MODELLING METHOD OF PREDICTION MODEL FOR SALT FIELD ION CONCENTRATION UNDER SOLAR THERMAL SYSTEM USING RANDOM FOREST

by

Jun LIU[a, b], Supei ZHANG[b*], Zihan XU[b], Siqi SUN[b], Aowen XIAO[b], and Zhuang DU[b]

[a] Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan, China
[b] School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China

Original scientific paper
https://doi.org/10.2298/TSCI181128151L

At present, the underground brine deposit of Lop Nor salt lake in Xinjiang, while is rich in solar energy resource and can be one high efficient solar thermal utilization area, has become an important potash production base in China. During the development of salt lake brine, the collection methods are different according to the concentration of ions in different locations. In order to solve the problems of low accuracy and high calculation cost in prediction of salt field ion concentration, a data mining method based on random forest is applied in this paper. To build the model, we collected K^+, SO_4^{2-}, Cl^-, and other two kinds of ions, among which the features included the collection time, collection locality and the number of salt pond. We used several methods to train and test the sample data, evaluated the experimental results using a variety of performance metrics and compared it with other methods at the same time. The results revealed that the optimal random forest model yielded the mean square error and coefficient of determination values of 0.073 and 0.940, which performed relatively better than support vector machine and extremely randomized trees.

Key words*: ion concentration, random forest, prediction model, data mining*

Introduction

Potash is an important material food security concern for China for it is of great significance of improving soil environment and promoting agricultural production in China. While the deficiency of potassium in cultivated soil and mineral resources in our country is so serious that it has become one of vital limiting factors in the crop production. China has a large consumption of potassium and a serious shortage, 70% of which depends on foreign imports [1]. Where supply is dependent on foreign supply because of a lack of potassium in China. It is strategically significant for China to increase the potassic fertilizer production and to develop the potassic fertilizer industry. In 2002, the State Development and Investment Corporation began to develop potash in Lop Nor [2]. As one of the largest dry salt lakes in the world, Lop Nor Salt Lake has become an important potash production base in China, for it is rich in solar energy resource and can be one high efficient solar thermal utilization area. It also has formed a variety of industrial models from single potash development to salt lake chemical industry, energy and chemical industry [3].

* Corresponding author, e-mail: zhangsupei@wit.edu.cn

The salt field mainly uses the solar thermal energy to evaporate brine to obtain the ion resources, including K^+, SO_4^{2-}, Cl^-, and the like. Sodium chloride can be used to make alkali chemical products, potassium-containing carnallite is used to produce potash (such as potassium sulphate and potassium chloride), and bischofite is the main raw material of magnesium industry [4-9]. So the modelling method of prediction model for the concentration of ions at a certain location in the salt field can effectively improve the collection efficiency, cost reduction and energy saving can be achieved.

At present, many different methods have been proposed in the field of ion concentration prediction. Xu *et al.* [10] predicted chloride concentration in concrete based on radial basis function network, which improved the accuracy and stability compared with the traditional back-propagation (BP) network. Chen *et al.* [11] used the BP algorithm to optimize the scale and translation parameters of the Morlet wavelet function, the weight coefficients, threshold values in WNN structure. Parveen *et al.* [12] applied Grey-Markov process prediction method to analyze the trend of several major ionic concentrations in Jilantai salt lake brine. Based on the combination of Grey system and Markov process, the prediction accuracy of data with large random fluctuation was improved. Xiong *et al.* [13] developed a SVR-based model to predict the sorption capacity of Cr (VI), compared with multiple linear regression and ANN, the SVR model is more accurate than other two models. Suykens and Vandewalle *et al.* [14] built a least squares version of least squares support vector machine. Compared with the least-squares SVM, it speeded up the calculations and provided better results. We used the SVR model to predict the content of camelina protein using FT-IR spectroscopy [15]. We applied the SVR model to get the accurate prediction of potassium ion concentration in salt pools for the actual production of potash fertilizer, too [16].

