Paper 23

Mining BIG Data: The Future of Exploration Targeting Using Machine Learning

Desharnais, G. [1], Paiement, J.P. [1], Hatfield, D. [1], Poupart, N. [1]

1. SGS Canada Ltd. Geological Services, 203 de la Seigneurie, Blainville Québec, Canada

ABSTRACT

Exploration expenditures will increasingly be shifting to deeper domains and blind targets as the "easy discoveries" are progressively exhausted. The mining industry's discovery rate has fallen significantly in the past decade because of this, which begs the question: "have we fully optimized the exploration targeting process?" The application of the rapidly evolving science of machine learning coupled with increasingly powerful computers have resulted in process optimizations and breakthroughs in many other industries such as medicine and transportation. The minerals industry is poised to drag their datasets into the 21st century to unlock the predictive capability of these powerful algorithms. Many mining or exploration companies have large amounts of historical data, within which clues to mineralized systems are hiding. Unfortunately, much of these data are in a poor state, often on paper, and require a significant investment to digitize and validate them. In terms of new data capture, the advancement of geophysical-survey technology and hyperspectral core logging tools have been in-step with our technological capability to store and process this rich data. Significant advancements have been made in the past decade to convert this 1- or 2-dimensional data into the 3D realm through inversions, stochastic modelling and implicit modelling. Our capability to harness these massive databases and establish vectors to ore, however, has been limited by the human brain's limited capacity to see patterns in multidimensional data. However, this is a particular strength of machine learning. Unlike other industries where data is collected where it is most useful, we do not have drill hole data where we want to discover new deposits. The clustered nature of the data around known deposits is a major challenge for the application of the many algorithms readily available, and we must rely more heavily on indirect measurements such as geophysics, geochemistry, extrapolation and interpretation. Further research is needed to establish what are the most robust and productive algorithms that will enable prediction of ore bodies. Careful consideration of the inputs by human geologists is required to ensure that the model does not merely predict what is already known, or provide spurious results. This requires high quality geoscientific data, solid interpretations, a good dose of common sense, and in most cases several iterations to understand what the software is predicting. This paper proposes a standard workflow for the effective application of supervised machine learning to exploration targeting.

INTRODUCTION

The past decade has seen an acceleration of several persistent trends in exploration investment and exploration success rates. This is well illustrated in a recent compilation by Shodde (2017):

1. A transfer of investment from major mining companies to junior exploration companies
2. A transfer of investment from greenfields projects to brownfields
3. A reduced rate of discovery of Tier 1 deposits
4. A reduced investment efficiency for the discovery of significant deposits

The decreasing efficiency with which we are finding new ore bodies can be expected to continue unless a sufficiently disruptive exploration methodology or technology is applied to the problem. Past examples include geophysics and geochemical tools that have been used to directly or indirectly detect hidden mineralization. We can expect further developments in these technologies with the advancement of miniaturization, electronics, inversion modelling and drones; these are discussed elsewhere in this volume.

The phenomenal increase in the physical capacity of computers to crunch data coupled with major developments in artificial intelligence, particularly in the subfield of machine learning has caused major disruptions in several industries. Notable examples of the disruptive impact of machine learning to other fields include autonomous vehicles, genetic engineering, targeted customer marketing, financial markets, military weaponry, medicine, agriculture and various language applications (Freund, 2017). The capability of machine learning tools to sift through data to find patterns and classify a targeted result has in many instances surpassed expectations and continues to evolve rapidly. This paper provides an overview of the challenges specific to the application of machine learning to the problem of predicting the location of ore bodies.

The authors of this paper are mining professionals and data scientists that formed the core of the SGS Geostat team, winning the 2016 Integra Gold Rush Challenge. This crowd funded competition released historic data on the Sigma-Lamaque gold property in Val D'Or, Canada, and challenged the public to find innovative ways to identify new drill targets. A strategy that combined a traditional weights of evidence approach coupled with machine learning and target vetting in virtual reality was applied in the winning submission (Figure 1). This paper discusses the particular challenges facing the application of machine learning or other artificial intelligence techniques to the problem of exploration targeting. The specific algorithms to be used and the mechanics of those tools are not explored here. The proposed workflow is applicable to supervised learning regardless of the chosen algorithm: i.e., cases where we can

In "Proceedings of Exploration 17: Sixth Decennial International Conference on Mineral Exploration" edited by V. Tschirhart and M.D. Thomas, 2017, p. 319–323

identify a training set of data that can be tagged as either ore or waste.

