CHEMICAL ENGINEERING TRANSACTIONS

VOL. 62, 2017

A publication of AIDIC

The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Fei Song, Haibo Wang, Fang He

ISBN 978-88-95608- 60-0; **ISSN** 2283-9216

DOI: 10.3303/CET1762126

Data Preprocessing Technology in Chemical Process Data Mining

Feifei Yang

Shandong Youth University of Political Science, Jinan 250103, China
362189196@qq.com

This paper studies the data preprocessing technology in chemical process data mining. It mainly studies the real-time chemical process data in-depth from the perspective of software development and explores various kinds of real-time data preprocessing methods. The enterprise pre-processes real-time data based on the MES data mining system and other systems associated with real-time data to develop a real-time data preprocessing system based on the theory. In addition to acquisition and storage of real-time data, this system corrects false data and fills missing data to provide accurate and credible data for data mining tools. Finally, the real-time data preprocessing system integration is applied to data mining systems and other information integration systems associated with real-time data.

1. Introduction

In the process industry, the process is complex. Due to the constantly changing of production associated with real-time data and relational data, technologists need a variety of MES tools for data mining of vast amounts of information in real-time database and relational database, performance parameter calculation and condition analysis to guide the adjustment of operating conditions (Kumar et al., 2016, Yao et al., 2016, Tugizimana et al., 2016). The accurate and complete real-time data is the basis for MES-based data mining tools. Some data mining results are based on the accurate measurement of required parameters (Kobryn and Prystrom, 2016, Ahmed, 2016). The correctness of data acquisition directly affects online calculation results and normal operation of data mining system (D'Amico et al., 2016, Giuliani et al., 2016). Chemical instruments often work in high temperature, vibration, corrosion and other harsh environments, easy to fail, resulting in wrong data collected by the data acquisition system (Tajinder and Madhu, 2016, Choi and Song, 2016, Immohr et al., 2016). In addition, the measurement may be affected by disturbances, drift and the environment, and the real-time measurement may be inaccurate (Liland et al., 2016, Nayak and Kanive, 2016). Wrong measurement can cause serious consequences. The research shows that even 1% of measurement of key parameters drift can cause significant heat loss and short service life of device (Chaharmahali et al., 2016, Vaidya and Anand, 2016, Reddy, 2016). Such deviation is difficult to check visually (Yamac et al., 2016, Dayarathna et al., 2016, Rao et al., 2016). Therefore, real-time data pre-processing before being used in the system is necessary.

2. Methods

2.1 Data mining flow chart

Step 1: define the issues of data mining, that is, define the business issues clearly to determine the purpose of data mining.
Step 2: data preparation, selecting data to extract the target dataset for data mining in large databases and data warehouse. Data preprocessing includes checking data integrity and consistency, de-noising, filling in lost domains, deleting invalid data and so on.
Step 3: data mining. The algorithm is selected according to the types of functions and features of the data, and data mining is performed on the cleaned and converted datasets.
Step 4: result analysis. The results of data mining are interpreted and evaluated, transforming into knowledge that can be understood by the user.

Please cite this article as: Feifei Yang, 2017, Data preprocessing technology in chemical process data mining, Chemical Engineering Transactions, 62, 751-756 DOI:10.3303/CET1762126

Finally, application of knowledge. The analytic knowledge is integrated into the organization structure of business information system.

2.2 Regression analysis

The relationships between variables include deterministic relation and non-deterministic relation. In deterministic relation, when the independent variable is given, the value of dependent variable is determined accordingly. For example, the relationship between voltage, resistance, and current can be described by an exact function. Non-deterministic relations between variables are called correlations. Variables with correlations cannot be described by exact functions, but fluctuate around certain functions. Regression analysis deals mainly with the correlation, including how to determine the regression model between dependent variables and independent variables; how to estimate and test the regression model and position parameters based on the observed data; determine which of the independent variables has significant influence on dependent variables and which independent variables are not significant; the dependent variable's value is estimated and predicted from the known or given values of the independent variables. Figure 1 shows the regression realization.

