Pergamon

Expert Systems With Applications, Vol. 10, No. 2, pp. 209–221, 1996

0957-4174/96 $15.00 + 0.00

0957-4174(95)00047-X

The Impact of Measurement Scale and Correlation Structure on Classification Performance of Inductive Learning and Statistical Methods

INGOO HAN*

Department of Management Information Systems, Korea Advanced Institute of Science and Technology, Seoul 130-650, Korea

JOHN S. CHANDLER

Department of Accountancy, University of Illinois at Urbana-Champaign, Champaign, IL 61820, U.S.A.

TING-PENG LIANG

Department of Information Management, National Sun Yat-Sen University, Kaoshiung 47907, Taiwan

Abstract—*This is a comparative study of inductive learning and statistical methods using the simulation approach to provide a generalizable results. The purpose of this study is to investigate the impact of measurement scale of explanatory variables on the relative performance of the statistical method (probit) and the inductive learning method (ID3) and to examine the impact of correlation structure on the classification behavior of the probit method and the ID3 method. The simulation results show that the relative classification accuracy of ID3 to probit increases as the proportion of binary variables increases in the classification model, and that the relative accuracy of ID3 to probit is higher when the covariance matrices are unequal among populations than when the covariance matrices are equal among populations. The empirical tests on ID3 reveal that the classification accuracy of ID3 is lower when the covariance matrices are unequal among populations than when the covariance matrices are equal among populations and that the classification accuracy of ID3 decreases as the correlations among explanatory variables increases.*

1. INTRODUCTION

Classification refers to separating distinct sets of objects or observations and allocating new objects or observations into previously defined groups. Classification needs an algorithm to separate and allocate objects or observations. This algorithm is called a classification technique. The ultimate goal of a classification method is to provide the relevant outcome or to replicate the expert's judgment. The relative performance of different classification techniques may depend on data conditions.

Classification studies in business have traditionally used statistical techniques. Recently, inductive learning, a subfield of artificial intelligence (AI), began to be applied to the classification research. Examples include stock market prediction (Braun & Chandler, 1987), scholarship/fellowship grant case (Garrison & Michaelsen, 1989), accounting inventory method choice (Liang et al., 1992), bankruptcy prediction (Chung & Tam, 1993), asset writedown (Ragothaman & Naik, 1994), executive compensation planning (Michaelsen and Swigger, 1994), and text-to-speech mapping (Dietterich, Hild & Bakiri, 1995).

Inductive learning uses a data set of examples and determines a relationship between these examples via inductive inference. The induced rules can then be used to predict outcomes or to replicate judgments. The inductive learning approach is different from the statistical approach in many aspects though both approaches are inductive in nature. The key difference is that traditional statistical methods use continuously varying parameters to express classification criteria through

* To whom all correspondence should be addressed.

numerical manipulation while inductive learning methods describe concepts through the manipulation of symbolic representations (Rendell, 1986; Greene, 1987).

Most studies applying inductive learning to classification chose an application domain and selected attributes appropriate for the domain of their choice. The selection of attributes was typically done through interviews with domain experts. These studies applied the inductive learning method to the classification problem and compared the predictive accuracy of the inductive learning method with the results from statistical methods or judgments of domain experts. Prior studies provided evidence regarding the performance of the inductive learning method in certain domains. However, they did not provide a general guideline to researchers as to the conditions under which inductive learning methods are preferable to traditional techniques. Messier & Hansen (1988) recognized the limited generalizability of results from previous studies and suggested that future studies should examine the performance of inductive learning algorithms in problem domains of different types and complexity.

Mingers (1987) compared inductive learning and linear regression with respect to stochastic (noisy) data which can be characterized by random variations in additions to variation due to systematic effect of other factors. Arinze & Subbanarasimha (1994) chose the data set which were noisy and non-normally distributed (skewed) to test the relative efficiency of rule based induction to regression model.

This study investigates the relative performance of inductive learning and statistical methods in order to identify the data conditions under which one approach has a relative strength over the other. Statistical methods frequently employed in classification research include multiple discriminant analysis (MDA), logit, and probit methods. The probit method is chosen in this study because it is a frequently used technique in recent classification studies.[1] The ID3 method is chosen among various inductive learning algorithms because it has been the most widely used algorithm in inductive learning applications.

The ID3 method (generally inductive learning methods) assumes nominal attributes whereas the probit method (generally statistical models) assumes numeric attributes. The ID3 algorithm was developed to deal with nominal attributes. On the other hand, the probit method was developed for numeric variables. These basic assumptions of probit and ID3 lead to the hypothesis that the ID3 method may perform relatively better when nominal variables are used and that the probit method may perform relatively better when numeric variables are used.

The impact of correlation structure on the classification performance has been a major concern of classification researchers. Statistical studies show that linear statistical models such as LDA (linear discriminant analysis), logit, or probit performed worse for unequal covariance matrices than for equal covariance matrices (Moore, 1973; Dillon & Goldstein, 1978) and that linear models could not satisfactorily cope with situations where the correlations were large (Schmitz et al., 1983; Schmitz, Habbema, & Hermans, 1985). There are few empirical evidence regarding the impact of correlation structure on the classification accuracy of the ID3 method.[2] The ID3 method, a nonparametric algorithm, is expected to be less sensitive to the inequality of covariance structure and the magnitude of correlations than the probit method.

This paper intends to study the relative performance of the ID3 and probit, which are representative of inductive learning and statistical methods respectively, with respect to the measurement scale of attribute and the correlation structure. Simulation allows researchers to manipulate data conditions and enhance internal validity, and this makes the results more generalizable. This paper uses the simulation approach to provide the generalizable results. The rest of this paper is organized as follows. Section 2 presents a theoretical comparison of probit and ID3 together with empirical evidences of previous studies. In Section 3, the simulation design is developed and the ANOVA models are presented for the empirical test. Section 4 discusses the results. The last section provides a summary of major findings and the implications of this study, and suggests limitations and some future research issues.

