PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 :		(11) International Publication Number:
G06F 15/40	A2	WO 92/17853
		(43) International Publication Date: 15 October 1992 (15.10.92)

(21) International Application Number: PCT/US92/02757

(22) International Filing Date: 6 April 1992 (06.04.92)

(30) Priority data:
681,115 5 April 1991 (05.04.91) US

(71) Applicant: PATTERN RECOGNITION, L.P. [US/US]; 1890 Maple, Evanston, IL 60201 (US).

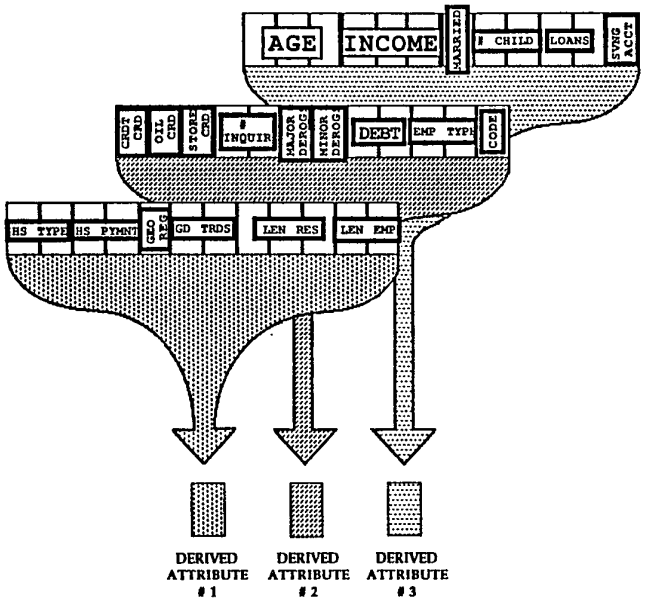
(72) Inventor: FREY, Peter, W. ; 1317 Livingston, Evanston, IL 60201 (US).

(74) Agents: MEYERS, Gerson, E.; Dressler, Goldsmith, Shore, Sutker & Milnamow, Ltd., Two Prudential Plaza, 180 N. Stetson, Suite 4700, Chicago, IL 60601 (US) et al.

(81) Designated States: AT (European patent), AU, BE (European patent), CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, LU (European patent), MC (European patent), NL (European patent), SE (European patent).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: DIRECT DATA BASE ANALYSIS, FORECASTING AND DIAGNOSIS METHOD



(57) Abstract

A method for analyzing records of a data base by selecting a target measure related to a selected outcome, identifying data in known records of the data base for use as predictor variables, grouping selected ones of the predictor variables, producing derived values of the target measure for different combinations of the predictor variables for each group, identifying the derived values for a test record, identifying a selected number of known records that are most similar to the test record with respect to the derived values, identifying the value of the selected outcome of the selected most similar known records, and using that value for predicting a selected outcome for the test record.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	**FI**	Finland	**ML**	Mali
AU	Australia	**FR**	France	**MN**	Mongolia
BB	Barbados	**GA**	Gabon	**MR**	Mauritania
BE	Belgium	**GB**	United Kingdom	**MW**	Malawi
BF	Burkina Faso	**GN**	Guinea	**NL**	Netherlands
BG	Bulgaria	**GR**	Greece	**NO**	Norway
BJ	Benin	**HU**	Hungary	**PL**	Poland
BR	Brazil	**IE**	Ireland	**RO**	Romania
CA	Canada	**IT**	Italy	**RU**	Russian Federation
CF	Central African Republic	**JP**	Japan	**SD**	Sudan
CG	Congo	**KP**	Democratic People's Republic of Korea	**SE**	Sweden
CH	Switzerland			**SN**	Senegal
CI	Côte d'Ivoire	**KR**	Republic of Korea	**SU**	Soviet Union
CM	Cameroon	**LI**	Liechtenstein	**TD**	Chad
CS	Czechoslovakia	**LK**	Sri Lanka	**TG**	Togo
DE	Germany	**LU**	Luxembourg	**US**	United States of America
DK	Denmark	**MC**	Monaco		
ES	Spain	**MG**	Madagascar		

DIRECT DATA BASE ANALYSIS, FORECASTING AND DIAGNOSIS METHOD

Field of the Invention

The present invention relates to databases and more particularly to methods for forecasting and diagnosis based on direct analysis of data in such data bases.

Background

One technique for evaluating data in order to produce useful information is to produce a model of the data and attempt to derive parameters for the model from the data. Desirably, useful information such as forecasts, predictions, and diagnoses can then be derived from the model.

One problem with this approach is that a model is only an approximation of the actual data. Therefore, its value depends on the quality of the approximation. Furthermore, the data employed to create a model often become dated. As a result, a model can age rapidly, and its approximation of current data relationships becomes increasingly less accurate with time. As the assumptions upon which a model is based become less valid, and as the parameters estimated become dated, predictions based on the model can be inaccurate and unreliable. Since data relationships generally change over time, the predictive utility of each specific model deteriorates as the model ages.

Most models assume that predictor variables bear a linear relationship to the outcome being estimated. Furthermore, it is also common to assume that predictor variables combine in an additive fashion. Linear models based on these assumptions have distinct limitations in dealing with higher order interactions. They do not reflect the proper

relationships among the variables in these situations. This reduces the accuracy of their projections.

Models, by their very nature, are based on assumptions relative to the general relationships among the variables. These assumptions are often inconsistent with the properties of the data being examined. In cases where the assumptions are violated, the predictive accuracy is reduced.

Furthermore, the imposition of these assumptions tends to place a limit on the asymptotic degree of accuracy of the predictions independent of the number of observations in the existing data base. Increasing the sample size initially improves predictions because a larger number of observations permits more precise specification of the model parameters. At some point, however, additional data provide no further enhancement. For example, with linear regression or discriminant analyses, it is generally believed that further improvements are not observed when the sample size is larger than about ten thousand.

For several reasons, model building is currently the standard approach for working with large sets of data, in spite of these limitations. Historically, data bases tended to be much smaller than they are today. Model building is an effective technique for deriving useful information within the context of small and medium size data bases.

Furthermore, for many years, the available computational resources were inadequate for the direct analysis of large amounts of data in a reasonable time frame. Direct methods, although discussed from time to time in the academic literature, were simply not cost effective. The rapid evolution of low cost, high-speed, large volume computers may change these limitations. Many

computational methods which were impractical less than a decade ago may become potentially feasible.

It would be desirable to be able to analyze and use large amounts of data to produce information for practical ends. Modeling techniques, however, are often not adequate to provide the desired capabilities.

Many data bases are organized as two-dimensional flat files. In this type of structure, the data is typically organized in rows and columns in which the rows represents individual records in the data base, e.g., persons, households, accounts, or events, and the columns represent fields describing attributes, e.g., age, weight, symptom, or outcomes, e.g., diagnoses, credit risk, which comprise each record. Often information is acquired which can be used to establish a partial record in which the attribute information is present but the outcome information is missing.

