ICIC Express Letters
Part B: Applications
Volume **14**, Number **4**, April **2023**

ICIC International ©2023 ISSN 2185-2766
pp. **359–367**

FEATURE SELECTION BASED ON COUNTERFACTUAL REASONING

Lili Tao*, Zhihua Hu and Miao Huang*

School of Intelligent Manufacturing and Control Engineering
Shanghai Polytechnic University
No. 2360, Jinhai Road, Pudong District, Shanghai 201209, P. R. China
*Corresponding authors: { lltao; huangmiao }@sspu.edu.cn; zhhu@sspu.edu.cn

Received September 2022; accepted December 2022

Abstract. *The high-correlation features with target variable are often selected in prediction in the industrial process. And the distribution of the target variable is predicted by observing the distribution of the correlation features. However, feature selection based on correlation cannot explain well when making decisions and judgments. The influence of different features on the target variable can be shown by the causality. Through causal analysis, the change of the target variable can be explained reasonably, which is helpful for decision-making and judgment. In this paper the counterfactual reasoning feature selection (CRFS) algorithm is proposed to select more representative features based on counterfactual logic. Based on the operating data of a petrochemical polyethylene plant, CRFS is compared with the commonly used feature selection methods. The results indicate that CRFS has better performance in prediction, and the causal structure determined by this method is also reasonable.*

Keywords: Industrial process, Causal analysis, Counterfactual reasoning, Feature selection

1. **Introduction.** Polyethylene is a high-yield polymer material product. Recently, the global polyethylene production capacity is steadily increasing. In order to enhance the competitiveness of products, operating optimization is an important aspect. One of the key issues in the optimization process is how to select operating parameters that can affect production capacity and energy consumption [1]. Feature selection is a preprocessing step of data mining [2]. Generally speaking, redundant or irrelevant features can be filtered out through the preprocessed data [3], which can simplify the process model and make the model reasonable. It is noticed that data preprocessing is very important during modeling the chemical process, and how to select reasonable features is also a problem that needs to be considered. Features related to target variables are often chosen by algorithms such as principal component analysis (PCA), and maximal information coefficient (MIC) [4,5]. However, feature selection methods based on correlation or information theory cannot determine the impact of input features on output, and thus may make wrong decisions and judgments. The causal analysis can filter the characteristics to find the cause of the change of the target variable, so as to achieve the control of the target variable.

This paper aims to conduct feature selection through causal analysis of observational data, and the CRFS method is proposed. First, the direct influence between features and target variables is analyzed through counterfactual reasoning in causal analysis [6], and then the features are selected based on this method. After that a directed causal structure diagram is determined. Correlation and causality have certain differences and connections. Causality can explain some correlation problems, but it is not the only explanation for correlation. Therefore, additional counterfactual dependencies are needed to confirm causality. Secondly, in order to verify the effectiveness of the proposed feature

DOI: 10.24507/icicelb.14.04.359

selection method, this paper conducted several experiments using operating data of a real petrochemical polyethylene plant. The experiment includes comparison of prediction accuracy with currently popular feature algorithms and comparison of selected feature numbers.

2. Causal Reasoning.

Feature selection is a very important preprocessing process in the field of machine learning and data mining which can directly affect the quality of the model. There are two main methods of feature selection: filters and wrappers [7]. The filter sorts the attributes according to some requirements, and filters out some of the top attributes. The wrapper selects certain attributes through iterative search to gradually improve model performance. Wrappers such as PCA and information-based MIC are usually used [8-10]. Due to the huge number of chemical process parameters, it is better to use the filter method for feature selection in this paper.

