Technical Report, 2003
School of Informatics
University of Wales, Bangor

Examining the Relationship Between Majority Vote Accuracy and Diversity in Bagging and Boosting

C.J. Whitaker and L.I. Kuncheva
School of Informatics, University of Wales, Bangor
Bangor, Gwynedd, LL57 1UT, United Kingdom
e-mail: {c.j.whitaker,l.i.kuncheva}@bangor.ac.uk

Abstract

Much current research is undertaken into combining classifiers to increase the classification accuracy. We show, by means of an enumerative example, how combining classifiers can lead to much greater or lesser accuracy than each individual classifier. Measures of diversity among the classifiers taken from the literature are shown to only exhibit a weak relationship with majority vote accuracy. Two commonly used methods of designing classifier ensembles, Bagging and Boosting, are examined on benchmark datasets. Bagging is shown to produce teams with little diversity or improvement in accuracy, while Boosting tends to produce more diverse classifier teams showing an improvement in accuracy.

Keywords Classifier combination; Diversity measures; Majority vote; Bagging; Boosting.

1 Introduction

A classifier is a rule that uses observations made on an object and from these assigns a label to that object. For example, collecting information in the form of questions and biochemical tests, and from these diagnosing the illness of a patient. As no rule is perfect, there will be misclassified objects. The accuracy of a classifier is the probability that the rule gives a correct classification for an object.

Although it is natural, there is no reason why we should restrict ourselves to only using one classifier. We are at liberty to derive more than one rule to assign labels to the objects. For example, we can approach more than one medical expert for a diagnosis in the example above. We then need to combine the labels from all the rules used to give a combined output from the set of classifiers. This area of research is known as Multiple Classifier Systems and uses some ideas from the statistical literature.

The main reason why we would want to combine classifier outputs is to achieve a more accurate decision from the set of classifiers than can be achieved from the best of the classifiers. Classifiers should be different from each other, otherwise there is no gain in combining them. However this difference, also called diversity, is not a simple concept, and has been recognized to be an important research direction in combining classifiers ([15]).

It is assumed that independence is 'good' and dependence is 'bad' in combining classifiers. This arises out of the following observation.

Consider an odd number of classifiers L. The *majority vote rule* labels the object correctly if at least $\lfloor L/2 \rfloor + 1$ classifiers in the team "vote" for the correct class label, where $\lfloor L/2 \rfloor$ denotes the largest integer less than or equal to $L/2$. Assume that all L classifiers have the same accuracy p.

The majority vote method with independent classifier decisions gives an overall correct classification accuracy calculated by the binomial formula

$$P_{maj} = \sum_{m=0}^{\lfloor L/2 \rfloor} \left(\begin{array}{c} L \\ m \end{array} \right) p^{L-m}(1-p)^m. \quad (1)$$

The majority vote method with independent classifiers is guaranteed to give a higher accuracy than individual classifiers when $p > 0.5$, i.e., $P_{maj} > p$ under these conditions ([16, 17]).

While the above argument shows that independence is good, we will show in Section 2 that it is possible to have dependence which is even better. Clearly the concept of dependence (or diversity as it is called) needs to be developed so that its relationship with the majority vote can be examined.

In this paper, Section 2 gives an enumerative example. Section 3 introduces 10 measures of diversity. In Section 4 we continue the example and show how to calculate the diversity measures. Section 5 describes Bagging and Boosting, two popular methods for generating classifier ensembles for a data set. We then quote an experimental result demonstrating the two different ways Bagging and Boosting address diversity. Section 6 contains a discussion of the practical issues of "inserting" diversity in the ensemble, and offers our conclusions.

2 An example

If we think of identical classifiers having a positive dependence and independent classifiers as having no dependence then we might be able to do better if we have negatively dependent classifiers. We can demonstrate this by an example which shows that the majority vote can perform noticeably better and noticeably worse than a single classifier.

Let $\mathcal{D} = \{D_1, D_2, D_3\}$ be a set of three classifiers with the same individual probability of correct classification $p = 0.6$. Suppose that there are 10 objects in a hypothetical data set, and so each classifier labels correctly exactly 6 of them. Each classifier output is recorded as correct (1) or wrong (0). Given these requirements, *all* possible combinations of distributing 10 elements into the 8 combinations of outputs of the three classifiers are shown in Table 1. The penultimate column of Table 1 shows the majority vote accuracy of each of the 28 possible combinations. It is obtained as the proportion (out of 10 elements) of the sum of the entries in columns '111', '101', '011' and '110' (two or more correct votes). The rows of the table are ordered by the majority vote accuracy.

To clarify the entries in Table 1, consider as an example the first row. The number 3 in the column under the heading '101', means that exactly 3 objects are correctly recognized by D_1 and D_3 (the first and the third 1's of the heading) and misclassified by D_2 (the zero in the middle).

The table offers at least two interesting facts

- There is a case where the majority vote produces 90 % correct classification. Although purely hypothetical, this vote distribution is *possible* and offers a dramatic increase over the individual rate $p = 0.6$.
- On the other hand, the majority vote is not guaranteed to do better than a single member of the team. The combination in the bottom row has a majority vote accuracy of 0.4.

Kuncheva et al. [14] have named and studied the best and the worst possible cases illustrated above as "the pattern of success" and the "pattern of failure".

Table 1: All possible combinations of correct/incorrect classification of 10 objects by three classifiers so that each classifier recognizes exactly 6 objects. The entries in the table are the number of occurrences of the specific binary output of the three classifiers in the particular combination. The majority vote accuracy P_{maj} and the improvement over the single classifier, $P_{maj}-p$ are also shown. Three characteristic classifier ensembles are marked.

