# Methods for Combining Heterogeneous Sets of Classifiers

**Dennis Bahler and Laura Navarro**
Artificial Intelligence Laboratory
Department of Computer Science
North Carolina State University, Raleigh NC

**Abstract**

The combination of classifiers has long been proposed as a method to improve the accuracy achieved in isolation by a single classifier. In contrast to such well-explored methods as boosting and bagging, we are interested in ensemble methods that allow the combination of heterogeneous sets of classifiers, which are classifiers built using differing learning paradigms. We focus on theoretical and experimental comparison of five such combination methods: majority vote, a Bayesian method, a Dempster-Shafer method, behavior-knowledge space, and logistic regression. We have developed an upper bound on the accuracy that can be obtained by any of the five methods of combination, and can show that this estimate can be used to determine whether an ensemble may improve the performance of its members. We have conducted a series of experiments using standard data sets and learning methods, and compared experimental results to theoretical expectations.

## Introduction

The reasons for combining the outputs of multiple classifiers are compelling, because different classifiers may implicitly represent different useful aspects of a problem, or of the input data, while no one classifier represents all useful aspects. In the context of pattern recognition, the idea of combining the decisions of several classifiers has been well explored. A wide variety of methods have been used to combine the results of several neural networks [Cho 95, Krogh & Vedelsby 95, Perrone & Cooper 93, Schaal & Atkeson 95, Tumer & Ghosh 94], decision trees [Ali & Pazzani 95b, Kong & Dietterich 95, Bauer & Kohavi 99], sets of rules [Ali & Pazzani 95a], and other models [Ho, Hull, & Srihari 94; Huang & Suen 95; Wernecke 92; Xu, Krzyzak, & Suen 92; Xu & Jordan 93; Xu, Jordan, & Hinton 95]. Among the well-known general methods proposed have been bagging [Breiman 96] and Adaboost [Freund & Schapire 97].

In contrast to approaches which combine models derived from multiple versions of the same learning method, the specific focus of this paper is on ensemble methods that are able to combine the decisions of multiple classifiers of different types, so-called heterogeneous sets of classifiers. Different classifiers typically express their opinions in different ways. For example, an expert may report probabilities, distances between the concept described by an example and already defined concepts, or simply a label representing the predicted class of the example.

We can formalize our problem as follows. Assume a pattern space $U = C_1 \cup C_2 \cup \ldots \cup C_M$ of $M$ mutually exclusive sets, where each $C_i$, for all $i \in \{1, \ldots, M\}$ represent a set of patterns called a **class**. Each **expert** or **classifier** (denoted by e) assigns to a sample $x \in U$ an index $j \in \{1, \ldots, M+1\}$ that represents x as being from class $C_j$, if $j \neq M+1$, and $j = M+1$ represents that the expert has no idea about which class x belongs to, i.e. x is **rejected** by e. The decision of expert e is denoted by $e(x) = j$.

Based on this definition of the decision of an expert, our problem can be defined as the **combination** of $D$ different classifiers $e_k$, $k = 1, \ldots, D$, each of which assigns x to a label $j_k$ denoted $e_k(x) = j_k$, into an **integrated classifier** E, which gives x one definitive label j. The performance of a classifier can be described by its recognition, substitution, and rejection rates.

To investigate systematically various approaches to combination, we performed the following set of experiments. We began with the only four data sets from the UCI repository [UCI 96] that were suitable for our purposes: Tic-Tac-Toe Endgame, Wisconsin Breast Cancer, Pima Indians Diabetes, and Credit Screening. These data sets were each used to train a heterogeneous set of three classifiers using standard techniques: a decision tree, a Bayesian belief network, and a back-propagation neural network. To obtain an integrated classifier, the results of the classifiers for each data set were then combined into an ensemble using five methods: majority vote, a method based on Bayes' rule, a method based on Dempster-Shafer evidence combina-

tion, behavior-knowledge space, and logistic regression. Next, by examining conditions under which the combination of classifiers does not improve the accuracy obtained by a single expert, we developed a method to predict the theoretical maximum recognition rate that can be achieved by each of the ensemble methods. Finally, we compared the experimentally observed changes in accuracy of the ensemble models with the theoretical bounds.

## Methods of Combining Multiple Classifiers

Combination methods arising in statistics [French 85] and operations research [Bates & Granger 69] typically combine not only classification but also measures of confidence in these opinions. These combination techniques require that measures be homogeneous, that is, of the same kind for all the experts in the group; for example, all of them represent probabilities, or distances, or they need to be converted to equivalent scales. If the experts are not able to supply the same measure of confidence, or not able to supply any measure at all, we are confronted with a different problem. This is the case of interest to us, when several heterogeneous classifiers are to be combined, where the only common information that is available from all the experts is their opinion. Methods have been developed to deal with this problem, and they are described in this section.

