Science of the Total Environment 490 (2014) 182–190

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Uncertainty quantification and integration of machine learning techniques for predicting acid rock drainage chemistry: A probability bounds approach

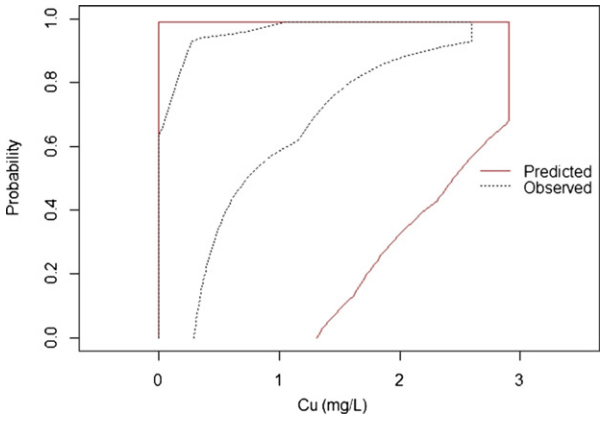
Getnet D. Betrie [a,*], Rehan Sadiq [a], Kevin A. Morin [b], Solomon Tesfamariam [a]

[a] School of Engineering, UBC, Kelowna, BC, Canada
[b] Minesite Drainage Assessment Group, Surrey, BC, Canada

HIGHLIGHTS

- A method to quantify the predictive uncertainty of machine learning was developed.
- Two machine learning techniques were integrated to improve their predictions.
- The sources of uncertainty in model prediction were identified.
- A possible way for reducing prediction uncertainty was suggested.
- A better technique to evaluate the performance of models is found and recommended.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 February 2014
Received in revised form 19 April 2014
Accepted 29 April 2014
Available online 21 May 2014

Editor: F.M. Tack

Keywords:
Acid rock drainage
Machine learning
Artificial neural network
Support vector machine
Uncertainty analysis

ABSTRACT

Acid rock drainage (ARD) is a major pollution problem globally that has adversely impacted the environment. Identification and quantification of uncertainties are integral parts of ARD assessment and risk mitigation, however previous studies on predicting ARD drainage chemistry have not fully addressed issues of uncertainties. In this study, artificial neural networks (ANN) and support vector machine (SVM) are used for the prediction of ARD drainage chemistry and their predictive uncertainties are quantified using probability bounds analysis. Furthermore, the predictions of ANN and SVM are integrated using four aggregation methods to improve their individual predictions. The results of this study showed that ANN performed better than SVM in enveloping the observed concentrations. In addition, integrating the prediction of ANN and SVM using the aggregation methods improved the predictions of individual techniques.

* Corresponding author.
E-mail address: getnet.betrie@ubc.ca (G.D. Betrie).

http://dx.doi.org/10.1016/j.scitotenv.2014.04.125
0048-9697/

1. Introduction

Globally acid rock drainage (ARD) is a major pollution problem that poses severe adverse risks to the environment (Gray, 1996, 1998; Azapagic, 2004). The probable global area covered with mine waste is in the order of 100 million ha that contain several hundred thousand million tonnes of mine wastes, and 20,000–25,000 million tonnes of solid waste is added every year (Lottermoser, 2010). The associated liability costs of potentially acid-generating wastes at minesites are estimated to be US$ 1.2–20.6 billion in USA, US$ 1.3–3.3 billion in Canada, and US$ 530 million in Australia (Miller et al., 2006). However, the authors suspect that the estimate for Canada should be at least an order of magnitude higher.

ARD is generated when a sulfide-bearing material is reacted with oxygen and water during mining activities (Morin & Hutt, 2001; Price, 2009). The reaction results in oxidation and other weathering processes, which changes relatively insoluble chemical species in sulfide minerals into more easily dissolved free ionic species (e.g., Cu, Cd and Zn) or secondary minerals (e.g., sulfate, carbonates and oxyhydroxides). Moreover, the oxidation of some sulfide minerals produces acid that may lower the drainage pH. A lower drainage pH could increase the rate of sulfide oxidation, solubility of many products of sulfide oxidation, and rate of weathering for other minerals.

