www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

Arch. Min. Sci. **68** (2023), 2, 251-266

Electronic version (in color) of this paper is available: http://mining.archives.pl

DOI: https://doi.org/10.24425/ams.2023.146178

DAN ZHAO[1], ZHIYUAN SHEN[1*], ZIHAO SONG[1], LINA XIE[2]

SDAE CLEANING MODEL OF WIND SPEED MONITORING DATA IN THE MINE MONITORING SYSTEM

The effective utilisation of monitoring data of the coal mine is the core of realising intelligent mine. The complex and challenging underground environment, coupled with unstable sensors, can result in "dirty" data in monitoring information. A reliable data cleaning method is necessary to figure out how to extract high-quality information from large monitoring data sets while minimising data redundancy. Based on this, a cleaning method for sensor monitoring data based on stacked denoising autoencoders (SDAE) is proposed. The sample data of the ventilation system under normal conditions are trained by the SDAE algorithm and the upper limit of reconstruction errors is obtained by Kernel density estimation (KDE). The Apriori algorithm is used to study the correlation between monitoring data time series. By comparing reconstruction errors and error duration of test data with the upper limit of reconstruction error and tolerance time, cooperating with the correlation rule, the "dirty" data is resolved. The method is tested in the Dongshan coal mine. The experimental results show that the proposed method can not only identify the dirty data but retain the faulty information. The research provides effective basic data for fault diagnosis and disaster warning.

Keywords: intelligent ventilation; monitoring data; data cleaning; association rules; stacked denoising autoencoder

1. Introduction

With the development of intelligent construction of coal mines, the monitoring data of mine sensors show a trend of considerable growth [1,2]. The sensor monitoring data provides basic information for the ventilation system fault diagnosis and disaster warning. However, the underground environment is complex and harsh, and the sensor signal transmission process is easily

[1] LIAONING TECHNICAL UNIVERSHITY, COLLEGE OF SAFETY SCIENCE & ENGINEERING, FUXIN 123000, CHINA
[2] SHENYANG INSTITUTE OF TECHNOLOGY, SHENYANG 110000, CHINA

* Corresponding author: szy0207@foxmail.com

affected by various underground interference sources to generate false signals [3,4]. In addition, abnormal or missing instantaneous values are caused by instrument failure, specific operation, dust, sensor failure, power failure, network transmission failure, manual adjustment, and other factors [5]. These outliers and missing values are collectively referred to as "dirty" data. Furthermore, ventilation parameters will change due to roadway blockage, accumulation of debris, improper support, fire, water inrush, large-scale roof fall, etc., which lead to abnormal horizontal migration or abnormal trend change in sensor monitoring data [6]. This part of the data does not belong to the cleaning category because they reflect the conditions of the ventilation system. Therefore, it is essential to eliminate and repair the "dirty" data while retaining the effective data when the ventilation system fails. Improving the quality of the basic data is the key to the research.

The research on data cleaning is mainly concentrated in the field of power. Gao et al. [7] proposed a multi-level cleaning and identification method of measured data based on the temporal and spatial correlation characteristics of the data. Yan et al. [8] applied time series analysis to establish transformer monitoring data cleaning. An effective data-cleaning model can both extract effective features from the data and repair outliers [9,10]. In 2008, Vincent and Bengio et al. [11,12] proposed a denoising autoencoder (DAE) deep neural network model. A stacked denoising autoencoder (SDAE) is composed of multiple DAEs, which have powerful feature extraction and data reconstruction capabilities. With the development of a deep neural network, the SDAE model has been widely used in industry and academia. Xu et al. [13] applied the SDAE model to predict the life of lithium-ion batteries, and the prediction results were better than SVM, BP, and RF. Dai et al. [14,15] constructed the SDAE cleaning model for the status data of power transmission and transformation equipment. This method can effectively identify the "dirty" data in the status data of power transmission and transformation equipment and realise data reconstruction.

At present, there are many studies on gas data processing in coal mines. Kozielski et al. [16] collected data from 28 different sensors placed at various locations around the coal mine and processed the data using LSTM. This data set can be used in a variety of analytical tasks, and the results are satisfactory. Ślęzak et al. [17] focused on feature extraction and feature selection, in the case of underground coal mine sensors, derivation of multivariate series of simple window-based statistics to deal with noisy and incomplete data sources, better reflect temporal drifts and correlations, and reliably describe real situations using higher-level data characteristics. However, there are few studies on cleaning methods of mine speed monitoring data in coal mines. Huang et al. [18] used the Laser Doppler Velocimetry system to obtain wind speed data and applied the adaptive Kalman Filter model to clean outliers of wind speed data. However, when the data is missing for a long time, the data after cleaning by this method has a large deviation from the original. Zhang et al. [19] proposed three cleaning methods for wind speed sensor data, including FCM, S-G, and Rloess, however, these three methods have different application scenarios and are not universal. Qu [20] applied the improved Laida criterion to determine the abnormal value of the wind speed sensor monitoring data, but this method cannot repair outliers.

Current data cleaning methods only delete or reduce "dirty" data, which destroys the continuity and integrity of data. In addition, they ignore the correlation between monitoring data, which is not conducive to data mining in fault diagnosis. As such, we propose to use association rules to mine state parameters with a strong correlation with ventilation monitoring data, and then build the SDAE model. The correlation between data series is used to further distinguish whether ventilation monitoring data belongs to the cleanable category. The cleanable data is reconstructed

through the SDAE model to correct the outliers and fill in the missing values. Compared with the traditional method of removing outliers, the proposed method can automatically repair the "dirty" data and retain the effective status information of the ventilation system, which provides a new concept for processing intelligent ventilation data on a large scale.

