Identification Method of Iron-Ore Type in Open Pit Based on Measured Spectra and Random Forest

Hongqi Xiong[1], Pengfei Pan[1], Hui Luan[1], Shanjun Liu[2], Yachun Mao[2], Yue wang[2]
[1]Anqian Mining Limited Liability Company of Anshan Iron and Steel Group Corporation, Anshan, China
[2] College of Resources and Civil Engineering, Northeastern University, Shenyang, China
Email: liusjdr@126.com

KEYWORDS: Spectroradiometer, Iron ore, Random Forest method, Classification

ABSTRACT: Iron ore is the main mineral resources to support China's national economy, and open-pit mining is the main way. Chemical analysis has its disadvantages in terms of the determination of iron-ore type, such as slow speed and high cost. Therefore this study has put forward a method to determine iron-ore type through field spectrum analysis. First various types of iron-ore samples were collected in Anqian mine of Liaoning Province，China, including 48 magnetite and 84 hematite samples. Then, through the hole drilling and cutting, the circular sheets of 1~2 cm thick are formed. After that the chemical composition, visible/near-infrared spectra of these samples were tested and analyzed. The principal component analysis is used to reduce the data dimension of spectrum, and the random forest algorithm is applied to classify the iron ore type of samples, while the accuracy of classification results is validated. The results show that the random forest method can classify different types of iron ore, and the successful rate is over 95%, which lay the foundation for the determination of the iron- ore type in open pit based on the portable spectroradiometer or the airborne imaging spectrometer.

1. INTRODUCTION

Iron ore is the main mineral resources to support China's national economy, while the magnetite and hematite are the main types of iron ore, accounting for 73.5% of the total amount. Open pit mining is the main mining way of iron ore. At present, the determination of the boundary between the ore and surrounding rock is mainly dependent on the field sampling and laboratory test, which is slow, high cost, and difficult to meet the requirements for fast, efficient, accurate, real-time delineation of ore type in open pit mines.

In recent years, the visible/near infrared spectroscopy test technology has been greatly developed and its application is increasingly widely, with the boom in machine learning. In this study, random forest of machine learning is applied to sample classification, based on visible/near infrared tests for different types of ore samples in open pit mine. In addition, the classification effect is validated.

2. SAMPLE COLLECTION AND TEST

2.1 Survey of research area

This study area was chosen at the open pit of Anqian Mining Limited Liability Company of Liaoning province, China. By the end of 2011, The company's iron ore reserves was 11.8 tons, as well as 2.1 tons of open pit mine. The major ore species are hematite and magnetite， and the surrounding rock mainly includes some metamorphic rocks, such as chlorite schist, mica schist, phyllite and mixed granite, etc.

2.2 Sample collection and chemical test

A total of 132 samples were collected in the study area, including 84 hematite and 48 magnetite samples. The sample was drilled and cut, and finally circular sheets of 1~2 cm thick was formed (figure 1). After that the chemical composition of the samples was tested.

Figure 1. cutting iron ore sample image

2.3 Visible and near infrared spectra measurement

This study adopted HR - 1024 portable object spectrometer made from the Spectra of SVC Vista company, whose wave band is 0.35~2.5μm and spectral resolution is less than or equal to 8.5 nm. When measuring, the surface of the samples was kept level and the spectrometer lens was perpendicular to the observation surface. Each sample received repeated observation three times, and the average spectral reflection were adopted.
Because of the large amount of samples, the spectral curves of all samples are not reflected in a diagram. Some typical spectra of magnetite and hematite samples are showed in Figure 2.

