An empirical investigation of some data effects on the classification accuracy of probit, ID3, and neural networks*

TING-PENG LIANG *Purdue University*

JOHN S. CHANDLER *University of Illinois at Urbana-Champaign*

INGOO HAN *Kookmin University*

JINSHENG ROAN *National Chung-Chen University*

Abstract. This paper reports an investigation of some data and method effects on the predictive accuracy of LIFO/FIFO classification models. The methods compared were probit, ID3, and neural networks. Experiments were conducted to study the effect of data characteristics on classification accuracy and the situations under which a particular method performs better. Hold-out samples were used to calculate the predictive accuracy. The results indicate that (1) different methods identify different factors that affect the LIFO/FIFO choice and (2) in hold-out tests, neural network models have the highest average predictive accuracy, whereas ID3 models have the lowest. Neural network models are the best when dominant nominal variables are present; otherwise, probit models are the best.

Résumé. Les auteurs rapportent les résultats d'une analyse de l'incidence de certaines données et de certaines méthodes sur le pourcentage de prévisions exactes dérivées des modèles de classification selon l'épuisement à rebours et l'épuisement successif. Ils comparent la méthode probit, la méthode ID3 et la méthode des réseaux neuronaux et procèdent à des expériences destinées à l'étude de l'incidence des caractéristiques de certaines données sur ce pourcentage et des situations dans lesquelles une méthode particulière donne de meilleurs résultats. Les auteurs ont recours, pour la démonstration, à des échantillons à partir desquels est calculé le pourcentage de prévisions exactes. Les résultats révèlent que 1) les facteurs influant sur le choix de l'épuisement à rebours ou de l'épuisement successif diffèrent selon la méthode utilisée et que 2) dans les tests ayant servi à la démonstration, les modèles de réseaux neuronaux présentent le meilleur pourcentage moyen de prévisions exactes, alors que les modèles ID3 ont le plus faible pourcentage de prévisions exactes. Les modèles de réseaux neuronaux donnent les meilleurs résultats lorsqu'il y a des variables nominales dominantes, faute de quoi les modèles probit sont les meilleurs.

Introduction

In the past decade, probit has been one of the primary methods used in studying accounting classification problems such as accounting method choice or bankruptcy prediction (e.g., Dopuch and Pincus, 1988; Hagerman and Zmijewski, 1979; Lee and Hsieh, 1985). Although probit has been argued to be theoretically superior to both multivariate discriminant analysis (MDA) and ordinary least square regression (e.g., Dietrich and Kaplan, 1982)[1] in classification research, limitations exist when nominal variables are involved. In this case, dummy variables are often used to represent different values of the nominal variables, which may result in a violation of the assumption that the error term has a cumulative normal distribution (Aldrich and Nelson, 1984). In addition, the assumption that the dependent variable is a linear function of the independent variables may be questionable when nominal variables exist.

Recently, nonparametric decision-tree techniques, such as recursive partitioning algorithm (RPA), iterative dichotomizer 3 (ID3), and neural networks (NN) have been considered as alternatives to traditional parametric methods for classification. For example, Braun and Chandler (1987) found ID3 to be better than discriminant analysis in predicting stock market behavior. Messier and Hansen (1988) indicated ID3 to be better than discriminant analysis in predicting loan defaults and bankruptcies. Garrison and Michaelsen (1989) and Parker and Abramowicz (1989) reported ID3 to be better than both discriminant analysis and probit in tax decisions. Although the positive evidence indicates the potential of the ID3 approach, it is not without questions. For example, one common element in these studies is the dominance of nominal variables. Furthermore, some research findings indicate that RPA, a decision-tree technique similar to ID3, is not better than probit or logit, especially when the data do not include nominal variables (Elliott and Kennedy, 1988; Marais, Patell, and Wolfson, 1984). NN is a nonlinear modeling technique that has gained popularity recently. Literature has reported excellent performance in pattern recognition and other classification problems (Fisher and McKusick, 1989; Khanna, 1990; White, 1989). It is interesting to investigate the performance of these methods in accounting settings and the interactions existing between data characteristics and method performance.

In this research, experiments are conducted to investigate the sensitivity of each method to the training sample size and the nature of the data set. The particular accounting problem studied is the LIFO/FIFO decision because "it has occupied an important place in financial reporting and the accounting literature for almost 50 years" (Lindahl, Emby, and Ashton, 1988). In addition, the industry-specific feature of LIFO/FIFO choice allows the effect of a nominal variable to be examined using real world data. Our empirical findings include the following.

First, different methods identify different factors that affect LIFO/FIFO choice. This raises a concern about the effect of research methods on the interpretation of research findings, although for ID3 and NN models, confirmatory analysis that tests theories is not yet available.

Second, probit dominates ID3 and NN in holdout tests when all variables

* The authors thank James Gentry; Bill Scott, the editor; and reviewers for their helpful comments on earlier versions of the paper.

1 Counterarguments exist. For example, Noreen (1988) shows that (1) the rejection regions for the probit test statistics are not well specified for small samples and (2) the ordinary least square regression seems to perform at least as well as probit for the cases considered.

are non-nominal financial ratios, whereas NN performs better when variables include both nominal and non-nominal. This indicates that, compared to probit, NN can handle nominal variables better. This may be due to the existence of nominal variables that causes violations of some probit assumptions.

Third, ID3 is less sensitive to the decrease of training sample sizes. When the variables include both nominal and non-nominal, NN performs better in holdout tests if the sample size of the input data set is small relative to the total population, whereas probit performs better when the sample size is relatively large.

The remainder of this article is organized as follows. The next section briefly compares probit, ID3, and NN, and then some methodological issues in LIFO/FIFO research are examined. The experiments and the results are discussed, and the conclusion presents the findings and some implications.

A comparison of probit, ID3, and neural networks

The probit method uses statistical inference procedures to derive a linear model from a set of input data. The model estimates the likelihood that, given the input data, each case falls in a particular class. It has several assumptions. First, the dependent variable is categorical. Second, the error term has a cumulative normal distribution. Third, no two or more independent variables are perfectly correlated. Fourth, there is no serial correlation of the dependent variable among the cases. Based on these assumptions, probit estimates the parameters of the linear model by the maximum likelihood estimation (MLE) procedures (e.g., Aldrich and Nelson, 1984).

Unlike probit analysis that constructs linear models, decision-tree methods derive decision trees from data automatically.[2] ID3 is one of such methods originally developed by Hunt, Martin, and Stone (1966) and later implemented and expanded by Quinlan (1979, 1982). It assumes that the entire space of possible events begins as a single category and then applies specialization operators to recursively partition the space to maximize the likelihood of containing positive events (see Braun and Chandler, 1987, and Garrison and Michaelsen, 1989, for a detailed explanation of the ID3 algorithm).

