Machine Learning Applied to Geo-Archaeological Soil Data

O. S. Farrington
School of Mathematical and Computer Sciences,
Heriot-Watt University, Edinburgh, UK

N. K. Taylor
School of Mathematical and Computer Sciences,
Heriot-Watt University, Edinburgh, UK

Abstract

'Decision tree learning' is one of the most versatile and practical methods of inductive inference developed in applied artificial intelligence. The 'Machine Learning' technique selected for this study was the ID3 algorithm, and represents a new application for this simple classification building technique.
The data, to which the ID3 algorithm was applied, consists of a collection of data sets relating to specific soil analytical procedures (e.g. heavy mineralogy, bulk geochemistry, fabric analysis, etc.) originally collected to address the dark earth conundrum.
In a previous study, PCO analysis proved adept at identifying contributions from individual variables such that it was possible to identify the geo-archaeological components that together constituted the 'character profile' of a typical dark earth.
In this study the classification building abilities of the ID3 algorithms, have been used to explore the possibility of establishing a classification (based on soil data) of not only dark earth's but other archaeological deposits characteristic of urban contexts.

Introduction

A unique database of geo-archaeological soil data was put together several years ago from the results of analyses of soil samples collected from selected urban archaeological deposits at excavation. The database at present consists of a collection of multi-variate ordination plots: the product of the application of principal coordinate analysis (PCO) to soil data from a uniquely comprehensive range of (standard) soil analytical procedures that were applied to a large collection of predominantly urban archaeological soil samples. It has only recently become possible to resume work on this data, which originally formed the basis for a University of London doctoral thesis, soon to be completed.

In the intervening period, the first author has been introduced to some of the newer techniques of Artificial Intelligence, and became aware of their potential suitability for further mining of the database. During preliminary assessment of the principal coordinate plots, it was noted that the urban archaeological deposits studied appeared to be made up of two components; one derived from the local geology and the other an 'anthropogenic' component. This is a new observation for an archaeological deposit. However new studies now in progress (located in the Department of Environmental Science, University of Stirling, under D.A. Davidson), and based on rather different types of archaeological soils than we intend to use, have recently reported a similar finding: that the archaeological soil ('plaggen' soil) may similarly have an 'anthropogenic' component.

We intend to examine in greater detail the supposition that archaeological deposits are in general made up of material from the local geology and an anthropogenic component. We are particularly interested in the composition of the anthropogenic fraction in our urban soils and its stability between different types of archaeological deposit. To this end the geo-archaeological database of urban archaeological material should prove invaluable. This very large archaeological soil database is probably the largest and most comprehensive dataset available of different types of archaeological soils. We have selected in the first instance to work on the mineralogy of the fine sand fractions of a collection of archaeological deposits that included dark earths, occupation deposits, cultivation, and other putative deposit types including comparative material from the local surficial geologies of the archaeological samples in the study because PCO plots show a clear separation between local geology and 'anthropogenic source material' in the form of a number of species of 'authigenic' minerals.

Initially we will look to use various artificial intelligence data-mining techniques, possibly in conjunction with more multivariate ordination statistical methods.

This new study will form the basis of a short research project in the School of Mathematical and Computer Sciences at Heriot-Watt University, Edinburgh.

The Database

The data to be used for data-mining concerns the mineralogy of the fine sand fraction of a range of urban archaeological deposits, summarised in a pair-wise multi-variate ordination plot between PC1 and PC2. This plot demonstrated a clear separation between minerals derived from the local geology and a putative anthropogenic component.

Several of the archaeological deposit samples were observed to be rich in authigenic minerals, such as the synthetic olivines (fayalite). Their presence in these deposits is easily predicted to represent the result of anthropogenic activities, such as metal processing, which produced these synthetic minerals.

The Proposed Data-Mining Methods

The Artificial Intelligence techniques which appear initially to be of most value are the Induction Tree techniques, of which the simplest is the ID3 Algorithm (Quinlan 1986), and Artificial Neural Networks, such as the Multi-Layer Perceptron (Rumelhart et al. 1986).

The Archaeological Context

The soil database was prepared from soil samples collected from urban archaeological sequences between 1984 and 1988, specifically for the purpose of investigating the composition provenance and depositional history of the enigmatic urban dark earth horizons.

