ACTA STEREOL 1996; 15/1: 15-24
ORIGINAL SCIENTIFIC PAPER

CLASSIFICATION PROCEDURES FOR DIAGNOSIS BASED ON MULTIPLE MORPHOMETRIC PARAMETERS

Andrew J. EINSTEIN, Joan GIL

Departments of Biomathematical Sciences and Pathology, Box 1023, Mount Sinai School of Medicine of the City University of New York, One Gustave L. Levy Place, New York, NY 10029, USA

ABSTRACT

A diagnosis in pathology is a classification based upon multiple parameters. Therefore, an important challenge in morphometry is to determine what the most effective classification procedure is for a given setting; in practice, this applies to a narrow differential diagnosis between predetermined choices of similar morphology. To resolve this problem, a number of approaches have been recommended, both statistical and non-statistical. Among the statistical classificatory procedures are discriminant analysis, logistic regression, k-nearest neighbor analysis, and recursive partitioning. Among the non-statistical procedures, the community has expressed a preference for artificial neural networks. Experience in our laboratory using neural networks in the diagnosis of histology and cytology specimens suggests the tentative conclusion that in comparison with other classificatory algorithms, neural networks are the method of choice.

Key words: classification, diagnosis, image analysis, morphometry.

INTRODUCTION

Diagnosis **is** classification. In pathology, the clinical applicability of image analysis is dependent on prescribing a specific method of classifying new cases based on multiple parameters. Each case is represented by a vector of parameters; a sample of these vectors with known diagnoses, referred to as a training set, is used to determine a general method for assigning arbitrary parameter vectors to diagnostic groups. Although most morphometric classificatory studies to date have employed stepwise parameter selection and discriminant analysis to classify cases, there is no reason to assume that they are the most effective of the numerous methods which exist. The problem of choice and optimization of classification methods is a crucial one, yet a review of the literature shows that this issue has rarely been addressed and never been the subject of systematic study. Mathematically, the problem can be framed in terms of how to best partition a multidimensional parameter space, where "best" can be understood in terms of various measures of diagnostic effectiveness, such as accuracy, sensitivity, specificity, and false negative rate.

It is important to note that the classifications we are dealing with are between very similar diagnostic groups, such as breast epithelial cell lesions, and not among broad categories such as cancer. The role of computers in diagnosis in the foreseeable future will be to assist with objective classification among narrow differentiations, not to replace physicians. While a cytopathologist has no trouble identifying breast tissue and hyperplastic breast epithelial cells within that tissue, he may find it difficult to reproducibly distinguish between mild and moderate hyperplasia. Moreover, as stressed by Rosai (1991), one cytologist's mild hyperplasia is the next one's moderate hyperplasia. Image analysis and classificatory methodology enable us to objectify this portion of the diagnostic process. This consultative role of computers in medicine is discussed with great insight and lucidity by Blois (1980). As

illustrated in Figure 1, he describes the cognitive span required during diagnosis as a series of judgments, each of which narrows the possibilities for diagnosis. While at the beginning of this diagnostic process (Point A) the "totality of the world must be confronted," the role of computers is at the terminal stage represented by point B. Here, the "task domain has been structured through previous human effort, an abstraction is available, and little common-sense knowledge may be required."

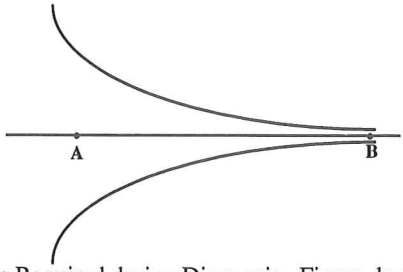

Fig. 1. The Cognitive Span Required during Diagnosis. Figure drawn after Blois (1980).

The classification problem is really a twofold one: 1) how to best select parameters, and 2) how to best select and optimize a classification algorithm.

SELECTION OF PARAMETERS

Parameters are known alternatively, depending on the context, as features, descriptors, inputs, variables, or covariates. They characterize properties of tissue deriving from morphometric or stereologic methods; in pathology these typically include measures of nuclear profile sizes and size distributions, chromatin texture, tissue architecture, etc. Besides from the parameter selection incorporated into some classification algorithms, numerous stand-alone methods are employed. These include both univariate approaches such as ANOVA and its nonparametric analog the Kruskal-Wallis test, Genchi-Mori ambiguity, area under an ROC curve (Bartels, 1979), and the Fisher ratio (de Meester et al., 1991), and multivariate methods, such as factor analysis and the popular stepwise procedure. While parameter selection merits treatment in its own right, the focus of the discussion here will be on classification algorithm selection and optimization. We shall treat, in order, types of classifiers, the details of some important algorithms, comparisons between classification methods, and overtraining/overfitting.

