Firm Bankruptcy Prediction: Experimental Comparison of Isotonic Separation and Other Classification Approaches

Young U. Ryu and Wei T. Yue

***Abstract*—A newly introduced method called isotonic separation is evaluated in the prediction of firm bankruptcy. Feature reduction methods are first applied to reduce the ratios used in the prediction. Then, various classification methods, including discriminant analysis, neural networks, decision tree induction, learning vector quantization, rough sets, and isotonic separation, are used with the reduced ratios. Experiments show that the isotonic separation method is a viable technique, performing generally better than other methods for short-term bankruptcy prediction.**

***Index Terms*—Bankruptcy prediction, isotonic separation, pattern classification, prediction method.**

I. Introduction

THE ability to have prior warnings regarding a distressed firm is desirable because it could serve to reduce the negative impacts resulting from the fallen firm. Those benefited by prior warnings include the creditors, shareholders, employees, and other participants of the related firm. One major approach of determining the health of a firm is to monitor the financial information from the firm's financial statement. Firm bankruptcy predictions in the past have relied on these financial indicators to discern distressed firms from healthy firms. Over the last few decades, there have been continuous improvements in creating better prediction techniques [1], [48], [61], [63]. Essentially, the problem of bankruptcy prediction is a type of the classification problem. The principal goal is to classify distressed firms based on a set of financial variables. Prior classification techniques in the problem of bankruptcy prediction include discriminant analyses [1], [11], [17], [19], [48], neural networks [61], [63], decision tree induction methods [23], [46], and rough sets [26], [59], to name a few. In addition, given the vast amount of financial information associated with a firm, there have been many discussions in the past regarding the appropriate selection of the effective financial ratios to identify a distressed firm [4], [9].

The objectives of this paper are to establish that a newly introduced method called the isotonic separation method [14] is a viable technique for firm bankruptcy prediction, and to understand which financial variables are effective in bankruptcy prediction. Isotonic separation, which was previously applied in an information filtering problem [30] and a disease recurrence time-line prediction problem [53], is a linear programming technique that can be applied to classify data in an order restricted domain [5], [6]. Such order restrictions may be known in advance or must be obtained from data. This method when used in bankruptcy prediction minimizes the number of misclassified firms in prediction. Other methods minimize impurity measures of entropy or the total distance between misclassified data and the estimator in an Euclidean space, or maximize the conditional likelihood. We would like to show that this direct and simple classification objective used by isotonic separation leads to a good bankruptcy prediction system.

In order to validate the predictive capability of the isotonic separation method, the outcome is compared against the results of other classification methods, including discriminant analyses, linear programming discrimination methods [10], [60], neural networks, learning vector quantization [36], [37], decision tree induction methods [47], [51], [52], and rough set analyses [49], [50]. A total of 23 financial ratios from various literatures [1], [17], [19], [23], [64] were selected in the current study. Based on these ratios, variable reductions techniques such as sequential elimination, stepwise discriminant analysis [20], [27], [31], and mutual information based feature selection [7], [12], [39], were applied to identify the best set of ratios for prediction.

Three datasets were collected to conduct one-year, two-year, and three-year bankruptcy prediction experiments. The study outcome indicated that it was difficult to identify a universally best set of ratios for prediction; in fact the selection of ratios was heavily dependent on specific prediction models. For instance, in our study, the debt to asset and the quick asset to total asset ratios were among the best predictors in discriminant analyses; the liability to asset, the equity to debt, and the sales to asset ratios were among the best predictors in neural network methods; the equity to debt and quick asset to sales ratios were among the best predictors in isotonic separation.

By comparing the isotonic separation method with other classification methods, we observed that the isotonic separation method offered better accuracy in identifying bankrupt firms than all other methods for the one-year and two-year prediction cases. For the three-year prediction case, the isotonic separation method performed as well as other methods. Though a generalized claim of the superiority of the isotonic separation method in firm distress prediction would be too much to say based on one set of experiments, this study gives encouraging indications of isotonic separation being a promising method which deserves further investigations.

II. Firm Failure Studies

Since Beaver's [8] pioneering work on firm failure prediction based on financial ratios and Altman's [1] subsequent seminal

Manuscript received January 23, 2003; revised April 8, 2004. This paper was recommended by Associate Editor R. G. Mathieu.

The authors are with the University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: ryoung@utdallas.edu; wei.yue@utdallas.edu).

Digital Object Identifier 10.1109/TSMCA.2005.843393

study, firm bankruptcy prediction has received tremendous attention in the fields of accounting, finance, and, more recently, quantitative/computational sciences. Early bankruptcy studies focused on the validity of using financial ratios with statistical methods, and identified the best sets of financial ratios to predict firm bankruptcy [1], [9], [11], [17], [19]. The discussions later extended to the discovery of more superior classification methods. The earliest studies involved using the linear discriminant analysis [1] developed by Fisher [22], immediately followed by the use of the logistic discriminant analysis method [48]. Recently, nonstatistical classification methods such as decision tree/rule induction methods [23], [46], neural networks [61], [63], genetic algorithms [55], and rough set methods [26], [59] were applied to categorize bankrupt firms. Thus far, we have seen these techniques often provided better bankruptcy predictions than linear and logistic discriminant analyses. Decision tree induction methods [23], [46] were shown to be promising; neural networks [61], [63] were found to perform better than discriminant analyses, nearest neighborhood methods, and decision tree induction methods. Even a hybrid method [40], which involves discriminant analyses, decision tree induction, and neural networks, was reported to result in good prediction outcomes.

We have seen in the past that financial ratios used in bankruptcy predictions vary from study to study. These ratios which individually represent different aspects of the firm were discovered by accounting and finance specialists to be effective in bankruptcy prediction. Beaver [9] used liquid asset variables as the main measuring ratios because they were known to be good short-term predictors. These ratios were divided into common denominator group of total assets, current debts, and net sales. Ohlson [48] included ratios based on previously identified ratios, which were similar to Beaver's ratios. Blum [11] investigated the financial variables using the cash-flow framework which treated the firm as a "reservoir of financial resources," and identified ratios that affected the reservoir state. Altman [1] started with 22 ratios from the financial categories of liquidity, profitability, leverage, solvency, and activity, and then narrowed down the list to five ratios that performed best in predicting bankruptcy from discriminant analyses. In this paper, we used 23 ratios commonly used in these and other bankruptcy prediction studies [1], [9], [11], [17], [19], [23], [48], [64] for the comparative study of isotonic separation and nine other methods.

III. Isotonic Separation

Isotonic separation [14] is a linear programming technique that separates d-dimensional data in an order restricted domain. In firm bankruptcy prediction, for instance, the order restriction can be formed by stating that a firm with a lower cash flow to total liabilities ratio, a higher current liabilities to current assets ratio, a lower net income to total assets ratio, a higher total liabilities to total assets, and a lower working capital to total assets ratio is more likely to go bankrupt [48]. This underlying idea of the order restricted domain in isotonic separation is borrowed from isotonic regression [5], [6]. For isotonic separation, the weakest form of order relation is sufficient; such an order relation is called the quasi order, which is a reflexive and transitive relation. It was previously applied and validated to be an effective method in an information filtering problem [30] and a disease recurrence time-line prediction problem [53].

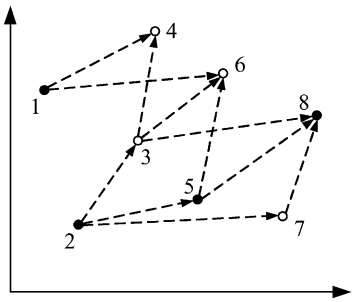


Fig. 1. Sample data.

