Neural Networks for Classification: A Survey

Article *in* IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews) · December 2000

DOI: 10.1109/5326.897072 · Source: IEEE Xplore

1 author:

Peter G. Zhang
Georgia State University
73 PUBLICATIONS **16,819** CITATIONS

Neural Networks for Classification: A Survey

Guoqiang Peter Zhang

***Abstract*—Classification is one of the most active research and application areas of neural networks. The literature is vast and growing. This paper summarizes the some of the most important developments in neural network classification research. Specifically, the issues of posterior probability estimation, the link between neural and conventional classifiers, learning and generalization tradeoff in classification, the feature variable selection, as well as the effect of misclassification costs are examined. Our purpose is to provide a synthesis of the published research in this area and stimulate further research interests and efforts in the identified topics.**

***Index Terms*—Bayesian classifier, classification, ensemble methods, feature variable selection, learning and generalization, misclassification costs, neural networks.**

I. Introduction

CLASSIFICATION is one of the most frequently encountered decision making tasks of human activity. A classification problem occurs when an object needs to be assigned into a predefined group or class based on a number of observed attributes related to that object. Many problems in business, science, industry, and medicine can be treated as classification problems. Examples include bankruptcy prediction, credit scoring, medical diagnosis, quality control, handwritten character recognition, and speech recognition.

Traditional statistical classification procedures such as discriminant analysis are built on the Bayesian decision theory [42]. In these procedures, an underlying probability model must be assumed in order to calculate the posterior probability upon which the classification decision is made. One major limitation of the statistical models is that they work well only when the underlying assumptions are satisfied. The effectiveness of these methods depends to a large extent on the various assumptions or conditions under which the models are developed. Users must have a good knowledge of both data properties and model capabilities before the models can be successfully applied.

Neural networks have emerged as an important tool for classification. The recent vast research activities in neural classification have established that neural networks are a promising alternative to various conventional classification methods. The advantage of neural networks lies in the following theoretical aspects. First, neural networks are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model. Second, they are universal functional approximators in that neural networks can approximate any function with arbitrary accuracy [37], [78], [79]. Since any classification procedure seeks a functional relationship between the group membership and the attributes of the object, accurate identification of this underlying function is doubtlessly important. Third, neural networks are nonlinear models, which makes them flexible in modeling real world complex relationships. Finally, neural networks are able to estimate the posterior probabilities, which provides the basis for establishing classification rule and performing statistical analysis [138].

On the other hand, the effectiveness of neural network classification has been tested empirically. Neural networks have been successfully applied to a variety of real world classification tasks in industry, business and science [186]. Applications include bankruptcy prediction [2], [96], [101], [167], [187], [195], handwriting recognition [61], [92], [98], [100], [113], speech recognition [25], [106], product inspection [97], [130], fault detection [11], [80], medical diagnosis [19], [20], [30], [31], and bond rating [44], [163], [174]. A number of performance comparisons between neural and conventional classifiers have been made by many studies [36], [82], [115]. In addition, several computer experimental evaluations of neural networks for classification problems have been conducted under a variety of conditions [127], [161].

Although significant progress has been made in classification related areas of neural networks, a number of issues in applying neural networks still remain and have not been solved successfully or completely. In this paper, some theoretical as well as empirical issues of neural networks are reviewed and discussed. The vast research topics and extensive literature makes it impossible for one review to cover all of the work in the filed. This review aims to provide a summary of the most important advances in neural network classification. The current research status and issues as well as the future research opportunities are also discussed. Although many types of neural networks can be used for classification purposes [105], our focus nonetheless is on the feedforward multilayer networks or multilayer perceptrons (MLPs) which are the most widely studied and used neural network classifiers. Most of the issues discussed in the paper can also apply to other neural network models.

The overall organization of the paper is as follows. After the introduction, we present fundamental issues of neural classification in Section II, including the Bayesian classification theory, the role of posterior probability in classification, posterior probability estimation via neural networks, and the relationships between neural networks and the conventional classifiers. Section III examines theoretical issues of learning and generalization in classification as well as various practical approaches to improving neural classifier performance in learning and generalization. Feature variable selection and the effect of misclassification costs—two important problems unique to classification

Manuscript received July 28, 1999; revised July 6, 2000.

G. P. Zhang is with the J. Mack Robinson College of Business, Georgia State University, Atlanta, GA 30303 USA (e-mail: gpzhang@gsu.edu).

Publisher Item Identifier S 1094-6977(00)11206-4.

problems—are discussed in Sections IV and V, respectively. Finally, Section VI concludes the paper.

II. Neural Networks and Traditional Classifiers

A. Bayesian Classification Theory

Bayesian decision theory is the basis of statistical classification methods [42]. It provides the fundamental probability model for well-known classification procedures such as the statistical discriminant analysis.

Consider a general M-group classification problem in which each object has an associated attribute vector $\mathbf{x}$ of d dimensions. Let ω denote the membership variable that takes a value of ω_j if an object is belong to group j. Define $P(\omega_j)$ as the prior probability of group j and $f(\mathbf{x}\,|\,\omega_j)$ as the probability density function. According to the Bayes rule

$$P(\omega_j\,|\,\mathbf{x}) = \frac{f(\mathbf{x}\,|\,\omega_j)P(\omega_j)}{f(\mathbf{x})} \tag{1}$$

where $P(\omega_j\,|\,\mathbf{x})$ is the posterior probability of group j and $f(\mathbf{x})$ is the probability density function: $f(\mathbf{x}) = \sum_{j=1}^{M} f(\mathbf{x}\,|\,\omega_j)P(\omega_j)$.

Now suppose that an object with a particular feature vector $\mathbf{x}$ is observed and a decision is to be made about its group membership. The probability of classification error is

$$\begin{aligned} P(\text{Error}\,|\,\mathbf{x}) &= \sum_{i\neq j} P(\omega_i\,|\,\mathbf{x}) \\ &= 1 - P(\omega_j\,|\,\mathbf{x}) \quad \text{if we decide } \omega_j. \end{aligned}$$

Hence if the purpose is to minimize the probability of total classification error (misclassification rate), then we have the following widely used Bayesian classification rule

$$\text{Decide } \omega_k \text{ for } \mathbf{x} \text{ if } P(\omega_k\,|\,\mathbf{x}) = \max_{i=1,2,\ldots,M} P(\omega_i\,|\,\mathbf{x}). \tag{2}$$

This simple rule is the basis for other statistical classifiers. For example, linear and quadratic discriminant functions can be derived with the assumption of the multivariate normal distribution for the conditional density $f(\mathbf{x}\,|\,\omega_j)$ of attribute vector $\mathbf{x}$. There are two problems in applying the simple Bayes decision rule (2). First, in most practical situations, the density functions are not known or can not be assumed to be normal and therefore the posterior probabilities can not be determined directly. Second, by using (2), the decision goal is simply to minimize the probability of misclassifying a new object. In this way, we are indifferent with regard to the consequences of misclassification errors. In other words, we assume that the misclassification costs for different groups are equal. This may not be the case for many real world applications where the cost of a wrong assignment is quite different for different groups.

If we can assign a cost to a misclassification error, we may use that information to improve our decision. Let $c_{ij}(\mathbf{x})$ be the cost of misclassifying $\mathbf{x}$ to group i when it actually belongs to group j. The expected cost associated with assigning $\mathbf{x}$ to group i is

$$C_i(\mathbf{x}) = \sum_{j=1}^{M} c_{ij}(\mathbf{x})P(\omega_j\,|\,\mathbf{x}), \quad i = 1, 2, \ldots, M. \tag{4}$$

$C_i(\mathbf{x})$ is also known as the conditional risk function. The optimal Bayesian decision rule that minimizes the overall expected cost is

$$\text{Decide } \omega_k \text{ for } \mathbf{x} \text{ if } C_k(\mathbf{x}) = \min_{i=1,2,\ldots,M} C_i(\mathbf{x}). \tag{5}$$

When the misclassification costs are equal (0–1 cost function), then we have the special case (2) of the Bayesian classification rule. Note the role of posterior probabilities in the decision rules (2) and (5).

From (1) and (4) and note that the denominator is common to all classes, Bayesian decision rule (5) is equivalent to: Decide ω_k for $\mathbf{x}$ if $\sum_{i=1}^{M} c_{ik}(\mathbf{x})P(\omega_i)f(x\,|\,\omega_i)$ is the minimum. Consider the special two-group case with two classes of ω_1 and ω_2. We should assign $\mathbf{x}$ to class 1 if

$$c_{21}(\mathbf{x})P(\omega_2)f(\mathbf{x}\,|\,\omega_2) < c_{12}(\mathbf{x})P(\omega_1)f(\mathbf{x}\,|\,\omega_1)$$

or

$$\frac{f(x\,|\,\omega_1)}{f(x\,|\,\omega_2)} > \frac{c_{21}(x)P(\omega_2)}{c_{12}(x)P(\omega_1)}. \tag{6}$$

Expression (6) shows the interaction of prior probabilities and misclassification cost in defining the classification rule, which can be exploited in building practical classification models to alleviate the difficulty in estimation of misclassification costs.

