Soybean crop disease classification using machine learning techniques

Rajashree Krishna
Dept of CSE, Manipal Institute of Technology
Manipal Academy of Higher Education
Manipal, India
raji.krish@manipal.edu

Prema K V
Dept of CSE, Manipal Institute of Technology
Manipal Academy of Higher Education
Manipal, India
prema.kv@manipal.edu

***Abstract*—Machine learning is very widely used for many applications like classification and regression. Diseases in the soybean crop are classified using machine learning techniques. Physic crop properties and weather parameters are used as a attributes for classification. K nearest neighbor, naive Bayes, decision tree, neural network algorithms are used for classification. The result is compared with the ensemble classifier called bagging.**

***Index Terms*—Machine learning, crop disease, decision tree, weather parameters**

I. Introduction

One of the applications of artificial intelligence is Machine learning. Machine learning will help the systems to get the ability of learning automatically and enhance the learning from experience without being programmed explicitly. Development of computer programs are focused by machine learning and those programs access data and use it themselves for learning. The most commonly and widely acquired definition is: machine learning is the potential for computers to learn and act without being explicitly programmed.

The field of machine learning has the concern about the development of computer programs that automatically enhance the experience. The intention behind the area of machine learning is that a system learns to perform a specific task by understanding and considering a training set of samples [1]. Then the computer performs the same job with unknown data which was not trained before. In recent years many machine learning applications have been developed. Agriculture crop yield is greatly affected by many pests and diseases. In order to improve the yield, early diagnosis and control of disease are very important. Since long time, soybean is considered as meat, milk, cheese, bread as well as oil for the people of China and East Asia. The early published writings of those countries called soybean as ”gold from the soil”. The main Brazilian agricultural product has been Soybean , contributing significantly to the country’s trade balance.

978-1-7281-9885-9/20/$31.00 © 2020 IEEE

Soybean is the rapid growing crop in India and it is considered as a Kharif crop. The leading producer of soybean in India is Madhya Pradesh, followed by Maharashtra and Rajasthan [2]. Around the world Soybean is recognized as one of the outstanding crops. It is majorly used for protein, animal feed and vegetable oil.

Soybean is considered to be an important food commodity because of its high protein content (greater than 40%) and high oil content (greater than 20%). Since the Soya Protein supplies a sufficient amount of amino acids, it is called complete protein. Soybean oil does not contain cholesterol. Low yield is the major problem with the soya industry of the country because diseases are one of the reason. Soybeans are affected by several diseases like downy mildew, pod and stem blight, phytophthora root and stem rot, brown spot, cercopsora leaf blight, purple seed stain, frogeye leaf spot and many more. The work deals with the classification of the soybean diseases based on weather data, physic crop properties, plant properties, and crop management properties. The dataset is available in the UCI machine learning repository. Machine learning techniques like Decision tree, Naive Bayes, k nearest neighbor and ANN will be used to develop the classification model. Ensemble classification model is also used for the comparison.

II. Literature Survey

In this section different crop and plant disease detection, classification and prediction model is discussed. The classification models which are developed using artificial neural network and machine learning techniques are also discussed. Jayraj Chopda et al [3] detected cotton crop disease using decision tree classifier. Here the temperature and soil data captured from sensors are considered as input for disease detection.

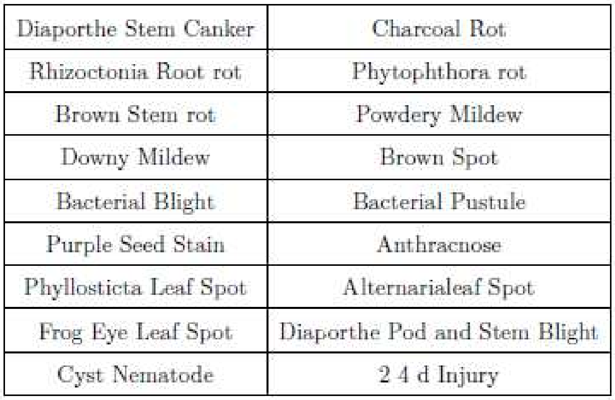
Sukmawati Nur Endah et al [11] selected different attributes to find out the disease and pest that appears in the soybean plant. Using Learning Vector Quantization2 method authors classified the disease with 91% accuracy. The paper [4] discusses about indoor environment classification using machine learning approach based on real-time measurements of the radio frequency (RF) signal in a practical environment. Different machine learning classification methods are discussed which includes support vector machine, decision trees, and k-nearest neighbor using different radio frequency features.

