geoconvention
Calgary • Canada • May 15-19 2017

Random forests on the Atlantic margin: lithology prediction from wireline logs

Matt Hall[1], Diego Castañeda[1], Evan Bianco[1], and Jason Moore[2]
[1]Agile, [2]Canstrat

Summary

High-resolution rock-type information in a well usually requires inspection of core or cuttings by a geologist. The problem of predicting similar descriptions in neighbouring wells from wireline logs, producing 'e-facies', is an old and tough problem. Training a statistical model to predict such lithologies requires a reliable, consistent suite of lithology labels in a reasonable number of wells, plus corresponding high-quality wireline logs. This case study describes how we formulated the prediction problem as a machine learning task. We started with a highly regular set of lithology labels and a much less regular set of wireline logs, to produce lithology predictions with cross-validation accuracies ranging from 0.5 to 0.8, and averaging around 0.7. We achieved this with a gradient boosted-tree model, a type of random decision-tree forest model.

Introduction

During the second half of the 20th century, Canstrat and Amstrat devoted substantial resources to developing highly regularized lithologic descriptions from core and cuttings samples in thousands of wells in North America. Their database contains robust lithological categories ('labels') in about 15 000 wells in Canada and a further 45 000 wells in the United States. The labels are useful as a consistent and reliable source of lithologic information, especially in an exploration setting. The only problem is the relatively small number of wells that have Canstrat or Amstrat data — only about 1.5% of the millions of wells on the continent. Canstrat would like to have similarly regular and consistent lithology labels in every well that has wireline log data, which is inexpensive and almost ubiquitous.

This scenario is an ideal supervised machine learning task — we have a strong business need, relatively accessible data, and high quality labels. In particular, getting such labels is usually a big problem, because they are often expensive to acquire. (It turned out that the wireline logs were the problem.)

Taking advantage of the recent explosion in open-source machine learning software, we designed a workflow to train a statistical model using wireline log data in wells that have Canstrat or Amstrat data, then to use that model to predict the Canstrat labels in wells that only have wireline logs.

Method

The overall workflow is illustrated in Figure 1. Our method is best described as a pipeline, consisting of components through which the dataset travels. The pipeline is flexible enough to handle data from different sources, train different models, and produce results in various formats. The pipeline was implemented in Python and depends on open source software.

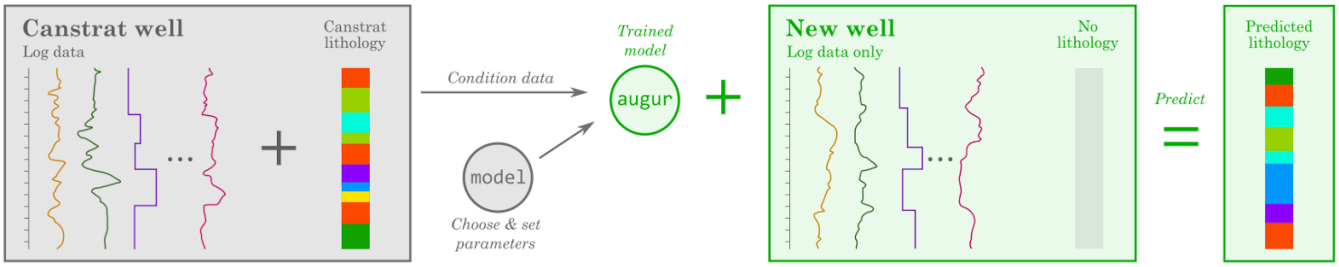

Figure 1. An overview of the workflow: we use a number of wells with logs and lithologies to train a statistical model. From the trained model 'augur', we can make predictions for wells that lack lithology labels.

The first part of the pipeline handles data loading, quality checking, and conditioning. The source files for the wireline logs are LAS files from various sources, read and processed by Agile's open source project `welly` (http://github.com/agile-geoscience/welly). The product is a regularized suite of wireline logs, constituting the feature matrix in the machine learning task, plus the lithologies, where known. The sources for the lithologies are Canstrat DAT files, read by Agile's `striplog` tool (Hall 2016). The lithologies become the label vector in the machine learning task.

