Association analysis in food sampling inspection data

Tongqiang Jiang, *Xin* Chen*, and *Huan* Jiang

School of E-Business and Logistics, Beijing Technology and Business University, Beijing, China

Abstract. At present, China exists a problem that the cost of food sampling inspection is too high. This paper attempts to reduce the number of sampling inspection items in the same food category, reduce the cost of food sampling inspection, and improve the work efficiency through the association analysis of national sampling inspection data. And this paper applies Apriori algorithm to analyse the association rules, which is based on the unqualified pastry sampling inspection data in the 2019 national food sampling inspection database. Finally, we obtain 10 strong association rules through experiments. The results show that this association analysis can reduce the workload of food sampling inspection effectively.

keywords: Apriori algorithm, Food sampling inspection data, Data mining.

1 Introduction

Food sampling inspection data is crucial for food safety. China spends a large number of people and property on food sampling inspection every year. It has gradually accumulated a great number of data. Through making a deep analysis of the data, it can not only effectively reduce the cost of food sampling inspection, but also improve the efficiency of food sampling inspection.

At first, association rules are proposed for shopping basket analysis. The purpose is to help retailers understand which goods are frequently purchased together by customers through shopping basket data mining, so as to study better marketing strategies [1]. At present, association analysis method has been widely used in food sampling inspection, such as the association analysis of multiple attributes in food safety testing data, but it still has not solved the problem of how to scientifically sample inspection and determine the sampling inspection items of the samples.

This paper selects the unqualified pastry samples sampling inspection data from the 2019 national sampling inspection database. Firstly, we preprocess the data, and then use the association rules mining method—Apriori algorithm to analyse the correlation of the sampling inspection items. The purpose is to find the relationship between the same sample

* Corresponding author: mschenxin@163.com

and different detection items, so as to reduce the workload of food sampling inspection, to achieve the purpose of saving labour costs.

2 Experiment and method

2.1 Data mining related theoretical knowledge

2.1.1 Association rules

Association rules were proposed by R. Agrawal and others at the SIGMOD conference in 1993. It is a method that is easy to find things, item sets in relational databases, and association rules, correlation or causal structure in objects [2].

Suppose $I = \{I_1, I_2, ..., I_M\}$ is a set of different items, $D = \{T_1, T_2, ..., T_N\}$ is a database composed of a series of transactions with TID. The individual attributes of each record in the dataset are called items, the collection of items is called an item sets, and the frequency with which the item sets occur in a transaction are called the support count. Each transaction T is a nonempty subset of I, that is, $T \subseteq I$. If A and B are two item sets contained in transaction T, $A \subseteq I$, $B \subseteq I$, and $A \cap B = \Phi$. Then the related association rule expression of A with respect to B is $A \Rightarrow B$, where A is referred to as the antecedent and B as the consequent [3].

Generally, the attributes of a pair of association rules can be described by Confidence, Support and Lift.

I) Confidence. The confidence of association rule $A \Rightarrow B$ is the ratio of the number of transactions containing {A, B} to the number of transactions containing {A}, which is recorded as confidence $(A \Rightarrow B) = P(B \mid A)$. This parameter reflects the certainty of association rules, with a value range of (0,100%].

II) Support. The support of association rule $A \Rightarrow B$ refers to the probability that both A and B are included in the item sets. It is the percentage of A and B transactions in D and it is recorded as support $(A \Rightarrow B) = P(A, B)$. This parameter reflects the usefulness of association rules, with a value range of (0,100%].

III) Lift. The lift of association rule $A \Rightarrow B$ denoted as lift $(A \Rightarrow B) = P(B \mid A) / P(B)$. This parameter reflects the importance of association rules, with a value range of $(0, \infty)$.

2.1.2 Apriori algorithm

Apriori algorithm, proposed by R. Agrawal and others in 1994 and based on association rules. It is a basic algorithm for mining frequent item sets required to generate association rules [4]. So far, Apriori algorithm is still the classic algorithm of association rule mining, which has been widely discussed and studied by scholars.

The Apriori algorithm has the following five steps.

Step 1: Determining the support of each item by scanning the data set once. Once this step is completed, we can obtain the set F_1 of all frequent 1-item sets.

Step 2: Using the frequent (k-1) – item sets found in the last iteration to generate new candidate k-item sets.

Step 3: Scanning the database again and using subset function to count the support of candidate set. The number determines all candidate k-item sets contained in each transaction.

