International Journal of Mineral Processing 155 (2016) 140–146

Contents lists available at ScienceDirect

International Journal of Mineral Processing

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

Explaining relationships among various coal analyses with coal grindability index by Random Forest

S.S. Matin [a], James C. Hower [b], L. Farahzadi [c], S. Chehreh Chelgani [d,*]

[a] Department of Environment and Energy, Science and Research Branch, Islamic Azad University, Tehran, Iran
[b] Center for Applied Energy Research, University of Kentucky, 2540 Research Park Drive, Lexington, KY 40511, USA
[c] Architecture Faculty, Dr. Shariaty Technical College, Tehran, Iran
[d] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form 20 August 2016
Accepted 31 August 2016
Available online 1 September 2016

Keywords:
Hardgrove Grindability Index
Random Forest
Variable importance
Proximate analysis
Ultimate analysis
Petrography

ABSTRACT

Application of Random Forest (RF) via variable importance measurements (VIMs) and prediction is a new data mining model, not yet wide spread in the applied science and engineering fields. In this study, the VIMs (proximate and ultimate analysis, petrography) processed by RF models were used for the prediction of Hardgrove Grindability Index (HGI) based on a wide range of Kentucky coal samples. VIMs, coupled with Pearson correlation, through various analyses indicated that total sulfur, liptinite, and vitrinite maximum reflectance (R_{max}) are the most importance variables for the prediction of HGI. These effective predictors have been used as inputs for the prediction of HGI by a RF model. Results indicated that the RF model can model HGI quite satisfactorily when the $R^2 = 0.90$ and 99% of predicted HGIs had less than 4 HGI unit error in the testing stage. According to the result, by providing nonlinear VIMs as well as an accurate prediction model, RF can be further employed as a reliable and accurate technique for the evaluation of complex relationships in coal processing investigations.

1. Introduction

The U.S. Energy Information Administration (EIA) estimated that total coal production will be increased by 27 million short-tons in 2017 (EIA, 2016). Increasing demand for high purity coal, as well as growing awareness about environmental pollution is associated with coal consumption. Therefore, developing technologies that make coal cleaner significantly have been considered to ensure it plays a part in our future clean energy. Ultimately, to liberate, and finally remove coal impurities (such as mineral matter), coal particles have to comminuted to fine particles (in size range of several microns) (Sengupta, 2002).

Comminution (crushing and grinding or pulverizing), as an essential step in coal treatment, is often the greatest energy consumer in coal washing plants (Sengupta, 2002). Therefore understanding the behavior of coal through comminution process would be important, and having more information on the subject could be effective for control and optimization of other treatment processes (combustion, gasification, carbonization, etc.) (Bhattacharya et al., 1998; Sengupta, 2002; Vuthaluru et al., 2003; Lee et al., 2003).

Grindability measurement of coal can demonstrate the above mentioned aspects where coal grindability as an essential physical property of coal reflects its relative hardness, tenacity, and fracture (effective parameters on comminution performance). Coal grindability is influenced by coal rank, petrography and mineral matter (Hower et al., 1987; Hower and Wild, 1988; Conroy, 1994; Barton et al., 1994; Bailey and Hodson, 1994; Hower, 1998; Rubiera et al., 1999; Bhattacharya et al., 1998; Sengupta, 2002; Vuthaluru et al., 2003; Trimble and Hower, 2003). Grindability of coal is usually measured by Hardgrove Grindability Index (HGI) (based on the standard test method ASTM D 409-71) (ASTM, 1971; Lee et al., 2003). The result of the HGI test is the most effective parameter in designing a coal mill for power plants. HGI also as a predictive index is used to estimate the performance capacity of industrial pulverizers in power station boilers (lower HGI will require a greater energy input and time to the desired size for pulverized-fuel combustion) (Mackowsky and Abramski, 1943; Peters et al., 1962; Hower and Lineberry, 1988; Hower, 1998; Vuthaluru et al., 2003; Peisheng et al., 2005).

Although the HGI test is not costly (albeit time consuming), due to inherent limitation through the test, HGI determination can be rather difficult. Some of the difficulties which can be considered for a HGI determination are: limitation of the developed methodology; various types of HGI machine and difference in grinding bowl and its material composition; different stages to get the required size; differences in sample preparation; and reliability, repeatability and reproducibility of the test. These difficulties could be due to heterogeneous properties of coal samples such as coal rank, maceral and microlithotype distribution, and mineral matter (Xuexin, 2001; Sengupta, 2002). To overcome these problems many researchers have investigated the prediction of HGI

* Corresponding author.
E-mail address: schehreh@umich.edu (S.C. Chelgani).

http://dx.doi.org/10.1016/j.minpro.2016.08.015
0301-7516/

based on various coal analyses; common coal analyses (proximate and ultimate analysis), petrography, and vitrinite maximum reflectance (R_{max}) by using regression and soft computing methods [artificial neural networks (ANNs), genetic algorithm (GA), Nero-fuzzy (such as ANFIS)] (Hower and Wild, 1988; Peisheng et al., 2005; Jorjani et al., 2008a, 2008b; Chehreh Chelgani et al., 2008; Chehreh Chelgani et al., 2011a, 2011b; Chehreh Chelgani and Makaremi, 2013).

