ATLANTIS PRESS

Advances in Engineering Research, volume 141

5th International Conference on Mechatronics, Materials, Chemistry and Computer Engineering (ICMMCCE 2017)

A Data Preprocessing Method for Food Detection Data Warehouse

Ying Han[1,a], Chunlong Yao[1,2,b,*], Qiuyan Xin[2], Xiaoqiang Yu[1] and Xu Li[1]

[1]School of Informaation Science and Engineering, Dalian Polytechnic University, Dalian 116034, China

[2]National Engineering Research Center of Seafood, Dalian 116034, China

[a] hying821@126.com, [b]yaocl@dlpu.edu.cn

Abstract: The food detection data plays an extremely important and indispensable role in the evaluation of food quality. However, there are a number of detection items based on different evaluation standards for a product. The detection items can be divided into four categories: optimum index, negative index, scope index and sampling index by the evaluation standards. Their measurement units and technical requirements are different, so it is difficult to construct a data warehouse by using these complex detection data. Data preprocessing is necessary before constructing the data warehouse. The existing data processing methods are difficulty in evaluating the food quality comprehensive and efficiently. There are two problems, one is the inconsistent function relationship between the detection data and the processing result, and the other is the sampling index can't be processed. To address these problems, this paper proposes an approach for data preprocessing by calculating the quality index for detection data. This method can not only evaluate the food quality comprehensively and efficiently, but also provide a scientific basis for constructing the data warehouse.

Keywords: detection data, food quality, data warehouse, data preprocessing.

1. Introduction

Recently, an increasing number of people begin to pay special attention to the quality of life. In terms of food, people attach great importance to the food safety. They lay stress on whether the food is natural, nutritional, pollution-free, harmless and green product. Food is the integral part of human life [1]. As a consequence, for the companies or institutions related food industry, it is particularly important to ensure the food safety in the process of production. The food quality and safety are related with the consumer's health and the quality of their life. Governments from all over the world have attached great importance to this issue [2][3]. It is an important work to carry out food quality supervision and inspection. At present, a lot of people are working on the study of food safety. The food detection laboratories at all levels do a lot of food testing work every day. It generates a large amount of food detection data every day. During this period, it accumulated a lot of resources of food safety detection data [4]. And the collection of a large amount of data contains a wealth of information on food safety [5]. Data warehouse technology lays a foundation for analyzing data resources deeply, using the data resources effectively and assisting the manager to make a decision [6]. But there are many data sources, and the data format of these data is not uniform. So a data preprocessing method is urgently needed. It can use the data resources effectively, and evaluate the food quality comprehensive and efficiently. Xu [7] proposed data screening, cleaning and conversion, when he constructed a framework of food safety warning data analysis system. But he did not illustrate the specific data processing scheme. Guo [4] analyzed the detection data of import and export food testing laboratory in Shandong district. And then he built up data labels and rules, discretized and layered the main data attributes. Then it formed the hierarchical dictionary table. However, it also faces some limitations and problems. It can't apply to other kinds of food, because

the different kinds of food have different detection items and standards. Consequently, it can't be used in all kinds of food to evaluate the food quality comprehensive based on this kind of preprocessing method.

In fact, the food safety is a comprehensive evaluation results. The safety of the product needs to be discriminated by testing a variety of items. At present, the system of food hygiene security evaluation index is relatively perfect [8]. In addition, there are also a lot of evaluation methods, such as index evaluation method [9][10][11], Nemerow comprehensive pollution index method[12], analysis hierarchy process etc [9][10]. In those methods, the most commonly used is the index evaluation method, because it is easy to calculate and grasp. And it is also the basis of other methods. But it can't deal with four kinds of detection items simultaneously by using the exiting evaluation methods to process the data. And in some normalization methods, the function relationship between the detection data and the processing results is not consistent. When considering the negative index, the detection data and processing results are direct proportion. However, when it comes to the optimum index, they are inverse proportion.

This paper proposes a data preprocessing method to solve the problem of universality. In this way, the data can be processed into a unified data format. This method is to calculate the quality index according to the index type of each test item. It provides a scientific basis for constructing the data warehouse model. And this data preprocessing method is ready for achieving data mining further.

2. Data Preprocessing Method

The detection data has many characteristics, such as there are many data sources. In addition, the data types are various, and the data formats are inconsistent. Therefore, analyzing the data comprehensively in the food industry, the data must be processed. In this paper, it uses a linear processing method to deal with the data. The result is called quality index, which noted as d. If the value of the quality index is between 60 and 100, it is regarded as qualified. And the closer to 100 the value, the better the quality. But if $d < 60$, the product is unqualified. Then the data preprocessing methods about four kinds of indexes are given as follows. And their function images are shown as figure 1.

