Machine learning in plant disease research
Xin Yang\textsuperscript{1*}, Tingwei Guo\textsuperscript{2}

\textsuperscript{1} Department of Information Systems, Statistics, and Management Science, Culverhouse College of Commerce and Business Administration, The University of Alabama, Tuscaloosa, AL, USA 35487
\textsuperscript{2} Department of Biological Sciences, The University of Alabama, Tuscaloosa, AL, USA 35487

Abstract: Plants are constantly exposure to pathogens such as virus, bacteria and fungi. Plant diseases caused by pathogens lead significant crop yield loss globally. Numerous researchers have been studying how to reduce the damage of plant diseases. Some researchers investigated the resistance genes in plants and attempt to enhance plants’ resistance against pathogens. Meanwhile, some other researchers developed identification and scoring system to monitor and predict plant diseases based on leaves images. The purpose of this review is to present the application of machine learning in plant resistance genes discovery and plant diseases classification.

Key words: Machine learning; plant diseases; algorithm.

Correspondence to: Xyang15@cba.ua.edu

Received: February 16 2017; Accepted: March 24 2017; Published: March 31 2017

doi: http://dx.doi.org/10.18088/ejbmr.3.1.2016.pp6-9

Introduction

Over the past decades, continuous studies have been performed to reveal the interactions between plant immune respond and pathogens. Large amount of data has been generated from those researches due to the tremendous advances in genomics and proteomics. Previously, scientists typically applied large-scale genetic screening and genomic approaches to identify genes and proteins of interest \textsuperscript{(1,2)}. But now, the development of machine learning algorithms, which are a collection of analytic methods that automate model building process and iteratively learn from data to gain insights without explicitly programming, provides more powerful and efficient tools to not only identify genes/proteins involved in plant-pathogen interactions, but also classified plant diseases from images of infected leaves. We here present a review of studies that utilize machine learning regarding the plant-pathogen interactions and plant disease identifications.

Machine learning techniques

Compared to statistical models, machine learning methods focus on data themselves and emphasize the performance of certain tasks. Machine learning can be applied to four areas based on the problems to be solved: 1) identification/detection 2) classification; 3) quantification; 4) prediction. On the other hand, based on whether outcomes are labeled, machine learning is divided into two categories: 1) supervised learning and 2) unsupervised learning. In the recent decade, machine learning has been used in various disciplines, such as computing, bioinformatics, marketing, medical diagnosis, game playing, etc. In the healthcare industry, machine learning algorithms are also frequently used for predicting healthcare cost and risks of diseases based on electronic health records (EHR) \textsuperscript{(3-5)}. A couple of machine learning algorithms are frequently used in researches, such as Naïve Bayes Classifier, K Means clustering, support vector machine (SVM), artificial neural networks (ANN), decision trees and random forests. A typical process of employing machine learning includes data collection, dataset preparation, feature extraction, preprocessing, feature selection, choosing and applying machine learning algorithms and performance evaluation. Up to now, machine learning methods are mainly applied to molecular biology and agriculture related to plant diseases.

Machine learning and predictions of plant or pathogen-related molecules

A complex cascade of defense responses is induced during plant-pathogen interaction via invading signals from invaders and/or plant themselves, and such signals could be detected by plants immune systems through different mechanisms. As a result, various defensive reactions, such as production of reactive oxygen species(ROS), reinforcement of plant cell wall, and synthesis of defense enzymes, will be activated and initiated through different signal transduction pathways \textsuperscript{(6,7)}. Due to the large number of plant resistance genes families, a high-throughput method is needed to identify genes involved in the resistance to pathogens. Although machine learning techniques have been used for various subjects, only a few studies have been conducted to predict plant pathogens related genes/proteins. One of the major interests in plant-pathogen interactions is to identify the plant resistance genes.

A good example of applying machine learning in plant-pathogen interaction research by Pal et al. showed that supporting vector machine (SVM), which was used to predict plant resistance proteins (R proteins) based on 10270 features extracted from amino acids sequences of proteins, achieved an accuracy of 91.11% on the test datasets \textsuperscript{(8)}. Many researchers focused on the virulent proteins...
from the pathogen, for example, the effectors that can target plant molecular components and facilitate pathogen infections. We found more studies using machine learning to predict the pathogen effectors than plant resistance genes/proteins. For instance, Sperschneider et al. compared the results of fungal effector predictions from several machine learning algorithms, including Naïve Bayes, Naïve Bayes-K, logistic regression, multilayer perceptron, C4.5 decision tree and random forest. Among the different machine learning methods, Naïve Bayes achieved highest performance of prediction, with sensitivity of 0.78, specificity of 0.76 and area under curve (AUC) of 0.85(9). Later, Sperschneider et al. developed a tool (LOCALIZER) based on SVM to predict effector subcellular localizations, which provide critical clues about the functions of effectors in plant cells. Impressively, LOCALIZER showed accuracy of 91.4% for chloroplast localization, 91.7% for mitochondria localization, and 73% for nucleus localization (7). Several other studies that employed machine learning, also obtained very good results. For example, Saunders et al. identified candidate effectors from rust fungi via Markov clustering (6) and Wang et al. predicted bacterial type 3 secreted effectors via SVM based on position-specific amino acid composition features. The SVM model showed a sensitivity of 90.97% and a specificity of 97.42% evaluated by 5-fold cross-validation (10). A summary of information is presented in table 1.

