INFORMS Journal on Computing
Vol. 17, No. 4, Fall 2005, pp. 462–474
ISSN 1091-9856 | EISSN 1526-5528 | 05 | 1704 | 0462

DOI 10.1287/ijoc.1030.0061

Isotonic Separation

R. Chandrasekaran
School of Engineering and Computer Science, P.O. Box 830688, The University of Texas at Dallas, Richardson, Texas 75083-0688, USA, chandra@utdallas.edu

Young U. Ryu
School of Management, P.O. Box 830688, SM 33, The University of Texas at Dallas, Richardson, Texas 75083-0688, USA, ryoung@utdallas.edu

Varghese S. Jacob
School of Management, P.O. Box 830688, SM 40, The University of Texas at Dallas, Richardson, Texas 75083-0688, USA, vjacob@utdallas.edu

Sungchul Hong
Department of Computer Science and Information Sciences, Towson University, 8000 York Road, Stephens Hall, Room 314, Towson, Maryland 21252, USA, shong@towson.edu

Data classification and prediction problems are prevalent in many domains. The need to predict to which class a particular data point belongs has been seen in areas such as medical diagnosis, credit rating, Web filtering, prediction, and stock rating. This has led to strong interest in developing systems that can accurately classify data and predict outcome. The classification is typically based on the feature values of objects being classified. Often, a form of ordering relation, defined by feature values, on the objects to be classified is known. For instance, the objects belonging to one class have larger (or smaller) feature values than do those in the other class. Exploiting this characteristic of isotonicity, we propose a data-classification method called *isotonic separation* based on linear programming, especially network programming. The paper also addresses an extension of the isotonic-separation method for continuous outcome prediction. Applications of the isotonic separation for discrete outcome prediction and its extension for continuous outcome prediction are shown to illustrate its applicability.

Key words: data classification; isotonic separation; linear programming; network programming; outcome prediction
History: Accepted by Amit Basu; received July 2002; revised June 2003, September 2003; accepted November 2003.

1. Introduction

The study of outcome prediction based on historical data has many applications in a wide range of disciplines including business management, e.g. firm-bankruptcy prediction (Altman 1968), and medicine, e.g., medical diagnosis/prognosis (Schenone et al. 1993, Burke 1994, Mangasarian et al. 1995). No matter which domain is considered, the historical data typically contain information about the outcome (e.g. bankruptcy status in firm-bankruptcy prediction) and the relevant attributes (e.g. debt-to-asset ratio, etc. (Altman 1968) in the case of firm-bankruptcy prediction) that allow one to make a prediction. Using these attributes, or features, it is desirable to classify firms systematically into two or more categories and further generalize the classification in order to predict other firms' status in the future.

Statistical discriminant analysis (Anderson 1972, Cox 1966, Fisher 1936), linear programming approaches to discriminant analysis (Smith 1968, Freed and Glover 1981, Glover 1990, Bennett and Mangasarian 1992), neural networks (Burke 1994), genetic algorithms (Siedlecki and Sklansky 1989, Güvenir and Sirin 1993, Punch et al. 1993), data envelopment analysis (Pendharkar and Kumar 1998), probability analysis (Detrano et al. 1989), decision-tree induction (Murthy et al. 1994; Quinlan 1986, 1993), multisurface separation (Mangasarian 1965, 1968; Mangasarian et al. 1990; Wolberg and Mangasarian 1990), and various other artificial intelligence techniques (Michalski et al. 1986, Gennari et al. 1989) were used for this purpose. Lim et al. (2000) conducted a very extensive study of 33 data-classification methods compared on 32 data sets. Among the 33 data-classification methods are 22 decision-tree-induction methods, nine variations of discriminant analysis, and two neural network algorithms. A nonlinear-mathematical-programming method called support vector machines (Vapnik 1998) is currently recognized as a promising data-classification method (Burges 1998).

In this paper, we propose a new mathematical-programming method, called *isotonic separation*. Given an isotonic consistency condition on historical data points, we build a maximum-flow network model for data separation. The isotonic consistency condition establishes a quasi-order on the set of data points, which becomes the major set of constraints of the network model. An isotonic consistency condition for a classification problem is a kind of domain knowledge of the problem. For instance, in breast-cancer diagnosis, it is known that when all other feature values are identical, a tumor with a bigger epithelial cell is more likely to be malignant than a tumor with a smaller epithelial cell. Such isotonic consistency conditions are known in many data classification and outcome prediction areas, including the cases of medical diagnosis (Mangasarian et al. 1990, 1995), bankruptcy prediction (Altman 1968), and Internet information filtering (Jacob et al. 1999). The isotonic-separation method takes advantage of the known isotonic consistency conditions and accurately classifies historical data points.

We start with the two-category isotonic separation and generalize it into three-category and multicategory separation techniques. Especially, the presentation of the two-category isotonic separation includes issues of problem-size reduction and outcome prediction based on data classification, which can be extended to cases of three-category and multicategory separations. Also, we propose a continuous outcome isotonic model, which is closely related to isotonic regression (Gebhardt 1970, Barlow et al. 1972, Dykstra and Robertson 1982, Block et al. 1994). Finally, we discuss the practical use of isotonic separation in medical diagnosis/prognosis and compare isotonic separation results with those of other methods.

2. Two-Category Separation of Two-Category Data

Suppose we are given a finite set A of data points in a d-dimensional real space $\Re^d$, a function p: $A \to \{0, 1\}$, and a reflexive and transitive binary relation S on $\Re^d$.

Definition 1 (Isotonic Consistency Condition). A quasi-ordering (i.e., reflexive and transitive) relation S is called an *isotonic consistency condition* of a function π on $A \subseteq \Re^d$ (and π is called *isotonic* with respect to S) if "$i, j \in A$" and "$(i, j) \in S$" imply "$\pi(i) \geq \pi(j)$." □

We want to obtain a classification function π^*: $A \to \{0, 1\}$ that is isotonic with respect to S, while minimizing misclassification (i.e., $p(i) \neq \pi^*(i)$ for $i \in A$). If $p(i) = 1$ but $\pi^*(i) = 0$, the misclassification is measured by α; similarly, if $p(i) = 0$ but $\pi^*(i) = 1$, the misclassification is measured by β.

Definition 2 (Two-Category Isotonic Separator). Suppose p is a given function on A to $\{0, 1\}$, and α and β are nonnegative real numbers. Then, a classification function π^* on A to $\{0, 1\}$ is called a *two-category isotonic separator* of A under p with weights α and β if π^* is isotonic with respect to S and minimizes

$$\sum_{i \in A} [p(i) - \pi(i)]w(i)$$

in the class of all functions π on A to $\{0, 1\}$ that are isotonic with respect to S, where

$$w(i) = \begin{cases} \alpha & \text{if } p(i) = 1 \\ -\beta & \text{if } p(i) = 0 \end{cases}$$

for $i \in A$. □

The process of two-category separation works as follows. First, an isotonic separator is constructed from undominated or boundary data points as shown in (2.7). Such undominated points are obtained by solving a linear program (2.1), whose dual is a maximum-flow network model (2.2). Obtaining an isotonic separator on the data set A is called *training* and the set A is called the *training data set*. Next, in order to classify new data points (e.g. testing data) that do not belong to A, we extend the isotonic separator to a function on $\Re^d$ using a nearness criterion (2.10).

2.1. The Isotonic Consistency Condition

The isotonic consistency condition of Definition 1 comes from domain knowledge of the problem, on which the isotonic-separation operation works. For points i and j in a d-dimensional real space whose coordinate vectors are **i** and **j**, in many cases, $(i, j) \in S$ when $\mathbf{i} \geq \mathbf{j}$. Any binary relation as an isotonic consistency condition would be sufficient for the purpose of estimating π^* on A. However, to estimate π^* on $\Re^d$, it must have a form of ordering relation. Thus, in general, an isotonic consistency condition is required to be a reflexive and transitive ordering relation (known as a quasi-order, which is the weakest form of ordering relation).

For instance, consider the following fine-needle aspirate data of tumor taken from the University of Wisconsin Hospitals breast cancer data set (Merz and Murphy 1998) in a nine-dimensional space of clump thickness (d_1), uniformity of cell size (d_2), uniformity of cell shape (d_3), marginal adhesion (d_4), single epithelial cell size (d_5), bare nuclei (d_6), bland chromatin (d_7), normal nucleoli (d_8), and mitoses (d_9):

Points	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
123	4	4	2	1	2	5	2	1	2
125	4	4	4	4	6	5	7	3	1
182	8	4	4	5	4	7	7	8	2
198	4	1	1	3	1	5	2	1	1

The known isotonic consistency condition in breast-cancer diagnosis states that a data point with bigger dimension values is more likely to be malignant than another with smaller dimension values. Thus, we have $\{(182, 198), (182, 123), (125, 198)\} \subset S$. Here, we expect point 182 to be more likely to be malignant than points 198 and 123, and point 125 to be more likely to be malignant than point 198.

We would like to classify these data points into two categories of malignant and benign tumors using the isotonic consistency condition S. Let "$\pi^*(i) = 1$" indicate the data point i's being classified to be malignant and "$\pi^*(i) = 0$" benign. Then, the pair of data points $(182, 198) \in S$, meaning that point 182 is more likely to be malignant than point 198, must satisfy a constraint "$\pi^*(182) \geq \pi^*(198)$." The details of the mathematical model will be addressed in the following section.

Isotonic separation is a data-classification method that exploits given isotonic consistency conditions. For problems with known isotonic consistency conditions, the proposed approach can be a viable and powerful tool for data classification. On the other hand, when no isotonic consistency conditions are known, they must be tested (Robertson et al. 1988) and discovered from the available data set for training. This issue is reported in current studies on heart-attack recurrence/survival prediction (Ryu et al. 1999) and extended bankruptcy prediction (Ryu and Yue 2004).

