Resource Management through Machine Learning

Justin Granek
University of British Columbia
Vancouver, BC V6T 1Z4
jgranek@eos.ubc.ca

Eldad Haber*
University of British Columbia
Vancouver, BC V6T 1Z4
haber@eos.ubc.ca

Elliot Holtham
NEXT Exploration Inc
Vancouver, BC V6N 2S6
elliot@nextexploration.com

**presenting author asterisked*

SUMMARY

In the modern era of diminishing returns on fixed exploration budgets, challenging targets, and ever-increasing numbers of multi-parameter datasets, proper management and integration of available data is a crucial component of any resource exploration program. Machine learning algorithms have successfully been used for years by the technology sector to accomplish just this task on their databases, and recent developments aim at appropriating these successes to the field of natural resource exploration. Numerous algorithms have been attempted for resource prospectivity mapping in the past, and in this paper we apply a modified support-vector machine algorithm to a test dataset from the QUEST region in central British Columbia, Canada, to target undiscovered Cu-Au porphyry districts. The modified algorithm is designed to properly handle the highly variable uncertainty associated with both the training data (ie: geophysics, geochemistry, geological mapping) as well as the training labels (known Cu-Au porphyry targets in the region). Support vector machines are introduced, the challenges of working with geoscientific datasets are discussed, and finally results from applying the modified algorithm to the QUEST dataset are presented.

Key words: Mineral prospectivity mapping, machine learning, support vector machine, data mining

INTRODUCTION

As new resource discoveries become increasingly difficult and costly to find, geoscientists must adapt and look for new technologies with which to explore. With the ever-increasing volume of available geoscience data, one such advance is the adoption of machine learning methods for resource prospectivity mapping (RPM). Facilitated by the widespread use of geographical information system (GIS) software packages, the analysis of spatial datasets including geological maps, geochemical assays and geophysical surveys has opened a new era in resource exploration.

Originating in the late 1980s, resource prospectivity mapping aims to derive a map of resource potential by learning the relationship between known resource occurrences and multi-parameter geoscience datasets. Numerous data mining approaches have been borrowed from the machine learning field, including weights of evidence (Bonham-Carter *et al*., 1989; Agterberg *et al*., 1990; Carranza, 2004), fuzzy logic (Porwal *et al*., 2003), feed-forward neural networks (Singer and Kouda, 1997; Barnett and Williams, 2006) and support vector machines (Zuo and Carranza, 2011; Abedi *et al*., 2012). Although much research has been devoted to algorithm development within the machine learning community, the RPM problem contains difficulties related to uncertainty management and computational efficiency that have yet to be fully addressed. As a result, we developed a new algorithm that we tested on a Canadian mineral exploration example dataset that covers parts of British Columbia, Canada. The available public datasets include airborne gravity, magnetics and electromagnetics, geochemical sampling, geological mapping and a database of known resource occurrences in the region.

METHOD AND THEORY

Resource prospectivity mapping was first developed in the late 1980s (Bonham-Carter *et al*., 1989) at the Geological Survey of Canada as a method for interpreting the spatial patterns in multiple thematic geoscience maps. In the most general sense, the goal is to learn the relationship between multi-parameter geoscience data (ie: geology, geochemistry and geophysics) and known resource occurrences such that resource potential can be estimated in new regions using available data. Mathematically, this can be represented as follows: given a series of geoscience data maps $\mathbf{X}^{\mathbf{train}}$ and a database of known resource occurrences $\mathbf{y}^{\mathbf{train}}$, learn a mapping function $f(\boldsymbol{\theta};\mathbf{X}^{\mathbf{train}})$ which can approximate the relationship between the two, such that resource potential can be estimated in new areas using existing data. In general $f(\cdot)$ can be any function, ranging from simple linear classifiers to complex non-linear functions such as neural networks.

