# USING CAUSAL KNOWLEDGE TO LEARN MORE USEFUL DECISION RULES FROM DATA

Louis Anthony Cox, Jr., U S WEST Technologies
4001 Discovery Drive, Boulder, Colorado, 80303
(303)-541-6043 (phone) (303)-541-6003 (Fax) tony@uswest.com

## INTRODUCTION

One of the most popular and enduring paradigms in the intersection of machine-learning and computational statistics is the use of recursive-partitioning or "tree-structured" methods to "learn" classification trees from data sets [Buntine, 1993; Quinlan, 1986]. This approach applies to independent variables of all scale types (binary, categorical, ordered categorical, and continuous) and to noisy as well as to noiseless training sets. It produces classification trees that can readily be reexpressed as sets of expert systems rules (with each conjunction of literals corresponding to a set of values for variables along one branch through the tree). Each such rule produces a probability vector for the possible classes (or dependent variable values) that the object being classified may have, thus automatically presenting confidence and uncertainty information about its conclusions. Classification trees can be validated by methods such as cross-validation (Breiman et al., 1984), and they can easily be modified to handle missing data by constructing rules that exploit only the information contained in the observed variables.

Despite these powerful advantages, classification tree technology, as implemented in commercially available software systems, is often more useful for pattern recognition than for decision support. Practical business and engineering decisions require some new considerations to be incorporated into the recursive partitioning paradigm. The most important ones include

*(i) Costs of information acquisition* (Cox and Qiu, 1994). If a classification tree requires tests that are infeasible or that are too expensive to perform in practice, then it will be rejected by practitioners no matter how well it performs on training samples.

*(ii) Ability to make changes based on the tree.* If the best tree for predicting a response uses information (e.g., about time-varying covariates) that does not become available in practice until after key decisions must be made, then the inferences supported by the tree, while perhaps very valuable for scientific research purposes, will not be suitable for real-time decision support and guidance of actions. An example from the domain of cancer risk prediction is as follows. The best predictor of liver carcinomas in mice exposed to certain chemicals turns out to be the presence of liver adenomas. This is useful to scientists studying the relation between benign and malignant tumors, but it is useless for predicting whether a specific mouse will develop a liver carcinoma, since neither adenomas nor carcinomas can be observed until autopsy.

*(iii) Pursuit of multiple objectives* that may be differently affected by the actions taken. Actions that affect the values of variables in a causal model may lead to (perhaps probabilistic) changes in several outcome measures. For example, in an analysis of employee survey data presented below, the multiple objectives include improving employee job satisfaction, increasing the productivity of work groups and the quality of work done, reducing job-related stress, and improving the perceived performance of management. Although these goals are largely consonant, there is some conflict among them, e.g., between reducing job stress and increasing productivity. The challenge in this setting -- to find a small core set of actions that is undominated by other available actions in improving measurements of all these performance dimensions simultaneously -- is typical of many real-world applications.

In summary, a theory is needed of how to learn *prescriptive,* rather than merely descriptive, classification trees and models from data. The goal is to identify and recommend changes (actions or decisions) that are predicted to have high impacts on improving the criteria. This contrasts with the usual goal of existing classification tree systems, which focus on recommending tests that will be most useful in accurately predicting the values of some quantities from observations of the values of other quantities.

This paper presents results of an applied research effort focused on how to modify conventional recursive partitioning programs (e.g., CART, Knowledge Seeker, ID3) so that they will learn useful decision rules (prescriptions for action) from data. The desired output is no longer a probabilistic prediction of the value of a dependent variable, based on the observed values of independent variables. Instead, it is a prescription for what potentially costly actions to take (i.e., what values to assign to different controllable input variables), based on (perhaps costly) measurements of multiple independent variables, so as to bring several dependent variables simultaneously into a desired "efficient" (undominated) target set of joint values. The resulting methodology has been applied to several real problems in the telecommunications industry, including selecting actions to improve customer service (Cox and Bell, 1995) and identifying a strategy for improving employee performance and morale, discussed in the last part of this paper.