2.2. The Mathematical Model

Define the variable

$$\pi_i = \begin{cases} 1 & \text{if } i \text{ is classified as 1} \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i \in A.$$

Let $a_i = p(i)$ and $b_i = 1 - p(i)$ for $i \in A$. (That is, $a_i = 1$ if $p(i) = 1$ and $b_i = 1$ if $p(i) = 0$.) Also let $S_A = \{(i, j) \in S \mid i, j \in A\}$. Then, a solution to the following math program will provide an isotonic separator of Definition 2 while satisfying the isotonic consistency condition of Definition 1:

$$\underset{\pi_i:\, i\in A}{\text{minimize}} \left\{ \begin{array}{l|ll} \alpha \sum_{i\in A} a_i(1-\pi_i) & \pi_i - \pi_j \geq 0 & \text{for } (i,j) \in S_A \\ \quad +\beta \sum_{i\in A} b_i \pi_i & \pi_i \in \{0,1\} & \text{for } i \in A \end{array} \right\}.$$

Here, $\sum_{i\in A} a_i(1 - \pi_i)$ and $\sum_{i\in A} b_i\pi_i$ measure the number of misclassified data points. The constraint matrix in "$\pi_i - \pi_j \geq 0$" consists of only 1, 0, and -1 and thus is unimodular. This implies that we can drop the integer requirement of the variables and still get an integer solution (Murty 1976, Shapiro 1979). Therefore, the above math program can be simplified to the following linear program:

$$\underset{\pi_i:\, i\in A}{\text{minimize}} \left\{ \sum_{i\in A} (\beta b_i - \alpha a_i)\pi_i \;\middle|\; \begin{array}{ll} \pi_i - \pi_j \geq 0 & \text{for } (i,j) \in S_A \\ 0 \leq \pi_i \leq 1 & \text{for } i \in A \end{array} \right\}. \tag{2.1}$$

Let $f_{i,j}$ be the dual variable of "$\pi_i - \pi_j \geq 0$" and u_i the dual variable of "$-\pi_i \geq -1$." Further, let $B_i = \{j \in A \mid (i, j) \in S_A\}$ and $C_i = \{j \in A \mid (j, i) \in S_A\}$ for $i \in A$. Then, we have the following dual formulation of (2.1):

$$\underset{\substack{f_{i,j}:(i,j)\in S_A \\ u_i: i\in A}}{\text{maximize}} \left\{ -\sum_{i\in A} u_i \;\middle|\; \begin{array}{ll} \sum_{j\in B_i} f_{i,j} - \sum_{j\in C_i} f_{j,i} - u_i \leq \beta b_i - \alpha a_i & \\ \qquad \text{for } i \in A & \\ f_{i,j} \geq 0 \quad \text{for } (i,j) \in S_A & \\ u_i \geq 0 \quad \text{for } i \in A & \end{array} \right\},$$

or equivalently,

$$\underset{\substack{f_{i,j}:\,(i,j)\in S_A \\ u_i, s_i:\, i\in A}}{\text{minimize}} \left\{ \sum_{i\in A} u_i \;\middle|\; \begin{array}{ll} \sum_{j\in B_i} f_{i,j} - \sum_{j\in C_i} f_{j,i} - u_i + s_i & \\ \qquad = \beta b_i - \alpha a_i \quad \text{for } i \in A & \\ \sum_{i\in A} u_i - \sum_{i\in A} s_i = \alpha \sum_{i\in A} a_i - \beta \sum_{i\in A} b_i & \\ f_{i,j} \geq 0 \quad \text{for } (i,j) \in S_A & \\ u_i, s_i \geq 0 \quad \text{for } i \in A & \end{array} \right\}. \tag{2.2}$$

This is a maximum-flow network model with $|A| + 2$ nodes and $|S_A| + 2|A|$ arcs.

For instance, let us consider a set A of eight data points in a two-dimensional space as shown in Figure 1, where $p(i) = 0$ for each bullet point (i.e., $i =$ 1, 2, 5, or 8) and $p(i) = 1$ for each circle point (i.e., $i = 3$, 4, 6, or 7). We have an isotonic consistency condition $S = \{(i, j) \mid \mathbf{i} \geq \mathbf{j}\}$ where $\mathbf{i}$ and $\mathbf{j}$ are coordinate vectors of i and j. Thus,

$$\begin{aligned} S_A = \{&(1,1), (2,2), (3,2), (3,3), (4,1), (4,2), (4,3), \\ &(4,4), (5,2), (5,5), (6,1), (6,2), (6,3), (6,5), \\ &(6,6), (7,2), (7,7), (8,2), (8,3), (8,5), \\ &(8,7), (8,8)\}, \end{aligned}$$

which yields the consistency constraints of the model (2.1). However, one can easily find that some consistency constraints are redundant. For instance, "$(1, 1) \in S_A$" corresponds to "$\pi_1 \geq \pi_1$," which is tautology; "$\pi_6 \geq \pi_2$" for "$(6, 2) \in S_A$" is deducible from "$\pi_6 \geq \pi_3$" and "$\pi_3 \geq \pi_2$" due to the transitivity property of "$\geq$." After eliminating all such redundant (i.e., reflexive or transitively implied) constraints, we have the following consistency constraints from S_A:

$$\begin{array}{llll} \pi_3 - \pi_2 \geq 0 & \pi_4 - \pi_1 \geq 0 & \pi_4 - \pi_3 \geq 0 & \pi_5 - \pi_2 \geq 0 \\ \pi_6 - \pi_1 \geq 0 & \pi_6 - \pi_3 \geq 0 & \pi_6 - \pi_5 \geq 0 & \pi_7 - \pi_2 \geq 0 \\ \pi_8 - \pi_3 \geq 0 & \pi_8 - \pi_5 \geq 0 & \pi_8 - \pi_7 \geq 0, & \end{array} \tag{2.3}$$

as shown by directed arcs in Figure 1.

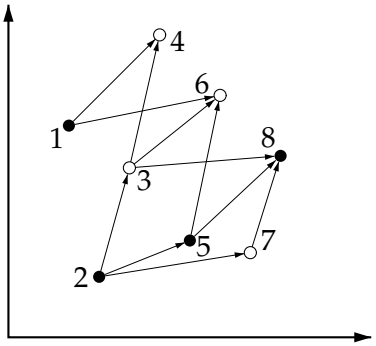

Figure 1 Sample Points in a Two-Dimensional Space

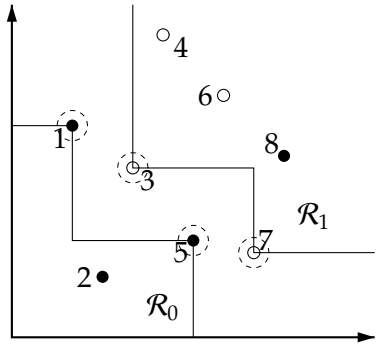

Figure 2 Separation of the Two-Dimensional Space of Figure 1

With $\alpha = \beta > 0$, we have the objective function of (2.1) as follows:

$$\text{minimize } \pi_1 + \pi_2 - \pi_3 - \pi_4 + \pi_5 - \pi_6 - \pi_7 + \pi_8. \quad (2.4)$$

The optimization of this linear-programming model with the boundary constraints of

$$0 \le \pi_i \le 1 \quad \text{for } i = 1, 2, \ldots, 8$$

gives the following optimal solution

$$\begin{aligned} &\pi_1^* = 0 \quad \pi_2^* = 0 \quad \pi_3^* = 1 \quad \pi_4^* = 1 \\ &\pi_5^* = 0 \quad \pi_6^* = 1 \quad \pi_7^* = 1 \quad \pi_8^* = 1. \end{aligned} \quad (2.5)$$

This optimal solution with $\alpha = \beta > 0$ indicates that points 1, 2, and 5 are classified as 0 and points 3, 4, 6, 7, and 8 are classified as 1. As a result, point 8 is misclassified. If we solved the problem with $\beta \ge 2\alpha > 0$, the optimal solution would be

$$\begin{aligned} &\pi_1^* = 0 \quad \pi_2^* = 0 \quad \pi_3^* = 0 \quad \pi_4^* = 1 \\ &\pi_5^* = 0 \quad \pi_6^* = 1 \quad \pi_7^* = 0 \quad \pi_8^* = 0, \end{aligned} \quad (2.6)$$

and points 3 and 7 would be misclassified.

Let $\Theta^* = \{\pi_i^* \mid i \in A\}$ be an optimal solution to (2.1). It gives a function π^* on A (i.e., $\pi^*(i) = \pi_i^*$ for $i \in A$), but we would like to extend it to a function on $\Re^d$. This is necessary for the purpose of actual applications of isotonic separation, that is, for the purpose of classification of new data points. To do it efficiently, first define undominated data points:

$$\begin{aligned} A_0^* = \{i \mid \pi_i^* = 0 \text{ and } \nexists\, \pi_j^* \in \Theta^* \text{ such that} \\ i \ne j,\ \pi_j^* = 0, \text{ and } (j, i) \in S_A\} \\ A_1^* = \{i \mid \pi_i^* = 1 \text{ and } \nexists\, \pi_j^* \in \Theta^* \text{ such that} \\ i \ne j,\ \pi_j^* = 1, \text{ and } (i, j) \in S_A\}. \end{aligned} \quad (2.7)$$

Also assume $A_0^* \ne \varnothing$ and $A_1^* \ne \varnothing$. Then, for any data point $i \in \Re^d$, if there exists $j \in A_0^*$ such that $(j, i) \in S$ then $\pi^*(i) = 0$; if there exists $j \in A_1^*$ such that $(i, j) \in S$ then $\pi^*(i) = 1$. In the example of Figure 1, with the optimal solution of (2.5) (i.e., $\alpha = \beta$), we have $A_0^* = \{1, 5\}$ and $A_1^* = \{3, 7\}$, as shown in Figure 2, in which each undominated point is surrounded by a dashed circle.