Sk Al Zaminur Rahman et al [5] did the classification of soil using machine learning techniques like weighted KNN, bagged trees and support vector machine. The intention of the proposed work is to design a suitable model for classifying various kinds of soil series data and then suggesting suitable crops for certain areas. K nearest neighbor, multilayer perceptron, Decision Trees machine learning algorithms are used in Microarray Breast Cancer Data Classification [6]. Maize crop disease classification using multi-class support vector machine is concluded by K.R.Aravind et al [7]. In this study, 2000 images of maize leaf are taken from PlantVillage dataset for classifying three different diseases namely common rust, Cercospora leaf spot, leaf blight as well as healthy leaves. Classification of Mobile Application Review for the Indonesian Language is designed using machine Learning Approach by Indra Budi and Yudo Ekanata [8]. The paper concludes that decision tree has the unacceptable performance as compared to support vector machine, Naïve Bayes, and logistic regression in the process of classification.

Everton Castelão Tetila et al [9] used different images for soybean diseases classification. They classified images into three classes like with no disease class, target spot class, and powdery mildew class. They used SVM, Random forest, kNN, Naïve Bayes, J48 and AdaBoost classifiers. Found SVM is the best classifier for soybean disease classification. Zhongqi Lin et al [10] developed classification model for Wheat Leaf Diseases using fine grained images. In this Unified Matrix-Based Convolutional Neural Network model is introduced. They proposed a matrix based convolution neural network model for classification. This approach shows the new path for the identification of crop disease.

III. MATERIALS AND METHODS

The soybean data set is taken from the UCI machine learning library. The dataset contains 307 samples, 35 attributes, and 19 class labels. Data set is divided into training set and testing set with 80%, 20% and 70%, 30% ratio. The figure 1 shows the different diseases of soybean as class labels in the dataset.



Diaporthe Stem Canker	Charcoal Rot
Rhizoctonia Root rot	Phytophthora rot
Brown Stem rot	Powdery Mildew
Downy Mildew	Brown Spot
Bacterial Blight	Bacterial Pustule
Purple Seed Stain	Anthracnose
Phyllosticta Leaf Spot	Alternarialeaf Spot
Frog Eye Leaf Spot	Diaporthe Pod and Stem Blight
Cyst Nematode	2 4 d Injury

Fig. 1. List of soybean diseases

Attributes plays very important role in the classification process. Based on the value of the attribute class labels are identified. In our dataset attributes are divided into two different categories. They are

Weather parameters: Date, Rainfall, Temperature, and Hail.

Physic crop parameters: leaf mild, crop hist, area damaged, severity, seed tmt, germination, plant growth, leaves, leafspots halo, leafspots marg, leafspot size, plant stand, leaf shread, leaf malf, stem, lodging, stem cankers, canker lesion, fruiting bodies, external decay, mycelium, int discolor, sclerotia, fruit pods, fruit spot, seed, seed-size, mold growth, seed-discolor, shriveling, and roots.

Data set contains some missing values which are filled with mean value of the column. Naive Bayes classifier, KNN, Decision tree, ANN and Bagging machine learning techniques are used for soybean disease classification.

A. Naïve Bayes

In machine learning, naive Bayes classifiers comes under simple probabilistic classifiers category. It is based on pertaining Bayes' theorem with strong (naive) independent assumptions within the features. This method is used to find the class of a given sample. Here the assumption is that every feature is class conditionally independent. This feature simplifies the computation. Naive Bayes works by applying Bayes theorem :

$$P(Classi|X) = \frac{P(X|Classi)P(Classi)}{\sum_{i=1}^{n} P(X|Classi)P(Classi)}. \qquad (1)$$

Here the X represents the sample, which needs to be classified.

B. K Nearest Neighbor

This is the simplest classification algorithm in supervised learning. This method classifies samples based on the majority vote of the class label of the closest training sample in the feature space. KNN gives more correct classification for boundary samples. It is a time-consuming algorithm to train a large volume of data. KNN is well suited for lower dimensional data. The KNN calculates the distance between samples using the following formula:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2)$$

Where x and y are the two sample points.

C. Decision Tree

The frequently used learning algorithm under supervision is decision tree where the split in data samples are done according to a certain parameter. Decision nodes and leaves are the two entities which will help to explain the tree. Decisions or final outcomes are mentioned by leaves of the tree. Data samples split always happens at the decision nodes. The decision tree is used to make a decision at the nodes and reach class label at leaf nodes. Decision rules which is essential for classification are paths from the root to a leaf node. To construct decision tree many algorithms are available in the field of machine learning. Among them ID3 is the best algorithm which stands for Iterative Dichotomiser 3. A decision tree is moving from impurity to purity, and purity can be calculated using entropy and information gain. The Entropy is the measure of impurity in data and information gain measures the relative change in entropy with respect to the independent variables.

