Geoscience Frontiers 7 (2016) 3–10

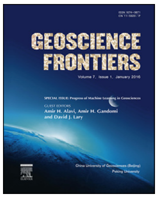
HOSTED BY

Contents lists available at ScienceDirect

China University of Geosciences (Beijing)

Geoscience Frontiers

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



Research paper

Machine learning in geosciences and remote sensing

David J. Lary [a], Amir H. Alavi [b,*], Amir H. Gandomi [c], Annette L. Walker [d]

[a] Hanson Center for Space Science, University of Texas at Dallas, Richardson, TX 75080, USA
[b] Department of Civil and Environmental Engineering, Michigan State University, East Lansing, MI 48824, USA
[c] BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI 48824, USA
[d] Aerosol and Radiation Section, Naval Research Laboratory, 7 Grace Hopper Ave., Stop 2, Monterey, CA 93943-5502, USA

ARTICLE INFO

Article history:
Received 15 April 2015
Received in revised form
17 June 2015
Accepted 17 July 2015
Available online 12 August 2015

Keywords:
Machine learning
Geosciences
Remote sensing
Regression
Classification

ABSTRACT

Learning incorporates a broad range of complex procedures. Machine learning (ML) is a subdivision of artificial intelligence based on the biological learning process. The ML approach deals with the design of algorithms to learn from machine readable data. ML covers main domains such as data mining, difficult-to-program applications, and software applications. It is a collection of a variety of algorithms (e.g. neural networks, support vector machines, self-organizing map, decision trees, random forests, case-based reasoning, genetic programming, etc.) that can provide multivariate, nonlinear, nonparametric regression or classification. The modeling capabilities of the ML-based methods have resulted in their extensive applications in science and engineering. Herein, the role of ML as an effective approach for solving problems in geosciences and remote sensing will be highlighted. The unique features of some of the ML techniques will be outlined with a specific attention to genetic programming paradigm. Furthermore, nonparametric regression and classification illustrative examples are presented to demonstrate the efficiency of ML for tackling the geosciences and remote sensing problems.

1. Introduction

Machine learning (ML) is an effective empirical approach for both regression and/or classification (supervised or unsupervised) of nonlinear systems. Such systems can be massively multivariate involving a few or literally thousands of variables. In ML, a comprehensive 'training dataset' of examples is constructed covering as much of the system parameter space as possible. Typically, a random subset of the data is put aside for a completely independent validation. ML is ideal for addressing those problems where our theoretical knowledge is still incomplete but for which we do have a significant number of observations and other data. In an ideal world, if we had complete theoretical understanding, ML would be superfluous.

ML has proven useful for a very large number of applications in many parts of the earth system (land, ocean, and atmosphere) and beyond, from retrieval algorithms, crop disease detection, new product creation, bias correction and code acceleration (e.g. Yi and Prybutok, 1996; Atkinson and Tatnall, 1997; Carpenter et al., 1997; Lary et al., 2004, 2009; Brown et al. 2008; Azamathulla, 2012; Zahabiyoun et al., 2013; Madadi et al., 2015). The types of the ML algorithms commonly used are artificial neural networks (ANN), support vector machines (SVM), self-organizing map (SOM), decision trees (DT), ensemble methods such as random forests, case-based reasoning, neuro-fuzzy (NF), genetic algorithm (GA), multivariate adaptive regression splines (MARS), etc (e.g., Shahin et al., 2001; Shahin and Jaksa, 2005; Das and Basudhar, 2008; Samui, 2008a,b; 2012; Azamathulla and Wu, 2011; Azamathulla et al., 2011, 2012; Garg et al., 2014a,b,c). The ML-based methods have been widely applied to the science and engineering problems for near two decades. This is while the application of these techniques in the geosciences and remote sensing area is fairly new and limited. Herein, a number of relevant and documented applications of ML will be summarized. The unique features of some of the ML techniques for dealing with the geosciences and remote sensing problems will be reviewed. Moreover, two very different but complementary illustrative examples are presented: one using multivariate nonlinear nonparametric regression, and the other using multivariate nonlinear unsupervised classification. For these two illustrative cases, we will start with the scientific motivation that makes clear the real need for ML and then demonstrate how ML addresses this need.

* Corresponding author. Tel.: +1 (517) 526 1455.
E-mail addresses: alavi@msu.edu, ah_alavi@hotmail.com (A.H. Alavi).
Peer-review under responsibility of China University of Geosciences (Beijing).

http://dx.doi.org/10.1016/j.gsf.2015.07.003

2. Overview of ML applications in geosciences and remote sensing

The ML algorithms are "universal approximators". That is, they learn the underlying behavior of a system from a set of training data. Another interesting feature of the ML-based techniques is that they do not need a prior knowledge about the nature of the relationships between the data. The application of ML may be categorized into three areas (Lary, 2010):

(1) The system's deterministic model is computationally expensive and ML can be used as a code accelerator tool.
(2) There is no deterministic model but an empirical ML-based model can be derived using the existing data.
(3) Classification problems.

As mentioned before, ML includes a variety of algorithms ANN, SVM, SOM, and DT. Over the last decade, there has been considerable progress in developing ML-based methodologies for many of Earth Science applications (Lary, 2010). Some of these studies have received special recognition as a NASA Aura Science highlight (Lary et al., 2007) and commendation from the NASA MODIS instrument team (Lary et al., 2009). ANN and SVM are the most commonly used ML techniques for dealing with geoscience problems. A comprehensive review of application of ANN and SVM in geoscience and remote sensing can be found in Lary (2010). Also, Nikravesh (2007) presented an inclusive review study of the application of neuro-computing, fuzzy logic and evolutionary computing in geosciences and oil exploration. That study also covers the successful application of hybrid methodologies such as NF, neural-genetic, fuzzy-genetic and neural-fuzzy-genetic in the field. Nikravesh (2007) discussed the major impact of these techniques for tackling problems in geophysical, geological and reservoir engineering (e.g., intelligent reservoir characterization and exploration, seismic data processing, and characterization, well logging, reservoir mapping, etc.).