This scheme divides the d-dimensional space into three areas:

$$\mathcal{R}_0 = \{i \in \Re^d \mid \exists\, j \in A_0^* \text{ such that } (j, i) \in S\}, \quad (2.8)$$

where alls points are classified as 0,

$$\mathcal{R}_1 = \{i \in \Re^d \mid \exists\, j \in A_1^* \text{ such that } (i, j) \in S\}, \quad (2.9)$$

where all points are classified as 1, and an unclassified area between $\mathcal{R}_0$ and $\mathcal{R}_1$, in which no data points of A exist and the isotonic consistency condition cannot classify the area. To complete the classification, we will use a nearness criterion, measured by the weighted distance to the closest points on the boundaries of $\mathcal{R}_0$ and $\mathcal{R}_1$. The classification penalties α and β will serve as the weights in the distance measure. For a real number a, let $(a)_*$ be a function returning 1 if $a > 0$ and 0 if $a \le 0$. For a d-dimensional vector $\mathbf{i}$ whose k-th element is a_k, let $(\mathbf{i})_+$ be a function returning a d-dimensional vector whose k-th element is a_k if $a_k > 0$ and 0 if $a_k \le 0$. When $(i, j) \in S$ for $\mathbf{i} \ge \mathbf{j}$, we finally have π^* defined as:

$$\pi^*(i) = \left(\alpha \min_{j \in A_0^*} \mathbf{e}^T (\mathbf{i} - \mathbf{j})_+ - \beta \min_{j \in A_1^*} \mathbf{e}^T (\mathbf{j} - \mathbf{i})_+ \right)_*, \quad (2.10)$$

where $\mathbf{i}$ and $\mathbf{j}$ are d-dimensional coordinate vectors of i and j respectively, and $\mathbf{e}^T$ is the transpose of the d-dimensional unit vector. When the above nearness criterion is used, if scale differences among feature values are pronounced, data must be normalized. The final separator for the example of Figure 1, obtained using (2.10), is illustrated in Figure 3 as a dashed line.

2.3. Reduction of the Problem Size

As discussed in the previous example, removing tautological and implied constraints in S_A reduces the size of the problem by reducing the number of constraints in the linear-programming model (2.1) or the number of arcs in the network (2.2). The reduced

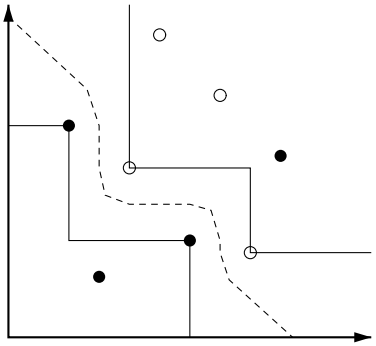

Figure 3 The Final Separator for Figure 1

isotonic consistency condition S'_A can be obtained as follows:

$$S'_A = \{(i,j) \in S_A \mid i \neq j \text{ and } \nexists k \in A \text{ such that } \mathbf{i} \neq \mathbf{k}, \mathbf{j} \neq \mathbf{k}, (i,k) \in S_A, \text{ and } (k,j) \in S_A\},$$

where **i**, **j**, and **k** are coordinate vectors of i, j, and k, respectively. Here, by the first component ($i \neq j$), we remove tautological constraints (e.g. $\pi_i \geq \pi_i$); by the second component (there does not exist k such that $(i,k) \in S_A$ and $(k,j) \in S_A$), we remove transitively implied constraints (e.g., $\pi_i \geq \pi_j$ if $\pi_i \geq \pi_k$ and $\pi_k \geq \pi_j$ are constraints of the model).

Now, we will discuss two additional reduction methods, which reduce not only the number of constraints (i.e., arcs in the network) but also the number of variables (i.e., nodes in the network). Observe bullet points 1, 2, and 5 (whose p function value is 0) in Figure 1. No circle points (whose p function value is 1) are located at their lower left-hand side in the coordinate system. That is, for $i \in \{1,2,5\}$, there does not exist $j \in A$ such that $p(j) = 1$ and $(i,j) \in S_A$ (i.e., $\pi_i - \pi_j \geq 0$). Thus, setting $\pi_i = 0$ for all $i \in \{1,2,5\}$ satisfies (2.3); furthermore, the objective-function value of (2.4) when $\pi_i = 0$ for all $i \in \{1,2,5\}$ is less than that when $\pi_i = 1$ for some $i \in \{1,2,5\}$. Similarly, there are no bullet points (whose p function value is 0) at the upper right-hand side of circle points 4 and 6 (whose p function value is 1). That is, for $i \in \{4,6\}$, there does not exist $j \in A$ such that $p(j) = 0$ and $(j,i) \in S_A$ (i.e., $\pi_j - \pi_i \geq 0$). Thus, setting $\pi_i = 1$ for all $i \in \{4,6\}$ satisfies (2.3); furthermore, the objective-function value of (2.4) when $\pi_i = 1$ for all $i \in \{4,6\}$ is less than that when $\pi_i = 0$ for some $i \in \{4,6\}$. Solutions to this example (2.5) and (2.6) with $\alpha = \beta > 0$ and $\beta \geq 2\alpha > 0$ (and in fact, with any $\alpha > 0$ and $\beta > 0$) confirm this observation. As a result, we may exclude these data points when constructing (2.1):

$$\begin{aligned} \text{minimize} \quad & -\pi_3 - \pi_7 + \pi_8 \\ \text{subject to} \quad & \pi_8 \geq \pi_3 \\ & \pi_8 \geq \pi_7 \\ & 0 \leq \pi_3, \pi_7, \pi_8 \leq 1, \end{aligned}$$

whose dual is a maximum-flow network with five nodes and eight arcs.

We generalize this observation as follows.

Lemma 1. *Let* $A_0 = \{i \in A \mid p(i) = 0\}$; $A_1 = \{i \in A \mid p(i) = 1\}$. *Define*

$$\mathcal{A} = \{i \in A_0 \mid \exists\, j \in A_1 \text{ such that } (i,j) \in S_A\} \cup \{j \in A_1 \mid \exists\, i \in A_0 \text{ such that } (i,j) \in S_A\},$$

or equivalently

$$\mathcal{A} = \{i \in A_0, j \in A_1 \mid (i,j) \in S_A\}. \tag{2.11}$$

Then, for every optimal solution $\Theta^* = \{\pi_i^* \mid i \in A\}$ *to* (2.1), $\pi_i^* = p(i)$ *for all* $i \in A \backslash \mathcal{A}$. (*Note, "\" denotes the set subtraction operator.*)

Proof. If $\pi_i^* \neq p(i)$ for some $i \in A \backslash \mathcal{A}$, then Θ^* would not be an optimal solution to (2.1) because changing π_i^* to $p(i)$ for all such i still satisfies all constraints of (2.1) and improves its objective function. □

As a result of Lemma 1, we can reduce (2.1) to

$$\underset{\pi_i : i \in \mathcal{A}}{\text{minimize}} \left\{ \sum_{i \in \mathcal{A}} (\beta b_i - \alpha a_i)\pi_i \;\middle|\; \begin{array}{ll} \pi_i - \pi_j \geq 0 & \text{for } (i,j) \in S_{\mathcal{A}} \\ 0 \leq \pi_i \leq 1 & \text{for } i \in \mathcal{A} \end{array} \right\}, \tag{2.12}$$

where $S_{\mathcal{A}} = \{(i,j) \in S_A \mid i,j \in \mathcal{A}\}$. That is, if $\{\pi_i^* \mid i \in \mathcal{A}\}$ is an optimal solution to (2.12), then $\{\pi_i^* \mid i \in \mathcal{A}\} \cup \{\pi_i^* = p(i) \mid i \in A \backslash \mathcal{A}\}$ is an optimal solution to (2.1). The dual formulation of (2.12), which can be obtained by replacing A by $\mathcal{A}$ in (2.2), results in a maximum-flow network model with $|\mathcal{A}|+2$ nodes and $|S_{\mathcal{A}}|+2|\mathcal{A}|$ arcs.

When the data set A (or the reduced data set $\mathcal{A}$) is very large and several points have the same set of coordinates in the d-dimensional space, we can further reduce the problem size by having only one variable for each set of such data points. The following lemma shows this.

Lemma 2. *Let* A' *be a maximal subset of* A *that contains data points with different coordinate vectors. Furthermore, for* $i' \in A'$, *let* $a_{i'} = |B_{i',1}|$ *and* $b_{i'} = |B_{i',0}|$, *where* $B_{i',1} = \{i \in A \mid p(i) = 1 \text{ and } \mathbf{i} = \mathbf{i}'\}$ *and* $B_{i',0} = \{i \in A \mid p(i) = 0 \text{ and } \mathbf{i} = \mathbf{i}'\}$. (*Here,* **i** *and* **i**′ *are coordinate vectors of* i *and* i', *respectively.*) *That is,* $a_{i'}$ *and* $b_{i'}$ *denote the numbers of actual data points on the coordinates* **i**′ *whose p function values are* 1 *and* 0, *respectively. Then, the optimization problem* (2.1) *can be reduced to*:

$$\underset{\pi_{i'} : i' \in A'}{\text{minimize}} \left\{ \sum_{i' \in A'} (\beta b_{i'} - \alpha a_{i'})\pi_{i'} \;\middle|\; \begin{array}{ll} \pi_{i'} - \pi_{j'} \geq 0 & \text{for } (i',j') \in S_{A'} \\ 0 \leq \pi_{i'} \leq 1 & \text{for } i' \in A' \end{array} \right\}, \tag{2.13}$$

where $S_{A'} = \{(i',j') \in S \mid i',j' \in A'\}$.

