A New Model for Measuring the Accuracies of Majority Voting Ensembles

Xueyi Wang
Department of Mathematics and Computer Science
Northwest Nazarene University
Nampa, ID, USA
xwang@nnu.edu

Abstract— **Good ensemble methods require accurate and diverse individual classifiers, but the relationship between the diversity of individual classifiers and the accuracy of an ensemble method is not clear. In this paper, we propose a novel model called COB (core, outlier, and boundary) to quantitatively measure the accuracies of majority voting ensembles for binary classification. In this model, we first divide data items into three subsets, core, outlier, and boundary, based on the prediction correctness of these items from individual classifiers in an ensemble method. Then we measure the accuracy of the ensemble method for each subset and combine the results together. We tested the performance of the COB model on 32 datasets from the UCI repository. The experiments use three different ensemble methods (bagging, random forests, and a randomized ensemble), two different numbers of individual classifiers (7 and 51), and three different individual machine learning algorithms (decision trees, *k*-nearest neighbors, and support vector machines). All 24 experiments showed less than 5% average absolute errors for 32 datasets between the accuracies by the COB model and the actual accuracies of ensembles. Also the experiments showed that the COB model performed significantly better than the binomial model. The COB model suggests that to achieve a high accuracy for an ensemble method, weak individual classifiers should be partly diverse instead of fully diverse, that is, be diverse on correctly predicted items but in agreement on some incorrectly predicted items.**

Keywords-ensemble methods; majority voting; measurement, accuracy

I. INTRODUCTION

Ensemble methods make predictions by combining the predictions from a set of individual classifiers. To achieve high prediction accuracy, traditionally it is believed that ensemble methods should have accurate and diverse individual classifiers. "Accurate classifiers" means the prediction accuracy of each classifier should be better than random, that is, larger than 0.5 for a binary classifier. "Diverse classifiers" means each classifier should make prediction independently, so that a combination of these predictions will result in high prediction accuracy for ensemble methods.

Research has showed that ensemble methods achieved better performance than individual classifiers in many practical problems. For example, the three most popular ensemble algorithms, bagging [2], random forests [3], and boosting [11], are very effective in practical uses [9] [15] [16]. Bagging and random forests ensembles use a majority voting rule, where all individual classifiers have the same weight and the prediction of an ensemble goes with the majority. Boosting ensemble uses a weighted majority voting approach, where each individual classifier is assigned to a different weight and the prediction of an ensemble goes with the weighted majority. We consider only the majority voting rule in this paper.

While ensemble methods have shown much success on many practical problems, a thorough theoretical study of ensemble methods is still lacking. One key problem is how to find a set of individual classifiers that can maximize the prediction accuracy of an ensemble method on a set of data. In this paper, we focus on the problem of measuring the accuracies of ensemble methods based on the prediction accuracies of individual classifiers. If we can find a good measure for ensemble methods, then it will help us understand the performance of ensemble methods and build better ensemble methods in the future.

Traditionally, an ensemble method is assumed to follow a binomial distribution, in which case individual classifiers are considered independent of each other [8] [12]. When individual classifiers are accurate and have the same accuracy, theoretically the binomial model shows that the accuracy of the ensemble method is always better than those of the individual classifiers, as long as the individual classifiers are independent.

When applying the binomial model to actual ensembles, one issue is that this model can give only a qualitative reason of why ensemble methods are better but not a quantitative measure of the actual accuracies of ensemble methods. In previous studies, although most of studies showed that ensemble methods are better than individual classifiers, the actual accuracies of ensemble methods are usually worse than the accuracies predicted from the model. Furthermore, research has even showed that in some cases ensemble methods achieved worse performance than individual classifiers [6] [14].

To obtain a quantitative measure of the performance of ensemble methods, various measures have been proposed for measuring the diversity of individual classifiers. For example,

Supported by NIH Grant P20 RR0116454

Banfield et al. [1] proposed a Percentage Correct Diversity Measure, and Kuncheva and Whitaker [13] and Caruana et al. [5] proposed a set of ten measures each. But in the meantime, the usefulness of applying these measures for the diversity of individual classifiers has also been questioned [13]. A few recent studies have provided more insights into the diversity of individual classifiers. For example, Brown and Kuncheva [4] proposed "good" diversity and "bad" diversity in trying classify the prediction error of ensemble methods. Wang and Davidson [17] showed the upper and lower bounds of an ensemble method when given the prediction accuracies of individual classifiers in a binary classification.

