Asian Journal of Water, Environment and Pollution, Vol. 19, No. 1 (2022), pp. 41-46.
DOI 10.3233/AJW220007

Evaluating the Efficiency of the Classifier Method When Analysing the Sales Data of Agricultural Products

Yuxin Wang, Svetlana Avdeenko[1]* and Yuriy Shmidt[2]

College of Innovation and Management, Suan Sunandha Rajabhat University, Bangkok, Thailand
[1]Department of Agriculture and Technology of Storage of Crop Production, Don State Agrarian University, Persianovsky, Russian Federation
[2]Department of Business Informatics and Economic-Mathematical Methods, Far Eastern Federal University, Vladivostok, Russian Federation
✉ avdeenkosvet9@rambler.ru

Received August 2, 2021; revised and accepted August 24, 2021

Abstract: Data classification as a method of input analysis is of the greatest interest and necessity for proper distribution and quality evaluation of agricultural products. The use of classification methods allows predicting whether a selected sample from the data set will fit into a particular class or group, which is necessary for the process of sorting products. This study presents the results of a comparative analysis of high-performance classifiers for assessing the effectiveness of further use in the sorting of agricultural products. The study was carried out utilising the classifiers of k-nearest neighbours, naive Bayesian classifiers, and artificial neural networks for data analysis during apple fruit sorting. It has been established that the greatest accuracy 99% of the results is demonstrated by the classifiers of k-nearest neighbours, but, at the same time, they show the lowest calculation speed (0.47 s). The best performance at any data size (65-100%) is shown by the neural network. A comprehensive review of the features and restrictions of the studied classification algorithms, as well as their applications in various areas of agriculture, has been performed.

Key words: Artificial neural network, classifier of k-nearest neighbours, intelligent data analysis, naive Bayesian classifier.

Introduction

The quality of agricultural products traditionally plays a fundamental role in almost all quality assessments in the food industry. The traditional method for evaluating the quality of agricultural products based on the operator's sensory analysis during manual sorting, or the division into several assessment stages for each criterion, is tedious and expensive. Besides, it is easily influenced by physiological factors that cause subjective and conflicting assessment results (Ireri et al., 2019; Safonov et al., 2020). The use of innovative computer technologies to evaluate the quality of agricultural products allowed establishing automatic sorting systems (Fan et al., 2020), which are a combination of software and hardware. Information in the form of images and signals from devices may contain a different set of data characterising the product (Prabhakar and Mohana, 2018). The sorting of products using machine vision can be randomly divided into four stages: image acquisition, segmentation, interpretation, and fruit classification. Placing fruits on rollers and capturing their images via a camera while they rotate is a wide and most common method (Narendra et al., 2020). At the same time, the image of the entire object is rolled out as a single scene on a computer monitor, depending on its location.

*Corresponding Author

Following the complexity of a data set, the information processing should correspond to the set task and ways of its decision, which are explicit in the program. The sorting process can be based on various product criteria such as size, weight, degree of maturity, colour, density, surface roughness, etc., which affect the demand and pricing of the product, as well as its purpose (MacDonald, 2020). Increased demand for e-commerce, excellent prospects, competitive prices for agricultural products, and internationally comparable standards facilitate good trading opportunities in the agricultural sector. It creates a need to develop an intelligent e-commerce system that will serve as a means by which each user can meet the requirements and take advantage of agro-industrial buying and selling, as well as other business opportunities.

Intelligent data analysis consists of applying machine learning methods to detect new models and relationships in a large data set (Hemmatian and Sohrabi, 2019). In general, the tasks of intelligent data analysis can be divided into two categories, namely, methods of descriptive and predictive classification (Findawati et al., 2019). Data classification is a process of arranging information in categories relative to the similarity of data objects for one group and variation for different groups (Rajesh and Karthikeyan, 2017). Classification methods can handle large volumes of data and predict categorical class labels, as well as classify data based on a model composed using the tutorial set and associated class labels to be used for arranging newly available test data (Safonov et al., 2019).

