Classifier Combining: Analytical Results and Implications

Kagan Tumer and **Joydeep Ghosh**
kagan@pine.ece.utexas.edu, ghosh@pine.ece.utexas.edu
Department of Electrical and Computer Engineering,
University of Texas, Austin, TX 78712-1084

Abstract

Several researchers have experimentally shown that substantial improvements can be obtained in difficult pattern recognition problems by combining or integrating the outputs of multiple classifiers. This paper summarizes our recent theoretical results that quantify the improvements due to multiple classifier combining. Furthermore, we present an extension of this theory that leads to an estimate of the Bayes error rate. Practical aspects such as expressing the confidences in decisions and determining the best data partition/classifier selection are also discussed.

Keywords: Linear combining, order statistics combining, Bayes error, error correlation, error reduction, ensemble networks, performance limits.

Introduction

Given infinite training data, consistent classifiers approximate the Bayesian decision boundaries to arbitrary precision, therefore providing similar generalizations (Geman, Bienenstock, & Doursat 1992). However, often only a limited portion of the pattern space is available or observable (Duda & Hart 1973; Fukunaga 1990). Given a finite and noisy data set, different classifiers typically provide different generalizations by realizing different decision boundaries (Ghosh & Tumer 1994). For example, when classification is performed using a multilayered, feed-forward artificial neural network, varying weight initializations, or network architectures (number of hidden units, hidden layers, node activation functions etc.) result in differences in performance. Therefore, when approaching a classification problem, it is necessary to train a multitude of networks to ensure that a good model/parameter set is found. However, selecting the "best" classifier at this point is not necessarily the ideal choice, since potentially valuable information may be wasted by discarding the results of less-successful classifiers.

This observation motivates the concept of "combining" wherein the outputs of all the available classifiers are pooled before a decision is made. This approach is particularly useful for difficult problems, such as those that involve a large amount of noise, limited number of training data, or unusually high dimensional patterns. The concept of combining has been studied in the connectionist framework in several forms, including "stacked generalization" (Wolpert 1992), and ensemble methods (Hampshire & Waibel 1992; Lincoln & Skrzypek 1990; Perrone & Cooper 1993; Tumer & Ghosh 1996c). Concepts such as "weighted averaging" (Benediktsson *et al.* 1994; Hashem & Schmeiser 1993; Jacobs 1995; Lincoln & Skrzypek 1990), "rank based combining" (Al-Ghoneim & Vijaya Kumar 1995; Ho, Hull, & Srihari 1994), "belief based combining" (Rogova 1994; Xu, Krzyzak, & Suen 1992; Yang & Singh 1994) have also been analyzed. Combining has also been studied in other fields such as econometrics (Granger 1989), and machine learning (Barnett 1981; Garvey, Lowrance, & Fischler 1981).

This paper summarizes our recent theoretical results that quantify the improvements due to multiple classifier combining. First, we summarize linear combiners, and provide expressions for the error rates. Then we discuss order statistics combiners, and analyze their properties. Thirdly, we apply the linear combining results to estimating the Bayes error rate. Then we present a data-based performance limit, followed by a method that assesses the confidence each classification decision carries. We also address the practical problem of reducing the correlations among classifiers for improved combining results. Finally, we discuss the implications of these results.

Linear Combining

Consider a reasonably well trained classifier, whose outputs approximate the corresponding *a posteriori* class probabilities (Richard & Lippmann 1991). The

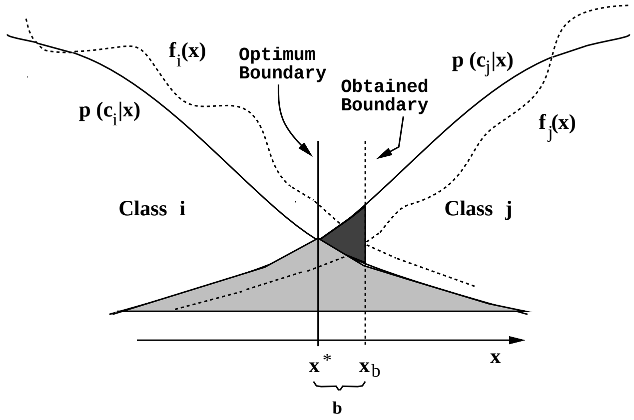

Figure 1: Error regions associated with approximating the *a posteriori* probabilities.