Neural networks (NN) are a family of modeling techniques with origins in cognitive sciences. The format of a neural network model is a set of connected nodes. Each node in the network is called a neuron. A neuron contains functions for the summation of inputs and transfer of inputs to outputs. Neurons are organized into layers. The most popular architecture is a three-layer design that includes input, output, and a hidden layer. Each neuron in the input layer is connected to all neurons in the hidden layer, and each neuron in the hidden layer is connected to all neurons in the output layer. Each connection has a weight called "connection weight." The output value of a neuron is a function of the weighted sum of its inputs. In fact, probit can be represented as a two-layer (no hidden layer) network with the normal density function as its transfer function. The addition of a hidden layer, however, increases the flexibility of network models and allows nonlinearity to be modeled. A discussion of the statistical properties of neural networks can be found in White (1989), and a review of different neural networks paradigms can be found in Khanna (1990).

Probit, ID3, and NN methods are different in at least the following aspects. First, ID3 and NN methods make no assumption on data distribution, but probit analysis assumes that the error term is normally distributed. Therefore, ID3 and NN may be more appropriate when the normality assumption is likely to be violated, and probit may be more appropriate otherwise.[3]

Second, these methods generate models in different functional forms. The ID3 method generates decision tree models in which the effects of different factors are not compensatory. In other words, ID3 partitions the entire event space into several discrete, nonlinear discriminant regions. This allows data whose discriminant space is nonlinear to be classified. NN also assumes nonlinear relationships between independent and dependent variables, but the functional form is sigmoid (the most popular one). Probit analysis assumes a linear compensatory relationship among independent variables. This implies that ID3 and NN may be more appropriate when the problem involves nominal variables that make a linear model inappropriate and probit may be more appropriate otherwise.

Third, these methods have different model construction processes. ID3 constructs models in an exhaustive decomposition process. It continuously divides the input observations to increase the homogeneity within subsets. NN uses a trial-and-error process (the most popular approach is called "feedforward with error backpropagation") to find a set of connection weights that minimize the sum of squared errors. The probit method focuses on optimizing the probability of correct classification. Therefore, ID3 and NN are more likely to overfit the sample data and hence may be more sensitive to the noise in the input data set, though some attempts have been made to reduce the overfitting problem.

Finally, the criteria used for selecting variables in different methods have different biases. The entropy function used in ID3 is a logarithmic function generally biased toward variables with more levels and against variables with less levels (Mingers, 1987). In other words, variables with more levels are more likely to have a higher priority in the model construction process. NN and probit models do not have this bias in processing numerical variables, but probit analysis may favor attributes with less levels when dummy variables are used to handle nominal variables.

2 ID3 and RPA are the two most popular decision-tree methods. They are similar in nature but different in the criteria they used in constructing a tree. ID3 minimizes the overall entropy, whereas RPA minimizes the misclassification costs. ID3 has many different versions. The one we used for this research was the original algorithm. Readers interested in RPA should read Breiman, Friedman, Olshen, and Stone (1984).

3 This and some of the following statements are in fact hypotheses that can be tested. The empirical study to be presented later explores some of them. To reach a conclusion, however, more research is necessary.

Given these differences, it would be interesting to know whether these three methods have different properties when they are applied to accounting classification problems. Are the variables identified by the three methods different? Do the models have different levels of accuracy? Which method is better? When and why does a particular method outperform the other? In the remaining sections, we describe experiments investigating some of these issues in the context of LIFO/FIFO choices.[4]

Background of LIFO/FIFO research

Choice of inventory accounting methods has been a research issue for the past several decades (Lindahl et al., 1988). Theoretically the LIFO method has tax advantages when inflation exists. In practice, however, a majority of firms still adopt FIFO as their primary inventory accounting method. As a result, much research has been conducted to investigate the factors affecting the adoption of a certain method (e.g., Biddle, 1980; Cushing and LeClere, 1988; Dopuch and Pincus, 1988; Lee and Hsieh, 1985; Morse and Richardson, 1983).

Previous literature has examined at least three potential theories of LIFO/FIFO choice: Ricardian costs, agency costs, and political costs (Lee and Hsieh, 1985). The Ricardian hypothesis assumes that the inventory method choice is based on a firm's comparative advantage in tax minimization associated with the production-investment opportunity set. A particular method (e.g., LIFO) will be adopted if its tax savings exceed the implementation costs. Therefore, LIFO may be the optimal choice for some firms, whereas FIFO is the optimal choice for others. The agency cost hypothesis assumes that some firms remain on FIFO to report higher earnings because of managers' concerns about the impact of a LIFO switch on the securities market or their compensation contracts (e.g., Abdel-khalik, 1985; Dhaliwal, Salamon, and Smith, 1982; Ricks, 1982). Managers are willing to forgo potential tax savings to obtain other benefits. The political costs hypothesis assumes that a method will be chosen based on political costs as well as potential tax savings (Daley and Vigeland, 1983). For example, the dominating firm in an industry may choose LIFO to reduce its reported earnings to avoid being the target of antitrust laws.

Probit has been the major method used in previous studies to test these hypotheses. Empirical findings, however, are inconclusive in many aspects. For example, the relative frequency of price increases was found to be significantly different between LIFO and FIFO firms by Lee and Hsieh (1985); but the effect was insignificant in Dopuch and Pincus (1988).

Because of the significance of LIFO/FIFO research, we selected this domain to compare the ID3, NN, and probit methods. In addition, the industry type allows us to examine the effect of nominal variable dominance.[5] In this study, we examine some data and method effects on the classification accuracy of LIFO/FIFO choices. Holdout data are used to investigate how training sample size and the nature of data affect the predictive accuracy of these methods.

4 These questions are general and, of course, cannot be answered conclusively by simply testing one data set. Analytical and simulation studies are necessary. Some simulation results can be found in Han (1990) and Liang (1992).

The experiments

Data collection

Based on theories and previous research findings, 12 explanatory variables affecting LIFO/FIFO choices were selected, which included 1 nominal and 11 numerical variables.[6] Each variable is related to one or more of the following concerns: nature of industry, inflation and its variability, inventory and its variability, inventory controllability, capital intensity, and debt/equity ratio. Our purpose in variable selection is not to determine whether previous LIFO/FIFO research findings are correct but to develop a set of LIFO/FIFO data on which the effect of different techniques can be compared.

After selecting the variables, data were collected from the COMPUSTAT data base. The inflation data necessary for calculating the growth and the variance of growth of input prices were collected from the Data Resource Institute (DRI) tape. The criterion for selection was that the firms must have used LIFO or FIFO only for at least 10 consecutive years. Since many firms switched from FIFO to LIFO in 1974 in response to the oil crisis, we set 1976 as the starting year to obtain samples. Data were collected from 1975 to 1984 and aggregated over 10 years for calculating means and variances. Ten-year average figures were used for variables such as net sales and total assets. Initially, 220 FIFO firms and 60 LIFO firms were identified. Three of them were later eliminated because of missing data. These firms were distributed in 23 industries, as listed in Table 1.

Since more than one surrogate variable may reflect the same theoretical factor in our initial data base, high correlations exist among them (see Table 2). To test the effect of different variables and methods in classification research, we compiled six data sets of different variables from the initial data set. First, after considering the multicollinearity issue, we compiled three sets of data with eight numerical variables each.[7] For example, since net sales and total assets have a high degree of correlation, only one of them is included in the data sets. This allows us to examine the effect of using different surrogate variables in model construction. Second, we added the nominal variable, industry type, to the three

5 Due to the industry domination observed in LIFO/FIFO decisions, the results obtained from this study may not be readily generalizable to other domains.

