Plant disease detection using Multi-class Support Vector Machine

Dhanush Dinesh[1*]

[1*]Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, 24060, Virginia, USA.

Corresponding author(s). E-mail(s): dhanushnd@vt.edu;

Abstract

A large number of crops are lost every year due to crop diseases. It is essential to prevent such crop loss for the food security of the world. In the project computer vision as well as machine learning techniques are used to identify and classify the diseases. We use the credible data-set of "plant village" for training and testing of our model. The approach used in this work is to extract the features, classify using multi-SVM technique and visualization the outcomes using PCA and t-SNE. The approach is towards automating plant disease diagnosis on a large scale.

Keywords: Leaf Detection, Image processing, SVM, ML, segmentation, feature extraction, agriculture, crop disease detection, PCA, t-SNE

1 Introduction

Agriculture employs the largest number of people in India about 42.6 % [1] of the workforce but contributes only 20.2 % [2] of the GDP. The main reason for crop loss is pests and diseases, around 35 % of crop is lost due to diseases. We can mitigate crop loss by detecting diseases in plants at an early stage and notifying the farmer to take corrective measures. The proposed project tries to classify the leaves and try to alert the farmer to take corrective measures to stop the spread of the disease.

If human interaction is removed or limited. The automation will accelerate the detection of disease in the plants in advance and make it easy for the

farmer to start the treatment at an early stage hence reducing the severity of the disease.

2 Specific Aim

The main aim of this project is to classify/ detect if a given leaf sample has been infected with a disease (if so which disease) or it's a healthy leaf sample. The technique used to determine the classification is a multi-class support vector machine (MSVM). The impact of the different algorithms used to perform the classification will be evaluated and compared.

3 Background

Agriculture is the backbone of any civilization. It is very important in this age of population explosion to make the agriculture sector more productive. There have been a lot of advances in the field of image processing and image detection. There have been many research papers written on the classification of leaves. Some of the commonly used techniques used to classify the images are Logical regression, KNN, SVM and CNNs. In this work, we try to implement multi-SVM for classifying the leaves.

3.1 Literature review

In this paper [3] the author Yang, Yan et al. discusses improving SVM classifier with prior knowledge in microcalcification they use two techniques to incorporate rotation invariant a prior into the SVM classifier first one being virtual support SVM and the second one being the tangent vector SVM. It is proven that tangent vector SVM works well for applications with a low false-positive rate

In this paper [4] the binary and multi class SVMs are compared. The author describes how a binary classification requires a series of parameter optimization where as a multi class SVM required only a small training set for accurate classification.

In this [5] paper Pushkar Sharma, et al. compares different types of models on plant leaf such as Logical regression, KNN, SVM, CNN and concludes that the CNNs have best accuracy. Although the previous paper concluded the CNNs were a perfect fit for leaf detection this paper [6] talks about the comparison between CNNs and Multi-SVMs. Jenifa, A. et al. appreciate how multi-SVMs are global, unique and are less prone to overfitting compared to CNNs which suffers from multiple local minima.

In this paper [7] Wiwart et al. try to classify the image based on the number of minerals present in the leaves which intern determine the color of the leaves. The RGB image is converted into HSI and L*a*b* color spaces. The differences are quantified by calculating the Euclidean distances in both the image spaces and the images are classified.

4 Research design and Methods

In the project we treat leaf detection as a supervised machine learning problem. This requires the data set to be classified into two sets as shown in Figure 1. The training and testing data-set are randomly chosen from the data-set to make the model more reliable and more accurate.

To extract the maximum information from the leaf image we first run the segmentation on the leaf images as shown in Figure 1. As we are only interested in the diseased part of the leaves, we can mask the background as well as the green region of the leaf. Thereby we can obtain the region of interest (RoI). For training the model we will use a Multi-class Support Vector Machine (MSVM) algorithm as implemented in [4]. By running the data-set through the MSVM we will classify the leaves. The basic steps in classifying a leaf into diseased and non-diseased will be carried out by the following flowchart.

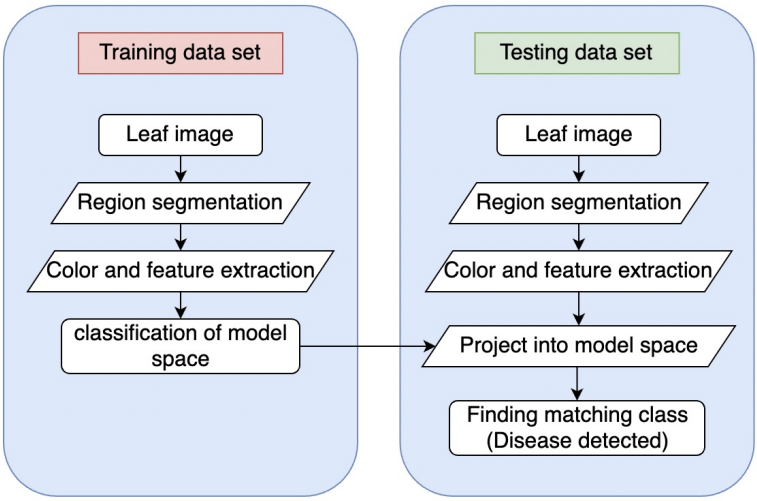

Fig. 1 Flowchart of the proposed method

4.1 Feature extraction

A human requires three basic features to identify a given sample. The features required are colors (spectrum), texture and contextual features. A machine will require more than the above features to classify the samples. Therefore, we extract mean, standard deviation, skewness, kurtosis, entropy, contrast, correlation, homogeneity and RMS values.

