Section 6.3 gives the results of an enumeration of *all possible classifier 0/1 outputs* with accuracy 0.6 for all classifiers. The data set consists of 30 objects which leads to 563 possible classifier teams. In this way, we show the whole possible range of diversities. This time, however, the distribution of the number of teams over the diversity will not be uniform.

Uniform generation of outputs is an unlikely scenario if we build classifier teams in a real-life problem. The same holds for the simulation experiment in Section 6.3. It is reasonable to expect that the individual classifiers will be positively correlated. To find out how related diversity and accuracy are in this case, we used a real data set, the Wisconsin breast cancer data from the UCI Machine Learning Repository Database.

Section 6.4 describes an experiment where we use different subsets of features for each of $L = 3$ classifiers. We built 14700 teams by *enumerating* all possible partitions of the 10 features into three subsets. We considered all subsets of cardinality 4, 3, and 3 (total of 4200 teams) and 4, 4, and 2, (total of 3150 teams). For each team we used three linear or three quadratic discriminant classifiers. This time, the classifier outputs were estimates of the posterior probabilities, so a series of combination methods were applied along with the majority vote.

Finally, in Section 6.5, we studied how average individual accuracy is related to diversity and to the ensemble accuracy. Two data sets were used: Phoneme data (UCI) and Cone-torus data (Kuncheva, 2000). We applied both the bagging (Breiman, 1996) and the random weak-classifiers methods (Ji & Ma, 1997).

A summary of the experiments is given in Table 3.

6.2. *Simulation experiment (uniform distribution)*

A Matlab program was designed to randomly generate L binary classifier outputs $\mathbf{y}_1, \dots, \mathbf{y}_L$ for N objects, so that the individual accuracy is approximately p (i.e., $\sum_j y_{i,j} \approx N \times p,\ i = 1, \dots, L$), and with a (symmetric) matrix of dependencies $Q = [Q_{i,k}]$. The parameters L, N, p and Q are specified by the user. The generating algorithm is beyond the scope of this paper and will be discussed in a separate publication. The experiments were organized as follows.

1. Six sets of experiments were carried out with the individual accuracy $p \in \{0.6, 0.7\}$ and the number of classifiers $L \in \{3, 5, 9\}$. In all experiments $N = 1000$.

2. For each of the 6 combinations of p and L, 15 classifier teams $\mathcal{D}$ were generated for each of the 21 values of the averaged pairwise dependency $Q_{av} \in \{-1.0, -0.9, \dots, 0.9, 1.0\}$, giving a total of 315 classifier teams. In each generation, all off-diagonal elements of the target matrix Q were set to the specified Q_{av}. Generating identical or approximately independent classifiers is trivial unlike generating classifier teams of a fixed accuracy and dependency. The algorithm is not always able to generate data with the target value for Q. To overcome this, we split the interval from -1 to 1 into bins of width 0.1. We tallied the number of Q_{av} in each bin. For the bins with 10 or fewer points, we used all of them. For the bins with more than 10 values of Q_{av} we randomly sampled 10 of them. Thus, the distribution for Q_{av}, although without spanning the whole interval, followed an approximate uniform distribution.

3. To examine the relationship between the measures, rank correlation coefficients were calculated between each pair. The rank correlation coefficient was chosen because the relationships were not linear. The minimum absolute value of the correlation coefficients for the 6 cases of L and p are displayed in Table 4. All these coefficients are high. This is due, in part, to the way we have artificially generated the classifier teams to have Q_{av} spanning the whole range -1 to 1.

A relationship between the measures of diversity was sought by applying a cluster analysis procedure. An *average-linkage* relational clustering was run to find out groups among the diversity measures.[2] The procedure was executed three times, for 2, 3, and 4 clusters respectively, for each of the 6 combinations of L and p. The "distance" matrix needed for the input of the program was formed as one minus the absolute value of the rank correlation. Interestingly, different clusters were found for the different L and p. These results are combined with further results and summarized in Table 8.

4. Let P_{mean} and P_{max} be the observed mean and maximum accuracies respectively of the generated team $\mathcal{D}$. For each combination of L and p, we calculated the correlation between

Table 3. A summary of the experiments: N is the number of objects, L is the number of classifiers, p is the individual accuracy.

	Section 6.2. Simulation
Data type	Generated 0/1 outputs
N	1000
L	3, 5, and 9
No. of teams	315
Classifier type	N/A
p	$\approx$0.6 and $\approx$0.7
Description	Classifier teams are generated and then selected so that Q had an approximately uniform distribution over -1 to 1.
	Section 6.3. Enumeration
Data type	Generated cell counts (0/1 outputs)
N	30
L	3
No. of teams	563
Classifier type	N/A
p	0.6
Description	All partitions of 30 objects into $2^3 = 8$ cells of possible 3 votes of 0 and 1.
	Section 6.4. Feature subspace method
Data type	Breast cancer data, UCI repository
N	569
L	3
No. of teams	14700
Classifier type	LDC and QDC
p	Not specified
Description	The set of first 10 features was partitioned into all possible subsets of 4, 4, 2 and 4, 3, 3
	Section 6.5. Bagging and random weak classifiers
Data type	Phoneme data, UCI repository; Cone-torus data
N	5404 (Phoneme); 800 (Cone-torus)
L	9 with both methods and both data sets
No. of teams	500 with both methods and both data sets
Classifier type	LDC with both methods and both data sets plus neural networks (NN) for bagging
p	Not specified
Description	Random sampling with replacement to form the 9 training data sets for bagging. Randomly generated LDC for the random weak-classifiers method (this intuitively leads to more diverse classifiers than trained classifiers.) 500 splits into training/testing subsets. For the Phoneme data, the split was 1000/4404 and for Cone-torus 400/400.

Table 4. Minimum, by absolute value, rank correlation coefficients between pairs of diversity measures for the simulation experiment with $L = 3$, 5 or 9 individual classifiers, each of individual accuracy $p = 0.6$ or 0.7.

