On Voting Ensembles of Classifiers (extended abstract)

Ofer Matan
Department of Computer Science
Stanford University, CA 94305
ofer@cs.stanford.edu

Abstract

We study the classification ability of majority-vote ensembles of classifiers. A majority ensemble classifies a pattern by letting each member of the ensemble cast a single vote for the correct class and decides according to a simple majority or a special majority vote. We give upper and lower bounds on the classification performance of a majority ensemble as a function of the classification performances of its individual members.

Keywords: Ensembles, Multiple Models, Boosting, Majority Gates, Voting Committees, Classification, Representation.

1 Introduction

We study the classification ability of majority ensembles of classifiers. A majority ensemble classifies a pattern by letting each member of the ensemble cast a single vote for the correct class and decides according to a simple majority or a special majority vote.

In this work we concern ourselves with the representational power of majority classifiers; That is, what are the best and worst combinations of individual classifiers, and what are the qualities of such combinations. Our results give insight valuable to either inductive or deductive algorithms that search this representation space. These results are of particular relevance to boosting algorithms that search the space of majority ensembles (Schapire 1990; Freund 1992; Freund & Schapire 1995).

2 A Motivating Example

Common sense leads us to believe that "Two heads are better than one". As the following example displays, multiple views do not always lead to improvement:

Example 1 (Generalized XOR) Consider the target concept of a generalized XOR function as shown in Figure 1. Let us study a certain ensemble $\mathcal{E}_1$ of 3 classifiers shown below the target concept. The concept described by each classifier is shown. Each of the three ensemble members perform at a correctness rate of 50%. However, the concept described by the majority classifier, has a correct rate of only 25% which is lower than any individual's performance. Figure 1 shows another ensemble, $\mathcal{E}_2$, in which each individual performs at a 50% correct rate. The concept described by this majority ensemble classifier achieves a performance rate of 75%.

Why do the two ensembles perform differently? Both ensembles are composed entirely of members performing at 50% corrrect. To answer this question we study the correctness indicator functions of the ensemble members. These functions are plotted below their respective classifiers in Figure 1. Areas for which the classifier is correct are colored grey and areas for which it is in error are colored white. Using this representation, it is easy to read off the performance rate as the relative area of grey.

In ensemble $\mathcal{E}_1$, the upper left quadrant is classified correctly by all ensemble members while all other quadrants are classified correctly by only one ensemble member. In both cases this is a "waste" of resources. First, the upper left quadrant needs only two ensemble members to classify correctly by majority. Second, the other correct answers of each ensemble member are not utilized due to participating in a minority. $\mathcal{E}_2$ which is an optimal combination "spends" only two ensemble members on each quadrant. Since each ensemble member is correct in only two quadrants (50% correct), a three member ensemble can classify 75% correctly by majority.

Ensembles $\mathcal{E}_1$ and $\mathcal{E}_2$ demonstrate the best and worst combinations of three "50%" classifiers. The following section presents the upper and lower bounds for arbitrary ensembles.

