A Novel ML Approach for Plant Disease Identification

ANDE DEVIKA [#1] & D.D.D.SURIBABU [#2] & V.SARALA [#3]

[#1] M.Sc Student, Master of Computer Science, D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

[#2] Head & Associate Professor, Dept of CSE, D.N.R. College of Engineering, Bhimavaram, AP, India.

[#3] Assistant Professor, Master of Computer Science, D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

Received: January 12, 2019 Accepted: February 21, 2019

ABSTRACT: *Machine Learning (ML) field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. Identification of the plant diseases is the key to prevent the losses in the yield and quantity of the agricultural product. The studies of the plant diseases mean the study of visually observed patterns seen on the plant. Health monitoring and disease detection on the plant is very critical for the substantial growth. It is very difficult to identify the diseases on the plant manually and provide the treatment for that appropriate disease. It requires a tremendous amount of work experience and should be expertise in the plant diseases and also requires excessive time for processing. Here in this proposed application we try to find out the disease of the plant based on the inputs which we observe physically on any plant. For any plant there are 5 levels of disease occurrences: Stem Level, Leaves Level, Seed Level,Lessions Level, Plant Level. So any disease on the plant can be either of these five levels. We try to design an medical dictionary in which all the physical inputs are substituted according to any of the level and then try to detect which disease plant is suffered with and it will try to provide cure for that appropriate disease. Our evaluation results on the proposed method using ML approach for identifying diseases on plant able to identify the diseases accurately and try to provide a solution for the end users.*

Key Words*: Pattern Matching, Lesions Level, Health Monitoring System, Disease Detection, Momentum.*

I. INTRODUCTION

Data Mining is the process of extracting the useful data from a large data collection into a useful manner. In the process of data mining there are many algorithms which are used for clustering, classification, rule mining and visualization. Of all classification is a main job for in allocation of objects with a set of attribute like, Symptoms, diseases, or by any other. In the process of classification, each and every individual algorithm takes multiple instances (training data) and predicts which of several classes each instance belongs to. Each instance consists of several attribute also known as symptoms, through which each of which takes on one of many possible values. The attribute consists of several predictors attributes and one target attributes. Each of the target attribute's possible values is a class to be predicted on the basis of that case's predictor attribute values. Regression takes numerical dataset and constructs a mathematical formula that fits the data[1].

Although we can able to define the rule that are important for diagnosing a disease in the Natural plants and Garlic plants using rule based and machine learning expert system[2]. Here, the rules and rule combinations are prepared according to the data given by the subject experts and stored in the database in a table format. Here machine learning algorithm is applied to get better optimization results in the present Natural expert system[3] . This paper focuses on the optimization algorithm which gives higher searching efficiency, better optimized and high quality results. The present application is consisting of Machine learning system[4].

Main Tasks and Input Data Sets

The two main important tasks that are undertaken in this paper provide the basis for the design of an information technology framework that is capable to identify and disseminate healthcare information. The first task is to gather all the sensitive attributes related to plant disease in terms of symptoms and then try to add those symptoms in one database.In the second level try to add the levels of plant which cause diseases[5]. The first task (task 1 or sentence selection) is mainly gathered from various nurseries for collecting the collaborative information about various plant and their disease and treatments. The second task (task 2 or relation identification) has a deeper semantic dimension and it is focused on identifying disease-treatment relations in the sentences already selected as being informative (e.g., task 1 is applied first). We focus on two relations: Cure, Prevent with the solution for that disease[6].

II. Related Work

In this section we mainly try to discuss about the background work that is carried out in order to identify the plant diseases and their prevention and cure for their diseases.

Motivation

As we try to use classification algorithms, Normally there are six representative models for classification: decision-based models (Decision trees), probabilistic models (Naı¨ve Bayes (NB) and Complement Naı¨ve Bayes (CNB), which is adapted for text with imbalanced class distribution), adaptive learning (Ada-Boost), a linear classifier (support vector machine (SVM) with polynomial kernel), and a classifier that always predicts the majority class in the training data (used as abaseline). In this proposed thesis we try to use Naı¨ve Bayes classifiers because this is the best learning algorithms in the literature and were shown to work well on both short and long texts. Also we try to apply the Decision trees are decision-based models similar to the rule-based models that are used in handcrafted systems, and are suitable for short texts. Probabilistic models, especially the ones based on the Naı¨ve Bayes theory, are the state of the art in text classification and in almost any automatic text classification task[7]. Adaptive learning algorithms are the ones that focus on hard-to-learn concepts, usually underrepresented in the data, a characteristic that appears in our short texts and imbalanced data sets. SVM-based models are acknowledged state-of-the-art classification techniques on text. All classifiers are part of a tool called Weka.9 One can imagine the steps of processing the data (in our case textual information—sentences) for ML algorithms as the steps required to obtain a database table that contains as many columns as the number of features selected to represent the data, and as many rows as the number of data points from the collection[8].

