Data Driven Methods for Finding Pattern Anomalies in Food Safety

S. Anantha Krishna, Amal Soman, and Manjusha Nair(✉)

Department of Computer Science and Applications, Amrita Vishwa Vidyapeetham, Amritapuri, Kollam, India
manjushanair@am.amrita.edu

Abstract. The indigenous part of all living organisms in the world is food. As the world population increases, the production and consumption of food also increases. Since the population progresses in a rapid manner, the productivity of the food materials may not be sufficient for feeding all the people in the world. There rises the cause of food adulteration and food fraud. Adulteration is the process of adding a foreign substance to the food material which affects the natural quality of the food. As the amount of adulterants increases, the toxicity also increases. Machine learning techniques has been used previously to automate the prediction of food adulteration under normal scenarios. In this paper, we use different machine learning technique for finding food adulteration from milk data sets. This paper surveys the different concepts used in automating the detection of food adulteration and discusses the experimental results obtained by applying machine learning algorithms like Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Artificial Neural networks (ANN), Linear Regression, and Ensemble methods. The accuracy of the models ranged from 79% to 89%. Ensemble method outperformed other algorithms with an accuracy of 89% and Linear Regression showed least accuracy of 79%. Artificial Neural networks showed an accuracy of almost 87%. SVM and Naïve Bayes showed accuracy 84% and 80% respectively.

Keywords: Machine learning · Food safety · Food fraud · Adulteration · Neural networks

1 Introduction

There has been a dramatic change in the food production and consumption worldwide. According to the High-level Expert forum, the world in the 21st century is not very capable of mapping the food production to food consumption, as the labor force and other natural resources are not as adequate. The demand for cereals is expected to grow by over 3 billion tons in the year 2050 in which the current need is 2.1 billion tons [1]. From these, it is clear that the production cannot meet the requirement. Hence the use of adulterants is raised significantly over the past few years. Adulteration in food mostly happens due to the unhygienic or unhealthy treatment of food items during production, storage or delivery, which is done mainly for financial benefits. Unhealthy

S. M. Thampi et al. (Eds.): SoMMA 2020, CCIS 1366, pp. 122–130, 2021.
https://doi.org/10.1007/978-981-16-0419-5_10

food consumption results in an unhealthy civilization, which is a main threat in the current century for the world to face.

There has been a number of studies initiated on food safety through improved analytic and machine learning tools in detection of adulteration in food materials. There were several machine learning methods applied in previous studies, including Artificial Neural Networks (ANN), Time series analysis, SVM, Fuzzy logic etc. [2]. The main application of Artificial Neural Network was on prediction and fraud detection. It can be also seen that 90% of the tools were used to build models based on yield, concentration and quantity of food safety parameters. The current study uses similar machine learning tools to predict the presence of foreign substance in milk [3].

Machine learning algorithms like PCA, Naıve Bayes, K-nearest neighbor, Linear Discriminate Analysis, Decision Tree, ANN and Support vector machines were used previously for quality control of olive oil [4]. Similar algorithms were used previously to predict drug likeliness in compounds [5]. In the studies conducted in [4], there were two methodologies involved in the process and Artificial Neural Network (ANN) had the highest accuracy of 65.83% for the test data while Naıve Bayes had the lowest accuracy of 45.83% among the algorithms, using the first method. The test results of the second method told a different story as Naıve Bayes produced the highest accuracy with 70.83% and Linear Discriminate Analysis (LDA) had the lowest accuracy with 56.67%. When the prediction of food fraud was done using the Bayesian Network (BN) modelling approach [2], accuracy of prediction was increased to 91%. A food fraud early warning system using European Media Monitor (EMM) was used in the study along with a media monitoring system called MedISys. A similar text classification method was used to extract data from Wikipedia texts and articles in another study [6]. They used natural language processing and recurrent neural networks to establish an automated system that helped to detect the adulteration in food.