TYPES OF CLASSIFIERS

Classifiers can be grouped as supervised or unsupervised, statistical or non-statistical, and parametric or nonparametric. In supervised learning, we know the "state of nature" (e.g., normal, adenoma, carcinoma) for each sample, while in unsupervised learning we don't (Duda and Hart, 1973). Unsupervised classifiers such as clustering methods, which are not generally used in morphometry, use the underlying structure of the data to assign group numbers to individual cases. The groups, or clusters, which are formed are not defined *a priori* but rather are suggested by the data, although they can be compared to, and identified with, states of nature (SAS Institute, 1989). Numerous clustering methods exist. They vary in the nature of the clusters, which may be disjoint, hierarchical, overlapping, or fuzzy. Most classifiers are statistical, but some use alternative methods, e.g., neural networks and rule-based expert systems. The distinction between statistical and non-statistical classifiers is an artificial one; it is easily seen that the statistical method CART effectively generates a rule-based expert system. Statistical classifiers assuming a normal distribution of covariates are called parametric. While they are generally robust, non-normality may require a transformation or rank ordering of covariates. Often, a simple logarithmic tranform will yield covariates with a Gaussian distribution. Properties of a number of popular classifiers are summarized in Table 1, and we discuss the rudiments of a few of these approaches below.

Table 1. Properties of Some Classification Methods.

Classifier	Supervised?	Statistical?	Parametric?
Artificial Neural Network (ANN)	Yes	No	No
Bayesian Belief Network	Yes	Yes	—
CART (Recursive Partitioning)	Yes	Yes	No
Cluster Analysis	No	Yes	No
Discriminant Analysis	Yes	Yes	Yes
k-Nearest Neighbor	Yes	Yes	No
Logistic Regression	Yes	Yes	No
Rule-based Expert System (LISP)	Yes	No	No

Discriminant Analysis

Discriminant analysis is a parametric classification method and as such is most appropriate when covariates have approximately normal within-class distributions. Nevertheless it is robust and may still serve as an effective classifier in the absence of normality (normality can be determined with the Shapiro-Wilk W test). The basic approach is to produce a mathematical rule, called a discriminant function, that determines the posterior probability that an unknown belongs to a particular group. Owing to the nature of the resultant discriminant function, the method is called linear discriminant analysis if we assume equal within-group covariance matrices, and quadratic discriminant analysis if we assume unequal within-group covariance matrices. Diagnosis can be associated with the group of highest posterior probability. The procedure for computing posterior probabilities is given as follows (SAS Institute, 1989):

The generalized squared distance from an unknown parameter vector $\boldsymbol{x}$ to group i is given by

$$D_i^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{m}_i)' \boldsymbol{V}_i^{-1} (\boldsymbol{x} - \boldsymbol{m}_i) + g_1(i) + g_2(i)$$

where

p_i	is the prior probability of membership in group i,
$\boldsymbol{m}_i$	is the vector containing the variable means in group i,
$\boldsymbol{V}_i$	is the covariance matrix within group i if we assume unequal within-group covariance matrices, or the pooled covariance matrix if we assume equal within-group covariance matrices,
$g_1(i) = ln \mid V_i \mid$	if the within-group covariance matrices are used, or
$g_1(i) = 0$	if the pooled covariance matrix is used,
$g_2(i) = -2 \ln p_i$	if the prior probabilities are not all equal, or
$g_2(i) = 0$	if the prior probabilities are all equal.

Letting n denote the number of groups, the posterior probability of unknown $\boldsymbol{x}$ belonging to group i is defined as

$$p(i \mid \boldsymbol{x}) = exp\ (-0.5\ D_i^2(\boldsymbol{x})) \ /\ \Sigma_{j=1,2,\ldots,n}\ exp\ (-0.5\ D_j^2(\boldsymbol{x}))\ .$$

