

(Mar. 1996) John F. Elder, A review of *Machine Learning, Neural and Statistical Classification* (eds. Michie, Spiegelhalter & Taylor; Ellis Horwood, 1994), *Journal of the American Statistical Association* **91**, no. 433: 436-437.

Machine Learning, Neural, and Statistical Classification.

Donald Michie, David J. Spiegelhalter, and Charles C. Taylor (eds.). New York: Ellis Horwood, 1994. xiv + 289 pp. \$68.

Algorithms which construct classifiers from sample data -- such as *neural networks*, *radial basis functions*, and *decision trees* -- have attracted growing attention for their wide applicability. Researchers in the fields of Statistics, Artificial Intelligence, Machine Learning, Data Mining, and Pattern Recognition are continually introducing (or rediscovering) induction methods, and often publishing implementing code. It is natural for practitioners and potential users to wonder, "Which classification technique is best?", or more realistically, "What subset of methods tend to work well for a given type of dataset?". This book provides perhaps the best current answer to that question.

The book grew out of the European *StatLog* project, in which a team of 6 University and 6 Industrial research groups performed a controlled evaluation of a score of procedures on as many example datasets "to determine to what extent the various techniques met the needs of industry" (p. 4). By highlighting strengths and weakness of popular approaches on realistic problems, the project also hoped to contribute to their improvement. Procedures were drawn from the three fields of the book title and their models trained on most of each data set, then tested on the remaining, unseen data. (All fine-tuning of the procedures was to have relied solely on the training sets, with the test data secreted in a separate site, though one lapse of these controls was noted.) The evaluation results were summarized from several angles and then, interestingly, were themselves made a dataset for the purpose of what could be called *meta-modeling*. That is, a rule-based model (an *Application Assistant*) was trained to forecast, from summary features of a dataset, the most promising method(s) for that problem.

That program and a utility facilitating comparison studies (an *Evaluation Assistant*), were deposited in the public domain. All of the datasets, and almost all of the classification algorithms are similarly available, with pointers provided in the Appendices. Though the book has over a dozen contributors (with 60 mentioned as part of the *StatLog* project), the editors apparently required several iterations and revisions to avoid duplication and make it more unified. This effort was largely successful (with exceptions noted below) and greatly enhances its readability. The book should prove to be of strong interest to researchers -- for whom it will no doubt spark a number of studies extending its results -- as well as to practitioners, who will welcome its practical tone, and find much in it of immediate utility.

Synopsis

The book is organized as follows. A brief Chapter 1 (5 pgs.) introduces the aim of the study and its procedures, and provides working definitions of what classifies an algorithm as "statistical" (reliance on an underlying probability model, with human intervention assumed!), "machine learning" (using logical operations to be interpretable), or "neural" (employing layers of interconnected nonlinear nodes). Chapter 2 (11 pgs.) illustrates the task of classification on the well-known Fisher Iris dataset, using three archetypal procedures: linear discrimination, decision trees, and k -nearest neighbors. Each of the 23 procedures studied, the authors maintain, could trace its ancestry back to one of these. (Later, in Chapter 8, it is suggested that a minimal toolbox of algorithms -- essentially, a set spanning "method space" -- should include these 3 as well as projection pursuit and radial basis functions. Elder and Pregibon (1995) further add polynomial networks and adaptive splines.)

Seven of the methods are grouped with linear discrimination: linear, quadratic, and logistic discrimination, multi-layer perceptrons (trained either by backpropagation or cascade correlation), DIPOL92 (a locally-optimized piecewise linear classifier), and projection pursuit. Decision tree methods are the most numerous, represented by 9 programs: NewID, AC², Cal5, CN2, C4.5, CART (the best known to statisticians), IndCART, Bayes Tree, and ITrule. The remaining 7 methods are less homogeneous, but are all related to local probability density estimation: k -nearest neighbors, radial basis functions (RBF), Naive Bayes, Polytrees (a simple causal network), learning vector quantizers (LVQ), kernels, and a Kohonen self-organizing network (actually a clustering method which should not have been included).

This interesting approach to grouping methods is subsequently dropped however, in favor of pedigree. Chapters 3 through 6 briefly describe, respectively, methods from classical statistics (12 pgs.), modern statistics (21 pgs.), machine learning (34 pgs.), and neural networks (24 pgs.), including some which are not tested (ACE, MARS, CHAID, directed acyclic graphs, Gaussian mixture models, and RAMnets) -- due, in some cases, to storage limitations of the then available implementations. Most descriptions are brief but well done; for procedures with which I was familiar, the key "knobs" (parameters worth adjusting) were usually emphasized. Still, the algorithms were basically run on their default values whenever possible in the tests, with the exception of neural networks (as discussed below).

Chapter 7 (18 pgs.) describes how the methods were compared, outlining the error estimation techniques of train-and-test, cross-validation, and the bootstrap (though the last was not used). To

investigate the relation between algorithm performance and type of dataset, several dataset features were calculated. These ranged from simple counts of cases and attributes, to statistical measures, such as tests for homogeneity of covariances and collinearity of the class means, and also included "yardstick" results from the simplest statistical classification methods. The chapter also briefly mentions ways in which other key modelling issues were handled in the tests, including tuning of parameters, dealing with missing values, selecting and extracting features, enforcing class proportionality in training and testing sets, and coding categorical and hierarchical data. These important issues are often neglected in comparative tests (or at least in their reports) but can critically influence the results. A brief Chapter 8 (6 pgs.) mentions many previous empirical comparisons of classification methods, and notes their usual drawbacks (resisted, if not completely eliminated in this study). These can be summarized as: lack of breadth in methods and/or datasets employed, and bias in choice of cost criteria, dataset, "straw-man" opponent, or level of "tender loving care" bestowed during training.