Deep learning and ensemble methods were used previously to find the adulteration in milk samples [3]. In adulteration, the milk industry tops the market, making milk and milk products as one of the highly adulterated foods. The study used Fourier Transformed Infrared Spectroscopy (FTIR) for accessing the milk quality by producing the spectral data on the samples. The compositional information is fed into the machine learning algorithms-neural networks and decision trees. The goal of the project was to perform a binary classification of raw milk and adulterated milk, and also a multi class classification of predicting the ingredient present in the adulterated sample. Using ensemble machine learning methods and Convolutional Neural Networks (CNN), the system showed an accuracy of 98.67% which was much higher than the classical learning methods in the dairy industry. In the current study, we are using six machine learning algorithms, namely, Naive Bayes, Support Vector Machines (SVM), Artificial Neural Networks (ANN), K-Nearest Neighbor (KNN), Linear Regression, and Ensemble methods to predict the presence of adulterants from milk data set.

2 Machine Learning Algorithms Used

Six different machine learning algorithms, as detailed below (Sect. 2.1 to 2.6), are used in this study. WEKA tool is used to get the statistical output of the data and helps in inspecting the data. Initially it is dome in WEKA and later the algorithms are implemented using Python language.

2.1 Naive Bayes

The Bayes classifiers are used widely in the food sectors, food supply chain, food fraud and in health sectors [7–10]. The Naive Bayes classifier follows probabilistic machine learning approaches using Bayes theorem, assuming strong, or naive, independence between the features in the feature vector. The probabilities are estimated as,

$$P(Y|X_1, X_2, \ldots X_n) = \frac{P(X_1|Y)P(X_2|Y)\ldots P(X_n|Y)P(Y)}{P(X_1).P(X_2)\ldots P(X_n)} \quad (1)$$

$P(Y|X_i)$: Posterior Probability
$P(Y)$: Prior probability of the class variable
$P(X_i|Y)$: Likelihood
$P(X_i)$: Predictor Prior Probability

This type of classifiers are proved to be successful in many real world applications with lesser training data set, irrespective of the oversimplification and assumptions of the variables. It is also defined as a classification technique based on Bayes Theorem with the assumption of the independence among the predictors. The Bayes classifiers assume the presence of a specific feature in a class is unrelated to the presence of any other feature. This model is actually easy to build and is used mostly in larger dataset.

2.2 Support Vector Machines

The Support Vector Machines (SVMs) are supervised, non-probabilistic, binary learning models with related learning algorithms that analyses data for classification purpose mostly [11, 12]. It is a discriminative classifier formally defined by a separating hyperplane. There are different hyperplanes that can classify the data, choosing the best hyperplane is the one that can distinguish the two classes with a wide separation. These created hyperplanes are used for the classification and regression. Linear and non-linear classifiers are used for transforming the feature space. The SVMs are used to solve some real-world problems like text categorization, classification of images and it is being applied in the biological and other sciences. SVMs are also used in rice yield predictions in India [13]. The machine learning technique is now widely used for the prediction of crop yield under different climate scenarios. SVM uses classification algorithm for two-group classification problems. After giving the SVM model sets of the labelled training data for reach category, they will be able to categorize new text.

2.3 Artificial Neural Network

Artificial Neural Networks are an advanced machine learning technique that works on the principle of neurons of the brain [14]. A neuron gets its input from other neurons through the synapses and dendrites and the processed information is transmitted from the soma to the next level of neurons through axons [15]. Analogues to that, there are three different layers in a simple neural network: the input layer, hidden layers and the output layer. The perceptron in the neural network receives multiple input values, say feature values of the dataset. The weighted sum of all inputs after processed by an activation function, is fed to the output layer of the neural network. Given a unit j in a hidden or output layer, the net input, I_j, to unit θj is

$$I_j = w_{ij}O_j + \theta_j \tag{2}$$

where, w_{ij} = weight of connection from unit i in the previous layer to unit j in the hidden layer, O_j = i$^{\text{th}}$ output from previous layer and θ_j = bias of unit.