This research aimed to assess the effectiveness of the various classification methods most commonly used in intelligent data analysis for the distribution of agricultural products. A comparative study of the algorithms was carried out between the classifiers of the Bayesian network, k-nearest neighbours, and artificial neural networks. Evaluation of the accuracy for each classification algorithm in terms of efficiency and time complexity, as well as analysis of advantages and disadvantages of one method over the other, will help to identify the restrictions of each method. Comparative analysis of the efficiency of the method is given on the example of fruit sorting.

Materials and Methods of Research

Data Sources

As sources of data to be classified, the databases of physical characteristics (weight, shape, linear parameters), sensory data (surface colour, presence of defects, etc.), for fruit (apples) from publicly available reporting data of agricultural companies (fruit-tree-nuts data) are used. For the training set, the data were entered according to the Fresh Produce Traceability Guidelines (FPT) used for food labelling. Classification is done in three categories (highest, medium, and lowest). Each element contains a data set of nine criteria characterised by a set of attributes. The assessment model is based on six main nominal criteria (physical properties) defined by one attribute for each class. Thus, there should be at least six sample examples for each category. The remaining three criteria relate to a variety, date of harvesting, and storage life. The number of attributes can be set depending on the complexity of the sorting assessment and the task required.

Preliminary Data Processing

Data available in the database may consist of noisy data, irrelevant attributes, and missing data. Before applying any kind of intelligent data analysis algorithm, data processing is required, which is carried out in the following stages:

(1) Integration of data, where the elimination of inconsistencies in attribute names or attribute value names between data sets is performed from databases of different sources.
(2) Data cleansing, including the detection and correction of data errors, fill-in missing values, etc.
(3) Discrete data to convert a continuous attribute to a categorical attribute, taking only a few discrete values.
(4) Select the attribute relevant to the subset.

Data processing was performed using Toad for SQL Server 6.1. After processing, the database consisted of 4500 copies, where each contained nine input attributes including not only physical characteristics but also variety, the month of harvest, and storage time. The database contained three different varieties of apples harvested in July-September of the harvest year and a storage time of 3, 6, and 9 months. The summary of the results for the implementation of different classifiers using the WEKA tool. The effectiveness of the classifiers is compared by the number of processed specimens and the number of evaluation criteria.

Methods of Data Classification

Bayesian Network (BN)

Bayesian classifiers use a combination of conditional probability and a posteriori probability that can be

estimated from a training set to classify any information. This study uses a method based on a special class of the Bayesian Network Model (BN).

Bayesian network (Dale, 2012) $B = \langle G, \Theta \rangle$ is the oriented acyclic graph G, which simulates the probabilistic relations between a set of random variables $U = \{X_1, ..., X_n, \Omega\} = \{U_1, ..., U_{n+1}\}$, where each variable in U has certain conditions or values denoted by lower case letters $\{x_1, ..., x_n, \omega\}$. In this case, n corresponds to the number of attributes in the classifier. Each node of the graph is a random variable, and the arcs fix direct dependencies between the variables. The coding of conditional independent relations occurs according to the fact that each node is not dependent on its descendants. Conditional independent data reduce the number of parameters that are necessary to represent the probability distribution. Symbol Θ denotes a quantitatively defining set of parameters. Thus, each node of U_i will be represented as a local conditional distribution of probabilities considering the ancestors of A_{Ui}. The joint distribution of network probabilities is determined by local conditional distributions of probabilities as follows:

$$P(U) = \prod_{i=1}^{n+1} P(U_i \mid A_{U_i}) \quad (1)$$

According to Duda et al. (2012), the naive Bayesian decision rule assumes that all attributes are conditionally independent, considering the class label. This independence between functions ignores any correlations between them. Values of attributes X_i and X_j ($X_i \neq X_j$) are conditionally independent taking into account the node class label. Hence, x_i is conditionally independent of this class x_j whenever $P(x_i \mid \omega, x_j) = P(x_i \mid \omega)$ for all $x_i \in X_i$, $x_j \in X_j$, $\omega \in \Omega$, and when $P(x_j, \omega)>0$.

