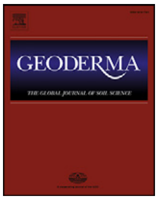
Contents lists available at ScienceDirect

Geoderma

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



An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping

Brandon Heung [a,*], Hung Chak Ho [b], Jin Zhang [a], Anders Knudby [c], Chuck E. Bulmer [d], Margaret G. Schmidt [a]

[a] Soil Science Lab, Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada
[b] Remote Sensing and Spatial Predictive Modeling Lab, Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada
[c] Department of Geography, University of Ottawa, 60 University, Ottawa, ON, K1N 6N5, Canada
[d] British Columbia Ministry of Forests Lands and Natural Resources Operations, Natural Resource Sciences Section, Vernon, BC, V1B 2C7, Canada

ARTICLE INFO

Article history:
Received 11 May 2015
Received in revised form 8 August 2015
Accepted 10 November 2015
Available online 1 December 2015

Keywords:
Digital soil mapping
Machine-learning
Soil classification
Data-mining
Model comparison

ABSTRACT

Machine-learning is the automated process of uncovering patterns in large datasets using computer-based statistical models, where a fitted model may then be used for prediction purposes on new data. Despite the growing number of machine-learning algorithms that have been developed, relatively few studies have provided a comparison of an array of different learners — typically, model comparison studies have been restricted to a comparison of only a few models. This study evaluates and compares a suite of 10 machine-learners as classification algorithms for the prediction of soil taxonomic units in the Lower Fraser Valley, British Columbia, Canada.
A variety of machine-learners (CART, CART with bagging, Random Forest, *k*-nearest neighbor, nearest shrunken centroid, artificial neural network, multinomial logistic regression, logistic model trees, and support vector machine) were tested in the extraction of the complex relationships between soil taxonomic units (great groups and orders) from a conventional soil survey and a suite of 20 environmental covariates representing the topography, climate, and vegetation of the study area. Methods used to extract training data from a soil survey included by-polygon, equal-class, area-weighted, and area-weighted with random over sampling (ROS) approaches. The fitted models, which consist of the soil-environmental relationships, were then used to predict soil great groups and orders for the entire study area at a 100 m spatial resolution. The resulting maps were validated using 262 points from legacy soil data.
On average, the area-weighted sampling approach for developing training data from a soil survey was most effective. Using a validation of $R = 1$ cell, the *k*-nearest neighbor and support vector machine with radial basis function resulted in the highest accuracy of 72% for great groups using ROS; however, models such as CART with bagging, logistic model trees, and Random Forest were preferred due to the speed of parameterization and the interpretability of the results while resulting in similar accuracies ranging from 65–70% using the area-weighted sampling approach. Model choice and sample design greatly influenced outputs. This study provides a comprehensive comparison of machine-learning techniques for classification purposes in soil science and may assist in model selection for digital soil mapping and geomorphic modeling studies in the future.

1. Introduction

Data mining may be defined as the automated or semi-automated process of uncovering patterns from large electronic datasets using trained models, where the patterns may then be used on new data for the purposes of prediction (Witten and Frank, 2005). The process of 'training' a model is also synonymously described as a type of 'learning' where 'machine-learning' can be defined as the process of discovering the relationships between predictor and response variables using computer-based statistical approaches (Witten and Frank, 2005; Hastie et al., 2009).

In soil science, machine-learning techniques have most commonly been used in the subfield of pedometrics for the development of predictive or digital soil maps (DSM; Scull et al., 2003; McBratney et al., 2003) due to developments in geographical information systems, availability of digital spatial data, and constantly advancing computer technology (McBratney et al., 2003). In DSM, the workflow for the environmental-correlation approach (McKenzie and Austin, 1993; McKenzie and Ryan, 1999) entails the collection of soil point or polygon data that are co-located with a suite of *clorpt* soil-environmental variables (Jenny, 1941) in order to develop the training dataset (McBratney et al., 2003). The relationships between the soil and environmental covariates are fitted with a model, and the learned relationships are then applied to locations

* Corresponding author.
E-mail addresses: bha4@sfu.ca (B. Heung), hohung@sfu.ca (H.C. Ho), jza18@sfu.ca (J. Zhang), aknudby@uottawa.ca (A. Knudby), Chuck.Bulmer@gov.bc.ca (C.E. Bulmer), Margaret_schmidt@sfu.ca (M.G. Schmidt).

http://dx.doi.org/10.1016/j.geoderma.2015.11.014
0016-7061/© 2015 Elsevier B.V. All rights reserved.

where soil data are not available. This generic procedure, a form of supervised learning, may be applied to the prediction of quantitative outputs (i.e. soil organic matter content, clay content, pH, or electrical conductivity) using regression, or the prediction of qualitative outputs (i.e. soil taxonomic units) using classification (McBratney et al., 2003; Hastie et al., 2009).

Numerous machine-learning algorithms are available for use, including the commonly used tree-based learners such as the classification and regression tree (CART) learner proposed in Breiman et al. (1984) and its extensions using bagging (Breiman, 1996) or boosting (Breiman, 1998) and subsequently, the development of Random Forest (RF; Breiman, 2001). Other learners less commonly used in DSM include support vector machines (Kovačevic et al., 2010; and Priori et al., 2014), artificial neural networks (Aitkenhead et al., 2013; Priori et al., 2014; and Silveira et al., 2013), *k*-nearest neighbor (Mansuy et al., 2014), and linear approaches (Kempen et al., 2009; Vasques et al., 2014). With the notable exceptions of Brungard et al. (2015) and Taghizadeh-Mehrjardi et al. (2015), the number of models compared in DSM studies have generally been restricted to a few models for each study (i.e. Cavazzi et al., 2013; Ließ et al., 2012; Bourennane et al., 2014; Priori et al., 2014; Collard et al., 2014) rather than an expansive comparison where some learners, commonly used in other fields, have yet to be tested for DSM.

