Working Paper Series
ISSN 1170-487X

Machine Learning and Statistics: A matter of perspective.

by Sally Jo Cunningham

Working Paper 95/11
April 1995

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Machine Learning and Statistics: A matter of perspective

Sally Jo Cunningham
Department of Computer Science
University of Waikato
Hamilton, New Zealand
email: sallyjo@waikato.ac.nz

Abstract: Information has become an important commercial commodity–indeed, possibly the most important product of the future. While we have well-developed technologies to store data, the analysis to extract information is time-consuming and requires skilled human intervention. Machine learning algorithms augment statistical analysis by providing mechanisms that automate the information discovery process. These algorithms also tend to be more accessible to end-users and domain experts. The two analysis methods are converging, and the fields have much to offer each other.

1. Introduction

With the increasing availability of computers comes an ever-expanding amount of scientific and commercial data stored in machine-readable form. In 1989, it was estimated that $5x10^6$ computer databases existed in the world. Given that it is also estimated that the amount of information in the world doubles every 20 months, it is likely that both the size and number of databases have grown dramatically since then (Frawley et al, 1991). Unfortunately, at the same time the gap between data *generation* and data *understanding* has increased as well. Accumulating data is not usually the problem–in fact, many organisations literally have more data than they know what to do with! Traditionally, statistical techniques have been used to extract implicit information from data. But effective statistical analysis requires a mathematical background that few database managers or domain experts have. Moreover, statistical analysis is time-consuming, as the analyser must formulate and test each hypothesis individually–a daunting task, given the number of possibilities implicit in even a moderately-sized database. The techniques of *machine learning* have been developed in response to the current pressing need to automate the information discovery process. Typically, machine learning algorithms permit the user to specify the types of information desired, with the analysis conducted autonomously or with minimal human guidance. Machine learning automates (at least partially, if not wholly) the generation of hypotheses, as well as their testing.

The two approaches to data analysis are complementary, rather than contradictory. Machine learning algorithms have a sound mathematical basis, and many directly incorporate statistics into their algorithms. Statistical techniques (notably the CART algorithm) have been independently developed that are fundamentally similar to machine learning, and that produce similar output (decision trees and rule descriptions of a domain). Model validation techniques are the same for both types of analysis.

This paper explores the overlaps between machine learning and statistical analysis of data. Section 2 discusses the commonalities between the two methods, and Section 3 considers the problem of selecting the best analysis technique for a given data set. Section 4 presents our conclusions.

2. Commonalities of machine learning and statistical research

Data analysis methods can be generally classified as exploratory or confirmatory. Exploratory techniques look for "interesting" or "unusual" patterns in data, whereas confirmatory analysis is just that—a pattern is hypothesised to exist in the data, and the analysis confirms or denies its presence [Parsaye and Chignall, 1993]. T-tests and

analysis of variance are examples of confirmatory tests, and factor analysis is a common exploratory technique.

Machine learning algorithms are primarily exploratory. This paper will focus on the most common type of algorithm: supervised learning schemes, which generally produce decision trees or if-then rules that classify instances (tuples of data) by a single attribute. Figure 1 illustrates this paradigm (Witten et al, 1993). Input to the algorithm is a flat table of instances, in which each line of the table corresponds to a unique "case" or "example". Here, each line represents an ophthalmologist's patient, and the various characteristics of that patient are used to determine whether that person should wear hard contact lenses, soft contacts, or no contacts at all.