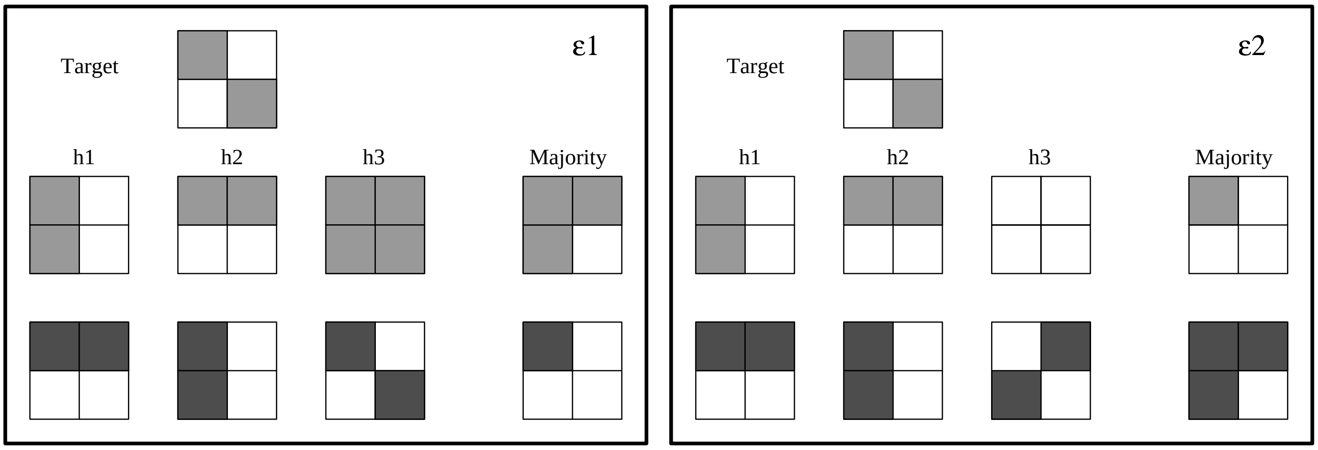

Figure 1: Two majority ensembles composed of classifiers performing at 50%. $\mathcal{E}_1$'s combined performance is 25% while $\mathcal{E}_2$'s combined performance is 75%. The top row shows the classifiers, the bottom row plots the respective correctness indicator function.

3 Performance Bounds on majority ensembles

Let X be a set called the *domain*. The *target concept* is a boolean function $c : X \to \{0, 1\}$. A *classifier* is a boolean function $h : X \to \{0, 1\}$.

The *performance* of a classifier $h(x)$ is the expected value

$$g = E_{x\ \mathcal{D}}([h(x) = c(x)]) \tag{1}$$

where $\mathcal{D}$ is a fixed and unknown distribution over the domain X and $[p]$ is the indicator (Iversonian) function for predicate p that returns 1 if p is true and 0 otherwise.

Given an ensemble of n classifiers $\{h_1, h_2, ..., h_n\}$, where n is an odd integer, the majority classifier is defined as:

$$h(x) = \begin{cases} 1 & \sum_1^n h_i(x) \geq \frac{n}{2} \\ 0 & Otherwise \end{cases} \tag{2}$$

We can generalize the notion of the majority classifier to a classifier with special majority by extending the range of the majority classifier to include an "unknown" response, $h : X \to \{0, 1, \mathtt{U}\}$. The definition of performance remains the same as in eq. 1.

A k-of-n special majority classifier for $k > \frac{n}{2}$ is defined as:

$$h_k^n(x) = \begin{cases} 1 & \sum_1^n h_i(x) \geq k \\ 0 & \sum_1^n h_i(x) \leq n-k \\ \mathtt{U} & Otherwise \end{cases} \tag{3}$$

In the following theorem we present tight upper and lower bounds on the performance (g_k^n) of a k-of-n special majority classifier as a function of the classification performance of the individual members in the ensemble. Without loss of generality we label the individual classifiers in ascending order according to their performance. *i.e.* $g_1 \leq g_2 \leq ... \leq g_{n-1} \leq g_n$.

Theorem 1 *Given an ensemble of n classifiers $\{h_1, h_2, ..., h_n\}$ with respective performances $g_1 \leq g_2 \leq ... \leq g_{n-1} \leq g_n$ under domain distribution $\mathcal{D}$, The performance g_k^n of an k-of-n special majority classifier under the same distribution is bounded by:*

$$\max(g_k^n) = \min(1, \Sigma(k), \Sigma(k-1)\ldots, \Sigma(2), \Sigma(1)) \tag{4}$$
$$\min(g_k^n) = \max(0, \xi(k), \xi(k-1)\ldots, \xi(2), \xi(1)) \tag{5}$$

where

$$\Sigma(m) \equiv \frac{1}{m}\sum_{i=1}^{n-k+m} g_i \qquad m = 1\ldots k \tag{6}$$

$$\xi(m) \equiv \frac{1}{m}\sum_{i=k-m+1}^{n} g_i - \frac{n-k}{m} \qquad m = 1\ldots k \tag{7}$$

Proof (sketch)