No	111	101	011	001	110	100	010	000	P_{maj}	$P_{maj}-p$	
1	**0**	**3**	**3**	**0**	**3**	**0**	**0**	**1**	**0.9**	**0.3**	← Pattern of
2	2	2	2	0	2	0	0	2	0.8	0.2	success
3	1	2	2	1	3	0	0	1	0.8	0.2	
4	0	2	3	1	3	1	0	0	0.8	0.2	
5	0	2	2	2	4	0	0	0	0.8	0.2	
6	4	1	1	0	1	0	0	3	0.7	0.1	
7	3	1	1	1	2	0	0	2	0.7	0.1	
8	2	1	2	1	2	1	0	1	0.7	0.1	
9	2	1	1	2	3	0	0	1	0.7	0.1	
10	1	2	2	1	2	1	1	0	0.7	0.1	
11	1	1	2	2	3	1	0	0	0.7	0.1	
12	1	1	1	3	4	0	0	0	0.7	0.1	
13	6	0	0	0	0	0	0	4	0.6	0.0	← Identical
14	5	0	0	1	1	0	0	3	0.6	0.0	classifiers
15	4	0	1	1	1	1	0	2	0.6	0.0	
16	4	0	0	2	2	0	0	2	0.6	0.0	
17	3	1	1	1	1	1	1	1	0.6	0.0	
18	3	0	1	2	2	1	0	1	0.6	0.0	
19	3	0	0	3	3	0	0	1	0.6	0.0	
20	2	1	1	2	2	1	1	0	0.6	0.0	
21	2	0	2	2	2	2	0	0	0.6	0.0	
22	2	0	1	3	3	1	0	0	0.6	0.0	
23	2	0	0	4	4	0	0	0	0.6	0.0	
24	5	0	0	1	0	1	1	2	0.5	-0.1	
25	4	0	0	2	1	1	1	1	0.5	-0.1	
26	3	0	1	2	1	2	1	0	0.5	-0.1	
27	3	0	0	3	2	1	1	0	0.5	-0.1	
28	**4**	**0**	**0**	**2**	**0**	**2**	**2**	**0**	**0.4**	**-0.2**	← Pattern of
											failure

Table 2: 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 Measures of diversity

The problem we are trying to solve is the following. We will be presented with N objects, on each of which we have available n measurements. So the data for each object is an n-dimensional vector $\mathbf{z}_j \in \Re^n$. Together with this data vector will be a preassigned (presumably correct) label ω out of the possible label set $\Omega = \{\omega_1, \ldots, \omega_c\}$. Thus we have a labelled data set $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ to use for training.

Let $\mathcal{D} = \{D_1, \ldots, D_L\}$ be an ensemble of classifiers built on the data set Z such that $D_i : \Re^n \to \Omega$ for $\mathbf{x} \in \Re^n$. For each classifier D_i, we record whether it correctly classifies $\mathbf{z}_j$ (the label it produces matches the true label) or not. Thus we construct an 'oracle' output of each classifier as the N-dimensional binary vector $\mathbf{y}_i = [y_{1,i}, \ldots, y_{N,i}]^T$, such that $y_{j,i} = 1$ if D_i correctly recognizes z_j and 0 otherwise for the L classifiers D_i, $i = 1, \ldots, L$.

To date no "best" measure of diversity or dependency has been found. Broadly, the measures in the literature fall into two types. There are pairwise measures which are calculated for each pair of classifiers in $\mathcal{D}$ and then averaged; and non-pairwise measures that either use the idea of entropy or correlation of individual outputs with the averaged output of $\mathcal{D}$, or are based on the distribution of "difficulty" of the data points.

In this study we present 10 measures of classifier diversity for oracle classifier outputs: 4 pairwise and 6 non-pairwise.

3.1 Pairwise diversity measures

Consider two classifiers D_i and D_k, and a 2×2 table of probabilities that summarizes their combined outputs as in Table 2.

Many pairwise statistics have been proposed as measures of similarity in the numerical taxonomy literature (e.g., [22]), 4 of which are shown in Table 3.

$Q_{i,k}$ and $\rho_{i,k}$ both vary between -1 and $+1$, and for statistically independent classifier outputs equal 0. The disagreement and double-fault measures vary from 0 to $+1$ and the value for independent classifiers depends on the accuracies of the two classifiers (Kuncheva and Whitaker (2001)).

For all four measures, when there are L classifiers, we calculate the mean of all the $L(L-1)/2$ pairwise values.

3.2 Non-pairwise diversity measures

We denote by $l(\mathbf{z}_j) = \sum_{i=1}^{L} y_{j,i}$ the number of classifiers in $\mathcal{D}$ that correctly recognise $\mathbf{z}_j$.

Kohavi-Wolpert variance. We can consider $\mathbf{y}$ as a Bernoulli variable taking values 0 and 1. As suggested by Kohavi and Wolpert (1996), we can use the average of the variance for each object in Z as a measure of diversity. This leads to the following measure

Table 3: Pairwise measures of diversity

Measure	Reference	Abbreviation	Formula
Q statistic	Yule (1900) [23]	$Q_{i,k}$	$\frac{ad-bc}{ad+bc}$
Correlation coefficient	Sneath and Sokal (1973) [22]	$\rho_{i,k}$	$\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$
Disagreement measure	Skalak (1996) [21] Ho (1998) [10]	$D_{i,k}$	$b+c$
Double-fault measure	Giacinto and Roli (2000) [8]	$DF_{i,k}$	d

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

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

$$KW = \frac{L-1}{2L} D_{av}. \tag{3}$$

(The proof of the equivalence is given in (Kuncheva and Whitaker, 2002, submitted).)

