International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

Classification of Different Wheat Varieties by Using Data Mining Algorithms

Kadir Sabanci [*1], Mustafa Akkaya [2]

Accepted 30th May 2016 **DOI: 10.18201/ijisae.62843**

Abstract: There are various applications using computer-aided quality controlling system. In this study, seed data set acquired from UCI machine learning database was used. The purpose of the study is to perform the operations for separation of seed species from each other in the seed data set. Three different seed whose data was acquired from the UCI machine learning database was used. Later it was classified by applying the methods of KNN, Naive Bayes, J48 and multilayer perceptron to the dataset. While wheat seed data received from the UCI machine learning database was classified, WEKA program was used. By changing the number of neurons, the highest classification success rate was achieved when the number of neuron was 7. The success rate with 7 neurons was 97.17%. When the classification success rate was calculated according to KNN for the different number of neighbors, the highest success rate was obtained as 95.71% for 4 neighbors.

Keywords: *WEKA, Multilayer Perceptron, KNN, J48, Naive Bayes*

1. Introduction

In Classification, training examples are used to learn a model that can classify the data samples into known classes. The Classification process involves following steps: Create training data set, identify class attribute and classes, identify useful attributes for classification (relevance analysis), learn a model using training examples in training set and use the model to classify the unknown data samples [1].

There are many studies in the literature in which data mining classification algorithms are used. The main areas are medical, food and agriculture.

(Jiang et al; 2013) used WEKA software to classify 11 fruits under varying pose and lighting conditions. (M. Omid; 2011) used the J48 decision tree method to classify by use of the acoustic properties of open and closed-shelled pistachios. The dataset was divided into two groups. 210 of 300 pistachios were assigned for training and rest of them were for test. (S. G. Ceballos-Magaña et al.; 2013) purposed to classify the silver and gold varieties of aged and extra-aged tequila. They used multilayer perceptron method. And by this method the highest truth rate was achieved. (E.M. de Oliveira et al.; 2016) used Bayes algorithm with artificial neural networks to classify in terms of evaluation and marketing of the colour of green coffee. In this study, Bayes algorithm was conducted for a group of 4 coffee bean. 1.15% generalization error was acquired with artificial neural network modelling. In Bayes algorithm, the classification was done for the colours: whitish, green, green of cane, bluish green.

(Karthikeyan et al; 2015) used classification algorithms such as j48, naïve bayes, multilayer perceptron, rondom forest through datasets from the UCI database by taking data from the hepatitis occurring in the liver. At the end of study the most successful percentage was acquired as a result of naive Bayes algorithm in the classification by patient cells in hepatitis patients. (Nowakowski et al; 2009) developed a neural model depending on digital photography for the determination of mechanical damage on corn. Primarily, the properties which separate damaged and healthy kernels from each other, were determined. At the end of study, an artificial neural network which is similar to the multilayer perceptron close to the human capacity to define, was created. (D. A. Aguiar et al.; 2010) studied on the pastures that were degraded in different levels in the state of Mato Grosso do Sul, in Brazil. In this study, MODIS time series were used to obtain fractional images and determine vegetation's. Input parameters required for Weka J48 classifier method, was acquired using small wave technique in various decomposition levels. Thus, Pastures were selected from Cerrado successfully. The distinction between different Pastures led to lower performance; the best results were acquired in pastures containing common plants followed by good grass.

In this paper, the dataset of seed species, obtained from UCI database which was consist of 3 different type of seeds, were used. The open-source WEKA software was used for the classification. The success rates and error values were presented for K-Nearest Neighbour Algorithm, Multilayer Perceptron, J48, and Naive Bayes classification methods.