2.1. Causality.

Causality is an attempt to describe the relationship between two events, that is, an event makes a certain result more likely to appear, and a certain result will not appear without it or that it can produce a certain result under certain conditions. If this event causes a certain result, then it is considered that there is causality between this event and the result. Figure 1 shows a simple relationship network. It can be seen in this network, X_2 and X_4 have direct effects on the output Y, while X_1 and X_3 have a certain correlation with Y but not have a substantial effect on Y. Since X_1, X_2, X_3 and X_4 have strong correlation with Y in the observation data set, if the analysis method based on correlation is used for feature selection, it is very likely that X_1, X_2, X_3 and X_4 will be selected as the selection result. So the disadvantage is that the two relationships cannot be distinguished. However, in the industrial process, if you want to make decisions and judgments, it is valuable to distinguish these two types of relationships. The existence of causality usually means that the event and the result are related, but correlation does not necessarily mean that the event and the result are causal. Therefore, the causality can explain the correlation to some extent, but it is not the only explanation for the correlation. There are many methods of causal reasoning, including observation method, calculation method and experimental method. Those who are interested in some related methods can refer to [4] and the cited literature.

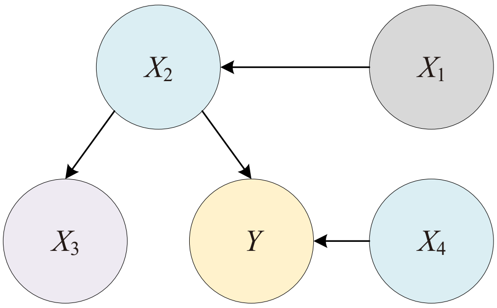


Figure 1. Relational network

For the two variables X and Y, the causal relationship can be divided into four situations, namely, X is the cause of Y, Y is the cause of X, X and Y have no causal relationship, and X and Y are mutually causal.

2.2. Counterfactual reasoning.

The basis of counterfactual reasoning is that if E is caused by C, then two conditions must be satisfied: the appearance of E is determined by C, that is, if C appears, then E will also appear; if C does not appear, then E will not appear. The cause variable C and the result variable E are both binary. These two conditions include the sufficiency and necessity of the cause. Usually, probabilistic counterfactual reasoning is often used to dig out causality from observational data. Formulae (1) and (2) are used to express the necessity and sufficiency of the reason [6]:

$$P(y_{x=0} = 0|x = 1, y = 1) \tag{1}$$

$$P(y_{x=1} = 1|x = 0, y = 0) \tag{2}$$

Formula (1) describes the probability that there is no event y if there is no event x (x is the reason of y), which is the necessity of the reason. Formula (2) describes the probability that x causes y to occur, which is the sufficiency of the cause. The necessity and sufficiency of cause are different aspects of causality. $y_x = 0/y_x = 1$ means the value of y when $x = 0$ ($x = 1$), which is a different concept from the conditional probability $P(y|x = 0)$. The former is also called the "do" operation, which means the probability of y when only x is considered with other variables fixed under experimental conditions. While the conditional probability represents the probability of y with different conditions of x under natural conditions. This natural condition does not need to fix other variables, that is, there is no restriction on the value of other variables. Under the "do" operation, since the values of other variables are controlled, the pure causal relationship between x and y can be seen. Under natural conditions, since other variables are not controlled, the change of y comes from two aspects, one is directly caused by the change of x, and the other is indirectly caused by other variables or caused by other reasons. This phenomenon is called confounding.

The calculation of Formulae (1) and (2) is an important research content in counterfactual reasoning. However, since there are many research directions, there is no general calculation method. [4] introduces a "monotonicity" assumption that can be satisfied in many situations:

$$y_{x=1} \geq y_{x=0} \tag{3}$$

The meaning of monotonicity is that the effect y after taking a certain measure will not be lower than the effect of not taking a certain measure. Therefore, counterfactual reasoning is carried out through inequalities. Under the condition that the "do" operation can be performed, if the inequality can be satisfied, then x is the cause of y. Unfortunately, in the actual observation data set, the "do" operation is difficult or impossible to achieve since confounding phenomena will interfere with the calculation of the real cause. Therefore, how to identify the confounding in the observation data to find out the real cause and effect is of practical significance.