	$p = 0.6$	$p = 0.7$
$L = 3$	0.9880	0.9836
$L = 5$	0.9858	0.9809
$L = 9$	0.9536	0.9644

Table 5. Extreme values of the rank correlation coefficients between measures of diversity and the improvement on the single best accuracy and the mean accuracy of the team for the simulation experiment.

	$P_{maj} - P_{max}$		$P_{maj} - P_{mean}$	
	$p = 0.6$	$p = 0.7$	$p = 0.6$	$p = 0.7$
Minimum	0.9371	0.9726	0.9652	0.9826
Maximum	0.9870	0.9923	0.9909	0.9949

each of $P_{maj} - P_{mean}$ and $P_{maj} - P_{max}$ with the 10 measures. All measures exhibited high (by absolute value) correlation as summarized in Table 5.

The relationship between Q_{av} and $P_{maj} - P_{max}$ is illustrated graphically in figure 2. Each point in the scatterplot corresponds to a classifier ensemble.

Figure 2 shows the relationship between diversity and majority vote accuracy of the team for this experiment. Smaller Q (more diverse classifiers) leads to higher improvement over the single best classifier. Negative Q (negative dependency) is better than independence ($Q = 0$) as it leads to an even bigger improvement.

On the other hand, total positive dependency (identical classifiers, $Q = 1$) is *not the worst case*. The worst case is obtained for positively dependent but not identical classifiers. This agrees with our theoretical limits on majority vote (Kuncheva et al., 2000). To see this phenomenon more clearly, the zero-improvement is marked with a horizontal line in figure 2. The points below the line correspond to classifier ensembles that fared worse than the single best classifier. In all these cases, the ensembles consist of positively related but not identical classifiers.

We have to be cautious when drawing conclusions from experiments. In this case, the correlation was calculated on generated data with the following characteristics:

1. All individual classifiers have approximately the same accuracy (pre-specified).
2. The pairwise dependency was approximately the same (pre-specified).
3. Not all negative values of Q_{av} were possible for all L. This means that the distribution of Q_{av}, which was intended to range uniformly from -1 to $+1$ spanned a shorter range from $-a$ to 1, where $a \in (0, 1)$.

If we generated the classifier outputs randomly, the range of all measures of diversity would be much smaller. Instead of being uniform, the distribution of Q_{av} would peak at

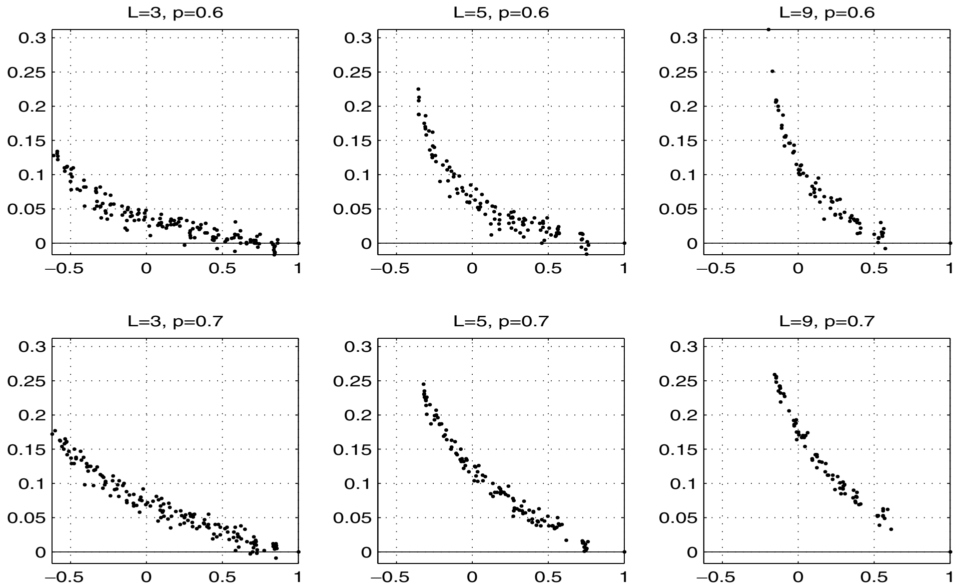

Figure 2. Plot of $(P_{maj} - P_{max})$ versus Q_{av} for the simulation experiment.

0, and the correlation might not show up as it did in our experimental set up. There is no "true" distribution of Q_{av} as it depends on the classifier design procedure.

6.3. *Enumeration experiment*

To look at a wider picture, we carried out an enumeration experiment with $L = 3$, $p = 0.6$ and $N = 30$. *All* possible distributions of the classifier votes for the 30 elements of $\mathbf{Z}$ were generated so that each of the three classifiers recognizes exactly $0.6 \times 30 = 18$ objects. In other words, we generated all partitions of 30 into $2^3 = 8$ cells, each cell corresponding to a combination of votes. For example, one possible partition is shown below

D_1, D_2, D_3	111	101	011	001	110	100	010	000
Number of objects	7	1	2	8	7	3	2	0

The majority vote is $(7 + 1 + 2 + 7)/30 \approx 0.5667$, and the diversity measures are

$$\begin{aligned} Q_{av} &= -0.1580 & E &= 0.1704 \\ \rho_{av} &= -0.0648 & KW &= 0.7767 \\ Dis_{av} &= 0.5111 & \kappa &= -0.0648 \\ DF_{av} &= 0.1444 & \theta &= 0.0696 \\ & & GD &= 0.6389 \\ & & CFD &= 0.7174 \end{aligned}$$

The three classifiers in this example exhibit negative correlation (indicating high diversity) but the majority vote is less than the individual accuracy of 0.6.

In summary, the characteristics of this set of experiments are

1. All three individual classifier have exactly the same accuracy, $p = 0.6$.
2. The pairwise dependency is not restricted. As the example shows, classifiers of very different pairwise Q are generated too.
3. *All* possible classifier outputs with $L = 3$, $p = 0.6$ and $N = 30$ were generated (no special arrangements for uniform distribution of Q or any other measures have been made). The total number of possible classifier ensembles is 563.