In these cases, it may be desirable or necessary to forecast, predict, or diagnose the unknown outcome information for a new record by making use of the other information in the data base. Traditional methods for predicting outcomes have been based on linear regression or discriminant analyses. More recently, other approaches have been employed, such as rule-based expert systems.

An alternative method which, in theory, is appropriate for solving this problem is the nearest neighbor method (see Duda & Hart, Pattern Classification in Scene Analysis. New York: Wiley, 1973). This method is used traditionally to identify the record in the data base which is most similar to the test case or test record, i.e., a new, partial record. The outcome for this "nearest neighbor" is assumed to be the best prediction for the new case. Although this method is straight forward

conceptually, there are often implementation difficulties which limit its use.

The fields in a data base commonly represent information of three different types. Some of the fields represent variables that are boolean, such as, e.g., true-false, yes-no, agree-disagree, like-dislike. Other fields represent variables that are categorical, such as, e.g., marital status -- single, married, separated, divorced, widowed; and employment status -- full-time, part-time, retired, student, unemployed. Still other fields represent variables that are numerical, such as, e.g., annual income, age, months at current job. Determination of the similarity of two records is greatly complicated by these multiple data types.

If none of the fields are numerical, the similarity between any two records might be calculated by counting the number of fields for which the two records have identical values. The nearest neighbor would be the record which has the most fields with values in common with the target record (see e.g., Stanfil & Waltz, Toward memory-based reasoning. Communications of the ACM, 1986, 29, 1213-1228).

If all of the fields are numerical, which is rarely the case, each field can be considered as one dimension of a multi- dimensional hyperspace, and the distance between any two records can be computed by assuming that the hyperspace has Euclidean properties. By computing the distance between the test case and every other case in the data base, the nearest neighbor is identified as the record with the smallest Euclidean distance from the test record.

It would be desirable to be able to analyze and use information in large multiple type data bases to formulate accurate analyses and predictions of

expected outcomes. Existing techniques do not provide the desired capabilities and reliability.

Summary

In accordance with the present invention there is provided a method of organizing large sets of data in a data base to facilitate evaluation of a test record for producing information about the test record. A method incorporating the present invention extracts information from large sets of data which is useful for prediction, forecasting, and diagnosis.

In accordance with the present invention, it is possible to organize selected information from a large data base in a way, for example, which is predictive of a desired outcome, i.e., so the information can be used to predict an expected event or characteristic for a test case directly from the information in the data base. Such methods compare favorably with prior techniques based on model building.

The method incorporating the present invention involves the prediction of an expected outcome based on an analysis of the similarity of a test case or test record in a data base to each of the prior cases, i.e., existing records, which have been stored in the data base. In general, the method incorporating the present invention is an outgrowth and refinement of the classical nearest neighbor method. Methods incorporating the present invention provide techniques for transforming a raw data base into a numerical representation which permits utilization of the power which is inherent in the nearest neighbor approach. Additionally, methods incorporating the present invention, when compared to the classical technique, greatly enhance the effectiveness of the process.

Typically, the various data fields of the records for most large data bases include a mixture

of numerical, boolean, and categorical information. Under these circumstances, the appropriate method for determining similarity becomes problematic. There is no obvious method which is clearly best for measuring inter-record similarity. This is a serious impediment to the application and use of the nearest neighbor approach to real-world data.

Another difficult problem relates to the relative predictive value of the information in different fields of each record. For any given outcome to be predicted, some of the fields may have high predictive value while others may be totally irrelevant. For example, in diagnosing a medical problem, some symptoms (attributes) may be much more important than others in selecting the correct diagnosis (outcome). In forecasting credit risk for a bank, some applicant attributes (e.g., prior credit history, salary level, job security) may be much more important than others (e.g., age, gender, time at current residence) in determining credit risk.

The effectiveness of a procedure for determining the similarity between records would be enhanced by the ability to take into account and reflect the relevance, value or weight of the data of the various fields of the data base in correctly forecasting the desired outcome.

Yet another important aspect of making effective use of the nearest neighbor approach is to determine when and how to aggregate individual fields to form one or more composite traits which can then be employed to determine similarity between records of the data base. In this regard, fields which may provide little useful information when considered by themselves can be useful when evaluated in combination with, or on a relational basis with, other fields.

Yet another difficulty in measuring the similarity between records is that numerical fields are often not scaled in a way which faithfully reflects the relationship between the values of a given trait and the values of the outcome which is being predicted or diagnosed. For example, although credit risk may vary with age, the difference between records due to differences in values of this data field may differ for different values. The risk may be great for certain values, e.g., during certain age ranges such as the younger ages of 18 to 30. On the other hand the difference may be small for other age ranges, e.g., the senior years of 55 to 70.

For example, the similarity on the basis of the age attribute (the difference in age) for two individuals whose ages differ by eight years are different for different age values. Thus, analysis of the data would show that two individuals who are 56 and 64 years of age are considered to be highly similar. In contrast, two individuals who are 20 and 28 years of age would be considered to be somewhat different.

A simple mathematical treatment of these data would conclude that both cases are equally similar, namely eight years difference in age. A method in accordance with the present invention attempts to optimize the predictive validity of a nearest neighbor system by scaling each predictor value such that the numerical values reflect an accurate relationship to the outcome measure. Current nearest neighbor technology does not provide a solution for this problem.

The method incorporating the current invention addresses these problems and thereby is an advance over classic nearest neighbor technology. The solution is based on the use of a recursive binary classification process (cf., Breiman,

Friedman, Olshen, & Stone, Classification and Regression Trees. Monterey, CA:Wadsworth, 1984; Quinlan, I. R. Induction of decision trees. Machine Learning, 1986, 1 81-106). This process has been used by itself to classify or categorize data and is a powerful tool for these purposes.

A method in accordance with the present invention, utilizes numerical scaling of similarities created by use of non-linear mapping functions. Numerical scaling as used in accordance with the present invention combines a number of original variables to produce a derived variable.

In accordance with the current invention, a binary classification tree is used in a novel way as a preparatory step to translate (i.e., map) diverse information in raw data base form into a new, derived representation having equal or equivalent units of measurement which are appropriate for applying the nearest neighbor approach.

The binary classification tree provides a means for mapping the fields in the original records of the data base into a new set of derived fields which can be employed for making determinations of the similarity between records. This mapping process generally involves a reduction in the number of fields contained in each derived record. This results from the elimination of fields which provide no relevant information, and from the combination of two or more original fields into a single new derived field.

In accordance with the present invention, this mapping process involves the following steps:

(1) Select an outcome measure which is identical to or highly related to the primary outcome which will be the focus of the nearest neighbor forecast or diagnosis.

This outcome measure is to be available as one of the fields in the original record layout and is to be numerical or boolean in type (i.e., not categorical).

(2) Identify the fields in the original record layout which can potentially be employed as predictor variables. Partition these fields into groupings such that the original fields making up each group share common properties or provide information which has common characteristics or is related in some fashion. These groupings can be based on statistical measurements, on engineering requirements or restrictions, on suspected cause and effect relationships, or on knowledge available from individuals who have had prior experience with the application in question. The number of such groupings usually varies between two and twelve. The number of original fields forming any one group usually varies between one and eight.