- (Upper Bound)
  First we prove that the expression on the right side of Equation 4 is an upper bound. The performance of each classifier is defined by eq. 1 and may be written as
  $$g_i = \int \mathcal{I}_i(x)\rho(x)dx \tag{8}$$
  Where $\mathcal{I}_i(x)$ is the indicator function $[h_i(x) = c(x)]$ and $\rho(x)$ is the probability density function for distribution $\mathcal{D}$. We can write the following expression for g_k^n:
  $$g_k^n = \int_{\sum \mathcal{I}_i(x) \geq k} \rho(x)dx \tag{9}$$

We find that g_k^n is bounded by a series of inequalities. The first states that g_k^n can be no larger than unity: $g_k^n \leq 1$. The following inequality follows from the fact that at every point where the majority classifier is correct there are at least k individual classifiers correct at that point:

$$\begin{aligned} \sum_1^n g_i &= \sum_1^n \int \mathcal{I}_i(x)\rho(x)dx & (10) \\ &\geq \int_{\sum \mathcal{I}_i(x) \geq k} k\rho(x)dx = kg_k^n & (11) \end{aligned}$$

The resulting inequality is $g_k^n \leq \frac{1}{k}\sum_1^n g_i$. If we remove h_n, we can follow similar reasoning to prove the following inequality: $g_k^n \leq \frac{1}{k-1}\sum_1^{n-1} g_i$. Similarly can remove additional classifiers to achieve the following bound for any m between 1 and k:

$$g_k^n \leq \frac{1}{m}\sum_{i=1}^{n-k+m} g_i \equiv \Sigma_k^n(m) \tag{12}$$

We will write $\Sigma(m)$ without the modifiers while it is clear from context. Now, we can claim that g_k^n is bounded by the most restrictive of the above bounds:

$$g_k^n \leq \min(1, \Sigma(k), \Sigma(k-1)\ldots, \Sigma(2), \Sigma(1)) \tag{13}$$

This can be reformulated to explicitly specify the conditions under which each of the terms is minimal:

$$g_k^n \leq \begin{cases} 1 & \Sigma(k) \geq 1 \\ \Sigma(k) & \Sigma(k-1) \geq g_n \\ \Sigma(k-1) & \Sigma(k-2) \geq g_{n-1} \\ . & \\ . & \\ \Sigma(k-j) & \Sigma(k-j-1) \geq g_{n-j} \\ . & \\ . & \\ \Sigma(2) & \Sigma(1) \geq g_{n-k+2} \\ \Sigma(1) & \Sigma(1) < g_{n-k+2} \end{cases} \tag{14}$$

The above expression should be read in an *if-then-else* fashion: g_k^n is bounded by the left-side expression corresponding to the topmost right-side inequality which is true.

To complete the proof, we show that there exists a solution at the upper bound of Equation 4. This is done by induction on both n and k. The induction passes through values of k that are smaller than $n/2$ and do not correspond to an intuitive use of a majority vote classifier. However, g_k^n is well defined for these values and is neccesary for the induction.

The induction is performed by noting that Equation 14 has the following structure:

$$g_k^n \leq \begin{cases} 1 & \Sigma(k) \geq 1 \\ \Sigma(k) & \Sigma(k-1) \geq g_n \\ g_{k-1}^{n-1} & \Sigma(k-1) < g_n \end{cases} \tag{15}$$

The proof sketch for the three cases is the following:

The main case is the second. We construct a solution for a k-from-n classifier based on h_n and the inductive hypothesis' solutions for h_k^{n-1} and h_{k-1}^{n-1} constructed from the remainder of the ensemble.

The first case is based on the the first case when $\Sigma(k) = 1$. and the fact that increasing the performance of any classifier cannot decrease the upper bound.

The final case follows from removing h_n from the ensemble and noting that a $(k-1)$-of-$(n-1)$ majority-vote ensemble is a solution for this case.

The base cases for the induction, g_1^n and g_{n-1}^n, are easy to prove by direct construction.