Machine learning is promising in the several ways to discover new insights and knowledge in molecules involved in plant-pathogen interactions. In addition to plant resistance genes/proteins, plant susceptible genes/proteins identifications are often ignored but are also important. Meanwhile, protein partners that physically interact with resistance genes are critical to elucidate the plant defense pathways and should be investigated using machine learning methods. Another good question for pathologists is whether a specific pathogens effector target a variety of plant molecules. Machine learning tools can be potentially used to predict and understand how different plants genes/proteins response to pathogen effectors.

### Table 1. Examples in machine learning in prediction of plant-pathogen interactions.

<table>
<thead>
<tr>
<th>Machine learning algorithm</th>
<th>Purpose</th>
<th>Accuracy</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>Predict plant R proteins based on amino acids sequence</td>
<td>91.11%</td>
<td>(8)</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>Fungal effectors prediction</td>
<td>sensitivity 0.78; specificity 0.76; AUC of 0.85</td>
<td>(9)</td>
</tr>
<tr>
<td>SVM</td>
<td>LOCALIZER for predicting effector subcellular localization in plant cells</td>
<td>73-91.7%</td>
<td>(7)</td>
</tr>
<tr>
<td>Markov clustering</td>
<td>Identify candidate effectors</td>
<td>sensitivity 90.97%; specificity 97.42%</td>
<td>(6) (10)</td>
</tr>
<tr>
<td>SVM</td>
<td>Prediction of type 3 effectors</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Another important application of machine learning is classification, which classify plant diseases into different stages or different types. The machine learning algorithms were intensively applied on the classification of powdery mildew, a fungal pathogen that causes yield loss in varieties of crops (18). Raza et al. extracted both local and global statistics from thermal and visible light image data and used SVM to identify the powdery mildew-inoculated tomato leaves. They showed that the machine learning system was able to identify the tomato leaves infected naturally by powdery mildew (19). An integrated non-supervised learning algorithms with Bayesian classifiers was presented by Hernandez-Rabadan et al. to separate healthy and infected tomatoes in a uncontrolled environment (20). Mokhtar et al. reported a machine learning approach using SVM with different kernel functions to differentiate two tomato’s viruses and the accuracy was as high as 92% based on quadratic kernel function (21). In addition to powdery mildew, other type of classifications was also implemented by machine learning algorithm. In the study by Wahabzada et al., authors proposed methods combining Bayes factors, interpretable matrix factorization and Di-
richlet aggregation regression to map three plant disease progress (Pyrenophora teres, Puccinia hordei and Blumearia graminis hordei) (22). Besides crops, machine learning has also been applied to study interactions between the model plant Arabidopsis and pathogens. Machine learning algorithms such as SVM, Bayesian classifier and random forest were utilized to identify bacterial pathogens with high prediction precisions (23-25). By using deep convolutional neural networks (CNN), which is the latest generation of machine learning methods, Sladojevic et al. has successfully recognized 13 different plant pathogens and achieved precision between 91% and 98% (26). Furthermore, another group also applied deep CNN to classify 14 crop species and 26 diseases with best F1 score of 0.9934 (27).

So far, only a few studies have been done to predict the disease development onset (28) and quantification of plant diseases stress (29,30). However, prediction and quantification of plant diseases are potentially more important than identification and classification of diseases in the future due to the implications to precision agriculture. Studies of such could lead to preventing crop diseases at early stage and cutting cost for pesticides.

In conclusion, in the big data era, machine learning provides a powerful tool to analyze tremendous amount of data. Careful selection of pre-processing data methods and machine learning tools is critical to obtain highest accuracy of classification. Meanwhile, compared to traditional methods of identifying genes involved plant pathogen interactions, methods integrating machine learning approaches are relatively scarce in the literature. Thus more machine learning based tools are needed to predict important plant resistance genes, as well as make contribution to the agriculture. With aerial imaging platforms and sensor technology, collecting field data becomes easier and more precise, which is critical for improving machine learning accuracy. More sophisticated methods such as deep learning algorithms will be applied in detecting plant diseases and discovering plant resistance genes.

References
12. Emmi L, Gonzalez-de-Soto M, and Gonzalez-de-Santos P. Configuring a fleet of ground robots for agricultural tasks. ROBOT2013: First Iberian Robotics Conference. 2014: 505-517
20. Hernández-Rabadán D L, Ramos-Quintana F, and Guerrero Juk J. Integrating SOMs and a Bayesian classifier for segmenting diseased

Table 2. Examples of Machine learning in plant diseases detection.

<table>
<thead>
<tr>
<th>Machine learning algorithm</th>
<th>Purpose</th>
<th>Accuracy</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>Detect HLB</td>
<td>61-90%</td>
<td>(14)</td>
</tr>
<tr>
<td>SVM, ANN</td>
<td>Differentiate HLB from Zinc deficiency</td>
<td>92.8% for SVM and 92.2% for ANN</td>
<td>(15)</td>
</tr>
<tr>
<td>SVM with kernel</td>
<td>Detect HLB on citrus trees</td>
<td>67-85%</td>
<td>(16)</td>
</tr>
<tr>
<td>QDA</td>
<td>Detect citrus HLB</td>
<td>95%</td>
<td>(17)</td>
</tr>
<tr>
<td>SVM</td>
<td>Detect powdery mildew</td>
<td>75%</td>
<td>(19)</td>
</tr>
<tr>
<td>SVM</td>
<td>Detect tomato viruses</td>
<td>92%</td>
<td>(21)</td>
</tr>
</tbody>
</table>