## PROBLEM FORMULATION

Let Y be a vector of dependent variables, X a set of controllable independent variables, and Z a set of observable (empirically measurable) but not controllable variables. Values of these variables may differ across the individuals ("cases" or "instances") to which the decision rule is applied. A *reduced causal model* describing the (in general probabilistic) dependence of Y on X and Z is defined by a pair of functions [p(y | x, z), f(z)], where p(y | x, z) is the conditional probability density that Y = y, given that X = x and Z = z. The function f(z) is the marginal frequency distribution of Z values in a population of individuals or cases under study, i.e., f(z) = Pr(Z = z). Such a "reduced" form can always be obtained from a directed graph representation of a corresponding "structural" causal model, such as a path diagram or an influence diagram. The idea of causality in this context is that X is not merely passively associated with Y -- e.g., because X and Y have a joint frequency distribution in the population that is not the product of their marginals -- but in fact that Y will change when X is changed (Spirites and Glymour, 1994). Learning how to choose X so as to obtain desired Y values is an exercise in valid prediction of how changes will propagate from controlled variables to outcome variables -- a task more difficult than statistical inference alone.

Let $v_j(y, z)$ denote the "value", as measured on criterion scale j, of obtaining response y from an individual described by covariate vector z. Let $\mathbf{v}(y, z)$ be the value vector whose jth component is $v_j(y, z)$, for j = 1, 2, ..., K, where K is the number of criteria. Then the problem addressed in this paper is to learn

directly from a data set how to choose x so that the frequency distribution of $\mathbf{v}(y, z)$ induced by x in the population of individuals is undominated (in the sense of multivariate first-order stochastic dominance, FSD). The desired output is thus a prescribed choice of x from a set A of feasible alternatives such that no other choice of x in A yields a higher probability of simultaneously obtaining at least as much of every criterion (assuming that criteria are oriented so that more is better). According to the standard von Neumann-Morgenstern (NM) theory of "rational" decision making (and to many more recent normative theories with less stringent axioms), *all* rational decision makers who prefer more to less of each criterion will prefer x to x' if and only if x dominates x' by multivariate first-order stochastic dominance (Hirschleifer and Riley, 1992). We will abbreviate this relation as x FSD x'.

If the reduced model (p, f) were known with certainty, then the preceding problem could be formulated as a standard multicriteria statistical decision theory problem, e.g., as the vector optimization problem

$$\max_{x \in A} E_{p,f}[\mathbf{v}(y, z) \mid x] \tag{1}$$

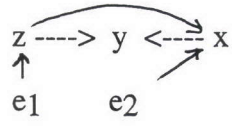
or as the scalar optimization problem

$$\max_{x \in A} E_{p,f}[u(\mathbf{v}(y, z)) \mid x] \tag{2}$$

where u is a multiattribute utility function mapping the vectors $\mathbf{v}(y, z)$ into "utilities" representing risk-adjusted individual preferences for outcomes (Hirschleifer and Riley, 1992). Costs of control (i.e., the cost needed to assign a specific value to a controllable variable) may be included in either formulation either as a penalty term or as another criterion. (Either leads to the same set of undominated actions.) If the detailed individual preferences and risk attitudes required to construct a multiattribute utility function are not available, then the vector optimization problem (1) can still be solved by identifying values of x that are undominated with respect to the FSD criterion. The resulting solution is typically a set of several x values, each of which may be preferred by some rational decision maker to any of the others.

The significance of the concept of causality in this setting is that the standard statistical decision theory formulation may generate incorrect recommendations, as we shall now show.

Consider a causal model represented by the diagram




and described by the equations

$$y = z - x + e_1 \tag{3}$$

$$x = z/3 + e_2, \quad (4)$$

which together imply that

$$y = 2x + (e_1 - 3e_2). \quad (5)$$

Assume that $E(e_1) = E(e_2) = 0$ and that these error terms are uncorrelated. Then the regression relation between x and y is found from equation (5) to be

$$E(y \mid x) = 2x. \quad (6)$$

On the other hand, the expected change in y when x is increased from 0 to $x^* > 0$ directly (without changing z) is found from equation (3) to be

$$E(y \mid x = x^*) - E(y \mid x = 0) = [E(z) - x^*] - [E(z) - 0] = -x^*. \quad (7)$$

Thus, what might be termed the "causal regression line" relating y to x is

$$E(y \mid x) = - x \quad (8).$$

The causal association between x and y has the opposite sign from their statistical association. This continuous version of Simpson's paradox, adapted from Lindley (1990), can readily be extended to probabilities by interpreting y as the logit for occurrence of a binary event (e.g., cancer among individuals in a population exposed occupationally to a dose, x, of an anticarcinogenic chemical whose ambient concentration is positively correlated with the ambient level of a carcinogenic chemical, z.) Treating conditional probabilities (or conditional expected values of y) given x as the values that would be obtained by setting values of x will in general lead to incorrect decisions because of misspecification errors in the model relating decisions to their probable consequences.

