Applying machine learning methods to predict geology using soil samples

TIMOTHY C. C. LUI*, DANIEL D. GREGORY, AND SHARON A. COWLING

Department of Earth Sciences, University of Toronto, Toronto, ON, Canada M5S 3B1
(*correspondence: timmeh.lui@mail.utoronto.ca)

The first step in many projects, whether it be mineral exploration or infrastructure development, involves understanding the underlying geology in order to highlight areas of mineral potential or infrastructure areas of weakness. Unfortunately, sometimes it is difficult to determine geologic unit when bedrock is not exposed at the surface. The issue gets complicated further for regions where drilling is not possible or cost efficient. The purpose of this study is to explore the effectiveness of using machine learning models to predict underlying geology through soil geochemistry.

The study area is within the Klaza Property in the southwestern part of Yukon, Canada. The underlying geology is mainly of igneous origin, mostly Mid-Cretaceous granodiorite. While most previous studies have involved regional scales, in this study we apply machine learning techniques to predict underlying geology at the property scale. Over 6700 soil samples were collected and analyzed for the chemical concentrations of 32 elements. We also include topographical data which is not commonly used but is important for soil processes. These data are used in the machine learning models.

The study tests 10 different machine learning algorithms and their accuracies. The 10 algorithms being studied are logistic regression, quadratic discriminant analysis, nearest neighbors, linear support-vector machine, radial basis function support-vector machine, Gaussian processes, naïve Bayes, artificial neural network, random forest, and AdaBoost. This allows us to identify which algorithms perform well in this type of classification problem.

The study also tests the effectiveness of 4 methods to deal with unbalanced class sizes: to simply use unbalanced classes in the training set, to under-sample majority classes, to over-sample minority classes, and to synthesize new minority class instances. Additionally, the study tests the performance of a multi-classifier system that ensembles all 10 of the previously mentioned algorithms.

The results of this study will help identify the optimal machine learning techniques to predict underlying geology. This will aid in the creation of geologic maps especially in regions that are challenging to map.