Fruit industry is a major industry which contributes 20% of the nation's growth. Increase in the production and productivity is largely due to the adoption of improved technologies, which include quality planting material, balanced nutrients and timely protection against major insect-pests and diseases. India is the second largest producer of fruits with a production of 44.04 million tonnes from an area of 3.72 million hectares. This accounts 10% of the world fruit production. A large variety of fruits are grown in India of which apple, citrus, banana, grape, mango, guava, are the

[1] *Karamanoglu Mehmetbey University, Faculty of Engineering Department of Electrical and Electronics Engineering, Karaman, Turkey*
[2] *KMU, Faculty of Engineering Department of Energy Systems Engineering, Karaman,Turkey*
** Corresponding Author: Email: kadirsabanci@gmail.com*

major ones. Also, India is a large low cost producer of fruit, and horticulture has huge export potential.

In spite of the fact that India is blessed with a wide range of soil and climatic conditions for growing large number of horticultural crops, there are still several constraints which adversely affect development of a sound horticulture industry. Due to improper cultivation of fruits, lack of maintenance and manual inspection there has been a decrease in production of good quality of fruits. Farmers are finding difficulty, especially in finding the fruits affected by diseases which results in huge loss of revenue to the farmers and the nation. Non adoption of adequate and timely control measures against pests and diseases also cause major fruit losses. In the absence of comprehensive knowledge, disputes over costs, benefits, and the potential for harm of chemical pesticides easily become polarized [31]. Farmers are also concerned about the huge costs involved in these activities and severe loss. The cost intensity, automatic correct identification of diseases based on their particular symptoms is very useful to farmers and also agriculture scientists. Detection of diseases is a major challenge in horticulture / agriculture science. Development of proper methodology, certainly of use in these areas. One of the main concerns of scientists is the automatic disease diagnosis and control [15].

Computer vision systems developed for agricultural applications, namely detection of weeds, sorting of fruits in fruit processing, classification of grains, recognition of food products in food processing, medicinal plant recognition etc. In all these techniques, digital images are acquired in a given domain using digital camera and image processing techniques are applied on these images to extract useful features that are necessary for further analysis. To know the state-of-the-art in automation of the task/activities in horticulture field and automatic detection of fruit disease using computer vision techniques, a survey is made. The gist of a survey which carried out is given as follows.

(Jagadeesh D.Pujari et al; 2013) proposed grading and classification of anthracnose fungal disease in mangoes. Different types of segmentation techniques were used to separate and grade percentage of affected areas. GLRM was used to extract texture features and further classified fungal affected mango images from normal using Artificial Neural Network (ANN) classifier. (Sudheer reddy bandi et al; 2013) proposed machine vision and image processing techniques in sleuthing the disease mark in citrus leaves. Citrus leaves were investigated using texture analysis based on the Color Co-occurrence Matrix (CCM) and classified using various classifiers. (Shiv Ram Dubey et al; 2012) proposed image processing based approach to evaluate diseases of apple. Local binary features were extracted from the segmented image, and finally images were classified using a multi-class Support Vector Machine (SVM). (Patil et al; 2012) describes the method for extraction of color & texture features of diseased leaves of maize. The textures features like correlation, energy, inertia and homogeneity were obtained by computing GLCM. (Jayamala K. Patil and Raj Kumar, 2011) have provided advances in various methods used to study plant diseases/traits using image processing. The methods studied were for increasing throughput and reducing subjectiveness arising from human experts in detecting the plant diseases. (D. Moshou et al; 2011) developed a prototype system for detection of plant diseases in arable crops automatically at an early stage of fungal disease development and during field operations. Hyperspectral reflectance and multi-spectral imaging techniques were developed for simultaneous acquisition of images. An intelligent multi-sensor fusion decision system based on neural networks was developed to predict the presence of diseases. A robust multi-sensor platform integrating optical sensing, Geostationary Positioning System (GPS) and a data processing unit was constructed and calibrated. (D.S.Guru et al., 2011) have presented a novel algorithm for extracting lesion area and application of neural network to classify tobacco seedling diseases. First order statistical texture features were extracted from lesion area and Probabilistic Neural Network (PNN) is employed to classify anthracnose and frog-eye spots present on tobacco seedling leaves. (H. Al-Hiary et al., 2011) have evaluated a software solution for automatic detection and classification of plant leaf diseases. The affected area was segmented and texture analysis was done using CCM. Neural network classifier was used to classify various plant diseases. (Di Cui et al; 2010) reports research outcomes from developing image processing methods for quantitatively detecting soybean rust severity from multi-spectral images. To achieve automatic rust detection, an alternative method of analysing the centroid of leaf color distribution in the polar coordinate system was investigated. Leaf images with various levels of rust severity were collected and analysed. (Qing Yao et al., 2009) presented an application of image processing techniques and SVM for detecting rice diseases using shape and texture features. (Dae Gwan Kim et al; 2009) investigated the potential of using color texture features for detecting citrus peel diseases. Classification models were constructed using the reduced texture feature sets through a discriminant function based on a measure of the generalized squared distance. (Geng Ying et al., 2008) have provided various methods of image preprocessing techniques for recognition of crop diseases. (Di Cui et al; 2008) proposed a method to detect the infection and severity of of soybean rust. The test performed using multispectral image sensor could quantitatively detect soybean rust compared to laboratory-scale research. (Kuo-Yi Huang, 2007) have presented an application of neural network and image processing techniques for detecting and classifying phalaenopsis seedling diseases. The texture features using GLCM and color features were used in the classification procedure. A Back Propagation Neural Network (BPNN) classifier was employed to classify phalaenopsis seedlings diseases.