- (Lower Bound)

  The lower bound can be easily calculated from the upper bound for binary concepts.The lower bound of the performance is met when the upper bound on the negated concept is achieved. Equation 4 becomes:

$$\max(1 - g_k^n) = \min(1, \Sigma(k), \ldots, \Sigma(1)) \tag{16}$$

$$\Sigma(m) \equiv \frac{1}{m}\sum_{i=n}^{k-m+1} (1 - g_i) \qquad m = 1\ldots k \tag{17}$$

Note that we have replaced all g's with $1-g$, and reversed the order of the indices i. After solving for g_k^n we arrive at Equation 5.

4 Examples & Ramifications

Example 2 (3-member ensemble $(g_1 \leq g_2 \leq g_3)$)

$$\max(g_2^3) = \begin{cases} 1 & \frac{1}{2}(g_1+g_2+g_3) \geq 1 \\ \frac{1}{2}(g_1+g_2+g_3) & g_1+g_2 \geq g_3 \\ g_1+g_2 & Otherwise \end{cases} \tag{18}$$

$$\min(g_2^3) = \begin{cases} g_2+g_3-1 & g_2+g_3-1 > g_1 \\ \frac{1}{2}(g_1+g_2+g_3-1) & (g_1+g_2+g_3) > 1 \\ 0 & Otherwise \end{cases} \tag{19}$$

Example 3 (n classifiers where $\forall i. g_i = g$)
In this case

$$\max(g_{(n+1)/2}^n) = \begin{cases} 1 & \frac{2n}{n+1}g \geq 1 \\ \frac{2n}{n+1}g & Otherwise \end{cases} \tag{20}$$

$$\min(g_{(n+1)/2}^n) = \begin{cases} \frac{2n}{n+1}g - \frac{n-1}{n+1} & \frac{2n}{n-1}g \geq 1 \\ 0 & Otherwise \end{cases} \tag{21}$$

This example leads to some interesting results:

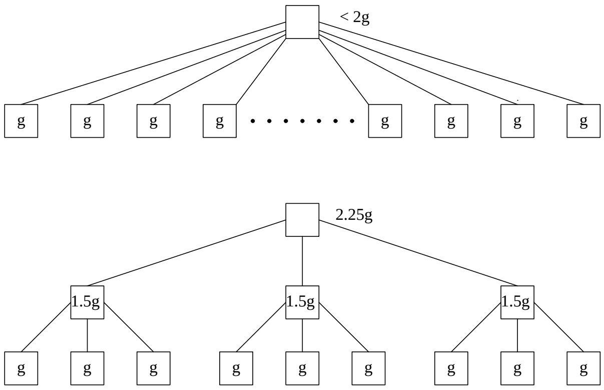

Figure 2: (Example 4) Combining an arbitrary number of classifiers of performance g can give no more than performance $2g$ using a flat majority-vote classifier. However, using a hierarchical structure, 9 classifiers can be combined to achieve a performance of $2.25g$.

- A majority ensemble can improve the performance by at most a factor of two. i.e. One needs over 50% performance to achieve arbitrary performance (this in accordance with the boosting results (Schapire 1990; Freund 1992; Freund & Schapire 1995).
- One can compute the lower bound on the number of ensemble members needed to achieve a given performance G :

$$n \geq \frac{G}{2g - G} \tag{22}$$

  If we follow the notation used in boosting literature where a weak learners' performance is $1/2 + \gamma$, the size of a majority classifier achieving performance of $1 - \epsilon$, must be at least:

$$n \geq \frac{1 - \epsilon}{2\gamma + \epsilon} \tag{23}$$

Example 4 (Hierarchical structures)
Using Theorem 1 we may analyze not only flat structures: Figure 2 demonstrates the fact that hierarchical structures are more expressive and gives the upper bound on the performance for such a structure.

