Accuracy bounds for ensembles under 0 – 1 loss

Remco R. Bouckaert
Xtal Mountain Information Technology &
Computer Science Department, University of Waikato,
New Zealand
rrb@xm.co.nz, remco@cs.waikato.ac.nz

June 24, 2002

Abstract

This paper is an attempt to increase the understanding in the behavior of ensembles for discrete variables in a quantitative way. A set of tight upper and lower bounds for the accuracy of an ensemble is presented for wide classes of ensemble algorithms, including bagging and boosting. The ensemble accuracy is expressed in terms of the accuracies of the members of the ensemble.

Since those bounds represent best and worst case behavior only, we study typical behavior as well, and discuss its properties. A parameterized bound is presented which describes ensemble behavior as a mixture of depentent base classifier and independent base classifier areas. Some empirical results are presented to support our conclusions.

1 Introduction

Ensemble algorithms like bagging [4], boosting [8], arcing [5] and their variations are widely researched and used. There is a good body of emperical work on ensembles [1, 6].

Ensemble behavior for real valued class variables, like for example in regression problems, are reasonably well understood [4, 12]. However, when the class variable has discrete values, analysis is not very outspoken and no strong quantitative results are available.

Various approaches based on variance-bias decomposition [7, 10] and diversity-loss [3, 9] describe the behavior of ensembles, but interpretation for discrete classes is cumbersome.

In this article, we concentrate on accuracy of ensembles with discrete classes by deriving tight bounds on the ensemble accuracy in terms of the mean accuracy of the ensemble members. Starting simple with a binary class and so-called

uniform democratic voting as in standard bagging, we generalize the bounds for increasingly wider classes of ensembles including non-uniform voting, multi-valued classes and probabilistic voting. Unfortunately, it turns out that the bounds are rather wide, which explains the difficulty in finding good quantitative analysis of ensemble behavior.

In the following section, terminology is introduced. In the following section we derive upper and lower bounds on ensemble accuracy for various situations and voting schemes. Upper and lower bounds only indicate best and worst case behavior, so we proceed considering typical behavior. We conclude with a summary, some final remarks and directions for further research.

2 Terms and definitions

We consider a set of variables $\mathbf{x} = \{x_1, \ldots, x_n\}$, $n \geq 0$ with domain $\mathbf{X} = \{X_1, \ldots, X_n\}$ called *attributes* and a single variable y with domain Y called *class variable* or just *class*. We will consider finite discrete classes only in this paper. A *classifier* C is a function $\mathbf{X} \rightarrow [0,1]^{|Y|}$ that maps an instantiation of $\mathbf{x}$ $in\mathbf{X}$ to a probability distribution $P_C(y|\mathbf{x})$ of y. The value of y with the highest probability is the *prediction* of the classifier on $\mathbf{x}$, denoted by $C(\mathbf{x}) = argmax_{y' \in Y} P_C(y = y'|\mathbf{x})$. Note that classifiers that just output a single value y^* of y (such as simple nearest neighbour and support vector machines) still fit the model: their output can be interpreted as a degenerate distribution with $P(y = y^*) = 1$ and $\forall_{y' \neq y^*} P(y = y') = 0$.

An *ensemble* is a classifier that combines the results of a set of classifiers (called *base classifiers*) $C_1, \ldots, C_k$, $k > 1$. There are various ways an ensemble can combine the results of base classifiers: each classifier has a vote with *weight* $w_1, \ldots, w_k$ $\sum_{i=1}^{k} w_i = 1$.

Democratic voting combines votes by letting each base classifier assign a single vote to its most likely class, and selecting the class value with the most votes. The prediction under democratic voting is for instance $\mathbf{x}$ is

$$argmax_{y \in Y} \sum_{i=1}^{k} w_i I(y = C_i(\mathbf{x}))$$

where $I(.)$ is the indicator function ($I(true) = 1$, $i(false) = 0$) and $P_i(.|.)$ the probability distribution of classifier C_i.

Probabilistic voting combines votes by letting each base classifier assign a single vote to its most likely class, and selecting the class value with the most votes. The prediction under democratic voting is for instance $\mathbf{x}$ is

$$argmax_{y \in Y} \sum_{i=1}^{k} w_i P_i(y|\mathbf{x})$$

Figure 1: Upper bound construction for $k = 4$ and 5. 'c' is correctly classified, '.' incorrect.

x	C_1	C_2	C_3	C_4	C
1	c	.	.	c	c
2	c	c	.	.	c
3	.	c	c	.	c
4	.	.	c	c	c
5	.	.	.	.	.

x	C_1	C_2	C_3	C_4	C_5	C
1	c	.	.	c	c	c
2	c	c	.	.	c	c
3	c	c	c	.	.	c
4	.	c	c	c	.	c
5	.	.	c	c	c	c
6	.	.	.	.	.	.

Other more exotic voting methods exists[3] but these are outside scope of this paper. *Uniform voting* is voting where all weights are equal, $\forall_{i=1}^{k} w_i = \frac{1}{k}$, otherwise it is *non-uniform voting*. Combined with the voting schemes, we get four combinations: uniform and non-uniform democratic voting and uniform and non-uniform probabilistic voting.

