User satisfaction with machine learning as a data analysis method in agricultural research

Robert J. McQueen
Department of Management Systems
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

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

Lyn Hunt
Department of Statistics
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Abstract Machine learning has potential use in the understanding of information hidden in large datasets, but little is known about user perceptions about the use of the technology. In this study, a number of datasets were solicited from agricultural researchers and processed using a machine learning workbench. The results were reported to the researchers, and then interviews were conducted with some of them to determine their perceptions about the use of machine learning as an additional analysis technique to traditional statistical analysis. A number of themes about their satisfaction with this technique were constructed from the interview transcripts, which generally indicate that machine learning may be able to contribute to analysis and understanding of these kinds of datasets.

Keywords machine learning; user satisfaction; statistical analysis; agricultural research

INTRODUCTION

This study reports on the user satisfaction of scientists with machine learning analysis of their experimental data in the domain of agricultural research. Machine learning is an information analysis tool that has potential for aiding the understanding of data collected from research experiments, but is seldom used in this environment in comparison with traditionally accepted methods of statistical analysis. In part, this may be a result of a low awareness of the characteristics of machine learning by these scientists, as well as the absence of a strong motivation to move away from the traditional, and well accepted statistical techniques.

Statistical methods for collecting, classifying, and analysing data from experimental situations have a strong tradition in agricultural research. The principles governing the planning and analysis and interpretation of experimental data are well documented in the literature, and with the availability of personal computer software packages for performing statistical analyses, many researchers routinely use statistical techniques in their research. However, statistical software may sometimes be used mechanically and without an understanding of the principles and the assumptions of the technique that is used (Maindonald & Cox 1984). There is a need for an informed attitude for data analysis, presentation, and interpretation for those involved (Camden et al. 1982).

With the ease of automatic collection and recording of data, large amounts of data may be collected, which creates problems for the researcher such as determining the quality of data sets, and being overwhelmed with large numbers of outliers in regression. It is considered prudent that before an experiment is commenced and data are collected, a statistician is consulted to clarify the objective of collecting the data and to advise on the experimental design.

Statistical approaches to classification are generally characterised by having an explicit underlying probability model, which gives the probability of an individual being in a class as opposed to a classification. They require human intervention with regard to variable selection. Multivariate analysis techniques are still used relatively infrequently in agronomic research, partly because of lack of knowledge of how to deal correctly with the data collection for multivariate analysis and with associated problems which can arise using multivariate analysis (Matthew et al. 1994).

Machine learning is a new analysis technique that has potential to assist agricultural researchers, and is often compared and contrasted with traditional statistical analysis methods. It is at the analytical stage, after the experimental design has been determined and data collected, that the two fields of statistics and machine learning appear to diverge. At this stage, data analysis can be either *confirmatory*, where a pattern has been previously hypothesised before data collection, and analysis confirms or denies its existence, or *exploratory*, where interesting or unusual patterns are sought in data, often after the data collection has occurred. In statistical data analysis, t-tests and analysis of variance are examples of confirmatory analysis, and factor analysis is a common exploratory technique. Machine learning algorithms are predominantly exploratory.

The two fields share a surprisingly similarity and have, at the research level, come together more in the last five years. Decision trees (Quinlan 1986) are similar to classification and regression trees which came from the statistical community (Breiman et al. 1984), and many statistical techniques are commonly incorporated into machine learning algorithms. The statistical community has provided frameworks for validating structural descriptions such as *n*-fold cross-validation. Machine learning provides methods for making computationally expensive algorithms more practical for real world applications. The divergence in the two fields is partially in the techniques employed but mainly in the perception held by users that statistical techniques are the only true analytical tools for their experiments.

The fundamental goal of machine learning algorithms is to discover meaningful relationships in a body of training data, presented as individual examples, and to produce a generalisation of those relationships that can be used to interpret subsequently presented test data. We concentrate on the most commonly performed exploratory task in machine learning—classification. The examples (often called instances, and presented as rows of data in a spreadsheet file) each comprise a set of attributes and a classification (presented as column entries for each instance). For example, in learning the factors that determine apple bruising the attributes might be energy level—a measure of the drop height of the apple, location—which part of the apple was bruised, and radius of curvature—a measure of the shape of the apple, and the classification would be a record of the actual bruise size for the apple in the experiment—B0 indicates that the bruise size is under the commercially accepted limit of 1.0 cm^2, B1 indicates significant bruising. An example of a few instances (from a large table) with three attributes and a classification is given in Table 1.

Table 1. Examples of training data instances.