ANN is the foundation of artificial intelligence which solves the problems which prove to be difficult for human or statistical standards. It also has the self-learning capabilities which allows them to get better result as more data becomes available. An ANN has a number of artificial neurons called the processing unit, which are interconnected by the nodes and these processing units are made up of input and output units.

2.4 K-Nearest Neighbor

K-Nearest neighbor algorithm is a powerful machine learning technique that is used in pattern recognition for classification and regression. It is a non-parametric method [16]. In both cases of classification and regression the input set contains K closest training examples which is placed in the feature space with the basic assumption that similar data points in the feature space are close to each other. The value of K is initialized to the chosen number of neighbors and the distance between the test data and each row of training data, is calculated. From the ordered and sorted collection of K such distances, the most frequent class is estimated as the resulting class in the algorithm. This supervised, non-parametric machine learning technique, irrespective of its simplicity, finds wide applications in the field of machine learning. Even though KNN can be used for both classification and regression predictive problems, it is more widely used in the classification problem in the industry. This algorithm fairs around all parameters of consideration and it is commonly used for its ease of interpretation and low calculation time.

2.5 Linear Regression

Regression is a term used for describing the models that examines the variable's interrelationship. This model learns one-to-one relationships among dependent variables and one or more independent variables. If one independent variable is present, it is called simple linear regression and if more than one independent variables are present, it is referred to as multiple linear regression [17]. The main purpose of the linear regression is to predict the relationship between input variables (X) and output variables (Y).

The simple linear regression model is represented as:

$$Y = \beta_0 + \beta_1 X + \in \tag{3}$$

where β values are the bias coefficients and $\in$ represents the error term in the model. In a linear regression model for classification, the values of the coefficients are estimated from data. Then the learned model is used for predictions.

2.6 Ensemble Learning

We have discussed several machine learning methods that are used to build and predict the models. Apart from that learning model's discussed above, Ensemble method is a technique which combines multiple base models which can produce an optimal single predictive model [18]. The advantage of the ensemble learning mechanism is that the resultant classifier produces more accurate predictions than any of the single classifier used in the ensemble learning. Bagging and Bboosting methods are the two latest and most commonly used ensemble methods [19]. Bagging (Bootstrap Aggregating) works on the principle of attaining a number of base learners from bootstrap sample and trains that base learners. The bootstrap sample is generated from the training data set by sub sampling them with replacement. Boosting deals with the utilization of set of algorithms with each of its weighted average for making the weak learners stronger. Unlike the bagging mechanism, boosting builds the model with further combination of features: like after running a model it finds out what feature needs to be run next for better classification.

2.7 Dataset

This study used machine learning techniques on spectral data of milk samples, downloaded from public repository [3]. The data set contained milk samples of 1000 instances with information on different properties of milk such as whether it is raw milk, whether it contains lactose, milk urea, fat, protein, free fatty acid, solid content in the milk, somatic cell counting, casein content; freezing point of the milk, and Qvalue which shows the quality of milk. The data set is split into two parts in which the training set contained 66% of the original data and the rest is used for testing the algorithms. The dataset was carefully examined and arranged in a manner that it contains 14 parameters and these are the key factors in classifying the milk sample as raw or contaminated. The algorithms we prescribed correlates these values form the dataset and perform the analysis through training and testing the data. After the dataset is trained with parameters it then tests the data and produces the final result which shows the instances having contaminants.

3 Results and Discussion

3.1 Result Evaluation

The possible classifications of the instances were converted into 4 categories:

TP (True Positive): Both condition and prediction are true.
TN (True Negative): Both condition and prediction are false.
FP (False Positive): Here the condition is false but the prediction is true.
FN (False Negative): Here the condition is true but prediction is false.