6 Papers consulted during the variable selection process include Abdel-khalik (1985); Cushing and LeClere (1988); Dopuch and PIncus (1988); Eggleton, Penman, and Twombly (1976); Hagerman and Zmijewski (1979); Lee and Hsieh (1985); and Morse and Richardson (1983).

7 Eight variables in each data set were arbitrary. One factor we considered, however, was that these data sets should share some common variables that appeared significant in previous research and contain a few variables unique to individual data sets.

TABLE 1
Distribution of sample firms

SIC code*	Description	FIFO	LIFO
20	Food and Kindred Products	6	0
22	Textile Mill Products	3	3
23	Apparel and Other Finished Products made from Fabrics and Similar Materials	14	0
24	Lumber and Wood Products except Furniture	5	1
25	Furniture and Fixtures	1	0
26	Paper and Allied Products	3	2
27	Printing, Publishing, and Allied Industries	10	3
28	Chemicals and Allied Products	13	4
29	Petroleum Refining and Related Industries	1	3
30	Rubber and Miscellaneous Plastic Products	3	4
31	Leather and Leather Products	2	1
32	Stone, Clay, Glass, and Concrete Products	2	2
33	Primary Metal Industries	1	7
34	Fabricated Metal Products, except Machinery and Transportation Equipment	8	7
35	Industrial and Commercial Machinery and Computer Equipment	14	7
36	Electronic and Other Electrical Equipment and Components, except Computer Equipment	60	3
37	Transportation Equipment	15	1
38	Measuring, Analyzing, and Controlling Instruments; Photographic, Medical and Optical Goods; Watches and Clocks	20	2
39	Miscellaneous Manufacturing Industries	6	2
50	Wholesale Trade — Durable Goods	12	4
51	Wholesale Trade — Nondurable Goods	11	1
53	General Merchandise Stores	1	0
59	Miscellaneous Retails	6	3
	Total	217	60

*Two-digit industrial SIC code.

sets to form another three data sets. This allows us to examine the effect of nominal variables in model construction.

Data analysis

The data analysis is divided into two parts. In the first part, we examine the variables selected by probit and ID3 methods. The NN method was not used in this stage because we did not have a proper method to determine the relative importance of variables in NN models. In the second part, we use holdout samples to compare the predictive accuracy of probit, ID3, and NN methods.

Difference in variables. For each data set, probit and ID3 were applied to construct models to examine the difference due to using different surrogate

TABLE 2
Correlation matrix

	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X2	−.06	−.04	.98	−.10	−.01	−.16	−.20	.36	.45	.19
X3		−.20	−.05	.84	.01	.07	−.05	−.04	.08	−.08
X4			−.04	−.16	.15	.05	.07	−.02	−.01	.00
X5				−.08	−.01	−.13	−.22	−.41	.47	.19
X6					−.02	.11	.01	−.10	.00	−.02
X7						−.07	−.12	.06	−.13	.08
X8							.69	−.12	.04	−.30
X9								−.50	−.13	−.14
X10									.27	−.01
X11										−.30

X1 = Industry type
X2 = Net sales
X3 = CV (standard deviation/mean) of net sales
X4 = CV of net sales growth
X5 = Total assets
X6 = CV of inventory
X7 = Long-term debt/Equity
X8 = Inventory/Net sales
X9 = Inventory/Total assets
X10 = Gross capital intensity
X11 = Growth of input price
X12 = CV of growth of input price

variables.[8] The results of probit analysis, as shown in Table 3, indicate that there is no major difference in the variables identified as significant by probit when different surrogate variables are used, although the significance level of certain variables may change slightly. For example, long-term debt/equity is significant in model 1 but insignificant in models 2 and 3.

Another effect we observed is the impact of nominal variables. By comparing models 1, 2, and 3 with models 4, 5, and 6 in Table 3, we find that three variables become significant because of the existence of a nominal variable (industry type): net sales, long-term debt/equity, and growth of input price. However, the significance of gross capital intensity decreases. All the dummy variables for different industries are not statistically significant. In summary, the results in Table 3 suggest that the addition or deletion of a nominal variable may change the significance levels of other variables and hence affect the reliability of hypothesis testing in probit. This may be due to high correlations between some variables.

After probit analysis, ID3 was applied to the same data sets.[9] The variables in the resulting decision tree models are then compared with those in the probit

8 The industry type was coded as 22 dummy variables for probit analysis.
9 The software used to run the ID3 algorithm is called ACLS, which stands for Analog Concept Learning System. Complete decision trees derived from the data sets are omitted to save space.

TABLE 3
Probit models derived from six data sets

Variable	Model 1 (Set 1)	Model 2 (Set 2)	Model 3 (Set 3)	Model 4 (Set 4)	Model 5 (Set 5)	Model 6 (Set 6)
X1	—	—	—	Coefficient ≤ .01 and insignificant		
X2	0.953	—	0.912	1.781*	—	1.783*
X3	0.433	0.819	—	0.993	1.291	—
X4	—	—	−0.427	—	—	−1.119
X5	—	1.014	—	—	1.450	—
X6	−2.858†	−3.267‡	−3.758†	−2.742†	−3.082†	−3.154†
X7	−1.656*	−1.602	−1.558†	−3.188†	−3.017†	−2.837†
X8	−2.971†	—	−3.018‡	−3.193†	—	−3.221†
X9	—	−0.536	—	—	−1.097	—
X10	2.216*	1.808*	2.221*	1.471	1.254	1.331
X11	1.025	0.906	1.079	2.191*	2.121*	2.319*
X12	−1.126	−0.495	−1.184	1.328	1.333	1.380
Log likelihood ratio	57.92	48.70	57.93	127.27	115.94	128.30
Classification accuracy	83.03	81.59	82.67	86.28	87.00	85.92

The meaning of each variable is listed in Table 2.
*Significant at least at 5% level.
†Significant at least at 0.1% level.
‡Significant at least at 1% level.
— Variables not included in the model.

models. Assuming the variables chosen earlier by the ID3 algorithm[10] or tested more significant by probit to be more important, we can compare the resulting ID3 and probit models to find two effects. First, the factors selected by ID3 and probit were different. For example, SIC code was insignificant in probit but selected as the first variable for constructing trees in ID3 models (see the appendix). Second, different factors were identified by ID3 for different industries (also see the appendix). For instance, long-term debt/equity was found important in printing, publishing, and allied industries (SIC code 27) but irrelevant in the lumber (24) or chemical (28) industries. This implies that ID3 may be capable of identifying the industry-specific nature of inventory accounting choices.

In addition to the differences in model format and variables included in a model, the classification accuracies are different. The classification accuracy is calculated as the percentage of the cases in the input data set that are correctly classified by the model. The accuracy of the probit models is shown in Table 3. Since the ID3 algorithm tries to cover as many observations as possible in the process of model construction, a perfect classification accuracy is usually

10 Again, ID3 cannot test the significance of a variable, but it chooses variables based on their discriminant power measured by entropy. One risk is that if two variables provide similar partitions of the data set, then one may be completely shadowed by the other and never appear in the decision tree.

achieved unless conflicting data exist in the samples. A potential problem associated with the ID3 algorithm, however, is that it tends to overfit the input data and hence reduce the reliability of the resulting models.