2. COMPARISON OF PROBIT AND ID3

Qualitative response models are regression models in which a dependent variable takes discrete values. The probit model, which is one of qualitative response models, has been frequently used in recent business classification studies. The binary decision by the ith individual is represented by a random variable y_i that takes the value of 0 or 1. Let U_{i1} and U_{i0} denotes the utilities of the two choices. U is a linear function of explanatory variables. A univariate binary probit model is defined by

$$\Pr(y_i = 1) = \Pr(U_{i1} > U_{i0}) = F(X_i'b), \quad i = 1, 2, \ldots, n.$$

[1] Qualitative response models such as probit and logit provide slightly better classification accuracy than MDA in recent studies (Vlachonikolis & Marriott, 1982; Schmitz et al., 1983). Judge et al. (1985, p. 761) provided a theoretical argument for the use of probit in preference to logit while Amemiya (1981) noted that the use of probit or logit makes little difference except when data are heavily concentrated in the tails.

[2] Tu (1989) tested the impact of the magnitude of correlations on the path length of the decision tree generated from inductive learning algorithms though this is not a direct test on the classification accuracy.

where $\{y_i\}$ is a sequence of independent binary random variables taking the value 1 or 0, X_i is a vector of explanatory variables, b is a vector of unknown parameters, and F is a cumulative distribution function of the normal distribution (Judge et al., 1985)

Multinomial models are generalizations of the binary model. The utility that the ith individual derives from the choice of the jth alternative can be represented as

$$U_{ij}=X'_{ij}b+e_{ij}$$

where X_{ij} is a vector of variables representing the attributes of the jth choice to the ith individual, b is a vector of unknown parameters, and e_{ij} is a error vector. A representative multinomial probit model is the MZ polychotomous probit model developed by McKelvey & Zavoina (1971, 1975). MZ probit model assumes that X_{ij} is dichotomous or measured on at least an interval scale and that e_{ij} has a multivariate normal distribution.

Probit models (generally regression-based models) and inductive learning methods are similar in that they both induce a relationship between independent and dependent variables from a number of observations. However, the probit method is developed in the domain of statistics while inductive learning is a subfield of artificial intelligence.[3] Naturally, different terms tend to be used in the two different domains. First of all, the difference of terminology needs to be clarified.

Statistics	Inductive learning
case	example
independent variable	attribute
dependent variable	class
derivation (estimation) sample	training sample
validation (holdout) sample	testing sample

There are significant differences between probit and ID3 methods though their goals are similar. The Mz polychotomous probit model assumes that categories of a dependent variable are ordered. The issue of ordering does not arise in binary probit models. Probit models assume that independent variables are continuous with at least an interval scale of measurement though 0–1 dummy variables may be used to represent nominal variables. The ID3 method assumes that dependent variables are measured on a nominal scale. ID3 can also deal with ordinal dependent variables by regarding ordered categories as nominal categories. The ID3 method also assumes that independent variables are nominal though a variable of interval or ratio scale can be transformed into a nominal variable with two categories and be processed in the algorithm.

Probit models assume normality of the error term while the ID3 method does not assume any specific distribution. *Ex ante*, nonparametric techniques may be appropriate for classification research in business considering that the distributional assumptions of classification techniques are likely to be violated to some degree in most business research problems (Elliott & Kennedy, 1988).

The relationship between independent and dependent variables is represented as a linear function in probit models (generally regression-based models) while the ID3 method (generally inductive learning algorithms) induces a decision rule which is logical rather than functional. The parameters of a linear function are estimated by the maximum likelihood estimation method for probit models while the formation of a decision tree is based on the entropy measure of information theory by Shannon (1964). These comparisons are summarized in Table 1.

The differences in distribution assumption, relationship of independent and dependent variables, and modeling basis between probit and ID3 have, to a large extent, originated from the different assumptions on the measurement scale for independent variables between the two methods. Normality (generally a parametric characteristic), functional form, and maximum likelihood estimation have been developed for numeric variables while nonparametric characteristics, decision rules, and the entropy measure are tailored to qualitative measurement.[4]

Generally speaking, statistical methods attempt to represent a classification model as a combination of attribute weights. All the attributes are numerically coded and the search for parameter weights is conducted through the mathematical manipulation of means, frequencies, and variances. The evaluation function that directs the search is based on a measure such as mean squared error, Bayes theorem, or maximum likelihood estimation. On the other hand, the inductive learning method sequentially builds systems of production rules by symbolic representation (Greene, 1987).

The fundamental assumptions on the measurement scale for explanatory variables of the ID3 and probit methods lead to hypothetical statements that the ID3 method performs relatively better when explanatory variables are nominal and that the probit method performs relatively better when explanatory variables are numeric.

The empirical results in the AI literature are generally

[3] Refer to Amemiya (1981) for the detailed description of probit models. Braun & Chandler (1987) provided a detailed presentation of the ID3 algorithm. Mingers (1987) discussed the differences between inductive learning and regression models.

[4] Hays (1973) noted that the entropy measure is more appropriate for qualitative measurement situation than statistical methods developed for numerical data.

