Physicochemical reaction mechanism study and model optimization based on big data analysis

Abstract: The huge amount of data generated by information technology, intelligence and automation has serious redundancy and duplication, which becomes a hidden cost of consuming resources. In the field of physicochemical analysis, for the explanation of mechanism, the analysis of big data can provide reference and integration. The purpose of big data-based analysis is to remove redundancies and inconsistencies in the given raw characterization data to better improve its completeness and accuracy. The paper takes the density matrix of quantum bits, which is different from the classical real number representation, etc. as the representation of feature data, and investigates the cleaning of feature data based on quantum representation from the perspective of the intersection of theoretical physics and computer science. This paper comprehensively describes the basic theories and methods of data mining. On the basis of understanding and analyzing a variety of data mining techniques, it focuses on the BP neural network-based data mining techniques to provide in-depth analysis and elaboration. Further, this paper proposes an improved algorithm with multi-algorithm advantage integration and joint optimization for the deficiencies in the BP neural network algorithm, which is fully demonstrated in the nonlinear function simulation. In addition, based on the HA-BP algorithm, this paper has designed anomaly data detection model and variable factor analysis model, and their reliability and practicality are also fully demonstrated in the nonlinear function simulation in the analytical study of physicochemical reaction mechanism.

Keywords: Big data; Physicochemical reaction; Reaction mechanism; BP neural network.

I. INTRODUCTION

A. Significance of the study

The rapid development of the Internet of Things (IoT) technology in recent years has strongly contributed to the formation of a digital and informatized society. There is a large amount of redundant and inconsistent data in these complex multi-source heterogeneous big data, which increases the difficulty of data processing. However, for the data aspect of physical chemistry, the previous big data network is a very effective way of processing for the study of reaction mechanisms. Data fusion is the fusion of heterogeneous data representing the same target entity in the real world into a single, consistent and clear representation, thus improving the completeness and accuracy of the given raw data. In this context, data mining based on big data has become a field with great potential for future development. Data mining is often called data mining, usually refers to the processing or analysis of raw data through intelligent algorithms, from a large number of random data to explore the hidden in which there are some special and unknown internal laws, and then classify, correlate, predict the target problem, and ultimately realize the efficient "value-added" data. Therefore, the data mining technology based on big data comes into being and shows vigorous vitality in today's information age, and its process mainly consists of three steps: data preparation, data mining and result representation and evaluation. In data mining, the basic database based on scientific books, scientific and technological papers, etc. is not a high-quality "gold mine", but an "ore" with low gold content, so how to effectively develop these low value density, high-capacity data ore to make it play a higher value, has become a major challenge in the current data mining industry. How to effectively develop these low-value density, high-capacity data ores and make them play a higher value has become a problem and challenge in the field of data mining.

B. Current situation at home and abroad

In the early data mining work, researchers mostly used the joint approach of database and symbolic machine learning for their research. However, they are constrained by the problems of many data variables, complex nonlinear relationships, and noisy data interference in real cases, which in turn make these methods less generalizable. In recent years, through the continuous efforts of scholars at home and abroad, many different data mining related algorithms have been proposed. Among them, artificial neural network methods have significant advantages and good practicality in the field of data mining. In particular, the BP neural network algorithm is more widely used. The advantage of the BP neural network algorithm in the field of data mining is highlighted...
thanks to the fact that it has a signal processing ability similar to the function of real biological brain. This gives it good association, storage and self-learning capabilities. It is more suitable for nonlinear and complex problem solving. BP neural network can effectively explore the intrinsic law of the information system by learning the information architecture stored in the database, and then analyze and predict the information system without the need to understand the complex mechanism characteristics inside the information system. The existence of these advantages has led to its rapid promotion and application in the fields of industrial engineering, discipline intersection, and optimization of complex problems.