The pairwise rank correlation coefficients between the measures were calculated. The lowest coefficient by absolute value was that between Q and *CFD*, -0.9406. Except for rank correlations involving either of these two measures, all other correlations were either 1 or -1. This uniformity suggests that the measures of diversity behave exactly in the same way, hence further cluster analysis was not necessary.

The rank correlations between P_{maj} (same as $P_{maj} - 0.6$) and the diversity measures were -0.5127 for Q_{av}, 0.7379 for CFD, and identical, with absolute value of 0.5210, for the other eight measures. Figure 3 shows the scatterplot of P_{maj} versus Q_{av} (left), ρ_{av} (middle) and *CFD* (right). The other seven patterns were either the same as that of ρ_{av}, or a mirror image of it. The horizontal line at 0.6 marks the (best) individual accuracy.

This experiment shows that

1. The diversity measures were very similar to each other.
2. When the dependency between the pairs of classifiers in $\mathcal{D}$ is not approximately constant, the relationship between the diversity measures and the improvement offered by the majority vote deteriorates. The rank correlation coefficients in this case are much lower than when pairwise Qs are approximately equal. Therefore, if we use a diversity measure in the design of a classifier team, we have to take into account that if classifier pairs have substantially different pairwise dependencies, the diversity measure might not be useful.

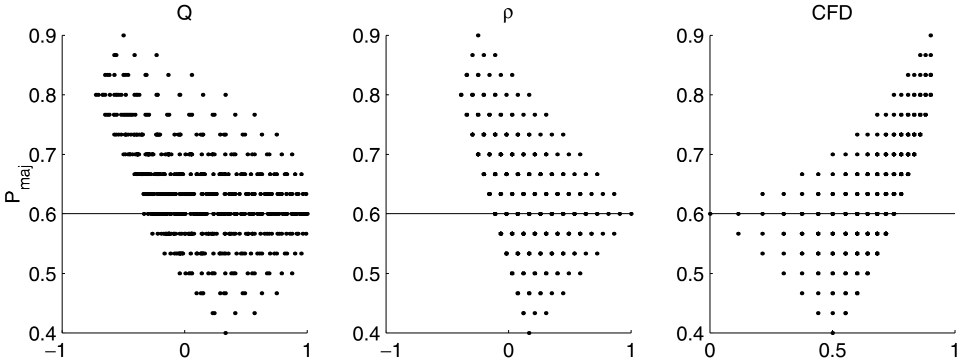

Figure 3. Scatterplot of P_{maj} versus Q_{av} (left), ρ_{av} (middle) and *CFD* (right).

3. There is a threshold value of each diversity measure such that above that value, the improvement over the individual error rate is *guaranteed*, irrespective of whether pairwise dependencies are symmetrical or not.

6.4. *Feature subspace method*

We used the Wisconsin Diagnostic Breast Cancer data base[3] taken from the UCI Repository of Machine Learning Database.[4] The set consists of 569 patient vectors with features computed from a digitized image of a fine needle aspirate of a breast mass. The breast masses are grouped into two classes: benign and malignant. Out of the original 30 features we used the first 10; these were the means of the relevant variables calculated in the image. The study was confined to 10 variables to enable a reasonable enumerative experiment. The data set was split randomly into two halves, one being used for training and one for testing. We considered $L = 3$ classifiers. All partitions of the 10-element feature set into $\langle 4, 4, 2\rangle$ (3150 partitions) and $\langle 4, 3, 3\rangle$ (4200 partitions) were generated. For each partition, three classifiers were built, one on each subset of features. Two simple classifier models were tried, linear and quadratic classifier, leading to 4 sets of experiments:

1. $\langle 4, 4, 2\rangle$ with linear classifiers;
2. $\langle 4, 4, 2\rangle$ with quadratic classifiers;
3. $\langle 4, 3, 3\rangle$ with linear classifiers;
4. $\langle 4, 3, 3\rangle$ with quadratic classifiers.

A preliminary study suggested that the four experiments were not different in any important respect, and a pooled data set of a total of 14700 teams was formed. The results in this and the next section are calculated on the test set, unseen at any stage during training of the individual classifiers or the ensemble. Based on this data, the rank correlations between the measures of diversity were calculated. Unlike in the previous experiments, the rank correlations between some of the measures were not very high. The average-linkage relational clustering algorithm was run. Two distinct groups were found: DF and θ in the one group, and the remaining measures in the other. Figure 4 shows in a pictorial form the relationships between the 10 measures. It is organized as a correlation matrix but instead of the correlation coefficients, a scatterplot of the respective entries is shown. Table 6 shows the correlation coefficients corresponding to the plots.

Since the classifiers had continuous valued outputs, we applied other combination methods besides the majority vote (MAJ) (Kuncheva et al., 2001). These were Naive Bayes (NB), the Behavior Knowledge Space method (BKS) (Huang & Suen, 1995); maximum (MAX), minimum,[5] average (AVR) and product (PRO) from the simple methods; and the decision templates (DT) method (Kuncheva et al., 2001). Table 7 shows the rank correlation coefficients between the measures and the diversity measures and the *improvement* of the accuracy, i.e., $P_{team} - P_{max}$. The values for E and KW are omitted as they are the same as Dis.

The absolute value of the correlation coefficients in Table 7 are so low that the diversity measures considered here have no useful predictive value. Shown in figure 5 is the

Table 6. Rank correlation coefficients (in %) between the diversity measures for the breast cancer data experiment.

	ρ	*Dis*	*DF*	*KW*	κ	*E*	θ	*GD*	*CFD*
Q	98	−97	65	−97	98	−97	36	−98	−95
ρ		−97	72	−97	100	−97	44	−100	−97
Dis			−56	100	−98	100	−26	98	96
DF				−56	70	−56	93	−72	−70
KW					−98	100	−26	98	96
κ						−98	41	−100	−98
E							−26	98	96
θ								−44	−42
GD									98

Table 7. Correlation in % between the improvement on the single best accuracy ($P_{team} - P_{max}$) and the 10 measures of diversity for the breast cancer experiment.