In this paper, we propose a new model called COB (Core, Outlier, and Boundary) to quantitatively measure the accuracies of majority voting ensembles for binary classification. In this model, we first divide a set of data items into three subsets, core, outlier, and boundary, based on the assumption that individual classifiers may make the same correct predictions on some items (the core subset), make the same incorrect predictions on some other items (the outlier subset), and make independent predictions on remaining items (the boundary subset). An example is shown in Figure 2. After obtaining these three subsets, we model and calculate the accuracy of each subset separately and then combine the accuracies together. As real datasets may not show a clear classification of these three subsets, we can set threshold values for classifying core and outlier subsets. For example, items correctly predicted by 85% of individual classifiers are classified into the core subset, items incorrectly predicted by 85% of individual classifiers are classified into the outlier subset, and the remaining items are classified into the boundary subset.

The COB model shows that with the presence of a nonempty core subset, the accuracy of an ensemble method is worse than the accuracy from a binomial model, but is still better than the average of the accuracies of individual classifiers. On the other hand, the presence of a nonempty outlier subset gives mixed results. For weak classifiers, the presence of a small nonempty outlier subset along with an empty core subset may make the accuracy better than the one predicted from the binomial model. For strong classifiers, the presence of a nonempty outlier subset always worsens the performance. Therefore, to achieve high accuracy for ensembles with weak classifiers (for example, $p < 0.68$), we should make the correct predictions diverse (i.e. let the boundary subset be big and let the core subset be small) but concentrate part of incorrect predictions (i.e. let the outlier subset be slightly big). For ensembles with strong classifiers, we should make both correct and incorrect predictions diverse (i.e. let the boundary subset be big and let both the core and outlier subsets be small).

We conducted 24 experiments on this new model using 32 datasets from the UCI repository [10]. The datasets have item sizes vary from 10^2 to 10^4 and feature sizes from 4 to 10^2. We tested three ensemble methods (bagging, random forests, and a randomized ensemble) with two different numbers of individual classifiers (7 and 51) and three different machine learning algorithms (decision trees, *k*-nearest neighbors, and support vector machines; each with different parameters). The results show that the COB model performs significantly better than the binomial model in measuring the actual accuracies of ensemble methods. Table II shows that for all 24 experiments, the average absolute errors between the prediction accuracies of the COB model and the actual accuracies of ensemble methods are within 5%, while average absolute errors between the prediction accuracies of the binomial model and the actual accuracies of ensemble methods are over 5%. Furthermore, for the total of 768 individual experiments (32 datasets in 24 experiments), when using the COB model, only 75 individual experiments have absolute errors larger than 5% and 7 individual experiments have absolute errors larger than 10%, but when using the binomial model, 433 individual experiments have absolute errors larger than 5% and 310 individual experiments have absolute errors larger than 10% (supplementary materials).

The rest of this paper is organized as follows. In Section 2, we first discuss the binomial model and then propose the COB model. In Section 3, we conduct the 24 experiments on 32 datasets using bagging, random forests, and a randomized ensemble and discuss the results. In Section 4, we conclude the paper and discuss some future work.

II. THE COB MODEL

We assume there is an ensemble method E with N ($N > 1$) individual binary classifiers $\{C_i, i = 1, 2, \ldots, N\}$. For the convenience of using the simple majority voting rule, we set N as an odd number: $N = 2K + 1$, where K is a natural number. We further assume there is a testing dataset $\mathbf{X}$ with n items $\{(\mathbf{x}_j, y_j), j = 1, 2, \ldots, n\}$. Each input item $\mathbf{x}_j$ is a vector with m features (variables) $\{x_{jk}, k = 1, 2, \ldots, m\}$ and each output y_j is a class label in $\{-1, 1\}$.

For each input item $\mathbf{x}_j$, each individual classifier C_i predicts an output c_{ij}. We set $z_{ij} = \begin{cases} 1 & c_{ij} = y_j \\ 0 & c_{ij} \neq y_j \end{cases}$, so the ensemble method E predicts the item $\mathbf{x}_j$ correctly if and only if $\left(\sum_{i=1}^{n} z_{ij}\right) > K$ by the majority voting rule.

We denote $p_i = \sum_{j=1}^{n} z_{ij} \Big/ n$ as the prediction accuracy of each classifier C_i and $p_i = \underset{j}{\text{count}}\left(\left(\sum_{i=1}^{n} z_{ij}\right) > K\right) \Big/ n$ as the prediction accuracy of the ensemble E.

We first show the accuracies of ensemble methods when individual classifiers follow the binomial model. Then we discuss the deficiency of the binomial model and propose the COB model.

