Expert Systems With Applications, Vol. 1, pp. 391–401, 1990
Printed in the USA.

0957-4174/90 $3.00 + .00

Integrating Statistical and Inductive Learning Methods for Knowledge Acquisition

TING-PENG LIANG

Purdue University, West Lafayette, IN, USA

JOHN S. CHANDLER AND INGOO HAN

University of Illinois, Champaign, IL, USA

Abstract—*Inductive learning is a method for automated knowledge acquisition. It converts a set of training data into a knowledge structure. In the process of knowledge induction, statistical techniques can play a major role in improving performance. In this paper, we investigate the competition and integration between the traditional statistical and the inductive learning methods. First, the competition between these two approaches is examined. Then, a general framework for integrating these two approaches is presented. This framework suggests three possible integrations:* (1) *statistical methods as preprocessors for inductive learning,* (2) *inductive learning methods as preprocessors for statistical classification, and* (3) *the combination of the two methods to develop new algorithms. Finally, empirical evidence concerning these three possible integrations are discussed. The general conclusion is that algorithms integrating statistical and inductive learning concepts are likely to make the most improvement in performance.*

1. INTRODUCTION

KNOWLEDGE ACQUISITION is a process by which expert knowledge is elicited and represented in a formal structure for decision making. It is a necessary and probably the most important step in developing expert systems. Traditionally, knowledge acquisition is considered a manual process in which knowledge engineers apply structured interviews or other techniques to communicate with experts and then document their findings (Kidd, 1987). Due to human cognitive limitations, however, this process has certain limitations. For example, it is well known that experts usually have difficulties in articulating their knowledge. In addition, the knowledge articulated by experts may be inconsistent and incomplete (Hoffman, 1987). Therefore, automating part of the manual process can significantly improve the productivity of expert systems development.

One approach to improving manual knowledge acquisition is to examine how decisions are made by experts and then apply an algorithm to induce the knowledge structure from this data. This process is usually called inductive learning, or rule induction, in machine learning. The major advantage of this approach is that knowledge is acquired based on existing evidence obtained from experts' real decisions. This reduces the effect of human cognitive biases and increases the efficiency of knowledge acquisition by automating the process.

Recently, much attention has been paid to this automated knowledge acquisition process (e.g., Carter & Cattlet, 1987; Chandrasekaran & Goel, 1988; Geene, 1987). Several well-known methods have been developed. For example, Quinlan (1979, 1983) modified Hunt, Martin, and Stone's (1966) induction mechanism to develop the ID3 method. Michalski and Stepp (1982) applied predicate logic to develop the AQ15 method. A typical inductive learning process includes three stages. First, experts identify the major factors (called attributes) that should be considered in the decision process and the possible decision outcomes (called classes). Second, the knowledge engineer collects existing cases and determines the attribute values and the actual outcome for each case. Finally, the data are analyzed by an induction mechanism and a set of rules is derived.

This process is quite similar to statistical processes such as regression or multiple discriminant analysis (MDA) that have been used by business researchers for decades. In fact, both inductive learning and statistical methods are tools for knowledge acquisition. They have a common goal of eliciting knowledge structures from data. The resulting structures can then be used to predict outcomes in new situations or to provide explanations for existing reality. The only difference between these two approaches is that different assumptions and algorithms are used to generate knowledge structures.

Requests for reprints should be sent to Ting-Peng Liang, Krannert Graduate School of Management, Purdue University, West Lafayette, IN 47907.

Statistical methods assume certain data distributions and focus on optimizing the likelihood of correct classification, whereas some existing inductive learning methods use criteria other than data distribution and maximum likelihood estimations. This difference also results in structures with different formats. An inductive learning method usually generates a decision tree or a set of decision rules, whereas a statistical method usually generates a linear function.

Given the same goal and different algorithms in statistical and inductive learning approaches, it is natural for knowledge engineers to consider them competitive methods. A number of studies in the literature have examined their differences and compared their performance under difference circumstances (e.g., Braun & Chandler, 1987; Chandler, Liang, & Han, 1989; Liang & Yu, 1989; Messier & Hansen, 1988; Mingers, 1989). While this comparative research provides certain insights into the selection of methods, a more important issue would be how these two approaches can be integrated to complement each other. In fact, the integration between statistical and artificial intelligence methods has been a research area for some time. For example, Gale (1986) edited a book on artificial intelligence and statistics. Lee, Oh, and Shim (1990) studied the application of a knowledge-based approach to assist statistical forecasting. Berzuini (1988) proposed an approach that used regression analysis as a preprocessor for constructing decision trees. Whitehall, Lu, and Stepp (1989) suggested an improvement of AQ15 by incorporating information of data distribution for processing continuous variables. Only through proper integration can new algorithms capable of generating highly accurate knowledge structures be developed.

Toward this end, this research studies how these two approaches can be integrated to improve the quality of the induced knowledge. There are three major motivations for this research. First, the integration of statistical and inductive learning approaches is likely to enhance the knowledge acquisition process. Previous research has found these two approaches to have different strengths and weaknesses in different areas. In an empirical study, for instance, Chandler, Liang, and Han (1989) found that Probit (a statistical method) outperformed ID3 (an inductive learning method) when the attributes were primarily numerical but ID3 outperformed Probit when the training sample size was small.[1] Therefore, a proper integration that takes advantage of the strengths of both methods should provide performance improvement.