Step 4: Calculating the support count of candidate item sets, the algorithm will delete all candidate item sets whose support count is less than the support threshold.

Step 5: Repeat step 2, 3, and 4. When no new frequent item sets are generated, the

algorithm ends.

The Apriori algorithm is a layer wise algorithm, which uses "produce-test" strategy to discover frequent item sets. In the process of generating k-item sets from (k-1) – item sets, it is necessary to make sure that all the (k-1) - proper subsets of the newly generated k-item sets are frequent. If one is not frequent, then it can be removed from the current candidate item sets.

2.2 Data mining tools

Weka is a comprehensive machine learning and data mining application platform, which integrates a large number of machine learning algorithms that can undertake data mining tasks. It includes a large number of classic algorithms for data preprocessing, data classification, feature selection, clustering analysis, association rules, association analysis, and visualization on the new interactive interface [5]. The general steps of association rule data mining using Weka are shown in Figure 1.

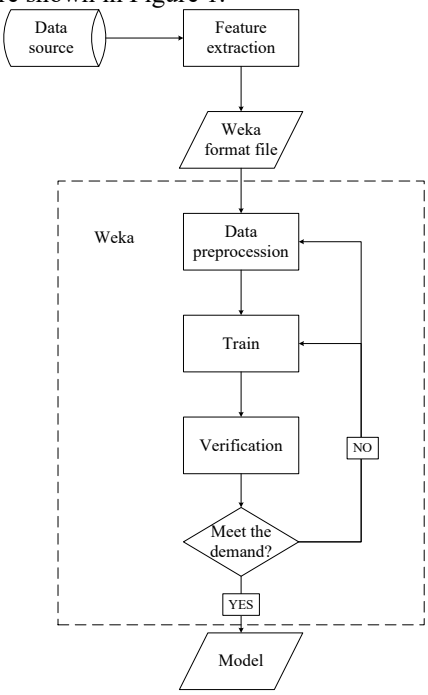

Fig. 1. Flow chart of data mining.

2.3 Experiment on correlation analysis of food sampling data

2.3.1. Data sources

The data of this study is from the 2019 national sampling inspection database. And it usually contains plenty of information, such as the name of the sampling sample, the ID of the sampling sample, the sampling region (province / city / county), the sampling time, and whether the sampling sample is qualified. This study mainly focuses on the association analysis between unqualified items, so more attention is paid to the unqualified attributes.

2.3.2. Data preprocessing

In this study, we select the sampling inspection data of unqualified pastry samples from the 2019 national sampling database, and the obtained data are not suitable for direct analysis, so we carry out the pretreatment first. For non-standard sample test results, the non-numerical data should be cleaned, including the treatment of vacancy value, etc. Furthermore, in the process of data preprocessing, we remove the attributes which is irrelevant with data mining, and only retain the information of unqualified items. Then two unqualified items in the same sample are divided into item1 and item2 fields, and summarized into a table. Due to the poor performance of Weka software in supporting Chinese characters, this paper presents the unqualified item names in the form of Chinese Pinyin. The converted table is shown in Table 1. Finally, save the ". xls" file as ". csv" file.

Table 1. Original data table (top 10 items) (in Chinese Pinyin).

item1	item2
suanjia	guoyanghuazhi
guoyanghuazhi	tangjingna
shanlisuanjiqijiayan	tuoqingyisuanjiqinayan
junluozongshu	meijun
tuoqingyisuanjiqinayan	fangfujihunheshiyong
suanjia	meijun
suanjia	dachangjunqun
dachangjunqun	meijun
guoyanghuazhi	fangfujihunheshiyong
lvdecanliuliang	tuoqingyisuanjiqinayan

