ELSEVIER

Information Fusion 3 (2002) 135–148

INFORMATION FUSION

www.elsevier.com/locate/inffus

Relationships between combination methods and measures of diversity in combining classifiers

Catherine A. Shipp *, Ludmila I. Kuncheva

School of Informatics, University of Wales, Bangor, Gwynedd, LL57 1UT, UK

Received 26 June 2001; received in revised form 7 November 2001; accepted 7 December 2001

Abstract

This study looks at the relationships between different methods of classifier combination and different measures of diversity. We considered 10 combination methods and 10 measures of diversity on two benchmark data sets. The relationship was sought on ensembles of three classifiers built on all possible partitions of the respective feature sets into subsets of pre-specified sizes. The only positive finding was that the Double-Fault measure of diversity and the measure of difficulty both showed reasonable correlation with Majority Vote and Naive Bayes combinations. Since both these measures have an indirect connection to the ensemble accuracy, this result was not unexpected. However, our experiments did not detect a consistent relationship between the other measures of diversity and the 10 combination methods. © 2002 Published by Elsevier Science B.V.

Keywords: Combining classifier; Diversity; Dependence

1. Introduction

Combining classifiers is an established research area shared between statistical pattern recognition and machine learning. It is variously known as committees of learners, mixtures of experts, classifier ensembles, multiple classifier systems, consensus theory, etc. If we have many different classifiers, it is sensible to consider using them in a combination in the hope of increasing the overall accuracy [1]. It is intuitively accepted that classifiers to be combined should be *diverse*. If they were identical, we could not gain any improvement by combining them. Therefore, *diversity* (negative dependence, independence, orthogonality, complementarity) among the team has been recognised as a key issue [2,3]. Theoretically, a group of independent classifiers improve upon the single best classifier when majority vote combination is used. A dependent set of classifiers may be either better than the independent set or worse than the single worst member of the team, so diversity can be both beneficial or harmful [4,5].

Several techniques exist which aim to improve the performance of classifier ensembles by manipulating the data set on which classifiers are trained. These include Bagging, Boosting, and Arcing [6,7] which can be perceived as guidelines in constructing classifier ensembles. The superiority of these ensemble-building techniques over a simple pooling of independently trained classifiers is attributed to boosting the classification margins [9] and reducing the variance of the error [6]. On the other hand, it has been found that Boosting can be paralysed [10], i.e., no further improvement is achieved when adding new classifiers to the team. It may be that these methods are in some way altering the diversity of the classifiers in the ensemble and this is the key to their success (or failure in some cases). We are interested in whether there is any connection between combination accuracy and diversity in the ensemble.

The proven relationship to date is the result due to Tumer and Ghosh [11,12] who showed that under certain assumptions, the *averaging* combination method produces accuracy which is related to the correlation between the classifier outputs. They extended this result to show similar relationship for combination by order statistics: *minimum*, *maximum*, *mean* [13]. In a previous study, we proved that there is a functional relationship between the Q statistic and the upper and the lower limits of the *majority vote* accuracy [14]. However, there is no theoretical proof of any relationship in the general case. Some authors have used a measure of correlation of the outputs to enforce diversity in the ensemble

* Corresponding author. Tel.: +44-1248-38-3661; fax: +44-1248-38-3663.

E-mail addresses: map802@bangor.ac.uk, mas00a@bangor.ac.uk (C.A. Shipp).

1566-2535/02/$ - see front matter © 2002 Published by Elsevier Science B.V.
PII: S1566-2535(02)00051-9

during training of the component classifiers. The negative correlation training of neural networks was developed [15–18] which showed promising practical results for both regression and classification. Again, the *average* combination method was used. Ensembles built through random subspace method and aggregated using *majority vote* are reported to correlate well with a measure of diversity based on entropy [3,19].

In this study, we examine several widely used combination methods and several diversity measures to try to establish whether or not there exists a relationship

- amongst the 10 combination methods;
- amongst the 10 diversity measures;
- between each of the 10 combination methods and each of the 10 diversity measures.

We first introduce 10 combination methods (Section 2) and 10 measures of diversity (Section 3), and then study their relationship experimentally (Sections 5 and 6). Finally we present our conclusions and ideas for future study (Section 7).

2. Combination methods

Let $\mathscr{D} = \{D_1, D_2, \ldots, D_L\}$ be a set of classifiers and $\Omega = \{\omega_1, \ldots, \omega_c\}$ be a set of class labels. Each classifier gets as its input a feature vector $\mathbf{x} \in \mathbb{R}^n$. The classifier output is a c-dimensional vector $D_i(\mathbf{x}) = [d_{i,1}(\mathbf{x}), \ldots, d_{i,c}(\mathbf{x})]^{\mathrm{T}}$, where $d_{i,j}(\mathbf{x})$ is the degree of "support" given by classifier D_i to the hypothesis that $\mathbf{x}$ comes from class ω_j, $j = 1, \ldots, c$. Without loss of generality we can restrict $d_{i,j}(\mathbf{x})$ within the interval $[0, 1]$, $i = 1, \ldots, L$, $j = 1, \ldots, c$, and call the classifier outputs "soft labels". Most often $d_{i,j}(\mathbf{x})$ is an estimate of the posterior probability $P(\omega_i|\mathbf{x})$.

Combining classifiers means we combine the L classifier outputs $D_1(\mathbf{x}), \ldots, D_L(\mathbf{x})$ to get a soft label for $\mathbf{x}$, denoted $D(\mathbf{x}) = [\mu_1(\mathbf{x}), \ldots, \mu_c(\mathbf{x})]^{\mathrm{T}}$.

If a crisp class label of $\mathbf{x}$ is needed, we can use the maximum membership rule: assign $\mathbf{x}$ to class ω_s iff,

$$d_{i,s}(\mathbf{x}) \geqslant d_{i,j}(\mathbf{x}) \forall j = 1, \ldots, c \text{ for individual crisp labels} \tag{1}$$

$$\mu_s(\mathbf{x}) \geqslant \mu_t(\mathbf{x}), \forall t = 1, \ldots, c \text{ for the final crisp label.} \tag{2}$$

Ties are resolved arbitrarily. The minimum-error classifier is recovered from (2) when $\mu_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$.

There are many different combination methods available. Here we consider nine combination methods and the Oracle (a favourable abstraction used as an upper limit for the performance of the other methods).

2.1. Majority vote (MAJ), Maximum (MAX), Minimum (MIN), Average (AVR), Product (PRO)

Once the classifiers in the ensemble are trained, these combination methods do not require any further training. For the majority vote combination, the class label assigned to $\mathbf{x}$ is the one that is most represented in the set of L crisp class labels obtained from $D_1(\mathbf{x}), \ldots, D_L(\mathbf{x})$. For the remaining simple combination methods

$$\mu_j(\mathbf{x}) = \mathscr{O}(d_{1,j}(\mathbf{x}), \ldots, d_{L,j}(\mathbf{x})), \quad j = 1, \ldots, c, \tag{3}$$

where $\mathscr{O}$ is the respective operation (maximum, minimum, average or product) and the class ω_j with maximum μ_j is the assigned class. For the case of two classes, it can be proven that maximum is always equivalent to minimum (Appendix A, Proposition 1). Table 1 shows an example of how these simple aggregation methods work.

2.2. Naive Bayes (NB)

Consider the crisp class labels obtained from $D_1(\mathbf{x}), \ldots, D_L(\mathbf{x})$ by (1), so in this case $D_i(\mathbf{x}) \in \Omega$, $i = 1, \ldots, L$. This scheme assumes that the classifiers are mutually independent; this is the reason we use the name "naive". Let $s_1, \ldots, s_L$ be the crisp class labels assigned to $\mathbf{x}$ by classifiers $D_1, \ldots, D_L$, respectively. The independence assumption leads to

$$\mu_j(\mathbf{x}) \propto \prod_{i=1}^{L} \hat{P}(\omega_j|D_i(\mathbf{x}) = s_i), \tag{4}$$

where $\hat{P}(\omega_j|D_i(\mathbf{x}) = s_i)$ are probability estimates calculated from the data.

$$\begin{aligned}\hat{P}(\omega_j|D_i(\mathbf{x}) = s_i) &= (\text{number of objects labelled } s_i \text{ by } D_i \\ &\quad \text{whose true label is } \omega_j) \\ &\quad /(\text{number of objects labelled } s_i \text{ by } D_i).\end{aligned}$$

Table 1
Example showing how the simple aggregation methods work

Classifier	Support for ω_1	Support for ω_2	Crisp decision
D_1	0.8	0.2	ω_1
D_2	0.4	0.6	ω_2
D_3	0.3	0.7	ω_2
D_4	0.6	0.4	ω_1
D_5	0.3	0.7	ω_2
MAJ			ω_2
MIN	0.3	0.2	ω_1
MAX	0.8	0.7	ω_1
AVR	0.48	0.52	ω_2
PRO	0.01728	0.2352	ω_2

The following example illustrates the NB combination method. Let $L = 3$ and $c = 2$. Suppose that the confusion matrices of the three classifiers, calculated on a data set $\mathbf{Z}$ with 100 objects are as shown in Table 2.