B. Posterior Probability Estimation via Neural Networks

In classification problems, neural networks provide direct estimation of the posterior probabilities [58], [138], [156], [178]. The importance of this capability is summarized by Richard and Lippmann [138]:

"Interpretation of network outputs as Bayesian probabilities allows outputs from multiple networks to be combined for higher level decision making, simplifies creation of rejection thresholds, makes it possible to compensate for difference between pattern class probabilities in training and test data, allows output to be used to minimize alternative risk functions, and suggests alternative measures of network performance."

A neural network for a classification problem can be viewed as a mapping function, $F : R^d \rightarrow R^M$, where d-dimensional input $\mathbf{x}$ is submitted to the network and an M-vectored network output $\mathbf{y}$ is obtained to make the classification decision. The network is typically built such that an overall error measure such as the mean squared errors (MSE) is minimized. From the famous least squares estimation theory in statistics (see [126]), the mapping function $F : \mathbf{x} \rightarrow \mathbf{y}$ which minimizes the expected squared error

$$E[\mathbf{y} - F(\mathbf{x})]^2 \tag{7}$$

is the conditional expectation of $\mathbf{y}$ given $\mathbf{x}$

$$F(\mathbf{x}) = E[\mathbf{y}\,|\,\mathbf{x}]. \tag{8}$$

In the classification problem, the desired output $\mathbf{y}$ is a vector of binary values and is the jth basis vector

$e_j = (0,\ldots,0,1,0,\ldots,0)^t$ if $\mathbf{x} \in$ group j. Hence the jth element of $F(\mathbf{x})$ is given by

$$\begin{aligned} F_j(\mathbf{x}) &= E[y_j \,|\, \mathbf{x}] \\ &= 1 \cdot P(y_j = 1 \,|\, \mathbf{x}) + 0 \cdot P(y_j = 0 \,|\, \mathbf{x}) \\ &= P(y_j = 1 \,|\, \mathbf{x}) \\ &= P(\omega_j \,|\, \mathbf{x}). \end{aligned} \tag{9}$$

That is, the least squares estimate for the mapping function in a classification problem is exactly the posterior probability.

Neural networks are universal approximators [37] and in theory can approximate any function arbitrarily closely. However, the mapping function represented by a network is not perfect due to the local minima problem, suboptimal network architecture and the finite sample data in neural network training. Therefore, it is clear that neural networks actually provide estimates of the posterior probabilities.

The mean squared error function (7) can be derived [143], [83] as

$$\begin{aligned} \text{MSE} = &\sum_{j=1}^{M} \int_{R^d} [F_j(\mathbf{x}) - P(\omega_j \,|\, \mathbf{x})]^2 f(\mathbf{x})\, d\mathbf{x} \\ &+ \sum_{j=1}^{M} \int_{R^d} P(\omega_j \,|\, \mathbf{x})(1 - P(\omega_j \,|\, \mathbf{x})) f(\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{10}$$

The second term of the right-hand side is called the approximation error [14] and is independent of neural networks. It reflects the inherent irreducible error due to randomness of the data. The first term termed as the estimation error is affected by the effectiveness of neural network mapping. Theoretically speaking, it may need a large network as well as large sample data in order to get satisfactory approximation. For example, Funahashi [53] shows that for the two-group d-dimensional Gaussian classification problem, neural networks with at least $2d$ hidden nodes have the capability to approximate the posterior probability with arbitrary accuracy when infinite data is available and the training proceeds ideally. Empirically, it is found that sample size is critical in learning but the number of hidden nodes may not be so important [83], [138].

That the outputs of neural networks are least square estimates of the Bayesian *a posteriori* probabilities is also valid for other types of cost or error function such as the cross entropy function [63], [138]. The cross entropy function can be a more appropriate criterion than the squared error cost function in training neural networks for classification problems because of their binary output characteristic [144]. Improved performance and reduced training time have been reported with the cross entropy function [75], [77]. Miyake and Kanaya [116] show that neural networks trained with a generalized mean-squared error objective function can yield the optimal Bayes rule.

C. Neural Networks and Conventional Classifiers

Statistical pattern classifiers are based on the Bayes decision theory in which posterior probabilities play a central role. The fact that neural networks can in fact provide estimates of posterior probability implicitly establishes the link between neural networks and statistical classifiers. The direct comparison between them may not be possible since neural networks are nonlinear model-free method while statistical methods are basically linear and model based.

By appropriate coding of the desired output membership values, we may let neural networks directly model some discriminant functions. For example, in a two-group classification problem, if the desired output is coded as 1 if the object is from class 1 and -1 if it is from class 2. Then, from (9) the neural network estimates the following discriminant function:

$$g(\mathbf{x}) = P(\omega_1 \,|\, \mathbf{x}) - P(\omega_2 \,|\, \mathbf{x}). \tag{11}$$

The discriminating rule is simply: assign $\mathbf{x}$ to ω_1 if $g(\mathbf{x}) > 0$ or ω_2 if $g(\mathbf{x}) < 0$. Any monotone increasing function of the posterior probability can be used to replace the posterior probability in (11) to form a different discriminant function but essentially the same classification rule.

As the statistical counterpart of neural networks, discriminant analysis is a well-known supervised classifier. Gallinari *et al.* [54] describe a general framework to establish the link between discriminant analysis and neural network models. They find that in quite general conditions the hidden layers of an MLP project the input data onto different clusters in a way that these clusters can be further aggregated into different classes. For linear MLPs, the projection performed by the hidden layer is shown theoretically equivalent to the linear discriminant analysis. The nonlinear MLPs, on the other hand, have been demonstrated through experiments the capability in performing more powerful nonlinear discriminant analysis. Their work helps understand the underlying function and behavior of the hidden layer for classification problems and also explains why the neural networks in principle can provide superior performance over linear discriminant analysis. The discriminant feature extraction by the network with nonlinear hidden nodes has also been demonstrated in Asoh and Otsu [6] and Webb and Lowe [181]. Lim, Alder and Hadingham [103] show that neural networks can perform quadratic discriminant analysis.

Raudys [134], [135] presents a detailed analysis of nonlinear single layer perceptron (SLP). He shows that during the adaptive training process of SLP, by purposefully controlling the SLP classifier complexity through adjusting the target values, learning-steps, number of iterations and using regularization terms, the decision boundaries of SLP classifiers are equivalent or close to those of seven statistical classifiers. These statistical classifiers include the Enclidean distance classifier, the Fisher linear discriminant function, the Fisher linear discriminant function with pseudo-inversion of the covariance matrix, the generalized Fisher linear discriminant function, the regularized linear discriminant analysis, the minimum empirical error classifier, and the maximum margin classifier [134]. Kanaya and Miyake [88] and Miyake and Kanaya [116] also illustrate theoretically and empirically the link between neural networks and the optimal Bayes rule in statistical decision problems.

Logistic regression is another important classification tool. In fact, it is a standard statistical approach used in medical diagnosis and epidemiologic studies [91]. Logistic regression is often preferred over discriminant analysis in practice [65],

[132]. In addition, the model can be interpreted as posterior probability or odds ratio. It is a simple fact that when the logistic transfer function is used for the output nodes, simple neural networks without hidden layers are identical to logistic regression models. Another connection is that the maximum likelihood function of logistic regression is essentially the cross-entropy cost function which is often used in training neural network classifiers. Schumacher *et al.* [149] make a detailed comparison between neural networks and logistic regression. They find that the added modeling flexibility of neural networks due to hidden layers does not automatically guarantee their superiority over logistic regression because of the possible overfitting and other inherent problems with neural networks [176].

Links between neural and other conventional classifiers have been illustrated by [32], [33], [74], [139], [140], [151], [175]. Ripley [139], [140] empirically compares neural networks with various classifiers such as classification tree, projection pursuit regression, linear vector quantization, multivariate adaptive regression splines and nearest neighbor methods.

A large number of studies have been devoted to empirical comparisons between neural and conventional classifiers. The most comprehensive one can be found in Michie *et al.* [115] which reports a large-scale comparative study—the StatLog project. In this project, three general classification approaches of neural networks, statistical classifiers and machine learning with 23 methods are compared using more than 20 different real data sets. Their general conclusion is that no single classifier is the best for all data sets although the feedforward neural networks do have good performance over a wide range of problems. Neural networks have also been compared with decision trees [28], [36], [66], [104], [155], discriminant analysis [36], [127], [146], [161], [193], CART [7], [40], k-nearest-neighbor [82], [127], and linear programming method [127].

III. Learning and Generalization

Learning and generalization is perhaps the most important topic in neural network research [3], [18], [157], [185]. Learning is the ability to approximate the underlying behavior adaptively from the training data while generalization is the ability to predict well beyond the training data. Powerful data fitting or function approximation capability of neural networks also makes them susceptible to the overfitting problem. The symptom of an overfitting model is that it fits the training sample very well but has poor generalization capability when used for prediction purposes. Generalization is a more desirable and critical feature because the most common use of a classifier is to make good prediction on new or unknown objects. A number of practical network design issues related to learning and generalization include network size, sample size, model selection, and feature selection. Wolpert [188] addresses most of these issues of learning and generalization within a general Bayesian framework.

