Research on food safety prediction method based on k-means clustering algorithm

Mengjuan Wu, and *Huan* Jiang*

National Engineering Laboratory for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing, 100048, China

Abstract. Aiming at the problem of food risk prediction, this paper proposes a method based on clustering algorithm to predict product risk by raw material risk. Firstly, based on the provincial supply chain closed-loop hypothesis, this paper proposes the selection method of clustering indexes for products and their raw materials. Secondly, this paper uses the k-means clustering algorithm to cluster the products and the corresponding raw materials respectively, then based on the clustering Class results automatically determine the high-risk categories of the products and their raw materials. Finally, the analysis of the experimental data of the 8 categories of products and their raw materials shows that the ratio of the high-risk categories of products and the ratios of the corresponding high-risk categories of raw materials have a strong positive correlation. The experimental results prove the rationality of the raw material clustering index selection method proposed in this paper and the correctness of the method of predicting product risk based on the raw material risk based on the clustering algorithm.

Keywords: Food safety, Food raw materials, K-means clustering, Pearson correlation coefficient.

1 Introduction

The World Health Organization defines food safety as "a public health problem that affects human health from toxic and harmful substances in food" [1]. With the continuous development of economy, the types of food are constantly enriched, and food safety issues are constantly emerging, which makes food safety the focus of attention.

This article aims to explore whether there is a linear correlation between the high risk ratio of raw materials and the high risk ratio of food products, so as to provide a basis for predicting the risk of food products using the data of the high risk ratio of raw materials, and then provide a basis for monitoring food safety.

Most of the current research is to establish an indicator system for the links of the entire food supply chain and calculate the weights to obtain indicators that have a greater impact on food safety. The food supply chain includes four links: raw material procurement,

* Corresponding author: 1412666583@qq.com

product processing, product transportation, and product sales. In previous research on food supply chain safety early warning, the raw material procurement link has the largest weight in the total security of the supply chain[4] , which shows that the safety of raw materials has the greatest impact on food safety. However, there is no research to explore the relationship between raw material safety and product safety. In addition, the difficulty of exploring the relationship between raw material risk and product risk is much less than that of exploring the whole supply chain. Therefore, it is of great significance to study the relationship between product risk and corresponding raw material risk.

The procurement of raw materials is the source of food safety. This article takes the closed-loop supply chain in the province as the premise (the product and its corresponding raw materials are produced in the same province), starting from the source of food safety, using k-means clustering and Pearson correlation analysis to explore whether there is a linear relationship between the high risk of raw materials and the high risk of corresponding products. Experiments show that there is a strong positive linear relationship between the high risk of raw materials and the high risk of corresponding products.

2 Data source

The data source of this study is the 2019 National Supplementary Database, and 8 food categories are selected for experimentation with food categories as the unit. When selecting food categories, a high proportion of unqualified data is the preferred selection condition. The final selected food categories are: Shaanxi potato and puffed food, Shaanxi Pastry, Hainan Cookies, Fujian Soy Products, Yunnan rice noodle products, Jiangsu Egg Products, Guangdong aquatic products and Jiangxi Seasoning.

3 Predictive models and algorithms

3.1 K-means model and algorithm

K-means clustering is an iterative solution clustering analysis algorithm, the tool to implement k-means clustering in this experiment is python3.8. The specific process of the algorithm is as follows [6]:

(1) Manually determine the K value of the clustering category number in advance, and then randomly select K points as the initial cluster center.

(2) Calculate the Euclidean distance between each object and each cluster center, and assign each object to the cluster center closest to it. Suppose the Euclidean distance between two p-dimensional data points $x_i=(x_{i1},x_{i2},\ldots,x_{ip})$ and $x_j=(x_{j1},x_{j2},\ldots,x_{jp})$, as in formula (1):

$$d(x_i,x_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{ip}-x_{jp})^2} \tag{1}$$

The average distance of all samples is equation (2):

$$Meandist(S)=\frac{2}{n(n-1)}*\sum_{i\neq j,j=1}^{n}d(x_i,x_j) \tag{2}$$

(3) After recalculating the mean value of each sample, return to step (2) until a certain termination condition is met. The termination condition can be that no (or minimum

number) of objects are reassigned to different clusters, no (or minimum number) of cluster centers change again, and the sum of error squares is locally minimum.

(4) The clustering ends, and K clusters are obtained.

3.2 Experiment

Set up two experiments in this research. The first experiment is to cluster food products; the second experiment is to cluster the raw materials corresponding to the products.

Before conducting the two experiments, it is necessary to select the indicators of product clustering and corresponding raw material clustering respectively, and each selected indicator is a dimension of clustering.

Taking Shaanxi potato and puffed food as an example, id is the index of this kind of food. The detection items of this kind of food include peroxide value, acid value, lead, saccharin sodium, and aluminum residues.

