Comparison of Missing Data Imputation Methods for Leaching Process Modelling

D.K. He, T.S. Chu, Y.B. Lang, G.X. Sun
Key Laboratory of Process Automation Industry, Ministry of Education
North eastern University
Shenyang, China

***Abstract*—As the original information of production process, industrial production data is the important basis and foundation of process modelling and optimization. However, the data acquisition operation, the restriction of instrument operation environment and malfunction often lead to data missing. Under this condition, the research on missing data imputation in the leaching process is vital and significant. In this paper, the leaching process mechanism model is presented firstly. Missing data characteristics, the basic principle of imputation methods are introduced in detail next. Based on the analysis of data deficiency and its features during the acid intermittent leaching of cobalt compound ore, this article will launch the research on the deficiency of crucial values, such as sulphur dioxide flow, PH value of leaching agent, leaching rate, and apply various data packing methods into leaching process modelling. According to the simulation results, this paper evaluates the application performance of different imputation and modelling methods in accuracy and concludes the method with which could pack the missing data effectively under different data missing condition.**

Keywords-missing data imputation; leaching process; modelling; multiple imputation (MI); gaussian mixture model (GMM)

I. INTRODUCTION

As a step of metal extracting, hydrometallurgical leaching process directly affects the metal recovery rate [1]. The production date of leaching process is particularly important to construct a yield-predicted model based on data [2-4]. However, leaching process is implemented in harsh environment with strong acid, high temperature and so on [5]. A few important process variables such as sulphur dioxide flow and PH value may be missing or abnormal due to equipment or human reasons. When the sample size is limited, deletion of abnormal data will lead to waste of system resource. In order to adapt to low quality and absence of data, the study on data imputation method based on the information of observed data have important theoretical significance.

As so far, several approaches for the imputation of missing data have been proposed. Rubin et al [6] proposed the Expectation Maximization (EM) algorithm which is an effective tool for missing data estimation. In order to make up for the defects of single fill methods, Rubin [7] proposed Multiple Imputation (MI) method in 1978.

In this paper, 4 data fill methods which are BP neural network (BP-NN), least squares support vector machine (LS-SVM), Gaussian mixture model based on expectation maximization (GMM-EM) and multiple imputation (MI), are applied into leaching process. Compare the performance of different methods and conclude the better method in different situation.

II. LEACHING MECHANISM AND TECHNOLOGY

Due to excellent characteristics such as low pollution and high efficiency, hydrometallurgy received extensive attention. Hydrometallurgy is a chemical metallurgy. The metal is extracted from ore through the chemical and physical reaction between ore and aqueous medium. Generally, hydrometallurgy consists of four major procedures which are pre-treatment, leaching process, solid-liquid separation and extraction process. This is shown in Figure 1. The product of leaching process is leachate. The quality of leachate directly determines the purity and quality of the metal that is extracted in the subsequent step. Therefore, the study on leaching process is particularly important.

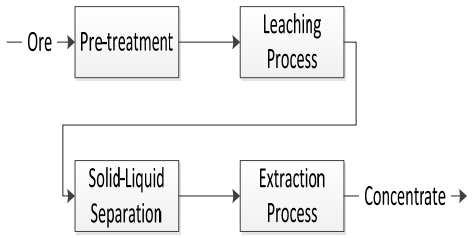


FIGURE Ⅰ. BASIC PROCEDURES OF HYDROMETALLURGY PROCESS.

Acidic intermittent leaching of cobalt compound ore is discussed in this paper. This process proceeds in an atmospheric and heated leaching tank. There is a screw propeller keeping stirring the ore slurry in order to make the slurry uniform. During the cobalt leaching process, temperature, PH value and sulphur dioxide flow can be detected online while leaching yield is tested offline. At the very beginning, ores are milled into mineral powder and mixed with water solvent. Before the production, slurry is injected into the leaching tank. Then sulphur dioxide and sulphuric acid are pumped into the leachate. Heat and stir the leachate during the reaction in order to make more useful metal components immersed in the solution and separate the metal and gangue effectively. Until the end of leaching, the qualified leachate is discharged and sent to next step.