Suppose a set A_0 of data belonging to a group 0, a set A_1 of data belonging to a group 1, and the order restriction (i.e., the isotonic consistency condition) S are given. Without using any generality, we assume that for a pair of data points i and j whose attribute vectors are $(a_{i1}, a_{i2}, \ldots, a_{id})$ and $(a_{j1}, a_{j2}, \ldots, a_{jd})$, respectively, $(i,j) \in S$ if and only if $a_{ik} \geq a_{jk}$ for $k = 1, 2, \ldots,$ and d. For each data point $i \in A_0 \cup A_1$, define a separation variable π_i such that if $\pi_i = 1$ then i is labeled as 1, and if $\pi_i = 0$ then i is labeled as 0. Then, the separation of data in $A_0 \cup A_1$ is achieved by solving the following linear program

$$\begin{aligned} &\text{minimize } \alpha \sum_{i \in A_1} (1 - \pi_i) + \beta \sum_{i \in A_0} \pi_i \\ &\text{subject to } \pi_i - \pi_j \geq 0 \quad \text{for } (i,j) \in S \\ &\qquad 0 \leq \pi_i \leq 1 \quad \text{for } i \in A_0 \cup A_1. \end{aligned} \tag{1}$$

Here, π_i is a binary variable, but it can be relaxed to a real variable in the range of 0 to 1, because the constraint matrix in "$\pi_i - \pi_j \geq 0$" is the transpose of a network type constraint matrix. In the objective function (1), $\sum_{i \in A_1}(1 - \pi_i)$ indicates the number of data points that are mislabeled as 0; $\sum_{i \in A_0} \pi_i$ indicates the number of data points that are mislabeled as 1; and $\alpha > 0$ and $\beta > 0$ are costs or penalties of misclassification. Often we set $\alpha = \beta$, or $\alpha = 1/|A_1|$ and $\beta = 1/|A_0|$. (Note $|A|$ denotes the cardinality, i.e., the number of elements, of set A.)

To illustrate, consider data points in a two-dimensional attribute space of Fig. 1, in which bullet data points belong to the group 0 (i.e., $A_0 = \{1, 2, 5, 8\}$) and circle data points belong to group 1 (i.e., $A_1 = \{3, 4, 6, 7\}$). Then, S includes pairs of data points such as (2, 2), (3, 2), (4, 3), (4, 2), etc. Here, (2, 2) is a reflexive pair giving a constraint "$\pi_2 - \pi_2 \geq 0$," which is a tautology; (4, 2) giving a constraint "$\pi_2 - \pi_4 \geq 0$" is transitively implied by constraints of (3, 2) and (4, 3). These reflexive and transitively implied pairs can be safely dropped. As the result, S includes pairs of points as shown by directed arcs in Fig. 1. Then, we have the following isotonic separation linear program (1) with $\alpha = \beta > 0$:

$$\begin{aligned} &\text{minimize } \pi_1 + \pi_2 - \pi_3 - \pi_4 + \pi_5 - \pi_6 - \pi_7 + \pi_8 \\ &\text{subject to } \pi_3 - \pi_2 \geq 0 \quad \pi_4 - \pi_1 \geq 0 \quad \pi_4 - \pi_3 \geq 0 \\ &\qquad \pi_5 - \pi_2 \geq 0 \quad \pi_6 - \pi_1 \geq 0 \quad \pi_6 - \pi_3 \geq 0 \\ &\qquad \pi_6 - \pi_5 \geq 0 \quad \pi_7 - \pi_2 \geq 0 \quad \pi_8 - \pi_3 \geq 0 \\ &\qquad \pi_8 - \pi_5 \geq 0 \quad \pi_8 - \pi_7 \geq 0 \\ &\qquad 0 \leq \pi_i \leq 1 \quad \text{for } i = 1, 2, \ldots, 8. \end{aligned}$$

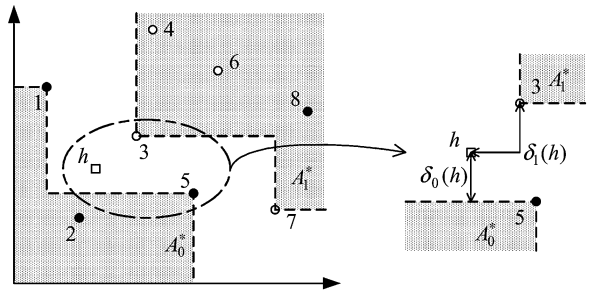


Fig. 2. Isotonic separation ($\alpha = \beta$) of sample data.

This linear program is optimized with the following solution:

$$\pi_1^* = 0 \quad \pi_2^* = 0 \quad \pi_3^* = 1 \quad \pi_4^* = 1$$
$$\pi_5^* = 0 \quad \pi_6^* = 1 \quad \pi_7^* = 1 \quad \pi_8^* = 1.$$

That is, points 1, 2, and 5 are separated into the group 0, and points 3, 4, 6, 7, and 8 are separated into the group 1, where point 8 is incorrectly separated.

Once the separation of data in $A_0 \cup A_1$ is done, the d-dimensional attribute space (i.e., the order restricted domain) is separated as follows. Let $A_0^* = \{i \in A_0 \cup A_1 \,|\, \pi_i^* = 0\}$ and $A_1^* = \{i \in A_0 \cup A_1 \,|\, \pi_i^* = 1\}$, where $\{\pi_i^* \,|\, i \in A_0 \cup A_1\}$ is an optimal solution to the linear program (1). For a point h whose attribute vector is $(a_{h1}, a_{h2}, \ldots, a_{hd})$, define its distances to A_0^* and A_1^*

$$\delta_0(h) = \alpha \min \left\{ \sum_{k=1}^{d} \max(a_{hk} - a_{ik}, 0) \,\middle|\, i \in A_0^* \right\}$$
$$\delta_1(h) = \beta \min \left\{ \sum_{k=1}^{d} \max(a_{ik} - a_{hk}, 0) \,\middle|\, i \in A_1^* \right\}$$

where $(a_{i1}, a_{i2}, \ldots, a_{id})$ is the attribute vector of i. If $\delta_0(h) < \delta_1(h)$ then h is allocated into the group 0; otherwise it is allocated into the group 1.

From the separation result of data points in Fig. 1, it can be determined $A_1^* = \{1, 2, 5\}$ and $A_1^* = \{3, 4, 6, 7, 8\}$, as shown in Fig. 2. If a new data point h lies in the area of A_0^*, then $\delta_0(h) = 0$ and $\delta_1(h) > 0$; thus, it is allocated into the group 0. If it lies in the area of A_1^*, then $\delta_0(h) > 0$ and $\delta_1(h) = 0$; thus, it is allocated into the group 1. If it lies in the area between A_0^* and A_1^* (e.g., h in Fig. 2), then $\delta_0(h) = \alpha(a_{h2} - a_{52})$ and $\delta_1(h) = \beta[(a_{31} - a_{h1}) + (a_{32} - a_{h2})]$, where (a_{h1}, a_{h2}), (a_{31}, a_{32}), and (a_{51}, a_{52}) are attribute vectors of points h, 3, and 5, respectively.

Let f_{ij} be the dual variable of "$\pi_i - \pi_j \geq 0$" and u_i be the dual variable of "$-\pi_i \geq -1$" of the linear program (1). Then, we have the following dual formulation of (1)

$$\text{minimize} \sum_{i \in A_0 \cup A_1} u_1$$
$$\text{subject to} \sum_{j:(i,j) \in S} f_{ij} - \sum_{j:(j,i) \in S} f_{ji} - u_i + s_i = \begin{cases} -\alpha & \text{for } i \in A_1 \\ \beta & \text{for } i \in A_0 \end{cases}$$
$$\sum_{i \in A_0 \cup A_1} u_i - \sum_{i \in A_0 \cup A_1} s_i = \alpha |A_1| + \beta |A_0|$$
$$f_{ij} \geq 0 \text{ for } (i,j) \in S$$
$$u_i, s_i \geq 0 \text{ for } i \in A_0 \cup A_1$$

which is a maximum flow network with $|A_0 \cup A_1| + 2$ vertices and $|\{(i,j) \in S\}| + 2|A_0 \cup A_1|$ edges. This implies that isotonic separation is a computationally efficient method, because the maximum flow problem of a network with v vertices and e edges can be solved with $O(v^3)$ [33], $O(ve \log(v^2/e))$ [25], or more efficient algorithms [24].

IV. Other Classification Methods

Discriminant analyses and neural networks have been two of the most frequently used methods in bankruptcy prediction. Linear programming discrimination methods [10], [60] are simple and computational efficient methods that have been verified to be good classification methods in other areas such as disease diagnoses. Learning vector quantization [36], [37], which is often implemented as a two-layer perceptron for competitive learning, captures the essence of the nearest neighborhood method. ID3/C4.5 [51], [52] and OC1 [47], two of well-known decision tree induction methods based on recursive partitioning, are included as they have been applied in various classification problems. The rough set theory [49], which has been recently verified to be an effective method in firm failure prediction [26], [59], is also tested and compared in the studies. In this section, we briefly describe discriminant analyses, linear programming methods, neural networks, and decision tree induction methods, while somewhat detailed descriptions of learning vector quantization and the rough set theory are provided as they are relatively new in the firm failure research.