Among the main subsets of ML, applications of genetic programming (GP) (Koza, 1992) in the geoscience and remote sensing domain are very new and restricted to a few areas. Despite the good performance of ANNs, SVM and many of the other ML methods, they are considered as black-box models. That is, they are not capable of generating practical prediction equations. GP is considered as an efficient approach to deal with this issue. GP uses the principle of Darwinian natural selection to generate computer programs for solving a problem. In fact, GP is a specialization of GA where the encoded solutions (individuals) are computer programs rather than binary strings (Alavi and Gandomi, 2011). A notable feature of GP and its variants is that they can produce prediction equations without a need to pre-define the form of the existing relationship (Alavi et al., 2010; Alavi and Gandomi, 2011; Alavi et al., 2011a; Gandomi and Alavi, 2011). Herein, we present an overview of a number of relevant and recent applications of GP in the field. The majority of applications of GP focus on the behavioral characterization of rock mass. The other few studies use GP as a tool for interpreting the remote sensing data. It is worth mentioning that there are some other studies mainly on the applicability of GP for analyzing geotechnical engineering problems such as liquefaction phenomenon, ground motion parameters, or ground movement patterns (e.g., Javadi et al., 2006; Shuhua et al., 2006; Lia et al., 2007; Cabalar and Cevik, 2009; Alavi et al., 2011b; Gandomi et al., 2011; Alavi and Gandomi, 2012; Gandomi and Alavi, 2013).

As mentioned before, most of the GP-based studies focus on estimating the properties of rock. Perhaps, one of the pioneer studies in the field was done by Baykasoglu et al. (2008). They applied GP-based approaches to the strength prediction of limestone. Different variants of GP, called multi expression programming (MEP), gene expression programming (GEP) and linear genetic programming (LGP) to the uniaxial compressive strength (UCS) and tensile strength prediction of chalky and clayey soft limestone. The models were developed using experimental data. The models had a good accuracy with determination coefficient (R^2) equal to 0.76 and 0.95 for tensile strength and UCS, respectively. Beiki et al. (2010) developed new models to determine the deformation modulus of rock masses using GP. Several parameters were used as the predictor variables such as modulus of elasticity of intact rock (Ei), uniaxial compressive strength (UCS), rock mass quality designation (RQD), the number of joint per meter (J/m), porosity, dry density, and geological strength index (GSI). Beiki et al. (2010) also found that the GP models give higher predictions over existing empirical models. Recently, Karakus (2011) employed GP to analyze laboratory strength and elasticity modulus data for some granitic rocks. Uniaxial compressive strength (σ_c), tensile strength (σ_t) and elasticity modulus (E) were formulated in terms of total porosity (n), sonic velocity (V_p), point load index (Is) and Schmidt Hammer values (SH). The results clearly indicated that GP is a potential tool for predicting the elasticity modulus and the strength of granitic rocks.

Rock mass modulus of deformation (E_m) plays a critical role in designing many structures on rock. Ravandi et al. (2013) performed a back analysis calculation to derive an equation for estimation of E_m using GP. The model was developed using a database of 40,960 datasets, including vertical stress (r_z), horizontal to vertical stresses ratio (k), Poisson's ratio (m), radius of circular tunnel (r) and wall displacement of circular tunnel on the horizontal diameter (d). The computer program (CP) generated by GP had a good accuracy with a correlation coefficient equal to 0.97. More recently, Ozbek et al. (2013) proposed models to estimate the UCS of rocks with different characteristics using a GP branch, i.e., GEP. They have considered five different types of rocks including basalt and ignimbrite (black, yellow, gray, brown) were prepared. UCS was formulated in terms of effective porosity (n), water absorption by weight (w_A), and unit weight (γ). It was shown that GP can be used for estimating the UCS of rocks successfully.

The ML-based techniques are increasingly used for interpreting the remote sensing images (RSIs). Conversely from the other ML methods, there are few GP-based studies in the field of remote sensing technology. Some typical examples are estimation of the typhoon rainfall over ocean using multi-variable meteorological satellite data (Chen et al., 2011), monitoring reservoir water quality using remote sensing images (Chen, 2003), mapping of base-metal deposits (Lewkowski et al., 2010), image thresholding for landslide detection (Rosin and Hervas, 2002), and soil moisture distribution analysis (Makkeasorn et al., 2006). As good examples in this context, let us consider the studies done by Makkeasorn et al. (2009) and dos Santos et al. (2010). RSIs are widely used as valuable tools in different real world applications. In the context of agribusiness applications, a major challenge is recognition of crop type regions. To cope with this issue, dos Santos et al. (2010) proposed a new GP-based approach for automatic recognition of coffee crops in RSIs. They combined texture and spectral information encoded by image descriptors. Fig. 1 shows the steps of the proposed classification process. As it is seen, this approach can be divided into two main phases: (1) the image description and (2) image classification. The first phase including the Step 1 to 3 is focused on the image content characterization. The remaining 4 steps belong to the image classification process. GP has been used by dos Santos et al. (2010) to identify relevant partitions by combining the similarities provided by descriptors. Later, dos Santos et al. (2010) proved that their GP-based method yields slightly better results than the traditional MaxVer approach.