Let the output of the three classifiers for some $\mathbf{x} \in \mathbb{R}^n$ be $[s_1, s_2, s_3] = [\omega_2, \omega_1, \omega_2]$. Majority vote would label $\mathbf{x}$ in ω_2. However, the support for that class label is weak, albeit hypothesised by two of the three classifiers. For the Naive Bayes combination

$$\begin{aligned}
&\hat{P}(\omega_1 | s_1 = \omega_2) = \frac{12}{66},\\
&\hat{P}(\omega_2 | s_1 = \omega_2) = \frac{54}{66},\\
&\hat{P}(\omega_1 | s_2 = \omega_1) = \frac{68}{73},\\
&\hat{P}(\omega_2 | s_2 = \omega_1) = \frac{5}{73},\\
&\hat{P}(\omega_1 | s_3 = \omega_2) = \frac{16}{34},\\
&\hat{P}(\omega_2 | s_3 = \omega_2) = \frac{18}{34},\\
&\mu_1(\mathbf{x}) \propto \frac{12}{66}\cdot\frac{68}{73}\cdot\frac{16}{34} \approx 0.08 > \mu_2(\mathbf{x})\\
&\qquad \propto \frac{54}{66}\cdot\frac{5}{73}\cdot\frac{18}{34} \approx 0.03.
\end{aligned} \tag{5}$$

Accordingly, class ω_1 will be assigned.

2.3. Behavior–knowledge space (BKS)

BKS is in fact a fancy name for the multinomial combination. Let again $\mathbf{s} = (s_1, \ldots, s_L) \in \Omega^L$ be the crisp class labels assigned to $\mathbf{x}$ by classifiers $D_1, \ldots, D_L$, respectively. We can consider $\mathbf{s}$ to be an L-dimensional random variable and estimate $\mu_i(\mathbf{x}) = \hat{P}(\omega_i | \mathbf{s})$. To do so, every possible combination of class labels (a value of $\mathbf{s}$) is regarded as an index to a cell in a look-up table (*BKS* table) [20]. The table is designed using a labelled data set $\mathbf{Z}$. Each $\mathbf{z}_j \in \mathbf{Z}$ is placed in the cell indexed by $D_1(\mathbf{z}_j), \ldots, D_L(\mathbf{z}_j)$. The number of elements in each cell are tallied and the most representative class label is selected for this cell. Ties are resolved arbitrarily and the empty cells are labelled appropriately (e.g., at random or by majority, if applicable). After the table has been designed, the *BKS* method labels an $\mathbf{x} \in \mathbb{R}^n$ to the class of the cell indexed by $D_1(\mathbf{x}), \ldots, D_L(\mathbf{x})$.

Table 2
The confusion matrices of classifiers D_1, D_2, and D_3

True label	Guessed label					
	D_1		D_2		D_3	
	ω_1	ω_2	ω_1	ω_2	ω_1	ω_2
ω_1	30	12	68	2	54	16
ω_2	4	54	5	25	12	18

For the example discussed in the previous sections, assume again that D_1, D_2 and D_3 produce output $(s_1, s_2, s_3) = (\omega_2, \omega_1, \omega_2)$. Suppose there have been 22 objects in $\mathbf{Z}$ for which this combination of labels occurred; 14 having label ω_1, and 8 ω_2. Hence the table cell indexed by $(\omega_2, \omega_1, \omega_2)$ will be labelled ω_1 no matter that the majority of the classifiers suggest otherwise.

2.4. Wernecke's method (WER)

The model is similar to the *BKS*. The difference is that in constructing the table, Wernecke [21] considers the 95% confidence intervals of the frequencies in each cell. If there is overlap between the intervals, the "least wrong" classifier among the L members of the team is identified and authorised to label $\mathbf{x}$. For this, L estimates of the probability $P(error \text{ and } D_i(\mathbf{x}) = s_i)$ are calculated. Then the classifier with the smallest probability is nominated for labelling the cell. For an $\mathbf{x} \in \mathbb{R}^n$, the cell is identified by the labels assigned by $D_1, \ldots, D_L$ and then either the cell label is recovered or the label of the nominated classifier is taken as the label of $\mathbf{x}$.

To continue the example illustrating *BKS* combination method, we calculate the 95% confidence intervals using Chebyshev's inequality (e.g., see [8]). The CI for ω_1 is $[4.49, 23.59]$, and for $\omega 2$, $[-1.59, 17.59]$. Since the CI are overlapping, estimates of $P(error \text{ and } D_i(\mathbf{x}) = s_i)$ have to be obtained. Using the data in Table 2,

$$\begin{aligned}
\hat{P}(error \text{ and } D_1(\mathbf{x}) = \omega_2) &= \hat{P}(\omega_1 | s_1 = \omega_2)\hat{P}(s_1 = \omega_2)\\
&= \frac{12}{66}\cdot\frac{66}{100} = \frac{12}{100},\\
\hat{P}(error \text{ and } D_2(\mathbf{x}) = \omega_1) &= \hat{P}(\omega_2 | s_2 = \omega_1)\hat{P}(s_2 = \omega_1)\\
&= \frac{5}{73}\cdot\frac{73}{100} = \frac{5}{100},\\
\hat{P}(error \text{ and } D_3(\mathbf{x}) = \omega_2) &= \hat{P}(\omega_1 | s_3 = \omega_2)\hat{P}(s_3 = \omega_2)\\
&= \frac{16}{34}\cdot\frac{34}{100} = \frac{16}{100}.
\end{aligned}$$

As $\hat{P}(error \text{ and } D_2(\mathbf{x}) = \omega_1)$ is the smallest of the three, classifier D_2 is authorised to label $\mathbf{x}$, and thus the assigned class is ω_1.

2.5. Decision templates (DT)

The classifier outputs can be conveniently organised in a *decision profile* as the following matrix [22]

$$DP(\mathbf{x}) = \begin{bmatrix} d_{1,1}(\mathbf{x}) & \ldots & d_{1,j}(\mathbf{x}) & \ldots & d_{1,c}(\mathbf{x}) \\ \ldots & & & & \\ d_{i,1}(\mathbf{x}) & \ldots & d_{i,j}(\mathbf{x}) & \ldots & d_{i,c}(\mathbf{x}) \\ \ldots & & & & \\ d_{L,1}(\mathbf{x}) & \ldots & d_{L,j}(\mathbf{x}) & \ldots & d_{L,c}(\mathbf{x}) \end{bmatrix}. \tag{6}$$

Using decision templates (DT) for combining classifiers is proposed in [22]. Given L (trained) classifiers in $\mathscr{D}$, c

decision templates are calculated from the data, one per class.

$$DT_i = \frac{1}{N_i} \sum_{\substack{\mathbf{z}_j \in \omega_i \\ \mathbf{z}_j \in \mathbf{Z}}} DP(\mathbf{z}_j), i = 1, \ldots, c. \tag{7}$$

DT_i can be regarded as the expected $DP(\mathbf{x})$ for class ω_i. The support for the class offered by the combination of the L classifiers, $\mu_i(\mathbf{x})$ is then found using a measure of *similarity* between the current $DP(\mathbf{x})$ and DT_i, e.g.,

$$\begin{aligned}\mu_i(\mathbf{x}) &= 1 - d_{\mathrm{E}}(DP(\mathbf{x}), DT_i) \\ &= 1 - \sum_{j=1}^{c} \sum_{k=1}^{L} (d_{k,j}(\mathbf{x}) - dt_i(k,j))^2,\end{aligned} \tag{8}$$

where $dt_i(k,j)$ is the k,jth entry in decision template DT_i. Here we use the squared Euclidean distance for calculating the similarity but other measures can also be applied.

