Crop Disease Recognition using Machine Learning Algorithms

Archana Chaudhary Thakur

Abstract*: Classification is a method of observing the features of a new object and assigning it to a known class. Machine learning classification problem consists of known classes and a vivid training set of pre-categorized examples. The work diagnoses groundnut diseases using outstanding machine learning algorithms namely simple logistic, decision tree, random forest and multilayer perceptron for accurate identification of groundnut diseases. Experiments are conducted with the help of 10-fold cross validation strategy. The results advocate that above mentioned classification algorithms diagnose the groundnut diseases with excellent accuracy level. Simple logistic and multilayer perceptron show outstanding performance than other algorithms and result in 96.37% and 95.80% disease classification accuracy. Random forest and decision tree algorithms provide fair accuracies in less time. These machine learning algorithms can be used in diagnosing other crop diseases also.

Keywords: Decision tree, Machine learning, Multilayer perceptron, Oilseed diseases, Simple logistic.

I. INTRODUCTION

Machine learning is a sub-field of Artificial Intelligence that deals with development and study of system that can learn from data. It deals with developing programs that learn from experience and is also a recent field of research. Machine learning methods are widely used as compared to various statistical methods as machine learning methods do not consider basic data assumption. The accurate prediction systems for identifying crop diseases widely use machine learning algorithms [5, 6, 7, 8, 9]. Appropriate recognition of groundnut diseases helps in reducing yield losses and also notifies the agriculturists to initiate competent disease prevention methods. Now day's simple logistic, naive bayes, genetic algorithms, multilayer perceptron are widely used for predictive modeling. Researchers have used machine learning algorithms for diagnosing various crop diseases [3, 4, 19]. The authors have employed logistic model, naïve bayes and random forest machine learning algorithms for recognition of oilseed diseases [3, 4] and the authors have also confirmed the results using UCI standard datasets [12]. Machine learning algorithms are also found useful for forecasting powdery mildew in mangos [15]. A classification problem in machine learning is a supervised learning problem. It consists of observing the input data (training set) and to develop an exact portrayal or model of every class with the help of attributes present in data [13]. The developed model is then used to catalog the test data with unknown class labels. The decision tree, multilayer perceptron and simple logistic machine learning classification algorithms are exercised in different applications [19]. Groundnut is the basis of valuable protein and is used widely in eatables and industries. Groundnuts are rich source of vegetable oil. Additionally groundnuts also contain major quantity of minerals, salts and vitamins. Timely recognition of groundnut diseases plays a vital role in enhancing the productivity and production of good quality crops. Clustering and Rough Set Theory (RST) based methods for leading to description of crop diseases was suggested in [10, 11]. The major focus was to show the appropriateness of the presented algorithms for developing various disease clusters and then to analyze features of a particular disease. The paper is arranged as follows: Section II consists of materials and methods used in the present work. Section III portrays results and discussions. Section IV finally presents the conclusions drawn from the present work.

II. MATERIALS AND METHODS

The real life groundnut disease dataset is used in the present work [4]. The disease classes in the dataset are Alternaria leaf spot, Charcoal rot, Collar rot, Cylindrocladium black rot, Early leaf spot, Fusarium rot, Late leaf spot, Myrothecium leaf blight, Powdery Mildew, Rust, Stem rot, Yellow mold, Zonate leaf spot. The dataset in [4] is enhanced by appending symptoms of 4 more groundnut diseases namely Anthracnose, Rhizoctonia foliar blight, Verticillium wilt and Scab. The new dataset has 17 disease classes and 2022 instances with no missing feature values. The dataset contains all the nominal disease influencing features. There are 26 disease influencing features and one feature as the disease target class representing groundnut diseases.

A. Machine Learning Algorithms

Machine learning algorithms namely decision tree, multilayer perceptron, simple logistic and random forest are used to recognize various groundnut diseases. Experiments are conducted using machine learning suite WEKA [18].

1) Decision Tree

It is a famous machine learning algorithm [17]. It is one of the best prediction ensemble algorithms used in various fields like statistics, data mining and machine learning. Decision trees use the principle of information gain for creating a decision tree from the training set. For each node the algorithm selects a feature which most promisingly partitions the data. The chosen feature has the greatest information gain. The same method is carried out until the entire decision tree is formed, having a representation of all the class examples based upon which test samples are classified.

Revised Manuscript Received on September 05, 2020.
***** Corresponding Author
Archana Chaudhary*, School of Computer Science & IT, Devi Ahilya University, Khandwa Road, Indore 452001, Madhya Pradesh, India. E-mail: archana_scs@yahoo.in

2) Multilayer Perceptron

It is one of the best machine learning algorithms. Feed forward neural network consists of one or more hidden layers which transforms a set of inputs into suitable outputs. The neural network contains three layers namely input layer, one or more hidden layer(s) and an output layer. A neuron or a perceptron signifies a node in the network. The network connects each neuron (node) of one layer to every other neuron (node) of adjacent layer(s). The training or test vectors are linked to the input layer and are also assessed by the hidden and output layers respectively [19].

