IDEA—A conversational, heuristic program for inductive data exploration and analysis*

by LAURENCE I. PRESS and MILES S. ROGERS

System Development Corporation
Santa Monica, California

INTRODUCTION

Classical induction algorithms such as factor analysis, multiple regression, and early forms of cluster analysis have have been employed successfully to reveal a limited set of structures for particular classes of data. Even contemporary forms of cluster analysis[1] and other tree-growing processes[2,3,4] are restricted either in the structure to which they can respond or in the type of data to which they apply, or both. The goal of this project is to improve the power and scope of computer routines that search for structure in a data base. IDEA (*I*nductive *D*ata *Ex*ploration and *A*nalysis) is a computer program that detects and represents inherent structure in multi-variage data.

When the structure or pattern in data is poorly defined and specified, man's experiences, hunches, and intuitions often enable him to proceed where existing pattern-detecting algorithms fail. To overcome the limitations and restrictions of pure machine induction, we are providing an opportunity for interplay between the investigator's judgment about and knowledge of his data and the computation power of a time-shared computer.

To this end, the IDEA program is designed either to run in an automatic mode or to allow the investigator to intercede at each major decision in the analysis. At each such juncture he is presented with information that permits him to concur or override a computer decision before the program continues to the next major decision. In addition, IDEA has two other distinguishing features: (1) Heuristic computational procedures are used for those cases where the combinational aspects of the analysis would require extensive computations, and (2) heuristics are selectively used for different types of data, enabling IDEA to operate on a mixture of nominal (categorical), ordinal (ranked), and interval- or ratio-scaled measurements.

The IDEA approach

Classical statistical induction assumes that any given value of the dependent variable can best be predicted by adding together weighted contributions from several independent variables, from transformations of several independent variables and/or from interactions (products) of two or more independent variables. In general every component is employed in predicting the dependent variable in a uniform way for all of the data points. In those cases where intraregional differences are found (usually by visual inspection of scatter plots and regression lines), dummy variables are used to break the model into several specialized models, each of which is tailored to fit a specific region of the data space.

Rather than explain the observed value of the dependent variable by summing the effects of several components, IDEA attributes variability in the dependent variable to the conjunction of several conditions, defined solely in terms of values of various independent variables, and not in terms of transformations of the original variables or in terms of interactions among the components. By using the conjunction of several conditions we are able to define several regions in the data space (i.e., the regions for which the conjunction of the conditions is "true"). We thus account for the observed dependent values by noting the region in which they occur.

In other words, IDEA searches heuristically for regions which are well fit by a simple model in which

*The work reported herein was supported by SDC and contract DA-49 083 OSA-3124, Advanced Research Projects Agency, U. S. Department of Defense.

the dependent variable in that region is assumed to be constant with random error. In future versions we hope to search for regions where more complex models (e.g. regression) fit the data. Thus IDEA produces a model which contains sub-models in each of the regions defined by the meta-model.

The representational problem

In addition to a unique means of discovering structure in a data base, IDEA includes an unusual method for representing such structure: the decision tree. Methods for representing complex structures may be broken down into two categories: those that focus upon the observations (points in the multivariate observation space), and those that focus upon the variables (dimension of the space).

The latter are exemplified by factor analysis and traditional cluster analysis. In these, structure is described in terms of vectors in a factor space, or in terms of clusters or variables of similar factorial composition.

Our approach falls into the former class in that the decision tree that IDEA produces represents a partition of the multivariate space into exhaustive, mutually exclusive regions in which the data have homogeneous values for the dependent variables. As an example, consider Figures 1a and 1b.

1. What is your age? (AGE)
2. What is your sex? (SEX)
3. What is your education level? (EDUC)
4. How often do you attend church? (RELIG)
5. Did you participate in the riots or not? (RIOT)

Figure 1(a) – Sample questionnaire (5 variables).

