Advance Journal of Food Science and Technology 8(10): 707-710, 2015
DOI:10.19026/ajfst.8.1593
ISSN: 2042-4868; e-ISSN: 2042-4876

Submitted: October 22, 2014 Accepted: January 2, 2015 Published: July 10, 2015

Research Article

Research of Food Safety Risk Assessment System Based on Data Mining

Liu Xin
Hunan Railway Professional Technology College, Hunan 412001, China

Abstract: Data mining is a new data analysis technology, playing an increasingly important role in many industries. Taking it into the food safety inspection data analysis can make food safety testing data analysis and early warning more intelligent and precise. In this study, a data mining subsystem of the CQS platform is detailed designed. A mining database is made from the data published by General Administration of Quality Supervision, Inspection and Quarantine. On basis of this database, mining model set used by the subsystem is established; meanwhile mining results and performance are analyzed.

Keywords: Data mining, food safety inspection, quality supervision

INTRODUCTION

Data mining is a process as well as a knowledge discovery in database, namely, it is from the large, incomplete, noisy, fuzzy and random data to extract implicit information that people do not know in advance, but it is potentially useful information and it is a non- trivial process for knowledge. Data mining is originated from many disciplines, including database, artificial intelligence, statistics, machine learning, etc. Among them, the most important three fields are database, machine learning and statistics. These different historical influences made the different scholars hold different views on the function of data mining.

The classification of data mining: Data mining is related to many disciplines and methods, therefore, there are data a variety of classification methods for data mining. According to the task of data mining, it can be divided into classification or warning model discovery, data summarization, clustering, association rules discovery, sequence pattern discovery and dependency relation or the dependent model discovery, exception discovery and trend discovery, etc.; according to the object of data mining, it can be including the relational database, object-oriented database, spatial database, temporal database, text database, multimedia database heterogeneous database, heritage database and Web, etc.; according to the method, it can be divided into the machine learning method, statistical method, neural network method and database method. While machine learning methods can be divided into inductive learning methods (decision trees, induction of rules, etc.) based on the case study, active learning, genetic algorithms, etc (Riden and Bollen, 2007). Statistical analysis methods can be divided into regression (multivariate regression and autoregressive regression, etc.), discriminant analysis (Bayesian discriminating, Fischer discriminant, nonparametric discriminant, etc.), cluster analysis (hierarchical clustering, clustering segmentation, etc.), exploratory analysis (principal component analysis, correlation analysis, etc.) and so on (Tan *et al*., 2006). The artificial neural network method can be divided into feed forward neural networks (BP algorithm), self organizing neural network (self-organizing feature map, competitive learning, etc.). The database method mainly includes the multidimensional data analysis, attribute-oriented induction method, etc.

MATERIALS AND METHODS

Selecting data: The data are obtained through different data sources, the data source can be included data warehouse, data mart or other existing databases. The goal of selecting data is to prepare for the data mining of the next step. Choosing how many data and what kind of data is very important for data mining. If the amount of the available data is very huge and there are lots of limitations in time and space during calculation, then it has to use a sampling technique. Using a portion of data or data sampling is perhaps the only way for the completion of the project within a specified time (Liang, 2006). The key of data sampling is to ensure that the data sampling can be a better representative sample of the entire data generally. The format of data and the type of data is the most basic component of data mining. As for the continuous variables, during the period of analyzing the quality of data, it needs to check the reasonable scope of each variable. Selecting data must firstly understand the purpose of data mining and then data mining can be performed on the original data to select the reasonable properties. The purpose of the

Table 1: Table of food data

The historical evaluating data sheet
Inspection number
The category of food
The name of food
The risk grade of enterprise
The risk grade of official regulatory
The grade of trade
The risk grade of food
Quarter
Month

early warning of food inspection is to analyze the properties of the inspection food information, mining through the historical evaluation data, so as to get the relation between the properties of inspection food and the rating of the risk assessment (Hand *et al*., 2001). The historical evaluating data mainly includes: categories of food, food, the risk grade of food safety, risk of food characteristics, enterprise risk, the official regulatory risk, etc. Because there the inspection product is not detected, so the selected historical evaluating data is without the risk of food characteristics. The selected historical evaluating data table is as shown in Table 1.

The method of associated rules mining is based on analyzing the evaluation of the risk grade of food and early warning of food safety and other factors, which has incomparable advantages over the method of mathematical statistics, thus the obtained rules are more intuitive. This study had in-depth exploration on the historical food evaluating data, using the method of associated rules to generate the rules database, completing the project risk evaluation and prediction and detection on the risk grade of the non-detected food, giving the early warning information. The purpose of doing so is not only can enable enterprises to see why food is with high risk, so as to improve it in the future and improve the economic benefits of food enterprises, as well as promote the development of economy; it also can reduce the cost of food inspection, improve the efficiency of supervision; it can reduce the exporting risk of national food so as improve the exporting credit for the country.