	Q	ρ	*Dis*	*DF*	κ	θ	*GD*	*CFD*
MAJ	−17	−21	33	18	−20	35	28	38
NB	−15	−20	32	20	−18	37	26	36
BKS	−15	−17	17	5	−15	16	18	18
WER	−15	−17	17	5	−16	17	19	18
MAX	−1	−0	20	38	0	45	7	11
AVR	−13	−15	34	33	−14	47	22	30
PRO	−11	−11	29	33	−11	44	18	24
DT	−12	−15	32	30	−14	44	22	29

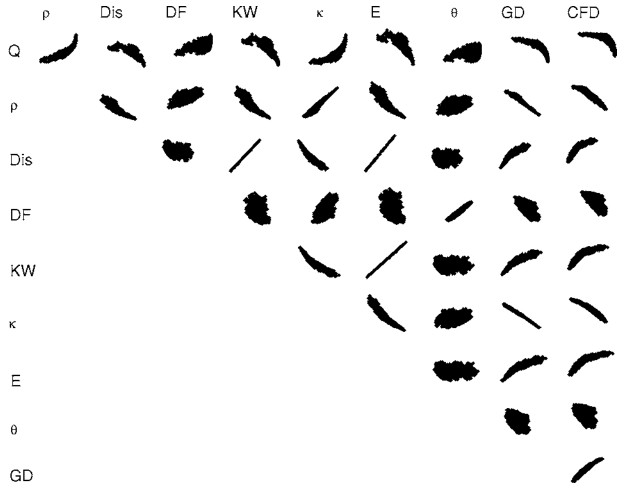

Figure 4. Pairwise scatterplots of the 10 measures of diversity for the breast cancer experiment.

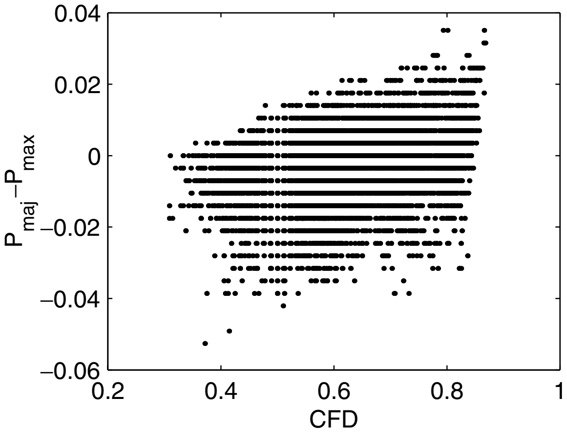

Figure 5. Scatterplot of the improvement of the majority vote combination over the single best classifier ($P_{maj} - P_{max}$) versus the Coincident Failure Diversity (*CFD*) for the breast cancer experiment.

scatterplot of $P_{maj} - P_{max}$ against *CFD* (one of the two "highest" correlations (38%) in the table).

The failure of the diversity measures could be attributed to the nature of the experiment. First, the improvement over the single best classifier was small anyway. It is possible that such small improvements (or deteriorations) cannot be detected well by the diversity measures. Second, only three classifiers were considered, while the number of classifiers L could also be a factor in the predictive values of the diversity measures. Third, no restrictions were imposed on either individual accuracy or mutual dependency between classifiers. Thus, the difference in the results of this experiment to the previous two is maybe not surprising.

6.5. *Bagging and random weak classifiers*

For this experiment we used two data sets.

- Phoneme data: Five-dimensional data characterizing two classes of phonemes: nasals (70.65%) and orals (29.35%).
- Cone-torus data: A three-class dataset with 400 2-d points generated from three differently shaped distributions: a cone, half a torus, and a normal distribution with prior probabilities 0.25, 0.25, and 0.5, respectively.[6]

Each data set was randomly split 500 times into training/testing. For the Phoneme data, the split was 1000/4404 and for Cone-torus data, 400/400.

We chose $L = 9$ classifiers for both methods.

For bagging, we sampled from the training subset with replacement 9 times to obtain the training data for the members of the ensemble. Two classifier models were used, linear and neural network (NN). The NN was implemented as a multi-layer perceptron with one hidden layer containing 15 neurons. Fast backpropagation training was used, as implemented in

Table 8. Coincidence matrix illustrating the results from 36 clusterings of the diversity measures from Sections 6.2 and 6.4. The higher the number, the more similar the measures.

	ρ	*Dis*	*DF*	*KW*	κ	*E*	θ	*GD*	*CFD*
Q	30	23	1	23	30	23	18	21	6
ρ		28	4	28	36	28	22	26	8
Dis			4	36	28	36	17	21	7
DF				4	4	4	9	5	11
KW					28	36	17	21	7
κ						28	22	26	8
E							17	21	7
θ								22	10
GD									13

Matlab Neural Network Toolbox. The training of each single neural network was stopped after 300 epochs.

For the random weak-classifiers method, 9 *weak* classifiers were generated by randomly assigning the coefficients of the linear discriminant functions. A classifier was accepted for the team if the training accuracy was greater than the highest prior probability (estimated from the training set).

As for the simulation experiment, the measures were clustered with the average linkage procedure into 2, 3, and 4 clusters. The procedure was run for the 6 combinations of a data set, classifier model and ensemble construction method. To summarize the results from the clusterings of the diversity measures, we introduce the coincidence matrix $C = \{c_{m_1,m_2}\}$ (Table 8) where m_1 and m_2 stand for diversity measures, and c_{m_1,m_2} is the number of times the two measures were grouped in the same cluster. There were 6×3 different clustering results from Section 6.2, and 6×3 obtained here, so the maximum possible entry in the coincidence matrix would be 36. This happens, for example, for ρ and κ, indicating their similarity.