A. Discriminant Analysis Methods

Two-class linear discriminant analysis [22] is a multivariate technique to find a linear discriminant function that converts multivariate data in two groups into univariate data such that means of univariate data in different groups are separated as much as possible relative to the population variance [32]. The linear discriminant function, then, leads to a classification rule (or a hyperplane) that can be used to allocate new data into proper groups. The linear discriminant analysis method assumes that data in each group are normally distributed and the covariance matrices of two groups are same. The two class logistic discriminant analysis [3], [15], [16], based on the cumulative logistic probability function, is a method without the assumptions of the normal distribution of data and the same covariance matrix of two groups. Instead, it assumes the log-linearity of the ratio of probability densities of two groups. The Probit model is yet another variation which is based on the cumulative normal probability function rather than the cumulative logistic probability function. We tested all these discriminant analysis methods in the firm failure studies.

B. Linear Programming Discrimination Methods

The idea of linear programming discrimination [10], [60] is very similar to that of linear discriminant analysis. Multivariate (i.e., multiattribute) data are transformed into univariate data on which a separating point is determined; as the result, a single linear hyperplane is drawn to separate data. The main difference is that linear programming discrimination minimizes misclassification errors that are measured as the distance between the hyperplane and the misplaced data. The method proposed by Smith

[60] sets the same misclassification error rate for both groups while the robust method [10] adjusts misclassification error rates by the sizes of groups. We tested both methods in the studies, but reported the result of the robust method mainly because it was consistently better than the result of Smith's method.

C. Neural Networks

An artificial neural network [21], [29] is a machine learning technique based on the intuition of the inner working of the human brain. The network is composed of units, or neurons, connected by directed arcs, or communication channels. An input unit receives an external signal and passes it to connected units. A noninput unit (such as a hidden unit or an output unit) receives inputs and invokes an output based on the inputs and the weights associated with arcs through which it receives inputs. A neural network can be presented as a feedforward network or a feedback network. In a feedforward network, data goes from input to output. In a feedback network, data can travel in both directions. In a popular backpropagation network, data first travel forward. The output is then compared to the target output to determine the error rate. This information is transferred, or propagated, back to adjust the weights of the arcs using the gradient decent algorithm and the chain rule. Out studies included the result from backpropagation networks.

D. Decision Tree Induction Methods

ID3/C4.5 [51], [52] is a top-down decision tree induction method based on the idea of entropy reduction. It induces an axis-parallel decision tree, in which each node contains one attribute variable and branches from the node have equality or inequality conditions on the attribute variable. OC1 [47], another top-down decision tree induction method, generates an oblique decision tree, in which each node contains a hyperplane separating the attribute space and each of its subsequent nodes further separates the half space with another hyperplane. We tested both methods and reported the results of firm failure prediction.

E. Learning Vector Quantization

Learning vector quantization [36], [37] is a competitive learning method for data clustering. Suppose a d-dimensional space containing n data points (each of which is known to belong to one of two or more classes or groups) is to be separated into m clusters. Let $(w_{k1}, w_{k2}, \ldots, w_{kd}|c_k)$ be the codebook vector or prototype representing (the center of) cluster k, where $(w_{k1}, w_{k2}, \ldots, w_{kd})$ is the coordinate vector in the d-dimensional space and c_k is the class that it belongs to. When all such codebook vectors are obtained from the given n data points, a new data point is allocated to the class of a nearest codebook vector. (That is, the partitioning of the d-dimensional space can be done by Voronoi tessellation.)

The learning algorithm starts with initial m codebook vectors which can be chosen randomly from the given data points or by some simple observation of the given data [37]. Let $\mathbf{m}_k(t)$ denote the codebook vector for cluster k at time t. For a data point whose coordinate vector is $(x_1, x_2, \ldots, x_d)$, belonging to class c_x, find a nearest codebook vector

$$k^* = \arg \min_{1 \le k \le m} (w_{k1} - x_1)^2 + (w_{k2} - x_2)^2 + \cdots + (w_{kd} - x_d)^2. \quad (2)$$

If the class of k^* is same as that of x (i.e., $c_{k^*} = c_x$), then move the codebook vector k^* toward x (by a fraction $\alpha_{k^*}(t)$ of the distance between them); otherwise, move it away from x

$$\mathbf{m}_{k^*}(t+1) = (\ldots, w_{k^*i} + s\alpha_{k^*}(t)(s_i - w_k^* i), \ldots, |\, c_{k^*}) \quad (3)$$

where $s = 1$ if $c_{k^*} = c_x$, $s = 0$ if $c_{k^*} \neq c_x$, and for $k \geq 1$

$$\alpha_{k^*}(t) = \frac{\alpha_{k^*}(t-1)}{1 + s\alpha_{k^*}(t-1)}$$

with $0 < \alpha_{k*}(0) < 1$. This process is iterated until a stopping criterion (e.g., a number of iterations or a threshold change in codebook vectors) is met.

The learning vector quantization method can be implemented as a neural network with d input nodes and m output nodes where the weight w_{ki} of the arc from input node i to output node k constitutes normalized codebook vector k. Since codebook vectors in the neural network are normalized, the nearest vector selection of (2) can be achieved by

$$k^* = \arg \max_{1 \le k \le m} w_{k1}x_1 + w_{k2}x_2 + \cdots + w_{kd}x_d.$$

When the output node k^* is selected, weights w_{k^*i} (for $i = 1, 2, \ldots,$ and d) of arcs connected to k^* are replaced by $w_{k^*i} + s\alpha_{k^*}(t)(x_i - w_{k^*i})$ as shown in (3).

F. Rough Set Analyses

The rough set theory [49] is an extension of the classical set theory used for representation of incomplete knowledge. One of problems addressed by the rough set theory is decision analysis, especially the multicriteria sorting problem [50]. Bankruptcy prediction, as a form of multicriteria sorting problem, was studied based on the rough set theory [26], [59]. Rough set decision analysis works as follows.

Consider a set A of objects with which a set Q of attributes are associated. For $I \in A$ and $q \in Q$, let v_{iq} denote i's value of the attribute q. When $P \subseteq Q$, a binary relation on A

$$\text{IND}(P) = \{(i, j) \in A \times A : v_{iq} = v_{jq} \text{ for all } q \in P\}$$

is called an indiscernability relation and $A\,|\,\text{IND}(P)$ denotes the partition of A induced by $\text{IND}(P)$. For $X \in A\,|\,\text{IND}(P), \text{Des}_P(X)$, defined as

$$\text{Des}_P(X) = \{(q, v_{iq}) : \text{for all } q \in P\}$$

where $i \in X$ denotes the description of X. Suppose $B \subseteq A$. The P-lower approximation of B (denoted by $\underline{P}B$) is defined as

$$\underline{P}B = \bigcup \{X \in A\,|\,\text{IND}(P) : X \subseteq B\}.$$

Let a partition $\Pi_A = \{B_1, B_2, \ldots, B_n\}$ of A be a classification of A. Then, the quality of approximation of classification Π_A by $P \subseteq Q$ is measured by

$$\gamma_P(\Pi_A) = \frac{\sum_{k=1}^{n} |\underline{P}B_k|}{|A|}.$$

A set R is called a Π_A-reduct of P if R is a minimal subset of $P \subseteq Q$ and $\gamma_R(\Pi_A) = \gamma_P(\Pi_A)$. Often there exist many reducts. Attributes belonging to all reducts are called cores. (Kumar [38] presented a relational algebra method to find reducts and cores.) When there exist many reducts, an attribute sorting method can be used to select one [57]. The method starts with the set of cores. It adds to the set of cores each

attribute; sorts the data by the set of attributes; groups the data points by the sorting result; and estimates the accuracy. The set of cores and an attribute whose sorting accuracy is the highest are chosen. This process is repeated until the set of chosen attributes and the cores become a reduct, which is selected as the best reduct.