PROOF. Note that $\bigcup_{i' \in A'}(B_{i',1} \cup B_{i',0}) = A$. For $i_1, i_2 \in B_{i',1} \cup B_{i',0}$ because their coordinate vectors are identical, $(i_1, i_2) \in S_A$ and $(i_2, i_1) \in S_A$. That is, (2.1) includes constraints of $\pi_{i_1} - \pi_{i_2} \geq 0$ and $\pi_{i_2} - \pi_{i_1} \geq 0$. Thus, $\pi_{i_1} = \pi_{i_2}$ for $i_1, i_2 \in B_{i',1} \cup B_{i',0}$. As a result, π_i for all $i \in B_{i',1} \cup B_{i',0}$ can be substituted by $\pi_{i'}$. Then, the objective function of (2.1) is identical to that of (2.13):

$$\begin{aligned}\sum_{i \in A}(\beta b_i - \alpha a_i)\pi_i &= \sum_{i \in A'}\left(\sum_{j \in B_{i,0}} \beta \pi_j - \sum_{j \in B_{i,1}} \alpha \pi_j\right) \\ &= \sum_{i \in A'}(\beta |B_{i,0}| \pi'_i - \alpha |B_{i,1}| \pi'_i) \\ &= \sum_{i \in A'}(\beta b'_i - \alpha a'_i)\pi'_i,\end{aligned}$$

and the constraints and boundary conditions of (2.1) are identical to those of (2.13). □

The same reduction operation can be applied to $\mathcal{A}$, which may reduce (2.12) to a smaller one.

3. Three-Category Separation of Two-Category Data

In practical applications of data separation (or example-based learning in general), we often do not want to classify certain data because of insufficient data used for training (or insufficient examples used for learning). For instance, in a problem of credit-application evaluation, if an applicant's income, debt ratio, etc. are not obviously low or high and there were no previous applicants with similar incomes, debt ratios, etc., then the automated evaluation system may refuse to recommend acceptance or denial, but refer the case to the human expert. In fact, this can be easily achieved by using the two-category isotonic separation result of the previous section. Instead of the two-category classification function π^*: $\Re^d \to \{0, 1\}$ of (2.10), the following three-category classification function π^*: $\Re^d \to \{-1, 0, 1\}$ can be defined:

$$\pi^*(i) = \left(\min_{j \in A_0^*} \mathbf{e}^T(\mathbf{i} - \mathbf{j})_+\right)_* - \left(\min_{j \in A_1^*} \mathbf{e}^T(\mathbf{j} - \mathbf{i})_+\right)_*, \quad (3.1)$$

where i is classified as 1 if $\pi^*(i) = 1$, i is classified as 0 if $\pi^*(i) = -1$, and i is unclassified if $\pi^*(i) = 0$. That is, $\pi^*(i) = -1$ for $i \in \mathcal{R}_0$ of (2.8); $\pi^*(i) = 1$ for $i \in \mathcal{R}_1$ of (2.9); and $\pi^*(i) = 0$ for $i \notin \mathcal{R}_0 \cup \mathcal{R}_1$.

In the above three-category separator obtained from two-category data, a data point $i \in \Re^d$ is unclassified ($\pi^*(i) = 0$) because there exist no data points in the two-category data set with which the isotonic consistency condition can classify i. This may not be the only practical reason not to classify certain data points. If many data points of different categories are mixed in an area, it may also be practical not to classify data points in that area. For instance, among current customers in a certain range of income, debt ratio, etc., some persistently defaulted their payments and others did not. Then, the automated evaluation system may refer to the human expert the applicants in this range of income, debt ratio, etc. In this section, we develop a generalized isotonic-separation method for such a data classification/prediction problem.

Suppose we are given a finite set A of data points in a d-dimensional real space $\Re^d$, a function p: $A \to \{0, 1\}$, and a reflexive and transitive binary relation S on $\Re^d$. While minimizing misclassification, we want to obtain two classification functions ξ^*: $A \to \{0, 1\}$ and μ^*: $A \to \{0, 1\}$, both of which are isotonic with respect to S, where $\mu^*(i) \geq \xi^*(i)$ for $i \in A$. (In fact, we may have two binary relations S_1 and S_2 such that ξ^* is isotonic with respect to S_1, while μ^* is isotonic with respect to S_2. We assume here that S_1 and S_2 are the same for the sake of simplicity.) Equivalently, we want to obtain a classification function π^*: $A \to \{-1, 0, 1\}$ that is isotonic with respect to S, where $\pi^*(i) = \xi^*(i) + \mu^*(i) - 1$. (If we have two binary relations S_1 and S_2 such that ξ^* is isotonic with respect to S_1 and μ^* is isotonic with respect to S_2, then π^* is isotonic with respect to $S_1 \cup S_2$.) Here, $i \in A$ is classified as 1 if $\pi^*(i) = 1$ (or $\xi^*(i) = 1$ and $\mu^*(i) = 1$), classified as 0 if $\pi^*(i) = -1$ (or $\xi^*(i) = 0$ and $\mu^*(i) = 0$), and unclassified if $\pi^*(i) = 0$ (or $\xi^*(i) = 0$ and $\mu^*(i) = 1$). For misclassification, we have the following nonnegative penalties: α if $p(i) = 1$ but i is classified as 0; β if $p(i) = 0$ but i is classified as 1; γ if $p(i) = 1$ but i is unclassified; and δ if $p(i) = 0$ but i is unclassified.

DEFINITION 3 (THREE-CATEGORY ISOTONIC SEPARATOR). Suppose p is a given function on A to $\{0, 1\}$, and α, β, γ, and δ are nonnegative real numbers. Suppose functions ξ^* and μ^* on A to $\{0, 1\}$ are isotonic with respect to S and minimize

$$\sum_{i \in A}\big([p(i) - \mu(i)]w_1(i) + [p(i) - \xi(i)]w_2(i) + [\mu(i) - \xi(i)]w_3(i)\big)$$

in the class of all functions ξ and μ on A to $\{0, 1\}$ (where $\mu(i) \geq \xi(i)$ for $i \in A$) that are isotonic with respect to S, where

$$w_1(i) = \begin{cases} \alpha & \text{if } p(i) = 1 \\ 0 & \text{if } p(i) = 0 \end{cases}$$

$$w_2(i) = \begin{cases} 0 & \text{if } p(i) = 1 \\ -\beta & \text{if } p(i) = 0 \end{cases}$$

$$w_3(i) = \begin{cases} \gamma & \text{if } p(i) = 1 \\ \delta & \text{if } p(i) = 0 \end{cases}$$

for $i \in A$. Then, a classification function $\pi^* = \xi^* + \mu^* - 1$ on A to $\{-1, 0, 1\}$ is called a *three-category isotonic separator* of A under p with weights α, β, γ, and δ. □

Define the variables

$$\xi_i = \begin{cases} 1 & \text{if } i \text{ is classified as 1} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i \in A;$$

$$\mu_i = \begin{cases} 0 & \text{if } i \text{ is classified as 0} \\ 1 & \text{otherwise} \end{cases} \quad \text{for } i \in A.$$

Let $a_i = p(i)$ and $b_i = 1 - p(i)$ for $i \in A$. Also let $S_A = \{(i, j) \in S\colon i, j \in A\}$. Then, we have

$$\underset{\xi_i, \mu_i : i \in A}{\text{minimize}} \left\{ \begin{array}{l|l} \alpha \sum_{i \in A} a_i(1-\mu_i) + \beta \sum_{i \in A} b_i \xi_i & \xi_i - \xi_j \geq 0 \quad \text{for } (i,j) \in S_A \\ \quad + \gamma \sum_{i \in A} a_i(\mu_i - \xi_i) & \mu_i - \mu_j \geq 0 \quad \text{for } (i,j) \in S_A \\ & \mu_i - \xi_i \geq 0 \quad \text{for } i \in A \\ \quad + \delta \sum_{i \in A} b_i(\mu_i - \xi_i) & \xi_i, \mu_i \in \{0,1\} \quad \text{for } i \in A \end{array} \right\}.$$

After dropping the integer requirements, we finally have the following formulation for three-category separation of the two-category data set A:

$$\underset{\xi_i, \mu_i : i \in A}{\text{minimize}} \left\{ \begin{array}{l|l} \sum_{i \in A} [(\beta - \delta) b_i - \gamma a_i] \xi_i & \xi_i - \xi_j \geq 0 \quad \text{for } (i,j) \in S_A \\ \quad + [\delta b_i - (\alpha - \gamma) a_i] \mu_i) & \mu_i - \mu_j \geq 0 \quad \text{for } (i,j) \in S_A \\ & \mu_i - \xi_i \geq 0 \quad \text{for } i \in A \\ & 0 \leq \xi_i,\ \mu_i \leq 1 \quad \text{for } i \in A \end{array} \right\}. \tag{3.2}$$

Let $f_{i,j}$ be the dual variable of "$\xi_i - \xi_j \geq 0$," $g_{i,j}$ the dual variable of "$\mu_i - \mu_j \geq 0$," u_i the dual variable of "$\mu_i - \xi_i \geq 0$," v_i the dual variable of "$-\xi_i \geq -1$," and w_i the dual variable of "$-\mu_i \geq -1$." Furthermore, let $B_i = \{j \in A\colon (i, j) \in S_A\}$ and $C_i = \{j \in A\colon (j, i) \in S_A\}$ for $i \in A$. Then, we have the dual formulation:

$$\underset{\substack{f_{i,j}, g_{i,j} : (i,j) \in S_A \\ u_i, v_i, w_i : i \in A}}{\text{minimize}} \left\{ \sum_{i \in A} v_i + \sum_{i \in A} w_i \;\middle|\; \begin{array}{ll} \sum_{j \in B_i} f_{i,j} - \sum_{j \in C_i} f_{j,i} - u_i - v_i & \\ \quad \leq (\beta - \delta) b_i - \gamma a_i & \text{for } i \in A \\ \sum_{j \in B_i} g_{i,j} - \sum_{j \in C_i} g_{j,i} + u_i - w_i & \\ \quad \leq \delta b_i - (\alpha - \gamma) a_i & \text{for } i \in A \\ f_{i,j}, g_{i,j} \geq 0 & \text{for } (i,j) \in S_A \\ u_i, v_i, w_i \geq 0 & \text{for } i \in A \end{array} \right\}.$$

This is a maximum-flow network model with $|A| + 2$ nodes and $2|S_A| + 4|A|$ arcs.

