Available online at www.sciencedirect.com

ScienceDirect

European Journal of Operational Research 181 (2007) 842–854

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

Computing, Artificial Intelligence and Information Management

Breast cancer prediction using the isotonic separation technique

Young U. Ryu [a,*], R. Chandrasekaran [b], Varghese S. Jacob [c]

[a] School of Management, The University of Texas at Dallas, P.O. Box 830688, SM 33, Richardson, TX 75083-0688, USA
[b] School of Engineering and Computer Science, The University of Texas at Dallas, P.O. Box 830688, Richardson, TX 75083-0688, USA
[c] School of Management, The University of Texas at Dallas, P.O. Box 830688, SM 40, Richardson, TX 75083-0688, USA

Received 18 February 2005; accepted 5 June 2006
Available online 7 September 2006

Abstract

A recently developed data separation/classification method, called isotonic separation, is applied to breast cancer prediction. Two breast cancer data sets, one with clean and sufficient data and the other with insufficient data, are used for the study and the results are compared against those of decision tree induction methods, linear programming discrimination methods, learning vector quantization, support vector machines, adaptive boosting, and other methods. The experiment results show that isotonic separation is a viable and useful tool for data classification in the medical domain.

Keywords: Data mining; Isotonic separation; Breast cancer diagnosis

1. Introduction

The task of data classification using knowledge obtained from known historical data has been one of the most intensively studied subjects in statistics, decision science, operations research, and computer science. It has been applied in problems of medicine, social science, management, and engineering. For instance, the StatLog project (Michie et al., 1994) and its extension (Lim et al., 2000) evaluated a number of separation/classification techniques applied in various problem domains such as disease diagnosis, image recognition, and credit evaluation. Linear programming approaches (Smith, 1968; Grinold, 1972; Freed and Glover, 1981; Bennett and Mangasarian, 1992) were also verified to be efficient and effective methods in medical and other domains. Currently support vector machines (Burges, 1998; Cristianini and Shawe-Taylor, 2000; Müller et al., 2001; Schölkopf et al., 1999; Vapnik, 1998, 2000) and AdaBoost (Freund and Schapire, 1997; Schapire, 1999; Schapire and Singer, 1999, 2000) are gaining attentions as techniques for solving classification and regression problems.

This paper applies a recently developed linear programming method called isotonic separation (Chandrasekaran et al., 2005) to breast cancer prediction, that is, categorical (binary in this case) prediction of breast cancer diagnosis. A critical assumption of the technique is monotonic consistency within a set of data points, which in turn establishes a partial order on data points. Such a

[*] Corresponding author. Tel.: +1 972 883 4065; fax: +1 972 883 2089.

E-mail addresses: ryoung@utdallas.edu (Y.U. Ryu), chandra@utdallas.edu (R. Chandrasekaran), vjacob@utdallas.edu (V.S. Jacob).

0377-2217/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2006.06.031

consistency condition may be known in advance. In the case of the Wisconsin breast cancer data set tested in the paper, if a patient is known to have a malignant tumor with a certain epithelial cell size, then another patient's tumor with a bigger epithelial cell size is diagnosed to be malignant, when all other features are identical. If no isotonic consistency condition is known, as in the case of the Ljubljana breast cancer data set also tested in the paper, it must be obtained from the data set.

The two cancer data sets (i.e., Wisconsin and Ljubljana breast cancer data sets) used in the paper are similar in that both have binary class variables. Besides the fact that one is a cancer diagnosis data set and the other is a cancer recurrence data set, they differ in that the Wisconsin data set has sufficient features to induce accurate classification systems, whereas the Ljubljana data set contains ambiguous data (Congdon, 2000) and the provided features are insufficient (Clark and Niblett, 1987). Following Lim et al. (2000) study, we also tested the method on the Wisconsin cancer data with artificially added noisy features. The testing was done under three possible scenarios: a clean data set with sufficient features, a data set with not only sufficient features but also noisy features, and a data set with insufficient features.

The isotonic separation method tested on the two data sets is compared against others, some of which were reported in previous studies and others are tested in this paper. The experiment results show that the isotonic separation method has lower testing error rates than all other methods and is statistically validated to be better than many of them.

2. Two-category isotonic separation

Consider a set P of data points in a d-dimensional space $\mathbb{R}^d$ that is partitioned into two disjoint sets B of blue points and R of red points. Given these two classes of points, we would like to design a system that can learn to separate these points as either blue or red with a minimum of errors. In this paper we propose a classification scheme that assumes the data set satisfies an isotonic consistency condition. The isotonic consistency condition yields a partial ordering relation S on $\mathbb{R}^d$. That is, $S = \{(i,j): \mathbf{a}_i \geqslant \mathbf{a}_j\}$ contains ordered pairs of data points for which if i is classified as red then j must be classified as red, and conversely, if j is classified as blue, then i must be classified as blue. Here, $\mathbf{a}_i$ and $\mathbf{a}_j$ are the coordinate vectors of i and j, respectively, and $\mathbf{a}_i \geqslant \mathbf{a}_j$ if and only if $a_{ki} \geqslant a_{kj}$ for all $k = 1, 2, \ldots, d$, where a_{ki} and a_{kj} are the kth element of $\mathbf{a}_i$ and $\mathbf{a}_j$, respectively.

For example, in the case of breast cancer diagnosis (Mangasarian et al., 1990, 1995), if fine needle aspirates taken from subjects with certain values of clump thickness, uniformity of cell size, uniformity of cell shape, etc. are diagnosed malignant, then those from other subjects with the same or higher values must be diagnosed malignant. In the case of firm bankruptcy prediction (Altman, 1968), if firms with certain values of capital-to-asset, earning-to-asset, and equity-to-debt ratios are predicted to go bankrupt, then other firms with the same or lower values must be predicted to go bankrupt.

Our analysis also takes into account penalties for misclassification. We define the following misclassification penalties: $\alpha > 0$ for each blue point that is classified as red by the system, and $\beta > 0$ for each red point that is classified as blue by the system. The proposed isotonic separation technique minimizes the total misclassification penalties, i.e., $\alpha n_1 + \beta n_2$ where n_1 is the number of misclassified blue points and n_2 is the number of misclassified red points.

Isotonic separation is closely related to isotonic regression (Gebhardt, 1970; Barlow et al., 1972; Dykstra and Robertson, 1982; Block et al., 1994). In both methods, isotonic consistency conditions are the main constraints. However, the main difference is that while isotonic separation minimizes numbers of misclassified categorical data points, isotonic regression minimizes errors measured as the distances between the actual outcomes and predicted outcomes of non-categorical data points. Other methods that use the isotonic consistency conditions include the monotonic decision tree induction technique (Ben-David, 1995), the UTA-DIS (Jacquet-Lagrèze, 1995), and the dominance-based rough set approach (Greco et al., 1998). A monotonic decision tree is built in such a way that monotonicity between branching conditions and the outcome preference is maintained. The UTA-DIS is a multicriteria decision making method in which utilities of multiple criteria are aggregated while their monotonicity aspects are explicitly maintained. The dominance-based rough set approach is a variation of the rough set method based on monotonic dominance relation over order attribute domains. These methods and isotonic separation all utilize the known nature of monotonicity of attributes or criteria but in different manners.