The instances of $(\mathbf{x}, y) \in \mathbf{X} \times Y$ occurs with a probability distribution $P(\mathbf{x}, y))$. A *loss function* l is a function $Y \times Y \rightarrow \Re$ that compares the prediction of a classifier with the true value of y and maps it onto a real value. In this paper, we consider the 0-1 loss function $l(y, y') = I(y \neq y')$, which is zero when the prediction matches the true value, and one otherwise. The *loss* L of a classifier C is the expected loss, that is,

$$L = \int_{\mathbf{X} \times Y} l(y, C(\mathbf{x}))P(\mathbf{x}, y)d\mathbf{x}, y.$$

The *accuracy* A of a classifier C is the one minus the 0-1 loss of the classifier, i.e $A = 1 - L$, or

$$A = \int_{\mathbf{X} \times Y} I(y, C(\mathbf{x}))P(\mathbf{x}, y)d\mathbf{x}, y.$$

3 Bounds under 0-1 loss

In this section, we provide bounds on the accuracy of an ensemble in terms of the accuracies of the individual base classifiers in the ensemble. We start with the simple case of uniform democratic voting with a binary class and subsequently generalize in the direction of

- binary to multinomial class,
- uniform to weighted voting, and
- democratic to probabilistic voting.

Finally, a general bound for weighted probabilistic voting with a multinomial class is given.

3.1 Uniform democratic voting with binary class

In this section, on domain X with binary class Y let $C_1, \ldots, C_k$, $k > 1$ be a set of classifiers with accuracies $A_1, \ldots, A_k$. Let $\bar{A}$ be the mean accuracy $\bar{A} = \frac{1}{k}\sum_{i=1}^{k} A_i$, and A the accuracy of the ensemble using uniform democratic voting.

LEMMA 3.1

$$A \leq 2\bar{A}$$

So, the accuracy of an ensemble never exceeds the mean accuracy by a factor of 2.

Proof: Let T be the maximum (by probability mass) set of instances $(\mathbf{x}, \mathbf{y})$ for which $C(\mathbf{x}) = y$. Because uniform democratic voting is used, for every $(\mathbf{x}, y) \in T$ the number of votes for y must be larger or equal $k/2$ (in case it is equal, the voting scheme is choosing a class value randomly, and since we are constructing an upper bound, we assume it happens to always make the correct choice).

So, every $(\mathbf{x}, y) \in T$ consumes at most $k/2.P(\mathbf{x}, y)$ votes. There is a total of $\sum_{i=1}^{k} A_i$ votes available. Therefore, the fraction of $(\mathbf{x}, \mathbf{y})$ in T is at most the number of available votes divided by the number of required votes $P(T) = (\sum_{i=1}^{k} A_i)/(k/2) = 2\frac{1}{k}\sum_{i=1}^{k} A_i = 2\bar{A}$. No better allocation of votes is possible, since for every move of a vote for y for instance $(\mathbf{x}, y)$, would make that instance misclassify and no other instance is helped by the moved vote. So this is the largest size T can take.

Note that A is the probability an instance in T occurring, $A = \int_{\mathbf{X}\times Y} I(y, C(\mathbf{x}))P(\mathbf{x}, y)d\mathbf{x}, y \leq \int_T P(\mathbf{x}, y)d\mathbf{x}, y = P(T)$, so $A \leq 2\bar{A}$. □

LEMMA 3.2

$$A > 2\bar{A} - 1$$

In words, the accuracy of an ensemble never decreases below twice the mean accuracy minus 1. So, if the mean accuracy drops below a half, the ensemble accuracy may drop to zero.

Proof: We try to find a lower bound by allocating the votes of the base classifiers in a way that lowers the accuracy as much as possible. There is a fraction of $k\bar{A}$ votes available. Let A_{min} be the minimum obtainable accuracy and T as in previous proof, then to spill as many votes as possible, each $(\mathbf{x}, y) \in T$ should get k votes, consuming kA_{min} votes. The fraction in F (i.e. all instances outside X) should consume $k/2$ votes, for a total of $k/2P(F) = k/2(1 - A_{min})$ votes.

Summing kA_{min} and $k/2(1-A_{min})$ gives the total set of votes $k\bar{A}$. And after some manipulation, we get,

$$\begin{array}{rrcl} & k/2(1-A_{min})+kA_{min} & = & k\bar{A} \\ \Leftrightarrow & 1/2(1-A_{min})+A_{min} & = & \bar{A} \\ \Leftrightarrow & 1/2(1)+1/2A_{min} & = & \bar{A} \\ \Leftrightarrow & A_{min} & = & 2\bar{A}-1 \end{array}$$

So A never decreases below $A_{min} = 2\bar{A} - 1$. □

The question now is whether the bounds in Lemma 3.1 and 3.2 are tight or loose. Figure 1 shows an example for $k = 4$ and $k = 5$ where the first column indicates a number of the instance $(\mathbf{x}, y)$, and the last column contains the classification of the ensemble. The columns in the middle show whether a classifier C_i has a correct (c) prediction or an incorrect (.) for the particular instance. So, for $k = 4$ and $k = 5$ there are cases that are close to the upper bound indeed. The following lemma shows this is true for general $k > 1$.