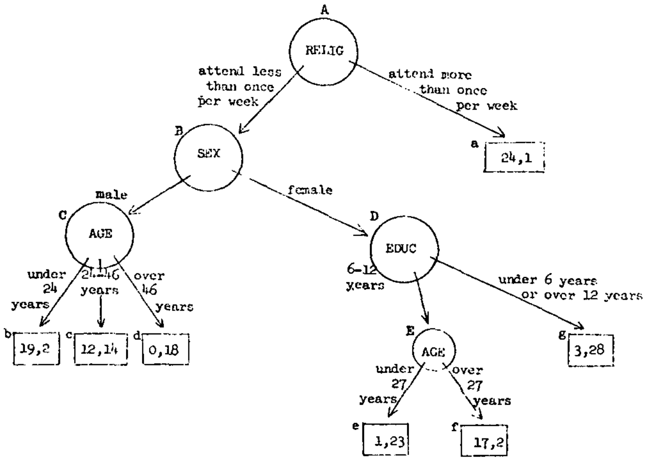

Figure 1(b) – Sample decision tree with RIOT as the dependent variable

The entities in this data base are a set of questionnaire responses obtained from residents of the Watts area of Los Angeles just after the 1965 riots. The variables are the questions asked in the questionnaire (Figure 1a). Thus each observation consists of the responses of a particular individual to all of the questions. Figure 1b illustrates a decision tree that might represent the structure of this data base. This particular decision tree results from treating a selected variable, participation in the riot (RIOT), as the dependent variable.

Each tree is made up of nodes (represented diagrammatically as circles), leaves (shown as rectangles), and the paths between them. The name associated with each node denotes the particular independent variable being used to try to account for the observed values of the dependent variable. The labels beside the paths leading from each node specify the particular values of the independent variable at the node. These were found to predict the dependent variable as accurately as possible for those observations partitioned by the decision tree to the prior node.

Note that these values partition the observations into exhaustive, mutually exclusive subsets.

At some point on each path it will be decided not to continue to partition. The path will then be terminated by a leaf (rectangle). The numbers inside the leaf denote the frequency of the dependent variable values for the observations in the data base that have been partitioned or sorted to that leaf.

For example, in Figure 1b, where participation in the riot is the dependent variable, the observations sorted to node D (females who are not highly religious) are partitioned into two subsets depending upon whether the education level is 6-12 years or over, or under that. Of the 74 observations sorted to node D, 31 are either under 6 or over 12 years in educational level. Leaf *g* tells us that 3 of the 31 were participants.

If it were possible to account for all differences on this dependent variable, each leaf would represent either all participants or all nonparticipants (e.g., leaf *d*). In this particular tree, leaf *c* is highly heterogeneous and might better have been replaced by a node for further partitioning.

In general, the decision tree is a graphic, easily understood representation of the data base structure, providing a simple description of many highly complex relationships. For instance, the relationship between two independent variables (x and y) and a nominal dependent variable (I or II), illustrated in Figure 2a, is summarized by the decision tree in Figure 2b. Many existing structure-seeking techniques would have difficulty in representing, let alone discovering, this relationship.

In addition to representing structure in a given data base, the tree produced by IDEA can be used to estimate dependent variable values for the sampled population. It may also be hypothesized as representing the true population structure, and this hypothesis can then be tested by additional research and analysis. The investigator may also gain insight while attempting to explain some characteristic of the structure, such as a complex interaction or a radical change in the dependent variable distribution between successive nodes.