A. Independent individual classifiers

A general assumption in ensemble learning is that individual classifiers are independent of each other since the items are sampled from a dataset uniformly. If the accuracies of all individual classifiers are the same, say, $p_i = p$, $i = 1, 2, \ldots,$

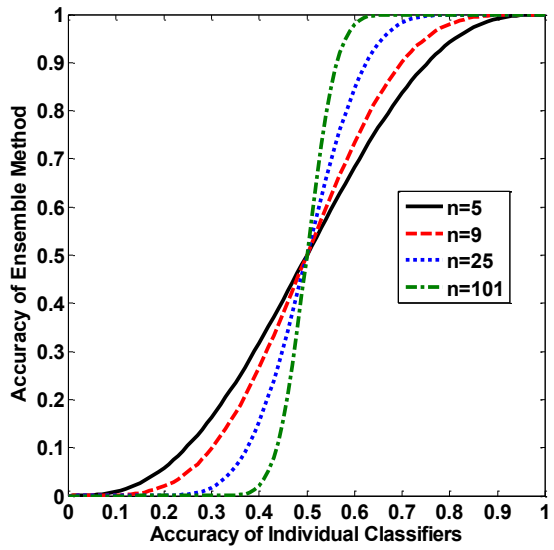


Figure 1. The relationship of the accuracy of individual classifiers p and the accuracy of an ensemble method p_B under the binomial distribution, given the number of classifiers n (n = 5, 9, 25, and 101) in the ensemble method.

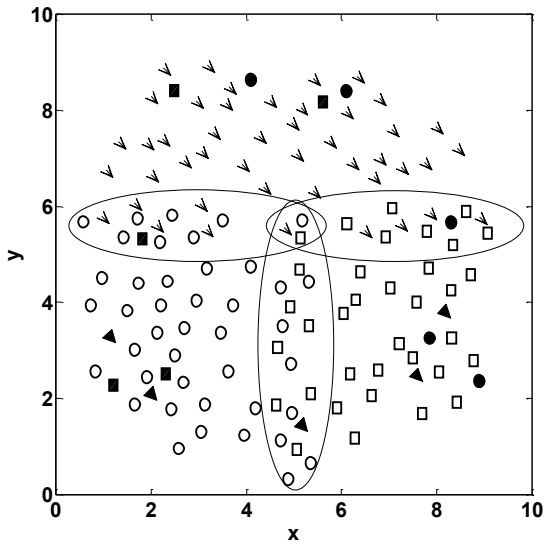


Figure 2. An example of core, outlier, and boundary subsets. The arrow, circle, and square denote three classes of data. The unfilled and non-encircled ones belong to the core subset, the filled ones belong to the outlier subset, and the unfilled but encircled ones belong to the boundary subset.

N, then these classifiers follow the binomial distribution and the accuracy of the ensemble method can be calculated as

$$p_B = \sum_{i=K+1}^{n} \binom{N}{i} p^i (1-p)^{N-i} \quad (1)$$

If we consider p_B as a function of p, it can be shown that for any given $n > 1$, p_B strictly increases when p increases. When $p > 0.5$, $p_B > p$, when $p < 0.5$, $p_B < p$, and when $p = 0.5$, $p_B = p$. So as long as individual classifiers are accurate ($p_i > 0.5$), the accuracy of an ensemble method is better than the average accuracy of individual classifiers. Figure 1 shows the relation of p and p_B when N = 5, 9, 25, and 101.

The assumption of independent individual classifiers and the binomial model have been widely used in the ensemble learning to illustrate the reason why ensemble methods achieve better accuracies than individual classifiers [8] [12]. But while the binomial model gives a qualitative reason for using ensemble methods, it fails to give a quantitative measure of the actual accuracies of ensemble methods. For example, with the binomial model, as long as the accuracy of individual classifiers $p > 0.5$, when $N \to \infty$, $p_B \to 1$. But in practical studies of majority voting ensemble methods, such as bagging or random forests, we rarely see accuracies approaching 100%, even if we use a large N [2] [3] [7]. Furthermore, studies show that ensemble methods may perform worse than single classifiers in some cases [6] [14].

One reason why the binomial model fails to quantitatively measure the accuracies of the ensemble methods is that the individual classifiers may not follow the binomial distribution. Binomial distribution requires individual classifiers be independent in predicting all items, but in practice some items may always be predicted correctly or incorrectly by all classifiers and in this case the individual classifiers will no longer be independent. For example, given a dataset with weights from 10-year-old kids and 25-year-old men, those with weights less than 90 pounds are almost exclusively 10-year-old kids, but those with weights more than 200 pounds are almost exclusively 25-year-old men. No matter how we build our individual classifiers, the binomial model will not fit for this dataset.

B. The COB Model

In the COB model, we assume that a dataset consists of three subsets: core, outlier, and boundary. An example of these three subsets is shown in Figure 2. For a dataset with items from multiple classes, some items may be buried by items from the same class (the unfilled and non-encircled points in the figure), some may be buried by items from other classes (the filled points in the figure), and some may be surrounded by mixed items from the same and other classes (the unfilled but encircled point in the figure). We classify those items in the first case as a core subset, those in the second case as an outlier subset, and those in the last case as a boundary subset. From another point of view, a core subset contains items that are clearly different from items in other classes, an outlier subset contains items that are classified by mistake, and a boundary subset contains items that are similar to some items in other classes and can be correctly predicted or misclassified. We note that it may not be possible to classify a highly noisy dataset into these three subsets, as the errors will be overwhelming.