LEMMA 3.3 *There exist sets of classifiers such that* $A = min(2\bar{A} + \epsilon, 1)$ *where* $\epsilon = \bar{A}/2k$

Proof: Let k be even. Let $C_1, \ldots, C_k$ have equal accuracies, $\forall_{i=1}^{k} A_i = \bar{A}$. First, assume $\bar{A}$ is less than a half. Let $(\mathbf{x}, y)$ be such that it can be split into sets of instances $i_1, \ldots, i_k, i_{k+1}$ such that for $1 \leq i \leq k$, $P(i_i) = 2\bar{A}/k$ and $P(i_{k+1}) = 1 - 2\bar{A}$

Now, let C_i vote correct for $i_i, \ldots, i_{i+k/2}$, [1] then a fraction of $\sum_{j=i}^{i+k/2} P(i_j) = k/2.2\bar{A}/k = \bar{A} = A_i$ portion of the vote is consumed for classifier C_i. Also, for each $1 \leq i \leq k$, $k/2$ of the base classifiers vote correct, hence the vote is correct for $\sum_{i=1}^{k} P(i_i) = \sum_{i=1}^{k} 2\bar{A}/k = 2\bar{A}$

If $\bar{A}$ is more than $\frac{1}{2}$, the same recipe as before with $\bar{A} = 1/2$ can be applied and we get $A = 1$. The procedure leaves a few extra votes to be assigned, which can be done at random.

If k is odd, a fraction $k + 1/2$ of the votes is required instead of $k/2$ which requires $\bar{A}/2k$ of the vote to be assigned to exceed the threshold. □

A similar construction exists for the lower bound, as illustrated in Figure 2 and is generalized in the following lemma.

LEMMA 3.4 *There exist a set of classifiers such that* $A = max(2\bar{A} - 1 + \epsilon, 0)$ *where* $\epsilon = \bar{A}/2k$.

Proof: Let k be even. Let $C_1, \ldots, C_k$ have equal accuracies, $\forall_{i=1}^{k} A_i = \bar{A}$.

If $\bar{A} > \frac{1}{2}$, let $(\mathbf{x}, y)$ be such that it can be split into sets of instances $i_1, \ldots, i_k, i_{k+1}$ such that $P(i_{k+1}) = 2\bar{A}-1$ and for $1 \leq i \leq k$, $P(i_i) = (2-2\bar{A})/k$ constant. Let instances in i_{k+1} be classified correct by each base classifier,

[1] In the following paragraph, all indices are to be interpreted modulo k.

Figure 2: Lower bound construction for $k = 4$ and 5

x	C_1	C_2	C_3	C_4	C
1	c	.	.	c	.
2	c	c	.	.	.
3	.	c	c	.	.
4	.	.	c	c	.
5	c	c	c	c	c

x	C_1	C_2	C_3	C_4	C_5	C
1	c	.	.	.	c	.
2	c	c	.	.	.	.
3	.	c	c	.	.	.
4	.	.	c	c	.	.
5	.	.	.	c	c	.
6	c	c	c	c	c	c

consuming $k(2\bar{A} - 1)$ of the votes and contributing $2\bar{A} - 1$ to the accuracy. Now, let C_i vote correct for $i_i, \ldots, i_{i+k/2}$, [1] then a fraction of $\sum_{j=i}^{i+k/2} P(i_j) = k/2.2\bar{A}/k = \bar{A} = A_i$ portion of the vote is consumed for classifier C_i. Also, for each $1 \leq i \leq k$, $k/2$ of the base classifiers vote *in*correct, hence the vote is *in*correct. Now, every classifier C_i has assigned $k/2.(2-2\bar{A})/k$ for the instances $i_1, \ldots, i_k$ and $(2\bar{A}-1)$ for instances i_{k+1}, which sums to $\bar{A} = C_i$ of its votes. No contribution to the accuracy is made by the instance $i_1, \ldots, i_k$ since the vote never exceeds $k/2$. So, the total accuracy is contributed by i_{k+1} alone, and this is only $2\bar{A} - 1$.

If $\bar{A} \leq \frac{1}{2}$, the same procedure applies, but now the set of instances i_{k+1} can be made empty, and some instances cannot get a correct vote, resulting in an accuracy of zero.

If k is odd, a fraction $k - 1/2$ of the votes is required instead of $k/2$ to make an instance misclassify, so this leaves $\bar{A}/2k$ of the vote to be assigned to instances in T. □

THEOREM 1 *Subject to* $0 \leq A \leq 1$,

$$2\bar{A} - 1 < A \leq 2\bar{A} \tag{1}$$

and there exist sets of classifiers that realize these bounds within $O(1/k)$.

Proof: Follows directly from Lemma 3.1 to 3.4 and the observation that the accuracy by definition is between zero and one. □

3.2 Uniform democratic voting with multinomial class

The results from the previous section can be generalized to multinomial classes with $|Y|$ values realizing that to get a correct ensemble prediction, in the best case only $|Y|/k$ of the base classifiers have to be correct. However, in the worst case $k/2$ of the base classifiers may be correct while the ensemble prediction is incorrect. This happens namely if the incorrect base classifiers all vote for the same incorrect class.

In this section, on domain X with multinomial class Y let $C_1, \ldots, C_k$, $k > 1$ be a set of classifiers with accuracies $A_1, \ldots, A_k$. Let $\bar{A}$ be the mean accuracy $\bar{A} = \frac{1}{k}\sum_{i=1}^{k} A_i$, and A the accuracy of the ensemble using uniform democratic voting.

THEOREM 2 *Subject to* $0 \leq A \leq 1$,

$$2\bar{A} - 1 < A \leq |Y|\bar{A} \tag{2}$$

and there exist sets of classifiers that realize these bounds within $O(1/k)$.

Proof: The lower bound follows from the same construction as in the proof of Lemma 3.2 where it is assumed that for an instance $\mathbf{x}$ all incorrect classifiers vote for the same class. Existence of a set of classifiers that realizes this bound is as in the proof of Lemma 3.4 under the same restriction.