Second, many statistical methods can be used to improve certain stages of the inductive knowledge acquisition process, but do not directly compete with the inductive learning approach. For example, correlation analysis can be used to determine the dependencies between attributes to facilitate the selection of the most appropriate set of attributes. Correlation analysis, however, is not designed for classification and hence there is no direct competition with inductive learning. In fact, statistical methods consist of a number of techniques for different purposes. Only a few of them that are frequently used for classification are competing with the inductive learning methods. Therefore, an investigation of opportunities to integrate nonclassification statistical analyses may significantly enhance the performance of inductive learning.

Third, a general framework for the integration is necessary to consolidate research findings and to guide future research. Although a few examples in the literature on possible integrations of these two approaches exist (e.g., Breiman, Friedman, Olson, & Stone, 1984; Liang, 1989a, 1989b; Mingers, 1987a, 1987b; Phelps & Musgrove, 1986; Rendell, 1986; Tu, 1989), they are largely ad hoc in nature. A framework that integrates previous findings and explores issues to be studied can lead to systematic research and expedite progress in the area.

In the remainder of this paper, we first review the studies concerning statistical and inductive learning approaches as competitive methods and discuss the relative advantages and drawbacks of each method. Then, we present a general framework for integrating statistical and inductive learning methods. The framework discusses three possible ways of integration: (1) statistical methods as preprocessors for inductive learning, (2) inductive learning methods as preprocessors for statistical analysis, and (3) a combination of the two methods to develop new rule induction algorithms. Finally, empirical findings concerning the integration of these methods and directions for future research are discussed.

2. COMPETITION OF STATISTICAL AND INDUCTIVE LEARNING METHODS

Although statistical classification and inductive learning have the same goal of eliciting knowledge structures from data, they have many differences. For example, statistical methods are usually based on some assumptions of data distribution, while inductive learning methods often ignore data distribution. Therefore, two issues need to be clarified in order to compare them.

First, *what methods are to be compared?* Since both statistical and inductive learning approaches consist of sets of different techniques, the comparison cannot be

[1] Probit is a prevailing statistical method for classification problems, and ID3, which stands for Iterative Dichomotizer 3, is the most popular inductive learning algorithm in machine learning. A brief description of ID3 can be found in Braun and Chandler (1987) and Messier and Hansen (1988); a description of statistical methods can be found in Judge, Hill, Griffiths, and Lee (1980). They are hence omitted in the paper.

performed on the approaches in general. Rather, representative techniques must be selected from each approach and then compared. For example, Braun and Chandler (1987) and Messier and Hansen (1988) compared multiple discriminant analysis (MDA)[2] and ID3. Chandler, Liang, and Han (1989) compared Probit and ID3. Mingers (1987b) compared statistical regression analysis and ID3. Parker and Abramowicz (1989) compared MDA, Probit and ID3. Garrison and Michaelsen (1989) examined MDA, Logit, and ID3. Elliott and Kennedy (1988) compared Probit and RPA (a decision tree induction algorithm). In general, ID3 is the most studied inductive learning technique, while Probit and MDA are the most studied statistical classification techniques. In addition, it should be noted that the findings obtained from existing comparisons only allow us to conclude that a certain statistical method is better or worse than a particular inductive learning method. We cannot conclude that, in general, statistical methods are better or worse than inductive learning methods.

Second, *how can these techniques be compared?* In other words, what are the major aspects to be compared? In general, there are two approaches that can be used to compare selected techniques: theoretical and empirical analyses.

Theoretical analysis focuses on the fundamental similarities and differences in the process of constructing knowledge structures. Since each approach is built on certain assumptions, uses certain criteria to select variables, and constructs models to optimize a measurement function, a theoretical analysis suggests that different techniques can be compared by their basic assumptions, measurement functions, criteria for variable selection, processes for variable selection, and resulting models.

For example, Table 1 shows a theoretical comparison of multiple discriminant analysis and ID3. The criteria used for comparison include the major assumptions underlying the method, the measurement used to determine the relevancy of a variable, the criteria used for selecting variables into the resulting model, the variable selection process, and the format of the resulting model. It shows that MDA assumes multivariate normal data distribution, no perfect correlation among independent variables, and equal covariance matrices for classes, whereas ID3 makes no such assumptions except that no conflicting data exists. This means that MDA has more rigid assumptions about the kind of data for which the method is designed. If the data distribution is not a multivariate normal distribution, then the power of the method is likely to decline.

TABLE 1
Comparison of Discriminant Analysis and ID3

(1) Major assumptions
Multiple Discriminant Analysis (MDA)
—Data population is multivariate normal distribution
—No perfect correlation among independent attributes
—Equal covariance matrices for classes
ID3 algorithm
—No conflict in the training data set

(2) Measurement, selection criteria, selection process, and resulting model

	MDA	ID3
Measurement	Covariance	Entropy
Selection criteria	Maximum likelihood estimation	Minimum entropy
Selection process	Matrix operation	Repetitive decomposition
Resulting models	Linear equations	Rule structures

Note: This table was adapted from Liang (1989a).

Concerning the measurement for the relevancy of a variable, MDA adopts the covariance matrix (a typical approach used in statistical methods), whereas ID3 uses entropy. Based on the relevancy measurement, MDA selects attributes to maximize the likelihood of correct classification, and generates a linear equation to capture the knowledge, whereas ID3 selects attributes to minimize the overall entropy, and generates a decision tree or rule structure to capture the knowledge.

