Food & Nutrition Journal
Jiang T, et al. Food Nutr J. 6: 247.
https://www.doi.org/ 10.29011/2575-7091.100147
www.gavinpublishers.com

Research Article

Correlation Analysis and Research on Sampling Inspection Data of China's Grain-processed Products Based on CARMA Algorithm

Tongqiang Jiang[1,2] , Yongchun Jiao[1,2],Tianqi Liu1[2], Wei Dong [1,2,*], Qi Yang[1,2]

[1]National Engineering Research Centre for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing100048, China

[2]School of E-business and Logistics, Beijing Technology and Business University, Beijing 100048, China

[*]**Corresponding author:** Wei Dong, National Engineering Research Centre for Agri-Product Quality Traceability, School of E-business and Logistics, Beijing Technology and Business University, Beijing100048, China

Citation: Jiang T, Jiao Y, Liu T, Dong W, Yang Q (2022) Correlation Analysis and Research on Sampling Inspection Data of China's Grain-processed Products Based on CARMA Algorithm. Food Nutr J 7: 247. DOI: 10.29011/2575-7091.100147

Received Date: 16 May, 2022; **Accepted Date:** 24 May, 2022; **Published Date:** 26 May, 2022

Abstract

At present, there are many problems (huge workload of sampling inspection, difficult to determine the focus of the sampling inspection, etc.) in the supervision and sampling inspection of food safety. This paper aims to determine the key or non-key testing objects in the sampling work by mining the related rules of the national sampling data, so as to improve the inspection frequency of key objects and reduce the inspection frequency of non-key testing items. In this way, it can achieve the reasonable allocation of sampling resources and improve the efficiency of food sampling inspection. After cleaning and standardizing the data of major categories of grain-processed products in the national food sampling inspection data from 2019 to 2021, CARMA algorithm is used to explore the association rules of the data of food category (sub-subclass), sampling province, types of food packaging,testing site, pollutant category, relative risk of pollutants and so on. Ten valuable and valid strong association rules are obtained after the experiment. The results show that some key and non-key testing objects can be determined through exploring the association rules for relevant sampling inspection data, which can provide some reference for the allocation of sampling inspection resources.

Keywords: Data Mining; Food Sampling Inspection Data; Correlation Analysis

Introduction

China's national food sampling inspection work refers to the process that the State Council or provincial/municipal /county food supervision and administration departments carry out quality sampling inspection of produced and sold food according to law. It is one of the important technical supervision means for all kinds of food supervision. China spends a lot of manpower, material and financial resources on food inspections every year. It is one of the most concerned and urgent problems of the food and drug safety regulatory departments that how to scientifically carry out sampling inspection, scientifically determine the inspection items and determine the inspection frequency of hazardous substances [1]. Through the mining and analysis of food sampling inspection data, its internal connection can be found to provide a reference for the reasonable allocation of sampling inspection resources. Most of the existing literature has only studied the main polluting factors in food data in individual regions, and few have analyzed the correlation in food in multiple provinces across the country [2].

Hu, K and others [3] analyzed the food production and food-processing data based on scientific metrology theory and measurement, and discussed the characteristics of these two fields. Liang Lanxin, et al. [4] used the support vector machine and regression model to analyze the sampling inspection data

in Shenzhen, China, established the sampling batch distribution process, and used the fuzzy optimization model to obtain the corresponding sampling distribution weight. Qian Ziqi et al. [5] used the dynamic weight method and clustering analysis method to analyze the local sampling inspection data of Tianjin Food and Drug Administration in 2017. Sewekow E [6] analyzed the differences in the food sampling inspection data in different states of Germany and analyzed the impact of these differences on food safety. Most existing literature only studies the food data of individual regions, and rarely conducts data mining and analysis of certain kinds of food in many provinces.

Correlation analysis of data can describe the laws and patterns of certain attributes in a transaction, among which algorithms such as Aprili, F-Growth and CARMA have been widely used in medical, sales and other fields, and these algorithms are also suitable for the study of the laws between food data [7-9].Chao Fengying et al. [10] introduced Apriori algorithm into the analysis of food safety testing data, and conducted the actual analysis of the food safety testing database provided by an entry-exit Inspection and Quarantine Bureau. The results showed that this method is more suitable for the analysis of multiple factors in food safety testing data than the previous mathematical statistical analysis methods. These rules can provide decision support for food safety supervision and improve the efficiency of supervision. Zong Wanli et al. [11] used Apriori algorithm to analyze the unqualified items of food safety sampling inspection data in Shandong Province, China, and obtained the correlation between multiple unqualified items of the same sample. The above study explores some of the rules of association between food products, but due to the different objectives and regulatory needs, there are differences in the scope of application, simplicity, and indicator settings.