III. Proposed Bag of Words (BoW) Representation for Classifying the Diseases Based on Individual Symptoms

In this section we will find out the BoW model to identify and classify the diseases based on individual symptoms.

Scope

The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable words as features. After the feature space is identified, each training and test instance is mapped to this feature representation by giving values to each feature for a certain instance. Two most common feature value representations for BOW representation[9] are: binary feature values—the value of a feature can be either 0 or 1, where 1 represents the fact that the feature is present in the instance and 0 otherwise; or frequency feature values—the value of the feature is the number of times it appears in an instance, or 0 if it did not appear. Here we try to collect all the plant diseases and place those things into the BoW,which is also act like a Support vector Machine for identifying the plant diseases.

Classification Algorithms

As classification algorithms, we use a 2 models:

1) probabilistic models (Naı¨ve Bayes (NB) for classifying the type of disease the plant is affected with &
2) a linear classifier (support vector machine (SVM) with polynomial kernel), which holds all the disease data set into it and try to provide support for the NB classifier algorithm in order to predict the disease symptoms.

One can imagine the steps of processing the data (in our case textual information—sentences) for ML algorithms as the steps required to obtain a database table that contains as many columns as the number of features selected to represent the data, and as many rows as the number of data points from the collection.

The two tasks that are undertaken in this paper provide the basis for the design of an information technology framework that is capable to identify and disseminate plant diseases. The first task identifies and extracts informative sentences on diseases and treatments topics, while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between diseases and cure for that plant[10].

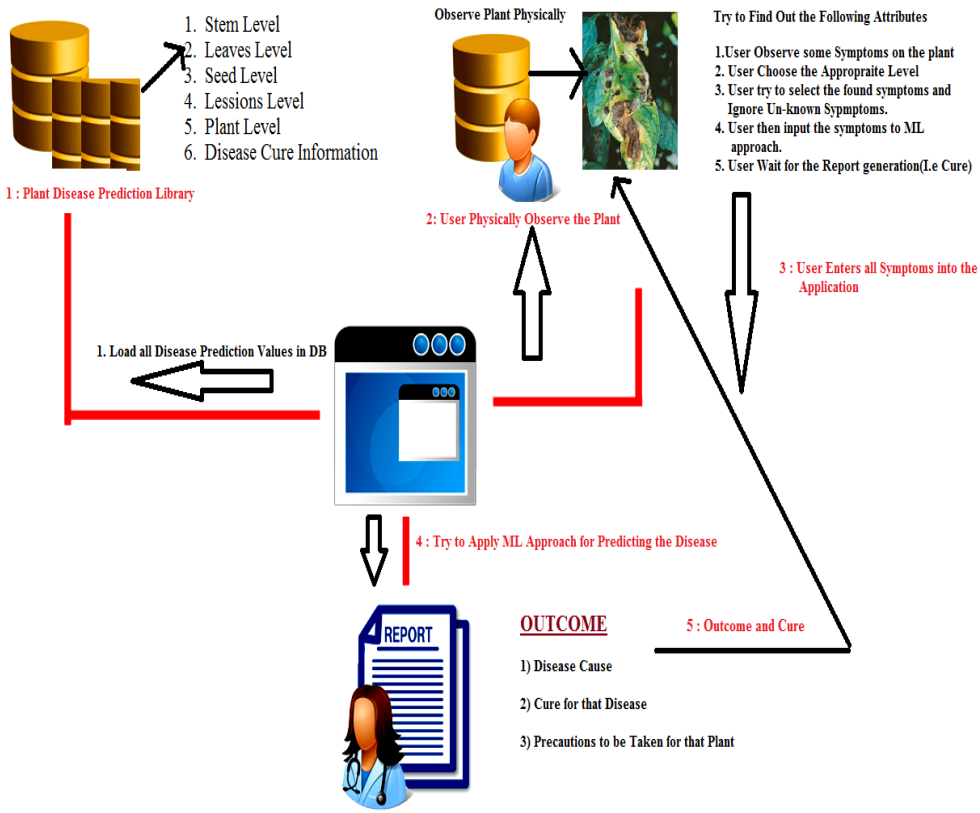

Figure 1.Represents the Proposed Method to Identify the Diseases in a Plant

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

- Consider a fictional dataset that describes the disease of a plant based on the symptoms of the given inputs. Each tuple classifies as Disease found("YES") and Disease not found ("NO").
- Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

