Paper Review: 01

Title: Comparison of multiclass classification techniques using dry bean dataset.

Published date: 20th June, 2023

Link: https://www.keaipublishing.com/en/journals/international-journal-of-cognitive-computing-in-engineering/

1. Introduction

Machine learning techniques are becoming more popular in the fields of medical, biostatistics, bioinformatics, agriculture, business etc. Many scientists have worked to develop and identify qualityful seeds in the agricultural sector using artificial intelligence algorithms. There are many computational equipments available for controlling the quality of foods and agricultural goods, but most of them are done with the use of conventional techniques. Machine learning and computer vision have been used to ensure high-quality food product in greater quantities. Dry beans are the most nutritious and widely cultivated vegetable found (Fabaceae-Leguminosae) all over the world. A technique to detect and categorize seed features rapidly and repeatedly is crucial to improve the response of plants and/or tolerance to environmental stimuli. The XGB classification technique shows better performance among the selected classifiers with the help of ADASYN algorithms, and the proposed ML-based system has improved accuracies of 1.4%, 5.5%, 0.40%, and 11.5% compared to RF, DT, KNN, MLP respectively.

2. Related works and motivations

In 2021, Oliveira et al. developed a fast and reliable computer vision system to classify fermented cocoa beans into four categories. In 2020, Sanl et al. evaluated the performance on different datasets using KNN, J48, SMO, NB, NBM, BAGGING and JRIP classification algorithms.

Koklu et al. introduced a Computer Vision System (CVS) for recognizing registered varieties of dry beans with comparable traits. They assessed their performance by comparing MLP, SVM, KNN, and DT classification algorithms using 10-fold cross validation strategy.

3. Research framework and relevant materials

The proposed framework consists of multiple stages, including data collection, data scaling, and classification using eight known classifiers. AUC, ACC, MSE, F 1 - score, FPR, Kappa SE and SP are computed to evaluate the performance of the classifiers.

3.1. Data collection

For experimental evaluation, a dry bean dataset was collected from the University of California, Irvine's Machine Learning Repository. The dataset comprises 13,611 items of seven distinct registered dry beans.

The collected dataset has sixteen distinct features, twelve dimensions and four distinct shapes, including area, perimeter, length of the major axis, aspect ratio, eccentricity, equivalent diameter, solidity and convex area. 11) Roundness, 12) Compactness, 13) ShapeFactor1, 14), ShapeFactor2, 15), ShapeFactor3), and 16) were calculated for dry bean classes.

3.2. Basic information and descriptive analysis

The frequency distribution of dry bean classes is shown in Table 1 and Fig. 2 . The Bombay class has the lowest number of seed samples and the highest average seed weight, while Derma-son has the most observations with an average weight 0.28.

3.3. Data pre-processing

We have detected missing values and outliers in the dry bean dataset using statistical method boxplot and interquartile range (IQR). The outliers are removed with the help of IQR in python program to the related variables as shown in Fig. 4 .

3.4. Classification models

In this study, state-of-the-art classifiers like RF, XGB, SVM and MLP outperform well-known classifiers in dry bean classification.

3.4.1. Logistic regression classifier (LR)

The Logistic Regression (LR) model is a probabilistic statistical classification technique used to build a categorical dependent variable or a categorical outcome variable. It makes assumptions about under- lying data distribution and predictor independence.

3.4.2. K-nearest neighbor classifier (KNN)

KNN is a distance based supervised machine learning technique that uses training data to categorize new data points. It returns an integer number representing the productivity (labels) of a classification algorithm output.

3.4.3. Decision tree classifier (DT)

A Decision Tree (DT) classifier uses the divide and conquer principle to classify patterns by filtering them through tree tests.

3.4.4. Random forest classifier (RF)

In RF method, we have utilized a few numbers of de-correlated DTs and Gini as impurity index to develop a classification technique.

3.4.5. Support vector machine classifier (SVM)

SVM is the most widely used classification technique for predicting the class label of unknown sample based optimal decision boundary. It uses a kernel function to train the SVM model and transfer the feature vectors into a higher-dimensional space.

3.4.6. Naïve Bayes classifier (NB)

NB is a probabilistic classifier that applies Bayes theorem to achieve the highest level of performance on large datasets with high dimensionality.

3.5. Performance measures

There are several evaluation metrics to measure the performance of a machine learning algorithm, including accuracy, error rate, sensitivity, specificity, mean square error, recall, false positive rate, kappa and F 1 -score.

4.1. Confusion matrix for selected different algorithms

In this study, the confusion matrix is adopted to visualize and summarize the performance of the classifiers. It reveals that the ADASYN algorithm improves the accuracy of the classifiers, with the exception of Dermason dry bean seeds.

The XGB classifier attains the highest performance measure when applying ADASYN algorithm, including an ACC that is more than 95% and an AUC that is 99.64%. Additionally, the rest of the performance measures are higher compared to other selected models when applying ADASYN algorithm.

The KNN model shows good performance with ADASYN algorithm, but the RF classification model presents a relatively poor performance in the absence of ADASYN algorithm. The MLP model shows the lowest accuracy with ADASYN algorithm.

XGB and RF exhibit the greatest performance in terms of their AUC values as well as the classification thresholds, while the NB model has somewhat worse classification performance. The KNN and SVM models have also indicated relatively superior accuracy measures.

Eight models were used to classify dry beans. The XGB approach performed better than other models in Seker, 97% in Barbunya, 100% in Bombay, 97% in Cali, 96% in Horoz, 85% in Sira and 99% in Dermason.

5. Discussion

In this framework, several machine learning techniques are applied to classify dry beans from crop production with low computational cost and overcome bean intra-class variations.

Our multiclass classification model outperformed prior competing methods on various datasets, with respect to precision, recall and accuracy. The accuracy of our method is 83.90% for 14 kinds of rice seeds, and 93.00% on 5 variants of seed.

We used 13,611 items with 7 different types of dry beans collected from various planting areas in Turkey under varying imaging conditions to evaluate the XGB classification model. The model performed well with possible highest accuracy when all classes have equal number of samples.

5.1. Reasons of better performance behind the XGB classifier

The machine learning algorithm with XGB classifier performs better due to gradient boosting, minimizing loss function, and avoiding overfitting. It also overcomes the tiring process using approximate greedy algorithm.

5.2. Strengths, limitations, and future scopes of the study

This study proposes a technique to identify automatically uniform seed varieties for more crop production with reducing computational cost.

The developed algorithm performs well on a dataset with 13,611 items, but may be reduced with poor data pre-processing and segmentation.

Although the classifiers have achieved satisfactory accuracy for this dataset, it still suffers from various real-time challenges in uniform bean identification. Feature fusion may improve the performance.

6. Conclusions

This work studies the classification performance of dry bean dataset with imbalanced and balanced distribution. XGB beats all other approaches with both balanced and imbalanced classes for the experimental dataset with ACC of 93% and 95% respectively, and KNN and RF algorithms also demonstrate superior performance in terms of accuracy.

Acknowledgments

The authors thank the Statistics Discipline, Science, Engineering and Technology School, Khulna University, Bangladesh, and Milon Sheikh for their contributions to this article.