2. The proposed method

2.1. Association analysis

Association rules are mainly used to mine the association relation between data attributes. Taking the monitoring state parameter sequences X and Y as an example, the Apriori algorithm is used for correlation analysis. The specific process is as follows:

(1) The selected state parameter sequence is symbolised. The Apriori algorithm requires that the input data type be Boolean symbols, and the monitoring data needs to be converted. X and Y of length L_{data} are truncated into N subsequences using a sliding window of length L. Subsequences are represented by x and y. The least-square method is used to fit each subsequence, and the slope of the fitted equation is normalised so that it is distributed in the interval [-1,1]. Finally, the symbolic transformation is completed according to Table 1.

(2) The Apriori algorithm is used to find the frequent itemsets in the two sequences, and the itemsets greater than the minimum confidence are used as the association rule.

(3) If there are m rules that meet the association rules. Equation (1) is used to calculate the correlation and confidence between X and Y.

$$\begin{cases} P_{cr}(X \Rightarrow Y) = \sum_{i=1}^{m} P_{\sup}(x_i \Rightarrow y_i) \\ P_{cf}(X \Rightarrow Y) = \sum_{i}^{m} P_{cof}(x_i \Rightarrow y_i) P_{\sup}(x_i \Rightarrow y_i) \end{cases} \tag{1}$$

Where $P_{\sup}$ is the support of a single rule; P_{cof} is the confidence of a single rule; P_{cr} is the correlation between sequences; P_{cf} is the confidence between sequences.

TABLE 1

The rule for symbol conversion

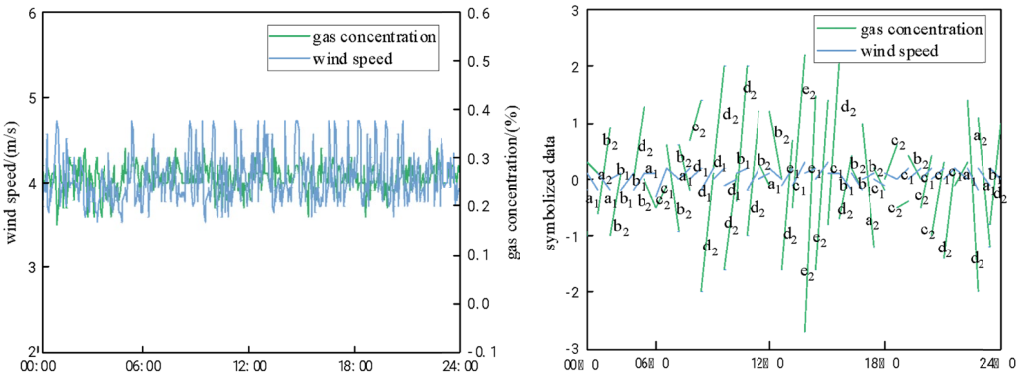
Interval	x	y
[–1,–0.6]	a_1	a_2
[–0.6,–0.2]	b_1	b_2
[–0.2,0.2]	c_1	c_2
[0.2,0.6]	d_1	d_2
[0.6,1]	e_1	e_2

For ventilation system monitoring data, the higher correlation and confidence of the two sequences are considered to have a strong correlation. According to previous experience and the data

characteristics, when considering the integrity and accuracy of association rules, the minimum correlation and confidence are set as 0.5 in this paper. When the calculated correlation and confidence of the two monitoring data series exceed the minimum threshold, it is considered that there is a strong correlation between the two data sequences. Analysing the relationship between the two data sequences should be done during the subsequent data processing step. Otherwise, it is considered that the two sequences are irrelevant.

2.2. Case of association analysis

Taking the Dongshan coal mine as an example, correlation analysis is made on the monitoring data of the wind speed sensor and gas sensor installed at the same monitoring point in the western mining area. Data collected from 00:00 to 24:00 on April 1 are selected as sample data. Where, $L_{data} = 1000$, sliding window $L = 25$. The wind speed data and gas data are respectively truncated into 40 subsequences. The original data is shown in Fig. 1(a), and the symbolised is shown in Fig. 1(b).



(a) (b)

Fig. 1. Monitoring original data and symbolised image

The results of association analysis using the Apriori algorithm are shown in Table 2.

TABLE 2

Results of association analysis

Frequent ItemSets	Support	Confidence
$c_1 \rightarrow c_2$	0.3400	0.7100
$d_1 \rightarrow d_2$	0.4100	0.8000

According to Table 2, it can be calculated from Equation (1) that the correlation between the wind speed and gas concentration is 0.75, and the confidence is 0.5694. The correlation and confidence both meet the minimum threshold rule, so it is considered that there is a strong correlation between the two monitoring data sequences at this monitoring point.

2.3. Realisation for data cleaning

2.3.1. SDAE algorithm

The SDAE model is a deep neural network structure, by adding noise to the sample data, enhances the robustness of the model by learning damaged data, and finally, the input data are restored through data reconstruction. The SDAE structure includes the input layer, hidden layer, and output layer. The hidden layer is made up of multiple DAEs. Each DAE unit consists of an encoder and a decoder. The SDAE structure is shown in Fig. 2. Where f_θ means encoding, g_θ means decoding, and q_D means randomly mapping part of the value of the input data x to 0 (stochastic mapping).