2. Material Method

2.1. Dataset

The purpose of this study is to be able to distinguish seed varieties named Rose, Canadian and Kaman from each other according to properties. In the study, for determining spices of a seed the data of the seed was processed by WEKA with KNN, Naive Bayes, J48 and multilayer perceptron algorithms. In the study, the seed data set was received from UCI [9]. The data set includes 3 classes named Rose, Canadian and Kama and 7 attributes. These wheat kernels are required to separate from each other due to the fact that they grow intertwined with each other and they have different financial returns. By getting random 70 pieces from each of the 3 types of seed, totally 210 samples were analysed. The properties of the data set consist of perimeter, area, core height, core length, core asymmetry and radius [10].

2.2. WEKA (Waikato Environment for Knowledge Analysis)

The system of WEKA was initially developed on JAVA language as open source at the University of Waikato in New Zealand. Machine learning over WEKA and many libraries related statistics come ready. Pre-processing of data, grouping and classification are some of the available library [11].

2.3. Multilayer Perceptron Algorithm

Multilayer Perceptron (multilayer sensors) is a sensor system

 This journal is © Advanced Technology & Science 2013

consisting of three layers. There are 3 layers called input layer, hidden layer and output layer. Input layer is the part that transfers the input data from outside of the neural network to hidden layer. There is no process on the data in the input layer. Any information entered into this layer is sent to hidden layer how it is as unprocessed. The intermediate layer sends info to the output layer by processing the information from the input layer. A multilayer perceptron may include multiple intermediate layer [11].
The output layer sends the data to outer world by processing information from hidden layer. The neural network creates a response to inputs [11].

2.4. K-Nearest Neighbor Algorithm

K-Nearest Neighbour algorithm is one of the algorithm system used to solve classification problems. While algorithm is applied, the matchings are done with the average of k-data appearing to be the closest as depended on the predetermined threshold value by comparing the similarity between data to be classified and normal behaviour data in the learning set [12].

2.5. J48 Decision Tree Algorithm

Decision tree algorithms is well known, widely used and powerful classification method. The strength of decision tree algorithms among the other classification methods is to be higher legibility of the model it manufactured and is that the evaluation process is higher than other techniques. The algorithm is in the form of trees as the name suggests and it is consist of leaf nodes and test nodes [13].