The structure of a naive Bayesian classifier is a network where each attribute conditionally does not depend on other attributes specified by the class label ω of a variable class. The variable of Ω class is the sole parent for each attribute X_i, designated as $V_{Xi}=\{\Omega\}$ for all $1 \leq i \leq n$. Therefore, the shared probability distribution for this network can be expressed through:

$$P(X_1, ..., Xn, \Omega) = P(\Omega)\prod_{i=1}^{n+1} P(X_i \mid \Omega) \quad (2)$$

Therefore, the conditional distribution of probability for classes from Ω can be expressed as:

$$P(\Omega \mid X_1, ..., X_n) = \alpha P(\Omega)\prod_{i=1}^{n+1} P(X_i \mid \Omega) \quad (3)$$

where α is the normalization constant. The selection of features is entered into this network by removing unnecessary features through the search algorithm.

Classification of k-Nearest Neighbours

The nearest neighbourhood (KNN) method is widely used due to its high effectiveness. The task of the k-nearest neighbourhood algorithm is to find the k-nearest neighbours for a given instance according to the class of most adjacent neighbours. The k-nearest neighbours should be found to classify the sample data X, and then X is assigned with a class label to which most of its neighbours belong. Selecting k also affects the performance of the k-nearest neighbour algorithm (Bhavsar and Ganatra, 2012). If the value of k is too low, the k-NN classifier may be vulnerable to re-election because of the noise present in the set of tutorial data. On the other hand, if the k value is too high, the nearest neighbourhood classifier may incorrectly classify a test pattern because its list of nearest neighbours may contain some data points that are far away from its neighbourhood. The k-NN algorithm is based on the belief that the data is linked in the features space. Consequently, all points are considered one after another to know the distance between the data points. At that, a single value of k is given, which is used to determine the total number of the nearest neighbours that define a class label for an unknown sample. If value k =1, it is called the nearest neighbour classification.

In the algorithm of k-nearest neighbour classification, the tutorial set $S = \{(x_1, y_1), ..., (x_n, y_n)\}$ should be determined in the phase of training, where $x_i = (x_i^1, ... x_i^d)$ is the d-dimensional vector of features of real numbers for all $i = 1,..., n$. Then the labels of y_i classes correspond to each x_i for all i, and $y_i \in C$, where $C = (1, ..., N_c)$ and N_c is the number of different classes. Therefore, the main task at this stage is to define y_{new} for x_{new}. In the next phase of testing, the nearest point from x_j to x_{new} should be established by calculating the Euclidian distance using the formula:

$$\sqrt{(x_j^1 - x_{new}^1)^2 + ..., + (x_j^d - x_{new}^d)^2} \quad (4)$$

and classified by $y_{new} = y_j$.

Artificial Neural Networks (ANN)

Artificial neural networks are self-trained adaptable statistical models based on an analogy with the human brain structure. Neural networks are constructed of neurons connected to each other by a set of weighted connections. The direct link network provides a common framework for representing nonlinear functional mappings between a set of input and output variables (Safonov, 2020). The network is constructed based on the input layer that is used to distribute inputs to a number of hidden layers whose outputs are connected

to the output layer. The unit outputs are connected to the next input via connection weights. The function at the node output in an artificial neuron can be represented as follows:

$$y_k = f_k\,(\alpha_k + \sum_{j\to k} w_{jk}\; f_j\,(\alpha_j + \sum_{i\to j} w_{ij} x_i)) \quad (5)$$

which displays the network inputs x_i on the target training output y_k for a network with one hidden layer, w_{ij} is the weight of the input connection to the hidden layer, α_j is the offset in the hidden layer nodes, and f_j is the activation function on the outputs of the hidden layer nodes. Pointers i, j, and k denote the type of connection from input to hidden layer and from hidden to the output layer, respectively.