Difference in predictive accuracy: Data preparation. The second part of the experiment uses holdout data to compare the predictive accuracy of probit, ID3, and NN models. Two factors that may affect the applicability of a particular method were investigated: nature of the data set and training sample size. The nature of data sets is differentiated by whether they have a nominal variable and, if they do, whether the nominal variable has a dominant effect. Therefore, the experimental design includes four independent variables: data analysis method, presence of a nominal variable, dominance of the nominal variable, and training sample size. They are organized into a $3 \times 2 \times 2 \times 3$ factorial design.

The methods investigated were probit, ID3, and NN. The presence of a nominal variable may be yes or no and the dominance of the nominal variable may be high or low. In this experiment, industry type is the only nominal variable used to differentiate data sets. A data set is said to be highly dominated by a nominal variable if the variable (i.e., the industry type) alone can correctly classify a significant portion (e.g., 80 percent of the input cases. The training sample sizes included three levels: large/small (L/S), medium/medium (M/M), and small/large (S/L). Large/small means using a large portion (2/3 in this research) of the sample to derive the model for predicting a small portion (1/3) of holdout observations. Medium/medium means using about half of the sample to predict the other half. Small/large means using a small portion (1/3) of the sample to predict the remaining two-thirds of the observations.

The dependent variable was the predictive accuracy of the model derived in a particular setting. It was defined as follows:

$$\text{Predictive accuracy} = \frac{\text{Number of holdout cases correctly predicted}}{\text{Total number of holdout cases}}$$

We examine the effects of three factors: method, existence of nominal variables, and training sample size.

1 *Effect of methods*—Because previous findings comparing probit, ID3, and NN are inconclusive, we assume that the average predictive accuracies of these methods are the same.
2 *Effect of nominal variables*—The effects of nominal variables include the existence and the degree of dominance of these variables. Since both ID3 and NN can better handle nominal variables, we expect them to perform better when a nominal factor has a significant effect on the decision outcome and probit to perform better otherwise.
3 *Effect of training sample size*—The normality assumption usually is true only when the training sample size is large. Since ID3 and NN do not require this assumption, we expect ID3 and NN to be less sensitive to the decrease of sample sizes.

TABLE 4
Composition of LDOM and HDOM data sets

LDOM			HDOM		
SIC code	FIFO	LIFO	SIC code	FIFO	LIFO
22	3	3	20	6	0
26	3	2	23	14	0
27	10	3	24	5	1
28	13	4	33	1	7
30	3	4	36	60	3
34	8	7	37	15	1
35	14	7	38	20	2
39	6	2	51	11	1
50	12	4			
59	6	3			
Total	78	39	Total	132	15

To examine these effects, the data were decomposed into two sets with different characteristics. One set was composed of firms in industries not dominated by a particular inventory accounting method, whereas the other set consisted of firms in industries dominated by a single method. The degree of industry dominance used to differentiate these two sets was 80 percent. In other words, industries with at least 80 percent of their firms using the same method were classified as high industry dominance (HDOM). The remaining industries were classified as low industry dominance (LDOM). The industries with less than five firms in the original data set were eliminated to avoid a potential problem that all firms in the industry may be selected in a training or testing data set. Table 4 lists the two-digit SIC codes and number of firms included in these data sets. If we define the degree of industry dominance as the percentage of the firms in the data set whose actual inventory method can be correctly classified by observing the industry type only, these two data sets have 93.9 percent and 67.5 percent industry dominance, respectively. In the case where industry type has no effect, the degree of dominance should be 50 percent, that is, LIFO and FIFO have an equal opportunity to be predicted.

After dividing the initial data by industry dominance, training and testing sets used for comparing probit, ID3, and NN were constructed. For both HDOM and LDOM data sets, 30 pairs of training and holdout data subsets were randomly compiled. These subsets have three different levels of sample sizes, large/small, medium/medium, and small/large, resulting in a total of six different settings as shown in Table 5.

For each setting, 10 pairs of training and holdout data sets were compiled. The variables selected in model 4 (see Table 3) were used in the analysis. The nominal variable was represented as several dummy variables (one dummy for each two-digit SIC code) for probit and NN, but as a single nominal variable

TABLE 5
Six different settings by sample size and industry dominance

	LDOM			HDOM		
Size	Training	Holdout	Total	Training	Holdout	Total
L/S	78	39	117	98	49	147
M/M*	58	58	116	78	78	146
S/L	39	78	117	49	98	147

* One company was randomly held out to make the training and holdout observations equal.

for ID3. Finally, the industry type variable in these data sets was eliminated to form another 60 pairs of data sets with only non-nominal variables, resulting in a total of 120 data sets.

Findings in predictive accuracy. For each pair of training and holdout data sets, ID3, probit, and NN analyses were performed.[11] Tables 6 and 7 show the average predictive accuracy under various settings. Tables 6(a) and 7(a) show the statistics involving a single factor. Tables 6(b) and 7(b) show the statistics of different methods and sample sizes. Tables 6(c) and 7(c) show the statistics involving the interaction of all factors. The average predictive accuracy ranges from .5923 to .9302.

The results indicate that the NN method has the highest average accuracy (.7919), slightly better than probit (.7853) and much better than ID3 (.7425). Concerning the performance of ID3 and probit, it turns out that, contrary to previous findings including Parker and Abramowicz (1989) and Garrison and Michaelsen (1989), the average predictive accuracy of probit is better than that of ID3. In fact, the average accuracy of ID3 is lower than the result of the naive model.[12] The likely explanation is that ID3 overfits the training data. This is supported by the perfect classification accuracy of ID3 on the training sample.

Further examination of the results in Table 6 indicates that NN is the only method that performs better than the naive model in all settings. Probit performs well when the industry dominance is low (LDOM) but has problems when the industry dominance is high (HDOM). Overall, ID3 performs poorly. The poor performance of the probit and ID3 methods in HDOM is not totally unexpected. Since the data in the HDOM sample are highly dominated by a single nominal variable, the variable alone accounts for more than 90 percent of the variance.

11 The NN software used in the experiment was NeuroShell by Ward Systems Group, Inc. We assigned an input neuron for each input and output variable. The number of hidden neurons was half of the input neurons. A constant training rate of 0.4 and a momentum of 0.6 were used in training the NN model. In a few cases, these values were changed to 0.1 and 0 when the training session failed to converge in a reasonable period.

12 The naive model simply assumes that all firms use LIFO (or FIFO) if there are more firms in the training dataset that use LIFO (or FIFO). The model is then applied to the corresponding holdout data to calculate predictive accuracy.