Measurement of interrater agreement κ. ([5]). If we denote $\bar{p}$ to be the mean of the classification accuracy from the L classifiers, then

$$\kappa = 1 - \frac{\frac{1}{L}\sum_{j=1}^{N} l(\mathbf{z}_j)(L - l(\mathbf{z}_j))}{N(L-1)\bar{p}(1-\bar{p})} \tag{4}$$

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

$$\kappa = 1 - \frac{L}{(L-1)\bar{p}(1-\bar{p})} KW = 1 - \frac{1}{2\bar{p}(1-\bar{p})} D_{av}. \tag{5}$$

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 $\mathbf{y}_j$ 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

$$E = \frac{1}{N}\frac{2}{L-1}\sum_{j=1}^{N}\min\left\{\sum_{i=1}^{L} y_{j,i}, L - \sum_{i=1}^{L} y_{j,i}\right\}. \tag{6}$$

E varies between 0 and 1, where 0 indicates no difference and 1 indicates the highest possible diversity. Let all classifiers have the same individual accuracy p. Then while value 0 is achievable for any number of classifiers L and any p, the value 1 can only be attained for $p \in \left[\frac{L-1}{2L}, \frac{L+1}{2L}\right]$.

The measure of difficulty θ . The idea for this measure came from a study by ([9]). 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 classified an input $\mathbf{x}$ drawn randomly from the distribution of the problem. The measure of *difficulty* θ is defined as

$$\theta = Var(X). \tag{7}$$

For identical classifiers X only takes the values 0 and 1 and so θ takes its maximum value, which leads to the higher the value of θ, the less diverse the classifier team.

Generalized diversity. This measure has been proposed in [20]. Let Y be a random variable expressing the proportion of classifiers (out of L) that **fail** on a randomly drawn object $\mathbf{x} \in \Re^n$. Denote by p_i the probability that $Y = \frac{i}{L}$, i.e., i classifiers fail simultaneously on a randomly drawn $\mathbf{x}$. (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 $\mathbf{x}$. The probability $p(1)$ that a randomly chosen classifier will fail on a randomly chosen $\mathbf{x}$ is

$$\begin{aligned} p(1) &= \sum_{i=1}^{L} Pr(\text{chosen classifier fails}|\text{exactly } i \text{ fail})Pr(\text{exactly } i \text{ fail}) \\ &= \sum_{i=1}^{L} \frac{i}{L} p_i. \end{aligned} \tag{8}$$

Suppose that two classifiers are randomly picked from $\mathcal{D}$. The probability $p(2)$ that they both will fail on a randomly chosen $\mathbf{x}$ is

$$\begin{aligned} p(2) &= \sum_{i=1}^{L} Pr(\text{both chosen classifiers fail}|\text{exactly } i \text{ fail})Pr(\text{exactly } i \text{ fail}) \\ &= \sum_{i=1}^{L} \frac{i}{L}\frac{(i-1)}{(L-1)} p_i. \end{aligned} \tag{9}$$

Partridge and Krzanowski argue that maximum diversity occurs when failure of one of the two randomly chosen classifiers is accompanied by correct labeling by the other classifier. Thus, for the maximum diversity case, the probability of both classifiers failing is $p(2) = 0$. Minimum diversity occurs when the two randomly picked classifiers are both correct or both wrong on a randomly picked $\mathbf{x}$, i.e., they behave as a single classifier. Thus, for the minimum diversity case, the probability $p(2)$ of both classifiers failing is the same as the probability of one randomly picked classifier failing, i.e., $p(1)$. So $p(2)$ spans the range from 0 (diverse) to $p(1)$ (nondiverse). Then the generalized diversity measure GD is defined as the normalized $p(2)$

Table 4: A distribution of the votes of three classifiers (row 27 from Table 1)

D_1, D_2, D_3	111	101	011	001	110	100	010	000
Frequency	3	0	0	3	2	1	1	0

Table 5: The three pairwise tables for the distribution in Table 4

D_1, D_2	
0.5	0.1
0.1	0.3

D_1, D_3	
0.3	0.3
0.3	0.1

D_2, D_3	
0.3	0.3
0.3	0.1

$$GD = \frac{p(1) - p(2)}{p(1)} = 1 - \frac{p(2)}{p(1)}. \tag{10}$$

Coincident failure diversity. This is a modification of GD proposed in (Partridge and Krzanowski (1999)).

If the classifiers in $\mathcal{D}$ are identical, then they will all be correct or all be wrong for a randomly drawn $\mathbf{x}$. Therefore $p_0 + p_L = 1$, and all other values will be $p_1 = p_2 = \ldots = p_{L-1} = 0$. This case corresponds to minimum diversity. On the other hand, consider the case where all misclassifications are "unique", i.e., there is at most one failure for any randomly drawn $\mathbf{x}$. Then $p_0 + p_1 = 1$, and $p_2 = \ldots = p_L = 0$. We can deem this to be the most diverse case, and require the diversity measure to attain its maximum. The Coincident Failure Diversity (CFD) is proposed as

$$CFD = \begin{cases} 0, & p_0 = 1.0; \\ \frac{1}{1-p_0} \sum_{i=1}^{L} \frac{L-i}{L-1} p_i, & p_0 < 1 \end{cases} \tag{11}$$

CFD is zero for $p_0 + p_L = 1$, and 1 for $p_0 + p_1 = 1$, $p_0 < 1$ (Note that $p_0 = 1$ means identical and completely accurate classifiers as there are no misclassifications.)

The diversity measures above have been categorized as pairwise or not. They can also be categorized in two further ways: (1) Symmetrical, if reversing the labels of 1 and 0 for correct and incorrect has no effect on the calculated measure, or nonsymmetrical otherwise; and (2) Ascending ($\uparrow$), if the larger the measure means a more diverse set of classifiers, or descending ($\downarrow$) otherwise.

4 The example continued

Take for example row 27 from Table 1. The 10 objects are so distributed that $P_{maj} = 0.5$ even though all three classifiers have accuracy $p = 0.6$. For an easier reference, the distribution of the votes (correct/wrong) of row 27 of Table 1 is duplicated in Table 4.

The three pairwise tables are shown in Table 5.