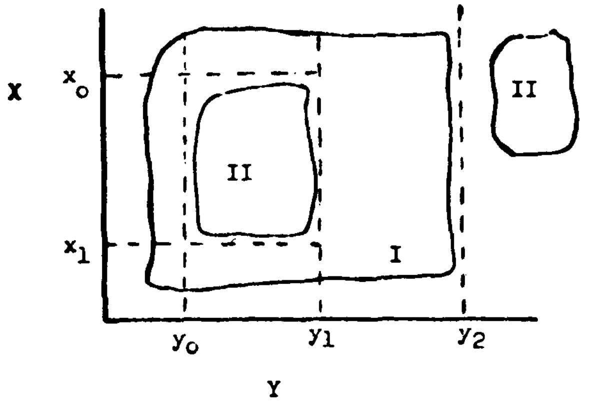

Figure 2(a) – Hypothetical multivariate data base

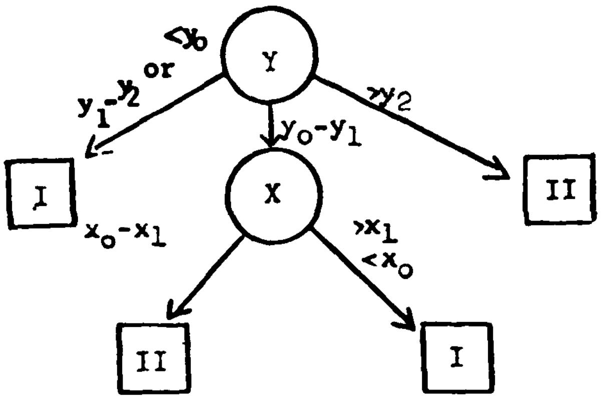

Figure 2(b) – Decision-tree summary of Figure 2(a)

The IDEA system

The IDEA system consists of three components: (1) A means for the investigator to input, edit, transform, and reconfigure his data base (data preparation); (2) a library of machine induction heuristics (heuristic induction routines); and (3) an interactive package allowing the investigator to obtain critical data about the program's impending decisions, to alter heuristic parameters, to add or delete data, and to accept the machine's decisions or impose his own (interaction). These components are considered separately below.

Data preparation

Component 1 above is currently provided via an interface with the TRACE[5] system. TRACE is a program with extensive capability to edit, reconfigure, and display multivariate data. In future versions we anticipate a requirement for a more continuous exchange between IDEA and TRACE. This capability would be a portion of component 3 above.

Heuristic induction routines

Component 2 consists of a library of heuristics and a monitor system. The heuristic applied at a given point in the analysis may be chosen by the investigator as long as it is appropriate to the level of measurement of the variables involved. For example, variables measured with a nominal scale (i.e., categorized) would be analyzed with statistics such as chi-square, which is appropriate to nominal scale data. Variables for which interval scale measurement is available could be analyzed with chi-square also. However, since this analysis would be insensitive to information involving order and distance, and since such information might be highly relevant to the underlying structure, IDEA would suggest that the heuristics used be based on variance components. On the other hand, heuristics involving components of variance would not be used with categorical data (nominal scale) because valid interpretation of such analyses presupposes order and distance relations in the data that may have no empirical basis.

Since we distinguish between the dependent and the independent variables measurement scales, considerations must be made separately for each. In terms of Stevens'[6] classification (nominal, ordinal, interval, ratio) there exists a 4 + 4 matrix of "cases" for which separate heuristics should be developed. Currently we have at least one operational heuristic appropriate to (i.e., which makes full use of all valid information concerning) the following cases: (a) nominal-nominal,* (b) nominal-ordinal, (c) nominal-interval, (d) interval-nominal, (e) interval-ordinal, and (f) interval-interval. These heuristics are also applicable to nominal-ratio, and interval-ratio cases, although with some loss in information. We are currently seeking appropriate heuristics for these cases and for the four cases involving an ordinal dependent variable.

In addition, we have developed alternate heuristics for those cases where a monotonic relation is known to exist between the dependent and independent variables, and this permits a reduction in the computation involved. Finally, we are also developing alternate heuristics for the nominal-nominal case in which we attempt to form parallel partitions instead of the sequential partitions discussed so far. Parallel

*Dependent variable scale is listed first, independent variable scale second.

partitioning involves more than one independent variable at a node as the basis for partitioning, and, hopefully, will complement sequential partitioning by being sensitive to aspects of underlying structure overlooked by the sequential mode.

The goal of the heuristics implemented initially is to produce a highly significant chi-square in the case of a nominal dependent variable and to produce a large reduction in unexplained variance in the interval case. These two criteria in themselves represent heuristic or judgment decisions, and are chosen because of several properties.

Our property is that they favor a partition into few subsets over a partition into many, even though the latter may yield a more precise estimate. While this property is pleasing when we are interested in a parsimonious structure, it may be inappropriate in an inferential setting, such as estimation of the dependent variable or pattern recognition. In these cases a heuristic which favors partitions into many subsets may be more useful (e.g., CLS – 10 in Hunt[7]).