### Voting

Generally speaking, the voting principle is just what we know as majority voting. Several variations of this idea have been proposed:

**Unanimity.** The combined classifier decides that an input pattern x comes from class $C_j$ if and only if all the classifiers decide that x comes from class $C_j$, otherwise it rejects x.

**Modified unanimity.** The combined classifier decides that an observation x comes from class $C_j$ if some classifiers support that x belongs to $C_j$, and no other classifier supports that x belongs to any other class (i.e. rejects x), otherwise it rejects x.

**(Weighted or unweighted) majority.** The majority rule, where the combined classifier decides an observation x belongs to class $C_j$ if more than half of the classifiers support that x belongs to $C_j$. A modification of this rule is to require a different proportion of classifiers to agree instead of half of them.

**Thresholded plurality.** The combined classifier decides for the observation x belonging to class $C_j$ if the number of classifiers that support it is considerably bigger than the number of classifiers that support any other class.

Combining classifiers with this method is simple; it does not require any previous knowledge of the behavior of the classifiers nor does it require any complex methodology to decide. It only counts the number of classifiers that agree in their decision and accordingly decides the class to which the input pattern belongs. This simplicity has a drawback, however: the weight of the decision of all the classifiers is equal, even when some of the classifiers are much more accurate than others.

### Bayesian Ensemble Methods

Voting methods are based solely on the output label computed by each classifier. No expertise or accuracy is considered. In these methods the decision of each classifier is treated as one vote, but what happens if one of the classifiers is much more accurate than any other? Should its accuracy not be considered in the combination? Let us suppose that we ask two experts about their opinions on whether a stock value will increase in the next month. We are likely to consider their opinions differently depending on how accurate they have been in the past.

To address this problem we can establish weights proportional to each expert's accuracy, so each classifier's output is considered according to its past performance and combining them using Bayes' theorem [French 85, Lee 97].

### Behavior-Knowledge Space Methods

One of the strongest conditions for the applicability of Bayes' rule to combine the decision of several experts is the fact that they must perform independently. The Behavior-Knowledge Space (BKS) method [Huang & Suen 95, Wernecke 92], by contrast, makes no assumption about the data to be combined.

A BKS is a D-dimensional space, where each dimension corresponds to the decision of one classifier. Each classifier has $M + 1$ possible decision values, for the M classes plus the rejection of the given sample by the classifier (class $M + 1$). The intersection of the decisions of individual classifiers occupies one point of the BKS.

### Dempster-Shafer Ensemble Methods

The use of Bayes' rule to combine the decision of several classifiers may be inappropriate in some cases. Bayes' rule requires the belief measures to behave as probabilities [Lee 97] and in the case of decision of experts this is a requirement often impossible to satisfy.

The Dempster-Shafer calculus is a system for manipulating degrees of belief that is more general than the Bayesian approach and does not require the additive probability assumption. Xu [Xu, Krzyzak, & Suen 92] introduced a method to combine multiple classifiers by adapting Dempster-Shafer's evidence theory.

In using Dempster-Shafer theory to combine the decisions of several classifiers, the exclusive and exhaustive possibilities that form the frame of discernment $\theta$ are the propositions $x \in C_i$, $i = 1, \ldots, M$ that denote that an input sample x comes from class $C_i$.

### Logistic Regression Ensembles

Another method that assigns weights to each classifier is the combination using logistic regression [Ho, Hull, & Srihari 94].

Consider $Y_i$, a binary variable that has value 1 when $x \in C_i$ and 0 otherwise. The goal of recognition is to predict the value of $Y_i$, for all $i = 1, \ldots, M$. Since $Y_i$ is a binary value, the combination problem may be reformulated in the context of logistic regression analysis, where $Y_i$ depends on the value of an explanatory variable.

## Analysis of Ensemble Methods

These methods have all been used in applications, but the empirical evidence of their effectiveness is mixed. In many studies they have improved the accuracy obtained by any single classifier [Ali & Pazzani 95a; Ho, Hull, & Srihari 94; Huang & Suen 95; Wernecke 92; Xu, Krzyzak, & Suen 92]. At the same time, in other analytical studies no improvement was seen from that of the single classifier [Hansen & Salamon 90, Meir 95]. Characteristics of the data to be combined determine the degree of improvement obtained by the combination.