The upper bound follows from realizing that only $k/|Y|$ of the base classifiers need to be correct, as long as none of the incorrect values of $|Y|$ gets more than $k/|Y|$ of the votes (and in a draw of the votes the ensemble always happens to select the correct value). The proofs of Lemma 3.1 and 3.3 apply with these modifications, giving the upper bound and existence of a set of classifiers realizes this bound. □

Note, Theorem 1 is a special case of Theorem 2 for $|Y| = 2$.

3.3 Non-uniform democratic voting with multinomial class

For non-uniform voting, the mean accuracy $\bar{A}$ needs to be generalized as follows: $\bar{A} = \sum_{i=1}^{k} w_i A_i$, which reduced to $\frac{1}{k}\sum_{i=1}^{k} A_i$ for uniform voting.

THEOREM 3 *Subject to* $0 \leq A \leq 1$,

$$2\bar{A} - 1 < A < |Y|\bar{A} \tag{3}$$

and there exist sets of classifiers that realize these bounds arbitrarily close.

Proof: Existence of the bounds follow from the observation that still an amount of $k\bar{A}$ votes need to be optimally distributed. The existence proofs follow the same pattern.

Realization of the bounds follows from the fact the bounds are realized for uniform voting, which is a special case of non-uniform voting for k even. For k odd, let $w_1 = 0$, and the classifiers $2, \ldots, k$ form a set with an even number of classifiers, as before. □

Note, Theorem 3 generalizes Theorem 1 and 2.

The theorem can be reformulated in terms of loss instead of accuracy as follows. Let L_i be the 0-1 loss for base class C_i for $1 \leq i \leq k$ and let the mean member loss be $\bar{L} = \sum_{i=1}^{k} w_i L_i$. Then we have the following property.

CONSEQUENCE 3.1 *Subject to* $0 \leq L \leq 1$,

$$|Y|\bar{L} + 1 - |Y| \leq L < 2\bar{L}$$

and there exist sets of classifiers that realize these bounds arbitrarily close.

Proof: Follows from definitions and some manipulation $\bar{L} = \sum_{i=1}^{k} w_i L_i = \sum_{i=1}^{k} w_i(1 - A_i) = \sum_{i=1}^{k} w_i - \sum_{i=1}^{k} w_i A_i = 1 - \bar{A}$. Now, substituting $L = 1 - A$ and $\bar{L} = 1 - \bar{A}$ in equation (3) gives

$$\begin{array}{rrcccl} & 2(1-\bar{L}) - 1 & < & (1-L) & \leq & |Y|(1-\bar{L}) \\ \Leftrightarrow & 1 - 2\bar{L} & < & 1 - L & \leq & |Y| - |Y|\bar{L} \\ \Leftrightarrow & -2\bar{L} & < & -L & \leq & |Y| - 1 - |Y|\bar{L} \\ \Leftrightarrow & 2\bar{L} & > & L & \geq & |Y|\bar{L} + 1 - |Y| \end{array}$$

which by switching terms and observing $0 \leq L \leq 1$ gives the desired result. □

3.4 Non-uniform probabilistic voting with multinomial class

For probabilistic voting, the member accuracy A_i needs to be generalized as follows: instead of considering the prediction of a classifier, consider the probability mass assigned to the correct class. More formally,

$$A_i^p = \int_{\mathbf{X} \times Y} P_i(y|\mathbf{x}) P(\mathbf{x}, y) d\mathbf{x}y$$

Now, let $\bar{A} = \sum_{i=1}^{k} w_i A_i^p$, then we have the final result of this section.

THEOREM 4 *Subject to* $0 \leq A \leq 1$,

$$2\bar{A} - 1 < A < |Y|\bar{A} \tag{4}$$

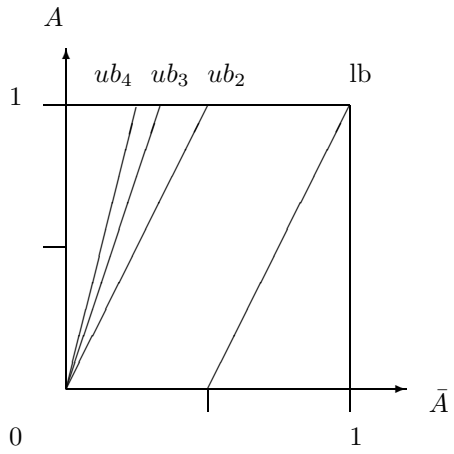
and there exist sets of classifiers that realize these bounds

Proof: Existence of the bounds follow from the observation that still an amount of $k\bar{A}$ votes need to be optimally distributed. The existence proofs follow the same pattern.

Realization of the bounds again follows from the fact the bounds are realized for democratic voting, which is a degenerate case of probabilistic voting. However, since there is a more refined control over distributing votes, because fractions of votes can be assigned, it is possible to select classifiers for the upper bound such that for correctly classified instance $\mathbf{x}$ $k/|Y| + \epsilon$ of the votes are correct, where ϵ can be made arbitrarily small. Same argument applies for the lower bound. □

Note, Theorem 4 differs from Theorems 1 to 3 in that the bounds now can be reached arbitrarily close, and in the definition of the mean member accuracy $\bar{A}$.

Figure 3: Graph of bounds (lb = lower bound, ub_i = upper bound)



Let L_i^p be $1 - A_i^p$, that is the probabilistic 0-1 loss for base class C_i for $1 \leq i \leq k$ and let the mean member loss be $\bar{L} = \sum_{i=1}^{k} w_i L_i$. Then we have the following property.

