Journal of Pharmacognosy and Phytochemistry

Available online at www.phytojournal.com

E-ISSN: 2278-4136
P-ISSN: 2349-8234
JPP 2019; 8(3): 47-49
Received: 03-03-2019
Accepted: 05-04-2019

Puneet Dheer
SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

Purshottam
Department of Genetics and Plant Breeding, N.D. University of Agriculture and Technology, Kumarganj, Ayodhya, Uttar Pardesh, India

Vinod Singh
Department of Genetics and Plant Breeding, N.D. University of Agriculture and Technology, Kumarganj, Ayodhya, Uttar Pardesh, India

Correspondence
Puneet Dheer
SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

Classifying wheat varieties using machine learning model

Puneet Dheer, Purshottam and Vinod Singh

Abstract
Machine learning models viz., Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbor's and Naïve-Bayes were studied for their accuracy, precision and recall accommodating 100 samples each of five most important metric traits namely, plant height, number of fertile tillers/plant, spike length, number of spikelets/spike and number of grains /spikes in seven diverse and promising wheat varieties (DBW 14, Halana, NW 1012, NW 2036, Raj. 3077, UP 2338 and WH 147) in order to use the best one model for classifying the varieties. The average accuracy of Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbor's and Naïve-Bayes was 97.32%, 98.57%, 99.38% and 98.78%, respectively. The precision and recall of test data set of all 7 varieties were 100%. The K-NN model was thus found to be out performed over other models under studied and could therefore effectively be utilized for characterizing, classifying and or identifying the wheat varieties.

Keywords: Classification, wheat varieties, machine learning, k-nn

Introduction
Wheat being a second largest staple food crop just after corn, is grown globally. It is produced around 734.74 million tonnes worldwide led by China and followed by India and Russia (www.statista.com, 2019). India accounts for about 8.7 per cent of total wheat production in the world. In India, the introduction of the "Green Revolution" plan led to a substantial increase in wheat production. India has produced ever high to the tune of 99.70 million tonnes during 2017-18 and also estimated to cross 100 million tonnes in the pursuing year 2018-19 (The Hindu Business line. Jan. 24, 2019). Thus, wheat is playing a fundamental role in food and nutritional security leading to sustainable agriculture worldwide. It's all due to the concerted efforts made for the development of high yielding varieties and improved cropping technologies (Curtis *et al*., 2002; Goel *et al*., 2017; Yadav, 1991) [5, 10, 15]. The success of high yielding varieties depends upon the availability of seed and /or produce with assured high quality which brought in the trading. Thus, characterizing the varieties led to distinguish to each other is rather important. A number of approaches namely morphological (Malik *et al*.,2014) [12], biochemical (Sharma *et al*., 2015) [14], molecular (Yadawad *et al*., 2017) are being used. Machine learning is a trending technology nowadays and it can be used in modern agriculture to improve the productivity and quality of the crops. Image-based results of some studies (Camelo-Mendez *et al*., 2012; Hobson *et al*., 2007; Kong *et al*.,2013) [3, 6, 11] have been promising, they have included a limited number of rice varieties for categorization, requires high-end imaging processing techniques and the respective test dataset that makes the method too costly and not frequently available to the consumer.
In very recent, Dheer (2019) [7] and Dheer and Singh (2019) [8] distinguished and identified the promising rice varieties based on the self-collected dataset using K-NN classifier.
The present study aimed to differentiate the seven wheat varieties using machine learning methods with the following steps:(1) Pre-processing the acquired data (2) Evaluation of different machine learning classifiers (Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors and Naïve Bayes) and (3) Testing of the best-selected model after cross-validation.

Materials and Methods
1. Sample Collection and Pre-processing
The present investigation was taken up with seven promising varieties namely, DBW-14, Halana, NW-1012, NW-2036, Raj-3077, UP-2338 and WH-147 of wheat (*Triticum aestivum*). One hundred random samples comprising five features of each variety were acquired during field inspection. All the collected data were further divided into training and testing data set in a 70:30 ratio and then normalized. All these samples were collected when each variety reached

Their respective stages. Five different features selected were: plant height, number of fertile tillers/plant, panicle length, number of spikelets/spike and number of grains /spikes.