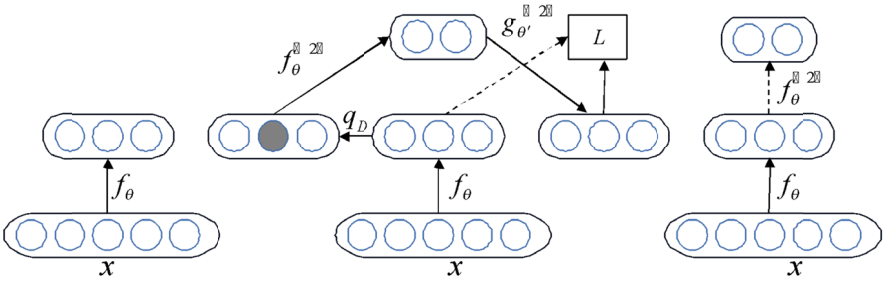


Fig. 2. Structure of SDAE

Taking the wind speed sensor monitoring data set x as input samples, $\tilde{x}$ is obtained by adding random noise to x. After encoding, the hidden layer feature y of $\tilde{x}$ can be expressed as Equation (2).

$$y = f_\theta(\tilde{x}) = s(W\tilde{x} + b) \tag{2}$$

Where, W and b are the weight matrix and bias vector of the encoding model, respectively; (W, b) updated by stochastic gradient descent algorithm (SGD); s is the active function, and the sigmoid function is selected in this paper. The expression of the s function is as Equation (3).

$$s(u) = 1/(1 + \exp(-u)) \tag{3}$$

After decoding, the reconstruction characteristic of y is expressed as Equation (4).

$$z = g_{\theta'}(y) = s(W'y + b') \tag{4}$$

Where z is the reconstruction value; W' and b' are the weight matrix and bias vector of the decoding model.

The reconstruction value z is not completely consistent with the input data x. The reconstruction error is used to characterise the training effect, which is calculated as Equation (5). The smaller the reconstruction error is, the higher the approximation degree between z and x is.

$$L(x, z) = \frac{1}{2}\sum \|x - z\|_2^2 \tag{5}$$

The SDAE network can not only reproduce the input results but also achieve noise reduction. This paper uses its characteristics to realise the cleaning and reconstruction of senior monitoring data of the mine ventilation system. In the training stage, the monitoring data of the mine ventilation system is taken as input, and the minimum reconstruction error is taken as the tuning criterion. In the testing phase, for the cleanable data, the SDAE reconstructed data is taken as the modified data. The SDAE cleaning algorithm of mine ventilation system monitoring data is described as follows (Algorithm 1).

Algorithm 1: The SDAE training of mine ventilation system monitoring data

Input: monitor data set x, $x_i \in R$
Output: Parameters θ, θ' and reconstructed data set z

Data preprocessing: $x \rightarrow \bar{x}$
Initialisation: number of network structure layers L, number of iterations K, denoising rate α, learning rate β, weight matrix W and W', bias vector b and b', number of neurons in visible layer and hidden layer, weight-decay λ
Stochastic mapping: $\alpha\bar{x} \rightarrow \tilde{x}$

for $j = 1$ to K

for $i = 1$ to L

$$y_i = f_\theta(\tilde{x}) = s(W_i\tilde{x} + b_i)$$
$$z_i = g_{\theta'}(y) = s(W_i'\tilde{x} + b_i')$$
$$\tilde{x} = y_i$$

end

end

for j =1 to K

$$\delta_L = -(\nabla y_{L-1} L(x,z)) \otimes f'(W_L\tilde{x} + b_L)$$

for $i = L$–1 to 1

$$\delta_i = f'(\tilde{x}) \otimes (W_{i+1}^T(y - (W_i\tilde{x} + b_i)))$$

end

for $i = 1$ to L

$$\nabla_{w_i} = \delta_i(f(\tilde{x}))^T, \nabla_{b_i} = \delta_i$$
$$W_i \leftarrow W_i - \alpha\nabla_{w_i} - \lambda W_i, b_i \leftarrow b_i - \alpha\nabla_{b_i}$$

end

end

2.3.2. Cleaning process

When the mine ventilation system runs normally, the monitoring data is usually distributed near the one-dimensional manifold (blue dot in Fig. 3). When abnormal or environmental disturbances cause considerable measurement errors or data loss, isolated singularities deviate from expected values in monitoring data, and these noise points or missing values will deviate from the manifold distribution of normal data (red dots in Fig. 3). When the normal monitoring data of the ventilation system is used for SDAE model training, part of the data will be randomly damaged to form noise staining data, which is similar to the red dot in Fig. 3. For randomly added noise data

subsets, the SDAE model extracts distribution features of undamaged data through continuous learning and then restores noise data through the prediction of undamaged data so that it has the ability to meet the probability distribution of training samples. The SDAE model can map deviations from data distribution to manifolds that meet expectations. Compared with the reconstruction errors of normally distributed monitoring data, the SDAE model can satisfy the equation.

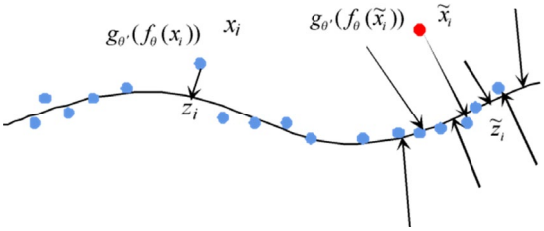


Fig. 3. Geometric description of data cleaning principle

The fault of the mine ventilation system is usually the change of trend in the monitoring status parameters, such as the abnormal opening and closing of the damper and wind window, and the monitoring data of the wind speed sensor will show a continuous upward or downward trend. The SDAE model is trained by using the monitoring data during the normal operation of the ventilation system. When the ventilation system fails and the trend of state parameters changes, the parameters and values of the trained SDAE model cannot meet the mapping relationship of the ventilation system fault state data. Therefore, when the fault samples are tested, there will be large data reconstruction errors with a long error duration. The steps of data cleaning are as follows. The flowchart is shown in Fig. 4.

(1) The sensor monitoring data during the normal operation of the mine ventilation system are obtained as the training sample is set and normalised. According to Algorithm 1, the SDAE model is trained with the training sample set as input, and the model parameter θ and θ' are determined.

(2) The reconstruction errors of the training sample set are calculated according to Equation (5), and the upper limit T_{hd} of reconstruction errors is determined by kernel density estimation (KDE) [21]. The error tolerance time T_W is determined by analysing the historical fault data.