3.1.1 Selection of product cluster detection indicators.

First determine the test items that cause the product to fail, and determine the test items as the necessary test indicators for the experiment. The test items that caused the failure of Shaanxi potato and puffed food were peroxide value, acid value and lead, these three items are necessary test items for Shaanxi potato and puffed food. Secondly, select the main test items of the product, each product will test this index. For example, if all 100 vegetable oil products will test the peroxide value, the peroxide value is the main test item for vegetable oil.Finally, determine whether the data of the main test items are valid and usable for the experiment. For example, if the detection values of the main detection items are all "undetected" or all are the same value, and such data does not affect the clustering, the detection value of the indicator is considered invalid and unusable.

3.1.2 Selection of product raw materials.

First confirm the main raw materials of each product in a certain food category. Secondly, by observing the main raw materials of each product to determine the main raw materials of this food category. If some raw materials in each product are consistent, the consistent raw materials are considered as the main raw materials of the food category. If the main raw materials of each product are different, because each product comes from the same food category, the food category of the different raw materials is consistent. For example, in Shaanxi potato and puffed food, the main raw materials of product A are rice, vegetable oil and edible salt; the main raw materials of product B are millet, vegetable oil and edible salt. The different raw materials between A and B are rice and millet, rice and millet belong to the same category of processed food products. Therefore, the main raw materials of Shaanxi potato and puffed food can be determined as processed food products, vegetable oil, and edible salt. Finally, make sure that the province where the raw material is produced is the same as the province where the product is produced.

3.1.3 Selection of indicators for raw material clustering detection items.

The selection process of the raw material inspection index can be based on the selection process of the product inspection index. First, determine the test items that cause the raw materials to be unqualified, and determine the test items as the necessary test indicators for the experiment. Second, determine the main inspection items of each raw material and

ensure that the inspection values of the main inspection items of the raw materials are valid and available. Third, if there are differences in the main inspection items of raw materials, for example, in this experiment, the main inspection items of the rice raw materials of product A are lead and inorganic arsenic, while the main inspection items of millet raw materials of product B are lead and cadmium. The detection items of A and B products are different heavy metals. In order to put the detection item data of raw material rice and millet into the same dimension, different heavy metal detection indicators must be converted into the same detection indicator. According to the literature, the pollution degree of heavy metals in crops is usually evaluated by the single factor index (P_i) and the Nemerow Composite Index (P) [3], so the Nemerow Composite Index can be used as a new unified detection item for rice and millet to replace the original detection items of different heavy metals. For example, the detection indicators of rice raw materials for product A and the detection indicators for millet raw materials of product B are both Nemerow Composite Index.

The single factor index (P_i) and Nemerow Composite Index(P) are specifically expressed as formulas (3) and (4):

$$P_{\mathrm{i}} = \frac{C_i}{C_o} \tag{3}$$

$$p = \sqrt{\frac{(p_A)^2 + (p_{\max})^2}{2}} \tag{4}$$

C_i is the heavy metal content of crops (mg·kg^{-1}), C_o is the limit standard for heavy metal elements (see Table 1 [7]for the reference standards of each element), and P_i is a single factor index, reflecting the degree of heavy metal enrichment. P_A is the average value of the single factor index, P_{max} is the maximum value of the single factor index, and P is the Nemerow Composite Index.

Table 1. Food sanitation standard limit of heavy metal elements in corps.

Element	Food category name	Limit standard (mg/kg)
Inorganic arsenic	brown rice, rice	0.2
Total arsenic	Cereals (except rice)	0.5
	Grain milled processed products (except brown Rice and rice)	0.5
Chromium	Cereals	1
	Processed grains	1
	Beans	1
Lead	Cereals and their products (except oatmeal, gluten, canned eight treasures, rice and noodles with fillings)	0.2
	Beans	0.2
	Potatoes	0.2
Cadmium	Cereals (except rice)	0.1
	Grain milled processed products (except brown rice and rice)	0.1
	Brown rice, rice	0.2
	Beans	0.2

The clustering dimensions of Shaanxi potato and puffed food are: heavy metal pollution of food crops (Nemerow Composite Index), edible salt (iodine), vegetable oil (acid value) and vegetable oil (peroxide value).

After determining the dimensions of the clustering, use the product id or raw material id as the index to cluster using the k-means clustering algorithm. Taking Shaanxi potato and puffed food as examples, the K value is determined to be 3 for clustering. After the clustering is completed, the Euclidean distance from each cluster center to the origin is automatically calculated by the algorithm as in formula (5), the class with the farthest distance is judged as the highest risk class, and the percentage of the highest risk class to all classes is calculated. Suppose a cluster center is $X_n=(X_{n1},X_{n2},...,X_{np})$, then the distance from the cluster center to the origin is:

$$\sqrt{\left(X_{n_1}-0\right)^2+\left(X_{n2}-0\right)^2+...+\left(X_{np}-0\right)^2} \tag{5}$$

4 Result analysis

According to the clustering results of products and corresponding raw materials, two sets of data can be obtained: the percentage of high-risk products and the percentage of high-risk raw materials, as shown in Table 2.

