On Piecewise-Linear Classification

Gabor T. Herman and K. T. Daniel Yeung

Abstract— **If two sets of vectors (in N-dimensional real Euclidean space R^N) do not have an element in common, then they can always be separated from each other by using a series of $N-1$ dimensional hyperplanes in R^N. In piecewise-linear classification, one finds such a series of hyperplanes using a training set containing elements from both classes. Efficient methods to find such a piecewise-linear separation for the training sets have been proposed in the literature. However, since complete separation of the training set fits the "noise" as well as the "signal" in the training set, the desirability of such a complete separation depends on the nature of the data. In this paper, we make use of a real data set (containing 9-D measurements of fine needle aspirates of a patient's breast for the purpose of classifying a tumor's malignancy) for which early stopping in the generation of the separating hyperplanes is not appropriate. We compare a piecewise-linear classification method (both with complete separation on the training set and with separation using only seven hyperplanes) with classification based on a single (but in a statistical sense optimal) linear separator. A precise methodology for comparing the relative efficacy of two classification methods for a particular task (including a way of providing the statistical significance of the results) is described and is applied to the comparison on the breast cancer data of the relative performances of the two versions of the piecewise-linear classifier and the classification based on an optimal linear separator. It is found that for this data set, the piecewise-linear classifier that uses all the hyperplanes needed to separate the training set outperforms the other two methods and that these differences in performance are significant at the 0.001 level. There is no statistically significant difference between the performance of the other two methods. We discuss the relevance of these results for this and other applications.**

Index Terms—**Malignancy detection, medical diagnosis, optimal linear separation, pattern recognition, performance evaluation, piecewise-linear classification.**

I. Introduction and Background

In the problems that are the subject matter of this article, each item of a data set is represented by an N-dimensional vector s of real numbers. We are assuming that we have to make a binary (yes or no) decision regarding s, such as "does s indicate that a malignancy is present?" We use the terms *normal* and *abnormal* to distinguish between the s's in the two classes. We refer to the abstract for an outline of that which follows.

The classification methods under consideration make use of *linear abnormality index functions.* These are linear functionals α on R^N, i.e., they associate with each N-dimensional vector s in R^N a real number $\alpha(s)$, where the operator α itself can be represented as a nonzero element of R^N whose nth component is α_n so that

$$\alpha(s) = \sum_{n=1}^{N} \alpha_n s_n \tag{1}$$

(s_n is the nth component of s). Further, we assume that the components α_n are scaled so that

$$\max_{1 \le n \le N} |\alpha_n| = 1. \tag{2}$$

Manuscript received September 27, 1990; revised August 21, 1991. Recommended for acceptance by Associate Editor S. Tanimoto. This work was supported by the National Institutes of Health under Grant HL28438.

The authors are with the Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104.

IEEE Log Number 9107015.

In the context of standard pattern classification theory [1], the vector of α_n's corresponds to the *weight vector* of a linear discriminant function. The reason why we introduced linear abnormality index functions, as opposed to using the classical linear discriminant functions, is that we are interested in the behavior of the latter when the weight vector is fixed, but the *threshold* (called the *threshold weight* in [1]) is allowed to vary. Geometrically, there is a one-to-one correspondence between weight vectors that satisfy (2) and sets of parallel hyperplanes and, once one such set is chosen, between the thresholds and the elements of the set. In the classical approach [1], [2], one typically searches for a linear discriminant function that is in some sense optimal. This involves searching for an optimal combination of the weight vector and the threshold. In our discussion, we find it more convenient to separate them.

For any particular application, one has to use a *training set* of normals and abnormals to find the appropriate α's for that application. Furthermore, the training set is also used to find "optimal thresholds," which are defined as follows.