CONSEQUENCE 3.2 *Subject to* $0 \leq L \leq 1$,

$$|Y|\bar{L} + 1 - |Y| \leq L < 2\bar{L}$$

and there exist sets of classifiers that realize these bounds arbitrarily close.

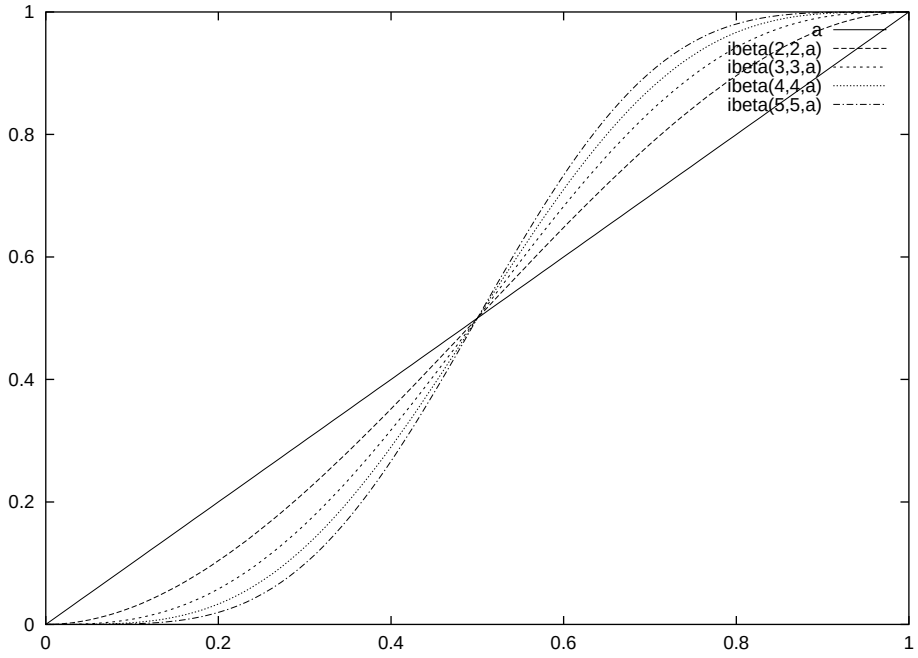
The proof follows closely that of Consequence 3.1 and is left to the reader.

3.5 Discussion

Figure 3 shows a graph of the bounds of A expressed in $\bar{A}$ for binary, ternary and four-valued classes, labeled as ub_2, ub_3 and ub_4 respectively. It shows that the bounds are fairly wide: for $\bar{A} = \frac{1}{2}$ in fact A can be anything between 0 and 1. Unfortunately, sets of base classifiers exists such that with $\bar{A} = \frac{1}{2}$ $A = 0$ and there exists sets of classifiers such that $A = 1$. Consequently, it will remain hard to perform quantitative analysis on ensembles.

One point highlighted by Figure 3 is that ensembles do not necessarily outperform a single best classifier. Indeed, an ensemble may not even outperform the best base classifier in the ensemble.

Figure 4: Graph of accuracy A against $\bar{A}$ for various values of k



4 Expected accuracy

The bounds in the previous section describe worst and best case behavior. In this section, we study typical behavior by looking at the expected accuracy.

4.1 Uniform democratic voting with binary class

Let $C_1, \ldots, C_k$ be a set of classifiers with the same accuracies $\forall_{i=1}^{k} A_i = \bar{A}$. Furthermore, assume all classifiers are independent. Though this is not such a realistic assumption it helps in illustrating the behavior of the ensemble. Now under uniform democratic voting with binary class, for an instance $\mathbf{x}$ we have the probability that $\mathbf{x}$ is classified correctly by the ensemble

$$A = P(C(\mathbf{x}) = y) = P(\sum_{i=1}^{k} I(C_i(\mathbf{x}) = y) \geq \lceil k/2 \rceil).$$

Since we assumed that the classifiers are independent, the probability that exactly j classifiers vote correct is $\bar{A}^j(1-\bar{A})^{k-j}$ which can be done in $\binom{j}{k}$ configurations. So, $P(\sum_{i=1}^{k} I(C_i(\mathbf{x}) = y) = j) = \bar{A}^j(1-\bar{A})^{k-j}\binom{j}{k}$ and hence the accuracy can be written as a sum over binomials with parameter $\bar{A}$,

$$A = \sum_{j=\lceil k/2 \rceil}^{k} \bar{A}^j(1-\bar{A})^{k-j}\binom{j}{k}. \tag{5}$$

For small k, the binomial distribution in (5) can be approximated[2] by (incomplete) Beta functions B, giving

$$A \approx \frac{B(\bar{A}; \lceil k/2 \rceil, k - \lceil k/2 \rceil)}{B(\lceil k/2 \rceil, k - \lceil k/2 \rceil)} = \frac{B(\bar{A}; \lceil k/2 \rceil, \lfloor k/2 \rfloor)}{B(\lceil k/2 \rceil, \lfloor k/2 \rfloor)}. \tag{6}$$

Figure 4 plots equation (6) as a function of $\bar{A}$ for various values of k. When A is over $\bar{A}$ it is better to use the ensemble, otherwise it is better to use the best base classifier in the ensemble. The crossing point, as shown in the plot, is when $\bar{A}$ is $\frac{1}{2}$. It shows that increasing the number of base classifiers increases the influence of the ensemble, but the effect becomes smaller with higher k.