The objectives of this study are (1) to provide an overview of the machine-learning techniques that have been or could be used for DSM; (2) to evaluate and compare a suite of 10 machine-learners as classifiers for the prediction of soil taxonomic units; and (3) to evaluate different methods for generating training data from a conventional soil survey. The evaluation and comparison between the modeling approaches are based on a case study for the Lower Fraser Valley region of British Columbia, Canada, where the various classifiers are used to learn the relationships between soil taxonomic units and environmental covariates through the data mining of a conventional soil survey as described in Heung et al. (2014). In order to make a fair comparison between the learners, model parameters were all optimized to the training data.

2. Overview of machine-learning techniques

Here, a brief overview of various machine-learning techniques is presented. The objective is not to provide a detailed explanation of each approach but rather to provide a summary of several, and their relevance in DSM. In addition to the learners used in DSM, we also explore approaches that have been used in other disciplines but have yet to be explored in DSM. As the objective of this study is to examine machine-learners for mapping soil taxonomic units, this overview is focused mainly on the learners as classifiers for mapping soil classes rather than for the numerical mapping of soil attributes.

2.1. Tree-based learners

Tree-based algorithms are perhaps the most commonly used learners in the DSM literature. Tree-based learners consist of nodes and leaves where each node is a partition of the training dataset that aims to maximize the within-node homogeneity and the between-node heterogeneity based on node splitting rules that are generated from a set of predictor variables — a type of *if–then* statement (Breiman et al., 1984). The leaves are the terminal nodes where a decision is made with regard to the response variable of interest. As a result of their hierarchical structure, tree-based learners are able to represent non-linear and non-smooth relationships between predictor and response variables as well as interaction effects where the relationship between a predictor and the response depends on one or more other predictors. In addition, tree-based learners are also flexible as they are able to handle numerical, ordinal, or discrete predictors, and do not require assumptions on normality (Hastie et al., 2009).

Tree-based learners have commonly been used for classification to map soil taxonomic units (i.e. Behrens et al., 2010; Bui and Moran, 2001, 2003; Bui et al., 1991; Grinand et al., 2008; Jafari et al., 2014; Moran and Bui, 2002; Nelson and Odeh, 2009; Schmidt et al., 2008; Scull et al., 2005; and Taghizadeh-Mehrjardi et al., 2014) or soil parent material classes (i.e. Bui and Moran, 2001; Lacoste et al., 2011; and Lemercier et al., 2012), and more recently, for the disaggregation of complex map units from conventional soil maps (i.e. Nauman and Thompson, 2014; Odgers et al., 2014; and Subburayalu et al., 2014). In addition, they have also been used to map soil attributes such as pH, soil depth, organic C, clay content, and total N and P using regression modeling (i.e. Bui et al., 2006, 2009; Henderson et al., 2005; and McKenzie and Ryan, 1999).

The RF learner is conceptually similar to tree-based learners and shares the same advantages; however, multiple decision trees are trained and the results are based on the predictions from an ensemble of the individual trees (Breiman, 2001). For the RF learner, each tree is trained from a randomized bootstrap sample of the entire training set and a subset of predictors used for the node-splitting rules is also randomly selected. Although the RF learner was adopted early on to analyze large datasets in the bioinformatics literature (i.e. Díaz-Uriarte and Alvarez de Andrés, 2006; Qi, 2012; and Svetnik et al., 2003), its usage in DSM appears to become increasingly more prominent. DSM applications of the RF learner, similar to those of the decision trees, have included the mapping of soil organic C (i.e. Grimm et al., 2008; Guo et al., 2015; and Wiesmeier et al., 2011), soil texture (Ließ et al., 2012) as well as for classification purposes such as the mapping of soil parent material classes (Heung et al., 2014) or the updating and disaggregation of conventional soil survey maps (Häring et al., 2012; and Rad et al., 2014). Despite the similarities between single tree-based learners and RF, few studies in DSM have compared the two, with the exception of Ließ et al. (2012) who compared them for the prediction of particle size fractions using regression and found that RF performed better.

2.2. Logistic regression

A review of DSM approaches by McBratney et al. (2003) identified that linear models (i.e. multiple linear regression and generalized linear models) have commonly been used for mapping soil attributes and have regularly been hybridized with kriging in regression kriging (i.e. Odeh et al., 1995; Hengl et al., 2007). For classification purposes, however, the most frequently used linear approach is through the use of multinomial logistic regression models (i.e. Kempen et al., 2009; Debella-Gilo and Etzelmüller, 2009; Collard et al., 2014; Jafari et al., 2012).

Logistic regression models are a type of generalized linear model that is well suited for datasets where the dependent variable is categorical. These models are able to describe the relationships between a set of predictor variables and a dichotomous dependent variable that has values of 0 or 1. In the binomial case, outputs of logistic regression are expressed in probabilistic terms where values close to 0 indicate a low probability of occurrence, and values close to 1 represent a high probability of occurrence (Kleinbaum et al., 2008).

In order to extend the logistic regression model approach to predict multinomial categorical response variables, both Kempen et al. (2009) and Debella-Gilo and Etzelmüller (2009) propose a multinomial logistical regression approach. In both cases, logistic regression models were developed for each soil class that was found in the study area. The relationships between topography and soil taxonomic units were determined from legacy soil data. In order to convert a set of binomial logistic regression models into a generalized multinomial model, the following equation is used:

$$p_i = \frac{\exp(p_i)}{\exp(p_1) + \exp(p_2) + \ldots + \exp(p_n)}, \qquad (1)$$