Regardless of the method employed, most of these solutions treat the learning problem as solving for some generic function given a matrix of training data and a vector of training labels. Though not incorrect, this ignores a number of characteristics that make the resource prospectivity mapping problem unique. Due to the difficulty in field testing algorithms for this application and the relatively slow adoption of these methods by industry, most of the work, both past and present, on RPM has been primarily academic (although examples of governments and major resource companies using related methods do exist). The primary challenge for this application is that geoscience data are prone to errors, but none of the work previously mentioned addresses this issue explicitly. This paper will

do so by following the SVM approach and reformulating the problem as a total least squares optimization problem. In this way, the errors in the data can be faithfully and robustly treated.

The basic principle of SVMs is to construct an optimal margin classifier that has complexity based not on the dimensionality of the feature space, but rather on the number of support vectors, thus allowing for sparse solutions in high dimensions. SVM is an optimal margin classifier because the goal is to learn the equation for a hyperplane that separates the different data classes with as large a margin as possible (see Figure 1).

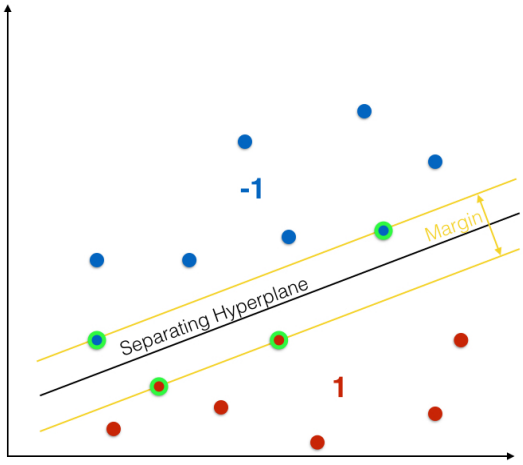

Figure 1: Optimal margin classification using SVM. The separating hyperplane (black) is determined by maximizing the margin (yellow) between a few support vectors (outlined in green).

SVM falls under the branch of machine learning known as supervised learning, in which a predictor is taught using training data and training labels. For the RPM problem, $\mathbf{X_i}$ would be a vector of different geoscience layers for a given sample location, and $\mathbf{y_i}$ would signify "resource" or no "resource" for that location. If the data is linearly separable, then one can define a separating hyperplane

$$f(x) = \mathbf{Xw} + \mathbf{b}$$

where $\mathbf{w}$ and $\mathbf{b}$ are weights with the normalization $|\mathbf{X_n w} + \mathbf{b}| = \mathbf{1}$. The problem then becomes one of maximizing the margin between training points of opposing classes. This is equivalent to the following optimization problem

$$\min \ \tfrac{1}{2}\, \mathbf{w}^T\mathbf{w} \quad \text{Subject to}$$

$$\mathbf{y_i}\,(\mathbf{X_i w} + \mathbf{b}) \ \geq \ \mathbf{1}, \quad \mathbf{i = 1,\ldots,n}$$

which can be solved using quadratic programming or a number of iterative gradient-based methods. Although many black-box software packages exist for various machine-learning algorithms, the optimization problem presented by RPM has a number of practical characteristics that should be considered when implementing a solver. In most supervised machine learning environments, unbiased training is achieved by approximately sampling uniformly from each class. When this is not true, the problem is termed "imbalanced" and can lead to poor generalization of the resulting predictor. As might be expected, resource occurrences are relatively rare, resulting in an extremely imbalanced set of training labels. On top of the imbalanced nature of the RPM problem, there is a large degree of uncertainty associated with the training labels. In this regard, there is a fundamental problem that, in most cases, a label of "no resource" simply means that a reserve has not been discovered, and not necessarily that one does not exist. It is clear that a classification of "no resource" has very different implications in the middle of a highly explored region than it does in a remote location with little exploration history.

As with any observed data, the training data in the RPM problem are associated with uncertainty from various sources. Some data, such as a magnetic total-field measurement, will have numerical uncertainties associated with detection limits and processing procedures. Others, such as geological mapping of lithological units, will have qualitative uncertainties associated with expert interpretation and sampling bias. Additionally, some data can have a spatially correlated uncertainty introduced by different exploration environments in the field (e.g., beneath thick deposits of overburden or seawater where it becomes prohibitively difficult to map bedrock). Unlike many machine-learning problems, where both the data and the labels are reliable, it is known in resource prospectivity mapping that both have associated errors.

