

UNIVERSITY OF
ILLINOIS LIBRARY
AT URBANA-CHAMPAIGN
BOOKSTACKS

CENTRAL CIRCULATION AND BOOKSTACKS

The person borrowing this material is responsible for its renewal or return before the **Latest Date** stamped below. **You may be charged a minimum fee of \$75.00 for each non-returned or lost item.**

Theft, mutilation, or defacement of library materials can be causes for student disciplinary action. All materials owned by the University of Illinois Library are the property of the State of Illinois and are protected by Article 16B of Illinois Criminal Law and Procedure.

TO RENEW, CALL (217) 333-8400.
University of Illinois Library at Urbana-Champaign

JAN 04 2000

NOV 16 1999

When renewing by phone, write new due date
below previous due date.

L162

330
B385
1534 COPY 2

STX

BEBR
FACULTY WORKING
PAPER NO. 89-1534

A Composite Approach to
Inducing Knowledge for
Expert Systems Design

THE LIBRARY OF THE

FEB 23 1989

LIBRARY OF THE
FEB 23 1989

Ting-peng Liang



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

BEBR

FACULTY WORKING PAPER NO. 89-1534

College of Commerce and Business Administration


University of Illinois at Urbana-Champaign

February 1989

A Composite Approach to Inducing Knowledge
for Expert Systems Design

Ting-peng Liang, Assistant Professor
Department of Accountancy

This research was supported by a summer research grant from Department of Accountancy, the University of Illinois at Urbana-Champaign. The author thanks James C. McKeown for providing the bankruptcy data and Ingoo Han for running ACLS and SAS packages.



Digitized by the Internet Archive
in 2011 with funding from
University of Illinois Urbana-Champaign

<http://www.archive.org/details/compositeapproac1534lian>

A COMPOSITE APPROACH TO INDUCING KNOWLEDGE FOR EXPERT SYSTEMS DESIGN

ABSTRACT

Knowledge acquisition is a bottleneck for expert system design. This paper presents an approach that automatically induces rules from data for developing expert systems. First, motivation for developing such a rule induction method is described. Then, three major components of the mechanism are discussed. They include a hypothesis generator, a probability calculator, and a rule scheduler. Finally, to evaluate the performance of this approach, an empirical study that compares it with the ID-3 method and discriminant analysis is presented. The results indicate that the approach outperforms both ID-3 and discriminant analysis in analyzing a set of bankruptcy data.

KEY WORDS: Knowledge Acquisition, Rule Induction, Expert Systems, Artificial Intelligence,

1. Introduction

Expert systems (ES) designed to support or replace human experts have drawn considerable attention in the past several years. Business applications have been reported in areas such as accounting (Connell 1987; Dungan and Chandler 1985; Hansen and Messier 1986; Steinbart 1987), finance (Duchessi and Belardo 1987; Kastner, Apte, et al. 1986), manufacturing (Brumi, Elia, and Laface 1986; Kanet and Adelsberger 1987), marketing (Steinberg and Plank 1987), taxation (Michaelson and Messier 1987; Shpilberg and Graham 1986), and others (Blanning 1985; Malmberg, et al 1987; Sathi, Morton, and Roth 1986). In general, evidences indicate that, under certain circumstances, expert systems outperform human experts (e.g., Yu, et al. 1979) and can be used as valuable decision aids (Liang 1988; Turban and Watkins 1986).

The process of developing an ES includes acquiring knowledge from human experts, representing and organizing the knowledge in production rules, storing the rules in a knowledge base, and then applying a deductive inference mechanism (usually called the inference engine) to the knowledge base for decision making. For most systems, the knowledge acquisition stage plays a key role in determining the quality of the resulting system (Buchanan, Barstow, et al. 1983; Denning 1986; Duda and Shortliffe, 1983; Freiling et al. 1986).

A knowledge acquisition process usually involves eliciting, analyzing, and interpreting the knowledge human experts use in solving a particular problem, and then transforming this knowledge into a proper representation. There are at least two approaches for knowledge acquisition.

First, knowledge engineers use techniques such as structured interviews and protocol analyses to elicit knowledge from human experts (called domain experts). The domain experts formulate their knowledge and the knowledge engineers encode this

knowledge for use by the system (see Kidd [1987] for an introduction to these techniques). A major problem with this approach is that human experts frequently have difficulty in articulating their knowledge accurately (Dreyfus and Dreyfus 1986; Hoffman 1987). In addition, it is time-consuming and expensive.

Instead of acquiring rules directly from domain experts, a second approach takes advantage of inductive inference mechanisms that induce decision rules from data. In the process, knowledge engineers collect data from previous decisions, identify key attributes (variables) with the help of domain experts, and then use an inductive program to construct a set of rules for decision making. The core of this approach is an inductive algorithm that accepts a set of data as inputs and produces "If-Then" rules capable of interpreting the data set. Compared to the first approach, the inductive knowledge acquisition (also called rule induction, inductive learning or learning from examples) generates more consistent rules (Braun and Chandler 1987) and the knowledge engineering process is more efficient. In addition, it requires less involvement of domain experts, which implies fast prototyping and cost savings.

The key to a successful inductive knowledge acquisition is the power of the rule induction program. Induction is an inferential process that develops a structure from instances (Holland, et al. 1986). It has been a standard methodology in business research for a long time. For example, most statistical methods such as regression analysis, discriminant analysis, Probit and Logit are inductive in nature. Rule induction mechanisms are different from statistical methods in two ways. First, the resulting structure is a set of "If-Then" rules rather than mathematical equations. Second, the rule induction algorithm may be based on criteria different from sample mean and variance.

Quinlan's ID-3¹, a popular rule induction algorithm, for instance, uses entropy to measure the information content of each attribute and then derives rules by a repetitive decomposition process that minimizes the overall entropy (Quinlan 1979). Although recent research findings indicate that rules generated by this approach outperform both expert judgments and models derived from statistical discriminant analysis in stock market prediction (Braun and Chandler 1987), loan default, and bankruptcy analysis (Messier and Hansen 1988), the algorithm has several limitations. First, since it uses a repetitive decomposition process, real numbers must be converted to integers. This may reduce the accuracy of the results. Second, the repetitive decomposition process is inefficient when the sample size is large. Third, the entropy does not consider the distribution of data and hence is difficult to assess the probabilities associated with rules. Finally, a single algorithm is used to process both nominal (also called categorical, e.g., male and female) and non-nominal (e.g., financial ratios) attributes with completely different properties.

In order to alleviate these shortcomings, a new approach to inducing rules for expert systems design, called a composite rule induction system (CRIS), is presented in this article. The approach assesses probabilities for rules and applies different methods to handle nominal and non-nominal attributes. It is different from the existing ones in the following aspects. First, instead of using a single measure such as entropy to handle both nominal and non-nominal attributes, it uses a cross-tabular approach to process nominal attributes and a statistical inference approach to handle non-nominal attributes. Second, instead of adopting a repetitive decomposition process, it uses a rule scheduling mechanism to determine the relative importances of the candidate rules and

¹ID-3 stands for interactive dichotomizer 3. It originated from Hunt, Marin and Stone's (1966) work on CLS (concept learning system) and is one of the most popular mechanisms in business applications. A good introduction to the algorithm can be found in Braun and Chandler (1987), Messier and Hansen (1988), Thompson and Thompson (1985), and is hence omitted here.

to select the rule set accordingly. Third, it uses sample distributions to infer the population for non-nominal attributes and then estimates the probabilities associated with rules accordingly. In the remainder of the paper, CRIS will be discussed in detail and illustrated with a set of bankruptcy data. Empirical results comparing CRIS with the existing ID-3 method and the statistical discriminant analysis will also be analyzed.