2.6. Naive Bayes Algorithm

Data mining is the process of reaching the information by processing of the available data. Naïve Bayes is one of the classification algorithms used in data mining. Naive Bayes is a measure of the probability of information taking part at the end of each stage of decision. The algorithm estimate related information by calculating the probability values [14].

3. Results and Discussion

To distinguish the wheat of Rose, Canadian and Kaman from each other, it was processed with Weka program. Classification success of wheat was obtained for KNN algorithm with different values of k-neighbour. Additionally the values of the root mean square error (RMSE) and mean absolute error (MAE) were found. Classification success rate obtained with the KNN algorithm and MAE and RMSE values are shown in (Table 1). The graph showing the change of MAE and RMSE error values depends on the number of neighbourhood in classification made with KNN algorithm was shown in Figure 1.

Table 1. The Success Rate and Error Values Obtained by using kNN Classifier

Neighborliness Number (k)	Classification accuracy (%)	MAE	RMSE
1	94.2857	0.0444	0.1938
2	91.9048	0.0539	0.1865
3	92.8571	0.0379	0.1275
4	**95.7143**	**0.0555**	**0.169**
5	92.381	0.0584	0.1702
6	93.3333	0.0608	0.1724
7	92.8571	0.0621	0.1706
8	92.8571	0.0666	0.1787
9	92.381	0.0694	0.1852
10	92.8571	0.0685	0.1818

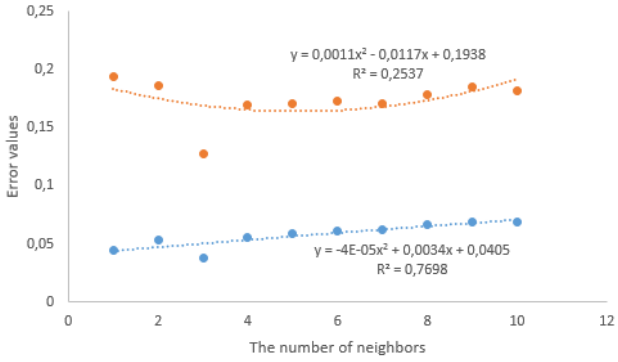

Figure1. Variation of error rate based on the number of neighborhood

Data at the same dataset was acquired classification success the wheat of the Rose, Canadian and Kama using multilayer perceptron model. Classification success rates among different numbers of neurons in the hidden layer and MAE, RMSE error rate was found. Classification success rates obtained using Multilayer perceptron and MAE and RMSE values are shown in (Table 2).

Table 2. Classification success rates obtained using multilayer perceptron and MAE, RMSE values

The number of neurons in hidden layer	Classification accuracy (%)	MAE	RMSE
3	96.6667	0.0335	0.1297
5	95.2381	0.0412	0.1511
7	**97.1429**	**0.0298**	**0.1181**
9	96.6667	0.029	0.1183
11	96.1905	0.0315	0.1243
13	95.2381	0.0345	0.1365
15	95.7143	0.0317	0.1291
17	96.1905	0.0308	0.1211
19	95.7143	0.0321	0.1305
21	96.6667	0.0319	0.1228
23	96.1905	0.0304	0.1177
25	95.7143	0.0338	0.1319

While the number of neurons in the hidden layer that the highest classification success was obtained is 7, multilayer perceptron model is shown in (Figure 3).