In this study, based on the more than half a million pieces of data grain processing product data in China's national food sampling data from 2019 to 2021, the CARMA algorithm in the correlation rules was used to establish a correlation analysis model between qualified data and unqualified data, and analyzed them separately, trying to explore potential factors affecting food safety, determining key and non-key test objects, and providing a reference for risk identification and priority supervision order for food sampling work.

Materials and Methods

Materials

Data Sources

The data source for this research is the National Food Sampling Inspection Database for 2019 - 2021, as shown in Table 1. The database mainly contains the following information: food id, food name, food category (category/sub-class, sub-subclass), time for sampling inspection,types of food packaging, testing site (province/city/district/county), sampling inspection sample No., pollutant category, detection value of pollutants, testing site, qualified item, etc. The data characteristics are as follows:

Field Name	Type	Field meaning	Primary key	Foreign key	Annotations
id	varchar(255)	Food id	Yes		
detect_object_19_id	varchar(255)	Test sample No.			
detect_object_19_sp_s_14	varchar(255)	Food name			
detect_object_19_created_at	varchar(255)	Time for sampling inspection			
detect_object_19_sp_s_3	varchar(255)	Province for sampling inspection			
detect_object_19_sp_s_4	varchar(255)	City for sampling inspection			

detect_object_19_sp_s_5	varchar(255)	District (county) for sampling inspection			
detect_object_19_sp_s_2	varchar(255)	Testing site			
detect_object_19_sp_s_13	varchar(255)	Food production license number			
detect_object_19_sp_s_33	varchar(255)	Food packaging			
detect_object_19_sp_s_61	varchar(255)	Food status			
detect_object_19_sp_s_71	varchar(255)	Whether it is a qualified sample			
detect_object_19_sp_s_17	varchar(255)	Food category			
detect_object_19_sp_s_18	varchar(255)	Food subclass			
detect_object_19_sp_s_19	varchar(255)	Food sub-subclass			
detect_object_19_sp_s_20	varchar(255)	Food class			
detect_result_19_id	varchar(255)	Sampling inspection No.			
detect_result_19_spdata_0	varchar(255)	Pollutant name			
wuranwu_leibie	varchar(255)	Pollutant category			
detect_result_19_sp_data_1	varchar(255)	Pollutant test results			
detect_result_19_sp_data_1_clear	varchar(255)	Detection value of pollutants			
detect_result_19_sp_data_2	varchar(255)	Whether it is qualified			
detect_result_19_spdata_3	varchar(255)	Pollutant determination basis			

detect_result_19_spdata_4	varchar(255)	Pollutant limit standard			

Table 1: Food sampling inspection database information.

(1) Different testing frequency of each category of food as shown in Figure 1, rice food receives the highest detection frequency, accounting for 50% of all the grain-processed products. The main reason is that rice food has the largest annual output, which has a great impact on people and needs to be paid great attention to; Wheat flour food includes general wheat flour and special wheat flour, which is mainly used to make steamed bread, bread and other flour products, accounting for 22% of all grain-processed products; vermicelli, processed grain products, grain milling products and grain flour manufactured products account for 6%, 4%, 8% and 10%, respectively. The inspection frequency is higher for the food with higher attention and higher output; The inspection frequency is lower for the food with lower attention and lower output;

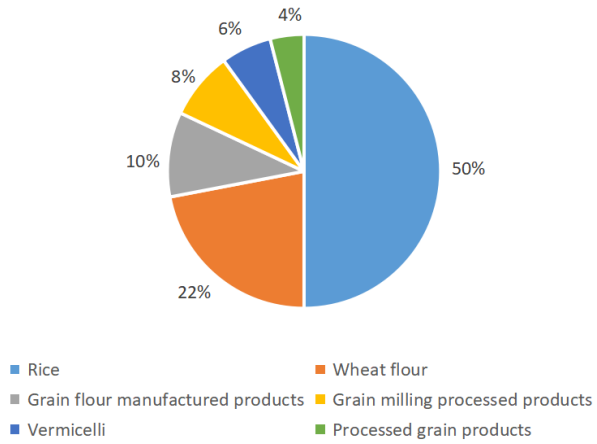

Figure 1: Distribution of detection frequency of grain-processed products.