2. CRIS: A Composite Rule Induction System

The goal of a rule induction algorithm is to construct an optimal structure² from a data set, which can interpret the behavior of the input data set and facilitate decision making when a new case is encountered. Similar to the requirements of discriminant analysis, the input data set for rule induction includes a number of cases, each of which has values for a dependent attribute and several independent attributes affecting the dependent one. The dependent attribute usually is nominal or ordinal, such as bankrupt or non-bankrupt and bull market or bear market. The independent attributes can be nominal or non-nominal. The resulting structure is composed of rules in the following format:

If $(X \alpha V)$ Then $(Y \beta C)$ With (Probability γ P)

Where: X = a certain independent attribute,
V = a hurdle value of the attribute,
Y = the dependent attribute,
C = a value of the dependent attribute,
P = a probability value,
 α , β , and γ = relational operators;
 $\alpha, \beta, \gamma \in \{=, \geq, \leq, >, <\}$

Example: If humidity \geq 0.95 Then weather = raining With prob \geq 0.8.

In order to generate rules, therefore, a rule induction mechanism must determine (1) the independent attribute to be considered, (2) the hurdle value of the independent

²By an optimum structure, we mean that the rule set correctly classifies the maximum number of cases in the input data set (internal validity). If the input data set is a true representative of the problem, then the resulting structure will also be optimum when new cases are encountered.

attribute, (3) the corresponding value of the dependent attribute, (4) the probability associated with the rule, and (5) three relational operators. Furthermore, in order to organize rules into a structure, the rule induction mechanism must be able to differentiate the interpretative power of different rules and select an optimum set of rules. In CRIS, these functions are provided by three major components:

- (1) A hypothesis generator that determines the proper relationship between independent and dependent attributes;
- (2) A probability calculator that determines the probability associated with each rule; and
- (3) A rule scheduler that determines how candidate rules should be organized to form a structure.

The interaction of the first two components generates candidate rules and the third component organizes the rules into a structure. They are discussed below.

2.1 Hypothesis Generation

The first step for CRIS to induce a decision rule is to generate hypotheses concerning possible rules for interpreting the input data. A hypothesis is a preliminary "If-Then" rule whose probability is to be determined by the probability calculator and whose interpretative power is to be determined by the rule scheduler. The primary purpose of this stage is to identify causal relationships between dependent and independent attributes. These relationships provide a basis for rule construction. Since different measurement scales of independent attributes have different characteristics, CRIS uses two different methods to generate hypotheses for nominal and non-nominal attributes.

2.1.1 Nominal attributes

For nominal attributes, the values are simply arbitrary identifications of different properties. Their mean and variance have no actual meaning. The attribute "bankruptcy", for example, may have values 1 (yes) and 0 (no). An average value of

0.5 has no meaning in this case. Therefore, CRIS adopts a cross-tabular approach to determine the relationship between nominal attributes and the dependent attribute. The approach includes three steps:

1. For each nominal attribute, classify all cases in the input data set by their attribute values v_i ($i = 1 \dots m$) and dependent attribute values c_j ($j = 1 \dots n$), and then count the number of cases (f_{ij}) in each combination. The result of this step is an occurrence frequency table:

		Y		
		c_1	c_j	c_n
X	v_1	f_{11}	\dots	f_{1n}
	v_i	\dots	f_{ij}	\dots
	v_m	f_{m1}	\dots	f_{mn}

2. For each $X = v_k$ ($k = 1 \dots m$), select a $Y = c_s$, where $f_{ks} = \max \{f_{kj} \mid j = 1 \dots n\}$, to formulate the hypothesis, "If $X = v_k$ Then $Y = c_s$." Since attribute X has m levels ($k = 1, \dots, m$), the total number of hypotheses to be generated for the attribute is m .

3. Repeat steps 1 and 2 until all nominal attributes are processed.

[Example] In order to illustrate the CRIS mechanism, a set of bankruptcy data shown in Table 1 is used as an example. The Sample data set includes 20 cases and 5 attributes. $V1$ is the dependent attribute indicating whether a firm went bankrupt. $V2$ is a nominal attribute indicating auditor's opinion on the previous year's financial statements; $V3$, $V4$, and $V5$ are financial ratios of a firm.

TABLE 1 HERE

After obtaining the data, the first step for CRIS to analyze V2 is to generate the occurrence frequency table. The result is shown in the following table:

		V1	
		0	1
V2	0	10	6
	1	0	4

It is obvious that, in the table, $10 > 6$ for $V2 = 0$ and $4 > 0$ for $V2 = 1$. Therefore, the following two hypotheses can be formulated for V2:

H1: If $V2 = 0$ Then $V1 = 0$.

H2: If $V2 = 1$ Then $V1 = 1$.

2.1.2 Non-nominal attributes

For non-nominal attributes, sample mean and variance provide valuable information about the population and hence are useful for hypothesis formulation. A basic assumption for processing non-nominal attributes is that, in order for an attribute to be able to differentiate cases in different classes (i.e., cases having different dependent attribute values), the population distribution of the attribute must be different for different classes. In other words, different classes of cases are drawn from different populations. Otherwise, a classification will not be possible. For example, if the current ratio (V5 in the example) is to be used to differentiate bankrupt firms from non-bankrupt ones, then the mean and variance of the the bankrupt firms must be different from those of the non-bankrupt.

Based on this assumption, we can formulate hypotheses to determine proper value ranges for each class. If the attribute of a case has a value within the range of a certain class, then we assume the case to be of that class. Figure 1 shows the basic concept.

FIGURE 1 HERE

For classes 1 and 2 (i.e., $Y = c_1$ and $Y = c_2$), their distributions of attribute X are $[\mu_1, \sigma_1^2]$ and $[\mu_2, \sigma_2^2]$, respectively³. Since \bar{X} and S^2 are unbiased estimators for μ and σ^2 , they are used to substitute μ and σ^2 . In order to differentiate these two classes, we first find a value X_c where cases are equally likely to be classified as either class. This value is called the cut between these two classes, which means that if the attribute value of a case is higher (lower) than the cut, then the case is likely to be of the class with the higher (lower) mean. By assuming that attribute values in both classes are normally distributed, the cut can be calculated by the following equation (see Appendix 1 for its derivation).

$$X_c = \frac{S_1 \bar{X}_2 + S_2 \bar{X}_1}{S_1 + S_2} \quad (1)$$

The cut provides a basic hurdle value for hypothesis formulation. If $\bar{X}_2 \geq \bar{X}_1$, for instance, then two hypotheses can be formulated:

- (1) If $X \geq X_c$ Then $Y = c_2$.
- (2) If $X < X_c$ Then $Y = c_1$.

Although these hypotheses can be directly used for classification, the accuracy is frequently lower than the desirable level. In order to ensure the quality of the resulting model, therefore, hypotheses with higher classification accuracy must be developed. In other words, the hurdle value needs to be higher than X_c for hypothesis (1) and lower than X_c for hypothesis (2). Because the sample means and variances do not provide adequate information to determine the optimal hurdle value, the approach adopted by

³ μ and σ^2 stand for population mean and variance; whereas \bar{X} and S^2 stand for sample mean and variance.

CRIS is to generate several values for each attribute to formulate multiple hypotheses and then choose the best one by the rule scheduler.