In order to meet the requirements of fast computing speed, high accuracy, and low computational cost, random forest regression is applied to the ion concentration prediction in salt field. The method is evaluated by a variety of methods to illustrate the effectiveness of non-linear prediction for small sample and noisy data. The results show that compared with support vector machine and other regression models`, the proposed method improves the accuracy and stability.

The principle of random forest

Random forest was originally proposed by Breiman [17] and is a kind of ensemble learning method. The implementation of the random forest model is simple and can be parallelized in the training process, which facilitates the calculation and simulation of a large amount of data in a relatively short time. It has been widely used in medical diagnosis, financial market and other fields [18-20].

The main idea of integrated learning is to build and combine multiple base learners to achieve better generalization capabilities. The random forest uses the classification and regression tree (CART) as the base learner, which is one of algorithms of decision tree. Structurally, decision tree is a tree structure that is recursively generated from top to bottom. The decision node represents the full set of samples, and the leaf nodes represent the decision results. It aims to establish a binary or multi-tree with the fastest decline in entropy as measured by information entropy, which represents a measure of uncertainty in a set and is a description of the degree of uncertainty [21, 22].

The CART uses the Gini index to classify attributes [23]. The Gini index, which represents the probability of randomly extracting two samples from the dataset, reflects the purity

of the dataset [24]. The smaller the Gini index, the higher the purity of the sample. In the classification problem, the mathematical formula of the Gini index:

$$\mathrm{Gini}(D)=\sum_{k=1}^{K}p_k\left(1-p_k\right)=1-\sum_{k=1}^{K}p_k^2 \tag{1}$$

where D is the data set, K – the number of sample types, and p_k – the proportion of the k^{th} sample. If the data set is divided into two parts D_1 and D_2 based on the value of the feature A. Then under the condition of the feature A, the Gini index of D can be computed:

$$\mathrm{Gini}(D,A)=\frac{|D_1|}{|D|}\mathrm{Gini}(D_1)+\frac{|D_2|}{|D|}\mathrm{Gini}(D_2) \tag{2}$$

Gini (D, A) indicates the uncertainty of D after $A = a$ division. The larger the Gini (D, A) , the higher the sample uncertainty. Generally, the attribute with the smallest Gini index after the division of the candidate set attribute A is selected as the optimal division attribute.

The basic step of random forest is to first perform m times of bootstrap sampling from the dataset to obtain datasets of m samples, and train the decision tree on each dataset. The traditional decision tree selects the optimal attribute in the current node when dividing nodes. When dividing nodes for each decision tree, random forest randomly selects some attributes of the current node, and then selects the optimal attribute from the partial attributes. For example, there are currently attributes, and the number of randomly selected attributes is generally $k = \log 2d$.

The predicted output of each decision tree is $T_i(x)$, where $i = 1, 2,...m$. For regression problems, the final predicted output:

$$Pre(x)=\frac{1}{m}\sum_{i=1}^{m}T_i(x) \tag{3}$$

Since bootstrap sampling is random sampling without replacement, some data are still unsampled after m times of sampling. The probability that the sample is never sampled:

$$P=\left(1+\frac{1}{m}\right)^m \tag{4}$$

The calculation shows that as m approaches infinity, p is approximately equal to 0.368. Therefore, about 36.8% of the data in the data set is not trained, and this part of data is called out of bag estimation [25], which can be used to detect the generalization ability of the model.

The principle of random forest

The data set

The experimental data in this paper were collected from a certain region of Xinjiang, including the ion measurement results of 12 salt fields, such as K^+, Na^+, SO_4^{2-}, Cl^-, Mg^{2+}, *etc.* The data dimensions include the collection time, the abscissa and ordinate of the collection locality, the number of salt field and the ion concentration of the collection locality. The acquisition time is mapped to 1-365. The experimental data are shown in tab. 1.