  Where A and B are events

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

IV. Implementation Modules

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. We have implemented the proposed concept on Java programming language with JEE as the chosen language in order to show the performance this proposed novel IPath protocol. The front end of the application takes JSP, HTML and Java Beans and as a Back-End Data base we took My-SQL Server. The application is divided mainly into following 5 modules. They are as follows:

1. Load Disease Data Set
2. User Module
3. Machine Learning Technique
4. Identify Disease using Rule Based Expert System
5. Identify Disease using Machine Learning Algorithm

Now let us discuss about each and every individual module in detail as follows:

1. Load Data Set Module

This is a predefined task which is done by the administrator in order to maintain a proper data set to classify the plant diseases. Whenever a new plant disease is invented, then that disease details should be maintained into the database by the administrator. This is nothing but collecting training data set information for the naïve bayes classification algorithm.

2. User Module

In this module the user is one who try to verify the diseases on a plant. he try to identify all the infected areas of that plant physically and then try to substitute those symptoms on that plant library. Here those symptoms which he found will be choose as yes and remaining those symptoms which he didn't observe on that plant will be selected as No. So once after choosing all the values the corresponding inputs is send to the ML approach. Here the user try to give test inputs for the Naïve Bayes classification algorithm. These test values should be matched with training data set and then probability of disease is identified by the ML approach.

3. Machine Learning Technique

In this module, the predefined disease data sets and user inputs are to be learned by the machine (computer). Machine learning is the study of how to make computers learn; the goal is to make computers improve their performance through experience. This ML is mainly used in clustering the diseases based on the type of symptoms. As we all know that all plants may not suffer with same type of disease and same level of complaints. So based on the individual problem ,the ML system need to guide the user to take cure on those conditions.

4. Identify Disease using Rule Based Expert System

Here the diseases can be identified based on rule based expert system which means user need to input the symptoms based on the one of the 5 levels and after the user choose the level and enters all the symptoms then only the user will get the appropriate disease name based on expert knowledge. Here if the user choose all symptoms as null, then the resultant output will be displayed as no disease for that plant. If the same user try to choose appropriate inputs then based on that the disease will be predicted and cure will be provided for that user.

5. Identify Disease using Machine Learning Algorithm

Here this module clearly tells that there is no need to choose individual category and then input the symptoms. If the user find some common symptoms which can be generally visualized and identified. Those symptoms he try to choose from this common list of attributes and based on those fields ,the ML approach will decide which type of disease the plant suffer from and how much percentage of infection occurred to the plant. We can calculate the percentage of infection on that plant.

V.Conclusion

In this paper, we for the first time have designed a novel method to find out the disease of the plant based on the inputs which we observe physically on any plant. For any plant there are 5 levels of disease occurrences. We try to design a medical dictionary in which all the physical inputs are substituted according to any of the level and then try to detect which disease plant is suffered with and it will try to provide cure for that appropriate disease. Our evaluation results on the proposed method using ML approach for identifying diseases on plant clearly demonstrate that this is best to identify the diseases accurately and try to provide a solution for the end users.

VI. References

1. R. Bunescu and R. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction," Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/ EMNLP), pp. 724-731, 2005.
2. R. Bunescu, R. Mooney, Y. Weiss, B. Scho¨ lkopf, and J. Platt, "Subsequence Kernels for Relation Extraction," Advances in Neural Information Processing Systems, vol. 18, pp. 171-178, 2006.
3. A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju, "Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage," Proc. 13th Text Retrieval Conf.(TREC), 2004.
4. M. Craven, "Learning to Extract Relations from Medline," Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.
5. I. Donaldson et al., "PreBIND and Textomy: Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine," BMC Bioinformatics, vol. 4, 2003.
6. C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky,"GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles," Bioinformatics, vol. 17, pp. S74-S82, 2001.

7. O. Frunza and D. Inkpen, "Textual Information in Predicting Functional Properties of the Genes," Proc. Workshop Current Trends in Biomedical Natural Language Processing (BioNLP) in conjunction with Assoc. for Computational Linguistics (ACL '08), 2008.
8. R. Gaizauskas, G. Demetriou, P.J. Artymiuk, and P. Willett, "Protein Structures and Information Extraction from Biological Texts: The PASTA System," Bioinformatics, vol. 19, no. 1, pp. 135-143, 2003.
9. C. Giuliano, L. Alberto, and R. Lorenza, "Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature," Proc. 11th Conf. European Chapter of the Assoc. for Computational Linguistics, 2006.
10. J. Ginsberg, H. Mohebbi Matthew, S.P. Rajan, B. Lynnette, S.S. Mark, and L. Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data," Nature, vol. 457, pp. 1012-1014, Feb. 2009.