Contents lists available at ScienceDirect

Chemical Engineering Research and Design

journal homepage: www.elsevier.com/locate/cherd

Study on missing data imputation and modeling for the leaching process

Dakuo He[a,b,*], Zhengsong Wang[a], Le Yang[a], Wanwan Dai[c]

[a] College of Information Science and Engineering, Northeastern University, Shenyang 110004, Liaoning, China
[b] State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110004, China
[c] School of Automation, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

ARTICLE INFO

Article history:
Received 20 December 2016
Received in revised form 24 May 2017
Accepted 26 May 2017
Available online 4 June 2017

Keywords:
Leaching process
Data imputation
Expectation maximization imputation
Gaussian mixture model
Multiple imputation

ABSTRACT

The leaching process is an important component in hydrometallurgy. A predictive model of the leaching rate lays the foundation for soft measurement and process optimization, and data collection is the key in such a modeling effort. However, because of the complexity and harshness of leaching process, data can only be collected sparsely, which results in data deficiency in the modeling process. Therefore, data imputation before modeling seems to be extremely significant. In this paper, expectation maximization imputation based on the Gaussian mixture model (GMM-EM) and multiple imputation (MI) are respectively applied to perform missing data imputation for leaching process under different data loss rates and data loss patterns, and then the imputation performances are evaluated. Simulation experiment results have shown that GMM-EM and MI both have advantages with regard to data imputation. Therefore, MI based on GMM (GMM-MI), which combines the advantages of GMM and MI, is proposed in this paper. The effectiveness of GMM-MI is verified by a series of simulations.

1. Introduction

Hydrometallurgical industry has attracted a wide spread attention due to the advantages of high efficiency and low energy consumption. Leaching process always plays an important role in hydrometallurgical processes as the central unit operation, and it directly determines the yield of recovered metal. So the leaching process has become a popular area of research (Hu et al., 2011; Veglio et al., 2001; Coudert and Blais, 2014; Grazyna et al., 2014; Liu and Hu, 2014; Zhang et al., 2015a). Currently, metal leaching rate, as the key production index, can only be measured offline; the online measurement is difficult to implement because of its chemical complexity. Therefore, modeling the leaching process to predict leaching rate is extremely important.

Regardless of what kind of modeling method is used, process data acquisition is the foundation for modeling. However, there is a problem of data deficiency in a practical industrial process for two reasons: first, leaching production is conducted in a harsh environment with a strong acid or base, high pressure and high temperature, which lead to data measuring instruments malfunctioning. Second, data deficiency and inaccuracy may also be derived from interference, drifting and human error. Data deficiency results in the loss of original information in sample data, which hampers the performance of predictive model constructed with incomplete data, especially the dynamic model. Thus, to develop a predictive model with high precision, missing data need to be first estimated and filled using imputation methods.

Data imputation, which is defined as the filling in of missing values for partially missing data, has always been a popular research topic and attracted extensive attention because of its availability and widespread application. From the point of the number of filled values for each missing observation, data imputation can be divided into two categories: single imputation (SI) and multiple imputation (MI). SI is the filling in of a single value for each missing observation with the disadvantage that imputing a single value does not capture the sample variability of the imputed value or the uncertainty associated with the model used for imputation (Lakshminarayan et al., 1999). MI suggests multiple choices for each missing value, and it needs more expensive compu-

[*] *Corresponding author at*: College of Information Science and Engineering, Northeastern University, Shenyang 110004, Liaoning, China.
E-mail address: hedakuo@ise.neu.edu.cn (D. He).

http://dx.doi.org/10.1016/j.cherd.2017.05.023

tation compared to SI but is not associated with the aforementioned drawbacks (Lakshminarayan et al., 1999). In view of the superiority of MI, it has been widely applied to industrial case studies. Gomez-Carracedo et al. (2014) compared the performance of four SI methods and a MI method on actual air quality datasets, and the conclusion proved that MI yielded more disperse imputed values. Bernhardt et al. (2014) proposed a computationally efficient MI method in modeling survival time of patients, and the Simulation studies demonstrated that the proposed MI method works well while alternative methods lead to estimates that are either biased or more variable. Jones et al. (2014) assessed the exposure to drinking water contaminants using the MI method, which appears to be an effective method for filling in water quality values between measures. Young and Johnson (2015) handled missing values in longitudinal panel data with MI, and the MI strategies with fixed effect, pooled time-series models and event-history models are examined. In addition to these practical applications, there are also some improvements for MI have been developed. For example, Gheyas and Smith (2010) proposed a novel nonparametric algorithm called the generalized neural network regression ensemble (GE) and concluded that GE for MI performed better than other conventional imputation algorithms; Zhang et al. (2016) proposed a new MI based validation (MIV) framework and corresponding MIV algorithms for clustering big longitudinal eHealth data with missing values. Besides, some other novel imputing approaches have also been developed to deal with the industrial missing data problem, and please refer to Wang et al. (2006) and Hron et al. (2010) for the detailed case studies. Apart from the innovations mentioned above, many other scholars have also made tremendous contributions to data imputation in both theory and applications, such as Shukur and Lee (2015), García-Laencina et al. (2015), Duan et al. (2016), Nishanth and Ravi (2016) and Fernandes et al. (2017), and it will not be covered here. To the best of the authors' knowledge, however, research on the applications of data imputation in the leaching process (Hu et al., 2011) to handle missing data has rarely been reported in the literature. In this work, a study on data imputation and modeling for the leaching process is conducted to solve the practical problem of data deficiency in process modeling.

