Machine learning for lithology classification, data fusion and evaluating prediction uncertainty

Matt Cracknell and Anya Reading - April 2013
CODES/School of Earth Sciences, University of Tasmania
Machine learning for lithology classification

Computational Geophysics and Data Visualisation
Dr Anya Reading (Project Leader)

- Machine learning for geological/geophysical data inference
  - Ambient seismic energy: Tectonic structures
  - Combining geophysical, petrophysical and geological data to constrain 3D geophysical models
  - Computational and human interaction strategies in knowledge generation from spatial information
Motivation

- Geological maps for the base input layer for a wide range of applications (e.g., ore deposit targeting, hydrology, geo-hazards risk assessment)
- Generally constructed from limited field observations based on subjective interpretations formulated by a domain expert – the geologist
- Uncertainty is difficult to quantify and rarely indicated

Presentation structure

- Machine learning → “Data Fusion” → supervised classification
- Comparison of machine learning algorithms (MLAs)
- Estimating/evaluating prediction uncertainty
- VHMS prospectivity example, west Tasmania
- Knowledge discovery → variable/class interactions
What is Machine Learning?

- Machine Learning (Pattern Recognition/Data Mining/Artificial Intelligence)
- Family of algorithms that attempt to mimic the ability of humans to recognise patterns in data and use this learning to identify similar patterns
- Tools for efficiently solving non-linear problems in high dimensional input space (Kanevski et al., 2009)
- Supervised and unsupervised approaches to inductive data-driven inference

Why use Machine Learning?

- Large digital datasets, continuous collection and range of data types
- Practically impossible for humans to effectively visualise and interpret
- Useful in situations where the underlying process is obscure and/or difficult to understand
- Facilitate understanding and objective interpretations of complex natural phenomena (aka Knowledge Discovery)
Machine Learning for lithology classification

Machine Learning for geophysical “inversion”

- Turning geophysical data into geological information (model)
- Quantify data and model uncertainty
- Geophysical inversion is the process of estimating some model $m$ based on data $d$ (containing noise $\eta$) which are related via the function $G$

\[
G(m) = d
\]
\[
d = d_{\text{true}} + \eta
\]

DETERMINISTIC
Partial Differential Equations

MACHINE LEARNING
- Partial Differential Equations
- e.g. Random Forests, Support Vector Machines

MULTIPLE MODEL ENSEMBLES
Monte Carlo

MODEL PARAMETER SAMPLING
MCMC
Supervised classification

- Prediction of discrete unordered categories from training data
- Input variables \( <x_1, x_2, ..., x_d> \) linked to a set of class labels \( \{y_1, y_2, ..., y_c\} \)
- Train a classification model that links variables to target classes via the function \( y' = f'(x) \)
- Machine learning induces links based on the information within data to guide the training of a classification model via minimisation of error/loss function

Generalised workflow

- Data pre-processing
- Training and parameter selection
- Prediction evaluation
Machine Learning strategies

Naïve Bayes (NB) – Statistical Learning Theory
• estimates class-conditional probability assuming that for a given class the variables are independent (Hastie et al., 2009)

$k$ - Nearest Neighbours (kNN) – Instance-based
• sample proximity in feature space (Fix and Hodges, 1951; Cover and Hart, 1967)

Random Forests (RF) – Logic-based
• majority vote based on the outcomes of multiple (ensemble) random decision trees (Breiman, 2001)

Support Vector Machines (SVM)
• maximum marginal decision boundaries with respect to support vectors in kernel space (Vapnik, 1998)

Artificial Neural Networks (ANN) – Perceptrons
• networks of primitive functions with weighted inputs and binary outputs (Rojas, 1996)
Lithology Classification – comparing algorithms

- Inferring the location of geological units
- Evaluation of performance against ground truthed geological map
- Which “off-the-shelf” MLAs are most accurate, efficient and easy to use?