(2) The pollutants detected in the database food are: polycyclic aromatic hydrocarbons, biotoxins, additives, heavy metals, and other contaminants. Most pollutants are limited class pollutants, mainly referring to the pollutants with the maximum residue limit stipulated, and most of the test results are numerical data.

(3) The detection values of pollutants are not completely correct and are not all in numerical form, often including characters, Chinese characters or mathematical symbols, such as '/', 'undetected', '<10', or value vacancies.

Data pre-processing

(1) This research mainly focused on food species (sub-subclass), types of food packaging, testing site, testing place, pollutant category and detection value of pollutants. Before the data analysis, the information of these aspects were extracted separately. All the unqualified data were placed in an Excel, and then the detected qualified data were put into a different Excel by sub-subclass of food.

(2) Firstly, clean the data of the detection value of pollutants, delete the records with empty detected value, replace the "non-detected" with "0", and replace the records of "< 10" with 0. If the same detection result contains multiple values, replace it with the average value of multiple results.

(3) Since the data required for association rule analysis are discrete, the numerical test results of limited class items in qualified food should be discretized. According to the 2017 National Food Safety Standard Limit of Pollutants in Food and the risk classification method of general hazardous substances, select 1/2 of the standard value of the limit and divide the test results into low and medium levels .This research is to separate different food subclasses. These relative risk levels are also for a single food category - pollutant combination, and are not applicable to the comparison between different food pollutants.

Methods

Association rule mining technology

The mining of association rules in data mining technology allows us to find the relationship between items from the datasets, which can be used to find the internal correlations of different items that appear in the same event. For example: Used to find connections between shopping basket data to facilitate cross-selling [12]; Text mining can be done; Some valuable connections can also be found in other fields such as bioinformatics, medical diagnosis, earth sciences, etc.

The strength of the association rules can be measured by the level of support and confidence. Support represents the probability that the item set {A, B} appears in the total item set. The meaning of confidence is the probability of deriving B from the association rule "A → B" when condition A occurs. Lift refers to the ratio of the probability of containing B under the condition of containing A to the overall probability of occurrence of A. The rules that satisfy the minimum support and the minimum confidence are the "strong association rules". If Lift (A→B)>1, rule "A→B" is a valid strong association rule. Instead, it is an invalid strong association rule [13].

CARMA algorithm

CARMA (Controlled Auto-Regressive Integrated Moving Average) algorithm used in this paper was first proposed in 1999 by Professor Christian Hidber of Berkeley University, which improves the traditional association rule mining algorithm [14]. Compared with other static association rule algorithms, CARMA algorithm has its obvious advantages:

(1) CARMA algorithm performs more efficiently, and the result set can be constructed by scanning the data twice [15];

(2) CARMA algorithm processes data in both Tabular and Transactional formats;

(3) CARMA algorithm can set the support for the antecedent and consequent of the rule. Meanwhile, CARMA allows rules with multiple consequents [16];

(4) CARMA algorithm runs with even less memory.

In view of the above advantages of CARMA algorithm, the characteristics of this experimental data, and the experiment and analysis of various association rules, CARMA algorithm is selected as the core algorithm of association rule mining.

A mathematical expression for CARMA algorithm [17]:

$$X(z^{-1})o(t) = Y(z^{-1})i(t) + C(z^{-1})e(t) \quad (1)$$

The delay operator is included in the expression; z^(-1)i(t) and o(t) are the system input and output, respectively; e (t) is the Gaussian white noise with zero mean variance [18], That is, the stochastic process of the zero-mean variance composed of the sampling point when the system randomly obtains the sampling value from a Gaussian distribution.

Wherein:

$$X(z^{-1}) = 1 + \sum_{1}^{n} a_i\, z^{-i} \quad (2)$$

$$Y(z^{-1}) = 1 + \sum_{1}^{n} b_i\, z^{-i} \quad (3)$$

$$C(z^{-1}) = 1 + \sum_{1}^{n} c_i\, z^{-i} \quad (4)$$

$X(z^{-1})X(z^{-1})$, $Y(z^{-1})Y(z^{-1})$ and $C(z^{-1})C(z^{-1})$are the polynomial of the delay operator.

Two stages of CARMA algorithm:

(1) The first stage produces the latent frequent item set, followed by the deletion of the latent frequent items to obtain the final set;

(2) Calculations on the basis of the frequent item set generated in the previous stage produce the final association rules [19].

