International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org 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, random 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. 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.

2. Material Method

2.1. Dataset

The purpose of this study is to be able to distinguish seed varieties named Rose, Canadian and Kama 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

[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*

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 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 Neighbours 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].

For classification of wheat varieties, 7 attributes have been obtained. By using this data set, machine learning algorithms like MLP, kNN, J48 and Native Bayes algorithms have been proposed for classification. The block diagram of the classification process is seen in Figure 1.

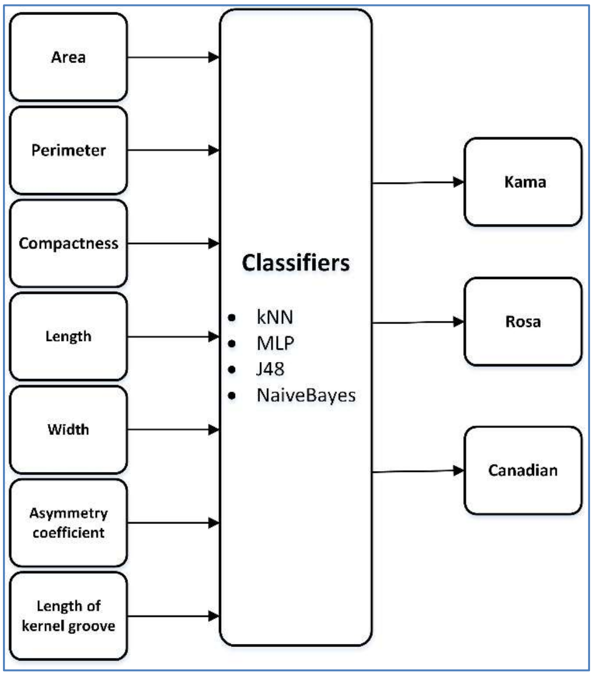

Figure 1. The block diagram of the classification process

3. Results and Discussion

To distinguish the wheat of Rose, Canadian and Kama from each other, it was processed with WEKA software. During training the 10 fold cross validation method was used as test option. 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 2.

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

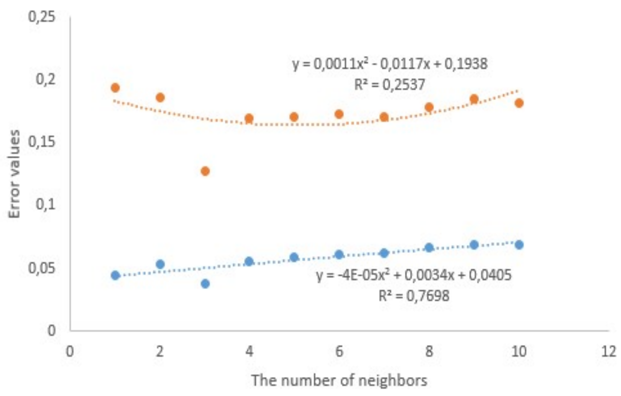

Figure 2. Variation of error rate based on the number of neighborhood

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.
The graph that shows the change of MAE and RMSE error values versus the number of neurons in the hidden layer in classification made with MLP algorithm was presented in Figure 3.

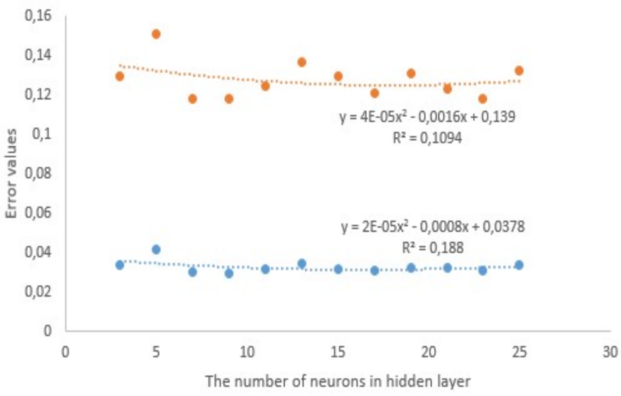

Figure 3. Variation of error rate based on the number of neurons in hidden layer

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 4).