(3) Use the fields within each group as the splitting variables in creating a binary classification tree (c.f., Breiman et al., 1984) with the selected outcome field as the focus of the classification. There are different ways in which a classification tree can be created (c.f., Mingers, I., An empirical comparison of selection measures for decision-tree induction. Machine Learning, 1989, 3, 319-342; and Mingers, I., An empirical comparison of pruning methods for decision tree induction. Machine Learning, 1989, 4, 227-243). Many

of the common procedures would be appropriate for the mapping process herein described.

(4) Once the classification tree has been constructed, each record in the data base can be associated with one, and only one, terminal node within the tree. The outcome value associated with the terminal node, i.e., the average value for the records within that terminal node, is the numerical value assigned to that record for the trait, i.e., the new field, derived from the original fields used to create the classification tree.

(5) Repeat this procedure such that a separate classification tree is created for each grouping, as defined in (2) above, of the fields in the original data base. The new derived representation for each original record has as many fields as the number of classification trees which have been created. The values for these derived fields are the outcome value of the terminal node which is appropriate for the original record, and are in equal or equivalent units of measurement. Each grouping of the original fields has an associated binary classification tree; each tree assigns every record in the original data base to one of its terminal nodes in each derived record of the derived data base. The value of the terminal node becomes the value of the corresponding new field for that record.

The mapping procedure as described above provides a solution for the problems discussed previously. A uniform numerical unit of measurement is derived which is independent of the original field type (boolean, categorical, or numerical). The relative importance of each of the original fields in the original data base is determined and weighted properly by the creation of the binary classification tree. The combination or elimination of fields is also a natural consequence of tree creation. Finally, the scaling of the new derived field values in respect to the target outcome is an integral aspect of the binary classification process.

Thus, the method incorporating the current invention provides an effective procedure for preparing a raw data base for nearest neighbor forecasting. The mapping procedure creates a new, derived representation in which the fields are all numerical and are in equal or equivalent units of measurement, in which the values are scaled properly, and in which each field provides meaningful predictive information. This new representation permits the use of the powerful Minkowski distance metric to determine inter-record similarity.

In accordance with one aspect of the method incorporating the present invention, selected variables are grouped or aggregated for combined processing and analysis. A plurality of such groupings may be utilized, each of which contains information or data derived from fields different from the fields used in other groupings in so far as the desired output is concerned. A grouping of data for analysis to provide an output predictive of one selected outcome could very well be different from a grouping or aggregation of data fields with respect to another outcome. Values of the data in a selected group are categorized with respect to the

selected outcome to produce multiple discrete combinations each having a predictive value for the desired outcome.

In addition to the mapping process which is employed to prepare the information acquired from the original data base, the current invention incorporates procedures which provide refinements of the classical nearest neighbor procedure (c.f., Duda & Hart, 1973). This is to be distinguished from existing procedures, such as those for determining the outcome for the test record by observing the outcome associated with its nearest neighbor in the data base (i.e., the record which is most similar to the test record). A common variation of the procedure is to determine the k nearest neighbors (where k typically ranges between 1 and 20) and designate the outcome for the test case as the statistical average of the outcomes observed for the k nearest neighbors when the outcome is numerical or boolean or as the most frequent value when the outcome is categorical. In both of these circumstances, each record within the subset of the k nearest neighbors has equal weight in determining the outcome for the test record.

The method incorporating the current invention refines the k-nearest neighbor approach in four specific ways:

(1) The number of neighboring records (i.e., the population of voters) which are used or participate in determining the outcome for the test record is greatly expanded. In accordance with the method incorporating the present invention, the number of voters or participating records typically ranges from about 50 to 800.

(2) The number of participating records varies, depending jointly on two independent criteria which are established to determine the eligibility of each potential participating record. Participating records are typically include or are selected from those records (a) within a pre-specified numerical distance of the test record, and (b) having a nearness rank (e.g., such as the 400th closest record) which is less than a pre-specified value. All records which fit these criteria may be included in the population of participating records. Records which do not meet both of the criteria above are not used as participating records, i.e., do not vote.

(3) Each of the participating records has a differential amount of influence on the prediction or diagnosis. The influence of each of the records used is proportional to the similarity of that record to the test record. The more similar records have greater influence. This contrasts with the one record, one vote method of the classical k-nearest neighbor approach.

(4) The voting process can consider the number of eligible participating records and both the central tendency and the variance of the distribution of outcome values for the eligible participating records, i.e., the nearest neighbors, in determining the appropriate outcome for the test record. Prior technology has focused on the central tendency, e.g., the mean, of this distribution.

Thus an applicant for credit might be analyzed on the basis of one or more criteria. For example, credit might be approved or denied on the basis of revenue potential and absence of risk. The mean of the participating records represents the best forecast of revenue potential. The standard deviation of this group represents a forecast of the risk associated with the forecasted revenue potential. A dual-criteria decision rule can be utilized which considers both measures in deciding to approve or deny credit to an applicant.

When the number of participating records is below some predetermined value, the system may respond with the mean outcome value for the entire data base. If the outcome is categorical, the response may be the most common category or alternatively, that there is not enough information to make a choice. When the number of participating records exceeds the pre-specified number, the system reports the mean and standard deviation for numerical or boolean outcome measures and reports the proportion of each category for categorical outcome measures.

Numerous other advantages and features of the present invention will become readily apparent from the following detailed description of the invention and the embodiments thereof, from the claims, and from the accompanying drawings in which the details of the invention are fully and completely disclosed as a part of this specification.

Brief Description Of the Drawings

FIGURES 1 is a logical flow chart illustrating the method of the present invention; and

FIGURES 2, 3 and 4 are diagrams of decision trees produced by analyses of records of a database

in accordance with the method incorporating the present invention.

Detailed Description

While this invention is susceptible of embodiment in many different forms, there is shown in the drawing and will be described herein in detail specific embodiments thereof with the understanding that the present disclosure is to be considered as an exemplification of the principles of the invention and is not intended to limit the invention to the specific embodiment illustrated.

As indicated above, data suitable for use with the method incorporating the present invention may be arranged in databases organized in the form of flat files. As represented in Fig. 1, a flat file data base is a set of records in which each record contains information about, i.e., represents, a single subject of the data base, e.g., a person, household, account, or event. Each of the records typically contains a plurality of fields. The fields in each record represent attributes (i.e., characteristics) of the subject of the record. These attributes can be boolean (e.g., yes, no), categorical (e.g., single, married, separated, divorced, widowed), or numerical variables. In some cases, the value of the attribute will not be known (missing values).

Each record also contains a field that represents a target event which is the object of the forecast, i.e., the desired outcome. The target event or outcome field can also be a boolean, categorical or numerical variable. Some of the records, i.e., the existing set, contain known values for the target event. One or more other records, i.e., the prediction or test record or set, have unknown values for the target event. A method

incorporating the present invention, uses the derived information relating the attributes to the outcome in the training set to "predict" the best value for the outcome for each record in the prediction set.