2. Models used

- **Logistic Regression:** A traditional statistical procedure, separates two classes by an S-shaped discriminant function through the decision space (Agresti, 1996) [1].

$$P(x) = \frac{1}{1 + e^{-y}}$$

Where *y* represents the output of the linear layer of a model and *P(x)* is a Probability of a given input *x*.

- **Fisher's Linear Discriminant Analysis**: It is a well-known classification technique that has been used successfully in many statistical pattern recognition problems. Aim is to project a dataset onto a lower dimensional space with good class separability in order to avoid overfitting. LDA determines the discriminant dimension in response-pattern space, on which the ratio of between-class over within-class variance of the data is maximized (Bishop 2007; Duda et al. 2000) [2, 9].
- **K-Nearest Neighbors:** is based on the principle that the samples within a dataset will generally exist in close proximity to other instances that have similar properties (Cover and Hart, 1967) [4]. If the samples are tagged with a classification label, then the value of the label of an unclassified sample can be determined by observing the class of its nearest neighbours. The K-NN locates the *k* nearest instances to the query sample and determines its class by identifying the single most frequent class label. The determination of K is crucial for K-NN. In this study, K was optimized by comparing K-NN models using K from 3 to 100 with a step of 1. Here, K distance was selected as 20 after cross validation.
- **Naïve Bayes Classifier:** The Naïve Bayes is a statistical classifier which is based on Bayes theorem (Mitchell, 1997) [13]. This method predicts probabilities of a given samples belonging to a specific class, which means that it provides the probability of occurrence of a given sample or data points within a particular class. The following equation is used to explain the principle of Bayes' theorem:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

where P(H|X) is the posterior and P(H) is the prior probability of class (target) whereas P(X|H) and P(X) are the likelihood and prior probabilities of predictor respectively.

3. Evaluation measure

The Accuracy of classification of the wheat varieties under study has been computed using the following expression which uses numerical details of correctly classified class from total samples of wheat in the dataset.

$$Accuracy = \frac{no.of\ indentified\ samples}{toal\ no.of\ samples} * 100$$

The Precision (What proportion of positive identifications was actually correct?) and Recall (What proportion of actual positives was identified correctly?) are also the important measure to consider for system evaluations which are calculated as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} * 100$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} * 100$$

Results and Discussion

The proposed plant classification system was tested on the dataset of seven different wheat varieties with 100 samples each. Each sample accompanied with five features. These data were trained and tested for four different classifiers (K-NN, LR, and LDA, NB). We have applied 10-fold cross-validation on the training set and selected the best suitable model based on average accuracy for further classification on unseen test data set. The Table 1 shows the average accuracy of all the models and 99.38% was the best average accuracy associated with the K-NN, and therefore this model was selected.

K-NN Classifier was trained on 70% of the collected dataset and tested on the remaining 30%. The Table 2 and Table 3 show that the Precision and Recall results for all varieties only with the best selected K-Nearest Neighbors classifier (K-NN) for training and testing data set respectively. The confusion matrix shows the number of correctly and incorrectly classified varieties against every variety in diagonal and non-diagonal elements respectively (Fig. 1). Here, incorrect classification contains both false positive and false negative test samples and correct classification includes all true positive and true negative values after application of K-NN and selected features. Accuracy, Precision and Recall scores calculated for analysis of the best-selected model. K-NN classifier gives an Accuracy of 99.38% and 100% on training and test dataset respectively. The Precision and Recall of test dataset are 100%. Although the above-mentioned other classifiers show different accuracies in comparison to each other and KNN outperforms all others for wheat classification. After experimenting with the proposed system, we conclude that KNN performs better than other classifiers for classification of these varieties.

Table 1: Cross Validation on Training data set in Wheat.

Classifier Models	Cross Validation (Average Accuracy)
Logistic Regression	97.32%
Linear Discriminant Analysis	98.57%
K-Nearest Neighbors	99.38%
Naïve Bayes	98.78%

Table 2: Precision and Recall of wheat Varieties under K-NN Model on Training data set.

Varieties	Precision	Recall
DBW-14	100%	98%
Halana	100%	100%
NW-1012	99%	100%
NW-2036	100%	99%
Raj-3077	100%	100%
UP-2338	99%	99%
WH-147	99%	100%

Table 3: Precision and Recall of wheat Varieties under K-NN Model on Test data set.