For the example, assume that the following decision templates have been obtained from a data set $\mathbf{Z}$ using Eq. (7),

$$DT_1 = \begin{pmatrix} 0.62 & 0.38 \\ 0.52 & 0.48 \\ 0.6 & 0.4 \end{pmatrix}, \quad DT_2 = \begin{pmatrix} 0.46 & 0.54 \\ 0.56 & 0.44 \\ 0.48 & 0.52 \end{pmatrix}.$$

Given an object $\mathbf{x}$ with decision profile

$$DP(\mathbf{x}) = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \\ 0.3 & 0.7 \end{pmatrix},$$

we use the squared Euclidean distance to calculate the dissimilarity to DT_1 and DT_2, and subsequently the support for the two classes

$$\begin{aligned} d_{\mathrm{E}}(DP(\mathbf{x}), DT_1) &= (0.6 - 0.62)^2 + (0.4 - 0.38)^2 \\ &\quad + (0.4 - 0.52)^2 + (0.6 - 0.48)^2 \\ &\quad + (0.3 - 0.6)^2 + (0.7 - 0.4)^2 \\ &= 0.2069, \end{aligned}$$

$$d_{\mathrm{E}}(DP(\mathbf{x}), DT_2) = 0.1552,$$

$$\mu_1(\mathbf{x}) = 1 - d_{\mathrm{E}}(DP(\mathbf{x}), DT_1) = 1 - 0.2096 = 0.7904,$$

$$\mu_2(\mathbf{x}) = 1 - 0.1552 = 0.8448.$$

Since $\mu_2 > \mu_1$ we assign class label ω_2 to $\mathbf{x}$.

2.6. Oracle (ORA)

This is an abstraction, which is only used as a possible upper limit on the classification accuracy. It works by correctly classifying an object provided at least one of the L classifiers correctly classifies the object.

2.7. Differences between the methods

The combination methods can be divided into groups by whether they require training or not and by the type of individual classifier output they require [1]. Majority Vote and the other simple combination methods of *MAX*, *MIN*, *AVR*, and *PRO* do not require any training whilst the remaining methods require training.

There are two main levels of classifier output that different combination methods may require: measurement and abstract [23]. At measurement level we have $DP(\mathbf{x})$ as the L by c matrix (6), and at the abstract level, we have the crisp classifier outputs $s_1, \ldots, s_L \in \Omega^L$. Decision templates, Maximum, Minimum, Average and Product all work at the measurement level of information. Majority vote, Behavior–knowledge space, Wernecke's method and Naive Bayes work with the abstract level of information. Of course since all outputs can be transformed from measurement to abstract label (assigning crisp class labels), the methods of the latter group will work for measurement outputs as well.

3. Measures of diversity

There are different diversity measures available from different fields of research. Some of these measures, such as the Q-statistic and the correlation coefficient have come directly from mainstream statistics whilst others have developed through the field of statistical pattern recognition, specifically for the problems of multiple classifier systems. Some of these measures work on the whole group of L classifiers whilst other measures consider the classifiers on a pairwise basis and then average the results.

3.1. Pairwise diversity measures

Consider two classifiers, D_i and D_k, and a 2×2 table that summarises their outputs as shown in Table 3. The entries in the table are the probabilities for the respective pair of correct/incorrect outputs.

There are various statistics to assess the similarity of two classifier outputs.

Table 3
The 2×2 relationship table with probabilities

	D_k correct (1)	D_k wrong (0)
D_i correct (1)	a	b
D_i wrong (0)	c	d
Total	$a + b + c + d = 1$	

3.1.1. The Q statistic (Q)

Yule's Q statistic [24] for two classifiers, e.g., D_i and D_k, is

$$Q_{i,k} = \frac{ad - bc}{ad + bc}. \tag{9}$$

For statistically *independent* classifiers, $Q_{i,k} = 0$. Q varies between -1 and 1. For a set of L classifiers, the averaged Q statistics of all pairs is taken.

3.1.2. The correlation coefficient (ρ)

The correlation between two binary classifier outputs (correct/incorrect) is

$$\rho_{i,k} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{10}$$

For any two classifiers, Q and ρ have the same sign, and it can be proved that $|\rho| \leqslant |Q|$.

3.1.3. The disagreement measure (D) (used in [25,26])

$$D_{i,k} = b + c. \tag{11}$$

3.1.4. The double-fault measure (DF) (used in [27])

$$DF_{i,k} = d. \tag{12}$$

We note that all these pairwise measures have been proposed as measures of (dis)similarity in the numerical taxonomy literature (e.g., [28]).

3.2. Non-pairwise diversity measures

For the non-pairwise measures we quote the formulae for L classifiers. Let $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ be a labelled data set, $\mathbf{z}_j \in \mathbb{R}^n$ coming from the classification problem in question. We can represent the output of a classifier D_i as an N-dimensional binary vector $\mathbf{y}_i = [y_{1,i}, \ldots, y_{N,i}]^{\mathrm{T}}$, such that $y_{j,i} = 1$, if D_i recognises correctly $\mathbf{z}_j$, and 0, otherwise, $i = 1, \ldots, L$.

3.2.1. Kohavi–Wolpert variance (kw)

We take the formula for the variance from Kohavi and Wolpert's paper [29]. They derived a decomposition formula for the error rate of a classifier, giving an expression of the variability of the predicted class label b for $\mathbf{x}$, across training sets, within a specific classifier model as

$$variance_x = \frac{1}{2}\left(1 - \sum_{i=1}^{c} P(b = \omega_i \,|\, \mathbf{x})^2\right), \tag{13}$$

where $P(b = \omega_i \,|\, \mathbf{x})$ is estimated as an average over different data sets. We use their general idea by looking 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 considering the class labels in Ω, we consider two possible classifier outputs: correct and incorrect. $P(b = 1 \,|\, \mathbf{x})$ and $P(b = 0 \,|\, \mathbf{x})$ will be obtained as an average over $\mathscr{D}$. If we denote by $l(\mathbf{z}_j)$ the number of classifiers from $\mathscr{D}$ that correctly recognise $\mathbf{z}_j$, i.e., $l(\mathbf{z}_j) = \sum_{i=1}^{L} y_{j,i}$ we obtain:

$$P(b = 1 \,|\, \mathbf{x}) = \frac{l(\mathbf{x})}{L} \text{ and } P(b = 0 \,|\, \mathbf{x}) = \frac{L - l(\mathbf{x})}{L}. \tag{14}$$

Substituting (14) into (13),

$$variance_x = \frac{1}{2}(1 - P(b = 1 \,|\, \mathbf{x})^2 - P(b = 0 \,|\, \mathbf{x})^2) \tag{15}$$

and averaging over the whole of the training set $\mathbf{Z}$, we obtain the kw measure of diversity as

$$kw = \frac{1}{NL^2} \sum_{j=1}^{N} l(\mathbf{z}_j)(L - l(\mathbf{z}_j)). \tag{16}$$

3.2.2. Measurement of interrater agreement (κ) (used in [30])

If we denote $\bar{p}$ to be the average individual classification accuracy in the ensemble, then

$$\kappa = 1 - \frac{(1/L)\sum_{j=1}^{N} l(\mathbf{z}_j)(L - l(\mathbf{z}_j))}{N(L-1)\bar{p}(1-\bar{p})} \tag{17}$$

and so κ can be shown to be related to kw and D as follows

$$\kappa = 1 - \frac{L}{(L-1)\bar{p}(1-\bar{p})} kw = 1 - \frac{1}{2\bar{p}(1-\bar{p})} D. \tag{18}$$

3.2.3. The entropy measure (Ent)

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 $\mathbf{y}_j$ 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. One possible measure of diversity based on this concept is

$$Ent = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{(L - \lfloor L/2 \rfloor - 1)} \min\left\{ \sum_{i=1}^{L} y_{j,i}, L - \sum_{i=1}^{L} y_{j,i} \right\}. \tag{19}$$

Ent varies between 0 and 1, where 0 indicates no difference and 1 indicates the highest possible diversity. While value 0 is achievable for any number of classifiers L and any p, value 1 can only be attained for $p \in [(L-1)/2L, (L+1)/2L]$.

It should be noted here that our measure *Ent* is a non-classical entropy measure because it does not use the logarithm function. A classical version of this measure is

proposed by Cunningham and Carney [3] (we denote it here as E_{CC}). [1] Taking the expectation over the whole feature space, letting the number of classifiers $L \to \inf$, and denoting by a the proportion of 1's (correct outputs) in the team, the two expressions become

$$Ent(a) = \frac{1}{2}\min\{a, 1-a\} \text{ and}$$
$$E_{CC}(a) = -a\log(a) - (1-a)\log(1-a). \quad (20)$$

Fig. 1 plots the two entropies versus a.