In accordance with the present invention, if the desired outcome is to predict the probable credit record of an applicant for credit, e.g., based on revenue to be earned, a data base of existing credit records can be analyzed. In one example, such records may represent information about individuals applying to receive a line of credit, often in the form of a credit card or loan, e.g., to banks, retail stores, or oil companies. The credit evaluation of an applicant can be predicted in accordance with the method incorporating the present invention by comparing the attribute information from an individual's application to the corresponding information contained in the records of prior applicants whose credit record is currently known. The similarity between new applicant's attributes and those of prior applicants provides a basis for approving or denying credit. For example, a similarity between the new applicant's attributes and those of prior applicants who have abused their credit privileges would provide a basis for denying credit.

For credit screening, the predictor attributes commonly consist of several numerical variables, several boolean variables, and several categorical variables. Examples of numerical values include monthly income, monthly credit payments, monthly housing payment, months at current job, months at current residence, number of 60-day delinquencies, number of balances past due, and number of recent inquiries to credit bureau. Examples of boolean variables include the existence or absence of a savings account, checking account,

loan account, bank credit card, and oil company card, and whether certain information has been provided, e.g., a job description. Examples of categorical variables include type of housing, source of application, educational background. There are typically twenty to forty fields which provide useful information.

In such an example, an appropriate outcome measure or target event may be the net revenue (positive or negative) to be derived from the applicant if credit is issued. In accordance with a method incorporating the present invention, a subset of the available fields relating to the existence and non-existence of certain financial parameters such as, e.g., bank card, department store card, oil company card, savings account, checking account, as well as housing type, e.g., owns, rents, with parents, can be grouped together.

The original information or data in this subset or group of attribute fields is evaluated to produce derived predictive values relative to the target event. Thus, in accordance with the present invention, the grouping of original data fields is subjected to a binary classification procedure to produce a binary classification tree structure shown in Fig. 2 which segments the database in a useful way.

The top box in Fig. 2 indicates that the entire data base represents 191,293 records of credit applicants which produce an average annual net revenue of -1.28. The first split of the data base is based on the existence or absence of a bank card. These two groups are represented by the two boxes in the second row of Fig. 2. As shown in Fig. 2, there are 106,221 records with no bank card when they applied for credit, and there are 85,072 with one or more bank cards when they applied for credit. The

group without bank credit cards produced an average net revenue of -6.91, while the group with one or more bank cards produce an average net revenue of 5.76.

As further illustrated in Fig. 2, applicants with bank credit cards are further split based on the existence or absence of a checking account, while those with a bank credit card a next split based on ownership of housing. These splits are shown in the third row of Fig. 2. As shown, there are over 32,000 applicants with no bank card and no checking account, producing an average revenue of -14.34; there are over 73,000 applicants with no bank card, but with a checking account producing average revenue of -3.41. There are over 47,00 applicants with a bank card who do not own housing and producing an average revenue of 2.86, while the over 37,000 applicants with a bank card who own housing produce an average income of 5.68.

As shown in Fig. 2, a binary classification tree represents a recursive process which continues to split the subset into groups as long as meaningful splits are possible. In this context, meaningful refers to the creation of two new groups which are significantly different from each other in a statistical sense.

In the example illustrated in Fig. 2, thirty terminal groups were produced, not all at the same level. Each group represents a specific segment of the original data base. The number of records representing individual applicants in each group varies from group to group as does the observed average net revenue produced. Each of the thirty terminal groups can be characterized as follows :Thus, the terminal group in the lower left portion of the Fig. 2 represents 1956 applicants with no bank card or checking account (including no answer), which

rent housing, with no department store card, and without an answer about a savings account. This first group produced an average net revenue of -25.70. The next terminal group is similar, except that the applicants in this group have savings accounts. The second group produces an average net revenue of -20.45.

GROUP	AVERAGE REVENUE	COMBINATION OF ATTRIBUTE DATA
1	-25.70	no bank card, no checking account, housing rented, no department store card, no answer savings accounts
2	-20.45	no bank card, no checking account, housing rented, no department store card, savings account
3	-13.33	no bank card, no checking account, housing rented, department store card
4	-19.81	no bank card, no checking account, housing other than rent, no department store card, no answer savings accounts
5	-11.62	no bank card, no checking account, housing other than rent, no department store card, savings accounts, lives in house
6	- 6.02	no bank card, no checking account, housing other than rent, no department store card, savings accounts, other than house
7	-15.35	no bank card, no checking account, housing other than rent, department store card, no savings accounts
8	- 3.21	no bank card, no checking account, housing other than rent, department store card, savings accounts

GROUP	AVERAGE REVENUE	COMBINATION OF ATTRIBUTE DATA
9	-20.17	no bank card, checking account, rents house, no answer savings accounts, no department store card
10	- 1.35	no bank card, checking account, rents house, no answer savings accounts, department store card
11	-12.15	no bank card, checking account, rents house, answer savings account, no department store card, no savings account
12	- 6.8	no bank card, checking account, rents house, answer savings accounts, no department store card, savings account
13	- 1.29	no bank card, checking account, rents house, answer savings accounts, department store card
14	-15.7	no bank card, checking account, not rent house, no department store card, housing with parents, no savings accounts
15	- 4.67	no bank card, checking account, not rent house, no department store card, housing with parents, savings accounts
16	- 0.51	no bank card, checking account, not rent house, no department store card, housing not with parents
17	- 0.49	no bank card, checking account, not rent house, department store card, housing with parents
18	0.02	no bank card, checking account, not rent house, department store card, not housing with parents, no oil company card
19	4.01	no bank card, checking account, not rent house, department store card, not housing with parents, oil company card

GROUP	AVERAGE REVENUE	COMBINATION OF ATTRIBUTE DATA
20	- 7.16	bank card, not own house, no checking account, no department store card
21	- 1.26	bank card, not own house, no checking account, department store card
22	1.29	bank card, not own house, checking account, no department store card
23	2.36	bank card, not own house, checking account, department store card, no travel/ent. card
24	5.91	bank card, not own house, checking account, department store card, travel/ent. card
25	1.52	bank card, owns house, no department store card, no checking account
26	6.26	bank card, owns house, no department store card, checking account
27	6.97	bank card, owns house, department store card, no answer checking account
28	10.51	bank card, owns house, department store card, answer checking account

The purpose of the binary classification tree is to produce a mapping relationship such that any set of responses for the six predictor items automatically places the applicant into one and only one of the 28 terminal bins. An applicant inherits the value of the derivative data assigned to the bin defined by the applicant's responses. In the example shown in Fig. 2, the bin values represent average net revenue produced by the individuals in the bin.

In the current example, five of the six predictor variables have 3 possible values (yes, no,

no answer) and one (housing type) has 4 possible values (own, rent, live w/parents, other). Therefore, there are a total of 972 ways in which an applicant could respond to the 6 items in the subset (3 x 3 x 3 x 3 x 3 x 4). The binary classification tree maps each of the 972 response patterns into one of the 28 terminal bins and thereby assigns a numerical value for each of the 972 response patterns. As is apparent from Fig.2, however, some of the bins encompass more than one of the possible response patterns.

