The Application of Big data Mining in Risk Warning for Food Safety

Yajie WANG, Bing YANG, Yan LUO, Jinlin HE, Hong TAN*

Guizhou Academy of Testing and Analysis, Guiyang 550002, China

Abstract Comprehensive evaluation and warning is very important and difficult in food safety. This paper mainly focuses on introducing the application of using big data mining in food safety warning field. At first, we introduce the concept of big data miming and three big data methods. At the same time, we discuss the application of the three big data miming methods in food safety areas. Then we compare these big data miming methods, and propose how to apply Back Propagation Neural Network in food safety risk warning.

Key words Food safety, Big data mining, Risk warning, Bayesian network, Decision tree, BP neural network

1 Introduction

In recent years, the new services about information industry have boomed, and the data type and scale of various industries show an exponential growth trend. The era of big data in China has started. The rise of big data concept provides us with a new way of looking at the world, and to make large amounts of data stored in the database become valuable, big data mining has become a common concern. The study on big data mining methods becomes a great challenge in today's society. At the same time, food safety incidents occur frequently in China, such as melamine incident and poisonous rice event, posing a serious threat to people's health and causing negative social effects. Therefore, the comprehensive evaluation and early warning of food safety are increasingly becoming the focus of food safety. An effective early warning method can greatly improve the level of food safety, and big data mining technology is such an effective way of early warning. In this paper, we analyze the basic concept of big data, and analyze three typical mining methods in big data mining as well as their application in food safety. Based on the comparative analysis of application of three big data mining methods in food safety risk warning, we select the best big data mining method and explore the application of it in food risk warning.

2 Overview of big data

The use of the term "big data" dates back to the open source project Nutch of apache org. At that time, big Data was defined as large amounts of data for batch processing or analysis required by web search update[1]. Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate[2]. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reductions and reduced risk. Grobelink thinks big data have the following three features (3V): Volume; Velocity; Variety[3]. It is the representative definition at present. In addition, based on 3V, some large enterprises make 4V definitions, namely adding a new feature on the basis of existing 3V. Currently, the definition of the fourth V has not yet been unified, and IDC believes that big data should also have "Value"[4], while IBM thinks big data are bound to have "Veracity"[5].

3 Overview of three big data mining methods and their application in the food safety industry

In recent years, with the steady development of computer hardware, a lot of powerful data collection devices and storage media are widely available on the market, which vigorously promotes the development of the database and makes large amounts of information and data stored in the database[6]. Big data mining is to extract potential hidden information from the data, and conduct automatic mining in the database by the development of computer programs. It is an effective means to find law or model[6]. Big data mining is to dig knowledge from big data[7]. The common machine learning data mining technologies include Bayesian network, decision tree, artificial neural network and the like.

3.1 Bayesian network Bayesian network, developed by Pearl in 1988[8], is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG)[9]. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Formally, Bayesian networks are DAGs whose nodes represent random variables in the Bayesian sense: they may

Received: April 21, 2015 Accepted: June 25, 2015

Supported by Soft Science Research Project of Guizhou Province (R20142023); Key Youth Fund Project of Guizhou Academy of Sciences (J201402).

* Corresponding author. E-mail: wyj@gzata.cn

be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies; nodes that are not connected represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables, and gives (as output) the probability (or probability distribution, if applicable) of the variable represented by the node. Similar ideas may be applied to undirected, and possibly cyclic, graphs; such are called Markov networks. Efficient algorithms exist that perform inference and learning in Bayesian networks. Bayesian networks that model sequences of variables are called dynamic Bayesian networks. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams[10]. Bayesian networks are used for modelling beliefs in computational biology and bioinformatics medicine, biomonitoring, document classification, information retrieval, semantic search, image processing, data fusion, decision support systems, engineering, sports betting, gaming, law, study design and risk analysis. As for the application in the food industry, Bayesian networks are mainly used for food product design[11]. For example, in the food Bayesian network modeling, if we know people generally prefer sweet foods and there are sweet and popular foods in the samples, then Bayesian network infers that the color of the food will affect its popularity. The traditional rule-based expert recommendation system can not deal with such similar problems because the system is modular, and some of the rules have nothing to do with other rules or content of data sources, while the conditional probability of the Bayesian network solves this problem. In addition, the Bayesian network model, the representative of probabilistic risk assessment model, was once used in risk probability estimates of the food supply chain[12]. Due to differences in the response of food supply chain to different initiating events, the development processes and results of events are also different. By obtaining the conditional probability value of each node of the Bayesian network, we can get the risk value of food by calculating the joint probability.