An optimal solution $\Theta^* = \{\xi_i^*, \mu_i^* \mid i \in A\}$ to (3.2) gives functions ξ^* and μ^* on A (i.e., $\xi^*(i) = \xi_i^*$ and $\mu^*(i) = \mu_i^*$), but we would like to extend them to functions on $\Re^d$. Using Θ^*, define sets of undominated data points:

$$\begin{aligned} A_0^* &= \{i \mid \xi_i^* = 0 \text{ and } \nexists\, \xi_j^* \in \Theta^* \text{ such that } i \neq j,\ \xi_j^* = 0, \text{ and } (j,i) \in S_A\} \\ A_1^* &= \{i \mid \xi_i^* = 1 \text{ and } \nexists\, \xi_j^* \in \Theta^* \text{ such that } i \neq j,\ \xi_j^* = 1, \text{ and } (i,j) \in S_A\} \\ B_0^* &= \{i \mid \mu_i^* = 0 \text{ and } \nexists\, \mu_j^* \in \Theta^* \text{ such that } i \neq j,\ \mu_j^* = 0, \text{ and } (j,i) \in S_A\} \\ B_1^* &= \{i \mid \mu_i^* = 1 \text{ and } \nexists\, \mu_j^* \in \Theta^* \text{ such that } i \neq j,\ \mu_j^* = 1, \text{ and } (i,j) \in S_A\}. \end{aligned} \tag{3.3}$$

Assume $A_0^* \neq \varnothing$, $A_1^* \neq \varnothing$, $B_0^* \neq \varnothing$, and $B_1^* \neq \varnothing$. Also, assume $(i, j) \in S$ for $\mathbf{i} \geq \mathbf{j}$. Then, we can define ξ^* and μ^* on $i \in \Re^d$:

$$\xi^*(i) = 1 - \left(\min_{i \in A_1^*} \mathbf{e}^T (\mathbf{j} - \mathbf{i})_+ \right)_*$$

$$\mu^*(i) = \left(\min_{i \in B_0^*} \mathbf{e}^T (\mathbf{i} - \mathbf{j})_+ \right)_*,$$

which can be combined into a single classification function $\pi\colon \Re^d \to \{-1, 0, 1\}$:

$$\begin{aligned} \pi^*(i) &= \xi^*(i) + \mu^*(i) - 1 \\ &= \left(\min_{i \in B_0^*} \mathbf{e}^T (\mathbf{i} - \mathbf{j})_+ \right)_* - \left(\min_{i \in A_1^*} \mathbf{e}^T (\mathbf{j} - \mathbf{i})_+ \right)_*, \end{aligned} \tag{3.4}$$

where i is classified as 1 if $\pi^*(i) = 1$, i is classified as 0 if $\pi^*(i) = -1$, and i is unclassified if $\pi^*(i) = 0$.

Lemma 3. *If $\alpha \leq \gamma$ or $\beta \leq \delta$, then $\xi_i^* = \mu_i^*$ for all $i \in A$ in some optimal solution $\Theta^* = \{\xi_i^*, \mu_i^* \mid i \in A\}$ to* (3.2).

Proof. Consider a solution $\Theta_1 = \{\xi_i, \mu_i \mid i \in A\}$ to (3.2), in which $\mu_j = 1$ and $\xi_j = 0$ for some $j \in A$. Let P be the set of all such points: $P = \{i \in A \mid \mu_i = 1 \text{ and } \xi_i = 0\}$.

- Suppose $\alpha \leq \gamma$. Obtain Θ_2 from Θ_1 by setting $\mu_i = 0$ and $\xi_i = 0$ for all $i \in P$. Since Θ_2 satisfies all constraints of (3.2) it is a solution. For $i \in P$, if $p(i) = 0$, then Θ_1 contributes the misclassification penalty of δ to the objective function of (3.2), but Θ_2 contributes no penalty; If $p(i) = 1$, Θ_1 contributes the misclassification penalty of γ, but Θ_2 contributes α. Thus, Θ_1 is not a better solution than Θ_2.
- Suppose $\beta \leq \delta$. Obtain Θ_2 from Θ_1 by setting $\mu_i = 1$ and $\xi_i = 1$ for all $i \in P$. Since Θ_2 satisfies all constraints of (3.2), it is a solution. For $i \in P$, if $p(i) = 0$, then Θ_1 contributes the misclassification penalty of δ to the objective function of (3.2), but Θ_2 contributes β; If $p(i) = 1$, Θ_1 contributes the misclassification penalty of γ, but Θ_2 contributes no penalty. Thus, Θ_1 is not a better solution than Θ_2.

That is, for any solution in which $\mu_i > \xi_i$ for some $i \in A$, there exists a solution with a same or better objective-function value in which $\mu_i = \xi_i$ for all $i \in A$. Thus, if $\alpha \leq \gamma$ or $\beta \leq \delta$, then there exists an optimal solution in which $\xi_i^* = \mu_i^*$, for all $i \in A$. □

If $\xi_i^* = \mu_i^*$ for all $i \in A$ in an optimal solution $\Theta^* = \{\xi_i^*, \mu_i^* \mid i \in A\}$ to (3.2), then $A_0^* = B_0^*$ and $A_1^* = B_1^*$ in (3.3), and thus (3.4) becomes equivalent with (3.1). Therefore, when creating the separation model of (3.2), we practically have $\alpha > \gamma$ and $\beta > \delta$.

For this three-category separation of two-category data, the same reduction method of §2.3 can be applied. Specifically, the reduction of A to $\mathcal{A}$ as in (2.11) of Lemma 1 can simplify the separation model significantly.

4. Multicategory Separation

Suppose a finite set A of points in a d-dimensional real space $\Re^d$, a function p: $A \to \{0, 1, \ldots, m\}$ where $m \geq 1$, and a reflexive and transitive binary relation S on $\Re^d$ are given. While minimizing misclassification, we want to obtain n classifications ξ_h^*: $A \to \{0, 1\}$ for $1 \leq h \leq n$, where $m \leq n \leq 2m$, all of which are isotonic with respect to S, where $\xi_{h-1}^*(i) \geq \xi_h^*(i)$ for $i \in A$. (We may have up to n binary relations for isotonic consistency. But, we assume that all of them are the same for the sake of simplicity.) Equivalently, we would like to obtain a classification function π^*: $A \to \{0, 1, \ldots, n\}$ that is isotonic with respect to S, where $\pi^*(i) = \sum_{h=1}^{n} \xi^*(i)$. (If we have n binary relations S_h for $1 \leq h \leq n$, then π^* is isotonic with respect to $\bigcup_{h=1}^{n} S_h$.) Here, $i \in A$ is classified as g if $\pi^*(i) = g$ (or $\xi_g(i) = 1$ and $\xi_{g+1}(i) = 0$). The misclassification-minimization and isotonicity requirements for π^* are as follows:

- For $i \in A$ where $p(i) = g$, we have a nonnegative classification penalty of $\alpha_{g,h}$ if $\pi(i) = h$.
  - — For each $g \in \{0, 1, \ldots, m\}$, there exists exactly one $h \in \{0, 1, \ldots, n\}$ such that $\alpha_{g,h} = 0$.
  - — $\alpha_{0,0} = 0$ and $\alpha_{m,n} = 0$.
  - — If $\alpha_{g,h} = 0$ where $0 \leq g < m$ then either $\alpha_{g+1,h+1} = 0$ or $\alpha_{g+1,h+2} = 0$.

The classification function π must be defined to minimize the total classification penalties for all $i \in A$.

- For $i, j \in A$ when $(i, j) \in S$, if i is classified as h, then j must be classified as h or $h-1$ or $\ldots 0$; if j is classified as h, then i must be classified as h or $h+1$ or $\ldots n$.

Definition 4 (Multicategory Isotonic Separator). Suppose p is a given function on A to $\{1, 2, \ldots, m\}$, and $\alpha_{g,h}$ for $0 \leq g \leq m$ and $0 \leq h \leq n$ are nonnegative real numbers. Suppose functions ξ_h^* on A to $\{0, 1\}$ for $1 \leq h \leq n$ are isotonic with respect to S and minimize

$$\sum_{i \in A} \sum_{g=0}^{m} \Bigg([1 - \xi_1(i)] w_{g,0}(i) + \sum_{h=1}^{n-1} [\xi_h(i) - \xi_{h+1}(i)] w_{g,h}(i) + \xi_n(i) w_{g,n}(i) \Bigg)$$

in the class of all functions ξ_h on A to $\{0, 1\}$ for $1 \leq h \leq n$ (where $\xi_{h-1}(i) \geq \xi_h(i)$ for $i \in A$) that are isotonic with respect to S, where

$$w_{g,h}(i) = \begin{cases} \alpha_{g,h} & \text{if } p(i) = g \\ 0 & \text{otherwise} \end{cases}$$

for $0 \leq g \leq m$, $0 \leq h \leq n$, and $i \in A$. Then, a classification function $\pi^* = \sum_{h=0}^{n} \xi_n^*$ on A to $\{0, 1, \ldots, n\}$ is called a *multicategory isotonic separator* of A under p with weight $\alpha_{g,h}$ for $0 \leq g \leq m$ and $0 \leq h \leq n$. □

