MODEL DEVELOPMENT

William Bricken
March 1984

A model is the interpretation given to a formal (syntactic) structure. We propose a nested hierarchy of frames as the formal structure which we interpret as the behavior of a system, at the levels of objects, collections, and strategies. In addition, specific parts of the formal data structure (such as the data that represents the location of a particular object) are interpreted as specific facts about the model (such as the actual location of an object).

Qualitative modeling in AI defines relations between the parts of a model using logical rules and constraints. Traditionally, expert knowledge is arduously transcribed into computational formalisms. Machine Learning, the generation of new rules and constraints by the computer, is currently a pure research topic, and is unsuitable for application projects.

Quantitative modeling by statistical methods, however, is a mature field which can provide powerful tools provided that the assumptions about the structure of the numerical data are met.

The synthesis of AI and statistical modeling techniques is at the forefront of modeling research. Our proposal for automated model development emphasizes statistical techniques under an AI control regime, but does not attempt basic research into the synthesis of the two techniques. We recognize the integral part that a human analyst must play in the interpretation of complex data.

THE STATISTICAL TOOLKIT

Statistical analysis serves to condense large amounts of data into simple models that represent the most probable description of the data. For elementary analysis, we propose the following tools that could be applied to all numerical components of a complex data structure:

1. Data Description and Management

A facility for detailed description of a data field, including:

a. Display of raw data
b. Data editing and altering
c. Data transformation,
   including combination of multiple fields
      into a composite data field
d. Data description, including sub-population description,
   means, standard deviations, ranges, frequency
      distributions, and histograms
e. Estimation of missing values

2. Data Comparison

A facility for directly testing the similarity of two or more groups of data, using t-tests or analysis of variance.

3. Diagnostic Plotting of Data

a. Scatter plots to display correlations,
   accompanied by regression equations
b. Normal probability plots, to indicate conformance
   to distribution assumptions and the need for transformation

4. Categorical Data Analysis

a. Frequency plots
b. Chi-square analysis
c. Log-linear modeling.

The above facilities form a basic vocabulary for elementary data description. AI control of the invocation of these facilities is achievable and recommended. The analysis facilities that follow are technically more sophisticated, and would require user request:

5. Regression Analysis

Simple regression analysis would be provided in the Plotting facility above. Components in this package would include:

a. Multiple Linear Regression,
   for testing the dependency of several variables
      against an independent variable
b. Polynomial Regression, for higher order relationships
c. All Possible Subsets Regression,
   for exploratory regression analysis

All of the above facilities would include display of correlation matrices, and of residuals for determining the appropriateness of the analysis.

6. Cluster and Factor Analytic Techniques

Facilities for determining related clusters of data, including:

a. Cluster Analysis, to determine the association or similarity of groups of variables and of groups of cases. Analysis would proceed from either correlation matrices or distance metrics.

b. Discriminant Analysis, to determine classification functions that discriminate between members of a group.

7. Time Series Analysis:

Facilities for displaying and analyzing the relationship between data measurements over time, including:

a. Variable by time display plots
b. Spectral Analysis, the representation of time series data
   as summed sine waves, including
   1) trend removal
   2) filtering
   3) smoothing
   4) confidence bands
   5) calculation of autoregressive coefficients
      and autoregressive filtering
c. Autoregressive-Integrated Moving Average (ARIMA) techniques
   and intervention and transfer function models, including
   1) model identification
   2) parameter estimation
   3) diagnostic checking
   4) forecasting

Although these latter two techniques usually require the judgment of a statistician for their application, structured situations may exist in which the system could call upon a particular analysis template automatically. These applications are a matter of research; their identification would require the use of standard expert system engineering techniques. Specifically, the expert system engineer and the subject domain expert would need to identify and model those situations where, say, autoregressive techniques are appropriate. The expert system engineer would then code these situations as rules within the control structure of the expert system. These rules would subsequently call upon the analysis package to perform and display the statistical analysis whenever the conditions of the rules become manifest.

SOME SPECIFIC APPLICATIONS

A. Identification of Outliers during Regression Analysis

An outlier is an extreme observation. In the context of SSBN analysis, an outlier may be an interesting event. Outliers are identified in the course of regression analysis by plotting and analyzing the residuals (the difference between the predicted and the observed measurements). Residuals are assumed to be normally distributed. Those lying, say, more than two standard deviations from zero are abnormal. Alternatively, abnormality of residuals can be formally tested (for example, Dixon's test).

B. Model Building during Time Series Analysis

A time-domain model represents the regression of a variable upon its past. In exploratory model building, the phase of an autocorrelation may be unknown. As an example, particular submarine behaviors may be seasonal, but this autocorrelation would not show up in comparing only monthly behaviors (that is comparing the regression residuals for a given month to the month before). The phase of the autocorrelation would be shifted by six observations (six months). Exploration of different phase shifts for autocorrelation constitute this kind of model building.

C. Determination of Policy Effects using an Indicator Variable

Assume that a particular regression relation is known. An environmental or political change can be factored into this regression by coding it dichotomously and recalculating the regression. If the resulting equations (with and without the indicator variable) have significantly different coefficients, then the indicated change has effected the known behavior.

D. Stability of Behavior

Stability of behavior is intimately connected with its predictability. However, different types of stability are indicated by different measurements [ref: Wohlwill].

1. Absolute invariance is behavior that remains unchanged over time; it is measured by the coefficient of variability.

2. Regularity of occurrence can be measured by chi-square tests to determine the deviation from equal distribution of events over time.

3. Regularity of regression shape is measured by the degree of the best fitting polynomial.

4. Conformance to a prototype function is determined by the least squares goodness of fit.

5. Constancy of position relative to a norm is measured by the variance of relativized measures over time.

6. Preservation of individual differences is indicated by the stability coefficient, the correlation between measures at different times.

REFERENCE

Wohlwill, J. F., The Study of Behavioral Development. New York: Academic Press, 1973.