Even if the causal relation between controllable variables and outcomes is known, however, there is another pragmatic difficulty. In applications ranging from statistical quality control to design of new products, neither the population "mixing distribution" f nor the probabilistic response function (or probabilistic choice function, in marketing applications), p, is likely to be well known. Instead, an approximation to the model must be learned directly from the data. Thus, the full problem is to examine a sample of cases with different (x, y, z) values, some of which may be missing, and then to make a decision based on the observed sample values -- and on any knowledge, beliefs, and information (e.g., on mental models of the causal relations among the x, y, and z variables) that are expected to lead to a more favorable subjective distribution of outcomes. The decision to be made is typically a choice of either a specific *value* of x to apply to the population as a whole -- for example, in decision contexts such as public health decision making, where x might represent a permissible exposure limit for a chemical that has been statistically associated with adverse health outcomes -- or else choice of a *decision rule* to be applied to individual cases -- for example, a rule for examining and treating patients arriving with certain symptoms. In either case, the novel challenge faced by practitioners is to select a decision rule or value based on an incompletely known probability model that has been partially revealed through sample values of the (x, y, z) variables.

Before turning to strategies for solving this problem, it is worth noting the following additional complexities that often arise in practice.

> The choice set, A, of possible values of x may be specified through a complex set of rules or constraints, e.g., stating that choosing some components of x restricts the permissible values of the remaining components.
>
> Some of the Z variables may be observable only if others have already been observed, or only of certain actions have already been taken.

More generally, the feasible sequences of activities (observations and actions) may be partially ordered by a "causally enables" relation between sets of completed activities and as-yet unattempted activities. Such constraints embody the types of temporal/causal restrictions that have been incorporated in most automatic planning systems since STRIPS, in which the "enables" relation is interpreted in terms of satisfaction of preconditions. If such constraints are important in an application, the task of discovering a minimum expected cost undominated strategy may be computationally intractable. Indeed, even without these restrictions, discovery of minimum expected cost decision rules is in general NP-hard (Cox and Qiu, 1994). Therefore, we turn next to a class of heuristics that have proved useful for solving the types of problems described here.

## A SOLUTION HEURISTIC

We next present an algorithm, somewhat similar in spirit to Box's EVOP methodology for combined experiment design and process improvement (Box and Draper, 1969), that has proved useful in learning robust, efficient decision rules (and reduced models supporting them) from data. The process requires searching through a space of classification trees to discover locally efficient (undominated) ones that are also robust to small changes in the observed data. Tree evaluation and selection criteria include

> o *Actionability.* The tree should consist of nodes that represent controllable variables, and the actions that it prescribes should not require information or other preconditions that violate causal knowledge.
>
> o *Causality and efficiency.* The actions in the prescribed tree should lead to an undominated improvement in the evaluation criteria of the multi-criteria design or decision problem.
>
> o *Robustness.* The set of recommended actions in the prescribed tree should remain the same or nearly the same if another "neighboring" efficient tree is selected instead. This criterion may be formalized in terms of "core strategies", i.e., sets of actions that constitute kernels or quasi-kernels of underlying binary comparison digraphs.

The focus of the following discussion is on methods that have been used in practical applications at U S WEST over the past several years. Therefore, our solution approach is presented as a general "meta-heuristic", i.e., a generalized algorithm whose steps can be instantiated in various ways. An implementation

that we have used to analyze several data sets consisting of responses to customer and employee surveys is then described in the context of an application to finding actions to improve employee survey results. Given the complexity of the class of problems addressed by this method, many interesting theoretical and empirical research questions remain open about the best techniques for carrying out the steps in the meta-heuristic. These are topics of ongoing research at U S WEST.