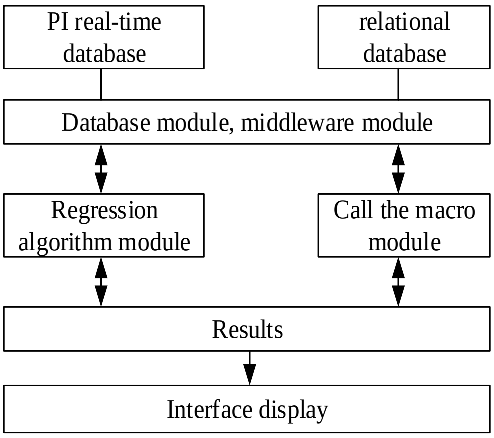

Figure 1: Regression realization

2.3 Energy consumption module

Raw materials and consumables in the process industry is continually input, products and intermediate products are continually output, the process operation conditions are constantly changing, and the energy consumption of products and intermediate products is constantly changing. In the MES environment, real-time databases offer the possibility of online monitoring and management of energy. The module automatically generates the monthly comprehensive energy consumption reports of methanol and phthalic anhydride based on the MES information bus, and conducts real-time monitoring and tracking of carbon-based products and intermediate products to realize the dynamic management of energy. The principle of energy consumption analysis is the input-output method. Figure 2 shows the energy consumption technical framework.

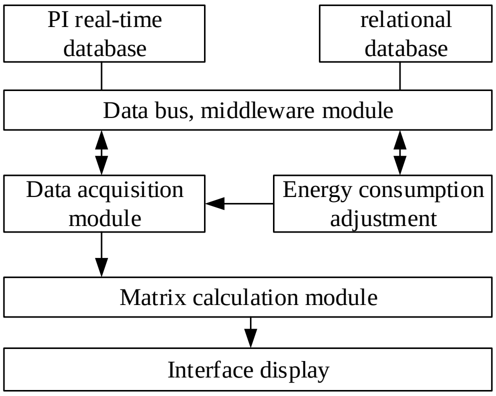
PI real-time database
relational database
Data bus, middleware module
Data acquisition module
Energy consumption adjustment
Matrix calculation module
Interface display

Figure 2: Energy consumption technical framework

3. Real-time data preprocessing algorithm

3.1 Data preprocessing

Data preprocessing is reprocessing of selected raw data, including checking data integrity and consistency, de-noising, filling in lost domains, deleting invalid data and so on. As can be seen from Figure 3, data preprocessing is a key step in data mining and plays an important role in the success of data mining.

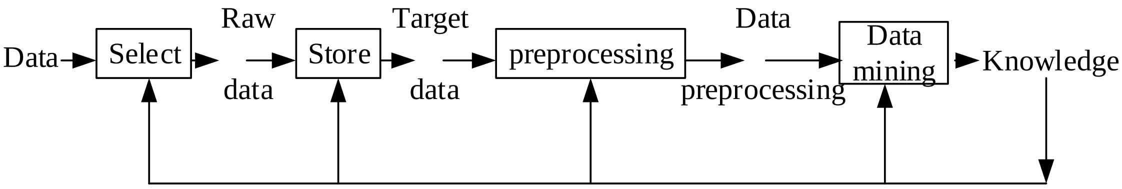

Figure 3: KDD Process

3.2 Instrument zero

Occasional zero of instrument is completed by the operator, but it is difficult for the operator to feedback zero information at first time; for full scale zero, it is difficult to grasp the zero information. In case of zero of instrument, determine the time of zero first before handling.

The purpose of zero is to extract the sampling value of the instrument when it is zero. This paper uses the iterative method to process. Iterative method is a process for recurrence of new value of the variable using the old value constantly. The direct method is on the contrast of the iterative method, which solve the problem one time. The iterative method is divided into precise iteration and approximate iteration. "Dichotomy" and "Newton's iterative method" belong to the approximate iterative method. Figure 4 shows the instrument zero example.