The COB model models these three subsets separately. We define the numbers of items in the core, boundary, and outlier subsets as n_1, n_2, and n_3, respectively, so $n_1 + n_2 + n_3 = n$. For an ensemble method E with N individual classifiers $\{C_i, i = 1, 2, \ldots, N\}$, all classifiers always predict the items in the core subset correctly and predict the items in the outlier subset incorrectly. We define the accuracy on the core subset as q

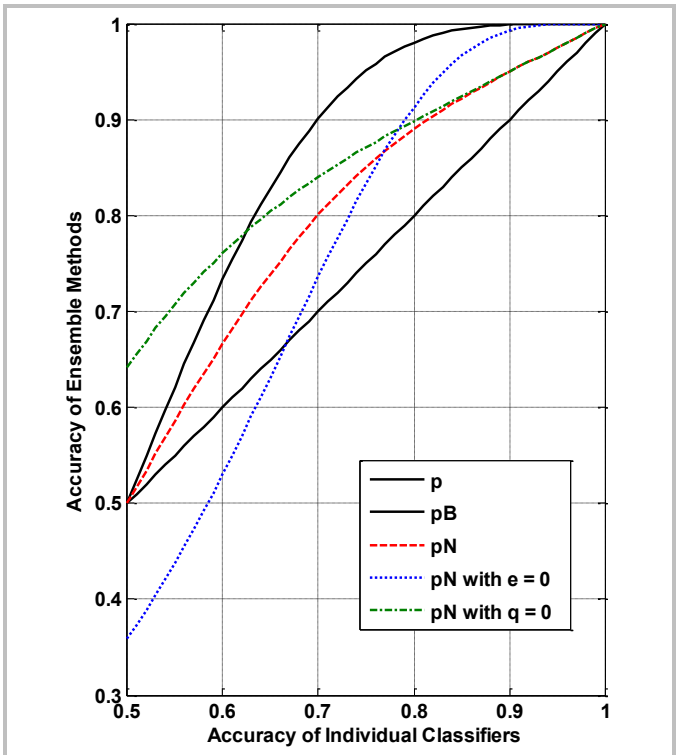


Figure 3. Comparison of the p_E from the COB model to the p_B from the binomial model and the accuracy of individual classifiers p.

(note that $q \le \min_i(p_i)$), the error on the outlier subset as e (note that $e \le \min_i(1-p_i)$), and the accuracy of each classifier C_i on the boundary subset as p_{ib}, so $q = n_1/n$, $e = n_3/n$, and $p_{ib} = p_i - q$.

We assume each individual classifier C_i is independent in making predictions for items in the boundary subset. If the accuracies p_{ib} of all individual classifiers are the same, say, $p_{ib} = p_b = p - q$, $i = 1, 2, \ldots, N$, then all classifiers C_i follow the binomial distribution. Since the adjusted accuracy of each classifier C_i on the boundary subset is $p_{ba} = p_b n/n_2 = p_b n/(n - n_1 - n_3) = (p-q)/(1-q-e)$ and the adjusted accuracy of the ensemble method on the boundary subset is $p_{Ba} = \sum_{i=K+1}^{n} \binom{N}{i} (p_{ba})^i (1-p_{ba})^{N-i}$, the accuracy of ensemble method is

$$p_N = (p_{Ba} n_2 + q n_1)/n = p_{Ba}(1-q-e) + q \quad (2)$$

For any given $N > 1$, p_N is a function of p, q, and e. Given any q, e, and N that $0 \le q \le p$ and $0 \le e \le 1 - p$, p_N strictly increases when p increases. When $q = e = 0$ (both core and outlier subsets are empty), we have $p_{ba} = p$ and $p_N = p_B$ and the COB model becomes the binomial model.

When $q \ne 0$ and $e = 0$ (the outlier subset is empty), we have $p_{ba} = (p - q) / (1 - q)$ and

$$p_N = p_{Ba}(1-q) + q \quad (3)$$

When $q = 0$ and $e \ne 0$ (the core subset is empty), we have $p_{ba} = p / (1 - e)$ and

$$p_N = p_{Ba}(1-e) \quad (4)$$

Figure 3 shows the accuracies p_N in equations (2), (3), and (4) when $N = 9$ and compares them to the accuracy of p_B (from the binomial model) and the accuracy of p (from individual classifiers). We use $q = 0.5p$ and $e = 0.5(1 - p)$ and generate p_N and p_B for $0.5 \le p \le 1$.