The algorithm flow is shown in Figure 2:

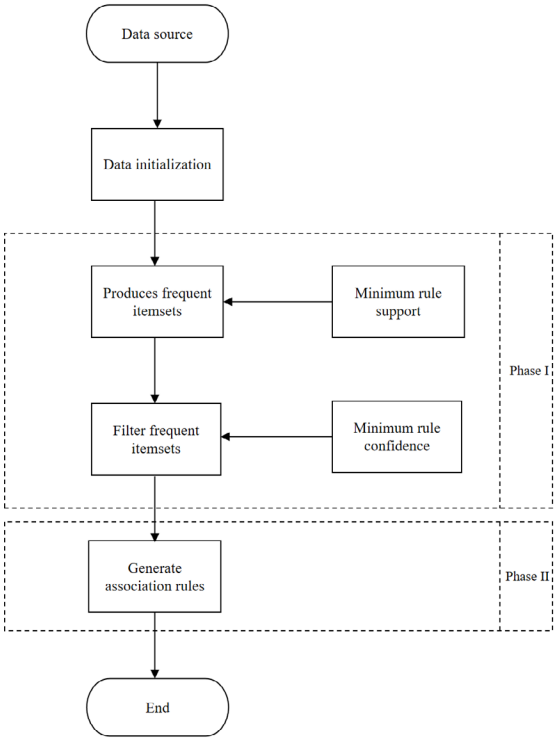

Figure 2: CARMA algorithm flow chart.

Experiment

SPSS Modeler software used in this research is an advanced data mining and predictive analysis platform software, which can realize the functions of association modeling, predictive analysis and cluster analysis of data [20].The specific modeling process is as follows:

Step 1: Add an "Excel type" source file to SPSS Modeler software to import Excel data for different food types separately;

Step 2: Add the "type" option to SPSS Modeler software, and set the food category (sub-subclass), testing site, testing place and pollutant category as "arbitrary" for the qualified category data of grain-processed products. The relative risk grade is "target"; For the unqualified data of grain-processed products, each index data is set as "arbitrary" type;

Step 3: Add the "mark" option to SPSS Modeler software, and set the sampling inspection number as the summary keyword;

Step 4: Add the "Carma model" option to SPSS Modeler software, and set the minimum rule support of 10%, and the minimum rule confidence of 70%, to establish the association analysis model; The valid strong association rules obtained after the experiment run are shown in Tables 2 and 3.

Food category	Consequent	Antecedent	Support degree /%	Confidence /%
Rice	Relative risk level = low	Pollutants = chromium (in Cr) and lead (in Pb) and total mercury (in Hg) and cadmium (in Cd)	99.39	100
Rice	Relative risk level = low	Province = Hunan and testing site = supermarket and pollutant = aflatoxin B1	44.06	100
Grain flour products	Relative risk level = low	Province = Sichuan and pollutant = lead	10.01	100
Grain processed products	Relative risk level = low	Pollutant = aflatoxin B1	99.38	100
Grain processed products	Relative risk level = medium	Province = Anhui and Pollutant = lead (in Pb) and Food packaging = plastic bags	10.06	72.13

Table 2: Association rules for qualified grain-processed products.

Consequent	Antecedent	Support degree /%	Confidence /%
Testing site = supermarket and Pollutant = EDTA-2Na	Pollutant = Deoxynivalenol and Food packaging = plastic bags	20.1	71.42
Pollutant = Deoxynivalenol and Province = Anhui	Food category=Vermicelli	13.94	73.22
Pollutant = Sorbic acid and its potassium salts and Testing site= supermarket	Food category= Grain processed products and Province = Yunnan and Pollutant = Sodium dehydroacetate monohydrate and Food packaging = plastic bags	18.5	90.32
Pollutant = Deoxynivalenol and Food packaging = other	Food category=wheat flour and Pollutant = Ochratoxin A and Province = Shanxi	11.2	82.35
Pollutant = Sulfur dioxide	Food category=Grain processed products and Province = Yunnan and Pollutant =Aluminum	15.3	75.5

Table 3: Association rules for nonqualified grain-processed products.

Results

Rules 1: In the records of qualified grain-processed products, the food category is rice. 99.39% of the tested samples contain pollutants such as chromium, lead, total mercury and cadmium with a low relative risk; 100% of the tested samples contain pollutants such as chromium, lead, total mercury and cadmium with a low relative risk. It shows that if the sampled rice food contains chromium, lead, total mercury and cadmium, the relative risk level of these four pollutants is low.