Since, unlike all our other combination methods. majority voting is not based on the performance of the members of the ensemble, we analyze the performance of majority voting separately.

### Majority Voting

When multiple classifiers are combined using majority vote, we expect to obtain good results based on the belief that the majority of experts are more likely to be correct in their decision when they agree in their opinion. So if the decision of $D$ classifiers are combined, and more than half of them decide that observation x belongs to class $C_i$, the ensemble decides that $x \in C_i$.

Even when common sense leads us to believe that combining expert's decisions in this way will improve the correctness of the integrated result, this is not always true. Hansen and Salamon [Hansen & Salamon 90] show that if all of the classifiers being combined have an error rate less than 50%, we may expect the accuracy of the ensemble to improve when we add more experts to the combination. However, their proof is restricted to the situation when all of the classifiers perform independently and have the same error rate.

Ali and Pazzani [Ali & Pazzani 95b] show that the degree of error reduction obtained by the ensemble of classifiers that uses majority voting is related to the degree to which the members of the ensemble make errors in an uncorrelated manner. Matan [Matan 96] has established lower and upper bounds for the ensemble of classifiers, showing that majority voting might, in certain conditions, perform worse than any of the members of the ensemble.

### Ensembles Based on the Performance of their Members

The majority voting combination does not consider that some classifiers of the ensemble may be more accurate than others and treats the opinions of each expert as a vote, independently of the fact that this classifier may be always wrong or right.

In attempt to improve the performance of the combination over majority voting, we explored other ways of combining different classifiers that consider individual classifier performance. Other methods to combine the expert decisions have different characteristics and rely on different assumptions, so each may combine some kinds of data better than others. We also show how to estimate an upper bound for the performance of any ensemble, regardless of the combination method used.

**Independence.** Combinations based on Bayes' theorem and Dempster-Shafer theory assume independence in the decisions of the members of the ensemble. Xu *et al.* [Xu, Krzyzak, & Suen 92] proposed that this independence can be achieved if the experts are trained over different data sets. However, considering that all the experts are modeling the same function that defines the real class of each observation, it is unlikely that they behave independently as required by these ensemble methods.

Ensembles using behavior-knowledge space (BKS) do not assume that the decisions of the classifiers are independent, and when they are, this method and the Bayes method should obtain the same performance, since they are based on the same principle.

When the BKS method is used with $D$ experts that decide between $M$ classes, then there exist $(M+1)^D$ possible combinations of decisions to be taken into account. We also require enough samples from each possible combination of decisions to obtain a representa-

tive number of observations. For example, consider an ensemble of five experts and three classes using 10 samples per combination. This will require at least 10240 observations to train, which may or may not be reasonable.

**Maximum accuracy of an ensemble.** The section above outlines when each kind of combination scheme is likely to perform better than others. When there is not enough data available for training it is better to use Bayesian ensembles than BKS, and the Dempster-Shafer combination is better suited than other combination techniques to manipulate uncertain decisions of the individual classifiers.

This analysis could lead us to believe that even when some ensemble methods might not obtain better accuracies than the single classifiers, careful selection of the combination method, or creation of other combination methods, will always allow us to find a way to improve the performance obtained by any single expert. There is, however, a limit in the accuracy ensembles can achieve regardless of the combination method used.

Let $\theta$ be the set of all possible classes $\{1, \ldots, M\}$ to which an observation can belong. To obtain the maximum accuracy achievable by any ensemble of $D$ experts, we define function G:$(\theta \cup \{M+1\})^D \rightarrow \theta$. Given the decision of $D$ classifiers regarding the class of an observation x, $e_1(x) = j_1, \ldots e_D(x) = j_D$, G returns the class with the most observations for that combination of decisions. Given G, we can then define the maximum accuracy that can be obtained by an ensemble as:

$$\frac{\sum_{j_1=1}^{M+1} \cdots \sum_{j_D=1}^{M+1} n_{j_1 \ldots j_D G(j_1 \ldots j_D)}}{N}$$

where $n_{j_1,\ldots,j_D,i}$ represents the number of samples of class $C_i$ that have been classified as $e_k(x) = j_k$ by each expert $k = 1, \ldots, D$, and N is the total number of samples considered.

## Experiments

We carried out several experiments to test the predictive value of the estimates for the performance of the ensemble methods, as well as the expected behavior of these methods considering their reliability and the independence of their decisions.