TABLE 1
Comparisons of Probit and ID3 Methods

	Probit	ID3
Measurement scale for dependent variable[1]	Ordinal	Nominal[2]
Measurement scale for independent variable	Interval or ratio[4]	Nominal[3]
Distribution assumption	Normal	Nonparametric
Relationship between independent and dependent variables	Linear function	Decision rule
Basis of modeling	Maximum likelihood estimation	Entropy measure of information theory

[1] In binary cases, the measurement scale for the dependent variable does not matter.
[2] ID3 method can deal with an ordinal dependent variable assuming that the ordered categories are nominal.
[3] ID3 method can handle a variable with interval or ratio scale by transforming it into a binary nominal variable.
[4] Nominal variables can be represented by 0–1 dummy variables in probit models.

consistent with the above hypothetical statements.[5] When attributes are all nominal, the classification accuracies of inductive learning methods are very high (Quinlan, 1980, 1983; Shapiro & Niblett, 1982),[6] better than expert judgments (Michalski & Chilausky, 1980), or better than statistical results (Garrison & Michaelsen, 1989). On the other hand, the inductive learning algorithms perform worse than statistical methods when attributes are mostly numeric (Liang et al., 1992). These results are difficult to generalize because the studies selected different domains and the characteristics of the domains chosen are expected to affect the classification accuracies.[7] The statistics literature indicates that linear models perform worse when explanatory variables are all binary than when explanatory variables are all continuous or mixed (Bayne et al., 1983)

Some AI studies advocated transforming continuous variables into nominal variables for inductive learning algorithms. Paterson & Niblet (1982) noted that nominalization of continuous variables may improve the classification accuracy of the inductive learning method when noise is present to a substantial degree. Hoff, Michalski & Stepp (1986) suggested that continuous variables be categorized into a reasonably small number of discrete values before they are used by inductive learning methods.

3. RESEARCH DESIGN

The impact of measurement scale of independent variable, which is a focus of this study, is to be manipulated. In addition, the correlation structure, which has been a major issue in statistical classification studies, is to be manipulated. The simulation design is to discriminate between two populations. The multivariate normal distribution has been a standard procedure used to generate the multivariate continuous data in simulation. In addition, the technique to generate the multivariate discrete data from multivariate normal distribution has been well developed (Schmitz et al., 1983). Hence, the multivariate normal distribution is chosen to generate continuous and discrete data simultaneously. The data are generated from two eight-dimensional normal distributions using the RNMVN subroutine of IMSL STAT/LIBRARY:[8]

$$\text{Population } \pi_1 : N_8(\mu_1, \Sigma_1)$$

$$\text{Population } \pi_2 : N_8(\mu_2, \Sigma_2)$$

where

$$\mu_1' = (0, 0, 0, 0, 0, 0, 0, 0)$$

and

$$\mu_2' = (0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2).$$

[5] Studies using the small training samples (<30) are not considered here due to the lack of reliability.
[6] These studies did not provide comparisons with statistical results or expert judgments. Classification accuracies achieved by these studies were above 98%.
[7] The prediction accuracies of inductive learning achieved in scientific domains (Buchanan et al., 1976; Michalski & Chilausky, 1980) and in the chess endgame (Quinlan, 1980, 1983; Shapiro & Niblett, 1982) were higher than those achieved in business domains (Braun & Chandler, 1987; Garrison & Michaelsen, 1989; Liang et al., 1992) because the scientifc domains and the chess endgame are more static, deterministic, and almost enumerative while business domains are more dynamic, probabilistic, and noisy.
[8] IMSL STAT/LIBRARY is a collection of Fortran subroutines and functions for statistical analysis. RNMVN is a subroutine of IMSL to generate multivariate normal random numbers. Refer to IMSL STAT/LIBRARY users' manual (1987) for details.

The mean differences of explanatory variables between two populations are set to vary to give a structure to the data. Three values of correlation structure (Σ_1, Σ_2) are used in this study: (I, I), $(I, 4I)$ and (Σ_5, Σ_5). I is an identity matrix. All the variances of the eight explanatory variables are 1 and every correlation coefficient is 0 in I. $4I$ is equal to I except that all the variances of the eight variables are 4. The variances of eight variables are 1 and every correlation coefficient is 0.5 in Σ_5.[9] The correlation structure (I, I) represents the situation in which variables are not correlated at all. $(I, 4I)$ represents the inequality of variances among two population. (Σ_5, Σ_5) represents the data situation in which variables are highly correlated.

The selection of the values for mean and covariance structure was based on previous simulation studies. As is the case with most simulation studies, such selection may look arbitrary on surface. However, the issue of importance in the simulation design of this study is the magnitude of correlations and the equality of covariance structure among two populations with different mean vectors rather than the particular nature of the mean and covariance structure. Thus, the foregoing selection of values does not affect the validity of any results.

After generating the continuous variables, the variables need to be made discrete into two categories in order to provide a nominal classification model. Let the mean difference with respect to the variable i be denoted by α_i. The cut-off value to be used in this study is $1/2\ \alpha_i$ (Schmitz et al., 1983; Tu, 1989) Let X_i denote the continuous variable and Z_i denote the respective binary variable.

$$Z_i = 1 \text{ if } X_i > 1/2\alpha_i$$

$$0 \text{ otherwise.}$$

The model consisting of all continuous variables is called a numeric model. The 6–2 mixed model includes six continuous variables and two binary variables. The 4–4 mixed model includes four continuous variables and four binary variables. The 2–6 mixed model includes two continuous variables and six binary variables. The explanatory variables to be dichotomized into binary variables will be selected randomly. The model consisting of all binary variables is called a nominal model. Therefore, five classification models are used in this study. Fifty runs will be made with samples of size 100 in each cell.

The ID3 method is chosen to represent the inductive learning algorithms and the probit method is chosen to represent the statistical methods. When classification accuracy is estimated from the same sample used for model specification, the estimate of classification accuracy is biased upward because the classification model is tailored to the data. Techniques to avoid the overfitting problem include the holdout technique, jackknife procedure, and bootstrapping. The holdout technique that is frequently used in classification research is used in this study. Half of the sample will be used as a training set and the other half as a testing set. The performance measure used in this study is the classification accuracy which is defined as the number of holdout cases correctly predicted divided by the total number of holdout cases.