For large enough k, (5) can be approximated[2] by a normal with mean $\bar{A}k$ and variance $\sigma^2 = (1-\bar{A})\bar{A}k$, giving

$$A \approx \int_{j \geq \lceil k/2 \rceil} N(\bar{A}k, (1-\bar{A})\bar{A}k). \tag{7}$$

[2]See for example http://www.mathworld.com/ under topic Binomial distribution. [Accessed 24 April 2002]

Figure 5: Graph of accuracy A against $\bar{A}$ with a normal approximation

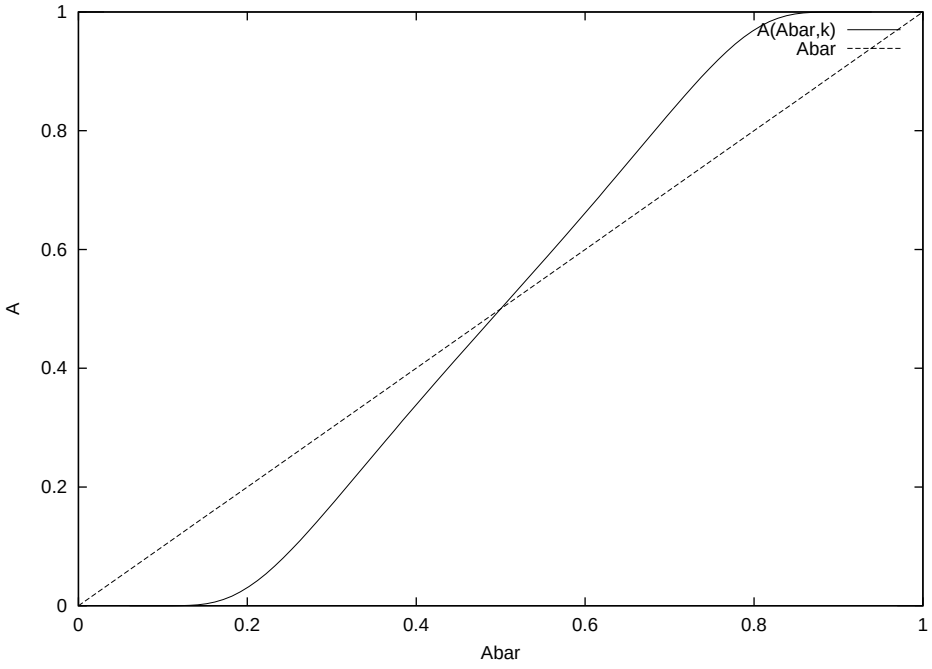


Figure **??** plots equation (**??**) as a function of $\bar{A}$, together with $\bar{A}$. Still, when A exceeds $\bar{A}$ it is better to use the ensemble, otherwise it is better to use the best base classifier in the ensemble.

Note that the size of the ensemble k (here chosen 10) does not have an impact on the shape of the plot. The reason is that (**??**) can be rewritten through scaling by k giving

$$A = \int_{j \geq \lceil k/2 \rceil / k} N(\bar{A}, (1-\bar{A})\bar{A}) \approx \int_{j \geq \frac{1}{2}} N(\bar{A}, (1-\bar{A})\bar{A})$$

where the approximation $\lceil k/2 \rceil / k \approx \frac{1}{2}$ holds for large k. The last formula has no dependency on k, so the behavior of an ensemble is not expected to change much. This does not mean that the ensemble can be made arbitrarily small because the approximation (**??**) holds for large enough k only.

4.2 Mixtures

In practice, ensemble classifiers with $\bar{A} < \frac{1}{2}$ can outperform single base classifiers, which indicates that the assumption we made that all base classifiers are independent does not in general hold. Apparently, the mutual dependence of base classifiers forms the strength of the ensemble.

This indicates the typical behavior of an ensemble can be interpreted as a mixture of classifiers behaving the same for the straightforward cases, and behaving more or less independent for the difficult cases. In other words, we can expect A to be a mixture of $\bar{A}$ and (6).

$$A = (1-x)\bar{A} + x(5)$$

In the binomial (5) the dependence between classifiers cannot be discarded, and one way to cater for this is adding a small constant c_1 to the mean accuracy $\bar{A}$, giving an approximation of (6) by

$$\frac{B(\bar{A} + c_1; \lceil k/2 \rceil, \lfloor k/2 \rfloor)}{B(\lceil k/2 \rceil, \lfloor k/2 \rfloor)}. \tag{8}$$

Intuitively, the mixture depends on $\bar{A}$: lower $\bar{A}$ puts more emphasis on the dependence between classifiers, while higher $\bar{A}$ puts more emphasis on the independe between classifiers. Together, we arrive at the following approximation:

$$A \approx (1 - c_2\bar{A})\bar{A} + c_2\bar{A}\frac{B(\bar{A} + c_1; \lceil k/2 \rceil, \lfloor k/2 \rfloor)}{B(\lceil k/2 \rceil, \lfloor k/2 \rfloor)}. \tag{9}$$

The constants c_1 and c_2 can be varied to obtain mean, and typical upper and lower bounds. In the experiments section, we see that (8) fits mean behavior properly using $c_1 = \frac{1}{4}$ and $c_2 = \frac{1}{2}$. Further, $c_1 = 0$, $c_2 = 0$ gives a lower bound and $c_1 = 0.5$, $c_2 = 1$ gives an upper bound.

4.3 Generalizations

We would like to generalize the results from the previous to non uniform voting, multinomial classes and probabilistic voting. However, a quantitative description is going to be increasingly complex without providing much insight, leaving us to just a few qualitative remarks.

