CHEMICAL ENGINEERING TRANSACTIONS

VOL. 66, 2018

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

Guest Editors: Songying Zhao, Yougang Sun, Ye Zhou
Copyright © 2018, AIDIC Servizi S.r.l.
ISBN 978-88-95608-63-1; **ISSN** 2283-9216

DOI: 10.3303/CET1866151

Application of Data Mining Technology in Chemical Engineering Optimization

Yongmei Niu

Nanyang Institute of Technology, Henan 473000, China
yongmeiniu7456@21cn.com

To find a fast and efficient fault diagnosis method for induction motor using limited data samples, the modelling and application of deep learning in induction motor fault diagnosis was studied. Firstly, the application of traditional convolution neural network (CNN) in fault diagnosis of induction motor was analysed, and a convolutional discriminative feature learning (CDFL) method based on improved CNN training mode used for discriminative learning of induction motor fault characteristics was put forward. The method mainly used BP neural network discriminative learning ability to learn local features of convolutional layer filter, so that the convolution pool model could learn the default features with the fault characteristic invariant of the induction motor vibration signal not adjusting the network parameters. In addition, the correct classification of fault type was achieved by selecting support vector machine SVM. At last, experiment was designed to compare CDFL method and signal processing wavelet packet transform method and it was compared with other deep learning methods. The experimental results showed that the classification results of CDFL fluctuated the minimum; when the filter window size was 200, the classification effect of model was the best and when the pool domain size was 20, it achieved the best effect of classification. To sum up, CDFL has the highest classification accuracy and good robustness, and it can learn the fault characteristics of induction motor more quickly, intelligently and effectively. To study the application of data mining technology in chemical engineering optimization, the data mining technology was adopted in this paper to discretize and analyze the experimental data. Results have shown that after the data discretion by the database technology, the expansion rate was below the optimal value when the deposition rate and recovery rate of the chemical equipment were the best. It was then concluded that the application of data mining technology in chemical engineering optimization, far from unrealistic, can greatly improve economic benefits for the company, for which it can be highly promoted and widely applied.

1. Introduction

Evolving with the Internet, the data mining technology has been applied to certain extent. Currently, as a hot topic in the chemical industry, the technology has been studied by a number of domestic and foreign scholars. Traced back to as early as the 1980s to 1990s, the data mining technology mainly refers to finding the target information timely and accurately from a large amount of information through certain technologies. Affected by several factors in the chemical manufacturing, the production equipment under operations may suffer disturbance from the environment, leading to inability of the original design scheme or procedure. However, it is difficult to accurately understand the relationship between variables by optimizing the parameters manually, which hinders the chemical companies to further enhance the production efficiency. To this end, the application of data mining technology to the chemical engineering optimization is very essential. Energy consumption and recovery rate are important indicators for evaluating the production efficiency of chemical companies, while the two indicators are often affected by product quality and manufacturing time.

With a large number of domestic and foreign related documents as the reference, this paper used the data mining technology to discretize and analyze the experimental data, and studied the application of data mining technology in chemical engineering optimization, with the purpose of further understanding the optimization of parameters in the chemical industry for enhanced economic efficiency of companies. Therefore, this study is of important practical significance.

Please cite this article as: Niu Y., 2018, Application of data mining technology in chemical engineering optimization, Chemical Engineering Transactions, 66, 901-906 DOI:10.3303/CET1866151

2. Literature review

The research methods of data mining technology are mainly established based on the theories and methods of artificial intelligence, computational intelligence and statistical methods, which mainly include: computational intelligence (neural network and genetic algorithm), statistical method (principal component analysis and partial least squares), fuzzy theory method, rough set theory, machine learning method (decision making) and so on. In the procedure of chemical process modelling, neural network (ANN), genetic algorithm (GA), principal component analysis (PCA) and partial least squares (PLS) are widely used.