TABLE 6
Predictive accuracy for data with industry type

Factor	Level	N	Mean
(a) Single factor			
Size	L/S	60	.8005
	M/M	60	.7954
	S/L	60	.7705
Method	ID3	60	.7552
	Naive	60	.7672
	Probit	60	.7953
	NN	60	.8160
Data	LDOM	90	.6859
	HDOM	90	.8918

(b) Interaction between size, method, and data dominance

		LDOM		HDOM	
Factor	Level	N	Mean	N	Mean
Size	L/S	30	.6948	30	.9062
	M/M	30	.6869	30	.9041
	S/L	30	.6759	30	.8650
Method	ID3	30	.6193	30	.8910
	Naive	30	.6355	30	.8990
	Probit	30	.7244	30	.8663
	NN	30	.7140	30	.9180

(c) Interaction of three factors

		LDOM		HDOM	
Size	Method	N	Mean	N	Mean
L/S	ID3	10	.6230	10	.9000
	Naive	10	.6256	10	.8959
	Probit	10	.7666	10	.8980
	NN	10	.6949	10	.9206
M/M	ID3	10	.6261	10	.8940
	Naive	10	.6373	10	.9000
	Probit	10	.7000	10	.8918
	NN	10	.7345	10	.9302
S/L	ID3	10	.6088	10	.8827
	Naive	10	.6436	10	.9011
	Probit	10	.7064	10	.8092
	NN	10	.7125	10	.9033

It is difficult for a method to make significant improvements, especially when the prior probabilities of LIFO and FIFO firms in the sample are not taken into consideration. What is interesting, however, is the superior performance of the NN method when the industry type is present. This indicates that its classification capability in domains including both nominal and non-nominal

TABLE 7
Predictive accuracy for data without industry type

Factor	Level	N	Mean
(a) Single factor			
Size	L/S	60	.7691
	M/M	60	.7622
	S/L	60	.7416
Method	ID3	60	.7299
	Naive	60	.7672
	Probit	60	.7752
	NN	60	.7679
Data	LDOM	90	.6859
	HDOM	90	.8669

(b) Interaction between size, method, and data dominance

		LDOM		HDOM	
Factor	Level	N	Mean	N	Mean
Size	L/S	30	.6667	30	.8715
	M/M	30	.6505	30	.8740
	S/L	30	.6282	30	.8551
Method	ID3	30	.6005	30	.8594
	Naive	30	.6355	30	.8990
	Probit	30	.6776	30	.8729
	NN	30	.6673	30	.8684

(c) Interaction of three factors

		LDOM		HDOM	
Size	Method	N	Mean	N	Mean
L/S	ID3	10	.5923	10	.8572
	Naive	10	.6256	10	.8959
	Probit	10	.6949	10	.8837
	NN	10	.7128	10	.8737
M/M	ID3	10	.6103	10	.8740
	Naive	10	.6373	10	.9000
	Probit	10	.6931	10	.8967
	NN	10	.6482	10	.8712
S/L	ID3	10	.5987	10	.8470
	Naive	10	.6436	10	.9011
	Probit	10	.6449	10	.8582
	NN	10	.6410	10	.8602

variables is stronger than ID3 and probit. When the industry type is removed from the data sets, all methods perform more poorly than the naive model in HDOM, but NN and probit perform better than the naive model in LDOM (Table 7(b)). Probit analysis has the best performance among the three.

To examine how individual methods respond to the change of data charac-

teristics. the difference between LDOM and HDOM and between the industry type and without industry type are compared. In Tables 6(b) and 7(b), we can find that the predictive accuracy of all models improves when the industry dominance changes from LDOM to HDOM. When the industry type is dropped from the data sets, the average predictive accuracy decreases (Tables 6(a) and 7(a)). The average reduction in predictive accuracy is approximately the same for ID3 and probit (.025 and .020, respectively), but much higher for NN (.048). ID3 models have large accuracy decreases in HDOM, probit models have large accuracy decreases in LDOM, and NN models have approximately the same amount of decrease in both cases. In fact, the average accuracy of probit models increases in the S/L setting of HDOM (from .8092 to .8582). More interestingly, the large accuracy changes for probit models occur in the L/S or S/L settings (i.e., the sizes of the training and holdout data are different).

One explanation for these observations is that different methods have different sensitivities to the loss of information provided by a key nominal variable and the gain from the lower degree of multicollinearity due to the removal of industry type. Removing the industry type results in a significant loss of information that will lower the predictive accuracy for all methods. However, because the removal of industry type also removes the multicollinearity existing between industry type and other variables, it may relax a constraint for methods with a high sensitivity to data multicollinearity (such as probit). Since probit can take advantage of the reduced data multicollinearity to compensate for some of the information loss due to the removal of industry type, its decrease in predictive accuracy is less than ID3 and NN models in HDOM. In LDOM, however, the loss of industry type information outweighs the reduced multicollinearity, which results in similar accuracy decreases for probit, ID3, and NN models.

To test the effect of the factors discussed previously, a four-way ANOVA was performed. The results, as shown in Table 8, indicate that four main effects and five interaction effects (B × D, C × D, A × B × D, A × C × D, and A × B × C × D) are significant at least at .05 level. Since the significance of main effects cannot be explained when significant interaction effects exist, we further test the significance of the main effects (i.e., whether the difference between the full model and the interaction model "$y = \mu + \alpha_{B\times D} + \alpha_{C\times D} + \alpha_{A\times B\times D} + \alpha_{A\times C\times D} + \alpha_{A\times B\times C\times D} + \epsilon$" is significant). The results indicate that the difference between the full model and the interaction model is indeed significant,[13] which means that the significance of the main effects and the interaction effects shown in Table 8 is statistically meaningful.

The main effects indicate that the predictive accuracy is affected by the existence of industry type, the dominance of nominal variables, sample size, and the modeling methods. Since the NN models have the highest mean accuracy, we

13 The F-value was calculated by $[(SSE_1 - SSE_2)/(df_1 - df_2)]/(SSE_1/df_1)$, where SSE_1 stands for the error sum of squares of the full model, SSE_2 stands for the error sum of squares of the interaction model, and df_1 and df_2 stand for the degree of freedom for the full and interaction models, respectively.

TABLE 8
Results of four-way ANOVA

Source	DF	SS	MS	F	P	R^2
A	1	0.0874	0.0874	37.46	.0001	.860
B	1	4.0508	4.0508	1735.39	.0001	
C	2	0.0553	0.0276	11.84	.0001	
D	2	0.1724	0.0862	36.93	.0001	
A × B	1	0.0035	0.0035	1.50	.2216	
A × C	2	0.0003	0.0001	0.06	.9387	
A × D	2	0.0138	0.0067	2.87	.0584	
B × C	2	0.0030	0.0015	0.64	.5256	
B × D	2	0.1444	0.0722	30.94	.0001	
C × D	4	0.0245	0.0061	2.62	.0350	
A × B × C	2	0.0078	0.0039	1.67	.1900	
A × B × D	2	0.0191	0.0095	4.09	.0176	
A × C × D	4	0.0281	0.0070	3.01	.0185	
B × C × D	4	0.0044	0.0011	0.47	.7593	
A × B × C × D	4	0.0250	0.0062	2.67	.0320	
Error	324	0.7563	0.0023			
Total	259	5.3956				

A = Industry type
B = Dominance of nominal variable
C = Size of data set
D = Method

can conclude that, on average, NN models are slightly more accurate than probit, and both NN and probit models are more accurate than ID3 models in classifying LIFO/FIFO data. This is consistent with Elliott and Kennedy (1988) and Marais et al. (1984) that indicate probit is more accurate than RPA, a decision-tree method similar to or probably better than ID3, but contradicts Garrison and Michaelsen (1989) and Parker and Abramowicz (1989). The statistical significance of industry type indicates that the existence of a nominal variable affects the classification accuracy of a method. By comparing the data in Tables 6(a) and 7(a), we find that it is true for all three methods. The statistical significance of data dominance indicates that the average predictive accuracy increases with the increased degree of dominance, whereas the statistical significance of sample sizes indicates that the average predictive accuracy decreases when the training sample size decreases.