The regions considered for RPM are often quite large, which combined with the large number of data and the ever-growing number of predictive data sources (e.g., seismic data, potential field data, electromagnetic data, fault locations, bedrock period, etc.) often leads to very large and dense data matrices. Algorithms for solving this problem need to be able to handle large-scale learning of non-linear relationships without prohibitive computational requirements. Before one can contemplate how best to integrate numerous geoscience datasets, an understanding of the data is required (see Figure 2 for a survey of practical challenges when working with real geoscientific data). A typical exploration program will employ data from three primary disciplines: geology, geochemistry and

geophysics. The variety of data within each of these is broad, comprising qualitative and quantitative measurements, inferred or interpreted values, and a large range of data resolutions and uncertainties. In an idealized exploration environment, all datasets would be densely sampled at the same locations, giving uniform coverage of the area of interest with an associated known uncertainty for each survey. The reality is more commonly represented by a scenario in which each survey is run independently, with highly varied sampling schemes, areas of coverage and target resolutions. The uncertainty on the data is typically a combination of numerous factors that also vary from survey to survey many of which are either qualitative, inferred or simply unknown. In most exploration environments, the available data will span multiple exploration programs, each potentially having a different area of interest. Even within a single exploration program, each survey will likely have a different sampling scheme based on the specific parameter being measured (e.g., bedrock geology may only be known at drill hole locations, whereas geophysical data are typically collected at predetermined grid points).

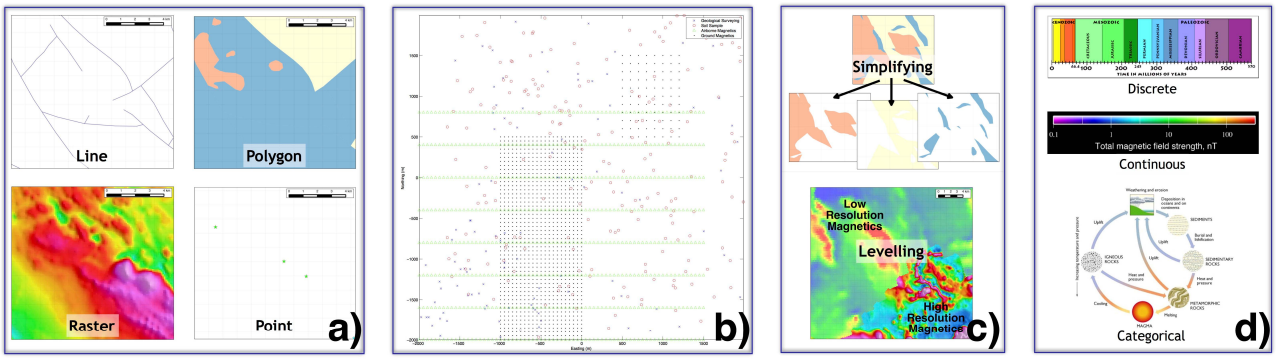

Figure 2: Practical challenges of working with real geoscientific data. a) Data have different spatial representations b) Surveys have different layouts and spatial coverage c) Some datasets require pre-processing to better represent the values d) Data values can be discrete or continuous numbers, as well as categorical labels.

To train the learning algorithm, the different measurements must have some geographic basis and, therefore, a domain must be specified. Traditional methods of prospectivity mapping, such as weights of evidence, handled this by reducing all field measurements - be they point measurements or polygons - to a series of overlapping polygons, each with a single value for each parameter. The approach used in this paper is to define a sample grid that specifies the target resolution of the prospectivity map. Continuous point measurements (i.e., geophysical fields or geochemical assays) can then be interpolated and resampled at the grid points, and discrete or categorical polygon layers (i.e., geological units) can simply be sampled at the specified locations. In this way, it is possible to define a spatial uncertainty for each data layer based on the distance from the field measurements to each grid point. Under this approach, sparse or missing data are easily handled as well, since they will simply have very large uncertainty, thus effectively removing their impact on the training of the algorithm.