The two measures are equivalent up to a (nonlinear) monotonic transformation. This means that they will have a similar pattern of relationship with the team accuracy. As *Ent* is easier to handle and quicker to calculate, we use it in the experiment.

3.2.4. The measure of difficulty (θ)

The idea for this measure came from a study by Hansen and Salomon [31]. 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 $\mathscr{D}$ that correctly classify an input $\mathbf{x}$ drawn randomly from the distribution of the problem. The measure of *difficulty* θ is defined as

$$\theta = Var(X). \quad (21)$$

The higher the value of θ, the worse the classifier team.

3.2.5. Generalised diversity (GD)

This measure has been proposed in [32]. Let Y be a random variable expressing the proportion of classifiers (out of L) that *fail* on a randomly drawn object $x \in \mathbb{R}^n$. Denote by p_i the probability that $Y = i/L$. (Note that $Y = 1 - X$, where X is the variable introduced for θ). Denote by $p(i)$ the probability that i randomly chosen classifiers will fail on a randomly chosen $\mathbf{x}$. Then

$$p(1) = \sum_{i=1}^{L} \frac{i}{L} p_i \quad (22)$$

and

$$p(2) = \sum_{i=1}^{L} \frac{i}{L} \frac{(i-1)}{(L-1)} p_i. \quad (23)$$

The generalised diversity measure, *GD*, is

$$GD = 1 - \frac{p(2)}{p(1)}. \quad (24)$$

3.2.6. Coincident failure diversity (CFD)

This is a modification of *GD* proposed in [33].

$$CFD = \begin{cases} 0, & p_0 = 1.0, \\ \frac{1}{1-p_0} \sum_{i=1}^{L} \frac{L-i}{L-1} p_i, & p_0 < 1. \end{cases} \quad (25)$$

[1] We wish to thank the anonymous reviewer C for pointing this reference to us.

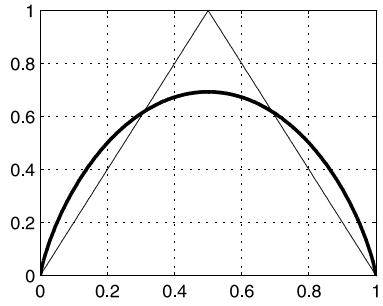


Fig. 1. The two entropy measures $Ent(a)$ (thin line) and $E_{CC}(a)$ (thick line) plotted versus a.

4. Commonalities and differences between the measures

For the case of correct/incorrect (1/0) classifier outputs (oracle-type outputs), *kw* differs from the averaged disagreement measure D by a coefficient [34]. Also for the case with $L = 3$ classifiers, *kw* and *Ent* differ by a coefficient (Appendix A, Proposition 2). This in turn means that the disagreement measure and Entropy differ by a coefficient for the three classifier case, with correct/incorrect, outputs, i.e.,

$$kw = \frac{L-1}{2L} D \quad (1/0 \text{ outputs})$$
$$\Rightarrow \quad kw = \frac{1}{3} D \quad (1/0 \text{ and } L = 3)$$
$$kw = \frac{2}{9} Ent \quad (1/0 \text{ and } L = 3)$$
$$\Rightarrow \quad D = \frac{2}{3} Ent \quad (1/0 \text{ and } L = 3).$$

We can consider the measures of diversity in two groups:

- measures looking for diversity: the higher the value the more diverse (↑);
- measures looking for similarity: the higher the value the less diverse (↓).

D, *kw*, *Ent*, *GD* and *CFD* belong to the first group. *Q*, ρ, *DF*, κ and θ belong to the second group. Also, *Q*, ρ and κ can take negative values, indicating negative correlation between the classifiers.

4.1. Upper and lower limits for diversity measures

The upper and lower limits for the measures of diversity depend on the number of classifiers, L, and the value of p, the individual classifier accuracy. The limits for the case with two classifiers with equal individual accuracy p have been determined in [35]. Figs. 2 and 3

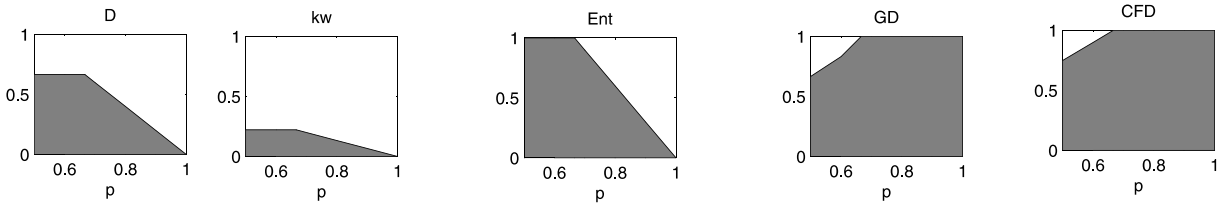


Fig. 2. The possible range of values (grey areas) for the five (↑) measures of diversity for $p \in [0.5, 1.0]$ individual classifier accuracy and $L = 3$ classifiers.

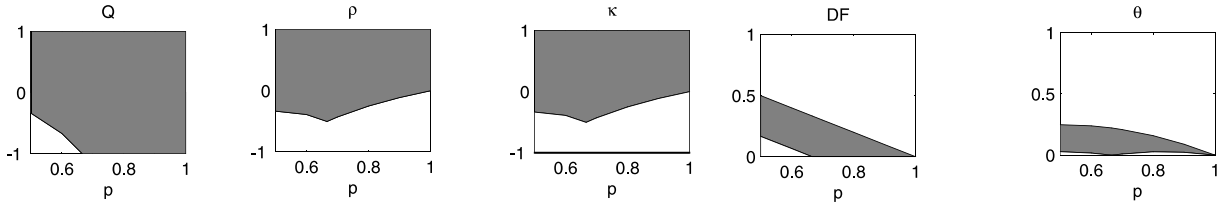


Fig. 3. The possible range of values (grey areas) for the five (↓) measures of diversity for $p \in [0.5, 1.0]$ individual classifier accuracy and $L = 3$ classifiers.

show the upper and lower limits for the ten measures of diversity for the case with $L = 3$ and $p \in [0.5, 1.0]$. They have been derived from the two classifier case limits by considering the possible range of each measure for $p = 0.5, 0.6, \frac{2}{3}, 0.7, 0.8, 0.9, 1.0$. It was found that there was often a change of direction of the curve at the point $2/3$.

To summarise, the 10 measures of diversity used in this study are:

- The Q-statistic (Q), (↓)
- The correlation coefficient (ρ), (↓)
- The disagreement measure (D), (↑)
- The double-fault measure (DF), (↓)
- The Kohavi–Wolpert variance (kw), (↑)
- The measurement of interrater agreement (κ), (↓)
- The entropy measure (Ent), (↑)
- The measure of difficulty (θ), (↓)
- The generalised diversity (GD), (↑)
- The coincident failure diversity (CFD), (↑)

4.2. Cross-relationship between combination methods and diversity measures

In our study one of the issues we are considering is the cross-relationship between the combination methods and the diversity measures. If we can find a strong correlation between any of the diversity measures and the accuracy of any of the combination methods then this will allow us to use the diversity of a set of classifiers as an indication of the ensemble accuracy obtained by combining them. This would allow us to select the 'best' subset of classifiers from a larger group or even use the diversity directly in the generation of a 'diverse' set of classifiers for an ensemble.

5. Experimental set-up

We used two databases summarised in Table 4, both taken from the UCI Repository of Machine Learning Database available at http://www.ics.uci.edu/~mlearn/MLRepository.html: The Wisconsin Breast Cancer Database [2] and the Pima Indian Diabetes Database.

The initial experimental protocol is also displayed in Table 4. From the original 30 features for the Breast Cancer data we used the first 10 so that we could run an exhaustive experiment with all possible partitions. We chose the first 10 because the features in this data set were logically grouped into 1–10, 11–20, 21–30. All partitions for three classifiers of the form $4, 3, 3$ (4200) and $4, 4, 2$ (3150) were generated so that the first classifier has four features as input, the second classifier has 3(4) features as input and the third classifier has 3(2) features as input. For each partition we designed, one ensemble of three linear classifiers and one ensemble of three quadratic classifiers. This is why the total number of ensembles for the Breast Cancer data is twice the total number of partitions. Our preliminary studies showed that there are no substantial differences between the four cases, so we pooled the data, thereby creating a set of 14,700 classifier teams.