In the field of physical chemistry, the reaction properties of molecules are one of the molecular conformational relationships that researchers focus on. This is mainly due to the fact that this property determines the progress and effect of the reaction and has a great influence on the prediction of the reaction mechanism. Therefore, in recent years, many studies have focused on the establishment of predictive models associated with the molecular structure for this kind of index. For the modeling of molecular structure, some scholars chose molecular topological indices, such as molecular surface area, electron accessibility and connectivity index, as the input parameters and established an artificial neural network model for molecular weight prediction. Some scholars have proposed optimization schemes through this model pair. In a recent work, Guo et al. also built an artificial neural network model for predicting pure compounds based on the QSPR strategy. In terms of input parameter selection, the study chose topological indices that are highly correlated with the physicochemical reaction process to characterize the physicochemical properties of the molecular reaction mechanism during the actual operation. The study pointed out that the topological indices related to molecular bonding energies have a relatively large influence on the reaction process. A predictive model of the research method for hydrocarbons as well as oxygen- and nitrogen-containing compounds was developed using two databases. The model was based on the support vector machine regression method and also the topological indices were screened by a multi-objective feature selection method, which gave the best input set in terms of the minimum set of topological indices combined with the best prediction accuracy. In a recent work, Wang et al. used the CME-L-Isomap flow learning algorithm based on the combination of dimensionality reduction algorithm and artificial neural network algorithm to downscale the octane data of NIR spectra, and then substituted them into BAS-BP neural network for the prediction model building. The prediction model outperforms other models in terms of mean square error, mean absolute percentage error, correlation coefficient and running time.

Besides, in other molecular physical and thermochemical properties that are also crucial to the reaction system, domestic and foreign researchers have also carried out rich big data analysis studies. For physical properties, some scholars have used artificial neural networks combined with QSPR methods to establish predictive models for properties such as boiling point, flash point, liquid density, surface tension, and other properties of reactants. Some scholars also conducted cross-validation through consensus modeling to reduce the influence of uncertainty on the prediction accuracy of the model. For reaction energy balance calculations, Zhang et al. used a fast error back-propagation algorithm with 25 widely used molecular topological indices as input parameters for molecular structure description to build a predictive model for the standard generation of braising for 185 acyclic alkane molecules. Subsequently, the study used a genetic algorithm to optimize the prediction model by simplifying the input parameters to 17 topological indices. In addition to single molecular properties, in a recent study, Shen et al. used a combination of genetic algorithms and multiple linear regression to model and predict the upper limit of reaction of a mixture from the molecular structures of the reaction mixture components for data analysis and gave the optimal set of molecular topological indices as input parameters.

C. Big data analytics research in reaction mechanism construction

In research fields such as combustion, catalysis, biochemistry, and electrochemistry, chemical reaction mechanisms (networks) are a fundamental tool in reaction studies, supporting researchers in understanding the principles behind processes such as pollution, diseases, and processes. As shown in Fig. 1, a chemical reaction mechanism is mathematically modeled as a set of nonlinear ordinary differential equations describing the dynamic behavior of a system, characterizing the route from reactants to products, and elucidating the intermediates involved in the route as well as the reaction rates at each step. Thus, chemical reaction mechanisms have been extensively studied since the formulation of the law of mass action for the purposes of reaction path investigation and optimization of reaction processes. As early as the 1960s, linear algebra was used to analyze the nature of chemical reaction networks, establishing the basic mathematical tools that can be used to model and analyze reaction mechanisms, and the target factor analysis method subsequently proposed by Bonvin et al. has
been widely used for the identification of linearly independent reaction quantities and the determination of stoichiometry of chemical reactions.

![Generic equation](image)

**Fig. 1** Schematic representation of the chemical reaction mechanism and its differential equation forms

Traditionally, the establishment of chemical reaction mechanisms has relied on expert knowledge, requiring the structure of the mechanism, i.e., the reaction pathway, to be first proposed based on specialized chemical knowledge, and then fitted or calculated to determine parameters such as rate constants based on the given available data. A great deal of work has been carried out by researchers in this area to invent chemical reaction mechanisms and kinetic models from experimental data. Some scholars have developed methods such as the Reaction Mechanism Generator, which uses reaction flux analysis to identify important reaction pathways. However, the existing knowledge of the reaction pathways of these processes and the values of parameters such as rate constants is insufficient, which hinders the in-depth study of the reaction mechanisms of the relevant processes. In this context, many researchers have proposed that the reaction pathways can be revealed and the rate constants determined by fitting the experimental data to obtain the coefficients to be determined in the equations based on the law of mass action from the canonical form of the differential equations of chemical reaction mechanisms. This approach is based on big data analysis and predictive modeling of species concentration data, an early example of which is Burnham’s statistical analysis work based on concentration data and hypothetical reaction models. There has also been work on optimizing response models based on least squares and regularized least squares.