One way to find hurdle values with higher accuracy is to control the probability that a case falls in a particular class. The rationale for this approach is that the lower the probability that a case belongs to a certain class, the higher the probability it will belong to other classes. In Figure 2, for example, $X_1(0.90)$ is the 90th percentile of X_1 , which indicates that if the attribute value of a case is greater than $X_1(0.90)$, then the chance that the case falls into class 1 is less than 10%. Therefore, replacing X_c in hypothesis (1) with $X_1(0.90)$ will increase the accuracy of the hypothesis⁴. For the same reason, replacing X_c in hypothesis (2) with the 10th percentile of X_2 , $X_2(0.10)$, will increase the accuracy of hypothesis (2). In addition to $X_1(0.90)$ and $X_2(0.10)$, more hurdle values such as $X_1(0.80)$ and $X_2(0.20)$ may also be generated for hypothesis formulation by specifying the desired probability.

FIGURE 2 HERE

Hurdle values identified by the above approach usually increase the classification accuracy for one class at the price of the other. Therefore, only one of the two potential hypotheses is useful. For example, $X_1(0.90)$ can be used to replace X_c in hypothesis (1) to improve accuracy but not in hypothesis (2). Equations (2) and (3) show how hurdle values can be calculated from sample mean, variance, and a specified probability, P . The $z(P)$ in the equation is the z -value at probability P of a standard normal distribution.⁵ Equation (2) applies to the class with the lower mean; whereas Equation

⁴In order to determine the actual probability that the case with attribute value greater than $X_1(0.90)$ falls into class 1, we also need to consider the probability that the case falls into class 2, and then make adjustments accordingly (see Section 2.2 for details). Therefore, the resulting probability may not be exactly 0.90 or 0.10.

(3) applies to the one with the higher mean.

$$X_i(P) = \bar{X}_i + z(P) * S_i \quad (2)$$

$$X_i(1 - P) = \bar{X}_i + z(1 - P) * S_i \quad (3)$$

According to the previous discussion, procedures for hypothesis formulation for non-nominal attributes can be summarized as follows:

1. For the attribute to be analyzed, calculate the mean and variance for each class.
2. Calculate the cut, X_C , to generate two basic hypotheses.
3. Specify the desired probabilities and, then, generate more hypotheses based on the hurdle values calculated by Equations (2) and (3).
4. Repeat steps 1 to 3 until all non-nominal variables have been processed.

[Example] The above procedures allow V3, V4, and V5 in the bankruptcy example to be analyzed. First, for each attribute, sample means and variances of bankrupt firms ($V1 = 1$) and non-bankrupt firms ($V1 = 0$) are calculated separately. Then, the cut values are calculated from sample means and variances. Finally, by specifying the desired probabilities for hypothesis formulation, say 90% and 85%, hurdle values, $X_1(0.90)$, $X_2(0.10)$, $X_1(0.85)$ and $X_2(0.15)$, can be calculated. Table 2 shows the results of these three steps.

TABLE 2 HERE

Based on the data in Table 2, the following hypotheses are formulated:

V3 (Net income/total assets)

- H3: If $V3 \geq 0.0128$ Then $V1 = 0$
- H4: If $V3 < 0.0128$ Then $V1 = 1$
- H5: If $V3 \geq 0.0327$ Then $V1 = 0$
- H6: If $V3 \geq 0.0535$ Then $V1 = 0$
- H7: If $V3 < -0.0051$ Then $V1 = 1$
- H8: If $V3 < -0.0239$ Then $V1 = 1$

⁵When the sample size is small, the z-value, $z(P)$, can be replaced by a t-value, $t(df, P)$, where df = degree of freedom.

V4 (Current assets/total assets)

H9: If $V4 \geq 0.4688$ Then $V1 = 1$
H10: If $V4 < 0.4688$ Then $V1 = 0$
H11: If $V4 \geq 0.5842$ Then $V1 = 1$
H12: If $V4 \geq 0.6281$ Then $V1 = 1$
H13: If $V4 < 0.1609$ Then $V1 = 0$
H14: If $V4 < 0.0438$ Then $V1 = 0$

V5 (Current assets/current liabilities)

H15: If $V5 \geq 1.8881$ Then $V1 = 0$
H16: If $V5 < 1.8881$ Then $V1 = 1$
H17: If $V5 \geq 2.5220$ Then $V1 = 0$
H18: If $V5 \geq 2.7759$ Then $V1 = 0$
H19: If $V5 < 1.3964$ Then $V1 = 1$
H20: If $V5 < 1.1994$ Then $V1 = 1$

2.2 Probability Assessment

After a hypothesis is generated, the probability calculator determines its probability. This probability is a conditional probability indicating the likelihood that the conclusion is true if the condition of the hypothesis is met. In CRIS, the probability is determined by bayesian calculus.

For a problem with n classes, c_1, \dots, c_n , the probability of a "greater-than" hypothesis, "If $X \geq v$ Then $Y = c_k$," is the conditional probability, $P(Y = c_k \mid X \geq v)$, which can be calculated from the prior probability of the class and other conditional probabilities. Two kinds of information usually are available from the input data: (1) the prior probability of class i , $P(Y = c_i)$, where $i = 1 \dots n$, and (2) the conditional probability that, given the class i , the attribute value falls in a certain range, $P(X \geq v \mid Y = c_i)$. These two kinds of probabilities allow the desired probability to be calculated by the following equation derived from the Bayesian Theorem:

$$P(Y = c_k \mid X \geq v) = \frac{P(Y = c_k) * P(X \geq v \mid Y = c_k)}{\sum_{i=1}^n P(Y = c_i) * P(X \geq v \mid Y = c_i)} \quad (4)$$

2.2.1 Nominal Attributes

For nominal attributes, the conditional probability in a situation is equal to its occurrence frequency divided by the total number of occurrences. Since both the numerator and denominator are divided by the same constant (i.e., total number of occurrence), Equation (4) can be simplified as follows (f_{vk} stands for the frequency in the situation where $X = v$ and $Y = c_k$):

$$P = \frac{f_{vk} * P(Y = c_k)}{\sum_{i=1}^n f_{vi} * P(Y = c_i)} \quad (5)$$

[Example] Assuming that the prior probability is 0.5 for either class in the bankruptcy example, then the probabilities associated with H1 and H2 can be assessed as 0.625 (10/16) and 1.0 (4/4), respectively.

2.2.2 Non-nominal Attributes

For non-nominal attributes, the conditional probability $P(X \geq v \mid Y = c_i)$ is determined by the distribution of X for class i ($i = 1 \dots n$). Assuming that the mean and standard deviation of the distribution is \bar{X}_i and S_i , then the probability $P(X \geq v \mid Y = c_i) = 1 - P(z = \frac{v - \bar{X}_i}{S_i})$. Hence, Equation (4) can be transformed to:

$$P(Y = c_k \mid X \geq v) = \frac{P(Y = c_k) * (1 - P(z = \frac{v - \bar{X}_k}{S_k}))}{\sum_{i=1}^n P(Y = c_i) * (1 - P(z = \frac{v - \bar{X}_i}{S_i}))} \quad (6)$$

Similarly, the equation for calculating the probability associated with a less-than hypothesis, "If $X \leq v$ Then $Y = c_k$," is:

$$P(Y = c_k \mid X \leq v) = \frac{P(Y = c_k) * P(z = \frac{v - \bar{X}_k}{S_k})}{\sum_{i=1}^n P(Y = c_i) * P(z = \frac{v - \bar{X}_i}{S_i})} \quad (7)$$