(3) The test samples are input into the SDAE model to obtain the reconstruction errors R_e and error duration E_t. Combining the SDAE reconstruction errors corresponding to the strong correlation monitoring series in the same period, the data types are determined as follows:

A. $R_e \leq T_{hd}$: The test data does not contain "dirty" data, which are generated during the normal operation of the ventilation system, so there is no need to clean.

B. $R_e > T_{hd}$, $E_t \leq T_w$ and $R_e \leq T_{hd}$ of the strong correlation data series in the same period: The "dirty" data in the test set are outliers, which are generated during the normal operation of the ventilation system, so there is a need to clean.

C. $R_e > T_{hd}$, $E_t \leq T_w$ and $R_e > T_{hd}$ of the strong correlation data series in the same period: The test data is generated when the ventilation system fails, which can provide effective information for roadway fault diagnosis, so there is no need to clean.

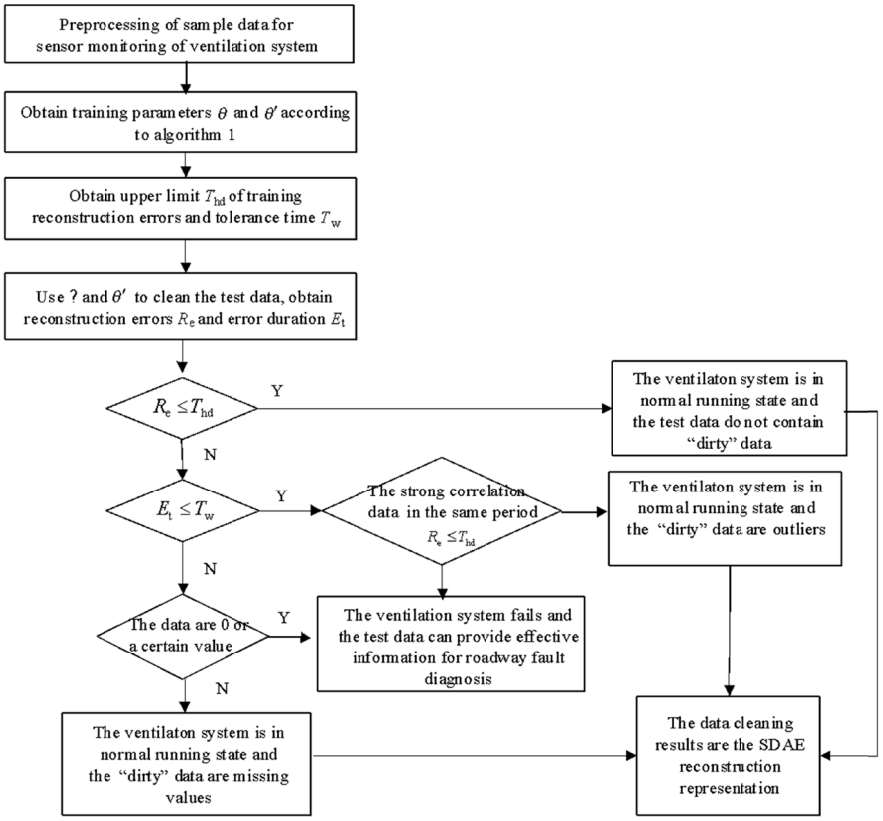


Fig. 4. The flowchart of data cleaning

D. $R_e > T_{hd}$, $E_t > T_w$, and this part of the data is 0 or a certain value: The "dirty" data in the test set are missing values.

E. $R_e > T_{hd}$, $E_t > T_w$ and $R_e > T_{hd}$, $E_t > T_w$ of the strong correlation data series in the same period: The test data is generated when the ventilation system fails, which can provide effective information for roadway fault diagnosis, so there is no need to clean.

(4) The SDAE model repairs the outliers and missing values through data reconstruction, and the non-destructive data and the repaired data constitute the effective information of the ventilation system monitoring data.

2.4. Performance evaluation

In this paper, mean absolute error(MAE) and root mean square error(RMSE) are used to evaluate the performance of the SDAE model and are calculated as Equation (6) and Equation (7).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - z_i\right| \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - z_i)^2} \tag{7}$$

Where n is the total number of input data.

3. Experimental analysis

In this paper, an SDAE model is constructed by taking the monitoring data of the wind speed sensor installed in the Dongshan Coal Mine to verify the effectiveness of the proposed method. All wind speed sensor data collected from April 1st to April 7th are used as training samples X_{train}. The monitoring data of the wind speed sensor 84G09F12A in the #1 wind roadway in the west mining area was collected on April 8th is taken as the test sample X_{test}, which is generated from the normal running state of the ventilation system. On April 9th, on the premise of ensuring product safety, the damper of the 1# return airway was opened, and the data generated by the 84G09F12A wind speed sensor are used as the ventilation system fault test sample X_{fault}. The SDAE model parameters are set as follows: There are 15 wind speed sensors installed in the whole mine, and the input data is 15-dimensional. Therefore, the number of neurons in the input layer is 15, and the number of neurons in the output layer is 1. There are 3 hidden layers, and the structure is {20, 10, 20}. The connection weight is set to a random number obeying the normal distribution N(0, 0.01), and the bias term is initialised to 0. The weight-decay is 0.001. The experimental environment is Windows 10 operating system, the processor is Intel Core i7, 16 GB RAM, and the software is MatlabR2017b.

3.1. Determination of key hyperparameters

The selection of iteration numbers, de-noising rate, and learning rate directly affect the performance of the SDAE model. Therefore, the above three key hyperparameters are studied and the optimal combination parameters are selected. X_{train} is randomly divided into a training set and a validation set in a ratio of 8:2. The experiment is repeated twice, using different training sets and validation sets, denoted as {Training1,Validation1}, {Training2,Validation2}, respectively.