3. **Feature Selection Method Based on Counterfactual Reasoning.**

When making decisions and judgments, the first step is to analyze which features have an impact on the results. Features can be selected through observation data by using counterfactual reasoning methods. In the industrial process, it is difficult to keep other variables constant to conduct a control experiment. Meanwhile, not only whether the feature changes should be concerned, but also the change state of the feature. In a complete device operation process, the feature state should have three states: stable, rising and falling. Therefore, for the different states of x, based on the principle of monotonicity formula (3), if x is a positive cause of y, with the mixed effects of other variables being considered, x and y should satisfy Formula (4):

$$\begin{cases} P(y = 0|x = 0) \geq aP(y = 0|x = 1) \\ P(y = 1|x = 1) \geq aP(y = 1|x = 0) \end{cases} \tag{4}$$

If x is the cause of a negative effect on y, then Formula (5) is satisfied:

$$\begin{cases} P(y = 0|x = 0) \leq 1 - (aP(y = 0|x = 1)) \\ P(y = 1|x = 1) \leq 1 - (aP(y = 1|x = 0)) \end{cases} \tag{5}$$

where $x = 1$ indicates that the state of variable x is rising, $x = 0$ indicates that the state of variable x is falling, and a is the confounding influence coefficient, the purpose of which is to consider the confounding factors in the observed data. Generally speaking, if the confounding phenomenon is stronger, then the analysis of a feature will be more disturbed. Therefore, the conditions for the establishment of the inequality need to be looser, i.e., the confounding influence coefficient is smaller. The meaning of the inequality is that if the state of the feature x is rising, the probability that the state of the output y will rise is greater than α times the probability that the state of the output y rises when x falls. When the state of feature x is falling, the probability that the state of output y will also fall is greater than a times the state of output y when x rises. In this way, x is considered to be a positive cause of y. The coefficient a is calculated according to Formula (6) [6]. In the industrial process, the state of characteristic variables includes 1) $y = 1$ is caused by $x = 1$, not $x = 0$; 2) $y = 1$ is not only caused by $x = 1$, but also by other variables.

$$a = \frac{N(y = 1)}{N(y = 1) - N(y = 1|x = 1) + \varepsilon} \tag{6}$$

where $N(y = 1)$ is the statistical number of the rising state of the target variable y, $N(y = 1|x = 1)$ is the statistical number of $y = 1$ in the case of $x = 1$, $N(y = 1|e)$ represents the number of $y = 1$ caused by other conditions except $x = 1$, and ε is a small value added to avoid zero denominator. Through this calculation, a coefficient can be obtained, which can describe the degree of influence of other variables in the analysis process. For the features that satisfy this type of inequality, it is considered that the feature is the cause of the output Y to a certain extent, so the feature needs to be retained.

It should be noticed that the retained feature set does not fully reflect the causal structure of the process. To determine the cause variable of Y (the selected feature set), it is also necessary to confirm the relationship of the features. Counterfactual reasoning is also used to analyze the causal relationship between features and finally determine the causal structure of the process. The selected feature subset can not only provide more accurate prediction accuracy for the target variable, but also provide an explanation for the change of the target variable.

4. **Experimental Simulation.**

4.1. **Data set.**

The data in this study comes from the real operating data of a petrochemical polyethylene plant in the first half of 2013, with a total 3700 data sample. The entire polyethylene plant includes multiple reactors, such as prepolymerization reactor R301, and loop reactor R302. Take the reactor R301 for example; there are 28 input features of the R301, as shown in Table 1. In order to reduce the impact of lost data on the model, the lost data samples are deleted and some abnormal points and outliers are eliminated through the 3σ criterion. The data is converted into the difference with the sampled data at the previous time, and the fluctuation of each feature is analyzed through its distribution, and the continuous value of each data sample is converted into discrete values 0, 1, 2. Among them, 1 means rising, 2 means falling, and 0 means stable.