3.2 Decision tree A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. A decision tree can better learn noise data, learn the law and extract expression[13]. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. In decision analysis a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated. A decision tree consists of 3 types of nodes: decision nodes-commonly represented by squares; chance nodes - represented by circles; end nodes- represented by triangles. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities. Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods. As for the application in the food industry, decision tree is used for the assessment of food safety based on agricultural products[14]. Based on the data features and dimensionality reduction approach, it conducts preprocessing to identify the main eigenvalues affecting the quality and safety, establishes the agricultural quality and safety criterion models on the basis of combinatorial optimization decision tree, and chooses different factors influencing agricultural products (such as heavy metal content in groundwater, soil pH and planting scale) as the attributes of decision tree. The data samples are divided into training set and test set, and the rule set is obtained through training. The data samples in test set are input into the decision tree model to calculate the rate of accuracy, thereby determining whether the decision tree can assess quality and safety risk of agricultural products. Decision tree is also used to assess the quality of instant fried noodles[15].

3.3 Artificial neural network Artificial neural networks, as a data mining approach with high learning accuracy, simulate the complex network consisting of interconnected neurons in biology for modeling. Artificial neural networks are a family of statistical learning models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which send messages to each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read. Like other machine learning methods- systems that learn from data - neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition. Currently, there are about a dozen of artificial

neural network models, and the common artificial neural network is BP (Back Propagation) neural network[13]. BP neural network has the mapping function of nonlinear function, and strong information storage capacity as well as large-scale parallel processing capabilities. Its good adaptability and capacity of resisting disturbance make it have a strong ability to learn. BP neural network is an effective method to deal with uncertainty problems in artificial intelligence, and it was once combined with the principal component analysis for the study of near infrared spectroscopy apple variety identification[16]. The study first uses principal component analysis to conduct clustering on apple and gets the near infrared fingerprints of apple. For the characteristic wave bands sensitive to apple varieties, the characteristic band map is regarded as the input of neural network, and varieties as output. The model is established, the training is carried out, and then the unknown samples are forecasted. This variety identification accuracy rate reaches 100%. In addition, BP neural network is also used for water consumption of winter wheat[17], and forecast of rice amylose content[18].