[Example] Assuming that the prior probability of bankruptcy or non-bankruptcy is 0.5, then the probability associated with hypotheses H3 to H20 can be assessed. The results are shown in Table 3. For example, the probability of hypothesis H6, "If $V3 \geq 0.0535$ Then $V1 = 0$," is calculated as follows (since the sample size is small, t-values are used to replace the z-values in Equation (6)):

$$P(V1 = 0) = 0.5; \quad P(V1 = 1) = 0.5;$$

$$P(V3 \geq 0.0535 \mid V1 = 0) = 1 - P(t(9, \frac{0.0535 - 0.0679}{0.0664})) = 0.58;$$

$$P(V3 \geq 0.0535 \mid V1 = 1) = 1 - P(t(9, \frac{0.0535 - (-0.0483)}{0.0736})) = 0.10.$$

Therefore,

$$P(V1 = 0 \mid V3 \geq 0.0535) = \frac{0.58}{0.10 + 0.58} = 0.85.$$

TABLE 3 HERE

Because the sample means and variances of different classes may be different significantly, it is possible that the assessed probability for a certain hypothesis is lower than that of the cut hypothesis. In this case, the hypothesis needs to be modified. For example, the probabilities associated with hypotheses H13 and H14 are 0.32 and 0.17, respectively (Table 3). These indicate that it would be more appropriate to hypothesize that $V1 = 1$ when $V4$ is less than 0.1609 or 0.0438. The probabilities of the new hypotheses are 0.68 and 0.83, respectively.

2.3 Structure Construction

A hypothesis, along with its associated probability, is called a candidate rule. For example, "If $V3 \geq 0.0535$ Then $V1 = 0$ With prob ≥ 0.85 ," is a candidate rule. General guidelines for determining the relational operators α , β , and γ for a candidate

rule are: (1) γ is "=" when α is "="; (2) γ is " \geq " when α is otherwise⁶; and (3) β usually is "=" if the dependent attribute is nominal. Appendix 2 lists the candidate rules generated from the bankruptcy data.

Candidate rules are the basic elements of the knowledge base of an expert system. Because more than one candidate rule is generated for each attribute in the previous process, these rules may be redundant or inconsistent. Additionally, these rules are generated based on information concerning a single attribute. Therefore, a mechanism that evaluates the relative importance of these candidate rules and forms a structure to correctly classify a maximum number of cases is necessary.

Unlike the ID-3 algorithm that selects attributes based on their entropy values, the rule scheduler of CRIS examines the extent to which these rules cover the cases in the input file and then uses a heuristic to schedule them based on their saliency. The saliency of a candidate rule is defined as the difference between the number of cases correctly covered and those incorrectly interpreted by the rule. These numbers are called the hit value and miss value of the rule, respectively. The cases used for determining the saliency of a rule are called training cases⁷. The resulting structure is a decision tree with rules as its nodes. The construction process includes:

1. Determination of rule saliency. Apply all rules to the training cases to determine their hit and miss values.

2. Selection of a rule. The rules generated from cut values (called cut rules) and high accuracy rules (called regular rules) have different properties. The former provides an equal-likelihood split between classes, whereas the latter specifies hurdle values for higher accuracy in classifying a certain class. Therefore, the heuristic for rule

⁶In practice, γ usually is "=", which means that the probability of the rule is at least equal to the specified value. This simplifies the representation of rules.

⁷Training cases are all input cases at the beginning, but are reduced gradually when more and more of them are covered by the rules already scheduled.

scheduling includes two steps. First, the regular rules are selected to interpret as many training cases as possible. Then, the cut rules are applied to cover the remainder in order to guarantee the completeness of the resulting structure⁸. Guidelines for rule selection are:

- 2.1. If there are rules whose miss values are zero and hit values are positive, then select the one with the highest hit value;
- 2.2. If all rules have positive miss values, then calculate the saliency for each rule by deducting its miss value from its hit value and select the one with the highest positive saliency value.
- 2.3. If more than one rule has the same saliency value, then choose the one with the highest probability.
- 2.4. If more than one rule has the same saliency value and probability, then choose the one associated with the most significant attribute. The significance of an attribute is measured by the following formula. The higher the value is, the more significant the attribute is.

$$\text{Significance} = \frac{\sum_{i=1}^n (\bar{X}_i - \bar{X})}{\sum_{i=1}^n \sqrt{\frac{S_i^2}{n_i}}} \quad (8)$$

Where: \bar{X}_i = mean of attribute X for class i;

\bar{X} = overall mean of attribute X;

S_i^2 = variance of attribute X for class i;

n_i = number of cases for class i; and

i = number of classes in the data set.

3. Redefinition of the Training Cases. The selected rule splits the original set of training cases into two subsets: cases covered by the rule (both correctly and incorrectly) and the remainder.

- 3.1. The covered set: If all cases covered by the rule are correctly interpreted, then add the rule to the final structure and stop processing this subset. Otherwise, add the rule to the structure, assign the cases covered by the rule to be the new training set, and then go to step 1 for further analysis.

⁸The completeness of a structure means that the structure is capable of covering all cases. Since the cut rules are paired, they guarantee that if a case fails to meet the condition of at least one regular rule, it will be covered by one of the cut rules.

- 3.2. The remainder: If no case is left after applying a rule, then keep the existing training set and go to step 5 to find a pair of cut rules for completeness. Otherwise, assign the remainder to be the new training set and go to step1.

4. Iteration of the process. Repeat steps 1 to 3 for the regular rules until the termination conditions stated in 3.1 and 3.2 are met or no regular rules that have positive saliency value exist.

5. Application of cut rules. The cut rules are used when no regular rule is available for further classifying the training set. The procedures are the same as applying the regular rules except that the cut rules must be applied in pair and hence their saliency value is the sum of their individual values. It is possible that more than one set of cut rules are applied to interpret a training set, as long as the number of cases correctly interpreted increases. The whole process stops when further improvement is impossible.

[Example] Following the procedures, we can generate a rule structure for interpreting the bankruptcy data. First, hits and misses of the candidate rules, as shown in Table 4, can be counted. The numbers in the hit and miss columns are case ID's shown in Table 1. By comparing the values in the iteration-1 column, rule R6 that hits six cases and misses none has the highest saliency value and is selected (Step 2.1).

TABLE 4 HERE

Rule R6 splits the training cases into two sets: [1,2,5,6,7,8,9] and [3,4,10,11,12,13,14,15,16,17,18,19,20]. Since the rules covered no miss, there is no need to further analyze the set [1,2,5,6,7,8,9] (Step 3.1) and the remainder is assigned as the new training cases. The new training set allows the hit and miss values to be updated, as shown in the iteration-2 column in Table 4. Based on the updated hit and miss values in iteration-2 and iteration-3 in Table 4, rules R2 and R8 are chosen subsequently.

Rule R8 splits the training cases into [10,14,19,20] and [3,4,12,16,17]. The first set contains one misclassified case and hence needs to be further studied. Further examination of the rules, however, shows that no rule can correctly interpret case #10 without introducing more misses (e.g., R1 covers #10, but will also wrongly cover #19 and #20). Therefore, no further exploration is possible and the branch stops with one misclassification (Step 4).

For the remaining cases, [3,4,12,16,17], rule R12 covers case #16, which reduces the uninterpreted cases to [3,4,12,17]. At this point, no regular rule with positive saliency value exists (see the iteration-5 column, only R19 has one hit, but it also has two misses). Therefore, the cut rules are applied to finalize the construction (Step 5). Among three pairs of cut rules, rules R3 and R4 that hit three cases and miss only one is selected because it has the highest combined saliency value (see iteration 6 in Table 4). Therefore, the resulting structure is as follows:

	If $V3 \geq 0.0535$ Then $V1 = 0$ With prob = 0.85;
Else	If $V2 = 1$ Then $V1 = 1$ With prob = 1.0;
Else	If $V3 < -0.0239$ Then $V1 = 1$ With prob ≥ 0.86 ;
Else	If $V4 \geq 0.6281$ Then $V1 = 1$ With prob ≥ 0.83 ;
Else	If $V3 \geq 0.0128$ Then $V1 = 0$ With prob ≥ 0.79 ;
Else	If $V3 < 0.0128$ Then $V1 = 1$ With prob ≥ 0.79 .