The neural network imitates the biological neural network from the structure, and is composed of a large number of simple neurons to form a network system according to certain rules, so as to achieve the goal of simulating human image intuition. Neural network uses its idea of nonlinear mapping and parallel processing, and uses its own structure to express input and output related knowledge. The neural network method has more advantages when it is used for nonlinear data and data containing noise. It can accomplish a variety of data mining tasks, such as classification, clustering, feature mining and so on. Because neural networks can approximate arbitrary nonlinear mappings with arbitrary accuracy and bring a non-traditional expression tool to modelling, it is widely used in the field of nonlinear chemical process modelling. Many researchers have done a lot of research on how to improve the predictive ability of neural networks. Saon used combinatorial generalization to propose a combined neural network method, which could significantly improve the predictive ability of the model, in which the selection of combined weights was very critical for the good performance of the composite neural network (Saon, 2018). Therefore, in recent years, many methods have been proposed to reasonably select combination weights, such as multiple linear regression, principal component regression and information inference combination. Qiu and others used combined neural network to predict the mass of polymer in batch reactor, that is, a mathematical model that uses a combined neural network to represent the relationship between the polymerization formula and the trajectory of the mass variable of the polymer. The predictive confidence interval of the combined neural network model was calculated to improve its generalization ability and was successfully applied to the study of an intermittent isobutylene methyl ester polymerization reactor (Qiu et al., 2016). A hybrid neural network modelling for an industrial wastewater treatment process was carried out by Ostad-Ali-Askari and others. First of all, a simplified mechanism model was established based on the experience and knowledge of the process. Then, a neural network model was established according to the actual operation data. Finally, the neural network model and the simplified mechanism model were combined together by adopting the parallel method. It showed that the hybrid neural network model had better predictive power and extrapolation performance by comparing with conventional methods (Ostad-Ali-Askari et al., 2016).

The genetic algorithm is an adaptive heuristic probability iterative global search algorithm, which has robustness and global optimality in solving nonlinear problems, and does not depend on the characteristics of the problem model in the process of solving the problem. In addition, it has the characteristics of parallelism and high efficiency. Due to the novelty of the genetic algorithm and the rapid development of computer technology, it has been widely used. For instance, the optimization of complex problems, pattern recognition, engineering design, and control system optimization have achieved good results. In chemical process, genetic algorithm is mainly used in modelling, control and optimization. Liu and others used genetic algorithms for the steady state modelling of the chemical process system, that is, the genetic programming method was adopted to establish the input and output model of the complex chemical process (Liu et al., 2016). The advantage of this method is that a simple model which can accurately reflect the characteristics of the process can be obtained without any modelling hypothesis, and the complexity of the model can be reduced by using the genetic algorithm. The method is successfully applied to the steady modelling of two typical chemical processes. Ivanov and so on also used genetic programming to establish the dynamic model of a chemical system, and proposed a nonlinear model predictive control strategy. The results showed that the performance of the predictive control strategy exceeded that of the usual linear model (Ivanov et al., 2016).

Principal component analysis and partial least squares are both new methods of multivariate statistical data analysis. Principal component analysis can achieve the following objectives: data simplification, data compression, modelling and variable selection. Kamilov and Mansour applied iterative nonlinear PLS method in the modelling of nonlinear chemical process, and successfully applied the method to 3 typical chemical processes. By comparing with other similar methods, it is proved that the method is more suitable for nonlinear process modelling, and the prediction ability of the built model can be significantly improved by using this method (Kamilov and Mansour, 2016).

Since each data mining technology is proposed for a specific background, there are some shortcomings that cannot be overcome by some methods themselves. Therefore, these data mining techniques can be combined to achieve a better effect than the use of certain data mining technology alone. The following two kinds of data mining integration technologies and their applications in chemical process modelling are introduced. Velásco-Mejía and others carried out the modelling research on a complex biochemical process by combining neural network with genetic programming. By comparing the result with the result of mechanism modelling, it is pointed out that the advantage of the method is that some process experience and knowledge are not required in the process of modelling, but the model established can still accurately predict the system response (Velásco-Mejía et al., 2016). Pablo and others applied neural network and genetic algorithm to model the complex process of unknown or complex mechanism and optimize its operating conditions. The genetic algorithm was used to quickly search the optimized region and the results proved the feasibility of the proposed method (Pablo et al., 2016).

In recent years, data mining technology has been widely applied in chemical process optimization, and new methods with great application prospect, such as ANN, GA, PLS and PCA, are proposed. But data mining, as a brand-new technology, is at the initial stage both at home and abroad at present, and its theories, methods and applications are not mature enough. In general, many data mining techniques are proposed for specific background. There are some limitations in the application of chemical process optimization, and many data mining techniques have not been widely used in chemical industry, such as rough set theory, decision tree and so on. However, data mining is a hot research field. With the further development of chemical process monitoring and control technology, a large number of process data are collected and stored by computer control system. New applications of data mining technology in chemical process optimization will continue to emerge.