The pairwise measures of diversity are calculated as follows

$$\begin{aligned} Q_{1,2} &= \frac{5 \times 3 - 1 \times 1}{5 \times 3 + 1 \times 1} = \frac{7}{8} \\ Q_{1,3} = Q_{2,3} &= \frac{3 \times 1 - 3 \times 3}{3 \times 1 + 3 \times 3} = -\frac{1}{2} \\ Q &= \frac{1}{3}\left(\frac{7}{8} - \frac{1}{2} - \frac{1}{2}\right) = -\frac{1}{24} \approx \mathbf{-0.04}; \end{aligned} \tag{12}$$

$$\begin{aligned}
\rho_{1,2} &= \frac{5\times 3-1\times 1}{\sqrt{(5+1)(1+3)(5+1)(1+3)}}=\frac{7}{12}\\
\rho_{1,3}=\rho_{2,3} &= \frac{3\times 1-3\times 3}{(5+1)(1+3)}=-\frac{1}{4}\\
\rho &= \frac{1}{3}\left(\frac{7}{12}-\frac{1}{4}-\frac{1}{4}\right)=\frac{1}{36}\approx \mathbf{0.03};
\end{aligned} \tag{13}$$

$$D = \frac{1}{3}((0.1+0.1)+(0.3+0.3)+(0.3+0.3))=\frac{7}{15}\approx \mathbf{0.47}; \tag{14}$$

$$DF = \frac{1}{3}(0.3+0.3+0.1)=\frac{1}{6}\approx \mathbf{0.17}. \tag{15}$$

The nonpairwise measures KW, κ and E are calculated by

$$\begin{aligned}
KW &= \frac{1}{10\times 3^2}(3\times(1\times 2)+2\times(2\times 1)+1\times(1\times 2)+1\times(1\times 2))\\
&= \frac{7}{45}\approx \mathbf{0.16};
\end{aligned} \tag{16}$$

$$\begin{aligned}
\kappa &= 1-\frac{D}{2\times 0.6\times(1-0.6)}=1-\frac{7/15}{12/25}\\
&= \frac{1}{36}\approx \mathbf{0.03};
\end{aligned} \tag{17}$$

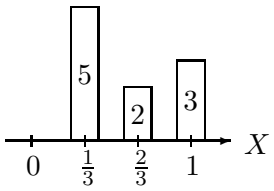
$$\begin{aligned}
E &= \frac{1}{10}\times\frac{2}{(3-1)}\times(3\times\min\{3,0\}+3\times\min\{1,2\}\\
&\quad +2\times\min\{2,1\}+1\times\min\{1,2\}+1\times\min\{1,2\})\\
&= \frac{7}{10}=\mathbf{0.70}.
\end{aligned} \tag{18}$$

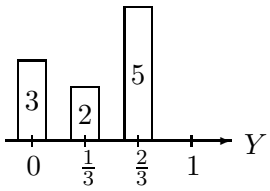
The distribution of the random variables X and Y needed for calculating θ, GD, and CFD are depicted in Figure 1.

The mean of X is 0.6, and the mean of Y ($p(1)$) is 0.4. The respective measures are calculated as follows

$$\begin{aligned}
\theta = Var(X) &= (1/3-0.6)^2\times 0.5+(2/3-0.6)^2\times 0.2+(1-0.6)^2\times 0.3\\
&= \frac{19}{225}\approx \mathbf{0.08};
\end{aligned} \tag{19}$$

$$\begin{aligned}
p(2) &= \frac{2}{3}\times\frac{(2-1)}{(3-1)}\times 0.5=\frac{1}{6};\\
GD &= 1-\frac{1/6}{0.4}=\frac{7}{12}\approx \mathbf{0.58};
\end{aligned} \tag{20}$$





X	0	1/3	2/3	1
$Pr(X)$	0	0.5	0.2	0.3

Y	0	1/3	2/3	1
$Pr(Y)$	0.3	0.2	0.5	0

Figure 1: The frequencies and the probability mass functions of the variables X and Y needed for calculating the diversity measures θ, GD and CFD.

$$\begin{aligned} CFD &= \frac{1}{1-0.3}\left(\frac{(3-1)}{(3-1)} \times 0.2 + \frac{(3-2)}{(3-1)} \times 0.5\right) \\ &= \frac{9}{14} \approx \mathbf{0.64}. \end{aligned} \tag{21}$$

Calculated in this way, the values of the 10 diversity measures for all distributions of classifier votes from Table 1 are shown in Table 6. To enable cross-referencing, the last two columns of the table show P_{maj} and the improvement $P_{maj} - P$. The rows are arranged in the same order as in Table 1.

With 10 objects, it is not possible to model pairwise independence. The table of probabilities for this case will contain $a = 0.36$, $b = c = 0.24$, and $d = 0.16$. To use 10 object, we have to round so that $a = 0.4$, $b = c = 0.2$, and $d = 0.2$, but instead of 0, this gives a value of Q

$$Q = \frac{0.08 - 0.04}{0.08 + 0.04} = \frac{1}{3}.$$

In this sense, "closest to independence" are rows 2, 17 and 23 in Table 1. It is not easy to spot by eye in Table 6 any relationship between diversity and accuracy for any of the 10 diversity measures. Instead of trying to quantify the relationship, we give a scatterplot of diversity versus improvement in Figure 3.

5 Bagging and Boosting

5.1 Bagging

Bagging ([2]) is a strategy for creating a team of classifiers. Bagging works by taking a bootstrap sample from the data set and building a classifier on the sample. Further bootstrap samples are taken and a classifier built on each. The classifier can be, e.g., Fisher's linear discriminant classifier, but many classifiers can be used ([4]). The final classification decision for an unlabeled data point $\mathbf{x}$ is made by taking the majority vote over the class labels produced by the L classifiers.

5.2 Boosting

Boosting ([6]) is another strategy for creating a team of classifiers. There are a number of variations of the Boosting algorithm. One of the most successful is the AdaBoost method with the reweighting

Table 6: The 10 diversity measures for the 28 distributions of classifier votes in Table 1. The characteristic ensembles are separated with lines: (row 1) pattern of success, (row 13) identical classifiers and (row 28) pattern of failure.