These values are represented as a confusion matrix to compare the actual observed value to the predicted value. The true positive and true negative fields, placed on the diagonal of the table showed the correct predictions that the algorithm had made. Accuracy is the most common metric used in the result and it is formulated as:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN}) \tag{4}$$

Sensitivity and specificity are the other two metrics used here. Sensitivity is also called recall and specificity is also known as precision. Sensitivity deals with the percentage of correctly classified instances and specificity is the percentage of incorrectly classified instances. Sensitivity measures the proportion of actual positives correctly classified as positive. Specificity measures the exactness of the models and measures the proportion of actual negatives correctly classified as negatives.

$$\text{Sensitivity or Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{5}$$

$$\text{Specificity or Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{6}$$

3.2 Accuracy Estimates of the Algorithms

The accuracy of the models ranged from 79% to 89% (Fig. 1). Ensemble method outperformed other algorithms with an accuracy of 89% and Linear Regression with least accuracy of 79%. Artificial Neural networks showed an accuracy of almost 87%. SVM and Naïve Bayes showed accuracy 84% and 80% respectively. Thus it has been concluded that ensemble methods performed better than other classification methods used.

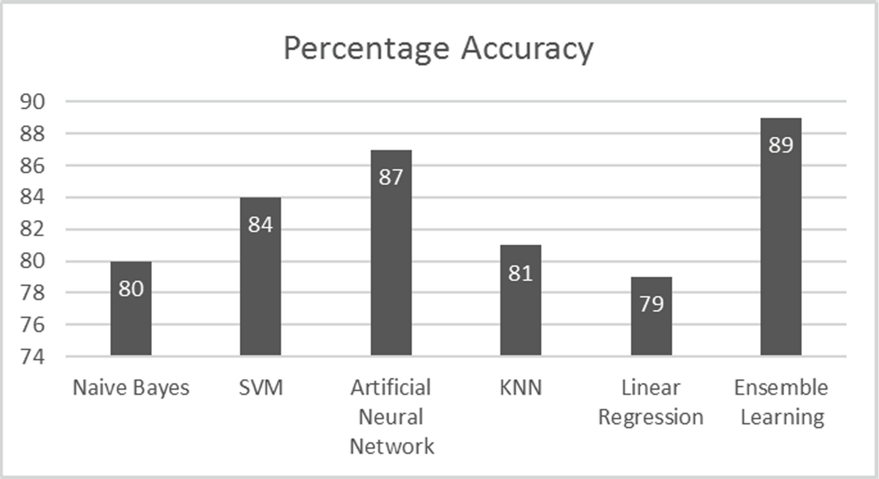

Fig. 1. Accuracy of prediction of different algorithms. Among all methods used, the Ensemble method showed highest accuracy of prediction (89%) and Linear regression showed the lowest (79%).

3.3 Sensitivity and Specificity

Since accuracy of a machine learning algorithm shows the number of correctly predicted instances from all the predictions made, and is not the only measure to evaluate the performance of algorithms, two other measures like sensitivity (recall) and specificity (precision) were also estimated from the data set and plotted (Fig. 2). Sensitivity was calculated to be maximum for ANN (84.14%) and minimum for KNN (81.37%) Specificity was calculated to the maximum for Naïve Bayes model (88.57%) and minimum for linear regression (83.81%).