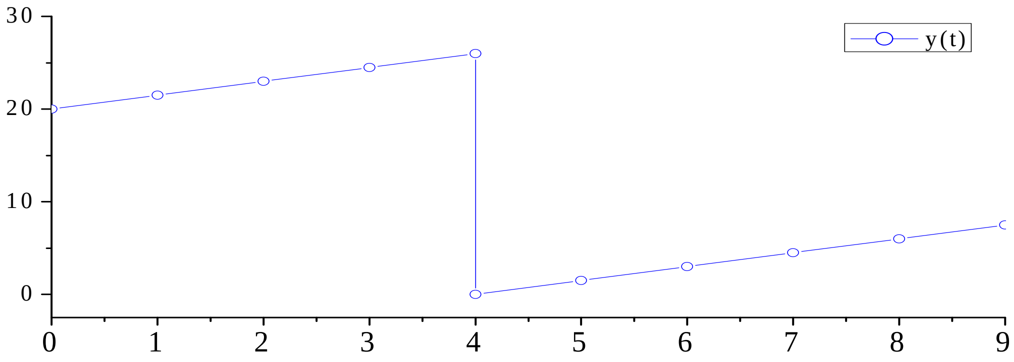

Figure 4: Instrument zero example

3.3 Integrated preprocessing algorithm based on SVM regression

SVM can maximize the promotion ability of learning machine, even the discriminant function obtained from finite dataset. A small error can be got for independent test set. This algorithm can effectively estimate the pressure and flow data in the chemical plant so as to substitute inaccurate real-time data; Filtering is effective in eliminating high frequency noise in pressure and flow signals and inhibiting cyclical interference. Integrated filtering algorithm and SVM regression algorithm can be adopted to effectively reduce the noise interference, while estimate and substitute wrong data. This paper introduces an integrated preprocessing algorithm based on SVM regression in the process of preprocessing.

4. Architecture of system

4.1 Development method

The real-time data preprocessing system is developed using server programs. The system connects to the PI real-time database as a client, and real-time data is read from the PI periodically according to the process task and stored in the relational database after proper preprocessing. The data preprocessing system is a client of PI real-time database. Essentially the data preprocessing system reads, processes and stores real-time data and works as the data source of MES upper layer software such as data mining, energy consumption analysis, dynamic cost and other information integration system. Therefore, the real-time data preprocessing system is in fact a server program of the database server. The design of development framework and implementation should be in accordance with the standard of server programs.

4.2 Specific system architecture

The system architecture is the structure of the entire system, which includes the components of the system and how the components are integrated. The real-time data preprocessing system consists of the real-time database, relational database, system display and system logic. Figure 5 shows the system architecture.

