GOSPODARKA SUROWCAMI MINERALNYMI – MINERAL RESOURCES MANAGEMENT

2025 Volume 41 Issue 4 Pages 175–192

DOI: 10.24425/gsm.2025.155880

ZHI HONG JIANG[1], AO CHEN[2]

Research on incremental semi-supervised ore particle size prediction algorithm

Introduction

Ore particle size is one of the critical parameters for the control of the ore dressing process, and it is also an essential reference to measure the crushing effect; real-time detection of ore particle size is of great significance for the optimization of the ore crushing process. In complex working conditions, the traditional ore particle size prediction model in the face of the actual ore size sample data, there are insufficient samples in the early stage, which cannot train the model, the later dynamic changes in the sample after the decline in model performance, retraining the model is inefficient and difficult to make the traditional prediction algorithms are complex to adapt to environmental changes.

✉ Corresponding Author: Ao Chen; e-mail: 1012558903@qq.com

[1] Jiangxi University of Science and Technology; Jiangxi Mining and Metallurgy Electromechanical Engineering Technology Research Center, China; e-mail: jzhee_mail@163.com

[2] Jiangxi University of Science and Technology, China; ORCID iD: 0009-0005-1579-6886; e-mail: 1012558903@qq.com

Incremental learning methods have been proposed to cope with conceptual drift in machine learning and to solve the problem of recognizing samples with changing distributions in machine learning models. It is a machine learning technique that learns new knowledge or handles new tasks by updating the model parameters step by step without re-training the whole model (Jiang et al. 2022; Gama et al. 2014). Si-si et al. divided the self-encoder into two parts. The first three layers are used to extract the samples' common features. The last three layers are used to extract the samples' private features, and the reconstruction error is used to determine whether the concept drift has occurred or not. After the concept drift occurs, the last three layers are learned and updated to avoid catastrophic forgetfulness of the incremental learning process (Si-si et al. 2021). Zeng et al. proposed an integrated learning method for incremental learning of unbalanced data, which constructs an integrated classification model by training multiple decision trees and, after obtaining a certain number of new samples, resampling the new samples based on the Bagging algorithm in order to balance their distributions and divide them into multiple sample sets, and then employing these sample sets to train all the decision trees in order to update the integrated classification model (Zeng et al. 2020). Zhu et al. realized parallel incremental learning by training a series of limiting support vector machines. They obtained the final results by weighting and summing the results of these limiting support vector machines. The experiments showed that the method achieved consistent results with batch learning (Zhu et al.2018). However, the above incremental learning methods solve the learning problem of increasing sample categories and cannot deal with the problem of changing the distribution of similar samples. They are all based on supervised learning, which is unable to cope with the problem of scarcity of labeled samples in the ore granularity prediction process.

The core concept of semi-supervised learning is to use the information of unlabeled samples to expand the limited set of labeled samples by obtaining pseudo-labels to improve the model's generalization ability (Hu et al. 2021). Kang et al. established a self-training-based semi-supervised support vector regression (SS-SVR) model to expand the dataset using a probabilistic local reconstruction (PLR) model to obtain pseudo-labeling (Kang et al. 2016). Mao et al. proposed a hyperspectral soil heavy metal mass concentration inversion model based on semi-supervised regression, and experiments showed that in the case of fewer labeled samples, the model inversion accuracy could be effectively improved by introducing a large number of unlabeled samples for semi-supervised regression analysis (Mao et al. 2022). However, the above method uses a traditional single learner to construct the prediction model, and the prediction accuracy is limited when there are dynamic changes in the samples, and its generalization ability needs to be further improved.

This paper proposes an incremental semi-supervised particle size prediction algorithm. The semi-supervised learning mechanism expands a limited number of ore particle size samples; high-quality pseudo-labeled samples are used to incrementally learn the model. This gives the ore particle size prediction model the ability to dynamically learn new knowledge and ensures that the model always has good prediction performance.

1. A high-confidence pseudo-label design method for ore particle size detection with the label sample scarcity problem

In the field of ore particle size prediction, the insufficiency of labeled samples often limits the training effect of the model. Traditional supervised learning methods rely on many labeled samples to ensure the model's accuracy. However, the acquisition of labeled samples of ore particle size mainly relies on the sieving method, which requires a large workforce and material resources, is inefficient, and has a high cost. In order to reduce the dependence of the ore particle size prediction model on labeled samples, it is proposed to utilize the pseudo-labeling technique to improve the performance and generalization ability of the model.

The core idea of the pseudo-labeling technique is to use an existing model to predict unlabeled samples and use the prediction results as "pseudo-labels" to expand the training set (Xu et al.2024). However, to ensure the credibility of pseudo-labeling, effective screening and calibration of pseudo-labeling are required to prevent the negative impact of misinformation on the model. A common strategy is a screening method based on model confidence, where pseudo-labels are adopted only when the confidence of the prediction results is higher than a certain threshold.

Based on the pseudo-labeling technique, the semi-supervised learning method further optimizes the model's training process. Semi-supervised learning combines a limited number of labeled samples with many unlabeled samples, which improves the model's performance by effectively utilizing the potential information of unlabeled samples (Wang et al.2024). The Tri-training algorithm is a multi-classifier combination method, and the primary purpose is to combine the outputs of individual classifiers to obtain better performance than a single classifier (Fu et al.2024). In this paper, we propose an ore particle size prediction algorithm based on semi-supervised learning based on the Tri-training algorithm. The flowchart is shown in the author's already-published article (Zhihong et al.2024).

The algorithm flow is as follows.

1. Build three regression prediction models by a stratified sampling of the original ore particle size labeled samples M_a and constructing the training set D_a; $M_a = \left(x_{m_a}, y_{m_a}\right)$ and m_a is the number of ore-size samples.
2. Fuse the prediction results of the three regression models using the VotingRegressor algorithm to pseudo-label the unlabeled ore particle size samples U_u and screen out the high-confidence pseudo-labeled samples.
3. Construct a new regression prediction model using the mixed sample set $M_{new\text{-}train}$ regression prediction model.

