Pattern Discovery from Big Data of Food Sampling Inspections Based on Extreme Learning Machine

Yi Liu[1], Xin Li[2], Jianxin Wang[2], Feng Chen[3], Junyu Wang[1](✉),
Yiwei Shi[1], and Lirong Zheng[1]

[1] State Key Laboratory of ASIC & System, Fudan University, Shanghai 200433, China
{liuyi13,junynuwang}@fudan.edu.cn
[2] Beijing Forestry University, Beijing 100083, China
wangjx@bjfu.edu.cn
[3] Information Center for China State Food & Drug Administration, Beijing 100053, China
chenfeng@cfda.gov.cn

Abstract. Food sampling programs are implemented from time to time in local areas or throughout the country in order to guarantee food safety and to improve food quality. The hidden patterns in the accumulated huge amount of data and their potential values are worthy to research. In this paper, Extreme learning machine (ELM) is employed on real data sets collected from the food safety inspections of China in recent two years, in order to mine the relationship between food quality and food category, manufacturing site and season, inspection site and season, and many other attributes. Experimental results indicate that the ELM approach has better prediction precision and generalization ability than Logistic regression that was adopted in preceding work. The patterns obtained are helpful for making more effective food sampling plans and for more targeted food safety tracing.

Keywords: Food sampling inspection · Big data · Extreme learning machine
Logistic regression

1 Introduction

Food safety issues have aroused world-wide attention since it is closely related to public and household health and interests [1]. Most countries have implemented systems for food safety supervision and inspection, in order to reduce the quantity, strength, and impact of food safety incidents, and to improve the quality of food finally delivered to the end users [2]. However, food quality testing and food safety inspections are time-consuming, labor-intensive tasks, and they could sometimes be a heavy financial burden. Therefore, much research work has been done in order to improve inspection efficiency and effectiveness without increasing inspection quantity and strength, or even with reduced quantity and strength of food safety inspections [1, 3].

In China, many food safety incidents have occurred in recent years [3]. To deal with these problems, China government has taken a lot of measures to guarantee food safety and quality, and all levels of food testing laboratories in China carry out every day a

A. M. Tjoa et al. (Eds.): CONFENIS 2017, LNBIP 310, pp. 132–142, 2018.
https://doi.org/10.1007/978-3-319-94845-4_12

great deal of testing work. As a result, a large amount of food testing data is accordingly recorded and collected, and as a matter of fact, after years of accumulation, a huge data warehouse has come into being with rich information about food quality and safety and with many other properties. Initially, these accumulated data were only a matter of recording, and gradually they were utilized for inquiry and statistical purposes. The accumulated data, however, were found to be much more valuable than that [3], since the obtained patterns or rules underlying the data did provide us useful and helpful knowledge about the relationship among the attributes, which are able to help us make more effective and powerful inspection plans to expose more food safety problems, and hence to reduce consumption of time, labor, and financial burden.

Nevertheless, with the size of the data growing steadily in the course of food production, processing distribution and trading, the huge amount of data cannot be handled by conventional computing methods, which are by and by replaced by the technology of big data [4]. After the technologies of cloud computing and internet of things, big data technologies are another profound revolution that have penetrated into a variety of areas and given rise to dramatic changes in these areas. Big data is an abstract concept, with the characteristics of great quantity, rich variety, semi-structured and unstructured data, fast-growing, and that the traditional database management software cannot process it pragmatically with single-node computing resource. Consequently, distributed computing is the core method and key means in the bunch of big data technologies. Reference [5] examined the potential for big data application in the agriculture sector, including the variety and velocity characteristics in the sector and the integration of data and analysis that will be needed for successful implementation.

With the big data of food safety inspections accumulated, managed, preprocessed and analyzed, a variety of applications could be implemented, including dynamic and comprehensive food safety analysis, foodborne disease study, early warning and assessment of food safety, and so on. Fulfillments of these applications are helpful for boosting food quality level and improving food safety tracking. In order to implement these applications, however, a bunch of approaches are needed such as data preprocessing, statistical analysis, machine learning, and data mining [3, 6, 7].