Let X and Y be any two finite nonempty subsets of R^N, α be an arbitrary linear abnormality index function, and t be any real number. We define $M(X, Y, \alpha, t)$ to be the number of x's in X for which $\alpha(x) > t$ plus the number of y's in Y for which $\alpha(y) \le t$. We think of $M(X, Y, \alpha, t)$ as the number of misclassified items for *normal set* X, *abnormal set* Y, linear abnormality index function α, and *threshold* t. We define the *optimal range* $T(X, Y, \alpha)$ to be the closure of the set of real numbers t for which $M(X, Y, \alpha, t)$ assumes its minimal value. (It is easy to see that for any τ, there exists an ϵ, such that the value of $M(X, Y, \alpha, t)$ is constant for all t in the range $[\tau, \tau + \epsilon]$. This is why we defined an "optimal range" rather than an optimal value.) We will assume that X, Y, and α are such that there exists a positive integer J and, for $1 \le j \le J$, real numbers $\ell_j(X, Y, \alpha)$ and $u_j(X, Y, \alpha)$ satisfying

$$\ell_j(X, Y, \alpha) < u_j(X, Y, \alpha), \quad \text{for } 1 \le j \le J \tag{3}$$

$$u_j(X, Y, \alpha) < \ell_{j+1}(X, Y, \alpha), \quad \text{for } 1 \le j < J \tag{4}$$

$$T(X, Y, \alpha) = \bigcup_{j=1}^{J} [\ell_j(X, Y, \alpha), u_j(X, y, \alpha)] \tag{5}$$

i.e., that $T(X, Y, \alpha)$ is the union of a finite number of closed intervals.

Given a training set consisting of a set X of known normals and a set Y of known abnormals and given a linear abnormality index function α, any threshold t in the interior of the optimal range $T(X, Y, \alpha)$ will minimize the number $M(X, Y, \alpha, t)$ of misclassifications in the training set. Since future decisions have to be based on a fixed threshold rather than on a range, we need to select one particular t. We do this by making further use of the training set as follows. Motivated by the assumption that the elements of X and Y were randomly selected from the population, we define the *prevalence* $p_X(X, Y)$ of normals and the *prevalence* $p_Y(X, Y)$ of abnormals as

$$p_X(X, Y) = |X|/(|X| + |Y|) \tag{6}$$

$$p_Y(X, Y) = |Y|/(|X| + |Y|) \tag{7}$$

where we use $|S|$ to denote the number of elements in a finite set S. If $T(X, Y, \alpha)$ consists of a single interval, then the more prevalent the

abnormals are, the nearer the threshold to be used for classification should be to the lower end of the interval. Thus, the function

$$t(X,Y,c,d) = p_Y(X,Y) \times \ell(c,d) + p_X(X,Y) \times u(c,d) \quad (8)$$

where $\ell(c,d)$ is the smaller and $u(c,d)$ is the larger of c and d, provides us with our sought-after threshold in case $T(X,Y,\alpha)$ consists of a single interval with end points c and d. If there are multiple intervals, then we combine them into one according to their lengths, i.e., we define

$$\ell(X,Y,\alpha) = \frac{\sum_{j=1}^{J} \ell_j(X,Y,\alpha) \times (u_j(X,Y,\alpha) - \ell_j(X,Y,\alpha))}{\sum_{j=1}^{J}(u_j(X,Y,\alpha) - \ell_j(X,Y,\alpha))} \quad (9)$$

$$u(X,Y,\alpha) = \frac{\sum_{j=1}^{J} u_j(X,Y,\alpha) \times (u_j(X,Y,\alpha) - \ell_j(X,Y,\alpha))}{\sum_{j=1}^{J}(u_j(X,Y,\alpha) - \ell_j(X,Y,\alpha))}. \quad (10)$$

We then define the *optimal threshold* $t(X,Y,\alpha)$ based on the normal set X, abnormal set Y, and linear abnormality index function α by

$$t(X,Y,\alpha) = t(X,Y,\ell(X,Y,\alpha),u(X,Y,\alpha)). \quad (11)$$

To summarize, for a training set consisting of normals X and abnormals Y and a fixed linear abnormality index function (alternatively, weight vector) α, (11) provides a threshold t, which is optimal in the sense that 1) the number of misclassifications in the training set using the linear discriminant function based on the weight vector α and the threshold t will not be more than the number of misclassifications using a linear discriminant function based on the same weight vector and any other threshold and 2) among all thresholds t that have this optimality property, the one provided by (11) matches best the prevalances of normals and abnormals in the training set.

II. Piecewise-Linear Classification

The following methodology was designed to achieve complete separation of disjoint sets of normals and abnormals. It iteratively builds up a (possibly nonconvex) piecewise-linear classifier that, after a finite number of iterative steps, is guaranteed to distinguish correctly between the normals and abnormals in the training set (provided only that there is no s that occurs both as a normal and as an abnormal in the training set). For a background to this method, see [3] and its references. (There are alternative approaches to piecewise-linear classification; for an example, see [4].)