ANOVA (analysis of variance) models are versatile statistical tools for investigating the relation between a dependent variable and one or more independent variables. ANOVA is used, for the most part, to compare means of dependent variables in studies involving more than two groups. ANOVA has been the major methodology for comparative studies of classification techniques to investigate the main and interaction effects of blocking and treatment factors which are mostly qualitative. The ANOVA technique is used in this study to investigate the statistical significance of the main and interaction effects of factors on the classification accuracy.

In summary, there are three factors in this design:

(1) measurement scale of explanatory variables: numeric, 6–2 mixed, 4–4 mixed, 2–6 mixed, and nominal classification models,
(2) correlation structure: (I, I), $(I, 4I)$, and (Σ_5, Σ_5).
(3) classification methods: probit and ID3.

The measurement scale and correlation structure are used as blocking factors and the classification method is the treatment factor to be studied. Descriptive statistics of classification accuracy are provided for each of thirty situations (5 classification models $\times$ 3 correlation structures $\times$ 2 classification methods). First, a three-way ANOVA is performed to investigate the statistical significance of all the main effects and interaction effects. Given the purpose of the research, the interaction effect between the measurement scale and the treatment is of special importance. The interaction effects between the equality of covariance matrices and the treatment and between the magnitude of correlations and the treatment are also of interest.

For the detailed analysis, a two-way (measurement scale and correlation structure) ANOVA is applied to probit and ID3 in order to investigate the statistical significance of the impact of the measurement scale, the equality of covariance matrices, and the magnitude of correlations on the classification accuracies of probit and ID3.

[9] The linear procedures performed reasonably in populations tructures characterized by low correlations (i.e. $<|0.2|$); poor results were obtained with use of linear procedures in population structures in which the correlations were high (i.e. $>|0.3|$) (Moore, 1973; Dillon & Goldstein, 1978). 0.5 is chosen as the value of considerably high correlation coefficient.

4. RESULTS AND ANALYSIS

The following notations are introduced to denote the respective blocking and treatment factors of the experimental design;

(1) MEAS: the blocking factor representing the measurement scale of explanatory variables. This factor can take on five levels which are INT (numeric), M6–2 (6–2 mixed), M4–4 (4–4 mixed), M2–6 (2–6 mixed), and NOM (nominal).
(2) CORR: the blocking factor that represents the correlation structure. This factor has three levels which are *I–I* (I, I), *I–4I* $(I, 4I)$, and *H–H* (Σ_5, Σ_5).
(3) METH: the treatment factor denoting the classification method. This factor can take on two levels which are PROBIT (probit method) and ID3 (ID3 method).

The dependent variable, ACCU, represents the classification accuracy which was previously defined as the number of holdout cases correctly predicted divided by the total number of holdout cases.

The descriptive statistics for the respective blocking factor levels are presented in Table 2. The examination of Table 2 shows that the classification accuracy of the probit method decreases as the proportion of binary explanatory variables increases and that the classification accuracy of the ID3 method increases as the proportion of binary explanatory variables increases. The mean classification accuracy of the probit method decreases by 0.0280 when all numeric explanatory variables are replaced by nominal variables. On the other hand, the mean classification accuracy of the ID3 method increases by 0.0212 when all numeric explanatory variables are replaced by nominal variables. The measurement scale has a consistent effect on the classification accuracy of the probit method across three levels of correlation structure.

Table 2 reveals that both probit and ID3 perform the best under the correlation structure of (I, I), the second best under that of $(I, 4I)$, and the worst under that of (Σ_5, Σ_5). The classification accuracy of the two methods behaves in the same direction across the correlation structure though the impact of correlation structure on the classification accuracy of the probit method is different than that of the ID3 method.

The three-way ANOVA results for the effects of measurement scale, correlation structure, and classification method are given in Table 3. The main effects of the treatment factor METH and the blocking factor CORR are significant. This indicates that the classification accuracy is significantly different between the two classification methods and that the classification accuracy differs as the correlation structure changes. The main effect of the blocking factor MEAS is not significant.

The interaction effect between MEAS and METH is significant, which implies that the differences in classification accuracy of probit and ID3 vary as the measurement scale changes. That is, the average classification behavior of two classification methods varies for different levels of measurement scale. The results confirm the hypothesis that the relative classification accuracy of the ID3 method to the probit method increases as the proportion of binary variables increases in the classification model.[10]

The interaction effect between CORR and METH is significant (at the 5% level), which implies that the differences in classification accuracy of probit and ID3 vary depending on the correlation structure. The interaction effect between MEAS and CORR is not significant, which implies that the differences in classi-

[10] The relative classification accuracy of ID3 to probit is the classification accuracy of ID3 over that of probit.

TABLE 2
Means (Standard Deviations) of Classification Accuracy by Measurement Scale, Correlation Structure, and Method

	Probit				ID3			
	I–I	*I–4I*	*H–H*	Total	*I–I*	*I–4I*	*H–H*	Total
INT	0.9116	0.8464	0.8320	0.8633	0.8544	0.8164	0.7940	0.8216
	(0.0573)	(0.0531)	(0.0577)	(0.0659)	(0.0599)	(0.0747)	(0.0616)	(0.0703)
M6-2	0.9008	0.8380	0.8252	0.8547	0.8552	0.8204	0.7956	0.8237
	(0.0614)	(0.0537)	(0.0651)	(0.0687)	(0.0606)	(0.0676)	(0.0598)	(0.0673)
M4–4	0.8940	0.8188	0.8140	0.8423	0.8688	0.8200	0.7956	0.8281
	(0.0607)	(0.0592)	(0.0648)	(0.0717)	(0.0606)	(0.0541)	(0.0718)	(0.0696)
M2–6	0.8832	0.8188	0.8124	0.8381	0.8692	0.8212	0.7964	0.8289
	(0.0643)	(0.0580)	(0.0542)	(0.0671)	(0.0574)	(0.0679)	(0.0626)	(0.0697)
NOM	0.8784	0.8172	0.8104	0.8353	0.8864	0.8264	0.8156	0.8428
	(0.0564)	(0.0643)	(0.0649)	(0.0691)	(0.0576)	(0.0521)	(0.0519)	(0.0623)
TOTAL	0.8936	0.8278	0.8188	0.8467	0.8668	0.8209	0.7994	0.8290
	(0.0613)	(0.0591)	(0.0621)	(0.0693)	(0.0604)	(0.0640)	(0.0624)	(0.0683)

