Geographic origin identification of coal using near-infrared spectroscopy combined with improved random forest method

Meng Lei, Xinhui Yu, Ming Li, Wenxiang Zhu

PII: S1350-4495(18)30182-8
DOI: https://doi.org/10.1016/j.infrared.2018.05.018
Reference: INFPHY 2569

To appear in: *Infrared Physics & Technology*

Received Date: 20 March 2018
Revised Date: 16 May 2018
Accepted Date: 21 May 2018

Please cite this article as: M. Lei, X. Yu, M. Li, W. Zhu, Geographic origin identification of coal using near-infrared spectroscopy combined with improved random forest method, *Infrared Physics & Technology* (2018), doi: https://doi.org/10.1016/j.infrared.2018.05.018

Geographic origin identification of coal using near-infrared spectroscopy combined with improved random forest method

Meng Lei[1,2], Xinhui Yu[1], Ming Li[1*], Wenxiang Zhu[1]

1. School of Information and Electrical Control Engineering, China University of Mining and Technology, Xuzhou, China
2. Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

HIGHLITGHTS

- The research provides a rapid identification method of geographic origin of coal by near infrared spectroscopy
- Establishing the quantitative analytical model for origin identification based on random forest algorithm
- Improving the random forest model using synthetic minority oversampling technique to overcome imbalanced data set

ARTICLE INFO

Article history:
Received ***
Revised ***
Accepted **
Available online ***

Keywords:
Geographic origin identification of coal
Near-infrared spectroscopy
Random forest
Synthetic minority oversampling technique

ABSTRACT

Traditional identification methods of coal origin have the drawbacks of complex operation, samples damage and environmental pollution. Near infrared spectroscopy is a new method which is used to solve the problems effectively. However, the coal samples spectra had the features of high dimension, redundancy and noise. Also the data set was small and imbalanced. Therefore, this study chose Random Forest(RF) algorithm as the basic modeling algorithm. Besides, the K-means algorithm was introduced to improve the Synthetic Minority Oversampling Technique (SMOTE) to overcome imbalanced data set. A comparison of the Support Vector Machine(SVM) model, the RF model and the improved RF model indicated that the improved RF model reached an overall accuracy of 97.92%, a G-mean value of 0.9696, and an average voting rate of 83.09%. These results were 6.25%, 7.03%, 6.94% higher than the counterparts of RF model respectively. Simultaneously, they were 8.34% and 5.86% higher than SVM model in accuracy and G-mean. The results suggested that the improved RF model produced reliable accuracy, validity and stability. Its results were conformed to the analysis of the coal-forming factors. Consequently, the algorithm is applicable to identify the geographic origin of coal rapidly.

1. Introduction

With the progress of globalization and the reform of industrial structure, both the import and export of coal are on the rise in China, which brings the identification of geographic coal origin into the limelight [1-2]. The traditional classification methods involve a series of instruments for varied measurements, such as ash content, volatile content, heat expansion coefficient, etc [3-4]. Characterized by their complex procedures, these methods are less automated and slow. Therefore they fail to meet the needs of rapid classification of coal.

The qualitative analysis based on the conjunction of Near Infrared Spectroscopy (NIRS) and pattern recognition offers a new method of coal classification [5-6]. The NIRS has the advantages of fast and lossless analysis, simple sample processing, and multi-component detection, which is superior to traditional analysis methods [7-9]. Li et al [10], utilized NIRS to quickly identify the origin of coal samples. The accuracy of identification model based on Learning Vector Quantization (LVQ) algorithm was 80%, and the model based on Support Vector Machine (SVM) could reach 90%. Wang et al [11], combined SVM and NIRS to classify four types of coal: rich coal, lean coal, coking coal and meager-lean coal, which produced a desirable classification result (93.5% in accuracy).

In this paper, a Random Forest (RF) model is established for coal classification using the acquired spectra of coal samples from five countries: Australia, Russia, Canada, Indonesia and China. Considering the weakness of imbnlanced sample in RF model, we proposed an improved SMOTE-RF model based on Synthetic Minority Oversampling Technique (SMOTE) algorithm. Finally, compared with the SVM model, the SMOTE-RF model can identify the coal origin accurately and effectively.

2. Theoretical Basis

2.1. Decision Tree

* Corresponding author. Tel.:+86 18361226726
E-mail address:limingsiee@126.com