The optimum index is used to calculate a kind of detection item which is given a minimum value about technical index. It is regarded as qualified when the detected value is not less than the minimum value, such as the energy in table 1. Let x denote the actual detection value of the detection item. S_1 represents the lower limit of the standard. When $x = S_1$, it is regarded as the $d = 60$. And the bigger the x, the better the quality. Assuming when $x = 2S_1$, the $d = 100$. And when $x > 2S_1$, the $d = 100$. The preprocessing method of this type of index is presented as formula (1). Its function image is shown as figure a in figure 1.

$$d = \begin{cases} \frac{60}{S_1}x, & x \in [0, S_1) \\ \frac{40}{S_1}x + 20, & x \in [S_1, 2S_1) \\ 100, & x \in [2S_1, +\infty) \end{cases} \tag{1}$$

The negative index is used to calculate a kind of detection item which is given a maximum value about technical index. It is regarded as qualified when the detected value is not more than the maximum value, such as the fat in table 1. Let S_2 denote the upper limit of the standard. Assuming when $x = S_2$, it is regarded as the $d = 60$. And the smaller the x, the better the quality. The $d = 100$ when $x = 0$. And when $x > 2S_2$, the $d = 0$. Then the preprocessing method of this type of index is shown as formula (2). Its function image is shown as figure b in figure 1.

The scope index is used to calculate a kind of detection item which is given the minimum value and the maximum value about technical index. It is regarded as qualified when the detected value between the minimum value and the maximum value, such as the vitamin A in table 1.

$$d=\begin{cases}-\dfrac{40}{S_2}x+100, & x\in[0,S_2)\\ -\dfrac{60}{S_2}x+120, & x\in[S_2,2S_2)\\ 0, & x\in[2S_2,+\infty)\end{cases} \tag{2}$$

S_m represents the mid-value between S_1 and S_2. Assuming when $x = S_1$ or $x = S_2$, it is regarded as the $d = 60$. And when $x = S_m$, the $d = 100$. When $x > S_1+S_2$, the $d = 0$. Then the preprocessing method of this type of index is shown as formula (3). Its function image is shown as figure c in figure 1.

$$d=\begin{cases}\dfrac{60}{S_1}x, & x\in(0,S_1)\\ \dfrac{40}{S_m-S_1}(x-S_1)+60, & x\in[S_1,S_m]\\ -\dfrac{40}{S_2-S_m}(x-S_2)+60, & x\in[S_m,S_2]\\ -\dfrac{60}{S_1}(x-S_2)+60, & x\in(S_2,S_1+S_2]\\ 0, & x\in(S_1+S_2,+\infty)\end{cases} \tag{3}$$

Finally, the sampling index is used to calculate a kind of detection item which is given a sampling plan. Let *n* denote the number of samples taken from the same batch of products. The letter of *c* represents the maximum number of samples allowed to exceed *m*. The lowercase letter of *m* represents the limited value of microbial index of the acceptable level. *M* represents the highest safety limits of microbial index. It is regarded as qualified when the detection results of all samples are less than *m* or there are *c* results between *m* and *M*. Otherwise it is not qualified. For example, the colonies number is this type of index in table 1. The preprocessing method of this type of index is shown as formula(4). $R=\{X_1, X_2,\ldots, X_n\}$ is a set of the detection value of *n* samples. Then make $R_1=\{X_i \mid X_i \in R,\ X_i\leqslant m\}$, $R_2=\{X_j \mid X_j\in R,\ m<X_j\leqslant M\}$, $R_3=\{X_k \mid X_k\in R,\ X_k>M\}$. When the detection item is qualified, the maximum detection value is $(n\text{-}c)m + cM$. There are *c* samples with the detection value of *M*, and the others are *m*. According to the figure d in the figure 1, it can be seen that when $\sum X_i = (n\text{-}c)m + cM$, the $d = 60$. The figure d in the figure 1 is the function image of the qualified sample. The detection item is unqualified when one sample is unqualified or c+1 samples' detection value exceed *m*. So the minimum detection value of unqualified sample is $\min\{(c+1)m, M\}$. Assuming when $\sum X_i = nM$, the $d = 0$. The figure e is the function image of the unqualified sample.

$$d=\begin{cases}-\dfrac{40}{(n-c)m+cM}\sum_{i=1}^{n}X_i+100, & |R_1|=n,\ \ or\ |R_2|\leq c\ \ and\ \ |R_3|=0\\ -\dfrac{60}{nM-\min\{(c+1)m,M\}}\sum_{i=1}^{n}X_i+\dfrac{60nM}{nM-\min\{(c+1)m,M\}}, & |R_2|>c,\ \ or\ |R_3|>0\end{cases} \tag{4}$$