1.1. Algorithmic implementation

Three regression prediction models of Tree, BP neural network and GBDT are selected to train D to obtain three initial regressors with significant differences h_1, h_2, h_3, calculate the maximum error $E_{(h_v)}$, $v = 1, 2, 3$, the relative error $e_{(h_v)}$, of each sample, the regressor coefficients $\alpha_{(hv)}$, and take the logarithm of the regressor coefficients according to the regressor coefficients and normalize it to get the regressor's weight $W_{(hv)}$, as shown in the following Equation:

$$W_{hi} = \frac{\ln \alpha_{hv}}{\ln \alpha_{h1} + \ln \alpha_{h2} + \ln \alpha_{h3}} \tag{1}$$

The VotingRegressor algorithm is used to fuse the models, combined with GridSearchCV (grid tuning method) in order to determine the optimal combination of weights, obtain the model coefficients β_{hv}, and output the final model:

$$h(x) = \sum_{hv=1}^{3} \beta_{hv} model_{(hv)} \tag{2}$$

Initial prediction of unlabeled ore grain size sample U_u using $h(x)$ yields a pseudo-labeled sample predicted $D'_a = \left(x_{m_u}, \hat{Y}_{x_{m_u}}\right)$ for U_u.

After obtaining the pseudo-labeled sample, it is necessary to determine whether the performance of the regression prediction model can be improved after the pseudo-labeled sample is added to the training setD_a. In this paper, we evaluate the regression model by calculating the mean square error of the regression model before and after the addition of $\left(x_{m_u}, \hat{Y}_{x_{m_u}}\right)$.

The mean square error of the regression model before the pseudo-labeled sample $\left(x_{m_u}, \hat{Y}_{x_{m_u}}\right)$ is added to the training set D_a is:

$$MSE = \frac{1}{m_a} \sum_{i=1}^{m_a} \left(Y_i - \hat{Y}_i\right)^2 \tag{3}$$

↳ Y_i – manually screened ore size distribution on the test set,
$\hat{Y}_i$ – predicted size distribution on the test set.

After $\left(x_{m_u}, \hat{Y}_{x_{m_u}}\right)$ is added to the training set, the mean square error of the regression model is denoted as *MSE'*, Let $\Delta_u = MSE' - MSE$ to obtain the difference Δ_u of the mean square error of the regression model before and after $\left(x_{m_u}, \hat{Y}_{x_{m_u}}\right)$ is added to the training set, whose computational expression is shown in Equation (4).

$$\Delta_u = \sum_{x_i \in M_a} \left(\left(y_i - f'(x_i)\right)^2 - \left(y_i - f(x_i)\right)^2 \right) \tag{4}$$

↳ $f(x_i)$ – initial regression model,
$f'(x_i)$ – regression model after adding the pseudo-label sample $\left(x_{m_u}, \hat{Y}_{x_{m_u}}\right)$.

Obviously, the larger the value of Δ_u, the greater the positive impact of the pseudo-labeled sample $\left(x_{m_u}, \hat{Y}_{x_{m_u}}\right)$ on improving the performance of the regression model. The threshold of the above iterative process is noted as the confidence level θ If $\Delta_u > \theta$, then it is a pseudo-labeled sample with a high confidence level, and this pseudo-labeled sample $\left(x_{m_u}, \hat{Y}_{x_{m_u}}\right)$ is added to the training set D_a. If $\Delta_u < \theta$, then the unlabeled sample U_u is returned (Zhihong et al. 2024).

Repeating the above evaluation process for a single unlabeled sample, the pseudo-labeled samples added to the training set D_a through the above process are noted as $\left(X_{sel}, \hat{y}_{sel}\right)$, yielding the mixed sample set $M_{new-train} = \left\{D_a, \left(X_{sel}, \hat{y}_{sel}\right)\right\} = \left(X_{new}, Y_{new}\right)$.

In this paper, the final prediction of the mixed sample set is performed using the stacked integration learning algorithm. Stacking is an integration strategy combining multiple base models through metamodeling, which is a serial-structured multilayer learning system (Zhu et al. 2024; SVN et al. 2024). The stack-integrated learning regression model can integrate many base regression models, such as single linear regression, machine learning regression models, etc., to produce more robust estimation results and provide better generalization capabilities in the granularity prediction process. The framework of its prediction model is shown in Figure 1.

When selecting the base learner for the first layer, choosing a base model with solid learning ability helps the overall prediction effect. The prediction performance of the regression model can be evaluated by the deviation of the ore particle size prediction data from the manual sieving size data, using the root mean square error (RMSE), the mean absolute error (MAE), and the coefficient of determination (R^2) as the evaluation criteria.