The following terms will be used. An *action* consists of a *measurement* that can be performed (its preconditions are satisfied) or an *act* that can be taken. (Thus, "actions" and "activities" are synonymous in this paper.) Actions may be partially ordered by precedence constraints reflecting causal enablement. (For example, mouse liver tumors cannot be observed -- a measurement -- until the mouse has been sacrificed -- an act.) There may be a cost (dollars, time, or other resources) required in order to undertake an act. Effective heuristics for identifying low-cost inspection strategies when inference, rather than action, is the goal have been presented by Cox and Qiu (1994). The following discussion focuses on cases in which economic cost is only one of many criteria and the goal is to discover effective sequences of actions (i.e., "plans" that tend to produce desired outcomes).

Our proposed heuristic approach is as follows. (This version is for decision problems in which a set of actions is to be selected to apply to an entire population, e.g., an employee body or a set of customers. A similar heuristic applies when decision trees incorporating diagnostic tests or other possibly expensive observations are allowed and can be applied to individuals one at a time.)

## HEURISTIC SEARCH FOR UNDOMINATED ACTION PLANS

0. Start with an empty set of actions. Search over action sequences of length 1 (i.e., individual actions) to identify actions that leads to an undominated (via FSD) improvement in the causal distribution of consequences. (As suggested previously, a "causal distribution" refers to the probability distribution that is expected to hold once the action has been taken. This distribution typically cannot be derived from observed data alone, but requires use of a hypothesized causal model, e.g., represented as an influence diagram or Bayesian belief net, relating actions to their probable consequences.) An action that leads to an undominated improvement in the performance criteria is called an "undominated action". It may also be called an *undominated partial plan of length 1.*

**Phase 1: Generate undominated partial plans.** For each undominated partial plan of length k, consider appending each action that is feasible once the actions in the undominated partial plan have been taken. If an action can be found that leads to a stochastically dominating (multivariate FSD) improvement in the performance criteria (with respect to the estimated causal distribution of consequences induced by the plan), then append it to the partial plan to create a new partial plan of length k + 1. Repeat until no further FSD improvements can be found. Delete all partial plans that have been improved (in the sense of FSD) by one or more extensions. The result of this phase is a set of undominated partial plans of various lengths.

**Phase 2: Identify a "core set" of actions for initial implementation.** Each undominated partial plan identified in Phase 1 consists of an ordered set of actions. To identify which specific actions to take first (continuing with the case where one or more actions must be selected for application to an entire population of cases), each initially feasible action is assigned a numerical score equal to the number of "votes" that it receives from the set of undominated partial

plans (i.e., the number of such plans that it appears in). The initially feasible actions with the highest scores are recommended for implementation.

In practice, both phases are often carried out using only the benefits-related criteria, and then the top-scoring actions are evaluated for cost , practical feasibility, and confidence that the predicted consequences will occur -- i.e., that conditional frequency distributions obtained from data are causally predictive). The output of Phase 2 is a set of top-ranked actions recommended for implementation. By construction, they meet the criteria of actionability (they are actions), efficiency (it is not easy to find plans that dominate the best plans starting with the recommended actions), causality (judgments about causal distributions have been incorporated into the search for FSD-improving extensions of partial plans), and robustness (since FSD is used as the criterion in generating plans for consideration).

Many refinements may be made in the implementation of this basic framework. In applications to large data bases (involving thousands of survey questionnaires -- the "cases" in the training sample -- and typically several dozen to over 100 variables, some actionable and some merely observable ), we use a "best-first" rule to select partial plans for possible extensions. Then, we search among actions that have already been automatically rank-ordered in terms of their ability to improve prediction of one of the criteria, using the KNOWLEDGE SEEKER(TM) recursive partitioning algorithm (Biggs et al., 1991). Undominated distributions are screened for by looking for distributions that maximize the proportion of cases giving extremely high performance on the selected criterion. The process is repeated for one criterion at a time, since the KNOWLEDGE SEEKER(TM) program is constructed to deal with only one dependent variable at a time, and the intersection of the resulting recommended action sets is used to make a final recommendation. This multi-pass implementation is admittedly only an approximation to the ideal two-phase approach based on multivariate FSD, but it has proved successful in identifying small sets of recommended actions that have won immediate credence and strong support from decision makers knowledgeable about the selected application domains.

