ENSEMBLE CLASSIFICATION BY CRITIC-DRIVEN COMBINING

David J. Miller and Lian Yan

Department of Electrical Engineering
The Pennsylvania State University
University Park, Pa. 16802

ABSTRACT

We develop new rules for combining estimates obtained from each classifier in an ensemble. A variety of combination techniques have been previously suggested, including averaging probability estimates, as well as hard voting schemes. We introduce a *critic* associated with each classifier, whose objective is to predict the classifier's errors. Since the critic only tackles a two-class problem, its predictions are generally more reliable than those of the classifier, and thus can be used as the basis for our suggested improved combination rules. While previous techniques are only effective when the individual classifier error rate is $p < 0.5$, the new approach is successful, as proved under an independence assumption, *even when this condition is violated* – in particular, so long as $p + q < 1$, with q the critic's error rate. More generally, critic-driven combining achieves consistent, substantial performance improvement over alternative methods, on a number of benchmark data sets.

1. INTRODUCTION

In recent years, ensemble classification has emerged as one of the most promising and one of the most actively investigated classification paradigms, e.g. [4]. In this approach, probabilistic class estimates or decisions from each classifier in a collection are pooled, using some rule of combination, in order to make the ultimate decision for a new datum. Typically, individual estimates are combined through soft averaging, or hard voting. This paradigm has been motivated as an alternative to the more conventional "stand-alone" classification system by arguing that individual classifiers are often suboptimal and that combining estimates obtained from multiple classifiers can improve upon the stand-alone performance of any classifier in the collection [4]. The effectiveness of combining can be given some simple analytical justification for majority-based voting, assuming that the classifiers all make errors or correct decisions independent of each other. Under this assumption the number of correct voters is given by the binomial distribution. If the individual classifier error rate is $p < 0.5$, it is a known result (Condorcet's theorem) [1] that for odd (or even) number of classifiers (voters) N, the correct decision rate for the voting system increases with increasing N, going to one as $N \rightarrow \infty$[1]. Alternatively, if $p > 0.5$, the correct decision rate *decreases* with increasing N. Since the Bayes error rate (typically positive) represents the *actual* ultimate classification performance, one must conclude that the independence assumption, which predicts a lower bound of zero, cannot hold as $N \rightarrow \infty$. Moreover, it is difficult in practice to design classifiers that even approximately achieve independent errors, even for small N. Still, this simple analysis does provide some qualitative characterization of what can be hoped for in practice, *i.e.*, that performance should improve with increasing N. It also identifies important system properties (approximate classifier independence, $p < 0.5$) typically required for the success of majority-based voting schemes and believed required for ensemble classification methods in general[4]. The restriction $p < 0.5$ serves as particular motivation for the present work as, in the sequel, we will suggest new combination paradigms which actually allow *removal* of this requirement. We will thus significantly extend the applicability of ensemble classification techniques.

We can categorize classifier combination methods in terms of the training scenarios under which they can be applied. In the most ideal situation, there is a common training set available for the design of the entire ensemble system. In this case, there are two basic training strategies: i) separate design of the individual classifiers, followed by joint optimization of the structure which combines their decisions (which we will refer to as the *combiner*) based on either the training set e.g. [8] or a validation set; ii) joint optimization of the entire system.

While the above are strategies of choice given a common training set, there are situations where such training is either impossible or impractical. As one instance, consider the case where the combiner pools the decisions of proprietary classifiers produced by individual companies. To the designer of the combiner, each individual classifier is an immutable "black box", trained on the company's (inaccessible) data set in some unknown way. In this case, the classifiers are pre-trained and there is no common training set for optimizing the combiner. Thus, joint optimization is not possible.

The above circumstance necessitates schemes that combine the decisions of pre-trained classifiers. In this case, a common training set is not required; nor, in fact, is training of the combiner required. Here, we restrict consideration to such schemes and seek to develop new, improved combina-

This work was supported in part by the NSF under Career Award IRI-9624870. The authors would like to thank Moonseo Park for a valuable suggestion.