3) Simple logistic

It is a regression model that is exercised for classification purposes [14]. It is an outstanding machine learning algorithm. Assuming that class variables are of binary types, the probability of instance membership in a class type employing simple logistic algorithm is given by p1(x), as in Eqn. (1) for the binary classification problem where α and β are the variables of the model as shown below –

$$p1(x) = \frac{e^{\alpha+\beta' x}}{1+e^{\alpha+\beta' x}} \quad (1)$$

4) Random forest

It is a popular ensemble machine learning algorithm. It is helpful in various classification problems [4]. It has multiple random trees [2]. The ensemble is useful for big datasets. It uses arbitrary samples to create each tree in the forest. The algorithm combines multiple random trees which vote/output on a particular outcome. In the algorithm each vote is given equal weight. The ensemble selects the classification output that has maximum votes.

B. Performance Assessment Measures

The performance assessment measures considered in the present work are classification accuracy, f- measure, recall, precision, and model build time (in seconds). The measures are expressed using the following formulations [1, 16] as under –

$$Classification\ Accuracy = \frac{T_P + T_N}{F_P + F_N + T_P + T_N} \times 100 \quad (2)$$

$$Recall = \frac{T_P}{F_N + T_P} \quad (3)$$

$$Precision = \frac{T_P}{F_P + T_P} \quad (4)$$

$$F - Measure = 2 \times \frac{recall \times precision}{recall + precision} \quad (5)$$

In Eqns. (2) to (4) T_P, T_N, F_P and F_N represent True Positive, True Negative, False Positive and False Negative respectively.

III. RESULTS AND DISCUSSION

Ten fold cross validation strategy is used in the present work for experiment conduction. Experiments are conducted using 34% testing samples and 66% training data related to groundnut diseases. Performance assessment of machine learning algorithms on real life groundnut disease dataset with respect to the performance indicators – disease classification accuracy and model build time is shown in Table – I. It is clear from Table – I that simple logistic attains the greatest classification accuracy as 96.37%. Next outstanding performance is shown by multilayer perceptron with classification accuracy 95.80%. It is a vital observation that all the classification algorithms considered in the present work, result in significant disease classification accuracy. Simple logistic and multilayer perceptron outperform other classification algorithms.

Table–I: Comparison of machine learning algorithms

Algorithm	Disease classification accuracy (in %)	Model Build time (in sec.)
Simple logistic	96.37	0.65
Multilayer Perceptron	95.80	3.08
Random forest	94.87	0.03
Decision tree	91.65	0.05

Performance assessment of classification algorithms for performance measures namely precision, f-measure and recall is shown in Fig. 1. It is evident from Fig.1 that for above mentioned performance measures simple logistic and multilayer perceptron show outstanding performance as compared to the other algorithms.

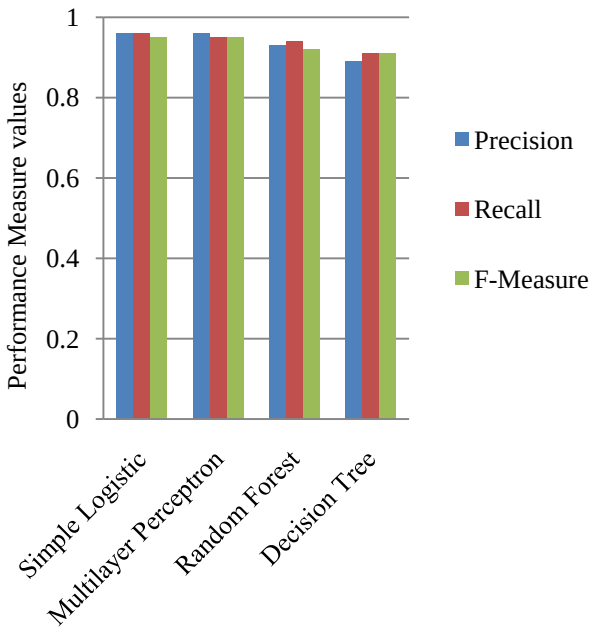

Fig. 1. Performance assessment of machine learning algorithms.

But one major limitation of multilayer perceptron is that the model build time is significantly high as compared to the other algorithms. Classification using decision tree results in a tree-like structure which permits greater user understanding. It is a vital observation that both random forest and decision tree algorithms have very little model build time as compared to the other algorithms. These results can easily act as important milestones for experts and non experts working in the domain of machine learning.