4 Comparative study on the application of the three big data mining methods in food safety risk warning

Bayesian network, decision tree and BP neural network are the most efficient classification methods for data mining. By modeling and training, the model learns the classification rules, and when there are new and unknown types of data, the model can have the ability to identify according to the learning experience. We call it forecasting ability. In Bayesian classification, all the properties will be involved in the calculation and classification. Decision tree is a two-branch or multi-branch tree for discrete variables, with a top-down recursive structure. Each leaf node of the tree represents a category. BP neural network is the classifier based on perceptron, and produces linear or non-linear discriminant function by the iteration of training mode and learning algorithm. BP neural network carries out training by positive propagation, weight adjustment and reverse propagation. Neural networks put the knowledge and laws learned in the memory of network weight, in order to identify the implicit data rules. The weight value of BP artificial neural network is not calculated but completed by the network training. The larger the amount of data, the more the training set, and the more accurate the classifiers. Bayesian network and BP neural network have high accuracy, while the accuracy of decision tree to a large extent depends on the completeness of the data, and the default values in certain fields will affect its accuracy. The more the default values, the less accurate the decision tree. For food safety testing data, there are many testing indicators, and many testing results are not judged or detected, leading to excessive default values, which will have a great impact on the decision tree learning. From the training rate, in the big data environment, there are many detection indicators for a food, that is, there are too many attributes. Since Bayesian network relies on probability calculation, the complexity of calculation of attribute set will increase, making it more difficult to forecast and requiring longer time. Decision tree implements depth-first search, and its algorithm is limited by memory size, so it is difficult to handle large training set. With the growth of amount of data, the processing speed of decision tree will slow down a lot. The perceptron-based neural network has a good ability to deal with uncertainties, and a large number of neuron onlooker activities constitute the overall macro effects of neural networks. It has good adaptability, and as the amount of data increases, the model will be more accurate. Unlike Bayesian network and decision tree, neural network learns the rules by self-adjustment of weight values, so it is better than the first two algorithms in terms of training speed. In terms of robustness, there are often default values or noises in the food inspection data, and Bayesian network can not accurately define the probability of noise or default values, which will affect the model training. Due to poor fault tolerance, decision tree is strongly dependent on data, and data noise or incompleteness will affect the building of decision tree model. The neural network itself has high fault tolerance, and if part of the data are incomplete, then the neural network can learn the hidden laws from another part of data, and adjust its own weight for law learning, to construct a robust model. In summary, BP neural network has high accuracy and high training speed, and its robustness is better than that of other two kinds of data mining methods. And it has advantages in parallel processing, self-learning and fault toleration. BP neural network has a strong degree of flexibility, and the new training data set can be used for model training, thereby increasing the accuracy of the model, so it is suitable for the application in food safety risk warning areas.

5 The application of BP neural network in food risk warning

BP neural network is a kind of artificial neural network and an important artificial intelligence tool. Through a lot of sample training, it gets the implicit laws of model. In food testing, we tend to get simple judgment of pass or fail. This test result is clear, but it is not conducive to the control of food safety risks. If we grade the degree of food risks based on the specific detection value, it will help to provide decision support for the relevant risk assessment departments. Traditional risk rating methods include expert scoring[19], risk matrix[20] and so on. These methods are accurate, but the expert scoring is based on human assessment, and the risk matrix consumes considerable manpower and physical resources, so both of them do not apply to the risk warning of big food safety data. Based on the underlying laws of BP neural network, we make it learn the risk rating of expert scoring method. When there are new test data, it can evaluate the data according to the laws learned. We believe that in the context of big data, BP neural network is very suitable for the food safety risk warning based on certain types of food. First of all, it screens the different dimensions that affect the test results of certain foods, such as chemical contamination, pesticide residues, veterinary drug residues, heavy

metals, bacteria and the like. Using expert scoring method, the risk rating is conducted based on the experts combined with the test results of test items of different dimensions. Then the testing values of the above dimensions are regarded as the input neurons of neural network, and the number of neurons is determined by the selected dimension. And the rating obtained through expert scoring is taken as the target output neuron for training. A lot of input and output samples are sent to the neural network to make it conduct potential rule learning through forward propagation, back propagation and weight adjustment. When there are unknown data, by inputting the samples to neural network, the neural network can simulate experts for rating. Since the neural network has high flexibility, the new data and rating can be also regarded as the training set of the neural network. So, with the increasing amount of data, the neural network model will be more accurate, in order to reduce errors caused by human factors as well as labor costs.

6 Conclusions

Firstly, this paper analyzes the basic concept of big data, machine learning in big data mining and three typical mining methods, and explores their use in food products. Secondly, this paper conducts a comparative analysis of the application of three big data mining methods in food safety risk warning. Finally, this paper comes up with the method of applying BP neural network to food risk early warning and explains why we believe that BP neural network is superior to the other two data mining technologies.

References

[1] MENG XF, CI X. Big data management: Concepts, techniques and challenges[J]. Journal of Computer Research and Development, 2013, 50(1): 146 -169. (in Chinese).

[2] Big data[EB/OL]. [2012 - 10 - 02] http://en. wikipedia. org/wiki/Big_dat.