5 Related work

Combining estimators has been studied both theoretically and experimentally. In (Hansen & Salamon 1990), the performance of an ensemble of classifiers is studied under a probabilistic modeling of the individual votes. Much of the work in combining ensembles has been done with combining regression classifiers. For example in (Perrone & Cooper 1993; Krogh & Vedelsby 1995) the model considered is one

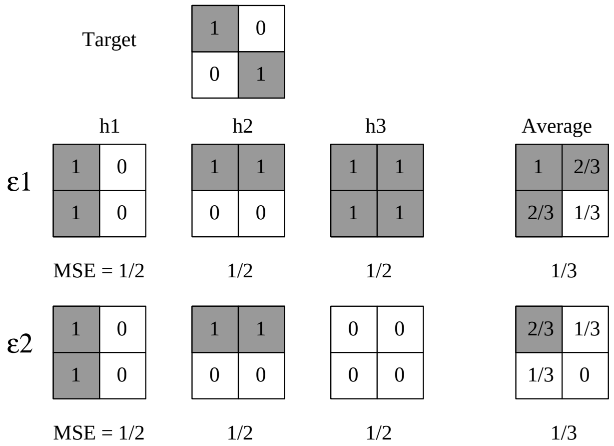

Figure 3: Combining $\mathcal{E}_1$ and $\mathcal{E}_2$ as regression classifiers.

where the ensemble's individual results are averaged:

$$h_{ens}(x) = \frac{1}{n}[h_1(x) + h_2(x) + \ldots + h_n(x)] \tag{24}$$

The most common performance measure used for regressors is the Mean Square Error(MSE) measure:

$$MSE(h) \equiv \int_{\mathcal{D}} (h(x) - c(x))^2 dx \tag{25}$$

It is well known that the MSE of an averaged ensemble is no larger than the average individual MSE:

$$MSE(h_{ens}) \leq \overline{MSE(h_i)} \tag{26}$$

This states that whenever one uses averaging one can improve over the average performance of a single learner. This seems a powerful result which may mislead one to believe that combining regression classifiers alway give better results. In fact, if the relevant measure is the classification performance (zero-one loss error) combining regressors may lead to inferior performance. To demonstrate, we return to the generalized XOR example.

Figure 3 shows both $\mathcal{E}_1$ and $\mathcal{E}_2$ combined as regressors. The classification behavior is the same as before. The MSE of the combined classifier is smaller than the average MSE. In fact the MSE for both is $\frac{1}{3}$. This demonstrates that the reduction in MSE for combining regressors does not necessarily bear on the whether the combination will result in superior performance.

6 Discussion

We have given upper and lower bounds on majority ensembles of classifiers as a function of the performance of the individual members. The results are best and worst case scenarios and not necessarily typical: for example, the results state that a majority ensemble of classifiers all performing at 90% should be able to achieve performance of 100%. Any practitioner knows that combining three 90% classifiers rarely changes the overall performance by much. In (Matan 1996) we give more realistic bounds by incorporating a measure of agreement, A, between the the members of the ensemble. For example, consider a majority classifier composed of three $g = 90\%$ classifiers which they are unanimous in their vote on 91% of the domain. We show that the bounds on the majority classifier h_2^3 are $87\% \leq g_2^3 \leq 93\%$.

The bounds are of interest to relatively weak classifiers and the representational analysis gives insight into the way one should combine weak representations into strong representations; One may argue that the boosting algorithm by Freund(Freund 1992) does exactly the right thing. It focuses new classifiers mainly on areas of the domain where only a slight effort is needed to gain majority. Areas that are all in agreement (correct or in error) are given smaller weight.

References

Freund, Y., and Schapire, R. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroColt-95*.

Freund, Y. 1992. An improved boosting algorithm and its implications on learning complexity. In *Proceedings of the 5th Workshop on Computational Learning theory*, 391–398. Morgan Kaufman.

Hansen, L. K., and Salamon, P. 1990. Neural network ensembles. *IEEE Trans. Pattern Anal. Machine Intell.* 12(10):993–10001.

Krogh, A., and Vedelsby, J. 1995. Neural network ensembles, cross validation and active learning. In Tesaro, G.; Touretzky, D.; and Alspector, J., eds., *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Matan, O. 1996. (to appear).

Perrone, M. P., and Cooper, L. N. 1993. When networks disagree: Ensemble methods for hybrid neural networks. In Mammone, R. J., ed., *Neural Networks for speech and Image Processing*. Chapman-Hall.

Schapire, R. E. 1990. The strength of weak learnability. *Machine Learning* 5:197–227.