Empirical analysis, on the other hand, considers the performance of the resulting model the most important criterion for method comparison. There are several possible performance measures. The most common one is the predictive power of the resulting models derived from different approaches. The approach that generates models with a higher prediction accuracy is considered better. Another measure is to compare the complexity of the resulting knowledge structures. The approach that generates a simpler knowledge structure is considered better. Given a selected performance measure, both statistical and inductive learning methods can be applied to the same set of training data to derive models. The resulting models are then applied to the same testing data for comparison.

For example, Braun and Chandler (1987) applied both discriminant analysis and ID3 to predict stock market behavior and found that ID3 outperformed both discriminant analysis and expert prediction. A similar result was confirmed by Messier and Hansen (1988), Parker and Abramowicz (1989), and Garrison and Michaelsen (1989). They found ID3 to be better than MDA in various domains including loan default,

[2] Multiple discriminant analysis is a statistical classification technique widely used in financial and marketing research. It derives from data a set of discriminant equations to classify different classes. A detailed description can be found in most statistics books and is omitted here.

tax decisions, and bankruptcy prediction. In a later study, however, Liang (1989a) found the difference in predictive power between ID3 and MDA was not statistically significant.

Instead of choosing discriminant analysis, Mingers (1987b) compared regression and ID3 and concluded that both techniques provided similar predictive power. Chandler, Liang, and Han (1989) extended previous research by investigating not only which technique is better but also the circumstances under which a particular technique is better. They controlled two data characteristics to compare the predictive power of Probit and ID3 in classifying LIFO/FIFO firms. They found that Probit outperformed ID3 when the training sample size was relatively large or the training samples were not dominated by categorical variables, whereas ID3 outperformed Probit otherwise. The most interesting implication of this finding is that statistical methods such as Probit and discriminant analysis may be better when their assumptions are satisfied (e.g., a large training sample size may result in a normal distribution that satisfies the data assumption of Probit), but may be worse otherwise. These results are not surprising, however, because the nonparametric nature of ID3 trades its prediction accuracy for efficiency (i.e., sacrificing optimum accuracy in some cases to reduce the minimum sample size for obtaining satisfactory accuracy and to broaden its applicability to other situations).

So far, neither theoretical nor empirical research concludes that one approach is better than the other. In fact, this conclusion may never be reached since there are so many factors that cannot be completely controlled by researchers. The major contribution of these comparative studies is, therefore, to provide insights into these different approaches and to motivate better integration of them.

3. INTEGRATION OF STATISTICAL AND INDUCTIVE LEARNING METHODS

There are at least three ways in which statistical and inductive learning methods can be integrated. First, statistical methods may be used as preprocessors for inductive learning methods. In other words, a statistical technique is applied to the data set before an inductive learning method is applied. The rationale behind this approach is that an inductive learning method usually is inaccurate in handling a large number of numerical variables because of its nonparametric nature. Therefore, a statistical method can be applied as a preprocessor to combine the numerical variables into a few attributes. The inductive learning method can then derive a knowledge structure from the original categorical attributes and the reduced set of numerical attributes generated from the statistical method.

The second approach to integration is to use an inductive learning method as a preprocessor for statistical methods. In contrast to the previous approach; an inductive learning method is applied to the data before a statistical method is applied. The rationale behind this approach is that categorical attributes usually violate the normality assumption associated with many statistical methods. Therefore, applying an inductive learning method to reduce the number of categorical attributes may be able to increase the accuracy of the resulting model.

In addition to the previous two straightforward approaches, a third approach is to combine statistical concepts into certain stages of inductive learning. In other words, the basic process of inductive learning remains unchanged, but statistical methods may be incorporated into selected stages to improve the performance. The rationale behind this approach is that a sequential processing of data with different methods in the previous two approaches may lead to the suboptimization of the resulting knowledge structures. In order to pursue the best knowledge structure, therefore, a maximum penetration of statistical concepts in the inductive learning process must be allowed. In order to differentiate the third approach from the previous two, we may call it "deep" integration and call the previous two "surface" integration.

3.1. A Framework for Deep Integration

An important question associated with deep integration is "where can integration occur?" To answer this question requires an examination of the functions of statistical methods and the steps in inductive learning.

The primary purpose of statistical methods is to infer further properties of populations from information available in sample data. In general, these methods fall into four functional categories: sampling, data analysis, classification, and hypothesis testing. Sampling techniques focus on constructing a set of unbiased samples to ensure the validity of data analysis. For example, a random sampling procedure and a proper experimental design can reduce systematic errors. Random number generators can also be used to create simulated data bases.

Data analysis techniques are usually used to provide statistics useful for inferring properties about populations. For example, calculation of mean and standard deviation provides unbiased and efficient estimators for probability estimation. Correlation analysis provides information concerning the dependencies among attributes.

Statistical classification techniques take advantage of information generated from data analysis to construct models for explaining different classifications and predicting possible outcomes for new cases. Typical examples include regression analysis, multiple dis-

criminant analysis, Probit, Logit, factor analysis, and cluster analysis.

Hypothesis testing techniques are useful in verifying whether a particular situation is the same as originally assumed. Typical examples include Chi-square test, p test, F test, and Z test, among others.