4.2. **Sliding test.**

Considering that in the actual operation of the chemical process, the operating state at the previous moment may have an impact on the operation at the next moment, so when performing regression prediction, the operating data at the previous moment should be considered. In this study, the sliding window method is used to divide the data samples into different training sets and test sets, and the time series method long short-term memory (LSTM) is used for experiments. Based on the analysis of the data, the window size is set to 5. The detailed division process is shown in Figure 2. The

Table 1. Input feature information

ID	Feature
1	Inlet ethylene flowrate of reactor R301
2	Inlet hydroge flowrate of reactor R301
3	Inlet cycle propane flowrate of reactor R301
4	Fresh propane flushing flowrate of catalyst feed valve xv3004
5	Fresh propane flushing flowrate of catalyst feed valve xv3005
6	Total catalyst feed to reactor
7	Cocatalyst feed of reactor R301
8	Temperature of reactor R301
9	Pressure of reactor R301
10	Total catalyst feed
11	Butene feed rate of prepolymerization reactor
12	R301 feed system-R301PRESS
13	R301 feed system-R301TEMP
14	Total feed flow of prepolymer
15	Prepolymerized polyethylene feed
16	Prepolymerized propane feed
17	Prepolymerized polybutene feed
18	Prepolymerized hydroge feed
19	Prepolymerized outflow
20	Prepolymerized polyethylene outflow
21	Prepolymerized propane outflow
22	Prepolymerized butane outflow
23	Prepolymerized hydroge outflow
24	Outflow of prepolymer liquid phase
25	Outflow of prepolymer solid phase
26	Residence time of prepolymer
27	Residence time of prepolymer liquid phase
28	Total prepolymer

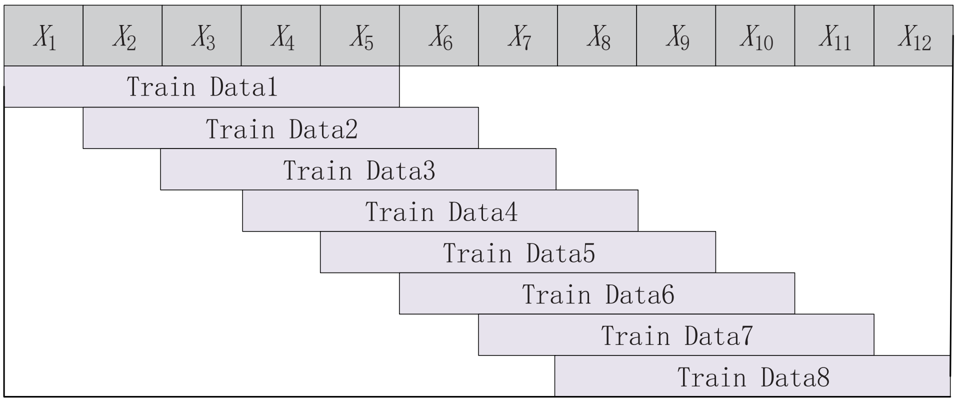


Figure 2. Sliding window partition data set

data set is divided by sliding window, and the last 20% of the divided data set is taken as the test set and the first 80% is taken as the training sample.

4.3. **Experimental results.** In this study, in order to ensure the reliability of the training results, the prediction model was trained and tested for many times, and the average value of the five prediction results of the data set was taken. Other feature selection algorithms were used for comparison. In Table 2, the mutual information coefficient between the features is calculated by the MIC. The results are compared with the feature number selected based on Formulae (4) and (5). Table 3 shows the feature selected by counterfactual reasoning and the feature selected based on the maximum information coefficient with a correlation coefficient greater than 0.8. The black font feature is the consistent feature selected by the two methods. The italic font feature is the unique feature of the correlation selection, while the bold font is a unique feature of counterfactual reasoning selection. The features selected by MIC (greater than 0.8 in Table 2), CRFS, PCA dimensionality reduction, original features based on Pearson correlation coefficient (PCC) and no feature selection (NoFC) are performed regression analysis. Then the causal structure of the selected sub-feature set is determined, and finally the prediction accuracy is given on the basis of counterfactual reasoning and selection based on correlation features. The measure of prediction accuracy used in this paper is mean square error (MSE). MSE refers to the expected value of the square of the difference between the estimated value and the true value of the parameter, and can evaluate the change degree of the data. The smaller the MSE value, the better the accuracy of the prediction model in describing the experimental data.