Logistic Regression

Logistic regression (Hosmer and Lemeshow, 1989) is a nonparametric statistical approach to classification. We want to discriminate among $n+1$ diagnostic categories numbered $0, 1, \ldots, n$ where category 0 is designated as the reference category. Let y be the outcome variable, taking the value of the diagnostic category, and let $\boldsymbol{x}$ be the $m+1$ element vector of covariates, where $x_0 = 1$ is constant and x_i represents the ith covariate for $i = 1, 2, \ldots, m$. Define the n logit functions as

$$\begin{aligned} f_i(\boldsymbol{x}) &= ln\ [\ P(y = i \mid \boldsymbol{x}) \ /\ P(y = 0 \mid \boldsymbol{x})] \\ &= \beta_{i0} + \beta_{i1}\, x_1 + \beta_{i2}\, x_2 + \ldots + \beta_{im}\, x_m \\ &= \boldsymbol{x}^T\, \boldsymbol{\beta}_i \end{aligned}$$

for $i = 1, 2, \ldots, n$. Letting $f_0(x) = 0$, we then have that the conditional probability for an outcome given a covariate vector is

$$p(y = i \mid x) = exp\ (f_i(x)) \,/\, \Sigma_{j=0,1,\ldots,n}\ exp\ (f_j(x))\ . \qquad (1)$$

To compute maximum likelihood estimates of the β_{ij} we must set the partial derivatives of the log-likelihood function equal to zero and solve the resultant equations, typically by the Newton-Raphson method. Estimates of the information matrix and of its inverse, the covariance matrix, are similarly determined by computing the second partial derivatives of the log-likelihood function. These, in turn, are used to determine the standard errors of the estimates. To test the null hypothesis that a component of β is zero, and thereby determine if a variable contributes to the model, we can use the likelihood ratio test, Wald's test, or the score test. The score test is the least popular of these methods. Of the other two, the likelihood ratio test is regarded as the better method; both Hauck and Donner (1977) and Jennings (1986) have demonstrated that Wald's test behaves aberrantly. All of these calculations are performed by commercial packages such as SYSTAT. Having determined MLEs of the β_{ij} from the training set, we can now use them to assign posterior probabilities to a new case by applying Eq. 1. Like in discriminant analysis, we can assign a diagnosis to a new case by choosing the outcome with the greatest posterior probability.

k-*Nearest Neighbor Analysis*

Another nonparametric classificatory algorithm is based on a case's k nearest neighbors in parameter space. Given an unknown case, we identify the k points (e.g., 10) nearest to it in parameter space and determine to which diagnostic category they belong (Unger et al., 1993). Nearness may be in terms of Euclidean or Mahalanobis distance. If all k nearest neighbors belong to the same group, then that is the classification for the unknown. If the neighbors belong to different groups, then posterior probabilities are computed for membership in each of the groups. Let n be the number of diagnostic groups, p_i $(i = 1, 2, \ldots, n)$ be the prior probabilities of membership in each group, and f_i $(i = 1, 2, \ldots, n)$ be the number of k-nearest neighbors of the unknown case belonging to each group, satisfying $\Sigma f_i = k$. Then the posterior probability $p(i \mid x)$ that the unknown belongs to group i is given by

$$p(i \mid x) = f_i p_i \,/\, \Sigma_{j=1,2,\ldots,n} f_j p_j\ .$$

A diagnosis can be associated with the group of maximal posterior probability.

Recursive Partitioning with Classification and Regression Trees (CART)

A third nonparametric approach to classification is recursive partitioning, generally associated with the Classification and Regression Trees (CART) method of Breiman and Friedman (Breiman et al., 1993). Recursive partitioning is based on the process of binary stratification, done by analyzing each parameter to find its optimal cutpoint, which best separates patients into diagnostic groups (Lacher, 1991). Cutpoint determination may be achieved by a number of splitting criteria (Segal, 1988). The test maximizing classification accuracy is chosen as the first partition, and the binary stratification is repeated on each half of the partitioned parameter space. The recursive partitions can be described by a tree of decision rules.

Unconstrained partitioning will lead to a tree with as many terminal nodes as cases, decreasing the generalizability of the classifier. For this reason, the decision tree is "pruned," removing non-informative nodes. Here too there are a number of methods that can be used, such as the cross-validation method. Thus, a CART classifier is characterized by its splitting and pruning criteria. Alternative classifiers can be generated by varying the cost associated with an incorrect diagnosis, expressed in terms of the misclassification cost ratio.

CART's counterpart and competitor is the Fast Algorithm for Classification Trees (FACT) of Loh and Vanichsetakul (1988). FACT attempts to combine the advantages of CART with linear discriminant analysis; this parametric method places more emphasis on linear combination and non-binary splits, does not use cross-validation but rather employs a stop-

splitting rule, and incorporates methods for dealing with missing values (Stewart and Stamm, 1991). Despite the case made by FACT's proponents, CART remains the standard for recursive partitioning.