Rules 2: In the records of qualified grain-processed products, the records with a food category of rice, with the sampling province in Hunan, with the testing site in the supermarket, with Aflatoxin B1 as the pollutant category and with a low relative risk level account for 44.06%; On the premise that the sampling province is Hunan, the testing site is supermarket and the pollutant category is Aflatoxin B1, the relative risk level of Aflatoxin B1 (100%) is low. This shows that the rice sold in supermarkets in Hunan province contains less Aflatoxin B1.

Rules 3: In the records of qualified grain-processed products, the records with the rice as the food category, with the sampling inspection province in Sichuan, with the lead as the pollutant category and with the low relative risk account for 10.01%; On the premise that the sampling province is Sichuan and the pollutant category is lead, the relative risk level of pollutant lead (100%) is low, indicating that the relative risk of lead in grain flour products in Sichuan Province is low.

Rules 4: In the records of qualified grain-processed products, the records with the grain-processed products as the food category, with the Aflatoxin B1 as the pollutant category and with the low relative risk account for 98.38%; When the pollutant category is Aflatoxin B1, the relative risk level 100% is low, indicating the less Aflatoxin b1 in the grain processed products.

Rules 5: In the records of qualified grain-processed products, the records with the grain-processed products as the food category, with the sampling province in Anhui, with the lead as the pollutant category,with the plastic bag as food packaging and with the medium relative risk account for 10.06%; Under the premise that the sampling province was Anhui, the pollutant category was lead and the food packaging was plastic bags, The relative risk level of lead in processed cereals was 72.13% moderate, indicating that the high lead content in grain processed products in Anhui province may be partly due to the high lead content in unqualified plastic bags.

Rules 6: In the unqualified grain processing product data, the pollutants are Deoxynivalenol and EDTA-2Na, and the food packaging is a plastic bag and the testing site is a supermarket, accounting for 20.1%; under the premise that the pollutant Deoxynivalenol is unqualified and the food packaging is a plastic bag, the probability that the testing place is a supermarket and the pollutant EDTA-2Na is also unqualified is 71.42%, indicating that the sample of Deoxynivalenol in the grain processing products packaged in the supermarket plastic bag is unqualified. EDTA-2Na may also be substandard.

Rule 7: In the unqualified grain processing product data, the food category is noodles, the pollutant category is Deoxynivalenol and the province is Anhui, accounting for 13.94%; Under the premise that the food category is noodles, the probability that the pollutant category is Deoxynivalenol and the province is Anhui is 73.22%, indicating that the unqualified samples of the noodles due to contamination with Deoxynivalenol mainly appear in Anhui Province.

Rule 8: In the unqualified grain processing product data, the food category is grain processed products, the sampling province is Yunnan, the pollutant is Sodium dehydroacetate monohydrate and Sorbic acid and its potassium salts, and the testing place is supermarket, accounting for 18.5%; in the food category is grain powder finished products, the sampling province is Yunnan, and the contaminant Sodium deoxyacetic acid and its sodium salt in the sampling sample are unqualified, and the probability of the contaminant Sorbic acid and its potassium salts is also unqualified and the testing site is a supermarket is 90.32%. It shows that while the Sodium dehydroacetate monohydrate in grain processed products in the supermarkets in Yunnan Province are unqualified, the possibility of Sorbic acid and its potassium salts being unqualified is also high.

Rule 9: In the unqualified grain processing product data, the food category is wheat flour, the contaminants are Ochratoxin A and Deoxynivalenol, the province is Shanxi and the food packaging is other records account for 11.2%; in the food category is wheat flour, the contaminant Ochratoxin A in the sampling sample is unqualified, and the contaminant Deoxynivalenol is also unqualified, and the probability of the province being Shanxi and the food packaging is other records is 82.35%; indicating that the packaging type in Shanxi Province is other wheat flour, When Ochratoxin A is not qualified, the likelihood that Deoxynivalenol will also fail is higher.

Rule 10: In the data on unqualified grain processed products, the food category is grain processed products, the province is Yunnan and the pollutants are Aluminum and Sulfur dioxide account for 15.3%; in the food category is grain processed products, the sampling province is Yunnan, and the probability of pollutant Sulfur dioxide being unqualified in the sampling sample is 75.5%, indicating that the Aluminum in the grain powder finished products in Yunnan Province may also be unqualified.