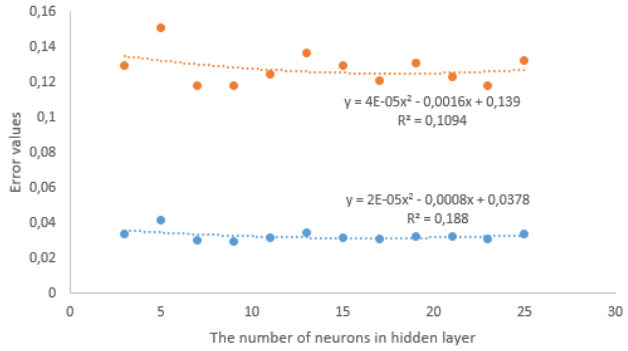

Figure 2. Variation of error rate based on the number of neurons in hidden layer

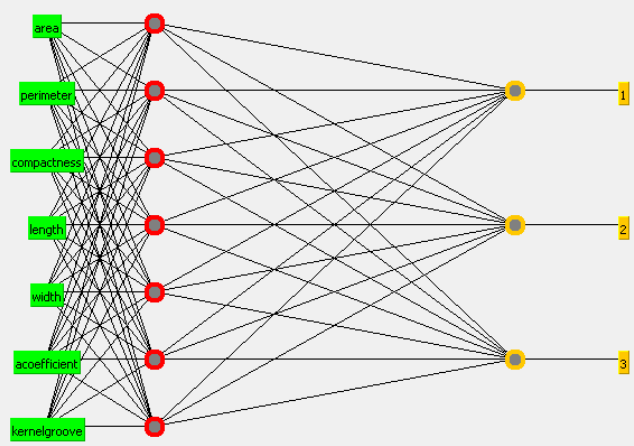

Figure 3. The structure of Multilayer Perceptron

J48 classification algorithm for the percentage of success and failure rates are as follows:

Correctly Classified Instances : 91.9048%
Mean absolute error : 0.0657
Root mean squared error : 0.2328

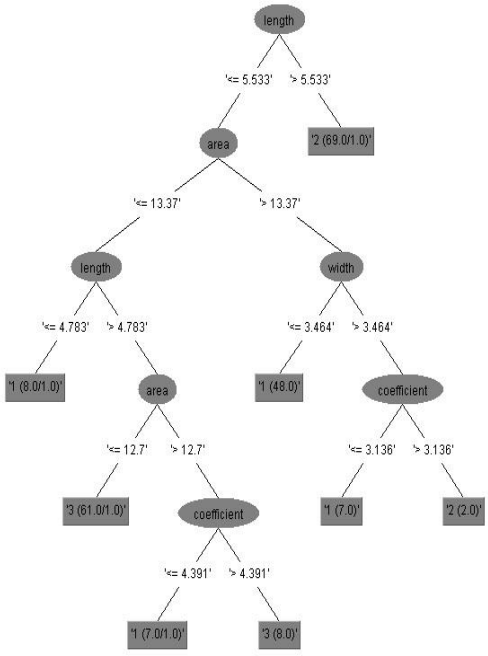

Figure 4. The structure of J48 tree

Table3. Success Rate Obtained By Using Various Data Mining Algorithms

Type Of Algorithm	Classification accuracy (%)	MAE	RMSE
K-Nearest Neighbor	95.7143	0.0555	0.169
Multilayer Perceptron	97.1429	0.0298	0.1181
J48 Decision Tree	91.9048	0.0657	0.2328
Naïve Bayes	91.4286	0.0635	0.2272