In general, a simple or inflexible model such as a linear classifier may not have the power to learn enough the underlying relationship and hence underfit the data. On the other hand, complex flexible models such as neural networks tend to overfit the data and cause the model unstable when extrapolating. It is clear that both underfitting and overfitting will affect generalization capability of a model. Therefore a model should be built in such a way that only the underlying systematic pattern of the population is learned and represented by the model.

The underfitting and overfitting phenomena in many data modeling procedures can be well analyzed through the well-known bias-plus-variance decomposition of the prediction error. In this section, the basic concepts of bias and variance as well as their connection to neural network classifiers are discussed. Then the methods to improve learning and generalization ability through bias and/or variance reductions are reviewed.

A. Bias and Variance Composition of the Prediction Error

Geman *et al.* [57] give a thorough analysis of the relationship between learning and generalization in neural networks based on the concepts of model bias and model variance. A prespecified model which is less dependent on the data may misrepresent the true functional relationship and have a large bias. On the other hand, a model-free or data-driven model may be too dependent on the specific data and have a large variance. Bias and variance are often incompatible. With a fixed data set, the effort of reducing one will inevitably cause the other increasing. A good tradeoff between model bias and model variance is necessary and desired in building a useful neural network classifier.

Without loss of generality, consider a two-group classification problem in which the binary output variable $y \in \{0, 1\}$ is related to a set of input variables (feature vector) $\mathbf{x}$ by

$$y = F(\mathbf{x}) + \varepsilon$$

where $F(\mathbf{x})$ is the target or underlying function and ε is assumed to be a zero-mean random variable. From (8) and (9), the target function is the conditional expectation of y given $\mathbf{x}$, that is

$$F(\mathbf{x}) = E(y \,|\, \mathbf{x}) = P(\omega_1 \,|\, \mathbf{x}). \tag{12}$$

Given a particular training data set D_N of size N, the goal of modeling is to find an estimate, $f(\mathbf{x}; D_N)$, of $F(\mathbf{x})$ such that an overall estimation error can be minimized. The most commonly used performance measure is the mean squared error

$$\begin{aligned} \text{MSE} &= E[(y - f(\mathbf{x}; D_N))^2] \\ &= E[(y - F(\mathbf{x}))^2] + (f(\mathbf{x}; D_N) - F(\mathbf{x}))^2. \end{aligned} \tag{13}$$

It is important to notice that the MSE depends on the particular data set D_N. A change of the data set and/or sample size may result in a change in the estimation function and hence the estimation error. In most applications, the training data set D_N represents a random sample from the population of all possible data sets of size N. Considering the random nature of the training data, the overall *prediction* error of the model can be written as

$$\begin{aligned} &E_D\{E[(y - f(\mathbf{x}; D_N))^2]\} \\ &\quad = E[(y - F(\mathbf{x}))^2] + E_D[(f(\mathbf{x}; D_N) - F(\mathbf{x}))^2] \end{aligned} \tag{14}$$

where E_D denotes the expectation over all possible random samples of sample size N. In the following, D will be used to represent the data set with the fixed sample size N for convenience. Since the first term on the right hand side, $E[(y -$

$F(\mathbf{x}))^2] = E[\varepsilon^2]$, is independent of both the training sample and the underlying function, it reflects the irreducible estimation error because of the intrinsic noise of the data. The second term on the right hand side of (14), therefore, is a nature measure of the effectiveness of $f(\mathbf{x}; D)$ as a predictor of y. This term can be further decomposed as [57]

$$E_D[(f(\mathbf{x}; D) - E(y \mid \mathbf{x}))^2] = \{E_D[f(\mathbf{x}; D)] - E(y \mid \mathbf{x})\}^2 + E_D\{(f(\mathbf{x}; D) - E_D[f(\mathbf{x}; D)])^2\}. \quad (15)$$

The first term on the right hand side is the square of the bias and is for simplicity called model bias while the second one is termed as model variance. This is the famous *bias plus variance* decomposition of the prediction error.

Ideally, the optimal model that minimizes the overall MSE in (14) is given by $f(\mathbf{x}; D) = E(y \mid \mathbf{x})$, which leaves the minimum MSE to be the intrinsic error $E[\varepsilon^2]$. In reality, however, because of the randomness of the limited data set D, the estimate $f(\mathbf{x}; D)$ is also a random variable which will hardly be the best possible function $E(y \mid \mathbf{x})$ for a given data set. The bias and variance terms in (15) hence provide useful information on how the estimation differs from the desired function. The model bias measures the extent to which the average of the estimation function over all possible data sets with the same size differs from the desired function. The model variance, on the other hand, measures the sensitivity of the estimation function to the training data set. Although it is desirable to have both low bias and low variance, we can not reduce both at the same time for a given data set because these goals are conflicting. A model that is less dependent on the data tends to have low variance but high bias if the model is incorrect. On the other hand, a model that fits the data well tends to have low bias but high variance when applied to different data sets. Hence a good model should balance well between model bias and model variance.

The work by Geman *et al.* [57] on bias and variance tradeoff under the quadratic objective function has stimulated a lot of research interest and activities in the neural network, machine learning, and statistical communities. Wolpert [190] extends the bias-plus-variance dilemma to a more general bias-variance-covariance tradeoff in the Bayesian context. Jacobs [85] studies various properties of bias and variance components for mixtures-of-experts architectures. Dietterich and Kong [41], Kong and Dietterich [94], Breiman [26], Kohavi and Wolpert [93], Tibshirani [168], James and Hastie [86], and Heskes [71] have developed different versions of bias-variance decomposition for zero-one loss functions of classification problems. These alternative decompositions provide insights into the nature of generalization error from different perspectives. Each decomposition formula has its own merits as well as demerits. Noticing that all formulations of the bias and variance decomposition in classification are in additive forms, Friedman [48] points out that the bias and variance components are not necessarily additive and instead they can be "interactive in a multiplicative and highly nonlinear way." He finds that this interaction may be exploited to reduce classification errors because bias terms may be cancelled by low-variance but potentially high-bias methods to produce accurate classification. That simple classifiers often perform well in practice [76] seems to support Friedman's findings.

B. Methods for Reducing Prediction Error

As a flexible "model-free" approach to classification, neural networks often tend to fit the training data very well and thus have low bias. But the potential risk is the overfitting that causes high variance in generalization. Dietterich and Kong [41] point out in the machine learning context that the variance is a more important factor than the learning bias in poor prediction performance. Breiman [26] finds that neural network classifiers belong to unstable prediction methods in that small changes in the training sample could cause large variations in the test results. Much attention has been paid to this problem of overfitting or high variance in the literature. A majority of research effort has been devoted to developing methods to reduce the overfitting effect. Such methods include cross validation [118], [184], training with penalty terms [182], and weight decay and node pruning [137], [148]. Weigend [183] analyzes overfitting phenomena by introducing the concept of the effective number of hidden nodes. An interesting observation by Dietterich [39] is that improving the optimization algorithms in training does not have positive effect on the testing performance and hence the overfitting effect may be reduced by "undercomputing."

Wang [179] points out the unpredictability of neural networks in classification applications in the context of learning and generalization. He proposes a global smoothing training strategy by imposing monotonic constraints on network training, which seems effective in solving classification problems [5].

Ensemble method or combining multiple classifiers [21], [8], [64], [67], [87], [128], [129], [192] is another active research area to reduce generalization error [153]. By averaging or voting the prediction results from multiple networks, the model variance can be significantly reduced. The motivation of combining several neural networks is to improve the out-of-sample classification performance over individual classifiers or to guard against the failure of individual component networks. It has been shown theoretically that the performance of the ensemble can not be worse than any single model used separately if the predictions of individual classifier are unbiased and uncorrelated [129]. Tumer and Ghosh [172] provide an analytical framework to understand the reasons why linearly combined neural classifiers work and how to quantify the improvement achieved by combining. Kittler *et al.* [90] present a general theoretical framework for classifier ensembles. They review and compare many existing classifier combination schemes and show that many different ensemble methods can be treated as special cases of compound classification where all the pattern representations are used jointly to make decisions.

An ensemble can be formed by multiple network architectures, same architecture trained with different algorithms, different initial random weights, or even different classifiers. The component networks can also be developed by training with different data such as the resampling data. The mixed combination of neural networks with traditional statistical classifiers has also been suggested [35], [112].

There are many different ways of combining individual classifiers [84], [192]. The most popular approach to combining multiple classifiers is via simple average of outputs from individual classifiers. But combining can also be done with weighted averaging that treats the contribution or accuracy of component classifiers differently [68], [67], [84]. Nonlinear combining methods such as Dempster–Shafer belief-based methods [141], [192], rank-based information [1], voting schemes [17], and order statistics [173] have been proposed. Wolpert [189] proposes to use two (or more) levels of stacked networks to improve generalization performance of neural network classifiers. The first level networks include a variety of neural models trained with leave-one-out cross validation samples. The outputs from these networks are then used as inputs to the second level of networks that provide smoothed transformation into the predicted output.

