Plant Archives Vol. 19 No. 1, 2019 pp. 155-158 **e-ISSN:2581-6063 (online), ISSN:0972-5210**

IDENTIFICATION OF INDIAN RICE VARIETIES USING MACHINE LEARNING CLASSIFIERS

P. Dheer[1] and R. K. Singh[2]

[1]Depart. of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai-603203 (T.N.)
[2]Depart. of Genetics and Plant Breeding, N.D. Univ. of Agri. and Technology, Kumarganj, Ayodhya-224229 (U.P.)

Abstract

Rice has always been one of the most globally consumed foods which contributes significantly more than 60 per cent of world population.A large number of rice varieties have been cultivated, imported andexported throughout the world. Rice varieties can be admixture during their production and preprocessing.Numerous studies have been done for classifying plant types and identifying diseases of various crops particularly using imaging techniques. The plant type identification problem is further complicated by common object recognition constraints due to light, pose and orientation. The present study was undertaken to distinguish the eight different Indian rice varieties by their respective collected features and applying machine learning models to develop a rice variety inspection system. The study was carried out with simple classifier models like Linear Discriminant Analysis, Logistic Regression, K-Nearest Neighbor's and Naïve-Bayes method. KNN out performed over the other methods with an Accuracy, Precision and Recall of 99.16%, 99.12% and 99.12%, respectively.

Key words: Identification, rice varieties, machine learning, K-NN

Introduction

India is the second largest producer of rice in the world after China. Rice (*Oryza sativa* L.) is an essential staple food for more than half of the global population. The quality and variety of rice grains are usually determined based on its various quantitative and qualitative traits. The classification of rice varieties into a specific category is the most important interest in the domain of specific professionals to quantify the level of genetic purity and also to improve the quality of rice exports in the country. Typically to distinguish the different varieties requires various sampling by inspection on the agricultural fields by skilled workers. Rice varieties are being ad mixture during the cultivation, harvesting, and processing which reduce authenticity and the quality of the products. However, the existence of a large number of various varieties makes it quite difficult to analyze and classify it by the novice worker. The varieties can be distinguished in the field by recognizing their plant height, panicle density, grain types and color etc.

Molecular marker-based methods have been applied for identifying rice varieties. Steele *et al.* (2008) selected insertion and deletion markers to distinguish Basmati rice grains from some other fragrant rice varieties. Further, Applied Random Amplified Polymorphic DNA (RAPD) approach was used to fingerprint rice grains of 13 Italian accessions (Cirillo *et al.*, 2009). Determined the genetic variability of certain Chilean and foreign commercial rice cultivars using Simple Sequence Repeat (SSR) markers (Becerra *et al.*, 2015). Nevertheless, SSR markers were applied to distinguish 36 varieties of rice grains from different countries (Chuang *et al.*, 2011). These genetic marker-based methods are accurate but, apart from other reasons, costly for real-time applications.

Besides, Image-based methods and image processing techniques have also been amplified for identifying the different varieties of rice based on their size, shape and color (Hobson *et al.*, 2007). The rice grains of 9 Mexican cultivars were classified by performing Principle Component Analysis (PCA) and hierarchical analysis (Camelo-Méndez *et al.*, 2012). Classified the rice seeds of 4 accessions using a near-infrared hyperspectral imaging system and various machine learning algorithms (Kong *et al.*, 2013). Image-based results of these studies have been promising, they have included a limited number

of 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, distinguished the self-collected dataset using K-NN classifier showed promising results for categorizing the 4 rice varieties without using sophisticated image techniques for novice field workers (Dheer, 2019). However, they have included a very limited number of varieties.

The current study aimed to differentiate the rice varieties of 8 Indian varieties using machine learning methods. The following were the specific objectives of the study: (1) Pre-processing the acquired data; (2) Different machine learning classifiers are evaluated including Linear Discriminant Analysis, Logistic Regression, K-Nearest Neighbors and Naïve Bayes method (3) Tested on the best-selected model after cross-validation.