Four data sets were taken from the UCI repository: Tic-Tac-Toe Endgame, Wisconsin Breast Cancer, Pima Indians Diabetes, and Credit Screening databases. These data sets were used to train three different classifiers, whose results were then combined to obtain the integrated classifier.

### Data

The data sets chosen were the only ones suitable at UCI, because of their number of observations, number of classes, noise, and missing features. In general we chose data sets that had more than one hundred observations per class. We wanted the groups chosen to be a mixture of noisy/not noisy, with/without missing features. All the data sets chosen contain two classes.

Each data set was partitioned into four subsets. The single classifiers were then trained using one of the groups, and the accuracy of the resulting classifier was measured using the examples in the other three groups.

The performance of majority voting was tested using the three partitions not used to train the classifiers. The behavior of the ensembles requiring knowledge of the performance of their members was evaluated using one of the remaining groups to train the ensemble and the other two to test them. The training of the ensemble methods involved obtaining:

- Bayes: the confusion matrices for each of the members of the ensemble;
- BKS: the Behavior-Knowledge Space matrix;
- Dempster-Shafer: the recognition, substitution, and rejection rates for each individual classifier;
- Logistic Regression: the constant parameters of the logit function for each class.

This procedure was repeated for all permutations of the four groups in each data set to train the single classifiers and the combination method, and to test the ensembles. Thus, except for majority voting, the ensembles were tested in twelve different ways.

### Classifiers

The ensemble of classifiers requires, as data, the decision of several experts with respect to a given observation. For our experiments we used three different classifiers, each of them based in a particular machine learning paradigm. The classifiers used are:

- Bayesian belief network. No modification or selection of features was done except discretization of continuous data by means of minimum entropy. This kind of classifier returns the probabilities that a sample belongs to each class. To obtain the result of the classifier we take the class with greater probability.
- Backpropagation neural network. This classifier returns a number between 0 and 1. We considered a sample as belonging to a class if the difference between the output produced upon the analysis of that sample and number between zero and one assigned

to that class is the minimum. Since all the data sets used have only two classes, the numbers assigned to the classes are 0 and 1.

- Decision tree. This kind of classifier returns a label representing the predicted class of the ensemble.

## Conclusions

Combining multiple classifiers requires a uniform representation of their decisions with respect to an observation. In order to assure the ability of the ensemble methods to combine the decisions of different types of classifiers, we considered only methods that use a label for each classifier that indicates that the expert assigned the sample to the class represented by the corresponding label. We described several methods of combining classifiers that consider this restriction and use different information about the performance of the experts, to obtain the class that represents with greatest accuracy the set of samples described through each combination of decisions.

**The impact of ensemble method and rejection rate.** Majority voting ensembles, which do not use any information about the behavior of their members, were shown to perform almost as well as other more complex ensemble techniques when the recognition rate of all the classifiers is approximately the same and they do not reject any sample. However, when the performance of the ensemble members is not uniform, or samples are rejected by them, the performance of majority voting is affected negatively.

The experiments showed that from the ensembles that use measures of the performance of their members, the ones based on Bayes' theory, behavior-knowledge space and Dempster-Shafer theory of evidence perform similarly in all the cases. However, when the members of the ensemble have rejection rates greater than zero we observed an improvement in the performance of Dempster-Shafer combination, while the accuracy of the ensembles based on BKS was decreased. The accuracy of the Bayesian ensembles was not affected by rejection rate.

The ensembles based on logistic regression proved to be unreliable. We observed that in some cases they performed much better than any other ensemble method, while in other cases they performed worse than the least accurate of their members.

**The impact of the upper bound.** Even though it has been shown that ensembles of classifiers sometimes improve the accuracy of their members, this certainly is not always the case. We proposed an upper bound to the recognition rate achievable by any ensemble that uses only the output class of each classifier. The experimental results showed that the ensembles can improve the accuracies of their members only when this bound is much greater than the accuracy of the best classifier. We also showed that this upper bound is a tighter approximation to the real accuracy of the ensemble than the one proposed in [Matan 96] for majority voting ensembles. However, even when the ensembles of classifiers may not always improve the recognition rate of their members, they can often improve the reliability of their results.

**The impact of statistical independence.** One of the major restrictions for the applicability of the Bayesian combination is that the decisions of the experts involved in the ensemble must be independent, or at least conditionally independent given the class. We showed that this ensemble method can also be used when this condition is not satisfied. In the experiments the accuracy of the Bayesian ensemble was comparable to the accuracy of ensembles based on BKS, which do not make any assumption about the characteristics of the data to be combined. In fact, the Bayesian ensemble sometimes outperformed the BKS ensemble.