We first give an intuitive description of the method. Given any (finite) training set of normals and abnormals, from any set of parallel hyperplanes (defined by a fixed linear abnormality index function), one can always select a pair of hyperplanes such that all the normals lie on one side of one of them, and all the abnormals lie on the opposite side of the other one. If the sets of normals and abnormals are linearly separable, then a single hyperplane can be used for both hyperplanes of this pair, and it provides a classifier that is error free for the training set. If the set of normals and abnormals is not linearly separable, then both hyperplanes of the pair will correctly classify all elements in the training set that do not lie between them, but they may misclassify elements that do lie between them. For any linear abnormality index function, there will be a pair of such hyperplanes for which the distance between them (and, hence, the width of the region of potential misclassification) is minimal. We select, among all possible linear abnormality index functions, the one that (in a rather specific sense) minimizes this just-defined minimum distance. This linear abnormality index function and the associated pair of hyperplanes provide a partial classifier that makes a decision for all points that are not between the hyperplanes. For points that are between the hyperplanes, we do the following. We repeat the process just described for those normals and abnormals in the original training set that fall between the hyperplanes. This provides a new pair of hyperplanes, which can be used for (partially) classifying points that could not be classified before. We keep repeating this process until the leftover normals and abnormals become linearly separable, and a final single hyperplane can be used in the classification process. We now give a precise version of this intuitive description.

For any finite nonempty subsets X^k and Y^k of R^N, we can find a linear abnormality index function α^k such that

$$\max_{x \in X^k} \alpha^k(x) - \min_{y \in Y^k} \alpha^k(y) \quad (12)$$

is as small as possible. The existence of such a minimizing α^k follows from the continuity of the functional in (12) and the condition on the unknowns α_n as given by (2). (An alternative precise statement is given in Theorem 3.2 of [3]. The method proposed there for finding α^k uses linear programming and therefore works in polynomial time when the components of x in X^k and y in Y^k are rational, which is not a restriction in any application. In our implementation, we have been using a multidimensional biased random search technique [5] to estimate the α^k.)

Given a finite training set consisting of nonempty sets of normals X and abnormals Y, we define a nonnegative integer $\bar{K}$ and sequences X^k, Y^k, α^k, c^k, d^k $(0 \le k \le \bar{K})$ as follows. The method makes use of two logical variables called *stuck* and *sept*.

PROCEDURE

Step (0)

$k = 0$;

$X^0 = X$ and $Y^0 = Y$;

$stuck$ = .false. and $sept$ = .false.

Step (i)

While $stuck$ = .false. and $sept$ = .false.;

find α^k which minimizes (12);

$$c^k = \min_{y \in Y^k} \alpha^k(y); \quad (13)$$

$$d^k = \max_{x \in X^k} \alpha^k(x); \quad (14)$$

if $d^k < c^k$;

then

$sept$ = .true.;

end then;

else

if for all $x \in X^k$, $\alpha^k(x) \ge c^k$ and for all $y \in Y^k$, $\alpha^k(y) \le d^k$;

then

$stuck$ = .true.;

end then;

else

$$X^{k+1} = X^k - \{x \in X^k \mid \alpha^k(x) < c^k\} \quad (15)$$

$$Y^{k+1} = Y^k - \{y \in Y^k \mid \alpha^k(y) > d^k\} \quad (16)$$

increase k by 1;

end else;

end if;

end else;

end if;

end while.

Step (ii)

$\bar{K} = k$.

end PROCEDURE.

Several comments are in order about this procedure.

First of all, the procedure is well defined in the sense that every time Step (i) needs to be executed, neither X^k nor Y^k is empty. This is proved by induction. The statement is valid for $k = 0$. Suppose now that is is valid for a particular k. In order to get to $k+1$, it has to be the case that $d^k \geq c^k$. Consider an $\bar{x}$ in X^k and a $\bar{y}$ in Y^k such that $d^k = \alpha^k(\bar{x})$ and $c^k = \alpha^k(\bar{y})$. Then, $\alpha^k(\bar{x}) \geq c^k$, and therefore, $\bar{x} \in X^{k+1}$ and $\alpha^k(\bar{y}) \leq d^k$, and therefore, $\bar{y} \in Y^{k+1}$.

