BEBR
FACULTY WORKING
PAPER NO. 89-1615

# Integrating Statistical And Inductive Learning Methods for Knowledge Acquisition

*Ting-peng Liang*
*John S. Chandler*
*Ingoo Han*

College of Commerce and Business Administration
Bureau of Economic and Business Research
University of Illinois Urbana-Champaign

BEBR

FACULTY WORKING PAPER NO. 89-1615

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

November 1989

# Integrating Statistical and Inductive Learning Methods for Knowledge Acquisition

Ting-peng Liang
Assistant Professor

John S. Chandler
Associate Professor

Ingoo Han
Ph. D. Candidate

Department of Accountancy
University of Illinois
1206 South Sixth Street
Champaign, IL 61820

# INTEGRATING STATISTICAL AND INDUCTIVE LEARNING METHODS FOR KNOWLEDGE ACQUISITION

## 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 to improve 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 pre-processors for inductive learning, (2) inductive learning methods as pre-processors 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).

An alternative to manual knowledge acquisition is to examine how decisions are made by experts and then induce the knowledge structure from the existing 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., Chandrasekaran and Goal, 1988; Geene, 1988). 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 are derived.

This process is quite similar to some 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. 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 as competitive methods. A number of studies in the literature have examined their differences and compared their performance under difference circumstances (e.g., Braun and Chandler, 1987; Chandler, Liang, and Han, 1989; Liang and Yu, 1989; Messier and Hansen, 1988). 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 sometime. For example, Gale (1986) edited a book on artificial intelligence and statistics. Lee, Oh, and Shim (1990) studied the application of knowledge-based approach to assist statistical forecasting. 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 that these two approaches have different strengths and weakness in different areas. In an empirical study, for instance, Chandler, Liang and Han (1989) found that Probit outperformed ID3 when the attributes were primarily numerical but ID3 ourperformed Probit when the training sample size was small. 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 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 non-classification 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 and Olson, 1984; Liang, 1989a, 1989b; Mingers, 1987a, 1987b; Phelps and 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 a pre-processor for inductive learning, (2) inductive learning methods as a pre-processor 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 is to be compared? Since both statistical and inductive learning approaches consist of a set of different techniques, the comparison cannot be performed on the approaches in general. Rather, representative techniques must be selected from each approaches and then compared. For example, Braun and Chandler (1987) and Messier and Hansen (1988) compare discriminant analysis (a statistical technique) and ID3 (an inductive learning technique). Chandler, Liang and Han (1989) compares Probit (a statistical technique) and ID3. Mingers (1987) compares statistical regression analysis and ID3. In general, ID3 is the mostly studied inductive learning technique, while Probit and discriminant analysis are the mostly studied statistical classification techniques. In addition, the findings obtained from comparison only allow us to conclude that a certain statistical classification technique is better or worse than a particular inductive learning technique. 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, process for variable selection, and the resulting model.

For example, Table 1 shows a theoretical comparison of discriminant analysis and ID3. It shows that discriminant analysis (DA) 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. In addition, DA adopts the covariance matrix to measure the relevancy of attributes, selects attributes to maximize the likelihood of correct classification, and generates a linear equation to capture the knowledge, whereas ID3 uses the entropy function to measure the relative importance of attributes, selects attributes to minimize the overall entropy, and generates a decision tree or rule structure to capture the knowledge.

TABLE 1 HERE

Empirical analysis, on the other hand, considers the performance of the resulting model to be 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). They found ID3 to be better than discriminant analysis in loan default and bankruptcy analysis. In a later study, however, Liang (1989a) found that the difference in predictive power between ID3 and discriminant analysis 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) extend 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 sample includes a number of categorical attributes, 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 non-parametric 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 a 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. The major contribution of these comparative studies is to provide insight 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 a pre-processor 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 large number of numerical variables because of its non-parametric nature. Therefore, statistical methods can be applied as a pre-processor to combine the numerical variables into a single attribute. The inductive learning method can then derive a knowledge structure from the original categorical attributes and the numerical attribute generated from the statistical method.

The second approach to integration is to use an inductive learning method as a pre-processor 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 straight-forward 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 advantages 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 discriminant 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, 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.

FIGURE 1 HERE

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 errors. In addition, a proper training sample size may be determined by properly balancing type I and type II 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 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 (Konomenko, Bratko, and Roskar, 1984; Fisher, 1987).

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 four or five significant factors out of a set of twenty 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.

### 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 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 and Ray, 1986; Liang, 1989a, Rendell, 1986), and other statistics such as chi-square or G-statistics may be used to replace the entropy as information measures for constructing the knowledge structure (e.g., Hart, 1984; Mingers 1987a; Race and 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. 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 a F-test may be used to test the signigicance of misclassification.

### 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 to 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, 1987).