For the Pima Diabetes data, we took all partitions of the form $3, 3, 2$ using 10-fold cross-validation to obtain a total of 560 ensembles. We note that our main experiment was on the Breast Cancer data because of the larger number of ensembles generated, and the Pima Indian Diabetes data was used mainly for re-confirmation and validation of the results. For space reasons, we

[2] Created by Dr. William H. Holberg, W. Nick Street and Olvi L. Mangasarian, University of Wisconsin.

Table 4
Summary of the data sets and the experiments

Name	c	N	n	(n_1, n_2, n_3)	Total number of ensembles	Training/testing
Wisconsin Breast Cancer	2	569	10	$(4,4,2)$	6300	Hold-out (random halves)
				$(4,3,3)$	8400	
Pima Indian Diabetes	2	768	8	$(3,3,2)$	560	10-fold cross-validation

Key: c: number of classes; N: number of objects in the data set; n: number of features used; (n_1, n_2, n_3): partition sizes; D_1 uses n_1 of the n features, D_2 uses n_2, and D_3 uses n_3 features.

decided that the results with the Breast Cancer data will be displayed in detail, whereas the results with the Pima Diabetes data will be discussed if there are relevant differences.

We then considered:

1. The overall accuracies of the combination methods and their improvement over the single best classifier.
2. The range of values for the measures of diversity.
3. The correlation between each method of combination and all other methods of combination.
4. The correlation between each measure of diversity and all other measures of diversity.
5. The correlation between each of the methods of combination and each of the measures of diversity.

The correlation coefficient used was Pearson's Product Moment correlation coefficient.

6. Results

6.1. Overall accuracies

Fig. 4(a) shows the accuracy on the testing data for the Breast Cancer Database: the single best classifier (best on the testing set!), the individual classifiers and the ensemble. The dashed horizontal line is the average accuracy of the single best classifier. Fig. 4(b) shows the improvement over the single best classifier accuracy (in %). The dashed line at 0 represents accuracy identical to the single best classifier. For both graphs, the lower end of each bar is the minimum value, the upper end of each bar is the maximum value, and the circled point is the average value.

Fig. 4 shows that all of the combination methods (excluding Oracle) are of similar accuracy, and have only a slight improvement over the average accuracy of the single best classifier. Not surprisingly, being a favourable abstraction, the Oracle shows an improvement over the single best classifier in all cases.

As we would expect, D_3 has poorer results than D_1 and D_2 since it only has two features to work with for the $(4,4,2)$ partitions and three for the $(4,3,3)$ partitions. Similarly, D_1 performs better than the other two classifiers since it always has four features to work with.

As Fig. 4 indicates, the lower limit of the individual accuracies is greatly improved by combining the three classifiers. The behaviour–knowledge space method and Wernecke's method have lower minimum values than the other combination methods suggesting variability in their performance.

The results using the Pima Diabetes data do not show any significant difference in the range of values obtained

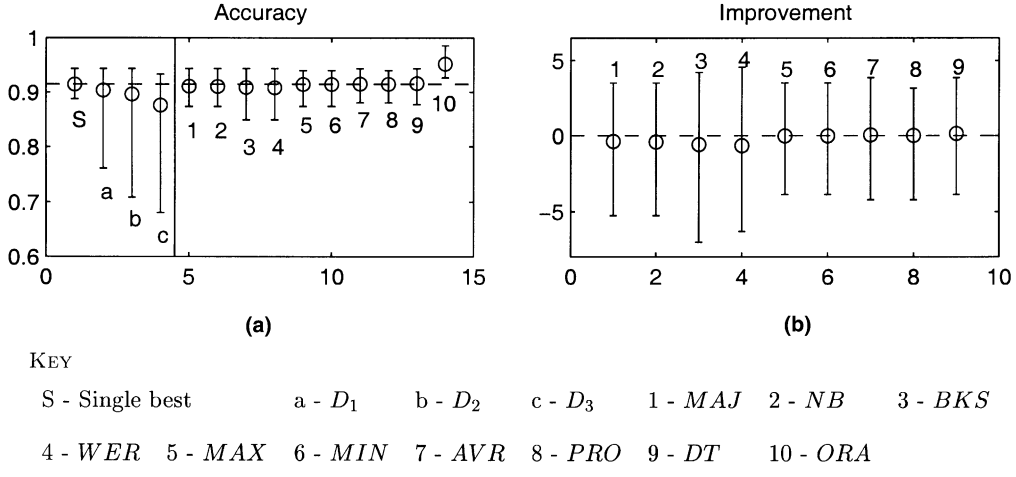


Fig. 4. Accuracy (a) and improvement (b) on the testing set for the individual classifiers and the ensemble.

for combination accuracy to those displayed above, found using the Breast Cancer data.

The failure of the ensemble to improve on the single best classifier can be explained by the hypothesis that all features were relevant and breaking them into subsets diminishes the chances of finding three *good* classifiers. More importantly, by "single best" we assumed the classifier with the best *testing* accuracy. Thus we gave a hard task to the ensemble compared to surpassing the single best classifier identified on the training data. The simple reason is that the "best training" classifier will not always be the "best testing" classifier and the averaged "single (training) best" accuracy on the testing set would be lower than the averaged "single (testing) best" accuracy.

In a way, the lack of improvement is not a bad result for our study. We would like to find a relationship between diversity and accuracy, which will help in uncertain situations like this. Ideally, we would like diversity to be sensitive enough to predict the improvement or the lack of it.

6.2. Overall diversities

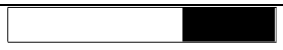
With the Breast Cancer data, the minimum observed value of the individual classifier accuracy, p, was 0.6807, the maximum was 0.9439 and the overall mean was 0.8922. Table 5 shows the observed range of values for the ten measures of diversity compared with their theoretical limits for the observed values of p, assuming equal p. The theoretical limits were deduced from the graphs shown previously in Figs. 2 and 3.

Q, ρ and κ all take negative values when the classifiers are negatively correlated. Given that none of these measures has any negative values, we can conclude that the classifiers are not very diverse. The measures where low values indicate high diversity, ($\downarrow$), except DF and θ, have high values , toward the right end of the range, as shown in Table 5. The measures where high values indicate high diversity, ($\uparrow$), except for CFD, have low values. This suggests overall, that the classifiers are less diverse than they could theoretically be, if identical accuracies p are assumed.

It is interesting to note that even though the measures do not indicate identical or close to identical classifiers, the average accuracy of the team was similar to the average best individual accuracy. Thus a range of values of diversity did not span a similar range of improvement/degradation of team accuracy. This is an early indication of the lack of any strong relationship between diversity measures and team accuracy in real-life classification problems.

Results from the Pima Diabetes data do not show any significant difference in the range of values obtained for the diversity measures to those above found using the Breast Cancer data.

Table 5
The observed range of values for the diversity measures compared with the theoretical limits possible for the observed values of p

Measure	Observed p	Theoretical limits for $\bar{p}$	Observed span	Graphical representation
Q ($\downarrow$)	0.7–0.9	$[-1.00, 1.00]$	$[0.30, 0.99]$	
ρ ($\downarrow$)	0.7–0.9	$[-0.43, 1.00]$	$[0.16, 0.83]$	
DF ($\downarrow$)	0.7–0.9	$[0.00, 0.2]$	$[0.03, 0.08]$	
κ ($\downarrow$)	0.7–0.9	$[-0.43, 1.00]$	$[0.12, 0.82]$	
θ ($\downarrow$)	0.7–0.9	$[0.02, 0.16]$	$[0.04, 0.08]$	
D ($\uparrow$)	0.7–0.9	$[0.00, 0.6]$	$[0.03, 0.26]$	
kw ($\uparrow$)	0.7–0.9	$[0.00, 0.20]$	$[0.01, 0.09]$	
Ent ($\uparrow$)	0.7–0.9	$[0.00, 0.60]$	$[0.04, 0.38]$	
GD ($\uparrow$)	0.7–0.9	$[0.00, 1.00]$	$[0.12, 0.75]$	
CFD ($\uparrow$)	0.7–0.9	$[0.00, 1.00]$	$[0.31, 0.87]$	

(□) theoretical; (■) observed range of values.