Artificial Neural Networks (ANNs)

Artificial neural networks (ANNs) are a type of artificial intelligence (Lawrence, 1993; Dytch and Wied, 1990; Deligdisch et al., 1995). They can generalize from a training set, making few assumptions as to the underlying data structure. While several neural network designs have been studied and employed for classification (Tourassi and Floyd, 1995; DaPonte and Sherman, 1991), backpropagation networks remain the standard, and the focus of this section. The fundamental unit of structure in artificial neural networks, like their biological counterparts, is the neuron. A backpropagation ANN has its neurons arranged in a multilayered architecture, with each neuron connected to the neurons in its adjacent layers. There is a layer of input neurons, one or more layers of hidden neurons, and a layer of output neurons. A value is associated with each neuron, and a weight with each connection. In addition, hidden and output neurons have biases associated with them; these may be regarded as weights of special neurons having constant values of one. The value of a hidden or output neuron is computed by passing the biased weighted sum of values from the previous layer through a "transfer function," most typically a sigmoidal function. More formally, let n be the number of layers, n_k be the number of neurons in layer k ($k = 1, 2,\ldots, n$), w_{ijk} be the connection weight between the ith neuron in layer k and the jth neuron in layer $k-1$ ($i = 1, 2,\ldots, n_k$; $j = 1, 2,\ldots, n_{k-1}$; $k = 2, 3,\ldots, n$), and b_{ik} and v_{ik} be the bias and value of the ith neuron in layer k, respectively ($i = 1, 2,\ldots, n_k$; $k = 2, 3,\ldots, n$). Then the values v_{ik} are determined from the formula

$$v_{ik} = 1 / (1 + exp(-\Sigma_{j=1,2,\ldots,n_{k-1}} (w_{ijk} * v_{j(k-1)} + b_{ik}))) .$$

The architecture of a typical backpropagation network for use in morphometry is illustrated in Fig. 2. As is shown, input neurons represent morphometric parameters, while output neurons code for the diagnosis. The values taken by output variables are generally constrained to the interval [0, 1]; thus, e.g., a value near one for the "Malignant" neuron is identified with a diagnosis of malignancy.

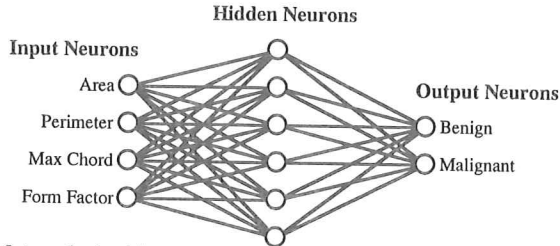

Figure 2. Neural Network Architecture. Example of a backpropagation neural network with four input neurons, six hidden neurons, and two output neurons.

The classificatory power of neural networks is contained in the weights and biases, known together as the connection matrix (Astion and Wilding, 1992b). Determination of the optimal connection matrix for a particular set of training facts is a computationally difficult problem; for even very simple networks, the problem has been mathematically demonstrated to be NP-complete (not computable in polynomial time) (Blum and Rivest, 1991). The approach generally used in ANNs is an adaptive one, in which the weights and biases are "learned" over repeated iterations through the training set. The connection matrix is initially set at random, and the values of hidden and output neurons are computed for a parameter vector. Output neuron values are compared with their target values, e.g., (0,1) for malignancy in the example. If the difference between the output value and its target value, known as the error factor, is less than a specified training tolerance, the case represented by the parameter value is regarded to be

correctly classified, and the next case is considered. If the error factor is too large, then the connection matrix is modified using some learning rule, most typically the generalized delta rule or a modification thereof. These learning rules reflect factors such as the transfer function, neuron values, error factors, weights, and biases. Initially, output biases and connection weights between output neurons and the last layer of hidden neurons are adjusted. These changes are then backpropagated, one layer at a time, until the whole connection matrix has been modified. Cycling through the training set continues until all cases are correctly classified or, barring this, some other specified stopping point.

Using the final connection matrix, a new set of inputs can be used to predict output values. Numerous factors are incorporated into the design of a neural network, and these may affect its training and ability to generalize. Such factors include network topology, noise, order of training facts, choice of transfer function, initial weights, range specification, learning rate, training tolerance, testing tolerance, and the training set.

Other Methods

Besides the popular approaches discussed above, there exist numerous other classification methods, several of which are described in Duda and Hart's (1973) classic book. Bartels and Thompson (Bartels et al., 1992) have championed the use of Bayesian Belief Networks as a classifier for quantitative pathology. Rule-based expert systems, often developed in the computer language LISP, have fallen out of favor for artificial intelligence applications. Other methods include Parzen Windows, the Furthest-Neighbor Algorithm, and linear programming.

