Fuel 182 (2016) 754–760

Contents lists available at ScienceDirect

Fuel

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

Full Length Article

Explaining relationships between coke quality index and coal properties by Random Forest method

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

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

HIGHLIGHTS

- The prediction of Free Swelling Index (FSI) as a coke quality parameter were studied.
- The importance of variables for FSI prediction was measured by Random forest (RF).
- RF method can build accurate model to assess complex relationships in coal processing.
- Results indicated that RF models can be applied for prediction of other fuel factors.

ARTICLE INFO

Article history:
Received 3 May 2016
Received in revised form 6 June 2016
Accepted 7 June 2016

Keywords:
Coke quality
Swelling index
Coal rank
Random forest
Variable importance

ABSTRACT

In this study was shown that random forest (RF) can be used as a sensible new data mining tool for variable importance measurements (VIMs) through various coal properties for prediction of coke quality (Free Swelling Index (FSI)). The VIMs of RF within coal analyses (proximate, ultimate, and petrographic analyses) were applied for the selection of the best predictors of FSI over a wide range of Kentucky coal samples. VIMs assisted by Pearson correlation through proximate, ultimate, and petrographic analyses indicated that volatile matter, carbon, vitrinite, and R_{max} (coal rank parameters) are the most effective variables for the prediction of FSI. These important predictors have been used as inputs of RF model for the FSI prediction. Outputs in the testing stage of the model indicated that RF can predict FSI quite satisfactorily; the R^2 was 0.93 and mean square error from actual FSIs was 0.15 (had less than interval unit of FSI; 0.5). According to the result, by providing nonlinear inter-dependence approximation among parameters for variable selection and also non-parametric predictive model RF can potentially be further employed as a reliable and accurate technique for the determination of complex relationship through fuel and energy investigations.

1. Introduction

Demands for fossil fuels, which are the main parts of the world energy consumption, are increasing [1]. Coal as a fossil fuel has been used for the thermal power generation; iron making; and in the cement, paper, and textile industries [2,3]. Metallurgical coke is commercially produced by carbonization of coals at temperature up to 1400 K. Through the coking procedure in an oven, the coal sample softens, fuses, and resolidifies to form a macro-porous carbon material [3–6].

Coke is an important fuel in steel making, and coke-making is an important use for coal. Coke is an expensive component, and although some other fuels (oil, granulated, and pulverized coal [7], and plastic wastes [8–11]) have been proposed as substitutes for coke, realistically there is no available replacement that can substitute for the metallurgical performance of coke to support the blast furnace charges [6,12–14].

In steel industry, coke has been used to provide heat for the melting of slag and metal (as a fuel), reduce iron ore to elemental iron (as a reduction agent), and maintain permeability in the blast furnace (as a permeable support) [6,13,15]. Quantitatively coke is the largest materials feed into the blast furnace. Therefore, as a permeable support, consistent quality is needed to economically drive the furnace and provide reliable steel quality [6,13,14,

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

http://dx.doi.org/10.1016/j.fuel.2016.06.034
0016-2361/

16,17]. There are many parameters which affect coke performance in the blast furnace. The coal rank parameters (Moisture, volatile matter, carbon, etc.), petrographic composition (macerals and vitrinite maximum reflectance (R_{max})), and impurities (ash, sulfur, phosphorous, and alkali contents) together control the quality of coals and fundamentally effects on coking coal quality [3,6,13,17–21]. The evaluation of coking quality from parent coal samples can be assisted by fluidity, dilatation, or Free Swelling Index (FSI) methods [12,13,17,22,23].

In the USA, a coal's cokeability is evaluated by FSI [1] among other test. According to ASTM D 720, FSI measurement involves heating a small sample of coal (1 g of fresh powdered coal sample (<250 μm)) for 2.5 min in a standard sized silica crucible to around 1100 K [24]. After cooling, FSI is evaluated by comparing the size and shape of the resulting solid "cross sectional profile of the coke button" with a series of standards and assigning a value from zero to nine at intervals of 0.5. Based on the result, standard coke quality (FSI) is classified into weakly (0–2), medium (2–4), and strongly (4–9) caking ranges [1,12,16,21,24,26–29]. There are some parameters that potentially could contribute to errors in FSI determination such as the proper heating rate, oxidation or weathering of the coal sample, and an excess of fine coal in the analysis sample [21,25,27,29]. Therefore, prediction of FSI from coal properties could provide this opportunity to assess coke quality by easy and accurate estimation and save all efforts involving the experimental procedure.