To investigate the effect of different iteration numbers, we set the de-noising rate to 0.1 and the learning rate to 0.01 and 0.05. The results of MAE and RMSE varied as iteration numbers are shown in Fig. 5.

It can be seen from Fig. 5 that the general trends of MAE and RMSE varied, with iteration numbers being the same. When the learning rate is 0.01, the MAE and RMSE are relatively stable between the 300th to 1100th iteration. At the 1200th iteration, the MAE and RMSE increase significantly. At the 1400th iteration, they decrease to the minimum value and then increase again. At the 1200th iteration, the model updates the network parameters and reaches a local optimum. When the learning rate is 0.05, the MAE and RMSE also show the same trend. Therefore, considering the model accuracy and running time comprehensively, the number of iterations is 1400.

The learning rate determines whether the model can converge to the global optimum. We set the de-noising rate to 0.1 and the number of iterations to 1400. The results of MAE and RMSE varied as the learning rate is shown in Fig. 6.

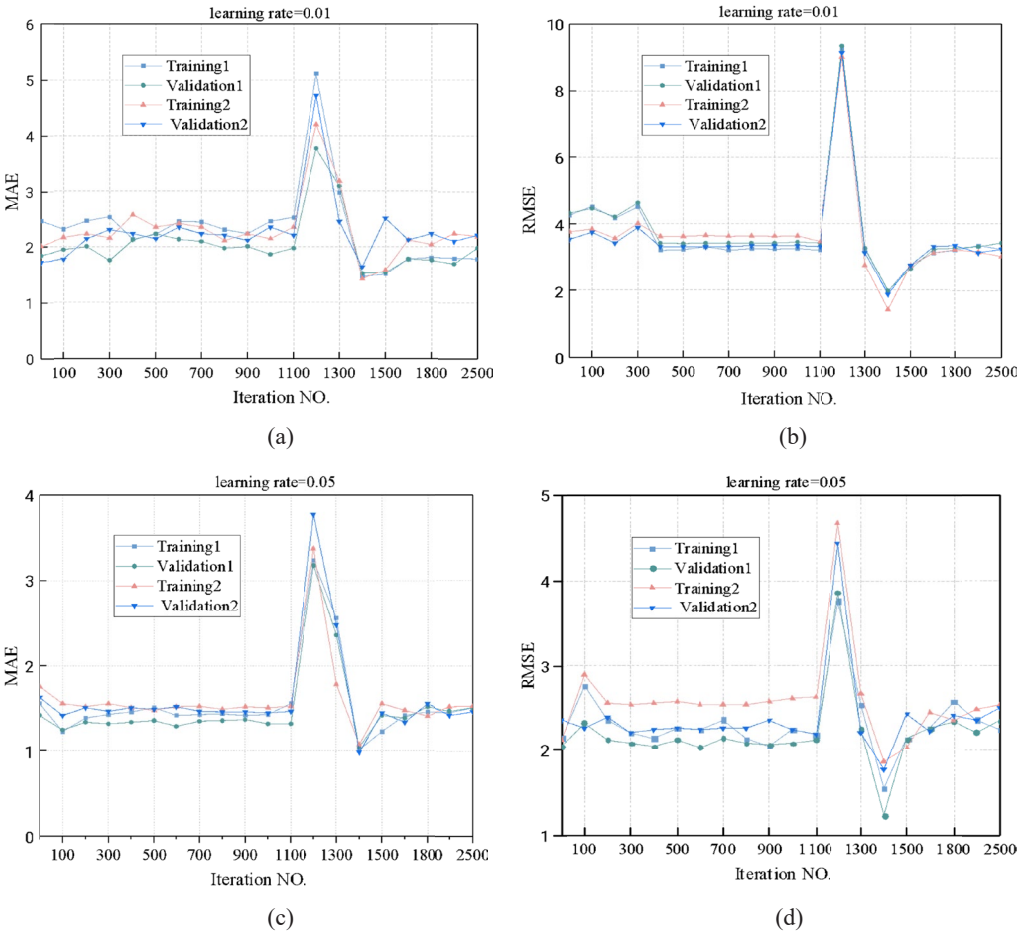
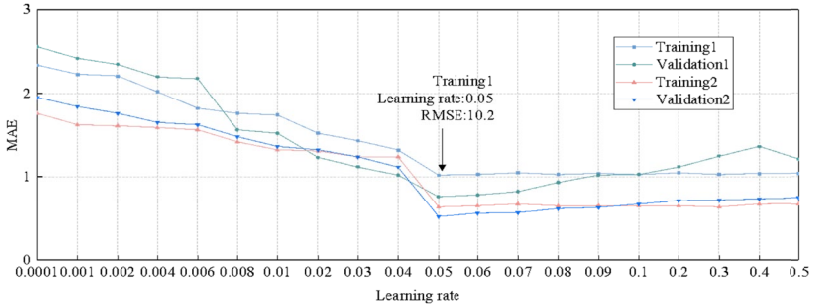


Fig. 5. Results of MAE and RMSE varied with the iteration number

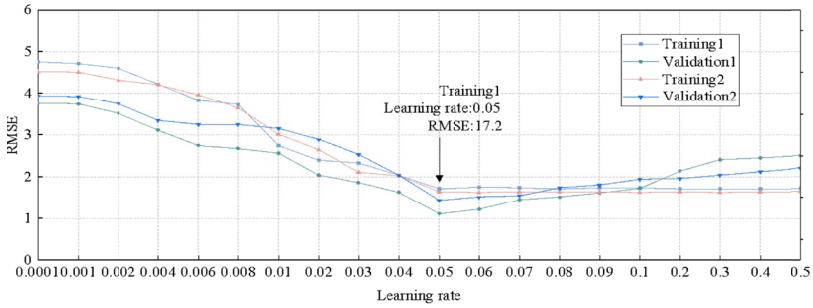
When the learning rate is between 0.001 and 0.05, the MAE and RMSE decrease gradually. When the learning rate exceeds 0.05, the MAE and RMSE results of the training set tend to be stable, but the MAE and RMSE results of the validation set show an upward trend. This means that when the learning rate exceeds 0.05, the prediction accuracy of the test set will decrease. Therefore, considering the model prediction accuracy, the learning rate is determined to be 0.05.