Figure 3 shows the resulting decision structure graphically. In this example, two cases, [10, 12], are misclassified. In other words, the resulting structure correctly classifies 90% (18/20) of the cases. Please note that CRIS allows conflicting cases in the training set and hence does not require that the resulting structure correctly classify all cases. In addition, the accuracy will be different when the resulting structure is applied to predict new cases.

FIGURE 3 HERE

In summary, this section presents the CRIS mechanism that induces rules for

classification from data. The induction process includes the following procedures. First, a set of data containing a nominal dependent attribute and several independent attributes are entered. Then, hypotheses are generated by the hypothesis generator of the system. Based on different properties of nominal and non-nominal attributes, different algorithms are developed for hypothesis generation. Third, the hypotheses are converted to candidate rules by assessing their probabilities and making necessary modification. Finally, the resulting candidate rules are evaluated and selected to form a decision structure that can interpret the existing cases and facilitate future decision making.

3. An Empirical Evaluation

A question concerning the CRIS mechanism is how good is the new approach. In order to understand the performance of the mechanism, a preliminary study was conducted to compare CRIS, ID-3, and discriminant analysis. Theoretically, these three approaches have different assumptions on data distribution, use different criteria to evaluate the relative importance of attributes, and generate different models from data, as summarized in Table 5. It is reasonable to assume that they have different performance in solving different types of problems.

TABLE 5 HERE

Since the ID-3 algorithm is based on an exhaustive decomposition process, unless there are conflicts in the data set, it always classifies all training cases correctly. Comparison on internal validity, i.e., the extent to which the cases in the training data set are correctly interpreted, is meaningless. Therefore, the empirical evaluation focuses on the predictive validity, i.e., the accuracy of the resulting models in predicting decisions in other contexts or in hold-out samples (Messier and Hansen, 1988).

3.1 Data and Procedures

The data used for the empirical comparison were a set of bankruptcy data⁹. It contained fifty cases. Each case included four nominal and five non-nominal attributes as defined in the following. X9 is the dependent attribute indicating the outcome of the case.

X1 = consistency exception opinion, 0 = no and 1 = yes;
 X2 = subject-to opinion, 0 = no and 1 = yes;
 X3 = going-concern opinion, 0 = no and 1 = yea;
 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;
 X9 = bankrupt, 0 = no, 1 = yes.

Twelve experiments as described below were conducted. In each experiment, the data set was randomly divided into a training set and an testing set. The training set contained cases used for inducing the model; whereas the testing set contained the hold-out cases for evaluating the predictive validity of the resulting model. For each pair of data sets, all three methods were applied to derive models from the training set. The induced models were then used to predict the cases in the testing set separately. The accuracy of a model was measured by the number of cases correctly predicted by the model divided by the total number of cases in the testing set.

Twelve observations were obtained for each method. The tools used for running the ID-3 algorithm and discriminant analysis were ACLS (Analog Concept Learn System)¹⁰ and the DISCRIM procedure in the SAS package, respectively. The sample size of the training sets had two different levels for examining whether different sample sizes may have effect on prediction accuracy. Six of them had 20 cases and six of them had 30 cases. All testing sets included 20 cases.

⁹The bankruptcy data was obtained from James C. McKeown.

¹⁰ACLS is an implementation of the ID-3 algorithm. See Braun and Chandler (1987) for details.

3.2 Data Analysis

It was not surprising that different methods generated different models from the same set of data. Figure 4 shows an example. These different models were also found having different prediction accuracy.

FIGURE 4 HERE

The average prediction accuracy over twelve experiments was 82.9% for CRIS, 77.1% for ACLS, and 75.8% for discriminant analysis. The results of a two-way ANOVA, as shown in Table 6(a), indicated that the effect of method used for deriving the model was significant at 10% level. Tables 6(b) and 6(c) show the results of comparing CRIS with ACLS and comparing CRIS with discriminant analysis directly. In the former case, CRIS outperformed ACLS at a significance level of 10%. In the latter case, CRIS outperformed discriminant analysis at a significance level of 5%. The difference between ACLS and discriminant analysis was not significant.

TABLE 6 HERE

A major reason for CRIS outperforming ACLS and discriminant analysis is that it takes into consideration the characteristics of nominal and non-nominal attributes and handles them differently. The ID-3 algorithm implemented in ACLS can handle nominal attribute easily, but it fails to consider the nature of interval attributes. Discriminant analysis, on the other hand, treats all attributes as normally distributed and is, hence, difficult to process nominal attributes properly. Therefore, handling different kinds of attributes differently gives CRIS an edge in solving problems involving both nominal and non-nominal attributes. Another possible reason is that the rule

scheduling method of CRIS is more tolerant to attribute correlation. In the rule selection process, cases already covered by previous rules are not considered in selecting the next rule. This can significantly reduce the effect of attribute correlation.

At this point, it is too early to conclude that CRIS is better than ID-3 and discriminant analysis. More empirical studies are needed. However, this preliminary analysis does provide encouraging evidence for pursuing this new rule induction approach.

4. Concluding Remarks

This article presents a composite approach for inducing rules from data. It can be used to acquire knowledge for developing expert systems. The major features that make it different from existing approaches are (1) it uses different techniques to generate hypotheses for nominal and non-nominal attributes; (2) it uses sample distribution (for non-nominal attributes) and frequency table (for nominal attributes) approaches to estimate the probabilities associated with rules; and (3) it uses a rule scheduling technique to determine the relative importance of different attributes and to construct the optimum rule structure. The results of the empirical study indicate that the new approach outperforms the traditional rule induction algorithm ID-3 and the statistical discriminant analysis in prediction accuracy.

Given the increased use of expert systems in various business areas, this work is a step toward improving the knowledge acquisition process for expert system design. Further research including conducting empirical and theoretical investigations to compare different induction approaches and developing new rule induction methods will help us find better knowledge acquisition tools and, more importantly, know which tool is better under what situations.

5. References

- Blanning, R. W. (1985), "Expert Systems for Management: Research and Development," Journal of Information Science, 9, pp. 153-162.
- 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.
- Bruni, G., Elia, A. and Laface, P. (1986), "A Rule-based System to Schedule Production," IEEE Computer, 19:7, pp. 32-40.
- Buchanan, B. G., Barstow, D., et al. (1983), "Constructing an Expert System," in F. Hayes-Roth, D. A. Waterman and D. B. Lenat (eds.) Building Expert Systems, Reading, MA: Addison-Wesley, pp. 127-167.
- Connell, N. A. D. (1987), "Expert Systems in Accountancy: A Review of Some Recent Applications," Accounting and Business Research, 17: 67, pp. 221-233.
- Denning, P. J. (1986), "Towards a Science of Expert Systems," IEEE Expert, 1:2, pp. 80-83.
- Dreyfus, H. L. and Dreyfus, F. E. (1986), "Why Expert Systems Do Not Exhibit Expertise?" IEEE Expert, 1:2, pp. 86-90.
- Duchessi, P. and Belardo, S. (1987), "Lending Analysis Support System (LASS): An Application of a Knowledge-based System to Support Commercial Loan Analysis," IEEE Transactions on Systems, Man, and Cybernetics, SMC-17:4, pp. 608-616.
- Duda, R. O. and Shortliffe, E. H. (1983), "Expert Systems Research," Science, 220, pp. 261-276.
- Dungan, C. W. and Chandler, J. S. (1985), "Auditor: A Micro-computer-based Expert Systems to Support Auditors in the Field," Expert Systems, 2:4, pp. 210-221.
- Freiling, M., et al. (1986), "Starting a Knowledge Engineering Project: A Step-by-step Approach," AI Magazine, 6:3, pp. 150-163.
- Hansen, J. V. and Messier, W. F., Jr. (1986), "A Preliminary Investigation of EDP-XPRT," Auditing: A Journal of Practice & Theory, 6:1, pp. 109-123.
- Hoffman, R. R. (1987), "The Problem of Extracting Knowledge of Experts From the Perspective of Experimental Psychology," AI Magazine, 8:2, pp. 53-67.
- Holland, J. H., et al. (1986), Induction: Processes of Inference, Learning, and Discovery, Cambridge, MA: MIT Press.
- Hunt, E. B., Marin, J. and Stone, P. T. (1966), Experiments in Induction, New York: Academic Press.
- Kanet, J. J. and Adelsberger, H. H. (1987), "Expert Systems in Production Scheduling,"