Figure 1: Oblique-view looking north of the targets defined by the SGS Geostat team for the 2016 Integra Gold Rush Challenge showing 50 m wide blocks colour-coded by exploration prospectivity: hotter colours are more likely to host gold mineralization. Numbers show the individual targets that were vetted and put forward by the team. The black shapes near the center of the image are the underground workings from Integra Gold Corp's historic Sigma-Lamaque gold mine.

REQUIREMENTS FOR APPLICATION OF MACHINE LEARNING TO EXPLORATION

There are a plethora of machine learning algorithms available, and some have been applied to exploration targeting for more than twenty years (e.g. Singer and Kouda 1996, Porwal et al., 2004). The application of these methods has been met with some apprehension in the mining industry in part due to the lack of transparency of some of the "black box" approaches. In particular, deep learning neural networks have shown immense promise in many other industries; however, these methods are completely opaque with no possibility of understanding the logic behind decisions or conclusions that are made. To increase the confidence of exploration budget decision makers, it is recommended to apply other more transparent methods in parallel, such as weight of evidence (Fallara et al., 2006) or machine learning methods which give some qualitative information on the relative weighting, often called "grey box" methods.

There is a wide breadth of capability and efficiency among the various machine learning techniques, this variability is further increased when combining techniques or tweaking the algorithm parameters to achieve a balanced result. Furthermore, each given problem has an optimal algorithm to solve it, and careful selection of the appropriate machine learning approach is needed to ensure the best possible results. To provide confidence that the algorithms are doing much better than random guessing or predictions by geologists using traditional methods, we must try to quantify how good they are at predicting the state of a given block or node as either ore or waste. Most algorithms have some "control knobs" that can be set to increase selectivity of the method and drive the result towards a higher precision or have a higher recall (Figure 2). *Precision* is an explicit term used in statistical analysis, which in the case of exploration targeting refers to the capability of the algorithm to limit the number of wasted drill holes. *Recall* on the other hand attempts to quantify the number of ore bodies that will be missed by wrongly predicting it as waste. A selected algorithm at a given set of parameters will tend to favour either the recall or the precision. This is a classic trade-off for which either extreme (missed ore bodies, or wasted drill holes) is undesirable. More comprehensive tools to measure the efficacy of the prediction include the use of an *F score* or *Receiver Operating Characteristics* (ROC) analysis and plots (Powers, 2007). These analyses are carried out on the *test set* as described below.

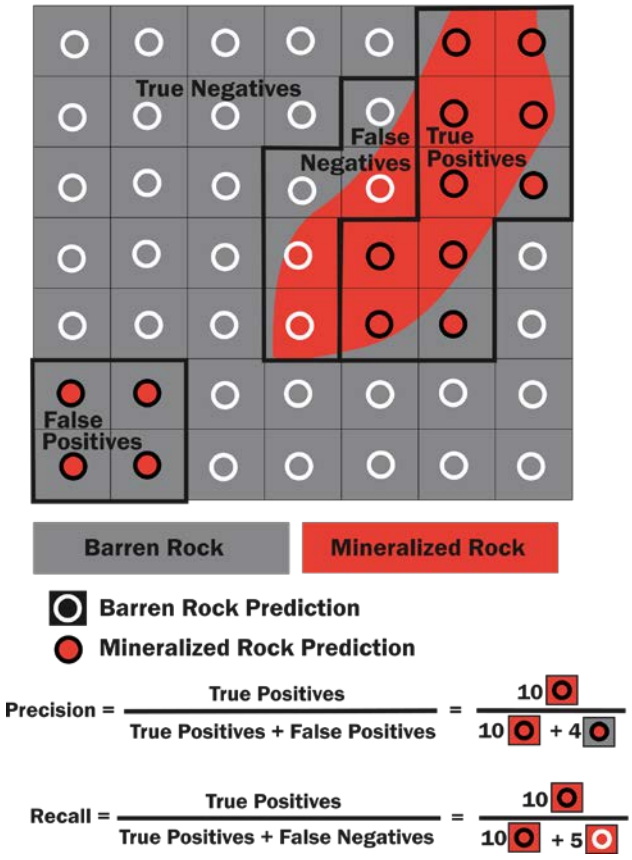

Figure 2: Cartoon explaining the possible combination of predictions and outcomes as applied to mineral exploration. The formulas for precision and recall are also shown.