In addition to the available techniques, we need to know where these techniques can be applied. A typical inductive learning process includes three stages: (1) construction of a training data set, (2) development of the knowledge structure, and (3) refinement of the knowledge structure. As illustrated in Figure 1, statistical techniques may be applied to all of these three stages.

3.1.1. *Construction of Training Set.* In the first stage, a set of training data must be collected by the knowledge engineer. This includes selection of relevant cases, determination of the sample size, and selection of proper attributes. In most inductive learning literature, the training data set is considered given. Therefore, discussion of the training set construction is extremely inadequate.

Statistical techniques applicable to this stage of inductive learning include the following. First, sampling techniques can be used to determine which and how many cases need to be included in the training set. For example, a knowledge engineer may use random or systematic sampling techniques to compile an unbiased training data set to reduce systematic errors.

Second, data analysis techniques can be used to determine what attributes to include in the training set. For example, some attributes are highly correlated and may be dropped without affecting the quality of the resulting model. This would require a correlation analysis to be performed on all attributes before they are selected. In addition, Bayesian and other estimators may also be used to estimate the missing values in the training set (Fisher, 1987; Konomenko, Bratko, & Boskar, 1984).

Third, statistical classification techniques can be applied to transform several attributes into more meaningful ones. This is necessary when the original data set consists of too many attributes or some attributes are highly correlated. For instance, we may use factor analysis to identify 4 or 5 significant factors out of a set of 20 attributes.

Fourth, testing techniques can be used to determine how much bias the training set may introduce. Since construction of a training set is a resampling process that selects a subset out of a set of samples, there are chances that biases may be introduced in this resampling process. For example, a training set that makes a 50–50 split of bankrupt and healthy firms when in reality, the ratio is probably 1 to 50, may result in a model that tends to overestimate the likelihood of bankruptcy.

3.1.2. *Development of Knowledge Structure.* The second stage of inductive learning is to develop a knowledge structure from the training data set. This includes operations that determine the relative importance of attributes, identify the causal relationships between attributes and classes, assess the probability associated with the causal relationships, and build the final

Functions of statistical techniques	Stages in Inductive Learning		
	Construction of training sets	Development of structures	Refinement of structures
Sampling	What and How many cases to include? (A)	How to model incrementally? (E)	What training cases to reuse? (I)
Analysis	What attributes to include? (B)	What attributes to select? (F)	What rule to refine? (J)
Classification	How can attributes be transformed? (C)	How do they compete? (G)	How to rebuild structures? (K)
Hypothesis testing	How much bias does the training set have? (D)	Is the model good? (H)	Is refinement necessary? (L)

FIGURE 1. A framework for deep integration.

knowledge structure. Statistical techniques applicable to this stage include the following.

First, sampling techniques can be applied to determine how incremental learning can be performed. Incremental learning is an important concern in implementing an inductive learning algorithm. It makes the learning process more efficient. For example, bootstrap or jackknife procedures may be applied to cross-evaluate the knowledge structure during the incremental learning process.

Second, data analysis techniques can be applied to determine the relative importance of attributes and to select the most appropriate ones. For example, Tu (1989) applies correlation analysis to determine the dependency among attributes and uses a look-ahead heuristic to improve the knowledge development process. The integration is reported capable of reducing the complexity of the induced knowledge structure. Furthermore, causal modeling techniques allow causal relationships to be identified, statistical estimation and Bayesian statistics allow probability associated with each relationship to be assessed (Lee & Ray, 1986; Liang, 1989a, Rendell, 1986), and other statistics such as Chi-square or G statistics may be used to replace entropy as information measures for constructing the knowledge structure (e.g., Hart, 1984; Mingers, 1987a; Race & Thomas, 1988).

Third, statistical classification techniques may be used as an alternative to decision trees or decision rules. For example, after identifying key attributes and their causal relationships with the dependent variable, a linear decision model, instead of a decision tree or decision rules, may be built.[3] Although no existing literature has indicated this integration, it remains a possibility, however.

Fourth, hypothesis testing techniques can be used to evaluate the knowledge structure generated from the training data. For example, O'Leary (1987) developed an approach that used a Chi-square test to validate the performance of expert systems. This same technique can be used to validate the resulting knowledge structure. In addition, other techniques such as an F test may be used to test the significance of misclassification.

3.1.3. *Refinement of Knowledge Structure.* After a knowledge structure is developed, it can be used to support decision making. Sometimes, however, the structure may not be good enough. For example, it may be too complex or proven invalid when applied to real cases. In addition, knowledge usually is dynamic and evolving over time. Therefore, refinement of a knowledge structure is often necessary. In the knowledge refinement process, there are several issues that can use statistical techniques, including when a refinement is necessary, what rules to refine, and whether the refinement is significant.

First, similar to the construction of a training data set, statistical methods can be applied to select a set of cases for refinement. They can help the knowledge engineer determine how many cases are necessary and whether an addition or deletion of attributes may be necessary.

Second, data analysis techniques can be applied to determine what rules to refine, which branch of the decision tree to prune, and how to assign responsibility when misclassification occurs. For example, a frequency analysis may be used to analyze the performance of each rule and then refine the rules proven inaccurate (Liang, 1989a) or to prune or simplify the decision tree (Quinlan, 1983, 1986, 1987a, b, c).

Third, statistical classification techniques can be used to rebuild knowledge structures and determine what is in error. To determine what is wrong with the existing knowledge structure is itself a classification problem. Therefore, regression analysis or other statistical classification techniques may be used in the process.