The analysis along any one branch of such a decision tree may be terminated when certain criteria are no longer met, e.g., the size of the group at the end of the branch falls below a selected value, or the quality of a proposed split does not meet certain criteria. Alternatively, the analysis can be forced beyond such limits if meaningful information can be extracted.

This value provides a derived numerical index reflecting the relative profitability of each person based on the information contained in the six predictor variables. The value represents a transformation from a heterogeneous set of categorical characteristics in the reference data base to a single dimension in which the unit of measurement (revenue) has ratio scale properties. This transformation is significant since it converts heterogeneous data into a homogeneous data in equal or equivalent units of measurement which are suitable for the nearest neighbor algorithm.

These transformed or derived values are appropriate for the determination of similarity. By measuring distance in terms of the difference in revenue produced by two individuals, the new values permit an application of a Minkowski distance metric.

The use of the binary classification tree as described above can be repeated with a different set of predictor variables. Fig. 3 presents a second binary classification tree. In this case, the outcome measure is also net revenue but the splitting variables are orthogonal to, have no overlap with, the variable of the first tree. The variables used in Fig. 3 include the number of good trades (transactions), time at current residence, income, and months at current job. This tree produces a second mapping relationship which results in the assignment of a ratio scale or derived value to each record based on the responses to the second set of predictor variables.

In Fig.3, the variables have been classified in a number of different ways. Thus even though only four attributes are used, several are used more than once as a result of different values for a given attribute value. Thus the terminal bin in the lower right corner of Fig. 3 is based on attributes of more than one good trade, more than two good trades, long time at current residence, income greater than 8, not long time at current job, and income greater than 22. Applicants so classified, produce average revenue of 11.19.

As shown in Fig.4, another group that can be evaluated on a scaler basis for similar information is age. The revenue produced by age can be evaluated, and segregated by age ranges. The benefit of scaling in this regard is, as indicated above, that changes as a function of age may differ for different values of age, i.e., similar age changes may or may not result in similar changes in revenue performance.

Thus as shown in Fig. 4, the average revenue produced by applicants aged 19 and under is -7.97. The average revenue produced by applicants

aged 20 - 28 (nine year range) is -5.31, while applicants aged 29 - 31 (three year range) produce an average income of -2.09. Over the next six years, applicants aged 32 - 37 produced an average income of 1.37 while applicants in the 38 - 48 age bracket (eleven year range) produce an average income of 4.8. Finally, applicants aged 49 and over (a large range) produce average revenue of 6.92.

Such an analysis separates ages, not by some predetermined age model, but from the actual data that indicates the segregation as a result of actual attribute data.

When the assignments based on the third set of variables has been completed, a another set can be chosen and the process repeated. This continues until all of the desired predictor variables have been included into one of the mapping sets.

When this process is complete, the 20 to 40 attributes associated with each application which are predictive of the expected outcome or desired result have been converted into four to six derived attributes based on the mapping sets. The new derived attributes have all of the properties that a Euclidean distance metric requires. Each value is based on the same measuring unit (e.g., in the above example, revenue). Each value has true ratio scale properties. Problematic issues such as variable selection, variable weighting, and variable scaling have been dealt with. Missing values are included since the binary classification tree treats missing values like other values, i.e., groups missing values with other values when they produce similar outcomes and splits them into a separate group when they are associated with unique outcomes.

For the purpose of the forgoing example, it is assumed that the credit application and credit bureau values have been mapped (using the binary

classification tree) into a new set of three attributes. The geometrical representation of our forecasting problem has now been reduced to a 3-dimensional hyperspace. To measure the similarity of two records in this 3-dimensional space, one can apply a standard Minkowski distance metric. In the forgoing example, the exponent for the Minkowski metric is set to equal 2 and thus assume Euclidean properties for measuring distance. The distance between any two records can be determined by measuring the difference in values of each of the newly derived attributes, squaring these difference values, adding the squared values, and taking the square root of the sum. This calculation can be applied to all records in the data base in order to determine for any given test record which of its neighbors are closest (i.e., most similar).

The result of grouping variables in accordance with the present invention is to enable the use of a small number of dimensions, each of which is substantially orthogonal to the others. In other words, groups of information are selected which do not share characteristics with each other and therefore lend themselves to analysis separate from the other groups.

By comparing the values of information in the variables of a record for which it is desired to predict an outcome with respect to a selected result, e.g., revenue, a predictive response can be produced based on the similarities of the values of the test record, the record for which a prediction is sought, and the values produced by the analysis of the data base. Since the comparison is based on the analysis of the actual data, the reliability and accuracy of the response, as compared to existing techniques, can be improved.

In the credit screening example, a subset of records in the data base which are most similar to the test record are identified. This subset is usually about 1/2 of 1% of the records in the entire database. For example, if the database consisted of 100,000 records, the 500 records which are most similar to the target item are identified. The forecast for the new applicant would be based on a statistical analysis of this subset. The mean value would be the best estimate of the expected value for the new applicant. The standard deviation of this subset would provide an estimate of the stability of the expected value. For credit screening, the standard deviation of the subset provides a direct measure of risk. In performing analyses on a large collection of data, it is first appropriate to determine the nature of the information that is desired. The data is then organized or categorized in a plurality of groupings of data, each of which has the capability of providing information with respect to the desired analysis.

Thus, in the example of a collection of data records based on applications for credit, the various information provided by credit applicants can be categorized or grouped into a plurality of groups, each of which consists of categories of data capable of providing information with respect to revenue.

Data respecting certain like information can be aggregated and processed as a unit to provide output in the form of a set of values of various combinations of the data which are predictive and are related to the ultimate question being investigated. Various combinations of data for each aggregation can be evaluated and values produced respecting the value of the combinations as a function of the outcome being processed. Initially there may be a large number of predictive variables which have a

relationship to each other. These are aggregated, and processed together.

In the example described above, the predictive variables of credit cards, e.g., types of credit cards, and numbers of credit cards can be aggregated and processed together to produce a set of values for various combinations of credit cards, each being a scaler value having a number which relates to the answer being sought. Thus, in conjunction with processing of credit applications, each combination of predictors will have a value corresponding to credit worthiness or similar function. In processing the combinations of data, the various predictors are combined in a way that will be the most helpful to producing information with respect to the desired outcome.

Initially the data may be categorized in an effort to split the records substantially equally on either side of the split. Each successive categorization is selected in a way to render the successive level using the predictive values that give the highest quality split. This general rule can be modified so that split is made to achieve the effectiveness of the variable. Thus, for a particular variable the number of records might be small. A split based on that information might occur earlier in the tree in order that the number of records affected by that split have an effect. At the lower end of the tree, a split based on such a small number of records might result in the effect of the particular variable on the decision tree being so minimal as to have no effect at all.

In addition, the split can be further modified in an effort to maintain the size of the groups. Each of the splits are somewhat equal once again to avoid a grouping that is so off center or small as to ultimately be ignored.

Thus, in the evaluation of credit card ownership, the initial split can be taken utilizing a particular type of credit card in which the number of credit card holders is sufficiently small so that if a split was made later on or lower down in the decision tree, the effect of that split would be dissipated. An ultimate decision tree is produced in with various combinations of credit cards having a series of values representative of revenue, as indicated above.