In this paper, the mechanism model of leaching process proposed by Hu [8] is taken as the physical device which generated the production data.

III. DATA FILLING METHOD

Generally, data filling divides into single imputation and multiple imputations. Single imputation is a method that constructs one alternative value to replace the miss data. The other way is multiple imputations which constructs m values for each missing data. Then take the mean of m values as the alternative value. Four methods referred to this paper will be introduced in detail.

A. BP Neural Network

In 1956, McClelland and Rumelhart proposed multilayer feed-forward neural network which is the basis of back propagation neural network (BP-NN) [9]. The structure of BP-NN has three or more layers including input layer, intermediate layer and output layer. When the input signal is supplied to the network, the output value is calculated through neuronal function of each layer. Then the threshold and weight of each neuronal function are modified according to the error between network output and expected output. And repeat this process until the error convergent.

B. Least Squares Support Vector Machine

Least squares support vector machine (LS-SVM) is an extension of standard support machine (SVM) which substitutes equality constraints for inequality constraints in SVM [10]. For linear problems, suppose the training set is D where

$$D = \{(\mathrm{x}_i, \mathrm{y}_i) \mid \mathrm{i} = 1, 2, \cdots, n\} \tag{1}$$

Here, x_i denotes input and $x_i \in R^m$. y_i denotes the output and $y_i \in R$. LS-SVM can generate a function to separate the training set where the function is

$$f(x) = \omega \bullet x + b \tag{2}$$

Here, ω denotes hyperplane normal vector and b denotes bias.

C. Expectation Maximization Algorithm

Expectation maximization (EM) is an iterative algorithm for solving maximum likelihood estimation problem concerned hidden variables. The main idea of this algorithm is to calculate the values of hidden variables according to the initial parameters and observed variables. Then update the initial parameters according to the complete data set that has been calculated. Repeat the two steps until convergence.

D. Gaussian Mixture Model Based on Em

As long as there are enough Gaussian components, Gaussian mixture model could fit any continuous probability density function. Therefore Gaussian mixture model can be used to fit the non-normal distribution of industrial process data. Gaussian mixture model consists of multiple Gaussian components and its probability density function is depicted as following [11].

$$\begin{aligned} P(x \mid \pi_k, \mu_k, \Sigma_k) &= \sum_{k=1}^{K} \pi_k N(\mu_k, \Sigma_k) \\ N(\mu_k, \Sigma_k) &= \frac{e^{\left[-1/2(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right]}}{(2\pi)^{m/2} \left|\Sigma_k\right|^{1/2}} \\ \sum_{k=1}^{K} \pi_k &= 1 \cap \left(\pi_k \succ 0, \forall k \in \{1, 2, \cdots, K\}\right) \end{aligned} \tag{3}$$

Here, x denotes m-dimensional input data. The mean and covariance of each Gaussian distribution are denoted by μ_k and Σ_k. Here, π_k denotes the weight of each Gaussian distribution and indicates the probability that x belong to the k-th Gaussian distribution. The letter k means the number of Gaussian components in Gaussian mixture model and determined by Bayesian information criterion (BIC).

After the establishment of GMM, the EM algorithm is applied to calculate the missing data.

E. Multiple Imputation

Multiple imputation (MI) is first proposed by Rubin [6]. The main idea is that m values are constructed and mean of these values is taken the value to be filled. Generally, MI contains two main steps which are depicted as following.

The first step is imputation. The current parameters are estimated to be $\theta^{(s)}$ and estimation of missing data is calculated according to observed variables $y_{obs,\mathrm{i}}$ and the parameters $\theta^{(s)}$. After imputation, the current full data set is $Y^{(s+1)} = (Y_{obs}, Y_{mis}^{(s+1)})$.

The second step is posteriority. Update the parameters $\theta^{(s)}$ according to the complete data set that just been filled. The new parameters $\theta^{(s+1)}$ are applied to the next imputation. The detail is described in Rubin's paper.

IV. SIMULATION AND ANALYSIS

In this section, data types in leaching process and missing pattern are introduced briefly. The original data set is generated by mechanism model and the corresponding values are deleted according to different missing patterns. The four method described before are applied to fill the incomplete data set. Analysis is given to each result. Then the full data set after fill is applied to construct a predicted model based on data.