The error reduction of ensemble method is mainly due to the reduction of the model variance rather than the model bias. Since the ensemble method works better if different classifiers in the ensemble disagree each other strongly [95], [111], [129], [141], some of the models in the ensemble could be highly biased. However, the averaging effect may offset the bias and more importantly decrease the sensitivity of the classifier to the new data. It has been observed [59] that it is generally more desirable to have an error rate estimator with small variance than an unbiased one with large variance. Empirically a number of studies [41], [93] find that the prediction error reduction of ensemble method is mostly accounted for by the reduction in variance.

Although in general, classifier combination can improve generalization performance, correlation among individual classifiers can be harmful to the neural network ensemble [69], [129], [172]. Sharkey and Sharkey [154] discuss the need and benefits of ensemble diversity among the members of an ensemble for generalization. Rogova [141] finds that the better performance of a combined classifier is not necessarily achieved by combining classifiers with better individual performance. Instead, it is more important to have independent classifiers in the ensemble. His conclusion is in line with that of Perron and Cooper [129] and Krogh and Vedelsby [95] that ensemble classifiers can perform better if individual classifiers considerably disagree with each other.

One of the ways to reduce correlation among component classifiers is to build the ensemble model using different feature variables. In general, classifiers based on different feature variables are more independent than those based on different architectures with the same feature variables [73], [192]. Another effective method is training with different data sets. Statistical resampling techniques such as bootstrapping [45] are often used to generate multiple samples from original training data. Two recently developed ensemble methods based on bootstrap samples are "bagging" [26] and "arcing" classifiers [27]. Bagging (for **b**ootstrap **agg**regation and combin**ing**) and arcing (for **a**daptive **r**esampling and **c**combin**ing**) are similar methods in that both combine multiple classifiers constructed from bootstrap samples and vote for classes. The bagging classifier generates simple bootstrap samples and combines by simple majority voting while arcing uses an adaptive bootstrapping scheme which selects bootstrap samples based on previous constructed ensemble's performances with more weights giving to those cases mostly likely to be misclassified. Breiman [27] shows that both bagging and arcing can reduce bias but the reduction in variance with these approaches is much larger.

Although much effort has been devoted in combining method, several issues remain or have not completely solved. These include the choice of individual classifiers included in the ensemble, the size of the ensemble, and the general optimal way to combine individual classifiers. The issue about under what conditions combining is most effective and what methods should be included is still not completely solved. Combining neural classifiers with traditional methods can be a fruitful research area. Since ensemble methods are very effective when individual classifiers are negatively related [85] or uncorrelated [129], there is a need to develop efficient classifier selection schemes to make best use of the advantage of combining.

IV. Feature Variable Selection

Selection of a set of appropriate input feature variables is an important issue in building neural as well as other classifiers. The purpose of feature variable selection is to find the smallest set of features that can result in satisfactory predictive performance. Because of the curse of dimensionality [38], it is often necessary and beneficial to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model. Out-of-sample performance can be improved by using only a small subset of the entire set of variables available. The issue is also closely related to the principle of parsimony in model building as well as the model learning and generalization discussed in Section III.

Numerous statistical feature selection criteria and search algorithms have been developed in the pattern recognition literature [38], [52]. Some of these statistical feature selection approaches can not be directly applied to neural classifiers due to nonparametric nature of neural networks. Recently there are increasing interests in developing feature variable selection or dimension reduction approaches for neural network classifiers. Most of the methods are heuristic in nature. Some are proposed based on the ideas similar to their statistical counterparts. It is found under certain circumstances that the performance of a neural classifier can be improved by using statistically independent features [49].

One of the most popular methods in feature selection is the principle component analysis (PCA). Principle component analysis is a statistical technique to reduce dimension without loss of the intrinsic information contained in the original data. As such, it is often used as a pre-processing method in neural network training. One problem with PCA is that it is a kind of unsupervised learning procedure and does not consider the correlation between target outputs and input features. In addition, PCA is a linear dimension reduction technique. It is not appropriate for complex problems with nonlinear correlation structures.

The linear limitation of the PCA can be overcome by directly using neural networks to perform dimension reduction. It has been shown that neural networks are able to perform certain nonlinear PCA [70], [125], [147]. Karhunen and Joutsensalo

[89] have discussed many aspects of PCA performed by neural networks. Battiti [16] proposes to use mutual information as the guide to evaluate each feature's information content and select features with high information content.

A number of heuristic measures have been proposed to estimate the relative importance or contribution of input features to the output variable. One of the simplest measures is the sum of the absolute input weights [150] to reflect the impact of that input variable on the output. The limitation of this measure is obvious since it does not consider the impact of perhaps more important hidden node weights. Another simple measure is the sensitivity index [150] which is the average change in the output variable over the entire range of a particular input variable. While intuitively appealing, these measures are not useful in measuring nonlinear effect of the input variable since they don not take consideration of hidden layer weights.

Several saliency measures of input variables explicitly consider both input and hidden weights and their interactions on the network output. For example, pseudo weight [133] is the sum of the product of weights from the input node to the hidden nodes and corresponding weights from the hidden nodes to the output node. An important saliency measure is proposed by Garson [55] who partitions the hidden layer weights into components associated with each input node and then the percentage of all hidden nodes weights attributable to a particular input node is used to measure the importance of that input variable. Garson's measure has been studied by many researchers and some modifications and extensions have been made [22], [56], [60], [114], [123]. Nath *et al.* [123] experimentally evaluate the Garson's saliency measure and conclude that the measure works very well under a variety of conditions. Sung [162] studies three methods of sensitivity analysis, fuzzy curves, and change of mean square error to rank input feature importance. Steppe and Bauer [158] classify all feature saliency measures used in neural networks into derivative-based and weight-based categories with the former measuring the relative changes in either neural network output or the estimated probability of error and the latter measuring the relative size of the weight vector emanating from each feature.

Since exhaustive search through all possible subsets of feature variables is often computationally prohibitive, heuristic search procedures such as forward selection and backward elimination are often used. Based on Garson's measure of saliency, Glorfeld [60] presents a backward elimination procedure to select more predictive feature variables. Steppe and Bauer [159], Steppe *et al.* [160], and Hu *et al.* [81] use the Bonferroni-type or likelihood-ratio test statistic as the model selection criterion and the backward sequential elimination approach to select features. Setiono and Liu [152] also develop a backward elimination method for feature selection. Their method starts with the whole set of available feature variables and then for each attribute variable, the accuracy of the network is evaluated with all the weights associated with that variable set to zero. The variable that gives the lowest decrease in accuracy is removed. Belue and Bauer [22] propose a confidence interval method to select salient features. A confidence interval on the average saliency is constructed to discriminate whether a feature has significant contribution to the classification ability. Using two simulation problems, they find that the method can identify relevant features on which a more accurate and faster learning neural classifiers can be achieved.

Weight elimination and node pruning are techniques often used to remove unnecessary linking weights or input nodes during the network training. One of the earlier methods is the optimal brain damage (OBD) [99]. With this approach, a saliency measure is calculated for each weight based on a simplified diagonal Hessian matrix. Then the weights with the lowest saliency can be eliminated. Based on the idea of OBD, Cibas *et al.* [34] develop a procedure to remove insignificant input nodes. Mozer and Smolensky [119] describe a node pruning method based on a saliency measure that is the difference of the error between when the node is removed and when the node is present. Egmont-Petersen *et al.* [46] propose a method for pruning input nodes based on four feature metrics. Reed [137] presents a review of some pruning algorithms used in neural network models.

All selection criteria and search procedures in feature selection with neural networks are heuristic in nature and lack of rigorous statistical tests to justify the removal or addition of features. Hence, their performance may not be consistent and robust in practical applications. Statistical properties of the saliency measures as well as the search algorithms must be established in order to have more general and systematic feature selection procedures. More theoretical developments and experimental investigations are needed in the filed of feature selection.

V. Misclassification Costs

In the literature of neural network classification research and application, few studies consider misclassification costs in the classification decision. In other words, researchers explicitly or implicitly assume equal cost consequences of misclassification. With the equal cost or 0–1 cost function, minimizing the overall classification rate is the sole objective. Although assuming 0–1 cost function can simplify the model development, equal cost assumption does not represent many real problems in quality assurance, acceptance sampling, bankruptcy prediction, credit risk analysis, and medical diagnosis where uneven misclassification costs are more appropriate. In these situations, groups are often unbalanced and a misclassification error can carry significantly different consequences on different groups.

Victor and Zhang [177] present a detailed investigation on the effect of misclassification cost on neural network classifiers. They find that misclassification costs can have significant impact on the classification results and the appropriate use of cost information can aid in optimal decision making. To deal with asymmetric misclassification cost problem, Lowe and Webb [107], [108] suggest using weighted error function and targeting coding to incorporate the prior knowledge about the relative class importance or different misclassification costs. The proposed schemes are shown effective in terms of improved feature extraction and classification performance.