Before applying methods mentioned above, complex processes should be taken for data preparation, including data cleaning, data normalization, and missing data imputation. In most cases, the phenomenon of missing data is inevitable in a real data set, and therefore missing value imputation is an essential preprocessing step in data mining and machine learning. The imputation methods of kNNI [8] in recent years have been widely applied because of its easy operating. The result and hence the accuracy of kNNI, however, are dependent of the parameter k, which means that each k should be tried in order to get an optimal one. Moreover, the result of kNNI is a biased estimation since the neighbors of the targeted point with missing value may lie unevenly around the point. Two variations [9, 10] of kNNI were proposed to overcome the defects of previous versions and they both perform satisfactory. Only after the preprocessing step, are the data sets of food safety inspections ready for further analyzing and mining.

The rest of the paper is organized as follows. Section 2 introduces research work related to this paper, including those on missing data imputation, Logistic regression, neural network, and extreme learning machine. In Sect. 3, the ELM framework is described in detail that is employed to mine the patterns hidden in the big data of food safety inspections. Section 4 presents the experiments on real data sets and the corresponding results. And Sect. 5 concludes the paper.

2 Related Work

A variety of methods and technologies have been studied, tested and/or implemented to analyze and utilize the data collected from food safety inspections, and many exciting results and conclusions have been obtained.

Khosa and Pasero [6, 7] used an artificial neural network (ANN) as a classifier to predict at an early stage of processing or manufacturing whether important food ingredients, pine and pistachio nuts, are healthy. X-ray images of the nuts were used, and texture features were extracted from the images. In that work, the texture features and the sample labels were used as the training data, and the texture features were independently used as the basis for making predictions and classifications. As a result, the ANN classifier achieved false negative rates of 0% and 6.8% for the pine nuts and pistachio nuts, respectively. The results imply that food quality has good predictability and good describability, at least in certain cases.

Reference [1] focused on a safety risk assessment of dairy products for a single corporation, also in the background of big data. That work used a classic classifier, the support vector machine (SVM). However, instead of using a serial algorithm for the SVM, a parallel cascade SVM was implemented on the platform of Apache Hadoop [11], which is an open-source distributed computing framework that is typically used to process big data by distributing the data in a large-scale cluster platform. The results from [1] demonstrate that when the number of cluster nodes increases steadily, the saved run time decreases steadily compared with the runtime for a single node. The SVM has been a successful classifier in many cases and in many areas due to its good classification accuracy, generalizability and stability. Despite this success, SVM does not perform satisfactorily when the positive and negative samples have more detailed relationships.

Statistical methods are most frequently used to analyze the data obtained from food safety inspections, with [3] being a typical study. Based on the food sampling results of the city of Shenzhen, China, that study first investigated the annual and inter-annual changing tendency of 11 food categories and analyzed the data using the t-test. Then, a logistic regression model was constructed, and the quantitative relationships between food quality and four attributes (namely, food origin, inspection season, sales site, and food packaging) were established. Instead of the result category (qualified/unqualified), the concept of “exceeded percentage” was used to measure the degree of unqualified food. Logistic regression is a powerful classifier that can be applied to both continuous and discrete variables. Although that work is a good application of logistic regression to predict which food products are most likely to be unqualified, the data for both training and testing are simulated data sets, not real data sets, which indicates that the work remains unsatisfactory.

Logistic regression, like many other regression methods, is essentially linear regression; it is aided by some nonlinear transformations, and it can capture the nonlinear relationships between the dependent variable and causative (independent) variables. The ANNs in [1, 7] used a considerable number of nonlinear transformations to capture more detailed relationships. However, as demonstrated earlier, the learning speed of feed-forward ANNs is considerably slower than that of regression learning algorithms, which take the least squares method (LSM) as the core technique. Considering both the speed advantage of the LSM and the nonlinearity advantage of the ANN, Huang et al. proposed the ELM with a single layer of hidden nodes in their two pioneering works [12, 13]. Compared with its predecessor learning techniques, the ELM improves the training speed by hundreds of times by randomly setting the weights between the input nodes and hidden nodes and by computing the weights between the hidden nodes and output nodes using the LSM. Other researchers have supported their work, particularly the random assignment of weights, by mathematical proofs, such as in [14], which provides a geometric perspective.