The central part of the pipeline comprises the machine learning task, implemented in Python's `scikit-learn` package (Pedregosa et al. 2011). The wireline logs are used as features and undergo a series of transformations in an effort to separate each lithology in this feature space as best as possible. Only the most complete data instances are used in training — i.e. we avoid wells with missing logs — a decision we can afford to make because we have a fairly large amount of data. We train a boosted tree model as implemented in the `XGBoost` package (Chen & Guestrin 2016). Gradient boosting of decision trees is an idea that originated in the late 1990s in the work of Jerome Friedman and others (e.g. Friedman 1999). It has since proven highly effective in classification problems in other domains.
The last part of the pipeline transforms the predicted labels back into lithologies, so that they can be exported as LAS files or Canstrat's ASCII-based DAT files.

Example

Our pilot study draws logs from a collection 73 wells on the Scotian Shelf. For the training dataset, we used only wells that have a complete suite of high-quality wireline measurements: gamma-ray (GR), spontaneous potential (SP), deep resistivity (RES), neutron porosity (NPHI), density (RHOB), and sonic (DT). Of the 73 wells, 25 meet these requirements. On a current generation laptop, the time for training a model on 800 000 data instances — corresponding to a 0.1 m sample interval — is around 10 minutes.

The results from one well are shown in Figure 2. The accuracy score is 78% for this well, and the qualitative result is very good — the prediction looks a lot like the actual lithologies. When we look closely at the misclassifications, we note that almost no siltstone was correctly predicted — almost all of it was classified as shale. Along with a substantial amount of sand being misclassified as shale, these two misclassifications account for most of the error.

We use cross-validation to assess how we are doing. This entails making predictions in wells that do have Canstrat labels, omitting them temporarily from the training, then comparing those predictions to the known lithologies in those wells. The 78% accuracy is at the upper end of what we achieve in cross-validation. We see accuracies from 47% to 78%, with most wells around 70%. Generally speaking, prediction accuracy improves in deeper stratigraphic intervals where there is more data support, and higher log data quality due to smaller drill bit sizes and cleaner hole conditions.