Experiment method

Experiment results were evaluated using the hold-out method. Firstly, the experimental data are divided into a training set and a test set, and random Numbers are generated by the system to ensure the random distribution of data. Due to the large difference in the numer-

Table 1. Sample of data

Time	Abscissa axis	Ordinate axis	Number	Na^+	K^+	Mg^{2+}	SO_4^{2-}	Cl^-
3	900	300	2	3.0	4.2	8.13	14.67	21.33
4	1100	400	2	1.99	4.62	8.34	13.85	21.36
4	1100	500	2	1.99	4.62	8.34	13.85	21.36
5	1200	400	2	2.28	4.55	8.34	14.85	21.01
5	1200	200	2	2.09	5.52	8.3	13.19	22.69
7	1263	800	2	2.87	5.82	8.04	10.7	25.24
9	1100	600	2	2.64	6.19	8.37	13.22	24.36
12	1350	700	2	6.83	4.61	7.28	10.83	27.95
13	1400	370	2	8.55	4.4	6.27	11.36	27.07
14	1400	900	2	6.73	4.98	7.21	9.96	28.59

ical range of each data dimension, for example, the abscissa range of the collection locality is [0.3000], and the ion concentration range is [0.10]. To achieve data standardization, the data is mapped to a space with a mean of 0 and a standard deviation of 1 by standard deviation. Standardization can make the characteristics of different measurements comparable without changing the distribution of the original data [26]. The transformation function:

$$x^* = \frac{x - \mu}{\sigma} \tag{5}$$

where μ represents the mean of the data, and σ represents the standard deviation of the data.

After preprocessing the data, the algorithm model is constructed. All the data in the training set are put into the model for training. After the training, independent variables in the test set are input into the model. The model learner calculates the predicted values of dependent variables and compares the predicted values with the true values to measure the actual performance of different algorithm models [27-36].

Evaluation method

In order to evaluate the generalization ability of the model, the mean squared error (MSE) and the coefficient of determination, R^2, were selected as indicators.

The formula for mean squared error:

$$E(f;D) = \frac{1}{m}\sum_{i=1}^{m}\left[f(x_i) - y_i\right]^2 \tag{6}$$

where $D = (x_1, y_1), (x_2, y_2),...(x_m, y_m)$ represents the data set, and y_i represents the true value of the independent variable x_i. The f denotes the trained algorithm model, and $f(x_i)$ is the predicted value calculated by the model based on the independent variable x_i.

It can be known from the formula that the MSE is non-negative. The smaller the MSE value, the higher the fitness of the algorithm model and the better the overall performance.

The R^2 reflects the degree to which the regression equation explains the change of the independent variable. It is also the statistic of the goodness of fit of the regression equation [37] at the same time. Its mathematical formula:

$$R^2(f;D) = 1 - \frac{\sum_{i=1}^{m}[f(x_i) - y_i]^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{7}$$

where $\bar{y}$ represents the average of the dependent variable y_i.

The variation range of the coefficient of determination is $0 \leq R^2 \leq 1$. When $R^2 = 1$, the regression equation completely fits the sample data. The magnitude of the R^2 indicates the percentage of the change in the dependent variable that can be explained by the independent variable. Generally, the higher the R^2, the better the regression fits data. However, if the R^2 gets too high, it may represent the overfitting of the algorithm model.

Experiment results

The trained model is validated on the test set, and the predicted value of the model is compared with the true value of the test set. The comparison of the results on K^+ is shown in fig 1.

If a change in a feature increases the error of the model, the feature is important. Breiman [17] introduced the concept of permutation feature importance measurement in 2001. In this experiment, the feature importance of random forest method on K^+ is shown in fig 2.