Table 2. Correlation coefficient between features and output

Feature_ID	Real_α	α	MIC	Feature_ID	Real_α	α	MIC
1	1.72	1.64	0.93533903	15	1.60	1.62	0.9374706
2	0.94	1.39	0.72711892	16	0.54	1.29	0.69803401
3	1.01	1.30	0.64974575	17	2.56	1.79	0.62614839
...				...			
14	2.39	1.79	0.92736246	28	0.86	1.45	0.73154304

Table 3. Results of CRFS and MIC feature selection

Feature_ID	MIC	Feature_ID	MIC	Feature_ID	MIC
1	0.93533903	13	**0.71190448**	19	0.82324013
6	0.88979186	14	0.92736246	20	*0.9374706*
8	**0.70840817**	15	*0.9374706*	22	**0.62549919**
10	0.88260175	17	**0.62614839**	24	0.95658882

To carry out further causal result confirmation on the selected feature set, each feature is traversed and performed by counterfactual reasoning. As mentioned in Section 2.1, the causality is divided into four situations. Therefore, the final causal structure is adjusted for different situations as shown in Figure 3, and this sub-feature set is selected for output prediction. The prediction results are shown in Figure 4.

The cause to the result is drawn by the one-way arrow (from cause to result), and the two-way arrow describes the causal relationship between the two variables. Mutual causality can be described in industrial processes as under stable operating conditions. In order to ensure the full progress of the chemical reaction of the reactants, it is necessary to adjust the other reactants in proportion, and they interact with each other.

Figure 5 shows that the number of feature selections based on counterfactual reasoning may not be the smallest. In some cases, it will be slightly more than MIC, but it will not affect its prediction accuracy. The prominent advantage based on causal analysis lies in the ability to describe the process. And under the premise prediction accuracy, the structure diagram determined by the causal analysis can provide knowledge source for decision-making and judgment. According to the causal structure in Figure 3, it can be seen that

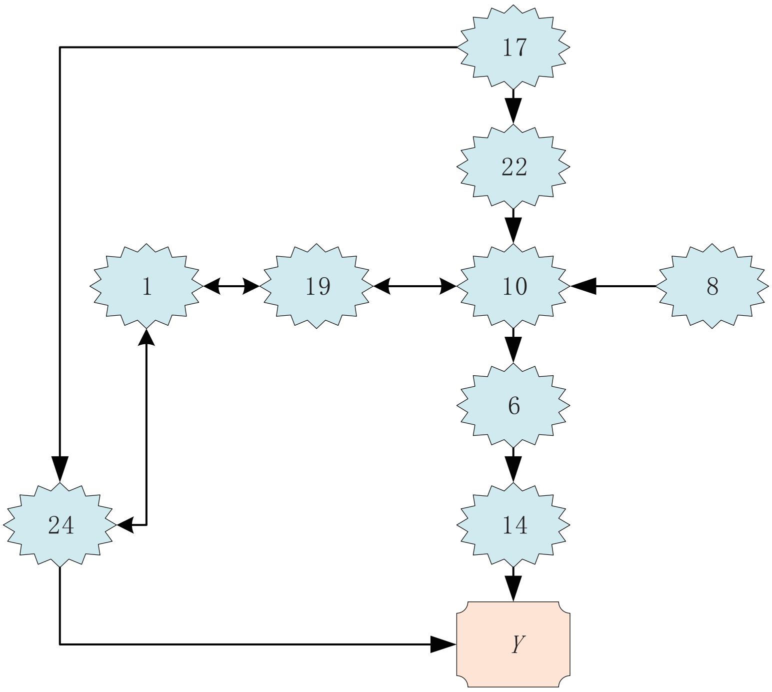
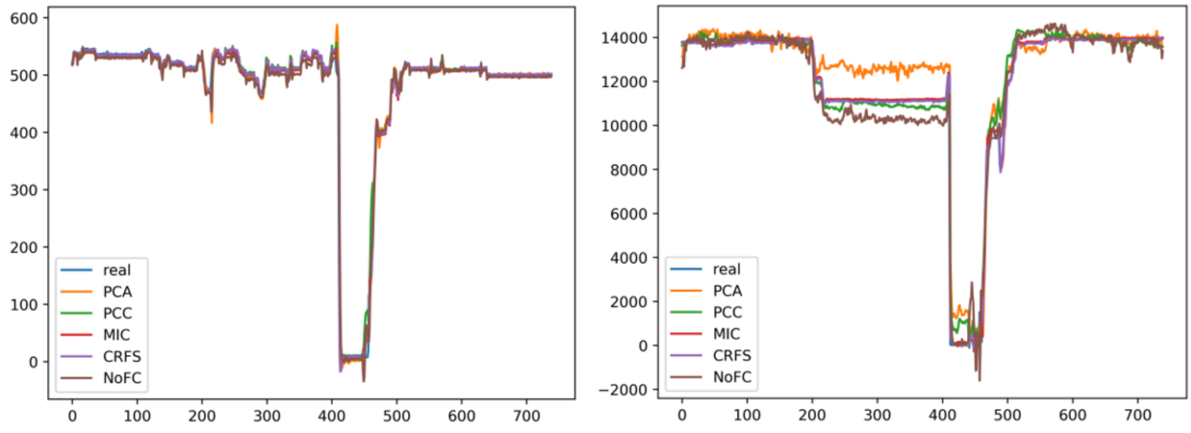