Varieties	Precision	Recall
DBW-14	100%	100%
Halana	100%	100%
NW-1012	100%	100%
NW-2036	100%	100%
Raj-3077	100%	100%
UP-2338	100%	100%
WH-147	100%	100%

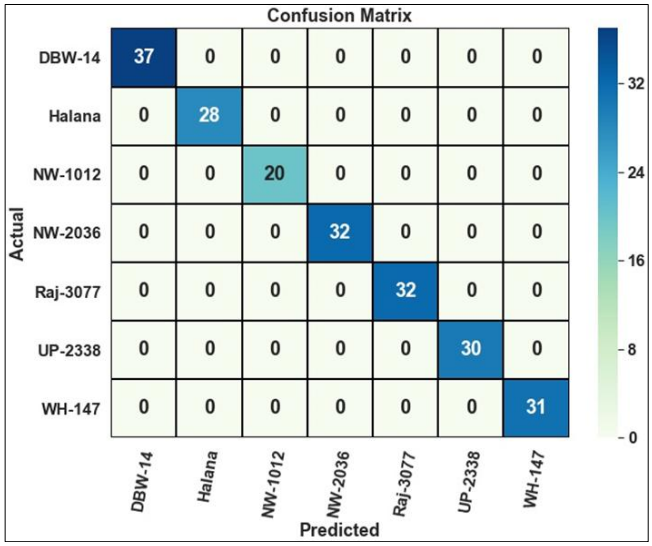

Fig 1: Confusion matrix of wheat varieties under K-NN Model on Test data set.

Conclusion

The present results obtained based on employing feature normalization revealed that K-NN model is quite promising for classification of wheat varieties. The precision and recall scores of collected datasets were 100%. Thus, this model can provide an accurate solution to the wheat varieties for their classification and/or identification problem as alternative to sophisticated image segmentation techniques. Further, this can also be used for mobile application, where even an occupational worker on the field can take a measurement the required features of wheat varieties to find the specific category that the wheat belongs to avoid admixture. Our future thrust will be being more focused towards our self-collected dataset of more wheat as well as other field crops varieties.

References

1. Agresti A. An Introduction to Categorical Data Analysis. Wiley, New York, 1996.
2. Bishop CM. Pattern Recognition and Machine Learning. Springer, New York, 2007.
3. Camelo Méndez GA, Camacho Díaz BH, del Villar Martínez AA, Arenas Ocampo ML, Bello Pérez LA, Jiménez Aparicio AR. Digital image analysis of diverse Mexican rice cultivars. Journal of Science, Food and Agriculture. 2012; 92:2709-2714.
4. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967; 13(1):21-7.
5. Curtis BC, Rajaram S, Macpherson HG. Bread wheat: Improvement and production. FAO Plant Production and protection Series No. 2002; 30:554.
6. Hobson DM, Carter RM, Yan Y. Characterization and Identification of Rice Grains through Digital Image Analysis. IEEE Instrumentation and measurement technology conference. IMTC 2007. Warsaw, 1-5.
7. Dheer P. Distinguishing of Rice Varieties by Using Machine Learning Models. International Journal of Advanced Research in Computer and Communication Engineering. 2019; 8(1):55-57.
8. Dheer P, Singh RK. Identification of Indian rice varieties using machine learning classifiers under press, 2019, 19(1).
9. Duda RO, Hart P, Stork DG. Pattern Classification. 2000. 2nd ed. John Wiley and Sons, New York.
10. Goel S, Singh K, Singh NK. Wheat improvement in India: Present and Future. Methods Molecular Biology. 2017; 1679:61-82.
11. Kong W, Zhang C, Liu F, Nie P, He Y. Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. Sensors. 2013; 13(7):8916-8927.
12. Malik Rekha, Sharma H, Sharma I, Kundu S, Verma A, Sheoran S *et al*. Genetic diversity of agro-morphological characters in Indian wheat varieties using GT biplot. Australian Journal of Crop Sciences. 2014; 8(9):1266-1271.
13. Mitchell T. Machine Learning. McGraw Hill, NY, 1997/
14. Sharma D, Saharan V, Joshi A, Jain D. Biochemical characterization of bread wheat (Tritticum aestivum L.) genotypes based on SDS-PAGE. Triticeae Genomics and Genetics. 2015; 6(2):1-7.
15. Yadav RDS. Breeding thrust for increasing the adaptability and yield in wheat under diara land condition. Crop Research. 1991; 4(2):258-263.