Define the variables

$$\xi_{i,h} = \begin{cases} 1 & \text{if } i \text{ is classified as } h \text{ or } h+1 \text{ or } \ldots \text{ or } n \\ 0 & \text{else} \end{cases} \quad \text{for } i \in A \text{ and } 1 \leq h \leq n$$

$$\xi_{i,0} = 1 \quad \text{for } i \in A$$

$$\xi_{i,n+1} = 0 \quad \text{for } i \in A.$$

For $i \in A$, let $a_{g,i} = 1$ if $p(i) = g$, otherwise $a_{g,i} = 0$. Also let $S_A = \{(i, j) \in S \mid i, j \in A\}$. Then, we have

$$\underset{\xi_{i,h}: i \in A, 0 \leq h \leq n+1}{\text{minimize}} \left\{ \sum_{g=0}^{m} \sum_{h=0}^{n} \sum_{i \in A} \alpha_{g,h} a_{g,i} \cdot (\xi_{i,h} - \xi_{i,h+1}) \;\middle|\; \begin{array}{ll} \xi_{i,h} - \xi_{j,h} \geq 0 & \text{for } (i,j) \in S_A \text{ and } 1 \leq h \leq n \\ \xi_{i,h-1} - \xi_{i,h} \geq 0 & \text{for } i \in A \text{ and } 1 \leq h \leq n \\ \xi_{i,h} \in \{0,1\} & \text{for } i \in A \text{ and } 1 \leq h \leq n \\ \xi_{i,0} = 1 & \text{for } i \in A \\ \xi_{i,n+1} = 0 & \text{for } i \in A \end{array} \right\}.$$

After dropping the integer requirements, we finally have the following formulation for multicategory separation:

$$\underset{\xi_{i,h}: i \in A, 0 \leq h \leq n+1}{\text{minimize}} \left\{ \sum_{g=0}^{m} \sum_{h=0}^{n} \alpha_{g,h} \sum_{i \in A} a_{g,i} \cdot (\xi_{i,h} - \xi_{i,h+1}) \;\middle|\; \begin{array}{ll} \xi_{i,h} - \xi_{j,h} \geq 0 & \text{for } (i,j) \in S_A \text{ and } 1 \leq h \leq n \\ \xi_{i,h-1} - \xi_{i,h} \geq 0 & \text{for } i \in A \text{ and } 1 \leq h \leq n \\ 0 \leq \xi_{i,h} \leq 1 & \text{for } i \in A \text{ and } 1 \leq h \leq n \\ \xi_{i,0} = 1 & \text{for } i \in A \\ \xi_{i,n+1} = 0 & \text{for } i \in A \end{array} \right\}. \tag{4.1}$$

The dual of the multicategory separation problem is also a maximum-flow network model, as described below:

$$\underset{\substack{f_{i,j,h}: (i,j) \in S_A, 1 \leq h \leq n \\ u_{i,h}: i \in A, 0 \leq h \leq n \\ v_{i,h}: i \in A, 1 \leq h \leq n}}{\text{minimize}} \left\{ \sum_{h=1}^{n} \sum_{i \in A} v_{i,h} \;\middle|\; \begin{array}{ll} \sum_{j \in B_i} f_{i,j,h} - \sum_{j \in C_i} f_{j,i,h} + u_{i,h-1} - u_{i,h} - v_{i,h} \leq \sum_{g=0}^{m} a_{g,i} (\alpha_{q,h} - \alpha_{g,h+1}) & \text{for } i \in A \text{ and } 1 \leq h \leq n \\ f_{i,j,h} \geq 0 & \text{for } (i,j) \in S_A \text{ and } 1 \leq h \leq n \\ u_{i,h} \geq 0 & \text{for } i \in A \text{ and } 0 \leq h \leq n \\ v_{i,h} \geq 0 & \text{for } i \in A \text{ and } 1 \leq h \leq n \end{array} \right\}.$$

Let $\Theta^* = \{\xi_{i,h}^* \mid i \in A, 0 \leq h \leq n+1\}$ be an optimal solution to (4.1) and define $\xi_h^*(i) = \xi_{i,h}^*$ for $0 \leq h \leq n$. Then, we have functions ξ_h^* on A, but we would like to extend them to functions on $\Re^d$. For $1 \leq h \leq n$,

define sets of undominated data points

$$A^*_{h,0} = \{i\colon \xi^*_{i,h} = 0 \text{ and } \nexists\, \xi^*_{j,h} \in \Theta^* \text{ such that } i \neq j,\ \xi^*_{j,h} = 0, \text{ and } (j,i) \in S_A\}$$

$$A^*_{h,1} = \{i\colon \xi^*_{i,h} = 1 \text{ and } \nexists\, \xi^*_{j,h} \in \Theta^* \text{ such that } i \neq j,\ \xi^*_{j,h} = 1, \text{ and } (i,j) \in S_A\}.$$

First, for $i \in \Re^d$,

$$\xi^*_0(i) = 1.$$

For $0 \le g \le m-1$, suppose $\alpha_{g,h} = 0$ and $\alpha_{g+1,h+1} = 0$. (Note that if $g = m-1$, then $h = n-1$.) Then, for $i \in \Re^d$,

$$\xi^*_{h+1}(i) = \left(\alpha_{g+1,h} \min_{j \in A^*_{h+1,0}} \mathbf{e}^T(\mathbf{i}-\mathbf{j})_+ - \alpha_{g,h+1} \min_{j \in A^*_{h+1,1}} \mathbf{e}^T(\mathbf{j}-\mathbf{i})_+\right)_*.$$

For $0 \le g \le m-1$, suppose $\alpha_{g,h} = 0$ and $\alpha_{g+1,h+2} = 0$. (Note that if $g = m-1$, then $h = n-2$.) Then, for $i \in \Re^d$,

$$\xi^*_{h+1}(i) = \left(\min_{j \in A^*_{h+1,0}} \mathbf{e}^T(\mathbf{i}-\mathbf{j})_+\right)_*$$

$$\xi^*_{h+2}(i) = 1 - \left(\min_{j \in A^*_{h+2,1}} \mathbf{e}^T(\mathbf{j}-\mathbf{i})_+\right)_*.$$

From ξ^*_h for $1 \le h \le n$, we can obtain the classification function

$$\pi^*(i) = \sum_{h=1}^{n} \xi^*_h(i).$$

5. Continuous Outcome Cases

So far, we considered the classification/prediction problems of data points with binary or discrete outcomes. In this section, we will consider cases with continuous outcomes. Suppose we have a finite set A of points in $\Re^d$, a function $p\colon A \to [0,1]$, and a reflexive and transitive binary relation S on $\Re^d$. We want to obtain a function $\pi^*\colon A \to [0,1]$ that is isotonic with respect to S and minimizes $\sum_{i \in A} \phi(p(i) - \pi(i))$ where ϕ is a piecewise convex penalty function that may look like Figure 4.

This problem is known as *isotonic regression* (Gebhardt 1970, Barlow et al. 1972, Dykstra and Robertson 1982, Block et al. 1994) if $\phi(p(i) - \pi(i)) = w_i(p(i) - \pi(i))^2$ with weight constants $w_i \ge 0$ and S is a partial order on A. The main difference between this continuous outcome isotonic model and isotonic regression is that isotonic regression requires a stronger condition S (i.e., at least a partial order) due to the computability. Specifically, efficient algorithms are known to exist (Wyatt 1997) only when S is a special type of partial order such as a linear order (i.e., an antisymmetric, transitive, and strongly complete order), a simple tree order (i.e., S constructs a simple tree), and a matrix order (i.e., the rectangular array of data being monotonic in rows and columns). The continuous-outcome isotonic model has a simpler form of penalty function; thus without tight requirement of the isotonic consistency condition, it can be efficiently solved. As shown in the following, the continuous-outcome isotonic model is a minimum-cost network problem, of which efficient algorithms are known (Tardos 1985, 1986; Ahuja et al. 1993).

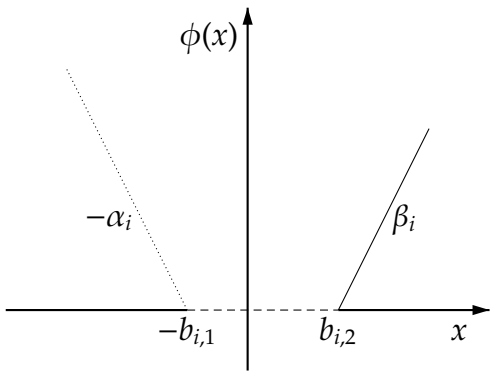

Figure 4 A Convex Penalty Function

Let $\pi_i = \pi(i)$ and $p_i = p(i)$ for $i \in A$. Then, with the piecewise convex penalty function of Figure 4, we have the following continuous-outcome isotonic model:

$$\underset{\pi_i, s_{i,1}, s_{i,2}, t_{i,1}, t_{i,2}: i \in A}{\text{minimize}} \left\{ \sum_{i \in A} \alpha_i s_{i,2} + \sum_{i \in A} \beta_i t_{i,2} \;\middle|\; \begin{array}{ll} \pi_i - s_{i,1} - s_{i,2} + t_{i,1} + t_{i,2} = p_i & \text{for } i \in A \\ \pi_i - \pi_j \ge 0 & \text{for } (i,j) \in S_A \\ 0 \le \pi_i \le 1 & \text{for } i \in A \\ b_{i,1} \ge s_{i,1} \ge 0 & \text{for } i \in A \\ s_{i,2} \ge 0 & \text{for } i \in A \\ b_{i,2} \ge t_{i,1} \ge 0 & \text{for } i \in A \\ t_{i,2} \ge 0 & \text{for } i \in A \end{array} \right\}.$$

Here $b_{i,1}$ and $b_{i,2}$ represent the flat portion of the curve drawn as a dashed line in Figure 4; $-\alpha_i$ and β_i represent the slopes of the dotted line and the solid line, respectively. This problem is a kind of monotone network (Minty 1960) and can be solved by network techniques.

