Advance Journal of Food Science and Technology 10(12): 970-973, 2016
DOI: 10.19026/ajfst.10.2296
ISSN: 2042-4868; e-ISSN: 2042-4876

Submitted: July 13, 2015 Accepted: August 30, 2015 Published: April 25, 2016

Research Article

A Novel Risk Warning Approach to Food Enterprises Using Data Mining and Information Fusion Technologies

Zhen-Feng Jiang
School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277160, China

Abstract: The food enterprise risk warning problem is an important issue both in the theoretical and practical aspects. In order to improve the risk warning level of food enterprises, a novel risk warning approach is studied using some emerging technologies, such as, data mining and information fusion. For this reason, a novel risk warning technology is proposed to food enterprises based on data mining and information fusion. Experimental results suggest that this technology is feasible, correct and valid. Experimental results and trend analysis has many practical significance for assisting government management and decision of import and export food agricultural product safety.

Keywords: Data mining, food enterprise, information fusion, risk warning

INTRODUCTION

Data mining aims to analyze, extract rules and development trend among data, discover relations among discrete data, exclude interference of rubbish data and form useful knowledge. Some believe this is an advanced stage of Online Analytical Processing (OLAP). Data mining tool can utilize existing analysis tools to analyze relations among data in mass information and establish effective analysis models (Dai *et al*., 2014; Demetis and Angell, 2007). These models and relations can be used to carry out trend analysis. It is the process of seeking potential and useful information and knowledge among mass, noisy and irregular practical application data with incomplete information (Garcia, 2013; Pang *et al*., 2015).

In the face of mass import and export food and agricultural product inspection result information, entry and exit inspection and quarantine hopes to establish data warehouse of food monitoring information through integrating inspection results of testing organizations in each place and combine statistical analysis and data mining to achieve real-time monitoring and early warning of food safety state, evaluate food safety in an scientific and effective way, accurately predict food safety development trend and provide scientific decision basis for supervision institutions (Yager, 1987; Agus *et al*., 2010; Ross and Hann, 2007). For example, big data mining can discover regional and seasonal distribution rule information of a pesticide in food and calculate the development trend according to time change.

With the development of emerging technologies, some scientific approaches should developed to the risk warning of food enterprises. For this reason, a novel risk warning technology is proposed to food enterprises based on data mining and information fusion.

PROPOSED METHODS

In mass data gained from food and agricultural product safety inspection, much potential and useful information is hidden. Entry-exit inspection and quarantine department hopes to extract information for decision making, early warning and trend analysis through data mining, discover unqualified items of food and agricultural products and spatial distribution rules and predict development tendency. If pesticide application dosage in a region continues to rise and positive results increase year by year, the early warning information of export onions can be proposed.

To be more specific, multi-dimensional analysis of food name, type, production enterprise, region, inspection item, inspection result, time, disqualified items is carried out first. Main method involved in this study is to use SAS tool to mine association rules and establish data models so as to find the relevance or relation among item sets in mass inspection data. Association rules refer to inspection of correlative dependence among information. The discovery rules can be used to discover strong rules from data warehouse that Conk-dente ad support are the threshold value set in advance. SAS system is used to construct a complex but useful data structure-data cube according to data analysis demand. Association rule mining is applied to analyze effectiveness of inspection items, frequency and casual inspection, find casual inspection effectiveness of food and agricultural products and analyze regional distribution feature of positive result and time distribution rules. According to distribution

characteristics discovered, spatial and temporal distribution rules, discovery rules are utilized to analyze the trend of safety risk factor. Trend prediction exceeding the preset value may serve as the early warning information.

Information fusion model adopted in this study is established on the basis of Back Propagation (BP) neural network, Support Vector Machine (SVM) and Dempster-Shafer (DS) evidence theory (it is called as IFCA). DS evidence theory is a method to handle uncertain information, which develops on the basis of Bayesian theory. It can handle uncertainty caused by both randomness and fuzziness. DS method can continuously narrow hypothesis set by depending on credibility distribution function accumulation and differentiate 'the unknown' and 'uncertainty'. Potential advantage of DS method is that it does not need prior probability and conditional probability density. In view of advantage in this aspect, it can be completely applied in uncertain assessment independently. DS evidence inference method is as follows.