3. Methods

The database technology simply organizes and stores the data in the database efficiently, and makes some simple analysis of the data, while a lot of useful information hidden inside the data cannot be obtained. In the field of machine learning, pattern recognition, and statistics, there are a large number of methods for knowledge extraction, but they largely play a role in experimental data or academic research as they are not combined with the massive data used in practical applications. Data mining combines the fields of database technology, machine learning, pattern recognition, and statistics from a new perspective, and explores a deeper, effective, novel, potentially useful, and ultimately understandable patterns that exist within data. The commonly used methods of data mining are as follows: (1) Decision tree. Decision tree technology is mainly used for predictive modeling of classification, clustering, and prediction. It uses the mutual information (information divergence) in the information theory to find the field with the largest amount of information in the database, establishes a node of the decision tree, then establishes the branch of the tree according to the different values of the field, and repeats the establishment of lower nodes and branches in each branch sub-set, so that a decision tree is generated. (2) Pattern recognition is one of the main methods of data mining. It is a kind of mathematical statistical method that uses computer to process information and judge classification. (3) Artificial neural network method is used for classification, clustering, feature mining, prediction and pattern recognition, as shown in Figure 1. The neural network method mimics the brain neuronal structure of animals based on the M-P model and Hebb learning rules. In essence, it is a distributed matrix structure that gradually calculates (including iterative iteration or cumulative calculation) the weights of neural network connections through the mining of training data.

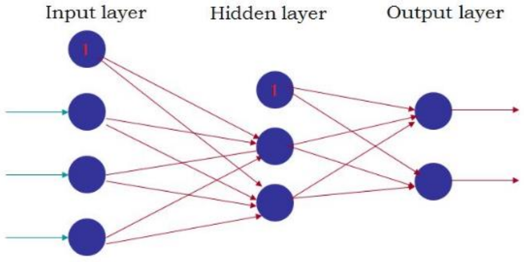
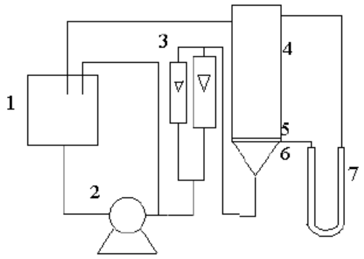


Figure 1: Artificial neural network method

In this experiment, acidic copper sulfate aqueous solution was used as the electrolyte to electrolyze copper ions. The copper ion concentration was 0.011 mol/l and the sulfuric acid concentration was 0.714 mol/l. The electrolysis process of the narrow component of the spherical copper powder was studied when the copper solution was flowing through the fluidized bed. Also, the porosity of the bed during the process of the fluidized bed electrode and the other parameters of the electrolysis process of copper power with different diameters were measured, such as current efficiency, recovery, deposition rate, and power consumption. With narrowgrained spherical copper powder as conductive particles, constant-potential electrolysis was used to electrolyze low-concentration copper ions, and the effects of current efficiency, recovery rate, deposition rate, and energy consumption under different bed expansion rates were examined. The average particle size of the particles was 0.33, 0.40, 0.52, 0.69 and 1.00mm, a total of 5 commonly used particle sizes. The bed mass was 100 g, and the bed expansion ratio was 10 to 50%, accounting as 5 expansion ratios. The experimental device is shown in Figure 2.



1. Sink; 2. Magnetic pump; 3. Rotameter; 4. Fluidized bed; 5. Distribution board; 6. Predistributor; 7. U-tube differential pressure meter

Figure 2: Experimental apparatus

Recovery rate: $\theta = 1 - \frac{C_1}{C_0}$
In the formula, C0 refers to the concentration of the reactants at the beginning of the reaction;
Ct means the concentration of the reactants at the end of the reaction;

4. Research results and discussion

Table 1 shows the experimental results obtained when the copper particles in commonly used five particle sizes in electrolytic copper processing are fluidized particles.