In some recent works researchers tried to get this class of data-driven chemical reaction mechanism models by using more appropriate algorithms and optimization strategies. Overseas scholars developed a chemical reaction network construction method using mixed integer linear programming and validated the method using the Van de Vusse reaction and the methanol decomposition reaction of glycerol triglyceride cool. Other scholars have inferred the reaction network structure, i.e., stoichiometric number, by analyzing potential reaction vectors and their likelihood measures. Based on big data analysis methods, many automated methods for constructing reaction mechanisms have been proposed, and the related work has made great progress in the breadth of application. With the booming development of topology and rigid equation solving in recent years, the data-driven construction of chemical reaction mechanisms will surely make further breakthroughs in the depth of research, and play a greater role in the fields of physical and chemical reactions and biology.

II. **BP NEURAL NETWORK ALGORITHM AND OPTIMIZATION**

**A. Overview of BP Neural Networks**

BP neural network, is a typical multi-layer feed-forward (usually three-layer) artificial neural network. Structurally, it consists of input, hidden and output layers with nodes in each layer, and the nodes in neighboring layers are connected by weights, but the nodes in each layer are independent of each other. Theoretically, it has been proved that a BP neural network with a single hidden layer can approximate any nonlinear function with arbitrary accuracy. In practice, the BP neural network with a single hidden layer can also meet the needs of computation very well. The structure of the BP neural network is shown in Figure 2.
Fig. 2 Sketch of the structure of BP neural network

In the figure, Xn is the input variable, Wij and Wjk are the number of nodes in the input, implicit and output layers respectively, Yn is the output value of the output layer.

B. Overview of BP neural network operation

The operational framework of the BP neural network can be specifically categorized into the following parts;

1) Determination of the network topology (input one implied one output) and initialization of the associated parameters;

2) Learning of BP neural network, the information is passed from input layer to output layer through implicit layer, and the actual output of this process network is output after calculation by output layer;

3) Calculate the error between the actual output of the network data and the expected output of the sample data, and back-propagate the error information to the input layer, and adjust the weights of each layer;

4) Cyclic iteration (2-3) of the two processes to gradually reduce the calculation error until the error reaches the set target error or the number of cyclic iterations reaches the set maximum number;

5) The optimal weights are obtained;

6) Extract the input information of the test samples, and with the optimal weights obtained in step (5), the predicted output of the test samples can be calculated.

C. Optimization analysis of BP neural network algorithm

1) Additional momentum terms

When the network calculates the error, the error begins to back propagate from the output layer to the input layer. In the process, the network will automatically adjust the weights and the gap value between the layers according to the adverse effect of the error to the situation, in order to effectively avoid the error flat zone generally used to improve the adjustment of the weights and the gap value by attaching a momentum term. Its specific expression is as follows:

$$\Delta \omega_{jk}(n+1) = -\eta \frac{\partial E(n)}{\partial \omega_{jk}(n)} + \lambda \Delta \omega_{jk}(n)$$

$$\Delta b_k(n+1) = -\eta \frac{\partial E(n)}{\partial b_k(n)} + \lambda \Delta b_k(n)$$

$$\Delta \omega_{ij}(n+1) = -\eta \frac{\partial E(n)}{\partial \omega_{ij}(n)} + \lambda \Delta \omega_{ij}(n)$$

$$\Delta b_j(n+1) = -\eta \frac{\partial E(n)}{\partial b_j(n)} + \lambda \Delta b_j(n)$$
is the momentum factor, usually a constant from 0 to 1.