Second, we see that the procedure is bound to terminate. This follows from the fact that for $k < \bar{K}$, either there is an x in X^k such that $\alpha^k(x) < c^k$ or there is a y in Y^k such that $\alpha^k(y) > d^k$ (or both). Hence, for all $k < \bar{K}$, either $|X^{k+1}| < |X^k|$ or $|Y^{k+1}| < |Y^k|$ (or both). Since both X^0 and Y^0 are finite and we have just proven that every time Step (i) is entered, neither X^k nor Y^k is empty, it follows that Step (i) can only be entered a finite number of times, and therefore, the procedure must terminate.

Third, we note that if the procedure terminates with *sept* being true, then for any t such that $d^{\bar{K}} < t < c^{\bar{K}}$, we have that for all $x \in X^{\bar{K}}$ and for all $y \in Y^{\bar{K}}, \alpha^{\bar{K}}(x) < t < \alpha^{\bar{K}}(y)$; therefore, a linear separation of $X^{\bar{K}}$ and $Y^{\bar{K}}$ can be achieved. On the other hand, if the procedure terminates with *stuck* being true, that implies that the sets $X^{\bar{K}}$ and $Y^{\bar{K}}$ are rather intermingled. One could now use alternative methods to further separate $X^{\bar{K}}$ and $Y^{\bar{K}}$ (see, e.g., the so-called degeneracy procedure in [3]), but we feel that such need not be introduced in this paper for two reasons. First, as noted in [3], for most real problems, the condition that results from *stuck* being true does not occur. (This is also what we found in the experiments reported below.) Second, we feel that if the condition does occur, it indicates a genuine overlap between normals and abnormals in the underlying distributions and further separation of the $X^{\bar{K}}$ and $Y^{\bar{K}}$ would be rather artificial.

Based on the sequences defined by the above procedure, the following algorithm can be used to classify an arbitrary element s of R^N. The algorithm uses a nonnegative integer K, the choice of which is discussed below. The only restriction is that $K \leq \bar{K}$.

ALGORITHM

Step (0)

$k = 0$.

Step (i)

While $k < K$;

if $\alpha^k(s) < c^k$, **then** classify s as normal **and** stop;

if $\alpha^k(s) > d^k$, **then** classify s as abnormal **and** stop;

increase k by 1;

end while.

Step (ii)

if $\alpha^K(s) \leq t(X^K, Y^K, c^K, d^K)$, **then** classify s as normal **and** stop;

classify s as abnormal.

end ALGORITHM.

In Step (ii), we use the t defined by (8). Note that the algorithm uses a total of $2K + 1$ hyperplanes.

The method that is recommended in [3] is essentially the same as the above algorithm with $K = \bar{K}$ and with a degeneracy procedure to avoid getting "stuck." Our preference is not to exclude the possibility of stopping prior to complete classification of the training set. In general, K should be chosen large enough to make use of all the relevant information in the training set but not so large that we start fitting irrelevant information (i.e., noise). A methodology for choosing the K is described in Section VI.

III. Classification Based on an Optimal Linear Separator

Since some of our early experiments (see Section VI below) indicated that often a very small value of K is an appropriate stopping point, it appeared to us reasonable to consider, as an alternative to the classification method described in the previous section, another one that uses a single linear abnormality index function α but one that is an optimal linear separator. This concept comes from statistical pattern recognition theory [2] and is defined as follows.

For any finite nonempty subset S of R^N and any linear abnormality index function α, we define a mean and a variance by

$$m(S, \alpha) = \frac{1}{|S|} \sum_{s \in S} \alpha(s) \tag{17}$$

$$v(S, \alpha) = \frac{1}{|S|} \sum_{s \in S} (\alpha(s) - m(S, \alpha))^2. \tag{18}$$

If X and Y are finite nonempty subsets of R^N such that at least one of $v(X, \alpha)$ and $v(Y, \alpha)$ is not zero. then we define the *separability of* X and Y using α by

$$\sigma(X, Y, \alpha) = \frac{p_X(X, Y) p_Y(X, Y)(m(X, \alpha) - m(Y, \alpha))^2}{p_X(X, Y) v(X, \alpha) + p_Y(X, Y) v(Y, \alpha)}. \tag{19}$$

A linear abnormality index function α is said to be an *optimal linear separator* for X and Y if for all linear abnormality index functions β

$$\sigma(X, Y, \alpha) \geq \sigma(X, Y, \beta). \tag{20}$$

We assume without further discussion that for a given training set consisting of X and Y, one can find an optimal linear separator α and that for this α, the conditions expressed in (3)–(5) hold. In practice, we have been using a multidimensional biased random search technique [5] to estimate an optimal linear separator for given sets of normals and abnormals.