Based on these facts, a few statistical models based on various coal analyses (proximate, ultimate, and petrographic analyses) have been developed to predict FSI by regression and artificial neural networks (ANNs) [21,25]. Generally these methods are only capable of capturing complex relationships among variables to predict an output, and do not give any insight into the interdependences among inputs and output variables. In addition, due to heterogeneous character of coals, a variable could be a relatively strong contributor to FSI, but would be more strongly correlated to the other variables which were influencing coke quality. Including these variables as inputs in a model inflates the correlation coefficient (R^2) of the model, but does not necessarily mean that the model more accurately describes FSI [30–32]. Hence, it would be an essential to use a method which can identify the individual effects of explanatory variables.

As an ensemble of multiple decision trees, the random forest (RF) method can overcome these drawbacks. RF is a predictive model based on decision by a collection of classification or regression trees. In the RF approach to variable importance measurements (VIMs) and predict a target value, a group of trees is used to establish relevant predictor variables with the target, and based on importance predictors estimate the target. Within the past few years, RF has been widely used in many investigations for prediction and interpretation purposes. Their popularity is based on their ability to deal with missing values and high-dimensional data, identify complex interactions between variables and the most important predictors measurements (VIMs), high predictive accuracy (low-bias models and low-variation in results), and robust against over-fitting [33–38].

Despite the growing literature on using tree-based methods in various fields, to our knowledge there is no investigation has been done to examine interrelationship (VIMs) between coal properties with FSI by RF and predict the coke quality (FSI) with those variables. The aim of the present study is the assessment of properties of over 900 coal samples from Kentucky, USA, to identify the variable importance and predict the coke quality with the most important variables (VIMs) based on various coal analyses (ultimate and proximate analysis, oxides, and petrographic analysis) of samples by using RF method.

2. Materials and methods

2.1. Experimental data

Valuable modeling by soft computing methods for the estimation of a target, and also for the examination of interrelationships between inputs and outputs, requires a high-dimensional robust database to cover a wide variety of sample properties. Such a model will be valuable for predicting of a target with a high degree of accuracy. In this study, data used to explain relationship between properties of coal samples with their coke qualities comes from studies conducted at the University of Kentucky, Center for Applied Energy Research. A total of >900 sets of data were used. The results of various analyses (input variables for the prediction) and their representative FSIs are shown in the Supplementary database. Analyses were performed according to the standard ASTM test methods.

2.2. Random forest

2.2.1. Variable importance measurements (VIMs)

In the learning step of models in which there are many variables in a high-dimensional data set, all of the variables are not important for prediction of the target. Irrelevant variables potentially can make bias on the model accuracy. By a variable selection technique, irrelevant variables could be eliminated, the prediction accuracy would be improved and the risk of overfitting might be decreased. Therefore, variable selections that can determine the most important variables have been receiving increased attention. For random forests, "Permutation accuracy importance (PAI)" is the most advanced variable measurement (VIM) [33,39–42].

The PAI is assessed by comparing the difference between prediction accuracy from a tree, before and after random permutation of the predictor variable (a non-linear assessment). In other words, PAI is comparing the prediction accuracy of a tree with and without the presence of this predictor variable. The average of differences through all trees indicates the final importance rank. Consequently, a strong relationship between the predictor variable and the output demonstrates with a large value of the PAI, and other values ranked after that. The advantage of PAI in RF by means of VIMs over other available methods is that PAI covers the impacts of each variable individually and simultaneously assesses the multivariate inter-correlations between other predictor variables [33,37,39,43,44]. For modeling and statistical analyses, the reference implementation of PAI is available in the "R" software package (a free software package for statistical computing) which has been used in this study.

2.2.2. Prediction by RF

Fundamentally RF models as ensemble methods are relied on the CART procedure (classification and regression trees), where the feature space is disjoint into separate nodes, and then by a simple model (like fitting constant to each region) the output estimated in each region. The splitting algorithm is hierarchical and designed in a binary fashion. Therefore in each step it determines the optimal "s" (splitting variable) and the best split-point along that variable. In each node of a tree, the best splitting variable is determined over all input data and all possible "s" by calculating the following minimization of the residual sum of squares (RSS):

$$RSS = \sum_{x_i \in R1_{ls}} (y_i - \bar{y}1_{ls})^2 + \sum_{x_i \in R2_{ls}} (y_i - \bar{y}2_{ls})^2 \quad (1)$$

$$argmin_{l \in \{1,2,\ldots,p\},s}(RSS) \quad (2)$$