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, some empirical evidence about different integrations will be provided.

## 4. EMPIRICAL ANALYSIS

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

The data for empirical comparisons were twelve pairs of bankruptcy data sets originally compiled in Liang (1989a). Each pair of data set included a training and a testing set. Six pairs consisted of thirty cases in the training set and the other six pairs consisted of twenty cases. All testing sets consisted of twenty cases. Each case included a class (i.e., bankrupt or not), three categorical and five numerical variables.

### 4.2. Experimental Procedures

The experiments included two parts: one examined surface integrations and the other investigated deep integration. For surface integration, multiple discriminant analysis (MDA) was selected as the representative of statistical methods and ID3 was chosen as the representative of inductive learning methods. For deep integration, we examined one possibility, which was applying factor analysis to select attributes and then applying ID3 to develop the decision tree. It falls in box C in Figure 1. In other words, for each pair of data sets, three analyses were conducted.

(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. 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 form a simplified 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 attributes. 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.

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

### 4.3 Data Analysis and Discussion

Following the previous procedures, twelve observations were obtained for each situation. These results are compared with those obtained from using MDA or ID3 alone. Table 2 summarizes the prediction accuracy in various settings. It seems that MDA, ID3, and MDA + ID3 have similar performance, whereas ID3 + MDA performs better and FACTOR + ID3 performs worse.

TABLE 2 HERE

In order to confirm the superior performance of ID3 + MDA, an ANOVA test was performed to compare it with MDA. Unfortunately, the results were not significantly (p= .299). Then, we separated the results from 30-case and 20-case training sets and further conducted a pairwise t-test on the data collected from 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). Analysis on the 20-case training sets was insignificant. Although the results are mixed, they do provide encouraging evidence that a proper integration of ID3 and MDA may generate a model more accurate than the individual methods alone.

One reason that may explain the superiority of ID3 and MDA is that ID3 screens out the attributes dominated by others and hence reduces the dependency among attributes. This allows the MDA algorithm to derive a more accurate model. 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.

Another reason that may explain the superiority or inferiority of a 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 3, indicate two findings.

TABLE 3 HERE

First, there exists a weak negative relationship between internal validity and prediction accuracy (p=0.14). However, since this insignificance may be due to the small sample size, it is still worth noting that the increase of internal validity tend to overfit the training data and hence jeopardize the prediction accuracy.

Second, there exists a strong negative relationship between tree complexity and prediction accuracy (p=0.03). 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 can be a very fruitful area for future research.

In addition to the above analysis, the data collected from this research can also be compared to existing literature to derive some interesting observations. In previous research, Liang (1989a) developed a deep integration algorithm called CRIS that applies different statistical methods to derive rules for numerical and categorical data. A rule scheduling approach is then applied to the resulting rules to form a decision structure. It was found that the deep integration algorithm outperformed both MDA ($p < 0.10$) and ID3 ($p < 0.05$). The performance of the ID3 + MDA, however, is comparable to that of CRIS (which was 0.83). Since both ID3 + MDA and the rule scheduling algorithm of CRIS screen out dominanted attributes in the training data set, an interesting implication of this finding is that the selection of attributes, which is often ignored in inductive learning literature, is probably the most important step for improving the performance of inductive learning.

In another previous research, 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) Much more research is necessary to explore promising integration and identify factors affecting the performance of integration.

## 5. CONCLUDING REMARKS

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 focus 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

Braun, H. and Chandler, J. S. (1987), "Predicting Stock Market Behavior Through Rule Induction: An Application of the Learning-From-Example Approach," Decision Sciences, 18:3, pp. 415-429.

Breiman, L., Friedman, J., Olson, R., and Stone, C. (1984), Classification and Regression Trees, Belmont, CA: Wadworth International Group.

Carter, C. and Cattlet, J. (1987), "Assessing Credit Card Applications Using Machine Learning," IEEE Expert, Fall, pp. 71-79.

Chandler, J. C., Liang, T. P., and 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. and 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, pp. 415-424.

Cheng, J., Fayyad, U. M., Iran, K. B., and Qian, Z. (1988), "Improved Decision Trees: A Generalized Version of ID3," Proceedings of the Fifth International Conference on Machine Learning, San Mateo, CA: Morgan Kaufmann, 100-106.

Fisher, D. H. (1987), "Conceptual Clustering, Learning from Examples, and Inference," Proceedings of the Fourth International Conference on Machine Learning, San Mateo, CA: Morgan Kaufmann, pp. 38-49.

Fisher, D. H. and Schlimmer, J. C. (1988), "Concept Simplification and Prediction Accuracy," Proceedings of the Fifth International Conference on Machine Learning, San Mateo, CA: Morgan Kaufmann, 22-28.

Gale, W. A. (1986), Artificial Intelligence and Statistics, ed., Reading, MA: Addison-Wesley.