After these pioneering studies, a variety of variations and improvements in the ELM were presented. Reference [15] proposed an inverse-free ELM that further improved the computational speed of the training process, as computing the inverse of a square matrix is the most time-consuming part of the LSM. Accounting for the architecture of the sub-network nodes, Y. Yang and Q. M. Jonathan Wu designed a variation of the ELM, ML-ELM, that exhibits competitive accuracy and speed compared with other conventional feature learning methods with sub-network nodes [16, 17].

The ELM has a notable defect, namely, that the number of hidden nodes must be manually assigned or assigned by other state-of-the-art methods. In fact, the optimal number of hidden nodes plays a decisive role in the ELM, as an insufficient number of hidden nodes could lead to underfitting, whereas an excessive number of hidden nodes could cause overfitting. Based on this observation, [18] presented an adaptive and automatic selection algorithm that can obtain a suitable or even an optimal number of hidden nodes for each learning case. This method can markedly reduce the degree of artificial participation and hence reduce the burden of human operators.

In addition, there are many applications of the ELM to different types of domains. The ELM was applied to predict soil moisture in an apple orchard [19], taking both the weather factors and the time series of the soil moisture as inputs. Compared to the conventional method of the SVM, the ELM exhibits a higher prediction accuracy over a larger forecast range with a higher speed. Reference [20] proposed a new classification algorithm for food classification based on both spectroscopy and the ELM, and the experimental results indicated that the ELM is typically more precise and robust than its competitors, including k-nearest neighbor, partial least-squares discriminant analysis, back propagation ANNs, and least-squares support SVMs.

3 ELM Approach Specification

In this section, we will present in detail the ELM-based classifier for predicting whether a sample food to be inspected is qualified or not. Firstly, in Sub-Sect. 3.1, the cause variables are selected according to whether it is likely to affect the food quality. And then data preprocessing techniques are presented in Sub-Sect. 3.2. After that in Sub-Sect. 3.3, the main framework of the ELM method is described based on the discussion of the former two sub-sections.

3.1 Selecting Relevant Factors

According to the food safety inspection data, the result variable that we are most interested in is quite simple: it has binary values for whether the food is qualified or not. However, the causative variables are more complex and involve many factors. We eliminate the factors that are not related to the ability to predict the food quality, such as the sampling number and name of the manufacturer. After the elimination operation, 9 causative variables are retained, as listed in Table 1.

Table 1. Causative (dependent) variables selected for the model[a].

Selected factor/variable	Meaning of the variable
Food category	There are 6 categories[a] in the inspection data
Manufacturing date	The date when the product was manufactured
Manufacturing site	The place where the product was manufactured
Inspection date	The date when the product was sampled and inspected
Inspection site	The place where the product was sampled and inspected

[a] The 6 categories are T0, dairy products; T1, aquatic products; T2, infant formula; T3, meat products; T4, liquor; T5: edible oil.

3.2 Preprocessing Technique

When all the factor variables are determined, the data are processed to eliminate the noise data. The missing values are completed with the imputation techniques proposed in [10] while considering the representative point and the densities of the points in each quadrant compared to the targeted point for the missing values.

New causative variables can be generated based on the variables listed in Table 1. For example, from the manufacturing date and inspection date, a new variable, the elapsed days, can be generated; this variable refers to the time span between the manufacturing time and inspection time. Another example is “Whether in the same province”, which is generated from the two variables “Manufacturing site” and “Inspection site”; this variable indicates whether the two sites are in the same province or not.

To date, certain variables have not yet been useful for the models of either logistic regression or ELM, as they are category variables, not numeric variables. For example, the manufacturing date appears to be a numeric variable, but in fact, it is more likely to be a categorical variable because it implies the seasonal information of the manufacturing time. Thus, we transform the variable “Manufacturing date” into four variables, namely, “Spring”, “Summer”, “Autumn” and “Winter”, with each having binary values of true or false. The four new variables are called dummy variables, and they are generated in the same manner as described in [21].

After the preprocessing stage, the data are ready for training, testing, and predicting using both the ELM method and its competitors.

3.3 Framework of ELM Method

In this paper, we also use one-hidden-layered nodes, as shown in [18–20]. The structure of the network is illustrated in Fig. 1.