2) Dynamic adjustment of adaptive learning rate

From the above theoretical analysis of the operation mechanism, it can be seen that the learning rate is positively related to the adjustment amount of the weights. That is, when the learning rate parameter is selected too small, the adjustment amount of the network to the weights will also become smaller. The operation of the network is extremely prone to oscillation. In order to effectively solve the above problems, the method of adaptive learning rate is proposed to improve the insufficiency of fixed learning rate parameter. The method is to adjust the learning rate in real time according to the error calculated by the network after each iteration, i.e., the learning rate changes its value during the operation of the network. The specific tuning rules for the adaptive learning rate used in this paper are as follows:

\[
\eta(n + 1) = \begin{cases} 
0.7\eta(n) & \text{mse}(n) > \text{mse}(n - 1) \\
\eta(n) & \text{mse}(n) = \text{mse}(n - 1) \\
1.2\eta(n) & \text{mse}(n) < \text{mse}(n - 1)
\end{cases}
\]

where \(\text{mse}()\) represents the error between the output value calculated by the network and the desired output value.

D. Test of optimization of BP neural network algorithm

Based on the theoretical analysis of BP neural network algorithm, operation mechanism and many common improvement algorithms, HA-BP neural network model with the advantage of integrating multiple algorithms is designed at the theoretical level, which not only has the excellent global optimization seeking performance of verifying PSO and GA, but also has the fast convergence ability of A-BP algorithm. In order to the performance of HA-BP, the commonly used and typical nonlinear trigonometric functions are used for testing. Table 1 shows the data of the test samples and Figure 3 shows the validation view.

![MSE graph for the number of nodes in the hidden layer](image)

**Fig. 3** MSE graph for the number of nodes in the hidden layer

**Table 1.** Function test sample data

<table>
<thead>
<tr>
<th>training sample</th>
<th>X1</th>
<th>Y1</th>
<th>test (machinery etc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.542</td>
<td>13.256</td>
<td>191.228</td>
</tr>
<tr>
<td>2</td>
<td>12.458</td>
<td>14.542</td>
<td>205.767</td>
</tr>
<tr>
<td>3</td>
<td>13.685</td>
<td>15.227</td>
<td>254.378</td>
</tr>
<tr>
<td>4</td>
<td>14.986</td>
<td>16.884</td>
<td>256.587</td>
</tr>
<tr>
<td>5</td>
<td>15.223</td>
<td>17.112</td>
<td>275.346</td>
</tr>
<tr>
<td>6</td>
<td>16.745</td>
<td>18.265</td>
<td>278.321</td>
</tr>
<tr>
<td>7</td>
<td>17.123</td>
<td>19.524</td>
<td>301.254</td>
</tr>
<tr>
<td>8</td>
<td>18.985</td>
<td>20.0365</td>
<td>335.264</td>
</tr>
</tbody>
</table>
III. REACTION MECHANISM CONSTRUCTION OF CHEMICAL REACTION NEURAL NETWORK BASED ON BIG DATA ANALYSIS

A. Reaction mechanism construction of chemical reaction neural network

In fields such as combustion biochemistry and electrochemistry, chemical reaction mechanisms are a fundamental tool in reaction studies, supporting researchers in their understanding of the principles behind processes such as contamination, disease and processes. Traditionally, the establishment of chemical reaction mechanisms has relied on expert knowledge, requiring reaction pathways to be first proposed based on specialized chemical knowledge, and then fitted or calculated to determine parameters such as rate constants based on given available data. However, the reaction pathways and rate constants of many processes are currently not well understood by researchers, and it is difficult to empirically study systems such as combustion and bioreactions that may contain tens of thousands of candidate reactions with a large number of intermediates. In this context, many researchers have suggested that it is possible to start from the canonical form of power law equations for chemical reaction mechanisms and fit experimental data based on power law equations to reveal reaction pathways and rate constants. Chemical reaction neural network is such a class of fitting method, but its training process has some difficulties due to reaction stiffness, data characteristics and other reasons. Therefore, a training strategy combining stochastic gradient descent method and genetic algorithm is developed based on the characteristics of the weight values in the chemical reaction neural network.