Once an optimal linear separator α has been determined based on the normals X and abnormals Y in the training set, we use the optimal threshold defined by (11) to classify an arbitrary element s of R^N as normal if and only if

$$\alpha(s) \leq t(X, Y, \alpha). \tag{21}$$

Note that this method of classification is very similar to using the algorithm of the last section with $K = 0$ and replacing α^0 with an optimal linear separator.

IV. Methodology of Comparison

Since we intend to compare the performance of the classification methods of the last two sections, in this section, we discuss a general methodology for comparing two classification techniques. We adopt the (reasonably standard) comparison procedure of [6] and extend it with a (reasonably standard) test for statistical significance.

We assume that we have two (hopefully large) sets X and Y of known normals and abnormals. We partition X and Y by randomly assigning (with equal probability) elements of X to subsets $X_1, X_2, \cdots, X_{10}$ and elements of Y to subsets $Y_1, Y_2, \cdots, Y_{10}$. For any classification method, we do the following, for $1 \leq i \leq 10$. We train the method on the normal set $X - X_i$ and abnormal set $Y - Y_i$. Then, we use the so-trained method to classify elements in X_i and in Y_i. Note that when we are done, each element of X and each element of Y has been classified exactly once. We define the *accuracy* of the method as the number of correctly classified items in X and Y divided by $|X| + |Y|$.

If we find that one method is more accurate than another one according to this definition, then we still need to decide whether or

TABLE I
SIGNIFICANCE (P) OF ALL THE DIFFERENCES BETWEEN THE THREE METHODS ON THE WISCONSON BREAST CANCER DATA. (METHOD 1 IS PIECEWISE-LINEAR CLASSIFICATION WITH COMPLETE SEPARATION OF THE TRAINING SET (ACCURACY 0.990). METHOD 2 IS PIECEWISE-LINEAR CLASSIFICATION BASED ON FOUR PAIRS OF HYPERPLANES, i.e., USING SEVEN HYPERPLANES FOR SEPARATION (ACCURACY 0.963). METHOD 3 IS CLASSIFICATION BASED ON AN OPTIMAL LINEAR SEPARATOR (ACCURACY 0.959). FOR DEFINITION OF Q AND q, SEE THE TEXT. P IS THE PROBABILITY OF OBSERVING AS HIGH OR HIGHER VALUE OF q FOR THE GIVEN VALUE OF Q IF THE NULL HYPOTHESIS OF EQUAL PERFORMANCE OF THE TWO METHODS WERE CORRECT.

	Q	q	p
1 vs. 2	17	15	<0.001
1 vs. 3	21	18	<0.001
2 vs. 3	20	11	0.412

not our finding is statistically significant. We adopt the sign test [7] to provide a level of significance for rejecting the null hypothesis that the two classification methods are equally good in favor of the hypothesis that the one with the greater accuracy is better.

In this approach to statistical significance, we look only at those elements of X and Y that have been classified differently by the two classification methods. Let Q be the total number of such elements, and let q be the number of such elements that have been correctly classified by the classification method with greater accuracy. The null hypothesis that the methods perform equally well implies that q is a random sample from a binomial distribution with total number of items Q and equal probabilities assigned to the two classes. We use this binomial distribution to determine the probability of randomly selecting an element from it with value q or higher. This probability provides the level of significance for rejecting the null hypothesis.

V. PERFORMANCE ANALYSIS FOR DIAGNOSIS OF BREAST CANCER

The performance of piecewise-linear classification when applied to 9-D data of measurements of a fine needle aspirate taken from a patient's breast is briefly discussed in [3]. Additional data have been collected since that time, and now, there is available from the same source a data set of 294 normals (no breast malignancy) and of 193 abnormals (confirmed breast malignancy). We refer to this data set as Wisconsin Breast Cancer Data (WBCD).