These developed models (regression and soft computing) dependent upon the quality of the input data into the generation of the models, and are used to yield promising descriptive results. The essential point is that, variation in a parameter such as HGI cannot be understood without a thorough knowledge of the fundamental coal properties. In addition, in many cases, a variable would be a relatively strong contributor to HGI, but would have stronger correlation to the other variables that were influencing HGI. Including these variables inflates the correlation (R^2) of the model, but does not necessarily mean that the model describes HGI more accurately. Therefore, before developing complex models, necessary caution has to be used in selecting of variables to study the great inter-dependence between coal properties, and then HGI (Trimble and Hower, 2003; Hower, 2006). Generally regression and soft computing methods are just capable of capturing complex relationships among large numbers of variables to predict a target, but they do not necessarily give any particular insight into the interrelationships among inputs and target variables. This major problem led to the development of so-called variable importance measures (VIMs) which can be used to identify the individual effects of explanatory variables (Auret and Aldrich, 2012).

A recently developed method, Random Forests (RFs), can overcome this drawback by providing attractive addition to nonlinear approximation of statistical relationships among inputs and outputs (Breiman, 2001; Strobl et al., 2008; Archer and Kimes, 2008; Hallett et al., 2014). RF as an ensemble of multiple decision trees is a type of a high-dimensional, non-parametric predictive model consisting of a collection of classification or regression trees (Breiman, 2001). RFs also have been successfully applied to various prediction models within the last decades and through this short period of time they have become a major data analysis tool which performs well in comparison with many standard methods (Díaz-Uriarte and Alvarez de Andrés, 2006; Heidema et al., 2006). RF models have several advantages over other statistical modeling techniques: they are able to deal with missing values and high-dimensional data, identify complex interactions between variables and the most important variables measurements (VIMs), predict with high accuracy (low-bias models and low-variation in results), and they are robust against over-fitting (Hopwood et al., 1994; Díaz-Uriarte and Alvarez de Andrés, 2006; Biau et al., 2008; Archer and Kimes, 2008). Although there is a widespread usage of RF models in various fields (RF method should be considered by well-informed experts in the field) (Auret and Aldrich, 2012; Biau et al., 2008; Archer and Kimes, 2008; Hallett et al., 2014; Chehreh Chelgani et al., 2016a), to our knowledge there are rarely used to explore interrelationship among coal properties or for predictions (Matin and Chehreh Chelgani, 2016; Chehreh Chelgani et al., 2016b).

The aim of the present investigation is to assess the properties of over 900 coal samples from Kentucky, USA, in order to estimate the HGI with the most important parameters based on ultimate and proximate analysis, oxides, and petrographic analysis of samples by using RF. To our best knowledge, no tree or RF based methods have been proposed for the estimation of coal grindability.

2. Materials and methods

2.1. Experimental data

A soft computing model for the HGI prediction requires a robust database to cover a wide variety of coal types. Such a model will be capable for predicting HGI with a high degree of accuracy. Data used to test the proposed approaches are from studies conducted at the University of Kentucky Center for Applied Energy Research. Samples were prepared from Western and Eastern Kentucky Southwest, Hazard and Big Sandy coals. A total of more than 900 sets of data were used. The results of various analyses (input variables for HGI prediction) and their representative HGIs are shown in the supplementary database. Analyses were performed according to the standard ASTM test methods (ASTMD 409-71: Hardgrove, ASTM D3172: Proximate, and ASTM D3176: Ultimate analyses). For petrology, all samples were previously prepared as particulate pellets.

2.2. Random Forest

2.2.1. Variable importance measurements (VIMs)

RF methods aside from accurate prediction have another extremely useful output which is variable importance measures (VIMs) (Breiman, 2001; Svetnik et al., 2003; Liaw and Wiener, 2002; Bylander, 2002). VIMs for RFs have been receiving increased attention as a means of variable selection in many non-parametric regression tasks (Wang et al., 2016). VIMs provide insight into the interactions between predictors and by a group of tree computed relationships between a target and predictors to indicate which variables have the significant effect on the target (Hallett et al., 2014). The most popular and advanced VIM available in RFs is the permutation accuracy importance (PAI) measure (Strobl et al., 2007; Hapfelmeier et al., 2014). For variable selection purposes, the main advantage of the PAI in RF as compared to other tree-based methods is that it covers the impact of each predictor variable individually as well as in multivariate interactions with other predictor variables (Strobl et al., 2007). PAI is broad applicable and unbiased through the consideration of multivariate interactions among variables (Breiman, 2001; Strobl et al., 2007).

In PAI for VIMs, "out of bag" (OOB: computations based on observations that were not part of the sample used for constructing the respective tree) dataset accuracy is always applied to evaluate the performance. OOB achieves higher accuracy with low bias and variance than other tree structured algorithms (Kulkarni and Sinha, 2013). The OOB data can be permuted, without required to train new forests (Breiman and Cutler, 2003; Archer and Kimes, 2008). In summary, the computation of the PAI consists of the following steps:

1) Calculating the mean square error (MSE) of a decision tree,
2) Permuting the values of explanatory variable in the OOB observations,
3) Recalculating the OOB MSE of that decision tree,
4) Calculating the difference between the MSE values which were calculated in step 1 and 3, and
5) Repeating the above steps for each decision tree and use the average difference over all trees as the overall importance score (Strobl et al., 2008; Hapfelmeier et al., 2014; Wang et al., 2016).

The reference implementation of PAI is available in the "R" software package for statistical computing which has been used in this study (https://www.r-project.org/). VIM is determined based on the "IncNodePurity". The IncNodePurity parameter of the RF is average overall nodes in all trees in the forest.

2.2.2. Prediction by RF

As mentioned, RFs are broadly used in many investigations for prediction of complex models (complicated relationships). Through prediction by RF, the model combines a number of trees by taking the same number of bootstrap samples (random samples of the original data with replacement and with the same length) from the database, and building a tree based on each bootstrap sample (Hallett et al., 2014). The procedure of taking a bootstrap sample from the original training data to establish the training dataset for each tree is called bagging of decision trees (Archer and Kimes, 2008; Wang et al., 2016). For prediction, an estimated label is provided by the average over all trees