Table 1: Electrolysis of copper in a fluidized bed electrode containing monocomponent particle

Expansion ratio %	Current efficiency%	Recovery rate%	Deposition rate g/h	Particle size mm
7	57	32	1.3	0.3
24	71	47	1.3	0.3
33	85	56	1.7	0.3

To evaluate a data mining tool, consider the following aspects:
1. The number of patterns generated and the ability to solve complex problems. With the increase of data volume and higher requirements for fineness and accuracy of the model will increase the complexity of the problem. The data mining system can provide the following solutions to complex problems. The first is the multiple modes. The combination of multiple category modes helps discover useful patterns and reduces problem complexity. For example, grouping data by clustering and then mining predictive patterns for each group will be more efficient and accurate than simply performing operations on the entire dataset. The second is multiple algorithms. Many models, especially those related to classification, can be implemented with different algorithms, each with its own advantages and disadvantages, applicable to different needs and environments.
Data selection and conversion patterns are often hidden by a large number of data items. Some data is redundant, and some is completely irrelevant. The existence of these data items will affect the discovery of valuable models. A very important function of the data mining system is to be able to deal with data complexity, provide tools, select the correct data items and convert data values.
2. Easy to operate. Easy operation is an important factor. Some tools have graphical interfaces that guide users to perform tasks semi-automatically, and some use scripting languages. Some tools also provide data mining APIs that can be embedded into programming languages such as C, VisualBasic, and PowerBuilder. Patterns can be applied to existing or newly added data. Some tools have a graphical interface, and some allow the schema to be exported to a program or database by using a programming language such as C or a set of rules in SQL.

iDA provides support for business and technical analysis by providing a complete set of visual data mining tools, including a preprocessor, 3 data mining tools, and a report generator. iDA provides a Microsoft Excel with user interface. The components of IDA are shown in Figure 3.

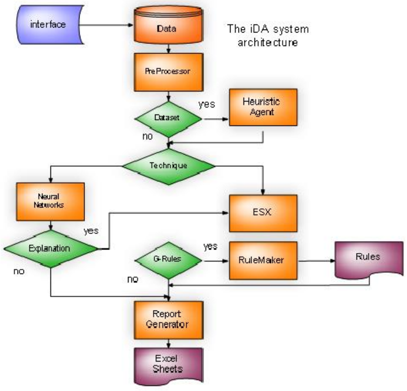


Figure 3: iDA system structure

Before the preprocessor submits the data in the file to the iDA's mining engine, it scans the file for several types of errors, including illegal values, blank lines, and null values. The preprocessor corrects several errors but does not fix errors in the numeric data. The preprocessor outputs a file waiting to be mined, or informs us of the location of the error that we cannot resolve.
The heuristic agent is responsible for the display of data files containing thousands of records, and allow us to decide whether to extract a representative data set from a data set for analysis or analyze all data sets.
The neural network. iDA, contains two neural network architectures: one is a backward propagating neural network that supports supervised learning, and the other is an unsupervised clustering self-organizing functional map.
In the experimental data of the narrow fluidized bed electrode electrolysis experiment we selected, the 21st and 22nd sets of data were isolated due to experimental conditional control failure and were manually cleared. The initial bed height and fluidized bed height were used to calculate the voidage, the flow rate was used to calculate the apparent flow rate, and the power consumption was linearly related to the current efficiency. Therefore, when the target data set was created, these three attributes were cleared, and the selected data set is shown in Table 2:

Table 2: Data set after data cleaning for narrow component fluidized bed electrode electrolysis experiments

Expansion ratio %	Current efficiency%	Recovery rate%	Deposition rate g/h	Particle size mm
7	57	32	1.4	0.5
24	71	47	1.5	0.5
33	85	56	1.9	0.5

Table 3: Discretized Data Sets for Narrow Fraction Fluidized Bed Electrolysis Experiments

Expansion ratio %	Current efficiency%	Recovery rate%	Deposition rate g/h	Particle size mm
0-20	50-60	30-40	M	0.3
20-40	70-80	40-50	M	0.3
20-40	80-90	50-60	H	0.3