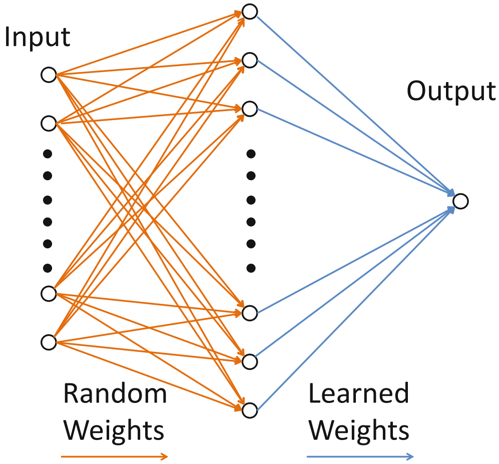

Fig. 1. Structure of the ELM. There is only one hidden layer and only one output node.

Each input node in Fig. 1 represents a causative variable. The causative variables selected and the variables generated by them are each represented by an input node. There are considerably more hidden nodes than input nodes; but it is not a fixed number. Instead, it varies according to the number of inputs and the structure of the training data based on the adaptive strategy given in [18]. As described in [13], the weights between the input nodes and the hidden nodes are set to random values (see Fig. 1), which implies that the output value of one hidden nodes may be proportional (or nearly proportional) to that of another hidden node. Therefore, at least one of them is useless to capture the relationship between the input and the output. The optimization algorithm in [18] first generates a large number of hidden nodes and then selects the nodes one by one, making the newly selected one least linear-correlated to the previously selected node. By taking into account the input data and the output data an optional number of hidden nodes can be obtained. We employ this optimization method to form the structure of ELM. The single output node represents the result of a record, which means whether a food sample is qualified.

The weights between the input nodes and hidden nodes are assigned randomly as described in [12, 13], and they are all set to be in the range $[-1, 1]$. However, all weights between the hidden nodes and output node are obtained by learning and computing based on the training data.

Suppose that the number of input nodes is n and the number of hidden nodes is h, the input of the jth hidden node is calculated as follows:

$$G_j = \beta \sum_{i=1}^{n} W_{i,j} \tag{1}$$

where β is a parameter that will be discussed later, $W_{i,j}$ is the weight between the ith input node and the jth hidden node.

Each hidden node processes its input by the following equation and then outputs the following:

$$H_i = \frac{2}{1 + e^{-G_i}} - 1 \tag{2}$$

H_i will always lie between −1 and 1, which makes its value distribution approximately symmetric about the y-axis. Equation (2) is often called the activation function, which is a highly nonlinear function. All activation functions in the hidden nodes together make the system capable of approximating nearly any nonlinear relationship between the input nodes and output node.

Parameter β in Eq. (1) will affect the effectiveness of the system. If β is not sufficiently large, the relationship between the input and output will degenerate to a linear relationship. However, if β is excessively large, all the inputs of the hidden nodes will be transformed by the activation function into either −1 or 1. Thus, we set parameter β in this paper according to the following empirical formula:

$$\beta = \frac{10}{n} \tag{3}$$

where n is the number of input nodes.

In the step of the LSM for calculating the weights between the hidden nodes and output node, the inverse of a square matrix must be computed, which will not be executable if the matrix is irreversible. If this problem occurs, we will change the square matrix slightly and make it reversible by using the method suggested in [22], which overcomes a significant shortcoming of the ELM.

The overall framework of the ELM approach is shown in Fig. 2.