[3] GROBELINK M. Big data computing: Creating revolutionary breakthroughs in commerce, science and society[R]. 2012.

[4] Hamish Barwick. The 'four Vs' of Big Data. Implementing information infrastructure symposium[EB/OL]. [2012 - 10 - 02].

[5] IBM. What is big data [EB/OL]. [2012 - 10 - 02]. http://www - 01. ibm. com/software/data/bigdata/.

[6] WITTEN IAN H., EIBE FRANK. Data Mining: Practical machine learning tools and techniques[M]. Morgan Kaufmann, 2005.

[7] HAN JW, KAN B. Data mining: Concepts and techniques[M]. Beijing: China Machine Press, 2001: 100 - 103. (in Chinese).

[8] PEARL JUDEA. Probabilistic reasoning in intelligent systems: networks of plausible inference[M]. San Mateo, Galif: Morgan Kaufmann Pub, 1988.

[9] LIN SM, TIAN FZ, LU YC. Construction and applications in data mining of bayesian networks[J]. Journal of Tsinghua University(Science and Technology), 2001, 41(1): 49 - 52. (in Chinese).

[10] JI JZ, LIU CN, SHA ZQ. Bayesian belief network model learning, inference and applications[J]. Computer Engineering and Applications, 2003, 39(5): 24 - 27. (in Chinese).

[11] CORNEY D. Designing food with bayesian belief networks[C]// ACDM 2000 Fourth International Conference on Adaptive Computing in Design and Manufacture. Springer London, 2000: 83 - 94.

[12] ZHANG L, TENG F, WANG P. Research on food supply chain risk assessment based on bayesian network[J]. Food Research and Developent, 2014, 35 (18): 179 - 182. (in Chinese).

[13] MITCHELL, TOM M. Machine learning[M]. WCB, 1997.

[14] ZHAO JX. Research on food safety evaluation based on decision tree[J]. Journal of Anhui Agricultural Sciences, 2011, 39(32): 20259. (in Chinese).

[15] OUYANG YF, XUE D, GAO HY, *et al*. Study on evaluation of fried instant noodle quality using decision tree[J]. Food Science, 2009, 30 (5): 27 - 31. (in Chinese).

[16] HE Y, LI XL, SHAO YN. Discrimination of varieties of apple using near infrared spectra based on principal component analysis and artificial neural network model[J]. Spectroscopy and Spectral Analysis, 2006, 26(5): 850 - 853. (in Chinese).

[17] CHEN B, OUYANG Z. Prediction of winter wheat evapotranspiration based on BP neural networks[J]. Transactions of the Chinese Society of Agricultural Engineering, 2010, 26(4): 81 - 86. (in Chinese).

[18] LIU JX, WU SY, FANG RM. Determination of apparent amylose content in rice by neural networks based on near infrared spectroscopy[J]. Transactions of the Chinese Society of Agricultural Machinery, 2001, 32(2): 55 - 57. (in Chinese).

[19] HAO SC, JIANG YN. Evaluation model of food suppliers based on improved principal component analysis[J]. Logistics Technology, 2010, 29 (8): 62 - 64. (in Chinese).

[20] LIU QJ, CHEN T, ZHANG JH, *et al*. Risk matrix - based risk monitoring model of food safety[J]. Food Science, 2010(5): 86 - 90. (in Chinese).

(From page 82)

[6] LIU HN, CHEN ZJ. An empirical study on the food safety credit rating of livestock products based on Fuzzy - AHP method[J]. Science and Technology Management Research, 2008, 28(5): 116 - 119. (in Chinese).

[7] SHI JP, WANG S, CHEN FS, *et al*. Food security risk assessment[M]. Beijing: China Agricultural University Press, 2010. (in Chinese).

[8] QU J. Food safety contrology[M]. Beijing: Chemical Industry Press, 2011. (in Chinese).

[9] MAO SY. On credit imperfection from the Confucian "Five Constant Virtues"[J]. Human Resources, 2011(9): 22 - 23. (in Chinese).