The significant interaction effects (all involve method effects) enable us to further examine the factors affecting the relative accuracy of different methods. The significant interaction between industry dominance and method shows that the predictive accuracy is different for different methods in HDOM and LDOM. Averaging the predictive accuracies in Tables 6(b) and 7(b) indicates that NN is the best when the data set is dominated by a nominal variable (0.8932) and probit is the best otherwise (0.7010). ID3 is better than probit only in the setting where a nominal variable exists, the degree of dominance is high, and the training sample is relatively small.

The significant interaction effect between training sample size and method indicates that the reduction of training sample size has different impact on probit, ID3, and NN. The data in Tables 6(c) and 7(c) show that, compared to ID3 and NN, probit is more sensitive to the reduction of the training sample size. For example, the average accuracy of probit reduces from .7666 (L/S) to .7000 (M/M for LDOM and from .8919 (M/M) to .8092 (S/L) for HDOM in Table 6(c). A similar reduction can be found in Table 7(c) between M/M and S/L in Table 7(c) for both LDOM and HDOM. The average accuracy of ID3 and NN models, however, remains relatively stable in different sample sizes except in the case of LDOM and without industry type, where the accuracy of NN models drops form .7128 (L/S) to .6482 (M/M) and .6410 (S/L).

In summary, we have observed the following results. First, the average predictive accuracy of NN and probit is better than that of ID3 in analyzing our LIFO/FIFO data. The NN model has slightly higher average accuracy than probit. Second, the relative predictive accuracy of the models developed by a method is affected by the nature of the data (including whether it includes industry type and the level of industry dominance), and the relative size of the training and holdout data. Finally, NN performs well when the data set includes dominant nominal variables, whereas probit performs well otherwise. ID3 performs well only in a few special situations.

A sensitivity analysis has also been conducted on the results. Instead of using the 80 percent rule for composing the HDOM and LDOM data sets, we used a 75 percent rule to recompile the data sets (i.e., move industries 27, 28, 39, and 50 in Table 4 from LDOM to HDOM) and repeated the previous experiment. The procedures were the same except that we skipped the M/M case and repeated 5 times (instead of 10 times) in each setting. The results shown in Table 9 support the above arguments that (1) NN models have the highest average predictive accuracy, slightly higher than probit (.7566 for NN and .7503 for probit on average), (2) NN performs well when the data sets includes a dominant nominal variable (.7732 for NN, .7588 for probit), while probit performs well otherwise (.7399 for NN, .7418 for probit), (3) none of the methods performs well in HDOM, and (4) ID3 is less sensitive to the reduction in sample sizes.

Two observations in Tables 6 and 7 do not hold in Table 9. First, the drop in accuracy as training sample size decreases is greater for NN than for probit in Table 9. Second, probit has higher accuracy in the HDOM case (0.8519 for probit, 0.8299 for NN), and NN has higher accuracy for the LDOM case (.6488 for probit, .6833 for NN). These results contradict the observations in Tables 6 and 7. Furthermore, the substantial accuracy decrease of NN for S/L-HDOM in Table 9(b) is unexpected. One reason that may explain the result is that the NN structure we chose may not fit the data. In the experiment, we used the same network structure for all data sets to ensure comparability. In the real world, however, people often tailor NN structures to application data by trial and error. Unfortunately, there are no generally applicable guidelines for selecting the optimal structure yet.

TABLE 9
Results of the sensitivity analysis

		LDOM		HDOM	
Size	Method	*N*	Mean	*N*	Mean
(a) With industry					
L/S	ID3	5	.5619	5	.8030
	Naive	5	.6095	5	.8418
	Probit	5	.6952	5	.8537
	NN	5	.7333	5	.8537
S/L	ID3	5	.5810	5	.8090
	Naive	5	.5762	5	.8710
	Probit	5	.6476	5	.8388
	NN	5	.6523	5	.8537

		LDOM		HDOM	
Size	Method	*N*	Mean	*N*	Mean
(b) Without industry					
L/S	ID3	5	.5619	5	.7672
	Naive	5	.6095	5	.8418
	Probit	5	.6381	5	.8627
	NN	5	.6857	5	.8239
S/L	ID3	5	.5810	5	.7985
	Naive	5	.5762	5	.8710
	Probit	5	.6143	5	.8522
	NN	5	.6619	5	.7881

Implications and limitations

The findings in the experiments allow us to answer, at least partially, the questions raised at the end of the second section about comparing the three methods. First, are the variables selected by the methods different? Based on the results, we find that ID3 and probit select different variables in their models (NN was not compared because the current software does not test variable importance). This is primarily because these two methods use different criteria for variable selection (one uses maximum likelihood estimation and the other uses entropy). The implication of this observation is that it is necessary to pay attention to method selection for data analysis because different variables may be selected into the resulting models. Currently, probit is the only method that allows hypothesis testing.

Second, do the classification models have different accuracy and which method is better? The results suggest that NN and probit models have significantly higher predictive accuracy than the ID3 models. As indicated before, the conclusion on probit and ID3 is consistent with Elliott and Kennedy (1988) and Marais et al. (1984) but contradicts Garrison and Michaelsen (1989) and Parker and Abramowicz (1989). These seemingly conflicting findings, however,