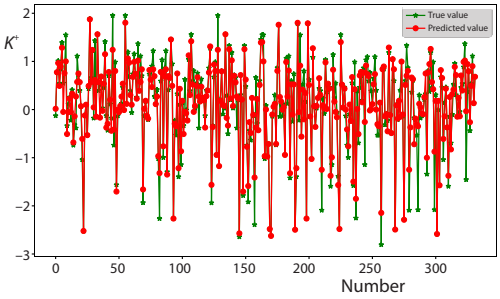

Figure 1. The comparison between the predicted results and the actual values

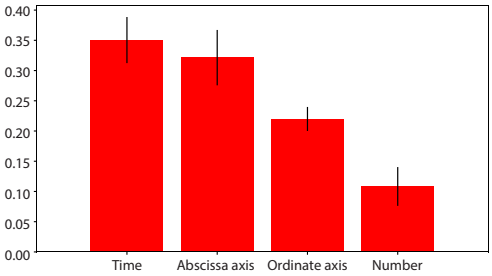

Figure 2. Feature importance of K^+

Results and discussion

In order to accurately and objectively judge the accuracy of the random forest algorithm, in this paper we compare it with SVM and Extra-Trees (extremely randomized trees) in the experimental stage. The MSE of different algorithm models on different ion test sets were shown in tab 2.

Table 2. The MSE of different models on different ion

Model	K^+	SO_4^{2-}	Cl^-	Mg^{2+}	Na^+
Random forest	0.229533	2.646299	3.073057	0.072694	0.945927
SVM	0.722956	6.153887	8.583963	0.080361	0.946203
Extra-trees	0.602676	3.831369	2.848140	0.117593	1.620384

It can be seen from the table that MSE varies with different ions, but the MSE of random forest model is kept to a minimum at most time, indicating that the deviation between the predicted value and the true value is the smallest.

The R^2 of different models on different ions were shown in fig 3. The values of random forest are all above 0.84, but they do not reach 0.99, indicating that the degree of fitting is good, and there is no over-fitting condition, and the degree of fitting of other models is relatively poor.

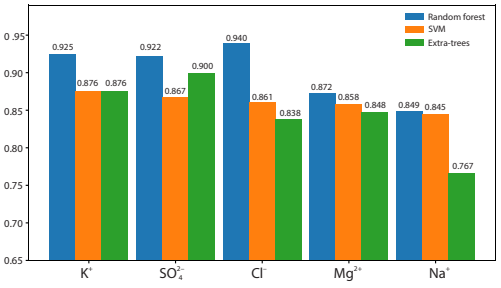

Figure 3. The R^2 of different models on different ions

It can be seen from the experimental results that the random forest has a certain degree of improvement in the MSE and the compared to other models. Therefore, the random forest algorithm achieves higher prediction accuracy and better generalization ability.

Conclusions

In the process of ion collection in Lop Nor salt lake, the ion concentration at the collection locality directly affects the method of removing impurities. Therefore, it is of great significance to predict the ion concentration at the collection site for the actual industrial production efficiency. In this paper, aiming at the problems of high computational cost and low prediction accuracy in the prediction of salt lake ion concentration, a random forest algorithm model was proposed for regression analysis. By simulating on a set of real data and using a variety of evaluation indicators to judge the performance of the model, the results show that the random forest achieves a higher prediction accuracy than SVM regression when implemented on the sample with less data and more noise, which has a certain significance for the collection process of salt lake resources in actual production.

In this paper, the parameters of random forest is not well tuned, so the accuracy of partial ion prediction is not good, which needs to be further studied. In addition, the attributes of the data are relatively small, and the simulation in the real environment may still have problems. Seeking more and better data attributes for regression analysis is also a problem that needs to be solved in the future.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61172150, 61803286), the Foundation of Hubei Provincial Key Laboratory of Intelligent Robot (HBIR 201802) and the tenth Graduate Innovation Fund of Wuhan Institute of Technology (CX2018197, CX2018200, and CX2018212).