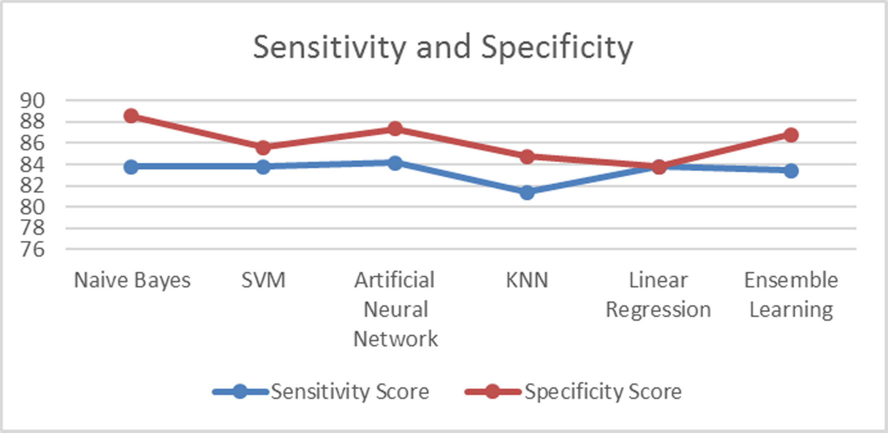

Fig. 2. Sensitivity and specificity scores

4 Conclusion

In this paper, we used different machine learning algorithms to classify the adulterated milk data set. The Classification accuracies of the algorithms ranged from 79% to 89%, the highest being reported in the Ensemble method. The current Ensemble methods can be further improved using Random Forest and Gradient boosting regressing trees. Only binary classification is used in the models and an extended study can be used to predict the adulterant present in the milk sample [20]. Also less attention was given in this study to capture the actual data set from a region of interest or subject of interest. Future studies should also focus on how automated techniques can be used for optimal utilization of resources, in food industry. Safety can be monitored more rigorously using automated techniques such as machine learning combined with data analytics in cloud computing and block chain technologies.

References

1. Executive Summary: How to Feed the World in 2050, pp. 1–35. https://www.fao.org/
2. Bouzembrak, Y., Marvin, H.: Development of early warning systems to detect, predict and assess food fraud (2018)
3. Neto, H.A., Tavares, W.L.F., Ribeiro, D.C.S.Z., Alves, R.C.O., Leorges, M., Campos, S.V. A.: On the utilization of deep and ensemble learning to detect milk adulteration. BioData Min. **12**, 1–13 (2019)
4. Ordukaya, E., Karlik, B.: Quality control of olive oils using machine learning and electronic nose. J. Food Qual. **2017**, 1–7 (2017)
5. Ani, R., Manohar, R., Anil, G., Deepa, O.S.: Virtual screening of drug likeness using tree based ensemble classifier. Biomed. Pharmacol. J. **11**(3), 1513–1519 (2018)
6. Gou, Y.: Food adulteration detection using neural network (2016)
7. Smith, J.Q., Barons, M.J., Zhong, X.: Bayesian Networks for Food Security What and Why Bayesian Networks? Influence Diagram, no. 2009 (2013)
8. Kubade, H.M.: The overview of Bayes classification methods. Int. J. Trend Sci. Res. Dev. (IJTSRD) **2**, 2801–2802 (2018)
9. Stein, A.: Bayesian networks and food security – an introduction, no. 1996, pp. 107–116 (2001)
10. Jha, K., Doshi, A., Patel, P., Shah, M.: A comprehensive review on automation in agriculture using artificial intelligence. Artif. Intell. Agric. **2**, 1–2 (2019)
11. Kavitha, K.R., Syamili Rajendran, G., Varsha, J.: A correlation based SVM-recursive multiple feature elimination classifier for breast cancer disease using microarray. In: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (2016)
12. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2008). p. 134
13. Armstrong, J.: Rice crop yield prediction in India using support vector machines. In: 2016 13th International Joint Conference on Computer Science and Software Engineering, no. 2010, pp. 1–5 (2016)
14. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U.S.A. **79**(8), 2554–2558 (1982)

15. Nair, M., Surya, S., Kumar, R.S., Nair, B., Diwakar, S.: Efficient simulations of spiking neurons on parallel and distributed platforms: towards large-scale modeling in computational neuroscience. In: 2015 IEEE Recent Advances in Intelligent Computational Systems, RAICS (2015)
16. Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. **46**(3), 175–185 (1992)
17. Yan, X.: Linear Regression Analysis: Theory and Computing, pp. 1–2. World Scientific, Singapore (2009)
18. Dietterich, T.G.: Ensemble methods in machine learning (1990)
19. Opitz, D., Maclin, R.: Popular ensemble methods: an empirical study. J. Artif. Intell. Res. **11**, 169–198 (1999)
20. Menon, R., Aswathi, P.: Document classification with hierarchically structured dictionaries . In: Berretti, S., Thampi, S.M., Dasgupta, S. (eds.) Intelligent Systems Technologies and Applications. AISC, vol. 385, pp. 387–397. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-23258-4_34