The situations of unequal misclassification costs often occur when groups are very unbalanced. The costs of misclassifying subjects in smaller groups are often much higher. Under the

assumption of equal consequences of misclassification, a classifier tends to bias toward the larger groups that have more observations in the training sample. For some problems such as medical diagnosis, we may know the prior probabilities of group memberships and hence can incorporate them in the training sample composition. However, a large training sample is often required in order to have enough representatives of smaller groups. Barnard and Botha [13] find that while neural networks are able to make use of the prior probabilities relatively efficiently, the large sample size can improve performance. An alternative approach in selecting training set is using equal number of examples from each group. The results can be easily extended to test sets with unbalanced groups by considering the different prior probabilities in training and test sets [24]. Lowe and Webb [107] propose a weighted error function with a weighting factor to account for different group proportions between the training set and the test set. In a bankruptcy prediction study, Wilson and Sharda [187] investigate the effect of different group compositions in training and test sets on the classification performance. They conclude that the neural network classifier can have better predictive performance using balanced training sample. However if the test set contains too few members of the more important group, the true model performance may not be correctly determined.

Although classification costs are difficult to assign in real problems, ignoring the unequal misclassification risk for different groups may have significant impact on the practical use of the classification. It should be pointed out that a neural classifier which minimizes the total number of misclassification errors may not be useful for situations where different misclassification errors carry highly uneven consequences or costs. More research should be devoted to designing effective cost-based neural network classifiers.

VI. Conclusion

Classification is the most researched topic of neural networks. This paper has presented a focused review of several important issues and recent developments of neural networks for classification problems. These include the posterior probability estimation, the link between neural and conventional classifiers, the relationship between learning and generalization in neural network classification, and issues to improve neural classifier performance. Although there are many other research topics that have been investigated in the literature, we believe that this selected review has covered the most important aspects of neural networks in solving classification problems.

The research efforts during the last decade have made significant progresses in both theoretical development and practical applications. Neural networks have been demonstrated to be a competitive alternative to traditional classifiers for many practical classification problems. Numerous insights have also been gained into the neural networks in performing classification as well as other tasks [23], [169]. However, while neural networks have shown much promise, many issues still remain unsolved or incompletely solved. As indicated earlier, more research should be devoted to developing more effective and efficient methods in neural model identification, feature variable selection, classifier combination, and uneven misclassification treatment. In addition, as a practical decision making tool, neural networks need to be systematically evaluated and compared with other new and traditional classifiers. Recently, several authors have pointed out the lack of the rigorous comparisons between neural network and other classifiers in the current literature [43], [47], [131], [145]. This may be one of the major reasons that mixed results are often reported in empirical studies.

Other research topics related to neural classification include network training [12], [15], [62], [124], [142], model design and selection [50], [72], [117], [121], [122], [180], [194], sample size issues [51], [135], [136], Bayesian analysis [102], [109], [110], [120], and wavelet networks [165], [166], [196]. These issues are common to all applications of neural networks and some of them have been previously reviewed [4], [10], [29], [120], [137], [192]. It is clear that research opportunities are abundant in many aspects of neural classifiers. We believe that the multidisciplinary nature of the neural network classification research will generate more research activities and bring about more fruitful outcomes in the future.

References

[1] K. Al-Ghoneim and B. V. K. V. Kumar, "Learning ranks with neural networks," in *Proc. SPIE Applications Science Artificial Neural Networks*, vol. 2492, 1995, pp. 446–464.

[2] E. I. Altman, G. Marco, and F. Varetto, "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)," *J. Bank. Finance*, vol. 18, pp. 505–529, 1994.

[3] B. Amirikian and H. Nishimura, "What size network is good for generalization of a specific task of interest?," *Neural Networks*, vol. 7, no. 2, pp. 321–329, 1994.

[4] U. Anders and O. Korn, "Model selection in neural networks," *Neural Networks*, vol. 12, pp. 309–323, 1999.

[5] N. P. Archer and S. Wang, "Fuzzy set representation of neural network classification boundaries," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, pp. 735–742, 1991.

[6] H. Asoh and N. Otsu, "An approximation of nonlinear discriminant analysis by multilayer neural networks," in *Proc. Int. Joint Conf. Neural Networks*, San Diego, CA, 1990, pp. III-211–III-216.

[7] L. Atlas, R. Cole, J. Connor, M. El-Sharkawi, R. J. Marks II, Y. Muthusamy, and E. Barnard, "Performance comparisons between backpropagation networks and classification trees on three real-world applications," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1990, vol. 2, pp. 622–629.

[8] R. Avnimelech and N. Intrator, "Boosted mixture of experts: An ensemble learning scheme," *Neural Comput.*, vol. 11, pp. 483–497, 1999.

[9] A. Babloyantz and V. V. Ivanov, "Neural networks in cardiac arrhythmias," in *Industrial Applications of Neural Networks*, F. F. Soulie and P. Gallinari, Eds, Singapore: World Scientific, 1998, pp. 403–417.

[10] P. F. Baldi and K. Hornik, "Learning in linear neural networks: A survey," *IEEE Trans. Neural Networks*, vol. 6, pp. 837–858, 1995.

[11] E. B. Barlett and R. E. Uhrig, "Nuclear power plant status diagnostics using artificial neural networks," *Nucl. Technol.*, vol. 97, pp. 272–281, 1992.

[12] E. Barnard, "Optimization for training neural nets," *IEEE Trans. Neural Networks*, vol. 3, pp. 232–240, 1992.

[13] E. Barnard and E. C. Botha, "Back-propagation uses prior information efficiently," *IEEE Trans. Neural Networks*, vol. 4, pp. 794–802, 1993.

[14] R. Barron, "Statistical properties of artificial neural networks," in *Proc. 28th IEEE Conf. Decision Control*, vol. 280–285, 1989.

[15] R. Battiti, "First- and second-order methods for learning; between steepest descent and Newton's method," *Neural Comput.*, vol. 4, pp. 141–166, 1992.

[16] ——, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[17] R. Battiti and A. M. Colla, "Democracy in neural nets: Voting schemes for classification," *Neural Networks*, vol. 7, no. 4, pp. 691–709, 1994.

