Classifying Rocks using Geochemistry? Use An Objective Data-Driven Approach

Multivariate Geochemical Tectonic Discrimination: Practical Approaches, Limitations and Opportunities

Morgan Williams, Jens Klump & Steve Barnes

MORE INFORMATION
Morgan Williams
Mineral Resources
morgan.williams@csiro.au
@metasomite

- **A machine learning approach to geochemical discrimination overcomes limitations of classical methods**
- **Free and open source tools for building classifier workflows are readily available**
- **Dimension-reducing techniques are particularly useful for visualisation**

Check out the extended abstract and live examples via links below.

ABSTRACT

tinyurl.com/aegc2019MWabstract

LIVE EXAMPLES

tinyurl.com/aegc2019MWlive

This poster is accompanied by a set of examples in Jupyter notebooks, which you can browse, modify and execute using Binder.

REPOSITORY

tinyurl.com/aegc2019MWgithub

This public repository contains the notebooks, abstract and a copy of this poster.

PYROLITE DOCS

pyrolite.readthedocs.io

Much of the tooling for this project is contained within the open source python package pyrolite; this link will take you to the online documentation.

Interpretations regarding the source and tectonic context of magmatic rocks are commonly centred around the abundance of a small number of relatively incompatible trace elements[1,2]. These classic methods for tectonic discrimination have been practically useful, but most underutilise available information, resulting in uncertain classification[3]. Modern public data repositories allow new opportunities for systematic analysis on a previously inaccessible scale, and provide a variety of data with global scope. This study investigates a machine learning approach to the problem of tectonic discrimination, focusing on the practical elements of building classification models using geochemistry or other compositional data.

Machine Learning for Tectonic Discrimination

Supervised machine classification models for tectonic discrimination of magmatic rocks improve resolution of geochemical contrast by using more geochemical features[4,5], and use training data from public databases (e.g. EarthChem[6]) to build classifiers overall accuracies of approximately 90%. The approach used here follows extends these methods, and achieves similar or better classification accuracies (Figure 1A), while exploring the nature of classification uncertainty. The classifier workflows used in this study are built around the scikit-learn[7] and pandas[8] python packages, with geochemistry-specific components from pyrolite. You can find links to the live demonstrations and documentation above.

The majority of the effort required to use machine classification models is for data preparation and pre-processing, but this is often where the largest gains for performance are to be found. Compositional data typically requires transformation prior to the use of any statistical measures (e.g. log ratio transforms), but most transforms are incompatible with missing data. This often results in a principal trade-off between the volume of valid training data and dimensionality. Hence the inclusion of compositional features in classification models should be based on the value they add. Training data are also commonly strongly biased towards particular classes, and resampling classes prior to training models can minimise this effect.