TABLE 3
Three-way ANOVA

Class	Levels	Values
MEAS	5	INT, M6–2, M4–4, M2–6, NOM
CORR	3	*I–I, I–4I, H–H*
METH	2	Probit, ID3

Dependent Variable: ACCU

Source	d.f.	SS	MS	*F*	Pr
Model	29	1.6858	0.0581	15.43	0.0001
Error	1470	5.5380	0.0038		
Corrected total	1499	7.2237			

Source	d.f.	ANOVA SS	*F*	Pr
MEAS	4	0.0151	1.00	0.4060
CORR	2	1.4005	185.87	0.0001
METH	1	0.1176	31.21	0.0001
MEAS*CORR	8	0.0080	0.27	0.9769
MEAS*METH	4	0.1103	7.32	0.0001
CORR*METH	2	0.0251	3.33	0.0360
MEAS*CORR*METH	8	0.0092	0.31	0.9640

fication accuracy under the three levels of correlation structure have the same size and sign across the different levels of measurement scale that are applied. In other words, the average classification behavior of each level of correlation structure is consistent across different levels of measurement scale.

The factor of correlation structure incorporates two components of interest; the magnitude of correlations and the equality of the covariance matrices across categories. The preceding analysis summarized in Table 3 examined the combined effect of the magnitude of correlations and the equality of covariance matrices. The effect of the equality of covariance matrices can be investigated by including only two levels of correlation structure, *I–I* and *I–4I*. The effect of the magnitude of correlations can be analyzed by including *I–I* and *H–H* for the correlation structure.

The ANOVA analysis is repeated with two levels *I–I* and *I–4I* for the correlation structure and the results are presented in Table 4. The main effect of CORR is

TABLE 4
Three-way ANOVA (Corr: *I–I* and *I–4I*)

Class	Levels	Values
MEAS	5	INT, M6–2, M4–4, M2–6, NOM
CORR	2	*I–I, I–4I, H–H*
METH	2	Probit, ID3

Dependent Variable: ACCU

Source	d.f.	SS	MS	*F*	Pr
Model	19	0.9840	0.0518	13.96	0.0001
Error	980	3.6356	0.0037		
Corrected total	999	4.6196			

Source	d.f.	ANOVA SS	*F*	Pr
MEAS	4	0.0094	0.63	0.6388
CORR	1	0.7795	210.13	0.0001
METH	1	0.0712	19.20	0.0001
MEAS*CORR	4	0.0064	0.43	0.7854
MEAS*METH	4	0.0863	5.82	0.0001
CORR*METH	1	0.0246	6.63	0.0102
MEAS*CORR*METH	4	0.0065	0.44	0.7793

TABLE 5
Three-way ANOVA (Corr: I–I and H–H)

Class	Levels	Values
MEAS	5	INT, M6–2, M4–4, M2–6, NOM
CORR	2	I–I, I–$4I$, H–H
METH	2	Probit, ID3

Dependent Variable: ACCU

Source	d.f.	SS	MS	F	Pr
Model	19	1.5036	0.0791	21.06	0.0001
Error	980	3.6823	0.0038		
Corrected total	999	5.1859			

Source	d.f.	ANOVA SS	F	Pr
MEAS	4	0.0084	0.56	0.6921
CORR	1	1.2631	336.15	0.0001
METH	1	0.1332	35.44	0.0001
MEAS*CORR	4	0.0024	0.16	0.9600
MEAS*METH	4	0.0880	5.85	0.0001
CORR*METH	1	0.0035	0.92	0.3375
MEAS*CORR*METH	4	0.0051	0.34	0.8527

significant. The statistical significance of CORR implies that the equality of covariance matrices affects the classification accuracy. The interaction between CORR and METH is significant (at the 5% level), which implies that the differences in classification accuracy of probit and ID3 are not consistent across the equal and unequal covariance matrices. The mean classification accuracy of the probit method decreases by 0.0658 when I–I is replaced by I–$4I$ while the mean classification accuracy of the ID3 method decreases by 0.0459. This shows that the relative classification accuracy of the ID3 method to the probit method is higher when the covariances are unequal among populations than when the covariances are equal among populations. In other words, the classification performance of the ID3 method is less sensitive to the inequality of covariance matrices than the probit method while the classification accuracies of both probit and ID3 are lower under unequal covariance matrices than under equal covariance matrices.

