The Power of Machine Learning in Petroleum Geoscience: Biostratigraphy as an Example

M.D. Simmons[1]*, O. Adeyemi[1], M.D. Bidgood[2], P. Maksymiw[1], P. Osterloff[3], D. Possee[4], I. Prince[3,]
C.M. Routledge[5], B. Saunders[1], F.S.P. van Buchem[1]

[1]Halliburton, [2]GSS (Geoscience) Ltd, [3]Shell, [4]University of Southampton, [5]University College London

Summary

The application of data science and machine learning is transforming petroleum geoscience workflows. Routine, yet time-consuming and important tasks can be made more efficient by the application of machine learning-based assisted interpretation, freeing the geoscientist to carry out tasks with greater value. Accuracy, reproducibility, and understanding of uncertainty are also improved and greater insight can be gained. Biostratigraphic data is very common in the industry but requires deep specialised knowledge and significant time to interpret, hence it can be underutilized. However, the form of the data makes it suitable for the application of machine learning techniques. The applications of machine leaning have been tested on biostratigraphic data from a set of typical industry wells to facilitate the interpretation of biozone/age and paleoenvironment. Application of Random Forest and Naïve Bayesian algorithms achieved results comparable to standard human interpretation, although pre-processing of the data (e.g. removal of spurious reworked or caved data) proved beneficial. Critical to the success of the project was the close working relationship between data scientists and subject matter experts in order to capture the nuances of biostratigraphic data and its interpretation. The work forms a case study for application to other geoscience data types.

Introduction

Data science, and machine learning in particular, has the potential to revolutionize the practice of petroleum geoscience. Routine tasks such as the interpretation of lithology from wireline log data or the interpretation of faults within seismic datasets easily lend themselves to machine learning techniques enabling huge gains in efficiency. Tasks that previously took several hours/days can now be accomplished accurately in minutes/seconds, freeing the geoscientist to concentrate on interpretive tasks that are of higher value. Time saving is not the only gain to be made. Bond et al. (2007) have succinctly demonstrated uncertainty in geoscience interpretation and how a scientist is influenced by his specialised skills and experience. In contrast, machine learning applications can provide greater accuracy and capture the range of uncertainty with clear reproducibility. Most importantly of all, new insights may be generated that can add huge value in the exploration and production workflow.

In order to test the application of machine learning to petroleum geoscience, data was utilized from one of the most abundant data types in the industry – well-based biostratigraphy. In most wells, cuttings and core samples are routinely analysed for microfossil content. Interpretation of this data provides information on age (by calibration of the biozones and bioevents recognised) and depositional environment that in turn contributes to subsurface correlation and petroleum systems element (reservoir, source, seal) mapping and modelling (Jones, 2011). The application of statistical techniques to palaeontological data is not new. The subject has long benefitted from statistical approaches to taxonomic classification and palaeoenvironmental or stratigraphic interpretation (Hammer & Harper, 2006). What is new is a specific machine learning approach enabled by recent advancements in computing power and the democratization of vast volumes of data and machine learning algorithms in open source software libraries.

The interpretation of biostratigraphic data is a highly specialised skill. Different fossil groups (e.g. foraminifera, nannofossils, and palynomorphs) of different ages require deep subject matter expertise, often gleaned over many years of specialised study, in order to provide accurate and reliable interpretations. It can also be a very time consuming process. A single well with tens or hundreds of samples with a rich and diverse microfossil content, can take several days to interpret. This does not include the time to collect and identify the microfossils, which is another area where machine learning and automation can play a role (Gard et al., 2016).

Because the interpretation of biostratigraphic data can be time consuming and requires specialised skills, there is a huge volume of underutilized data within the industry. Releasing value from this historical data alongside accelerating the interpretation of new data will be a significant boon to the industry. Furthermore, it can be considered a test case of how machine learning can be applied to an industry data type that is not dissimilar to other data types (e.g. geochemistry) where machine learning may release value.

Biostratigraphy

Many Phanerozoic sedimentary rocks, no matter if marine or non-marine, contain microfossils. The process of evolution means that particular species (or even higher taxonomic groups) of fossils are adapted to particular paleoenvironmental niches and have a particular stratigraphic range. This makes many microfossils encountered, once recovered from well samples, useful for correlation, age interpretation, and interpretation of depositional setting of any given interval in a well. The value of fossils for correlation has been known for over 200 years and routinely applied to well samples for around 100 years (Jones, 2011).