Broken Hill
Metamorphosed sedimentary, volcanic and intrusive rocks

Cracknell and Reading (under review), A Comparison of Machine Learning Algorithms for Data Inference in Applied Geosciences: Predicting Lithology from Airborne Geophysics and Satellite Data, Computers and Geosciences
Machine Learning for lithology classification

**Data pre-processing**
- Input variables: spatial coordinates, airborne geophysics and Landsat ETM+ imagery band ratios
- Removal of correlated/redundant variables
- 10% random sample for training data ~ 6500 samples

**Classification model training**
- 10-fold cross-validation for parameter selection, sensitivity to parameters

**Prediction evaluation**
- Overall accuracy on independent test data

![Graph showing cross-validation accuracy and test accuracy for different models (NB, SVM, kNN, ANN, RF)]
Comparing classifiers

Why do we see significant differences in performance between Random Forests and the other algorithms trialled?

- ensemble classifier – less sensitive to noise
- spatial distribution and number of training samples?
- essentially an adaptive form of kNN

<table>
<thead>
<tr>
<th>NB</th>
<th>ANN</th>
<th>kNN</th>
<th>SVM</th>
<th>RF</th>
</tr>
</thead>
</table>

PREDICTIONS

ERROR
Prediction Uncertainty

Use of predicted class membership probabilities $p_c$ to quantify uncertainty using a modified version of “variance” (Kohavi and Wolpert, 1996)

$1 - \sum p_c^2$

Are the statistical and spatial distributions of error related to estimates of prediction uncertainty?

Cracknell and Reading (accepted) The upside of uncertainty: Identification of lithology contact zones from airborne geophysics and satellite data using Random Forests and Support Vector Machines, *Geophysics*
Identifying “unclassified” samples

Assumption: high uncertainty indicates high probability that sample is misclassified

Identify uncertainty thresholds that remove maximum number of misclassified samples while preserving maximum number of correct predictions

![Graph showing Proportion vs. Uncertainty (Variance) for Random Forests and Supp. Vector Mach. with thresholds 0.63 and 0.59 respectively. Correct and misclassified curves are indicated.]
Identifying “unclassified” samples

Significant increase in Random Forests test accuracy after the exclusion of “unclassified” samples (i.e. ambiguous classifications)
VHMS mineralisation prospecting

**Hellyer – Mt Charter (Western Tasmania)**

- Three known economic Pb-Zn VHMS ore deposits
- Dense temperate rainforest vegetation and thick soil profiles
- Abundant geochemical and geophysical data
- Que-Hellyer Volcanics (QHV) suite of Cambrian sea-floor calc-alkaline volcanics and pervasive regional hydrothermal alteration

Cracknell, Reading and McNeill, to be submitted AJES April 2013, GEOLOGICAL MAPPING OF THE HELLYER – MT CHARTER REGION, TASMANIA, USING RANDOM FORESTS AND SELF-ORGANISING MAPS
Machine Learning for lithology classification

“Data Fusion”

All (41) input variables:
- Soil geochemistry
- Geophysics
- Landsat ETM+

1. Removal of correlated/redundant variables

2. Select most important/relevant variables (recursive backward selection) using Random Forest variable importance measures

CV accuracy

0.840±0.021
0.849±0.020
0.843±0.019
Results

- ~78% overall accuracy
- Majority of misclassifications along geological contacts
- Additional regions classified as “Host Horizon” (i.e. H and YA) indicate potential VHMS targets
Knowledge Discovery

Random Forests partial dependency plots for QHV
- Indicates the relative influence of a single variable on class membership probabilities
- Does not represent interactions between multiple variables
Knowledge Discovery

Self-Organising Maps (SOM)
- Identification of intra-class clusters
- Visualisation of the statistical and spatial distributions of clusters
Discussion and Conclusions

Data Fusion
- Demonstrated practical use of MLAs for robust non-deterministic geophysical inversion (modelling) for multivariate and multiclass problems in complex geological terranes

Random Forests - identified as best MLA in terms of:
- performance, stability and efficiency
- relationship between uncertainty and misclassified samples
- interpretability/understanding decision structures

MLA performance considerations
- Application (targets/scale) and available data – data pre-processing
- Training data: number and statistical/spatial distributions

Spatially aware classifiers?
- textural derivatives
- multiple local classification models
- spatial autocorrelation
Questions?