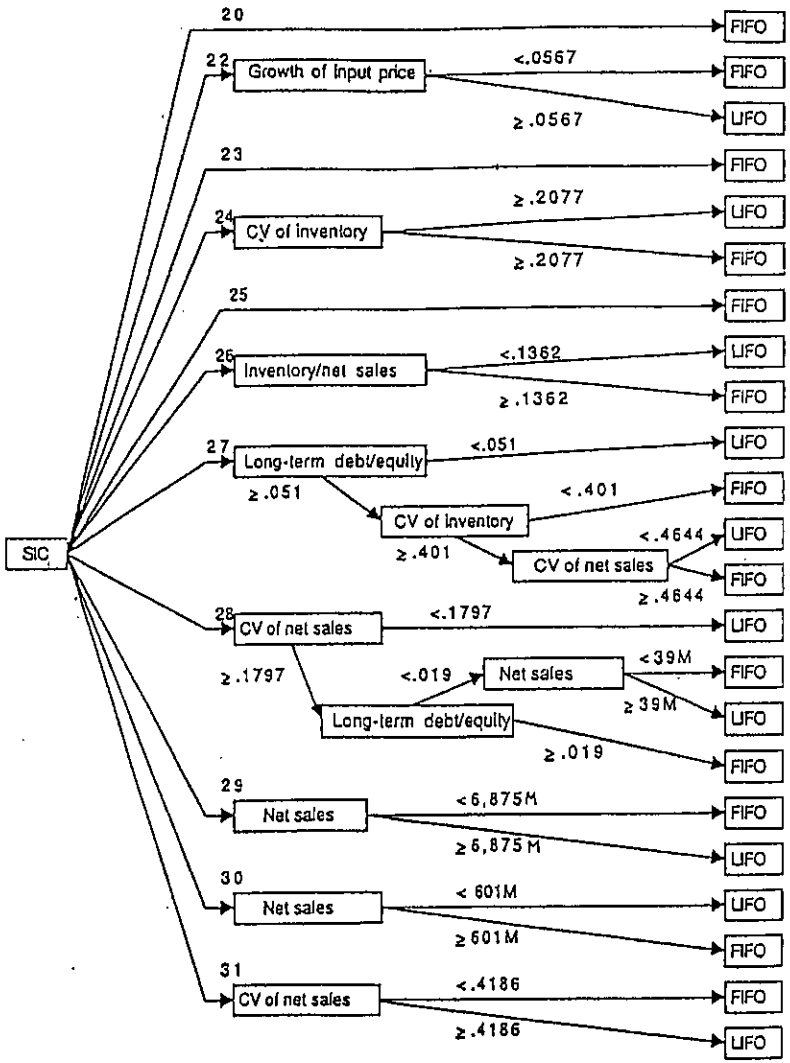
may be explained by comparing the nature of the data used in these different experiments. In Garrison and Michaelsen (1989) and Parker and Abramowicz (1989), ID3 performed better than probit because all variables used for analysis were nominal. In Elliott and Kennedy (1988), logit performed better than RPA on a set of non-nominal data. In Marais et al. (1984), probit was found to be better than RPA on numerical data, but RPA was found to be better on data including a combination of 6 nominal and 20 non-nominal variables. Therefore, the previous findings may be interpreted as indicating that probit is better if the data are dominated by non-nominal variables, and ID3 is better if the data are dominated by nominal variables. Our results also show that NN is a promising new technique for accounting classification. It is particularly useful when the data include nominal variables.

Finally, when and why does a particular method outperform the other? The results of the experiment provide primitive guidelines for selecting methods. If the data include a mixture of nominal and non-nominal variables, then NN should be used. Otherwise, probit is a good choice. One problem with NN models is that they are difficult to explain. The black box nature may reduce the user's confidence in the model. Another issue worth investigating is the good performance of the naive method in the HDOM case, especially when the industry type is excluded from the data. This result reveals two issues that need further studies. First, the three modeling techniques examined in the paper have limited capabilities in handling data strongly dominated by a nominal variable. Second, there may be factors unique to different industries that must be examined in future LIFO/FIFO studies.

In summary, we compared probit, ID3, and NN methods on a set of LIFO/FIFO data in this paper. It is, of course, impossible to answer the above questions conclusively by simply examining one domain. The inventory accounting choice data may have some characteristics (e.g., industry type) not existing in other domains such as audit judgments, loan evaluation, or bankruptcy prediction. Therefore, cautions are necessary when generalizing the results. Nonetheless, these findings do provide useful insights into the issues and clear evidence for future research. Possible directions for future work include at least the following:

1 *Other data characteristics.* In this work, we examined only the existence of one nominal variable and the degree of dominance of the nominal variable. The cases of multiple nominal variables and other criteria for classifying data characteristics should be investigated. Theoretical analysis and simulation studies that allow a better control of data characteristics should also provide useful insights.
2 *Other accounting problems.* Concerning real world data, further work may be done in other accounting classification problems to examine the generalizability of the findings. Bankruptcy prediction from financial reports, for example, may also include both nominal and numerical variables and have similar effects.
3 *Improved decision tree methods.* ID3 performs poorly in our experiment. This, however, does not mean that the method is useless. ID3 can be improved in several ways. First, tree-pruning algorithms may be introduced to alleviate the problem of overfitting the data (Quinlan, 1987). Although pruning often improves the predictive accuracy, there is no guarantee. It involves trade-offs between the risks of overfitting the data and overpruning the tree (i.e., overgrowing the tree versus cutting useful branches from the tree). Therefore, what are proper criteria for tree pruning and how to optimize the ID3 tree pruning process are interesting issues for further research.
4 *Integration of methods.* As stated previously, ID3 is only a representative of decision tree methods. There are other induction algorithms, such as Michalski's AQ approach (Michalski and Chilausky, 1980), that may also be useful for accounting research and need to be examined. In addition, given the different strengths of different methods, it is interesting to see how these methods can be integrated to create better models. In a related work, Liang et al. (1990) found that the integration of ID3 and discriminant analysis significantly outperformed individual methods alone. ID3 can be used as an exploratory technique to screen variables for statistical classification techniques. Liang (1992) also reported significant performance improvement by treating nominal and non-nominal variables separately in tree induction. It is also possible to integrate ID3 and NN methods. Research concerning how and when to integrate different methods, whether the integrated method can perform better, and if it does, why the integrated method is better, are also interesting for the future.

Appendix: A sample decision tree constructed by ID3



Note: This is only a subset of the tree for data set 4. Full trees are available from the authors.

References

Abdel-khalik, A.R., "The Effect of LIFO-Switching and Firm Ownership on Executives' Pay," *Journal of Accounting Research* (Autumn 1985) pp. 427–447.

Aldrich, J.H. and F.D. Nelson, *Linear Probability, Logit and Probit Models* (Beverly Hills, CA: Sage Publications, 1984).

Biddle, G.C., "Accounting Methods and Management Decisions: The Case of Inventory Costing and Inventory Policy," Supplement to the *Journal of Accounting Research* (1980) pp. 235–280.

Braun, H. and J.S. Chandler, "Predicting Stock Market Behavior through Rule Induction: An Application of the Learning-from-Examples Approach," *Decision Sciences* (Summer 1987) pp. 415–429.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees* (Monterey, CA: Wadsworth, Inc., CA, 1984).

Cushing, B. and M. LeClere, "Evidence on the Determinants of Inventory Accounting Policy Choice," Working Paper, Pennsylvania State University (1988).

Daley, L.A. and R.L. Vigeland, "The Effects of Debt Covenants and Political Costs on the Choice of Accounting Methods: The Case of Accounting for R&D Costs," *Journal of Accounting and Economics* (1983) pp. 195–211.

Dhaliwal, D., G. Salamon, and E.D. Smith, "The Effect of Owner versus Management Control on the Choice of Accounting Method," *Journal of Accounting and Economics* (July 1982) pp. 41–53.

Dietrich, R. and R. Kaplan, "Empirical Analysis of the Commercial Loan Classification Decision," *The Accounting Review* (1982) pp. 18–38.

Dopuch, N. and M. Pincus, "Evidence on the Choice of Inventory Accounting Methods: LIFO versus FIFO," *Journal of Accounting Research* (Spring 1988) pp. 28–59.