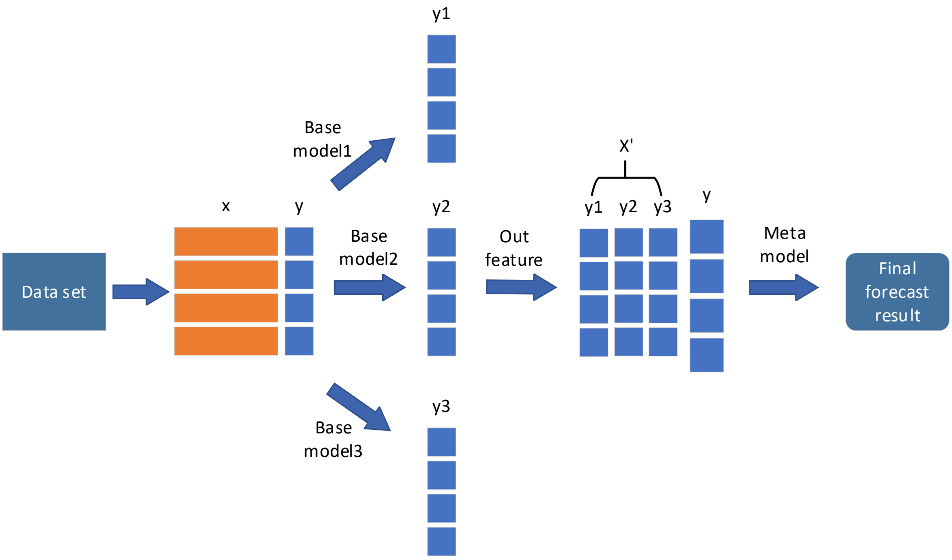

Fig. 1. The framework of the Stacking Ensemble Learning Prediction Model

Rys. 1. Struktura modelu predykcyjnego uczenia zespołowego typu stacking

When the predicted particle size distribution of the ore matches perfectly with that of the manual sieving, the RMSE and MAE are equal to 0, and R^2 is equal to 1, i.e., the established regression model is perfect; the more significant the error is, the larger the RMSE and MAE are, and the smaller the R^2 is (Liao et al. 2024; Gavin et al. 2023; Schenk et al. 2024). The expression of root mean square error is shown in Equation (5), the mean absolute value error is shown in Equation (6), and the expression of coefficient of determination is shown in Equation (7):

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2} \tag{5}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|\left(y_i - \hat{y}_i\right)\right| \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{m}\left(\overline{y}_i - y_i\right)^2} \tag{7}$$

↳ m – number of samples,
y_i – particle size distribution of manually sieved ore,
$\hat{y}_i$ – predicted particle size distribution of the model.

Applying the mixed sample set $M_{new-train}$, the RMSE, MAE, and R^2 of eight commonly used learners are selected, as shown in Table 1.

Table 1. Evaluation metrics of 8 commonly used learners on mixed sample sets

Tabela 1. Wskaźniki oceny 8 powszechnie stosowanych algorytmów uczenia się na mieszanych zestawach próbek

Model	RMSE	MAE	R^2
Random forest	0.047	0.032	0.795
BP neural network	0.040	0.033	0.767
GBDT	0.030	0.024	0.872
XGBoost	0.049	0.035	0.655
Lightgbm	0.070	0.045	0.287
Ridge Regression	0.044	0.714	0.036
Tree	0.026	0.022	0.904
SVM	0.093	0.073	–0.237

The table shows that Tree, GBDT, BP Neural Network, and Random Forest are the better ones in the single model. Therefore, this paper selects these four learners as candidate-based learners.

The selection of the second layer meta-learner generally tends to favor models with more vital generalization ability or simpler models to reduce overfitting. Among the above learners, XGBoost is a model with more vital generalization ability by comprehensively applying regularization, pruning, feature selection, missing value processing, efficient computation, and an early-stopping mechanism to improve the model prediction performance; ridge regression, as an optimization model for linear regression, is a simpler model (Aamir et al. 2024; Deng and Lumley 2024). Therefore, this paper compares using XGBoost and ridge regression as candidate meta-learners.

Experiments are designed to observe the prediction effect of the Stacking model under different combination methods, and some experimental results are shown in Table 2.

Table 2. Predictive effects of stacking models in different combinations

Tabela 2. Efekty predykcyjne modeli stackingowych w różnych kombinacjach

Combinatorial approach	Base model	Meta model	RMSE	MAE	R^2
1	Tree + GBDT + BP	XGBoost	0.0620	0.0401	0.4439
2	Tree + GBDT + BP	ridge regression	0.0234	0.0168	0.9207
3	Tree + GBDT + RF	ridge regression	0.0306	0.0252	0.8646
4	Tree + GBDT	ridge regression	0.0318	0.0263	0.8533
5	Tree + GBDT + BP + RF	ridge regression	0.0515	0.0379	0.6157

According to the table, there is a significant difference in the prediction effect of the Stacking model in different combinations. The prediction effect of the combination way 1 is better than the combination way 2, which indicates that ridge regression as a meta-model performs better than XGBoost and is more conducive to improving the model's generalization ability. Comparing the combination ways 2, 4, and 5, it can be obtained that the number of base models should be chosen appropriately, and in the case of this experiment, the prediction performance of the 3 base models is the best, and increasing or decreasing will reduce the prediction performance of the model. Therefore, in this paper, Tree, GBDT, and BP neural networks are selected as the base models of the Stacking model and the Ridge regression model as the Meta model.

2. Incremental semi-supervised ore particle size prediction algorithm

In the actual mine production process, the ore particle size samples are generated gradually over time, and the distribution law of the samples may also change over time; the traditional static model training method makes it difficult to cope with the continuous evolution of the sample distribution, in order to address this problem, this paper expands the training sample library through the high-quality pseudo-labeled samples generated by the semi-supervised learning method. It adopts incremental learning to update the model parameters step by step when the model accepts new data to adapt to the changes in sample distribution, optimize the stability of the model in a dynamic environment, and improve its prediction performance.

Deep Neural Network (DNN), the most active and successful machine learning model in artificial intelligence, is versatile and scalable, applicable to various types of machine learning problems, and adapted to different scenarios by adjusting model parameters (Ail et al. 2023; Bhadely et al. 2024). In this paper, an incremental semi-supervised ore granularity prediction algorithm is proposed based on the DNN small batch gradient descent algorithm; in the process of minimizing the DNN error function, the iterative solution is carried out by the small batch gradient descent method, and the minimized error function values and model parameter values can be obtained after a certain number of steps.

The DNN structure is designed to ensure that the neurons directly, quickly, and accurately transfer effective information. The established DNN structure system is shown in Figure 2.

The feature vector X of the high-quality pseudo-label sample data judged by confidence in the semi-supervised algorithm framework is imported into the input layer. $X_{(k)}$ is the first element in X, the feature vector of the kth high-quality pseudo-label sample data, $k = 1, 2, 3, \ldots, m$ and is the number of samples. According to the input value X, the output feature vector Y of the high-quality pseudo-label sample is calculated by Equation (10).