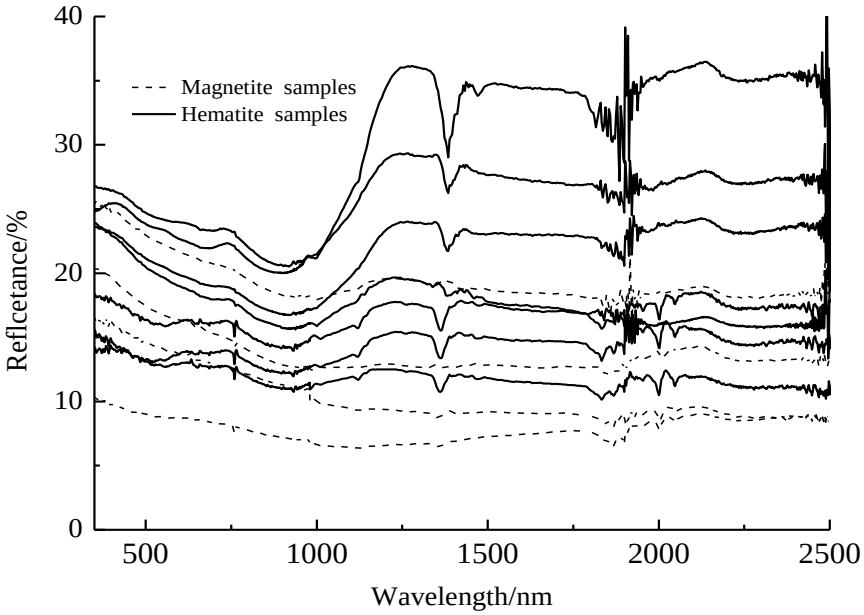

Figure 2. Part of the visible near infrared spectral curves of magnetite and hematite

From Figure 2, it can be seen that the spectral reflectance of hematite samples takes on a decreased trend in 343nm~978nm band, where it appears to be a weak trough in the vicinity of 500nm. The curves present on a prominent upward trend in the 978nm~1255nm wavelength range. After 1255nm, it shows a gentle trend and a weak trough at 2200nm. Two distinct troughs in 1400nm and 1900nm are respectively caused by the absorption of hydroxyl and atmospheric water vapor.

The spectrum characteristics of magnetite samples are: the spectral reflectance in the 343nm~980nm band decreases obviously, and the curves among the 980nm~1750nm band remain unchanged. Similarly, two distinct troughs in 1400nm and 1900nm are respectively related to the absorption of hydroxyl and atmospheric water vapor.

The spectral characteristics of hematite and magnetite samples are markedly different, mainly including the following two points:

(1) In the vicinity of 978nm, spectral curves of hematite samples appear an obvious wave valley, but magnetite samples do not.

(2) In the range of 978nm~1255nm band, spectral curves of hematite sample significantly increase, but those of magnetite samples remain unchanged.

3. DATA ANALYSIS

3.1 Data preprocessing

As is known to all, the principal component analysis[1-2] is one of the important methods in the dimension reduction of hyperspectral data. The principal component analysis is a multivariate statistical method, which integrates the multiple indexes into few comprehensive indexes, and transforms the original relevant variables into another set of unrelated variables by the linear combination. The total variance of variables remains unchanged in the transformation process, and the first variable has the biggest variation. So it is the first principal component, and the

send variable comes second called the second principal components, not associated with the first variable, and so forth.

Assuming that a *n*-dimensional vector include multiple indexes-x_1 $x_1, x_2, ..., x_n$, and they turn into a *m*-dimensional principal component vector- $y_1, y_2, ..., y_m$,and moreover, *m* is less than *n*, they will constitute the following formula:

$$\begin{cases} y_1 = l_{11}x_1 + l_{12}x_2 + ... + l_{1n}x_n \\ y_2 = l_{21}x_1 + l_{22}x_2 + ... + l_{2n}x_n \\ \quad . \\ \quad . \\ \quad . \\ y_m = l_{m1}x_1 + l_{m2}x_2 + ... + l_{mn}x_n \end{cases} \tag{1}$$

Coefficient $l_{ij}\,(i = 1,2,...,m;\ j = 1,2,...,n)$ in the formula above can be obtained from the score matrix through the analysis of the principal components.

In this method, new variables are selected to replace the original numerous variables for resolving the hyperspectral analytical problem – too many bands and spectral band overlapping, without missing the main spectral information. In this study, the principal component analysis is applied to the original data for 973 dimensions reducing to 5 dimensions, and the spectral information is reserved more than 99.5%. The cumulative contribution rates of the first 5 principal components are shown in Table 1.