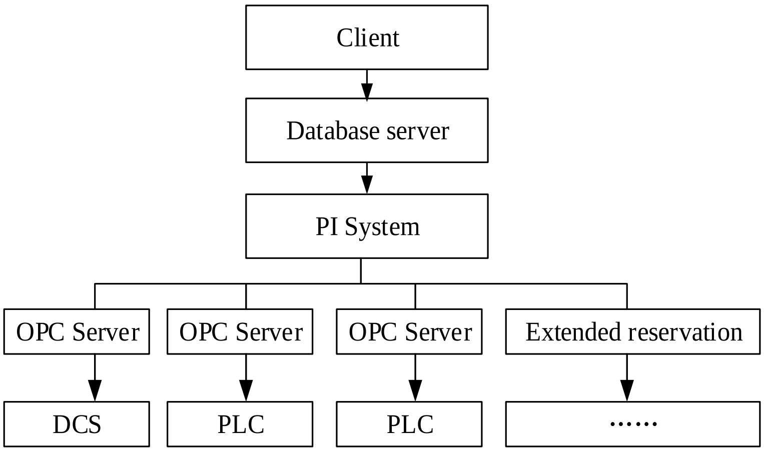

Figure 5: System architecture

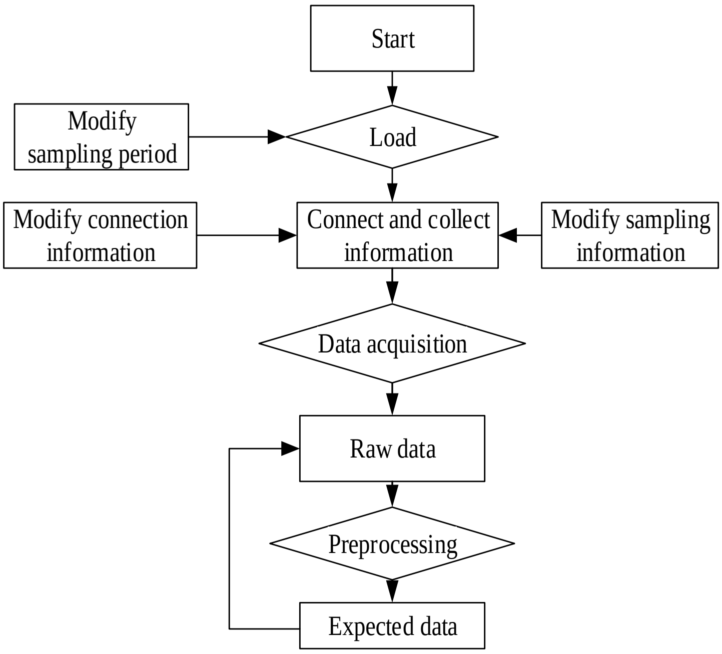
Start
Modify sampling period
Load
Modify connection information
Connect and collect information
Modify sampling information
Data acquisition
Raw data
Preprocessing
Expected data

Figure 6: System flow chart

4.3 Flow chart of real-time data preprocessing system

The system flow chart includes the system logic and analysis of functions, and description of each module, which is the prerequisite and necessary preparation for development of the system. An excellent system flow chart is a key step for successful system development, both to meet the customer satisfaction, but to ensure the simplicity and maintainability of development.

The real-time data preprocessing system includes information loading, data acquisition, data storage, data preprocessing, data re-storage and other functional modules. Figure 6 shows the system flow chart.

4.4 Function modules

(1) OPC connection module

OPC is a software interface standard for connecting a data source (OPC server) to a data user (PI System). Data source can be PLC, DCS, bar code reader and other control equipment.

(2) PI connection module

The PI real-time database stores real-time data collected from devices such as the underlying DCS and PLC via the OPC interface. The PI connection module is a programming interface program, so that the real-time data preprocessing system can access the PI database safely and stably.

(3) Database access information setting module

The system also accesses the relational database through the server IP address for the SQL Server relational database and the name of relational database.

(4) Tag point setting module

There are large numbers of sampling points added by the project implementation company to the PI real-time database. The real-time data preprocessing system only pre-processes part of the Tag points.

(5) Data storage module

Data storage module is called to store collected real-time data and preprocessed data.

(6) Data preprocessing module

This module is a core part of the real-time data preprocessing system.

5. Conclusion

This paper studies the MES real-time data preprocessing technology. The theory of preprocessing algorithm is studied in depth. The suitable preprocessing algorithm is selected according to the preprocessing needs of the enterprise and the actual chemical process. Concerning the realization, the predetermined requirements are met. The OPC connection module, PI connection module, database information setting module, Tag point setting module, data storage module and data pre-processing module are developed. The real-time data preprocessing algorithm based on support vector machine regression theory is studied. The algorithm detects the fault information by using the filtering algorithm combined with the deviation band test, and establish the support vector machine regression training model for the fault point using the redundant information. The example shows that the fault information can be effectively detected using this method, and the accurate estimation of the fault point can be given.