[1] This result is attributed to the Marquis de Condorcet (18th century) [1], who addressed the problem in a jury context.

tion rules. The difference between our approach and previous work on simple averaging and voting schemes is our use of a *critic*, specific to each classifier (expert), which evaluates expert opinions. Given the same input feature vector used by the expert, the critic tries to predict whether the expert's decisions are *valid* or *bogus*. The critic is trained after its expert, on the expert's training set, with supervision information indicating whether or not the expert's decisions agree with the true class labels – essentially, the critic is an expert *on* the expert. Since the expert is assumed to tackle a multiclass (> 2 class) problem and the critic only needs to solve a two-class problem, the critic's estimates should be more accurate than the expert's on average. The suggestion is then to incorporate the critic's opinion within the rule of combination, so as to achieve more reliable decisions. We next introduce several approaches.

2. CRITIC-DRIVEN COMBINATIONS

We propose both hard voting and soft averaging schemes. Consider classification of a feature vector $\underline{x} \in \mathcal{R}^d$ into one of C classes. Assume there are N experts.

2.1. Critic-driven Voting

We propose several critic-based extensions of standard voting methods: 1. Each expert votes for the class which it predicts if the critic assesses that the expert prediction is valid. Otherwise, the expert *abstains* from voting. The total number of votes (K) equals N minus the number of abstentions. All votes are given equal weight. If any class receives $l > \frac{K}{2}$ votes, then that class is the one predicted by the ensemble. Otherwise the datum is rejected. 2. An alternative rule is motivated by the following fact: the probability of correct decision for majority-based voting with N voters, N even, is *less than* the probability of correct decision with $N-1$ voters. This suggests a modification of the critic-based scheme just described wherein, when K is even, one voter is *dropped* prior to vote tallying. In practice, the voter with least confidence from its critic could be dropped. This modified scheme has two advantages over the former one. First, as one might expect, it achieves a greater correct decision rate. Second, it greatly simplifies the performance analysis of section 3. We also suggest a critic-driven version of plurality voting.

2.2. Critic-driven Averaging of Probabilities

In this case, each expert produces estimates of the *a posteriori* class probabilities, i.e. $P_e^{(j)}[k|\underline{x}]$, $k = 1, \ldots, C$ $j = 1, \ldots, N$, with "e" denoting expert. Each critic also produces probabilities $P_c^{(j)}[b|\underline{x}]$, where $b \in \{0, 1\}$, with "1" indicating a *valid* assessment and "0" a *bogus* assessment. Here "c" denotes the critic. A loosely stated objective for the combiner is to "agree with" expert probabilities, to the extent that they are valid, as estimated by the critic. Information theory suggests in this case use of a cross entropy (Kullback-Leibler distance) criterion – a measure of dissimilarity between probability mass functions that has been given axiomatic justification [7]. However, since cross entropy is an asymmetric cost, there are two possible objectives.

The Geometric Average Rule: If we view the expert probabilities as *priors*, then we will choose the combined probabilities $\{P[k|\underline{x}]\}$ as the posteriors minimizing the average cross entropy cost:

$$\sum_{j=1}^{N} w_j(\underline{x}) D(\{P[k|\underline{x}]\}||\{P_e^{(j)}[k|\underline{x}]\}). \tag{1}$$

Here, $D(\{P[k|\underline{x}]\}||\{P_e^{(j)}[k|\underline{x}]\}) \equiv \sum_{k=1}^{C} P[k|\underline{x}] \log\left(\frac{P[k|\underline{x}]}{P_e^{(j)}[k|\underline{x}]}\right)$, the standard cross entropy cost between pmfs, with the weighting function $w_j(\underline{x}) = \frac{P_c^{(j)}[1|\underline{x}]}{\sum_{l=1}^{N} P_c^{(l)}[1|\underline{x}]}$, a probabilistic measure of the critic's confidence in its expert[2]. After minimizing (1) over $\{P[k|\underline{x}]\}$ subject to constraints ensuring a pmf solution, we obtain the "geometric average" estimates:

$$P[k|\underline{x}] = \frac{\prod_{j=1}^{N} (P_e^{(j)}[k|\underline{x}])^{w_j(\underline{x})}}{\sum_{m=1}^{C} \prod_{j=1}^{N} (P_e^{(j)}[m|\underline{x}])^{w_j(\underline{x})}} \qquad k = 1, \ldots, C. \tag{2}$$

The chosen class is then the one with maximum *a posteriori* probability.