For multinomial classes with $|Y|$ class values, the threshold under which an ensemble makes a correct prediction depends on the distribution of the incorrect votes: the number of correct votes required may be as low as $k/|Y|$ or as high as $k/2$. This suggests that to get a good ensemble result, selecting a set of base classifiers that is more diverse in its predictions can minimize the required threshold, hence requiring base classifiers that are less accurate overall.

Analysing probabilistic voting could help by realizing that for a classifier C the class probabilities $P_C(y|\mathbf{x})$ can be normalized to for example percentages (100 values). Instead of treating C as a single base classifier we now can treat C as a collection of 100 base classifiers where $100 \times P_C(y|\mathbf{x})$ vote for class y on input $\mathbf{x}$.

5 Experiments

To get an impression how well approximation (8) works, some experiments were done measuring $\bar{A}$ and A. Ensembles were learned using bagging and base classifier C4.5 as implemented in Weka[3] [13]. A new evaluator method was implemented on top of the existing Weka package and the bagging class was made configurable so that the voting method (democratic or probabilistic) could be selected. Otherwise, for all algorithms default settings as in Weka 3.2 were used and uniform voting ($\forall_{i=1}^{k} w_i = \frac{1}{k}$) was applied. All algorithms were run ten times using ten fold cross validation. Reasonable estimates for the accuracy are obtained by collecting the accuracies for the ten folds and averaging over the ten runs [11]. The datasets are from the UCI repository [2] and are provided with Weka [4].

Figure 5 and 6 show $\bar{A}$ and A for the various datasets with democratic and probabilistic voting respectively. Also the line $A = \bar{A}$ and the approximation (8) are plotted. All datapoints are between those two lines, except for the datapoint for primary tumor.

To get more datapoints, 110 datasets were generated randomly generating a Bayesian network, populating it with randomly selected probability tables. The data was generated by instantiating the variables one by one according to the probability tables for the value of the parents. Datasets of 100 cases were generated and the cardinality of the variables was varied from 2 to 12,

[3] Weka can be obtained from http://www.cs.waikato.ac.nz/ml/

[4] The following datasets were used: autos, balance-scale, breast-cancer, breast-w, horse-colic, credit-rating, german-credit, pima-diabetes, glass, heart-c, heart-h, heart-statlog, hepatitis, iris, labor, lymphography, primary-tumor, segment, vehicle, vote, vowel, and zoo.

Figure 6: Graph of measured accuracy A against $\bar{A}$ for democratic voting with UCI datasets

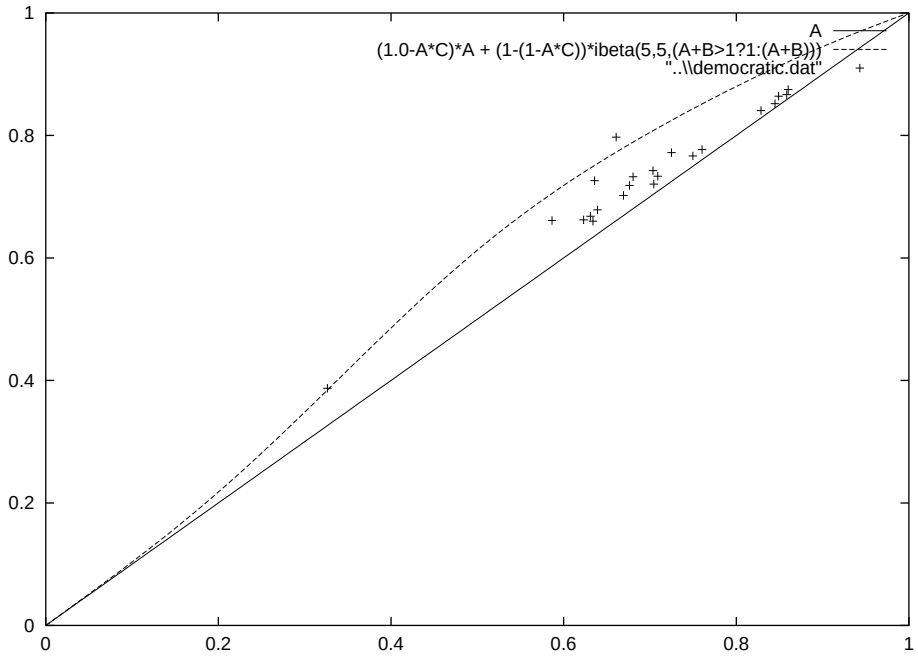


Figure 7: Graph of measured accuracy A against $\bar{A}$ for probabilistic voting with UCI datasets