References

[1] Lv, F. L., *et al*., The Discussion on Sedimentary Characteristics, Phased Evolution and Controlling Factors of Saline Lake in Asia Interior: Records from Deep Drill Cores of LDK01 in Lop Nur, Xinjiang, North-Western China, *Acta Petrologica Sinica*, *31* (2015), 9, pp. 2770-2782
[2] Xu, H., *et al*., Distribution Characteristics of the Main Ions in Luobei Area of Luobupo Salt Lake, *Shandong Land and Resources*, *29* (2013), 1, pp. 24-31
[3] Sun, M.-G., *et al.*, Potassium-Rich Brine Deposit in Lop Nor Basin, Xinjiang, China, *Scientific Reports*, *8* (2018), 1, 7676
[4] Kong, D. Y., *et al*., Seasonal Change of Water Absorption Capability and Moisture Content of the Top Salt-Crust in Lop Nur Dry Lake, *Acta Geoscientica Sinica*, *37* (2016), 2, pp. 185-192
[5] Qiu, X., *et al*., Removal of Borate by Layered Double Hydroxides Prepared through Microwave-Hydro-Thermal Method, *Journal of Water Process Engineering*, *17* (2017), June, pp. 271-276
[6] Chen, F., *et al*., Ionothermal Synthesis of Fe_3O_4 Magnetic Nanoparticles as Efficient Heterogeneous Fenton-Like Catalysts for Degradation of Organic Pollutants with H_2O_2, *Journal of Hazardous Materials*, *322*, *Part A* (2017), Part A, pp. 152-162
[7] Mao, Y. W., *et al*., Microstructure Analysis of Graphite/Cu Joints Brazed with (Cu-50TiH2)+B Composite Filler, *Fusion Engineering and Design*, *100* (2015), Nov., pp. 152-158