Let $I = \{I_1, I_2, I_3 ... I_m\}$, which is a set of items. Let D the task related data is a set of database transactions, among them, T is a set of each transaction item, Let $T \subset I$, let A be a set, then transaction T contains A, if and only if $A \subseteq T$. The associated rules are shaped like $A \Rightarrow B$, among them, $A \subset I$, $b \subset I$, moreover, $A \cap B = O$. The rule $A \Rightarrow B$ can be set up in set of D with support for s, among them, s is the percentage of D contains AUB (i.e., P (AUB)). Rule $A \Rightarrow B$ in set of D has confidence grade c, c is the all transactions in set of D containing A, based on it, the probability of containing B is (also known as P (AUB)). If the supporting grade of the set of I can meet the predefined minimum support threshold, then I is called frequent item set.

The algorithm of the frequent mining association rule can be described as follows.

The initial state:

$L = U_K L_K; AR = \phi; // L$ is a frequent item set, AR is a set of frequent association rules

for all λk (λk is the element of L, which is a frequent item set of k-, the size of c is n) {

for αk (αk is a nonempty and proper subset of λk) {

if (as well as is the confidence grade of $\alpha k \rightarrow \beta m >$ = minCont) {//

Here, $m+k = n$, meanwhile, $\alpha k \rightarrow \beta m$ is an association rule

$AR = AR \cup (\alpha k \rightarrow \beta m)$;

}

}

}

Return AR;

Model construction: Constructing the model is the core of data mining. Once the data cleaning and the transformation of variables are completed, the model construction is began. Before the construction of model, it must understand the target if the data mining and the types of data mining. In this stage, it needs for the cooperation with the relevant analysts who had the relevant knowledge. After understanding the task of data mining, selecting the appropriate algorithm becomes relatively easy (Henson and Caswell, 1999). Each data mining task should be corresponding to the appropriated algorithm. In most cases, we don't know what kind of algorithm is the most suitable before constructing the model. The accuracy of data depends on the properties of the algorithm. The correct way is to use different algorithms to construct multiple models, and then use the tools to evaluate the accuracy of these models. Even with the same algorithm, it should also set different parameters to build multiple models, so that it is conducive to adjust the accuracy of the model.

RESULTS AND DISCUSSION

Clustering analysis model: Clustering analysis model must meet the requirements:

- Each model must contain a unique key column, it can be a numeric or text column, which is used to identify each record uniquely.
- Each model must contain at least one input column, the input column should contain the value which is used to generate the classification. The input column can be set more at random, while adding additional input column will increase the time for the model setting.
- As far as this model is concerned, the predictable column is not required necessarily, but the predictable column can be added.

The value of the predictable column can be considered as the input of a clustering model, which also can be specified only for the purpose of prediction (Fig. 1).

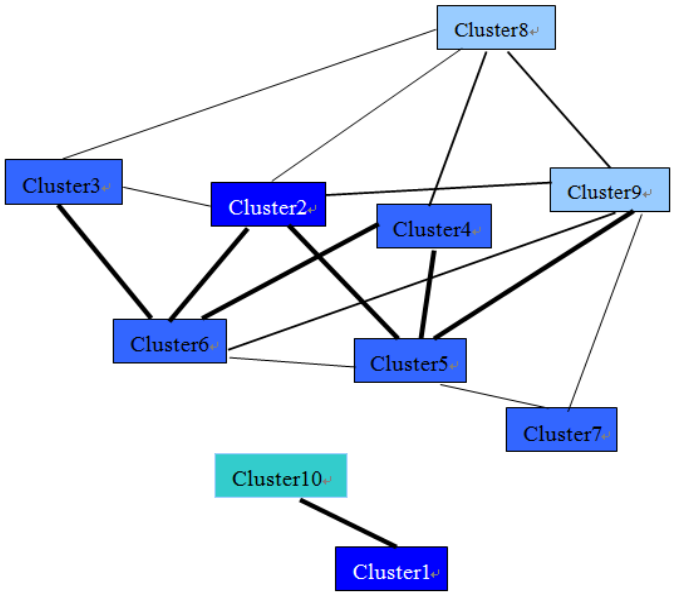

Fig. 1: Clustering links

Verifying mining model: Verification is the process of assessing the implementation of mining model over the real data. Before setting the mining model in the production environment, it must verify it through the understanding of its quality and characteristics. It can use a variety of methods to assess the quality and feature of data mining model. First of all, it can use various statistical validity information to determine whether there are problems in data or models. Secondly, the data can be divided into the set of training and set of test so as to test the accuracy of prediction. Finally, it can ask the business experts to check the result of the data mining model to determine whether the pattern of findings in the target business scheme is meaningful. All of these methods in data mining methods are very useful in creating, testing and optimizing models to solve specific problems which can be used repeatedly.