Results of Classifier Efficiency Assessment

The results of accuracy using WEKA tools of different classifiers are presented in Table 1. The table shows the best results with different tutorial and test sets.

As can be seen from the data in the table, the productivity of all investigated methods decreases depending on the number of classified objects, and the k-NN and ANN algorithms demonstrate the highest efficiency for small data values (99 and 100 %, respectively). The smallest values 65 % correspond to BN and ANN classifiers for large data sets. The algorithm of the k-nearest neighbours is demonstrated based on the analysis of performance results (90%). However, in terms of computation speed, it turns out to be the slowest of all. The best indicators of computational time are demonstrated by the algorithm of neural networks, where the computational speed (0-0.07 s) is higher at any data size in comparison with other results.

Thus, the result of the efficiency analysis for the same data set provided showed that the best efficiency and accuracy are achieved with the k-NN algorithm (99%), but the calculations take more time (0.03-0.47 s), which increases with increasing the data volume. In terms of better performance (100%), the neural network is the right candidate for fast and sufficiently accurate results. The Bayesian classification has the same accuracy as a neural network, but another method of forecasting is close to k-NN, which also reduces its productivity.

Discussion

According to the above-mentioned comparison results and characteristics of different classification algorithms, it is necessary to derive the advantages and disadvantages for each of them.

The advantages of the naive Bayesian algorithm include high computational efficiency and computational speed, which require short learning times. While classifying images into components of plants, weeds, and soil using colour and position characteristics, Marchant and Onyango (2003) showed that the Bayesian classifier is optimal in terms of general classification error. Implementation of the Bayesian classifier is carried out both through simple database search and correct function selection making the indexing trivial, which gives less configurable parameters compared to the neural network (Kim et al., 2018). A study of a large database set of soil profiles from different locations in India (Bhargava and Bansal, 2021) showed that naive Bayesian classifications, using different data collection methods, demonstrated the most effective results. However, despite numerous significant advantages of this method, a number of disadvantages limit the applicability and productivity. A naive Bayesian classifier requires a large number of records to obtain good results. The accuracy of the algorithm decreases if the data volume is smaller.

The algorithm of the k-nearest neighbours is characterised by an easy way for understanding and implementation, as well as good training speed and reliability for noisy training data. According to Vaishnnave et al. (2019), this algorithm works well in applications with many class labels. The k-NN algorithm

Table 1: Results of efficiency estimation and time of algorithms calculation of the investigated classifiers

Data type	*Database size, number of elements*	*Number of criteria*	*Efficiency, %*			*Model training time, sec.*			*Model testing time, sec.*		
			BN	*k-NN*	*ANN*	*BN*	*k-NN*	*ANN*	*BN*	*k-NN*	*ANN*
Nominal	12	6	93	99	100	0	0	0.03	0	0.03	0
Segmental	1500	7	82	98	99	0.09	0	0.18	0.3	0.45	0.07
Large (market)	4500	9	65	90	65	0.07	0.02	0.07	0.29	0.47	0.03

is used for prediction into their respective categories, under different values of *K* and normalisation where the accuracy of predictions has varied from 100 to 66 % (Pandey and Jain, 2017). As the studies have shown, this algorithm does not require classes to be linearly separated but has a zero-price learning curve. It is well suited for multimodal classes and is sometimes reliable concerning noisy training data. For small numbers of classes, it shows good performance and calculation speed. The disadvantages of the k-nearest classifier include expensive calculations with a large number of potential neighbours, with which this unmarked sample can be compared (Karthikeya et al., 2020). Besides, it is sensitive to noisy or irrelevant attributes. It is characterised by memory limitations and slow operation. The performance of the algorithm depends on the number of measurements used.