[8] Miao, L., *et al*., Cooking Carbon with Protic Salt: Nitrogen and Sulfur Self-Doped Porous Carbon Nanosheets for Supercapacitors, *Chemical Engineering Journal*, *347* (2018), Sept., pp. 233-242
[9] Chao, J. I., et al., Separation Properties of Magnesium and Lithium from Brine with High Mg~ (2+)/Li~+ Ratio by DK Nanofiltration Membrane, *Membrane Science and Technology*, *42* (2014), 4, pp. 607-615
[10] Xu, C. U., *et al*., Predict Chloride Concentration in Concrete Based on Neural Network, *Concrete*, (2010),
[11] Chen, H., *et al*., Fault Identification of Gearbox Degradation with Optimized Wavelet Neural Network, *Shock and Vibration*, *20* (2013), 2, pp. 247-262
[12] Parveen, N., *et al*., Development of SVR-Based Model and Comparative Analysis with MLR and ANN Models for Predicting the Sorption Capacity of Cr (VI), *Process Safety and Environmental Protection*, *107* (2017), Apr., pp. 428-437
[13] Xiong, F. Q., *et al*., Integrated Prediction Model of Iron Concentration in Goethite Method to Remove Iron Process, Control and Decision, *27* (2012), 3, pp. 329-334.
[14] Suykens, J. A. K., Vandewalle, J., Least Squares Support Vector Machine Classifiers, *Neural Processing Letters*, *9* (1999), 3, pp. 293-300
[15] Liu, J., *et al*., Predicting the Content of Camelina Protein Using FT-IR Spectroscopy Coupled with SVM Model, *Cluster Computing*, first on-line, https://doi.org/10.1007/s10586-018-1838-3
[16] Liu, J., *et al*., Intelligent Predicting of Salt Pond's Ion Concentration Based on Support Vector Regression and Neural Network, *Neural Computing and Applications*, first on-line, https://doi.org/10.1007/s00521-018-03979-9,2019
[17] Breiman, L., Random Forests, *Machine Learning*, *45* (2001), 1, pp. 5-32
[18] Azar, A. T., *et al*., A Random Forest Classifier for Lymph Diseases, *Computer Methods and Programs in Biomedicine*, *113* (2014), 2, pp. 465-473
[19] Huynh, T., *et al*., Estimating CT Image from MRI Data Using Structured Random Forest and Auto-Context Model, *IEEE Transactions on Medical Imaging*, *35* (2016), 1, pp. 174-183
[20] Keramati, A., *et al*., Developing a Prediction Model for Customer Churn from Electronic Banking Services Using Data Mining, *Financial Innovation*, *2* (2016), 1, 10
[21] Pandya, R., Pandya, J., The C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning, *International Journal of Computer Applications*, *117* (2015), 16, pp. 18-21
[22] Singh, S., Gupta, P., Comparative Study ID3, Cart and C4.5 Decision Tree Algorithm: A Survey, *Int. J. Adv. Inf. Sci. Technol., (Internet)*, *7* (2014), 3, pp. 47-52
[23] Song, Y.-Y., Ying, L., Decision Tree Methods: Applications for Classification and Prediction, *Shanghai Archives of Psychiatry*, *27* (2015), 2, 130
[24] Manek, A. S., *et al*., Aspect Term Extraction for Sentiment Analysis in Large Movie Reviews Using Gini Index Feature Selection Method and SVM Classifier, *World Wide Web*, *20* (2017), 2, pp. 135-154
[25] Toma, T., *et al*., Learning Predictive Models that Use Pattern Discovery – A Bootstrap Evaluative Approach Applied in Organ Functioning Sequences, *Journal of Biomedical Informatics*, *43* (2010), 4, pp. 578-586
[26] Wang, Z. L., *et al*., Flood Hazard Risk Assessment Model Based on Random Forest, *Journal of Hydrology 527* (2015), Aug., pp. 1130-1141
[27] Hlihor, R. M., *et al*., Experimental Analysis and Mathematical Prediction of Cd (II) Removal by Biosorption Using Support Vector Machines and Genetic Algorithms, *New Biotechnology*, *32* (2015), 3, pp. 358-368
[28] AndrE, S., *et al*., Developing Global Regression Models for Metabolite Concentration Prediction Regardless of Cell Line, *Biotechnology and Bioengineering*, *114* (2017), 11, pp. 2550-2559
[29] Yuntao W. U., *et al*., HOSVD-Based Subspace Algorithm for Multidimensional Frequency Estimation without Pairing Parameters, *Chin J Electron*, *23* (2014), 4, pp. 729-734
[30] Wang, Z., *et al*., Trilateral Constrained Sparse Representation for Kinect Depth Hole Filling, *Pattern Recognit Lett*, *65* (2015), Nov., pp. 95-102
[31] Liu, H., *et al*., Distributed Source Localization under Anchor Position Uncertainty, *Chin. J. Electron*, *23* (2014), 1, pp. 93-96
[32] Zhong, L., *et a*l., OHRank: An Algorithm Integrating Mentality and Influence of Opinion Holder for Opinion Mining, *Chin. J. Electron*, *22* (2013), 4, pp. 655-660
[33] Peng, Li., *et al*., A Robust Method for Estimating Image Geometry with Local Structure Constraint, *IEEE Access*, *6* (2018), Feb., pp. 20734-20747
[34] Lu, T., *et al*., Robust Face Super-Resolution Via Locality-Constrained Low-Rank Representation, *IEEE Access*, *5* (2017), June, pp. 13103-13117

[35] Yun Tao, W., *et al*., Utilizing Principal Singular Vectors for 2-D DOA Estimation in Single Snapshot Case with Uniform Rectangular Array, *International Journal of Antennas and Propagation*, 2015 (2015), ID 691251
[36] Yun Tao, W., *et al.*, HOSVD-Based Subspace Algorithm for Multidimensional Frequency Estimation Without Pairing Parameters, *Chinese Journal of Electronics*, *23* (2014), 4, pp. 729-734
[37] Nagelkerke, N. J. D., A Note on a General Definition of the Coefficient of Determination, *Biometrika*, *78* (1991), 3, pp. 691-692

Paper submitted: November 28, 2018
Paper revised: January 25, 2019
Paper accepted: February 5, 2019