output of such a classifier can be written as:

$$f_i(x) = p(c_i|x) + \eta(x)\,, \tag{1}$$

where $p(c_i|x)$ is the *a posteriori* probability distribution of the ith class given input x, and $\eta(x)$ is the error associated with the ith output[1]. For the Bayes optimum decision, a vector x is assigned to class i if $p_i(x) > p_k(x)\,,\ \forall k \neq i$. Therefore, the Bayes boundary between classes i and j is given by $p_i(x^*) = p_j(x^*)$. Since classifier m provides $f_i^m(\cdot)$ instead of $p_i(\cdot)$, in general the decision boundary obtained, x_b, will vary from the optimum boundary (see Figure 1).

If N classifiers are available, the ith output of a simple averaging combiner provides an approximation to $p(c_i|x)$ given by:

$$f_i^{ave}(x) = \frac{1}{N} \sum_{m=1}^{N} f_i^m(x) \;=\; p(c_i|x) \;+\; \bar{\eta}_i(x)\,, \tag{2}$$

where:

$$\bar{\eta}_i(x) = \frac{1}{N} \sum_{m=1}^{N} \eta_i^m(x)\,.$$

Figure 1 shows the *a posteriori* probabilities obtained by a non-ideal classifier, and the associated added error region. The lightly shaded area provides the Bayesian error region. The darkly shaded area is the added error region, $A(b)$, associated with selecting a decision boundary, x_b, that is offset by b, since patterns corresponding to the darkly shaded region are erroneously assigned to class i by the classifier. The added error region can be obtained from the differences of the posteriors. Computing the expected value of this region leads to the expected added error or E_{add} given by:

$$E_{add} = \int_{-\infty}^{\infty} A(b) f_b(b) db, \tag{3}$$

where f_b is the density function for b. Although direct computation of this integral is not always possible, we show in (Tumer & Ghosh 1996a) how the added error region can be computed using the moments of the variance of the boundary offset, b. Furthermore, we show how the variance of the boundary offset obtained by a combiner is related to the variance of the boundary offset of a single classifier. Then, by coupling these results, we obtain (Tumer & Ghosh 1996c):

$$E_{add}^{ave} \;=\; \frac{1+\rho(N-1)}{N}\, E_{add}\,, \tag{4}$$

where ρ is the correlation among the classifier errors in estimating the *a posteriori* class probabilities. Equation 4 quantifies the often observed rule of thumb that combining relatively independent classifiers provides better gains. The effect of ρ on the combiner's error is readily apparent from this equation. If the errors are independent, then the second part of the reduction term vanishes and the combined error is reduced by N. If on the other hand, the error of each classifier has correlation 1, then the error of the combiner is equal to the initial errors and there is no improvement due to combining. In general, the correlation values lie between these two extremes, and some reduction is achieved. It is important to understand the interaction between N and ρ in order to maximize the reduction.

Order Statistics Combining

In this section we will discuss some basic concepts and properties of order statistics combiners. For a given input x, let the network outputs of each of the N classifiers for each class i be ordered in the following manner:

$$f_i^{1:N}(x) \le f_i^{2:N}(x) \le \;\cdots\; \le f_i^{N:N}(x).$$

The kth order statistic, which will be denoted $f_i^{k:N}(x)$, is the kth value in this progression. An excellent introduction to order statistics is provided in (David 1970). Now, let us define the *max*, *med* and *min* combiners as follows:

$$f_i^{max}(x) \;=\; f_i^{N:N}(x), \tag{5}$$

$$f_i^{med}(x) \;=\; \begin{cases} \dfrac{f_i^{\frac{N}{2}:N}(x) + f_i^{\frac{N}{2}+1:N}(x)}{2} & (N \text{ even}) \\ f_i^{\frac{N+1}{2}:N}(x) & (N \text{ odd}), \end{cases} \tag{6}$$

$$f_i^{min}(x) \;=\; f_i^{1:N}(x). \tag{7}$$

[1] In this article, we will assume that the classifiers are unbiased. The details for classifiers with bias can be found in (Tumer & Ghosh 1996c).