European Journal of Operational Research, 29, pp. 51-59.

Kastner, J., Apte, C., et al. (1986), "A Knowledge-based Consultant for Financial Marketing," AI Magazine, 7:5, pp. 71-79.

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

Liang, T. P. (1988), "Expert Systems as Decision Aids: Issues and Strategies," The Journal of Information Systems, 2:2, pp. 41-50.

Malmborg, C. J., Agee, M. H., Simons, G. R., and Choudhry, J. V. (1987), "A Prototype Expert System for Industrial Truck Type Selection," Industrial Engineer, vol. 19, pp. 58-64.

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

Michaelsen, R. H. and Messier, W. F., Jr. (1987), "Expert Systems in Taxation," The Journal of the American Taxation Association, (Spring), pp. 7-21.

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.

Sathi, A., Morton, T. E. and Roth, S. F. (1986), "Callisto: An Intelligent Project Management System," AI Magazine, 7:5, pp. 34-52.

Shpilberg, D. and Graham, L. E. (1986), "Developing ExpertTAX (sm): An Expert System for Corporate Tax Accrual and Planning," Auditing: A Journal of Practice & Theory, 6:1, pp. 75-94.

Steinbart, P. J. (1987), "The Construction of a Rule-based Expert Systems as a Method for Studying Materiality Judgments," The Accounting Review, vol LXII, no. 1, pp. 97-116.

Steinberg, M. and Plank, R. E. (1987), "Expert Systems: The Integrative Sales Management Tool of the Future," Journal of the Academy of Marketing Science, 15:2, pp. 55-62.

Thompson, B. and Thompson, W. (1986), Finding Rules in Data," Byte, (November), pp. 149-158.

Turban, E. and Watkins, P. R. (1986), "Integrating Expert Systems and Decision Support Systems," MIS Quarterly, 10:2, pp.121-136.

Yu, V. L., et al. (1979), "Antimicrobial Selection by Computers: A Blinded Evaluation to Infectious Disease Experts," Journal of the American Medical Association, 242:21, pp. 1279-1282.

Appendix 1: Derivation of Equation (1)

Assuming that $\bar{X}_2 > \bar{X}_1$, then, for a case C with a value of V, its probabilities of belonging to class 1 and class 2 are $P(z_1)$ and $P(z_2)$, respectively; where z distribution is the standard normal distribution and z_1 and z_2 are two z-values with respect to classes 1 and 2.

$$z_1 = \frac{V - \bar{X}_1}{S_1}$$

$$z_2 = \frac{\bar{X}_2 - V}{S_2}$$

Since at the cut point, X_c , the probability are equal at both sides. In other words, $z_1 = z_2$. That is,

$$\frac{X_c - \bar{X}_1}{S_1} = \frac{\bar{X}_2 - X_c}{S_2}$$

$$S_1(\bar{X}_2 - X_c) - S_2(X_c - \bar{X}_1) = 0$$

Therefore,

$$X_c = \frac{S_1\bar{X}_2 + S_2\bar{X}_1}{S_1 + S_2}$$

Appendix 2: Candidate Rules for the Bankruptcy Example

R1: If $V2 = 0$ Then $V1 = 0$ With prob = 0.625.

R2: If $V2 = 1$ Then $V1 = 1$ With prob = 1.0.

R3: If $V3 \geq 0.0128$ Then $V1 = 0$ With prob ≥ 0.79 .

R4: If $V3 < 0.0128$ Then $V1 = 1$ With prob ≥ 0.79

R5: If $V3 \geq 0.0327$ Then $V1 = 0$ With prob ≥ 0.82 .

R6: If $V3 \geq 0.0535$ Then $V1 = 0$ With prob ≥ 0.85 .

R7: If $V3 < -0.0051$ Then $V1 = 1$ With prob ≥ 0.82 .

R8: If $V3 < -0.0239$ Then $V1 = 1$ With prob ≥ 0.86 .

R9: If $V4 \geq 0.4688$ Then $V1 = 1$ With prob ≥ 0.68 .

R10: If $V4 < 0.4688$ Then $V1 = 0$ With prob ≥ 0.68 .

R11: If $V4 \geq 0.5842$ Then $V1 = 1$ With prob ≥ 0.78 .

R12: If $V4 \geq 0.6281$ Then $V1 = 1$ With prob ≥ 0.83 .

R13: If $V4 < 0.1609$ Then $V1 = 1$ With prob ≥ 0.68 .

R14: If $V4 < 0.0438$ Then $V1 = 1$ With prob ≥ 0.83 .

R15: If $V5 \geq 1.8881$ Then $V1 = 0$ With prob ≥ 0.65 .

R16: If $V5 < 1.8881$ Then $V1 = 1$ With prob ≥ 0.65 .

R17: If $V5 \geq 2.5220$ Then $V1 = 0$ With prob ≥ 0.67 .

R18: If $V5 \geq 2.7759$ Then $V1 = 0$ With prob ≥ 0.67 .

R19: If $V5 < 1.3964$ Then $V1 = 1$ With prob ≥ 0.79 .

R20: If $V5 < 1.1994$ Then $V1 = 1$ With prob ≥ 0.86 .

Among them, R3, R4, R9, R10, R15, and R16 are cut rules.