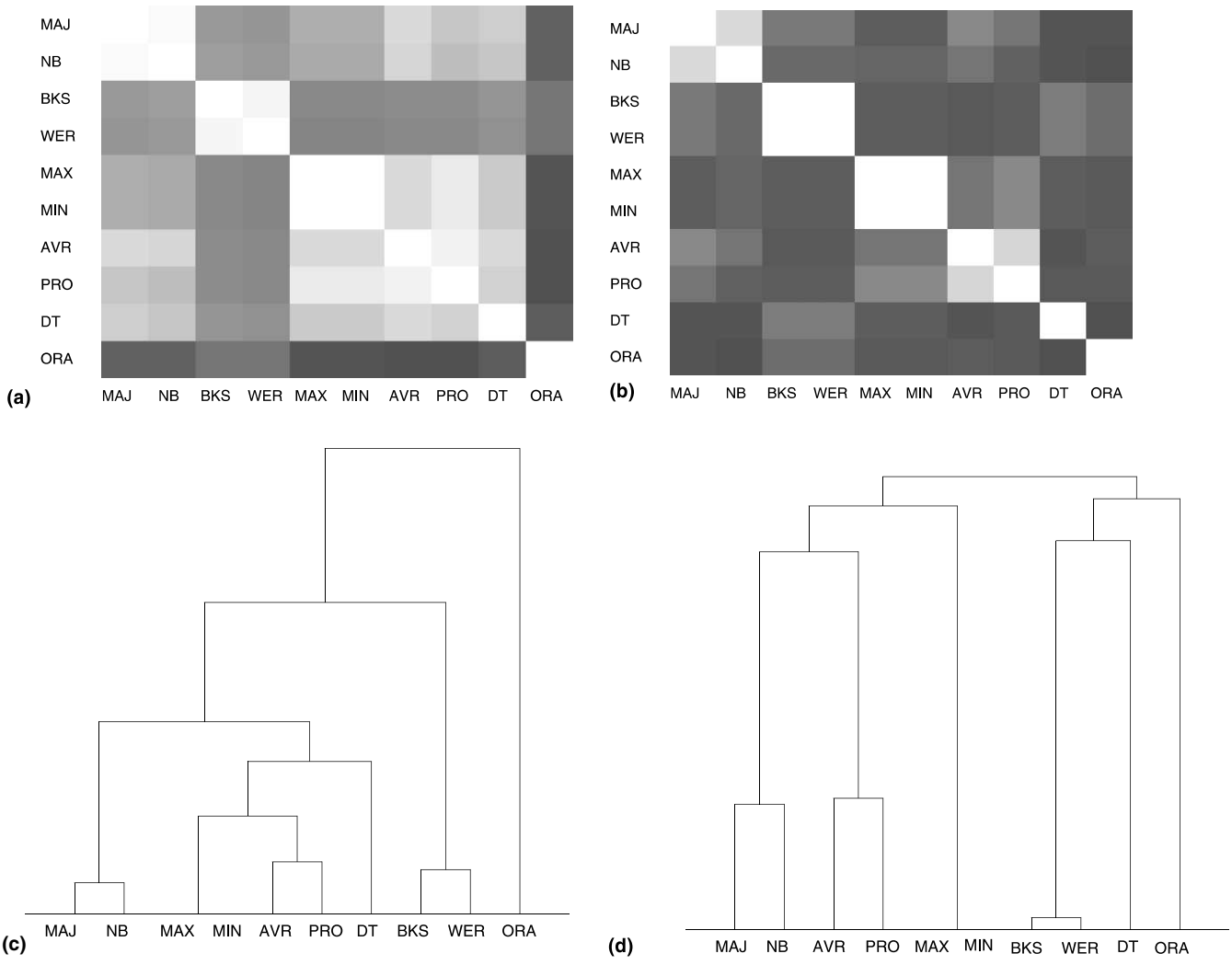


Fig. 5. The overall correlation between the combination methods and the cluster dendrograms for Breast Cancer data ((a) and (c)) and Pima Indian Diabetes data ((b) and (d)).

6.3. Correlation amongst the combination methods

Fig. 5(a) and (b) illustrate the correlation between the combination methods. The intensity of the colour is determined by the correlation. The stronger the correlation the lighter the colour. We found that the combination methods show only positive correlation amongst themselves. Fig. 5(c) and (d) show the dendrograms formed when we cluster the combination methods using average-linkage relational clustering. [3] The lower the branches joining the different combination methods the stronger the relationship between them. Using all graphs from Fig. 5 we found the following:

1. Majority vote is highly positively correlated with Naive Bayes. In fact, for the $(4, 3, 3)$ partitions with *QDC*, *MAJ* and *NB* are almost equivalent as they have a correlation of 0.999.
2. Behaviour–knowledge space is highly positively correlated with Wernecke's method, which can be expected, knowing that Wernecke's method is a "regularised" version of *BKS*.
3. Average is highly positively correlated with Product, and Minimum and Maximum are identical (proof in the Appendix A).
4. While the above three tendencies are common for both sets of experiments, the overall correlation between the combination methods with the Pima Diabetes data was much lower (darker shades in plot (b)). Looking at the dendrogram in (c), we cannot identify a "true" number of clusters in the set of methods because there are no big "jumps" of the

[3] The clustering routine and the dendrogram drawing routine are from the package PRTOOLS for Matlab [36].

clustering criterion value. The dendrogram in (d) reinforces the findings in 1–3, suggesting the following grouping of the methods $((MAJ, NB), (AVR, PRO), (MIN, MAX), (BKS, WER), (DT), (ORA))$.

5. For the Breast Cancer data, the Oracle appears to be different from all the other combination methods, which correlate well among themselves. With the Pima Indian Diabetes data, the methods are different from each other and yet none of them has a high correlation with the Oracle.

6.4. Correlation amongst the diversity measures

Figs. 6(a) and (b) illustrate the correlation between the diversity measures. The absolute values of the correlation coefficients have been taken to illustrate *any* correlation, whether it is positive or negative. The stronger the correlation the lighter the colour. Figs. 6(c) and (d) show the dendrograms formed when we cluster the diversity measures using average-linkage relational clustering. The lower the branches joining the different diversity measures are, the stronger the relationship between them. Using all graphs from Fig. 6 we found that the relationships are more complicated than for the combination methods with the dendrogram producing quite different clusters for the two data sets. The results we found using both data sets are:

- $D = kw = Ent$ is highly correlated with κ.
- GD is strongly correlated with κ and ρ.
- DF only shows strong correlation with θ.
- ρ is correlated with CFD and Q for both data sets as well.

Again, the number of "true" clusters is unclear because whilst both dendrograms appear to show *two* distinct clusters, the diversity measures within those clusters are not the same for both data sets.