For the p_N in equation (3), given any p and q that $0.5 \le p \le 1$ and $0 \le q \le p$, we can show that $p_N \le p_B$. It means that when all individual classifiers make the same correct predictions on some items, the accuracy of an ensemble method will be worse than the accuracy generated from the binomial model. For example, Figure 3 shows that when $p = 0.7$ and $q = 0.5p$, we have $p_B = 0.9$, but $p_N \approx 0.73$. When $p = 0.5$ and $q = 0.5p$, we have $p_N \approx 0.37 < 0.5$! Furthermore, for any p that $0.5 \le p \le 1$, given two q_1 and q_2 that $0 \le q_1 \le q_2 \le p$, we can show that $p_{N2} \le p_{N1}$. It means if all individual classifiers make the same correct predictions on more items, then the accuracy of an ensemble method will be worse. Based on this new model, if we want to achieve high accuracy for an ensemble method, it would better to find individual classifiers that are diverse on correctly predicted items, especially for weak learners.

For the p_N in equation (4), given p and q that $0.5 \le p \le 1$ and $0 \le q \le p$, curve p_N crosses curve p_B. When p is close to 0.5, $p_N > p_B$, and when p is close to 1, $p_N < p_B$. It shows that for weak individual classifiers, it is possible to make the accuracy of an ensemble method better than that of binomial model, if all individual classifiers make incorrect predictions on some items. For example, Figure 3 shows that when $e = 0.5(1 - p)$ and $0.5 \le p < 0.62$, we have $p_N > p_B$. When $e = 0.1(1 - p)$ and $0.5 \le p < 0.68$ we have $p_N > p_B$. So based on this new model, the accuracy of an ensemble method can be better than the binomial model if all weak individual classifiers make incorrect predictions on some items (that is, concentrates on these items). This is also one possible reason why sometimes ensemble methods show a much better performance with weak classifiers.

It should be noted that an arbitrary large e may make the accuracy of an ensemble method worse. For example, when $0.5 \le p \le 1$, the two curves of p_N for $e = 0.1$ and $e = 0.5$ will cross each other and p_N ($e = 0.5$) < p_N ($e = 0.1$) when $p > 0.61$, so a larger e means worse accuracy for the ensemble. One further study shows that when $0.5 \le p \le 1$, p_N ($e = 0.5$) > p_N ($e > 0.62$), so a good rule is to keep $e \le 0.5$.

When both $e \ne 0$ and $q \ne 0$, the p_N will be affected by both e and q. If e has stronger effect on p_N than p, we will still see an ensemble method with weak classifiers to have better accuracy than the accuracy from the binomial model.

One issue in practice is that it is highly unlikely that a dataset will clearly show the three subsets. Most data items are neither correctly predicted by all individual classifiers nor

independent but are something in between. To make an approximation, we use threshold values h_q and h_e for the core and outlier subsets. For example, if $h_q = 0.95$ and $h_e = 0.9$, then we classify those items as correctly predicted by ≥ 95% of classifiers into the core subset, those incorrectly predicted by ≥ 90% of classifiers into the outlier subset, and the remaining into the boundary subset.

III. Experiments and Discussion

A. Datasets

We collected 32 datasets from the UCI Machine Learning Repository [10]. The datasets have item sizes vary from 10^2 to 10^4 and feature sizes from 4 to 10^2, as shown in Table I. A few datasets have missing values and we replaced them with negative values. The nominal data types are changed to integers and are numbered starting from 1 based on the order of the appearance. For those dataset with multiple classes, we use class 1 as the positive class and all other classes as the negative class.

TABLE I. The list of 32 datasets

Australia, Cleveland, Diabetes, Ecoli, German, Glass, Heart, HillValley, HillValley_Noise, ImageSeg, Ionosphere, Iris, Kp_vs_k, Kr_vs_kp, Led_24, Letter, Libras, Liver, Pendigits, Pima, Promoters, Satimage, Segment, Sonar, Soybean, Shuttle, Thyroid, Vehicle, Votes, Vowel, Waveform, WaveformNoise

B. Majority voting ensembles and individual machine learning algorithms

We tested the COB model on three majority voting ensembles: bagging, random forests, and a randomized ensemble. In the randomized ensemble, for each individual classifier, we generated a new training set by uniformly sampling $\lceil n/2 \rceil$ items without replacement from an original training dataset, as shown in Algorithm 1.