The Arithmetic Average Rule: Alternatively, we can interpret the experts as *posterior* probabilities and seek a *prior* probability agreeing with each posterior, to the extent that it is valid. In this case, we obtain the "arithmetic average" rule:

$$P[k|\underline{x}] = \sum_{j=1}^{N} w_j(\underline{x}) P_e^{(j)}[k|\underline{x}]. \tag{3}$$

Equation (3) is a generalization of simple averaging [2], and specializes to it with the choice $w_j(\underline{x}) = \frac{1}{N}$. Both (2) and (3) are effective schemes, with neither dominating the other in all cases. In particular, we have found that while (2) often achieves better results than (3), it may produce less reliable results when some experts give probabilities close to zero.

An Improved Rule: While both (2) and (3) outperform simple averaging, neither approach gleans all the information contained in the ensemble. In particular, if we consider the case where $P_c^{(j)}[1|\underline{x}] = 0$, we see that in both (2) and (3), expert j effectively abstains from contributing its estimates. However, a zero probability from a critic is actually quite informative – it reasonably indicates that the expert's predicted ("winning") class should be excluded. This suggests the following approach: conditioned on critic j's validation of its expert, the pmf $\tilde{P}_e^{(j)}[k|\underline{x}, b_j = 1] = P_e^{(j)}[k|\underline{x}]$ is posited; conditioned on the critic's rejection, a uniform pmf is posited over all classes excluding the expert's predicted winner:

$$\tilde{P}_e^{(j)}[k|\underline{x}, b_j = 0] = \begin{cases} \frac{1}{C-1} & \text{if } k \neq c^* \\ 0 & k = c^*, \end{cases} \tag{4}$$

[2]Other choices for the weights $w_j(\underline{x})$ that are monotonically increasing in $P_c^{(j)}[1|\underline{x}]$ have also been found to be effective.

where $c^* = \arg\max_c P_e^{(j)}[c|\underline{x}]$. Now, the average cross entropy cost sums over $2N$ terms and the resulting estimator, assuming experts as posteriors, is:

$$P[k|\underline{x}] = \sum_{j=1}^{N} \sum_{l=0}^{1} w_{jl}(\underline{x}) \tilde{P}_e^{(j)}[k|\underline{x}, l]. \quad (5)$$

3. ANALYSIS OF VOTING METHODS

Assume experts make independent errors, with common rate p. Then, for standard majority-based voting, the probability of l correct votes from a total of N experts is given by a binomial distribution. The correct decision rate for the majority-based scheme is then

$$P_{c_m}(N) = \sum_{l > \frac{N}{2}}^{N} \binom{N}{l} (1-p)^l p^{N-l}. \quad (6)$$

By Condorcet's theorem, if $p < 0.5$, P_{c_m} increases with odd(even) N[3]. Likewise, if $p > 0.5$, P_{c_m} *decreases* with increasing odd(even) N[1]. Now, consider critic-driven majority voting, wherein a voter is dropped, if required, to achieve an odd-sized voting ensemble. Assume critics make errors at common rate q, independent of their experts and independent of other experts and critics. For convenience, assume q is the rate both for false "valid" and false "bogus" critic decisions. The probability of M voters ($N - M$ abstentions) is binomial, i.e.

$$P_N[M] = \binom{N}{M} (pq+(1-p)(1-q))^M ((1-p)q+p(1-q))^{N-M}. \quad (7)$$

Further, the distribution over the number of correct voters is also binomial. Thus, the probability of correct decision given a critic-driven voting ensemble of size M is

$$P[\text{correct}|M] = \sum_{l > \frac{M}{2}}^{M} \binom{M}{l} (1-\tilde{p})^l \tilde{p}^{M-l}, \quad (8)$$

with $\tilde{p} = \frac{pq}{pq+(1-p)(1-q)}$. Averaging over all ensemble sizes, we get the overall critic-based correct decision rate:

$$P_{c_c}(N) = \sum_{M=1}^{N} P_N(M) P[\text{correct}|\text{odd}(M)]. \quad (9)$$