While Apriori algorithm is used for the Boolean data, so the historical evaluation data need to be converted. The reflection of food category is *S*, wherein *x* represents the number of the category, *y* represents the number of sub-class. The reflection of time and quarter is J_1-J_4. The reflection of month is M_1-M_{12}, the enterprise risk level is set Q_L, Q_M, Q_S respectively as high risk, medium risk, low risk, for example, the official risk is set as G_L, G_M, G_S, the trade risk is set as T_M, T_L, T_S, the result of the risk assessment is set as P_L, P_M, P_S, which is encoded as Table 2.

Using Apriori algorithm for mining experiment, having inspection on food safety risk assessment of food inspection and recording the assessment of food inspection, setting the minimum supporting degree is 20%, the minimum confidence degree is 80%, which can get 1050 mining association rules, the obtained

Table 2: Data after being encoded

S_{01}, S_{015}, Q_S, G_S, T_S, J_3, M_9, P_S
S_{02}, S_{021}, Q_M, G_M, T_L, J_2, M_4, P_M
S_{02}, S_{022}, Q_M, G_S, T_L, J_2, M_4, P_M
S_{01}, S_{15}, Q_L, G_M, T_S, J_3, M_8, P_M
S_{02}, S_{012}, Q_M, G_M, T_M, J_2, M_5, P_M
S_{02}, S_{021}, Q_M, G_L, T_M, J_2, M_6, P_M
S_{02}, S_{022}, Q_S, G_M, T_M, J_1, M_1, P_S
S_{01}, S_{15}, Q_L, G_L, T_M, J_2, M_5, P_L

association rules are too large, in fact, the association rules that we are interested in are small parts, according to the association rules and Apriori information, we select the related information. As for the predicting grade of evaluating safety of food inspection, the constraint of the association rules is as followed: the later must have one and only one result of the risk assessment. The former one should include time category, together with some other attributes, the more the content includes, the better the result is, therefore, the result of the safety risk evaluation on the inspection of food is more accurate. After the constraints and deletion of the association rules, the amount of the association rules is less 70%, finding out the association rules mainly as follows:

- The kind of food = plant food, enterprise risk = high risk, the official risk = high risk, quarter = the second quarter --> the grade of risk evaluation = high risk
- The food sub-class = cabbage, enterprise risk = high risk, the official risk = medium risk, trade risk = low risk, quarter = third quarter --> the grade of risk evaluation = medium risk
- The food sub-class = ginger, venture enterprise risk = medium risk, month = August --> the grade of risk evaluation = medium risk

The first rule: Some enterprises have inspection on the plant food of the second quarter, if the risk of the enterprise is high and the official risk is high, under this circumstances, through the rules that we can know the plant food is the food with high risk.

The second rule: If the enterprise exports cabbage in the third quarter, the enterprise is the enterprise with high risk, while the official department is not in strict accordance with the policy regulation, but the exported amount is reasonable, then the food safety risk is medium risk.

The third rule: If the enterprise exports ginger in August, the enterprise is the enterprise with medium risk, but the exported amount exceeded the standard exported amount too much, the food safety risk is regarded as medium risk.

It can be seen from the above examples, rules generated by association rules mining revealed the hidden relation between the index data in the risk assessment, the produced rule is meaningful, on one hand, these rules are based on a large amount of historical evaluation data by mining, which is in line with the practical significance; on the other hand, the risk of the enterprise has a certain impact on the export of food, if an enterprise does not have the exporting credit, which is often informed, with this kind of phenomenon, the exported food naturally should suffer from the strict management, the risk of the exported food is high.

CONCLUSION

The official risk is also the same, if the official department can go on with food sampling in the process of food production, so that enterprises cannot make rectification with high risk food during the process of the food production, which will increase the food risk of the enterprise. Therefore, the exported risk in the actual application should be combined with professional knowledge to analyze the rules produced by mining, at the same time; these rules also provide some clues and inspiration for the professional research.

REFERENCES

Hand, D., H. Mannila and P. Smyth, 2001. Principles of Data Mining. The MIT Press, Cambridge, MA.

Henson, S. and J. Caswell, 1999. Food safety regulation: An overview of contemporary issues. Food Policy, 24: 589-603.

Liang, X., 2006. Data mining, model, algorithm and application and system. Comput. Technol. Dev., 1: 14-25.

Riden, C.P. and A.F. Bollen, 2007. Agricultural supply system traceability, Part II: Implications of packhouse processing transformations. Biosyst. Eng., 98: 401-410.

Tan, P.N., M. Steinbach and V. Kumar, 2006. Introduction to Data Mining. Pearson Addison Wesley, Boston, Vol. 16.