RC23198 (W0404-174) April 30, 2004
Other

IBM Research Report

Tree Models

Richard Daniels
Terry College of Business
University of Georgia
Athens, GA 30602

Dailun Shi
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Tree Models

Richard Daniels, Terry College of Business, University of Georgia, Athens, GA 30602
Dailun Shi, IBM T. J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532

Tree models, also known as multinomial process tree models, are data-analysis tools widely used in behavioral sciences to measure the contribution of different cognitive processes underlying observed data. They are developed exclusively for categorical data, with each observation belonging to exactly one of a finite set of categories. For categorical data, the most general statistical distribution is the multinomial distribution, where observations are independent and identically distributed over categories, and each category has associated with it a parameter representing the probability that a random observation falls within that category. These probability parameters are generally expressed as functions of the statistical model's parameters, i.e., they redefine the parameters of the multinomial distribution. Linear (e.g., analysis of variance) and nonlinear (e.g., log-linear and logit) models are routinely used for categorical data in a number of fields in the social, behavioral, and biological sciences. All that is required in these models is a suitable factorial experimental design, upon which a model can be selected without regard to the substantive nature of the paradigm being modeled.

In contrast, tree models are tailored explicitly to particular paradigms. In tree models, parameters that characterize the underlying process are often unobservable, and only the frequencies in which observed data fall into each category are known. A tree model is thus a special structure for redefining the multinomial category probabilities in terms of parameters that are designed to represent the underlying cognitive process that leads to the observed data. Tree models are formulated to permit statistical inference on the process parameters using observed data.

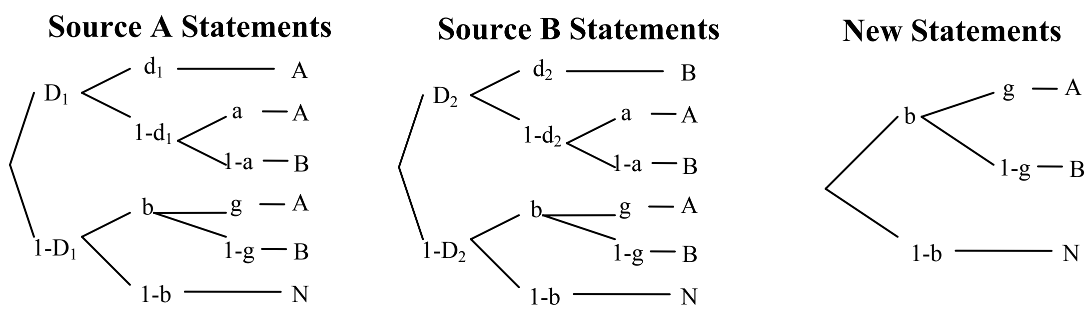
Tree models reflect a particular type of cognitive architecture that can be represented as a tree, i.e., a graph having no cycles. In a tree that depicts the underlying cognitive process, each branch represents a different sequence of processing stages, resulting in a specific response category. From one stage to the next immediate stage in a processing sequence, one parameter is assigned to determine the link probability. The probability associated with a branch is the product of the link probabilities along that branch. Each branch must correspond to a category for which the number of observations is known; however, there can be more than one branch for a given category. The observed response patterns can thus be considered as the final product of a number of different cognitive processes, each of which occurs with a particular probability.

A key characteristic of tree models is that category probabilities are usually nonlinear polynomial functions of the underlying process parameters (in contrast to the classical models for categorical data mentioned above, which all have linearity built in at some level). On the other hand, tree models are much less detailed than more sophisticated cognitive models like neural networks. Thus, while tree models capture some, but not all, of the important variables in a paradigm, they are necessarily approximate and incomplete, and hence are confined to particular paradigms. Despite this disadvantage, the statistical tractability of a tree model makes it an attractive alternative to standard, multipurpose statistical models.

A comprehensive review of the theory and applications of tree models is given in Batchelder and Riefer (1999)[1]. For readers interested in learning more about tree models and statistical inference, Xiangen Hu has developed an informative website at http://irvin.psyc.memphis.edu/gpt/.

An Example: "Who Said What" Task

To illustrate the structure of a tree model, consider the "Who Said What" task. Perceivers first observe a discussion involving members of two categories (e.g., men and women). In a subsequent recognition test, subjects are shown a set of discussion statements and asked to assign each statement to its speaker. Apart from statements that occurred in the discussion (called old statements), new statements are also included in the assignment phase. For each statement, participants must assign Source A (male), Source B (female), or N (new statement). The figure below depicts a tree model for the three types of statements. Note that there are a total of 7 process parameters $\{D_1, D_2, d_1, d_2, a, b, g\}$, 15 branches, and 9 response categories (A, B, and N for each tree).



The model assumes that a participant first detects whether a statement is old or new with probability D_1, D_2, or b for source A, B, or new statements, respectively. If an old statement is correctly detected as old, then d_1 and d_2 capture the capacity to correctly assign the old statement to source A and B, respectively. If the participant cannot directly attribute a statement to a source (with probability $1\text{-}d_i$, i=1,2), a guessing process determines the statement's source – the effectiveness of this process is measured by parameter a. If a statement is new, then another guessing process (the effectiveness of which is measured by parameter g) is used to determine the statement's source. Finally, if an old statement is not detected as old (with probability $1\text{-}D_i$, i=1,2), it is treated as a new statement; as such, the branches emanating from $1\text{-}D_i$, i=1,2, reproduce the new statement tree.

[1] W. H. Batchelder and D. M. Riefer, Theoretical and empirical review of multinomial process tree modeling, Psychonomic Bulletin & Review, 1999, 6(1), page 57-86.

Several observations emerge from this example. First, the sequential nature of the process is based on both cognitive theory and assumptions about how statements are assigned to sources. Second, some parameters (e.g., a, g and b) appear in more than one tree, implying, e.g, that the probability of assigning a statement that is incorrectly detected as new to Source A is equal to the probability of assigning an incorrectly-identified new statement to Source A. Since most of the parameters can be interpreted as conditional probabilities (i.e., conditional on the success or failure of other processes), it would perhaps be more appropriate to use different parameters to represent the same cognitive process in different trees. However, if S denotes the number of process parameters and J the number of resulting data categories, S must be no larger than J-1 for the model to be statistically well defined. As a result, model realism may be traded off to gain model tractability and statistical validity.

Finally, note that the category probabilities are the sums of the products of the underlying processing parameters. For example, the probability of correctly identifying a statement from Source A is $P(A \mid A) = D_1 d_1 + D_1(1-d_1)a + (1-D_1)bg$. Similarly, the probability that a random observation falls into each of the other eight categories can be expressed as a function of the seven process parameters (D_1, D_2, d_1, d_2, a, b, g). As such, the objective of tree modeling is to draw statistical inference on the process parameters using the sample frequencies of observations that fall into each data category, thus providing insight into the unknown cognitive processes.