2.1. The mathematical model

It is possible that multiple data points have the same set of coordinates; furthermore, some of these points could be blue belonging to B and the others red belonging to R. We will consider all such data points as one data point i such that r_i and b_i indicate the actual numbers of red and blue points, respectively, represented by i. Further, suppose there exist only n distinct data points (i.e., those with different sets of coordinates). Define a variable π_i for each data point i:

$$\pi_i \in \{0,1\} \quad \text{for } 1 \leqslant i \leqslant n, \tag{2.1}$$

meaning that

$$\pi_i = \begin{cases} 1 & \text{if } i \text{ is classified as blue by the system,} \\ 0 & \text{otherwise.} \end{cases}$$

As a result of the isotonic consistency condition, we have the following consistency constraints:

$$\pi_i - \pi_j \geqslant 0 \quad \text{for } (i,j) \in S. \tag{2.2}$$

If i is classified as red by the system (i.e., $\pi_i = 0$), then there will be a misclassification penalty of $b_i\alpha$. Similarly, if i is classified as blue (i.e., $\pi_i = 1$), then there will be a misclassification penalty of $r_i\beta$. These lead to the objective function

$$\text{minimize} \quad \sum_{i=1}^{n} b_i\alpha(1-\pi_i) + \sum_{i=1}^{n} r_i\beta\pi_i,$$

or equivalently,

$$\text{minimize} \quad \sum_{i=1}^{n} (-b_i\alpha + r_i\beta)\pi_i. \tag{2.3}$$

Even though the above formulation of (2.1)–(2.3) appears to be an integer program, the constraint matrix of (2.2) consists of only 1, 0, and −1, and thus is totally unimodular (Hoffman and Kruskal, 1956; Papadimitriou and Steiglitz, 1998). Therefore, we can drop the integer requirement in (2.1)

$$0 \leqslant \pi_i \leqslant 1 \quad \text{for } 1 \leqslant i \leqslant n, \tag{2.1$'$}$$

and still get integer solutions (Murty, 1976; Shapiro, 1979).

2.2. Separation of the d-dimensional space

As discussed above, the system is able to effectively separate the given data. However, we perceive the value of such a system as being able to classify new points once it has been trained on the given data set. Thus, in this section, we provide a framework to partition the d-dimensional space $\mathbb{R}^d$. New data points, therefore, can be classified based on which area they are located in. Let $\{\pi_i^* : 1 \leqslant i \leqslant n\}$ be an optimal solution to the problem of (2.1′), (2.2), and (2.3). We can divide the d-dimensional space $\mathbb{R}^d$ into three regions:

$$\mathscr{S}_{\mathrm{r}} = \{p \in \mathbb{R}^d : (i,p) \in S \quad \text{for some } \pi_i^* = 0\}, \tag{2.4a}$$

$$\mathscr{S}_{\mathrm{b}} = \{p \in \mathbb{R}^d : (p,i) \in S \quad \text{for some } \pi_i^* = 1\}, \tag{2.4b}$$

and

$$\mathscr{S}_{\mathrm{w}} = \mathbb{R}^d \setminus (\mathscr{S}_{\mathrm{r}} \cup \mathscr{S}_{\mathrm{b}}). \tag{2.4c}$$

The region $\mathscr{S}_{\mathrm{b}}$ of (2.4b) is classified as blue, because for every point $p \in \mathscr{S}_{\mathrm{b}}$, there exists $i \in B \cup R$ such that $\pi_i^* = 1$ (i.e., i is classified as blue) and $(p,i) \in S$ (i.e., since i is classified as blue, p must be classified as blue). Similarly, the region $\mathscr{S}_{\mathrm{r}}$ is classified as red. The region $\mathscr{S}_{\mathrm{w}}$, however, is the area in which the training sample set does not have data points and therefore, the system cannot partition the region. If the penalty for a blue point classified as red by the system is substantially greater than that for a red point classified as blue, then test data that lie in $\mathscr{S}_{\mathrm{w}}$ would be classified as blue; in the opposite case, they would be classified as red. Otherwise, a test data point is classified based on its weighted distance from the boundary. Thus, if a test data point in $\mathscr{S}_{\mathrm{w}}$ is closer to the boundary of $\mathscr{S}_{\mathrm{r}}$ than that of $\mathscr{S}_{\mathrm{b}}$, it is classified as red; otherwise, it is classified as blue. This is performed as follows.

Let F_{r} and F_{b} be sets of boundary corner points (i.e., undominated points) of $\mathscr{S}_{\mathrm{r}}$ and $\mathscr{S}_{\mathrm{b}}$, respectively. That is,

$$F_{\mathrm{r}} = \{i : \pi_i^* = 0 \text{ and } \nexists \pi_j^* = 0 \text{ such that } i \neq j \text{ and } (j,i) \in S\},$$
$$F_{\mathrm{b}} = \{i : \pi_i^* = 1 \text{ and } \nexists \pi_j^* = 1 \text{ such that } i \neq j \text{ and } (i,j) \in S\}.$$

When $S = \{(i,j)\text{: } \mathbf{a}_i \geqslant \mathbf{a}_j\}$ where $\mathbf{a}_i$ and $\mathbf{a}_j$ are the coordinate vectors of i and j, a point $p \in \mathscr{S}_{\mathrm{w}}$ (or more generally any point $p \in \mathbb{R}^d$) is classified as blue if

$$\min_{i \in F_{\mathrm{b}}} \left\{ \beta \sum_{h=1}^{d} \max\{a_{hi} - a_{hp}, 0\} \right\} \leqslant \min_{i \in F_{\mathrm{r}}} \left\{ \alpha \sum_{h=1}^{d} \max\{a_{hp} - a_{hi}, 0\} \right\}, \tag{2.5}$$

where $(a_{1p}, a_{2p}, \ldots, a_{dp})$ and $(a_{1i}, a_{2i}, \ldots, a_{di})$ are the coordinate vectors of p and i, respectively; otherwise it is classified as red.

2.3. Reduction of the problem size

It was shown that the isotonic separation technique formulates the classification problem as a linear programming model. The elimination of redundant pairs (i.e., reflexive and transitively implied ones) in S reduces the problem by reducing constraints in the linear programming model. However, a further investigation of the isotonic separation problem allows significant reduction of the problem size. Define a maximal subset $\overline{R}$ of R and a maximal subset $\overline{B}$ of B:

$$\overline{R} = \{i \in R : \nexists j \in B \text{ such that } (i,j) \in S\},$$
$$\overline{B} = \{i \in B : \nexists j \in R \text{ such that } (j,i) \in S\}.$$

That is, $\overline{R}$ is the biggest subset of R such that for $i \in \overline{R}$ there does not exist $j \in B$ with $(i,j) \in S$; similarly, $\overline{B}$ is the biggest subset of B such that for $i \in \overline{B}$ there does not exist $j \in R$ with $(j,i) \in S$. Then, in every optimal solution

$$\pi_i^* = 0 \quad \text{for } i \in \overline{R},$$
$$\pi_i^* = 1 \quad \text{for } i \in \overline{B}.$$

If not, by changing the solution by modifying only these variables, we would get a better solution. As a result of this observation, we can reduce the set of data points P to P':

$$P' = P \setminus (\overline{R} \cup \overline{B}),$$

and then, we can build and solve a linear programming model for P'.