[18] E. B. Baum, "What size net gives valid generalization?," *Neural Comput.*, vol. 1, pp. 151–160, 1989.
[19] W. G. Baxt, "Use of an artificial neural network for data analysis in clinical decision-making: The diagnosis of acute coronary occlusion," *Neural Comput.*, vol. 2, pp. 480–489, 1990.
[20] ——, "Use of an artificial neural network for the diagnosis of myocardial infarction," *Ann. Internal Med.*, vol. 115, pp. 843–848, 1991.
[21] ——, "Improving the accuracy of an artificial neural network using multiple differently trained networks," *Neural Comput.*, vol. 4, pp. 772–780, 1992.
[22] L. M. Belue and K. W. Bauer, "Determining input features for multilayer perceptrons," *Neurocomputing*, vol. 7, pp. 111–121, 1995.
[23] J. M. Benitez, J. L. Castro, and I. Requena, "Are artificial neural networks black boxes?," *IEEE Trans. Neural Networks*, vol. 8, pp. 1156–1164, 1997.
[24] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.
[25] H. Bourlard and N. Morgan, "Continuous speech recognition by connectionist statistical methods," *IEEE Trans. Neural Networks*, vol. 4, pp. 893–909, 1993.
[26] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 26, no. 2, pp. 123–140, 1996.
[27] ——, "Arcing classifiers," *Ann. Statist.*, vol. 26, no. 3, pp. 801–823, 1998.
[28] D. E. Brown, V. Corruble, and C. L. Pittard, "A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems," *Pattern Recognit.*, vol. 26, pp. 953–961, 1993.
[29] W. L. Buntine and A. S. Weigend, "Computing second derivatives in feed-forward networks: A review," *IEEE Trans. Neural Networks*, vol. 5, pp. 480–488, 1993.
[30] H. B. Burke, "Artificial neural networks for cancer research: Outcome prediction," *Sem. Surg. Oncol.*, vol. 10, pp. 73–79, 1994.
[31] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell, J. R. Marks, D. P. Winchester, and D. G. Bostwick, "Artificial neural networks improve the accuracy of cancer survival prediction," *Cancer*, vol. 79, pp. 857–862, 1997.
[32] B. Cheng and D. Titterington, "Neural networks: A review from a statistical perspective," *Statist. Sci.*, vol. 9, no. 1, pp. 2–54, 1994.
[33] A. Ciampi and Y. Lechevallier, "Statistical models as building blocks of neural networks," *Commun. Statist.*, vol. 26, no. 4, pp. 991–1009, 1997.
[34] T. Cibas, F. F. Soulie, P. Gallinari, and S. Raudys, "Variable selection with neural networks," *Neurocomput.*, vol. 12, pp. 223–248, 1996.
[35] C. S. Cruz and J. R. Dorronsoro, "A nonlinear discriminant algorithm for feature extraction and data classification," *IEEE Trans. Neural Networks*, vol. 9, pp. 1370–1376, 1998.
[36] S. P. Curram and J. Mingers, "Neural networks, decision tree induction and discriminant analysis: An empirical comparison," *J. Oper. Res. Soc.*, vol. 45, no. 4, pp. 440–450, 1994.
[37] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Contr. Signals Syst.*, vol. 2, pp. 303–314, 1989.
[38] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1982.
[39] T. G. Dietterich, "Overfitting and undercomputing in machine learning," *Comput. Surv.*, vol. 27, no. 3, pp. 326–327, 1995.
[40] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. Artif. Intell. Res.*, vol. 2, pp. 263–286, 1995.
[41] T. G. Dietterich and E. B. Kong, "Machine learning bias, statistical bias, and statistical variance of decision tree algorithms," Dept. Comput. Sci., Oregon State Univ., Corvallis, Tech. Rep., 1995.
[42] P. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[43] R. P. W. Duin, "A note on comparing classifiers," *Pattern Recognit. Lett.*, vol. 17, pp. 529–536, 1996.
[44] S. Dutta and S. Shekhar, "Bond rating: A nonconservative application of neural networks," in *Proc. IEEE Int. Conf. Neural Networks*, vol. 2, San Diego, CA, 1988, pp. 443–450.
[45] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. London, U.K.: Chapman & Hall, 1993.
[46] M. Egmont-Petersen, J. L. Talmon, A. Hasman, and A. W. Ambergen, "Assessing the importance of features for multi-layer perceptrons," *Neural Networks*, vol. 11, pp. 623–635, 1998.
[47] A. Flexer, "Statistical evaluation of neural network experiments: Minimum requirements and current practice," in *Proc. 13th Eur. Meeting Cybernetics Systems Research*, R. Trappl, Ed., 1996, pp. 1005–1008.
[48] J. H. Friedman, "On bias, variance, $0/1$-loss, and the curse of the dimensionality," *Data Mining Knowl. Disc.*, vol. 1, pp. 55–77, 1997.
[49] L.-M. Fu, "Analysis of the dimensionality of neural networks for pattern recognition," *Pattern Recognit.*, vol. 23, pp. 1131–1140, 1990.
[50] O. Fujita, "Statistical estimation of the number of hidden units for feedforward neural networks," *Neural Networks*, vol. 11, pp. 851–859, 1998.
[51] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 873–885, 1989.
[52] K. Funahashi, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1990.
[53] ——, "Multilayer neural networks and Bayes decision theory," *Neural Networks*, vol. 11, pp. 209–213, 1998.
[54] P. Gallinari, S. Thiria, R. Badran, and F. Fogelman-Soulie, "On the relationships between discriminant analysis and multilayer perceptrons," *Neural Networks*, vol. 4, pp. 349–360, 1991.
[55] G. D. Garson, "Interpreting neural network connection weights," *AI Expert*, pp. 47–51, 1991.
[56] T. D. Gedeon, "Data mining of inputs: Analysis magnitude and functional measures," *Int. J. Neural Syst.*, vol. 8, no. 1, pp. 209–218, 1997.
[57] S. Geman, E. Bienenstock, and T. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 5, pp. 1–58, 1992.
[58] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, 1990, pp. 1361–1364.
[59] N. Glick, "Additive estimators for probabilities of correct classification," *Pattern Recognit.*, vol. 10, pp. 211–222, 1978.
[60] L. W. Glorfeld, "A methodology for simplification and interpretation of backpropagation-based neural networks models," *Expert Syst. Applicat.*, vol. 10, pp. 37–54, 1996.
[61] I. Guyon, "Applications of neural networks to character recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 5, pp. 353–382, 1991.
[62] M. T. Hagan and M. Henhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Networks*, vol. 5, pp. 989–993, 1994.
[63] J. B. Hampshire and B. A. Perlmutter, "Equivalence proofs for multilayer perceptron classifiers and the Bayesian discriminant function," in *Proc. 1990 Connectionist Models Summer School*, D. Touretzky, J. Elman, T. Sejnowski, and G. Hinton, Eds. San Mateo, CA, 1990.
[64] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 10, pp. 993–1001, 1990.
[65] F. E. Harreli and K. L. Lee, "A comparison of the discriminant analysis and logistic regression under multivariate normality," in *Biostatistics: Statistics in Biomedical, Public Health, and Environmental Sciences*, P. K. Sen, Ed, Amsterdam, The Netherlands: North Holland, 1985.
[66] A. Hart, "Using neural networks for classification tasks—Some experiments on datasets and practical advice," *J. Oper. Res. Soc.*, vol. 43, pp. 215–226, 1992.
[67] S. Hashem and B. Schmeiser, "Improving model accuracy using optimal linear combinations of trained neural networks," *IEEE Trans. Neural Networks*, vol. 6, no. 3, pp. 792–794, 1995.
[68] S. Hashem, "Optimal linear combination of neural networks," *Neural Networks*, vol. 10, no. 4, pp. 599–614, 1997.
[69] ——, "Treating harmful collinearity in neural network ensembles," in *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, A. J. C. Sharkey, Ed. Berlin, Germany: Springer-Verlag, 1999, pp. 101–125.
[70] J. Hertz, A. Grogh, and R. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 1991.
[71] T. Heskes, "Bias/variance decompositions for likelihood-based estimators," *Neural Comput.*, vol. 10, no. 6, pp. 1425–1434, 1998.
[72] M. Hintz-Madsen, L. K. Hansen, J. Larsen, M. W. Pedersen, and M. Larsen, "Neural classifier construction using regularization, pruning and test error estimation," *Neural Networks*, vol. 11, pp. 1659–1670, 1998.
[73] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, pp. 66–75, Jan. 1994.
[74] L. Holmstrom, P. Koistinen, J. Laaksonen, and E. Oja, "Neural and statistical classifiers-taxonomy and two case studies," *IEEE Trans. Neural Networks*, vol. 8, pp. 5–17, 1997.
[75] M. J. Holt and S. Semnani, "Convergence of back propagation in neural networks using a log-likelihood cost function," *Electron. Lett.*, vol. 26, no. 23, pp. 1964–1965, 1990.
[76] R. C. Holte, "Very simple classification rules perform well on most commonly used data sets," *Mach. Learn.*, vol. 11, pp. 63–90, 1993.
[77] J. J. Hopfield, "Learning algorithms and probability distributions in feedforward and feedback networks," *Proc. Nat. Acad. Sci.*, vol. 84, pp. 8429–8433, 1987.