Predicting the future drainage chemistry is important to assess potential environmental risks of ARD and implement appropriate mitigation measures that reduce adverse environmental risks (Betrie et al., 2012). Predictive models are one of the approaches used to predict the future drainage chemistry of minesites. These models are classified as process-based and empirical (data-driven) models (Maest et al., 2005; Price, 2009). Process-based models describe the ARD system in terms of chemical and/or physical processes that are believed to control ARD generation (Betrie et al., 2012). Nevertheless, the physical/chemical processes that govern generation of ARD are not fully understood (Price, 2009). Subsequently, uncertainty is introduced in the prediction of drainage chemistry due to poor representation of the ARD system. In addition, process-based models introduce uncertainty due to data because they use database information (e.g., solubility product) that might not match a given site (Price, 2009). On the other hand, data-driven models (e.g., machine learning, soft-computing, computational intelligence) describe the time-dependent behavior of one or more variables of the ARD system in terms of observed data trends obtained from years of monitoring at a minesite (Betrie et al., 2012). Therefore, these models are prone to uncertainties in the data that arise due to epistemic (e.g., measurement errors and limited sample size) and aleatory (e.g., temporal and spatial variations) uncertainties, where these uncertainties arise due to incomplete knowledge and natural stochasticity, respectively (Sentz & Ferson, 2002).

The literature review shows that machine learning techniques (e.g., ANN and SVM) have been used to predict the ARD drainage chemistry. Khandelwal & Singh (2005) compared ANN and multiple regression analysis (MRA) to predict chemical parameters (sulfate, chloride, total dissolved solid (TDS) and others) as function of physical parameters (pH, temperature, and hardness). They reported that ANN provided acceptable results compared to MRA. Rooki et al. (2011) evaluated two types of ANN (back propagation neural network (BPNN) and general regression neural network (GRNN)) and MRA to predict heavy metals (Cu, Fe, Mn, Zn) as function of physical/chemical parameters (pH, sulfate, and Mg) in the Shur River near Sarcheshmeh Copper mine, Iran. They reported that the predictive accuracy of BPNN is the best followed by GRNN and MRA. For the Shur River and the same input–output variables, Aryafar et al. (2012) applied SVM and compared to their GRNN model results. The results showed that the predictive accuracy of SVM was slightly better than ANN.

Betrie et al. (2012) evaluated the predictive accuracy and uncertainty of four machine learning techniques (ANN, SVM, mode trees, and K-nearest neighbors) to predict copper concentration as a function of physical/chemical parameters and their time lags. The authors reported that SVM performed best followed by ANN, model trees and K-nearest neighbors both in terms of predictive accuracy and uncertainty. The prediction accuracy refers to the difference between observed and predicted values, whereas the predictive uncertainty refers to the variability of the overall error around the mean error (Betrie et al., 2012).

Although identification and quantification of uncertainties are integral parts of ARD assessment and risk mitigation (Price, 2009), previous studies have not addressed uncertainty issues except a minor attempt by Betrie et al. (2012). In this paper, predictive uncertainties of ANN and SVM due to input data are quantified using the probability bounds approach. The probability bounds approach is an uncertainty analysis method that combines probability theory and interval arithmetic to produce probability boxes, which allow the comprehensive propagation of both variability and uncertainty rigorously (Tucker & Ferson, 2003). Furthermore, predictions of ANN and SVM are integrated using four aggregation methods in order to improve the prediction of the individual technique. Aggregation methods are used to combine information obtained from various sources in order to improve the reliability of information (Sentz & Ferson, 2002).

The remainder of this paper is structured as follows. The next section presents the descriptions of ANN and SVM techniques and the method used for data preprocessing including treating missing and outlier values, defining modeling variables and conducting uncertainty analysis. The Results and discussion section presents the main findings of this study and discusses these findings, respectively. The Summary and conclusions section of this study completes this paper.

2. Material and methods

The methodology followed in this study is depicted in Fig. 1. It shows that the methodology consists of five blocks. In the first block, data pre-processing is done that includes filling missing values and outlier analysis. In the second block, variables that control drainage chemistry of ARD are identified and used to develop model using the machine learning techniques. In the third block, the dataset is divided into training and testing sets using ten-fold cross-validation technique. The training dataset is used to optimize parameters of the models, whereas the test dataset is used for predicting drainage chemistry. In the fourth block, first the predictive accuracy of training models is evaluated using four statistical techniques. If the results of training are not acceptable based on the obtained statistics, the modeling process would be reinitiated from the second block. However, the predictive accuracy and uncertainty for test models would be initiated if the training models provide acceptable results. In the last block, the uncertainties due to data and model are quantified using probability bounds approach. Also, predictions from ANN and SVM are integrated using four aggregation methods to reduce the predictive uncertainty of individual models.

2.1. Machine learning techniques and uncertainty analysis

Machine learning is an algorithm that estimates an unknown dependency between mine waste geochemical system inputs and its outputs from the available data (Betrie et al., 2012). In this study, ANN and SVM techniques are used since they performed well in our previous studies. These two techniques are implemented using WEKA 3.6.4 Software (Bouckaert et al., 2010). The concept of machine learning and the detailed evaluation of various machine learning techniques can be seen in Betrie et al. (2012). The description of ANN and SVM techniques is described in detail, consistent with Betrie et al. (2012), in the following subsections.