Chapter 9 (44 pgs.) describes the datasets, and (finally) a table of results for each. Datasets fall in the categories of credit risk estimation (2), image recognition (7), image segmentation (3), and those with misclassification cost matrices (11) (which include some medical and control applications). The results are analyzed several ways in Chapter 10 (38 pgs.), and conclusions are drawn about the efficacy of each method in Chapter 11 (15 pgs.). Then, when the book should be over, there instead appear two more papers: Chapter 12 on "Knowledge Representation" (18 pgs.), and Chapter 13 on "Learning to Control Dynamic Systems" (16 pgs.). The claim of the first, that "the choice of knowledge representation formalism is just as important as the choice of learning algorithm" (p. 245), is worth considering in any expansion of this empirical study. Yet, sadly, neither chapter has any connection to the testing project which is the focus of the book. Despite their individual quality, other readers may find it as hard as I did to push through them after the crescendo of results (and their implications) from Chapters 10 and 11.

Discussion of Results

The book has a Machine Learning (decision tree) slant, as evidenced by the relative length of that chapter, the use of more than enough (9) tree-building algorithms, tests on largely ML datasets (which the authors acknowledge occasionally have a structure explicitly set up for threshold-based methods), and use of a tree method, C4.5, on the performance results to build the meta-classifier (though that step is largely justified by the desire for interpretable rules). But they seem to have compensated somewhat to be fair to other techniques. For instance, while the tree methods were largely run with default parameter values, the neural networks had the benefit of using cross-

validation in training to select the number of hidden nodes (though the time might have been more profitably spent selecting variables). This was rather time-consuming, since such networks are already the slowest method to train. Though the neural network approaches were reported to provide "the best or near to best predictive performance in nearly all cases [without cost matrices]" (p. 221), the authors lump a number of methods into this class, and the common backpropagation version did not, in general, fare well. (Though this is largely a function of the datasets employed, it is also my experience that such networks, though a useful general tool, are rarely the winners whenever a suite of sufficiently diverse algorithms are tested.)

Some general conclusions are extracted from the results; i.e., that decision trees do well on credit datasets, k -nearest neighbors excel on problems concerning images, and applications with misclassification cost matrices require algorithms explicitly incorporating them. (Perhaps more could have been learned too, had the dataset features been more robust in the presence of created variables, or better distinguished between the likely efficacy of "global" vs. "local" methods.) Most interesting however, is the study of similarities between the algorithms, using clustering and multi-dimensional scaling of the performance matrix of scaled error rates (22 datasets by 23 methods). Almost all the method-wise variability could be captured in 3 descriptive dimensions (corresponding to the three "strands" of algorithms perhaps?). Also, as one outside of the Machine Learning community might have suspected, the 9 decision tree algorithms were quite close in performance, if not detailed behavior. For most purposes, use of a single leading version (e.g., CART) should suffice.

Readers will naturally wish other interesting algorithms will be added to this study, including commercial tools, which were unfortunately ignored. I'd suggest Polynomial Networks (Farlow, 1984; Elder & Brown, 1994), Multivariate Adaptive Regression Splines (MARS), Generalized Additive Models (Hastie & Tibshirani, 1990), multilinear trees (e.g., OC1; Murthy et al., 1994), and piecewise methods such as Constrained Topological Maps (Mulier, 1994). Also, a promising use of the results would be to examine the benefits of combining the outputs of different methods, which can often lead to particularly robust models.

The editors and other authors (particularly, R. J. Henery of the University of Strathclyde, who wrote several useful chapters) have delivered a readable and concise report of an important study, and it is anticipated that their efforts will be rewarded by strong interest in the book from both researchers and practitioners in this growing field.

John F. Elder IV
Rice University

REFERENCES

- Elder, J. F., IV, and Brown, D. E. (1992). "Induction and Polynomial Networks," Technical Report IPC-TR-92-9, Department of Systems Engineering, University of Virginia. Forthcoming as Chapter 3 in *Advances in Control Networks and Large-Scale Parallel Distributed Processing Models (Vol. II)*, M. D. Fraser (ed.). Norwood, NJ: Ablex.
- Elder, J. F., IV, and Pregibon, D. (1995, in press). "A Statistical Perspective on Knowledge Discovery in Databases," Chapter 4 in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). AAAI/MIT Press.
- Farlow, S. J. (ed.) (1984). *Self-Organizing Methods in Modeling: GMDH Type Algorithms*. New York: Marcel Dekker.
- Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.
- Mulier, F. (1994). "Statistical Analysis of Self-Organization," Ph.D. dissertation, Electrical Engineering Department, University of Minnesota.
- Murthy, S. K., Kasif, S., and Salzberg, S. (1994). "A System for Induction of Oblique Decision Trees," *Journal of Artificial Intelligence Research* **2**, 1--32.