The ANOVA results with two levels I–I and H–H for the correlation structure are summarized in Table 5. The main effect of CORR is significant and greater than that in Table 4. This implies that both the magnitude of correlations and the equality of covariance matrices affect the classification accuracy and that the magnitude of correlations has more impact on classification accuracy than does the equality of covariance matrices. The interaction effect between CORR and METH is not

TABLE 6
Two-way ANOVA for Probit

Class	Levels	Values
MEAS	5	INT, M6–2, M4–4, M2–6, NOM
CORR	3	I–I, I–$4I$, H–H

Dependent Variable: ACCU

Source	d.f.	SS	MS	F	Pr
Model	14	0.9229	0.0659	18.05	0.0001
Error	735	2.6842	0.0037		
Corrected total	749	3.6071			

Source	d.f.	ANOVA SS	F	Pr
MEAS	4	0.0844	5.77	0.0001
CORR	2	0.8334	114.11	0.0001
MEAS*CORR	8	0.0051	0.17	0.9942

significant. Though the mean classification accuracy of the probit method decreases by 0.0748 which is greater than 0.0674 by which the mean classification accuracy of the ID3 method decreases as *I–I* is replaced by *H–H*, the impact of the magnitude of correlations on the relative performance of ID3 to probit is not statistically significant. ID3 builds the classification tree by sequentially selecting new attributes. Newly added attributes may not improve the existing classification performance if they are highly correlated to already selected attributes.

A two-way ANOVA is conducted on probit and ID3 separately to provide a detailed analysis of the impact of measurement scale and correlation structure on the classification behavior of each of classification methods. The results of two-way ANOVA for the probit method are summarized in Table 6. The main effects of the treatment factor MEAS and the blocking factor CORR are significant. This confirms that the classification accuracy of the probit method significantly decreases as the proportion of binary variables increases in the classification model. The classification accuracy of the probit method decreases gradually from 0.8633 to 0.8353 as the numeric variables are replaced by the nominal variables. This trend is consistent across three levels of correlation structure as shown in Table 2. The statistical significance of CORR shows that the correlation structure affects the classification performance of the probit method. The interaction effect of MEAS*CORR is not significant. This statistical insignificance of MEAS*CORR confirms the consistent trend.

The analysis reported in Table 6 examined the combined effect of the magnitude of correlations and the equality of covariance matrices on the classification performance of the probit method. The separate analysis of the impact of the magnitude of correlations and the equality of covariance matrices is conducted. Table 7 summarizes the ANOVA study in which the correlation structure consists of two levels *I–I* and *I–4I*. The main effect of MEAS is significant as in Table 6. The main effect of CORR is significant. This shows that the classification accuracy of the probit method is significantly lower when the variances are unequal among populations than when the variances are equal among populations. The interaction effect of MEAS*CORR is not significant.

The impact of the magnitude of correlations and measurement scale on the classification performance of the probit method can be analyzed by the ANOVA results in Table 8 in which the correlation structure consists of two levels *I–I* and *H–H*. The main effect of MEAS is statistically significant (at 5% level) but less significant than that in Table 7. The main effect of CORR is statistically significant and more significant than that in Table 7. This implies that the classification accuracy of the probit method significantly decreases as the correlations among explanatory variables increase. The interaction effect of MEAS*CORR is not significant. The ANOVA results from Tables 7 and 8 imply that both the magnitude of correlations and the equality of covariance matrices affect the classification performance and that the magnitude of correlations has more impact on the classification performance of the probit method than does the equality of covariance matrices.

The two-way ANOVA results for ID3 are presented in Table 9. The main effect of MEAS is significant (at 5% level). The mean classification accuracy of the ID3 method increases gradually from 0.8216 to 0.8428 as the numeric variables are replaced by the nominal variables. The trend of increasing performance of ID3, as the numeric variables are replaced by the nominal variables, persists across three levels of correlation structure. The main effect of MEAS is less significant than that in case of probit (Table 6). The main effect of CORR is significant, which implies that the correlation structure

TABLE 7
Two-way ANOVA for Probit (Corr: *I–I* and *I–4I*)

Class	Levels	Values
MEAS	5	INT, M6–2, M4–4, M2–6, NOM
CORR	2	*I–I, I–4I, H–H*

Dependent Variable: ACCU

Source	d.f.	SS	MS	*F*	Pr
Model	9	0.6125	0.0681	19.18	0.0001
Error	490	1.7384	0.0035		
Corrected total	499	2.3509			

Source	d.d.	ANOVA SS	*F*	Pr
MEAS	4	0.0690	4.86	0.0008
CORR	1	0.5405	152.36	0.0001
MEAS*CORR	4	0.0030	0.21	0.9313

TABLE 8
Two-way ANOVA for Probit (Corr: *I–I* and *H–H*)

Class	Levels	Values
MEA-S	5	INT, M6–2, M4–4, M2–6, NOM
CORR	3	*I–I*, *I–4I*, *H–H*

Dependent Variable: ACCU

Source	d.f.	SS	MS	*F*	Pr
Model	9	0.7526	0.0836	22.1	0.0001
Error	490	1.8486	0.0038		
Corrected total	499	2.6012			

Source	d.f.	ANOVA SS	*F*	Pr
MEAS	4	0.0504	3.34	0.0103
CORR	1	0.6994	185.39	0.0001
MEAS∗CORR	4	0.0028	0.19	0.9451

significantly affects the classification performance of the ID3 method like it does the probit method. The interaction effect of MEAS∗CORR is not significant, which confirms that the classification behavior of ID3 is consistent across different levels of measurement scale.

The preceding analysis for ID3 examined the combined effect of the magnitude of correlations and the equality of covariance matrices. The effect of the equality of covariance matrices can be investigated separately from the effect of the magnitude of correlations. The results of ANOVA study in which the correlation structure consists of two levels *I–I* and *I–4I* are summarized in Table 10. The main effect of MEAS is not significant. The main effect of CORR is significant, which implies that the equality of covariance matrices affects the classification performance of the ID3 method like the case of the probit method. The interaction effect of MEAS∗CORR is not significant.