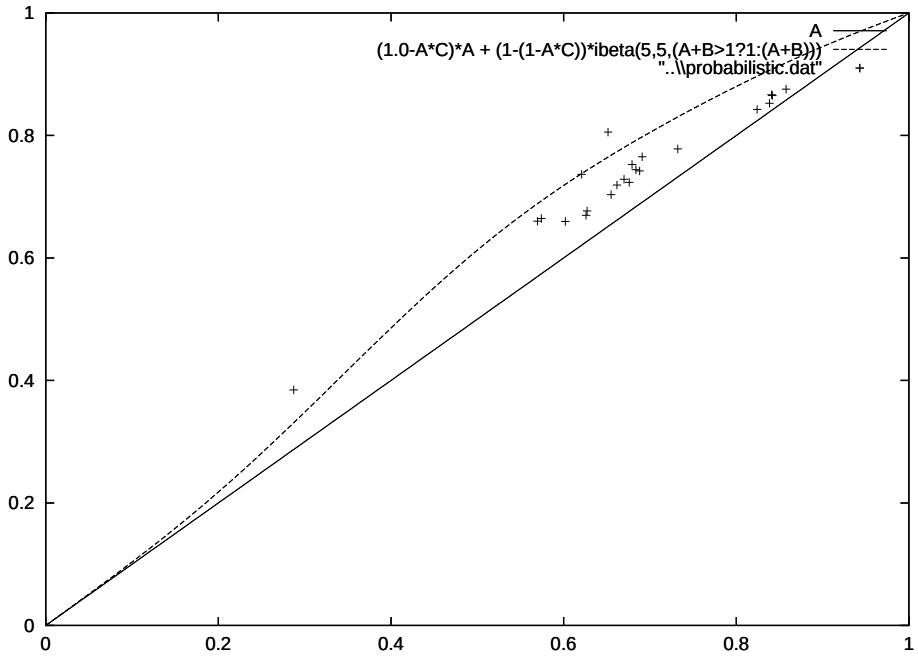
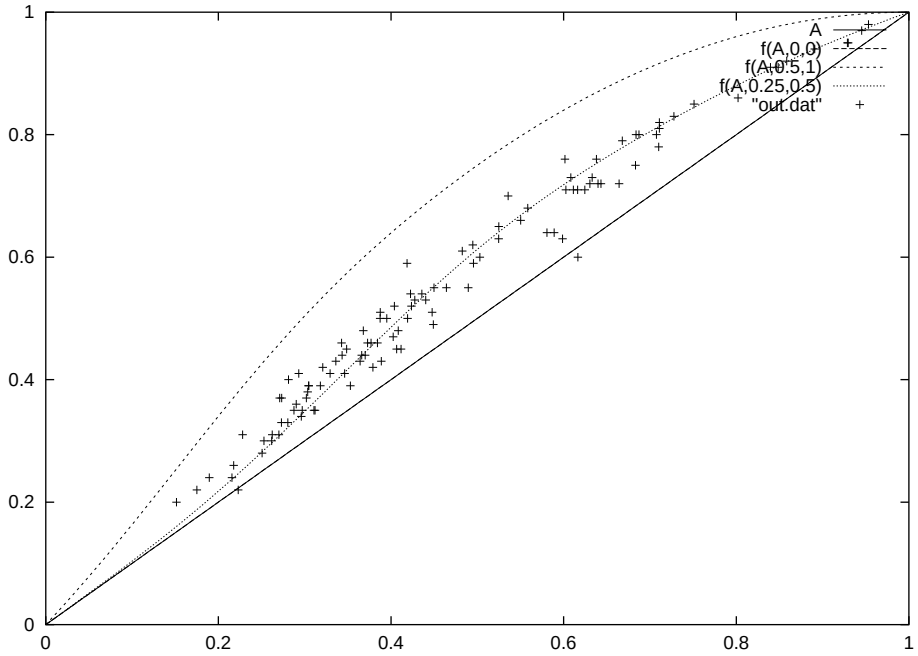


Figure 8: Graph of measured accuracy A against $\bar{A}$ for probabilistic voting with random data



10 datasets for each. Figure 7 shows the same graph for randomly generated datasets with one run ten fold cross-validation. Also, the mean, upper and lower bound based on (8) is pictured, showing that they provide good indications of the upper and lower bounds.

6 Conclusion

For 0-1 loss, accuracy upper and lower bounds are given for a variety of ensemble algorithms, including bagging and boosting. We express the accuracy in terms of the accuracies of the member classifiers and show that these bounds can (almost) be realized indicating the bounds are very tight. However, the bounds are pretty wide, hence it makes sense to concentrate on finding quantitative descriptions of typical behavior only.

Some progress was made on describing typical ensemble behavior for uniform voting with independent base classifiers. A parameterized approximation

explains typical ensemble behavior as a mixture of base classifiers voting completely dependent and completely independent. By selection of appropriate parameters, upper and lower bounds are obtained for most ensemble behavior.

Further research into quantitative descriptions of more general ensemble methods could give better insight in ensemble behavior.

References

[1] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, Boosting, and variants. Machine Learning, 36:105–139, 1999.

[2] C.L. Blake and C.J. Merz. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. 1998.
http://www.ics.uci.edu/∼mlearn/MLRepository.html[5]

[3] R.R. Bouckaert, M. Goebel and P. Riddle. A generalized unified decomposition of ensemble loss. Submitted to JMLR, 2002.

[4] L. Breiman. Bagging predictors. Machine Learning 24(2): 123-140, 1996.

[5] L. Breiman, L. Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California at Berkeley, 1996.

[6] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning, 40(2): 139-158, 2000.

[7] P. Domingos. A Unified Bias-Variance Decomposition and its Applications. Proceedings of the Seventeenth International Conference on Machine Learning, 564–569, 2000.

[8] 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.

[9] M. Goebel, P. Riddle, M. Barley. A unified decomposition of ensemble loss for predicting ensemble performance. Proceedings of the International Conference on Machine Learning, 2002.

[10] R. Kohavi and D. Wolpert. Bias plus variance decomposition for zero-one loss functions. Proceedings of the International Conference on Machine Learning, 275–283, 1996.

[5] All URLs mentioned in this article were accessed and available at 21 April 2002.

[11] R. Kohavi. A study of cross validation and bootstrap for accuracy estimation and model selection. IJCAI, 1137–1145, 1995.

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

[13] I.H. Witten and E. Frank. Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, 2000.