B. Principles of chemical reaction neural networks

For the primitive reaction \( \Delta v_A + v_B \Rightarrow v_C + v_D \), the rate constant \( k \) is known from the law of mass action, and the concentrations \( [A], \ [B], \ [C], \ [D] \) of substances \( A, \ B, \ C, \ D \) and the stoichiometric number of the chemical reaction \( v_A, v_B, v_C, v_D \) and the reaction rate \( r \) satisfy the relationship:

\[
 r = k[A]^{v_A}[B]^{v_B}
\]

And the relationship between substance yield and reaction rate is satisfied.

\[
\begin{align*}
\frac{d[A]}{dt} &= -v_A r \\
\frac{d[B]}{dt} &= -v_B r \\
\frac{d[C]}{dt} &= v_C r \\
\frac{d[D]}{dt} &= v_D r
\end{align*}
\]

In chemical reaction mechanisms, it is common to use the three-parameter form of the Arrhenius formula for the rate constant.

Make a representation:

\[
k(T) = AT^b \exp \left( -\frac{E_a}{RT} \right)
\]

Thus, the yield of a species can be expressed as an equation for temperature and concentration:

\[
\begin{align*}
\frac{d[A]}{dt} &= -v_A AT^b \exp \left( -\frac{E_A}{R_u T} \right) \cdot [A]^{v_A}[B]^{v_B} \\
\frac{d[B]}{dt} &= -v_B AT^b \exp \left( -\frac{E_A}{R_u T} \right) \cdot [A]^{v_A}[B]^{v_B} \\
\frac{d[C]}{dt} &= v_C AT^b \exp \left( -\frac{E_A}{R_u T} \right) \cdot [A]^{v_A}[B]^{v_B} \\
\frac{d[D]}{dt} &= v_D AT^b \exp \left( -\frac{E_A}{R_u T} \right) \cdot [A]^{v_A}[B]^{v_B}
\end{align*}
\]

As a result, the three empirical parameters required to calculate the number of reaction steps (stoichiometries) and rate constants are obtained by fitting this equation to the experimentally or computationally obtained data for species concentration, yield, and temperature, and the mechanism of the reaction is constructed.

The basic idea of a chemical reaction neural network is to represent the above equation through neurons. Corresponding the number of reaction levels and rate constants to the values of the weight gaps between neurons allows us to solve for these unknown quantities through the neural network training process. A neural network containing a single hidden layer neuron is represented.
\[
\begin{align*}
\frac{d[A]}{dt} &= - v_A \exp \left( v_A \ln [A] + v_B \ln [B] + 0 \ln [C] + 0 \ln [D] + b \ln (T) - \frac{E_a}{RT} + \ln (A) \right) \\
\frac{d[B]}{dt} &= - v_B \exp \left( v_A \ln [A] + v_B \ln [B] + 0 \ln [C] + 0 \ln [D] + b \ln (T) - \frac{E_a}{RT} + \ln (A) \right) \\
\frac{d[C]}{dt} &= v_C \exp \left( v_A \ln [A] + v_B \ln [B] + 0 \ln [C] + 0 \ln [D] + b \ln (T) - \frac{E_a}{RT} + \ln (A) \right) \\
\frac{d[D]}{dt} &= v_D \exp \left( v_A \ln [A] + v_B \ln [B] + 0 \ln [C] + 0 \ln [D] + b \ln (T) - \frac{E_a}{RT} + \ln (A) \right)
\end{align*}
\]

When the chemical reaction mechanism is unknown, the neural network can be trained on random initial values of the weights based on the data of species concentration, yield and temperature. After convergence of the errors the neural network represents a set of differential equations for the chemical reaction mechanism behind the data, which can then be backpropagated to obtain the reaction mechanism. Further, by stacking a plurality of neurons into a hidden layer, the chemical reaction neural network can represent a multi-step chemical reaction mechanism.