Here, odd(M) returns M if M is odd; else, it returns $M-1$. Similar to P_{c_m}, P_{c_c} increases with N for $p < 0.5$, assuming $q < p$. However, unlike P_{c_m}, if q is sufficiently small, then even if $p > 0.5$, P_{c_c} *still* increases with increasing N (where N now ranges over all positive integers). Essentially, the critics force abstentions so as to reduce the individual *voter's* error rate below 0.5, even when the expert rate p is above 0.5. In particular, we have the following result which shows that critics extend achievable voting performance:

Theorem: The critic-driven correct decision rate increases with increasing integer N if $p + q < 1$.
Proof : See [5].

In summary, we get a much less stringent condition for successful voting in the critic-driven case than in the standard majority-based case, under an independence assumption (as will be further confirmed experimentally).

4. EXPERIMENTAL RESULTS

We have evaluated the various combination schemes using radial basis functions (RBFs) [6] and decision trees [3] as the basic classifier structures. These structures were used to form both the experts and the critics. For a particular training/test split, we generated performance curves for several methods to demonstrate how the conventional approaches and the critic-driven ones fare for particular choices of p and, in the critic-based case, q.

In Figure 1, we compare simple averaging and the critic-driven "arithmetic averaging" for RBF-based classification of *glass*. The 214 sample data set was equally split into training and test sets. For simple averaging, we designed experts with 16 RBF components. For the critic-based scheme, we designed (expert,critic) pairs with (16,20) RBF components. The critic networks are actually less complex than the experts, since they only have two outputs, one per class. For this experiment, we observed $p \simeq 0.51$ and $q \simeq 0.47$. Thus, $p > 0.5$ and $p + q < 1$. Here, the analytical results of section 3.1 (extrapolated to the soft averaging case) are essentially validated. The trend for critic-based performance is a decrease in the error rate for increasing N, while there is no improvement (and some degradation) with increasing N for standard probability averaging. Moreover, the benefit of critic-based combining over averaging is substantial.

As a second example, we consider hard plurality voting on Deterding's *vowel* set. The 990 samples in this set were split into 525 training and 465 test. In this case, we used decision tree classifiers with both experts and critics consisting of 47 nodes. For this difficult example we observe $p = 0.66$ and $q = 0.46$, i.e. $p > 0.5$ and $p + q > 1$. Therefore, based on the analysis of section 3, we expect that the performance of both methods will degrade with increasing N. However, we see in Figure 2 that the standard voting error rate[4] stays roughly constant (even increasing a little) with increasing N, while the critic-driven rate *decreases* significantly. Moreover, critic-driven combining achieves a substantial performance advantage for increasing N. These results, which (in this case) prove the independent analysis pessimistic, can be explained from the standpoint of expert dependence. For the critic-based scheme, we can argue that even though $p + q > 1$, there must be regions of significant probability mass in the feature space where $p + q < 1$, and where the experts are roughly independent (thus allowing correct decision rate improvement for increasing N). Further, it may be the case that where $p + q > 1$, the experts are dependent – thus, the correct decision rate will not necessarily decrease for increasing N, even in regions where

[3] For N even, $P_{c_m}(N-2) < P_{c_m}(N) < P_{c_m}(N-1)$. Thus, while $P_{c_m} \to 1$ as $N \to \infty$, the performance curve is jagged, dipping on odd-to-even transitions.

[4] For hard majority and plurality voting, errors include rejections in our experiments.

$p + q > 1$. A corresponding argument can be applied to explain the standard voting performance.

We also give a third example of $p > 0.5$ for majority voting on the *yeast* data set, in Figure 3. The training/test split for this set was 742/742. We designed the CART classifiers and critics with 31 and 47 nodes, respectively. Here $p = 0.53$ and $q = 0.41$. Again the critic-driven trend is a decreasing error rate. Also, the performance is significantly better than the standard majority curve.

5. CONCLUSION

In this work we advanced the paradigm of critic-driven ensemble classification. Under an independence assumption, it was proved (in [5]) that critic-driven performance improves with increasing number of experts so long as $p+q < 1$, with p and q the expert and critic error rates. Moreover, the potential performance benefits relative to simple voting and averaging techniques were demonstrated on benchmark sets from the UC Irvine repository.