The difference in the grouping can be explained by the fact that all correlation coefficients were high, and small (noise) differences could lead to different grouping. Nevertheless, it is interesting to observe the general tendency for the symmetrical and non-symmetrical measures to be similar. As discussed earlier, the non-symmetrical measures tend here to form their own group. The double fault measure *DF* was the first to seed a separate cluster in most of the experiments. Also, the equivalence between *KW* and *Dis*, and the similarity to *E* showed up as the three were consistently standing as a separate cluster for $p \approx 0.7$. If we "cut" the coincidence matrix at, say, 15 co-occurrences, the clusters for the diversity measures would be

$$\boxed{DF}\,\boxed{CFD}\,\boxed{Q\ \ \rho\ \ Dis\ \ KW\ \ \kappa\ \ E\ \ \theta\ \ GD}$$

The rank correlation between diversity on the one hand and accuracy on the other hand was calculated. Table 9 shows the average and the extreme values of the rank correlation coefficients: the minimum by absolute value rank correlation coefficients between diversity

Table 9. Averaged and extreme values of the rank correlation coefficients for the bagging and weak-classifiers experiments. For diversity/diversity, the minima by absolute values are shown, for diversity/accuracy, the maxima by absolute values are shown.

	Bagging (linear)	Bagging (NN)	Weak classifiers
Phoneme data			
Average (Div/Div)	0.90	0.81	0.97
Extreme (Div/Div)	$-0.66\,(E/DF)$	$-0.28\,(Q/CFD)$	$0.87\,(Q/DF)$
Average (Div/Acc)	0.24	0.29	0.11
Extreme (Div/Acc)	$-0.43\,(DF/MAJ)$	$-0.95\,(DF/MAJ)$	$-0.54\,(DF/MAJ)$
Cone-torus data			
Average (Div/Div)	0.86	0.56	0.77
Extreme (Div/Div)	$-0.54\,(E/DF)$	$-0.06\,(E/GD)$	$0.21\,(E/CFD)$
Average (Div/Acc)	0.16	0.30	0.08
Extreme (Div/Acc)	$-0.29\,(DF/NB)$	$-0.85\,(E/PRO)$	$0.70\,(CFD/MAJ)$

measures, and the maximum by absolute value rank correlation coefficients between diversity and accuracy. Both of the bagging and the weak-classifiers experiments confirmed the results obtained in the feature subspace experiment: the relationships among the measures themselves are strong whereas the relationship between diversity and accuracy is weak. In general, the relationship between accuracy and diversity was strongest for the *DF* and *CFD* measures with the majority vote accuracy P_{maj}. Occasionally other diversity measures showed high rank correlations when a particular combination method was used. The high correlation in Table 9 of the entropy measure (E) and the accuracy obtained using the product combination (*PRO*) is one of these and should probably be ignored.

The objective of this experiment however was to investigate how individual accuracy is related to ensemble accuracy and ensemble diversity. Figure 6 illustrates the relationship in the following manner. The scatterplots show Q versus the averaged *individual* accuracy of the ensembles $\bar{p}$. Each point on the plot corresponds to one of the 500 ensembles generated for the respective selection of a data set and an ensemble construction method.

We also illustrate the triple relationship between Q, $\bar{p}$ and P_{maj} in figure 6. The points printed with '$+$' denote ensembles where P_{maj} is a substantial improvement on $\bar{p}$ ($P_{maj} > \bar{p} + \epsilon$), and points printed with $\bullet$ correspond to ensembles which are worse than the individual average ($P_{maj} > \bar{p} - \epsilon$). The gray points on the background correspond to ensembles in the middle. The threshold ϵ was taken to be one standard deviation of the distribution of accuracies $\bar{p}$.

The results shown in figure 6 are somewhat unexpected. It seems that there is little or no relationship between the averaged individual accuracy $\bar{p}$ and diversity Q. The intuition that less accurate classifiers will form more diverse ensembles is not supported by this experiment. This finding counters the observation by Dietterich (2000b) where the κ-error diagrams indicate a certain trade-off between diversity and accuracy for three ensemble design methods. We can offer two lines of explanation of this discrepancy.