COMPARISON OF CLASSIFICATORY APPROACHES

It is difficult to know which classifier will perform best for a given data set. Each method has its advantages and disadvantages. Neural networks probably make the least assumptions as to the surfaces that divide parameter space. They provide a number of options for fine-tuning. However, neural nets have been accused of being "black-box" classifiers, accurate but uninterpretable. While the effect of individual weights in a neural network is certainly less comprehensible than that of, for example, splits in CART, methods have been developed to determine how a neural network uses information. Ravdin and Clark (1992) examined a trained network by selectively inactivating its input units, both individually and in sets. They claim that the effect of using only a single input neuron is analogous to evaluating its univariate importance, while inactivating the neuron may determine its multivariate importance. By including or omitting pairs of neurons, interactions between parameters can be analyzed. Wilding et al. (1994) identify the most influential features in a jackknife analysis by individually varying inputs by ±10% of their range and determining the result on the output. These methods may also be used for parameter selection. The role of hidden neurons may be studied using a Hinton diagram (Qian and Sejnowski, 1988), which graphically represents connection weights. Apologetics aside, a neural network is undeniably more difficult to interpret than a linear discriminant function or CART tree. But this very complexity and nonlinearity is the source of neural networks' classificatory power.

Statistical approaches are more dependent on the validity of a model that is assumed. Linear discriminant analysis and CART should perform well if the "true" separating surfaces are roughly linear; whether this is a reasonable assumption for morphometric data is unclear. Quadratic discriminant analysis can model more complex separating surfaces than its linear counterpart, but it is more prone to overfitting of the training data (Hand, 1992). As mentioned, both forms of discriminant analysis may suffer in the absence of normality. Ironically, logistic regression may fail if categories are too well separated in space. Using a data set in which each case had previously been correctly diagnosed using a jackknifed neural network approach (Einstein et al., 1994), we attempted to make a classification using logistic regression, but log-likelihood functions did not converge. Interestingly, we observed that convergence may in fact be dependent on the reference group chosen. While some classifiers interpolate diagnostic information for regions of parameter space where they have no training points, logistic regression effectively refuses to venture a classification.

Breiman et al. (1993) suggest a number of advantages of their CART method. While a number of these could apply to most classifiers, some are specific to recursive partitioning.

These include that it is invariant under monotone transformations of parameters, and that it is very robust with respect to misclassified points and outliers. One feature of CART is that the classification tree it yields bears a strong resemblance to clinical practice guidelines (Hadorn et al., 1992). For applications where we want readily explainable classification rules, this is extremely beneficial. For applications where classification accuracy, regardless of complexity, is most important, this may be an unnecessary simplification.

Table 2. Selected Studies Comparing Classification Methods

Methods & Accuracies (%)				Reference
NN (93.3) >	LDA (84.4) >	QDA (82.2)		Erler et al., 1991
NN (47) >	Bayes (45)			Dawson et al., 1991
NN (100) =	LDA (100)			Unger et al., 1995
*NN (77.7) >	DT (76.5) >	LDA (75.0)		Palcic et al., 1992
LR (94.3) =	LDA (94.3)			Stenkvist & Strande, 1989
NN (93.8, 82.2, 77.3) >	WDD (93.5, 81.7, 77.2)			Errington & Graham, 1993
CART (91.3) >	LR (87.4) >	QDA (84.5) >	LDA (83.5)	Lacher, 1991
CART (83.3) >	NN (73.8)=	LDA (73.8)		Reibnegger et al., 1991
NN (80) >	QDA (75)			Astion & Wilding, 1992a
NN (94.1) >	LDA (85.7) >	QDA (78.0)		Erler et al., 1995
NN (80, 70) >	LDA (64, 65) >	QDA (69, 45)		Lamb & Niederberger, 1993
LR (72) >	CART (68), other models			Hadorn et al., 1992
NN (89.9) >	LR (88.4)			Doig et al., 1993
NN (90.6) >	kNN (81) >	LDA (78.125)		DaPonte & Sherman, 1991
Bayes (99.1) >	NN (95.4)			Tourassi & Floyd, 1995
kNN (83) >	NN (79) >	MLM (73)		Clarke et al., 1993
LR, LDA, CART (Not all accuracies given)				Stewart & Stamm, 1991

NN = Neural Network. LDA = Linear Discriminant Analysis. QDA = Quadratic Discriminant Analysis.
Bayes = Bayesian Classifier. DT = Decision Tree. LR = Logistic Regression.
WDD = Weighted Density Distribution Classifier, the best previously reported chromosome classifier.
kNN = k-Nearest Neighbors Method. MLM = Maximum Likelihood Method.
* But comparison differed for multi-stage classification: LDA (87.2) > NN (85.1) > DT (84.3).