Materials and Methods

1. Sample Collection and Pre-processing

The present investigation embodied eight Indian promising varieties namely, Jal Lahari, Kasturi, Madhukar, NDR-97, Sarjoo 52, Swarna, Taraori Basmati and Type-100. One hundred random samples comprising six features of each variety were acquired during field inspection. All the collected data were further divided into training and testing data set in 70:30 ratio and then normalized. All these samples were collected when each variety reached their respective maturity stages. Six different features selected were: plant height, panicle length, number of grains/panicles, number of effective tillers, grain length and grain breadth.

2. Models used

- **K-Nearest Neighbors:** The K-NN classifies test sample based on the majority of its K-Nearest Neighbors with minimum distance signifies most common attributes. 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 (Duda *et al.* 2000; Bishop 2007).
- **Naïve Bayes Classifier:** The Naïve Bayes is a statistical classifier which is based on Bayes theorem (Mitchell, 1997). 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.

- **Fisher's Linear Discriminant Analysis**: It is most commonly used as dimensionality reduction technique in the pre-processing step for classification. 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 (Duda *et al.* 2000; Bishop 2007).
- **Logistic Regression:** A traditional statistical procedure, separates two classes by an S-shaped discriminant function through the decision space (Agresti, 1996).

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

where y is a linear model and $P(x)$ is a Probability of a given input x.

3. Evaluation measure

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

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

The Precision and Recall are also the important measure to consider for system evaluations which are calculated as follows:

$$Precidion = \frac{\sum True\ Positive}{\sum Predicted\ Condition\ Posutive} \times 100$$

$$Recall = \frac{\sum True\ Positive}{\sum Condition\ Positive} \times 100$$

Results and Discussion

The proposed plant classification system was tested on the dataset of eight different rice varieties with 100 samples each. Each sample accompanied with six features. These data were trained and tested for four different classifiers (K-NN, LR, LDA, NB). We have applied 10-fold cross-validation on the training set and

Table 1: Cross Validation on Training data set.

Classifier Models	Cross Validation (Average Accuracy)
Logistic Regression	97.67%
Linear Discriminant Analysis	98.97%
K-Nearest Neighbors	99.28%
Naïve Bayes	98.57%

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

Varieties	Precision	Recall
Jal Lahari	98%	100%
Kasturi	99%	99%
Madhukar	100%	100%
NDR-97	99%	100%
Sarjoo-52	99%	100%
Swarna	100%	100%
Taraori Basmati	100%	96%
Type-100	100%	100%

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

Varieties	Precision	Recall
Jal Lahari	100%	100%
Kasturi	93%	100%
Madhukar	100%	100%
NDR-97	100%	100%
Sarjoo-52	100%	100%
Swarna	100%	100%
Taraori Basmati	100%	93%
Type-100	100%	100%

selected the best suitable model based on average accuracy for further classification on unseen test data set. Table 1 shows the average accuracy of all the models and 99.28% was the best average accuracy associated with the K-NN was selected.

K-NN Classifier was trained on 70% of the self-collected dataset and tested on the remaining 30%. Table 2 and Table 3 shows 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 Figure 1shows the number of correctly and incorrectly classified varieties against every variety in diagonal and non-diagonal elements respectively. 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

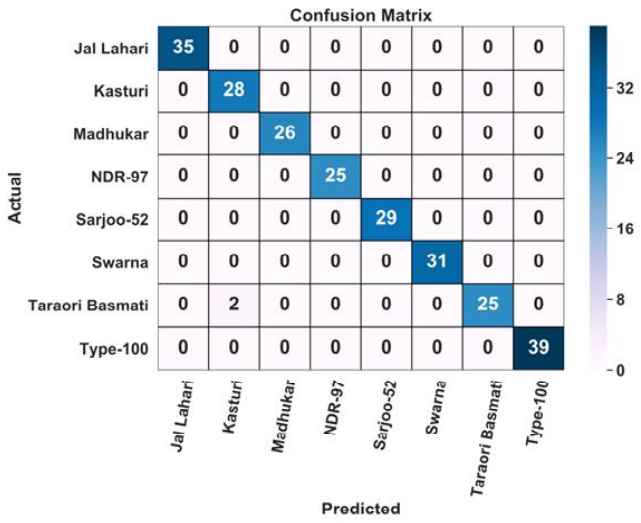

Fig. 1: Confusion Matrix of Rice Varieties Under K-NN Modelon Test data set.