Energy Level	Location	Radius of Curvature	Bruise Class
6	2	43	B1
6	2	46	B1
6	2	50	B1
1	3	27	B0
1	3	29	B0

To prepare the dataset for processing by a typical machine learning scheme, the dataset is randomised and then split into two partitions, a training set and a test set. The test set is conditionally chosen to reflect the distribution of class values in the dataset as a whole. The training set is processed by the machine learning scheme to produce a model, which is then tested for accuracy on the test set. This process is usually repeated a number of times and the results averaged to produce an overall accuracy. Typical splits are 90% training, and 10% test data, and the process is repeated 100 times (10 repetitions of 10 folds).

The training process constructs a classification model as a set of *production rules*. These rules take the form IF ‘L’ THEN ‘R’, where L (the left-hand side of the rule) is a conjunction of attribute-based tests and R (the right-hand side of the rule) is a class. These rules together with a default class (the most frequently encountered class label) are sufficient to predict the class of unseen cases or test data (see below)—data with attribute values but no class-attribute value. This is achieved by finding the first rule (the rules are ordered) whose left-hand side is satisfied by the case. The predicted class is then the one nominated by the right-hand side of this rule. If a case *falls through* all of the rules then we designate the class to be the default class. The following rules were generated by using 1536 instances with just the Energy Level and Location attributes to predict the bruise class:

Rule 1: IF Energy Level = 1
THEN Bruise Class = B0 (96.7%)

Rule 2: IF Location = 2 OR Location = 4 AND Energy Le**vel** = 2
THEN Bruise Class = B0 (86.6%)

Rule 3: IF Energy Level > 2 THEN Bruise Class = B1 (91.5%)

Rule 4: IF Location = 1 OR Location = 3 AND Energy Level > 1
THEN Bruise Class = B1 (88.6%)

Default Rule: Bruise Class = B1

The figures in parentheses are estimates of the accuracy of each rule based on the training data. To properly test the rules it is necessary to train and test random samples of the data a number of times (*n*-fold cross validation).

The rules would classify the first instance in Table 2 as B1 by Rule 3, the second as B1 by Rule 4, the third as B0 by Rule 1, the fourth as B0 by Rule 2 and the final case as B1 by Rule 3.

Table 2 Examples of instances to be classified.

Energy Level	Location	Bruise Class
5	1	?
2	3	?
1	2	?
2	4	?
6	1	?

For research scientists the attraction of this type of analysis is the descriptions they receive of their data. The attributes are typically measurements from an experiment, and the descriptions tell them which attributes are the most significant and how those attributes combine to determine a classification. The descriptions also focus the scientist on the outcome of the experiment because they are forced to state in the classification attribute exactly what it is that they are trying to do. Standard statistical techniques can be used after preliminary machine learning to verify the accuracy of any hypothesis that might be suggested in the structural description of the data.

While machine learning tends to focus on correlations rather than causality there is a sense in which causality is implicit to a given application. When collecting data an expert chooses properties or attributes that they expect to be causally related to the classification task. A classification at some time *t* from measurements made at that time have to be measurable at that time so that a classification can be made. This is another sense in which causality is implicitly defined. Explicit definition of causality is also addressed in other branches of machine learning where attempts are made to learn causal models (Pearl 1985), however, the software used in this study does not explicitly seek to uncover causal relationships.

Adoption of machine learning technology has been reported in a number of application areas. Langley & Simon (1995) described a number of cases of reported applied use of machine learning rule induction, some of which are shown in Table 3.

Table 3 Applied examples of machine learning.

ML Application	Rule induction for.....
Celestial objects (1995)	Automatic classification of Sky Survey images
Chemical process control (1984)	Manufacturing process settings to guide operators
Circuit boards (1987)	Diagnosis of faults
Credit card (1993)	Approval of credit card applicants
Credit decisions (1989)	Loan approval of credit applicants
Diagnosis of pumps	Vibration data analysis and prediction
Equipment configuration (1990)	Configuration of fire detection equipment
Gas from oil separation (1987)	Separation vessel configuration
Healthcare (1994)	Profiles of claims for a given diagnosis
Military	Prediction of capabilities of military units
Mortgages (1994)	Prediction of payment of overdue payments
Printing (1990)	Avoidance of banding on presses
Process monitoring (1989)	Quality of oil/water cooling emulsions
Sharemarket (1994)	Advice on share trading
Telephones (1994)	Fault diagnosis on public payphones
Transformer oil (1984)	Prediction of insulating oil breakdown

Although these have been reported as successful implementations of rules derived from machine learning rule induction, they are only positive examples. There seems to be no parallel evidence on machine learning systems that have been attempted, and discarded, or the reasons for success or failure. We have no evidence of why the process engineers and other users involved with these applications saw machine learning as a potential contributor to their problem solutions.