Table 2. High risk ratio of products and raw materials.

Category	High product risk ratio	High risk ratio of raw materials
Shaanxi potato and puffed food	3.60%	3.80%
Shaanxi Pastry	1.50%	3.80%
Hainan Cookies	13.50%	11.10%
Fujian Soy Products	3.50%	3.50%
Yunnan rice noodle products	9%	5.30%
Jiangsu Egg Products	7.30%	4.20%
Guangdong aquatic products	7%	9.80%
Jiangxi Seasoning	12%	12%

Through Figure 1, it can be clearly seen that there is a positive linear relationship between the high-risk ratio of products and the high-risk ratio of raw materials.

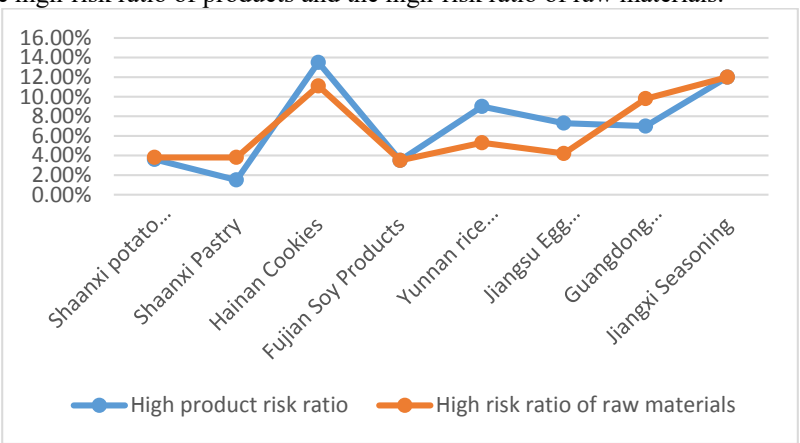

Fig. 1. High risk ratio of products and raw materials.

Using SPSS 26 to analyze the Pearson correlation coefficient of the two sets of data of the high-risk ratio of products and the high-risk ratio of raw materials, the results are shown in Table 3.

Table 3. Correlation.

		High product risk ratio	High risk ratio of raw materials
High product risk ratio	Pearson correlation	1	.824*
	Sig.		.012
	Number of cases	8	8
High risk ratio of raw materials	Pearson correlation	.824*	1
	Sig.	.012	
	Number of cases	8	8
*. *. At the 0.05 level (two-tailed), the correlation is significant.			

The Pearson correlation between the high-risk ratio of products and the high-risk ratio of raw materials is 0.824, which means that there is a strong positive correlation between the high-risk ratio of products and the high-risk ratio of raw materials. And the p value of 0.012 is less than the p value of 0.05, indicating that there is a significant linear relationship between the high risk ratio of products and the high risk ratio of raw materials.

5 Conclusion

From the perspective of food raw materials, this paper studies the linear relationship between the high-risk ratio of products and the high-risk ratio of raw materials, which provides a basis for using raw material inspection data to predict food product risks.

This study is supported by Beijing Natural Science Foundation (No.4202014), Natural Science Foundation of China (61873027), Humanity and Social Science Youth Foundation of Ministry of Education of China (No.20YJCZH229), the R&D Program of Beijing Municipal Education Commission (No.KM202010011011).

References

1. Li Shuguang, Chen Lili, Chen Bo. Analysis of food safety incidents exposed by media in my country from 2004 to 2012[J]. Chinese Journal of Food Science, 2014, 14(03):1-8.
2. Martin Cole, Mary Ann Augustin. Food Safety and Health [J]. Engineering, 2020, 6(04): 391-394.
3. Yang Jianzhou, Wang Zhenliang, Gao Jianweng, Yan Hui, Hu Shuqi, Tang Shixin, Gong Jingjing. Accumulation and health risks of heavy metals in grains, vegetables and fruits in intensive plantations in Hainan Province[J/OL]. Environmental Science: 1- 20[2021-06-07]
4. Jiang Fangtao, Song Ziling. Research on the Food Supply Chain Safety Early Warning Index System[J]. Chinese Seasonings, 2020, 45(02): 192-196+200.

5. Yang Xuemei, Wang Xiaoyi, Li Hongmin. Risk assessment of food safety emergencies in my country from the perspective of supply chain[J]. Food Science, 2017, 38(19): 309-314.
6. Lu Ding. Research on student performance data analysis and predictive modeling based on K-means clustering algorithm[J]. Microcomputer Applications, 2021, 37(05): 148-150.
7. GB 2762-2017, National Food Safety Standard Limits of Contaminants in Food[S].