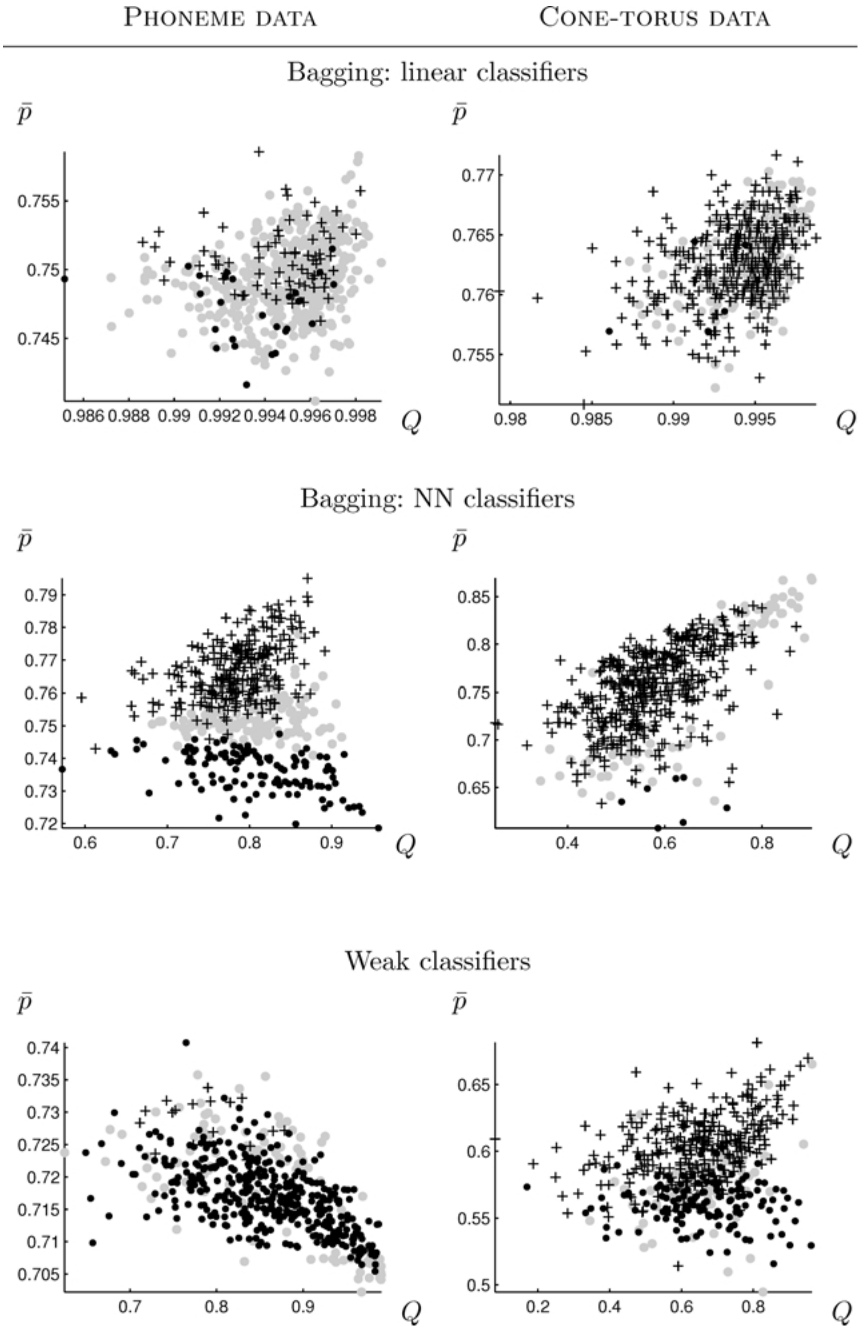

Figure 6. Scatterplot of Q versus individual accuracy of the ensemble $\bar{p}$. '+' denote ensembles whose majority vote was better than the individual average, '•' denote "worse" ensembles, and the gray dots denote ensembles in the middle.

First, the relationship might be dependent upon the diversity measure employed. We used an averaged measure for the whole ensemble, Q_{av}, and an averaged accuracy across the L classifiers. Dietterich (2000b) only considers pairs of classifiers, and not the ensemble as a whole. Thus, the variability of the individual accuracy within the ensemble could be responsible for the blurring of the trade-off pattern. This argument leads us again to question 1: how do we define and measure diversity. The explanation of the κ-error diagram (Dietterich, 2000b) suggests that diversity is perceived as statistical independence ($\kappa = 0$) but here we look also at alternative interpretations as discussed previously.

Second, as also pointed out by Dietterich (2000b), different methods for building classifier ensembles might lead to different relationship patterns.

What is even more discouraging is the lack of evidence of a relationship between the three variables. For the same $\bar{p}$ (an imaginary line parallel to the x-axis), we would expect that higher diversity will lead to higher ensemble accuracy. Thus, for smaller values of Q (better diversity), we would expect to see $+$'s and for the larger values, we would expect $\bullet$'s. Unfortunately, this pattern is not apparent.

7. Conclusions

We studied ten measures of diversity between the classifiers in an ensemble: four pairwise and six non-pairwise measures (some taken directly and some derived from the literature). Wherever a link between measures was spotted, it was carried forward to disclosing the formal relationship between the measures (e.g., between Dis_{av}, KW, and κ, and also between the "global" and the pairwise definition of κ).

Attempting to answer questions 2 and 3 from the introduction, four experimental setups were used to investigate the relationship between the diversity measures and the improvement of the majority vote (P_{maj}) over the single best classifier (P_{max}) and the average of the ensemble (P_{mean}).

In the simulation experiments, we generated classifier outputs (correct/incorrect votes) for $L \in \{3, 5, 9\}$ classifiers and accuracies $p \in \{0.6, 0.7\}$. The idea was to span the largest possible interval for the diversity measures and generate ensembles with classifiers of approximately the same accuracy (p) and symmetrical pairwise dependency (i.e., the dependency between classifiers D_i and D_k is approximately the same as between any other pair from the ensemble). We found that all measures had approximately equally strong relationships with the improvement of the majority vote. Also, the measures were strongly correlated between themselves.

In the enumeration experiments, we generated all possible classifier outputs for $L = 3$, $p = 0.6$ and $N = 30$. Thus, the symmetry of the pairwise dependency was no longer a restriction. We found that the relationship between the diversity measures and $P_{maj} - P_{mean}$ weakens because ensembles of the same individual accuracy (p) and the same diversity, could result in very different P_{maj} due to the imbalance of the pairwise dependencies. However, it was also found that even for the case of unequal pairwise dependency, there are threshold values for all diversity measures, such that sets of ensemble with diversity higher than the threshold offer a guaranteed improvement over the single best classifier. Again, the mutual relationships between the diversity measures were found to be strong.

In the third experiment we used the breast cancer data and enumerated all partitions of the set of 10 features into subsets of 4, 3, 3 and 4, 4, 2. Each subset was used as a classifier input. All teams of three linear or three quadratic classifiers were thereby generated, forming a set of 14700 classifier teams. The results from this experiment revealed the possible inadequacy of the diversity measures for predicting the improvement on the best individual accuracy. The low absolute values of the correlation coefficients between the measures on the one hand and the improvement, on the other hand, is discouraging. It raises doubts about the abilities of diversity measures to pick up small improvements and deteriorations. Practical problems are likely to be exactly in this class. This deficiency was confirmed in the fourth experiment where 9 classifiers were used with two data sets: Phoneme data and Cone-torus data. Two ensemble designing methods were tried: bagging and random weak classifiers, to the same effect. It appeared that there is no clear relationship between diversity and the averaged individual accuracy, which is counterintuitive, much as most of the other findings in our study.

We note here that our conjectures are based on a set of experiments, and it is possible that in different circumstances, diversity and accuracy could exhibit a stronger relationship. For example, incremental ensemble building through Adaboost (Freund & Schapire, 1997) is a methodology especially designed to enforce diversity in the ensemble. The ensembles designed through Adaboost consist of classifiers of greater variability in their overall performance but with good complementary behavior (Dietterich, 2000b). This variablity can lead to a larger spread of the values of the diversity measures, and consequently to a more definite relationship with accuracy. Our results in Section 6.2 indicate the possibility of such a relationship.