Let $Q = C \cup D$ and $C \cap D = \emptyset$, where C is the condition attribute set and D is the decision attribute set. (In bankruptcy prediction, C is the set of financial indicators under consideration and D contains the bankruptcy status variable.) Suppose C' be an $A \,|\, \text{IND}(D)$-reduct of C. For every $X \in A \,|\, \text{IND}(C')$ and $Y \in A \,|\, \text{IND}(D)$, if $X \cap Y \neq \emptyset$, then we have a decision rule

$$\text{Des}_{C'}(X) \Rightarrow \text{Des}_D(Y).$$

The generated decision rules are merged and pruned to minimal (i.e., simpler and cleaner) decision rules [58]. A new data point whose attribute values are closest to the condition of a decision rule is classified to have the decision attribute values of the closest decision rule. (The closeness measure is presented in [56].)

A variation of the above method for bankruptcy prediction [26] was proposed based on the notion of dominance relation. Suppose the domain V_q of every attribute $q \in Q$ is ordered. Assume the decision attribute set is singleton, i.e., $D = \{d\}$. (Note that the domain attribute set can be a nonsingleton set if the Cartesian product of domains of decision attributes is ordered.) Let $Cl_t = \{i \in A : v_{id} = t\}, Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s$, and $Cl_t^{\leq} = \bigcup_{s \leq t} Cl_s$. For $P \subseteq C$, let $D_P^+(i) = \{j \in A : v_{iq} \geq v_{jq} \text{ for all } q \in P\}$ and $D_P^-(i) = \{j \in A : v_{iq} \leq v_{jq} \text{ for all } q \in P\}$. Then, the qualify of approximation of classification $\Pi_A = \{Cl_t \subseteq A : t \in V_d\}$ by $P \subseteq C$ is measured by

$$\gamma_P(\Pi_A) = 1 - \frac{\left|\bigcup_{t \in V_d} Bn_P(Cl_t^{\geq}) \cup Bn_P(Cl_t^{\leq})\right|}{|A|}$$

where

$$Bn_P(Cl_t^{\geq}) = \bigcup_{i \in Cl_t^{\geq}} D_P^+(i) - \{i \in A : D_P^+(i) \subseteq Cl_t^{\geq}\}$$
$$Bn_P(Cl_t^{\leq}) = \bigcup_{i \in Cl_t^{\leq}} D_P^-(i) - \{i \in A : D_P^-(i) \subseteq Cl_t^{\leq}\}.$$

Reducts and decision rules are similarly generated. However, when dominance relations are used, conditions of decision rules contain inequality clauses rather than equality clauses. Thus, decision rules can be applicable in the classification of new data points without the use of a closeness measure [26].

V. Feature Selection

Feature selection for a prediction system is a process to find relevant features that would give the best result. Among various approaches (e.g., [35], [39], and [43]), some are integrated with induction processes of classification methods, some are standalone but developed for specific classification methods, and others are standalone and independent of classification methods. In this paper, we use three methods selectively in our experiments. The backward sequential elimination method is used together with the isotonic separation and the linear programming discrimination methods, mainly because of its simplicity and empirical validation of its usefulness (e.g., [44]). The stepwise discriminant analysis method [20], [27], [31] is specifically developed for discriminant analyses. The mutual information based feature selection method [7], [12], [39] has often been used for neural network learning and was claimed to be superior to the correlation based feature selection method [41]. Considering the similarity in the use of the impurity measure of entropy, the mutual information based feature selection method may also improve the classification accuracy of decision tree induction methods, especially for the ID3/C4.5 method.

A. Backward Sequential Elimination

Two instances of backward sequential elimination are implemented for isotonic separation and linear programming discrimination. The sequential elimination method for isotonic separation works as follows. Let A be the given set of d-dimensional data for training, which is partitioned to A_t for feature selection and A_v for validation of selected features. The process starts with the set F_0 of all features, with which isotonic separation is performed on A_t and testing is done on A_v. Next, for $i = 0, 1, \ldots,$ and $d-1$, using each subset F_{ik} of $|F_i| - 1$ features (where there are $|F_i|$ subsets), perform isotonic separation on A_t and test the accuracy on A_v. Find F_{ik^*} that results in the best testing accuracy and let $F_{i+1} = F_{ik^*}$. As the final feature set, select, among $F_0, F_1, \ldots,$ and F_{d-1}, one that results in the best testing accuracy on A_v.

The feature elimination method for linear programming discrimination has the same process, but the feature subset evaluation criterion differs. It starts with the set F_0 of all features. For $i = 0, 1, \ldots,$ and $d-1$, using the feature set F_i, perform linear programming discrimination on A_t and test the accuracy on A_v. Check the coefficients w_k (for $k \in F_i$) of the hyperplane "$w_1x_1 + w_dx_d + \cdots + w_dx_d = \gamma$" resulted from the linear programming separation to find w_{k^*} that is closest to 0, and then let $F_{i+1} = F_i - \{k^*\}$. As the final feature set, select, among $F_0, F_1, \ldots,$ and F_{d-1}, one that results in the best testing accuracy on A_v.

B. Stepwise Discriminant Analysis

Suppose a set A of d-dimensional data points partitioned into two or more classes is given. Let C be the set of all class values and A_i be the set of data points of class $i \in C$. For a set S of features and a data point $j \in A$, let $\mathbf{a}_j^S$ be j's vector of features in S; let $\boldsymbol{\mu}^S$ be the mean vector of features in S for all data points in A; let $\boldsymbol{\mu}_i^S$ be the mean vector of features in S for all data points in A_i (i.e., those of class i). Define within-class and between-class cross product matrices ($\mathbf{W}_S$ and $\mathbf{B}_S$, respectively) of S

$$\mathbf{W}_S = \sum_{i \in C} \sum_{j \in A_i} \left(\mathbf{a}_j^S - \boldsymbol{\mu}_i^S\right)' \left(\mathbf{a}_j^S - \boldsymbol{\mu}_i^S\right)$$
$$\mathbf{B}_S = \sum_{i \in C} |A_i| \left(\boldsymbol{\mu}^S - \boldsymbol{\mu}_i^S\right)' \left(\boldsymbol{\mu}^S - \boldsymbol{\mu}_i^S\right)$$

from which Wilks' lambda is defined as

$$\Lambda(S) = \frac{\det(\mathbf{W}_S)}{\det(\mathbf{W}_S + \mathbf{B}_S)}$$

where $\det(\cdot)$ denotes the determinant of a matrix, and the F-statistic is defined as

$$\mathrm{FS}(S) = \frac{(|A| - |C| - |S| + 1)(1 - \Lambda(S))}{(|C| - 1)\Lambda(S)}.$$

The stepwise discriminant analysis process works as follows. Let F be the set of all d feature variables and S be an empty set. Find

$$f^* = \arg\min_{f \in F} \Lambda(S \cup \{f\}) \quad \text{or} \quad f^* = \arg\max_{f \in F} \mathrm{FS}(S \cup \{f\}).$$

If the null hypothesis of the F-test is rejected, remove f^* from F, add it to S, and iterate the same process; otherwise stop. In the bankruptcy prediction experiments, we set the F-test confidence level to 0.99 (i.e., the significance level to 0.01).

C. Mutual Information Based Feature Selection

Suppose a set A of d-dimensional data points partitioned into two or more classes is given. The mutual information based feature selection (MIFS) method, which has been used for feature selection in supervised neural network learning, is a method based on the notion of entropy reduction, i.e., mutual information. The ID3/C4.5 decision tree induction method, which is also based on this notion, uses probabilities of classes and features approximated by histograms from the given dataset. The MIFS method [39] applied in this paper uses a nonparametric kernel density estimation approach [12] to the calculation of mutual information. Let X_i be a class or feature variable and for data point $j \in A$, let a_{ji} be its value for X_i. Then, the probability for a data point's having $X_1 = x_1, \ldots,$ and $X_m = x_m$ is estimated as

$$P(X_1 = x_1, \ldots, X_m = x_m) = \frac{1}{|A| h^m} \sum_{j \in A} K\left(\frac{x_1 - a_{j1}}{h}, \ldots, \frac{x_m - a_{jm}}{h}\right)$$

where the constant h is the windows radius or bandwidth that determines the degree of averaging in the estimate and the kernel function $K(\cdot)$ is the quadratic kernel, known as the Epanechnikov kernel

$$K(x_1, \ldots, x_m) = \prod_{k=1}^{m} \max\left\{\frac{3}{4}\left(1 - x_k^2\right), 0\right\}.$$

Note, a kernel is a continuous, bounded, and symmetric real function that integrates to one. The use of the quadratic kernel instead of others such as the triangular kernel and the normal (i.e., Gaussian) kernel is mainly due to the computational efficiency. In the bankruptcy prediction experiments, we set $h = 0.2$.