The results of each of the decision tree analyses is a set of data representing a trait plotted with respect to the outcome to be predicted. Although each characteristic or trait is established as a function of a particular set or aggregation of data evaluated and analyzed, the units of the traits as a function of the data analyzed is the same. The similarity between the test case and the individual cases making up the database can be determined based on the similarities between the values of the trait for each set of data analyzed.

The plot of each such trait is often referred to as a dimension of the data base. In accordance with the present invention data in a number of such dimensions is used to identify or characterize each of the records in the database. The location of each known record is thus determined based on the correspondence of each trait of each record to the plotted values of that trait based on the corresponding decision tree analysis. The location of the test case is similarly determined.

The test case is compared to a selected number N of most similar cases from the data base, i.e., to the N nearest neighbors. The degree of similarity or the distance between the test case and the known cases is determined by known computational techniques, such as by computing the "Minkowski

distance" between the test case and each of the other cases in the data base. The Minkowski distance is calculated in accordance with the following formula:

$$D = (w^r + x^r + y^r + z^r + \text{-----})^{1/r}$$

Each of the variables "w", "x", "y" and "z" represents the difference between a value of the test case and the corresponding value of the known record for each plot of a trait as determined by the decision tree analysis. If "r" = 2, the distance is called the Euclidean distance. The number so calculated represents the distance or similarity between the test case and a particular known case in the data base. The higher the number, the less similar or the greater is the distance being calculated.

The prediction for the test case is determined by taking the average of the values of the information being predicted, e.g., credit risk, for the "N" closest neighbors. The accuracy can be improved, as discussed above, by using a weighted average in which the values are weighted as a function of the distance of each known case from the test case.

Thus in accordance with the present invention, data can be evaluated with the purpose of predicting a selected outcome for a test case by aggregating similar data, manipulating the values of traits for combinations of each group of data to produce for each group scaler values in common units, determining the similarity between a test case and prior cases, and providing a prediction based on the central tendency, e.g., the mean value or the mode, of the target outcome for the most similar cases.

In accordance with the present invention, the reliability of the predictive values is improved as compared to predictions based on existing techniques. The predictive values in accordance with

the present invention are determined by comparison with the most similar records, those in the "local neighborhood" rather than on global estimates which occurs when other techniques, such as modeling, are used. The predictive values produced in accordance with the present invention are further enhanced since the test case is compared to the actual data making up the data base, and since data bases are typically updated to incorporate recent records.

From the foregoing, it will be observed that numerous variations and modifications may be effected without departing from the true spirit and scope of the novel concept of the invention. It is to be understood that no limitation with respect to the specific apparatus illustrated herein is intended or should be inferred. It is, of course, intended to cover by the appended claims all such modifications as fall within the scope of the appended claims.

WHAT IS CLAIMED IS:

1. A method for facilitating analysis of a test record and production of information about the test record respecting a selected outcome from data representing attributes of the test record comprising the steps of:

identifying the different attribute data capable of providing information respecting said selected outcome in a plurality of reference records forming a reference database;

selecting from the identified attribute data of the reference records one or more groups of said different attribute data;

producing a set of derived data expressed in equivalent units of measurement from each of said groups of different attribute data, each of said sets of derived data having values derived from different combinations of the different attribute data of the selected group;

said sets of derived data defining a derived record for each of the reference records of the reference database;

identifying values of the derived data for each of the reference records;

identifying values of the derived data for the test record;

identifying a selected number of reference records having values of derived data most similar to the values of the derived data for the test record;

identifying outcome data for the selected number of reference records; and

providing information about the selected outcome for said test record based on the outcome data for said selected number of reference records.

2. A method as claimed in Claim 1 wherein each set of derived data is expressed in units of

measurement having values corresponding to the expected outcome to be predicted.

3. A method as claimed in Claim 1 including the steps of:

identifying for each set of derived data different logical combinations of reference record attribute data; and

determining the value of the derived data in each set for each said different combination.

4. A method as claimed in Claim 3 including the steps of:

determining for each reference record the logical combination of attribute data of that record for each set of derived data, and

identifying for each reference record the value of the derived data in each set corresponding to said logical combination of attribute data.

5. A method as claimed in Claim 4 including the steps of:

determining for the test record the logical combination of attribute data of that record for each set of derived data, and

identifying for the test record the value of the derived data in each set corresponding to said logical combination of attribute data.

6. A method as claimed in Claim 5 including the step of:

comparing the values of the derived data in each set thereof for each reference record with the value of the derived data in the corresponding set for the test record.

7 A method as claimed in Claim 6 including the step of:

determining for compared values of derived data the similarity between the test record and each of the reference records.

8 A method as claimed in Claim 7 including the steps of:

selecting a number of the reference records most similar to the test record;

identifying the selected outcome data for each of the selected number of reference records.

9. A method as claimed in Claim 8 including the step of:

determining an average value of the selected outcome data for said selected reference records.

10. A method as claimed in Claim 9 including the step of:

adjusting the value of the selected outcome data for each of the selected number of reference records as a function of the similarity of each such reference record to the test record; and

determining an average value of the adjusted selected outcome data for said selected reference records.

11. A method for facilitating analysis a plurality of reference records forming a database and production of information respecting a selected outcome from data representing attributes of the reference records comprising the steps of:

identifying the different attribute data capable of providing information respecting said

selected outcome in said plurality of reference records;

selecting from the identified attribute data of the reference records one or more groups of said different attribute data;

producing a set of derived data from each of said groups of different attribute data, each of said sets of derived data having values derived from different combinations of the different attribute data of the selected group;

identifying values of the derived data for each of the reference records;

identifying outcome data for the reference records; and

providing information about the selected outcome based on the outcome data for said reference records.

12. A method as claimed in Claim 11 wherein each set of derived data is expressed in equivalent units of measurement.

13. A method as claimed in Claim 11 wherein each set of derived data is expressed in units of measurement having values corresponding to the expected outcome to be predicted.

14. A method as claimed in Claim 11 wherein said sets of derived data define a derived record for each of the reference records of the database.

15. A method as claimed in Claim 11 including the steps of:

identifying values of the derived data for a test record.

16. A method as claimed in Claim 15 including the steps of:

identifying a selected number of reference records having values of derived data most similar to the values of the derived data for the test record;

17. A method as claimed in Claim 11 including the steps of:

identifying for each set of derived data different logical combinations of reference record attribute data; and

determining the value of the derived data in each set for each said different combination.

18. A method as claimed in Claim 17 including the steps of:

determining for each reference record the logical combination of attribute data of that record for each set of derived data, and

identifying for each reference record the value of the derived data in each set corresponding to said logical combination of attribute data.

19. A method as claimed in Claim 18 including the steps of:

determining for a test record the logical combination of attribute data of that record for each set of derived data, and

identifying for the test record the value of the derived data in each set corresponding to said logical combination of attribute data.

20. A method as claimed in Claim 19 including the step of:

comparing the values of the derived data in each set thereof for each reference record with the value of the derived data in the corresponding set for the test record.