Fourth, hypothesis testing techniques can be applied to determine whether a refinement is necessary. Sometimes, misclassification is due to the noise in the problem domain. This kind of error is usually called random error. In this case, refinement of the knowledge structure is unnecessary. In order to differentiate random errors from systematic errors generated from an inaccurate knowledge structure, statistical testing techniques are essential. In addition, after a refinement is considered necessary, an optimal alternative must be selected from a number of alternatives. Statistical testing is also necessary to compare the relative contribution of the candidates and choose the one with the most significant contribution to maximize the effect of refinement. For example, Liang (1989b) proposes using a p test to test the significance of misclassification and to select the optimal refinement.

In summary, a framework for integrating statistical and inductive learning methods has been presented. The framework consolidates existing research findings and provides guidelines for future studies. In the following section, two empirical studies and their findings about different integrations will be presented.

4. EMPIRICAL ANALYSES

The previous framework describes three possible approaches for integration: (1) statistical methods as preprocessors, (2) inductive learning methods as preprocessors, and (3) deep integration. In order to understand which approach is more promising, two empirical studies have been conducted to compare them. In the

[3] The purpose of this point is to explore a possible integration. However, we have no intention to suggest that a linear model is better than decision rules or vice versa.

first experiment, the data include both nominal and numerical variables. In the second experiment, the data include numerical variables only. Since some statistical methods cannot handle data with nominal variables only, the case where only nominal variables present is not examined in this research.[4] Based on the findings reviewed in Section 2, however, it is likely that inductive learning may be better when the problems include nominal variables only.

Given the large number of possibilities, it is obviously impossible for the authors to compare all alternatives exhaustively. Therefore, this empirical work is more exploratory than conclusive. The findings, however, do provide some initial guidelines for future work.

4.1. The First Experiment

4.1.1. *Data Collection.* The data for the first empirical study were 12 pairs of bankruptcy data sets originally compiled in Liang (1989a). Each data set pair included a training and a testing set. Six pairs consisted of 30 cases in the training set and the other six pairs consisted of 20 cases. All testing sets consisted of 20 cases. Each case included a class (i.e., bankrupt or not), three categorical and five numerical variables, as follows:

X1 = consistency exception opinion, yes or no;
X2 = subject-to opinion, yes or no;
X3 = going-concern opinion, yes or no;
X4 = the ratio of net income/total assets;
X5 = the ratio of current assets/total assets;
X6 = the ratio of current assets/current liabilities;
X7 = the ratio of cash/total assets;
X8 = the ratio of sales/current assets.

4.1.2. *Experimental Procedures.* The experiments included two parts: One examined surface integrations and the other investigated deep integration. For surface integration, multiple discriminant analysis (MDA) and factor analysis were selected as the representative of statistical methods and ID3 was chosen as the representative of inductive learning methods.[5] They were chosen because of their popularity in literature. These three techniques allowed us to examine three alternatives. First, MDA was used as a preprocessor of ID3. Second, ID3 was used as a preprocessor of MDA. Third, factor analysis was applied to reduce the number of numerical variables for ID3. For deep integration, we examined the CRIS approach that integrated statistical inferences in rule induction process.[6] In other words, for each data set pair, the following four integrated methods were applied:

1. MDA + ID3: MDA was applied to the training data set to simplify the attributes and then ID3 was applied to the simplified training data set to derive a decision tree model.[7] The model was then used to predict the cases in the testing data set.
2. ID3 + MDA: ID3 was applied to the training data set to derive a knowledge structure. Then, the attributes excluded from the knowledge structure were dropped from the training set to simplify the training set. Finally, MDA was applied to the simplified training set to generate a linear classification model. The model was then applied to predict the cases in the testing data set.
3. FACTOR + ID3: Factor analysis was applied to the training data set to reduce the number of numerical variables. Based on the resulting factor loads, the training data set was modified and then used to derive a decision tree by ID3. The resulting decision tree was then used to predict the cases in the testing data set.
4. CRIS: The CRIS approach was applied to each training data set to generate a set of decision rules. The resulting rules were then applied to predict the corresponding testing cases.

The primary criterion used for comparing different combinations was the predictive power of the resulting model. It was measured by the percentage of the cases in the testing data set correctly predicted by the model.

4.1.3. *Data Analysis.* Following the previous procedures, 12 observations were obtained for each situation. These results are compared with those obtained from using MDA or ID3 alone. Table 2 summarizes the predictive accuracy in various settings. As indicated in the table, CRIS and ID3 + MDA results in the best average predictive accuracy; MDA, ID3, and MDA + ID3 have slightly worse accuracy; and FACTOR + ID3 is the worst.

A Freeman Two-way ANOVA test shows that the method effect is statistically significant ($p = 0.0242$). This means that different methods do show different predictive accuracy and the difference is statistically

[4] A dissertation project that varies the ratio of nominal/numerical variables and controls the degree of multicollinearity is underway (Han, 1990).

[5] Factor analysis is a statistical technique for reducing the number of numerical variables. It is popular in behavioral and social sciences research. It is selected to examine the effect of applying a statistical technique to simplify data for inductive learning methods. Since it is not a classification technique, using ID3 as a preprocessor of factor analysis is infeasible.