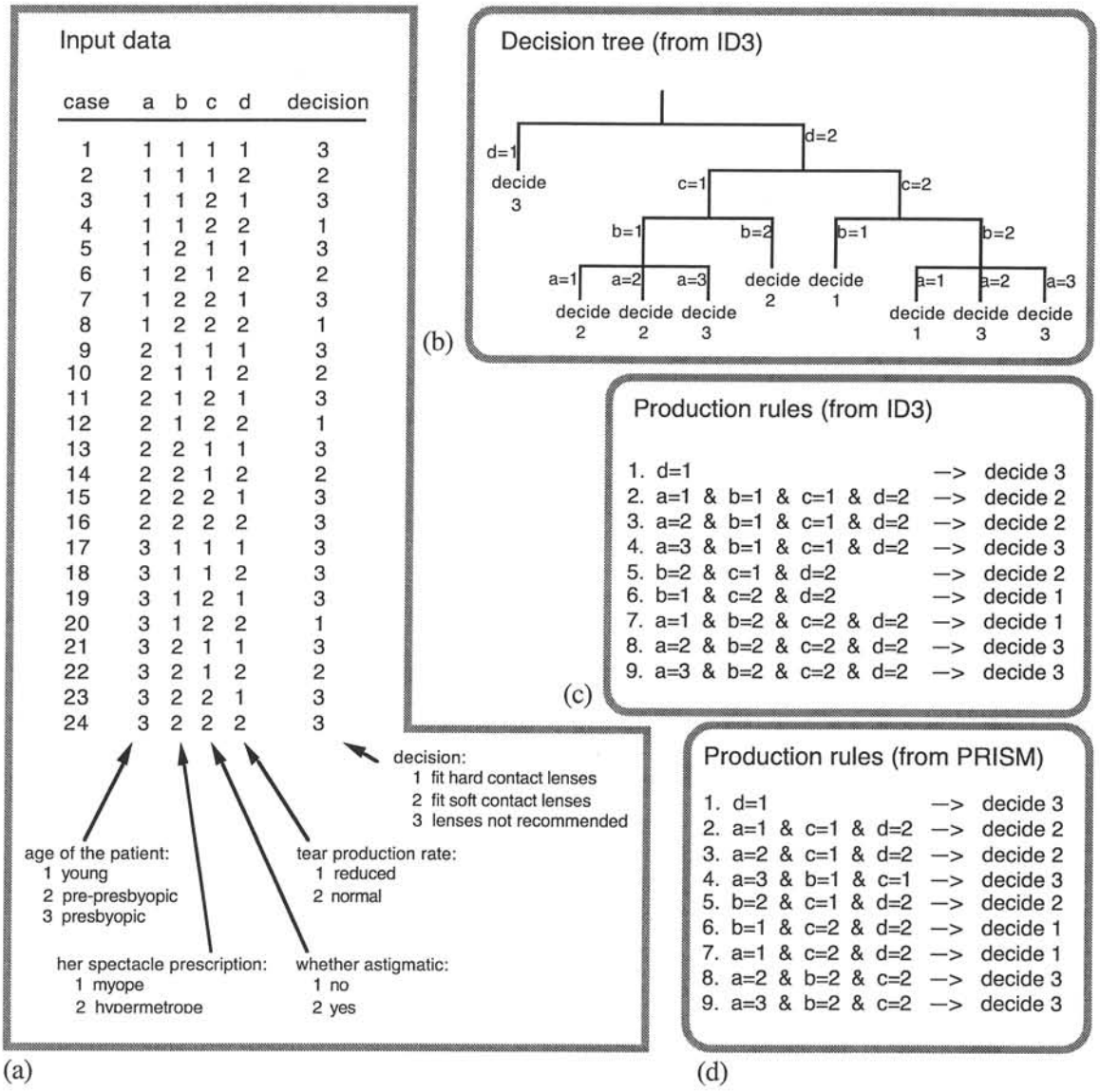

Figure 1. Decision tree and rules (taken from Witten et al, 1993)

The machine learning algorithms attempt to summarise or generalize the classification decisions of the ophthalmologist. A decision tree is read from top to bottom, with each branch of the tree representing a query about the value of the data instance being classified. For example, the first branch of the decision tree in (1b) asks the question, "is the patient's tear production rate reduced?" If the answer is "yes", then the diagnosis is made that the patient cannot tolerate contact lenses; otherwise, the "d=2"

branch is taken and the querying continues. Production rules are interpreted as simple "If..Then" statements, and the user classifies a new patient by matching that person's characteristics to the rule "If" portions. Note that different algorithms can produce different rule sets or decision trees from the same data set.

While these inductive, exploratory methods have been associated primarily with the artificial intelligence/machine learning research community, statisticians have had a minority interest in inductive learning for at least three decades. Statistical inductive learning has received more emphasis in statistics since the introduction of the CART analysis system (Breiman et al, 1984), an independently developed algorithm that is similar to ID3. Indeed, classification tree software has been incorporated into multipurpose statistical packages like SPSS and S very recently (within the last two years) (White and Liu, 1994).