It can be seen from Section 1.3.1 that the SDAE model randomly adds noise in the initial stage of training. In this paper, the learning rate is set to 0.005, and the number of iterations is set to 1400 to study the effect of different de-noising rates on the performance of the model. The results of MAE and RMSE varied with the de-noising rate are shown in Fig. 7.

With the change in the de-noising rate, there are only small fluctuations in MAE and RMSE, which indicates that the model cleaning results are not affected by the de-noising rate. Therefore, the de-noising rate is set to 0.1.

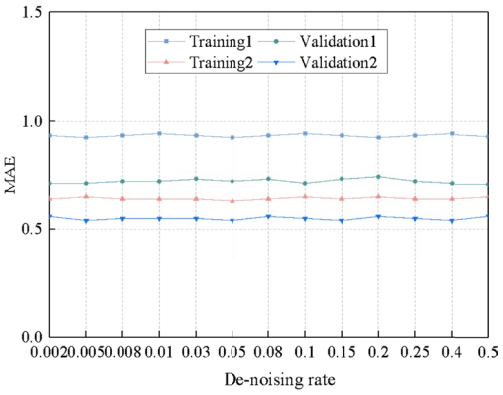


(a)

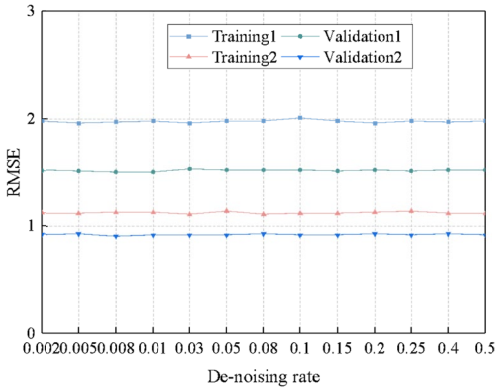


(b)

Fig. 6. Results of MAE and RMSE varied with learning rate



(c)



(d)

Fig. 7. Results of MAE and RMSE varied as the de-noising rate

3.2. Results of experimental

The X_{train} is normalised, and the SDAE model is constructed according to algorithm 1. After training, the MAE of the sample data is 5.52%. It can be considered that 94.48% of the sample

data can be repaired, and 5.52% of the sample data cannot be accurately repaired. According to Step 2 in Section 2.3.2, the KDE method is used to obtain the cumulative probability distribution of model reconstruction errors, as shown in Fig. 8. Meanwhile, due to the background noise, the accuracy of the model is affected. The confidence level is improved to avoid the background noise being judged as abnormal. Therefore, the confidence interval is set as 0.96. As shown in Fig. 8, when the confidence level is 0.96, the upper limit of reconstruction error T_{hd} is 0.01929. The error tolerance time (T_w) is set to 5.

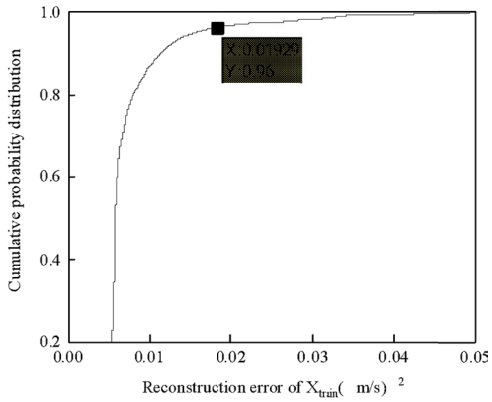


Fig. 8. The distribution of reconstruction errors

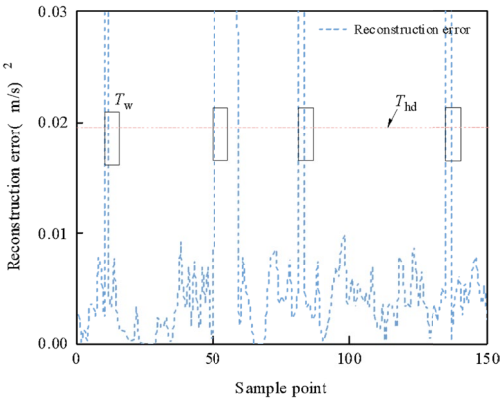


Fig. 9. Relationship between reconstruction error about X_{test} and T_{hd}, T_w

3.2.1. Cleaning data under a normal running state of the ventilation system

For the convenience of the display, some sample points are intercepted. The 11th and 51th-58th sample points are removed, and outliers are added to the 82nd and 136th sample points in Xtest. The normalised Xtest is entered into the trained SDAE model. To overcome the randomness of the algorithm, we repeated 10 experiments. The average reconstruction errors are shown in Fig. 9.

As can be seen from Fig. 9, the reconstruction errors at the 11th, 51st-58th, 82nd, and 136th sample points exceed the upper limit of reconstruction errors Thd, and these data points are "dirty" data. According to Section 2.2, there is a strong correlation between gas data and wind speed data at this monitoring point. Therefore, the reconstruction errors of gas data are analysed in collaboration to determine the types of sample data. At data points 11th, 51th-58th, 82nd, and 136th, the reconstruction errors of gas data do not exceed the error limit by the corresponding SDAE model. According to Step 3 of Section 2.3.2, the "dirty" data type is identified and corrected. The results are shown in Table 3. Where the cleaning value is the average of 10 experiments. The MAE and RMSE are the mean values after denormalisation.