The sampling programme adopted was fully comprehensive and involved collection not only of dark earths, but all other 'types' of archaeological deposit available in the same profile or near vicinity. Samples of the local geological substratum were also routinely collected.

A preliminary assessment of the data concerning soil samples taken from urban archaeological deposits (primarily dark earths) can be found in Farrington 1989, or Farrington and Bateman 1992.

Until recently, archaeologists have taken little interest in the routine analysis of the < 2 mm fraction of an archaeological deposit.

On the other hand, geomorphologists have over the past 100 years developed numerous methods of investigation of the <2mm fractions of soils and sediments. Quaternary soil scientists in particular have developed methods of identifying depositional processes and provenances of landscapes, which are especially suited to materials common in late Quaternary (Cenozoic sub-era) landscapes. As such these were deemed suitable for the investigation of the depositional history and provenance and their contribution to former urban landscape development.

The Adoption of Soil Science Analytical Procedures

This is the first attempt at a study based entirely on soil samples from archaeological deposits, using standard soil science analytical procedures. All soil analytical work was carried out in the Soils Division, Rothamsted Experimental Station, Harpenden, Hertfordshire. At the conclusion of the pilot study an internal report was sent to the sponsors; the Museum of London (Farrington 1983).

In 1984 a much more ambitious project was initiated as the basis for a doctoral research project shared jointly between the Department of Geography, Birkbeck College, University of London, and the Soils Division, Rothamsted. The topic was the application of soil analysis to an archaeologically and historically important archaeological deposit: Dark Earth.

A full list of all the analytical procedures used is inappropriate here, but will appear in the doctoral thesis.

When the plots were drawn, preliminary observations showed that the archaeological deposits which had been sampled were derived from two different fractions; one was the local geological substratum and the other was an anthropogenic component.

It is the aim of the present study to take as its starting point this observation and to examine more closely the distribution of minerals between the putative anthropogenic and geological components to identify the boundaries between the two.

This proposition is suitable for the two Artificial Intelligence procedures (ID3/MLP), as both test slightly different aspects of the problem.

Preliminary Statistical Analysis (Multivariate Ordinations)

A considerable volume of geo-archaeological data was summarised using the multivariate ordination procedure, know as principal coordinate analysis contained in the statistical pakkage Genstat developed at Rothamsted. Genstat Version 5 release 4.5 was used to produce scatter-grams based on pair-wise plots of the first four principal co-ordinates. These were then visually inspected for groups, patterns or trends. Gower Similarity Coefficients were also calculated for all pair-wise combinations of samples and minimum spanning trees, linking samples by maximum Gower Similarity Coefficient, were superimposed on selected PCO scatter-grams.

It was observed that in general the first two principal co-ordinates provided the best separations, both with and without the addition of the minimum spanning trees. One database, that of the mineralogical data, was selected as the basis of this study. The plot of PC1 against PC2 produced a clear visual separation between archaeological deposits according to locality and anthropogenic content.

When plotted on a scatter-gram, the first two components resulting from the PCO clearly clustered the samples according to location. This is readily discernible visually, but the clusters are not linearly separable. I.e. one cannot draw straight lines on the scatter-gram which separate the clusters and so their boundaries remain unknown. Such data is well suited to processing with a Multi-Layer Perceptron (MLP).

Artificial Neural Networks

An MLP is an Artificial Neural Network (ANN) which can be trained by example to perform a non-linear mapping from a set of input values to one or more output values, i.e. it can discover an "optimal" set of curves, which do separate the clusters in a scatter-gram. Subsequent samples can then be presented to the MLP for classification into the cluster which they best fit.

One problem with ANNs is their 'black-box nature'. It is very difficult to ascertain exactly what they have learnt during training. It is somewhat easier to extract rules from symbolic learning systems.

Inductive Decision Tree Methods (ID3)

One very popular symbolic learning system, ID3, is capable of inducing the rules which determine where the clusters will appear on a scatter-gram. This system will, therefore, not only be able to classify new samples into appropriate clusters, it will also provide information indicating why the clusters were chosen. Symbolic systems generally perform better with discrete data values rather than continuous variables and tend to fail less gracefully than ANN approaches.

Machine learning techniques such as these offer great benefits in the analysis and interpretation of geo-archaeological soil data and the authors are currently investigating the potential of a number of algorithms.