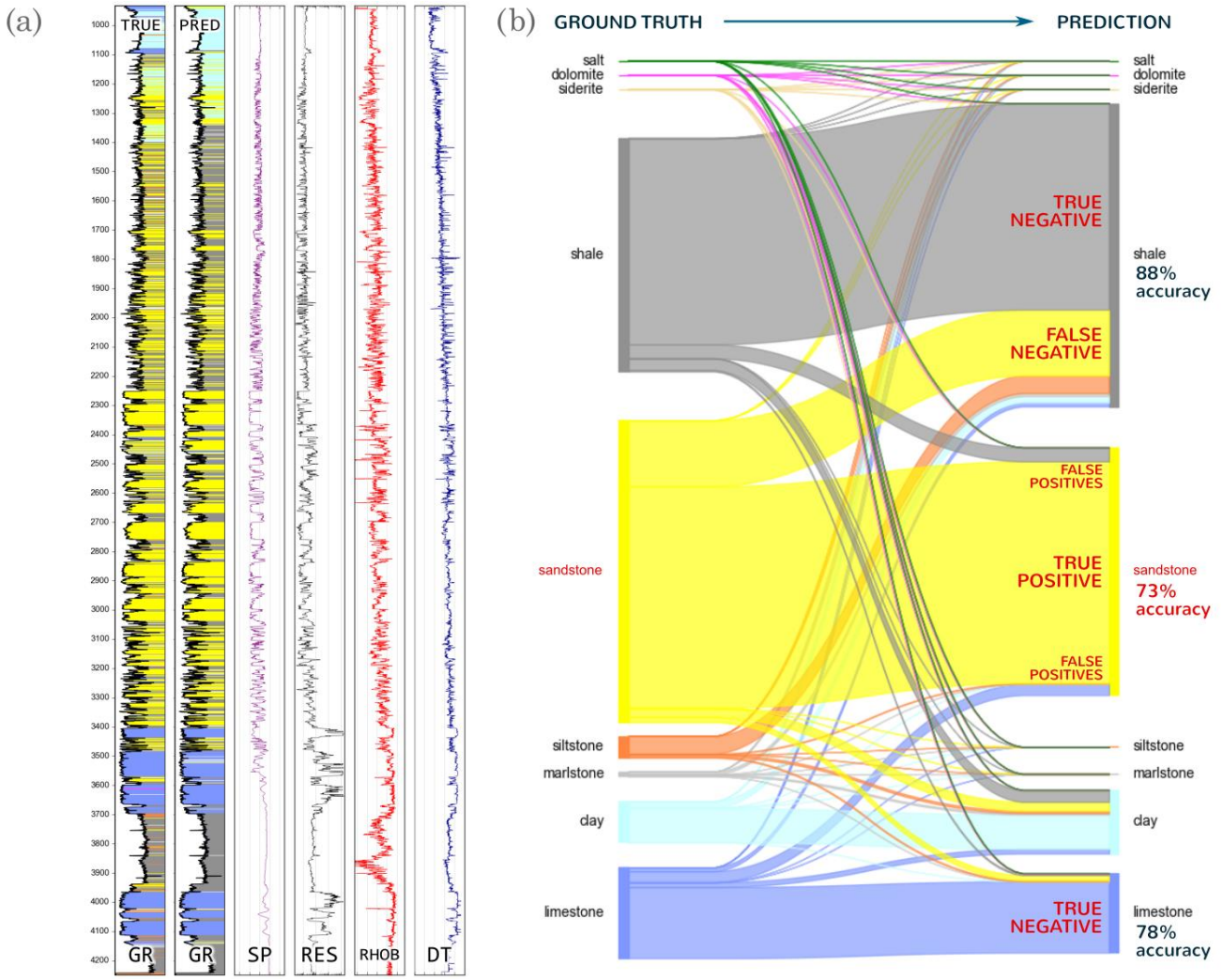

Figure 2. Cross validation. (a) True (leftmost) vs predicted lithologies in one of the wells in the Nova Scotia cohort of wells, with some of the logs used in training. The accuracy score for this well is 78%. (b) A visualization of how the true lithologies (left) were predicted. For example, most sandstone samples were correctly predicted. A substantial number were misclassified as shale (false negatives). Conversely, some of the predictions of sand were actually other lithologies (false positives).

Discussion

Our results so far are encouraging, but there are a great many open questions. For example, while we can be quite quantitative about model reliability, it's not clear yet what a useful level of certainty is. Visualizations of uncertainty, as well as a way to include it in other workflows, will help. But this is a general problem in subsurface analysis.

Among the many questions for further research in well-based lithology prediction are:

- Which model families are best at this task? We have had good results from random forests and especially boosted trees, but we have not tried every possible approach.
- What sort of features are most helpful in the model? It is usually up to the researcher to select and engineer features. There may be especially useful nonlinear combinations of features, or spatial treatments of the data instances.
- What labels are most predictable? Lithology is useful, but it's possible that other properties, such as porosity, grain-size, or saturation, or combinations of properties, are more predictable.

- How general can the model be? Is a model that has seen a large number wells from many different basins capable of predicting lithologies in wells that are far outside the training cohort?
- How can we incorporate updates to the labels — new descriptions in old wells, new wells, or human 'hints' to the best prediction — to update and improve the model?

Conclusions

- We have built a flexible machine learning pipeline that predicts Canstrat-like lithologies in wells that have only wireline logs.
- We implemented our pipeline in Python using open source software and public wireline log data.
- The pipeline has the following components: read the Canstrat lithology data with `striplog` software, read and condition the wireline log data with `welly`, engineer features with `scikit-learn`, train a gradient-boosted trees model with `XGBoost`, predict lithologies with the trained model.
- The pipeline achieves a prediction accuracy of about 70% in cross validation.
- This highly repeatable workflow makes lithologies consistent with existing Canstrat lithologies available in a much larger number of wells.

Acknowledgements

An open access version of this abstract is available at ageo.co/geocon2017-ml.

References

Chen, T, and C Guestrin (2016). XGBoost: A scalable tree boosting system. *SIGKDD Conference on Knowledge Discovery and Data Mining*, Aug 2016, San Francisco. DOI: 10.1145/2939672.2939785, arXiv:1603.02754 [cs.LG].

Friedman, J (1999). Greedy function approximation: A gradient boosting machine. IMS Reitz Lecture, February 1999. Online at http://statweb.stanford.edu/~jhf/ftp/trebst.pdf.

Hall, M (2016). Striplog: new open source software for handling and analysing discontinuous and qualitative data. Oral paper at the Atlantic Geoscience Society Colloquium, Truro, Nova Scotia, Canada, 5–6 February 2016.

Pedregosa, F, et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, pp. 2825-2830, 2011. HAL Id: hal-00650905.