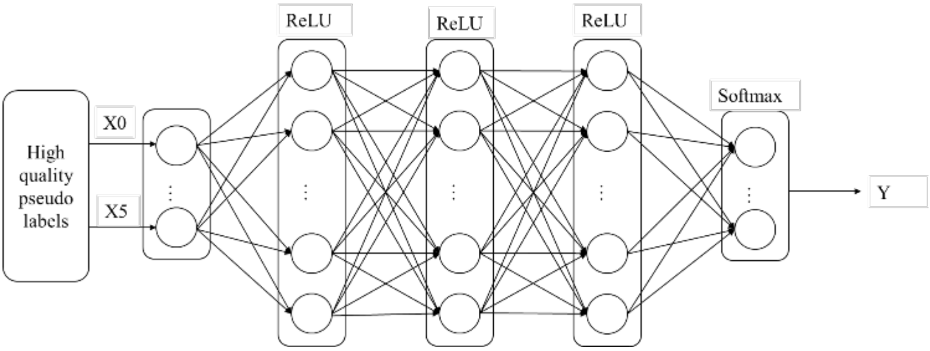

Fig. 2. DNN structure architecture diagram

Rys. 2. Schemat architektury struktury DNN

$$X = \left(X_{(1)}, X_{(2)}, \ldots, X_{(k)} \ldots, X_{(m)}\right)^{\mathrm{T}} \tag{8}$$

$$X_{(k)} = \begin{pmatrix} x_1 \\ \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{9}$$

$$Y = f\left(W_{ij}^{(l)} \ldots f\left(W_{ij}^{(2)} f\left(W_{ij}^{(1)} X + b_{ij}^{(1)}\right) + b_{ij}^{(2)}\right) + \ldots b_{ij}^{(l)}\right) \tag{10}$$

$$W_{ij}^{(l)} = \begin{bmatrix} W_{11}^{(l)} & \cdots & W_{1j}^{(l)} \\ \vdots & & \vdots \\ W_{i1}^{(l)} & \cdots & W_{ij}^{(l)} \end{bmatrix} \tag{11}$$

$$b_{ij}^{(l)} = \begin{bmatrix} b_{11}^{(l)} & \cdots & b_{1j}^{(l)} \\ \vdots & & \vdots \\ b_{i1}^{(l)} & \cdots & b_{ij}^{(l)} \end{bmatrix} \tag{12}$$

↳ n – number of data dimensions;
$W_{ij}^{(l)}, b_{ij}^{(l)}$ – weight feature vectors and deviation feature vectors from l to $l + 1$ layers, respectively;
i, j – *ith* and *jth* neurons from l to $l + 1$ layers, respectively;
l – number of hidden layers, $l = 1, 2, 3, \ldots, r$.

In order to avoid the phenomenon of gradient vanishing during the backpropagation of the DNN, to improve the computational speed and accuracy, and at the same time to reduce the dependence of the DNN parameters with the probability of overfitting, Relu (the gradient is always 0 or 1) is used as the activation function (Yang et al. 2024). Finally an L2 regularization term is added to prevent model overfitting.

$$g(x) = ReLU(x) = \max(0, x) \tag{13}$$

↳ x – denotes the input value;
$g(x)$ – denotes the output value.

The principle of the gradient-based optimization algorithm is to determine the forward direction and distance of each iteration by calculating the derivative information of the objective function, and finally obtain the global or local optimal solution of the unknown parameters (Sclocchi et al. 2024). Among them, Batch Gradient Descent (BGD), Stochastic Gradient Descent (SGD) and Mini-batch Gradient Descent (MBGD) are three typical variants of optimization algorithms. The BGD algorithm utilises the entirety of the training samples to calculate the gradient in each iteration, exhibiting excellent convergence properties but concomitant large computation volume. Furthermore, the training process does not conform to the principle of incremental learning algorithms. Conversely, the SGD algorithm employs a single training sample to calculate the gradient in each iteration. This results in the possibility of the SGD algorithm falling into a local optimum, which can be detrimental to the efficiency of the training process. The MBGD algorithm utilises a predetermined number of samples to calculate the gradient, thereby ensuring the capture of the global characteristics of the data and adhering to the fundamental principles of incremental learning algorithms. Therefore, in this paper, we choose small batch gradient descent to achieve incremental learning.

Assuming that the cost function of the DNN model is $l(\theta)$, the gradient descent algorithm updates the parameters using the derivative $\partial l(\theta)/\partial\theta$ of the cost function concerning the target parameters, as shown in Equation (14).

$$\theta = \theta - \eta \cdot \frac{\partial l(\theta)}{\partial \theta} \tag{14}$$

↳ η – learning rate parameter.

For the DNN model shown in Figure 2. Assuming that the output of the lth hidden layer is h^l, there is:

$$h^l = g\left(W^l h^{l-1}\right) \tag{15}$$

$$y = W^{r+1} h^r \tag{16}$$

↳ W^l – weight parameter between the $l - 1st$ implicit layer and the lth implicit layer;
g – activation function.

Then the cost function can be expressed as:

$$l(\theta) = \frac{1}{2}\sum_{l=1}^{n}\left\{y_l - W^{l+1}\left[g\left(W^l\left(\cdots g\left(W^1 x_l\right)\right)\right)\right]\right\}^2 \tag{17}$$

The small batch gradient descent algorithm uses a small sample set of a fixed number of training samples for each parameter update (Chen et al. 2024). Setting the number of samples in the small sample set to *z*, the network parameter update rule in DNN can be expressed as the following equation:

$$W^l = W^l - \eta \cdot \frac{1}{z} \sum_{j=1}^{z} \left. \frac{\partial l(\theta)}{\partial W^l} \right|_{(x_j, y_j)} \tag{18}$$

In order to get the optimal small sample set sample number *z*, calculate the model evaluation index under different conditions and compare, the calculated optimal small sample set sample number is 32, when the model effect is optimal.