The algorithm [39] (which is a modified version of the algorithm [7] based on a recursive partitioning method) works as follows. Let C denote the class variable. Let F be the set of all d feature variables and S be an empty set. Find

$$f^* = (\arg\max_{f \in F} I(C; S \cup \{f\}).$$

Note, $I(C; \{f_1, \ldots, f_p\}) = I(C; f_1; \ldots; f_p)$. If $I(C; S \cup \{f^*\})/I(C; S) < \alpha$, then the process stops. Otherwise, remove f^* from F, add it to S, and iterate the same process. The set S in the end of the process is the selected feature set. In the bankruptcy prediction experiments, we set $\alpha = 0.99$.

TABLE I
YEARLY NUMBERS OF BANKRUPT FIRMS

Bank. Year	Number of Bankrupt Firms		
	1-Year Data Set	2-Year Data Set	3-Year Data Set
1996	17	14	10
1997	18	19	13
1998	11	12	12
1999	18	21	23
2000	24	29	30
2001		14	16
Total:	88	109	104

TABLE II
COMPUSTAT INDUSTRIAL CLASSIFICATION OF FIRMS IN THE DATASETS

DNUM*	Description	1-Year Data Set		2-Year Data Set		3-Year Data Set	
		Bank. Firms	Non-B. Firms	Bank. Firms	Non-B. Firms	Bank. Firms	Non-B. Firms
1xx	Agriculture	0	0	1	1	1	1
1xxx	Mining, Oil, Gas, Const., etc.	6	6	5	6	5	7
2xxx	Consumer Products	10	7	15	9	14	7
3xxx	Plastic, Steel, Electronics, etc.	25	21	30	29	26	27
4xxx	Transp., Comm., Util., etc.	1	5	5	7	6	7
5xxx	Whole & Retail	24	15	26	19	25	17
6xxx	Financial Institutes	1	5	2	5	2	5
7xxx	Travel, Recreation, S/W, etc.	10	22	11	25	10	26
8xxx	Health & Education	9	6	10	7	10	6
9xxx	Others	2	1	4	1	5	1
Total:		88	88	109	109	104	104

* COMPUSTAT® Industrial Classification Codes

VI. EXPERIMENTS

A. Data

In our experiments, we considered one-year, two-year, and three-year bankruptcy predictions. The one-year, two-year, and three-year prediction experiments were to assess classification methods' accuracy of firm bankruptcy within one year, within two years, and within three years, respectively. Previous studies [9], [11] found financial data of up to five years prior to firm failure were useful for prediction. Twenty-three financial ratios used in previous studies [1], [9], [11], [17], [19], [23], [48], [64] were included in our experiments. (The list of financial ratios used can be found in Table V.) It would be ideal to have data of similar size firms in similar industries within a narrow time line. Collecting all 23 ratios of failed firms, however, was the major difficulty in the study.

After considering various options, we collected data on firms of various sizes in various industries that failed in years between 1996 and 2001. They were obtained from Standard & Poor's COMPUSTAT North American database. Firms with null data entries for the selected variables were eliminated. Eighty-eight, 109, and 104 failed firms were found in the sample period respectively for the one-year, two-year, and three-year bankruptcy experiments. Their bankruptcy years and industry specifications are shown in Tables I and II. The

TABLE III
COMPOSITION OF DATASETS

Data Year	Bank. Year	1-Year Data Set		2-Year Data Set		3-Year Data Set	
		Bank. Firms	Non-B. Firms	Bank. Firms	Non-B. Firms	Bank. Firms	Non-B. Firms
1995	1996	17		14		10	
	1997			11		5	
	1998					7	
	n/a		12		17		19
	Total:	17	12	25	17	22	19
1996	1997	18		8		8	
	1998			4		1	
	1999					8	
	n/a		19		24		26
	Total:	18	19	12	24	17	26
1997	1998	11		8		4	
	1999			10		7	
	2000					11	
	n/a		17		21		23
	Total:	11	17	18	21	22	23
1998	1999	18		11		8	
	2000			17		10	
	2001					9	
	n/a		12		12		16
	Total:	18	12	28	12	27	16
1999	2000	24		12		9	
	2001			14		7	
	n/a		28		35		20
	Total:	24	28	26	35	16	20
Total:		88	88	109	109	104	104

n/a: Bankruptcy not observed between 1996 and 2001

TABLE IV
NUMBER OF OVERLAPPING BANKRUPTCY DATA AMONG DATASETS

Data Sets	No. of Overlapping Data
1-Year & 2-Year	53
1-Year & 3-Year	39
2-Year and 3-Year	69
1-Year, 2-Year, and 3-Year	39

one-year dataset contained bankrupt firms' financial ratios of one year prior to bankruptcy. Approximately half of the bankruptcy data in the two-year dataset were firms' financial ratios of one year prior to bankruptcy and the remaining half were those of two years prior to bankruptcy. Similarly, in the three-year dataset, approximately one third of bankruptcy data were firms' financial ratios of one year prior to the bankruptcy, one third were those of two years prior to bankruptcy, and the remaining data were those of three years prior to bankruptcy. In each dataset, bankrupt firms were pooled together with an equal number of randomly selected healthy firms in the same period. (In the data source, we found approximately 2% of firms filed bankruptcy.) The detailed composition of the data are shown in Table III. Due to the limited number of bankrupt firms, there were firms overlapped in the three datasets. The numbers of overlapping data among the three datasets are listed in Table IV. Firms sampled from the COMPUSTAT North American database had an average asset of US $505 million, an average liabilities of US $142 million, and an average revenue of US $495 million.

Each dataset was randomly partitioned into similarly sized blocks for tenfold cross validations, in which each block had an equal number of bankrupt and healthy firms. That is, each block of the one-year dataset partition contained eight or nine bankrupt firms and an equal number of healthy firms; similarly each block of the two-year dataset partition and the three-year dataset partition contained ten or 11 bankrupt firms and an equal number of healthy firms. Feature selection and data separation (i.e., training of a classification system) were performed on nine blocks of each dataset and the testing error was measured on the remaining block. By this, ten sessions of training and testing were performed on each dataset, and the averages of testing errors were reported.

TABLE V
FINANCIAL RATIOS AND ORDER RESTRICTIONS

Financial Ratios	Domain Know.	Experiments with All Features			Experiments with Selected Features		
		1[a]	2[b]	3[c]	1[a]	2[b]	3[c]
Cash Flow/Total Assets	+						
Cash/Sales	−						
Cash Flow/Total Debt	+						
Current Assets/Current Liabilities	+						
Current Assets/Total Assets	+						
Current Assets/Sales	+						
Earning before Tax and Interests/Total Assets	+						
Retained Earnings/Total Assets	+						
Net Income/Total Assets	?	+	+	+	n/s	n/s	n/s
Total Debt/Total Assets	−						
Sales/Total Assets	?	−	−	+	n/s	n/s	n/s
Working Capital/Total Assets	?	+	+	+	n/s	n/s	n/s
Working Capital/Sales	?	−	−	−	n/s	n/s	n/s
Quick Assets/Total Assets	?	+	+	+	n/s	n/s	n/s
Quick Assets/Current Liabilities	?	+	+	+	n/s	n/s	n/s
Quick Assets/Sales	?	−	−	−	−	−	n/s
Market Value of Equity/Total Capitalization	?	+	+	+	n/s	n/s	n/s
Cash/Current Liabilities	+						
Current Liabilities/Equity	−						
Inventory/Sales	−						
Equity/Sales	−						
Market Value of Equity/Total Debt	+						
Net Income/Total Capitalization	+						

[a]1: One-Year Prediction [b]2: Two-Year Prediction [c]3: Three-Year Prediction
+: Beneficial −: Harmful ?: Unknown n/s: Not Selected

B. Setup

Isotonic separation experiments required an additional consideration to determine order restrictions. Common sense and previous studies [11], [48], which we call domain knowledge, suggest order restrictions on some ratios, as shown in the second column of Table V. For instance, a firm with a lower debt to asset ratio (labeled with "−") is less likely to go bankrupt, while a firm with a higher asset to liability ratio (labeled with "+") is less likely to go bankrupt. Ratios labeled with "?" are ambiguous in whether higher values suggest healthier firms. Especially, it was shown in previous research [45] that a firm with higher a net income to total asset ratio was less likely to go bankrupt, but unusually high values of the ratio (in firms' annual reports) were often found among bankrupt firms. For those ratios, we had to rely on what the data indicated, which involved considering various possibilities of order restrictions on these ratios.