One of the advantages of the neural network is its ease of use with several settings to configure. The algorithm of artificial neural networks has received a greater application in the field of agricultural production sorting (Kujawa and Niedbała, 2021). Comparative analysis (Gonzalez-Fernandez et al., 2019) with the Bayesian classification and fuzzy calculations on images of linear scanning of apple kernel (taken with the help of computer axial tomography) showed that ANN algorithm with 88% accuracy is better than fuzzy classes (80%) and Bayesian classifier (79%). It was noted by Chao et al. (2002) that the accuracy level of neural classifiers was 100% for calibration and 93.3% for verification while using combined spectral pixels of the image as input data is better compared to 93.4% for calibration and 90% for verification using fast Fourier transformation as input data for a neural network model through online multispectral imaging for separating useful and poor quality chicken carcasses. Moldes et al. (2017) proposed an improved method for identifying and classifying image treatment tools using the neural network approach in winemaking technologies. Due to better performance in classifying neural networks than those extracted from individual descriptors, the model gained excessive popularity (Utelbaeva et al., 2013). However, there are also disadvantages of this type of classification algorithm like long processing time in large neural networks, difficulty in determining the number of neurons and layers required, and slow learning (Zhou et al., 2019). Artificial neural networks were then used to quantify the samples and the experiments performed reached a classification level of 100 %.

Conclusions

The present study shows the classification methods for proper distribution and quality evaluation of agricultural products. The comparison of classifiers of k-nearest neighbours, naive Bayesian classifiers, and artificial neural networks for data analysis during apple fruit sorting were established, and the greatest accuracy of up to 99% for small data size was demonstrated by the classifiers of k-nearest neighbours at large calculation time of 0.47 s. At the same time, the neural network showed the best performance at any data size up to 100% at a calculation time of 0.03 s. It is shown that the most accurate and easy-to-implement are artificial neural networks due to their self-learning properties. However, in terms of lower error rates and simple algorithm structures, the k-NN and Bayesian classifiers are more preferable. Improvement of the algorithms of the latter can increase the quality of performance in the narrowest sets of classes. The results of this study demonstrate the importance of choosing the classification method correctly in order to increase the efficiency of sorting and quality assessment of various agricultural products.

References

Bhargava, A. and A. Bansal (2021). Fruits and vegetables quality evaluation using computer vision: A review. *Journal of King Saud University-Computer and Information Sciences*, **33(3):** 243-257.

Bhavsar, H. and A. Ganatra (2012). A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, **2(4):** 2231-2307.

Chao, K., Chen, Y.R., Hruschka, W.R. and F.B. Gwozdz (2002). On-line inspection of poultry carcasses by a dual-camera system. *Journal of Food Engi*neering, **51(3):** 185-192.

Dale, A.I. (2012). A history of inverse probability: From Thomas Bayes to Karl Pearson. Springer Science & Business Media.

Duda, R.O., Hart, P.E. and D.G. Stork (2012). Pattern classification. John Wiley & Sons.

Fan, S., Li, J., Zhang, Y., Tian, X., Wang, Q., He, X., Zhang, C. and W. Huang (2020). On line detection of defective apples using computer vision system combined with deep learning methods. *Journal of Food Engineering,* **286:**110102.

Findawati, Y., Astutik, I.I., Fitroni, A.S. Indrawati, I. and N. Yuniasih (2019). Comparative analysis of Naïve Bayes, K

Nearest Neighbor and C. 45 method in weather forecast. *Journal of Physics: Conference Series*, **1402(6):** 066046.

Gonzalez-Fernandez, I., Iglesias-Otero, M.A., Esteki, M., Moldes, O.A., Mejutoand J.C. and J. Simal-Gandara (2019). A critical review on the use of artificial neural networks in olive oil production, characterization and authentication. *Critical Reviews in Food Science and Nutrition*, **59(12):** 1913-1926.

Hemmatian, F. and M.K. Sohrabi (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, **52(3):** 1495-1545.