3. Wisconsin breast cancer diagnosis data

The Wisconsin breast cancer data set (Merz and Murphy, 1998) contains 699 data points on fine needle aspirates taken from patients' breasts. Each data point consists of one class variable (indicating benign or malignant tumors) and nine features of clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses measured in the integer range of 1–10, with a higher value corresponding to a more abnormal state of the tumor (Mangasarian et al., 1990). Of 699 data points, 458 were diagnosed to be benign and 241 malignant.

Wolberg and Mangasarian (1990) used first 369 data points (367 points out of 699 breast cancer data points plus two others that were removed later) for the multisurface separation method (summarized in Section 3.5). When 50% of the data points were used for training, two parallel planes were drawn and 6.5% of remaining data for testing were misclassified. When 67% of the data points were used for training, three parallel planes were drawn and 4.1% of remaining data for testing were misclassified. Mangasarian et al. (1990) used all 369 data points as the training data set for the multisurface separation method, which resulted in four parallel planes. When other 45 data points were tested, all were correctly classified. Mangasarian and Wolberg (1990) used the same 369 points for the multisurface separation and reported the testing result of other 70 data points (including the previous 45 points), in which only one was incorrectly classified. Bennett and Mangasarian (1992) used 566 data points of which 67% were used for training and remaining 33% for testing, and reported a 2.56% testing error rate of the robust linear programming discrimination method (summarized in Section 3.3), compared with a 6.10% testing error rate of the multisurface separation method with one pair of parallel planes and a 3.58% testing error rate of Smith's (1968) linear programming discrimination method.

Lim et al. (2000) performed a comprehensive study on 33 data classification methods experimented with sixteen data sets including the Wisconsin breast cancer data. They reported the testing accuracies with provided feature sets and artificially added noisy features on the data set 683 data points (among which 444 are benign and 239 are malignant) using 10-fold cross validation. Among them, the neural network method utilizing the learning vector quantization algorithm (Kohonen, 1992, 2001) tested on the original features performed the best with a 2.78% testing error rate. The Quest classification system (Loh and Shih, 1997) tested on the original features performed the second with a 3.08% testing error rate. When tested with the artificially added noisy features, the Quest classification system performed the best with a 2.93% testing error rate and the flexible discriminant analysis method (Hastie et al., 1994) performed the second with a 3.21% testing error rate.

We replicated Lim et al.'s (2000) 10-fold cross validation experiment environments with the nine original features and with the artificially added nine noisy features. That is, we used exactly the same partition and exactly the same noisy features of Lim et al.'s experiments. Each of those noisy features contained randomly generated integer values between 1 and 10. The disjointing 10 blocks of the partition were randomly generated so that each block contained a same ratio of benign and malignant tumor data. The proposed isotonic separation method, support vector machines (Burges, 1998; Cristianini and Shawe-Taylor, 2000; Müller et al., 2001; Schölkopf et al., 1999; Vapnik, 1998, 2000), Bennett and Mangasarian's (1992) robust linear programming discrimination method, AdaBoost (Freund and Schapire, 1997; Schapire, 1999; Schapire and Singer, 1999, 2000), and Wolberg and Mangasarian's (1990) multisurface separation methods were all tested on the data sets and compared against each other and against methods used in Lim et al.'s (2000) study. The details are as follows.

3.1. Isotonic separation experiments

Isotonic separation experiments started with the known isotonicity: a higher feature value corresponds to a more abnormal state of the breast tumor cells (Mangasarian et al., 1990). Suppose $\mathbf{a}_i$ and $\mathbf{a}_j$ are 9-dimensional (or 18-dimensional in the case of experiments with the noisy features) feature value vectors of data points i and j. This known isotonic consistency condition yields an ordering relation $S = \{(i,j)\colon \mathbf{a}_i \geqslant \mathbf{a}_j\}$. The separation variable π_i for each data point i has a binary value of either 0 or 1, where $\pi_i = 1$ denotes malignancy and $\pi_i = 0$ denotes benignity.

For the task of feature selection, that is, choosing a subset of features that would result in the best prediction, we adopted the backward sequential elimination method (Kittler, 1986; Marill and Green, 1963). During each session of 10-fold cross valid experiments, a 10-fold partition of the training data set was generated and used for feature selection. (Note that this 10-fold partition for feature selection contains only the training data not the testing data.) Using the set of all nine features (or 18 features when noise was added), the isotonic separation testing (i.e., 10-fold cross validation experiments using the partition of the training data set) was performed. Next, each subset containing eight features (or 17 features when noise was added) was used for isotonic separation testing and one with the smallest error rate was chosen. From this chosen set of eight features (or 17 features when noise was added), subsets containing seven features (or 16 features when noise was added) were considered for further testing. This series of testing was performed until only two features were left. Among all of these subsets of features with which isotonic separation testing was performed, one with the best prediction result was selected as the final feature subset.

During the experiments, we noticed that the problem reduction method of Section 2.3 improved the efficiency of isotonic separation significantly. In the 10-fold cross validation experiments on the original data set, the isotonic separation model of Section 2.1 contained 615 variables (2.1′) and 122,651 constrains (2.2) on average. When the problem reduction method of Section 2.3 was applied, the model was reduced to that with 23 variables and 40 constraints on average. During the 10 training sessions with the original data, isotonic separation showed an average of 0.99% misclassification error rate. When the additional noisy features were added, an average of 0.29% of training data were misclassified.

Among 683 data points (among which 444 are benign and 239 are malignant) with the original nine features, when tested by the 10-fold cross validation, 417 data points belonged to the benign tumor area $\mathscr{S}_\mathrm{r}$ (2.4a), 200 data points belonged to the malignant tumor area $\mathscr{S}_\mathrm{b}$ (2.4b), and 66 data points (9.66%) belonged to the unclassified area $\mathscr{S}_\mathrm{w}$ (2.4c). Data points falling in the unclassified area were classified using the criterion of (2.5). Among them, 3 data points in $\mathscr{S}_\mathrm{r}$, 6 data points in $\mathscr{S}_\mathrm{b}$, and 6 data points in $\mathscr{S}_\mathrm{w}$ were misclassified. When the additional nine noisy features were used, 334 data points belonged to the benign tumor area $\mathscr{S}_\mathrm{r}$, 101 data points belonged to the malignant tumor area $\mathscr{S}_\mathrm{b}$, and 248 data points (36.31%) belonged to the unclassified area $\mathscr{S}_\mathrm{w}$. Among them, 2 data points in $\mathscr{S}_\mathrm{r}$, 2 data points in $\mathscr{S}_\mathrm{b}$, and 14 data points in $\mathscr{S}_\mathrm{w}$ were misclassified. That is, with the original nine features, the isotonic separation method showed a 2.20% error rate; with the 18 features including the nine noisy features, it showed a 2.64% error rate.