[6] CRIS stands for a Composite Rule Induction System. It is an algorithm that applies statistical inference procedures to numerical variables and frequency tables to nominal variables to generate hypotheses. The hypotheses are then selected based on their saliencies to construct a rule base. A detailed description can be found in Liang (1989a).

[7] Decision trees and decision rules are interchangable. In other words, they are equivalent to each other. A tree can be converted to a set of rules, and vice versa.

TABLE 2
Prediction Accuracy Under Various Settings

	MDA	ID3	MDA + ID3	ID3 + MDA	Factor + ID3	CRIS
(a) Training sample size = 30						
	.85	.85	.80	.90	.75	.85
	.70	.80	.65	.95	.75	.85
	.70	.75	.90	.80	.65	.80
	.65	.80	.70	.60	.70	.75
	.80	.80	.80	.90	.65	.85
	.75	.65	.70	.80	.60	.70
M:	.74	.78	.76	.83	.68	.80
(b) Training sample size = 20						
	.85	.90	.85	.90	.90	.85
	.65	.80	.65	.75	.75	.80
	.70	.75	.75	.80	.70	.80
	.80	.80	.70	.65	.75	.80
	.85	.65	.70	.85	.70	.90
	.80	.70	.80	.70	.60	.75
M:	.78	.77	.74	.78	.73	.82
Global Mean	.76	.77	.75	.80	.71	.81

meaningful. We then separated the results from the 30-case and the 20-case training sets and conducted a pairwise t test on the data collected from the 30-case training data sets. The results indicate that ID3 + MDA is significantly better than MDA ($t = 2.08$, $p = 0.09$) and FACTOR + ID3 ($t = 2.79$, $p = 0.038$). A Wilcoxon Paired Rank Test on all methods indicates CRIS is significantly better than MDA ($p = 0.382$), MDA + ID3 ($p = 0.06$), and FACTOR + ID3 ($p = 0.04$), and ID3 + MDA is slightly better than MDA. The general conclusion that we can derive from this experiment is that an integration of statistical and inductive learning methods is better than the individual methods alone.

4.2. The Second Experiment

In order to study whether the nature of variables may have some effect on the results, a second experiment was conducted. The major difference between this experiment and the first one is that the data sets include only numerical variables. The data sets were obtained by dropping the three nominal variables from the data sets in the first experiment. Therefore, each case includes five numerical variables (i.e., X4–X8 described in the first experiment). Since the first experiment indicated that ID3 + MDA and CRIS outperformed the other two alternative integrations, MDA + ID3 and FACTOR + ID3, only two methods were examined. Their performances were then compared with those of ID3 and MDA.

The results, as shown in Table 3, confirm our previous findings that certain integrated methods can generate higher predictive accuracy. The performance of CRIS is the best among the methods compared. A Wilcoxon Paired Rank Test indicates that CRIS significantly outperforms MDA ($p = 0.028$) and ID3 ($p = 0.03$). The difference between CRIS and ID3 + MDA is not statistically significant ($p = 0.173$), although CRIS has a higher average predictive accuracy. The differences between ID3 + MDA, ID3, and MDA are also not statistically significant (all propabilities are lower than 0.20). These findings indicate that deep integration may be better than individual methods alone and surface integration.

TABLE 3
Prediction Accuracy Under Various Settings

	MDA	ID3	ID3 + MDA	CRIS
(a) Training sample size = 30				
	.80	.80	.90	.90
	.70	.80	.85	.85
	.80	.70	.85	.85
	.65	.80	.60	.80
	.90	.80	.80	.90
	.75	.65	.80	.70
M:	.77	.76	.80	.83
(b) Training sample size = 20				
	.85	.90	.90	.85
	.65	.85	.75	.85
	.75	.75	.80	.80
	.70	.80	.65	.80
	.80	.70	.85	.85
	.75	.70	.70	.70
M:	.75	.78	.78	.81
Global Mean	.76	.77	.79	.82

4.3. Discussions

In the previous two experiments, CRIS has been found consistently better than individual methods, whereas ID3 + MDA has been found slightly better than individual methods. This indicates that both statistical and inductive learning methods can benefit from integration. One reason that may explain the superiority of ID3 + MDA in the first experiment is that ID3 screens out the attributes dominated by others and hence reduces the dependency among attributes. In other words, ID3 may have reduced the multicollinearity existing in the raw data sets. This allows the MDA algorithm to derive a more accurate model. This observation is also supported by the superiority of CRIS that adopts a rule-scheduling approach to eliminate redundant rules. A possible reason for explaining the inferiority of FACTOR + ID3 is that some important information may be lost in the attribute aggregation process. In other words, instead of screening out useless attributes, factor analysis may have dropped out some important information. Hence, issues such as how do we know whether a screening process would rule out useful or useless information and why a particular method drops useless information while others miss useful information become interesting for future research.

Another reason that may explain the superiority or inferiority of a particular method is whether the model fit the training samples properly. Systematic errors due to an overfit or underfit of the training sample usually deteriorate the performance of the resulting model. Two criteria can be used to measure the extent to which the model fits the training data set. For methods generating linear decision models, this may be measured by the percentage of cases in the training set correctly classified by the model (called internal validity). The higher this percentage is, the more likely that there may exist an overfit. For methods generating decision tree models, this may be measured by the complexity of the tree. The more complex the tree is, the more likely that there may be an overfit. In this research, we use the number of nodes and leaves in a decision tree to represent the complexity of the tree.