Statistical methods are often directly incorporated into many machine learning algorithms. Statistical analysis is commonly the basis for optimising decision tree/rule construction algorithms; for example, Quinlan's ID3 uses chi-square tests to decide whether an attribute should be added to a classification tree (Quinlan, 1979). Once a model has been produced, it may be over-fitted to the training data (ie, it contains over-specific rules that produce accurate classification results on the training data, but does not contain sufficient generalisations to perform adequately on new data). To overcome this problem, methods have been developed for pruning over-fitted decision trees–methods generally based on the statistical performance of the branches of the tree (Quinlan, 1986). This is a particularly apt example of the benefits that can accrue from greater communication between the two disciplines, since the machine learning researchers were initially unaware of the issues of over-fitting and constructing too-large trees, and apparently re-discovered this problem and its solutions independently (White and Liu, 1994).

Perhaps the most direct benefit that machine learning receives from statistics is the model validation techniques that have been borrowed wholesale, such as n-fold cross-validation and bootstrapping. While earlier machine learning literature focussed on development of novel algorithms, as the field has matured the emphasis is coming to bear on understanding existing algorithms and applying these algorithms in a more principled fashion. Mature methods from statistics are directly applicable to these problems (White and Liu, 1994).

In return, the artificial intelligence community has techniques of its own to contribute to the field of statistics–primarily methods for making computationally expensive algorithms more practical for real world applications. For example, the main difficulty in automating hypothesis generation is that usually far too many generalisations can be produced, only a small proportion of which are of interest to the analyst or can be explored in a reasonable amount of time. Artificial intelligence can mitigate this problem by offering a deep understanding of state space search techniques and the use of heuristics to prune search spaces.

A second significant example of the application of AI techniques is the development of case-based reasoning and instance-based learning. Both techniques are based on the K-Nearest Neighbours algorithms, which can be traced back to the early 1950's and have been investigated deeply by the statistical community. But where the nearest neighbour algorithms have prohibitively large storage requirements, the machine learning techniques significantly reduce the storage costs at the expense of minor reductions in learning rate and classification accuracy (Aha et al, 1991; Kolodner, 1993). This type of trade-off is typical of machine learning research, which tends toward approximate algorithms that execute far more efficiently than exact algorithms, while producing classifications or predictions that are only slightly worse.

3. How can analysts choose between statistical and machine learning techniques?

Given the a choice between machine learning and statistical algorithms, which should be used to analyse a particular data set? Moreover, if machine learning is selected, which of the many existing algorithms will produce the most reliable results for that data set?

An immediate answer to the first question often lies in the nature of the desired result. Chiogna (1994) points out that, "The symbolic techniques developed in the machine learning field tend to give a deterministic answer to the problem of classifying new observations; the statistical techniques focus on the estimation of probabilities of the possible results." Further, statistical models tend to be tailored for continuous attributes, whereas machine learning models in general are suitable for discrete attributes (with continuous values clustered to form a limited set of discrete values). Thus statistical methods often provide interpolative or extrapolative estimates for classifications (eg, "for the given parameters, the expected wheat yield is 5.63 bushels/acre"), while machine learning models commonly provide predictive ranges (eg, "the expected wheat yield is low, where low is in the range [0, 8] bushels/acre").

The main performance criteria for a data analysis algorithm–statistical or machine learning–is the error rate of classifications produced by the induced system models. Most descriptions of novel machine learning algorithms indicate the effectiveness of a new technique by testing its classification ability on one or more of the databases in The Machine Learning Database Repository, a large testbed of data sets maintained by the University of Irvine.[1] Machine learning as a field has come under criticism for providing algorithms with relatively little evaluation, and for failing to compare machine learning techniques to other types of data analysis (known informally as the "one dataset, one algorithm; one algorithm, one dataset" syndrome). Recently, however, more formal, empirical studies of the relative effectiveness of machine learning and statistical techniques have been performed (Feng, 1993; Weiss and Kulikowsi, 1992; Michie et al, 1994). The most notable of these is the StatLog project (Michie et al, 1994). StatLog applied each of 20 machine learning, statistical, and neural network analysis procedures to 20 datasets, in an attempt to improve on the ad hoc or small-scale evaluations that had taken place previously.

Not surprisingly, no one technique emerged as the victor in the StatLog trials. The results of using a data analysis technique were found to be dependent on three factors: "the essential quality and appropriateness of the technique; the actual implementation of the technique as a computer program; and the skill of the user in coaxing the best out of the technique" (Michie, 1994, p. 5). The second factor is a problem that, like the poor, we will always have with us. It will be alleviated as machine learning toolboxes or workbenches become more widely available, allowing analysts to use standard and well-tested implementations of common algorithms (Holmes et al, 1994). The first factor is currently a subject for research. Even in the much more mature field of statistics, choosing an analysis technique can be an art. This problem is much more acute in machine learning, where the fundamental characteristics of the algorithms are not as well understood.