The examples above of fielded use of machine learning are encouraging, but much of the activity and publication in machine learning research is on algorithm development, rather than on how to develop successful applications in field domains. The objective of machine learning algorithm development is to create schemes that will generate a model based on a set of learning data which will produce low error rates when test data are processed against that model. Many reports on the accuracy and suitability of various schemes for various kinds of data have been published (Taylor et al. 1994; Kohavi et al. 1997). However, little has been published on measuring the effectiveness (as opposed to accuracy) of machine learning schemes.

The New Zealand-developed WEKA (Waikato Environment for Knowledge Analysis) machine learning workbench provides an integrated environment which gives easy access to a variety of machine learning techniques through an interactive interface (Holmes et al. 1994). The software has been designed primarily to support a research scientist as the end user. The workbench is not a single program, but rather a set of tools bound together by a common user interface. The current system runs on a variety of platforms under the Unix operating system.

The first step in using the workbench is to invoke the workbench conversion programs to prepare data (usually provided in spreadsheet or database format) into a file format which all of the different machine learning techniques recognise. It is then possible to run a variety of machine learning schemes on the data and to view and compare the results. The schemes work in many different ways and no single scheme will outperform the others on all datasets. The schemes currently supported by the workbench are given in Table 4.

Table 4 Machine learning schemes supported by the WEKA workbench.

Scheme	Description
1R & T2	Learns simple rules using one or two attributes only.
Autoclass	Unsupervised Bayesian clustering
C4.5	Learns decision trees
FOIL	Learns relational rules
IB1-4	Family of nearest-neighbour instance-based learners
Induct	Learns complex ripple-down rules
K*	Instance-based learner based on Kolmogorov complexity
M5	Learns regression model trees
PEBLS	Case-based learner

The workbench also provides facilities for data visualisation using the XGobi system (Swayne et al. unpubl. data), an attribute editor which allows users to transform existing attributes into new attributes (for example, converting date of birth to age), and an experiment management facility which allows a user to maximise their computing resources when running every scheme over a large dataset.

METHODS

About 100 New Zealand agricultural research scientists were sent a letter which gave a short explanation of machine learning, and asked whether they had any datasets of past research work that they would be willing to give to us for analysis. About 30 of these responded positively, and 14 useable datasets were eventually received. Selection as a "usable" dataset required at least 50 instances, as well as the condition that the number of attributes did not exceed the number of instances. The datasets were usually received in spreadsheet form, and machine learning technical staff converted them to the attribute relation file format (ARFF) used by the WEKA workbench, which is a header containing a description of each attribute type followed by the data in a comma separated variable format. The staff then ran these files through the WEKA workbench on several of the machine learning schemes, and chose one of these based primarily on output accuracy, but also on the ability of the scheme to provide an explanation of the results. The technicians who prepared and processed the datasets through the machine learning workbench had little domain knowledge about the meaning and significance of the dataset attributes (other than the attribute name) and no knowledge about expected results. This was consciously done to avoid "guiding" the machine learning schemes to generate the expected result. Tuning parameters in each of the schemes were adjusted through a number of runs, until what looked like a reasonable result was obtained. A short report was prepared for the researcher on each dataset processed which contained the results and decision trees. An overall summary of the results for all datasets was compiled (Thomson & McQueen 1996).

Typically, the datasets that were provided were from field research that had previously been undertaken, in which the field data were collected then analysed using statistical techniques, and the results published. Our objective in soliciting these datasets was to understand the nature of the data collected in typical agricultural research, with two main motivations.

1. We expected that these field data would have characteristics such as missing values, incomplete data, error, and noise, which might provide our machine learning research team with insight into how better machine learning schemes and user interfaces could be constructed to tackle these real world problems.
2. If the machine learning analysis on these test datasets generated results which were interesting, we hoped that this would stimulate the research scientist's interest in using machine learning techniques as part of their research methodologies, including their future design of field experiments and data collection to take advantage of machine learning capabilities.

Interviews were arranged with eight of these researchers, selected mainly through availability and interest in the study, and conducted either face-to-face or by telephone. The interviews were recorded and transcribed. The interviews were semi-structured, with a set of questions asking about the nature of the original investigation, original expected results, perceptions about the machine learning analysis results presented to them, initial and present perceptions about machine learning, and perceptions about machine learning as a possible future analytical tool for their research.