Algorithm 1 Given a dataset X with n items, build a randomized ensemble with m classifiers.

for i = 1 to m

Uniformly sample $\lceil n/2 \rceil$ items in x without replacement

Build a classifier C_i on the $\lceil n/2 \rceil$ items

end for

For the bagging and randomized ensemble, we used three different machine learning algorithms each: decision trees, *k*-nearest neighbors, and support vector machines. For all individual machine learning algorithms, we test on two different sizes of classifiers: a small number of classifiers $N = 7$ and a large number of classifiers $N = 51$. For the decision trees, we tested on both the unpruned full trees and the pruned trees with prune level = 2. For the *k*-nearest neighbors, we tested on two different sizes of *k*: a small $k = 3$ and a large $k = 41$. For the support vector machines, we tested on only the Radial Basis Function (RBF) kernel with the soft margin = 1.

For the random forests, we used decision trees with $N = 7$ and $N = 51$. For the number of variables (features) F at each node of the tree, we tested two numbers $F = \log_2(\text{number of features})$ and $F = \sqrt{\text{number of features}}$. In total, there were 24 experiments for all three ensembles.

C. Experiments

We used a 10-fold cross validation for each experiment. For the total of 10 rounds of cross validation for each dataset in each experiment, we recorded the mean of the average accuracy of individual classifiers $\overline{p}_i$, the accuracy of the ensemble method p_E, the accuracy predicted from a binomial model p_B, the accuracy predicted from the COB model p_N, and the q and e. All the detailed results for the 32 datasets in the MATLAB format are available at http://www.xwanglab.com/research/COB/SupplementaryMaterials.zip.

The COB model and the binomial model require all individual classifiers to have the same prediction accuracies, that is, $p_1 = p_2 = \ldots p_N = p$, so we approximated p by taking the average $p \approx \overline{p} \le \sum_{i=1}^{N} p_i / N$.

We used threshold values $h_q = h_e = 0.85$ to approximate actual datasets for the model. Items correctly predicted by ≥ 85% of individual classifiers are classified into the core subset, items incorrectly predicted by ≥ 85% of classifiers are classified into the outlier subset, and the remaining items are classified into the boundary subset. We tested a few threshold values h_q and h_e and the model shows good performance with $h_q = h_e = 0.85$.

D. Results and Discussion

Tables II and III summarize the 24 experiments by showing the mean absolute error $\overline{|p_N - p_E|}$ and mean relative error $\overline{|p_N - p_E|}/p_E$ between the predicted accuracies p_N from the COB model and the actual accuracies p_E and comparing them with the errors from the binomial model. The MATLAB code is available upon request.

From the two tables, the COB model shows significant better performance than the binomial model in quantitatively measuring the accuracies of the ensemble methods. All 24 experiments have $\overline{|p_N - p_E|} < 5\%$ and $\overline{|p_B - p_E|}/p_E < 5\%$, while in the binomial model, all 24 experiments have $\overline{|p_B - p_E|} > 5\%$ and $\overline{|p_B - p_E|}/p_E > 5\%$. For the performance of 768 individual experiments (per dataset experiment for the 32 datasets in 24 experiments), in the COB model, only 75 (9.8%) individual experiments have absolute errors $|p_N - p_E| > 5\%$ and 7 (0.9%) individual experiments have absolute errors $|p_N - p_E| > 10\%$, while in the binomial model, 433 (56.4%) individual experiments have absolute errors $|p_B - p_E| \ge 5\%$ and 310 (40.4%) individual experiments have absolute errors $|p_B - p_E| \ge 10\%$.

For all three machine learning algorithms, the COB model matches the actual performance very well, while support vector

TABLE II. THE MEANS AND STANDARD DEVIATIONS OF ABSOLUTE ERRORS $|P_N - P_E|$ AND $|P_B - P_E|$ OF THE 32 DATASETS IN THE 24 EXPERIMENTS. NOTE RANDOM FORESTS USES PARAMS $F = log_2(\#\text{FEATURES})$ AND $F = \sqrt{\#\text{FEATURES}}$.

	#CLASSIFIERS	PARAMS	BAGGING		RANDOM FOREST		RANDOMIZED ENSEMBLE	
			$\|P_N - P_E\|$	$\|P_B - P_E\|$	$\|P_N - P_E\|$	$\|P_B - P_E\|$	$\|P_N - P_E\|$	$\|P_B - P_E\|$
DECISION TREES	m = 7	Unpruned	2.76±2.40	5.93± 4.95	3.95±3.01	5.09± 4.78	2.70±2.27	6.02± 5.18
		Pruned	2.44±2.09	6.68± 5.22	2.89±2.48	5.75± 4.98	2.35±1.93	7.65± 5.96
	m = 51	Unpruned	2.57±3.29	9.87± 9.20	3.20±2.97	8.88± 8.41	2.76±3.66	9.81± 8.78
		Pruned	2.07±2.52	10.64± 8.93	2.45±3.05	9.91± 9.13	2.53±2.83	11.89± 9.41
K-NN	m = 7	k = 3	1.55±1.82	6.80± 5.68	N/A	N/A	1.17±1.07	7.12± 5.95
		k = 41	0.62±0.91	8.31± 5.62	N/A	N/A	0.85±1.42	8.29± 5.81
	m = 51	k = 3	1.36±1.51	11.87±10.93	N/A	N/A	1.07±1.31	11.38±10.62
		k = 41	0.76±1.11	12.87±10.80	N/A	N/A	0.88±1.54	13.88±10.95
SVM	m = 7	GRB kernel	0.56±0.64	6.39± 6.22	N/A	N/A	0.82±1.02	6.79± 6.49
	m = 51	GRB kernel	0.98±1.35	10.29±10.79	N/A	N/A	1.08±1.41	11.48±11.56