We performed isotonic separation training on the datasets with all 23 ratio values (as shown in the third to fifth columns of Table V) and isotonic separation training with feature selection (as shown in the sixth to eighth columns of Table V). The discovery of order restrictions on these ratios was consistent over the three different time line experiments, except on the sales to

TABLE VI
TESTING ERROR RATES: EXPERIMENTS WITH ALL 23 RATIOS

Classification Methods	Testing Error Rates (%)								
	1-Year Prediction			2-Year Prediction			3-Year Prediction		
	Type 1	Type 2	Total	Type 1	Type 2	Total	Type 1	Type 2	Total
Isotonic Separation	27.27	19.32	23.30	22.02	29.36	25.69	22.12	31.73	26.92
Linear DA	15.91	32.95	24.43	17.43	22.94	25.69	17.31	33.65	25.48
Logistic DA	11.36	62.50	36.93	22.94	29.36	26.15	18.27	37.50	27.88
Probit	12.50	62.50	37.50	26.61	30.28	28.44	17.31	36.54	26.92
LP Discrimination	20.45	28.41	24.43	16.51	31.19	23.85	18.27	34.62	26.44
Neural Net.	23.86	22.73	23.30	27.52	24.77	26.15	27.88	27.88	27.88
Vector Quantization	23.86	21.59	22.73	25.69	23.85	24.77	15.38	28.85	22.12
ID3/C4.5	21.59	29.32	20.45	23.85	26.61	25.23	28.85	27.88	28.37
OC1	19.32	22.73	21.02	31.19	16.51	23.85	25.00	27.88	26.44
Rough Set	18.18	20.45	19.32	25.69	15.60	20.64	23.08	23.08	23.08

TABLE VII
SELECTED FINANCIAL RATIOS

Financial Ratios	Step. Disc.			MIFS			Sequential Elimination					
							Iso. Sep.			LP Disc.		
	1[a]	2[b]	3[c]	1[a]	2[b]	3[c]	1[a]	2[b]	3[c]	1[a]	2[b]	3[c]
Cash Flow/Total Assets										✓	✓	
Cash/Sales										✓		
Cash Flow/Total Debt							✓					
Current Assets/Current Liabilities					✓							
Current Assets/Total Assets							✓	✓	✓		✓	✓
Current Assets/Sales												
Earning before Tax and Int./Total Assets										✓		
Retained Earnings/Total Assets					✓							
Net Income/Total Assets											✓	
Total Debt/Total Assets	✓	✓	✓								✓	✓
Sales/Total Assets				✓		✓						
Working Capital/Total Assets										✓	✓	
Working Capital/Sales			✓								✓	
Quick Assets/Total Assets	✓	✓	✓									✓
Quick Assets/Current Liabilities												
Quick Assets/Sales			✓				✓	✓			✓	
Market Value of Equity/Total Cap.												
Cash/Current Liabilities									✓			
Current Liabilities/Equity				✓	✓	✓						
Inventory/Sales							✓			✓	✓	
Equity/Sales												
Market Value of Equity/Total Debt				✓	✓	✓	✓	✓	✓			
Net Income/Total Capitalization												
Total Number of Selected Features	2	2	4	3	4	3	5	3	3	5	8	3

[a]1: One-Year Prediction [b]2: Two-Year Prediction [c]3: Three-Year Prediction

total asset ratio for the three-year prediction case which differed from others when all 23 ratios were considered. This ratio, however, was not chosen in the three-year prediction case when the feature selection process was applied. The net income to total asset ratio which was set to be unknown for isotonic separation due to a previous study [45] was omitted by the feature selection process, though the isotonic method performed better with the positive order when all features were included. This indicated that the net income to total asset ratio did not demonstrate a clear order restriction in the dataset and thus the isotonic method performed better without this feature. Similar observations were made regarding other ratios.

In our experiments, we considered misclassification costs and the prior probability of bankruptcy using the expected risk term [18], [23], which is defined as

$$c_1\pi_1\frac{n_1}{|A_1|}+c_0(1-\pi_1)\frac{n_0}{|A_0|}$$

where A_1 and A_0 are the sets of bankrupt and nonbankrupt firms in the validation data, n_1 and n_0 are the numbers of misclassifications among bankrupt and nonbankrupt firms, π_1 is the prior probability of bankruptcy, and c_1 and c_0 are misclassification costs of bankrupt and nonbankrupt firms. In our data source, approximately 2% of firms filed bankruptcy, i.e., $\pi_1 = 0.02$. Altman [2] estimated that the misclassification of a bankrupt firm would be 32 to 62 times more costly than the misclassification of a nonbankrupt firm, i.e., $32 \le c_1/c_0 \le 62$. By setting $c_1/c_0 \cong 49$, we had

$$c_1\pi_1 \cong c_0(1-\pi_1).$$

In our measure, $n_1/|A_1|$ was labeled as the Type 1 error rate, $n_0/|A_0|$ was labeled as the Type 2 error rate, and $0.5\cdot n_1/|A_1|+0.5\cdot n_0/|A_0| = (n_1+n_1)/(|A_1|+|A_0|)$ was the total error rate. (Note that in our datasets, $|A_1| = |A_0|$.)

C. Results and Discussion

The experiments started with tenfold cross validations of the ten classification methods on the three datasets that consist of all 23 financial ratios. The validation results are summarized in Table VI. The rough set method was the top performer followed by OC1 and the learning vector quantization methods; and neural networks, logistic discrimination, and Probit methods performed worse than others.

There are a number of known data characteristics that affect specific classification methods' performance [13], [34]. Among those are the density and the modality of data. By the density of data, we mean the number of training data points relative to the number of attributes. Previous studies showed that neural networks performed worse than decision tree induction when sparse data were used for training [13]. The training datasets with all 23 financial ratios, which consist of less than 70 data points each, were very sparse. The results of experiments with all 23 ratios shown in Table VI agreed with the previously discussed issue on the density between neural networks and decision tree methods (OC1 and ID3/C4.5, in this paper). The rough set method [49], [50] is affected much less by the scarcity of the data, mainly because of the fact that the rough set process involves an operation similar to feature reduction. Table VI also confirmed this observation. The performance of isotonic separation mainly depends on the quality of the assumed isotonic consistency condition. When all 23 ratios were used, the best isotonic consistency condition, as shown in the third to fifth columns of Table V, was not good enough. Some features such as the net income to total asset ratio made the data less isotonic.

Multimodal data can be placed with little ambiguity at multiple disjointing regions in the feature space. It was shown that neural networks performed clearly better than axis parallel decision tree methods (e.g., ID3/C4.5) with multimodal data, but oblique decision tree methods (e.g., OC1) performed reasonably well [13]. Nearest neighbor methods (e.g., learning vector quantization) were known to perform well with multimodal data, too [34]. We measured the modality of datasets based on the ratio between the within-class deviation and the between-class distance [28]. The multimodality in the datasets with all 23 ratios was not clearly observed. Thus, this factor did not appear to affect the result shown in Table V.