A coming paper [5] extends the work in this paper in several directions. First, we present more comprehensive experimental results which validate the critic-driven paradigm. Second, we develop novel combination rule for the case where critics are "weak", i.e., where critics do not condition on input features. Finally, while the analysis based on independence does provide insight, the inaccurancy of the independence assumption motivated us to develop an alternative analysis technique for predicting ensemble performance which incorporates prior knowledge on classifier dependence. This technique is based on maximum entropy statistical inference. The resulting predictions are more accurate than those assuming independence [5].

6. REFERENCES

[1] S. Berg. Condorcet's jury theorem, dependency among jurors. *Social Choice and Welfare*, 10:87–95, 1993.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[4] T. G. Dietterich. Machine-learning research: four current directions. *AI Magazine*, pages 97–136, Winter 1997.

[5] D. J. Miller and L. Yan. Critic-driven ensemble classification. (submitted to *IEEE Transactions on Signal Processing*), 1998.

[6] J. Moody and C. J. Darken. Fast learning in locally-tuned processing units. *Neural Computation*, 1:281–294, 1989.

[7] J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26:26–37, 1980.

[8] N. Ueda and R. Nakano. Combining discriminant-based classifiers using the minimum classification error discriminant. In *Neural Networks for Signal Processing*, pages 365–374, 1997.

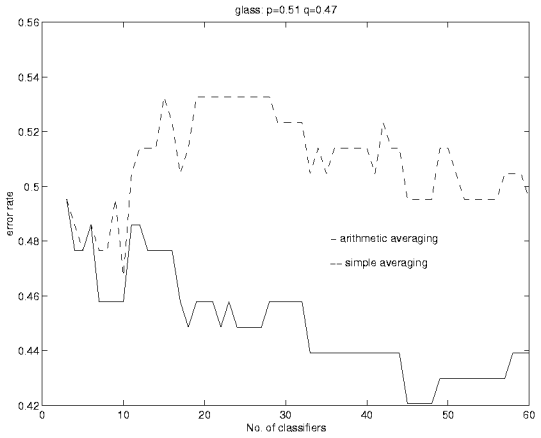

Figure 1: Error rates of RBF-based critic-driven arithmetic averaging and simple averaging for a single split of the *glass* data set.

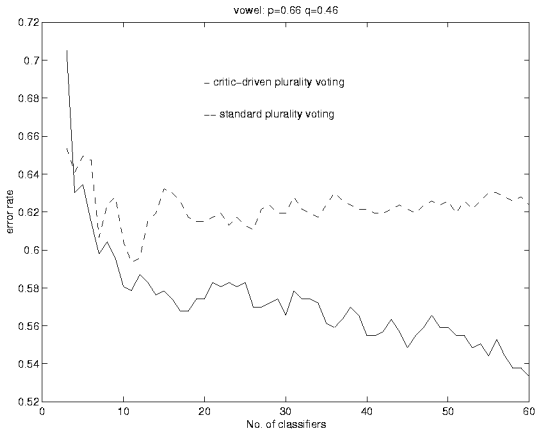

Figure 2: Error-plus-rejection rates of CART-based critic-driven plurality voting and standard plurality voting for a single split of the *vowel* data set.

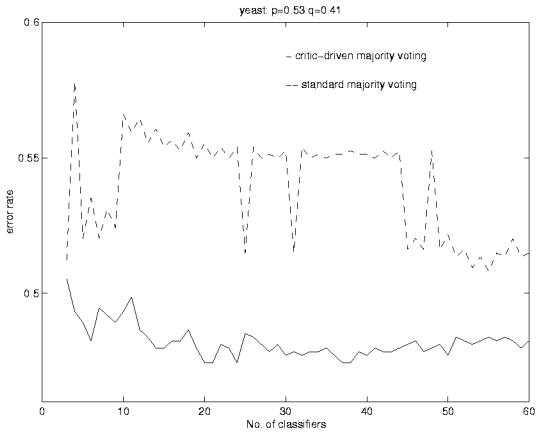

Figure 3: Error-plus-rejection rates of CART-based critic-driven majority voting and standard majority voting for a single split of the *yeast* data set.