Despite numerous papers using classificatory algorithms in a host of medical applications, there have been remarkably few comparing their accuracies. Table 2 summarizes some of this literature. We make no claims of comprehensiveness. The comparisons may not all be fair, in the sense that in a single study one method may be more optimized than the other(s). In some studies numerous classifications were compared and one representing each classifier had to be chosen. Perhaps the truth is as Hand (1992) contends, that "careful and sensitive use of any method will probably yield very similar results." A few papers have been omitted in which classifier comparison using different parameter sets may have biased the results, accuracy was based primarily on the training set, or the emphasis was on survival models or different classifiers than those above.

In diagnostic pathology, we are only aware of three papers that have employed both neural networks and a statistical classificatory method (Dawson et al., 1991; Erler et al., 1991; Unger et al., 1995); in all cases neural networks were equal or better. Classifier performance should be expected to vary depending on the data structure, and therefore results from one study may not be applicable to another. If it is possible to draw any conclusions, we would make the tentative claim that neural networks tend to perform better than statistical classifiers due to their minimal assumptions about data structure. Indeed, this flexibility has resulted in the application of neural networks to problems refractory to conventional approaches, such as speech pronunciation from text (Sejnowski and Rosenberg, 1987), which can be viewed as a classification problem with strings of letters as parameter vectors and phonemes as output. Our preference for neural networks is shared by other reviewers (Pun et al., 1994) and lent support by the American Joint Committee on Cancer. To replace the TNM cancer staging system, the Committee has developed a new prognostic system using an artificial neural network to combine prognostic factors (Burke, 1994). Nevertheless, the classification problem requires considerable further study.

CLASSIFIER TRAINING AND OVERTRAINING

Regardless of the classifier used, training can be performed by three methods. The simplest approach is simply to divide the data set into two groups, a training set and a testing set. The training set is used to determine the classifier's parameters (such as weights in a neural network or coefficients of a discriminant function) while the testing set is left to evaluate the trained classifier. If it is difficult to obtain a sufficiently large data set, cross-validation can be used, allowing training on a larger number of patients. The data set is divided into k groups, of which k-1 are used for training and the remaining group is used to test the trained classifier. This is repeated with each of the k groups left as the testing set. A composite measure of accuracy can be used to evaluate the classificatory scheme. When n-way cross-validation is performed on n patients, this is called jackknife analysis or leave-one-out analysis. Any cross-validation scheme yields multiple trained classifiers, but if the accuracy is sufficiently high for each of them we might propose the same classifier trained on the entire data set as a classificatory approach. While these training methods may seem obvious, the literature abounds with examples of poorly constructed studies in which training and testing sets overlap.

If we want to combine classifier training with optimization, the data can be divided into three parts: a training set, an optimization testing set, and an evaluation testing set (Astion and Wilding, 1992b). If desired, this may be combined with a scheme for cross-validation. A variety of classifier models can be compared, e.g., modifying tolerances, architecture, noise, and transfer function of a neural network. All the classifiers are tested against the optimization testing set, and the most accurate of the models is chosen. If we are cross-validating, the process is repeated for permutations of the three groups. The accuracy of this optimized classifier (or classifiers) may then be evaluated with the new testing facts left in the evaluation testing set (or sets). Accuracy determined by this method should generally be more realistic than accuracy estimated with a single testing set.

While at first glance it might seem reasonable to provide a classifier with as many parameters as are computationally tractable, leaving it to the algorithm to sort out which are important, this raises problems of overtraining, known in other contexts as overfitting. In any data classification scheme, two competing factors are at work as the classification rule is allowed to grow more complex: 1) identifying general features in a data set, those which are predictive of outcome, and 2) identifying the particular features of specific data points. If the ratio of the sample size to the number of weights or input parameters (the latter is called the S:I ratio) is low, a "clever" classifier may focus on the particulars and miss the big picture. Such a classifier will be able to classify its training cases with great accuracy, but not much more. While the problem of classifier complexity has been raised mostly as a criticism of overtraining neural networks (Clark et al., 1994), it takes equally problematic forms in other classificatory approaches, e.g., the number of splits allowed in CART and the type of discriminant function and number of covariates in discriminant analysis. In neural networks, overtraining may result from too many neurons or from too many cycles of training (a consequence of too small a training tolerance). Illustrating this was a study by Lamb and Niederberger (1993) using neural networks to predict fertility potential. Feeding a neural network their training set 10,000 times, it achieved a training classification rate of 92% and a testing rate of 81%. With 1.5 million cycles of training, the training rate increased to 98% but the testing rate decreased to 64%.