We applied three classification methods to this data set. One is the method of [3], which is the piecewise-linear classification algorithm with $K = \bar{K}$. (For all of the ten training sets that we generated according to the methodology of the last section, we found that the procedure that generates $\bar{K}$ terminated with $sept$ true and, hence, $stuck$ false, and therefore, our not including the degeneracy procedure of [3] makes no difference to the results of the experiments.) The second method is the piecewise-linear classification algorithm with the K chosen to be 3 since it was reported in [3] that four pairs of hyperplanes were necessary for complete separation of the subset of the WBCD, which was available at that time. (To completely separate the normal set $X - X_i$ from the abnormal set $Y - Y_i$, we needed four pairs of hyperplanes for two of the i's, we needed five pairs for three of the i's and six pairs for the remaining five i's.) The third method is classification based on an optimal linear separator. The respective accuracies of the three methods were found to be 0.990, 0.963, and 0.959. The difference in performance between the first method and either of the other two was found to be significant at the 0.001 level, but the difference in performance between the latter methods is not statistically significant. For details of the significance analysis, see Table I.

We note that in [3], a slightly different version of piecewise-linear classification is proposed from the one discussed in this paper. The difference is in Step (ii) of the algorithm, where [3] recommends $(c^K + d^K)/2$ as the threshold based on the final pair. We found that using this variant, there were some very minor changes in accuracy: that of Method 1 decreased to 0.988 and of Method 2 increased to 0.967.

VI. DISCUSSION

Classification of multidimensional data in medicine is an important topic since, if successful, it can lead to automated diagnosis or at least provide a tool to speed up and/or improve diagnosis.

The method of piecewise-linear classification will completely separate two disjoint finite point sets in R^N provided that a sufficient number of pairs of separating hyperplanes are used [3]. However, it occurred to us that this mathematically desirable property may not be very significant in practice since complete separation on a training set does not guarantee perfect performance on further data (e.g., on a testing set). In fact, after we have delineated the major parts of the clusters, the use of additional pairs of hyperplanes may even be counterproductive as opposed to using a single optimal-threshold hyperplane parallel to the last of the previous pairs since the additional pairs of hyperplanes would likely to be fitting the "noise" in the training set and would therefore be less relevant to the testing set than the prevalences of normals and abnormals on which the optimal threshold is based. This reasoning is reinforced by statements made in other papers that report on experimental comparison of methods that are trained on a training set. In [8], the authors say that "One surprising result of these experiments is how well the simple perceptron algorithm performs. The perceptron was largely abandoned as a general learning mechanism about 20 years ago because of its inherent limitations, such as its inability to learn concepts that are not linearly-separable [Minsky88]. Nevertheless, it performs quite well in these experiments. Except on NETtalk and in the presence of imperfect training data, the accuracy of the perceptron is hardly distinguishable from the more complicated learning algorithms. In addition, it is very efficient." They go on to say "Regardless of the reason, data for many "real" problems seems to consist of linearly separable categories. Since the perceptron is a simple and efficient learning algorithm in this case, using a perceptron as an initial test system is probably a good idea." Similarly, in [6], it is reported that although "in every case a logistic solution was found that exceeded the performance of solutions posed using different underlying models ... linear classifiers (with the assumption of a normal distribution) gave good performance in all cases except the thyroid experiment." Other experiments that we have performed, involving mathematically generated data intended to simulate the problem of lung tumor recognition in computerized tomography [9], also tend to confirm this conclusion. However, the nature of those data seems so particularly suited for classification by a single linear separator that we consider it not worthwhile to report on the details of those experiments.

Rather interestingly, our initial experience with the WBCD was also similar. Using the chronologically earlier items in the data as a training set and those obtained later as the testing set, we found that using a low value of K outperformed the version of the method that completely separated the training set. However, the full analysis reported here decisively shows that the initial result was an aberration. We found that for the WBCD, a classifier based on complete separation of the confirmed normals and abnormals is superior in a statistically very significant sense to others using fewer hyperplanes for classification. This experience confirms the need for

statistically valid analysis, as opposed to anecdotal data, in choosing a classifier.