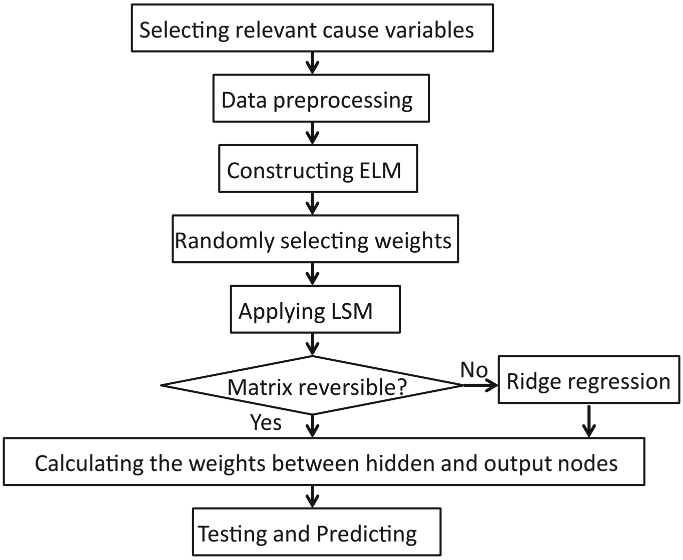
Selecting relevant cause variables
Data preprocessing
Constructing ELM
Randomly selecting weights
Applying LSM
Matrix reversible?
No
Ridge regression
Yes
Calculating the weights between hidden and output nodes
Testing and Predicting

Fig. 2. Overall framework of the ELM approach.

4 Experiments and Results

The data sets used are publicly available from the State Food and Drug Administration of China. For these samples, the manufacturing date ranges from November 26, 2014 to September 1, 2016, whereas the inspection date ranges from October 29, 2015 to September 9, 2016.

The two methods applied to the data sets are logistic regression presented in [3] and the ELM. The variable selection and data preprocessing are same for the two methods. For each category of food, all data are partitioned into training data and testing data. The training set and testing set are identical for the two methods. The testing results are listed in Table 2.

Table 2. Comparison of experimental results.

Category	Number of testing cases	Number of correct cases for LR	Number of correct cases for ELM	Accuracy of LR (%)	Accuracy of ELM (%)
T0	1376	1196	1212	86.9	88.1
T1	873	777	791	89.0	90.6
T2	1403	1260	1310	89.8	93.4
T3	4058	3633	3656	89.5	90.1
T4	2730	2615	2618	95.8	95.9
T5	2063	1785	1798	86.5	87.2

The data listed in Table 2 are shown in Fig. 3.

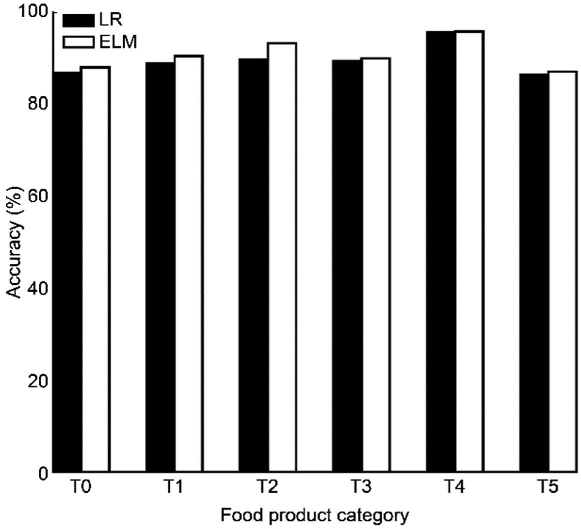

Fig. 3. Comparison of the experimental results from logistic regression and the ELM. The black bar represents the accuracy percentage of the logistic regression, whereas the white bar represents that of the ELM.

Figure 3 shows that the ELM has better accuracy than logistic regression for all food categories, although they perform nearly the same for certain categories, such as T4.

5 Conclusions

ELM is employed in this paper to describe the big data collected from the food safety inspections of China in recent two years. The trained model is used to predict the food quality and it performs better than Logistic regression that was implemented and tested on simulated data sets. Results from a series of experiments show that ELM is better in accuracy than Logistic regression for each of the 6 food categories. And both of the methods run very fast because they all take the advantage of optimized calculating steps. The success of the ELM owes much to the large number of hidden nodes and the nonlinear activation functions in them are able to capture the nonlinear components in the relationship between the inputs and the outputs.

With the ELM model and the according prediction system, food samples can be taken no longer randomly; on the contrary, food products could be filtered by the prediction system and only those with least qualification probabilities will be selected for sampling test. Therefore, aided by the ELM prediction and classification system, more effective inspection plans can be made which mean less labor input and more food safety problems exposed.

Acknowledgments. This work was supported by the National Science & Technology Pillar Program of China (2015BAK36B01), the National Natural Science Foundation of China (61402436), Shanghai Engineering Research Center of Product Traceability.