Complicating the matter even further is the fact that morphometric parameters, deriving from the same images, will in general be highly interdependent. Thus, formal treatments of sample size, based upon independent input variables, may not entirely apply, typically proposing more stringent sample size requirements than are necessary. Nevertheless, haphazard and nondiscriminatory parameter selection will still result in overtraining. The lack of a clear method of addressing this problem underscores the importance of using biologically relevant parameters, for a classificatory approach that simulates pathologist's decision-making processes and avoids multiple measures of biologic features should minimize some opportunities for the classifier to overtrain. Neural network overtraining may also be reduced by choosing a larger training tolerance and fewer hidden neurons.

In the literature, training problems in neural networks are often framed (Astion and Wilding, 1992b; Wilding et al., 1994) in terms of shrinkage, defined as the difference in the

classification rates for training and testing facts. This may just cloud the real issue of overtraining. Overtraining results in incorrect classifications for new testing facts. Shrinkage may be a second order measure of diagnostic accuracy, but ultimately what we care about is the accuracy itself, i.e., how well "unknowns" will be classified. (Of course sensitivity, specificity, and predictive values are similarly of concern, and the argument applies equally to these measures.) Training classification rate, which certainly should be at least as high as testing rate, is not particularly important in its own right. In sum, a classifier should be chosen based upon its ability to classify unknowns.

ACKNOWLEDGEMENTS

Andrew Einstein gratefully acknowledges support from a Hans Elias Bursary and from United States Public Health Service Grant 5-T32-GM 7280-16.

REFERENCES

Astion ML, Wilding P. Application of neural networks to the interpretation of laboratory data in cancer diagnosis. Clin Chem 1992a; 38: 34-8.

Astion ML, Wilding P. The application of backpropagation neural networks to problems in pathology and laboratory medicine. Arch Pathol Lab Med 1992b; 116: 995-1001.

Bartels PH, Thompson D, Bibbo M, Weber JE. Bayesian belief networks in quantitative histopathology. Analyt Quant Cytol Histol 1992; 14: 459-73.

Bartels PH. Numerical evaluation of cytologic data. III. Selection of features for discrimination. Analyt Quant Cytol 1979; 1: 153-9.

Blois MS. Clinical judgment and computers. N Engl J Med 1980; 303: 192-7.

Blum A, Rivest RL. Training a 3-node neural network is NP-complete. In: Touretzky DS ed. Advances in Neural Information Processing Systems I. San Mateo, CA: Morgan Kaufmann Publishers, 1989: 494-501.

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. New York: Chapman & Hall, 1993.

Burke HB. Increasing the power of surrogate endpoint biomarkers: the aggregation of predictive factors. J Cell Biochem Suppl 1994; 19: 278-82.

Clark GM, Hilsenbeck SG, Ravdin PM, De Laurentiis M, Osborne CK. Prognostic factors: rationale and methods of analysis and integration. Breast Cancer Res Treat 1994; 32: 105-112.

Clarke LP, Velthuizen RP, Phuphanich S, Schellenberg JD, Arrington, Silbiger M. MRI: stability of three supervised segmentation techniques. Magn Reson Imaging 1993; 11: 95-106.

DaPonte JS, Sherman P. Classification of ultrasonic image texture by statistical discriminant analysis and neural networks. Comp Med Imaging Graph 1991; 15: 3-9.

Dawson AE, Austin RE, Weinberg DS. Nuclear grading of breast carcinoma by image analysis: classification by multivariate and neural network analysis. J Clin Pathol Suppl 1991; 95: S29-S37.

de Meester U, Young IT, Lindeman J, van der Linden HC. Towards a quantitative grading of bladder tumors. Cytometry 1991; 12: 602-13.

Deligdisch L, Einstein AJ, Guera D, Gil J. Ovarian dysplasia in epithelial inclusion cysts: a morphometric approach using neural networks. Cancer 1995; 76: 1027-34.

Doig GS, Inman KJ, Sibbald WJ, Martin CM, Robertson JM. Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. Proc Annu Symp Comput Appl Med Care 1993: 361-5.

Duda RO, Hart PE. Pattern Classification and Scene Analysis. New York: John Wiley, 1973.