Let $B_i = \{j \in A \mid (i,j) \in S_A\}$ and $C_i = \{j \in A \mid (j,i) \in S_A\}$. The dual formulation is given as:

$$\underset{\substack{f_{i,j}: (i,j) \in S_A \\ g_i, h_i, k_i, u_i: i \in A}}{\text{maximize}} \left\{ \sum_{i \in A} p_i g_i - \sum_{i \in A} b_{i,1} h_i - \sum_{i \in A} b_{i,2} k_i - \sum_{i \in A} u_i \;\middle|\; \begin{array}{ll} \sum_{j \in B_i} f_{i,j} - \sum_{j \in C_i} f_{j,i} + g_i - u_i \le 0 & \text{for } i \in A \\ -g_i - h_i \le 0 & \text{for } i \in A \\ g_i - k_i \le 0 & \text{for } i \in A \\ -\alpha_i \le g_i \le \beta_i & \text{for } i \in A \\ f_{i,j} \ge 0 & \text{for } (i,j) \in S_A \\ u_i, h_i, k_i \ge 0 & \text{for } i \in A \end{array} \right\}.$$

Because $b_{i,1}$ and $b_{i,2}$ are nonnegative, at optimality the following relations hold, defining the values of h_i and k_i in terms of the values of g_i:

$$\begin{aligned} h_i &= \max\{0, -g_i\} \\ k_i &= \max\{0, g_i\} \end{aligned} \quad \text{for } i \in A.$$

Using these relations and by letting $g_i = g_{i,1} - g_{i,2}$ where $g_{i,1} \geq 0$ and $g_{i,2} \geq 0$, we can simplify the above problem to:

$$\underset{\substack{f_{i,j}:(i,j)\in S_A \\ g_{i,1},g_{i,2},u_i:i\in A}}{\text{maximize}} \left\{ \begin{array}{l|l} \displaystyle\sum_{i\in A}(p_i - b_{i,2})g_{i,1} & \displaystyle\sum_{j\in B_i} f_{i,j} - \sum_{j\in C_i} f_{j,i} + g_{i,1} \\ \quad \displaystyle -\sum_{i\in A}(p_i + b_{i,1})g_{i,2} & \quad -g_{i,2} - u_i \leq 0 \\ \quad \displaystyle -\sum_{i\in A} u_i & \quad \text{for } i \in A \\ & 0 \leq g_{i,1} \leq \beta_i \ \text{ for } i \in A \\ & 0 \leq g_{i,2} \leq \alpha_i \ \text{ for } i \in A \\ & f_{i,j} \geq 0 \ \text{ for } (i,j) \in S_A \\ & u_i \geq 0 \ \text{ for } i \in A \end{array} \right\}.$$

This is a minimum-cost network problem.

A solution to the above problem gives a value for each of the points in A, but no form for the function π on $\Re^d$. However, it provides a method for obtaining tight bounds for the π value of any point in $\Re^d$. If it is necessary to have a π value of each point in $\Re^d$, we may use some approximate interpolation on the tight bounds using the given isotonic consistency condition S. Even though the co-domain of p and π is $[0, 1]$ in the above model, it can be replaced by $\Re$ without sacrificing simplicity of the model. Then, the continuous-outcome isotonic model can be used for prognosis problems such as predicting whether the breast cancer will recur within two years after surgical removal of the malignant tumor or, more generally, when the cancer will recur.

6. Applications

The proposed isotonic-separation method is currently tested for breast-cancer diagnosis (Chandrasekaran et al. 1998), breast-cancer prognosis (i.e., cancer recurrence time prediction) (Ryu et al. 1999), heart-attack prognosis (i.e., recurrence and survival-time prediction) (Ryu et al. 1999), Internet information filtering (Jacob et al. 1999), firm-bankruptcy prediction (Ryu and Yue 2004), and other problems such as and consumers' brand-selection prediction.

In the Internet-information filtering study (Jacob et al. 1999), isotonic separation was shown to outperform ID3/C4.5 decision-tree induction (Quinlan 1986, 1993) and OC1 axis-parallel decision-tree induction (Murthy et al. 1994). In firm-bankruptcy prediction experiments (Ryu and Yue 2004), isotonic separation was shown to outperform linear and logistic discriminant analysis (Anderson 1972, Cox 1966, Fisher 1936), robust linear programming discrimination (Bennett and Mangasarian 1992), back-propagation neural networks (Bishop 1995), learning vector quantization (Kohonen 1992, 1995), ID3/C4.5 decision-tree induction, and OC1 oblique decision-tree induction. In heart attack recurrence and survival-time prediction (Ryu et al. 1999), the continuous outcome isotonic prediction model was shown to outperform incremental hyper-rectangle generation (Salzberg 1988) and recurrence surface approximation (Mangasarian et al. 1999, Street et al. 1995).

This section summarizes the experiments in breast-cancer diagnosis and prognosis.

6.1. Breast-Cancer Diagnosis

The breast-cancer diagnosis data set (Merz and Murphy 1998) used for isotonic separation contains 683 data points. Each data point in the data set on fine-needle aspirates taken from patients' breasts consists of nine input features: 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 the higher value corresponding to a more abnormal state of the tumor. Of 683 data points, 239 are diagnosed to be malignant and 444 benign.

We performed ten-fold cross-validation experiments (using the same data partition used in the Lim et al. 2000 study) with $\alpha/\beta = 1$, $\alpha/\beta = 2$, and $\alpha/\beta = 3$. The prediction accuracies calculated as the number of misclassified testing data points over the number of all testing data points were 97.81%, 96.64%, and 95.91%, respectively. However, when the accuracy measure considered differences in α and β, that is, when the following weighted accuracy measure was used:

$$\frac{\alpha * (\text{no. of type 1 errors}) + \beta * (\text{no. of type 2 errors})}{\alpha * (\text{no. of malignant data}) + \beta * (\text{no. of benign data})},$$

where a type 1 error occurs when a malignant data point is predicted as benign, and a type 2 error occurs when a benign data point is predicted as malignant, prediction accuracies were 97.81%, 97.29%, and 97.59%, respectively. We also performed ten-fold, eight-fold, six-fold, and five-fold cross-validation experiments with $\alpha = 1$ and $\beta = 1$ in order to see the effect of training-data-set size on prediction accuracy. The prediction accuracies were 97.81%, 97.66%, 97.51%, and 97.35%, respectively. As more training data were used, the overall accuracy increased.

The isotonic separation experiments were performed using AMPL/CPLEX 7.1.0 on a single Pentium 4 (2.0 GHz) processor system running Linux. The average elapsed time of the ten-fold cross-validation experiments (which included the CPU time, the tem-

porary file access time, etc.) was 0.0015 second for training alone, and 0.0018 second for training and testing together.

The comparative study of Lim et al. (2000) of 33 data-classification methods reported the prediction accuracies on the same breast-cancer-diagnosis data set using ten-fold cross-validation. Among the tested 33 methods, the learning vector quantization algorithm performed the best with a 97.22% accuracy; the logistic-discriminant-analysis method performed the best among discriminant-analysis methods with a 96.63% accuracy. Isotonic separation with 97.81% prediction accuracy performed better than all 33 methods.

The study by Lim et al. did not include the multisurface-separation method (Mangasarian 1965, 1968; Mangasarian et al. 1990; Wolberg and Mangasarian 1990), robust linear-programming method (Bennett and Mangasarian 1992), and the support vector machines (Burges 1998, Vapnik 1998). Therefore, we performed ten-fold cross-validation experiments on the same data set using these methods. Their results of 93.77%, 97.32%, and 96.73% prediction accuracies, respectively, were worse than the isotonic separation result.

6.2. Breast-Cancer Prognosis

The breast-cancer prognosis data set (Merz and Murphy 1998) contains 198 data points, among which 4 data points have a missing value. After removing these data points, we used 194 data points. Each data point consists of 32 input features, one class attribute (whether cancer recurred or not), and one recurrence/survival-time attribute. The input features consist of data on fine-needle aspirates taken at the time of diagnosis and data on the tumor observed at the time of surgery. Among 194 data points, 46 belong to the class A_r of cancer recurrence data with the recurrence time varying between 1 month and 125 months (with an average of 53.58 months) and 148 belong to the class A_s of nonrecurrence data with the survival-time varying between 1 month and 79 months. Note that the survival-time of the second class denotes the latest relative time after the surgery when a patient was observed to be cancer-free; because no further records were collected on these patients, it is not known whether cancer recurred later or not.

Mangasarian et al. (1995) and Street et al. (1995) conducted a breast-cancer recurrence prediction study on a subset of these data using a linear-programming method called the *recurrence surface approximation* (RSA) technique. Following the exact same experimental setup, we used the RSA technique on the whole data set and observed overall errors of 18.1 months and 14.7 months on two different parameter settings.

We performed isotonic prognosis experiments using a variation of the continuous outcome isotonic model (Ryu et al. 1999). The experiments were done with leave-one-out testing, in which the prediction model was created using all data points except one and tested on the left-out data point. Experiments were repeated for each of the 194 data points being left out for prediction-model creation and used for testing. This experimental setup was adopted from the RSA technique experiments of Mangasarian et al. (1995) and Street et al. The isotonic prognosis experiment resulted in overall average errors of 15.7 months (versus 18.1 months of RSA) and 13.4 months (versus 14.7 months of RSA) on two different parameter settings.

7. Concluding Remarks

The study of historical data classification, and outcome prediction based on it, has received both theoretical and practical attention. Statistical, mathematical-programming, and artificial-intelligence methods have been developed for this purpose. However, we have proposed another framework called isotonic separation, which models a data-classification problem as a maximum-flow or minimum-cost problem in a network. Isotonic separation is characterized by this network formulation, for which efficient algorithms are known, and the explicit utilization of domain knowledge that is expressed as an isotonic consistency condition.

We have obtained several data sets to implement isotonic separation and the work has begun to test it. The application of the two-category isotonic separation technique in breast-cancer diagnosis (Chandrasekaran et al. 1998) and the application of the continuous outcome isotonic model in breast-cancer prognosis (Ryu et al. 1999) were summarized in §6. A preliminary study on firm-bankruptcy prediction (Ryu and Yue 2004) showed that isotonic separation outperformed various other methods. A variation of isotonic separation was developed to address a marketing-management problem of product-brand-selection prediction based on customers' demographic and financial data and retailers' promotion tactics. We also proposed an isotonic-separation approach to Internet-information filtering with the PICS rating scheme (Jacob et al. 1999).