The mechanism model of the leaching process should first be developed to conduct such imputation and modeling research. Such a process model plays several significant roles: first, modeling analyzes the mechanism of the leaching process, identifies the important manipulated variables and production indices, and establishes the relationship between them. Second, such a model can be used as a simulator of a real production process to generate the required process data used for simulation and analysis. In this work, a widely accepted mechanistic model for the leaching process (Hu et al., 2011) is introduced to simulate the actual leaching process and produce the required original data.

The generated original data are ideal data that cannot accurately describe the characteristics of field data. In fact, it is impossible to produce such ideal data in the industrial production field. Therefore, the original data should first be roughened, for example, by noise addition or adulteration, to approximate or simulate the field data as far as possible. To simulate data deficiency, the full datasets are pruned artificially based on the data missingness mechanism in the actual leaching process. This pruned data is used for the study of data imputation and modeling.

In this work, we select two classical imputation methods, expectation maximization imputation based on the Gaussian mixture model (GMM-EM) and multiple imputation (MI), to study missing data imputation with designed experiments. Expectation maximization (EM) algorithm (Dempster et al., 1977) is a broadly used iterative algorithm to perform the maximum likelihood estimation and deal with incomplete-data problems (Ding and Song, 2016; Mustafa et al., 2012). The idea of the algorithm is to alternate between Expectation (E-step) and Maximization (M-step). The E-step is to compute the expectation of the complete data likelihood conditional on the previous parameter estimates and the M-step is to maximize this expectation regarding the desired parameters to obtain parameter estimates for the next recursion (Mustafa et al., 2012). The Gaussian mixture model (GMM) is a linear superposition of different Gaussian sub-models (Zhang et al., 2015b). It has been verified that the GMM could be used to fit an arbitrary continuous probability density function if there are sufficient Gaussian components (Chen et al., 2006). In view of its aforementioned advantages, GMM-EM has been successfully applied (Zio et al., 2007; Yan et al., 2015). However, GMM-EM is a single imputation approach with which just a single value is produced to estimate missing data and a greater imputed error may be generated by some uncertain factors. Therefore, MI is introduced to overcome the shortcomings of single imputation. The core superiority of MI is the generation of multiple estimations for missing data to handle uncertain factors. For the detailed procedure of MI please see Section 3.1.2.

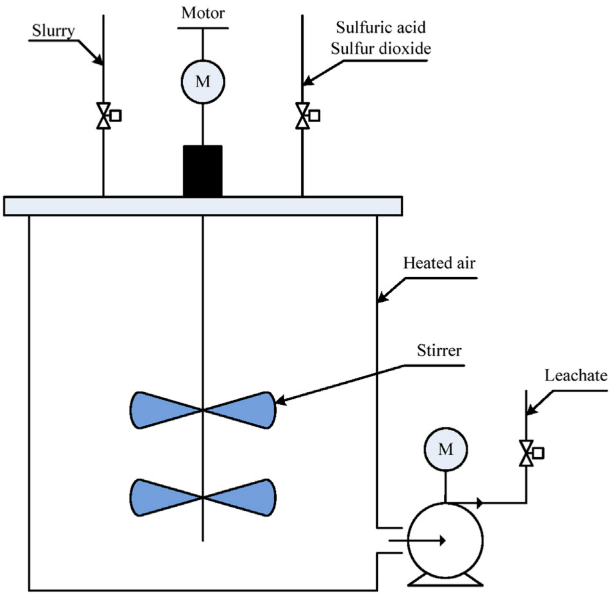

Fig. 1 – Schematic diagram of the leaching tank.

Imputation performances of GMM-EM and MI are assessed by comparing the true values and estimated values. Then two classical modeling methods, kernel partial least squares (KPLS) and least squares support vector machine (LSSVM), are chosen to develop the model of leaching rate. Modeling performance should also be evaluated. Through the in-depth analysis of the simulation results, it is indicated that GMM-EM and MI both have strengths under different data loss patterns and data loss rates. It is demanded that data imputation methods must accommodate various data loss patterns and data loss rates in the practice. Therefore, to further improve the overall imputation effect, the hybrid method, which is named MI based on GMM (GMM-MI) and combines the advantages of GMM and MI, is presented in this paper. The effectiveness of the hybrid method is validated by a series of simulations.

2. Mathematical mechanism model of the leaching process

2.1. *Brief introduction of the leaching process*

Generally, leaching is defined as the extraction of metals by dissolving them from solid ore. The acidic or basic reagent is injected into the leaching tank, in which the valuable metal components and impurities are separated by a chemical reaction. The entire process is a batch process. First, the ore is pulped in the slurry tank and transported to the leaching tank. Then, the ore slurry is stirred and heated, and reacted with the sulfuric acid and sulfur dioxide in leaching tank. The schematic diagram of leaching tank is shown in Fig. 1.