Geene, D. P. (1987), "Automated Knowledge Acquisition Overcoming the Expert System Bottleneck," Proceedings of the 8th ICIS Conference, Pittsburg, pp. 107-117.

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 University Press.

Hoffman, R. R. (1987), "The Problem of Extracting Knowledge of Experts From the Perspective of Experimental Psychology," AI Magazine, 8:2, pp. 53-67.

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

Kidd, A. L. (1987), Knowledge Acquisition for Expert Systems, New York, Plenum Press.

Konomenko, I., Bratko, I. and Boskar, E. (1984), Experiments in Automatic Learning of Medical Diagnostic Rule, (Technical Report), Ljubljana, Yugoslavia: Josef Stefan Institute.

Lee, J. K., Oh, S. B., and Shim, J. C. (1990), "UNIK-FCST: Knowledge Assisted Adjustment of Statistical Forecasts," Expert Systems with Applications: An International Journal, 1:1, forthcoming.

Lee, W. D. and Ray, S. R. (1986), "Probabilistic Rule Generator," Technical Report No. UIUCDCS-R-86-1263, Department of Computer Science, University of Illinois at Urbana-Champaign.

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.

Liang, T. P. and Yu, C. J (1989), "A Methodological Note on Examining Product Characteristics and Foreign Market Entry Strategies," Proceedings of the Second International Conference on Comparative Management, Kaohsiung, Taiwan, pp. 317-321.

Messier, W. F., Jr. and Hanson, J. V. (1988), "Inducing Rules For Expert Systems Development: An Example Using Default and Bankruptcy Data," Management Science, 34:12, pp. 1403-1415.

Michalski, R. S. and Stepp, R. (1982), "Revealing Conceptual Structure in Data by Inductive Learning," J. E. Hayes, D. Michie, and Y. H. Pao (eds.), Machine Intelligence 10, 173-196.

Minger, J. (1987a), "Expert Systems -- Rule Induction with Statistical Data," Journal of the Operational Research Society, 38:1, pp. 39-47.

Mingers, J. (1987), "Rule Induction with Statistical Data -- A Comparison with Multiple Regression," Journal of the Operational Research Society, 38:4, pp. 347-351.

Mingers, J. (1989), "An Empirical Comparison of Selection Measures for Decision-Tree Induction," Machine Learning, 3, pp. 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, pp. 253-279.

Phelps, R. I. and Musgrove, P. B. (1986), "Artificial Intelligence Approaches in Statistics," in Gale, W. A. (ed.), Artificial Intelligence and Statistics, Reading, MA: Addison-Wesley, pp. 159-171.

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, 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 and T. M. Mitchell (eds.), Machine Learning: An Artificial Intelligence Approach, Los Altos, CA: Morgan Kaufmann, pp. 463-482.

Quinlan, J. R. (1986), "Induction of Decision Trees," Machine Learning, 1:1, pp. 81-106.

Quinlan, J. R. (1987a), "Simplifying Decision Trees," Int'l J. of Man-Machine Studies, 27, pp. 221-234.

Quinlan, J. R. (1987b), "Decision Trees as Probabilistic Classifiers," Proceedings of the Fourth International Workshop on Machine Learning, Los Altos, CA: Morgan Kaufmann, pp. 31-37.

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. and Thomas, R. C. (1988), "Rule Induction in Investment Appraisal," Journal of the Operational Research Society, 39:12, pp. 1113-1123.

Rendell, L. (1986), "Induction, of and by Probability," in L.N. Kanal and J.F. Lemmer (eds.), Uncertainty in Artificial Intelligence, North-Holland, pp. 429-443.

Tu, P. (1989), "Toward an Intelligent Classification-Tree Approach to Problem-solving," Unpublished Ph. D. Dissertation, University of Illinois at Urbana-Champaign.

Utgoff, P. E. (1988), "ID5: An Incremental ID3," Proceedings of the Fifth International Conference on Machine Learning, San Mateo, CA: Morgan Kaufmann, 107-120.

Wirth, J. and Catlett, J. (1988), "Experiments on the Costs and Benefits of Windowing in ID3," Proceedings of the Fifth International Conference on Machine Learning, San Mateo, CA: Morgan Kaufmann, 87-99.

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

Figure 1. A Framework for Deep Integration

(1) Major assumptions

<u>Discriminant Analysis</u> (DA)

- Data population is multivariate normal distribution
- No perfect correlation among independent attributes
- Equal covariance matrices for classes

<u>ID3 algorithm</u>

- No conflict in the training data set

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

| | DA | 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)

Table 1. Comparison of discriminant analysis and ID3

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

Table 2. Prediction Accuracy Under Various Settings

(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 | | | | |

Table 3. Average Performance and Correlation Analysis