3.2. Support vector machine experiments

Suppose two sets B and R of malignant and benign data points in a d-dimensional space are given, where $|B \cup R| = n$. (Note, $|\cdots|$ denotes the cardinality of a set.) For each data point $i \in B \cup R$, let $\mathbf{a}_i$ be the vector of its feature values and c_i its class label such that $c_i = 1$ if $i \in B$, and

$c_i = -1$ if $i \in R$. Let $\mathbf{\Phi}$ be a function that transforms d-dimensional space to d'-dimensional space where $d' \geqslant d$. Define K, called a kernel function, to be the scalar product operation of vectors in the d'-dimensional space, that is, $K(\mathbf{x},\mathbf{y}) = \mathbf{\Phi}(\mathbf{x}) \cdot \mathbf{\Phi}(\mathbf{y})$. Typical forms of kernel function used for the support vector machine method include $\mathbf{x}\cdot\mathbf{y}$ (dot product kernel), $\exp\left(-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2/\sigma^2\right)$ (Gaussian, or radial basis function, kernel) where σ is a real number, and $(\mathbf{x}\cdot\mathbf{y}+\theta)^d$ (polynomial kernel) where θ is a real number and the degree term d is an integer. Note, $\|\mathbf{v}\|_n$ denotes the ℓ_n-norm of vector $\mathbf{v}$.

The support vector machine method finds a linear separator $\mathbf{w}\cdot\mathbf{x}+b=0$ for the transformed d'-dimensional space. whose coefficient vector $\mathbf{w}=(w_1,w_2,\ldots,w_{d'})$ and constant b are obtained by solving the following quadratic program:

$$\begin{aligned}
\text{minimize}\quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} y_i\\
\text{subject to}\quad & c_i(\mathbf{w}\cdot\mathbf{\Phi}(\mathbf{a}_i)+b) \geqslant 1-y_i\\
& \quad \text{for } i=1,2,\ldots,n,\\
& y_i \geqslant 0 \quad \text{for } i=1,2,\ldots,n.
\end{aligned}\tag{3.1}$$

Let non-negative λ_i for $i=1,2,\ldots,n$ be Lagrangian multipliers (i.e., dual variables) for the constraints. Then, we have the following dual quadratic program:

$$\begin{aligned}
\text{maximize}\quad & \sum_{i=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j c_i c_j K(\mathbf{a}_i,\mathbf{a}_j)\\
\text{subject to}\quad & \sum_{i=1}^{n}\lambda_i c_i = 0,\\
& 0 \leqslant \lambda_i \leqslant C \quad \text{for } i=1,2,\ldots,n.
\end{aligned}$$

A solution to this dual quadratic program with a specific kernel gives a separator for the d'-dimensional space.

For the support vector machine experiments, we used an implementation based on Joachims' (1999) algorithm. We trained support vector machines using the Gaussian kernel with $\sigma = 6$, the dot product kernel, and the polynomial kernel degree $d = 3$ and $\theta = 1$. The polynomial kernel could find perfect separators in all sessions of 10-fold experiments. When testing data were classified, it resulted in 4.97% testing error rate on the original data set and a 4.27% testing error rate on the data set with the additional noisy features. The Gaussian kernel showed 2.54% and 1.12% training error rates on the data sets without and with noisy features, respectively. When testing data were classified, it resulted in a 2.92% testing error rate on the original data set and a 2.78% testing error rate on the data set with the additional noisy features. Finally, the dot product kernel showed 2.55% and 2.28% training errors on the data sets without and with noisy features, respectively. When testing data were classified, it resulted in a 3.37% testing error rate on the original data set and a 2.78% testing error rate on the data set with the additional noisy features.

3.3. Linear programming discrimination experiments

Linear programming approaches to discriminant analysis (Smith, 1968; Grinold, 1972; Freed and Glover, 1981; Bennett and Mangasarian, 1992) are well-developed data classification methods. Among them, we used the robust linear programming (LP) discrimination method (Bennett and Mangasarian, 1992) which resolves the known problem of the null solution without any extraneous constraint. Given two sets B and R of malignant and benign data points in a d-dimensional space, the robust linear programming discrimination method finds a linear separator for the d-dimensional space:

$$w_1x_1 + w_2x_2 + \cdots + w_dx_d = \gamma \tag{3.2}$$

whose coefficients $w_1, w_2, \ldots, w_d$ and constant γ are obtained by the following linear program:

$$\begin{aligned}
\text{minimize}\quad & \alpha\sum_{i\in R} y_i + \beta\sum_{i\in B} z_i\\
\text{subject to}\quad & \sum_{k=1}^{d} a_{i,k}w_k + y_i \geqslant \gamma+1 \quad \text{for } i\in B,\\
& \sum_{k=1}^{d} a_{j,k}w_k - z_j \leqslant \gamma-1 \quad \text{for } j\in R,\\
& y_i, z_j \geqslant 0 \quad \text{for } i\in B \text{ and } j\in R,
\end{aligned}\tag{3.3}$$

where $(a_{i,1}, a_{i,2}, \ldots, a_{i,d})$ is the d-dimensional coordinate vector of point i. That is, it is expected that all malignant points are in one side of the separator and all benign points in the other side of the separator; however, if such a perfect separation is not possible, the linear program minimizes the total weighted distance of misplaced points. The robust LP method with pooled penalty has $\alpha = \beta$. The robust LP method with averaged penalty has $\alpha = 1/|R|$ and $\beta = 1/|B|$.

The robust LP method with pooled penalty showed a 2.55% training error rate and a 3.37%

testing error rate on the original data set and a 2.28% training error rate and a 2.78% testing error rate on the data set the additional noisy features. The robust LP with averaged penalty showed a 2.64% training error rate and a 2.78% testing error rate on the original data set and a 1.93% training error rate and a 2.93% testing error rate on the data set the additional noisy features.

For the feature selection task, we adopted the backward sequential elimination method (Kittler, 1986; Marill and Green, 1963). The robust LP method with pooled penalty showed a 2.64% training error rate and a 3.22% testing error rate on the original data set with the feature selection and a 2.85% training error rate and a 2.78% testing error rate on the data set with the additional noisy features after the features election. The robust LP method with averaged method showed a 2.73% training error rate and a 2.78% testing error rate on the original data set with the feature selection and a 3.29% training error rate and a 3.36% testing error rate on the data set with the additional noisy features after the features election.

3.4. AdaBoost

AdaBoost (Freund and Schapire, 1997; Schapire, 1999; Schapire and Singer, 1999, 2000) is a boosting method that improves the performance of a weak learning system. It runs the weak learning system multiple times with different weights on training data points. The final classification is done by combining multiple classification results of the weak learning system on the training data set with different weights.

