Abstract

Data-driven models that are built using machine learning (ML) algorithms are becoming increasingly common-place in subsurface science and engineering applications. The impetus for adopting this emerging technology comes from its success in multiple fields such as consumer marketing, finance, design and manufacturing, health care, etc. The use of ML is particularly well-suited for characterizing, describing, and forecasting the behavior of geologic carbon dioxide storage systems where typical data analysis challenges include: (a) incomplete and/or sparse data, (b) unreliable and/or immature physics-based models (if they exist), and (c) usage of data-driven models using conventional statistical methods that are not robust. In this context, ML methods are particularly well suited for capturing hidden patterns and complex input-output relationships automatically from the data without specifying an a priori functional model form for large multivariate datasets.

In this presentation, we will describe the application of ML for two specific problems: (1) geologic data integration, i.e., identification and prediction of electrofacies from well-log data, and (2) operational data analysis, i.e., prediction of bottomhole pressure and temperature in CO₂ injection wells from injection rate and wellhead pressure and temperature measurements. The identification of electrofacies helps build robust predictive models between well-log attributes and dynamic reservoir properties such as permeability. Also, bottom-hole gauges are not present in all wells (or they tend to malfunction from time to time), so the ability to predict pressure and temperature downhole from surface measurements is a valuable capability.

Our systematic ML workflow for building data-driven models involves the following steps: (a) exploratory data analysis to visually understand patterns, trends and outliers in the multivariate datasets, (b) statistical imputation to fill-in missing values (if any), (c) unsupervised learning to identify natural groupings (statistically homogeneous subsets) across the space of independent variables (predictors), and (d) supervised learning to fit predictive models between known predictors and responses (dependent variables). Unsupervised learning is typically carried out using principal component analysis (PCA), k-means clustering (KMC), hierarchical clustering (HC), etc. Supervised learning can be formulated either as a classification problem where the response is categorical, or a regression problem where the response is continuous. This is typically carried out using algorithms such as k-nearest neighbors (kNN), random forest (RF), artificial neural network (ANN), etc.

* Corresponding author. Tel.: +1-512-351-6038, E-mail address: mishras@battelle.org
For the geologic data integration problem in a CO$_2$ enhanced oil recovery and associated geological storage project, well-logs from 250+ oil wells in the Albion-Scipio field in Southern Michigan were collected in a database. The database was checked for outlier values which were either rectified or eliminated. Missing values for several well-logs were then imputed using a Random Forest algorithm to create a full database. Cluster analysis using both kMC and HC were used to identify the presence of 6 natural groups (or electrofacies) within the dataset (left panel, below). Finally, highly accurate models to predict the electrofacies based on well-log attributes were built and validated using both traditional statistical approaches (i.e., logistic regression) and machine learning (i.e., RF) approaches (right panel, below).

For the bottom-hole pressure and temperature prediction problem, data from 3 different CO$_2$ injection wells in different pinnacle reefs in Northern Michigan were collected. The dataset includes hourly values for wellhead pressure, wellhead temperature, wellhead density, injection rate, bottomhole pressure and bottomhole temperature. For all three wells, examination of the data revealed a bifurcation of the data around a wellhead density of 25 lb/ft$^3$. Therefore, different models were built for high- and low-density subsets of the data corresponding to this threshold. As in the previous case, missing values were filled in (imputed) using a random forest regression approach. Separate predictive models were built for bottom hole pressure and temperature as a function of the surface conditions for each well. The baseline model was a multivariate linear regression model with quadratic and cross terms. Machine learning options included a kNN model, an RF model, and an ANN model. The models were validated using three replicates of randomized 80-20 split sample testing (i.e., 80% training and 20% test data). The machine learning models were generally more successful, as demonstrated from the performance of the kNN model on the held-out data for one of the wells in the figure below.
In summary, a systematic workflow for ML applications has been demonstrated for two representative subsurface problems: (a) identification of electrofacies from basic well-log attributes, and (b) prediction of bottomhole pressure and temperature from surface measurements. Successful application of ML for these two different types of problems (one for geologic reservoir characterization side and one for operational data analysis) demonstrates the added value from these workflows. We will also discuss the transferability of such models to other wells in the vicinity, as well as across geologic basins.

*Keywords:* geologic storage; carbon dioxide; machine learning; electrofacies; bottomhole pressure; bottomhole temperature