FIGURE 3. Causal structure diagram



(a)

(b)

FIGURE 4. (color online) (a) Prediction results of R301 device by different methods; (b) Prediction results of R302 device by different methods

the complete causal structure contains multiple loops, such as $\langle 1, 6, 10, 14, 17, 19, 24, Y\rangle$ and $\langle 1, 10, 17, 19, 22, 24, Y\rangle$. Since the two-way arrows indicate that there is a mutual causal relationship between the features, it is necessary to give priority to the two-way arrows and then the one-way arrows when choosing a loop. The choice of the loop means that when you need to adjust Y, you can choose the scheme with the smallest two-way route, such as $\langle 1, 6, 10, 14, 17, 19, 24\rangle$, that is, if you need to make a decision on variable Y, all the variables of this loop should be adjusted accordingly. In this way, Y can be adjusted reasonably under sufficient reaction. Compared to other methods, the advantages of CRFS are that it can determine the causal structure and provide explanatory properties for the decision-making process.

In addition, it can be seen from Table 3 of whether the counterfactual reasoning or feature selection through MIC, have some same features, that is, features that can be selected by both methods. This also shows that causality and correlation are related to

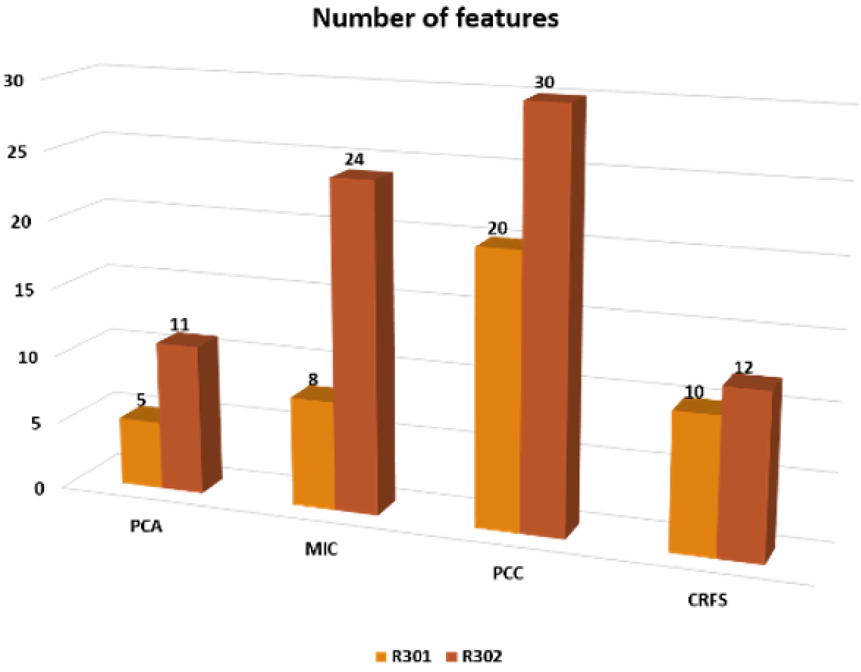


Figure 5. Number of features selecting for R301 and R302

a certain extent. Since mutual information analysis often needs to set the threshold of selected features, it is difficult to select some causal but weakly correlated features. For example, the mutual information coefficients of the features selected based on counterfactual reasoning in Table 3 is only 0.62. The reason is that in the chemical process, the effect of this type of features may offset the effect of other features. So the calculation result will not be very high, but in fact there is causality between them. Feature selection based on correlation cannot distinguish this, but counterfactual reasoning can directly identify some related features because it does not need to set a threshold or cumulative contribution. To a certain extent, CRFS is more interpretable than algorithms based on correlation. Therefore, it has better interpretation in the decision-making process. By determining the causal structure, you can know the characteristics that need to be noticed in the decision-making process.

