Applied Mathematics, Modeling and Computer Simulation
C.-H. Chen et al. (Eds.)

doi:10.3233/ATDE221104

Analysis of Food Sampling Data Based on CARMA Algorithm

Yongchun JIAO[a,b,1], Tongqiang JIANG[a,b,1], Tianqi LIU[a,b], Wei DONG[a,b,1], Qi YANG[a,b], Qingchuan ZHANG[a,b]

[a] *School of E-business and Logistics, Beijing Technology and Business University, Beijing, China*

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

Abstract. At present,China's food sampling inspection work has the problems of large workload and high cost. In this paper, we try to extract some valuable association rules from the national sampling data by analyzing them. Based on the data of spices major category in the 2019 national sampling database, this paper applies CARMA algorithm to mine association rules for the data studied in terms of food categories (sub-categories), sampling provinces, testing sites, contaminant categories, and relative risk levels of contaminants, and eight relatively valuable and effective strong association rules are obtained after the experiment, and some of them are interpreted. The results show that through association rule mining of relevant sampling data, the connection existing between some key and non-key testing objects and some unqualified items can be determined, which in turn can provide some reference for the allocation of resources for sampling work.

Keywords. Data mining; Food sampling data; Association rules.

1. Introduction

China's national food sampling and inspection work refers to the State Council, provincial, municipal and county food supervision and management departments to carry out quality sampling and inspection work on the production and sale of food products in accordance with the law, which is one of the important technical supervision means for the supervision of various types of food products. In recent years, society has been paying more attention to food safety issues, and government food safety supervisory departments have increased their efforts in food sampling and inspection, thus accumulating a large amount of sampling data. By mining and analyzing food sampling data, meaningful correlations or interrelationships between sets of items hidden behind and between sets of data can be discovered.

In this study, we use the CARMA algorithm in association rules to establish association analysis models for qualified data and unqualified data, and analyze them separately, in an attempt to explore potential factors affecting food safety, identify priority and non-priority testing targets, and provide a reference for risk identification

1 Corresponding Author, Yongchun Jiao, Tongqiang Jiang,Wei DONG; Fuchenglu Road, Haidian District, Beijing City; E-mail: jiaoyongchun1999@163.com, jiangtq@btbu.edu.cn,dongwei_06@126.com.

and establishing priority supervision order for food sampling work. This will provide reference for risk identification and prioritization of food sampling.

2. Materials and Methods

2.1. Materials

2.1.1.Data Sources

The data source for this study is the 2019 national food sampling database. 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.

2.1.2.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) Clean the data, delete the records with vacant detections and non-detections, replace "0" with "0" for "non-detections", replace "0" with "/" for records containing The average value is taken for multiple values.

(3) Since the risk level data required for the correlation rule analysis is discrete, the numerical test results of the limited items in the qualified food should be disaggregated according to the 2017 National Food Safety Standard Limits for Contaminants in Food. The data of the completed classification of spices were selected as the limit standard value 1/2, and the test results were classified into two relative risk levels: low and medium.

2.2.Methods

2.2.1.Association Rule Mining Technology

In the national food sampling database, each record represents a transaction, and each transaction contains a unique transaction identification number (e.g., sample id) and a list of items that make up the transaction (e.g., province, sampling time, sampling site, contaminant category, etc.). We call the set of items an itemset, and the set of items containing k items is called an itemset, e.g., the set {sample id, province, sampling site, contaminant name} is an itemset. The association rule mining algorithm allows us to discover relationships between items from a data set and can be used to find the intrinsic relevance of different items that appear in the same event.

Suppose X and Y are 2 sets of items contained in a transaction, i.e., X and Y are both true subsets of the transaction. If X is a non-empty subset, Y is also a non-empty subset, and the intersection of X and Y is the empty set, then X=>Y constitutes an association rule in the set of things. In other words, the association rule is an expression like X=>Y, X is called the front term and Y is called the back term[1].

The probability of having both X and Y in the set of terms is called the support of X=>Y, which is called support(X=>Y)=P(X,Y).