However, one should be careful not to confuse statistical significance with significance for the application at hand. We note that even the classifier based on a single (optimal) linear separator has accuracy 0.959, i.e., it misclassified only 4.1% of the cases. It can be further adjusted to meet the special requirements of the medical problem that we are trying to solve. For example, if a false positive only results in further (but more invasive) medical tests but a false negative would lead to not treating a cancer, it would be better to move the threshold in (21) to bias towards not misclassifying abnormals, even if the overall accuracy decreases as a result. Similar adjustments can of course be made to the piecewise-linear classifier, and the methodology described above can be applied to test the methods with the revised figure of merit in mind. In any case, one should carefully distinguish in experiments such as those described here between the estimated size of the difference in the performance of techniques (in our case 0.990 accuracy versus 0.959 accuracy) and the statistical significance of the conclusion that one method is better than the other (in our case 0.001, meaning that if the two methods were equally good, then experiments such as we have performed would indicate such superiority of performance less than once out of 1000 times).

We observe that if fewer hyperplanes are used, the classification will be less expensive. Our recommendation, therefore, is that for a new problem, piecewise-linear classification should be compared with classification based on an optimal linear separator, with the latter method being the method of choice until evidence indicates otherwise. Furthermore, in choosing the stopping point K for the piecewise-linear classification algorithm, the technique described in Section IV should be used. For the case of WBCD, this technique indicates that the stopping point should be based on the number of hyperplanes required to completely separate all confirmed normals and abnormals, but it is by no means certain that this conclusion would be valid for data sets obtained for other applications. There has been some discussion in the literature as to whether or not "overfitting" the training set causes a problem in the performance of a classifier [8], but in any case, one should try to use the minimum number of hyperplanes consistent with optimal performance on the currently confirmed cases according to an appropriate measure (such as accuracy).

There are techniques in the literature that appear to be competitive alternatives to piecewise-linear classification. An example is composite classification [10]. One way of using such an approach in conjunction with what is described above is to use a few (one or two, say) pairs of hyperplanes produced by our procedure for classification of (most) points and use a more expensive classifier with some desirable properties regarding the probability of error (such as the nearest-neighbor decision rule [2]) for those points for which this partial piecewise-linear classification fails. Such alternatives may turn out to have superior performance to the methods discussed in this paper. Although experimental comparison of all existing methods is a far too time consuming (and not particularly exciting) task, those who have faith in a particular classification method can use the methodology described in this paper to validate their claims and to assign statistical significance to their results.

ACKNOWLEDGMENT

The authors are grateful to O. L. Mangasarian and K. Bennett for providing them with the WBCD, as well as for many detailed discussions, to Y. Censor and I. Kapouleas for comments on an earlier version, and to M. A. Blue and J. Grossman for typing the manuscript.

REFERENCES

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

[2] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972.

[3] O. L. Mangasarian, R. Setiono, and W. H. Wolberg, "Pattern recognition via linear programming: Theory and applications to medical diagnosis," in *Large-Scale Numerical Optimization* (T. F. Coleman and Y. Li, Eds.). Philadelphia: SIAM, 1990, pp. 22–31.

[4] I. Foroutan and J. Sklansky, "Feature selection for piecewise linear classifiers," in *IEEE Proc. Comput. Vision Pattern Recog.* (San Francisco), 1985, pp. 149–154.

[5] E. Harth, T. Kalogeropoulos, and A. S. Pandya, "A universal optimization network," in *Proc. Spec. Symp. Maturing Technol. Emerging Horizons Biomed. Eng.* (New Orleans), 1988, pp. 97–107.

[6] S. M. Weiss and I. Kapouleas, "An empirical comparison of pattern recognition, neural nets, and machine learning classification methods," in *Proc. 11th Int. Joint Conf. Artificial Intell.* (Detroit, MI), 1989, pp. 781–787.

[7] R. F. Mould, *Introduction to Medical Statistics.* Bristol, England: Adam Hilger, 1989, 2nd ed.

[8] J. W. Shavlik, R. J. Mooney, and G. G. Towell, "Symbolic and neural learning algorithms: An experimental comparison," to be published in *Machine Learning.*

[9] G. T. Herman and K. T. D. Yeung, "Evaluators of image reconstruction algorithms," *Int. J. Imag. Syst. Techn.,* vol. 1, pp. 187–195, 1989.

[10] B. V. Basarathy and B. V. Sheela, "A composite classifier system design: Concept and methodology," *Proc. IEEE,* vol. 67, pp. 708–713, 1979.