Discussion

Based on the grain-processed products in the national food sampling inspection data of China from 2019 to 2021, this research used CARMA algorithm to mine the association rules of food category, sampling province, types of food packaging, testing site, relative risk level of pollutant category, and obtained ten valuable and valid strong association rules. Through the further interpretation of the association rules, it can be found that when some contaminants in some foods are present at the same time, the relative risk level of the food type-contaminant combination is low; the content of a certain contaminant in a certain type of food

may be related to factors such as sampling sites, food packaging, etc.; the same sample may produce multiple contaminants that are not qualified, such as when contaminant A is detected, the probability of contaminant B being unqualified is higher. So it can identify whether the food type and relative risk are the key objects of monitoring and reasonably determine the sampling items of the same type of food. Areas with high risk of pollution and food products should be put under the key supervision; for those areas and varieties with good quality and less quality problems, the sampling inspection cycle can be extended and the frequency of sampling inspection can be reduced.

References

1. Wang Haiping, Hu Xiaosong, He Su, Ling Rui, Si Wei (2017) Study on Correlation Rules Based on Food Inspection Data. Modern Food 24: 35-38.
2. Mao Jiaqi, Zheng Yunyun, Jiao Wenjing, Xie Huijun, Yan Guili, et al. (2022) Analysis and countermeasures of national food safety status based on multi-dimensional sampling data. Food and Fermentation Industry 48: 314-320.
3. Hu K, J Liu, Li B, et al. (2019) Global research trends in food safety in agriculture and industry from1991 to 2018: A data-driven analysis. Trends in Food Science & Technology 85: 262-276.
4. Liang Lanxian, Xie Wenxin, Sun Miao, Zhang Shaohui (2013) Analysis of Food Quality and Safety Sampling Inspection Data in Shenzhen City. Mathematical Modeling and Its Applications 2: 55-65+89.
5. Qian Ziqi (2018) Analysis of Local Sampling Data of Food in Tianjin Based on Power Query. China Food Safety Magazine 2018: 68-69.
6. Sewekow E, Schmidt-Faber B, Frost M. (2017).Official Food Control Data in Germany: Analysis and Manifestation of Differences. Journal für Verbraucherschutz und Lebensmittelsicherheit 5: 3-4.
7. ZHOU Hong (2017) Disease Correlation Analysis of Medical Big Data. Electronic Technology and Software Engineering 2017: 187-188.
8. Jiao Xinwei, Li Zhenjie, Liu Yuan, Bai Xiuling (2021) Application of association rule mining in the use of drugs for acute laryngitis. Modern Computer 27: 75-78+83.
9. Won D, Bo MS, Mcleod D (2006) An Approach to Clustering Marketing Data 2006.
10. Chao Fengying, Du Shuxin (2007) Food safety data mining method based on correlation rules. Food and Fermentation Industry 2007: 107-109.
11. Zong Wanli, Zhu Xijun (2020) Correlation rule mining of food sampling data based on Apriori algorithm. Journal of Food Safety 11: 1334-1337.
12. Dogan O, Kem FC, Oztaysi B (2022) Fuzzy association rule mining approach to identify e-commerce product association considering sales amount. Complex & Intelligent Systems 8: 1551-1560.
13. Wang Ruili, Zhu Jialiang, Wang Shuling, Huang Zhilu (2021) Unqualified analysis of national drug sampling in 2017-2019 based on association rules. Modern Applied Pharmacy of China 38: 1781-1787.
14. Yang Haiting (2012) An Empirical Study on CARMA Algorithm Mining Technology in Book Circulation. Library Journal 31: 70-75+25.
15. Xu Didi (2017) Research on cross-marketing of credit products of commercial banks based on correlation rules mining. Business Economics 2017: 103-106.
16. Zhao YL, Zheng DZ (2009) A new parameter estimation algorithm for CARMA models[C]//Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. Tianjin, China: IEEE Computer Society Press.
17. Guo L, Chen HF, Zhang JF (1989) Consistent order estimation for linear stochastic feedback control systems (CARMA model). Automat 25: 147.
18. Kaushik Mahata, Minyue Fu (2007) On the indirect approaches for CARMA modoel identification. Automatica 2007: 1457-1463.
19. Rong-Yao Ruan, Chang-Li Yang, Huixin Chen (2003) On-line order estimation and parameter identification for linear stochastic feedback control system (CARMA model). Automatica 2003: 243-253
20. Hidber C (1999) Online association rule mining. ACM SIGMOD Rec 28: 145.