	Diversity										Accuracy	
No	Q	ρ	D	DF	KW	κ	E	θ	GD	CFD	P_{maj}	$P_{maj}-p$
1	-0.50	-0.25	0.60	0.10	0.20	-0.25	0.90	0.04	0.75	0.90	0.9	0.3
2	0.33	0.17	0.40	0.20	0.13	0.17	0.60	0.11	0.50	0.75	0.8	0.2
3	-0.22	-0.11	0.53	0.13	0.18	-0.11	0.80	0.06	0.67	0.83	0.8	0.2
4	-0.67	-0.39	0.67	0.07	0.22	-0.39	1.00	0.02	0.83	0.90	0.8	0.2
5	-0.56	-0.39	0.67	0.07	0.22	-0.39	1.00	0.02	0.83	0.90	0.8	0.2
6	0.88	0.58	0.20	0.30	0.07	0.58	0.30	0.17	0.25	0.50	0.7	0.1
7	0.51	0.31	0.33	0.23	0.11	0.31	0.50	0.13	0.42	0.64	0.7	0.1
8	0.06	0.03	0.47	0.17	0.16	0.03	0.70	0.08	0.58	0.75	0.7	0.1
9	-0.04	0.03	0.47	0.17	0.16	0.03	0.70	0.08	0.58	0.75	0.7	0.1
10	-0.50	-0.25	0.60	0.10	0.20	-0.25	0.90	0.04	0.75	0.83	0.7	0.1
11	-0.39	-0.25	0.60	0.10	0.20	-0.25	0.90	0.04	0.75	0.83	0.7	0.1
12	-0.38	-0.25	0.60	0.10	0.20	-0.25	0.90	0.04	0.75	0.83	0.7	0.1
13	1.00	1.00	0.00	0.40	0.00	1.00	0.00	0.24	0.00	0.00	0.6	0.0
14	0.92	0.72	0.13	0.33	0.04	0.72	0.20	0.20	0.17	0.30	0.6	0.0
15	0.69	0.44	0.27	0.27	0.09	0.44	0.40	0.15	0.33	0.50	0.6	0.0
16	0.56	0.44	0.27	0.27	0.09	0.44	0.40	0.15	0.33	0.50	0.6	0.0
17	0.33	0.17	0.40	0.20	0.13	0.17	0.60	0.11	0.50	0.64	0.6	0.0
18	0.24	0.17	0.40	0.20	0.13	0.17	0.60	0.11	0.50	0.64	0.6	0.0
19	0.00	0.17	0.40	0.20	0.13	0.17	0.60	0.11	0.50	0.64	0.6	0.0
20	-0.22	-0.11	0.53	0.13	0.18	-0.11	0.80	0.06	0.67	0.75	0.6	0.0
21	-0.11	-0.11	0.53	0.13	0.18	-0.11	0.80	0.06	0.67	0.75	0.6	0.0
22	-0.21	-0.11	0.53	0.13	0.18	-0.11	0.80	0.06	0.67	0.75	0.6	0.0
23	-0.33	-0.11	0.53	0.13	0.18	-0.11	0.80	0.06	0.67	0.75	0.6	0.0
24	0.88	0.58	0.20	0.30	0.07	0.58	0.30	0.17	0.25	0.30	0.5	-0.1
25	0.51	0.31	0.33	0.23	0.11	0.31	0.50	0.13	0.42	0.50	0.5	-0.1
26	0.06	0.03	0.47	0.17	0.16	0.03	0.70	0.08	0.58	0.64	0.5	-0.1
27	-0.04	0.03	0.47	0.17	0.16	0.03	0.70	0.08	0.58	0.64	0.5	-0.1
28	0.33	0.17	0.40	0.20	0.13	0.17	0.60	0.11	0.50	0.50	0.4	-0.2

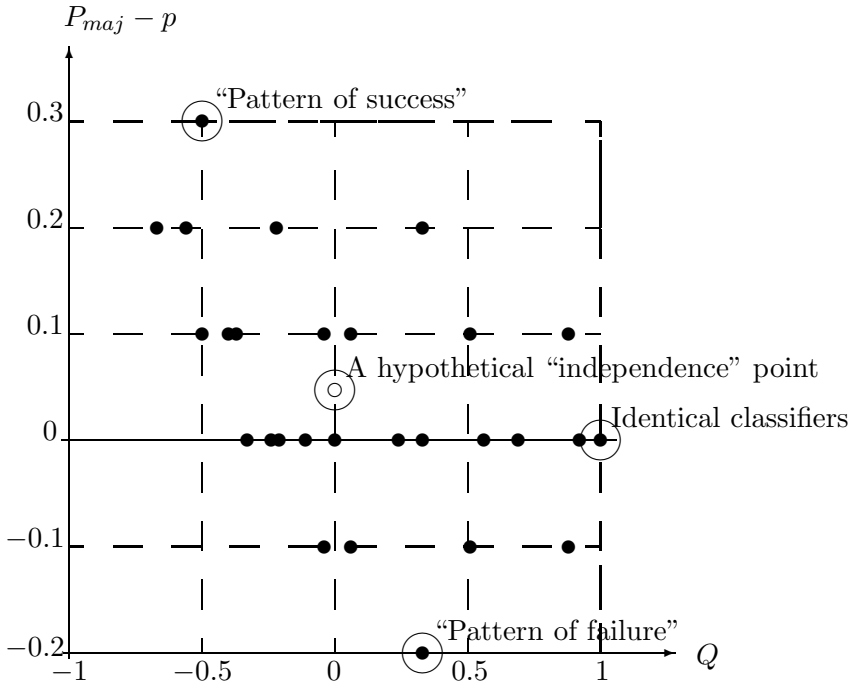


Figure 2: Improvement on the individual accuracy ($P_{maj} - p$) versus Q.

implementation (see [3, 7, 1] for a discussion of their relative merits).

Boosting, like Bagging, builds a classifier on each new sample taken from the dataset. However, while Bagging relies on random and independent changes in the samples taken to develop the classifiers, Boosting uses deterministic changes to ensure the sample contains more 'difficult to classify' objects. In this way it is expected that diversity will be induced into the classifier team.