[78] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, pp. 251–257, 1991.
[79] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.
[80] J. C. Hoskins, K. M. Kaliyur, and D. M. Himmelblau, "Incipient fault detection and diagnosis using artificial neural networks," in *Proc. Int. Joint Conf. Neural Networks*, 1990, pp. 81–86.
[81] M. Y. Hu, M. S. Hung, M. S. Shanker, and H. Chen, "Using neural networks to predict performance of sino-foreign joint ventures," *Int. J. Comput. Intell. Organ.*, vol. 1, no. 3, pp. 134–143, 1996.
[82] W. Y. Huang and R. P. Lippmann, "Comparisons between neural net and conventional classifiers," in *IEEE 1st Int. Conf. Neural Networks*, San Diego, CA, 1987, pp. 485–493.
[83] M. S. Hung, M. Y. Hu, M. S. Shanker, and B. E. Patuwo, "Estimating posterior probabilities in classification problems with neural networks," *Int. J. Comput. Intell. Organ.*, vol. 1, no. 1, pp. 49–60, 1996.
[84] R. A. Jacobs, "Methods for combining experts' probability assessments," *Neural Comput.*, vol. 7, pp. 867–888, 1995.
[85] ——, "Bias/variance analyzes of mixtures-of-experts architectures," *Neural Comput.*, vol. 9, pp. 369–383, 1997.
[86] G. James and T. Hastie, "Generalizations of the bias/variance decomposition for prediction error," Dept. Statistics, Stanford Univ., Stanford, CA, Tech. Rep., 1997.
[87] C. Ji and S. Ma, "Combinations of weak classifiers," *IEEE Trans. Neural Networks*, vol. 8, pp. 32–42, Jan. 1997.
[88] F. Kanaya and S. Miyake, "Bayes statistical behavior and valid generalization of pattern classifying neural networks," *IEEE Trans. Neural Networks*, vol. 2, no. 4, pp. 471–475, 1991.
[89] J. Karhunen and J. Joutsensalo, "Generalizations of principal component analysis, optimization problems and neural networks," *Neural Networks*, vol. 8, pp. 549–562, 1995.
[90] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 226–239, 1998.
[91] D. G. Kleinbaum, L. L. Kupper, and L. E. Chambless, "Logistic regression analysis of epidemiologic data," *Theory Practice, Commun. Statist. A*, vol. 11, pp. 485–547, 1982.
[92] S. Knerr, L. Personnaz, and G. Dreyfus, "Handwritten digit recognition by neural networks with single-layer training," *IEEE Trans. Neural Networks*, vol. 3, pp. 962–968, 1992.
[93] R. Kohavi and D. H. Wolpert, "Bias plus variance decomposition for zero-one loss functions," in *Proc. 13th Int. Conf. Machine Learning*, 1996, pp. 275–283.
[94] E. B. Kong and T. G. Dietterich, "Error-correcting output coding corrects bias and variance," in *Proc. 12th Int. Conf. Machine Learning*, 1995, pp. 313–321.
[95] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," *Adv. Neural Inform. Process.*, vol. 7, pp. 231–238, 1995.
[96] R. C. Lacher, P. K. Coats, S. C. Sharma, and L. F. Fant, "A neural network for classifying the financial health of a firm," *Eur. J. Oper. Res.*, vol. 85, pp. 53–65, 1995.
[97] J. Lampinen, S. Smolander, and M. Korhonen, "Wood surface inspection system based on generic visual features," in *Industrial Applications of Neural Networks*, F. F. Soulie and P. Gallinari, Eds, Singapore: World Scientific, 1998, pp. 35–42.
[98] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubberd, and L. D. Jackel, "Handwritten digit recognition with a backpropagation network," *Adv. Neural Inform. Process Syst.*, vol. 2, pp. 396–404, 1990.
[99] Y. Le Cun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1990, vol. 2, pp. 598–605.
[100] D. S. Lee, S. N. Sriharia, and R. Gaborski, "Bayesian and neural-network pattern recognition: A theoretical connection and empirical results with handwritten characters," in *Artificial Neural Networks and Statistical Pattern Recognition*, I. K. Sethi and A. K. Jain, Eds. New York: Elsevier, 1991, pp. 89–108.
[101] M. Leshno and Y. Spector, "Neural network prediction analysis: The bankruptcy case," *Neurocomput.*, vol. 10, pp. 125–147, 1996.
[102] M. S. Lewicki, "Bayesian modeling and classification of neural signals," *Neural Comput.*, vol. 6, pp. 1005–1030, 1994.
[103] G. S. Lim, M. Alder, and P. Hadingham, "Adaptive quadratic neural nets," *Pattern Recognit. Lett.*, vol. 13, pp. 325–329, 1992.
[104] T. S. Lim, W. Y. Loh, and Y. S. Shih, "An empirical comparison of decision trees and other classification methods," Dept. Statistics, Univ. Wisconsin, Madison, Tech. Rep. 979, 1998.
[105] R. P. Lippmann, "Pattern classification using neural networks," *IEEE Commun. Mag.*, pp. 47–64, Nov. 1989.
[106] ——, "Review of neural networks for speech recognition," *Neural Comput.*, vol. 1, pp. 1–38, 1989.
[107] D. Lowe and A. R. Webb, "Exploiting prior knowledge in network optimization: An illustration form medical prognosis," *Network Comput. Neural Syst.*, vol. 1, pp. 299–323, 1990.
[108] ——, "Optimized feature extraction and the Bayes decision in feed-forward classifier networks," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 355–364, 1991.
[109] D. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, pp. 415–447, 1992.
[110] ——, "A practical Bayesian framework for backpropagation networks," *Neural Comput.*, vol. 4, pp. 448–472, 1992.
[111] G. Mani, "Lowering variance of decisions by using artificial neural network portfolios," *Neural Comput.*, vol. 3, pp. 484–486, 1991.
[112] I. S. Markham and C. T. Ragsdale, "Combining neural networks and statistical predictions to solve the classification problem in discriminant analysis," *Decis. Sci.*, vol. 26, pp. 229–241, 1995.
[113] G. L. Martin and G. L. Pitman, "Recognizing hand-printed letter and digits using backpropagation learning," *Neural Comput.*, vol. 3, pp. 258–267, 1991.
[114] B. Mak and R. W. Blanning, "An empirical measure of element contribution in neural networks," *IEEE Trans. Syst., Man, Cybern. C*, vol. 28, pp. 561–564, Nov. 1998.
[115] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, Eds., *Machine Learning, Neural, and Statistical Classification*, London, U.K.: Ellis Horwood, 1994.
[116] S. Miyake and F. Kanaya, "A neural network approach to a Bayesian statistical decision problem," *IEEE Trans. Neural Networks*, vol. 2, pp. 538–540, 1991.
[117] J. Moody and J. Utans, "Architecture selection strategies for neural networks: Application to corporate bond rating prediction," in *Neural Networks in the Capital Markets*, A.-P. Refenes, Ed. New York: Wiley, 1995, pp. 277–300.
[118] N. Morgan and H. Bourlard, "Generalization and parameter estimation in feedforward nets: Some experiments," *Adv. Neural Inform. Process. Syst.*, vol. 2, pp. 630–637, 1990.
[119] M. C. Mozer and P. Smolensky, "Skeletonization: A technique for trimming the fat from a network via relevance assessment," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1989, vol. 1, pp. 107–115.
[120] P. Muller and D. R. Insua, "Issues in Bayesian analysis of neural network models," *Neural Comput.*, vol. 10, pp. 749–770, 1998.
[121] N. Murata, S. Yoshizawa, and S. Amari, "Learning curves, model selection and complexity of neural networks," in *Advances in Neural Information Processing Systems, 5*, S. J. Hanson, J. D. Cowan, and C. L. Giles, Eds. San Mateo, CA: Morgan Kaufmann, 1993, pp. 607–614.
[122] ——, "Network information criterion determining the number of hidden units for artificial neural network models," *IEEE Trans. Neural Networks*, vol. 5, pp. 865–872, 1994.
[123] R. Nath, B. Rajagopalan, and R. Ryker, "Determining the saliency of input variables in neural network classifiers," *Comput. Oper. Res.*, vol. 24, pp. 767–773, 1997.
[124] V. Nedeljkovic, "A novel multilayer neural networks training algorithm that minimizes the probability of classification error," *IEEE Trans. Neural Networks*, vol. 4, pp. 650–659, 1993.
[125] E. Oja, "Neural networks, principle components, and subspace," *Int. J. Neural Syst.*, vol. 1, pp. 61–68, 1989.
[126] A. Papoulix, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1965.
[127] E. Patwo, M. Y. Hu, and M. S. Hung, "Two-group classification using neural networks," *Decis. Sci.*, vol. 24, no. 4, pp. 825–845, 1993.
[128] M. P. Perrone, "Putting it all together: Methods for combining neural networks," in *Advances in Neural Information Processing Systems*, J. D. Cowan, G. Tesauro, and J. Alspector, Eds. San Mateo, CA: Morgan Kaufmann, 1994, vol. 6, pp. 1188–1189.
[129] M. P. Perrone and L. N. Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," in *Neural Networks for Speech and Image Processing*, R. J. Mammone, Ed. London, U.K.: Chapman & Hall, 1993, pp. 126–142.