However, the four calculation methods mentioned above are only the data process of a single detection item. For a kind of product, it has a large number of detection items. *t* indicates the number of detection items. So it needs to set weights $w_i \geqslant 0$ (i=1, 2,⋯, t) based on the importance of the detection item. w_i is the weight of the *i*th detection item. d_i (i=1,2,⋯,t) is the quality index value of the *i*th detection item. $Q=\{d_1, d_2,\cdots, d_t\}$, $Q_1=\{d_i \mid d_i<60,\ 1\leqslant i\leqslant t\}$, $Q_2=\{d_j \mid d_j>60,\ 1\leqslant j\leqslant t\}$. When all detection items are qualified, the overall quality index is the weighted average of all detection items. But the product is not qualified when there is a detection item is unqualified. Assuming the maximum detection value of unqualified product is *U*. i.e. the least weight of the detection item is

unqualified ($d_{\min} = 60$), the others are qualified ($d_i = 100$, $i \neq min$). The U is calculated as formula (5). The graph of unqualified product is shown as figure f in figure 1. The X-axis represents the weighted average of all detection items. Its calculation is shown as formula (6). It can get the overall quality index of samples as formula (7). The overall quality index is noted as H. For a product, it is regarded as qualified when $60 \leqslant H \leqslant 100$. But if $H<60$ it is regarded as unqualified.

$$U = \sum_{\substack{i=1, \\ i \neq \min}}^{t} \frac{w_i}{\sum_{k=1}^{t} w_k} \bullet 100 + \frac{w_{\min}}{\sum_{k=1}^{t} w_k} \bullet 60 \tag{5}$$

$$X = \sum_{i=1}^{t} \frac{w_i}{\sum_{k=1}^{t} w_k} d_i \tag{6}$$

$$H = \begin{cases} X, & |Q_1| = 0 \\ \frac{60}{U} * X, & |Q_1| \neq 0 \end{cases} \tag{7}$$