3. Experiments and analysis of results

In order to verify the effectiveness of the incremental semi-supervised ore particle size prediction algorithm, In this paper, we chose to collect the measured data sets on the fine crushing belt of a processing plant with three different ore grain sizes, namely +10 mm, +5~10 mm and –5 mm, for the experiments, using 400 high-quality pseudo-labeled data through the confidence judgment as the training set, and using 63 labeled samples as the test set to compare and analyze the results of several commonly used prediction models and incremental semi-supervised models. Test results, the model test comparison results, and evaluation indexes are shown in Table 3 and Figure 3.

From Table 3 and Figure 3, it can be found that the lightGBM model presents a significant error in the prediction of ore particle size. At the same time, Random Forest, GBDT, XGBoost, and BP neural network have a minor error in the prediction of ore particle size;

Table 3. Comparison of evaluation indexes of 6 prediction models

Tabela 3. Porównanie wskaźników oceny 6 modeli prognostycznych

Model	RMSE	MAE	R^2
Random forest	0.02986	0.00965	0.83186
BP neural network	0.02106	0.00779	0.91641
GBDT	0.02837	0.00881	0.84830
XGBoost	0.03220	0.01226	0.80452
LightGBM	0.03493	0.01426	0.76996
Incremental semi-supervised model	0.00354	0.00086	0.99344

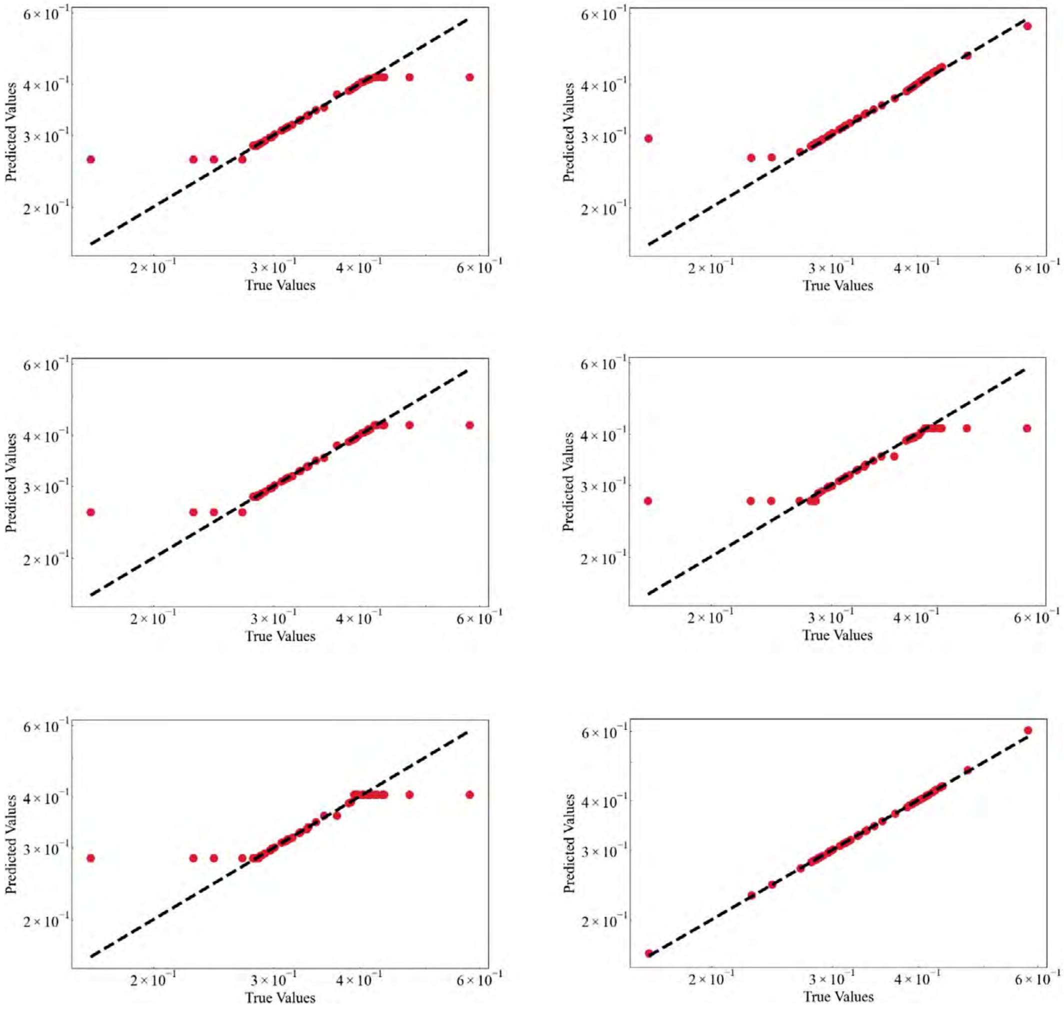

Fig. 3. Comparison of test results of 6 prediction models

Rys. 3. Porównanie wyników testów 6 modeli prognostycznych

the incremental semi-supervised ore size prediction algorithm proposed in this paper has a minor error, and the most significant coefficient of determination, with the RMSE, MAE, and R2 being 0.00354, 0.00086 and 0.99344, respectively. The RMSE, MAE, and R2 are 0.00354, 0.00086, and 0.99344, respectively. Therefore, compared with other commonly used machine learning methods, the incremental semi-supervised ore size prediction algorithm proposed in this paper has high accuracy.

In order to further analyze the performance of the algorithm, the 400 samples in the training set are divided into five small training sets, with the number of samples being

40, 100, 160, 200, and 300, respectively, to compare and analyze the test results of several commonly used prediction models as well as incremental semi-supervised prediction models on these five small training sets, and to use the RMSE, MAE, and R2 as the evaluation criteria, and to compute the results with the model evaluation indexes. Trends are shown in Table 4 and Figure 4.