5.3 Benchmark datasets

In order to examine Bagging and Boosting we report here the conclusions of a separate study (Kuncheva, Skurichina and Duin, 2002,[11]). Seven 2-class data sets were used:- five benchmark datasets from the UCI Machine Learning Repository and two synthetic datasets available from the University of Delft. Table 7 shows the number of features and cases for each dataset. Both Bagging and Boosting were used to generate teams of 3 classifiers. Figures 4 and 5 show the results of this. Each black dot on the plots corresponds to a classifier team. 124 teams were generated using a training sample size ranging from 3 to 200, with either the Nearest Mean or the Pseudo-Fisher Linear Discriminant classifier.

The grey dots show the results of an enumerative experiment, similar to the example in section 2, but performed using 40 objects. For this enumerative experiment we had to fix p to a constant. Since in the benchmark experiments the individual accuracies ranged from 0.54 to 0.98, the enumeration was run 6 times with $p \in \{0.54, 0.6, 0.7, 0.8, 0.9, 0.98\}$. In essence, the black dots of Figure 5 show the results of using Bagging and Boosting while the grey dots show *the possible values* for the combination of improvement and diversity. Some observed values of majority vote accuracy and diversity occur outside the grey background. This is because the enumeration was run under the assumption that

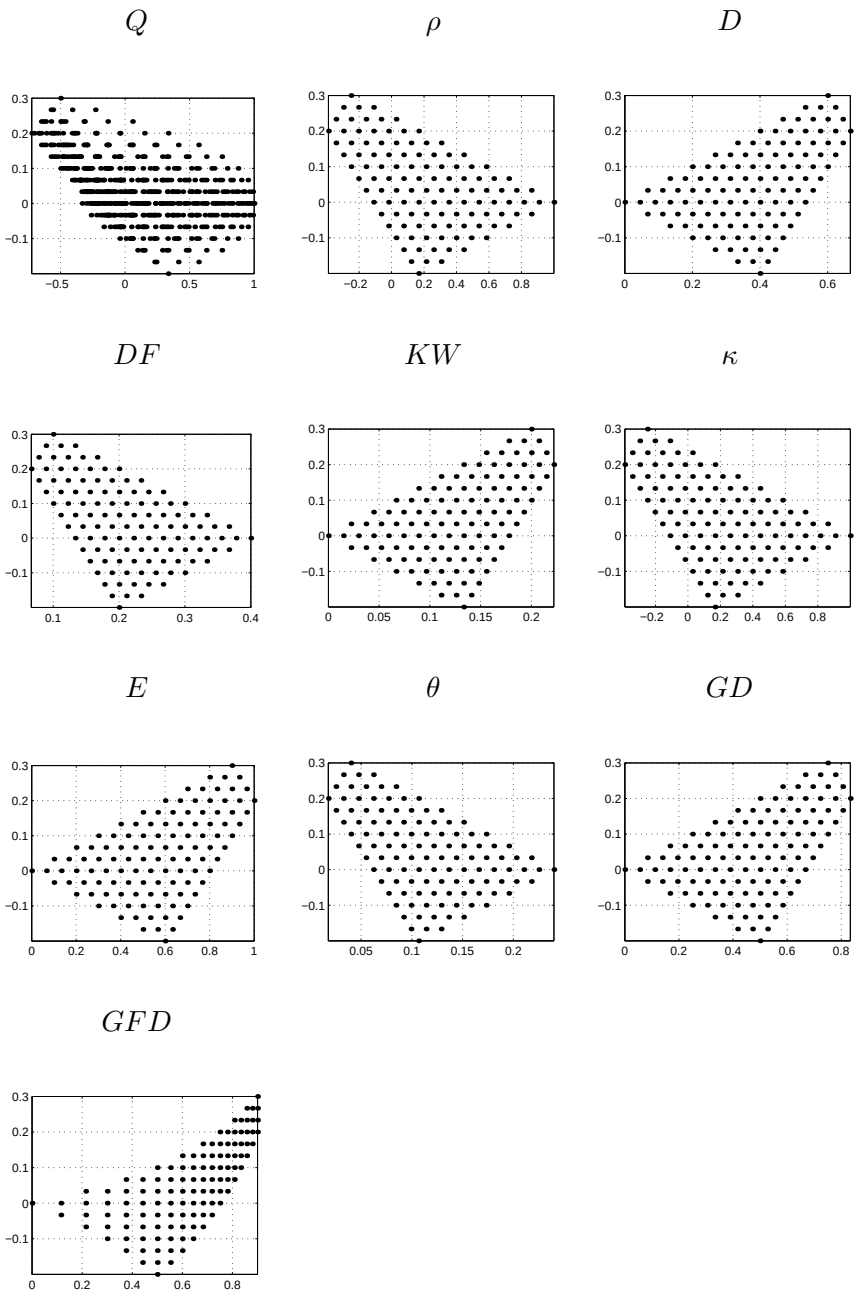


Figure 3: Improvement on the individual accuracy ($P_{maj} - p$) versus 10 diversity measures.

Table 7: Summary of the 7 two-class data sets used.