These three combiners are chosen because they represent important qualitative interpretations of the output space. Selecting the maximum combiner is equivalent to selecting the class with the highest posterior. Indeed, since the network outputs approximate the class *a posteriori* distributions, selecting the maximum reduces to selecting the classifier with the highest confidence in its decision. The drawback of this method however is that it can be compromised by a single classifier that repeatedly provides high values. The selection of the minimum combiner follows a similar logic, but focuses on classes that are unlikely to be correct, rather than on the correct class. Thus, this combiner eliminates less likely classes by basing the decision on the lowest value for a given class. This combiner suffers from the same ills as the *max* combiner, although it is less dependent on a single error, since it performs a min-max operation, rather than a max-max[2]. The median classifier on the other hand considers the most "typical" representation of each class. For highly noisy data, this combiner is more desirable than either the *min* or *max* combiners since the decision is not compromised as much by a single large error.

Combining N i.i.d. classifiers through order statistics leads to the following error reduction (Tumer & Ghosh 1996c):

$$E_{add}^{os} = \alpha \, E_{add}, \tag{8}$$

where α is a reduction factor that depends on both the order statistic chosen and the distribution of b. For most distributions, α can be found in tabulated form (Arnold, Balakrishnan, & Nagaraja 1992). For example, Table 1 provides α values for all three *os* combiners for a Gaussian distribution (Arnold, Balakrishnan, & Nagaraja 1992; Sarhan & Greenberg 1956).

Table 1: Reduction factors α for the *min, max* and *med* combiners for Gaussian error model.

N	OS Combiners	
	minimum/maximum	median
2	.682	.682
3	.560	.449
4	.492	.361
5	.448	.287
10	.344	.151
15	.301	.102
20	.276	.074

[2]Recall that the pattern is ultimately assigned to the class with the highest combined output.

Bayes Error Estimation

In many real world classification problems, generally, the classes are not fully separable. In such problems, it is unrealistic to expect perfect (100%) classification. The object of a statistical classification problem is to reach the "best possible" performance. The obvious question that arises of course, is how to determine the optimum classification rate. Since the Bayes decision provides the lowest error rates, the problem is equivalent to determining the Bayes error (Devijver & Kittler 1982; Schalkoff 1992; Young & Calvert 1974), which can be expressed as (Fukunaga 1990; Garber & Djouadi 1988):

$$E_{bay} \;=\; 1 - \sum_{i=1}^{L} \int_{C_i} P(c_i) p(x|c_i) dx \tag{9}$$

where C_i is the region where class i has the highest posterior, $P(c_i)$ is the *a priori* class probability of class i, $1 \leq i \leq L$, and $p(x|c_i)$ is the *likelihood* of class i.

Obtaining the Bayes error from Equation 9 entails evaluating the multi-dimensional integral of possibly unknown multivariate density functions over unspecified regions (C_i). Due to the difficulty of this operation, the Bayes error can be computed directly only for very simple problems, e.g., problems involving Gaussian class densities with identical covariances. For most real world problems, estimating the densities (e.g., through Parzen windows) and priors is the only available course of action. Although tedious, a numerical integration method can then be used to obtain the Bayes error. However, since errors are introduced both during the estimation of the class densities and the regions on which the classes are defined, and compounded by numerical integration, the results are in general not reliable.

Because direct computation of the Bayes error is often not possible, there are methods aimed at estimating or finding bounds for this error. For a 2-class problem for example, the Mahalanobis distance provides a distribution independent bound based on the means and covariances of the data (Devijver & Kittler 1982). Another bound for a 2-class problem can be obtained from the Bhattacharyya distance. However to obtain this bound, assumptions about the distributions are generally needed (Djouadi, Snorrason, & Garber 1990; Fukunaga 1990). Both of these bounds can be recursively extended to multi-class problems (Garber & Djouadi 1988). A non-parametric method based on the nearest neighbor classifier also provides also bounds for the Bayes error (Cover & Hart 1967; Fukunaga 1985).