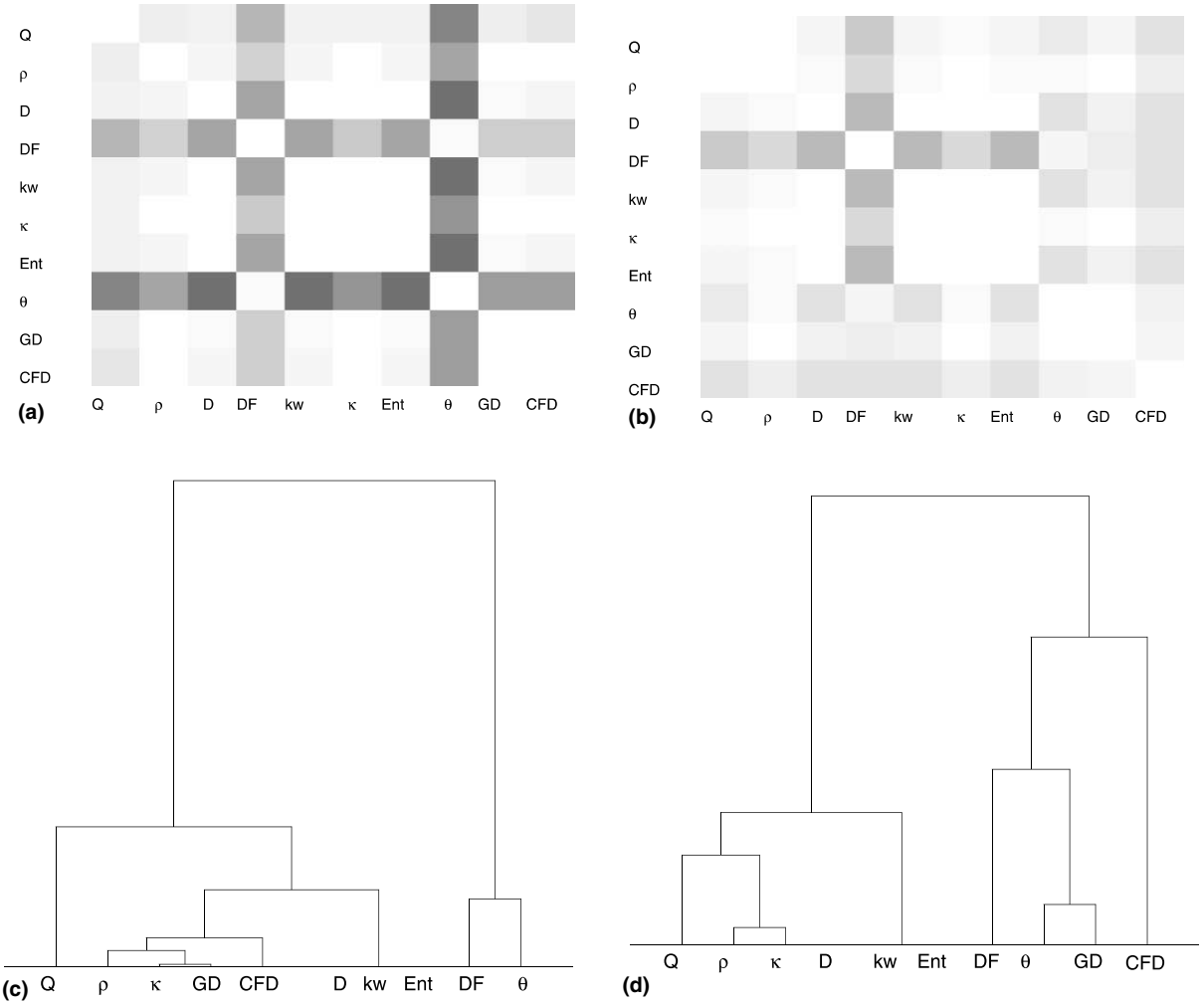


Fig. 6. The overall correlation between the diversity measures and the cluster dendrograms for Breast Cancer data ((a) and (c)) and Pima Indian Diabetes data ((b) and (d)).

Table 6
Breast Cancer: correlations between the combination methods and diversity

Comb\Div	Q	ρ	κ	GD	CFD	D	DF	θ
MAJ	−0.080	−0.099	−0.116	0.139	0.241	0.019	***−0.423***	***−0.596***
NB	−0.060	−0.080	−0.093	0.115	0.208	−0.002	***−0.495***	***−0.572***
MAX	0.056	0.095	0.056	−0.051	−0.008	−0.078	−0.197	−0.274
AVR	−0.088	−0.046	−0.080	0.091	0.168	0.035	***−0.349***	***−0.398***
PRO	−0.044	0.003	−0.039	0.045	0.108	0.009	−0.277	***−0.329***
BKS	−0.005	−0.031	−0.044	0.059	0.054	−0.020	***− 0.365***	***−0.437***
WER	0.005	−0.019	−0.034	0.047	0.039	−0.023	***−0.342***	***−0.413***
DT	0.055	−0.018	−0.028	0.047	0.123	−0.049	***−0.365***	***−0.443***

Table 7
Pima Diabetes: correlation between the combination methods and diversity

Comb\Div	Q	ρ	κ	GD	CFD	D	DF	θ
MAJ	−0.168	−0.254	−0.250	***0.355***	***0.647***	0.102	***−0.566***	***−0.423***
NB	−0.055	−0.128	−0.134	0.223	***0.440***	0.015	***−0.415***	−0.283
MAX	−0.100	−0.102	−0.107	0.082	0.061	0.142	−0.009	−0.054
AVR	−0.177	−0.215	−0.208	0.226	***0.328***	0.186	−0.246	−0.227
PRO	−0.170	−0.190	−0.187	0.188	0.257	0.184	−0.174	−0.180
BKS	0.057	0.001	−0.049	0.108	0.102	−0.029	−0.259	−0.158
WER	0.059	0.003	−0.046	0.107	0.103	−0.035	−0.262	−0.159
DT	0.085	0.080	0.065	−0.058	−0.049	−0.067	0.006	0.055

6.5. Correlation between the combination methods and the diversity measures

We took each of the combination method's results (a column of accuracies, range: 0–1) and each of the diversity measure's results (a column of values whose range depends upon the diversity measure in question) and calculated the correlation between the two. Table 6 shows the correlation between the combination methods and the diversity measures for the Breast Cancer data and Table 7 for the Pima Indian Diabetes data. Those correlations with absolute value greater than 0.3 are shown in bold italics. We can see that there are not many strong relationships consistent in both tables. The correlations between the combination methods and diversity measures are not as strong as those amongst the combination methods and diversity measures separately. The correlations show that many of the combination methods and diversity measures are even independent! Thus we have very little evidence of any relationships between the combination method accuracy and the diversity measure value. This means that we can hardly use these diversity measures as an indicator, guide, or predictor in designing classifier ensembles.

Oracle had stronger correlations (negative or positive) than all other methods of combination with every measure of diversity, but Oracle is not a true combination method and we did not show it in the table. We found that only DF and θ show significant correlations in both tables, but then only with MAJ and NB.

7. Analysis and conclusions

In this paper, we studied the relationships between different methods of classifier combination and measures of diversity. We considered 10 combination methods and 10 measures of diversity. We took a dataset of 10 feature values for 569 patients and using all partitions of the form $(4,4,2)$ and $(4,3,3)$ for two types of classifier, conducted a set of four enumerative experiments. The results from these four experiments were combined to give an overall set of 14,700 classifier teams. We also took a data set of eight feature values for 768 patients and conducted a set of 10-fold cross-validation experiments.

We then considered the overall accuracies of the combination methods and their improvement over the single best classifier. Also the range of values for the measures of diversity. Next we studied the correlation amongst the combination methods, the correlation amongst diversity measures, and the cross-correlation between the methods of combination and the measures of diversity.

We found that the classifiers were not very diverse and this meant that the combination methods did not improve notably over the single best classifier. We also found some interesting correlation amongst both the

combination methods and the diversity measures. In particular Majority vote's, (*MAJ*), strong correlation with Naive Bayes, (*NB*), and the measurement of interrater agreement's, (κ), very strong correlation with the generalised diversity, (*GD*).

We found very little correlation between the combination methods and the diversity measures, in fact, most of them showed independence. *GD*, *CFD*, *DF* and θ were the only measures to show any correlations with the combination methods greater than 0.3, but only double-fault, *DF*, and the measure of difficulty, θ, showed any such correlations consistent in *both* data sets. This result is discouraging because the measures of diversity are supposed to give an indication of classifier combination performance, and yet we found very little evidence of any correlation in these real-data problems. Of the correlations we *have* found, θ shows a stronger negative correlation with the combination methods than *DF* does for the Breast Cancer data but has less negative correlation than *DF* for the Pima Diabetes data. Also θ is more computationally expensive than *DF*. Perhaps the reason that *DF* and θ showed stronger correlations is because they are not exactly measures of diversity but have a subtle conceptual relationship with the accuracy?

DF and/or θ may be beneficial if we intend to use Majority vote and Naive Bayes which had the strongest correlations with the two measures. We could take *DF* and/or θ to guide us towards designing or selecting the classifiers in a team, trying to minimise the measures over a set of possible teams. In our experiment the measure with the strongest correlation depended upon which data set was being used, so maybe a combination of the two should be sought. However, *DF* is much simpler to calculate and so may be preferred for some tasks.

Since the correlation between these measures of diversity and combination methods is not *very* high or consistent, the question of the participation of diversity measures in designing classifier ensembles is still open. Directly calculating the accuracy for the chosen combination method makes more sense than calculating the diversity and trying to predict the accuracy, with the measures currently at our disposal. Even if the measure of diversity is easier to calculate than some combination methods, the ambiguous relationship between diversity and accuracy discourages optimising the diversity.

One avenue that might suggest a useful method for building classifier teams based on diversity is finding a more precise formulation of the notion of diversity and thereby constructing a more practical measure. Until then, different heuristics can be explored.

Acknowledgements

We gratefully acknowledge the suggestions given by the anonymous reviewers.

Appendix A. Proof of equivalence relationships

Proposition 1. *Let* $\mathscr{D} = \{D_1, \ldots, D_L\}, \Omega = \{\omega_1, \omega_2\}$. *Let* $a_1, \ldots, a_L$ *be the outputs of the classifiers for class* ω_1, *and* $1 - a_1, \ldots, 1 - a_L$ *be the outputs for class* ω_2, $a_i \in [0, 1]$. *Then the class label assigned to* $\mathbf{x}$ *by the MAX and MIN combination rules will be the same.*

Proof. Without loss of generality assume that $a_1 = \min_i a_i$, and $a_L = \max_i a_i$. Then the minimum combination rule will pick a_1 and $1 - a_L$ as the support for ω_1 and ω_2, respectively, and the maximum rule will pick a_L and $1 - a_1$. Consider the three possible relationships between a_1 and $1 - a_L$.