D. Neural Network

It is a massively parallel architecture for information processing to generate a knowledge base for the given problem domain which can be used for problem-solving. Neural network is the network of neurons and it consists of set of links used to connect different nodes. Weight is the numerical value associated with each link. The output of each neuron is the weighted sum of inputs. In the case study considered here multi-layer perceptron is used for soybean disease classification. The network comprises of single input, output and hidden layer as shown in figure 2.

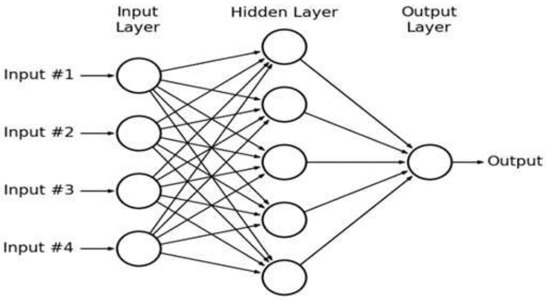


Fig. 2. Multilayer perceptron

E. Bagging

To improve the stability and accuracy of machine learning algorithms used in regression and classification ensemble meta algorithms are designed [5]. Bagging is sub sampling of training data as shown in Figure 3. If the new sample is an outlier or it is on the boundary of the class and it is difficult to classify, ensemble classifier will help to solve this. Bagging always re-samples the training set with replacement, some instances may be left out without using and some represented multiple times. Each member of ensemble is not exposed to the same set of samples, they are different from each other. By collecting the vote of these classifiers, bagging tries to reduce the error due to variance of the base classifier.

TABLE I
ACCURACY OF MACHINE LEARNING CLASSIFIER FOR 70 30 TRAIN TEST SPLIT

Bernoulli	60%
Gaussian	86%
K Nearest Neighbor	79.5%
Decision Tree Using Entropy	86%
Decision Tree Using Gini	81.7%
Multi layer perceptron	85%

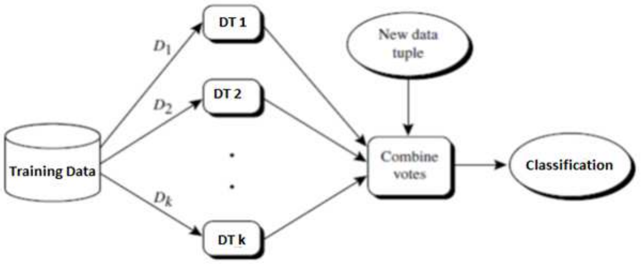


Fig. 3. Ensemble classifier

IV. RESULTS AND DISCUSSION

Preprocessing is done to fill the missing values. Naïve Bayes with Bernoulli distribution and Gaussian Naïve Bayes is used in the classification process which gives less accuracy. N nearest neighbour classififcation technique is used to classify the soybean disease. A decision tree can be created using Gini and entropy impurity method which is used for the comparison. Neural network model called multilayer perceptron method gives maximum accuracy when compared with other classification methods for soybean disease classification. The comparison of all the algorithms with its accuracy is shown below. We have considered the training data of 70% and testing data of 30% in the available storage. TableI shows the comparison of machine learning algorithms with accuracy value.

Figure 4 shows the bar chart depicting the Table I.

We have then considered 80% of data set for training and 20% of data for testing. With this distribution the machine learning algorithms have given following accuracy.

TABLE II
ACCURACY OF MACHINE LEARNING CLASSIFIER FOR 80 20 TRAIN TEST SPLIT

Bernoulli	58%
Gaussian	83.8%
K Nearest Neighbor	83.8%
Decision Tree Using Entropy	80.6%
Decision Tree Using Gini	82%
Multi layer perceptron	87%

Figure 5 shows the chart representation of Table II.