These algorithms may help one to analyze a particular disease and depict its influences and can get a good indicator of significant variables.

IV. CONCLUSIONS

Simple logistic, multilayer perceptron, decision tree and random forest machine learning algorithms were exercised in the present work for recognition of groundnut diseases. Experiments were conducted using a 10- fold cross validation strategy. The results confirm that all these classification algorithms recognize the groundnut diseases at a remarkable accuracy level. Simple logistic and multilayer perceptron show better performance as compared to other algorithms and yield 96.37% and 95.80% accuracy. Random forest and decision tree have little model build time as compared to the other machine learning algorithms.

REFERENCES

1. A.T. Azar, H.I. Elshazly, A.E. Hassanien, A.M. Elkorany, "A random forest classifier for lymph diseases", Comput. Meth. Programs Biomed, vol. 113, no. 2, (2014), pp. 465– 473.
2. L. Breiman, "Random Forests", (2001).
3. A. Chaudhary, S. Kolhe, R. Kamal, "A hybrid ensemble for classification in multiclass datasets: An application to oilseed disease dataset", Computers and Electronics in Agriculture, vol. 124, (2016), pp. 65–72.
4. A. Chaudhary, S. Kolhe, R. Kamal, "An improved random forest classifier for multi-class classification", Information Processing in Agriculture, (2016), pp. 215 – 222.
5. A. Chaudhary, S. Kolhe, R. Kamal, "Machine learning techniques for mobile intelligent systems: A study", In IEEE Ninth International Conference on Wireless and Optical Communications Networks (WOCN), (2012), pp. 1–55.
6. A. Chaudhary, S. Kolhe, R. Kamal, "Machine Learning Classification Techniques: A Comparative Study", International Journal on Advanced Computer Theory and Engineering, vol. 2, no. 4, (2013), pp. 21–25.
7. A. Chaudhary, S. Kolhe, R. Kamal, "Machine Learning Techniques for Mobile Devices: A Review", International journal of Engineering Research and Applications, vol. 3, no. 6, (2013), pp. 913–917.
8. A. Chaudhary, S. Kolhe, R. Kamal, "Performance Examination of Feature Selection methods with Machine learning classifiers on mobile devices", International Journal of Engineering Research and Applications, vol. 3, no. 6, (2013), pp. 587–594.
9. A. Thakur, R. Thakur, "Machine Learning Algorithms for Intelligent Mobile Systems", International Journal of Computer Sciences and Engineering, vol. 6, no. 6, (2018), pp. 1257–1261.
10. A. Arora, S. Upadhyaya, R. Jain, "Reduct based variable selection and knowledge discovery for disease clusters", Soybean Research, vol. 6, (2008), pp. 56–67.
11. A. Arora, S. Upadhyaya, R. Jain, "Post processing of clusters for pattern discovery: Rough set approach", Journal of the Indian Society of Agricultural Statistics, vol. 63, no. 2, (2009), pp.181–88.
12. University of California Irvine. UCI Machine Learning repository. Link: http://archive.ics.uci.edu/ml/.2010.
13. J. Han, M. Kamber, "Data Mining: Concepts and Techniques, Morgan Kaufmann", (2000).
14. T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning: Prediction, Inference and Data Mining", 2nd edition, Springer-Verlag, New York, (2009).
15. R. Jain, M. Sonajharia, R.V. Subramanian, "Machine learning for forewarning crop diseases", Journal of the Indian society of Agricultural Statistics, vol. 63, no. 1, (2009), pp. 97–107.
16. A. Özçift, "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis", Comput. Biol. Medicine, vol. 41, no. 5, (2011), pp. 265–271.
17. J.R. Quinlan, "Improved use of continuous attributes in c4.5", Journal of Artificial Intelligence Research, vol. 4, (1996), pp. 77–90.
18. I.H. Witten, E. Frank, "Data Mining: Practical machine learning tools and techniques", Second ed. Morgan Kaufman series in Data Management Systems, San Francisco, CA, USA, (2005).
19. S.H. Zak, "Systems and Control", Oxford University Press, New York, (2003).

AUTHORS PROFILE

Archana Chaudhary (Thakur) received her doctorate in Computer Science from Devi Ahilya University, Indore, India in 2016. She is working as an Assistant Professor at School of Computer Science & IT, Devi Ahilya University, Indore. She is currently guiding many doctoral scholars. She has published numerous research papers in different national and international journals. She has authored various papers in Elsevier journals. She has also reviewed Elsevier journal manuscripts. She has actively participated in different international and national conferences. She has also been esteemed speaker in IEEE international conferences. She has also chaired sessions in IEEE international conferences. Her research areas include Artificial Intelligence, Machine learning, Data Mining and Soft Computing.