calculated for analysis of the best-selected model. K-NN classifier gives an Accuracy of 99.28% and 99.16% on training and test dataset respectively. The Precision and Recall of test dataset are 99.12% and 99.12%, respectively. Although the above-mentioned other classifiers show different accuracies in comparison to each other and KNN outperforms all others for rice classification. After experimenting with the proposed system, we conclude that KNN performs better than other classifiers for classification of varieties.

Conclusion

In this study, previous work of different approaches was presented by their pros and cons. The results obtained based on employing feature normalization and K-NN model is quite promising for classification. The precision and recall scores of self-collected datasets remained above 99% and 99% respectively. The reported results suggest that this method can provide an accurate solution to the rice varieties for their classification and/or identification problem alternative to sophisticated image segmentation techniques as reported earlier. The proposed approach can be used for mobile application, where an occupational worker on the field can take a measurement of rice varieties features to find the specific class that the rice belongs to avoid admixture. Our future work is being more focused towards our self-collected dataset to include more rice varieties to address automatic identification of rice varieties in particular and other field crops varieties in general.

References

Agresti, A. (1996). An Introduction to Categorical Data Analysis. Wiley, New York.

Becerra, V., M. Paredes, E. Gutiérrez and C. Rojo (2015). Genetic diversity, identification, and certification of Chilean rice varieties using molecular markers. Chilean *J. Agric. Res*., **75**: 267–274.

Bishop, C.M. (2007). Pattern Recognition and Machine Learning. Springer, New York.

Camelo-Méndez, G.A., B.H. Camacho-Díaz, A.A. del Villar-Martínez, M.L. Arenas-Ocampo, L.A. Bello-Pérez and A.R. Jiménez-Aparicio (2012). Digital image analysis of diverse Mexican rice cultivars. *J. Sci. Food Agric*., **92**: 2709–2714.

Chuang, H., H. Lur, K. Hwu and M. Chang (2011). Authentication of domestic Taiwan rice varieties based on fingerprinting analysis of microsatellite DNA markers. *Botanical Stud*., **52**: 393–405.

Cirillo, A., S. Del Gaudio, G. Di Bernardo, U. Galderisi, A. Cascino and M. Cipollaro (2009). Molecular characterization of Italian rice cultivars. *Eur. Food Res. Technol*., **228**: 875–881.

Dheer, P. (2019). Distinguishing of Rice Varieties by Using Machine Learning Models. *International Journal of Advanced Research in Computer and Communication Engineering*, DOI:10.17148/IJARCCE.2019.8113

Duda, R.O., P. Hart and D.G. Stork (2000). Pattern Classification 2nd ed. John Wiley and Sons, New York.

Hobson, D.M., R.M. Carter and Y. Yan (2007). "Characterisation and Identification of Rice Grains through Digital Image Analysis," *IEEE Instrumentation & Measurement Technology Conference IMTC*, Warsaw, pp. 1-5.

Kong, W., C. Zhang, F. Liu, P. Nie and Y. He (2013). Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. *Sensors,* **13(7)**: 8916–8927.

Mitchell, T. (1997). Machine Learning. McGraw Hill, NY.

Steele, K.A., R. Ogden, R. McEwing, H. Briggs and J. Gorham (2008). In Del markers distinguish Basmatis from other fragrant rice varieties. *Field Crops Res*., **105**: 81–87.