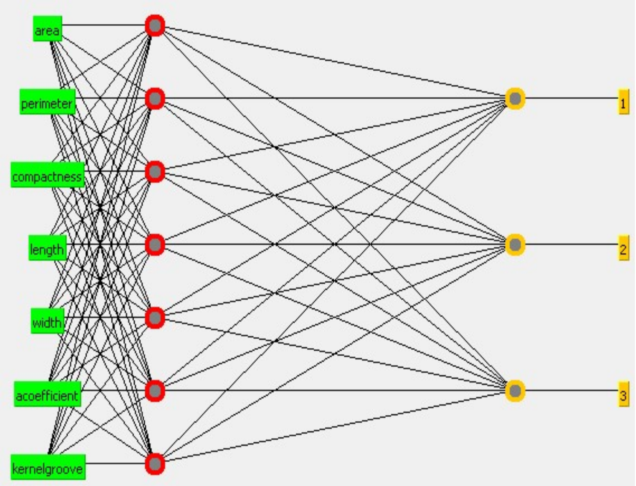

Figure 4. The structure of Multilayer Perceptron

The data in the same dataset are classified using the j48 and Naïve Bayes algorithms. By using these algorithms, classification success and error values are obtained. In Figure 5, the structure of the decision tree is presented.
Additionally, in Table 3 the classification successes and error values obtained with 4 different data mining algorithms are presented.
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

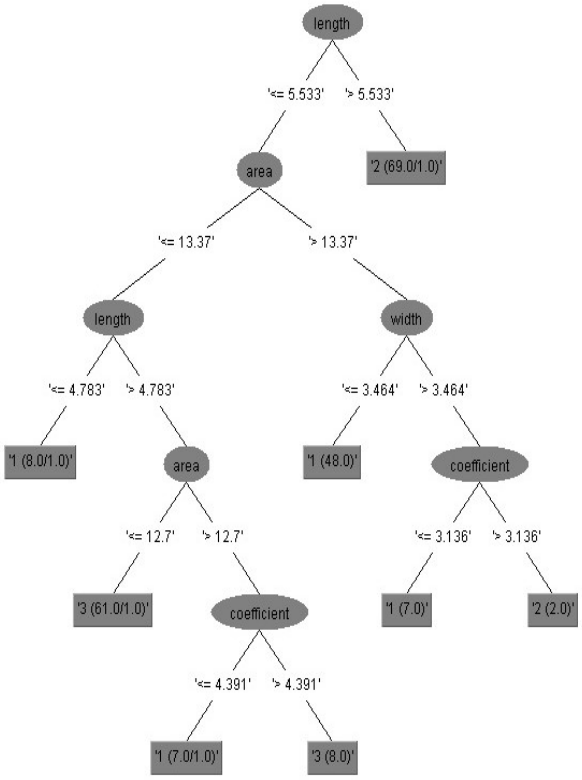

Figure 5. The structure of J48 tree

Table 3. Success rate obtained by using various data mining algorithms

Data Mining Algorithms	Classification accuracy (%)	MAE	RMSE
K-Nearest Neighbours	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

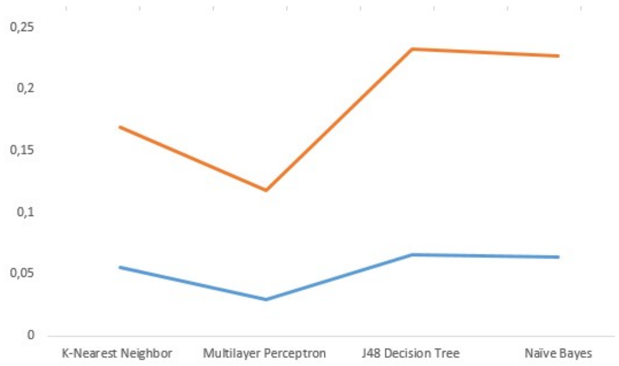

Figure 6. Variation of error rate based on data mining algorithms

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 Kama 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's 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 4 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 Biomimetic (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, "Characterization of tequila according to their major volatile composition using multilayer perceptron neural networks," Food chemistry, 136(3), 1309-1315, 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.