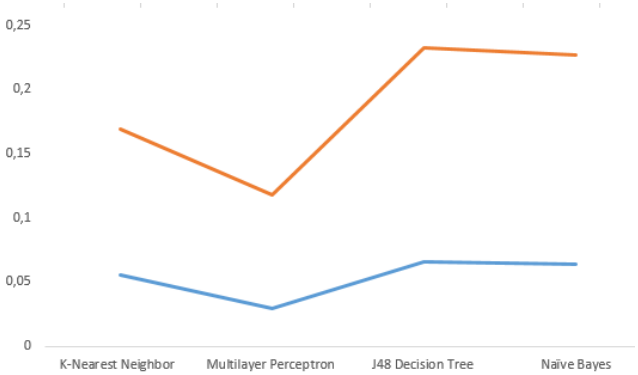

Figure 5. Variation of error rate based on data mining algotihms

4. Conclusion

In this study, by using classifiers of the data KNN, Multilayer Perceptron, J48 and Naïve Bayes in the data set including 7 attributes of rose, canadian and wedge wheat, classified and success rates were found. The success was found higher than when the classification was made by Multilayer perceptron algorithm. The greatest classification success was obtained with K-nearest neighbour algorithm and this value is 95.7143%. It was found that MAE error value is 0.0555 and RMSE error value is 0.169 in this neighbour value. While the number of neurons in the hidden layer is 7, the highest classification success was acquired and this value is 97.1749%. It was found that MAE error is 0.0298 and RMSE error is 0.1181. Classification success obtained with the J48 algorithm is 91.9048% and it was found that MAE error rate is 0.0657 and RMSE error rate is 0,2328. Classification success made by Naive Bayes was found as 91.4648% and it was found that MAE error rate is 0.0635 and RMSE error rate is 0.2272.

References

[1] T.C. Sharma, M. Jain, "WEKA approach for comparative study of classification algorithm," International Journal of Advanced Research in Computer and Communication Engineering, 2(4), 1925-1931, 2013.

[2] L. Jiang, A. Koch, S. A. Scherer, A. Zell, "Multi-class fruit classification using RGB-D data for indoor robots", Robotics and Biomimetics (ROBIO), 2013 IEEE International Conference on, 2013, pp. 587 - 592.

[3] M. Omid, "Design of an expert system for sorting pistachio nuts through decision tree and fuzzy logic classifier," Expert Systems with Applications, 38(4), 4339-4347, 2011.

[4] S. G. Ceballos-Magaña, F. de Pablos, J.M. Jurado, M.J. Martín, A. Alcázar, R.Muñiz-Valencia, R. Izquierdo-Hornillos, "Characterisation of tequila according to their major volatile composition using multilayer perceptron neural networks," Food chemistry, 136(3), 1309-1315,

This journal is © Advanced Technology & Science 2013

2013.

[5] E.M. de Oliveira, D.S. Leme, B. H. G. Barbosa, M. P. Rodarte, R. G. F. A Pereira, "A computer vision system for coffee beans classification based on computational intelligence techniques," Journal of Food Engineering,171, 22-27, 2016.

[6] T. Karthikeyan, P. Thangaraju, "Best First and Greedy Search based CFS-Naive Bayes Classification Algorithms for Hepatitis Diagnosis," Biosciences and Biotechnology Research Asia, 12(1), 983-90, 2015.

[7] K. Nowakowski, P. Boniecki, J.Dach, "The identification of mechanical damages of kernels basis on neural image analysis," In International Conference on Digital Image Processing (pp. 412-415), IEEE, 2009.

[8] D. A. Aguiar, M. Adami, W. Fernando Silva, B. F. T. Rudorff, M. P. Mello, J. D. S. V. Da Silva, "MODIS time series to assess pasture land," In Geoscience and Remote Sensing Symposium (IGARSS), IEEE International (pp. 2123-2126), IEEE, 2010.

[9] (Anonymous,2015a)UCI,Https://Archive.Ics.Uci.Edu/Ml/Datasets.Html Las Access: 22.12.2015.

[10] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, S. Żak, "Complete gradient clustering algorithm for features analysis of x-ray images," Information technologies in biomedicine (pp. 15-24), springer Berlin Heidelberg, 2010.

[11] (Anonymous, 2015b).WEKA, http://www.Cs.Waikato.Ac.Nz/~Ml/Weka/ Last Access: 19.12.2015.

[12] S. K. Caliskan, I. Sogukpinar, "Knn: K-Means and Methods K Nearest Neighbor Determination Of The Adoption Network," EMO, 120-124, 2008.

[13] C. Cengiz, "Data Mining Algorithm and Classification, Master' S Thesis," 2010.

[14] E. M. de Oliveira, D. S. Leme, B. H. G. Barbosa, M. P. Rodarte, R. G. F. A. Pereira, "A computer vision system for coffee beans classification based on computational intelligence techniques," Journal of Food Engineering, 171, 22-27, 2016.