Table 4. Comparison of evaluation metrics of 6 predictive models on 5 training sets

Tabela 4. Porównanie wskaźników oceny 6 modeli predykcyjnych na 5 zestawach szkoleniowych

Training set	Model	RMSE	MAE	R^2
Dataset 1	Random forest	0.0343	0.0133	0.77824
	BP neural network	0.03253	0.01962	0.80047
	GBDT	0.03385	0.01285	0.78399
	XGBoost	0.03572	0.01569	0.75950
	Light GBM	0.07375	0.05901	–0.02562
	Incremental semi-supervised model	0.00918	0.00386	0.98408
Dataset 2	Random forest	0.0313	0.01111	0.81565
	BP neural network	0.02441	0.01735	0.88768
	GBDT	0.03037	0.01062	0.82614
	XGBoost	0.03309	0.01381	0.79354
	Light GBM	0.03933	0.01889	0.70833
	Incremental semi-supervised model	0.01294	0.00364	0.96839
Dataset 3	Random forest	0.0326	0.01171	0.79928
	BP neural network	0.03626	0.02191	0.75212
	GBDT	0.03183	0.01126	0.80895
	XGBoost	0.03432	0.01353	0.77796
	Light GBM	0.03945	0.01887	0.70662
	Incremental semi-supervised model	0.01889	0.00524	0.93266
Dataset 4	Random forest	0.0335	0.01177	0.78899
	BP neural network	0.03375	0.02015	0.78519
	GBDT	0.03313	0.01166	0.79311
	XGBoost	0.03421	0.01301	0.77934
	Light GBM	0.03796	0.0171	0.72831
	Incremental semi-supervised model	0.0139	0.00349	0.96352

Training set	Model	RMSE	MAE	R^2
Datase t5	Random forest	0.0298	0.00959	0.83299
	BP neural network	0.0203	0.00654	0.92233
	GBDT	0.02837	0.00884	0.84827
	XGBoost	0.03235	0.01243	0.80272
	Light GBM	0.0354	0.01446	0.76368
	Incremental semi-supervised model	0.00916	0.00181	0.98414
average value	Random forest	0.0326	0.01139	0.80096
	BP neural network	0.02980	0.01616	0.83037
	GBDT	0.03195	0.01130	0.80763
	XGBoost	0.03469	0.01367	0.77384
	Light GBM	0.03803	0.01816	0.73063
	Incremental semi-supervised model	0.01306	0.0038	0.96104

The brown folded line in Figure 4 shows the incremental semi-supervised ore size prediction algorithm proposed in this paper, which can be seen that for the five small training sets with different numbers of samples, the incremental semi-supervised ore size prediction algorithm proposed in this paper shows significant advantages. Among them, the incremental semi-supervised ore particle size prediction algorithm has the lowest average RMSE and MAE on the five small training sets, which are 61.23% and 66.78% lower than that of the random forest, and 56.92% and 76.16% lower than the BP neural network, respectively, and has the highest average coefficient of determination, which is 15.34% higher than that of the GBDT model, and 23.04% higher than that of the lightGBM model 23.04%, the performance improvement is noticeable.

Conclusion

1. Aiming at the traditional prediction model in ore particle size detection, when facing the actual ore particle size sample data, there is a problem of insufficient samples in the early stage that cannot train the model, and the performance of the model decreases in the later stage after the sample changes, the semi-supervised algorithm based on the idea of Tri-training is utilized to screen out high-quality pseudo-labeled samples with integrated learning algorithms, such as VotingRegressor. It is used as incremental data to train the DNN-based small batch gradient descent model to improve the prediction accuracy of the prediction model.