Characteristics of the datasets

Table 5 gives an overview of the datasets tested. Instances refers to the number of observations, or rows in the dataset, while attributes refers to the number of variables, or columns recorded for each observation instance. In some cases, additional attributes were derived, or calculated from the original attributes in the provided dataset prior to machine learning runs, and these are indicated by *. The number of attribute types for each dataset are indicated according to the following brief descriptions.

Table 5 Datasets processed through the machine learning workbench.

Dataset Name	no. of Instances	no. of Attributes	no. Enumerated	no. Real	no. Integer
Apple	1662	16	7	6	3
Bulls	90	8	3	5	
Eucalypt	736	20	6		
Grub Damage	155	9	5	3	1
Grub Rain	19	6	1	5	
Grower	22	15	1	14	
Pasture	36	25	2	15	8
Pea Seed	51	15	1	13	1
Slugs	100	9*	5	3	1
Squash	261	24	3	6	16
Valley	878	15	1	13	1
Venison	21 448	23	13	1	9
Wasp	506	13	6	5	2
White Clover	63*	32*	5	27	

Enumerated: Data point is one of a set of discrete numbers or symbols, not necessarily in a continuous range. Numbers have no inherent meaning as a value other than symbolic. Comparison tests only on equality. The attribute value represented by the number 4 is not twice as big as the number 2. Examples: 1, 4, 9.2, high, a+, yes.

Real: Data point is in a continuous range of decimal numbers. Any number in the range is valid. Comparison tests include greater than, less than. Examples (for a range between -4.0 and 9.5): 4.1, 6.324, -2.5.

Integer: Non decimal number, usually positive. Comparison tests on greater than, less than. Numbers have order.

RESULTS AND DISCUSSION

Perceptions about machine learning from the interviews

This section extracts a number of themes about the effectiveness of machine learning from the transcripts of interviews with six of the agricultural research scientists who contributed datasets to this study. Seven themes illustrating perceptions of both strengths and weaknesses of machine learning are developed, paraphrased, and generalised from an analysis of their statements. For some, excerpts from the interviews are included to provide further illustration of the theme. It is not intended that these themes can be generalised and applied outside of the boundaries of this study, but they may provide some insight into what arm’s length users of machine learning perceive to be strengths and weaknesses of the technology. These themes may also prove useful in providing a starting point for similar studies in other application domains.

1. For machine learning, there may be a high return from a small amount of time put into preparing the data for machine learning analysis, especially when

compared with the large amount of time necessary to prepare data for conventional statistical analysis.

There is often a large front-end time commitment in preparing experimental data for statistical processing. Part of this time may be in narrowing the statistical analysis to a small area of potential interest, running the analysis, and then shifting to the next area of interest. Machine learning can usually take all of the data in one pass, and produce results, with relatively little front-end preparation and selection. The issue then becomes the increase of effective understanding of the data resulting from statistical or machine learning analysis, versus the time spent preparing for that analysis.

2. Machine learning can handle large sets of experimental observations more conveniently than statistical analysis.

Of the datasets processed in this research, half had 100 or less instances, which would normally be well suited to statistical analysis. Interestingly, this is often a typical size of dataset that is used as a test dataset in machine learning algorithm development research. However, the other half of our test base of 14 datasets had more than 100 instances, with the largest being over 21 000. The number of instances used has little impact, except for processing time, in machine learning analysis. Sometimes, the number of instances and variables measured in a piece of field research may be constrained because of the effort required to prepare large amounts of data for statistical processing.

3. Machine learning can help to narrow the focus of detailed investigation.

Apart from the use of machine learning as a single analysis approach, there may be the potential for using it to see the overall meaning in a set of data, or to locate a particular area of interest, and then focus in on that area using statistical analysis tools. One researcher saw potential in this dual approach.

> “[...we can use machine learning to] try and clarify the trends and things which might help us narrow our focus a bit more on what’s important and what parts of it wasn’t .... no one wants to do a hundred apples or two hundred apples if he can just do a small sample, measure certain things and be within a certain percentage of the real number. We’re very impressed with what we see [from the machine learning output]. It’s distinguishing between some things which ... we’d be guessing to get out of a statistical analysis ... there is a difference between [the] low energy drops and [the] high energy drops and then there’s a different set of rules for them. And that’s what I can see from looking at those [decision rulesets].”

4. Machine learning output has a sequence or hierarchy construction, which can be helpful to understanding the data and guiding further detailed statistical analysis.