Biostratigraphic data is typically gathered quantitatively or at least semi-quantitatively for each sample in a well. Thus, every species of a particular fossil group that is found is recorded in each sample either in absolute numbers or relative abundance. Data can be recorded in simple spreadsheets but is more easily interpreted when depicted using industry standard software. With the data displayed as a distribution chart (species on the x axis, depth on the y axis) (Figure 1), a biostratigrapher can interpret the data. This usually involves the recognition of key marker species (either first or last downhole occurrences, or intervals with relative high abundance) or assemblages of species that are thought to have particular stratigraphic and/or paleoenvironmental value. The recognition of such events or intervals can be fairly straightforward in extensively studied stratigraphy with good fossil recovery and where the marker species are well known. It can be more challenging in frontier basins where endemic taxa occur and the order of stratigraphic events is uncertain. Reworking of older microfossils into younger strata, or caving (the downhole collapse of sidewall material from higher in the well) can, in all cases, complicate the picture.

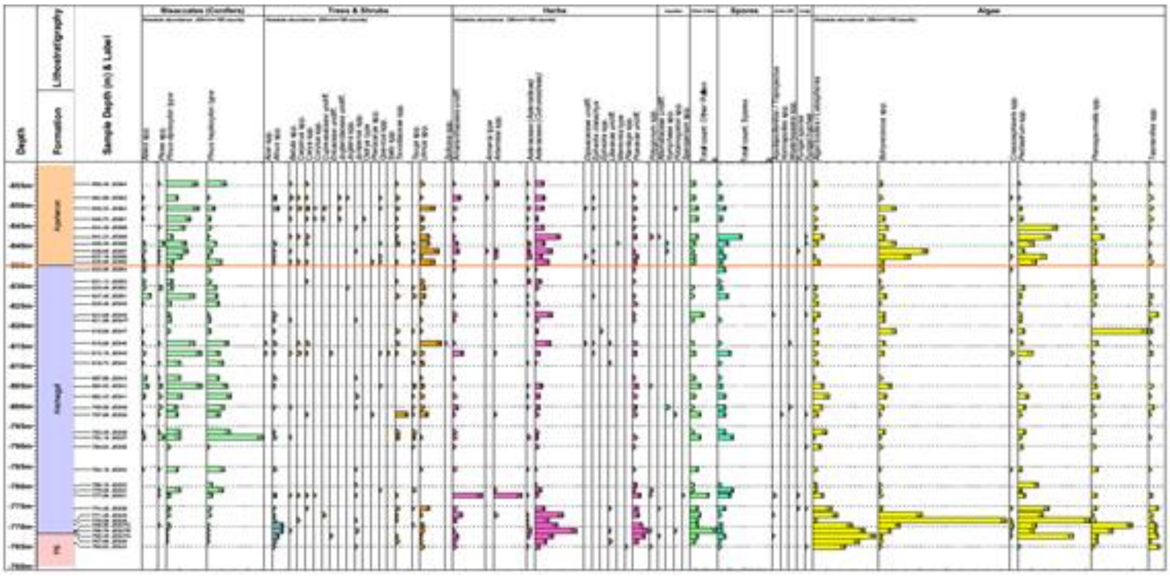

Figure 1: A typical example of biostratigraphic data from a well or outcrop section. Here, absolute numbers of individual species are recorded for each sample in the form of histograms. Data can be much more extensive than this example – wells can contain tens to hundreds of samples and hundreds of species.

Machine Learning

Machine learning is a field of research within computer science which looks to develop computer algorithms that “learn” from data rather than being specifically programmed. As such, machine learning algorithms have the capability to progressively improve the performance on a specific task when exposed to larger volumes of data. It can use supervised or unsupervised approaches. In supervised approaches, machine learning algorithms build a statistical model of labelled sample data, known as training data, to make predictions or decisions without being explicitly programmed to perform the task. Within the training dataset, an interpretation already exists in the form of labels (also known as classes). Machine learning algorithms then seek to identify the basis for the interpretation of each class and apply that model to uninterpreted data. In unsupervised application, machine learning algorithms study all the data without preconceptions to identify potential clusters and patterns; this can be extremely powerful for identifying patterns in multi-dimensional datasets where human interpreters are unable to visualize such complex problems.

An ideal dataset for a machine learning project should be clean (i.e. of a standard format and with minimal spurious data) and extensive such that the “signal” being extracted is visible over the noise. Biostratigraphic data is often suitable although care needs to be taken to avoid taxonomic complexities (e.g. the same species known by two different names) and extensive reworking or caving.

Case Study

The focus for this case study was biostratigraphic data from three wells from anonymous equatorial locations. Two wells form training data; with a third well the target for machine learning-based interpretation. Numerous samples had been analysed from each well and rich and diverse assemblages of foraminifera and calcareous nannofossils were present over an interval that ranged in age from Cretaceous to Neogene incorporating 60 biozones. In total, the three wells contained 768 species within 710 samples. However, in order to capture a full range of possible bioevents, the industry wells were supplemented with data from published sources (e.g. International Ocean Drilling Programme (IODP) reports on 16 wells with relevant stratigraphy) to form a more comprehensive training dataset.