An alternative method that estimates the Bayes error is based on the classifier combining strategy (Tumer

Table 2: Bayes Error Estimates for Artificial Data.

Data Set	Actual Bayes Error	Combiner Based Estimate	Nearest Neighbor Bounds	Mahalanobis Bound	Bhattacharyya Bounds
Data1	10.00	9.53	$8.17 \leq E_{bay} \leq 15.00$	≤ 18.95	$5.12 \leq E_{bay} \leq 22.04$
Data2	1.90	2.05	$2.46 \leq E_{bay} \leq 4.80$	≤ 14.13	$0.23 \leq E_{bay} \leq 4.74$

& Ghosh 1995a). As discussed above, the added error of a combiner can be related to the added error of a single classifier through Equation 4. Furthermore, the error of a classifier (E_{tot}) can be divided into the Bayes error and added error (E_{add}), which is the extra error due to the specific classifier (model/parameters) being used. Thus, the total error of a single classifier and the *ave* combiner are respectively given by:

$$E_{tot} = E_{bay} + E_{add} \; ; \tag{10}$$
$$E_{tot}^{ave} = E_{bay} + E_{add}^{ave}. \tag{11}$$

Note, that the Bayes error is not affected by the choice of the classifier. Therefore, by solving Equations 4, 10 and 11 simultaneously, the Bayes error can be isolated and evaluated. More precisely, solving these equations for E_{bay}, provides:

$$E_{bay} = \frac{N \, E_{tot}^{ave} - ((N-1)\rho \, + 1) \, E_{tot}}{(N-1)(1-\rho)}. \tag{12}$$

Equation 12 provides an estimate of the Bayes error as a function of the individual classifier error, the combined classifier error, the number of classifiers combined and the estimated correlation among them. Table 2 shows the combiner based Bayes error estimate, as well as some classical estimates for two well-known artificial problems detailed in (Fukunaga 1990; Tumer & Ghosh 1995a). Such experimental results suggest that the combiner based method provides reliable estimates of the Bayes error rate (Tumer & Ghosh 1995a).

The Plurality Limit

The previous section focused on estimating the Bayes error and was dependent on the properties of the *ave* combiner. In this section we present a "plurality limit" that is applicable to a larger variety of combiners (Tumer & Ghosh 1995a; 1995b). In combiners based on continuous classifier outputs, the decision of each individual classifier carries information that is not directly used in the combining step. Therefore, the classification decisions of individual classifiers can be used to determine the heuristics-based "plurality limit" which reflects the amount of separability that is present in the data.

In a general combining scheme with N classifiers, and L classes, a class will be called: *likely*, if at least two thirds of the classifiers have picked that class[3]; *unlikely*, if at most one third of the classifiers have picked that class; and *possible*, if it is neither likely, nor unlikely.

With these definitions, we specify four distinct types of patterns, namely, patterns where:

1. the correct class is possible or likely, and all the incorrect classes are unlikely;
2. the correct class is possible, but an incorrect class is also possible;
3. the correct class is unlikely, and all the incorrect classes are also unlikely;
4. The correct class is unlikely, but an incorrect class is possible or likely.

This characterization of the patterns allows an objective interpretation of the results[4]. In general, it is possible to correct some (or all) the errors occurring in patterns falling in the first three categories, because there is sufficient evidence pointing to the correct class. Errors caused by patterns in the fourth category, on the other hand, are in general not rectifiable. In these errors, most evidence points to a particular erroneous class[5]. Only rarely will it be possible for certain combiners to correct these errors. Therefore, the number of patterns in the fourth category provides a "plurality limit" on the number of errors after combining. In fact, combiners that rely directly on classification information (majority vote, plurality vote) cannot correct these errors, and for such classifiers, this value provides a lower bound on the attainable error. For other classifiers (e.g. simple averaging), the plurality limit can be reached under certain situations where a single correct decision can be strong enough to rectify the erroneous decisions of a larger number of classifiers.