[130] T. Petsche, A. Marcantonio, C. Darken, S. J. Hanson, G. M. Huhn, and I. Santoso, "An autoassociator for on-line motor monitoring," in *Industrial Applications of Neural Networks*, F. F. Soulie and P. Gallinari, Eds, Singapore: World Scientific, 1998, pp. 91–97.
[131] L. Prechelt, "A quantitative study of experimental evaluation of neural network algorithms: Current research practice," *Neural Networks*, vol. 9, no. 3, pp. 457–462, 1996.
[132] S. J. Press and S. Wilson, "Choosing between logistic regression and discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 73, pp. 699–705, 1978.
[133] M. Qi and G. S. Maddala, "Option pricing using ANN: The case of S&P 500 index call options," in *Proc. 3rd Int. Conf. Neural Networks Capital Markets*, 1995, pp. 78–91.
[134] S. Raudys, "Evolution and generalization of a single neuron: I. Single-layer perceptron as seven statistical classifiers," *Neural Networks*, vol. 11, pp. 283–296, 1998.
[135] ——, "Evolution and generalization of a single neurone: II. Complexity of statistical classifiers and sample size considerations," *Neural Networks*, vol. 11, pp. 297–313, 1998.
[136] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 252–264, Mar. 1991.
[137] R. Reed, "Pruning algorithms—A survey," *IEEE Trans. Neural Networks*, vol. 4, pp. 740–747, Sept. 1993.
[138] M. D. Richard and R. Lippmann, "Neural network classifiers estimate Bayesian *a posteriori* probabilities," *Neural Comput.*, vol. 3, pp. 461–483, 1991.
[139] A. Ripley, "Statistical aspects of neural networks," in *Networks and Chaos—Statistical and Probabilistic Aspects*, O. E. Barndorff-Nielsen, J. L. Jensen, and W. S. Kendall, Eds. London, U.K.: Chapman & Hall, 1993, pp. 40–123.
[140] ——, "Neural networks and related methods for classification," *J. R. Statist. Soc. B*, vol. 56, no. 3, pp. 409–456, 1994.
[141] G. Rogova, "Combining the results of several neural network classifiers," *Neural Networks*, vol. 7, pp. 777–781, 1994.
[142] A. Roy, L. S. Kim, and S. Mukhopadhyay, "A polynomial time algorithm for the construction and training of a class of multilayer perceptrons," *Neural Networks*, vol. 6, pp. 535–545, 1993.
[143] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a Bayes optimal discriminant function," *IEEE Trans. Neural Networks*, vol. 1, no. 4, pp. 296–298, 1990.
[144] D. E. Rumelhart, R. Durbin, R. Golden, and Y. Chauvin, "Backpropagation: The basic theory," in *Backpropagation: Theory, Architectures, and Applications*, Y. Chauvin and D. E. Rumelhart, Eds. Hillsdale, NJ: LEA, 1995, pp. 1–34.
[145] S. L. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Mining Knowl. Disc.*, vol. 1, pp. 317–328, 1997.
[146] M. S. Sanchez and L. A. Sarabia, "Efficiency of multi-layered feed-forward neural networks on classification in relation to linear discriminant analysis, quadratic discriminant analysis and r egularized discriminant analysis," *Chemometr. Intell. Labor. Syst.*, vol. 28, pp. 287–303, 1995.
[147] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Networks*, vol. 2, pp. 459–473, 1989.
[148] C. Schittenkopf, F. Deco, and W. Brauer, "Two strategies to avoid overfitting in feedforward networks," *Neural Networks*, vol. 10, pp. 505–516, 1997.
[149] M. Schumacher, R. Robner, and W. Vach, "Neural networks and logistic regression: Part I," *Comput. Statist. Data Anal.*, vol. 21, pp. 661–682, 1996.
[150] T. K. Sen, R. Oliver, and N. Sen, "Predicting corporate mergers," in *Neural Networks in the Capital Markets*, A. P. Refenes, Ed. New York: Wiley, 1995, pp. 325–340.
[151] I. Sethi and M. Otten, "Comparison between entropy net and decision tree classifiers," in *Proc. Int. Joint Conf. Neural Networks*, vol. 3, 1990, pp. 63–68.
[152] R. Setiono and H. Liu, "Neural-network feature selector," *IEEE Trans. Neural Networks*, vol. 8, no. 3, pp. 654–662, 1997.
[153] A. J. C. Sharkey, "Multi-net systems," in *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, A. J. C. Sharkey, Ed. Berlin, Germany: Springer-Verlag, 1999, pp. 1–30.
[154] A. J. C. Sharkey and N. E. Sharkey, "Combining diverse neural nets," *Knowl. Eng. Rev.*, vol. 12, no. 3, pp. 231–247, 1997.
[155] J. W. Shavlik, R. J. Mooney, and G. G. Towell, "Symbolic and neural learning algorithms: An empirical comparison," *Mach. Learn.*, vol. 6, pp. 111–144, 1991.
[156] P. A. Shoemaker, "A note on least-squares learning procedures and classification by neural network models," *IEEE Trans. Neural Networks*, vol. 2, pp. 158–160, 1991.
[157] J. Sietsma and R. Dow, "Creating artificial neural networks that generalize," *Neural Networks*, vol. 4, pp. 67–79, 1991.
[158] J. M. Steppe and K. W. Bauer, "Feature saliency measures," *Comput. Math. Applicat.*, vol. 33, pp. 109–126, 1997.
[159] ——, "Improved feature screening in feedforward neural networks," *Neurocomput.*, vol. 13, pp. 47–58, 1996.
[160] J. M. Steppe, K. W. Bauer, and S. K. Rogers, "Integrated feature and architecture selection," *IEEE Trans. Neural Networks*, vol. 7, pp. 1007–1014, 1996.
[161] V. Subramanian, M. S. Hung, and M. Y. Hu, "An experimental evaluation of neural networks for classification," *Comput. Oper. Res.*, vol. 20, pp. 769–782, 1993.
[162] A. H. Sung, "Ranking importance of input parameters of neural networks," *Expert Syst. Applicat.*, vol. 15, pp. 405–411, 1998.
[163] A. J. Surkan and J. C. Singleton, "Neural networks for bond rating improved by multiple hidden layers," in *Proc. IEEE Int. Joint Conf. Neural Networks*, vol. 2, San Diego, CA, 1990, pp. 157–162.
[164] H. Szu, B. Telfer, and S. Kadambe, "Neural network adaptive wavelets for signal representation and classification," *Opt. Eng.*, vol. 31, no. 9, pp. 1907–1916, 1992.
[165] H. Szu, X. Y. Yang, B. Telfer, and Y. Sheng, "Neural network and wavelet transform for scale-invariant data classification," *Phys. Rev.*, vol. 48, no. 2, pp. 1497–1501, 1994.
[166] H. Szu, B. Telfer, and J. Garcia, "Wavelet transforms and neural networks for compression and recognition," *Neural Networks*, vol. 9, pp. 695–708, 1996.
[167] K. Y. Tam and M. Y. Kiang, "Managerial application of neural networks: The case of bank failure predictions," *Manage. Sci.*, vol. 38, no. 7, pp. 926–947, 1992.
[168] R. Tibshirani, "Bias, variance and prediction error for classification rules," Dept. Statist., Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 1996.
[169] A. B. Tickle, R. Andrews, M. Golea, and J. Diederich, "The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks," *IEEE Trans. Neural Networks*, vol. 9, pp. 1057–1068, 1998.
[170] G. T. Toussaint, "Bibliography on estimation of misclassification," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 472–479, 1974.
[171] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognit.*, vol. 29, no. 2, pp. 341–348, 1996.
[172] ——, "Error correlation and error reduction in ensemble classifiers," *Connect. Sci.*, vol. 8, pp. 385–404, 1996.
[173] ——, "Classifier combining through trimmed means and order statistics," *IEEE Int. Conf. Neural Networks*, pp. 695–700, 1998.
[174] J. Utans and J. Moody, "Selecting neural network architecture via the prediction risk: Application to corporate bond rating prediction," in *Proc. 1st Int. Conf. Artificial Intelligence Applications Wall Street*, 1991, pp. 35–41.
[175] P. E. Utgoff, "Perceptron trees: A case study in hybrid concept representation," *Connect. Sci.*, vol. 1, pp. 377–391, 1989.
[176] W. Vach, R. Robner, and M. Schumacher, "Neural networks and logistic regression: Part II," *Comput. Statist. Data Anal.*, vol. 21, pp. 683–701, 1996.
[177] B. L. Victor and G. P. Zhang, "The effect of misclassification costs on neural network classifiers," *Decision Sci.*, vol. 30, pp. 659–682, 1999.
[178] E. Wan, "Neural network classification: A Bayesian interpretation," *IEEE Trans. Neural Networks*, vol. 1, no. 4, pp. 303–305, 1990.
[179] S. Wang, "The unpredictability of standard back propagation neural networks in classification applications," *Manage. Sci.*, vol. 41, no. 3, pp. 555–559, 1995.
[180] Z. Wang, C. D. Massimo, M. T. Tham, and A. J. Morris, "A procedure for determining the topology of multilayer feedforward neural networks," *Neural Networks*, vol. 7, pp. 291–300, 1994.
[181] A. R. Webb and D. Lowe, "The optimized internal representation of multilayer classifier networks performs nonlinear discriminant analysis," *Neural Networks*, vol. 3, no. 4, pp. 367–375, 1990.
[182] A. Weigend, D. Rumelhart, and B. Huberman, "Predicting the future: A connectionist approach," *Int. J. Neural Syst.*, vol. 3, pp. 193–209, 1990.

[183] A. Weigend, "On overfitting and the effective number of hidden units," in *Proc. 1993 Connectionist Models Summer School*, P. Smolensky, D. S. Touretzky, J. L. Elman, and A. S. Weigend, Eds. Hillsdale, NJ, 1994, pp. 335–342.
[184] S. M. Weiss and C. A. Kulilowski, *Computer Systems that Learn*. San Mateo, CA: Morgan Kaufmann, 1991.
[185] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural Comput.*, vol. 1, pp. 425–464, 1989.
[186] B. Widrow, D. E. Rumelhard, and M. A. Lehr, "Neural networks: Applications in industry, business and science," *Commun. ACM*, vol. 37, pp. 93–105, 1994.
[187] R. L. Wilson and R. Sharda, "Bankruptcy prediction using neural networks," *Decision Support Syst.*, vol. 11, pp. 545–557, 1994.
[188] H. Wolpert, "On the connection between in-sample testing and generalization error," *Complex Syst.*, vol. 6, pp. 47–94, 1992.
[189] ——, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.
[190] ——, "On bias plus variance," *Neural Comput.*, vol. 9, pp. 1211–1243, 1997.
[191] L. Xu, "Recent advances on techniques of static feedforward networks with supervised learning," *Int. J. Neural Syst.*, vol. 3, no. 3, pp. 253–290, 1992.
[192] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combing multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, pp. 418–435, 1992.
[193] Y. Yoon, G. Swales, and T. M. Margavio, "A comparison of discriminant analysis versus artificial neural networks," *J. Oper. Res. Soc.*, vol. 44, pp. 51–60, 1993.
[194] J.-L. Yuan and T. L. Fine, "Neural-network design for small training sets of high dimension," *IEEE Trans. Neural Networks*, vol. 9, pp. 266–280, 1998.
[195] G. Zhang, M. Y. Hu, E. B. Patuwo, and D. Indro, "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis," *Eur. J. Oper. Res.*, vol. 116, pp. 16–32, 1999.
[196] Q. Zhang and A. Benveniste, "Wavelet networks," *IEEE Trans. Neural Networks*, vol. 3, pp. 889–898, 1992.

Guoqiang Peter Zhang received the B.S. and M.S. degrees in mathematics and statistics from East China Normal University, Shanghai, China, and the Ph.D. degree in management science from Kent State University, Kent, OH.

He is an Assistant Professor of Decision Sciences at Georgia State University, Atlanta. His main research interests include neural networks and time series forecasting. His articles have appeared in *Computers and Industrial Engineering, Computers and Operations Research, Decision Sciences, European Journal of Operational Research, OMEGA, International Journal of Forecasting, International Journal of Production Economics*, and others.

Dr. Zhang is a member of INFORMS and DSI.