

StatSoft®
Business White Paper

**Boosting as a
General Technique**

Last Update: March 14, 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.

Boosting, TreeNets, and Stochastic Gradient Boosting

1. "Boosting" is a general technique (to successively apply a baselearner to those observations that were not well represented by the model generated in prior boosting steps). The best-known algorithm is the ADA-Boost, which came out of the AI tradition and was very popular about 5 years ago (i.e., researchers were very "excited" about this technique).

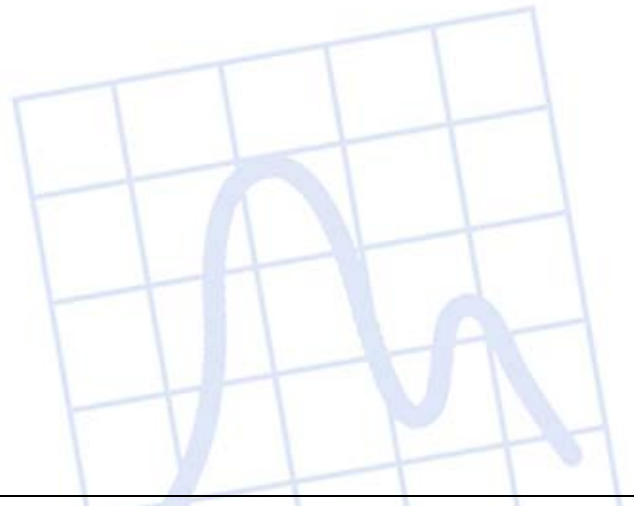
2. You can boost any base learner, e.g., linear regression or neural networks, etc. In particular, there were applications (software) to boost neural nets because of the "implied attraction" of power, i.e.: boost a really general and complex base learner, and thus represent even the most complex relationships.

3. As it turns out -- and our internal tests have also shown this when we experimented with other, more general boosting solutions-- there is no particular advantage to boosting complex (e.g., neural nets) or functionally constrained (e.g., linear models) base learners. In fact, excellent (the best?) solutions are obtained using a very simple base learner -- such as a simple 2-node-regression tree -- and applying it repeatedly to the data (residuals from prior steps). Interestingly, in cognitive science, there is a "parallel" notion based on findings suggesting that complex knowledge of human experts appears to be the accumulation of experiences that yielded a large number of simple rules, i.e., human experts have had a long "history of boosting a simple learner," which allowed them to learn (represent in memory) extremely complex knowledge (much of that research was done by Dr. Lewicki and his colleagues, and his research is now widely accepted among cognitive scientists worldwide).

4. This "insight" led Friedman and his colleagues to propose simple boosting of trees as a general and very powerful method for machine learning. Moreover, the method of stochastic gradient boosting -- applying the simple base learner to successive independent samples from the population-- adds additional safeguards against overlearning, and in practice yields very robust and valid results.

5. To the best of our knowledge, our implementation of stochastic gradient boosting is every bit as powerful and versatile as that available from Salford Systems, and we have encountered many real-world data sets that illustrate the utility of our method. I would not be surprised if over successive implementations, Salford "tweaked" their algorithms by introducing various additional parameters; however, we are not aware of any particular "real" improvements to the basic algorithms. Note that Salford's "problem" is that they do not have all the algorithms available for data mining (unlike *STATISTICA*); hence, I suppose they are likely trying to "adapt" the basic procedure to make it more efficient for instances that can be modeled more easily with, for example, generalized linear models.

In conclusion, if you have specific requests about features (or lack of features), we would very much like to learn about it. To the best of our knowledge, the implementation of stochastic gradient boosting in *STATISTICA Data Miner* is very complete, written to be highly scalable (to very large datasets), and deployable (in C, SVB, or PMML); I don't think that Salford's solution can be as flexibly deployed, or connected to external databases for efficient processing of large datasets.



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.