3 Experimental results and analysis

After data preprocessing, data mining experiments on food sampling inspection data are started. In this experiments, we set the minimum support threshold of 0.1, the minimum confidence threshold of 0.7, and the maximum number of displayed rules is 10. The data mining results are shown in Figure 2.

```
Best rules found:

 1. item1=dachangjunqun 8 ==> item2=meijun 8    <conf:(1)> lift:(2) lev:(0.08) [4] conv:(4)
 2. item1=tuoqingyisuanjiqinayan 6 ==> item2=fangfujihunhe 6    <conf:(1)> lift:(6) lev:(0.1) [5] conv:(5)
 3. item1=shanlisuanjiqijiayan 5 ==> item2=tuoqingyisuanjiqinayan 5    <conf:(1)> lift:(6.86) lev:(0.09) [4] conv:(4.27)
 4. item1=lvdecanliuliang 2 ==> item2=tuoqingyisuanjiqinayan 2    <conf:(1)> lift:(6.86) lev:(0.04) [1] conv:(1.71)
 5. item2=guoyanghuazhi 1 ==> item1=suanjia 1    <conf:(1)> lift:(9.6) lev:(0.02) [0] conv:(0.9)
 6. item2=tangjingna 1 ==> item1=guoyanghuazhi 1    <conf:(1)> lift:(24) lev:(0.02) [0] conv:(0.96)
 7. item2=bingerchun 1 ==> item1=sanlvzhetang 1    <conf:(1)> lift:(48) lev:(0.02) [0] conv:(0.98)
 8. item1=sanlvzhetang 1 ==> item2=bingerchun 1    <conf:(1)> lift:(48) lev:(0.02) [0] conv:(0.98)
 9. item1=bingsuanjiqinayangaiyan 1 ==> item2=junluozongshu 1    <conf:(1)> lift:(24) lev:(0.02) [0] conv:(0.96)
10. item1=junluozongshu 10 ==> item2=meijun 15    <conf:(0.83)> lift:(1.67) lev:(0.13) [6] conv:(2.25)
```

Fig. 2. Results of association rule mining.

Unqualified food testing items are mainly divided into heavy metals, pesticide residues, veterinary drug residues, additives and other pollutants. The causes of unqualified heavy metals may include environmental pollution, process pollution and organism accumulation. The causes of unqualified agricultural and veterinary drug residues may be excessive use of

agricultural and veterinary drugs, illegal use of agricultural and veterinary drugs banned by the government. And the causes of unqualified additives may be excessive use, beyond the scope of use. By mining association rules, we try to obtain the association relationship among multiple unqualified items in the same sample.

Interpretation of mining association rules:

(1) If the colonies number is unqualified, the mold may also be unqualified. The colonies number represents a large group of bacteria, and mold is one of them. The colonies number may contain bacteria that are harmful or harmless to human body. Mold is harmful to human health, so we should pay increasing attention to mold.

(2) When preservatives are mixed, the sum of the proportion of their respective dosage in the maximum dosage is unqualified. Meanwhile, sodium dehydroacetate are unqualified. Sodium dehydroacetate are commonly used as preservatives. And the reason why sodium dehydroacetate exceed the standard may be that enterprises increase the shelf life of the product, or make up for the poor sanitary conditions in the production process.

(3) The acid value of the samples with unqualified, peroxide value may also be unqualified. The unqualified acid value indicates that the foods have a high degree of oil rancidity, and the rancidity has been a long time. The most common reasons for the acid value exceeding the standard are that the manufacturers buy unqualified raw materials to save costs, the oil used for frying is unqualified, or the frying oil is not replaced and recycled for a long time.

4 Conclusion

In this study, based on the unqualified sampling inspection data of pastry food in the 2019 national sampling inspection database, the Apriori algorithm is used to analyze the association of the preprocessed food sampling inspection data. And we obtain the association rules between the unqualified items. Through in-depth interpretation of the rules, it is found that the same food sample may produce a number of unqualified indicators. For example, when test item A is unqualified, test item B is more likely to be unqualified, which further proves that there is a certain correlation between unqualified indicators. So in the future sampling inspection work, we can reasonably determine the same kind of food sampling inspection items, and try to reduce unnecessary item inspection, which not only reduces the cost of food sampling inspection effectively, but also improves the accuracy and scientific of food sampling inspection.

References

1. Zong Wanli, Zhu Xijun. Association rules mining of food sampling data based on Apriori algorithm [J]. Journal of food safety and quality inspection, 2020,11 (04): 1334-1337.
2. Wang Xuejun, Shen Yi, Yang Huiyuan. Exploration of data mining technology in food detection data [J]. China Pharmaceutical Science, 2019,33 (03): 259-262.
3. Wang Haiping, Hu Xiaosong, he Su, Ling Rui, Si Wei. Research on association rules based on food inspection data [J]. Modern food, 2017 (24): 35-38.
4. Faydd U M. Piatesky – Shapiro G. Smyth P. From Data Mining to Knowledge Discovery in Databases[J]. AI.Magazine.1966.17:37-54.
5. Chen Wenlian. Analysis of data mining technology and application of Weka software [J]. Information recording materials, 2020,21 (02): 142-144.