Name	n	N	Availability
80-D Correlated Gauss	80	1000	Delft
80-D Rotated Correlated Gauss	80	1000	Delft
Pima Indians Diabetes	8	768	UCI
Ionosphere	34	351	UCI
Wisconsin Diagnostic Breast Cancer	30	569	UCI
Sonar	60	198	UCI
German	24	1000	UCI

Notations:

n: number of features
N: number of cases in the database
UCI: http://www.ics.uci.edu/∼mlearn/MLRepository.html
Delft: The data is available by request from <marina@ph.tn.tudelft.nl>

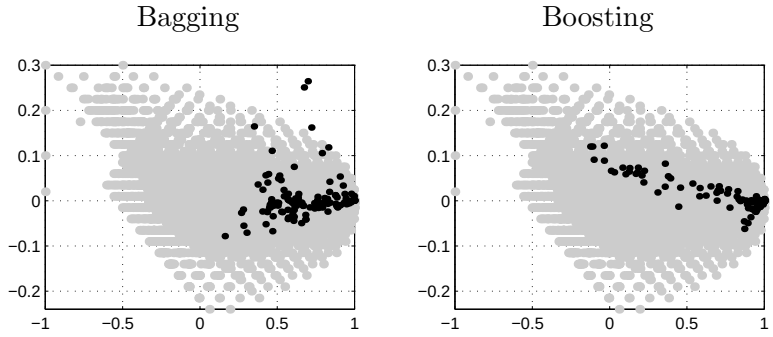


Figure 4: Improvement versus Q: an experiment

the individual accuracies in the team are equal. This special case cannot be expected to happen in practice, and the violation of the assumption explains the black points (classifier ensembles) outside the grey areas. Thus the grey background is only an approximate guideline for the position of the points in the real experiment rather than a firm region.

Only the diversity measures Q, ρ, D, θ and CFD are shown as the others have graphs that are very similar to one of the others.

6 Discussion and conclusions

By means of a simple enumerative example we have shown how the combination of three moderately accurate classifiers can result in a much more accurate classifier team. We have also shown that combining classifiers can also lead to less accuracy. This shows that there is scope for a great improvement in the accuracy of a team of classifiers but this is not guaranteed to be achieved. Our studies of the Patterns of Success and Failure (Kuncheva et al. (to appear)) have found that

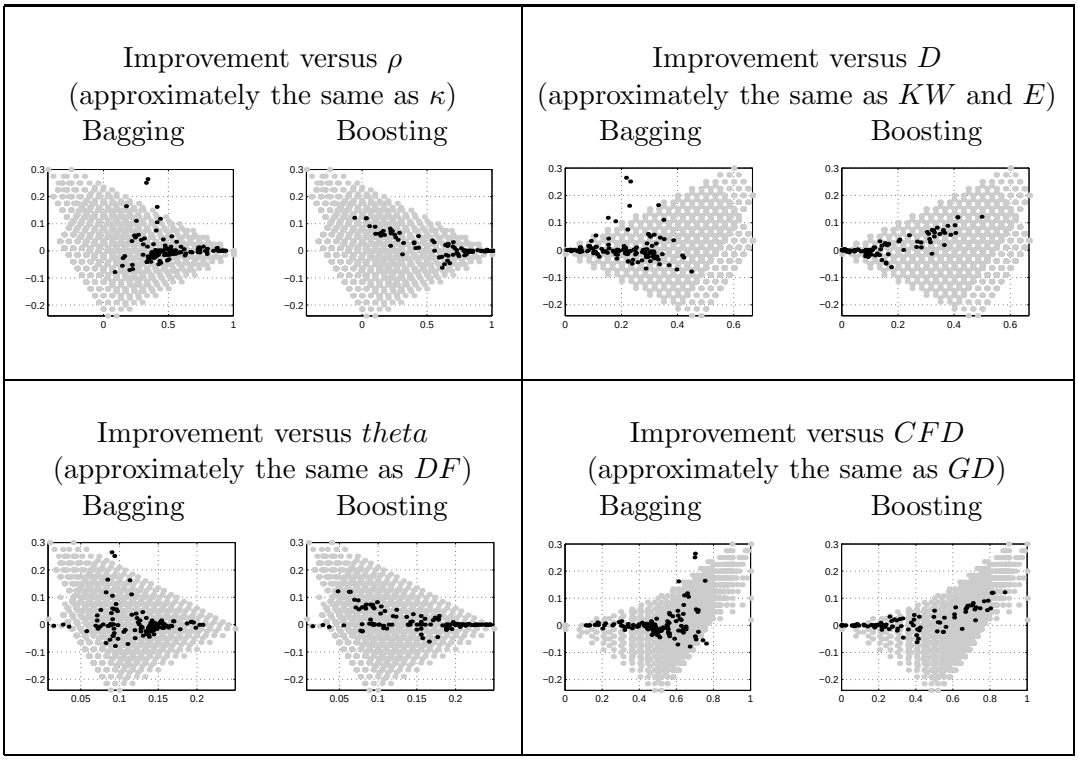


Figure 5: Improvement versus diversity: an experiment

if we combine three classifiers each of accuracy $p > \frac{2}{3}$, then the team can potentially give no misclassifications. All we have to be able to do is develop and combine appropriate classifiers together to form the team. However, developing classifiers independently of each other does not necessarily lead to statistically independent classifier outputs [18]. For example, bagging, where the samples where the samples are independently chosen, does not produce teams with independent classifier decisions as can be seen in Figures 4 and 5.

Measuring the diversity of a team of classifiers is the currently accepted idea that may result in a useful way to decide which classifiers to combine together to form the team. The hope and expectation is that if we can combine diverse classifiers then the team accuracy will be increased. After examining the literature we have found and defined ourselves (Kuncheva and Whitaker, submitted) ten diversity measures. The results of our enumerative experiments all show that there is no clear cut or strong relationship between any of the diversity measures and the majority vote accuracy. This somewhat negative finding could be interpreted in a number of ways:- (1) there is only a weak relationship between diversity and accuracy or (2) we do not have yet a good measure of diversity or (3) diversity is a multivariate rather than a univariate concept. As we noted in the Introduction diversity is not a simple concept, perhaps we have just confirmed this. Taking a more positive view of the findings, we could say that none of the ten diversity measures stand out as being particularly better or worse than any other. This means that for whatever purpose we are going to use a diversity measure, we are at liberty to use the simplest measure as they are all equally useful. This interpretation may suggest that a pairwise measure may be the statistic to use when forming classifier teams by a forward stepwise type algorithm.