**Open problems.** There remain many open problems, both theoretical and experimental. We conclude with two of the most interesting. First, we established an upper bound for the accuracy of any ensemble; however we still do not know how bad the performance of an ensemble can be. For this reason it would be helpful to obtain a lower bound for the accuracy of the ensemble. Second, our experiments used only three experts. It is possible that the use of more experts will obtain greater improvements in the accuracy of the ensembles with respect to their members. So important questions remain: how many classifiers are required in an ensemble to obtain a desired accuracy, and how many are needed for an improvement over the accuracy of the best single classifier?

## References

**Ali & Pazzani 95b** Kamal M. Ali and Michael J. Pazzani 1995. On the link between error correlation and error reduction in decision tree ensembles. Technical Report UCI TR-95-38, University of California, Irvine, California.

**Ali & Pazzani 95a** Kamal M. Ali and Michael J. Pazzani 1995. Error reduction through learning multiple descriptions. *Machine Learning*, 24:173–202.

**Bates & Granger 69** J. M. Bates and C. W. J. Granger 1969. The combination of forecasts. *Operations Research Quarterly*, 20:319–325.

**Bauer and Kohavi 99** E. Bauer and R. Kohavi 1999. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, 36(1/2):105-139.

**Breiman 96** L. Breiman 1996. Bagging Predictors. *Machine Learning*, 24(2):123-140.

**Cho & Kim 95** Sung-Bae Cho and Jin Kim 1995. Combining multiple neural networks by fuzzy integral for robust classification. *IEEE Transactions on Systems, Man and Cybernetics*, 25(2):380–384.

**French 85** Simon French 1985. Group consensus probability distributions: a critical survey. *Bayesian Statistics*, 2:183–202.

**Freund & Schapire 97** Y. Freund and R. Schapire 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Computer and System Sciences*, 55:119-139.

**Hansen & Salamon 90** L. K. Hansen and P. Salamon 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001.

**Ho, Hull, & Srihari 94** Tin Kam Ho, Jonathan Hull, and Sargur Srihari 1994. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75.

**Huang & Suen 95** Y.S. Huang and C.Y. Suen 1995. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):90–94.

**Kong & Dietterich 95** E.B. Kong and T. Dietterich 1995. Error-correcting output coding corrects bias and variance. *Proceedings of the 12th conference on Machine Learning*, pp 313–321.

**Krogh & Vedelsby 95** Anders Krogh and Jesper Vedelsby 1995. Neural network ensembles, cross validation, and active learning. *NIPS 7*.

**Lee 97** Peter M. Lee 1997. *Bayesian Statistics*. New York: Wiley.

**Matan 96** Ofer Matan 1996. On voting ensembles of classifiers. *AAAI-96 Workshop on Integrating Multiple Learned Models*, pp 84–88. [http://www.cs.fit.edu/ imlm/].

**Meir 95** R. Meir 1995. Bias, variance, and the combination of estimators; the case of least linear squares. *NIPS 7*.

**UCI 96**
*UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]*. University of California, Department of Information and Computer Science, Irvine, California, 1996.

**Perrone & Cooper 93** M.P. Perrone and L.N. Cooper 1993. When networks disagree: Ensemble methods for neural networks. *Artificial Neural Networks for Speech and Vision, Chapter 10*.

**Schaal & Atkeson 95** Stefan Schaal and Christopher G. Atkeson 1995. From isolation to cooperation: an alternative view of a system of experts. *NIPS 7*.

**Tumer & Ghosh 94** Kagan Tumer and Joseph Ghosh 1994. A framework for estimating performance improvements in hybrid pattern classifiers. *Proceedings of the World Congress on Neural Networks*, III:220–225.

**Wernecke 92** Klaus-D. Wernecke 1992. A coupling procedure for the discrimination of mixed data. *Biometrics*, 48:497–506.

**Xu & Jordan 93** Lei Xu and M.I. Jordan 1993. EM learning on a generalized finite mixture model for combining multiple classifiers. *Proceedings of World Congress on Neural Networks*, IV.

**Xu, Jordan, & Hinton 95** Lei Xu, M.I. Jordan, and G. E. Hinton 1995. An alternative model for mixtures of experts. *NIPS 7*.

**Xu, Krzyzak, & Suen 92** Lei Xu, Adam Krzyzakk, and Ching Y. Suen 1992. Several methods for combining multiple classifiers and their applications in handwritten character recognition. *IEEE Transactions on System, Man and Cybernetics*, 22(3):418–435.