More theoretical work remains to be done in order to improve the robustness and applicability of isotonic separation. The following list illustrates some issues currently under investigation.

7.1. Reduction of the Problem Size

We have already addressed the issue of several data points having the same coordinate vector. We have also dealt with the issue of redundant constraints.

Third, more significant methods for problem-size reduction were discussed in §2.3. However, in cases of a very large data set involved in a data-classification problem, we may need further reduction of the problem size. One possibility is to divide the region in $\Re^d$ into a number of cells (i.e., small regions) and think of an approximation in which points that are within a cell are considered to be one point having the same coordinates (say the center of the cell), and this new *heavy* point has weights corresponding to the numbers of original points belonging to different classes. Once we have the separation for this approximation, we may refine the division further, near the separation area to get a better approximation and so on.

7.2. Feature Reduction/Selection

Another important issue is to reduce the number of features used in the model. It is generally believed that the more parsimonious the model, the better its predictive ability. There is some work in this direction by Bradley and Mangasarian (1998) and Bradley et al. (1998), who have done a significant amount of work in planar separation and the use of neural networks in this area. We could use their ideas in our work as well for feature reduction/selection. Other possibilities include the use of genetic algorithms (Punch et al. 1993) and statistical methods (Liu 1997). For the breast-cancer-prognosis study (Ryu et al. 1999), we adopted the backward-sequential-elimination method (Marill and Green 1963, Kittler 1986), which gave reasonable outcome prediction results, as reported in §6. However, other methods are currently being tested for better outcome prediction in this and other problems.

7.3. On the Isotonic Consistency Condition

So far, the notion of consistency has been related to geometric ideas of vectors being ordered. Any quasi-order can be used for this purpose. For example, if the data sets are matrices, we could use submatrices to define a quasi-order. This idea has tremendous theoretical potential for discovering the smallest violators of various properties. We have begun some work in this area. Another possibility is to consider a rotation of the given points first, and then separate them by isotonic separators in the rotated space. For this, we need to find an orthonormal matrix R and then consider the set of points $\mathbf{j} = R\mathbf{i}$ and use them for defining the sets S. The number of possible sets S under such rotations is polynomial in the number of data points, but not in the number of features or dimensions. Using this idea, we can have a multilevel approach as Mangasarian (1968) did with linear separators. Clearly, the isotonic multilevel approach is stronger than linear multilevel approach.

Acknowledgments

The authors wish to thank Professor O. L. Mangasarian for encouragement and help in obtaining previous work and data sets. Thanks are also due to Professor R. Krishnan for very useful discussions on applications of this idea to be reported elsewhere and to Professor R. Madhavan for discussions on isotonic regression and its applications.

References

Ahuja, R. K., T. L. Magnanti, J. B. Orlin. 1993. *Network Flows.* Prentice-Hall, Englewood Cliffs, NJ.

Altman, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finance* **23** 589–609.

Anderson, J. A. 1972. Separate sample logistic discrimination. *Biometrika* **59** 19–35.

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

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

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford.

Block, H., S. Qian, A. Sampson. 1994. Structure algorithms for partially ordered isotonic regression. *J. Comput. Graphical Statist.* **3** 285–300.

Bradley, P. S., O. L. Mangasarian. 1998. Feature selection via concave minimization and support vector machines. *Proc. Fifteenth Internat. Conf. Machine Learning*, Morgan Kaufmann, San Mateo, CA, 82–90.

Bradley, P. S., O. L. Mangasarian, W. N. Street. 1998. Feature selection via mathematical programming. *INFORMS J. Comput.* **10** 209–217.

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

Burke, H. B. 1994. Artificial neural networks for cancer research: Outcome prediction. *Seminars Surgical Ontology* **10** 73–79.

Chandrasekaran, R., Y. U. Ryu, V. Jacob. 1998. Breast cancer diagnosis using an isotonic separation approach. Working paper, Department of Information Systems and Operations Management, School of Management, The University of Texas at Dallas, Richardson, TX.

Cox, D. R. 1966. Some procedures connected with the logistic qualitative response curve. F. N. David, ed. *Research Papers in Statistics: Festschrift for J. Neyman.* John Wiley and Sons, New York, 55–71.

Detrano, R., A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, V. Froelicher. 1989. International application of a new probability algorithm for the diagnosis of coronary artery disease. *Amer. J. Cardiology* **64** 304–310.

Dykstra, R. L., T. Robertson. 1982. An algorithm for isotonic regression of two or more independent variables. *Ann. Statist.* **10** 708–711.

Fisher, R. A. 1936. The use of multiple measurements in taxonomy problem. *Ann. Eugenics* **7** 179–188.

Freed, E., F. Glover. 1981. A linear programming approach to the discriminant problem. *Decision Sci.* **12** 68–74.

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

Gennari, J. H., P. Langley, D. Fisher. 1989. Models of incremental concept formation. *Artificial Intelligence* **40** 11–61.

Glover, F. 1990. Improved linear programming model for discriminant analysis. *Decision Sci.* **21** 771–785.

Güvenir, H. A., I. Sirin. 1993. A genetic algorithm for classification by feature partitioning. *Proc. Fifth Internat. Conf. Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA, 543–548.

Jacob, V., R. Krishnan, Y. U. Ryu, R. Chandrasekaran, S. Hong. 1999. Filtering objectionable Internet content. *Proc. Twentieth Internat. Conf. Inform. Systems*, Association for Information Systems, Atlanta, GA, 274–278.

Kittler, J. 1986. Feature selection and extraction. T. Y. Young, K. S. Fu, eds. *Handbook of Pattern Recognition and Image Processing.* Academic Press, New York, 59–83.

Kohonen, T. 1992. New developments of learning vector quantization and the self-organizing map. *Proc. 1992 Sympos. Neural Networks: Alliances and Perspectives in Senri* (*SYNAPSE'92*), Senri International Information Institute, Osaka, Japan.

Kohonen, T. 1995. *Self-Organizing Maps.* Springer-Verlag, Heidelberg, Germany.

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

Liu, H. 1997. Feature selection via discretization. *IEEE Trans. Knowledge Data Engrg.* **9** 742–645.

Mangasarian, O. L. 1965. Linear and nonlinear separation of patterns by linear programming. *Oper. Res.* **13** 455–461.

Mangasarian, O. L. 1968. Multisurface method of pattern separation. *IEEE Trans. Inform. Theory* **IT-14** 801–807.

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

Mangasarian, O. L., W. N. Street, W. H. Wolberg. 1995. Breast cancer diagnosis and prognosis via linear programming. *Oper. Res.* **43** 570–577.

Marill, T., D. M. Green. 1963. On the effectiveness of receptors in recognition systems. *IEEE Trans. Inform. Theory* **9** 11–17.

Merz, C. J., P. M. Murphy. 1998. UCI repository of machine learning databases. Department of Information and Computer Sciences, University of California, Irvine, CA.

Michalski, R. S., I. Mozetic, J. Hong, N. Lavrac. 1986. Multipurpose incremental learning system AQ15 and its testing application to three medical domains. *Proc. Fifth National Conf. Artificial Intelligence* (*AAAI-86*), AAAI Press, Menlo Park, CA, 1041–1045.

Minty, G. 1960. Monotone networks. *Proc. Royal Soc. London* **257A** 192–212.

Murthy, S. K., S. Kasif, S. Salzberg. 1994. A system for induction of oblique decision trees. *J. Artificial Intelligence Res.* **2** 1–32.

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

Pendharkar, P. C., S. Kumar. 1998. A DEA application for marginal cost assignment in certain case based expert systems. *Proc. Third INFORMS Conf. Inform. Systems Tech*, Montreal, Canada, 347–358.

Punch, W. F., E. D. Goodman, M. Pei, L. Chia-Shun, P. Hovland, R. Enbody. 1993. Further research on feature selection and classification using genetic algorithms. *Proc. fifth Internat. Conf. Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA, 557–564.

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

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

Robertson, T., F. T. Wright, R. L. Dykstra. 1988. *Order Restricted Statistical Inference.* John Wiley and Sons, New York.

Ryu, Y. U., R. Chandrasekaran, V. Jacob. 1999. Disease prognosis with an isotonic prediction technique. *Proc. Ninth Workshop Inform. Tech. Systems*, Charlotte, NC, 26–31.

Ryu, Y. U., W. T. Yue. 2004. Firm bankruptcy prediction: Experimental comparison of isotonic separation and other classification approaches. *IEEE Trans.* Forthcoming.

Salzberg, S. 1988. Exemplar-based learning: Theory and implementation. Technical Report TR-10-88, Aiken Computation Laboratory, Center for Research in Computing Technology, Harvard University, Cambridge, MA.

Schenone, A., L. Andreucci, V. Sanguinetti, P. Morasso. 1993. Neural networks for prognosis in breast cancer. *Physica Medica: Eur. J. Medical Phys.* **IX**(Supp. 1) 175–178.

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

Siedlecki, W., J. Sklansky. 1989. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Lett.* **10** 335–347.

Smith, F. W. 1968. Pattern classifier design by linear programming. *IEEE Trans. Comput.* **C-17** 367–372.

Street, W. N., O. L. Mangasarian, W. H. Wolberg. 1995. An inductive learning approach to prognostic prediction. *Proc. Twelfth Internat. Conf. Machine Learn*, Morgan Kaufmann, San Mateo, CA, 522–530.

Tardos, É. 1985. A strong polynomial minimum cost circulation algorithm. *Combinatorica* **5** 247–255.

Tardos, É. 1986. A strong polynomial algorithm to solve combinatorial linear programs. *Oper. Res.* **34** 250–256.

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

Wolberg, W. H., O. L. Mangasarian. 1990. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc. National Acad. Sci. USA* **87** 9193–9196.

Wyatt, G. J. 1997. Inference from partial orders: Central bank independence and inflation. Technical report, Department of Economics, Heriot-Watt University, Edinburgh, Scotland.