- First order moment (mean): Mean gives the average intensity value in an image.

$$\mu = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=N}^{N} P_{ij} \tag{1}$$

- Second order moment (Standard Deviation): Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean

$$\sigma = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (P_{ij} - \mu)^2} \tag{2}$$

- Third order moment (Skewness): Gives the degree of asymmetric 'leaning' to either left or right.

$$\theta = \sqrt[3]{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (P_{ij} - \mu)^3} \tag{3}$$

- Fourth order moment (Kurtosis): Gives the measure of central 'peakedness'.

$$\gamma = \sqrt[4]{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (P_{ij} - \mu)^4} \tag{4}$$

- Entropy : It is a measure of the degree of randomness in the image

$$E = -\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=N}^{N} P_{ij} log(P_{ij}) \tag{5}$$

We also extract texture-based features by first obtaining the Gray-Level Co-occurrence Matrices (GLCM) [8] from the segmented images-
(For an image P, we get the corresponding gray-Level co-occurrence matrix G with size $H \times H$)

- Energy.
- Entropy.
- Contrast.
- Homogeneity.

This gives us a total of 3 (channels) $\times$4 (color features) +4 (texture features)= 16 features.

4.2 Classification using SVM

For the project, we will be using a multi-class support vector machine. Support vector machine is a supervised learning algorithm where the data points are

divided using the hyperplanes. The following equation is used to classify the data-set.

$$\min_{\omega_{b\xi}} \frac{1}{2} W^T W + C \sum_{i=1}^{n} \xi_i \tag{6}$$

$$y_i(W^T \phi(x_i) + b) \geq 1 - \xi_i \tag{7}$$

Where ξ_i denotes the distance to the correct margin with $\xi \geq 0, i = 1, ..., n$
Where C denotes a regularization parameter
Where $W^T W = w^2$ denotes the normal vector
Where $\phi(x_i)$ denotes the transformation input space vector
Where b denotes a bias parameter
Where y_i denotes i-th target value

Using the above algorithm, we train the model with a predetermined labeled data-set. This will train the model to determine which leaves are healthy and which belong to disease-affected plants.

4.3 Method

Principal Compound Analysis(PCA) is used to reduce the higher dimensionality of the variables and identifying relationships among different variables. We use the t-Distributed Stochastic Neighbor Embedding (t-SNE) technique to visualize the data in higher dimensions.

5 Data-set and experiments

"Plant village" [9] data set is a publicly available data set. The data set consists of a variety of images (14 different classified leaves). The samples of leaves are shown in Image 2 The whole data set consists of around 54,306 images of healthy and diseased leaves we would be running the model with a couple of hundred images for this project.

Fig. 2 Sample of leaves from the data-set

"Plant Doc" [10] is another set of 20,900 images of leaves. The Plant Doc data-set is different compared to [9] in terms of number of classes (38) and the images are real-time with background noise.

5.1 Metrics

The following metrics are evaluated for each method and each hyperparameter to obtain an optimum model.

- **Precision:** It is the percentage truly positive of all predicted positive.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

- **Recall:** It is the percentage predicted positive of the total positive.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

- **F-1 Score:** It is the harmonic mean of precision and recall and takes both False Positives (FP) and False Negatives (FN) into consideration.

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

- **Confusion Matrix:** It is a tabular way of visualizing the model performance. The matrix is of size $n \times n$, where n is the number of labels.

Other than these metrics, we also get the t-SNE and PCA plots to visualize how the features would be distributed in lower dimensional space. This will give us insights about reducing the feature dimensions without affecting the accuracy too much extent.

6 Time Line

Week 1: Literature review data collection (Done)
Week 2: Pre-processing
Week 3: Feature extraction
Week 4: Creating the model and training
Week 5: Creating the model and training
Week 6: Validation
Week 7: Comparison of results from different algorithms
Week 8: visualization
Week 9: Conclusion
Week 10: Completing the report

References

[1] Government of India: Agri Census. https://censusindia.gov.in/census_and_you/economic_activity.aspx (2021)

[2] Government of India: Agri data. https://www.pib.gov.in/PressReleasePage.aspx?PRID=1741942 (2021)

[3] Yang, Y., Wang, J., Yang, Y.: Improving svm classifier with prior knowledge in microcalcification detection1. In: 2012 19th IEEE International Conference on Image Processing, pp. 2837–2840 (2012). https://doi.org/10.1109/ICIP.2012.6467490

[4] Mathur, A., Foody, G.M.: Multiclass and binary svm classification: Implications for training and classification users. IEEE Geoscience and Remote Sensing Letters **5**(2), 241–245 (2008). https://doi.org/10.1109/LGRS.2008.915597

[5] Sharma, P., Hans, P., Gupta, S.C.: Classification of plant leaf diseases using machine learning and image preprocessing techniques. In: 2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence), pp. 480–484 (2020). https://doi.org/10.1109/Confluence47617.2020.9057889

[6] Jenifa, A., Ramalakshmi, R., Ramachandran, V.: Classification of cotton leaf disease using multi-support vector machine. In: 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), pp. 1–4 (2019). https://doi.org/10.1109/INCOS45849.2019.8951356

[7] Wiwart, M., Fordoński, G., Zuk-Golaszewska, K., Suchowilska, E.: Early diagnostics of macronutrient deficiencies in three legume species by color image analysis. Computers and Electronics in Agriculture **65**, 125–132 (2009). https://doi.org/10.1016/j.compag.2008.08.003

[8] Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. IEEE Transactions on systems, man, and cybernetics (6), 610–621 (1973)

[9] Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Frontiers in plant science **7**, 1419 (2016)

[10] Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N.: Plantdoc: a dataset for visual plant disease detection. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, pp. 249–253 (2020)