The third factor–the skill of the user–is of particular significance for machine learning. One of the major strengths of this type of data analysis is that its implementations are particularly well-suited for use by domain experts, rather than data analysis experts. The decision tree and if-then rule outputs of machine learning algorithms are much more readable than the standard output of statistical packages. Consider, for example, the analysis of Fisher's classic Iris database presented in Figure 1 (taken from Parsaye

---

[1] At location <URL:http://www.ics.uci.edu/AI/ML/MLDBRepository.html>

and Chignell, 1993). This database contains 150 records, 50 for each of three species of iris (Setosa, Versicolor, and Virginica). Each record contains four attributes: petal and sepal length, and petal and sepal width. The objective is to use these characteristics to classify an iris by species.

(i) Sample rules derived by machine learning techniques:

Rule 1: If 4.8" <= petal length <= 6.7" and
1.8" <= petal width <= 2.5"
Then species = Virginica

Rule 7: If 1.7" <= petal length <= 4.9" and
0.6" <= petal width <= 1.7"
Then species = Versicolor

Rule 15: If 1" <= petal length <= 1.9"
Then species = Setosa

(ii) Output of analysis using a statistical package:

SUM OF PRODUCT MATRIX M = G'A' [A(X'X)-1)-1 AB (Hypothesis)

	S-LENGTH	S-WIDTH	P-LENGTH	P-WIDTH
S-LENGTH	61.332			
S-WIDTH	-15.583	14.193		
P-LENGTH	163.141	-52.047	417330	
P-WIDTH	73.197	23.239	175.126	84.230

MULTIVARIATERESULTS

HOTELLING-LAWLEY = 35.727
FSTAT = 584.923 DF = 8286 PROB = .000

WILKS' LAMBDA = .033
FSTAT = 196.491 DF = 8288 PROB = .000

PILLAI TRACE = 1.219
F-STAT = 56.636 DF = 8300 PROB = .000

THETA = .708 S = 3 M = .6 N = 70.1 PROB = .000

(SUM1-1) 2P2+3P-1
RHO = 1.0-(N(J)-1N-G) 6(P+1)(G-1)

Figure 1. Comparison of output of statistical and machine learning packages for Fisher's Iris Data (Parsaye and Chignell, 1993)

Obviously the rules are more easily understood by a layperson! Further, the generalisations for this database are more conveniently summarised by rules than equations. Rule descriptions are of the form "when these attributes have these values you can expect the flower to be species X", whereas the statistical package gives information in a less humanly-accessible format such as "if you multiply the petal length by three and add it to the petal width and the resulting number is between 12 and 15, then the flower is most likely species X" (Parsaye and Chignell, 1993, p. 197).

Parsaye and Chignell also provide a number of case studies illustrating the importance of human-friendly output. The underlying theme of these studies are that end-users can

perform useful analysis in a more timely fashion (and often more cheaply) than if a statistical consultant is involved. This is not, of course, to say that machine learning makes statisticians obsolete! Rather, the argument is that for well-defined problems it can be more efficient to permit domain experts to directly investigate their own data.

Finally, the difference between the two types of output is not merely a matter of presentation: machine learning rules are value-based, whereas the statistical results are trend-based ("variable A will tend to increase as variable B increases"). Clearly, neither approach is inherently superior. The suitability of a given approach to the problem depends on the nature of the problem.

4. Conclusions

Statistics and machine learning attempt to solve many of the same problems in data analysis, and have begun converging in their approaches. Since the late 1980's this convergence has been enhanced by formal efforts to disseminate new results across both fields, and to promote cross-disciplinary collaboration. Notable venues for these goals include the biannual Artificial Intelligence and Statistics conferences, the Machine Learning and Statistics Workshops sponsored by the European Conference on Machine Learning, and the numerous machine learning sessions included in statistics conferences.

Both fields have much to offer each other. In machine learning, preliminary statistical analysis of data is becoming more commonly reported in descriptions of practical machine learning applications. This trend was given impetus by the startling discovery that a classifier that bases its decisions on a single attribute often performs as well as more sophisticated algorithms! It appears that at least part of the reason for this result is that many of the standard test databases embody very simple underlying structures (Holte, 1993)–a conclusion that is confirmed by statistical analysis. Statistical visualisation and data smoothing/clustering methods are also being formally incorporated into machine learning toolboxes and workbenches (Tsatsarakis and Sleeman, 1993; Holmes et al, 1994). As with the experience with over-fitting of data described in Section 2, the machine learning community appears to be re-discovering problems long understood by statisticians. Perhaps this time the mathematical solutions will be adopted more quickly!