Table 1 Variance contribution rates of spectral data based on principal component analysis

Principal component	PC1	PC2	PC3	PC4	PC5
Variance contribution rate (%)	88.99	8.82	0.44	0.25	0.12
Cumulative contribution rate (%)	88.99	98.81	99.25	99.50	99.62

3.2 Random forest algorithm

Random forest is a statistical learning algorithm, proposed by United States scientist Leo Breiman, who combines the Bagging integrated learning theory and the random subspace method to form Random forest algorithm[3]. Random forest is a method which selects multiple samples from the original ones with the approach of bootsrap resampling and establishes the decision tree model for each extracted bootsrap samples, and then combines with the prediction of multiple decision tree, while the final classification result is determined by vote from the output results of the single decision tree [4].

The random forest algorithm overcomes the defect of overfitting of the decision tree, which also has a good tolerance to noise and outliers, and has good parallelism and scalability for high dimensional data classification. Due to the characteristics of random forest, it has been widely used in the fields of medicine, management, bioinformatics and etc [5].

The classification accuracy of the random forest algorithm depends on the user-defined parameters *N* (the quantity of trees) and *M* (the quantity of randomly selected feature variables). In this paper, 200 and 3 are chosen respectively for *N* and *M*. The algorithm is implemented by R2009 Matlab (a) programming.

3.3 Classification and validation

In this paper, we choose ten fold cross validation method to verify the accuracy of the classification. The 132 samples are randomly divided into 10 sub samples, while one separate sub sample is regarded as the validation sample and the remaining 9 sub samples are chosen for the training samples. Cross validation is repeated 10 times, so that each sub sample will be verified, and each test will output the corresponding correct rate. The mean of the

10 classification results is taken as the final classification accuracy and the estimate of the algorithm accuracy [6-7]. Ten fold cross validation results show that the accuracy of the random forest classification method can reach 96.21%, and the method can distinguish different types of iron ore, adequately. The classification results are shown in Table 2.

Table 2 Classification results of fold cross validation

Inspection times	Training sample size	Testing sample size	Classification accuracy	Inspection times	Training sample size	Testing sample size	Classification accuracy
1st	119	13	100%	2nd	119	13	100%
3rd	119	13	92.3%	4th	119	13	100%
5th	119	13	100%	6th	119	13	92.3%
7th	119	13	92.3%	8th	119	13	100%
9th	118	14	85.7%	10th	118	14	100%

4. CONCLUSIONS

This study exploits random forest to classify and identify the iron ore type based on the characteristics of visible and near infrared spectra, and the main conclusions are as following:

(1) The spectral characteristics of magnetite and hematite samples have obvious differences.

(2) The random forest method is applied to the classification for different types of iron ore, whose classification accuracy rate is up to over 95%, and it is a fast and effective method for the source survey.

(3) The method proposed by this paper shows better applicability in the detection of ore type based on the portable spectroradiometer or the airborne imaging spectrometer.

REFERENCE:

[1] Liu Xue-hua, Sun Yan, Wu Yan, *et al.* 2012,Reduction of Hyperspectral Dimensions and Construction of Discriminating Models for Identifying Wetland Plant Species[J]. Spectroscopy and Spectral Analysis. 32(2),pp. 459-464.
[2] Liu Zhan-Yu, Wang Da-Cheng, Li Bo, *et al*. 2009, Discrimination of Lodged Rice Based on Visible/Near Infrared Spectroscopy. 28(5), pp. 342-345.
[3] Leo Breiman. 2001, Random Forests[J]. Machine Learning, 45(1), pp. 5–32.
[4] Daniel R G, Wade A W, Andrew J F, *et al*. 2016, Predicting USCS Soil Classification from Soil Property Variables Using Random Forest[J]. Journal of Terra mechanics, 65, pp. 85-92.
[5] Chen X, Liu M. 2005, Prediction of Protein—protein Interactions Using Random Decision Rorest Framework [J]. Bioinformatics, 21(24), pp . 4394-4400.
[6] Li Pei-xian，Tan Zhi-xiang，Yan Li-li, *et al*. 2010 , Calculation Method of Probability Integration Method Parameters Based on Support Vector Machine[J]. Journal of China Coal Society, 35(8), pp .1247-1251.
[7] Yuan Jie, Li Xiao-jie, Chen Chao, *et al.* 2014, Study on Prediction of Compound-target-disease Network of Chuanxiong Rhizoma Based on Random Forest Algorithm[J]. China Journal of Chinese Material Medica, 39(12),pp. 2336-2340.