Based on these criteria, internal validity was measured for MDA and ID3 + MDA, and tree complexity was measured for ID3, MDA + ID3, and FACTOR + ID3. Then, correlation analysis was performed to detect the relationship between prediction accuracy and internal validity or tree complexity. The results, as shown in Table 4, indicate two findings.

First, a weak negative relationship exists between internal validity and prediction accuracy ($p = 0.14$ in Table 4a). Although this insignificance may be due to the small sample size, it is still worth noting that the increase of internal validity tends to overfit the training data and hence jeopardizes the prediction accuracy.

TABLE 4
Average Performance and Correlation Analysis

(a) Linear Decision Models

	Internal Validity		Prediction Accuracy	
Method	20-case	30-case	20-case	30-case
MDA	.900	.894	.775	.742
ID3 + MDA	.858	.822	.775	.825

Correlation coefficient = −.8556
Probability = .144

(b) Decision Tree Models

	Tree Complexity		Prediction Accuracy	
Method	20-case	30-case	20-case	30-case
ID3	7.67	10	.766	.775
MDA + ID3	8	12.3	.742	.758
FACTOR + ID3	12.83	19.33	.733	.683

Correlation coefficient = −.8445
Probability = .034

Note: All data in the Table are means. For example, 7.67 in (b) is the average ID3 tree complexity over six data sets for 20-cases data.

Second, a strong negative relationship exists between tree complexity and prediction accuracy ($p = 0.03$ in Table 4b). This implies that a simpler tree may be preferred over a more complex tree and overspecification must be avoided in designing a knowledge acquisition algorithm. In fact, this is where statistical methods can play a role in the inductive learning process. Since most inductive learning methods are based on repetitive decomposition, a certain degree of overspecification often exists. Applying statistical concepts to detect and reduce this possibility may be a very fruitful area for future research.

In another previous study, Tu (1989) developed a different deep integration algorithm that adopted a look-ahead heuristic to detect the dependency among attributes and then compared its tree complexity with that of ID3. She found that the heuristic significantly reduced the complexity of the resulting model. Although no comparison of prediction accuracy was performed, the negative relationship found in our research may suggest that her approach has a lower probability of overfitting the training set.

In summary, the empirical analysis has allowed us to explore certain insights into the integration of statistical and inductive learning methods. The general findings include the following:

1. Integration of statistical and inductive learning methods to detect and remove dominated attributes from the training data set is a key issue. A proper integration can significantly increase the prediction accuracy of resulting knowledge structures.

2. Overfitting the training data set tends to reduce the prediction accuracy. A proper use of statistical methods may prevent such overfitting.
3. Surface integration may not be able to generate any improvement unless it can remove redundancy or prevent overfitting. A poor integration may lose information and significantly deteriorate the prediction accuracy (such as the integration between factor analysis and ID3).
4. Algorithms heading toward deep integration (such as CRIS) are likely to create significant improvements.

Since this research is among the first that discusses the integration of statistical and inductive learning methods, it is more exploratory than conclusive. The above observations serve as a good starting place for further research. Inevitably, some statements would need more supporting research. Nonetheless, the empirical findings described above have provided many initial insights.

5. CONCLUSION

Statistical and inductive learning are two major approaches for inducing knowledge from data. Although their similarity in goal and data-processing process make many researchers consider them as competing methods, this research focused on the synergy that may be generated from their proper integration. In this paper, we first reviewed findings concerning the relative advantages and drawbacks of these two approaches. Then, we presented a conceptual framework for their integration. The framework classifies statistical methods into four categories: sampling, data analysis, classification, and hypothesis testing, and examines their potential applications in each of the following three inductive learning stages: construction of training set, development of knowledge structures, and refinement of knowledge structures. Finally, empirical findings were presented and analyzed to derive general guidelines. Given the complexity of the issue and the variety of possible integrations, the observations provided in this paper may not be conclusive. Much further research needs to be conducted. Nonetheless, these findings should provide a good starting point and trigger future works in this line of research.

REFERENCES

Berzuini, C. (1988). Combining symbolic learning techniques and statistical regression analysis. *AAAI-88 proceedings,* (pp. 612–617). San Mateo, CA: Morgan Kaufmann.

Braun, H., & Chandler, J.S. (1987). Predicting stock market behavior through rule induction: An application of the learning-from-example approach. *Decision Sciences,* **18**(3), 415–429.

Breiman, L., Friedman, J., Olson, R., & Stone, C. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks.

Carter, C., & Cattlet, J. (1987). Assessing credit card applications using machine learning. *IEEE Expert,* Fall, 71–79.

Chandler, J.S., Liang, T.P., & Han, I. (1989). An empirical comparison of probit and ID3 methods for accounting classification research. BEBR Working Paper No. 89-1592, University of Illinois at Urbana-Champaign.

Chandrasekaran, B., & Goel, A. (1988). From numbers to symbols to knowledge structures: Artificial intelligence perspectives on the classification task. *IEEE Trans. on Systems, Man, and Cybernetics,* **18**(3), 415–424.

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

Fisher, D.H. (1987). Conceptual clustering, learning from examples, and inference. *Proceedings of the Fourth International Conference on Machine Learning* (pp. 38–49). San Mateo, CA: Morgan Kaufmann.

Gale, W.A. (Ed.). (1986). *Artifical intelligence and statistics.* Reading, MA: Addison-Wesley.