Definition 1: We assume U is the set of propositions which consist of some mutually exclusive and exhaustive elements, called identification framework and ϕ means null proposition set. When the function m: $2^U \to [0, 1]$ meets the following conditions:

$$\begin{cases}(1)\, m(\phi) = 0 \\ (2)\sum\limits_{A \subseteq U} m(A) = 1\end{cases}$$

We call $m(A)$ is basic trust measure of proposition A and $m(A)$ means accurate trust degree of proposition A, i.e., direct explanation for A.

Definition 2: We assume U is an identification framework and m: $2^U \to [0, 1]$ is basic trust measure on U; definition:

$$Bel: 2^U \to [0,1]\ Bel(A) = \sum_{B \subseteq A} m(B), (\forall A \subseteq U)$$

Definition 3: We assume Bel_1 and Bel_2 are two trust functions on identification framework U; $m_1$and $m_2$are the trust degree respectively; focal elements are $A_1, A_2, \cdots, A_k$ and $B_1, B_2, \cdots, B_r$ and:

$$m(C) = \sum_{A_i \cap B_j = C} m_1(A_i) m_2(B_j) \Big/ (1-K)$$

$$K = \sum_{A_i \cap B_j = \phi} m_1(A_i) m_2(B_j) < 1, \forall C \subseteq U, C \neq \phi, m(\phi) = 0$$

Definition (3) describes D-S composition rule. Since it meets people's several basic desire characteristics for composition rule (focus, commutativity and modularity), it is evidence composition rule which is most widely applied. Meanwhile, it is also the foundation of studying other composition rules. The principles of information fusion assessment model established by use of BP network, SVM and DS evidence theory are as follows.

Specific design algorithm of credit risk assessment model based in BP network is as below. For data of commercial banks, BP neural network model may be used to establish 4-input and 5-output structure and 1 hidden layer (there are 9 nodes at hidden layer). Network output is set to a 5-dimension structure, including (1 0 0 0 0), (0 1 0 0 0), (0 0 1 0 0), (0 0 0 1 0) and (0 0 0 0 1). Besides, through setting the output layer to sigmoid function, each component of network output can be between (0, 1). Network output may cause that the sum of each number among the 5-dimension vector is not equal to 1. We assume vector of any output is (y_1, y_2, y_3, y_4, y_5). $\sigma = \sum_{i=1}^{5} y_i$ makes network output normalization be (y'_1, y'_2, y'_3, y'_4, y'_5), where $y'_i = \frac{y_i}{\sigma}, i = 1,2,3,4,5$. y'_i denotes the basic credibility for i^{th} condition.

Specific algorithm of SVM-based credit risk assessment model is as follows. To effectively relieve calculation complexity, SVM adopts least squares support vector machine. Kernel function adopts radial basis function. For SVM, it includes 2 outputs, +1 and -1 which represent corresponding classification situations. For recognition rate of sample classification, the following rules are adopted:

Rate of correct sample classification: $c_i = C_i/N_i$

Mean rate of correct sample classification:

$$c_i = \frac{\sum_j C_j}{\sum_j N_j}$$

Omission rate of sample: $l_i = L_i/N_i$.

In the above detection index formula, N_i means the number of samples of the i^{th} type; c_i means the samples correctly identified in the i^{th} type of samples; L_i means the samples which are not identified in the i^{th} type of defective samples. For SVM, classification results utilize fuzzy mathematics thought to test dependence degree of input samples on each type of samples. Basic credibility result for 5 situations (y''_1, y''_2, y''_3, y''_4, y''_5) and y'' is basic credibility of the i^{th} situation) is gained through one-by-one classification of SVM. In line with (y'_1, y'_2, y'_3, y'_4, y'_5) and (y''_1, y''_2, y''_3, y''_4, y''_5) gained in each situation, DS evidence inference is utilized to fuse (y'_1, y'_2, y'_3, y'_4, y'_5) and (y''_1, y''_2, y''_3, y''_4, y''_5). Normal, attention, secondary, suspicious and loss are expressed with A1, A2, A3, A4 and A5, respectively. Output result of data settled with BP network is

Table 1: Subject index risk comment set of food enterprise

		Risk comment set									
		High		Medium		Medium-low		Low			
Index weight										Uncertainty	
0.33	0.33	0.327	0.26	0.433	0.177	0.14	0	0	0	0.1	0.563
	0.67		0.09		0.54		0.27		0		0.1
0.1404		0.153		0.192		0.038		0		0.617	
0.1996		0.272		0.272		0		0		0.456	
0.33		0. 09		0.54		0.27		0		0.1	