Reference

Ahmed Z., 2016, Interactive quality and pre-processing pipeline for atac-seq data, Frontiers in Neuroinformatics, 10, DOI: 10.3389/conf.fninf.2016.20.00020

Chaharmahali I., Asadi S., Dousti M., 2016, A new pre-processing algorithm for analog circuit expression and simplification, International Journal of Numerical Modelling Electronic Networks Devices & Fields, DOI: 10.1002/jnm.2192

Choi K.Y., Song B.C., 2016, Linear sub-band decomposition–based pre-processing for perceptual video coding, Ieie Transactions on Smart Processing & Computing, 5(5), 366-373, DOI: 10.5573/ieiespc.2016.5.5.366

D'Amico G., Amodeo A., Mattis I., Freudenthaler V., Pappalardo G., 2016, Earlinet single calculus chain-technical: pre-processing of raw lidar data, Atmospheric Measurement Techniques, 8(10), 10387-10428, DOI: 10.5194/amt-9-491-2016

Dayarathna M., Li Y., Wen Y., Fan R., 2016, Energy consumption analysis of data stream processing: a benchmarking approach, Software Practice & Experience, DOI: 10.1002/spe.2458

Giuliani M., Vissa A., Driouchi A., Yip C.M., 2016, Effect of data pre-processing on super-resolution reconstruction and pattern recognition, Biophysical Journal, 110(3), 331a-331a, DOI: 10.1016/j.bpj.2015.11.1781

Immohr L.I., Turner R., Pein-Hackelbusch M., 2016, Data for a pre-performance test of self-developed electronic tongue sensors, Data in Brief, 9, 1090-1093, DOI: 10.1016/j.dib.2016.11.041

Kobryn A., Prystrom J., 2016, A data pre-processing model for the topsis method, Folia Oeconomica Stetinensia, 16(2), 219-235, DOI: 10.1515/foli-2016-0036

Kumar M.S., Mr N.D.C., 2016, A survey on improving classification performance using data pre processing and machine learning methods on nsl-kdd data, International Journal of Advanced Trends in Computer Science & Engineering, DOI: 10.18535/ijecs/v5i4.17

Liland K.H., Kohler A., Afseth N.K., 2016., Model-based pre-processing in raman spectroscopy of biological samples, Journal of Raman Spectroscopy, 47(6), 643-650, DOI:10.1002/jrs.4886

Nayak A.S., Kanive A.P., 2016, Survey on pre-processing techniques for text mining, International Journal of Advanced Trends in Computer, Science & Engineering, DOI: 10.18535/ijecs/v5i6.25

Rao S., Masilamani S., Sundaram S., Duvuru P., Swaminathan R., 2016, Quality measures in pre-analytical phase of tissue processing: understanding its value in histopathology, J Clin Diagn Res, 10(1), EC07-EC11, DOI:10.7860/jcdr/2016/14546.7087

Reddy G.N., 2016, Big data processing using hadoop in retail domain, International Journal of Advanced Trends in Computer Science & Engineering, DOI: 10.18535/ijecs/v5i9.65

Tajinder S., Madhu K., 2016, Role of text pre-processing in twitter sentiment analysis, Procedia Computer Science, 89, 549-554, DOI: 10.1016/j.procs.2016.06.095

Tugizimana F., Steenkamp P.A., Piater L.A., Dubery I.A., 2016, A conversation on data mining strategies in lc-ms untargeted metabolomics: pre-processing and pre-treatment steps: Metabolites, 6(4), 40, DOI: 10.3390/metabo6040040

Vaidya R., Anand S., 2016, Image processing assisted tools for pre-and post-processing operations in additive manufacturing, Procedia Manufacturing, 5, 958-973, DOI: 10.1016/j.promfg.2016.08.084

Yamac M., Cagatay D., Sankur B., 2016, Hiding data in compressive sensed measurements: a conditionally reversible data hiding scheme for compressively sensed measurements, Digital Signal Processing, 48, 188-200, DOI: 10.1016/j.dsp.2015.09.017

Yao J., Cao J., Zheng Q., Ma J., 2016, Pre-processing of incomplete spectrum sensing data in spectrum sensing data falsification attacks detection: a missing data imputation approach, Iet Communications, 10(11), 1340-1347, DOI: 10.1049/iet-com.2015.1111