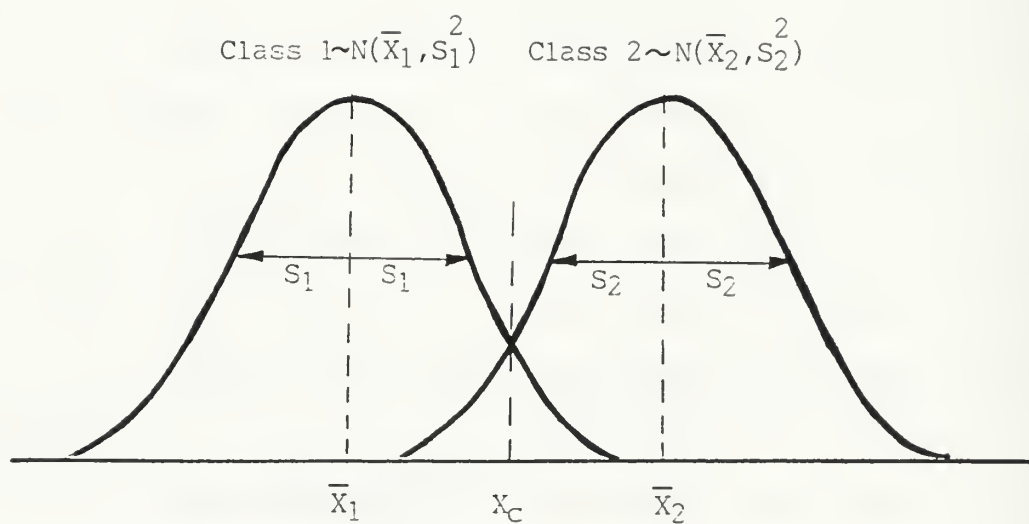


Figure 1. Basic Concept of Classification

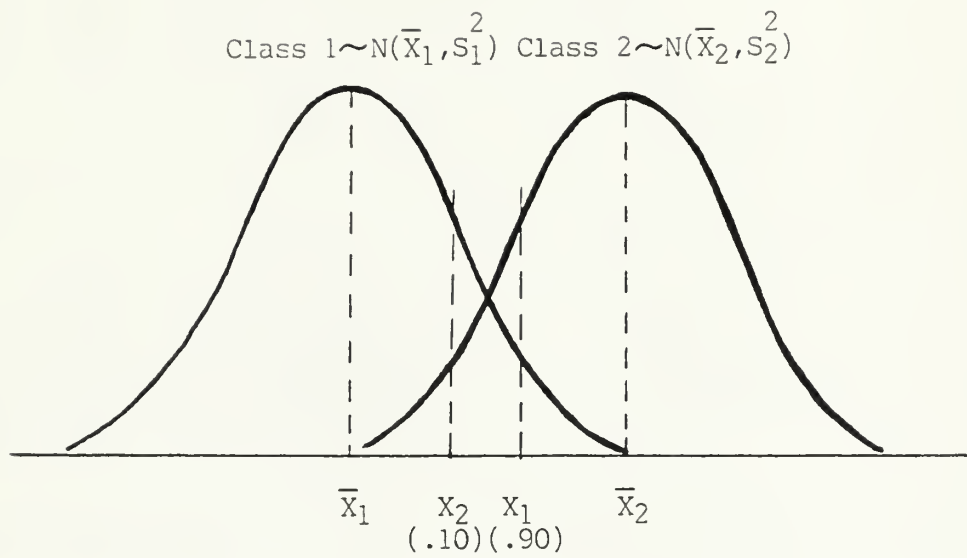


Figure 2. Hurdle Values with Higher Accuracy

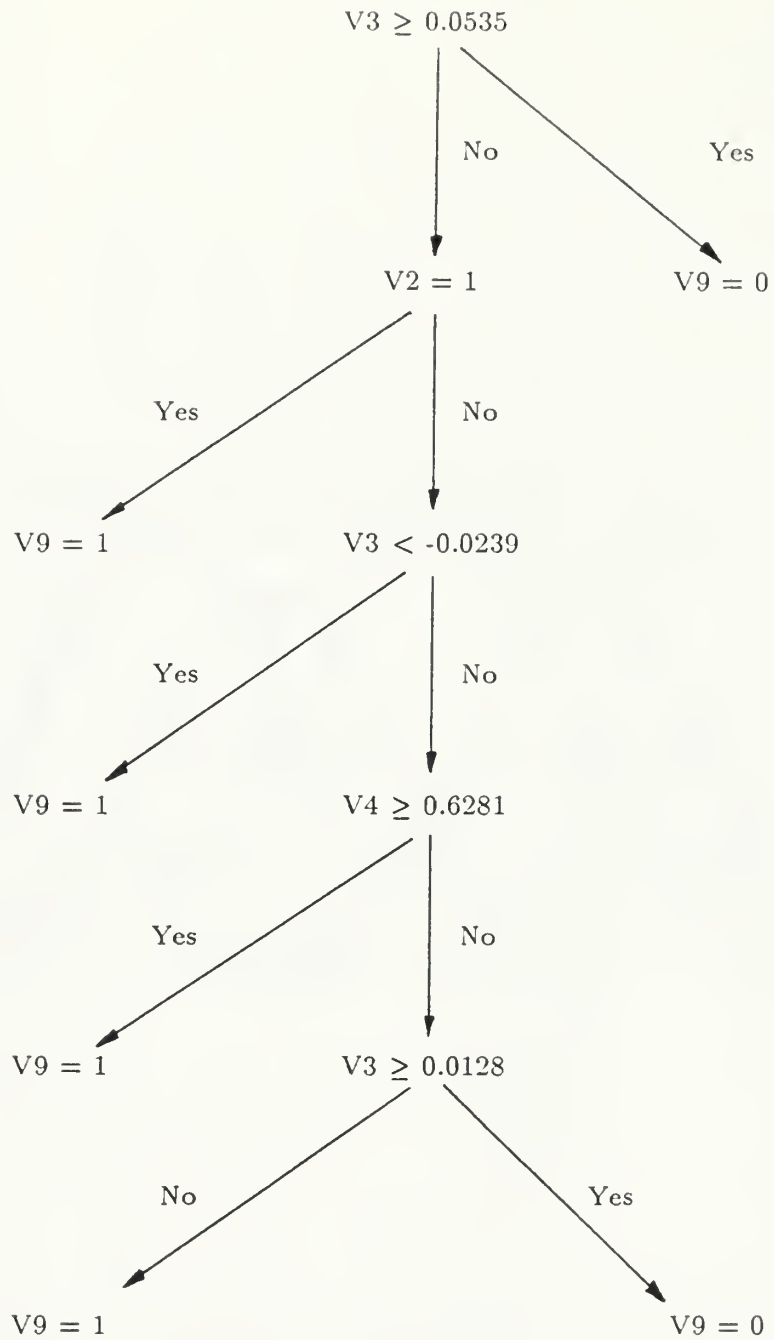


Figure 3. The Resulting Decision Tree for the Example

(1) CRIS Model

If $X_4 \leq 0.0026$ Then $X_9 = 1$ With prob ≥ 0.88
Else If $X_6 \geq 2.3522$ Then $X_9 = 0$ With prob ≥ 0.75
Else If $X_3 = 1$ Then $X_9 = 1$ With prob $= 1.0$
Else If $X_2 = 1$ Then $X_9 = 1$ With prob $= 1.0$
Else If $X_4 \geq 0.0622$ Then $X_9 = 0$ With prob ≥ 0.80
Else If $X_4 \geq 0.0297$ Then $X_9 = 0$ With prob ≥ 0.80
Else If $X_8 \geq 0.0387$ Then $X_9 = 0$ With prob ≥ 0.62
Else If $X_8 < 0.0387$ Then $X_9 = 1$ With prob ≥ 0.62

(2) ACLS Model

If $X_4 < 0.512$ Then $X_9 = 1$
Else If $X_6 \geq 1.289$ Then $X_9 = 0$
Else If $X_4 \geq 0.0517$ Then $X_9 = 1$
Else If $X_4 < 0.0517$ Then $X_9 = 0$

(3) Discriminant Analysis Model

$$\begin{aligned} X_9(0) &= -6.31 + 7.51X_1 - 3.16X_2 + 4.04X_3 + 3.47X_4 + 2.72X_5 + 2.41X_6 \\ &\quad + 9.32X_7 + 0.8X_8 \\ X_9(1) &= -8.99 + 8.54X_1 + 2.05X_2 + 6.05X_3 - 10.56X_4 + 6.66X_5 + 1.19X_6 \\ &\quad + 10.88X_7 + 3.01X_8 \end{aligned}$$