Suppose two sets B and R of malignant and benign data points in a d-dimensional space are given, where $|B \cup R| = n$. For each data point $i \in B \cup R$, let $\mathbf{a}_i$ be the vector of its feature values and c_i its class label such that $c_i = 1$ if $i \in B$, and $c_i = -1$ if $i \in R$. Consider the training of a learning system over T rounds. At round t, each training data point i is associated with a weight given by $D_t(i)$. At the initial round, $D_1(i) = 1/n$. Let

$$h_t : B \cup R \to \mathbb{R}$$

be the classification function of the learning system obtained at t. Weights of training data points at the next round are adjusted as follows. Let

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 + \sum_{i=1}^{n} D_t(i) c_i h_t(\mathbf{a}_i)}{1 - \sum_{i=1}^{n} D_t(i) c_i h_t(\mathbf{a}_i)} \right).$$

Then, weights for round $t+1$ is obtained as:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t c_i h_t(\mathbf{a}_i))}{Z_t},$$

where Z_t is a normalization factor. As a result, the weights of incorrectly classified data points are increased. Once h_t for $t = 1, 2, \ldots,$ and T are obtained, the final classification function H is defined as follows:

$$H(\mathbf{a}_i) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{a}_i) \right).$$

Experiments were performed with 500 boosting rounds using a simple one-level decision tree (Schapire and Singer, 2000). (The number of boosting rounds was chosen during the training phase. We tried various numbers of boosting rounds and chose one with the lowest training error. The level of decision tree was chosen similarly.) When the classifier was trained, it was able to find a perfect separator for both the original data and the data with additional noisy features. When the separator was tested, it resulted in a 3.66% testing error rate on the original data set and a 5.12% testing error rate on the data set with the additional noisy features.

3.5. Multisurface separation experiments

The multisurface separation method (Mangasarian, 1968; Mangasarian et al., 1990; Wolberg and Mangasarian, 1990) generates, until all data points are separated, multiple pairs of parallel planes in the d-dimensional space such that one side of the first plane contains all malignant data points and the other side of the second plane contains all benign data points. If the area containing all malignant data points and the area containing all benign points must overlap, either a closest pair of parallel planes or those with least data points between them are selected. We call the first method MSM-d (“d” indicating a distance measure) and the second method MSM-c (“c” indicating a count measure). A pair of such parallel planes can be obtained by solving $2d$ linear programs.

The multisurface separation algorithm runs until all training data are separated. Its training error becomes always 0. But, its testing performance is not always good mainly because the classifier can be overtrained. The testing error rates are as follows. When tested on the 10-fold partition of the breast cancer data set, MSM-d showed a 7.01%

error rate on the original data set and a 7.03% error rate on the data set with the additional noisy features; MSM-c showed a 4.83% error rate on the original data set and a 6.00% error rate on the data set with the additional noisy features.

The Wisconsin breast cancer diagnosis data set is known to have one outlying data point. (The outlying data point is a benign data point with clump thickness 8, uniformity of cell size 4, uniformity of cell shape 4, marginal adhesion 5, single epithelial cell size 4, bare nuclei 7, bland chromatin 7, normal nucleoli 8, and mitoses 2.) Since the multisurface separation method attempts to achieve a perfect separation during training, there might be a substantial negative effect from this outlying data point. In order to see the performance of the multisurface method on a cleaner data set, we removed this outlying data point and conducted the experiments. Then, MSM-d showed a 6.62% error rate on the original data set and a 6.01% error rate on the data set with the additional noisy features; MSM-c showed a 3.66% error rate on the original data set and a 4.27% error rate on the data set with the additional noisy features.

3.6. Experiment results

Table 1 shows average training error rates and standard deviations of classification methods obtained from the 10-fold cross validation experiment. Table 2 contains testing results of our experiments as well as best performers of Lim et al.'s (2000) experiments. The testing error rate of the isotonic separation method was smaller than those of other methods. The statistical validation summarized in Table 3 shows that isotonic separation on the original data set was better than most other methods, though significant performance difference was not observed (except in the case of comparisons with support vector machines with a polynomial kernel and AdaBoost) when they were tested on the data set with additional noisy features.

Interestingly, some methods performed better when noisy features containing randomly generated values were added. They include support vector machines and the robust LP method with pooled penalty in our experiments and Quest, radial basis function neural networks, discriminant analysis with the Gaussian mixture function, and the OC1 decision tree induction method in Lim et al.'s (2000) experiments. In order to further investigate this phenomenon, we generated five sets of nine attributes containing random values between 1 and 10 and performed experiments on them. That is, instead of Lim et al.'s noisy features, we used each of our own five sets of noisy features. The experiment results, summarized in Table 4, show that the improvement of testing accuracy when noisy features were added was incidental. Support vector machines with the dot kernel and the polynomial kernels tested on the data sets with our noisy features improved the testing accuracies in three cases out of five; the robust LP method with pooled penalty improved the testing accuracies in two cases; and support vector machines with the Gaussian kernel did not improve the testing accuracies. Other methods tested on the data sets with our noisy features did not show improved testing accuracies.

Table 1
Wisconsin breast cancer diagnosis experiment results: training error rates

Methods	Training error rates (%)	
	Original data	Data with noise
SVM[b] with		
Polynomial kernel	0.00 (0.00)[a]	0.00 (0.00)
Gaussian kernel	2.54 (0.29)	1.12 (0.18)
Dot product kernel	2.55 (0.24)	2.28 (0.21)
AdaBoost	0.00 (0.00)	0.00 (0.00)
Isotonic separation	0.99 (0.15)	0.29 (0.10)
Robust LP-A[c]		
Before feature selection	2.46 (0.24)	1.93 (0.26)
After feature selection	2.73 (0.20)	3.29 (0.35)
Robust LP-P[d]		
Before feature selection	2.55 (0.24)	2.28 (0.21)
After feature selection	2.64 (0.24)	2.83 (0.34)

[a] Values in (): standard deviations of error rates over 10-fold experiments.
[b] Support vector machine: polynomial kernel with $d = 3$, Gaussian kernel with $\sigma = 6$.
[c] Robust linear programming with averaged penalty.
[d] Robust linear programming with pooled penalty.

4. Ljubljana breast cancer recurrence data

The Ljubljana breast cancer data set (Merz and Murphy, 1998) contains 286 data points on the recurrence of breast cancer within five years after surgical removal of tumor. Each data point consists of one class variable (for recurrence or non-recurrence) and nine features of age, menopause status, tumor size, invasive nodes, node caps, degree of malignancy, breast, breast quadrant, and irradiation. All features except menopause status and breast quadrant are have either binary or ordered

Table 2
Wisconsin breast cancer diagnosis experiment results: testing error rates

Methods	Testing error rates (%)	
	Original data	Data with noise
Isotonic separation	2.20 (1.35)[a]	2.64 (1.71)
Robust LP-A[b]		
Before feature selection	2.78 (2.78)	2.93 (1.32)
After feature selection	2.78 (2.21)	3.37 (2.63)
SVM[c] with		
Gaussian kernel	2.93 (1.97)	2.78 (1.37)
Dot product kernel	3.37 (2.09)	2.78 (1.03)
Polynomial kernel	4.97 (2.45)	4.27 (1.22)
Robust LP-P[d]		
Before feature selection	3.37 (2.09)	2.78 (1.03)
After feature selection	3.22 (1.95)	2.78 (2.13)
AdaBoost	3.66 (2.18)	5.12 (2.01)
MSM-c[e]		
With all data	4.83 (2.12)	6.00 (2.95)
W/o outlier	3.66 (1.78)	4.27 (2.07)
MSM-d[f]		
With all data	7.01 (3.17)	7.03 (3.13)
W/o outlier	6.62 (2.88)	6.01 (3.02)
Learning vector quantization[g]	2.78 (N/A)	
Quest[g]	3.08 (N/A)	2.93 (N/A)
Flexible discriminant analysis[g]		3.21 (N/A)

[a] Values in (): standard deviations of error rates over 10-fold experiments.
[b] Robust linear programming with averaged penalty.
[c] Support vector machine: Gaussian kernel with $\sigma = 6$, polynomial kernel with $d = 3$.
[d] Robust linear programming with pooled penalty.
[e] MSM-c: multisurface method with count measure.
[f] MSM-d: multisurface method with distance measure.
[g] Top two performers of Lim et al.'s (2000) experiments.