The unscreened and unsupervised application of powerful algorithms to the problem of exploration targeting is likely to provide spurious results. Some machine learning techniques have shown immense promise of dealing with messy inputs and vague questions outside of the minerals industry (e.g. Kalyanpur and Murdock, 2015); however, these methods have yet to show how to deal with the many issues specific to mineral exploration data, as discussed below. Some prerequisites are necessary to ensure that the process is well controlled and to ensure that the results are not misleading.

MINERAL EXPLORATION DATA

Data produced by exploration is poorly suited to direct application of machine learning techniques. Whereas most other industrial applications have consistent datasets and often clear

positive and negative results, exploration data is plagued with issues requiring special care at several steps to avoid spurious results.

1. Historic data is often on paper, with poor documentation on the location and quality of the surveys or drill logs. Significant time and expense is required to properly digitize the data, this should always be done under the supervision of senior qualified professional to ensure that the key elements are captured with appropriate quality assurance procedures.
2. An unbalanced number of positive and negative examples is problematic for most machine learning techniques; most exploration projects will have this characteristic.
    a. Data in mature camps are often strongly clustered around known deposits with few data points in areas that remain to be explored. This is particularly true for drill data.
    b. In some cases no known ore deposits are located on the property in question, and therefore no examples of success are available. The direct application of an algorithm calibrated on a seemingly analogous project may prove useful, or completely misleading.
3. There may be several deposit types on a single property, each with data associations and probability distributions that can be contradictory in terms of the measured properties.
4. Data is of unequal quality over the area of interest. Special pre-processing or workflow structures are required to deal with missing attributes for certain data points:
    a. Drill data represents the ultimate truth, but is very localized.
    b. Geophysical or geochemical surveys sometimes only cover a fraction of the property.
    c. In several generations of work there may be data of uneven quality; the number and quality of geochemical analysis in particular.

The uneven and clustered nature of the data requires the geologist and data scientist to invest a significant amount of time to consider each data type and to decluster, interpolate, extrapolate or eliminate it.

PROPOSED WORKFLOW FOR APPLYING MACHINE LEARNING TO EXPLORATION TARGETING

This section proposes a workflow for the application of machine learning or other algorithms to exploration targeting. Future developments may facilitate certain steps in this process (particularly steps 2, 3, 4, 7 and 8); however, this rigorous and onerous work structure is necessary to ensure that resulting drill targets are optimal:

1. **Validate Raw Data**: Detection of errors or inconsistencies in the datasets ahead of any work will help ensure that spurious results are avoided. Inconsistencies as simple such as differences in units (e.g. g/tonne versus oz/ton) will be detrimental to the end result. In many cases it is best to ignore or discard unreliable or inconsequential data.
2. **Construct Robust Interpretations**: The regional and local geological, structural, metamorphic and geochemical domains can usually be interpreted and modelled in a relatively robust manner by compiling mapping and sampling from the field combined with geophysical tools. The resulting 3D model will provide a reliable framework to limit interpolation and extrapolation of the data in 3D space.
3. **Create a Declustered Framework**: For the algorithms to tell us where the most prospective targets are, it requires a system in which data can reside in 3D space upon which the inferences or classifications can be made. This framework requires locations where we have data as well as locations where we have yet to explore. These can be equidimensional blocks, or nodes; or a more flexible coordinate system that spans the volume upon which the targets will be created and evaluated. The distance between nodes should ideally be on the scale of a drill target (50–200 m). This approach is also required to decluster drilling information around areas of high information density.
4. **Interpolate and Extrapolate Data**: The garbage-in garbage out (GIGO) principle is particularly relevant when handing over trend identification tasks to a computer. Exploration professionals understand, or can estimate the distance beyond which extrapolation of data is unreasonable. The machine learning algorithms, in contrast, may not hesitate to extrapolate a single anomalous nickel assay for kilometres, ignoring lithological boundaries, if there are no other data points that contradict this interpretation. We must rely on existing, or develop new, geostatistical and geophysical inversion tools to appropriately redistribute available data beyond their initial collection location; all the while respecting, where appropriate, the boundaries defined in step 2.
5. **Isolate a Test Set**: It is important to have a portion of the declustered dataset put aside to test the success rate and potential overfitting of the machine learning algorithm (Figure 2). This subset of data should capture aspects from the main lithological, alteration, geophysical, geochemical and mineralization types to be relatively representative of the whole dataset. This is ideally accomplished through a random selection followed by a cross-validation method.
6. **Create a Training Set**: This is a second subset of the data upon which the selected algorithm will attempt to learn the patterns of data (model) that best classify blocks (or nodes) between positive (ore) and negative (waste) results. Just like the “test set” described above, the training set should be as representative as possible of the data as a whole. It should also have a relatively balanced number of positive and negative examples.