[3] A likely class, does not necessarily imply a correct class.

[4] For the special cases where either $N = 2$ or $L = 2$, only some error categories are possible.

[5] This situation typically indicates an outlier or a mislabeled pattern.

Confidences in Decision

Another advantage of combining classifiers is that it provides confidence values for classification decisions. The Bayesian context does not address the confidence or reliability of a classification decision. For example, the distinction between the lack of information and information about equi-probability among classes is lost if only the *a posteriori* information is available, as both cases would be represented by L outputs of value $1/L$. It is often desirable to know the confidence the classifier has in a decision. If all patterns *have to* be classified, it is unnecessary to pursue the confidence issue further. However, if errors are extremely costly (e.g., certain medical applications), it is more desirable to classify only patterns where the decision is accompanied by a high confidence.

A method based on the agreements/disagreements between individual classifiers can be used to extract confidence measures along with classification decisions. Indeed, the number of classifiers that select a given class provides a direct indication of a combiner's confidence in a decision. By letting all decisions stand, one obtains the standard classification results. However, by requiring higher confidences in the decision, the misclassification rates can be reduced. For combining N classifiers, there are $N+1$ levels of confidences, associated with the number of classifiers that "agree" with the final decision[6]. While combining three classifiers, a confidence of 3 means that all classifiers agreed on the decision whereas a confidence of 2 means that only two classifiers agreed on the decision and so on. In this manner, by progressively increasing the level of confidence, it is possible to achieve a predetermined level of classification accuracy.

Correlation Reduction Methods

As discussed in previous sections, the combiner error rates depend on the error rates of the individual classifiers and the correlation among them. When it is possible to extract different types of feature sets (e.g., power spectrum based vs. auto-regression based features from a time signal) from the raw data, or to use radically different types of classifiers, one can avoid combining highly correlated classifiers. However, when only a single feature set is available and only one type of classifier (say, a multi-layered perceptron) provides satisfactory results, care must be taken during the combining process. In such a case, alternatives to combining N instances of the same classifier trained on the same training set are needed.

[6]Note that it is theoretically possible to have a 0-confidence decision, where the combiner provides a decision different than either of the individual classifiers.

Statistical methods such as cross-validation (Friedman 1994; Weiss & Kulikowski 1991), and bootstrapping (Efron 1982; Jain, Dubes, & Chen 1987) that aim at deducing the "true error" of a classifier can also be used to artificially reduce the correlation among classifiers. For example, in (Tumer & Ghosh 1996b), we discuss $k-1$-of-k training, a cross-validation like technique that partitions the training set into k subset, and trains k classifiers on $k-1$ of these subsets (Tumer & Ghosh 1996b). Since each classifier sees a different partition of the data, the correlation among them is reduced, and combining the k classifiers provides the final decision. An alternative to this method is found in bootstrapping and combining (Breiman 1994; Tumer & Ghosh 1996b). In this scheme training sets are generated by resampling the original set with substitution, leading to classifiers trained on different data partitions.

In our experiments we found that when the amount of data to which a classifier has access is reduced, the classification performance suffers. Therefore, data partitioning methods need to be used with care, and a balance between the correlation reduction and individual error rate increase must be struck. Only then will these methods improve upon combining multiple classifiers trained on all the available data (Tumer & Ghosh 1996b).

Discussion

Combining the outputs of several classifiers before making the classification decision, has led to improved performance in many applications (Ghosh *et al.* 1996; Xu, Krzyzak, & Suen 1992; Zhang, Mesirov, & Waltz 1992). This paper summarizes our recent results that: quantify such improvements; lead to the estimation of the Bayes error rate; express the decision confidences; and motivate practical correlation reduction methods. For more details, including extensive experimentation, see (Tumer & Ghosh 1995a; 1996b; 1996c).