References

1. Ma, Y., Hou, Y., Liu, Y., Xue, Y.: Research of food safety risk assessment methods based on big data. In: 2nd IEEE International Conference on Big Data Analytics, Beijing, China, pp. 1–5 (2017)
2. Antunovic, B., Mancuso, A., Capak, K., Poljak, V., Florijančić, T.: Background to the preparation of the Croatian food safety strategy. Food Control **19**(11), 1017–1022 (2008)
3. He, L., Wang, Z., et al.: The method of food safety sampling inspection based on dynamic weight. Math. Model. Appl. **2**(3–4), 4–12 (2013)
4. Li, F., Lv, Y., Zhu, Q., Lin, X.: Research of food safety event detection based on multiple data sources. In: International Conference on Cloud Computing and Big Data, Shanghai, pp. 213–216 (2015)
5. Sonka, S.: Big data and the ag sector: more than lots of numbers. Int. Food Agribus. Manag. Rev. **17**(1), 1–20 (2014)
6. Khosa, I., Pasero, E.: Defect detection in food ingredients using multilayer perceptron neural network. In: 2014 World Symposium on Computer Applications & Research, Sousse, pp. 1–5 (2014)
7. Khosa, I., Pasero, E.: Artificial neural network classifier for quality inspection of nuts. In: International Conference on Robotics and Emerging Allied Technologies in Engineering, Islamabad, pp. 103–108 (2014)
8. Kung, Y.-H., Lin, P.-S., Kao, C.-H.: An optimal *k*-nearest neighbor for density estimation. Statist. Probab. Lett. **82**(10), 1786–1791 (2012)
9. Zhang, S.: Shell-neighbor method and its application in missing data imputation. J. Appl. Intell. **35**, 123–133 (2011)
10. Wang, J., Zhang, Z., Chen, Z., Yuan, Q.: Imputation missing values with distance- and density-weighted and quadrant-based nearest neighbors. J. Comput. Inform. Syst. [1] **11**(18), 6605–6613 (2015)
11. Singh, D., Reddy, C.K.: A survey on platforms for big data analytics. J Big Data **2**(1), 8 (2015)
12. Huang, G.B., Chen, L., Siew, C.K.: Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Trans. Neural Netw. **17**(4), 879–892 (2006)
13. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1–3), 489–501 (2006)
14. Cervellera, C., Maccio, D.: Low-discrepancy points for deterministic assignment of hidden weights in extreme learning machines. IEEE Trans. Neural Netw. Learn. Syst. **27**(4), 891–896 (2016)
15. Li, S., You, Z.H., Guo, H., Luo, X., Zhao, Z.Q.: Inverse-free extreme learning machine with optimal information updating. IEEE Trans. Cybern. **46**(5), 1229–1241 (2016)
16. Yang, Y., Wu, Q.M.: Extreme learning machine with subnetwork hidden nodes for regression and classification. IEEE Trans. Cybern. **46**(12), 2885–2898 (2016)
17. Yang, Y., Wu, Q.M.J.: Multilayer extreme learning machine with subnetwork nodes for representation learning. IEEE Trans. Cybern. **46**(11), 2570–2583 (2016)
18. Mesquita, D.P.P., Gomes, J.P.P., et al.: Pruning extreme learning machines using the successive projections algorithm. IEEE Lat. Am. Trans. **13**(12), 3974–3979 (2015)

19. Liu, Y., Mei, L., Ooi, S.K.: Prediction of soil moisture based on extreme learning machine for an apple orchard. In: IEEE 3rd International Conference on Cloud Computing and Intelligence Systems, Shenzhen, pp. 400–404 (2014)
20. Zheng, W., Fu, X., Ying, Y.: Spectroscopy-based food classification with extreme learning machine. Chemom. Intell. Lab. Syst. **139**, 42–47 (2014)
21. Changpetch, P., Lin, D.K.J.: Model selection for poisson regression via association rules analysis. Int. J. Statist. Probab. **4**(2), 1–9 (2015)
22. Li, G., Niu, P.: An enhanced extreme learning machine based on ridge regression for regression. Neural Comput. Appl. **22**(3), 803–810 (2013)