## AN EXAMPLE: USING EMPLOYEE SURVEY DATA TO PLAN IMPROVEMENTS

We conclude with insights gleaned from a recent application to employee survey data. The study had multiple goals: to understand drivers of employee job satisfaction, morale, and performance and to recommend management actions that would improve all three. The data were obtained from questionnaires filled out by several thousand U S WEST employees covering many aspects of work life, including perceptions of one's own job performance, rating of one's immediate supervisor along various dimensions, descriptions of one's work group and the quality of its work, ratings of top management on different dimensions, and attitudes and beliefs about the company as a whole, both in absolute terms and compared to

other companies. Each questionnaire contained approximately 110 detailed questions (including about a dozen demographic and career history questions). Most questions involved ratings on a conventional five-point scale (ranging from strongly disagree to strongly agree). Responses were treated as ordinal categorical variables. The sample was designed to be exhaustive and the response rates were high enough so that selection artifacts could not significantly affect the conclusions.

The large volume of data generated by this study was analyzed using the approach outlined in the previous section. Some instructive lessons from the analysis are as follows:

1. The automated KNOWLEDGE SEEKER(TM) procedure produced trees that clearly revealed the need to introduce causal knowledge into the analysis in order to achieve useful results. For example, a classification tree for predicting the rating of the GROUP'S WORK QUALITY variable (reflecting agreement with "My work group produces high-quality work") showed a mix of potentially actionable variables (such as "WE COOPERATE TO GET THE JOB DONE", which may be affected by strong team management) and factors that cannot be directly manipulated (such as "I LIKE THE KIND OF WORK I DO" or "MY WORK GIVES ME A FEELING OF PERSONAL ACCOMPLISHMENT").

2. We introduced *basic causal knowledge* in the simplest way: by separating variables into potentially actionable ones [represented by input nodes, i.e., nodes with only outward-directed arrows in a directed acyclic graph (dag) model of the causal relations among variables]; consequence or outcome measures (corresponding to output or "sink" nodes in the dag model), and intervening variables (represented by nodes with both inward and outward directed arrows.)

3. *A dag model representing causal relations* among variables was developed by an informal but useful process familiar to many researchers and practitioners in this field. First, a table of Spearman rank correlations was used to identify the pairs of variables that are most strongly associated. Then, "common sense" (i.e., knowledge of the meanings of the variables and how they might causally affect each other) was used to make tentative assignments of causal directions between strongly linked pairs. Next, other variables that might plausibly "explain away" the association were tested to see whether the binary association in fact disappeared when it was conditioned on the value of the new variable(s) (i.e., whether d-separability could be established). Conversely, third variables were interposed between strongly linked pairs (as in x --> z --> y, where z has been interposed between x and y) or introduced as common factors (as in x <-- z --> y) where doing so would successfully "explain away" the association between the two variables. (Human judgment was used to select variables to try interposing and to suggest other aspects of model structure, such as directions for arcs.) This process was used to grow larger and larger digraphs such that any pair of variables joined by a directed arc had a positive association not explained by any third variable. The largest dag generated is a model for the data. It is consistent both with the binary associations and the conditional independence structures implicit in the data, as well as with the causal knowledge used to suggest the graph structure (by selecting triples of variables to test for conditional independence.) A more formal approach in a similar spirit is the TETRAD program (Spirites and Glymour, 1994 and references therein).

4. The causal knowledge represented in the dag model allowed the number of distinct variables in the analysis to be reduced from over 100 to 24. The reduction process (described by the jargon name "homomorphic aggregation") is conceptually simple. If any two variables are related to the rest of the variables in the dag model by identical directed arcs, then the two variables are aggregated into one combined variable and interpreted as two measures of the same underlying construct. The final set of 24 variables included 13 input variables, 3 output variables (intention to stay with the company, quality of work produced, and rating given to one's supervisor), and 8 intervening variables (including job satisfaction, job stress, and possession of skills needed to do a good job). The abbreviated titles of the input variables include "job security", "teamwork", "can speak one's mind around here", "urgency of job", "employees are well trained", "can get needed skills", "my boss asks my opinion", and "my boss rewards improved work quality".) Some of these inputs affected (either directly or along a chain) many output variables. For example, "my boss asks my opinion" directly affects "supervisor rating" and indirectly affects "quality of work" through the intervening variable "I can take action". It also indirectly increases (in the sense of FSD)

"job satisfaction" and decreases "job stress" through paths that involve the intervening variable "my job makes good use of my skills and abilities".