RESULTS

To demonstrate the utility of such an algorithm for resource-prospectivity mapping, the following example from the QUEST (Quesnellia Exploration Strategy) project in central British Columbia, Canada is presented. This region is known to host a number of large, economic, copper-porphyry deposits. Through a government-sponsored program, a large amount of geoscience data (including airborne gravity, electromagnetic and magnetic data, age and lithology of bedrock, and geochemical analysis) was acquired between 2008 and 2012 in order to stimulate mineral exploration. Since each dataset was collected independently with its own sampling scheme, all layers were re-sampled to a base grid of 300 m by 300 m, resulting in more than 700, 000 sample points. When all data were assembled and properly processed for training, 91 distinct input layers were used, including both continuous and discrete values.

Uncertainty on these inputs can vary widely, depending on the data source. For example, most geophysical data can bear uncertainty in the form of a noise floor plus an acceptable standard deviation, whereas it is less obvious for geological data due to the subjective, interpreted nature of the measurements. In these cases, estimates can still be made based on confidence in the expert and the availability of field measurements. As previously mentioned, the labels for the RPM problem present a suite of practical issues. The lack of confident negative labels (no resource) results in an imbalanced learning problem, and the sparse subjective nature of the positive labels (resource) results in a large range in confidence that can be adequately quantified using a framework of uncertainty estimates. For the QUEST dataset, 155 alkalic copper-porphyry style mineral occurrences were used to generate a set of binary labels on the base grid. Each occurrence has associated with it a status ranging from “Showing” to “Producer” (six unique statuses are possible), indicating the confidence in the mineral occurrence being economic. Combining this with other factors such as the extent of the overburden (concealing potentially mineral-bearing bedrock), uncertainty estimates for the labels were generated, ranging from 1 (confident label) to 50 (not confident label). The final result is a predicted mineral prospectivity map (Figure 3) that indicates which regions are more favourable for copper-porphyry mineralization than others. We were able to train the algorithm on a (training) subset of the QUEST data and successfully predict the known prospective regions, as well as highlight potential new areas of exploration in a separate (validation) subset. The addition of uncertainty estimates in the algorithm provides a more robust framework for the incorporation of multidisciplinary data with varying spatial resolution and quality.