The 10 measures exhibited reasonably strong relationships among themselves. In the real classifier experiments, the uniformly high correlation was preserved between the majority of the measures. In these experiments, three distinct clusters of measures were discovered: the double fault measure (*DF*) and the coincident failure diversity (*CFD*) formed clusters on their own, and all the remaining measures formed the third cluster. This suggests that for real problems, the measures might behave differently, so they can be used as a complementary set. The more trivial conclusion that nonetheless might have a greater impact on the development of this research line is that the notion of diversity is not clear-cut. On the other hand, the great similarity between measures used by different authors suggests that there is an agreed upon general idea of diversity, at least on some intuitive level.

At this stage any answers to questions 4: "*Is there a measure that is best for the purposes of developing committees that minimize error?*", and 5: "*How can we use the measures in designing the classifier ensemble?*", can only be speculative.

A choice of measure to recommend can be based on ease of interpretation. Only Q_{av}, ρ_{av} and κ have the simple value of 0 for independence with negative values giving what the team should be striving for. Both ρ_{av} and κ have a minimum (negative) value which depends on the number of classifiers and their accuracies, whereas for Q_{av} it is unknown although our experiments suggest there is a minimum value. The relationship between Q_{av} and the limits of P_{maj} has been formally derived (Kuncheva et al., 2000), and for these limit cases, Q_{av} can take values -1 if the individual classifiers are all of accuracy $p \geq \frac{2}{3}$. Finally, Q_{av} is easy to calculate. For these reasons we recommend the use of Q_{av}.

The use of diversity measures for enhancing the design of classifier ensembles (question 5) is still an open question. Given the lack of a definitive connection between the measures and the improvement of the accuracy, it is unclear whether measures of diversity will be of any practical value at this stage. This prompts the question, is diversity such a key issue in classifier combination, as deemed in the literature? Our study suggests that the general motivation for designing diverse classifiers is correct but the problem of measuring this diversity and so using it effectively for building better classifier teams is still to be solved.

Appendix: Equivalence between the disagreement measure Dis_{av} and Kohavi-Wolpert variance KW

The Kohavi-Wolpert (1996) variance, in the case of two alternatives, 0 and 1, is

$$KW = \frac{1}{NL^2}\sum_{j=1}^{N} l(\mathbf{z}_j)(L - l(\mathbf{z}_j)) \tag{21}$$

$$= \frac{1}{NL^2}\sum_{j=1}^{N}\left(\sum_{i=1}^{L} y_{j,i}\right)\left(L - \sum_{i=1}^{L} y_{j,i}\right) = \frac{1}{NL^2}\sum_{j=1}^{N}\mathcal{A}_j, \tag{22}$$

where

$$\mathcal{A}_j = \left(\sum_{i=1}^{L} y_{j,i}\right)\left(L - \sum_{i=1}^{L} y_{j,i}\right). \tag{23}$$

The disagreement measure between D_i and D_k used in (Skalak, 1996) can be written as

$$Dis_{i,k} = \frac{1}{N}\sum_{j=1}^{N}(y_{j,i} - y_{j,k})^2. \tag{24}$$

Averaging over all pairs of classifiers i, k,

$$Dis_{av} = \frac{1}{L(L-1)}\sum_{i=1}^{L}\sum_{\substack{k=1\\ i\neq k}}^{L}\frac{1}{N}\sum_{j=1}^{N}(y_{j,i} - y_{j,k})^2 \tag{25}$$

$$= \frac{1}{NL(L-1)}\sum_{j=1}^{N}\sum_{i=1}^{L}\sum_{\substack{k=1\\ i\neq k}}^{L}(y_{j,i} - y_{j,k})^2$$

$$= \frac{1}{NL(L-1)}\sum_{j=1}^{N}\mathcal{B}_j, \tag{26}$$

where

$$\mathcal{B}_j = \sum_{i=1}^{L} \sum_{\substack{k=1 \\ i \neq k}}^{L} (y_{j,i} - y_{j,k})^2. \tag{27}$$

Dropping the index j for convenience and noticing that $y_i^2 = y_i$,

$$\mathcal{A} = L\left(\sum_{i=1}^{L} y_i\right) - \left(\sum_{i=1}^{L} y_i\right)^2 \tag{28}$$

$$= L\left(\sum_{i=1}^{L} y_i\right) - \left(\sum_{i=1}^{L} y_i^2\right) - \left(\sum_{i=1}^{L} \sum_{\substack{k=1 \\ i \neq k}}^{L} y_i y_k\right) \tag{29}$$

$$= (L-1)\left(\sum_{i=1}^{L} y_i\right) - \left(\sum_{i=1}^{L} \sum_{\substack{k=1 \\ i \neq k}}^{L} y_i y_k\right). \tag{30}$$

On the other hand,

$$\mathcal{B} = \sum_{i=1}^{L} \sum_{\substack{k=1 \\ i \neq k}}^{L} \left(y_i^2 - 2y_i y_k + y_k^2\right) \tag{31}$$

$$= 2(L-1)\left(\sum_{i=1}^{L} y_i\right) - 2\left(\sum_{i=1}^{L} \sum_{\substack{k=1 \\ i \neq k}}^{L} y_i y_k\right) \tag{32}$$

$$= 2\mathcal{A}. \tag{33}$$

Therefore,

$$KW = \frac{L-1}{2L} Dis_{av}. \tag{34}$$

Since the two diversity measures differ by a coefficient, their correlation with $P_{maj} - P_{mean}$ will be the same.

Notes

1. Averaged pairwise correlations between $P_i(\omega_s \mid \mathbf{x})$ and $P_j(\omega_s \mid \mathbf{x})$, $i, j = 1, \ldots, L$ are calculated for every s, then weighted by the prior probabilities $\hat{P}(\omega_s)$ and summed.
2. The package PRtools for Matlab was used (Duin, 1997).
3. Created by Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian, University of Wisconsin.
4. http://www.ics.uci.edu/~mlearn/MLRepository.html.
5. It can be shown that maximum and minimum combination methods coincide when there are two classes and the classifier outputs sum up to 1.

