Applications of Machine Learning on two agricultural datasets*

Stuart Yeates Kirsten Thomson

Computer Science Department
Waikato University

stuart@cosc.canterbury.ac.nz kthomson@cs.waikato.ac.nz

Abstract

The induction of decision trees from tabulated data is a field of machine learning which has been demonstrated successfully in several practical applications. This paper looks at the application of this technology to two datasets in the agricultural domain, and show why it was not possible to achieve the success obtained in other domains.

Introduction

Induction of decision trees is a new field of machine learning, involving the algorithmic acquisition of structured knowledge in forms such as concepts, decision trees, discrimination nets or production rules. A common goal of such methods is to predict the value of an attribute in the data based on other attributes. The output can then be used, by either a human or a system, to classify unseen examples. Good surveys can be found in [1] and [2]

Outlook	Temp. °F	%Hum -idity	Windy	Play ?
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	70	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play

Table 1 Golf data

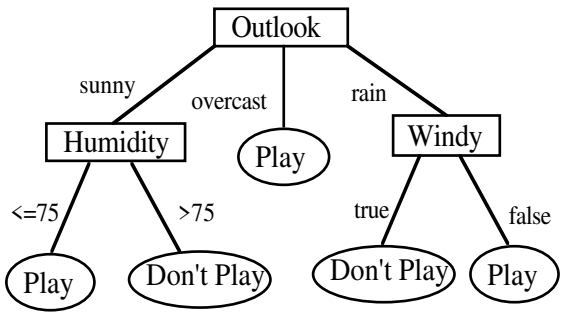

Figure 1 Decision tree derived from data in table 1 by C4.5

The canonical example in the area is golf, in which the algorithm is given a dataset containing weather data on a number of days and whether each of the days was suitable to play golf on, as shown in Table 1. The dataset is small, error free and contains a mixture of values in all the attributes. The algorithm, in our case C4.5 [3], is tested on previously unseen instances to determine whether it can differentiate suitable days. The C4.5 program produces the decision trees shown in Figure 1, which has been found to correctly classify unseen instances. The decision tree is intelligible by both humans and computer systems.

Agricultural Data

The Machine Learning Project at the University of Waikato is attempting to apply such learning techniques to real-world agricultural data. Much of the work under the project has focused on the Waikato Environment for Knowledge Analysis (WEKA)[4]. The workbench provides an interactive interface to existing machine learning algorithms and also pre- and post-processing tools for use by not only computer scientists, but also users of the data. In this paper we use these tools to examine two agricultural datasets. Both datasets were collected by Animal Behaviour and Welfare Research Centre, Ruakura, to whom we are indebted for the use of their data.

Venison Carcass Data

The Venison Carcass Data consisted of slaughter house records for more than 21,000 venison carcasses, containing such information as sex, weight, fat measurements, farm-of-origin, distance-from-farm, carrier, days-held-in-the-yards, date-of-slaughter and carcass-damage. The aim of our work was to predict carcass damage, a major problem in the venison industry, on the basis of the other attributes. It was thought that carrier - the company which trucked the animals to the slaughter house -

* Funding for this research was provided in part by FRST. Contract number UOW403

sex and fatness of the animals would be of importance.

After cleaning [4] we ran C4.5 extensively on the data.

We discovered three main trends: (a) many attributes were very closely linked to other attributes. For example farm-of-origin determines distance-from-farm, and since most farms only sent one consignment of animals to the works farm-of-origin also determines carrier and date-of-slaughter. It was also found that animals from the same farm tended to be in similar condition, so farm-of-origin also determined the weight and fat measurements. (b) carcass-damage was a relatively rare event, occurring in less than five percent of carcasses, which is comparable to the error in some of the attributes. (c) carcass-damage was most strongly linked to farm-of-origin, a result not found by statistical analysis, and not considered useful by the researchers.

Conventional statistics readily handles (a) and (b), removing their effect from the analysis. However, machine learning, particularly instance based methods such as C4.5 do not have access to the same range of techniques. In particular, while conventional statistics found significant relationships between fat and carcass-damage and between carrier and carcass damage, no such relationship was discovered by C4.5.

Bull Castration Data

The Bull Castration Data was derived from a field trial to compare physical and a chemical castration in juvenile bulls destined for beef production. The bulls were grouped into five trial groups: natural, physically castrated and three different forms of chemical castration. The data was a time-series over a number of sampling points, at which the bulls were tested for weight, testosterone levels etc. The aim of our work was to predict the most important time-period and attributes for development of the final weight.

Initially we ran C4.5 on the raw, cleaned data, but because the algorithm has no concept of time, or of time-series it treated each column as independent. In an attempt to capture some of the temporal nature of the data we tried creating extra columns which were the difference between successive samplings; this also largely failed. On the whole, the results from C4.5 on this dataset were abysmal, and considered of no use by the researchers who collected the data.

Conclusion

Datasets from the agricultural domain appear to be significantly more “complex” than the datasets traditionally used in machine learning. The reasons for this appear to be: (a) the extensive use of time series for modelling dynamic processes fundamental to agricultural systems (b) the high number of inter-attribute dependencies which are not related to the task at hand (c) the natural variation (noise) in agricultural systems, and (d) the experimental design strategies traditionally used in agricultural systems have developed in the light of traditional statistics and result in the data ideally suited to traditional statistics and less suited to machine learning.

There are three potential methods for overcoming these problems. Firstly to develop a separate experimental design methodology for machine learning as opposed to traditional statistics. This approach is likely to be divisive and keep large quantities of data beyond the reach of machine learning indefinitely. Secondly, to supplement experimental design to include additional attribute more appropriate for machine learning while maintaining traditional statistical robustness. This approach would give access to new data but exclude existing datasets. The third approach is to integrate some aspects of traditional statistics into machine learning systems such as WEKA. Such aspects might include *n*- way correlations between every pair of attributes to detect and eliminate the effects of very strongly linked attributes, or simple curve fitting methods to succinctly capture the shape of time series.

References

[1] R. J. McQueen, S. R. Garner, C. G Nevill-Manning and I. H. Witten, "Applying machine learning to agricultural data", *Journal of Computing and Electronics in Agriculture*, Vol 12(4), pp 275-293, 1995.

[2] P. Langley and H. A. Simon, "Applications of machine learning and rule induction", *Communications of the ACM*, Vol 38(11) 55-64, 1995.

[3] Quinlan J. R, (1992) C4.5: Programs for Machine Learning. Morgan Kaufmann

[4] G. Holmes, A. Dorkin and I. H. Witten, "WEKA: a machine learning workbench", *Proc. of Australia and New Zealand Conference on Intelligent Information Systems*, pp 357-361, 1994.