Aside from the visual characteristic of machine learning output, it also has a structure and a hierarchy, which can aid in understanding which relationships are at higher and lower levels, and not just correlated. One researcher found the structure of the machine learning output quite helpful:

> “... it’s got a sequence in it which isn’t present in a numerical analysis. A numerical analysis you just get the totals at the bottom or the scores but

> there’s not the sort of progression through ... statistical analysis doesn’t give you any sort of sequence”

5. Because machine learning is not broadly used, publication of research results might be difficult for other researchers to understand.

As the interviewees were research scientists, publication of results in peer reviewed journals and conferences is an important part of their activities. There was concern expressed that publication of machine learning based results would be more difficult, and might have reduced status, because of the traditional convention of using statistical analysis to present research results. One researcher commented...

> “One of the difficulties I would see with machine learning is that if you published it people wouldn’t be used to judging [it effectively], in the horticultural scene any way .... If you were to publish machine learning analysis ... you’re already facing some sort of barriers to people’s comprehension of what you’re doing ...”

6. Decision trees would be useful for presentations of results, but statistical analysis would remain as the predominant tool for journal publication of results.

The graphical decision tree output was seen as attractive, but perhaps more for presentation of results in meetings, rather than as the main analytical technique used for publication of results in papers.

7. Some researchers would be interested in using machine learning in future studies, provided they could learn more about how it should be used.

The first experience with machine learning had generated interest in using machine learning in subsequent studies. When asked whether they would be interested in using machine learning again, most of the interviewees said they would like to try it again.

> “Well, we’re really keen to use it again. We would consider designing research [to make use of machine learning]. I think it could be more appropriate than more traditional techniques for some particular purposes ...”

CONCLUSIONS

The perceptions of agricultural research scientists who participated in this study were both positive and negative about how effective machine learning would be in analysing experimental data, and presenting the outcomes in presentations and publications. This study involved “arms length” interaction between the machine learning technician and the research scientist, and future studies of this kind might be directed toward more closely tied, interactive collaborations between machine learning expert and research domain expert to see if some of the shortcomings identified here may be overcome.

Machine learning has a role alongside statistical techniques for data exploration and deserves wider use in agricultural studies. The adoption of machine learning techniques may well be best served by combining exploratory and confirmatory analysis tools that can generate hypotheses from data using machine learning and then test their significance using accepted statistical techniques. By producing more synergy between the two fields it may be possible to enjoy the benefits of data explanation and rapid hypothesis testing in an environment grounded in statistical rigour.

ACKNOWLEDGMENTS

This work was supported in part by the New Zealand Foundation for Research, Science and Technology.

REFERENCES

Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. 1984: Classification and regression trees. Belmont, CA., Wadsworth International Group.

Camden, M.; Jorgensen, M. A.; Vere-Jones, D. 1982: Statistical and other numerical skills required by technicians: A position paper of the New Zealand Statistical Association. *The New Zealand statistician 17(1)*:15-23.

Holmes, G.; Donkin, A. R.; Witten, I. H. 1994: WEKA: A machine learning workbench. *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems* Brisbane, Australia. Pp. 357-361.

Kohavi, R.; Sommerfield D.; Dougherty J. 1997: Data mining using MLC++, a machine learning library in C++. *International journal of artificial Intelligence tools 6*: 537-566.

Langley, P.; Simon, H. A. 1995: Applications of machine learning and rule induction. *Communications of the Association of Computing Machinery 38(11)*: 55-64.

Maindonald, J. H.; Cox N. R. 1984: Use of statistical evidence in some recent issues of DSIR agricultural journals. *New Zealand journal of agricultural research 27*: 597-610.

Matthew, C.; Lawoko, C. R. O.; Korte, C. J.; Smith, D. 1994: Application of canonical discriminant analysis, principal component analysis, and canonical correlation analysis as tools for evaluating differences in pasture botanical composition. *New Zealand journal of agricultural research: 37*: 509-520.

Pearl, J. 1985: Learning hidden causes from empirical data. *Proceedings of the Ninth International Conference on Artificial Intelligence*. San Mateo, CA., Morgan Kaufmann. Pp.567-572.

Quinlan, J. R. 1986: Induction of decision trees. *Machine learning 1*: 81-106.

Taylor, C., Michie, D.; Spiegalhalter, D. 1994: Machine learning, neural and statistical classification. Englewood Cliffs, New Jersey, Prentice Hall.

Thomson, K.; McQueen, R. J. 1996: Machine learning applied to fourteen agricultural datasets. Working paper 96/18. Dept of Computer Science, University of Waikato.