Table 2: Behavioral index risk comment set

		Risk comment set				
Index type	Index weight	High	Medium-high	Medium-low	Low	Uncertainty
Long-term behavior risk	0.5	0	0	0.386	0	0.614
Short-term behavior risk	0.5	0	0	0	0.9	0.1

Table 3: The food enterprise risk

		Comment set				
Criterion layer	Weight	High	Medium-high	Medium-low	Low	Uncertainty
Subject risk	0.3667	0.081	0.379	0.058	0.0035	0.4787
Behavior risk	0.6333	0	0	0.0548	0.7846	0.1606
The food enterprise risk		0.0246	0.1152	0.0733	0.7169	0.07

(y'_1, y'_2, y'_3, y'_4, y'_5). Make $m_1(A_i) = y'_i$, i = 1, 2, 3, 4, 5; similarly, make $m_2(A_i) = y''_i$, i = 1, 2, 3, 4, 5 and $m_1(A_i)$ means A_i credibility for an input gained through BP algorithm; $m_2(A_i)$ means A_i credibility for an input gained through SVM (A_i and A_j are mutually independent). According to DS composition formula:

$$K = \sum_{i \neq j} m_1(A_i) m_2(A_j), 1 \le i, j \le 5$$

$$m(A_i) = \sum m_1(A_i) m_2(A_i) / (1 - K), i = 1, 2, 3, 4, 5$$

Make $\hat{y}_i = m(A_i)$, i = 1, 2, 3, 4, 5. Then, is A_i credibility gained by DS inference. Then, recognition rate of the information fusion method can be obtained to make assessment decision.

EXPERIMENTAL RESULTS

To test accuracy and practicability of model, this study applies Matlab R2010a to process and classify real data under this model to gain final risk rating of users and known suspicious transactions so as to complete empirical test through comparison.

Information fusion: A food enterprise in the current province is chosen. Corresponding attributes of customers are extracted and scoring results of subject risk index correspond to them. In combination of elementary probability assignment function of customer subject risk index, subject index risk comment set of a customer is gained, as shown in Table 1.

According to data preparation work in the initial period of experiment, these attribute values of all enterprises under different time series are concluded and imported. Matlab programming tool is used to calculate clustering outlier and then classification result and outlier degree of behavior indexes within a time series are gained. In combination of elementary probability assignment function of behavior indexes, behavioral index risk comment set of the enterprise is gained, as shown in Table 2.

According to the proposed approach, elementary probability assignment and weight of each index, risk fusion is conducted. The food enterprise risk assessment result are listed as shown in Table 3. It is known from Table 3 that main comment of food enterprise risk is low in recent 10 months and owns small uncertainty.

CONCLUSION

A novel risk warning technology is proposed to food enterprises based on data mining and information fusion. Experimental results suggest that this technology is feasible, correct and valid.

REFERENCES

Agus, S., N. Sheela, Y. Ming, Z. Aijun, K. Daniel and C.D. Fernando, 2010. Statistical methods for fighting financial crimes. Technometrics, 52(1): 5-19.

Dai, Q., D.W. Sun, Z.J. Xiong, J.H. Cheng and X.A. Zeng, 2014. Recent advances in data mining techniques and their applications in hyperspectral image processing for the food industry. Compr. Rev. Food Sci. F., 13(5): 891-905.

Demetis, D.S. and I.O. Angell, 2007. The risk-based approach to AML: Representation, paradox and the third directive. J. Money Launder. Control, 10(4): 412-428.

Garcia, A.B., 2013. The use of data mining techniques to discover knowledge from animal and food data: Examples related to the cattle industry. Trends Food Sci. Tech., 29(2): 151-157.

Pang, Z., Q. Chen, W. Han and L. Zheng, 2015. Value-centric design of the internet-of-things solution for food supply chain: Value creation, sensor portfolio and information fusion. Inform. Syst. Front., 17(2): 289-319.

Ross, S. and M. Hann, 2007. Money laundering regulation and risk-based decision making. J. Money Launder. Control, 10(1): 106-115.

Yager, R.R., 1987. On the Dempster-Shafer framework and new combination rules. Inform. Sciences, 41(2): 93-137.