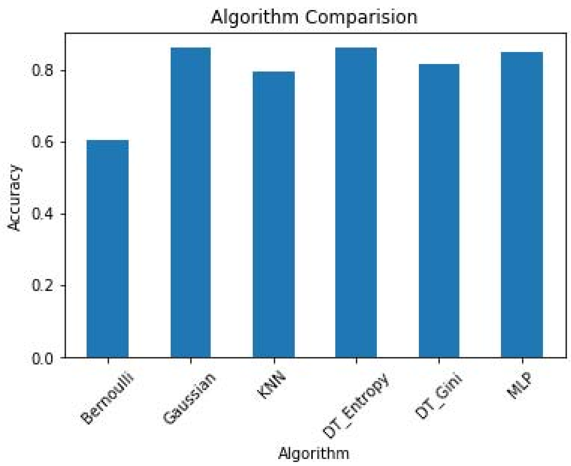


Fig. 4. Algorithm comparison for 70 : 30

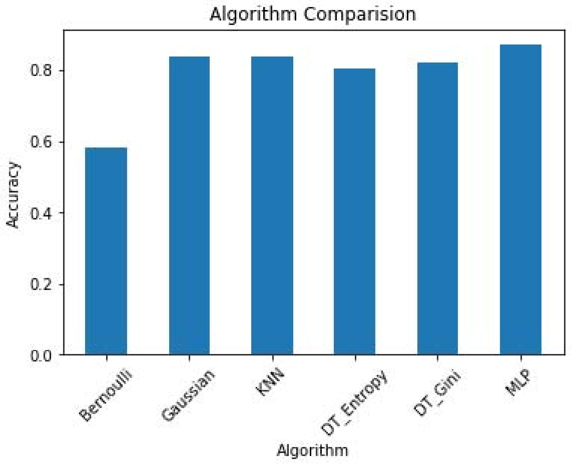
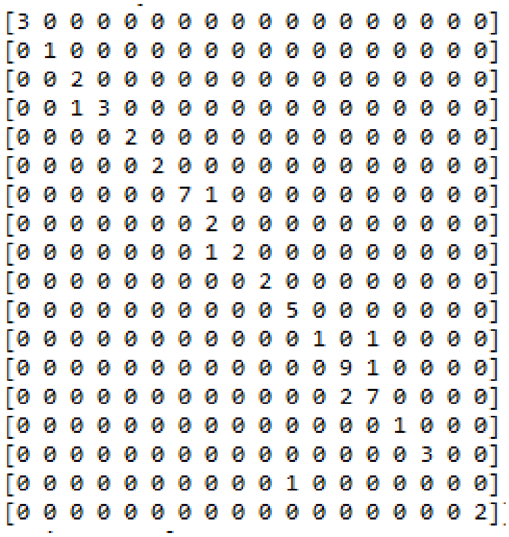


Fig. 5. Algorithm comparison for 80 : 20

Performance of the classification model on a set of test data is described by the table like structure called a confusion matrix. The amount of true and false classifications are encapsulated with count values and is shown by each class. Figure 6 shows the true false matrix designed for neural network model. The significance of this figure is it shows the correct and incorrect classification result, the primary diagonal gives the right classification count.



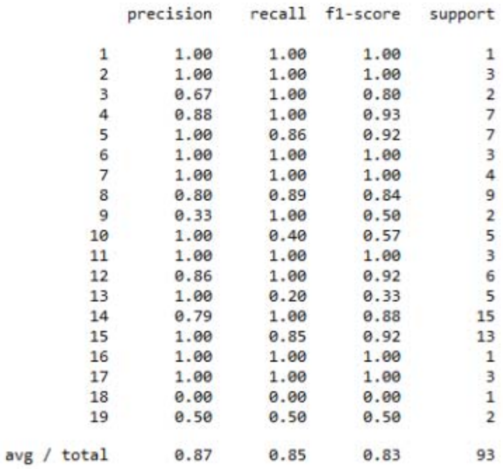
```
[3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 1 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 7 1 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 1 2 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 9 1 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 2 7 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0]
[0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2]
```

Fig. 6. Confusion Matrix for MLP

TABLE III
ACCURACY OF BAGGING CLASSIFIER

No. of trees	No. of Splits	Accuracy
25	50	82.7%
50	100	86%
80	150	88.4%

The figure 7 shows the classification report for multi layer perceptron classifier. Report includes Precision, Recall and F1-Score measures. **Precision** refers to the amount of results in percentage which are applicable. **Recall** refers to the percentage of total applicable results properly classified by given algorithm. **F1 score** is a measure of a accuracy of test. The F1 score is the harmonic average of the recall and precision, where an F1 score with value 1 is considered as best and value 0 is considered as unacceptable.



```
             precision    recall  f1-score   support

          1       1.00      1.00      1.00         1
          2       1.00      1.00      1.00         3
          3       0.67      1.00      0.80         2
          4       0.88      1.00      0.93         7
          5       1.00      0.86      0.92         7
          6       1.00      1.00      1.00         3
          7       1.00      1.00      1.00         4
          8       0.80      0.89      0.84         9
          9       0.33      1.00      0.50         2
         10       1.00      0.40      0.57         5
         11       1.00      1.00      1.00         3
         12       0.86      1.00      0.92         6
         13       1.00      0.20      0.33         5
         14       0.79      1.00      0.88        15
         15       1.00      0.85      0.92        13
         16       1.00      1.00      1.00         1
         17       1.00      1.00      1.00         3
         18       0.00      0.00      0.00         1
         19       0.50      0.50      0.50         2

avg / total       0.87      0.85      0.83        93
```

Fig. 7. Classification Report

In the literature different authors expressed that single classifier is not giving sufficiently accurate results, so we have used Bagging ensemble classifier with a decision tree to classify the diseases. Data set is divided into 80% training set and 20% testing set to train and test the ensemble model.