For linear combiners, if the errors of individual networks are zero-mean i.i.d., the reduction in added error is shown to be N, the number of classifiers that are combined. If the errors are correlated, this reduction is expressed in terms of N and ρ, the correlation factor. Order statistics combiners are presented as an alternative to linear methods, and are motivated by their ability to extract the "right" amount of information. Error reduction terms similar to those of linear combiners are provided for this family of combiners.

A direct application of this theory leads to the estimation of the Bayes error rate. Determining this value is of paramount importance for difficult classification problems. Such information is helpful in determining whether it is worthwhile to try a different classifier, or a different set of parameters for the chosen classifier, in order to get better classification rates. For example, knowing that certain errors (e.g., mislabeled patterns) are too deeply rooted in the original data to be eliminated with better classifiers or combiners, may prevent a fruitless search for non-existent "better" classifiers.

Experimental results we obtained on a real-life oceanic data set with four distinct classes support the above conclusions (Tumer & Ghosh 1995a; 1996b; 1996c). Training on qualitatively different feature sets provides the lowest correlations among classifiers, and combining such classifiers results in the lowest error rates. In contrast using different types of classifiers only provides marginal improvements over multiple runs of one type of classifier. Furthermore, when multiple runs of a single classifier trained on a single feature set are combined, data partitioning techniques provide modest improvements. Also, error rates can be significantly reduced if only high confidence decisions are accepted.

One final note that needs to be considered is the behavior of combiners for a large number of classifiers (N). Clearly, the errors cannot be arbitrarily reduced by increasing N indefinitely. This observation however, does not contradict the results presented in this analysis. As N increases, it becomes difficult to find lowly correlated classifiers, thus reducing the improvements due to each extra classifier. This has been recently observed by some researchers, such as Breiman, who developed bootstrap methods for achieving independence among estimators (Breiman 1993; 1994), and by Jacobs (Jacobs 1995).

Acknowledgements: This research was supported in part by AFOSR contract F49620-93-1-0307, NSF grant ECS 9307632, and ARO contracts DAAH 04-94-G0417 and 04-95-10494.

References

Al-Ghoneim, K., and Vijaya Kumar, B. V. K. 1995. Learning ranks with neural networks (Invited paper). In *Applications and Science of Artificial Neural Networks, Proceedings of the SPIE*, volume 2492, 446–464.

Arnold, B.; Balakrishnan, N.; and Nagaraja, H. 1992. *A First Course in Order Statistics*. New York: Wiley.

Barnett, J. 1981. Computational methods for a mathematical theory of evidence. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, 868–875.

Benediktsson, J.; Sveinsson, J.; Ersoy, O.; and Swain, P. 1994. Parallel consensual neural networks with optimally weighted outputs. In *Proceedings of the World Congress on Neural Networks*, III:129–137. INNS Press.

Breiman, L. 1993. Stacked regression. Technical Report 367, Department of Statistics, University of California, Berkeley.

Breiman, L. 1994. Bagging predictors. Technical Report 421, Department of Statistics, University of California, Berkeley.

Cover, T. M., and Hart, P. E. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13:21–27.

David, H. A. 1970. *Order Statistics*. New York: Wiley.

Devijver, P., and Kittler, J. 1982. *Pattern Recognition: A Statistical Approach*. Prentice-Hall.

Djouadi, A.; Snorrason, O.; and Garber, F. D. 1990. The quality of training-sample estimates of the Battacharyya coefficient. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(1):92–97.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. Wiley, New York.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: SIAM.

Friedman, J. H. 1994. An overview of predictive learning and function approximation. In Cherkassky, V.; Friedman, J.; and Wechsler, H., eds., *From Statistics to Neural Networks, Proc. NATO/ASI Workshop*, 1–55. Springer Verlag.

Fukunaga, K. 1985. The estimation of the Bayes error by the k-nearest neighbor approach. In Kanal, L. N., and Rosenfeld, A., eds., *Progress in Pattern Recognition 2*. Amsterdam: North-Holland. 169–187.

Fukunaga, K. 1990. *Introduction to Statistical Pattern Recognition*. (2nd Ed.), Academic Press.