C. Example of reaction mechanism construction

A hypothetical reaction mechanism was selected for an example of the specific process of mechanism construction. As shown in the following equation, the mechanism includes six chemical species labeled x1, x2, x3, x4, x5, x6, involving four reactions with rate constants of k1 = 0.2, k2 = 0.1, k3 = 0.15, and k4 = 0.05, where the fourth reaction is the inverse of the third.

\[
\begin{align*}
\frac{d[A]}{dt} &= - k_1 x_1^2 + x_4^2 = 0.0500 \\
\frac{d[B]}{dt} &= - k_2 x_2 + x_3 = 0.2001 \\
\frac{d[C]}{dt} &= k_3 x_1 + x_4 = 0.15 \\
\frac{d[D]}{dt} &= k_4 2x_6 = 0.05 \\
\end{align*}
\]

The initial reactants were set as x1 and x2 for the numerical calculations, and in addition, since the rate constants in the mechanism are independent of temperature, the temperature was not added as a variable to the input data when constructing the neural network model, i.e., the simplified chemical reaction neural network represented in Fig. 2 was used. A total of 20 experiments were simulated for the numerical calculations, with initial concentrations given randomly for each experiment. From the species concentrations and their yields were sampled every 2-time units in the interval up to 20-time units, thus generating a total of 300 sets of data. Out of these data, 200 sets i.e. 67% were used as training sets to train the neural network and the remaining 100 sets i.e. 33% were used as test sets to validate the prediction performance.

The reaction paths and stoichiometric numbers can be deduced from the weight matrix after the training is completed, and the rate constants for the corresponding reactions can be deduced from the threshold matrix. From the first row of the powder matrix, it can be seen that in a particular reaction x2, x3 are reactants with stoichiometric ratios all of 1 and x5 are products with stoichiometric ratio of 1. From the first column of the matrix, it can be seen that the rate constant of the reaction is exp^(-2.3025) = 0.1000. Therefore, this reaction can be written as.

\[
\begin{align*}
\frac{k=0.1000}{x_2 + x_3} &\Rightarrow x_5 \\
\end{align*}
\]

Similarly, we can write the other response equations corresponding to the weight threshold matrix are.

\[
\begin{align*}
\frac{k=0.0500}{2x_6} &\Rightarrow x_1 + x_4 \\
\frac{k=0.2001}{x_1 + x_2} &\Rightarrow x_3 + x_4 \\
\frac{k=0.1499}{x_1 + x_4} &\Rightarrow 2x_6 \\
\end{align*}
\]

D. Input parameter ordering and suggestions for optimization

The independent variables of a one-dimensional active subspace prediction model are linear combinations of the input parameters, so when the final prediction model is built with monotonic characteristics, the coefficients of each parameter in that linear combination can be considered as their importance weights on the output. The absolute value of the coefficients determines the importance of the parameter to the output, while the sign of the coefficients determines the direction of the influence. The values of the w1 element, i.e., the coefficients of the linear combination of the input parameters, in each prediction model are shown in Fig. 4, and their possible ranges of variation are represented as error bars in Fig. 5.

364
The effect of each input parameter on the rate of the physicochemical reaction is shown in Figure 5. Where the error bars indicate the variable range, it can also be seen that the accuracy of the model created in this study for the study of the physicochemical mechanism based on big data is high, which can reach more than 95%, and therefore it is of reference significance.

**Fig. 4** Coefficients of linear combination of input parameters

**Fig. 5** Predictive model performance corresponding to different samples of experimental data.

**IV. CONCLUSION**

The chemical reaction neural network method based on BP network can discover the reaction paths, calculate the kinetic parameters and construct the reaction mechanism from the time series data of species concentration.
yield and temperature independently. In this work, a genetic algorithm is combined with stochastic gradient descent method to propose a new strategy for chemical reaction neural network training by solving different characteristic parameters separately. Based on the chemical reaction neural network, a reaction mechanism for a hypothetical reaction was constructed by training the species concentration and yield data. The chromosome coding method in the genetic algorithm is simplified according to the relationship between the weight matrices during the training process, which significantly improves the training efficiency. It is verified that the combination of genetic algorithm and stochastic gradient descent method to solve different parameters can effectively reduce the negative influence of data and network structure on the training of chemical reaction neural network, which is important for the optimization of the mechanism of physical and chemical reaction.

REFERENCES