TABLE III. THE MEANS AND STANDARD DEVIATIONS OF RELATIVE ERRORS $|P_N - P_E| / P_E$ AND $|P_B - P_E| / P_E$ OF THE 32 DATASETS IN THE 24 EXPERIMENTS. NOTE RANDOM FORESTS USES PARAMS $F = log_2(\#\text{FEATURES})$ AND $F = \sqrt{\#\text{FEATURES}}$.

	#CLASSIFIERS	PARAMS	BAGGING		RANDOM FOREST		RANDOMIZED ENSEMBLE	
			$\|P_N - P_E\|$	$\|P_B - P_E\|$	$\|P_N - P_E\|$	$\|P_B - P_E\|$	$\|P_N - P_E\|$	$\|P_B - P_E\|$
DECISION TREES	m = 7	Unpruned	3.45±3.09	7.33± 6.33	4.79±3.77	6.25± 6.90	3.38±2.95	7.45± 6.62
		Pruned	3.11±2.83	8.29± 6.74	3.63±3.28	7.09± 6.30	3.03±2.60	9.76± 8.07
	m = 51	Unpruned	3.37±4.78	12.63±12.57	4.01±4.02	11.18±11.17	3.67±5.35	12.43±11.74
		Pruned	2.70±3.57	13.74±12.45	3.28±4.51	12.66±12.41	3.33±4.13	15.50±13.22
K-NN	m = 7	k = 3	2.03±2.34	8.89± 7.70	N/A	N/A	1.55±1.54	9.28± 8.09
		k = 41	0.84±1.24	10.82± 8.01	N/A	N/A	1.18±1.96	10.79± 8.19
	m = 51	k = 3	1.86±2.19	16.45±16.27	N/A	N/A	1.55±2.26	15.45±15.35
		k = 41	1.09±1.65	17.44±16.00	N/A	N/A	1.37±2.55	19.41±16.84
SVM	m = 7	GRB kernel	0.82±1.00	8.82± 8.69	N/A	N/A	1.23±1.62	9.02± 9.00
	m = 51	GRB kernel	1.64±2.68	14.75±15.90	N/A	N/A	1.74±2.68	16.48±11.16

machines and *k*-nearest neighbors has slightly better performance than decision trees, as shown in Tables II and III. The relatively worse performance of the decisions trees is possibly because of the overfitting problem, as we can see that the pruned trees has better performance than unpruned trees in the COB model.

Among the three ensemble methods, the random forests ensembles show the worst accuracy. Since the random forests use a random subset features in each node instead of using all features, it seems the COB model performs better when all features are presented.

For the two individual classifier sizes (m = 7 and 51), it seems the COB model performs slightly better for the smaller classifier size. For the classifier size m = 7, 37 individual experiments have absolute errors $|p_N - p_E| > 5\%$ and 1 individual experiment has absolute errors $|p_N - p_E| > 10\%$, while for the classifier size m = 51, 38 individual experiments have absolute errors $|p_N - p_E| > 5\%$ and 6 individual experiments have absolute errors $|p_N - p_E| > 10\%$. Overall the performance of the COB model on both classifier sizes is similar.

Figures 4a and 4b show relationship of dataset size or feature set size versus the relative error. The relative error stays almost the same when the dataset size or feature set size changes, so it shows the performance of the COB model is stable with different dataset sizes or feature sizes.