Figure 4. Three Models Generated by Different Methods From The Same Data Set

ID	V1	V2	V3	V4	V5
1	0	0	0.1113	0.3880	1.9862
2	0	0	0.0537	0.2087	1.6827
3	0	0	0.0178	0.4831	1.3325
4	0	0	0.0136	0.2014	0.7537
5	0	0	0.0975	0.4730	2.7911
6	0	0	0.1237	0.2982	2.8921
7	0	0	0.0539	0.5189	2.5375
8	0	0	0.1921	0.4395	2.9946
9	0	0	0.0777	0.3689	2.5478
10	0	0	-0.0621	0.7563	2.1047
11	1	1	-0.0656	1.5557	2.9152
12	1	0	0.0189	0.2409	1.2443
13	1	1	-0.1953	0.0113	0.0015
14	1	0	-0.1356	0.4794	2.4443
15	1	1	-0.0038	0.6956	1.9334
16	1	0	0.0118	0.9479	0.1530
17	1	0	0.0029	0.3398	1.8195
18	1	1	0.0448	0.8165	1.4482
19	1	0	-0.1046	0.7100	1.1111
20	1	0	-0.0569	0.3652	2.2768

Where: V1: bankruptcy; 0 = no; 1 = yes;

V2: auditor's opinion; 0 = unqualified, 1 = qualified opinion;

V3: the ratio of net income/total assets;

V4: the ratio of current assets/total assets;

V5: the ratio of current assets/current liabilities.

Table 1. A Set of Bankruptcy Data

Attribute	Class	Mean	St. Dev.	X_c	$X_i(P1)^{1,3}$	$X_i(P2)^{2,3}$
V3	V1 = 0	0.0679	0.0664	0.0128	0.0535	0.0327
	V1 = 1	-0.0483	0.0736	0.0128	-0.0239	-0.0051
V4	V1 = 0	0.4136	0.1551	0.4688	0.0438	0.1609
	V1 = 1	0.6162	0.4139	0.4688	0.6281	0.5842
V5	V1 = 0	2.1622	0.6962	1.8881	2.7759	2.5220
	V1 = 1	1.5347	0.8975	1.8881	1.1994	1.3964

- Notes: 1. The values are $X_i(0.90)$ for the class with higher mean (e.g., the first row in V3 is $X_{V3=1}(0.90)$ and $X_i(0.10)$ for the class with lower mean (e.g., the second row in V3 is $X_{V3=0}(0.10)$).
2. The values are $X_i(0.85)$ for the class with higher mean (e.g., the first row in V3 is $X_{V3=1}(0.85)$ and $X_i(0.15)$ for the class with lower mean (e.g., the second row in V3 is $X_{V3=0}(0.15)$).
3. Since the sample size was 10 for each class, t-values were used in calculating these hurdle values.

Table 2. Analysis of Three Non-nominal Attributes

Rule ID	Probability
H3	0.79
H4	0.79
H5	0.82
H6	0.85
H7	0.82
H8	0.86
H9	0.68
H10	0.68
H11	0.78
H12	0.83
H13	0.32
H14	0.17
H15	0.65
H16	0.65
H17	0.67
H18	0.67
H19	0.79
H20	0.86

Table 3. Probabilities of Rules R3 to R20

ID ¹	P ²	Hits ³	Misses ³	Type ⁴	Iter-1 ⁵		Iter-2		Iter-3		Iter-4		Iter-5		Iter-6	
					H	M	H	M	H	M	H	M	H	M	H	M
R1	0.62	1,2,3,4,5,6,7,8,9,10	12,14,16,17, 19,20	R	10	6	3	6	3	6	2	3	0	0		
R2	1.0	11,13,15,18	—	R	4	0	4	0	0	0						
R3	0.5	1,2,3,4,5,6,7,8,9	12,18	C	9	2	2	2	2	1	2	1	2	1	3	1
R4	0.5	11,13,14,15,16,17, 19,20	10	C	8	1	8	1	5	1	2	0	1	0		
R5	0.82	1,2,5,6,7,8,9	18	R	7	1	0	1								
R6	0.85	1,2,5,6,7,8,9	—	R	7	0	0	0								
R7	0.82	11,13,14,19,20	10	R	5	1	5	1	3	1	0	0				
R8	0.86	11,13,14,19,20	10	R	5	1	5	1	3	1	0	0				
R9	0.50	11,14,15,16,18,19	3,5,7,10	C	6	4	6	2	3	2	1	1	0	1	1	3
R10	0.50	1,2,4,6,8,9	12,13,17,20	C	6	4	1	4	1	2	1	2	1	2		
R11	0.78	11,15,16,18,19	10	R	5	1	5	1	2	1	1	0	0	0		
R12	0.83	11,15,16,18,19	10	R	5	1	5	1	2	1	1	0	0	0		
R13	0.68	13	—	R	1	0	1	0	0	0						
R14	0.83	13	—	R	1	0	1	0	0	0						
R15	0.50	1,5,6,7,8,9,10	11,14,15,20	C	7	4	1	4	1	2	0	0			2	2
R16	0.50	12,13,16,17,18,19	2,3,4	C	6	3	6	2	4	2	3	2	2	2		
R17	0.67	5,6,7,8,9	11	R	5	1	0	1								
R18	0.67	5,6,8	11	R	3	1	0	1								
R19	0.79	12,13,16,19	3,4	R	4	2	4	2	3	2	2	2	1	2		
R20	0.86	13,16	4	R	2	1	2	1	1	1	1	1	0	1		

Note: 1. ID = rule identification in Appendix 2.

2. P = probability of the rule.

3. The numbers in the hits and misses columns are the IDs of the training cases.

4. Type = the type of the rule; R = regular rule and C = cut rule.

5. Iter-1 to Iter-6 = iteration 1 to iteration 6; the values shown in H column are the hit values (the number of cases hit by the rules) and in M column are miss values (the number of cases misclassified by the rules).

Table 4. Evaluation of the Candidate Rules

(1) Major assumptions

Discriminant analysis (DA)

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

ID-3 algorithm

- No conflict in the training set

CRIS algorithm

- Non-nominal data are normally distributed for each class

(2) Selection criteria, processes, and resulting models

	DA	ID-3	CRIS
Selection criteria	Covariance matrix	Entropy	Rule saliency
Selection processes	Matrix operations	Repetitive decomposition	Rule scheduling
Resulting models	Linear equations	Rule structure	Rule structure

Table 5. Comparison of CRIS, ID-3 and Discriminant Analysis

(a) Overall

Source	DF	Sum of square	Mean square	F-value
Method	2	0.03431	0.01715	2.779*
Size	1	0.00028	0.00028	0.045
Interaction	2	0.00824	0.00412	0.667
Error	30	0.18500	0.00617	
Total	35	0.22806		

(b) CRIS and ACLS

Source	DF	Sum of square	Mean square	F-value
Method	1	0.02042	0.02042	3.288*
Size	1	0.00375	0.00375	0.604
Interaction	1	0.00167	0.00167	0.269
Error	20	0.12417	0.00621	
Total	23	0.15000		

(c) CRIS and Discriminant Analysis

Source	DF	Sum of square	Mean square	F-value
Method	1	0.03010	0.03010	4.894**
Size	1	0.00010	0.00010	0.016
Interaction	1	0.00844	0.00844	1.372
Error	20	0.12292	0.00615	
Total	23	0.16156		

Notes: "*" indicates significance at least at 10% level;
" **" indicates significance at least at 5% level.

Table 6. Results of F-tests

HECKMAN
BINDERY INC.



JUN 95

Bound - To Please

N MANCHESTER,
INDIANA 46962

UNIVERSITY OF ILLINOIS-URBANA



3 0112 045801468