Garber, F., and Djouadi, A. 1988. Bounds on the Bayes classification error based on pairwise risk functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10(3):281–288.

Garvey, T.; Lowrance, J.; and Fischler, M. 1981. An inference technique for integrating knowledge from disparate sources. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, 319–325.

Geman, S.; Bienenstock, E.; and Doursat, R. 1992. Neural networks and the bias/variance dilemma. *Neural Computation* 4(1):1–58.

Ghosh, J., and Tumer, K. 1994. Structural adaptation and generalization in supervised feedforward networks. *Journal of Artificial Neural Networks* 1(4):431–458.

Ghosh, J.; Tumer, K.; Beck, S.; and Deuser, L. 1996. Integration of neural classifiers for passive sonar signals. In Leondes, C., ed., *DSP Theory and Applications*. Academic Press. (to appear).

Granger, C. W. J. 1989. Combining forecasts—twenty years later. *Journal of Forecasting* 8(3):167–173.

Hampshire, J., and Waibel, A. 1992. The Meta-Pi network: Building distributed representations for robust multisource pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14(7):751–769.

Hashem, S., and Schmeiser, B. 1993. Approximating a function and its derivatives using MSE-optimal linear combinations of trained feedforward neural networks. In *Proceedings of the Joint Conference on Neural Networks*, volume 87, I:617–620.

Ho, T. K.; Hull, J. J.; and Srihari, S. N. 1994. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(1):66–76.

Jacobs, R. 1995. Method for combining experts' probability assessments. *Neural Computation* 7(5):867–888.

Jain, A.; Dubes, R.; and Chen, C. 1987. Bootstrap techniques for error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 9:628–633.

Lincoln, W., and Skrzypek, J. 1990. Synergy of clustering multiple back propagation networks. In Touretzky, D., ed., *Advances in Neural Information Processing Systems-2*, 650–657. Morgan Kaufmann.

Perrone, M., 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*. Chapmann-Hall. chapter 10.

Richard, M., and Lippmann, R. 1991. Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation* 3(4):461–483.

Rogova, G. 1994. Combining the results of several neural network classifiers. *Neural Networks* 7(5):777–781.

Sarhan, A. E., and Greenberg, B. G. 1956. Estimation of location and scale parameters by order statistics from singly and doubly censored samples. *Annals of Mathematical Statistics Science* 27:427–451.

Schalkoff, R. 1992. *Pattern Recognition: Statistical, Structural and Neural Approaches*. Wiley.

Tumer, K., and Ghosh, J. 1995a. Bayes error rate estimation through classifier combining. Technical Report 96-01-101, The Computer and Vision Research Center, University of Texas, Austin.

Tumer, K., and Ghosh, J. 1995b. Limits to performance gains in combined neural classifiers. In *Proceedings of the Artificial Neural Networks in Engineering '95*, 419–424.

Tumer, K., and Ghosh, J. 1996a. Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition* 29(2):341–348.

Tumer, K., and Ghosh, J. 1996b. Error correlation and error reduction in ensemble classifiers. *Conection Science*. (Submitted for publication.).

Tumer, K., and Ghosh, J. 1996c. Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. *IEEE Transactions on Neural Networks*. (Accepted for publication. Preprints can be obtained from URL *http://www.lans.ece.utexas.edu/~kagan/publications.html*).

Weiss, S. M., and Kulikowski, C. 1991. *Computer Systems That Learn*. Morgan Kaufmann.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* 5:241–259.

Xu, L.; Krzyzak, A.; and Suen, C. Y. 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics* 22(3):418–435.

Yang, J.-B., and Singh, M. G. 1994. An evidential reasoning approach for multiple-attribute decision making with uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics* 24(1):1–19.

Young, T. Y., and Calvert, T. W. 1974. *Classification, Estimation and Pattern Recognition*. New York: Elsevier.

Zhang, X.; Mesirov, J.; and Waltz, D. 1992. Hybrid system for protein secondary structure prediction. *J. Molecular Biology* 225:1049–63.