

StatSoft®
Business White Paper

Stability of Trees in *STATISTICA Data Miner*

Last Update: March 13, 2003

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Czech Republic: StatSoft Czech Rep. s.r.o.
France: StatSoft France

Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl

Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Poland: StatSoft Polska Sp. z o. o.

Portugal: StatSoft Iberica Ltda.
Russia: StatSoft Russia
Singapore: StatSoft Singapore
S. Africa: StatSoft S. Africa (Pty) Ltd.

Spain: StatSoft Espana
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Client Question

Word is circulating in the DM world that there is anecdotal evidence that Salford System's CART is the only tree method that produces stable trees.

Response

This all sounds very "silly," although perhaps this rumor (as many rumors) may have a kernel of truth (or half-truth as the case may be).

1. The CART (C&RT) algorithm is widely documented and rather straightforward. There are several numerical issues that must be solved when growing trees, that introduce some "instability". For example, which one of two predictors should one choose, when both are exactly of the same quality (lead to identical node purities)? While this is a rare occurrence (particularly in large datasets), this may happen and different programs may choose different predictors. However, in that case there simple is no stable tree, i.e., the ambiguity is inherent in the data.

2. The full implementation of the CART algorithm, as, for example, described in Breiman, Friedman, Olshen, and Stone (1984; "Classification and regression trees." Monterey, CA: Wadsworth), involves v-fold crossvalidation as a method for ensuring a stable final solution (please refer to our electronic manual descriptions for details). In practice, we have seen many applications where analysts (consultants) always use this option to ensure that the final tree is simple and robust. Among the implementations of the CART (C&RT) algorithm, only STATISTICA (!) and Salford's CART program have complete implementations of v-fold crossvalidation (in fact, we use the same technique in many places for precisely the same purpose, i.e., to guarantee stable and valid solutions for machine learning algorithms). Note that SPSS Answertree will only apply v-fold crossvalidation to a final specific solution, but will not apply the technique to the entire tree-building procedure, to find the best (most replicable) complexity of the tree. So we think that this is the issue that started the "rumor", which of course is false (i.e., the STATISTICA implementation is as complete in this instance, while Answertree is indeed deficient).

3. Regarding the claim of "secret black-box algorithms:" We feel that it is highly unethical to attempt to sell to gullible consumers of data mining solutions algorithms that are not understood, and purport to deliver miraculous results. If the methods are not applicable to the particular domain, how will the customer diagnose that? This is really dangerous! We prefer (1) to implement well documented and understood algorithms that have been reviewed by experts in various domains, and (2) to actually provide as much consulting with our data mining solutions as is necessary, to ensure

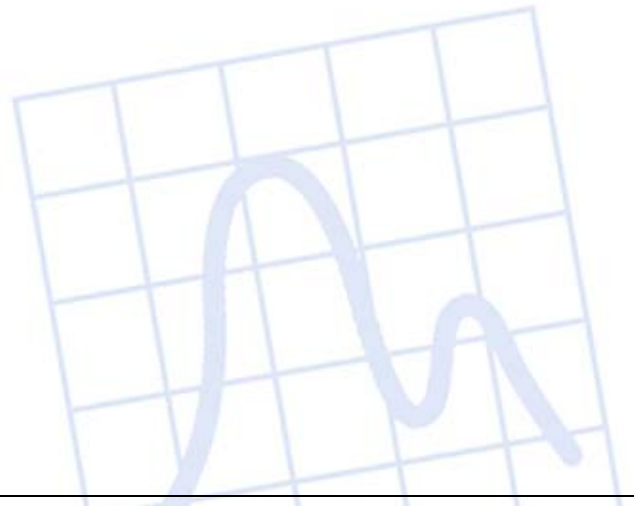
the customer is aware of the strengths and limitations of the different techniques. We are not trying to make a one-time sale of a "miracle box", but to establish a long term relationship with our clients to help them garner the best return on investment (ROI) for there area of application.

I hope we were able to address your concerns,

Regards, Thomas Hill, Ph.D.

PS: We would be most interested in learningsspecifics about the supposed "stability-enhancing " methods, i.e., what exactly is meant by "stability"here, and how is it measured. Further, we would invite competitive comparisons in this respect, i.e., if there are data sets that supposedly only Salford can properly analyze (I assure you there are many types of datasets that we can show where they couldnot derive useful solutions).

PPS. Note that CART is no longer considered a cuttingedge algorithm and while other manufacturers (e.g., SPSS) haven't still offered complete implementations of CART (with true v-fold crossvalidation), we focus on much moredemanding (and involving much more challenging performance optimization algorithms) methods such as our highly successful Stochastic Gradient Boosting Trees (see TreeNet...).



U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Czech Republic: StatSoft Czech Rep. s.r.o.
France: StatSoft France

Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl

Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Poland: StatSoft Polska Sp. z o. o.

Portugal: StatSoft Iberica Ltda.
Russia: StatSoft Russia
Singapore: StatSoft Singapore
S. Africa: StatSoft S. Africa (Pty) Ltd.

Spain: StatSoft Espana
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.