In practice it is not possible to examine all the possible classifiers that could be built for a real

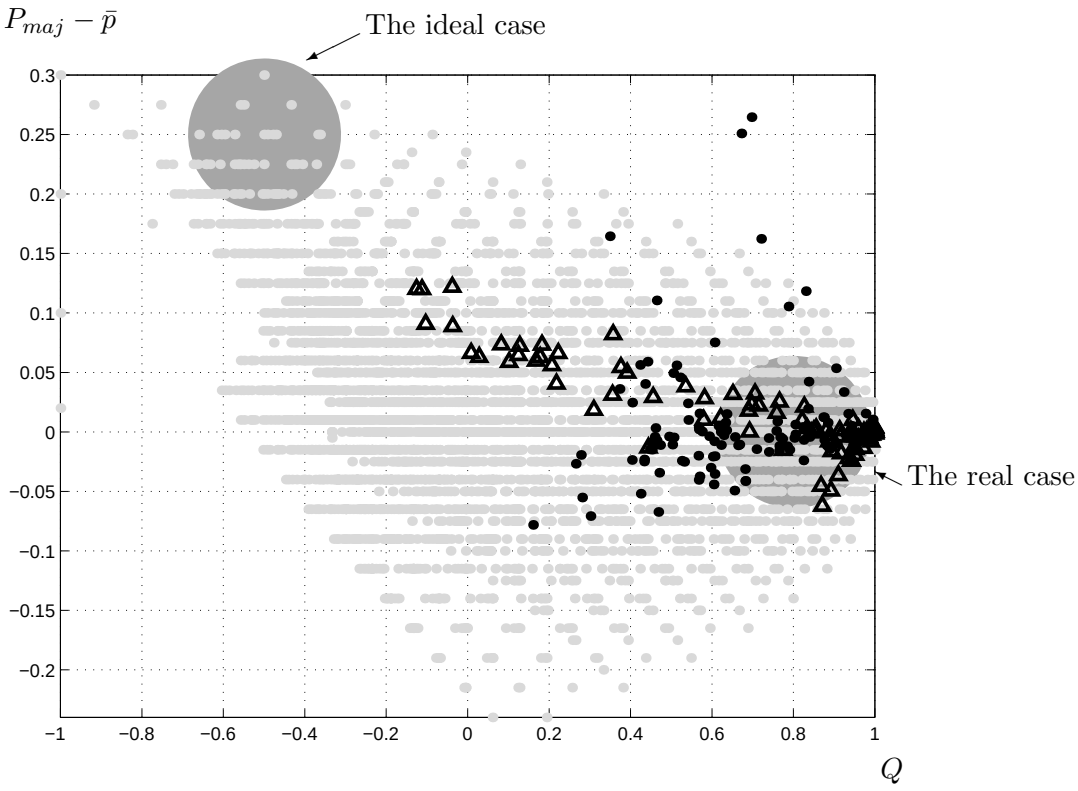


Figure 6: The desired and the real cases overlayed on the Q graph together with the Bagging ($\bullet$) and Boosting ($\triangle$) results.

dataset of a specified size. The Bagging and Boosting algorithms are currently accepted as the best available methods to generate teams that will lead to greater accuracy. Figures 4 and 5 show that for Bagging there is most often no improvement as the values are clustered around zero. However there are some notable exceptions. These exceptions are also outside the area of possible values (the grey dots), for the reasons stated above. The Boosting algorithm tends to give better results than Bagging as there are fewer negative and more positive values for the improvement, as shown in Figures 4 and 5. The other noticeable aspect of these Figures is that there is more evidence of a relationship between the improvement and all the diversity measures for Boosting rather than Bagging. It appears that taking bootstrap samples (Bagging) leads to classifiers with relatively little diversity, while samples designed to include more 'difficult to classify' objects (Boosting) leads, as the method was intended, to more diverse classifier teams.

Figure 6 shows the combined plot for Bagging and Boosting of Figure 4 for the Q diversity measure. Also shown are a circle centred at $Q = 0.8$ and improvement $= 0$ which contains the majority of the real data results (black dots). This means that the majority of classifier teams show little diversity and there is little evidence that the team does better than the best individual classifier. Also shown is a similarly size circle that is centred at $Q = -0.5$ and improvement $= 0.25$. This is the ideal where we would want the results to occur, but none do. Boosting does tend to lead to teams getting closer to the ideal but there is still room for an improvement in the algorithms for combining classifiers to achieve greater accuracy.

References

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

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

[3] L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–849, 1998.

[4] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, NY, second edition, 2001.

[5] J.L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 1981.

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

[7] Y. Freund and R.E. Schapire. Discussion of the paper "Arcing Classifiers" by Leo Breiman. *The Annals of Statistics*, 26(3):824–832, 1998.

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

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

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

[11] L.I. Kuncheva, M. Skurichina, and R.P.W. Duin. An experimental study on diversity for bagging and boosting. *Information Fusion*, 3:245–258, 2002.

[12] L.I. Kuncheva and C.J. Whitaker. Ten measures of diversity in classifier ensembles: limits for two classifiers. In *Proc. IEE Workshop on Intelligent Sensor Processing*, pages 10/1–10/6, Birmingham, February 2001. IEE.

[13] L.I. Kuncheva and C.J. Whitaker. Measures of diversity in classifier ensembles. *Machine Learning*, 51:181–207, 2003.

[14] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, and R.P.W. Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*, 6:22–31, 2003.

[15] L. Lam. Classifier combinations: implementations and theoretical issues. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 78–86, Cagliari, Italy, 2000. Springer.

[16] L. Lam and C.Y. Suen. Optimal combination of pattern classifiers. *Pattern Recognition Letters*, 16:945–954, 1995.

[17] L. Lam and C.Y. Suen. Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics*, 27(5):553–568, 1997.

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

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

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

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

[22] P.H.A. Sneath and R.R. Sokal. *Numerical Taxonomy*. W.H. Freeman & Co, 1973.

[23] G.U. Yule. On the association of attributes in statistics. *Phil. Trans., A*, 194:257–319, 1900.