7. **Learn Model on Training Set**: Feed the data into the selected machine learning algorithm and let it optimize the classification of ore blocks and waste blocks. See the discussion pertaining to Figure 2 to understand the inevitable trade-offs and the actual success metrics of your algorithm. Some adjustments to the algorithm can be done to optimize balance between recall and precision.
8. **Evaluate the Model Using the Test Set**: The predictions of the model at the test data locations are compared with the test data itself to assess the performance of the model. In contrast, comparing the predictions at training data points with the training data is a statistically unreliable metric of model performance and subject to a severe risk of the model over-fitting to the data.
9. **Professional Validation:** A review and validation by an experienced exploration professional is required to identify predictions that are not coherent with reality. Specific issues are identified and eliminated such as looking for patterns in predictions that are related to a specific generation of drilling or a spatial distribution which may suggest a bias in the input data or that violates known constraints.
10. **Apply to Rest of Data**: Once results of the algorithm are optimized and validated, we can apply the prediction model to the remainder of the dataset.
11. **Drilling Target Selection**: The final prediction of ore in the model's nodes should go through a rigorous vetting process where targets are ranked according to their probability of success, and judged against the objectives of the campaign. For example, we expect haloes of high probability targets surrounding known mineralization, but this may not be the primary goal of the project.
12. **Plan and Execute the Drill Campaign**: The drill plan should be planned by an experienced exploration professional in much the same way that would be done while targeting geophysical or geochemical anomalies.
13. **Start over at step 1.** Results from the drilling campaign (even misses) will contribute to improve the model's predictive capabilities.

FUTURE DEVELOPEMENT

The next decade of research will enhance our AI toolkit and our ability to apply existing methods to further enhance our capability to discover the next generation of ore bodies.

Data Capture and Validation

Developments in data capture and validation of paper reports would provide substantial savings of time and money required for transforming data from historic projects in a format suitable for analysis. Significant developments in optical character recognition (OCR) have been made in the past decade (e.g. Riedl et al. 2016); however, more intelligent systems are required to deal with the many inevitable and time consuming exceptions that occur (e.g. smudges, changes in unit, symbol for "below detection limit"). Data in historic exploration reports is often internally disconnected: for example drill hole log on one page, assay data in a separate table and the location only found on a map based on a local grid. Linking these pieces together and manually correcting the exceptions represents the single most onerous step in many targeting projects.

Inter-Node Trend Recognition

The work described herein considers each block or node as a single entity which contains the best estimate of each data-type. Each block is treated in isolation without for the most part considering its position within a bigger picture. Machine learning methods capable of recognizing the relative spatial position of nodes in space and find associations that vector towards ore could be developed and optimized. The textbook alteration mineralogy zonation around copper porphyry deposits effectively illustrates how linking information in space could be used as a 3D vector towards a blind target (Sillitoe, 2010). Creating a system whereby potential trends in data are correlated between nodes with directional tracking increases exponentially the mathematical problem. Consider for a moment the Integra Gold Rush dataset created by the SGS Geostat team, which was declustered into almost one million blocks; now consider that an algorithm would have to search for trends in the data, not only within each block but for each possible combination of pairs of resulting in 1,000,000n! ($8.3x10^{5,565,708}$) separate pairs. Each one of these pairs has 17 different data types. In the case of copper porphyry systems, one could imagine alteration in a sequence of four mineral assemblages pointing towards ore – this case is considerably worse. It is difficult to imagine that an algorithm could solve this problem on a human time scale. There are some obvious shortcuts that would considerably reduce the problem by confining the search to local neighborhoods, or simply artificially thinning the dataset. Regardless, this is an interesting problem that merits development and testing.