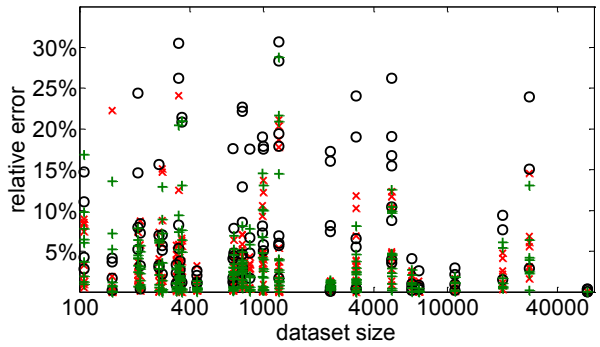
IV. CONCLUSION

We propose a new model called COB for quantitative measuring the accuracy of ensemble methods. Although the COB model is a simple expansion to the binomial model by adding two subsets (core and outlier), the experiments on 32 datasets show that the accuracies predicted from this model match the actual accuracies of ensemble methods very well, especially when we use ensemble methods with all features presented. The model is stable under different machine learning algorithms, dataset sizes, or feature sizes.

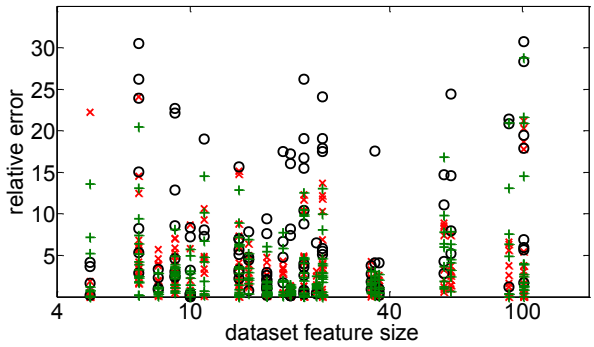
One interesting future work is to find some new ensemble methods that generate a larger boundary subset and a smaller core subset in order to achieve better performance. For example, methods that find a subset of features that minimize the core subset and maximize the boundary subset. Another interesting work is to see if the model works for those ensemble methods using heterogeneous classifiers (i.e. using two or more different machine learning algorithms).

REFERENCES

[1] B. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer Jr., "A New Ensemble Diversity Measure Applied to Thinning Ensembles," In *Proceedings of the Fourth International Workshop on Multiple Classifier Systems*, 2003, pp.306–316.



4a. Relationship of dataset size and relative error



4b. Relationship of dataset feature size and relative error

Figure 4. The dataset size/feature size versus the relative error $|p_N - p_E| / p_E$. x denotes bagging ensembles, o denotes random forests, and + denotes the randomized ensemble.

[2] L. Breiman, "Bagging Predictors," *Machine Learning*, 1996, 24(2):123–140.

[3] L. Breiman, "Random Forests," *Machine Learning*, 2001, 45:5–32.

[4] G. Brown and L. I. Kuncheva, "Good and Bad Diversity in Majority Vote Ensembles," In *Proceedings of the Ninth International Workshop on Multiple Classifier Systems*, 2010, pp.124–133.

[5] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble Selection from Libraries of Models," In *Proceedings of the 31st International Conference on Machine Learning*, 2004, pp.137–144.

[6] I. Davidson and W. Fan, "When Efficient Model Averaging Out-Performs Boosting and Bagging," *In Proceedings of 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Lecture Notes in Computer Science, 2006, 4213:478–486.

[7] T. G. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization," *Machine Learning*, 2000, 40(2):139–157.

[8] T. G. Dietterich, "Ensemble Methods in Machine Learning," In *Proceedings of the First International Workshop on Multiple Classifier Systems*. Lecture Notes in Computer Science, 2001, 1857:1–15.

[9] S. Dudoit and J. Fridlyand, "Bagging to Improve the Accuracy of a Clustering Procedure," *Bioinformatics*, 2003, 19(9):1090–1099.

[10] A. Frank and A. Asuncion, *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2001.

[11] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," In *Proceedings of the 13th International Conference on Machine Learning*, 1996, pp. 325–332.

[12] L. K. Hansen and P. Salamon, "Neural Network Ensembles," *IEEE Trans. Pattern Analysis and Machine Intell.* 1990, 12(10):993–1001.

[13] L. I. Kuncheva and C. J. Whitaker, "Measures of Diversity in Classifier Ensemble and Their Relationship with the Ensemble Accuracy," *Machine Learning*, 2003, 51(2):181–207.

[14] R. Maclin and D. Opitz, "An Empirical Evaluation of Bagging and Boosting," *In Proceedings of the 14th National Conference on Artificial Intelligence*, 1997, pp. 546–551.

[15] J. R. Quinlan, "Bagging, boosting, and C4.5," In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996, pp.725–730, Cambridge, MA.

[16] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric Bagging and Random Subspace for Support Vector Machine-Based Relevance Feedback in Image Retrieval," *IEEE Trans. Patt. Anal. mach. Intell*, 2006, 28(7):1088–1099.

[17] X. Wang and N. Davidson, "The Upper and Lower Bounds of the Prediction Accuracies of Ensemble Methods," In *Proceedings of the 9th International Conference on Machine Learning and Applications (ICMLA 2010)*, 2010, pp.373–378.