In general, the attributes in the database can be divided into two types. One is the continuous (quantitative) attribute that represents certain measurable properties of the object being described, with its value taken from a continuous interval, such as temperature, length, etc.; the other is discrete (qualitative) attributes, whose value is expressed in language or a small number of discrete values, such as gender, color, etc. In most cases, the same database contains both continuous and discrete attributes. The discretization process includes univariate discretization and multivariate discretization. The univariate discrete refers to a continuous attribute for a discretization, while the multivariate discretization can handle multiple continuous attributes at the same time. The following describes a typical univariate discretization process: (1) Rank data with continuous attributes to be discretized; (2) Preliminarily determine candidate points for continuous attributes;

(3) Continue to segment or merge candidate points according to some criteria; (4) If (3) reaches the suspension condition, the entire discretization process is aborted; otherwise, the step (3) is continued.
We use the “3-4-5” rule to discretize experimental data as shown in Table 3.
The expansion ratio is one of the important parameters of the fluidized bed electrode electrolyzer. When the expansion rate is very low, it may lead to the adhesion and agglomeration of the conductive particles, which is equivalent to a completely “on” state between copper particles in the packed bed. The discharge was reduced only near the semi-permeable membrane closest to the anode. When the expansion rate is too large, the particles may not be in contact with each other, reducing the effective use area so that the particle bed is in the "open circuit" state, and the copper ions are only discharged in the vicinity of the feed electrode. This will reduce the recovery rate and current efficiency, so the expansion rate should be properly controlled between 10% and 30% when copper removal has the fastest and highest efficiency with the current efficiency significantly increased, thereby reducing power consumption.

5. Conclusion

The rapid development of the information technology has led to the emergence of database technology. At present, database technology has been widely used in data information management to improve the efficiency of data information processing and storage. Today, facing severe environmental problems, the application of database technology into chemical companies can further optimize their production processes, save energy and reduce emissions, and help them better handle waste water and waste residues with heavy metal content. The results of this study show that after the database is discretized, the expansion rate is below the optimal value. At this time, the deposition rate and recovery rate of the chemical equipment are optimal. It can be seen that the application of data mining technology in chemical engineering optimization is actually practical and can greatly increase the economic efficiency of enterprises with high promotion and application value.
Owing to limited knowledge of the author, there may be some deficiencies in this paper, and the experimental design carried out in this paper is of a small scale, which may have a big gap with the actual chemical enterprise production situation, and requires further study.

Reference

Ivanov G., Nikolov N., Nikova S., 2016, Reversed genetic algorithms for generation of bijective s-boxes with good cryptographic properties, Cryptography & Communications, 8(2), 247-276, DOI: 10.1007/s12095-015-0170-5

Kamilov U.S., Mansour H., 2016, Learning optimal nonlinearities for iterative thresholding algorithms, Signal Processing Letters, 23(5), 747-751, DOI: 10.1109/LSP.2016.2548245

Liu S., Pan J., Yang M.H., 2016, Learning Recursive Filters for Low-Level Vision via a Hybrid Neural Network, Computer Vision–ECCV 2016, Springer International Publishing, 560-576, DOI: 10.1007/978-3-319-46493-0_34

Ostad-Ali-Askari K., Shayannejad M., Ghorbanizadeh-Kharazi H., 2016, Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran, Ksce Journal of Civil Engineering, 21(1), 1-7, DOI: 10.1007/s12205-016-0572-8

Pablo B.M.J., Piedad C.I.C.H., Santiago S.V., 2016, Neural Networks and Genetic Algorithms Applied for Implementing the Management Model “Triple A” in a Supply Chain, Case: Collection Centers of Raw Milk in the Azuay Province, 68, 06-08, DOI: 10.1051/matecconf/20166806008

Qiu J., Wang J., Yao S., Guo K., Li B., Zhou E., 2016, Going Deeper with Embedded FPGA Platform for Convolutional Neural Network, Acm/sigda International Symposium on Field-Programmable Gate Arrays, 26-35, DOI: 10.1145/2847263.2847265

Saon G.A., 2018, Speaker adaptation of neural network acoustic models using i-vectors, 55-59, DOI: 10.1109/ASRU.2013.6707705

Velásco-Mejía A., Vallejo-Becerra V., Chávez-Ramírez A.U., Torres-González J., Reyes-Vidal Y., Castañeda-Zaldivar F., 2016, Modeling and optimization of a pharmaceutical crystallization process by using neural networks and genetic algorithms, Powder Technology, 292, 122-128, DOI: 10.1016/j.powtec.2016.01.02