The figure 8 shows the chart for the data given in Table III.

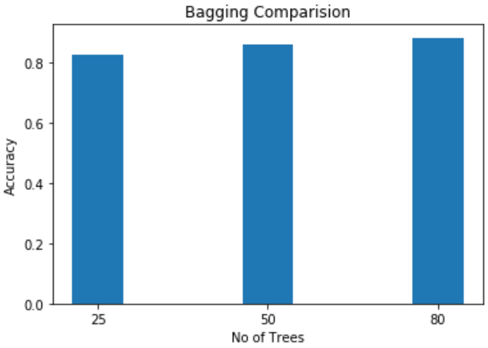


Fig. 8. Bagging algorithm comparison

V. CONCLUSION AND FUTURE SCOPE

The soybean disease classification is done using machine learning algorithms by considering the weather parameters and physic crop properties. Neural network model gives the best accuracy result when compared with other single classifier machine learning algorithms. But when it is compared with ensemble classifier, it gives more accuracy. In the future, deep neural network and network optimization should be used to increase the classification accuracy and to reduce the generalization error. The size of the dataset can be increased using data augmentation to work with deep learning techniques. Feature selection techniques can be used to verify the accuracy.

REFERENCES

[1] P. Louridas and C. Ebert. Machine learning. IEEE Software, 33(5):110–115, Sep 2016.

[2] Agarwal DK, Billore SD, Sharma AN, Dupare BU, Srivastava SK. Soybean: Introduction, improvement, and utilization in India—problems and prospects. Agricultural Research. 2013 Dec 1;2(4):293-300.

[3] J. Chopda, H. Raveshiya, S. Nakum, and V. Nakrani. Cotton crop disease detection using decision tree classifier. In 2018 International Conference on Smart City and Emerging Technology (ICSCET), pages 1–5, Jan 2018.

[4] M. I. AlHajri, N. T. Ali, and R. M. Shubair. Classification of indoor environments for iot applications: A machine learning approach. IEEE Antennas and Wireless Propagation Letters, 17(12):2164–2168, Dec 2018.

[5] Sk Al Zaminur Rahman, Kaushik Chandra Mitra, and S. M. Mohidul Islam. Soil classification using machine learning methods and crop suggestion based on soil series. 2018 21st International Conference of Computer and Information Technology (ICCIT), pages 1–4, 2018.

[6] S. Turgut, M. Dagtekin, and T. Ensari. Microarray breast cancer data classification using machine learning methods. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), pages 1–3, April 2018.

[7] K. R. Aravind, P. Raja, K. V. Mukesh, R. Aniirudh, R. Ashiwin, and C. Szczepanski. Disease classification in maize crop using bag of features and multiclass support vector machine. In 2018 2nd International Conference on Inventive Systems and Control (ICISC), pages 1191–1196, Jan 2018.

[8] Y. Ekanata and I. Budi. Mobile application review classification for the indonesian language using machine learning approach. In 2018 4th International Conference on Computer and Technology Applications (ICCTA), pages 117–121, May 2018.

[9] E. Castelão Tetila, B. Brandoli Machado, N. A. d. S. Belete, D. A. Guimarães, and H. Pistori. Identification of soybean foliar diseases using unmanned aerial vehicle images. IEEE Geoscience and Remote Sensing Letters, 14(12):2190–2194, Dec 2017.

[10] Z. Lin, S. Mu, F. Huang, K. A. Mateen, M. Wang, W. Gao, and J. Jia. A unified matrix-based convolutional neural network for fine-grained image classification of wheat leaf diseases. IEEE Access, 7:11570–11590, 2019.

[11] Endah SN, Sarwoko EA, Sasongko PS, Ulfattah RA, Juwita SR. Attribute Selection for Detection of Soybean Plant Disease and Pests. In2019 3rd International Conference on Informatics and Computational Sciences (ICICoS) 2019 Oct 29 (pp. 1-6). IEEE.