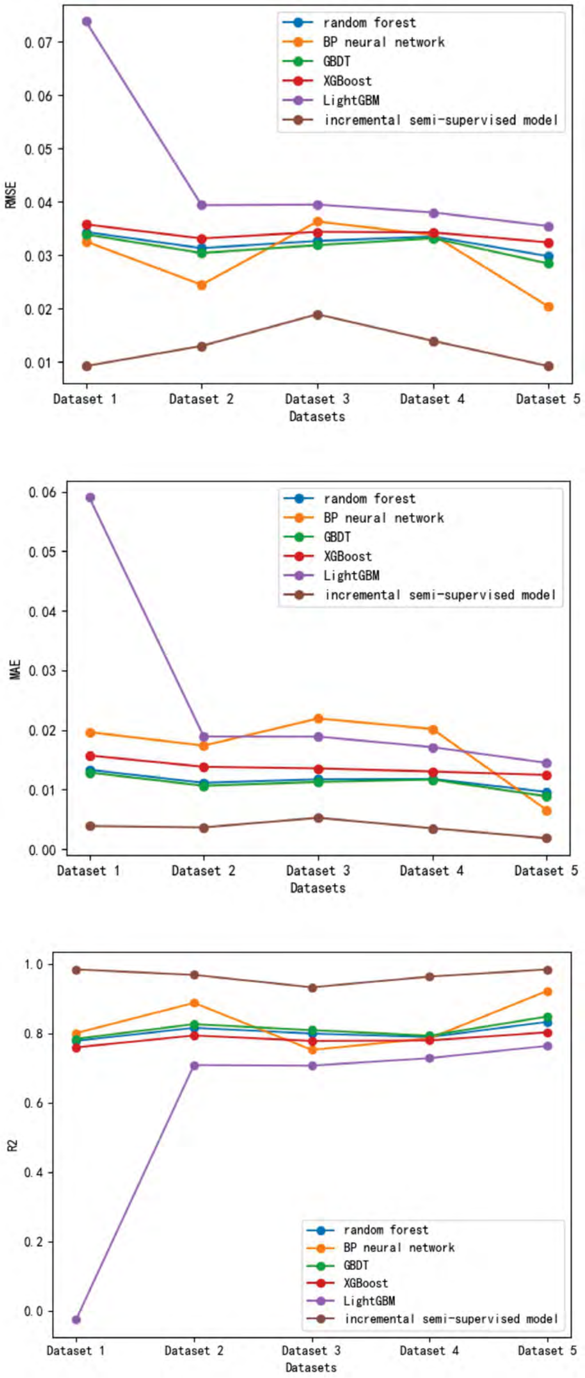

Fig. 4. Trends in metrics for 6 predictive models on 5 datasets

Rys. 4. Trendy w metrykach dla 6 modeli predykcyjnych na 5 zestawach danych

2. Compared with the traditional modeling methods such as random forest, BP neural network, XGBoost, etc., the incremental semi-supervised ore size prediction algorithm improved the model decision coefficients by 16.16%, 7.71%, and 18.89%, respectively, decreased the RMSE by 88.16%, 87.5%, and 89.09%, and the MAE by 91.09%, 91.96%, and 92.89%, respectively. 92.89%, verifying the feasibility and reliability of the incremental semi-supervised ore particle size prediction algorithm.
3. Aiming at the problem that traditional prediction algorithms are complex to adapt to environmental changes due to environmental variability and challenges, the proposed incremental semi-supervised ore particle size prediction algorithm is more suitable for actual ore size prediction than the traditional prediction model, and the accuracy rate has been significantly improved.

The Authors have no conflict of interest to declare.

This research was supported by the National Natural Science Foundation of China (52364025) and the National Natural Science Foundation of China (51464017).

REFERENCES

Aamir et al. 2024 – Aamir, S., Muhammad, A.,Walid, E. and Muhammad, F. 2024. New ridge parameter estimators for the quasi-Poisson ridge regression model. *Scientific Reports* 14(1), pp. 8489–8489, DOI: 10.1038/s41598-023-50085-5.

Bhadely, A.F. and İnan, A. 2024. An Innovative Approach for Enhancing Relay Coordination in Distribution Systems Through Online Adaptive Strategies Utilizing DNN Machine Learning and a Hybrid GA-SQP Framework. *Arabian Journal for Science and Engineering* 49(12), DOI: 10.1007/s13369-024-09291-0.

Chen et al. 2024 – Chen, L., Xiong, M., Ming, J. and He, X. 2024. Efficient mini-batch stochastic gradient descent with Centroidal Voronoi Tessellation for PDE-constrained optimization under uncertainty. *Physica D: Nonlinear Phenomena* 467, DOI: 10.1016/j.physd.2024.134216.

Deng, Y. and Lumley, T. 2024. Multiple Imputation Through XGBoost. *Journal of Computational and Graphical Statistics* 33(2), DOI: 10.1080/10618600.2023.2252501.

El Bilaliet al. 2023 – El Bilali, A., Abdeslam, T., Ayoub, N., Lamane, H., Ezzaouini, M.A. and Elbeltagi, A. 2023. An interpretable machine learning approach based on DNN, SVR, Extra Tree, and XGBoost models for predicting daily pan evaporation. *Journal of Environmental Management* 327, DOI: 10.1016/j.jenvman.2022.116890.

Fu et al. 2024 – Fu, R., Cheng, C., Yan, S., Wang, X. and Chen, H. 2024. Consistency-based semi-supervised learning for oriented object detection. *Knowledge-Based Systems* 304(9), DOI: 10.1016/j.knosys.2024.112534.

Gama et al. 2014 – Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, H. 2014. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)* 46(4), DOI: 10.1145/2523813.

Gavin et al. 2022 – Gavin, C., Devin, F. and Kellin, R. 2022. Bayesian projection pursuit regression. *Statistics and Computing* 34(1), DOI: 10.48550/arXiv.2210.09181.

Hu et al. 2022 – Hu, S., Miao, D. and Pedrycz, W. 2022. Multi granularity based label propagation with active learning for semi-supervised classification. *Expert Systems With Applications* 192, DOI: 10.1016/j.eswa.2021.116276.

Jiang, Z. and Chen, A. 2024. Semi-supervised Ore Granularity Prediction Algorithm Incorporating Fully Supervised Learning. *Gold Science and Technology* 32(3), DOI: 10.11872/j.issn.1005-2518.2024.03.040.

Jiang et al. 2022 – Jiang, W., Zhang, T., Qiu, H., Li, H. and Xu, G. 2022. Incremental Learning, Incremental Backdoor Threats. *IEEE Transactions on Dependable and Secure Computing* 99, DOI: 10.1109/TDSC.2022.3201234.

Kang et al. 2016 – Kang, P., Kim, D. and Cho, S. 2016. Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing. *Expert Systems with Applications* 51, DOI: 10.1016/j.eswa.2015.12.027.

Liao et al. 2024 – Liao, Z., Dai, S. and Kuosmanen, T. 2024. Convex support vector regression. *European Journal of Operational Research* 313(3), DOI: 10.48550/arXiv.2209.12538.

Mao et al. 2022 – Mao, G., Tu, Y., Cui, W, et al. 2022. Hyperspectral inversion of soil heavy metal mass concentration based on semi-supervised regression. *Journal of Applied Sciences – Electronics and Information Engineering* 40(6).

Schenk et al. 2024 – Schenk, A., Berger, M. and Schmid, M. 2024. Pseudo-value regression trees. *Lifetime data analysis* 30(2), DOI: 10.1007/s10985-024-09618-x.

Sclocchi, A. and Wyart, M. 2024. On the different regimes of stochastic gradient descent. *Proceedings of the National Academy of Sciences of the United States of America* 121(9), DOI: 10.1073/pnas.2316301121.