Eggleton, I.R.C., S.H. Penman, and J.R. Twombly, "Accounting Changes and Stock Prices: An Examination of Selected Uncontrolled Variables," *Journal of Accounting Research* (1976) pp. 66–88.

Elliott, J.A. and D.B. Kennedy, "Estimation and Prediction of Categorical Models in Accounting Research," *Journal of Accounting Literature* (1988) pp. 202–242.

Fisher, D.H. and K.B. McKusick, "An Empirical Comparison of ID3 and Backpropagration," *Proceedings of the International Joint Conference on Artificial Intelligence* (1989) pp. 788–793.

Garrison, L.R. and R.H. Michaelsen, "Symbolic Concept Acquisition: A New Approach to Determining Underlying Tax Law Constructs," *The Journal of American Tax Association* (Fall 1989) pp. 77–91.

Hagerman, R.L. and M.E. Zmijewski, "Some Economic Determinants of Accounting Policy Choice," *Journal of Accounting and Economics* (August 1979) pp. 141–161.

Han, I., "The Impact of Measurement Scale on Classification Performance of Inductive Learning and Statistical Approaches" (Ph.D. Dissertation, Department of Accountancy, University of Illinois at Urbana-Champaign, 1990).

Hunt, E.G., J. Martin, and P. Stone, *Experiments in Induction* (New York: Academic Press, 1966).

Khanna, T., *Foundations of Neural Networks* (Reading, MA: Addison-Wesley, 1990).

Lee, C. and D. Hsieh, "Choice of Inventory Accounting Methods: Comparative Analysis of Alternative Hypotheses," *Journal of Accounting Research* (Autumn 1985) pp. 468–485.

Liang, T.P., "A Composite Approach to Automated Knowledge Induction for Expert Systems Design," *Management Science* (January 1992), pp. 1–17.

———, J.S. Chandler, and I Han, "Integrating Statistical and Inductive Learning

Methods for Knowledge Acquisition," *Expert Systems with Applications* (1990) pp. 391–401.

Lindahl, F.W., C. Emby, and R.H. Ashton, "Empirical Research on LIFO: A Review and Analysis," *Journal of Accounting Literature* (1988) pp. 310–333.

Marais, M.L., J.M. Patell, and M.A. Wolfson, "The Experimental Design of Classification Models: An Application of Recursive Partitioning and Bootstrapping to Commercial Bank Loan Classification," Supplement to the *Journal of Accounting Research* (1984) pp. 87–114.

Messier, W.F., Jr., and J.V. Hansen, "Inducing Rules for Expert Systems Development: An Example Using Default and Bankruptcy Data," *Management Science* (December 1988) pp. 1403–1415.

Michalski, R.S. and R.L. Chilausky, "Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing Expert Systems for Soybean Disease Diagnosis," *International Journal of Policy Analysis and Information Systems* (1980) pp. 125–161.

Mingers, J., "Expert Systems—Rule Induction with Statistical Data," *Journal of the Operational Research Society* (1987) pp. 39–47.

Morse, D. and G. Richardson, "The LIFO/FIFO Decision," *Journal of Accounting Research* (Spring 1983) pp. 106–127.

Noreen, E., "An Empirical Comparison of Probit and OLS Regression Hypothesis Tests," *Journal of Accounting Research* (Spring 1988) pp. 119–133.

Parker, J.E. and K.F. Abramowicz, "Predictive Abilities of Three Modeling Procedures," *The Journal of American Taxation Association* (Fall 1989) pp. 37–53.

Quinlan, J.R., "Simplifying Decision Trees," *International Journal of Man-Machine Studies* (1987) pp. 221–234.

——— "Semi-autonomous Acquisition of Pattern-based Knowledge," in *Introductory Readings in Expert Systems*, ed. D. Michie (London: George & Breack, 1982).

——— "Discovering Rules from Large Collections of Examples: A Case Study," in *Expert Systems in the Micro Electronic Age*, ed. D. Michie (Edinburgh, Scotland: Edinburgh University Press, 1979).

Ricks, W., "The Market's Response to the 1974 LIFO Adoptions," *Journal of Accounting Research* (Autumn 1982) pp. 367–387.

White, H., "Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models," *Journal of the American Statistical Association* (1989) pp. 1003–1013.

Aggregate efficiency measures and Simpson's Paradox*

ABRAHAM MEHREZ *Kent State University*

J. RANDALL BROWN *Kent State University*

MOUTAZ KHOUJA *University of North Carolina at Charlotte*

Abstract. Much work has been directed to develop aggregate efficiency measures for firms or decision-making units (DMUs) in which we are able to observe only the outputs and inputs. Assuming that each DMU has the same type of observed outputs and inputs and using only this information, Farrell's technical efficiency and the CCR ratio can be used to assign an aggregate measure of efficiency to each DMU, which can then be used to compare the efficiency of the DMUs. This paper considers a subset of the general aggregate efficiency problem called the *matched output/input case* in which each output is matched to exactly one input, forming a subunit. Dividing the output by the input for each subunit within a DMU yields a subunit ratio that is the output per unit input. For a particular subunit, the subunit ratios for two DMUs can be compared directly. If all the subunit ratios of one DMU exceed the corresponding subunit ratios in another DMU, then we should reasonably expect that any aggregate efficiency measure has the efficiency of the first DMU greater than the efficiency of the other DMU. This requirement is defined as the Matched Output/Input Axiom, which is then shown to be violated for certain data sets satisfying Simpson's Paradox. Both Farrell's technical efficiency and the CCR ratio are then shown to violate the Matched Output/Input Axiom, which raises questions about the overall validity of both procedures.

Résumé. Les travaux visant l'élaboration de mesures globales du rendement des unités décisionnelles ou des entreprises, dans lesquelles il n'est possible d'observer que les extrants et les intrants, sont nombreux. En supposant que le même type d'extrants et d'intrants est observé pour chaque unité décisionnelle et que cette information est la seule qui soit utilisée, le rendement technique de Farrell et le ratio CCR (Charnes, Cooper et Rhodes) peuvent être utilisés pour attribuer une mesure globale de rendement à chaque unité décisionnelle, mesure qui peut ensuite permettre de comparer le rendement des différentes unités. Les auteurs étudient un sous-ensemble du problème général de rendement global, le cas de concordance extrant-intrant, dans lequel chaque extrant est associé à exactement un intrant, pour former un sous-ensemble. En divisant l'extrant par l'intrant pour chaque sous-ensemble d'une unité décisionnelle, on obtient, pour chacun d'eux, un ratio représentant l'extrant par unité d'intrant. Pour un sous-ensemble particulier, les ratios de deux unités décisionnelles peuvent faire l'objet d'une comparaison directe. Si la totalité des ratios des sous-ensembles d'une unité décisionnelle excède la totalité des ratios des sous-ensembles correspondants d'une autre unité décisionnelle, on est en droit de s'attendre à ce que l'application d'une mesure globale du rendement, quelle

* The authors would like to thank the anonymous reviewer of this paper's original draft for the constructive criticisms and comments which led to a revised paper.