The probability that the association result Y occurs under the condition that the precondition X of the association rule occurs, i.e., the probability that the set of items containing X contains both Y, is called the confidence of the association rule X=>Y, and is denoted as confidence(X=>Y)=P(Y|X)=P(X,Y)/P(X).

The lifting degree represents the ratio of the probability of containing X with Y at the same time, to the probability of X occurring overall. If Lift(X=>Y)>1, then the rule "X=>Y" is a valid strong association rule, and the opposite is an invalid strong association rule[2] .

2.2.2. 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.Compared with other static association rule algorithms, CARMA algorithm has its obvious advantages:

(1) The algorithm has a higher execution efficiency than other static association rule algorithms, and usually performs at most two traversals to achieve the data set;

(2) The algorithm can handle data in Tabular format, and can also transform data in Transactional format;

(3) The algorithm can set the support degree for the antecedent and postecedent of the rule separately, and CARMA allows rules with multiple postecedents;

(4) The algorithm has a lower memory consumption[3].

In view of the above advantages of CARMA algorithm, combined with the characteristics of large amount of data and different formats of this experiment, as well as the experiments and analysis of various association rule algorithms, CARMA algorithm is chosen as the core algorithm of association rule mining in this study.

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.

2.2.2.1. A mathematical expression for CARMA algorithm:

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

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.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.

Where in:

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

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

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

$X(z^{-1})$, $Y(z^{-1})$ and $C(z^{-1})$ are the polynomial of the delay operator[4].

2.2.2.2.Two stages of CARMA algorithm:

(1) The first stage uses an iterative approach to query frequent datasets in the dataset and calculate the support not lower than the set threshold to filter;

(2) The second stage performs mining based on the strategy of minimum confidence provided by the user.

The algorithm flow is shown in figure 1:

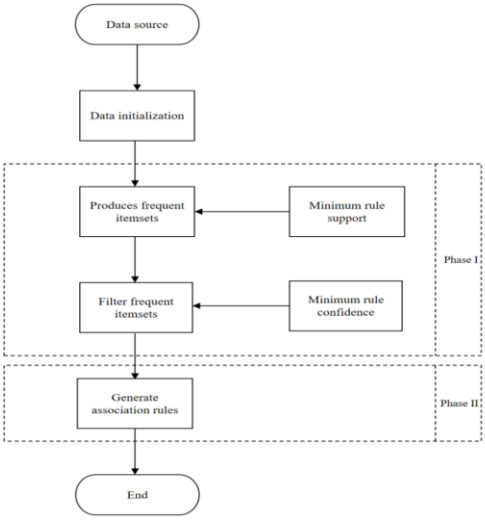

Figure 1 .CARMA algorithm flow chart

2.2.3.Experiment

In this study, we used SPSS Modeler software to establish the CARMA model and set the minimum rule support at 10% and the minimum rule confidence at 80%, and set the variable names and variable types as shown in table 1 and table 2.

Table 1. Qualified condiment variable setting

Variable Name	Meaning	Type
Province	The province where the samples were sampled	Input
Sampling sites	Sampling link premises, such as finished goods warehouses, supermarkets	Input
Packaging	Food outer packaging	Input
Food species (sub-subclass)	Food category	Input
Pollutant name	Name of the detected contaminant	Input
Relative risk level	Relative content of contaminants in samples	Objectives

Table 2. Unqualified condiment variable setting

Variable Name	Meaning	Type
Province	The province where the samples were sampled	Arbitrary type
Sampling sites	Sampling link premises, such as finished goods warehouses, supermarkets	Arbitrary type
Packaging	Food outer packaging	Arbitrary type
Food species (sub-subclass)	Food category	Arbitrary type
Pollutant name	Name of the detected contamina	Arbitrary type

3. Results

Some of the valid strong association rules obtained after the experiment are shown in table 3.