To verify the superiority of the SDAE cleaning model, the LSTM, SAE and adaptive Kalman Filter proposed in Reference [19] are included in the comparative experiment to predict the cleaning value. The experiments were repeated 10 times, and performance indexes are shown in Fig. 10.

TABLE 3

Data cleaning result of the SDAE model

Sample data	Normalised			Type	Denormalised	
	True	Outlier	Cleaned		MAE	RMSE
11	0.5629	0	0.5581	missing data	5.6%	1.87%
51	0.5154	0	0.5258	missing data		
52	0.5612	0	0.5613	missing data		
53	0.5138	0	0.5025	missing data		
54	0.5412	0	0.5124	missing data		
55	0.5961	0	0.5725	missing data		
56	0.5742	0	0.5241	missing data		
57	0.5741	0	0.5842	missing data		
58	0.5325	0	0.5241	missing data		
82	0.5856	0.1625	0.5748	outlier		
136	0.5458	0.8158	0.5612	outlier		

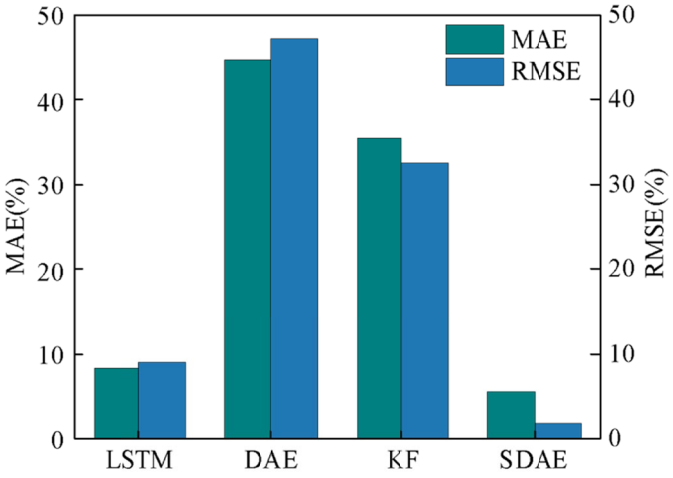


Fig. 10. Performance indicators of different algorithms

It can be analysed from Table 3 and Fig. 10:

(1) In terms of realisation, the SDAE model proposed in this paper is suitable for cleaning the wind speed sensor data of the mine ventilation system and can repair outliers and missing data. The MAE and RMSE are 5.6% and 1.87%, respectively.

(2) In terms of performance indicators, the SDAE model is lower than the three algorithms in both MAE and RMSE, with an average decrease of 75.42% and 74.98%, indicating that the SDAE model has more stable performance. Compared with the recently proposed adaptive Kalman Filter algorithm, the SDAE model is more suitable for the complex environment of coal mines.

3.2.2. Cleaning data under the ventilation system fault

The normalised X_{fault} is entered into the trained SDAE model. When the damper was opened after the 50th sample point, the wind speed showed an upward trend. The reconstruction errors of X_{fault} are shown in Fig. 11. The original data and cleaning results of X_{fault} are shown in Fig. 12.

From Fig. 11, the reconstruction error increases suddenly from the 50th sample point, exceeding the upper. The reconstruction errors of gas data are analysed collaboratively, and it is found that the reconstruction errors of gas data have the same trend as X_{fault}. According to Step 3 in Section 2.3.2, it is considered that the sample data at this time are generated when the ventilation system is at fault and does not need to be cleaned. It can be seen from Fig. 12 that the data cleaning of X_{fault} only repairs the outliers, while the roadway fault data are effectively preserved, which can provide effective information for the subsequent fault diagnosis of the ventilation system.

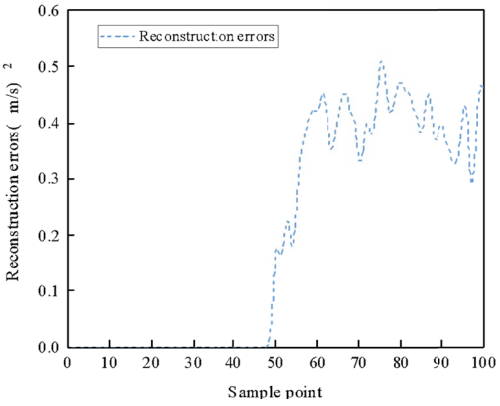


Fig. 11. The reconstruction error of X_{fault}

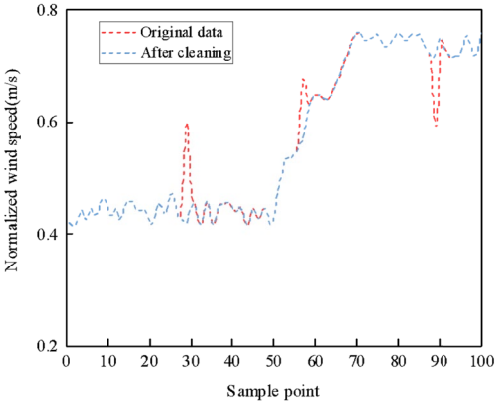


Fig. 12. Data cleaning results of X_{fault}

4. Conclusion

Based on the wind speed data of the mine ventilation system and the ability of the SDAE model to extract and restore "dirty" data, this paper proposes a cleaning method for mine ventilation system monitoring data based on SDAE. Association rules are used to determine the correlation between monitoring data series, and data types are determined in collaboration with the SDAE model. The following conclusions are drawn from the experimental analysis.

1. The SDAE model proposed is suitable for data cleaning of the mine ventilation system. The method can automatically identify outliers and missing values and repair "dirty" data. In addition, the proposed method can retain the effective status information of the ventilation system, which provides reliable data for subsequent ventilation system fault diagnosis and disaster warning.
2. The key hyperparameters of the model were determined, including a learning rate of 0.05, a denoising rate of 0.1, and an iteration number of 1400 for optimal SDAE cleaning. It provides a reference for the application of this model in other engineering fields.