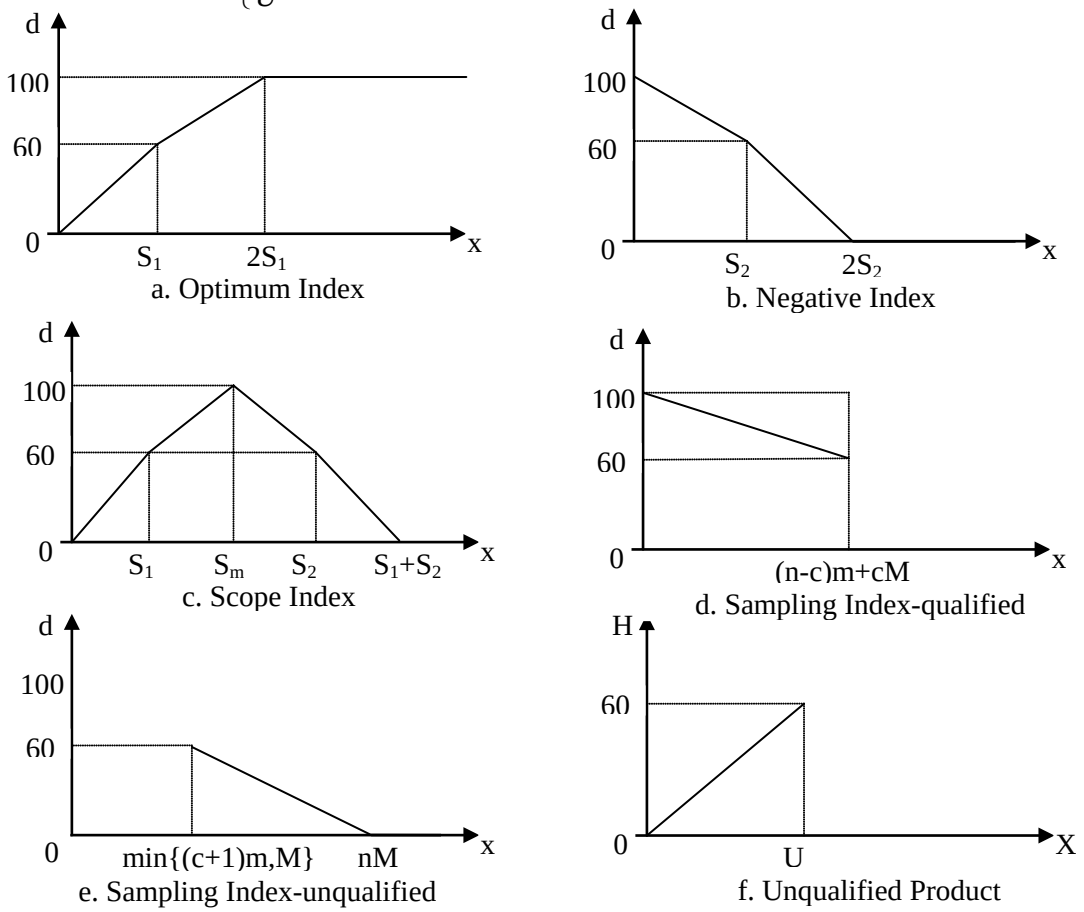

Fig. 1 The function images of four kinds indexes

According to the H, it is very easy to evaluate the food quality comprehensive and efficiently. The detection data and the evaluation results are linear relationship by using this method. This method can be applied to a variety of detection indicators. And it realizes the contrast among different products. This data preprocessing method unifies the format of the evaluation results.

3. Application

There are some detection data about three kinds of rice flour: product A, product B, product C. In this part, these data are dealt with linear processing according to the national standard of food safety, GB10769. And this paper analyzes the quality of these products from two perspectives. One is the

overall quality index, i.e. the *H* mentioned in the part II. The other one is Nemerow index. The results are shown in table 1.

Table 1 The Data Preprocessing of Three Kinds of Rice Flour

Product			Product A		Product B		Product C	
items	unit	requirement	detection	index	detection	index	detection	index
energy	KJ/100g	≥1250	1673.000	73.536	1648.000	72.736	1655.000	72.960
protein	g/100KJ	≥0.33	0.517	82.667	0.613	94.303	0.490	79.394
fat	g/100KJ	≤0.8	0.114	94.300	0.078	96.100	0.050	97.500
vitamin A	μg/RE/100KJ	14-43	37.000	76.552	25.000	90.345	20.000	76.552
vitamin D	μg/100KJ	0.25-0.75	0.627	79.680	0.568	89.120	0.430	88.800
lead	mg/kg	≤0.3	0.010	98.667	0.010	98.667	0.051	93.200
arsenic	mg/kg	≤0.2	0.050	90.000	0.060	88.000	0.083	83.400
aflatoxinB1	μg/kg	≤0.5	0.200	84.000	0.200	84.000	0.020	98.400
colonies number	CFU/g	n=5,c=2, $m=10^3$,$M=10^4$	140,40,140 ,40,90	99.217	35,35,50, 30,35	99.678	10,10,10, 10,10	99.913
the overall quality index (*H*)			--	86.513	--	90.327	--	87.791
Nemerow index			--	89.800	--	91.464	--	90.907

According to the information given in the table 1, we can see that the value of the detection data is very wide range, and the evaluation standards of different detection items are different. It is difficult to make an overall evaluation about the product by these raw detection data. So it is difficult to make a unified evaluation. With the linear processing method, the detection data are distributed between 0 and 100. It is convenient to make an overall evaluation about the product. From the table 1, we can see that it can compare the differences among the three products. When considering the overall quality index, the product B is best. And when considering the Nemerow index, the product B is also best. From this method, we can draw a conclusion that this data preprocessing method is more convenient for people to compare or evaluate the quality of food. Not only can compare the merits of a single detection item, it can also compare the product as a whole. It is easy to make an overall evaluation about the products. But there is a deficiency in this method. It is different about the influence of different detection items on one product. In the table 1, the weight of each detection item is taken to be 1. So it is necessary to further improve the weight value.