A. Data Types in Leaching Process and Missing Pattern

As the introduction to leaching process before, leaching yield in the acid intermittent leaching process of cobalt compound ore is determined by five factors which are temperature, sulphur dioxide flow, sulphuric acid concentration, granularity of ore and mineral components. In practical production, each batch lasts 15 hours and the temperature can be regarded as a constant. Though sulphuric acid concentration, mineral granularity and components influence the leaching yield badly, they cannot be detected online. So the PH value of leachate which can be detected online is substituted to represent the information that the three factors denotes. The leaching yield is the main production

quota and can be detected 1 hour each time offline. Therefore, the yield at one point is determined by the sulphur dioxide flow and PH value at this point and the yield at last point.

$$y(k) = f(\mathrm{SO}_2(k), \mathrm{PH}(k), y(k-1)) \qquad (4)$$

In this experiment, there are 6 leaching processes. Each leaching process lasts 15 hours totally and the variables are detected every hour. Therefore, 270 values are recorded in sum. Continuous deficiency of single variable is considered only. As the scale of data deficiency also influence the final results, the pattern is divided into three cases according to missing rate namely 10%, 20% and 30%. The detail of missing patterns and missing rate are depicted in Table 1.

TABLE Ⅰ.SINGLE PATTERNS OF MISSING DATA.

	Single SO_2	Single PH	Single Yield
10%	27	27	27
20%	54	54	54
30%	81	81	81

B. Simulation Of Imputation And Analysis

The variant methods introduced before are applied to fill missing data of different patterns. The packing error is determined by root mean square error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad (5)$$

Since different variables have different dimensions, the errors are normalised in the analysis.

$$y_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} \qquad (6)$$

(a) 10% deficient rate

When the deficient rate is 10%, the missing data of each pattern is filled by four methods. The results are depicted in Table 2. The performance of MI is better. GMM's advantages are not reflected when the scale of missing data is small.

TABLE Ⅱ.THE NORMALIZATION OF RMSE OF 10% MISSING DATA.

	BP	LS-SVM	MI	GMM-EM
Single SO_2	1	0.277128	0.097814	0
Single PH	0.876704	0.811512	0	1
Single Yield	1	0.451799	0	0.125297
Overall	0.958901	0.513480	0.032605	0.375099

According to results, the performance of each method to fill the deficient PH values is depicted in Figure 2.

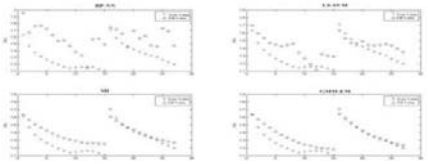


FIGURE Ⅱ.RESULT OF PACKING 10% MISSING DATA.

(b) 20% deficient rate

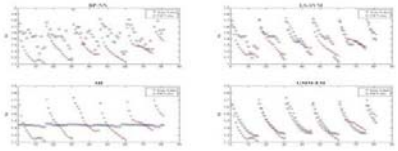
As the deficient rate rises up 20%, the performance of MI is still relatively ideal. The overall results are quite similar to the results before.

(c) 30% deficient rate

As the deficient rate become 30%, MI performs badly to fill the deficiency of single dimensional data while GMM-EM's performance is ideal for packing the deficiency of single dimensional data. The detail is depicted by Table 3 and the performance of packing missing data of Single PH is depicted by Figure 3.

TABLE Ⅲ.THE NORMALIZATION OF RMSE OF 30%SSING DATA.

	BP	LS-SVM	MI	GMM-EM
Single SO_2	1	0.266622	0.584934	0
Single PH	1	0.734341	0.709356	0
Single Yield	0.775331	0.801229	1	0
Overall	0.925110	0.600731	0.764763	0



FIGUREⅢ.RESULT OF PACKING 30% MISSING DATA.

C. Simulation of Modelling and Analysis

The complete data sets after imputation before are applied to construct predicted models for leaching process. LS-SVM and BP neural network are used as the method to construct model based on data. The results are depicted as following.