A second property is that these criteria are suitable for the discovery of any single valued functional relationship, again favoring the relatively simple. This property is mandatory in the analysis of complex data such as that in many practical estimation and pattern recognition problems, particularly when the problem is in fact difficult – as when assumptions such as linear separability of classes or knowledge of underlying distributions cannot be made (e.g., Figures 2a and 2b).

A third property is that these criteria also favor variables with strong main effects. However, when an interacting variable fails to have a significant main effect, the presence of a singificant interaction may not be discovered. While more complex heuristics could be designed to ferret out such occurrences, the investigator may be able to infer this state by observing structure of the tree produced. In such instances, human pattern recognition skills may be substituted for more complex programs.

Since these heuristic decision criteria are themselves the results of reasoned judgment, it is *not* reasonable to expend computational effort to guarantee optimal partitions. Therefore, the procedures implemented are also purposely heuristic. They consider only subsets of the possible partitions of a given variable. At times they do not seek a proven optimum even within that subset.

For illustrative purposes, we will describe the heuristic for the interval-interval case in greater detail.

Prior to implementing this heuristic, several alternatives were considered and rejected. One was an algorithm to search exhaustively for an optimal solution. It was hoped that a lemma by Fisher[8] would make this computationally feasible, but such was not the case (although Fisher's lemma is used in step 3, below). Next, an attempt was made to discover an analytic solution in literature concerning the identification of homogeneous strata in experimental design, but all this work imposed untenable restrictions as to underlying distributions. A third approach, that of embodying a one-dimensional special case of some existing cluster-seeking procedure (e.g.,[9,10], was ruled out due to the restriction that the order of the observations on the independent variable be respected (step 2, below).

The dissatisfaction with the above approaches led to the decision to develop an estimate of the optimal partition by computing a function, the within groups sum of squares (WSS), of a random sample of the observations. This approach, the six steps of which follow, is analogous to forming an estimate of the height of the tallest person in a population by observing the height of the tallest person in a sample.

1. Select a random sample of size N (N being a specified parameter) from the data blocks sorted to the current node (if less than N observations exist, use all of them).
2. Order these according to the value of their independent variable.
3. Find the minimum WSS for 2-, 3-, . . . , M-way partitions of the dependent variable (M being a specified parameter), while respecting the ordering of step 2. These will define partitions of the independent variable for the sample (call these n-way partitions p_n^* and call the associated WSS, WSS_n^*).
4. Find P*, which is the partition that minimized $\frac{WSS_n^*}{N_n}$ (where N_n is a normalizing factor), to account for the value of n.
5. Consider P*, P* with all partition points shifted down to the first unique independent value, and P* shifted similarly upward, as three candidates for the final choice.
6. Select the one that minimizes WSS (this time computed for the dependent variable for all of the observations at the node).

This procedure embodies as estimate of the partition that minimizes the WSS of the dependent vari-

This procedure embodies as estimate of the parittion that minimizes the WSS of the dependent variable for all data blocks at the node.

Statements about the power and efficiency of this technique require a knowledge of the frequency distribution of the WSS of all possible partitions. Until an analytical or empirical description of the distribution is made, the computational capacity of SDC's

Time-Sharing System will be used to determine the sample size (N).

A similar situation arises when we attempt to arrive at normalizing factors for comparing partitions into different numbers of subsets. Here, in the estimation of the mean n-way WSS ($\overline{WSS}_n$) for the population of observations, we desire to compute the mean WSS for the sample. Evoking the central limit theorem, we know that the distribution of sample WSS means is normal around the population WSS mean, with known dispersion.

Therefore, a sample technique to obtain an arbitrarily precise estimate of $\overline{WSS}_n$ as a normalizing factor could be implemented. This procedure would possibly yield different results than the current one of using the mean WSS of all partitions explicitly considered for final selection. As this latter approach deals with a screened (biased) sample, it produces a lower normalizing factor.

Interaction

The interaction package allows the investigator to apply his knowledge of the theory in his field, the precision of his data, the characteristics of the various IDEA heuristics, and his pattern recognition skills in the structure-seeking process. Figure 3 demonstrates the flow of control between investigator and computer during analysis.