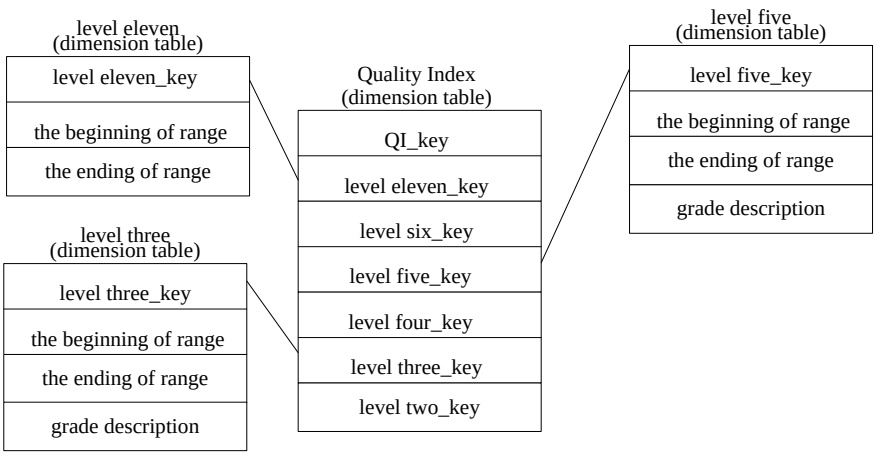

Fig.2 The Application in Data Warehouse

The quality index can be set into a dimension when constructing the data warehouse. And then it can be divided into different levels. So the situation of the product quality can be observed from different granularities. The application of the quality index in the data warehouse is shown in figure 2.

Table 2 The Structure of Level Five

Level five_key	Beginning	Ending	Description
1	0	60	unqualified
2	60	70	qualified
3	70	80	average
4	80	90	good
5	90	100	excellent

In figure 2, quality index is abbreviated as QI. In the figure 2, it can be seen that the quality index is processed by classification. The structure of level five is shown as table 2. According to this design, users can do better about statistics and evaluation of product quality.

4. Conclusions

The main work of this paper is to design a data preprocessing method for food detection data warehouse. It focuses on calculating the quality index from four aspects. And it realizes the linear processing. This method can compare the different batches of the same product and observe the change of the food quality. It provides a scientific basis for constructing the data warehouse model. And this method is ready for achieving data mining further. The linear processing can help enterprises manage the food safety as efficiently and effectively as possible. The manager can make a plan or adjust the relevant plan according to the change of the data. It provides a scientific basis for enterprise decision. Nevertheless, this design still has some deficiencies. The weight of each detection item is set to 1 in this paper. So it can't reflect the importance of the detection item. It needs to research further.

Acknowledgments

This work is funded by National Engineering Research Center of Seafood (2012FU125X03) and Key University Science and Technology Platform of Liaoning Province (No. 2011-191).

References

[1] Farhana R.Pinu, Metabolomics—The new frontier in food safety and quality research, J. Food Research International. vol.72 (2015) 80-81.
[2] Liu Jin-sheng, Li Zhe, Chu Cheng-shan and Miao Hui, Establishment and Application of Food Quality Index Model, J. Food Research And Development. vol.35, No.9 (2014) 1-4.
[3] Li Yang, Wu Guo-dong, Gao Ning, Study of food safety comprehensive assessment index based on intelligent calculation, J. Agriculture Network Information. No.4(2006) 13-14.
[4] Guo Shu-chao, Gong Fang, Ze Xiang-jun, Zhou Bao-hua and Yu Shi-chao, Applied Study on Food Test Data Warehouse Technology, J. Food Research And Development. vol.34, No.17 (2013) 125-128.
[5] Song Lianghui, Lou Xinai, Yang Zhong, Zheng Jiankun, A Lifu and Gao Jie, Data Warehouse Model Design based on Food Safety Detection, J. Journal of Henan Science and Technology. vol.568, No.7 (2015) 13-15.
[6] Leigh Ann Newman, Building Data Warehouse Starts With Definitions, J. Clinical Data Management. vol.6, No.2 (1999) 10-12.
[7] Xu Jianjun and Gao Shengpu, Study on the Construction of Data Analysis System to Food Safety Warning, J. Journal of Chinese Institute of Food Science and Technology. vol.11, No.2 (2011) 169-172.
[8] Li Zhemin, Research on the Connotation of Food Security and Evaluation Indicator System, J. Journal of Beijing Agricultural Vocation College. vol.18, No.1 (2004) 18-22.
[9] Deng Cong-wen, ZHU Xue-dong, and WANG Jun-neng, Brief Introduction to Food Safety Evaluation and Its Methods, J. Livestock and Poultry Industry. vol.242, No.6 (2009) 8-10.

[10] Liu Wen, Li Qiang, Liu Peng, Duan Min, DaiYue and Zheng Jiajia, Construction of Food Index and Its Empirical Analysis, J. Food Science. vol.36, No.11 (2015) 191-196.
[11] Yang Ying, Liu Yan-ming, Zhang Hui, Zhu Jian-hua and Hu Mei, Study on the Risk Index Evaluation of Food Nutrient, J. The Food Industry. vol.34, No.11 (2013) 187-189.
[12] SunYan-bin, Sun Ting, Dong Shu-xiang, Li Shi-kai, Zhong Qing and Zhang Jun, The application of contamination index method to evaluate heavy metal contaminations in dairy products, J. Chinese Journal of Food Hygiene. vol.27, No.4 (2015) 441-446.