Ireri, D., Belal, E., Okinda, C., Makangeand, N. and C. Ji (2019). A computer vision system for defect discrimination and grading in tomatoes using machine learning and image processing. *Artificial Intelligence in Agriculture*, **2:** 28-37.

Karthikeya, H.K., Sudarshanand, K. and D.S. Shetty (2020). Prediction of agricultural crops using KNN algorithm. *International Journal of Innovative Science and Research Technology*, **5(5):** 1422-1424.

Kim, S., Parhi, P., Junand, H. and J. Lee (2018). Evaluation of drought severity with a Bayesian network analysis of multiple drought indices. *Journal of Water Resources Planning and Management*, **144(1):** 05017016.

Kujawa, S. and G. Niedbała (2021). Artificial neural networks in agriculture. *Agriculture*, **11(6):** 497.

MacDonald, J.M. (2020). Tracking the consolidation of US agriculture. *Applied Economic Perspectives and Policy*, **42(3):** 361-379.

Marchant, J.A. and C.M. Onyango (2003). Comparison of a Bayesian classifier with a multilayer feed-forward neural network using the example of plant/weed/soil discrimination. *Computers and Electronics in Agriculture*, **39(1):** 3-22.

Moldes, O.A., Mejuto, J.C., Rial-Otero, R. and J. Simal-Gandara (2017). A critical review on the applications of artificial neural networks in winemaking technology. *Critical Reviews in Food Science and Nutrition*, **57(13):** 2896-2908.

Narendra, V.G., Prasad, G.S. and A.J. Pinto (2020). A framework for quality evaluation of edible nuts using computer vision and soft computing techniques. *In:* International Conference on Harmony Search Algorithm. Springer, Singapore, pp. 339-348.

Pandey, A. and A. Jain (2017). Comparative analysis of KNN algorithm using various normalization techniques. *International Journal of Computer Network and Information Security*, **11(11):** 36-42.

Prabhakar, C.J. and S.H. Mohana (2018). Computer vision based technique for surface defect detection of apples. *In:* Computer Vision: Concepts, Methodologies, Tools, and Applications. IGI Global, pp. 1627-1639.

Rajesh, P. and M. Karthikeyan (2017). A comparative study of data mining algorithms for decision tree approaches using weka tool. *Advances in Natural and Applied Sciences*, **11(9):** 230-243.

Safonov, V. (2020). Assessment of heavy metals in milk produced by black-and-white holstein cows from Moscow. *Current Research in Nutrition and Food Science Journal*, **8(2):** 410-415.

Safonov, V.A., Danilova, V.N.,. Ermakov, V.V. and V.I. Vorobyov (2019). Mercury and methylmercury in surface waters of arid and humid regions, and the role of humic acids in mercury migration. *Periodico Tche Quimica*, **16(31):** 892-902.

Safonov, V.A., Ermakov, V.V., Degtyarev, A.P. and N.N. Dogadkin (2020). Prospects of biogeochemical method implementation in identifying rhenium anomalies. *IOP Conference Series: Earth and Environmental Science*, **421(6):** 062035.

Utelbaeva, A.B., Ermakhanov, M.N., Zhanabai, N.Z., Utelbaevand, B.T. and A.A. MelDeshov (2013). Hydrogenation of benzene in the presence of ruthenium on a modified montmorillonite support. *Russian Journal of Physical Chemistry A*, **87(9):** 1478-1481.

Vaishnnave, M.P., Devi, K.S., Srinivasanand, P. and G.A.P. Jothi (2019). Detection and classification of groundnut leaf diseases using KNN classifier. *In:* 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN). IEEE, pp. 1-5.

Zhou, L., Zhang, C., Liu, F., Qiuand, Z. and Y. He (2019). Application of deep learning in food: A review. *Comprehensive Reviews in Food Science and Food Safety*, **18(6):** 1793-1811.