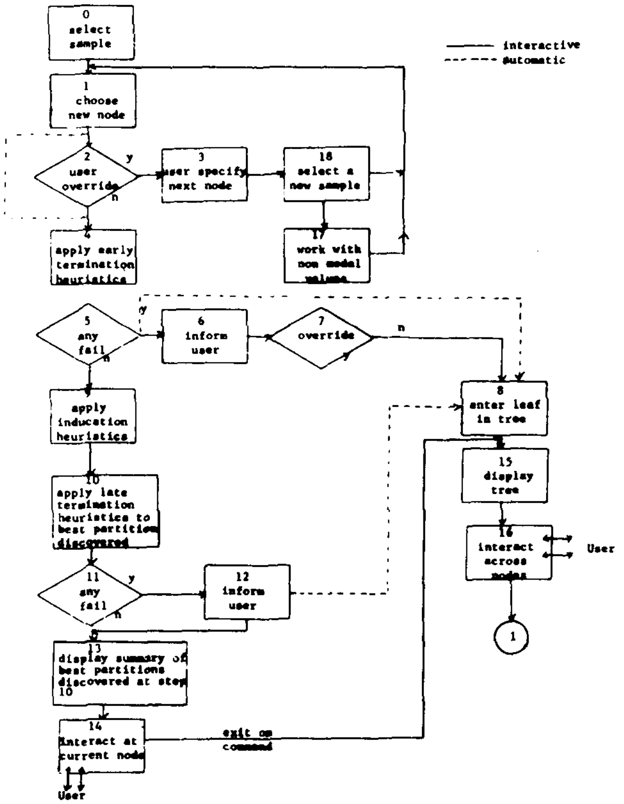

Figure 3 – Flow of control between investigator and computer

Most interaction occurs at step 14 where, in addition to a summary of the best partitions discovered (step 13), the investigator may request any information used by the heuristics in reaching the proposed decision. He may request detailed description of "good" partitions that the heuristics can discover for a specified variable, or of a given partition of a given variable.

In addition, IDEA provides for displays of specified variables. At present these include frequency distributions and contingency tables. We also intend to provide other plots, scatter diagrams, etc., with such aids as discriminant functions, regression analysis, iso-probability, and iso-distant curves derived under various assumptions.

The investigator may use these displays to decide which of several options to exercise at each node. Current options include: accepting the heuristics decision, overriding with a partition or a leaf, selecting a new portion of the emerging structure to analyze, "backing up" to a previously processed node, or continuing in the automatic node for N nodes before allowing further interaction.

SUMMARY

In summary, this project is concerned with the development of an interactive package of heuristic computer-induction programs. This program package has been used in the analysis of several multivariate data bases, including sociological questionnaires, projective test responses, and a sociopolitical study of Colombia. It is anticipated that the program will also prove useful in pattern recognition, concept learning, medical diagnosis, and so on. We believe that this approach to data analyses – a computer-mediated interplay between the investigator and his data – holds promise of a more effective inductive analysis than either man or algorithm could produce alone.

REFERENCES

1 G BALL
Data analysis in the social sciences
AFIPS Conf Proc 27 533 1955

2 E HUNT J MARIN and P STONE
Experiments in induction
Academic Press New York, 1966

3 J MORGAN and M SONQUIST
Some results of a non-symmetrical branching process that looks for interaction effects
J Amer Stat Ass 40 1963

4 J MORGAN and M SONQUIST
Problems in the analysis of survey data – a proposal
J Amer Stat Ass 58 415 1965

5 G H SHURE R J MEEKER and W H MOORE, JR.
TRACE time-shared routines for analysis, classification and evaluation
AFIPS Conf Proc Sp Jnt Comp Conf 30 525 1967

6 S S STEVENS, (Ed.),
Mathematics, measurement and psychophysics in handbook of experimental psychology
Wiley New York 1951

7 E HUNT
Utilization of memory in concept learning systems
West Mgmt Sci Inst UCLA 99 1966

8 W FISHER
On grouping for maximum homogeneity
J Amer Stat Ass 53 789 1958

9 E FORGY
Cluster analysis of multivariate data: efficiency vs. interpretability of classifications,
WNAR meetings, Univ of Calif Riverside 1965

10 J MacQUEEN
Some methods for classification and analysis of multivariate observations
West Mgmt Sci Inst UCLA 96 1966