5. Following these pre-processing steps, we applied the two-phase method of the previous section (using KNOWLEDGE SEEKER(TM) as a tool to search for FSD-undominated partial plans for one output variable at a time). The dag model was used to identify actionable input variables and to relate them not only to the three output variables, but also to intervening variables of interest such as job stress and job satisfaction.

The main results were that the method successfully identified a small set of core actions and implementable policies that simultaneously led to predicted improvements in all of the main outcome measures (job satisfaction, employee retention, perceived quality of work, and appraisal of management). Few undominated partial plans exceeded four actions in length. (Precedence constraints were not active in this problem, so plans were just unordered sets of activities.) A plan that remained undominated even when various subsets of variables were eliminated from the model (indicating a form of "robustness") emphasized the following three clusters of actionable principles (synthesized from items in the questionnaire):

> ***Challenge*** employees to find new and better ways of doing things that affect external customers positively. (Listen to and actively seek their ideas, then follow up with actions.)
>
> ***Enable*** employees to improve their skills so that they can do their jobs well. (Coach/develop them and keep them informed and involved in decisions that affect their work.)
>
> ***Reward*** employees for demonstrating continuous quality improvement.

These recommendations, extracted from the large volume of data and many competing items on the questionnaire, are remarkably consistent with recent popular books on effective management of research and technology organizations. The dag causal model behind them identifies *sense of personal accomplishment* as the key intervening variable mediating between these types of inputs and the outcome criteria of job satisfaction and employee retention, perceived work quality, and evaluation of U S WEST management.

Specific plans based on the preceding analysis were predicted to dramatically increase the proportion of employees expressing top levels of job satisfaction and work quality. (This prediction involved a causal judgment: that the selected factors actually affect scores on outcomes. An alternative hypothesis that was consistent with the data was that an unmeasured latent variable -- "toughness in grading" -- could explain away some or all of the association between these inputs and the various outcome measures.) In 1994, management changes based on the preceding analysis were trialed in a small group of approximately 50 employees. Results collected in the third quarter showed that the trial group outperformed the rest of its department and the rest of the company in almost all (25 out of 27) measured categories. Although causality has not been proved (e.g., because no pre-measures were performed), use of the heuristic approach outlined in this paper so far appears to be associated with dramatic improvements in performance metrics.

## REFERENCES


Biggs, D., B. de Ville, and E. Suen, "A method of choosing multiway partitions for classification and decision trees", *Journal of Applied Statistics,* **18**, 1, 49-62, 1991.

Box, G.E.P., and N.R. Draper. *Evolutionary Operations.* Wiley, 1969.

Buntine, W., 1993. Learning classification trees. In D.J. Hand (Ed), *Artificial Intelligence Frontiers in Statistics: Artificial Intelligence and Statistics IV.* Chapman and Hall, 182-201.

Cox, L.A., Jr., and Y. Qiu, Minimizing the expected cost of classifying patterns by sequential costly inspections. In P. Cheeseman and R.W. Oldford (eds), *Selecting Models from Data: Artificial Intelligence and Statistics IV.* Springer-Verlag, 1994.

Cox, L.A., Jr., and G.E. Bell, "A machine-learning approach to process improvement decision-making," *Annals of Operations Research* (D. O'Leary, ed., forthcoming, 1995).

Hirschleifer, J., and J.G. Riley, *The Analytics of Uncertainty and Information.* Cambridge University Press, New York, 1992.

Lindley, D.V., "Regression and correlation analysis" in J. Eatwell et al. (eds), The New Palgrave: Time Series and Statistics. W.W. Norton, New York, 1990.

Quinlan, J., 1986. Induction of decision trees. *Machine Learning,* 1, 81-106.

Spirites, P., and C. Glymour. Inference, intervention, and prediction. In P. Cheeseman and R.W. Oldford (eds), *Selecting Models from Data: Artificial Intelligence and Statistics IV.* Springer-Verlag, 1994.