Machine learning can contribute experience in tailoring output for end users with relatively little mathematical expertise. While it is obviously dangerous to expect a naive user to adequately analyse a variety of datasets, experience shows that domain experts can produce useful and timely analyses with a minimum of training (Parsaye and Chignell, 1993). And given the current vast accumulation of unexplored databases, the involvement of database managers and users is perhaps the best technique for extracting information from this raw data.

Intriguingly, machine learning may ultimately provide meta-information about the suitability of a variety of analysis algorithms to a given data set. Performance information is accumulated from evaluative programs such as StatLog that characterises datasets by certain features (such as attribute type, amount of unknown values, etc.) and associates these with classification accuracy for an analysis technique. This information can be easily expressed as a table, from which machine learning algorithms can "bootstrap" information about the features that are associated with enhanced results for each analysis technique. While this technique appears promising, it requires a more extensive set of test results than is currently available (Michie et al, 1994).

References

Aha, D.W., Kibler, D., and Albert, M.K. (1991) "Instance-based learning algorithms", *Machine Learning*, 6(1), pp. 37-66.

Breiman, L., Freidman, J.H., Olshen, R.A., and Stone, C.J. (1984) *Classification and Regression trees*, Wadsworth Press.

Chiogna, M. (1994) "Probabilistic symbolic classifiers: an empirical comparison from a statistical perspective", *Workshop on Machine Learning and Statistics*, workshop of the Seventh European Conference on Machine Learning, Italy.

Feng., C., Sutherland, A., King, S., Muggleton, S., and Henery, R. (1993) "Comparison of machine learning classifiers to statistics and neural networks", *Proceedings of the Third International Workshop in Artificial Intelligence and Statistics*, Ft. Lauderdale (FL, USA), pp. 41-52.

Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C.J. (1991) "Knowledge discovery in databases: an overview", in *Knowledge Discovery in Databases*, ed. by G. Piatetsky-Shapiro and W.J. Frawley, AAAI Press, 1991, pp. 1-27.

Feelders, A., and Verkooijen, W. (1995) "Which method learns most from the data? Methodological issues in the analysis of comparative studies", *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, Ft. Lauderdale (FL, USA), pp. 219-225.

Holmes, G., Donkin, A., and Witten, I.H. (1994) "WEKA: a machine learning workbench", *Working Paper 94/9*, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Holte, R.C. (1993) "Very simple classification rules perform well on most commonly used datasets", *Machine Learning*, 11, pp. 63-91.

Kolodner, J. (1993) *Case-based Reasoning*, Morgan Kaufmann.

Michie, D., Spiegelhalter, D.J., and Taylor, C.C., eds. (1994) *Machine learning, neural and statistical classification*, Ellis Horwood.

Nakhaezadeh, G. (1994) "Interaction between machine learning and statistics, an overview" (abstract only), *Machine Learning and Statistics Workshop of the ECML-94* (European Conference on Machine Learning), pp. 1-3.

Parsaye, K., and Chignell, M. (1993) *Intelligent database tools and applications*, John Wiley & Sons.

Quinlan, J.R. (1979) "Discovering rules by induction from large collections of examples", in *Expert Systems in the Micro-Electronic Age*, edited by D. Michie, Edinburgh University Press, pp. 168-201.

Quinlan, J.R. (1986) "Induction of decision trees", *Machine Learning*, 1(1), pp. 81-106.

Tsatsarakis, C., and Sleeman, D. (1993) "Supporting preprocessing and postprocessing for machine learning algorithms: a workbench for ID3", *Knowledge Acquisition*, 5, pp. 367-384.

White, A.P., and Liu, W.Z. (1994) "Statistical aspects of tree-based classification systems", *Machine Learnng and Statistics Workshop of the ECML-94* (European Conference on Machine Learning), pp. 1-12.

Witten, I.H., Cunningham, S.J., Holmes, G., McQueen, R. and Smith, L.A. (1993) "Practical Machine Learning and its Application to Problems in Agriculture", *Proceedings of the 13th Conference of the New Zealand Computer Society*, vol. 1, Auckland, New Zealand, pp. 308-325.