The data was provided without interpretation, thus an initial task was to carry out a human interpretation of biozone/age and palaeoenvironment which would then act as the "labelling" component of the training data. This then provided the context to develop the training dataset and a target to measure the success of the machine learning technique.

To facilitate both human and machine learning based interpretation, it was useful to identify and eliminate spurious data, for example, that resulting from reworking or caving. Such data was identified by reference to species dictionaries that identify the broad stratigraphic range of species and through statistical screening.

The project attempted to use machine learning to interpret both biozone/age and palaeoenvironment. For biozone/age interpretation, a Random Forest algorithm was initially used to predict the biozone probability of individual samples. Randomly dividing samples into training and test data lead to a model with an 80-90% classification accuracy (f1 score). However, when applied to a whole well, the accuracy was no more than 60%, partly a function of treating all occurrences with equal weight and partly a function of the limited training dataset. With this in mind, a Naïve Bayes approach was used to detect biozone presence. This calculated the 'likelihood' that a particular species belongs to a biozone using the training data. This was then used to predict the 'maximum likelihood' of biozones in the test well, paying particular attention to the top of a biozone, in the same way as a practicing biostratigrapher would. This led to much greater prediction accuracy. Importantly, the techniques highlighted the uncertainty in the assignment of a sample to a biozone in terms of a measurement of probability.

Palaeoenvironmental interpretation was expressed as a paleowater-depth curve for the test well. Machine learning outcomes using training data in a similar manner to biozone/age interpretation compared well with human interpretation (Figure 2). Both raw data and feature engineering were used. Feature engineering places fossil occurrences within groups related to their known broad palaeoenvironmental distributions (as a human biostratigrapher would). Interestingly, the outcomes between the two approaches were very similar.

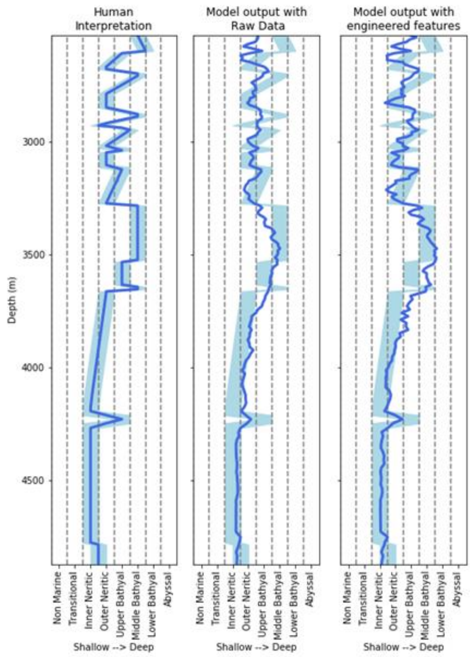

Figure 2: Palaeoenvironmental (paleo-water depth) interpretation of a test well, comparing human interpretation with machine learning outcomes. Shaded areas highlight uncertainty.

A key element in the success of this approach was the integration of subject matter expertise with statistical and coding skills. Biostratigraphers worked alongside those executing machine learning to ensure the results were meaningful. Palaeontological and biostratigraphic expertise proved vital to capture the nuances of interpretation and to ensure meaningful results.

Conclusions

A trial study has shown that machine learning techniques can be applied to biostratigraphic data in order to facilitate rapid, reasonably accurate interpretations of age/biozone and palaeoenvironment (paleo-water depth). Limitations were imposed by the restricted size of the training dataset, but nonetheless, the technique shows promise and merits further investigation. Learnings from the approach can be applied to other, non-biostratigraphic, geological data.

Acknowledgements

We are grateful to Shell for releasing datasets to us and to Dr Helen Smyth who thoughtfully led the Halliburton/Landmark internship programme through which much of this research was conducted.

References

Bond, C.E., Gibbs, A.D., Shipton, Z.K. and Jones, S., 2007. What do you think this is? “Conceptual uncertainty” in geoscience interpretation. *GSA today*, *17*(11), 4-11.

Gard, G., Prince, I., Crux, J.A., Shin, J.M. & Lee, B. 2016. Remote wellsite biostratigraphy and advances in automated fossil analysis. *AAPG Search and Discovery Article* 41930.

Hammer, O. & Harper, D. 2006. *Paleontological Data Analysis*. Blackwell Publishing, 351pp.

Jones, R.W. 2011. *Applications of Palaeontology: Techniques and Case Studies*. Cambridge University Press and Natural History Museum, 420pp.