21. A method as claimed in Claim 20 including the step of:

determining for compared values of derived data the similarity between the test record and each of the reference records.

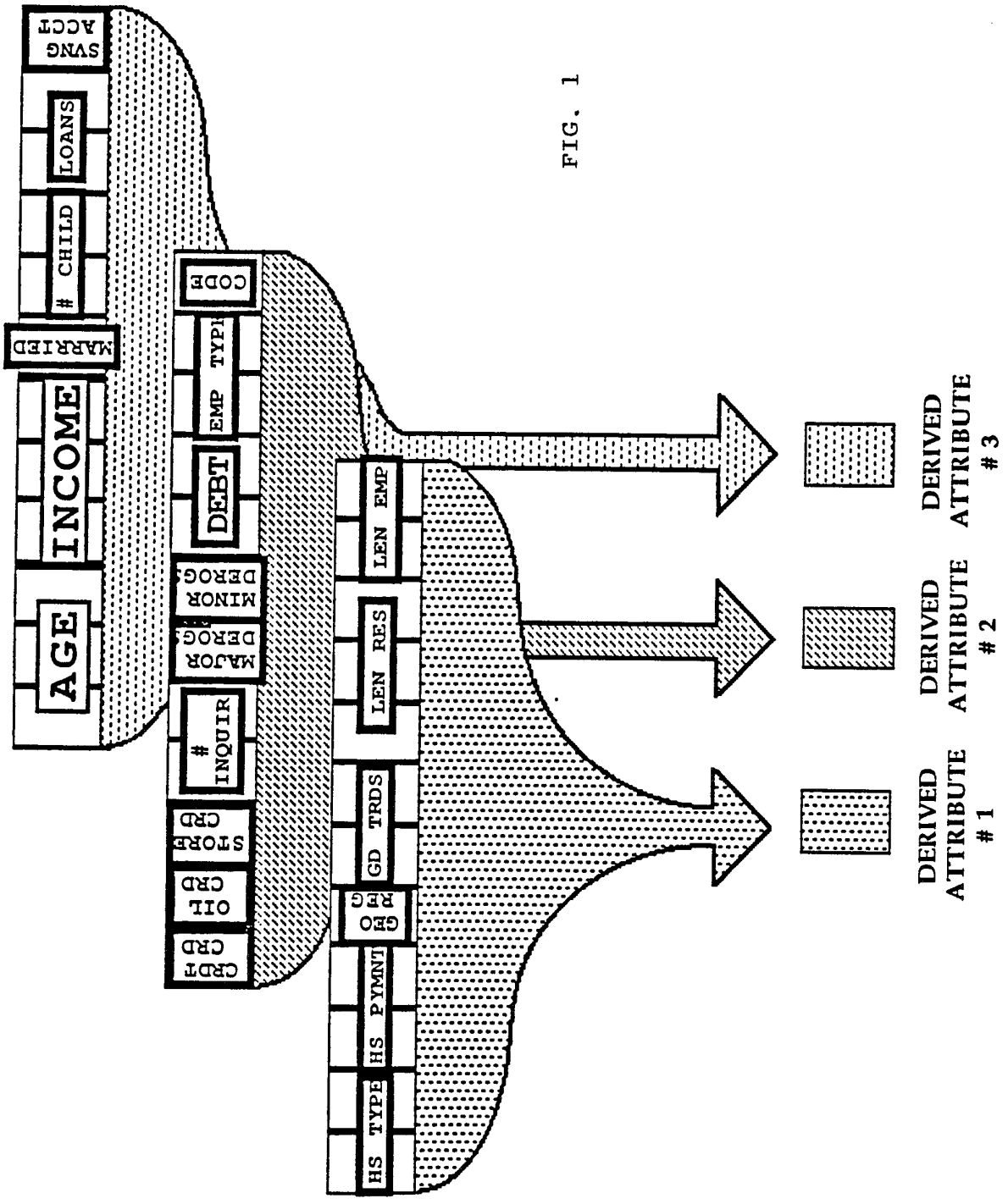

AGE
INCOME
MARRIED
CHILD
LOANS
SVNG ACCT
CODE
EMP TYPE
DEBT
MINOR DEROG
MAJOR DEROG
INQUIR
STORE CRD
OIL CRD
CRDT CRD
HS TYPE
HS PYMNT
GEO REG
GD TRDS
LEN RES
LEN EMP
DERIVED ATTRIBUTE # 1
DERIVED ATTRIBUTE # 2
DERIVED ATTRIBUTE # 3