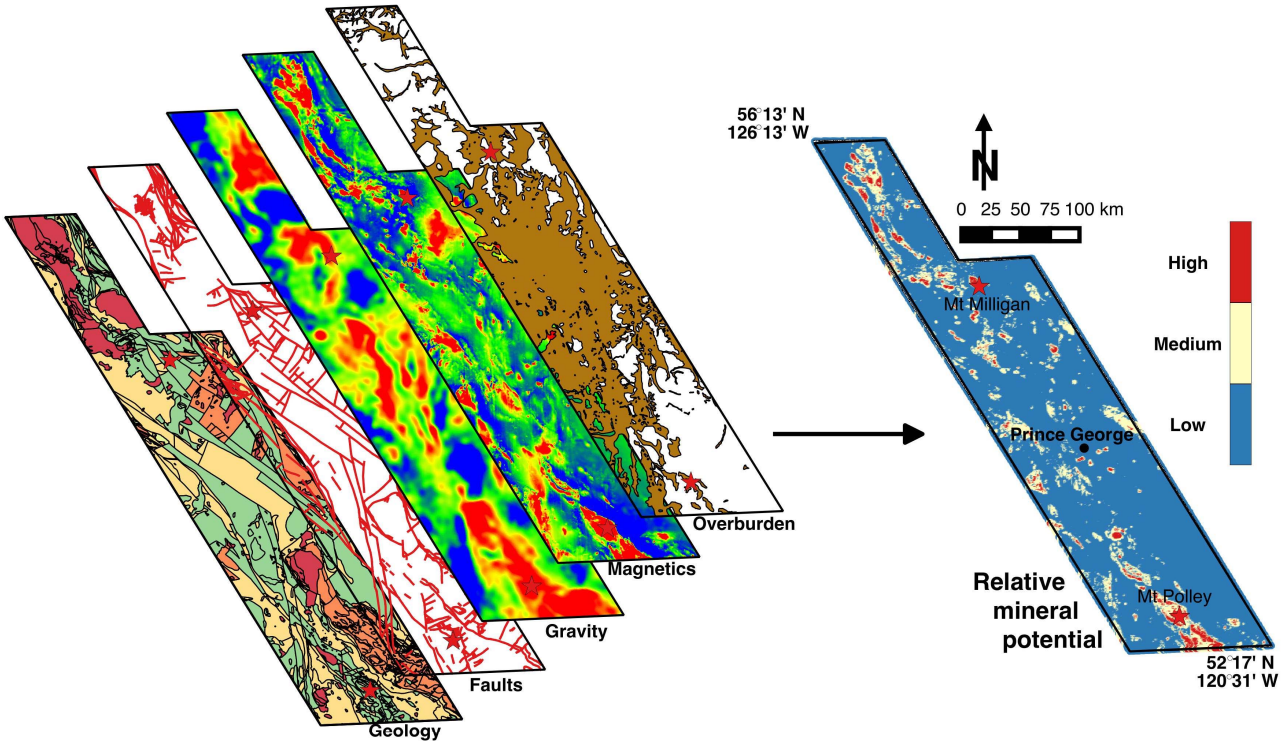

Figure 3: Resource prospectivity mapping combining multiple geoscience datasets into a single output map of resource potential to highlight regions most favourable for future exploration.

CONCLUSIONS

Although much work has been done in the fields of machine learning, geoscience and GIS independently, the intersection of all three fields has left an exciting opportunity to improve the resource exploration process. In this paper we have presented a quantitative approach to integrate geoscience datasets into a map of resource prospectivity. Using a subset of the data from the Geoscience BC Quest project, this paper addresses several challenges through the modification and application of a primal support-vector machine algorithm, incorporating uncertainties in both the labels and the data. In conjunction with this work, much effort has been directed at the proper understanding and use of data preparation prior to learning. Going forward, the authors are exploring different algorithms that are better able to handle the spatial and structured nature of many geoscience datasets; most notably, current research is working on developing a convolutional neural network for resource prospectivity mapping. Such algorithms show promise as tools for geoscientists, both in the early stages of exploration by allowing for critical appraisal of which datasets are most valuable, and in the later stages to extract maximum value from existing data-rich environments.

ACKNOWLEDGMENTS

Thanks to NEXT Exploration Inc for assistance in processing the QUEST data and input on efficient development of the machine learning algorithm.

REFERENCES

Abedi, M., Norouzi, G.H. and Bahroudi, A. [2012] Support vector machine for multi-classification of mineral prospectivity areas. Computers & Geosciences, 46, 272–283.

Agterberg, F., Bonham-Carter, G. and Wright, D. [1990] Statistical pattern integration for mineral exploration. Computer applications in Resource Estimation Prediction and Assement for Metals and Petroleum, 1–21.

Barnett, C. and Williams, P. [2006] Mineral Exploration Using Modern Data Mining Techniques. First Break, 24(7), 295–310.

Bonham-Carter, G., Agterberg, F. and Wright, D. [1989] Integration of geological datasets for gold exploration in Nova Scotia. Short Courses in Geology 10, 15–23.

Carranza, E.J.M. [2004] Weights of Evidence Modeling of Mineral Potential: A Case Study Using Small Number of Prospects, Abra, Philippines. Natural Resources Research, 13(3), 173–187.

Harris, D., Zurcher, L., Stanley, M., Marlow, J. and Pan, G. [2003] A Comparative Analysis of Favorability Mappings by Weights of Evidence, Probabilistic Neural Networks, Discriminant Analysis, and Logistic Regression. Natural Resources Research, 12(4), 241–255.

Porwal, A., Carranza, E. and Hale, M. [2003] Knowledge-driven and data-driven fuzzy models for predictive mineral potential mapping. Natural Resources Research, 12(1).

Schmitt, E. [2010] Weights of Evidence Mineral Prospectivity Modelling with ArcGIS .

Singer, D. and Kouda, R. [1997] Use of a neural network to integrate geoscience information in the classification of mineral deposits and occurrences. Proceedings of Exploration, 127–134.

Zuo, R. and Carranza, E.J.M. [2011] Support vector machine: A tool for mapping mineral prospectivity. Computers & Geosciences, 37(12), 1967–1975.