Table 3
Wisconsin data statistical validation against isotonic separation

Methods	Probability of t-test	
	Original data	Data with noise
Robust LP-A		
Before feature selection	0.023^{**}	0.264
After feature selection	0.051^{*}	0.066^{*}
SVM with		
Gaussian kernel	0.029^{**}	0.370
Dot product kernel	0.002^{***}	0.370
Polynomial kernel	0.000^{***}	0.000^{***}
Robust LP-P		
Before feature selection	0.002^{***}	0.370
After feature selection	0.004^{***}	0.370
AdaBoost	0.009^{***}	0.000^{***}

* Marginally significant ($p < 0.1$).
** Significant ($p < 0.05$).
*** Very significant ($p < 0.01$).

Table 4
Testing error rates in experiments on data with noise

Data sets	Support vector machines (%)			Robust LP-P (%)
	Gaussian	Dot product	Polynomial	
Original data	2.92	3.37	4.97	3.37
Noise-0[a]	2.78	2.78	4.27	2.78
Noise-1[b]	3.66	3.81	4.10	3.95
Noise-2[b]	3.22	2.93	3.95	2.93
Noise-3[b]	3.22	3.07	4.54	3.07
Noise-4[b]	3.51	3.51	5.12	3.51
Noise-5[b]	3.07	3.22	5.56	3.22

[a] Original data + Lim et al.'s (2000) nine noisy features.
[b] Original data + randomly generated nine noisy features.

values. The menopause status feature has values of "pre-menopause," "menopause before age 40," and "menopause at or after age 40." For experiments with isotonic separation, linear programming discrimination analysis, support vector machines, and learning vector quantization, we considered three possible ordered value assignments: that of 1, 2, and 3; that of 1, 3, and 2; and that of 1, 2, and 2. The feature of breast quadrant having five categorical values for detailed locations of tumor was not used for isotonic separation and linear programming discrimination analysis experiments. Of 286 data points, 201 are non-recurrent data and 85 are recurrent data.

The Ljubljana breast cancer data set is known to be very difficult to deal with. The data set is ambiguous (Congdon, 2000) in that some data points with same features have different class values. The provided features are not sufficient (Clark and Niblett, 1987) to produce high quality prediction systems. Assist 86 (Cestnik et al., 1987), a top-down decision tree induction method with improvements over ID3 (Quinlan, 1979, 1986), showed a 22% error rate tested on 30% of the data with a decision tree built with 70% of the data. IWN (Clark and Niblett, 1987), a bottom-up induction method of a network of multiple trees, showed a 26% testing error rate in a similar experiment environment. A genetic algorithm method (Congdon, 2000) showed a 28% error rate on average in five experimental runs with randomly selected 80% training and 20% testing data sets.

On this data set, we experimented the isotonic separation method, the robust linear programming method (Bennett and Mangasarian, 1992), the ID3/C4.5 decision tree induction system (Quinlan, 1993), the OC1 decision tree induction system (Murthy et al., 1994), support vector machines (Burges, 1998; Cristianini and Shawe-Taylor, 2000;

Müller et al., 2001; Schölkopf et al., 1999; Vapnik, 1998, 2000), AdaBoost (Freund and Schapire, 1997; Schapire, 1999; Schapire and Singer, 1999, 2000), and learning vector quantization (Kohonen, 1992, 2001) on a 70%/30% partition of the data set similar to the experiment environments for Assist 86 and IWN, and the Quest classification system (Loh and Shih, 1997) with its own built-in 3-fold cross validation option. The details are as follows.

4.1. Isotonic separation experiments

Not like for the Wisconsin data set used in the previous section, no isotonic consistency condition is known for the Ljubljana breast cancer data set. Thus, the feature selection operation for isotonic separation included isotonic consistency condition testing. To perform feature selection and isotonicity testing, we applied the backward sequential elimination method (Kittler, 1986; Marill and Green, 1963) on a string of 16 binary bits such that 2 bits determined the relevance and isotonicity of each feature. If a pair of bits are both 1s or 0s, the corresponding feature was considered to be relevant with positive isotonicity. If the pair of bits are 0 and 1, the corresponding feature was considered to be relevant with negative isotonicity. If the pair of bits are 1 and 0, the corresponding feature was considered to be irrelevant. For a feature with negative isotonicity, we multiplied feature values in the data set by -1; for a feature to be dropped, we multiplied feature values by 0.

The feature selection process reported that age, menopause status, node caps, degree of malignancy, and irradiation were relevant with positive isotonicity. For instance, the older is a patient, the more likely does cancer recur. For numeric value assignment of the menopause status feature, the process found that the assignment of 1, 2, and 3 to "premenopause," "menopause before age 40," and "menopause at or after age 40" was the best.

In the testing experiment on 87 data points (among which 61 are non-recurrent and 26 are recurrent), 70 data points belonged to the non-recurrence area $\mathscr{S}_{\mathrm{r}}$ (2.4a), 14 data points belonged to the recurrence area $\mathscr{S}_{\mathrm{b}}$ (2.4b), and 3 data points (3.45%) belonged to the unclassified area $\mathscr{S}_{\mathrm{w}}$ (2.4c). Data points falling in the unclassified area were classified using the criterion of (2.5). Among them, 14 data points in $\mathscr{S}_{\mathrm{r}}$, 2 data points in $\mathscr{S}_{\mathrm{b}}$, and 1 data point in $\mathscr{S}_{\mathrm{w}}$ were misclassified. That is, the isotonic separation method showed a 20% testing error rate.

4.2. Decision tree induction experiments

ID3/C4.5 (Quinlan, 1979, 1986) is a top-down decision tree induction method based on the idea of reducing entropy (Shannon, 1948). Given a set of data point A, a decision tree is induced by sequentially (i.e., from the root to leaves) placing at nodes the kth feature that results in the most reduction of uncertainty of class variable y, where reduction of uncertainty is measured as the difference between the entropy of A for y and the conditional entropy of A for y when the kth feature value is known. The ID3/C4.5 method induces an axis-parallel decision tree, in which each node contains one feature variable and branches from the node have equality or inequality conditions on the feature variable. OC1 (Murthy et al., 1994), another top-down decision tree induction method, generates an oblique decision tree, in which each node contains a hyperplane separating the d-dimensional feature space and each of its subsequent nodes further separates a half space. The hyperplane is derived by considering a certain impurity measure such as uncertainty reduction based on entropy. Quest (Loh and Shih, 1997) is also a top-town decision tree induction system, in which branches at a node of the decision tree split based on a form of quadratic discriminant analysis.