FIG. 1

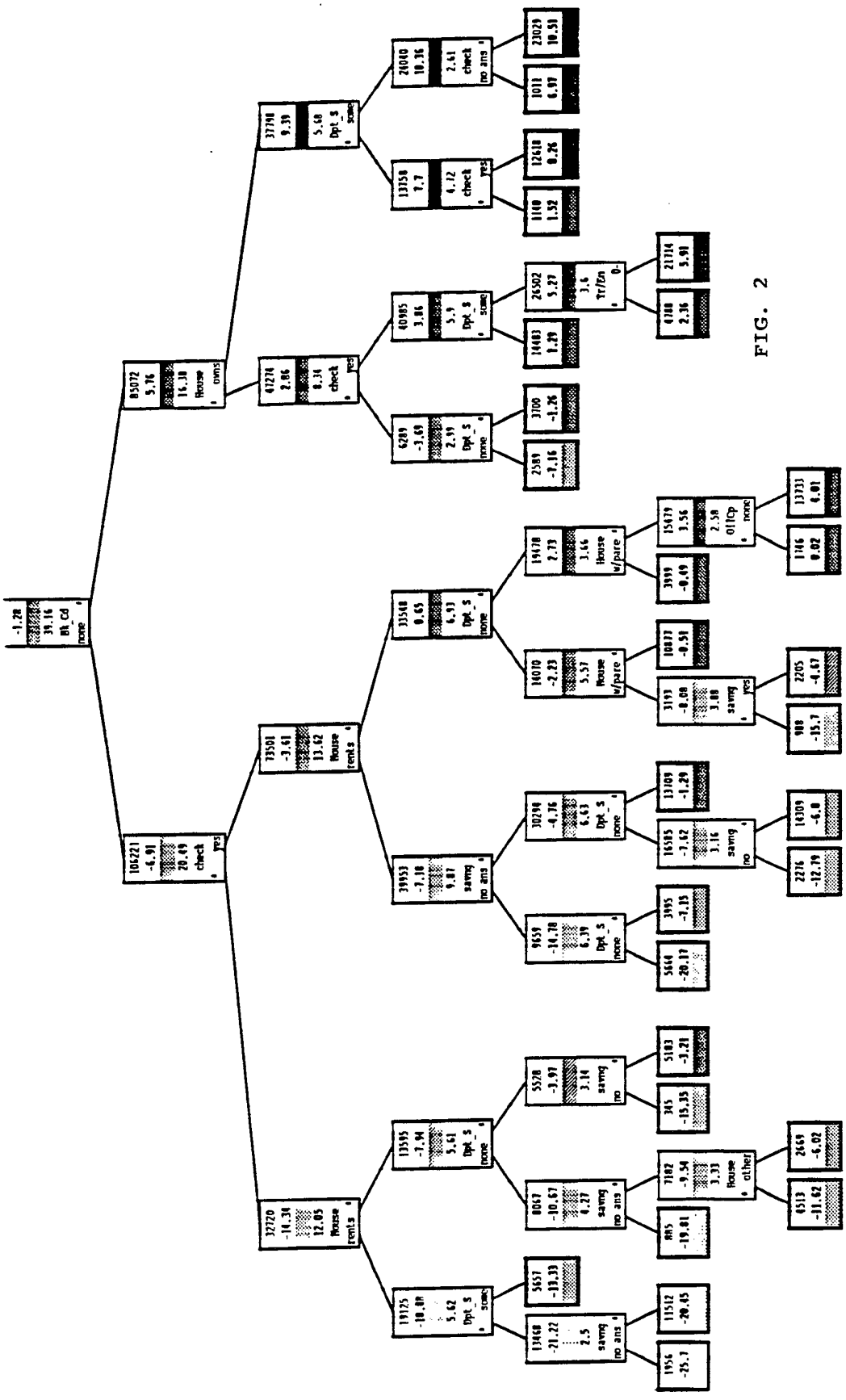


FIG. 2

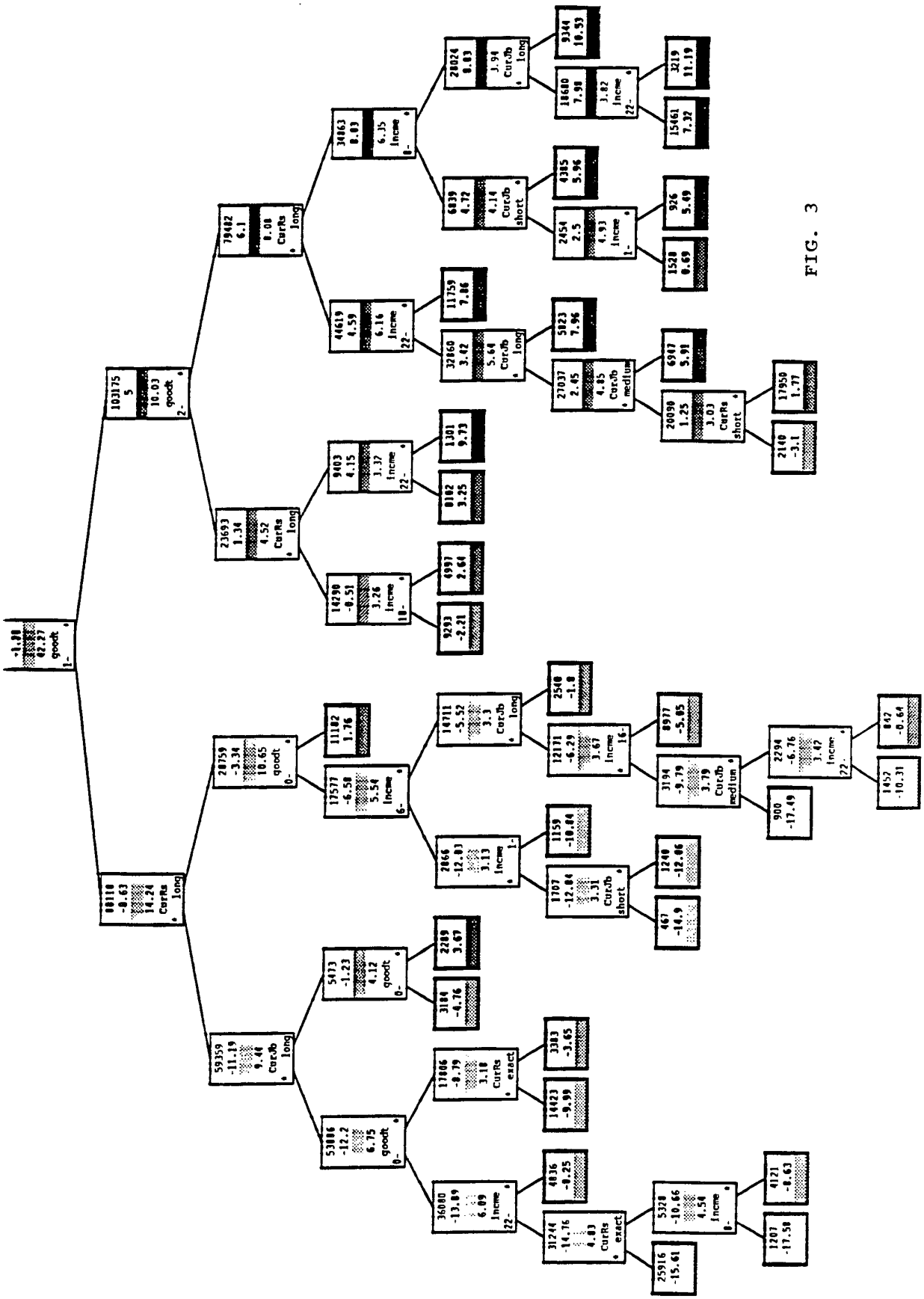
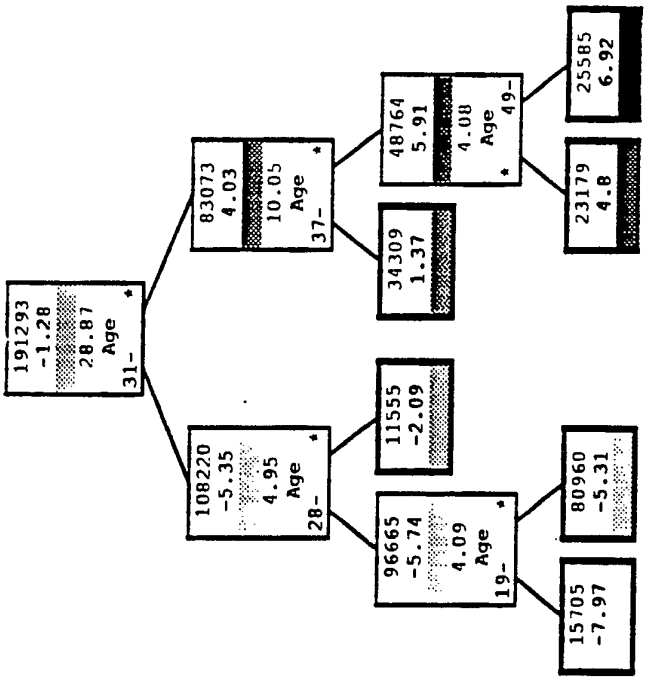


FIG. 3

FIG. 4


191293
-1.28
28.87
Age
31-
83073
4.03
10.05
Age
37-
48764
5.91
4.08
Age
49-
25585
6.92
23179
4.8
34309
1.37
108220
-5.35
4.95
Age
28-
11555
-2.09
96665
-5.74
4.09
Age
19-
80960
-5.31
15705
-7.97