Si-si et al. 2021 – Si-si, Z., Jian-wei, L. and Xin, Z. 2021. Adaptive online incremental learning for evolving data streams. *Applied Soft Computing Journal* 105(10), DOI: 10.1016/j.asoc.2021.107255.

Venkatesh et al. 2024 – Venkatesh, N.S., Sripada, D., Sugumaran, V. and Aghaei, M. 2024. Detection of visual faults in photovoltaic modules using a stacking ensemble approach. *Heliyon* 10(6), DOI: 10.1016/j.heliyon.2024.e27894.

Wang et al. 2024 – Wang, H., Li, Y., Li, S., Li, G., Sun, S., Sun, B., Cao, Y. and Shi, J. 2024. Tri-training algorithm based nuclear power systems semi-supervised fault diagnosis under multiple restricted data conditions. *Applied Soft Computing* 167(1), DOI: 10.1016/j.asoc.2024.112345.

Xu et al. 2024 – Xu, M.C., Zhou, Y., Jin, C., de Groot, M., Alexander, D.C., Oxtoby, N.P., Hu, Y. and Jacob, J. 2024. Expectation maximisation pseudo labels. *Medical image analysis* 94(3), DOI: 10.1016/j.media.2024.103125.

Yang et al. 2024 – Yang, D., Liu, H., Xu, B., Tang, C. and Cheng, T. 2024. A hybrid network with DNN and WGAN for supercontinum prediction. *Optical Fiber Technology*, DOI: 10.1016/j.yofte.2024.103816.

Zeng et al. 2020 – Zeng, L.,Wenchao, H., Yan, X., et al. 2020. Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm. *Knowledge-Based Systems* 195 (prepublish).

Zhu et al. 2018 – Zhu, L., Ikeda, K., Pang, S., Ban, T. and Sarrafzadeh, A. 2028. Merging weighted SVMs for parallel incremental learning. *Neural Networks* 100, DOI: 10.1016/j.neunet.2018.01.001.

Zhu et al. 2024 – Zhu, L., Lu, W., Luo, C., Xu, Y. and Wang, Z. 2024. An ensemble optimizer with a stacking ensemble surrogate model for identification of groundwater contamination source. *Journal of contaminant hydrology* 267(9), DOI: 10.1016/j.jconhyd.2024.104437.

RESEARCH ON INCREMENTAL SEMI-SUPERVISED ORE PARTICLE SIZE PREDICTION ALGORITHM

Keywords

ores, incremental learning, semi-supervisory, particle size testing, machine learning

Abstract

Aiming to improve ore particle size prediction accuracy in the beneficiation process, depending on the number of labeled samples and the fact that the traditional prediction model does not have continuous learning ability, the incremental semi-supervised ore size prediction algorithm is proposed. Taking the actual ore particle size data as the research object, we use semi-supervised learning and integrated learning to obtain high-quality pseudo-labeled samples, expand the limited number of labeled samples as incremental data for incremental training, and dynamically update the parameters of the prediction model to maintain good prediction ability. The incremental semi-supervised ore particle size prediction algorithm is validated using the ore particle size dataset obtained by the sieving method. The results show that the model coefficient of determination of the incremental semi-supervised ore particle size prediction algorithm reaches 0.9934. The root mean square error and the average absolute error are 0.00354 and 0.00086, the evaluation indexes after training on five different numbers of data sets are higher than the traditional prediction model, compared with the traditional prediction model prediction accuracy and generalization ability is significantly improved, in the face of the sample distribution changes in the face of the problem with the ability to dynamically learn new knowledge, to the intelligent management of the mining industry production field provides a good solution to improve the accuracy of the detection of the particle size of the ore to provide a strong technical support.

BADANIA NAD ALGORYTMEM STOPNIOWEGO, CZĘŚCIOWO NADZOROWANEGO PRZEWIDYWANIA WIELKOŚCI CZĄSTEK RUDY

Słowa kluczowe

rudy, uczenie przyrostowe, półnadzorowane, badanie wielkości cząstek, uczenie maszynowe

Streszczenie

W celu poprawy dokładności przewidywania wielkości cząstek rudy w procesie wzbogacania, w zależności od liczby oznaczonych próbek i faktu, że tradycyjny model przewidywania nie ma zdolności ciągłego uczenia się, proponuje się algorytm przyrostowego półnadzorowanego przewidywania wielkości rudy. Biorąc za przedmiot badań rzeczywiste dane dotyczące wielkości cząstek rudy, wykorzystujemy uczenie półnadzorowane i uczenie zintegrowane w celu uzyskania wysokiej jakości próbek z etykietami pseudonimowymi, rozszerzamy ograniczoną liczbę próbek oznaczonych etykietami jako dane przyrostowe do szkolenia przyrostowego oraz dynamicznie aktualizujemy parametry modelu prognozowania, aby utrzymać dobrą zdolność prognozowania. Algorytm przyrostowego półnadzorowanego przewidywania wielkości cząstek rudy jest weryfikowany przy użyciu zbioru danych dotyczących wielkości cząstek rudy uzyskanych metodą przesiewania. Wyniki pokazują, że współczynnik determinacji modelu algorytmu przyrostowego półnadzorowanego przewidywania wielkości cząstek rudy osiąga 0,9934. Średni błąd kwadratowy i średni błąd bezwzględny wynoszą odpowiednio 0,00354 i 0,00086. Wskaźniki oceny po szkoleniu na pięciu różnych zestawach danych są wyższe niż w przypadku tradycyjnego modelu prognozowania, a w porównaniu z tradycyjnym modelem prognozowania znacznie poprawiono dokładność prognozowania i zdolność uogólniania. W obliczu zmian w rozkładzie próbek model ten wykazuje zdolność do dynamicznego uczenia się nowej wiedzy, co stanowi dobre rozwiązanie dla inteligentnego zarządzania w dziedzinie produkcji przemysłu wydobywczego, poprawiając dokładność wykrywania wielkości cząstek rudy i zapewniając silne wsparcie techniczne.