If $a_1 > 1 - a_L$ then $a_L > 1 - a_1$, and we would select class ω_1 *with both methods*,

If $a_1 < 1 - a_L$ then $a_L < 1 - a_1$, and we would select class ω_2 *with both methods*.

If $a_1 = 1 - a_L$ then $a_L = 1 - a_1$, and we will pick a class at random *with both methods*. □

Proposition 2. *Let* $L = 3$ *so that* $\mathscr{D} = \{D_1, D_2, D_3\}$. *Then Ent and kw, calculated from a data set* $Z = \{\mathbf{z}_1, \ldots \mathbf{z}_n\}, \mathbf{z}_j \in \mathbb{R}^n$, *are equivalent up to a coefficient, i.e.,* $kw = 2/9 Ent$:

Proof.

For three classifiers :

$$kw = \frac{1}{9N} \sum_{j=1}^{N} l(\mathbf{z}_j)\big(3 - l(\mathbf{z}_j)\big),$$

$$Ent = \frac{1}{N} \sum_{j=1}^{N} \min\big\{l(\mathbf{z}_j), 3 - l(\mathbf{z}_j)\big\},$$

where $l(\mathbf{z}_j)$ is the number of classifiers that correctly classify object $\mathbf{z}_j$, therefore we need to show that:

$$\frac{1}{9N} \sum_{j=1}^{N} l(\mathbf{z}_j)\big(3 - l(\mathbf{z}_j)\big) = \frac{2}{9}\left(\frac{1}{N} \sum_{j=1}^{N} \min\big\{l(\mathbf{z}_j), 3 - l(\mathbf{z}_j)\big\}\right).$$

Consider the possible values of $l(\mathbf{z}_j)$ with three classes, and the respective values for *Ent* and *kw* in Table 8.

We can see that the sum of entries from column 4 of Table 8 will always be twice the sum of the

Table 8
Possible values for *Ent* and *kw* from the different values of $l(\mathbf{z}_j)$

$l(\mathbf{z}_j)$	$(3 - l(\mathbf{z}_j))$	*Ent* $\min\{l(\mathbf{z}_j), 3 - l(\mathbf{z}_j)\}$	*kw* $l(\mathbf{z}_j) \times (3 - l(\mathbf{z}_j))$
0	3	0	0
1	2	1	2
2	1	1	2
3	0	0	0

corresponding entries from column 3 of Table 8. Denote $\mathscr{B} = \sum_{j=1}^{N} b_j$ where $b_j = \min\{l(\mathbf{z}_j), 3 - l(\mathbf{z}_j)\}$.

Then $Ent = \frac{1}{N}\mathscr{B}$ and $kw = \frac{1}{9N}2\mathscr{B} = \frac{2}{9}Ent$. □

Note that this only holds for the case when there are three classifiers. If there are four or more classifiers there is no linear relationship between the values for *kw* and *Ent* as in the table.

References

[1] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (1) (2000) 4–37.

[2] L. Lam, Classifier combinations: implementations and theoretical issues, in: J. Kittler, F. Roli (Eds.), Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857, Springer, Cagliari, Italy, 2000, pp. 78–86.

[3] P. Cunningham, J. Carney. Diversity versus quality in classification ensembles based on feature selection. Technical report TCD-CS-2000-02, Department of Computer Science, Trinity College Dublin, 2000.

[4] S. Hashem, Treating harmful collinearity in neural network ensembles, in: A.J.C. Sharkey (Ed.), Combining Artificial Neural Nets, Springer, London, 1999.

[5] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P.W. Duin, Is independence good for combining classifiers?, in: Proceedings of 15th International Conference on Pattern Recognition, Barcelona, Spain, vol. 2, 2000, pp. 169–171.

[6] L. Breiman, Combining predictors, in: A.J.C. Sharkey (Ed.), Combining Artificial Neural Nets, Springer, London, 1999.

[7] H. Drucker, Boosting using neural networks, in: A.J.C. Sharkey (Ed.), Combining Artificial Neural Nets, Springer, London, 1999.

[8] S. Ghahramani, Fundamentals of Probability, second ed., Prentice Hall, New Jersey, 2000.

[9] R.E. Schapire, Y. Freund, P. Bartlett, W.S. Lee, Boosting the margin: a new explanation for the effectiveness of voting methods, The Annals of Statistics 26 (5) (1998) 1651–1686.

[10] J. Wickramaratna, S. Holden, B. Buxton, Performance degradation in boosting, in: J. Kittler, F. Roli (Eds.), Proceedings of the Second International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 2096, Springer, Cambridge, UK, 2001, pp. 11–21.

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

[12] K. Tumer, J. Ghosh, Analysis of decision boundaries in linearly combined neural classifiers, Pattern Recognition 29 (2) (1996) 341–348.

[13] K. Tumer, J. Ghosh, Linear and order statistics combiners for pattern classification, in: A.J.C. Sharkey (Ed.), Combining Artificial Neural Nets, Springer, London, 1999.

[14] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P.W. Duin, Limits on the majority vote accuracy in classifier fusion, Pattern Analysis and Applications, accepted.

[15] B.E. Rosen, Ensemble learning using decorrelated neural networks, Connection Science 8 (3/4) (1996) 373–383.

[16] Y. Liu, X. Yao, Negatively correlated neural networks for classification, in: Proceedings of the 3rd International Symposium on Artificial Life and Robotics (AROBIII'98), Japan, 1998, pp. 736–739.

[17] Y. Liu, X. Yao, Simultaneous learning of negatively correlated neural network, in: Proceedings of the 9th Australian Conference on Neural Networks (ACNN'98), Brisbane, Australia, 1998, pp. 183–187.

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

[19] P. Cunningham. Overfitting and diversity in classification ensembles based on feature selection, Technical report TCD-CS-2000-07, Department of Computer Science, Trinity College Dublin, 2000.

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

[21] K.-D. Wernecke, A coupling procedure for discrimination of mixed data, Biometrics 48 (1992) 497–506.

[22] L.I. Kuncheva, J.C. Bezdek, R.P.W. Duin, Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognition 34 (2) (2001) 299–314.

[23] L. Xu, A. Krzyzak, C.Y. Suen, Methods of combining multiple classifiers and their application to handwriting recognition, IEEE Transactions on Systems Man and Cybernetics 22 (1992) 418–435.

[24] G.U. Yule, On the association of attributes in statistics, Philosophy of Transactions A 194 (1900) 257–319.

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

[26] D.B. Skalak, The sources of increased accuracy for two proposed boosting algorithms, in: Proceedings of American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop, 1996.

[27] G. Giacinto and F. Roli, Design of effective neural network ensembles for image classification processes, Image Vision and Computing Journal (2000), to appear.

[28] P.H.A. Sneath, R.R. Sokal, Numerical Taxonomy, W.H. Freeman and Co, New York, 1973.

[29] R. Kohavi, D.H. Wolpert, Bias plus variance decomposition for zero–one loss functions, in: L. Saitta (Ed.), Machine Learning: Proceedings of the 13th International Conference, Morgan Kaufmann, Los Altos, CA, 1996, pp. 275–283.

[30] J.L. Fleiss, Statistical Methods for Rates and Proportions, Wiley, New York, 1981.

[31] L.K. Hansen, P. Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (10) (1990) 993–1001.

[32] D. Partridge, W.J. Krzanowski, Software diversity: practical statistics for its measurement and exploitation, Information and Software Technology 39 (1997) 707–717.

[33] D. Partridge, W. Krzanowski. Distinct failure diversity in multiversion software (personal communication).

[34] L.I. Kuncheva, C.J. Whitaker. Measures of diversity in classifier ensembles, Machine Learning (submitted).

[35] L.I. Kuncheva, C.J. Whitaker, Ten measures of diversity in classifier ensembles: limits for two classifiers, in: Proceedings of IEE Workshop on Intelligent Sensor Processing, Birmingham, February 2001, IEE, London, 2001, pp. 10/1–10/6.

[36] R.P.W. Duin, PRTOOLS (Version 2). A Matlab toolbox for pattern recognition. Pattern Recognition Group, Delft University of Technology, June 1997.