Experiments on ID3/C4.5, OC1, and Quest decision tree induction methods resulted in 28%, 25%, and 27% testing error rates, respectively.

4.3. Support vector machine experiments

Using the support vector machine method summarized in Section 3.2, we performed training and testing on the 70%/30% partition of the data set. The polynomial kernel with degree $d = 1$ and $\theta = 1$ resulted in a 26% error rate. The Gaussian kernel with $\sigma = 6$ resulted in a 28% testing error rate. The dot kernel resulted in a 29% testing error rate.

4.4. Linear programming discrimination experiments

Using the linear programming discrimination method summarized in Section 3.3, we performed training and testing on the 70%/30% partition of the data set. The robust linear programming method ($\alpha = 1/|R|$ and $\beta = 1/|B|$) showed a 37% error rate. When the backward sequential elimination method (Kittler, 1986; Marill and Green,

1963) was applied prior to training, the robust LP method with averaged penalty show a 30% testing error rate. The robust LP method with pooled penalty ($\alpha = \beta = 1$) showed a 28% testing error rate both with and without feature selection.

4.5. Learning vector quantization experiments

Suppose two sets B and R of recurrent and non-recurrent data points in a d-dimensional space are given, where $|B \cup R| = n$. For each data point $i \in B \cup R$, define its class label c_i such that $c_i = 1$ if $i \in B$, and $c_i = 0$ if $i \in R$. Learning vector quantization (Kohonen, 1992; Kohonen, 2001) is a competitive learning method that separates data points into m clusters. Let $\mathbf{w}_k$ be the coordinate vector representing the kth cluster and c_k be the class label of the cluster. That is, if $c_k = 1$ then the kth cluster belongs to the class of recurrent data, and if $c_k = 0$ then it belongs the class of benign data. We will call $(\mathbf{w}_k|c_k)$ the codebook vector of the kth cluster. When all such codebook vectors are learned from the given data points, the d-dimensional space is partitioned by Voronoi tessellation.

The learning algorithm starts at time $t = 1$ with initial m codebook vectors which can be chosen randomly from the given data points (or by some simple observations of the given data (Kohonen, 2001)). Let $\mathbf{m}_k(t)$ denote the codebook vector of the kth cluster at time t. For a data point i whose coordinate vector is $\mathbf{a}_i$, find a nearest codebook vector:

$$k^* = \arg\min_{1 \leqslant k \leqslant m} \|\mathbf{w}_k - \mathbf{a}_i\|_2. \tag{4.1}$$

Then, the codebook vector of the k^*th cluster at time $t + 1$ becomes

$$\mathbf{m}_{k^*}(t+1) = (\mathbf{w}_{k^*} + s\epsilon(\mathbf{a}_i - \mathbf{w}_{k^*})|c_{k^*}),$$

where $0 < \epsilon < 1$, and $s = 1$ if $c_{k^*} = c_i$ and $s = -1$ if $c_{k^*} \neq c_i$. This process is repeated until a stopping criterion (e.g., a number of iterations or a threshold change in codebook vectors (Kohonen, 2001)) is met. That is, if $c_{k^*} = c_i$, then the codebook vector of the k^*th cluster moves toward i, and if $c_{k^*} \neq c_i$, then it moves away from i.

When 20 vectors were learned over 500 iterations during training, the method resulted in a 29% testing error rate.

4.6. AdaBoost experiments

We performed experiments with AdaBoost, summarized in Section 3.4, using a simple one-level decision tree (Schapire and Singer, 2000). The experiment with 1000 boosting rounds resulted in the testing (error rate of 30%). (The number of boosting rounds was chosen during the training phase. We tried various numbers of boosting rounds and chose one with the lowest training error. The level of decision tree was chosen similarly.)

Table 5
Ljubljana breast cancer recurrence experiment results: training error rates

Methods	Training error rates (%)
SVM[a] with	
Polynomial kernel	15
Gaussian kernel	23
Dot kernel	28
Isotonic separation	21
AdaBoost	21
OC1	21
Learning vector quantization	23
Quest	24
ID3/C4.5	25
Robust LP-P	28
Robust LP-A	
Before feature selection	32
After feature selection	37

[a] Support vector machines: polynomial kernel with $d = 2$ and Gaussian kernel with $\sigma = 6$.

Table 6
Ljubljana breast cancer recurrence experiment results: testing error rates

Methods	Testing error rates (%)
Isotonic separation	20
OC1	25
SVM[a] with	
Polynomial kernel	26
Gaussian kernel	28
Dot kernel	29
Quest	27
ID3/C4.5	28
Robust LP-P	29
Learning vector quantization	29
AdaBoost	30
Robust LP-A	
Before feature selection	37
After feature selection	30
Assist 86	22[b]
IWN	26[c]
GA	28[d]

[a] Support vector machines: polynomial kernel with $d = 2$ and Gaussian kernel with $\sigma = 6$.

[b] Result cited from Cestnik et al. (1987).

[c] Result cited from Clark and Niblett (1987).

[d] Result cited from Congdon (2000).

Table 7
Ljubljana data statistical validation against isotonic separation

Methods	Probability of t-test
OC1	0.048**
SVM with	
Polynomial kernel	0.100*
Gaussian kernel	0.026**
Dot kernel	0.010**
ID3/C4.5	0.026**
Robust LP-P	0.017**
Neural nets with LVQ	0.022**
AdaBoost	0.019**
Robust LP-A	
Before feature selection	0.002***
After feature selection	0.010***

* Marginally significant ($p < 0.1$).
** Significant ($p < 0.05$).
*** Very significant ($p < 0.01$).

4.7. Experiment results

Table 5 shows training error rates of classifiers. Table 6 summarizes the results of all experiments performed for this paper and previously reported experiments and Table 7 contains statistical validation results of comparison between isotonic separation and other methods. Isotonic separation achieved the lowest testing error rate and was validated to be better than most other methods when tested on the Ljubljana breast cancer recurrence data set.

5. Conclusion

Isotonic separation, a linear programming method for data separation and classification, was applied to breast cancer prediction. The results of experiments on the two data sets showed that isotonic separation performed better than most other methods tested in this paper and previously published research reports. This signifies that isotonic separation is a viable and useful tool for data classification in the medical domain of breast cancer prediction.

For feature selection and isotonic consistency condition testing when the condition is unknown, we applied the backward sequential elimination heuristics (Kittler, 1986; Marill and Green, 1963) in this paper. We are investigating other approaches, such as an algebraic method and genetic algorithms. The algebraic method results in a mixed integer programming formulation, which is known to be computationally expensive to solve. We plan to use various relaxation techniques to reduce the computational complexity. With genetic algorithms, we hope the feature search space is more extensively but reasonably efficiently examined. The results of these approaches will be discussed in subsequent research reports.

Additional experiments are currently conducted to evaluate the applicability and usefulness of isotonic separation in other domains including organ transplant patient survival prediction. Preliminary experiment results show high accuracy rates. More intensive studies and comparisons against other methods will be reported in the future.

References

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance 23 (4), 589–609.