Table 11 summarizes the results of ANOVA study in which the correlation structure consists of two levels *I–I* and *H–H*. The main effect of MEAS is significant (at the 5% level). The smaller effect of MEAS in Table 10 is mainly because the classification performance of ID3 is not much affected by the measurement scale under *I–4I*. The classification accuracy of ID3 increases by 0.0100 under *I–4I* as the numeric variables are replaced by nominal variables while the classification accuracy of ID3 increases by 0.0320 under *I–I* and by 0.0216 under *H–H*. The main effect of CORR is significant and more significant than that in Table 10. This shows that both the magnitude of correlations and the equality of covariance matrices affect the classification accuracy of the ID3 method and that the magnitude of correlations has more impact on classification accuracy of ID3 than does the

TABLE 9
Two-way ANOVA for ID3

Class	Levels	Values
MEAS	5	INT, M6–2, M4–4, M2–6, NOM
CORR	3	*I–I*, *I–4I*, *H–H*

Dependent Variable: ACCU

Source	d.f.	SS	MS	*F*	Pr
Model	14	0.6453	0.0461	11.87	0.0001
Error	735	2.8538	0.0039		
Corrected total	749	3.4991			

Source	d.f.	ANOVA SS	*F*	Pr
MEAS	4	0.0411	2.64	0.0326
CORR	2	0.5921	76.25	0.0001
MEAS∗CORR	8	0.0121	0.39	0.9262

TABLE 10
Two-way ANOVA for ID3 (corr: *I–I* and *I–4I*)

Class	Levels	Values
MEAS	5	INT, M6–2, M4–4, M2–6, NOM
CORR	2	*I–I*, *H–H*

Dependent Variable: ACCU

Source	d.f.	SS	MS	*F*	Pr
Model	9	0.3003	0.0334	8.62	0.0001
Error	490	1.8972	0.0039		
Corrected total	499	2.1974			

Source	d.f.	ANOVA SS	*F*	Pr
MEAS	4	0.0268	1.73	
0.1425				
CORR	1	0.2636	68.06	0.0001
MEAS∗CORR	4	0.0099	0.64	0.6333

equality of covariance matrices. The interaction effect of MEAS∗CORR is not significant.

5. CONCLUSION

This is a comparative study of statistical and inductive learning methods. The focus of this study is to investigate the impact of measurement scale of explanatory variables on the relative performance of the ID3 method and the probit method and to examine the impact of correlation structure on the classification behavior of the ID3 method and the probit method.

The simulation results show that the relative classification accuracy of the ID3 method to the probit method increases as the proportion of binary variables increases in the classification model and that the relative classification accuracy of the ID3 method to the probit method is higher when the covariance matrices are unequal among populations than when the covariance matrices are equal among populations. The results also reveal that the classification accuracy of the ID3 method is lower when the covariance matrices are unequal among populations than when the covariance matrices are equal among populations and that the classification accuracy of the ID3 method decreases as the correlations among explanatory variables increase. The empirical results about the classification behavior of probit are consistent with those from previous studies.

The research findings of this paper provide a partial answer to why inductive learning methods performed better than statistical models in some of previous studies and why inductive learning methods performed worse than statistical methods in other studies. The implication of this study for classification research is that inductive

TABLE 11
Two-way ANOVA for ID3 (corr: *I–I* and *H–H*)

Class	Levels	Values
MEAS	5	INT, M6–2, M4–4, M2–6, NOM
CORR	2	*I–I*, *H–H*

Dependent Variable: ACCU

Source	d.f.	SS	MS	*F*	Pr
Model	9	0.6176	0.0686	18.34	0.0001
Error	490	1.8338	0.0037		
Corrected total	499	2.4515			

Source	d.f.	ANOVA SS	*F*	Pr
MEAS	4	0.0460	3.07	0.0162
CORR	1	0.5672	151.55	0.0001
MEAS*CORR	4	0.0046	0.31	0.8729

learning methods may be more appropriate when researchers deal with qualitative variables and that the inductive learning method may not be a good alternative to the statistical method when the explanatory variables under consideration are mostly quantitative. Another implication of this study is that the ID3 is likely to be a good alternative to statistical methods when the correlation structures significantly differ among groups.

This study has some limitations. First, this study uses only simulated data. Simulation allows researchers to manipulate data conditions and enhance internal validity, which makes the results more generalizable. However, the limitation of simulation is the lack of real world implications inherent in the artificial data. The results of this study using simulated data need to be tested by using real world data in the future study.

This study only considers the measurement scale of attribute and correlation structure. There are various factors which possibly affect the relative performance of inductive learning and statistical methods besides measurement scale of attribute and correlation structure. Examples include the number of attributes, the number of categories in an attribute, the number of classes, noise, the sample size, the underlying distribution of population, etc. Future studies are expected to compare inductive learning and statistical methods with respect to these factors and identify the relative strength and weakness of two approaches.

This study compares only two methods, the probit method and the ID3 method which are representative of statistical classification techniques and inductive learning methods respectively. There are various techniques for classifications available from statistics and inductive learning. A more extensive study including statistical and inductive learning methods other than probit and ID3 is expected to be conducted for generalization.

This study is about the exclusive choice of either method. An emerging issue of interest is the integration of inductive learning and statistical methods. Inductive learning and statistical methods have different strength and weakness. Therefore, a proper integration that takes advantages of the strengths of two approaches may enhance the classification performance.[11] The results of this study suggest a way of integrating inductive learning and statistical methods for mixed data. It is to apply the inductive learning method to the nominal variables, apply the statistical method to the numeric variables, and then combine the results from the two methods. Future studies need to develop an integrated algorithm or procedure which can perform better than the separate use of two approaches by utilizing comparative studies of inductive learning and statistical methods.

[11] Refer to Liang, Chandler, & Han (1990) for discussions of the general framework and research issues for integrating statistical and inductive learning approaches. Liang (1992) reported a significant performance improvement by an integration of two approaches called CRIS (Composite Rule Induction System).

REFERENCES

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, December, 1483–1536.