Garrison, L.R., & Michaelsen, R.H. (1989). Symbolic concept acquisition: A new approach to determining underlying tax law constructs. *Journal of American Taxation Association,* Fall, 77–91.

Geene, D.P. (1987). Automated knowledge acquisition overcoming the expert system bottleneck. *Proceedings of the 8th ICIS Conference* (pp. 107–117). Pittsburgh.

Han, I. (1990). The impact of measurement scale on classification performance of inductive learning and statistical approaches. Doctoral dissertation, University of Illinois at Urbana-Champaign.

Hart, A. (1984). Experience in the use of an inductive system in knowledge acquisition. In M. Bramer (Ed.), *Research and development in expert systems.* Cambridge, England: Cambridge University Press.

Hoffman, R.R. (1987). The problem of extracting knowledge of experts from the perspective of experimental psychology. *AI Magazine,* **8**(2), 53–67.

Hunt, E.B., Martin, J., & Stone, P.T. (1966). *Experiments in induction.* New York: Academic Press.

Judge, G.G., Hill, R.C., Griffiths, W.E., & Lee, T.C. (1980). *The theory and practice of econometrics,* New York: John Wiley & Sons.

Kidd, A.L. (1987). *Knowledge acquisition for expert systems.* New York: Plenum Press.

Konomenko, I., Bratko, I., & Boskar, E. (1984). *Experiments in automatic learning of medical diagnostic rule,* (Tech. Rep.), Ljubljana, Yugoslavia: Josef Stefan Institute.

Lee, J.K., Oh, S.B., & Shim, J.C. (1990). UNIK-FCST: Knowledge-assisted adjustment of statistical forecasts. *Expert Systems with Applications,* **1**(1), 39–49.

Lee, W.D., & Ray, S.R. (1986). *Probabilistic rule generator.* (Tech. Rep. No. UIUCDCS-R-86-1263). University of Illinois at Urbana-Champaign, Department of Computer Science.

Liang, T.P. (1989a). A composite approach to inducing knowledge for expert systems design. BEBR Working Paper No. 89-1534, University of Illinois at Urbana-Champaign.

Liang, T.P. (1989b). Empirical knowledge refinement in noisy domains. *Proceedings of the Fourth Knowledge Acquisition for Knowledge-based Systems.* Banff, Canada: University of Calgary.

Liang, T.P., & Yu, C.J. (1989). A methodological note on examining product characteristics and foreign market entry strategies. In *Proceedings of the Second International Conference on Comparative Management* (pp. 317–321). Kaohsiung, Taiwan.

Messier, W.F., Jr., & Hansen, J.V. (1988). Inducing rules for expert systems development: An example using default and bankruptcy data. *Management Science,* **34**(12), 1403–1415.

Michalski, R.S., & Stepp, R. (1982). Revealing conceptual structure in data by inductive learning. *Machine Intelligence,* **10**, 173–196.

Mingers, J. (1987a). Expert systems—rule induction with statistical data. *Journal of the Operational Research Society,* **38**(1), 39–47.

Mingers, J. (1987b). Rule induction with statistical data—a com-

parison with multiple regression. *Journal of the Operational Research Society,* **38**(4), 347–351.

Mingers, J. (1989). An empirical comparison of selection measures for decision-tree induction. *Machine Learning,* **3**, 319–342.

O'Leary, D.E. (1987). Validating the weights in rule-based expert systems: A statistical approach. *International Journal of Expert Systems,* **1**(3), 253–279.

Parker, J.E., & Abramowicz, K.F. (1989). Predictive abilities of three modeling procedures. *Journal of American Taxation Association,* Fall, 37–53.

Phelps, R.I., & Musgrove, P.B. (1986). Artificial intelligence approaches in statistics. In W.A. Gale, (Ed.), *Artificial intelligence and statistics* (pp. 159–171). Reading, MA: Addison-Wesley.

Quinlan, J.R. (1979). Discovering rules from large collections of examples: A case study. In D. Michie (Ed.), *Expert systems in the micro electronic age* (pp. 168–201). Edinburgh, Scotland: Edinburgh University Press.

Quinlan, J.R. (1983). Learning efficient classification procedures and their application to chess end-games. M.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (pp. 463–482). Los Altos, CA: Morgan Kaufmann.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning,* **1**(1), 81–106.

Quinlan, J.R. (1987a). Simplifying decision trees. *International J. of Man–Machine Studies,* **27**, 221–234.

Quinlan, J.R. (1987b). Decision trees as probabilistic classifiers. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 31–37). Los Altos, CA: Morgan Kaufmann.

Quinlan, J.R. (1987c). A case study of inductive knowledge acquisition. In J.R. Quinlan (Ed.), *Applications of expert systems.* Reading, MA: Addison-Wesley.

Race, P.R., & Thomas, R.C. (1988). Rule induction in investment appraisal. *Journal of the Operational Research Society,* **39**(12), 1113–1123.

Rendell, L. (1986). Induction, of and by probability. In L.N. Kanal & J.F. Lemmer (Eds.), *Uncertainty in artificial intelligence* (pp. 429–443). New York: North-Holland.

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

Whitehall, B.L., Lu, C.Y., & Stepp, R.E. (1989). *CAQ: Making AQ work with engineering problems.* Working paper, University of Illinois at Urbana-Champaign.