5. **Conclusions.** When making decisions and judgments on industrial processes, choosing appropriate features can ensure the rationality of decisions and judgments. Traditional feature selection algorithms are mainly based on the characteristics of data correlation. Although these algorithms can guarantee the predictive performance to a certain extent, they cannot describe the internal relationship between features and output, nor can they describe the potential causality between them. Therefore, such methods often fail to provide reasonable explanation during the decision-making process. This paper proposed a feature selection method based on counterfactual reasoning. By comparing with some feature selection methods, the superiority of the prediction accuracy of this method is verified. The method in this paper can not only effectively identify the potential associated features in the industrial process, but also ensure the prediction accuracy of the model. Therefore, under the required prediction accuracy, the features selected by this method have better interpretation. In summary, the method in this paper has great potential in the application of industrial process decision-making.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (Grant No. 62203291 and Grant No. 62003205).

REFERENCES

[1] E. El-Kenawy and M. Eid, Hybrid gray wolf and particle swarm optimization for feature selection, *International Journal of Innovative Computing, Information and Control*, vol.16, no.3, pp.831-844, 2020.

[2] Y. Min, M. Ye, L. Tian et al., Unsupervised feature selection via multi-step Markov probability relationship, *Neurocomputing*, vol.453, no.3, pp.241-253, 2021.

[3] E. Hancer, B. Xue and M. Zhang, A survey on feature selection approaches for clustering, *Artificial Intelligence Review*, vol.53, no.2, 2020.

[4] S. Wang, Y. Xue and W. Jia, A new population initialization of particle swarm optimization method based on PCA for feature selection, *Journal on Big Data*, no.1, 2021.

[5] G. L. Sun, J. B. Li, J. Dai et al., Feature selection for IoT based on maximal information coefficient, *Future Generation Computer Systems – The International Journal of Escience*, vol.89, pp.606-616, 2018.

[6] K. Kuang, L. Li, Z. Geng, L. Xu et al., Causal inference, *Engineering*, vol.6, no.3, pp.253-263, 2020.

[7] H. Sun, J. Jin, R. Xu et al., Feature selection combining filter and wrapper methods for motor-imagery based brain-computer interfaces, *International Journal of Neural Systems*, 2021.

[8] A. Yadav et al., Breast cancer prediction using SVM with PCA feature selection method, *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, pp.969-978, 2019.

[9] X. Qiao, J. H. Bao, H. Zhang, F. H. Wan et al., Underwater sea cucumber identification based on principal component analysis and support vector machine, *Measurement*, vol.133, pp.444-455, 2019.

[10] W. Pan, Feature selection algorithm based on maximum information coefficient, *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2021.