Arinze, B. & Subbanarasimha, P. (1994). An experimental investigation of the predictive accuracy of induction and regression. *Expert Systems with Applications*, **7**, 535–544.

Bayne, C., Beauchamp, J., Kane, V., & McCabe, G. (1983). Assessment of Fisher and logistic linear and quadratic discrimination models. *Computational Statistics & Data Analysis*, 1983, 257–273.

Braun, H. & Chandler, J. (1987). Predicting stock market behavior through rule induction: An application of the learning from example approach. *Decision Science*, 415–429.

Buchanan, B., Smith, D., White, W., Gritter, R., Feigenbaum, E., Lederberg, J., & Djerassi, C. (1976). Applications of artificial intelligence for chemical inference: Automatic rule formation in mass spectrometry by means of the meta-DENDRAL program. *Journal of American Chemical Society*, September, 6168–6178.

Chung, H. & Tam, K. (1993). A comparative analysis of inductive-learning algorithms. *International Journal of Intelligent Systems in Accounting, Finance, and Management*, January, 187–203.

Dietterich, T. & Hild, H. (1995). A comparison of ID3 and backpropagation for english text-to-speech mapping. *Machine Learning*, 51–80.

Dillon, W. & Goldstein, M. (1978). On the performance of some multinominal classification rules. *Journal of the American Statistical Association*, June 305–313.

Elliott, J. A. & Kennedy, D. B. (1988). Estimation and prediction of categorical models in accounting research. *Journal of Accounting Literature*, **7**, 202–242.

Garrison, L. & Michaelson, R. (1988). Symbolic concept acquisition: A new approach to determining underlying tax law constructs. *Journal of the American Taxation Association*, Fall, 77–91.

Greene, D. (1987). Automated knowledge acquisition: Overcoming the expert system bottleneck. *Proceedings of the 8th ICIS Conference*, Pittsburgh, pp. 107–17.

Hays, W. (1973). *Statistics for the Social Science*, 2nd edn., Holt, Rinehart & Winston.

Hoff, W., Michalski, R., & Stepp, R. (1986). INDUCE 3: A program for learning structural descriptions for examples. Artificial Intelligence Lab, University of Illinois at Urbana-Champaign.

IMSL STAT/LIBRARY users' manual. (1987). IMSL Inc.

Judge, G., Griffiths, W., Hill, R., & Lee, T. (1985). *The Theory and Practice of Econometrics*, 2nd edn. Wiley, New York.

Liang, T. (1992). A composite approach to automated knowledge induction for expert system design. *Management Science*, 1–17.

Liang, T., Chandler, J., & Han, I. (1990). Integrating statistical and inductive learning methods for knowledge acquisition. *Expert Systems with Applications*, **2**, 391–401.

Liang, T., Chandler, J., Han, I., & Roan, J. (1992). An empirical investigation of some data effects on classification accuracy. *Contemporary Accounting Research*, 306–328.

McKelvey, R. & Zavoina, W. (1971). An IBM FORTRAN IV program to perform N-chotomous multivariate probit analysis. *Behavioral Science*, March, 186–187.

McKelvey, R., Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 103–120.

Messier, W., Jr. & Hansen, J. (1988). Inducing rules for expert system development: An example using default and bankruptcy data. *Management Science*, December, 1403–1415.

Michalski, R. S. & Chilauski, R. L. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean diagnosis. *International Journal of Policy Analysis and Information Systems*, 125–161.

Michaelsen, R. & Swigger, K. (1994). Analysis of the effectiveness of machine learning in determining decision rules for executive compensation planning. *International Journal of Intelligent Systems in Accounting, Finance, and Management*, 263–278.

Mingers, J. (1987). Expert systems-rule induction with statistical data. *Journal of the Operations Research Society*, 39–47.

Moore, D. (1973). Evaluation of five discriminatory procedures of binary variables. *Journal of the American Statistical Association*, June, 399–404.

Paterson, A. & Niblet, T. (1982). *ACLS user manual*. Glasgow: Intelligence Terminal Ltd.

Quinlan, J. (1980). Discovering rules from large collections of examples: A case study. In D. Michie (Ed.) *Expert systems in the micro electronic age*. Edinburgh: Edinburgh University Press.

Quinlan, J. (1983). Learning efficient classification procedures and their application to chess end games. In R. Michalski, J. Carbonell, T. Mitchell (Eds), *Machine learning: An artificial intelligence approach*. Palo Alto, CA: Tioga.

Ragothaman, S. & Naik, N. (1994). Using rule induction for expert system development: The case of asset writedowns. *International Journal of Intelligent Systems in Accounting, Finance, and Management*, 263–278.

Rendell, L. (1986). A general framework for induction and a study of selective induction. *Machine Learning*, 177–226.

Schmitz, P., Habbema, J., Hermans, J., & Raatgever, J. (1983). Comparative performance of four discriminant analysis methods for mixtures of continuous and discrete variables. *Communications in Statistics—Simulation and Computation*, 727–751.

Schmitz, P., Habbema, J., & Hermans, J. (1985). A simulation study of the performance of five discriminant analysis methods for mixtures of continuous and binary variables. *Journal of Statistics, Computations, and Simulation*, 69–95.

Shannon, C. E. (1964). The mathematical theory of communication. In C. E. Shannon & W. Weaver, *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Shapiro, A. & Niblett, T. (1982). Automatic induction of classification rules for a chess endgame. In M. Clarke (Ed.), *Advances in computer chess*, Vol. 3. Edinburgh University Press.

Tu, P. (1989). Toward an intelligent classification tree approach to problem solving. Ph.D. dissertation, University of Illinois at Urbana-Champaign.

Vlachonikolis, I. & Marriott, F. (1982). Discrimination with mixed binary and continuous data. *Applied Statistics*, 23–31.