3. Compared with LSTM, DAE, and KF, the SDAE model proposed in this paper has lower MAE and RMSE indexes, which are more suitable for complex mine environments. Its data reconstruction results are closer to the actual data and have a more stable performance.

References

[1] M.A. Semin, L.Y. Levin, Stability of air flows in mine ventilation networks. Process. Saf. Environmen. Prot. **124**, 167-171(2019). DOI: https://doi.org/10.1016/j.psep.2019.02.006

[2] J.W. Cheng, S.Q. Yang, Data mining applications in evaluating mine ventilation system. Safety. Sci. **50** (4), 918-955 (2012). DOI: https://doi.org/10.1016/j.ssci.2011.08.003

[3] G.F. Wang, Y.X. Xu, H.W. Ren, Intelligent and ecological coal mining as well as clean utilization technology in China: Review and prospects. J. Int. J. Mining. Sci. Tec. **29** (2), 161-169 (2019). DOI:https://doi.org/10.1016/j.ijmst.2018.06.005

[4] L. Muduli, D.P. Mishra, P.K. Jana, Application of wireless sensor network for environmental monitoring in underground coal mines: A systematic review. J. Netw. Comput. Appl. **106**, 48-67 (2018). DOI: https://doi.org/10.1016/j.jnca.2017.12.022

[5] W.H. Wang, K.L. Shen, B.B. Wang, Failure probability analysis of the urban buried gas pipelines using Bayesian networks. Process. Saf. Environmen. Prot. **111**, 678-686 (2017). DOI: https://doi.org/10.1016/j.psep.2017.08.040

[6] D. Huang, J. Liu, L.J. Deng, A hybrid-encoding adaptive evolutionary strategy algorithm for windage alteration fault diagnosis. Process. Saf. Environmen. Prot. **136**, 242-252 (2020). DOI: https://doi.org/10.1016/j.psep.2020.01.037

[7] Z.N. Gao, F. Yang, S.B. Hu, et al., Pseudo-fluctuation data cleaning for state estimation of new energy power system. High. Voltage. Eng. **48** (06), 2366-2377 (2022). DOI: https://doi.org/10.13336/j.1003-6520.hve.20210591

[8] Y.J. Yan, G.H. Sheng, Y.F. Chen, et al., Cleaning method for big data of power transmission and transformation equipment state based on time sequence analysis. High. Voltage. Eng. **39** (7), 138-144 (2015). DOI: https://doi.org/10.7500/AEPS20140111003

[9] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks. Science **313**, 504-507 (2006). DOI: https://doi.org/10.1126/science.1127647

[10] Y. Bengio, Learning deep architectures for AI. Found. Trends. Mac. Lear. **2** (1), 1-127 (2009). DOI: http://dx.doi.org/10.1561/2200000006

[11] P. Vicent, H. Larochlle, Y. Bengio, et al., Extracting and composing robust features with denoising autoencoders. C. //25th International Conference on Machine Learning, June 5-9. Helsinki, Finland: 1096-1103 (2008). DOI: https://doi.org/10.1145/1390156.1390294

[12] P. Vicent, H. Larochlle, I. Lajoie, et al., Stacked denoising antoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11** (12), 3371-3408 (2010).

[13] F. Xu, F.F. Yang, Z.C. Fei, et al., Life prediction of lithium-ion batteries based on stacked denoising autoencoders. Reliab. Eng. Sys. Safe. **208**: 107396 (2021). DOI: https://doi.org/10.1016/j.ress.2020.107396

[14] J.J. Dai, H. Song, G.H. Sheng, et al., Cleaning method for status monitoring data of power equipment based on stacked denoising autoencoders. J. Ieee Access. **5**, 22863-22870 (2017). DOI: https://doi.org/10.1109/ACCESS.2017.2740968

[15] J.J. Dai, H. Song, Cleaning method for status data of power transmission and transformation equipment base on tacked denoising autoencoders. Automation of Electric Power Systems **41** (12), 224-230 (2017). DOI: https://doi.org/10.7500/AEPS2016201003

[16] M. Kozielski, M. Sikora, Ł. Wróbel, Data on methane concentration collected by underground coal mine sensors. Data in Brief. **39**, 107457 (2021). DOI: https://doi.org/10.1016/j.dib.2021.107457

[17] D. Ślęzak, M. Grzegorowski, A. Januszet, et al., A framework for learning and embedding multi-sensor forecasting models into a decision support system: A case study of methane concentration in coal mines. Inform. Sciences **451**, 112-133 (2018). DOI: https://doi.org/10.1016/j.ins.2018.04.026

[18] D. Huang, J. Liu, L.J. Deng, et al., An adaptive Kalman filter for online monitoring of mine wind speed. Arch. Min. Sci. **64** (4), 813-827 (2019). DOI: https://doi.org/10.24425/ams.2019.131068

[19] W. Zhang, Y.C. Li, H, Zhang, et al., Comparison of structured data noise reduction methods for airflow speed sensor of intelligent ventilation. Journal of Safety Science and Technology **17** (08). 70-76 (2021). DOI: https://doi.org/10.11731/j.issn.1673-193x.2021.08.011

[20] S.J. Qu, Real-time date processing method of wind speed sensor in roadway. Safety in Coal Mines **48** (02), 163-166 (2017). DOI: https://doi.org/10.13347/j.cnki.mkaq.2017.02.044

[21] S. Węglarczyk, Kernel density estimation and its application [C]//ITM Web of Conferences. EDP Sciences **23**, 00037 (2018). DOI: https://doi.org/10.1051/itmconf/20182300037