Table 3. Association rules in condiments

Consequent	Antecedent	Support degree /%	Confidence /%
Relative risk level = low	Contaminants = Acesulfame and Aspartame and Sudan Red I and Benzoic acid and its sodium salt	98.12	100
Relative risk level = low	Province=Jiangsuand Packaging=glass bottle and Contaminant=dehydroacetic acid and its sodium salt and Contaminant=sodium saccharin	52.23	96.33
Contaminants = monosodium glutamateand Food packaging = plastic bags	Sampling site = supermarkets and Contaminants = disodium presenting nucleotides	19.8	83.54

Rule 1: Among the qualified semi-solid compound seasoning records, 98.12% of the tested samples contained the contaminants acesulfame, aspartame, Sudan red Ⅰ, benzoic acid

and its sodium salt with low relative risk; 100% of the tested samples contained both contaminants acesulfame, aspartame, Sudan red Ⅰ, benzoic acid and its sodium salt with low relative risk level. It means that in the sampled semi-solid compound seasoning food, if the contaminants acesulfame, aspartame, Sudan red Ⅰ, benzoic acid and its sodium salt are contained at the same time, then the relative risk level of these four contaminants are low.

Rule 2: In the qualified seasoned wine records, the sampling province of Jiangsu, packaging for glass bottles, the category of contaminants for deoxyacetic acid and its sodium salt and sodium saccharin and the relative risk level of low records accounted for 52.23%; in the sampling province of Jiangsu, packaging for glass bottles, the category of contaminants for deoxyacetic acid and its sodium salt and sodium saccharin, the relative risk level of 96.33% for low. This means that in the seasoned wine sampled in Jiangsu Province, if the package is a glass bottle and contains both the contaminants deoxyacetic acid and its sodium salt and sodium saccharin, then the relative risk of these two contaminants is very likely to be low.

Rule 3: In the unqualified MSG data, the pollutants are monosodium glutamate and disodium nucleotide, food packaging for plastic bags and test sites for the supermarket records accounted for 19.8%; in the pollutant deoxynivalenol failed, the premise of food packaging for plastic bags, test sites for supermarkets and the pollutant disodium EDTA also failed the probability of 83.54%, indicating that supermarket plastic bags Packaging of monosodium glutamate in MSG failed the sample, the presentation of disodium nucleotide is likely to also failed.

4. Conclusion

By using the CARMA algorithm to classify the data of the major category of condiments in the national food sampling data during 2019 by food subcategories, association rules were mined for data in terms of province, sampling site, contaminant name, and relative risk level of contaminants in different categories of data, respectively, and eight effective strong association rules were obtained. Through further interpretation of the association rules, it can be found that: when several contaminants are present in certain condiments at the same time, the relative risk levels of the contaminants in such condiments are all low; the same sample may also have several contaminants failing at the same time, such as when contaminant A fails, contaminant B has a higher possibility of failing. Thus, it can be applied to identify the food category and the relative risk level of contamination, and reasonably determine the sampling items of the same type of food. For the higher risk of contamination of areas and food should focus on supervision, for those areas and food in good quality condition can extend the sampling cycle, reduce the frequency of sampling, reduce the detection of non-essential items.

Acknowledgments

This research was supported by the National Key Technology R&D Program of China (No.2019YFC1606401); the Beijing Natural Science Foundation (No.4202014); the Natural Science Foundation of China (No.61873027); the Humanity and Social Science Youth Foundation of Ministry of Education of China (No.20YJCZH229); and the Social Science Research Common Program of Beijing Municipal Commission of Education (No.SM202010011013).

References

[1] Chao, Feng-Ying, Du, Shu-Xin. An association rule-based approach to food safety data mining[J]. Food and Fermentation Industry, 2007, 33(4):3.

[2] Zong Wanli, Zhu Xijun. Association rule mining of food sampling data based on Apriori algorithm[J]. Journal of Food Safety and Quality Inspection, 2020(4):4.

[3] Yang, H. T. An empirical study of CARMA algorithm mining technology in book circulation[J]. Library Journal, 2012, 31(1):7.

[4] Zhao Y L , Zheng D Z . A New Parameter Estimation Algorithm for CARMA Models[C]// International Conference on Fuzzy Systems & Knowledge Discovery. IEEE, 2009.