Applying an Algorithm Without a Property-Specific Training Set

Grassroots exploration projects by definition have no examples of ore from which a machine learning algorithm can learn the data associations. More advanced projects would benefit from an algorithm that recognizes ore body types that have yet to be recognized on that property. A "master algorithm" could be imagined that could read the raw data from any given property and recognize any number of ore deposits regardless of geography or the data types available. This master algorithm would need to be able to manage all the shortcomings for mineral exploration data described herein, plus the amplified risk related to the complete detachment of the local data from the learning sets. Consider for example, the simple case of magnetic intensity (maximum, minimum and rate of change thereof) in a project with thick surficial cover compared to one with no cover. There will inevitably be a levelling problem for most of the data types.

DISCUSSIONS AND CONCLUSIONS

We have discussed herein the challenges and a potential workflow that can be applied to the most straightforward application of machine learning to exploration targeting: a node by node ore/waste classifier. The application of these techniques will undoubtedly result in new discoveries, particularly for deposit types that are created through complex systems such as

volcanogenic massive sulphide (VMS). These deposit types leave overlapping and diffuse signals that can be detected through lithology, structure, hydrothermal alteration mineralogy and chemistry, magnetism, conductivity, density, metal loss and endowment. Machine learning techniques thrive in this type of multidimensional data universe, because they are capable of seeing through the clutter to identify combinations of subtle associations that together point to a mineralizing system.

Each mineral property and dataset will have its own specific issues related to data quality and distribution. Thorough validation and verification will help limit spurious results, and the application of basic geological principles (e.g. cross-cutting relationships, scale, causes and symptoms of mineralizing systems) and common sense (e.g. anthropomorphic artifacts, reasonable ratio of ore/waste) will help ensure resulting drill targets are optimized for discovery.

REFERENCES

Fallara, F., M. Legault, and O. Rabeau, 2006, 3-D Integrated geological modeling in the Abitibi Subprovince (Québec, Canada): Techniques and Applications: Exploration and Mining Geology, 15, 27-41.

Kalyanpur, A., and J.W. Murdock, 2015. Unsupervised entity-relation analysis in IBM Watson: Proceedings of the Third Annual Conference on Advances in Cognitive Systems.

Porwal, A., E.J.M. Carranza, and M. Hale, 2004, A hybrid neuro-fuzzy model for mineral potential mapping: Mathematical Geology, 35, 803-826.

Powers, D.M.W., 2007, Evaluation: From Precision, Recal and F-Factor to ROC, Informedness, Markedness and Correlation: Technical Report School of Infomatics and Engineering Flinders University of South Australia, SIE-07-001.

Riedl, C., R. Zanlibbi, M.A. Hearst, and S. Zhu, S., 2016, Detecting figures and part labels in patents: competition-based development of graphics recognition algorithms: International Journal on Document Analysis and Recognition, 19, 155-172.

Shodde, R.C., 2017, MinEX consulting, 2017, Recent Trends and Outlook for Global Exploration, Presentation to the PDAC, http://www.minexconsulting.com/publications/mar2017.html accessed 01 April 2017.

Sillitoe, R.H., 2010, Prophyry Copper Systems: Economic Geology, 105, 3-41.

Singer, D. A., and R. Kouda, 1996, Application of a feedforward neural network in the search for Kuruko deposits in the Hokuroku district, Japan: Mathematical Geology, 28, 1017–1023.

Freund, K., 2017. Five things to watch in AI and Machine Learning In 2017, https://www.forbes.com/sites/moorinsights/2017/01/06/five-things-to-watch-in-ai-and-machine-learning-in-2017/#20e94d41455a, accessed 1 April 2017.