Dytch HE, Wied GL. Artificial neural networks and their use in quantitative pathology. Analyt Quant Cytol Histol 1990; 12: 379-93.

Einstein AJ, Barba J, Unger P, Gil J. Nuclear diffuseness as a measure of texture: definition and application to the computer-assisted diagnosis of parathyroid adenoma and carcinoma. J Microsc 1994; 176: 158-66.

Erler BS, Hsu L, Truong HM et al. Image analysis and diagnostic classification of hepatocellular carcinoma using neural networks and multivariate discriminant functions. Lab Invest 1994; 71: 446-51.

Erler BS, Vitagliano P, Lee S. Superiority of neural networks over discriminant functions for thalassemia minor screening of red blood cell microcytosis. Arch Pathol Lab Med 1995; 119: 350-4.

Errington PA, Graham J. Application of artificial neural networks to chromosome classification. Cytometry 1993; 14: 627-39.

Hadorn DC, Draper D, Rogers WH, Keeler EB, Brook RH. Cross-validation performance of mortality prediction models. Stat Med 1992; 11: 475-89.

Hand DJ. Statistical methods in diagnosis. Stat Methods Med Res 1992; 1: 49-67.

Hauck WW, Donner A. Wald's test as applied to hypotheses in logit analysis. J Am Stat Assoc 1977; 72: 851-3.

Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: John Wiley, 1989.

Jennings DE. Judging inference adequacy in logistic regression. J Am Stat Assoc 1986; 81: 471-6.

Lacher DA. Comparison of nonparametric recursive partitioning to parametric discriminant analyses in laboratory differentiation of hypercalcemia. Clinica Chimica Acta 1991; 204: 199-208.

Lamb DJ, Niederberger CS. Artificial intelligence in medicine and male infertility. World J Urol 1993; 11: 129-36.

Lawrence J. Introduction to Neural Networks: Design, Theory and Applications. Nevada City, CA: California Scientific Software, 1993.

Loh W, Vanichsetakul N. Tree-structured classification via generalized discriminant analysis. J Am Stat Assoc 1988; 83: 715-25.

Palcic B, MacAulay C, Shlien S, Treurniet W, Tezcan H, Anderson G. Comparison of three different methods for automated classification of cervical cells. Anal Cell Path 1992; 4: 429-41.

Pun T, Gerig G, Ratib O. Image analysis and computer vision in medicine. Comp Med Imaging Graph 1994; 18: 85-96.

Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. J Mol Bio 1988; 202: 865-84.

Ravdin PM, Clark GM. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. Breast Cancer Res Treat 1992; 22: 285-93.

Reibnegger G, Weiss G, Werner-Felmayer G, Judmaier G, Wachter H. Neural networks as a tool for utilizing laboratory information: comparison with linear discriminant analysis and with classification and regression trees. Proc Natl Acad Sci USA 1991; 88: 11426-30.

Rosai J. Borderline epithelial lesions of the breast. Am J Surg Pathol 1991; 15: 209-21.

SAS Institute Inc. SAS/STAT User's Guide, Version 6, Fourth Edition. Cary, North Carolina: SAS Institute Inc., 1989

Segal MR. Regression trees for censored data. Biometrics 1988; 44: 35-47.

Sejnowski TJ, Rosenberg CR. Parallel networks that learn to pronounce English text. Compex Systems 1987; 1: 145-68.

Stenkvist B, Strande G. Analysis of machine-selected cells with an image analysis system in normal and abnormal cervical specimens. Anal Cell Pathol 1989; 2: 1-13.

Stewart PW, Stamm JW. Classification tree prediction models for dental caries from clinical, microbiological, and interview data. J Dent Res 1991; 70: 1239-51.

Tourassi GD, Floyd CD. Lesion Size Quantification in SPECT using an artificial neural network classification approach. Comp Biomed Res 1995; 28: 257-70.

Unger PD, Hoon V, Stone N et al. Computerized interactive morphometry in the differential diagnosis of irradiated prostates. Analyt Quant Cytol Histol 1995; 17: 100-8.

Unger PD, Watson CW, Liu Z, Gil J. Morphometric analysis of neoplastic renal aspirates and benign renal tissue. Analyt Quant Cytol Histol 1993; 15: 61-6.

Wilding P, Morgan MA, Grygotis AE, Shoffner MA, Rosato EF. Application of backpropagation neural networks to diagnosis of breast and ovarian cancer. Cancer Lett 1994; 77: 145-53.

Presented at the 9th International Congress for Stereology, Copenhagen, August 20th to 25th, 1995.