6. Available on http://www.bangor.ac.uk/~mas00a/Z.txt (see for more experimental results with the same data (Kuncheva, 2000)). A separate data set for testing with 400 more points generated from the same distribution is also available: Zte.txt.

References

Afifi, A., & Azen, S. (1979). *Statistical analysis. A computer oriented approach*. New York: Academic Press.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning, 36*, 105–142.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 26:2*, 123–140.

Breiman, L. (1999). Combining predictors. In A. Sharkey (Ed.), *Combining artificial neural nets* (pp. 31–50). London: Springer-Verlag.

Cunningham, P., & Carney, J. (2000). Diversity versus quality in classification ensembles based on feature selection. Technical Report TCD-CS-2000-02, Department of Computer Science, Trinity College Dublin.

Dietterich, T. (2000a). Ensemble methods in machine learning. In J. Kittler, & F. Roli (Eds.), *Multiple classifier systems*, Vol. 1857 of Lecture Notes in Computer Science (pp. 1–15). Cagliari, Italy, Springer.

Dietterich, T. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning, 40:2*, 139–157.

Drucker, H., Cortes, C., Jackel, L., LeCun, Y., & Vapnik, V. (1994). Boosting and other ensemble methods. *Neural Computation, 6*, 1289–1301.

Duin, R. (1997). *PRTOOLS (Version 2). A Matlab toolbox for pattern recognition*. Pattern Recognition Group, Delft University of Technology.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. John Wiley & Sons.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55:1*, 119–139.

Giacinto, G., & Roli, F. (2001). Design of effective neural network ensembles for image classification processes. *Image Vision and Computing Journal*, *19:9/10*, 699–707.

Hansen, L., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12:10*, 993–1001.

Hashem, S. (1999). Treating harmful collinearity in neural network ensembles. In A. Sharkey (Ed.), *Combining artificial neural nets* (pp. 101–125). London: Springer-Verlag.

Ho, T. (1998). The random space method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20:8*, 832–844.

Huang, Y., & Suen, C. (1995). A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 17*, 90–93.

Ji, C., & Ma, S. (1997). Combination of weak classifiers. *IEEE Transactions on Neural Networks*, *8:1*, 32–42.

Kohavi, R., & Wolpert, D. (1996). Bias plus variance decomposition for zero-one loss functions. In L. Saitta (Ed.), *Machine Learning: Proc. 13th International Conference* (pp. 275–283). Morgan Kaufmann.

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation and active learning. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems* (Vol. 7, pp. 231–238). Cambridge, MA: MIT Press.

Kuncheva, L. (2000). *Fuzzy classifier design*. *Studies in Fuzziness and Soft Computing*. Heidelberg: Springer Verlag.

Kuncheva, L., Bezdek, J., & Duin, R. (2001). Decision templates for multiple classifier fusion: An experimental comparison. *Pattern Recognition, 34:2*, 299–314.

Kuncheva, L., Whitaker, C., Shipp, C., & Duin, R. (2000). Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*. accepted.

Lam, L. (2000). Classifier combinations: Implementations and theoretical issues. In J. Kittler, & F. Roli (Eds.), *Multiple classifier systems*, Vol. 1857 of Lecture Notes in Computer Science (pp. 78–86). Cagliari, Italy, Springer.

Littlewood, B., & Miller, D. (1989). Conceptual modeling of coincident failures in multiversion software. *IEEE Transactions on Software Engineering, 15:12*, 1596–1614.

Liu, Y., & Yao, X. (1999). Ensemble learning via negative correlation. *Neural Networks, 12*, 1399–1404.

Looney, S. (1988). A statistical technique for comparing the accuracies of several classifiers. *Pattern Recognition Letters*, *8*, 5–9.

Opitz, D., & Shavlik, J. (1999). A genetic algorithm approach for creating neural network ensembles. In A. Sharkey (Ed.), *Combining artificial neural nets* (pp. 79–99). London: Springer-Verlag.

Parmanto, B., Munro, P., & Doyle, H. (1996). Reducing variance of committee prediction with resampling techniques. *Connection Science, 8:3/4*, 405–425.

Partridge, D., & Krzanowski, W. J. (1997). Software diversity: Practical statistics for its measurement and exploitation. *Information & Software Technology, 39*, 707–717.

Rosen, B. (1996). Ensemble learning using decorrelated neural networks. *Connection Science, 8:3/4*, 373–383.

Ruta, D., & Gabrys, B. (2001). Application of the evolutionary algorithms for classifier selection in multiple classifier systems with majority voting. In J. Kittler, & F. Roli (Eds.), *Proc. Second International Workshop on Multiple Classifier Systems*, Vol. 2096 of Lecture Notes in Computer Science (pp. 399–408). Cambridge, UK. Springer-Verlag.

Schapire, R. (1999). Theoretical views of boosting. In *Proc. 4th European Conference on Computational Learning Theory* (pp. 1–10).

Sharkey, A., & Sharkey, N. (1997). Combining diverse neural nets. *The Knowledge Engineering Review, 12:3*, 231–247.

Skalak, D. (1996). The sources of increased accuracy for two proposed boosting algorithms. In *Proc. American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*.

Sneath, P., & Sokal, R. (1973). *Numerical Taxonomy*. W.H. Freeman & Co.

Tumer, K., & Ghosh, J. (1996a). Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition, 29:2*, 341–348.

Tumer, K., & Ghosh, J. (1996b). Error correlation and error reduction in ensemble classifiers. *Connection Science, 8:3/4*, 385–404.

Tumer, K., & Ghosh, J. (1999). Linear and order statistics combiners for pattern classification. In A. Sharkey (Ed.), *Combining artificial neural nets* (pp. 127–161). London: Springer-Verlag.

Yule, G. (1900). On the association of attributes in statistics. *Phil. Trans., A, 194*, 257–319.

Received July 7, 2000
Revised March 27, 2002
Accepted April 23, 2002
Final manuscript April 25, 2002