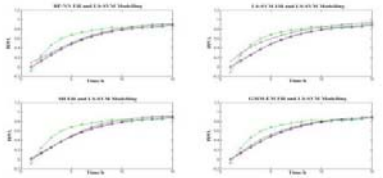
When the deficient rate is 10%, the performance of BP-NN modelling is better than LS-SVM modelling apparently. For single missing patterns, the combination of GMM-EMM packing and LS-SVM modelling is better. As the deficient rate rises up to moderate (20%), the combination of GMM-EM and LS-SVM modelling performs pretty well for single missing patterns. The detail is depicted by Table 4, Table 5 and Figure 4.As the deficient rate rises up to severe (30%), a good result can be obtained by the combination of LS-SVM and LS-SVM modelling for single missing patterns.

TABLE Ⅳ.THE NORMALIZATION OF RMSE OF MODERATE MISSING DATA MODELLING.

	BP-NN Modelling			
	BP	LS-SVM	MI	GMM-EM
Single SO_2	0.515532	0.464991	1	0.007634
Single PH	0.184712	0.108927	0.182009	1
Single Yield	1	0.071286	0.297081	0.333444
Overall	3.562852	1.887564	1.574011	2.143344

TABLE Ⅴ.THE NORMALIZATION OF RMSE OF MODERATE MISSING DATA MODELLING.

	LS-SVM Modelling			
	BP	LS-SVM	MI	GMM-EM
Single SO_2	0.049917	0.24033	0.024346	0
Single PH	0.000904	0.027534	0.031329	0
Single Yield	0.020182	0.036765	0.000599	0
Overall	0.108644	0.311301	0.058222	0.010767



FIGUREⅣ.LS-SVM MODELLING RESULT OF PACKING MODERATE MISSING DATA.

V. CONCLUSIONS

In this paper, the data imputation problem in leaching process is investigated. Totally four methods were applied to fill the values of single missing pattern. After the simulation, here are our conclusions.

For single patterns, the performance of MI is brilliant and the combination of MI and LS-SVM modelling works well if the deficient rate is low. As the deficient rate rise up to moderate, GMM-EM fills the missing data most effectively and the combination of GMM-EM and LS-SVM modelling is better than others. When the deficient rate becomes severe, GMM-EM is still the best method to fill the missing data whereas the LS-SVM modelling based on LS-SVM fill predicts the values more accurately.

ACKNOWLEDGEMENTS

This work is supported by the National Nature Science Foundation of Chinaunder Grant 61374147 and 61004083 and 973 Project No. 2009CB320601 and863 Project No. 2011AA060204 and the Fundamental Research Funds for theCentral Universities N120404014.

REFERENCES

[1] J. Fan, L. Yang, Control system of cobalt extracting, Automation in Petrochemical Industry, 2004, 6: 51-54.

[2] E. Li, J. Li, J.Yu, A genetic neural fuzzy system and its application in quality prediction inthe injection process. ChemEngCommun. 2004: 191: 335-355.

[3] N.Lu,F.Gao, Stage-based process analysis and quality prediction for batch processes. IndEngChem Res. 2005;44:3547-3555.

[4] D. Lee, J.H. Song, Weighted support vector machine for quality estimation in the polymerization process. IndEngChem Res. 2005: 44: 2101-2105.

[5] R.J. Ma, The new development of hydrometallurgy. Hydrometallurgy,2007, 26(1): 1-12.

[6] Rubin, D.B., Inference and missing data. Biometrika, 1976, 63(3): 581-592.

[7] Rubin, D.B., Multiple imputations: a primer. Statistical Methods in Medical Research, 1999, 8(1): 3-15.

[8] G.H. Hu, The modelling and optimization of leaching process, North-eastern University, 2011.

[9] HoPfield, J., Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of sciences, 1982, 79: 2544-2558.

[10] Y.N. Liu, S.L. Pang, J. Li, Mid-long term load forecasting based on rough set and LS-SVM. Electric Power, 2007, 40(10): 42-44.

[11] Bishop, C.M., Pattern recognition and machine learning. Springer, 2006, 1(1): 2-5.