Earlier studies [1], [11], [19], [48], [64] suggested that a relatively small number of financial ratios such as four to five ratios

TABLE VIII
TESTING ERROR RATES: EXPERIMENTS WITH SELECTED RATIOS

Classification Methods	Testing Error Rates (%)								
	1-Year Prediction			2-Year Prediction			3-Year Prediction		
	Type 1	Type 2	Total	Type 1	Type 2	Total	Type 1	Type 2	Total
Isotonic Separation-*c*	11.36	17.05	14.20	13.76	16.52	15.14	22.12	15.38	18.75
Linear DA-*a*	13.64	26.14	19.89**	16.51	28.44	22.48***	15.38	28.85	22.12*
Logistic DA-*a*	17.05	23.86	20.45***	19.27	26.61	22.94***	17.31	27.88	22.60*
Logistic DA-*d*	11.36	29.55	20.45**	17.43	27.52	22.48***	17.31	29.81	23.56**
Probit-*a*	18.18	23.86	21.02***	17.43	27.52	22.46***	16.35	27.88	22.12*
Probit-*d*	7.95	35.23	21.59***	17.43	28.44	22.94***	17.31	29.81	23.56**
LP Discrimination-*a*	18.18	25.00	21.59***	17.43	27.52	22.48***	14.42	27.88	21.15*
LP Discrimination-*d*	9.09	34.09	21.59***	21.10	26.61	23.85***	15.38	28.85	22.12*
Neural Networks-*b*	21.59	22.73	22.16***	25.69	15.60	20.64**	20.19	24.04	22.12*
Neural Networks-*c*	20.45	22.73	21.59***	22.02	16.51	19.27**	15.38	20.04	19.71
Vector Quantization-*b*	15.91	20.45	18.18*	23.85	19.27	21.56***	25.00	25.96	25.48***
Vector Quantization-*c*	21.59	14.77	18.18*	26.61	20.18	23.39***	21.15	24.04	22.60*
ID3/C4.5-*b*	25.00	17.05	21.02***	28.44	11.93	20.18**	28.85	24.04	26.44***
OC1-*b*	25.00	20.45	22.73***	32.11	19.27	25.69***	17.31	24.04	20.67
OC1-*c*	25.00	20.45	22.78***	22.94	22.94	22.94***	23.08	23.08	23.08*
Rough Set-*e*	18.18	20.45	19.32*	25.69	15.60	20.64*	23.08	23.08	23.08

-*a*: Stepwise Discriminant Analysis for Feature Selection (confidence level 0.99)
-*b*: Mutual Information Based Feature Selection (confidence level 0.99)
-*c*: Sequential Elimination with Isotonic Separation
-*d*: Sequential Elimination with LP Discrimination
-*e*: With All Features
Probability of *t*-test against isotonic separation: $^{***}p < 0.01$ (Very Significant) $^{**}p < 0.05$ (Significant) $^{*}p < 0.1$ (Marginally Significant)

were normally sufficient for prediction. We applied the previously described feature selection methods. A subset of data was sampled from each of one-year, two-year, and three-year prediction datasets on which the feature selection methods were applied. Stepwise discrimination was performed using SAS Statistical software, the MIFS process was programmed in C, and sequential elimination for isotonic separation and linear programming discrimination was conducted using the AMPLF/CPLEX system augmented with C programs. (For stepwise discrimination and MIFS, we set the confidence level to 0.99.) The results of feature selection are summarized in Table VII.

The financial ratios selected by the stepwise discriminant analysis largely overlapped with those of Deakin's study [17]; the mutual information based feature selection method and the Altman's study [1] selected a few common ratios. While stepwise discriminant and mutual information based methods are based on theoretically well-established statistical measures and were applied on the three datasets that consist of somewhat similar statistical profiles, the sequential elimination process works on an ad hoc basis. Thus, ratios of selected by the stepwise discriminant method were more consistent across the three datasets; this was also true with the mutual information based method. On the other hand, ratios resulted from sequential elimination with the linear programming discrimination differed across the datasets, while those from sequential elimination with the isotonic separation were somewhat consistent across the three datasets.

As discussed previously, the stepwise discriminant method's feature selection objective is more consistent with the data separation objective of discriminant analyses; and mutual information based method's feature selection objective is more consistent with the data separation objective of neural networks, learning vector quantization, and decision tree induction methods. Thus, they are expected to reduce testing accuracies of corresponding classification methods. On the other hand, sequential elimination specifically is designed for isotonic separation and linear programming discrimination, as a result, it is expected to work best for those methods.

Subsequently, the ten classification methods were evaluated using three datasets in the tenfold validation experiments with each of the selected feature sets. Table VIII lists testing error rates of all classification methods with the best sets of selected ratios. For most cases, the use of selected ratios reduced testing errors significantly. We believe that the improved accuracy was attributed by two factors: the increased density of training data due to reduced features and the factual observation made by earlier studies [1], [11], [19], [48], [64] that a relatively small number of financial ratios such as four to five ratios were normally sufficient for prediction. Especially, for isotonic separation, the feature selection improved the testing accuracies by large percentages. The probabilities of t-tests to evaluate the advantage of isotonic separation over other methods, listed in Table VIII, showed that the isotonic separation approach with sequentially eliminated ratios outperformed other methods for short-term (i.e., within two years) bankruptcy prediction. When three-year prediction was considered, the isotonic separation approach performed no worse than other methods.

D. Limitations

The experimental study presented in the previous section included the isotonic separation approach compared against nine other classification methods with three feature selection methods tested on three cross-industry datasets containing 23 ratios. The performance advantage of isotonic separation over other methods was observed for short-term bankruptcy prediction when selected features were used. There are numerous classification methods applied for various detection and learning problems [42]. The observation of this paper is restricted to isotonic separation and nine other methods that were used in previous studies. Our choice of MIFS and stepwise

discriminant methods for feature selection was based on the natures of the algorithms and previous studies. Various other feature selection methods could affect the experiment results presented in the previous section. Finally, the data characteristics (i.e., cross-industry data of 23 ratios) are another limiting factor of the research observation. The answer for whether the research observation of this paper will hold for a more homogeneous firm dataset (gathered within an industry) or a dataset with other financial ratios requires additional experimental studies.

VII. Concluding Remark

The isotonic separation method and nine other popular classification techniques were evaluated with the firm bankruptcy prediction problem. The results of experiments on three datasets showed that the isotonic separation method a viable technique for firm bankruptcy prediction. Part of the requirement of a good classification system is to possess the capability to provide high accuracy in prediction under diverse situations, and we demonstrated that the isotonic separation technique at least partially fulfilled this goal for short-term bankruptcy prediction.

The isotonic separation method can be extended to perform firm bankruptcy time-line prediction, i.e., prediction of at which point in time a distressed firm will go bankrupt. Using five year firm data, we can create a system that predicts not only whether a firm will go bankrupt but also when it will eventually happen. The underlying idea is similar to the statistical survival analysis, but the main goal is to estimate the explicit survival time-line rather than the survival/hazard function. This type of extension to isotonic separation method will be a worthwhile contribution to the current bankruptcy prediction studies.

References

[1] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *J. Finance*, vol. 23, no. 4, pp. 589–609, 1968.

[2] ——, "Commercial bank lending: Process, credit scoring, and costs of errors in lending," *J. Financ. Quant. Anal.*, vol. 15, pp. 813–832, 1980.

[3] J. A. Anderson, "Separate sample logistic discrimination," *Biometrika*, vol. 59, no. 1, pp. 19–35, 1972.

[4] B. Back, K. Sere, and M. C. van Wezel, "Choosing the best set of bankruptcy predictors," in *Proc. 1st Nordic Workshop on Genetic Algorithms and Their Applications*, 1995, pp. 285–299.

[5] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk, *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. New York: Wileys, 1972.

[6] R. E. Barlow and H. D. Brunk, "The isotonic regression problem and its dual," *J. Amer. Stat. Assoc.*, vol. 67, no. 337, pp. 140–147, 1972.

[7] R. Battiti, "Using mutual information for selection features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[8] W. H. Beaver, "Financial ratios as predictors of failure," *J. Account. Res.*, vol. 4, pp. 71–111, 1966.

[9] ——, "Alternative accounting measures as predictors of failure," *Account. Rev.*, vol. 43, no. 1, pp. 113–122, 1968.

[10] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Opt. Methods Softw.*, vol. 1, pp. 23–34, 1992.

[11] M. Blum, "Failing company discriminant analysis," *J. Account. Res.*, vol. 12, no. 1, pp. 1–25, 1974.

[12] B. V. Bonnlander and A. S. Weigend, "Selecting input variables using mutual information and nonparametric density estimation," in *Proc. 1994 Int. Symp. Artificial Neural Networks*, 1994, pp. 42–50.

[13] D. E. Brown, V. Corruble, and C. L. Pittard, "A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems," *Pattern Recognit.*, vol. 26, no. 6, pp. 953–961, 1993.

[14] R. Chandrasekaran, Y. U. Ryu, V. Jacob, and S. Hong, "Isotonic separation," *INFORMS J. Comput.*, 2005, to be published.

[15] D. R. Cox, "Some procedures associated with the logistic qualitative response curve," in *Research Papers in Statistics: Festschrift for J. Neyman*, F. N. David, Ed. New York: Wiley, 1966, pp. 55–71.