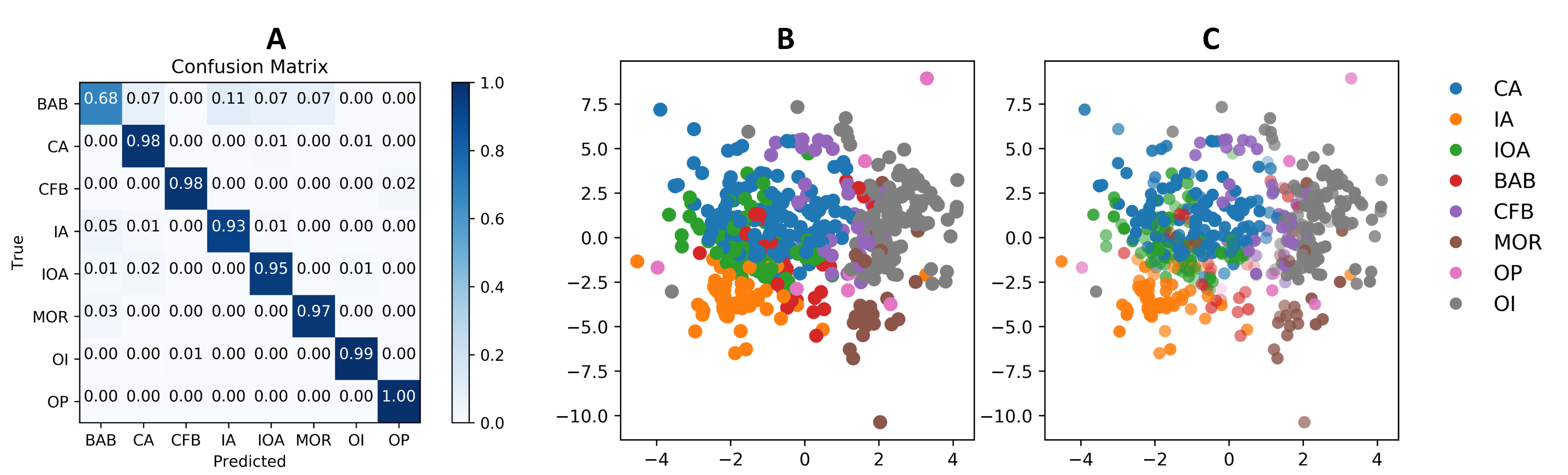

Figure 1: Example results from a support vector classification workflow using whole rock geochemistry for mafic rocks. **A** Confusion matrix indicating the proportion of test data which was correctly classified as belonging to the true tectonic setting class, and which classes account for the misclassifications. **B** Manifold embedding of 29-dimensional geochemical data coloured by predicted tectonic setting (axes in arbitrary units). **C** Manifold embedding of the same data, where the entropy of the multiclass probability is used as an uncertainty measure to modulate opacity (lighter colours indicate less certain classification). Tectonic setting classes are as follows: BAB – Back Arc Basin, CA – Continental Arc, CFB – Continental Flood Basalt, IA – Island Arc, IOA – Intra-oceanic Arc, MOR – Mid-Ocean Ridge, OI – Ocean Island, OP – Oceanic Plateau.

Overcoming Graphical Limitations for Visualisation

One of the historical limitations for geochemical classification methods has been visualisation. The multivariate classifiers effectively exploit the high-dimensional data relationships which remain difficult to visualise directly. One approach to approximating these relationships for visualisation is to use manifold dimensional reduction methods (e.g UMAP[9]), which preserve local scale data structure (Figure 1B). Information entropy is a useful measure which can reduce multi-class classification probabilities to singular values related to classification uncertainty (Figure 1C).

Results and Limitations

- Accurate classifiers for tectonic discrimination can be constructed using a variety of different classifier models.
- For this problem, with larger datasets the overall classification accuracy is typically approximately 90%, which reflects the significant geochemical overlap arising through natural variations and similarities in geological processes and sources across tectonic settings. Beyond certain thresholds, more data won't impact accuracy.
- There are limits to discrimination using a practical subset of elements, and features chosen will limit the future applicability of any classifier.
- These models are trained on data from modern tectonic settings, and applicability in deep time is complicated by rock-record biases and secular change.

Getting Started

- **Know your data, and get it in shape**; with curated datasets, building classifiers is relatively straightforward
- External public data with broad scope can be useful to **avoid biases** and reduce sensitivities to outliers/novel data
- **Start with simple models, and build multiple**
- Accuracy is often not the only objective, understanding the relative certainty of your predictions is useful
- Be wary of overfitting
- Know what question you trying to answer, and **adapt to your application;** targeted selection of features can increase the relevance and accuracy of predictions

REFERENCES

[1] Pearce, J.A., Cann, J.R., 1973. Tectonic setting of basic volcanic rocks determined using trace element analyses. Earth and Planetary Science Letters 19, 290–300. doi.org/10.1016/0012-821X(73)90129-5 **[2]** Pearce, J.A., 2008. Geochemical fingerprinting of oceanic basalts with applications to ophiolite classification and the search for Archean oceanic crust. Lithos 100, 14–48. doi.org/10.1016/j.lithos.2007.06.016 **[3]** Li, C., Arndt, N.T., Tang, Q., Ripley, E.M., 2015. Trace element indiscrimination diagrams. Lithos 232, 76–83. doi.org/10.1016/j.lithos.2015.06.022 **[4]** Petrelli, M., Perugini, D., 2016. Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. Contrib Mineral Petrol 171, 81. doi.org/10.1007/s00410-016-1292-2 **[5]** Ueki, K., Hino, H., Kuwatani, T., 2018. Geochemical Discrimination and Characteristics of Magmatic Tectonic Settings: A Machine-Learning-Based Approach. Geochemistry, Geophysics, Geosystems 19, 1327–1347. doi.org/10.1029/2017GC007401 **[6]** Lehnert, K., Su, Y., Langmuir, C.H., Sarbas, B., Nohl, U., 2000. A global geochemical database structure for rocks. Geochemistry, Geophysics, Geosystems 1. doi.org/10.1029/1999GC000026 **[7]** Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825−2830. **[8]** Wes McKinney, 2010. Data structures for statistical computing in python, in: van der Walt, S., Jarrod Millman (Eds.), Proceedings of the 9th Python in Science Conference. pp. 51–56. **[9]** McInnes, L., Healy, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.