Barlow, R.E., Bartholemew, D.J., Bremner, J.M., Brunk, H.D., 1972. Statistical Inference under Order Restrictions. John Wiley & Sons, New York.

Ben-David, A., 1995. Monotonicity maintenance in information-theoretic machine learning algorithms. Machine Learning 19, 29–43.

Bennett, K.P., Mangasarian, O.L., 1992. Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods and Software 1, 23–34.

Block, H., Qian, S., Sampson, A., 1994. Structure algorithms for partially ordered isotonic regression. Journal of Computational and Graphical Statistics 3 (3), 285–300.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167.

Cestnik, B., Kononenko, I., Bratko, I., 1987. Assistant 86: A knowledge-elicitation tool for sophisticated users. In: Bratko, N., Lavrač, N. (Eds.), Progress in Machine Learning: Proceedings of the Second European Working Session on Learning (EWSL-87). Sigma Press, pp. 31–45.

Chandrasekaran, R., Ryu, Y.U., Jacob, V., Hong, S., 2005. Isotonic separation. INFORMS Journal on Computing 17 (4), 462–474.

Clark, P., Niblett, T., 1987. Induction in noisy domains. In: Bratko, I., Lavrač, N. (Eds.), Progress in Machine Learning: Proceedings of the Second European Working Session on Learning (EWSL-87). Sigma Press, pp. 11–30.

Congdon, C.B., 2000. Classification of epidemiological data: A comparison of genetic algorithm and decision tree approaches. In: Proceedings of the 2000 Congress on Evolutionary Computation (CEC00). IEEE Press, pp. 442–449.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge.

Dykstra, R.L., Robertson, T., 1982. An algorithm for isotonic regression of two or more independent variables. The Annals of Statistics 10, 708–711.

Freed, E., Glover, F., 1981. A linear programming approach to the discriminant problem. Decision Sciences 12 (1), 68–74.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1), 119–139.

Gebhardt, F., 1970. An algorithm for monotone regression with one or more independent variables. Biometrika 57, 263–271.

Greco, S., Matarazzo, B., Słowiński, R., 1998. A new rough set approach to evaluation of bankruptcy risk. In: Zopounidis, C. (Ed.), Operational Tools in Management of Financial Risks, second ed. Kluwer, Dordrecht, pp. 121–136.

Grinold, R.C., 1972. Mathematical programming methods of pattern classification. Management Science 19 (3), 272–289.

Hastie, T., Tibshirani, R., Buja, A., 1994. Flexible discriminant analysis by optimal scoring. Journal of the American Statistical Association 89, 1255–1270.

Hoffman, A.J., Kruskal, J.B., 1956. Integral boundary point of convex polyhedra. In: Kuhn, H.W., Tucker, A.W. (Eds.), Linear Inequalities and Related Systems, vol. 33. Princeton University Press, Princeton, NJ, pp. 223–246.

Jacquet-Lagrèze, E., 1995. An application of the UTA discriminant model for the evaluation of R & D projects. In: Pardolos, Y., Siskos, Y. (Eds.), Advances in Multicriteria Analysis. Kluwer Academic Publisher, Dordrecht, pp. 203–211.

Joachims, T., 1999. Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (Eds.), Advances in Kernel Methods: Support Vector Learning. The MIT Press, Cambridge, MA, pp. 169–184.

Kittler, J., 1986. Feature selection and extraction. In: Young, K.S., Fu, K.S. (Eds.), Handbook of Pattern Recognition and Image Processing. Academic Press, New York, pp. 59–83.

Kohonen, T., 1992. New developments of learning vector quantization and the self-organizing map. In: Proceedings of the 1992 Symposium on Neural Networks: Alliances and Perspectives in Senri (SYNAPSE'92).

Kohonen, T., 2001. Self-Organizing Maps, third ed. Springer-Verlag, Heidelberg.

Lim, T.-S., Loh, W.-Y., Shih, Y.-S., 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning 4 (3), 203–228.

Loh, W.-Y., Shih, Y.-S., 1997. Split selection methods for classification trees. Statistica Sinica 7, 815–840.

Mangasarian, O.L., 1968. Multisurface method of pattern separation. IEEE Transactions on Information Theory IT-14 (6), 801–807.

Mangasarian, O.L., Wolberg, W.H., 1990. Cancer diagnosis via linear programming. SIAM News 23 (5), 1 and 18.

Mangasarian, O.L., Setiono, R., Wolberg, W.H., 1990. Pattern recognition via linear programming: Theory and application to medical diagnosis. In: Coleman, T.F., Li, Y. (Eds.), Large-Scale Numerical Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 22–31.

Mangasarian, O.L., Street, W.N., Wolberg, W.H., 1995. Breast cancer diagnosis and prognosis via linear programming. Operations Research 43 (4), 570–577.

Marill, T., Green, D.M., 1963. On the effectiveness of receptors in recognition systems. IEEE Transactions on Information Theory 9, 11–17.

Merz, C.J., Murphy, P.M., 1998. UCI Repository of Machine Learning Databases. Department of Information and Computer Sciences, University of California, Irvine.

Michie, D., Spiegelhalter, D.J., Tayor, C.C. (Eds.), 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood, London.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks 12 (2), 182–202.

Murthy, S.K., Kasif, S., Salzberg, S., 1994. A system for induction of oblique decision trees. Journal of Artificial Intelligence Research 2 (1), 1–32.

Murty, K.G., 1976. Linear and Combinatorial Programming. John Wiley & Sons, New York.

Papadimitriou, C.H., Steiglitz, K., 1998. Combinatorial Optimization: Algorithms and Complexity. Dover Publications, Mineola, NY.

Quinlan, J.R., 1979. Discovering rules by induction from large collections of examples. In: Michie, D. (Ed.), Expert Systems in the Micro Electronic Age. Edinburgh University Press, Edinburgh.

Quinlan, J.R., 1986. Induction to decision trees. Machine Learning 1, 81–106.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufman, San Mateo, CA.

Schapire, R.E., 1999. Theoretical views of boosting and applications. In: Proceedings of the 10th International Conference on Algorithmic Learning Theory.

Schapire, R.E., Singer, Y., 1999. Improved boosting algorithms using confidence-rated predictions. Machine Learning 37 (3), 297–336.

Schapire, R.E., Singer, Y., 2000. BoosTexter: A boosting-based system for text categorization. Machine Learning 39 (2–3), 135–168.

Schölkopf, B., Burges, C.J.C., Smola, A.J. (Eds.), 1999. Advances in Kernel Methods: Support Vector Learning. The MIT Press, Cambridge, MA.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423, 623–656.

Shapiro, J.F., 1979. Mathematical Programming: Structures and Algorithms. John Wiley & Sons, New York.

Smith, F.W., 1968. Pattern classifier design by linear programming. IEEE Transactions on Computers C-17 (4), 367–372.

Vapnik, V.N., 1998. Statistical Learning Theory. John Wiley & Sons, New York.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory, second ed. Springer, New York.

Wolberg, W.H., Mangasarian, O.L., 1990. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In: Proceedings of National Academy of Science of the United States of America, vol. 87, pp. 9193–9196.