[16] N. E. Day and D. F. Kerridge, "A general maximum likelihood discriminant," *Biometrics*, vol. 23, no. 2, pp. 313–323, 1967.

[17] E. B. Deakin, "A discriminant analysis of predictors of business failure," *J. Account. Res.* , vol. 10, no. 1, pp. 167–179, 1972.

[18] S. Dudoit and M. J. van der Laan, Asymptotics of cross-validated risk Estimation in model selection and performance assessment, Div. Biostatistics, Univ. Calif., Berkeley, Berkeley, CA, 2003.

[19] R. O. Edmister, "An empirical test of financial ratio analysis for small business failure prediction," *J. Financ. Quant. Anal.ysis*, vol. 7, no. 2, pp. 1477–1493, 1972.

[20] M. A. Efroymson, "Multiple regression analysis," in *Mathematical Methods for Digital Computers*, A. Ralston and H. S. Wilf, Eds. New York: Wiley, 1960, pp. 191–203.

[21] *Handbook of Neural Computation*, Oxford Univ. Press, Oxford, U.K., 1997. E. Fiesler and R. Beale.

[22] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 179–188, 1936.

[23] H. Frydman, E. I. Altman, and D.-L. Kao, "Introducing recursive partitioning for financial classification: The case of financial distress," *J. Finance*, vol. 40, no. 1, pp. 269–291, 1985.

[24] A. V. Goldberg, "Recent developments in maximum flow algorithms," NEC Res. Inst., Princeton, NJ, Tech. Rep. 98-045, 1998.

[25] A. V. Goldberg and R. E. Tarjan, "A new approach to the maximum flow problem," *J. ACM*, vol. 35, no. 4, pp. 921–940, 1988.

[26] S. Greco, B. Matarazzo, and R. Słowiński, "A new rough set approach to evaluation of bankruptcy risk," in *Operational Tools in Management of Financial Risks*, 2nd ed, C. Zopounidis, Ed. Dordrecht, The Netherlands: Kluwer, 1998, pp. 121–136.

[27] J. D. F. Habbema, J. Hermans, and K. van den Broek, "A stepwise discriminant analysis program using density estimation," in *Proc. 1974 Conf. Computational Statistics*, 1974, pp. 101–110.

[28] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 1991.

[29] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley, 1991.

[30] V. Jacob, R. Krishnan, Y. U. Ryu, R. Chandrasekaran, and S. Hong, "Filtering objectionable Internet content," in *Proc. 20th Int. Conf. Information Systems*, 1999, pp. 274–278.

[31] R. I. Jennrich, "Stepwise discriminant analysis," in *Statistical Methods for Digital Computers*, K. Enslein, A. Ralston, and H. S. Wilf, Eds. New York: Wiley, 1977, vol. 3, pp. 76–96.

[32] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

[33] A. V. Karzanov, "Determining the maximal flow in a network by the method of preflows," *Sov. Math. Doklady*, vol. 15, pp. 434–437, 1974.

[34] M. Y. Kiang, "A comparative assessment of classification methods," *Decision Support Syst.*, vol. 35, no. 5, pp. 441–454, 2003.

[35] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1/2, pp. 273–324, 1997.

[36] T. Kohonen, "New developments of learning vector quantization and the self-organizing map," in *Proc. 1992 Symp. Neural Networks: Alliances and Perspectives in Senri*, 1992.

[37] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Heidelberg, Germany: Springer-Verlag, 2001.

[38] A. Kumar, "New technique for data reduction in a database system for knowledge discovery applications," *J. Intell. Inf. Syst.*, vol. 10, no. 31–48, 1998.

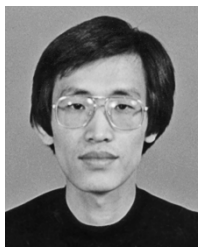
[39] P. Leary and P. Gallinari, "Feature selection with neural networks," *Behaviormetrika*, vol. 26, no. 1, 1999.

[40] K. C. Lee, I. Han, and Y. Kwon, "Hybrid neural network models for bankruptcy predictions," *Decision Support Syst.* , vol. 18, no. 1, pp. 63–72, 1996.

[41] W. Li, "Mutual information functions versus correlation functions," *J. Stat. Phys.*, vol. 60, no. 5/6, pp. 823–837, 1990.

[42] T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms," *Mach. Learn.*, vol. 4, no. 3, pp. 203–228, 2000.

[43] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA: Kluwer, 1998.
[44] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Oper. Res.*, vol. 43, no. 4, pp. 570–577, 1995.
[45] T. E. McKee and T. Lensberg, "Genetic programming and rough sets: A hybrid approach to bankruptcy classification," *Eur. J. Oper. Res.*, vol. 138, no. 2, pp. 436–451, 2002.
[46] W. F. Messier and J. V. Hansen, "Inducing rules for expert system development: An example using default and bankruptcy data," *Manage. Sci.*, vol. 34, no. 12, pp. 1403–1415, 1988.
[47] S. K. Murthy, S. Kasrif, and S. Salzberg, "A system for induction of oblique decision trees," *J. Artif. Intell. Res.*, vol. 2, pp. 1–32, 1994.
[48] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *J. Account. Res.*, vol. 18, no. 1, pp. 109–131, 1980.
[49] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht, The Netherlands: Kluwer, 1992.
[50] Z. Pawlak and R. Słowiński, "Decision analysis using rough sets," *Int. Trans. Oper. Res.*, vol. 1, no. 1, pp. 107–114, 1994.
[51] J. R. Quinlan, "Introduction to decision trees," *Mach. Learn.*, vol. 1, pp. 81–106, 1986.
[52] ——, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
[53] Y. U. Ryu, R. Chandrasekaran, and V. Jacob, "Prognosis using an isotonic prediction technique," *Manage. Sci.*, 2005, to be published.
[54] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1998.
[55] K.-S. Shin and Y.-J. Lee, "A genetic algorithm application in bankruptcy prediction modeling," *Expert Syst. Appl.*, 2005, to be published.
[56] R. Słowiński, "Rough set learning of preferential attitude in multi-criteria decision making," in *Methodologies for Intelligent Systems: Proc. 7th Int. Symp.*, 1993, pp. 642–651.
[57] K. Słowiński, R. Słowiński, and J. Stefanowski, "Rough sets approach to analysis of data from peritoneal lavage in acute pancreatitis," *Medican Inf.*, vol. 13, pp. 155–159, 1998.
[58] *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, R. Słowiński, Ed., Kluwer, Dordrecht, The Netherlands, 1992, pp. 445–456. R. Słowiński, J. Stefanowski, 'RoughDAS' and 'RoughClass' software implementations of the rough set approach.
[59] R. Słowiński and C. Zopounidis, "Application of the rough set approach to evaluation of bankruptcy risk," *Intell. Syst. Account., Finance, Manage.*, vol. 4, pp. 27–41, 1995.
[60] F. W. Smith, "Pattern classifier design by linear programming," *IEEE Trans. Comput.*, vol. C-17, no. 4, pp. 367–372, 1968.
[61] K. Y. Tam and M. Y. Kiang, "Managerial applications of neural networks: The case of bank failure predictions," *Manage. Sci.*, vol. 38, no. 7, pp. 926–947, 1992.
[62] S. Tyler, "Forecasting bankruptcy more accurately: A simple hazard model," *J. Bus.*, vol. 74, no. 1, pp. 101–124, 2001.
[63] R. L. Wilson and R. Sharda, "Bankruptcy prediction using neural networks," *Decision Support Syst.*, vol. 11, no. 5, pp. 545–557, 1994.
[64] C. V. Zavgren, "The prediction of corporate failure: The state of the art," *J. Account. Literature*, vol. 2, pp. 1–38, 1983.



Young U. Ryu received the Ph.D. degree in management science and information systems from the University of Texas, Austin, in 1992.

Since 1992, he has been affiliated with the Department of Information Systems and Operations Management, School of Management, University of Texas, Dallas, where he is currently an Associate Professor. His main interests of study include logic modeling, data mining, database, and information security.

Wei T. Yue received the Ph.D. degree in management science, concentration management information systems, from Purdue University, West Lafayette, IN, in 2003.

He is currently an Assistant Professor in the Department of Information Systems and Operations Management, University of Texas, Dallas. His research interests include classification methods and information security.