Acknowledgements

Maria Chamberlain provided many helpful suggestions during the drafting of this paper.

Jill Stirling acted as scribe to Orpah Farrington during the drafting of this paper to circumvent some of the difficulties of Dyslexia.
All soil samples from the archaeological deposits in the database used in this study were obtained with permission.

Training in the methods of examination of the mineral components of archaeological soil samples was received from John Catt. All analytical work was carried out in the Mineralogy Laboratory Suite of the Soils Division, Rothamsted Experimental Station, Harpenden, Hertfordshire.

Much helpful discussion on the interpretation of Principal Co-ordinate Analysis plots was provided by John Catt, and by the late James Rayner of the Statistics Division, Rothamsted.

Orpah Farrington is in receipt of a Research Student Scholarship from Heriot-Watt University.

References

DERCON, G., DAVIDSON, D.A., SIMPSON, I.A., DALSGAARD, K. and SPECK, T., 2003. The Nature and Formation on Anthrosols in NW Europe: a comparison between 3 countries. Extended Abstracts Second International Conference on Soils and Archaeology, Pisa:21-24.

DIGBY, P.G.N and KEMPTON, R.A., 1991. Multivariate analysis of ecological communities: population and community biology series. 3.5 Principal coordinates analysis:83-93. Chapman & Hall, London.

DORAN, J.E. and HODSON, F.R., 1975. Mathematics and computers in archaeology. Edinburgh University Press, Edinburgh.

DUCKE, B., 2003. Archaeological predictive modelling in intelligent network structures. In Doerr, M. and Sarris, A. (eds.), CAA2002: The Digital Heritage of Archaeology: Computer Applications and Quantitative Methods in Archaeology: Proceedings of the 30th Conference, Heraklion, Greece, April 2002. Archive of Monuments and Publications Hellenic Ministry of Culture:267-273.

FARRINGTON, O.S., 1983. The dark earth project. Unpublished Museum of London internal report.

FARRINGTON, O.S., 1989. Dark earth in Northwest Europe: application of geoanalytical techniques to a historically important archaeological deposit (abstract). 28th Int. Geol. Congr. Abst. 1:473.

FARRINGTON, O.S., 2001, unpubl. Machine learning (ID3), and artificial neural networks (MLP): A comparison of the properties of two different types of classification system applied to the same dataset: A selected subset of letters of the Latin Alphabet. MSc class report, Heriot-Watt University.

FARRINGTON, O.S. and BATEMAN, R.M., 1992. A holistic approach to the analysis of archaeological deposits, illustrated using a late Roman urban sequence from Northwest Europe. Mat. Res. Soc. Symp. Proc. Vol. 267:179-192.

GOWER, J.C, 1966. Some distance properties of latent roots and vector methods used in multivariate analysis. Biometrika 53:325-338.

GOWER, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857-872.

GOWER, J.C. and DIGBY, P.G.N., 1981. Expressing complex relationships in two dimensions. In Barnett, V. (ed), Interpreting multivariate data, Wiley, Chichester:83-118.

GOWER, J.C. and ROSS, G.J.S., 1969. Minimum spanning trees and single linkage cluster analysis. J. R. Stat. Soc. C 18:54-64.

MITCHELL, T., 1997. Machine learning. McGraw-Hill, NewYork.

QUINLAN, J.R., 1986. Induction of decision trees. Machine Learning 1(1):81-106.

RUMELHART, D.E., HINTON, G.E. and WILLIAMS, R.J., 1986. Learning internal represaentatioons by error propagation. In Rumelhart and McClelland (eds.), Parallel Distributed Processing Vol. 1, MIT Press, Cambridge Massachusetts:318-362.

SHENNAN, S., 1988. Quantifying archaeology. Edinburgh University Press, Edinburgh.

THOMAS, J., SIMPSON, I.A., DAVIDSON, D.A. and GAULD, J.H., 2003. GIS mapping of anthropogenic soils in Scotland: investigating the location and vulnerability of Scottish 'plaggen soil' deposits. Extended Abstracts Second International Conference on Soils and Archaeology, Pisa:127-128.

WILSON, C.E., DAVIDSON, D.A. and CRESSER, M.S., 2003. Multi-element soil analysis as an aid to archaeological interpretation. Extended Abstracts Second International Conference on Soils and Archaeology, Pisa:78-79.