ISSN 2348 – 8034
Impact Factor- 4.022

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

A REVIEW ON TECHNIQUES FOR PLANT LEAF CLASSIFICATION AND DISEASE DETECTION

Lakshmi V[*1] & Jinesh S[2]
[*1]M Tech Scholar, Department of ECE, College of Engineering, Thalassery, Kerala, India
[2]Assistant Professor, Department of ECE, College of Engineering, Thalassery, Kerala, India

ABSTRACT

Plants play a vital role in conserving the earth's ecology and the environment by preserving a healthy atmosphere. The categorization of leaves is a crucial process in some agricultural-based industries. The leaf contains the significant amount of information about the taxonomic plant characteristics. A leaf is categorized by its color, texture, and shape. The classification of plants is a technique, where the leaf is classified based on its features. There are various classification techniques such as k nearest neighbor classifier, Probabilistic Neural Network (PNN), Support Vector Machine (SVM). The classification decision is a difficult task since the quality of the results is different for different input data. The different classification methods which are used for plant recognition and also disease detection are discussed.

Keywords- *PNN, SVM, k-NN*

I. INTRODUCTION

Plant recognition or classification has a broad application prospective in agriculture and medicine, and is especially significant to the biology diversity research. Plant leaf classification finds application in botany and in tea, cotton and other industries. Plants are vitally important for environmental protection. However, it is an important and difficult task to recognize plant species on earth. Many of them carry significant information for the development of human society. The urgent situation is that many plants are at the risk of extinction. So it is very necessary to set up a database for plant protection. We believe that the first step is to teach a computer how to classify plants.

Leaf recognition plays an important role in plant classification. Plants are basically identified based on flowers and fruits. However these are three dimensional objects and increases complexity. Plant identification based on flowers and fruits require morphological features such as number of stamens in flower and number of ovaries in fruits. Identifying plants using such keys is a very time consuming task and has been carried out only by trained botanists. However, in addition to this time intensive task, there are several other drawbacks in identifying plants using these features such as the unavailability of required morphological information and use of botanical terms that only experts can understand. However leaves also play an important role in plant identification. Moreover, leaves can be easily found and collected everywhere at all seasons, while flowers can only be obtained at blooming season. Shape of plant leaves is one of the most important features for characterizing various plants visually. Plant leaves have two-dimensional nature and thus they are most suitable for machine processing.

Plants play a vital role in conserving the earth's ecology and the environment by preserving a healthy atmosphere. In the identification of a plant, the leaves play a paramount role due to its proximity throughout the year and it is easier to access through a computer. The shape of the leaf is one of the most significant features to characterize several plants visually. The categorization of leaves is a crucial process in some agricultural-based industries. In the area of botanical research, the classification of leaves is important to identify the classes and families of the plants. The leaf offers different features for identifying the species and the plant groups. In general, the identification of the plant is based on the observation of the plant's morphological characteristics such as the general character, the structure of stems, roots, and leaves by the known databases. The leaf contains the significant amount of information about the taxonomic plant characteristics. Furthermore, the leaves are available

ISSN 2348 – 8034
Impact Factor- 4.022

in the plants for various months in a year, whereas, the flowers and fruits may stay only for several weeks. A leaf is categorized by its color, texture, and shape. The classification of plants is a technique, where the leaf is classified based on its features.

Our paper presents survey of different classification techniques. Before classification can be done on basis of leaf some preprocessing is needed and most important step prior classification is feature extraction. For classification different techniques are available. Some of them are k-Nearest Neighbor Classifier, Probabilistic Neural Network and Support Vector Machine.

II. PLANT LEAF CLASSIFICATION TECHNIQUES

First step for plant leaf classification is image acquisition. Image acquisition includes plucking leaf from plant and then, the digital color image of the leaf is taken with a digital camera. After leaf image is obtained some pre-processing is needed. This stage includes grayscale conversion, image segmentation, binary conversion and image smoothing. The aim of image pre-processing is to improve image data so that it can suppress undesired distortions and enhances the image features that are relevant for further processing. Color image of leaf is converted to grayscale image. Variety of changes in atmosphere and season cause the color feature having low reliability. Thus it is better to work with grayscale image. Once image is converted to grayscale it is segmented from its background and then converted to binary. Using one of the edge detectors its contour is detected. Then certain morphological features are extracted from its contour image. This feature vector is then provided to the classifier. Fig.1 gives block diagram for plant leaf classification process.

A classification problem deals with associating a given input pattern with one of the distinct classes. Patterns are specified by a number of features (representing some measurements made on the objects that are being classified) so it is natural to think of them as d-dimensional vectors, where d is the number of different features. This representation gives rise to a concept of feature space. Patterns are points in this d-dimensional space and classes are sub-spaces. A classifier assigns one class to each point of the input space. The problem of classification basically establishes a transformation between the features and the classes. The optimal classifier is the one expected to produce the least number of misclassifications.

ISSN 2348 – 8034
Impact Factor- 4.022

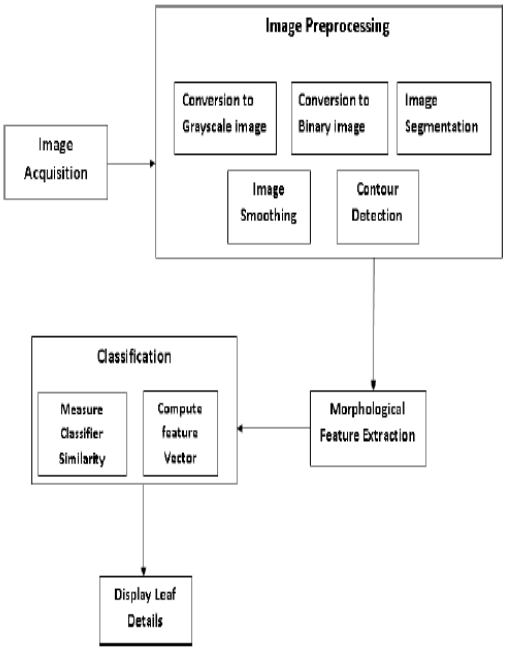

Figure 1: Block diagram for Plant Leaf classification

A. K-Nearest Neighbour Classifier

K Nearest Neighbor classifier calculates the minimum distance of a given point with other points to determine its class. Suppose we have some training objects whose attribute vectors are given and some unknown object w is to be categorized. Now we should decide to which class object w belongs.

Let us take an example. According to the k-NN rule suppose we first select k = 5 neighbors of w. Because three of these five neighbors belong to class 2 and two of them to class 3, the object w should belong to class 2, according to the k-NN rule. It is intuitive that the k-NN rule doesn't take the fact that different neighbors may give different evidences into consideration. Actually, it is reasonable to assume that objects which are close together (according to some appropriate metric) will belong to the same category. According to the k-NN rule suppose we first select k = 5 neighbors of w. Because three of these five neighbors belong to class 2 and two of them to class 3, the object w should belong to class 2, according to the k-NN rule.

ISSN 2348 – 8034
Impact Factor- 4.022

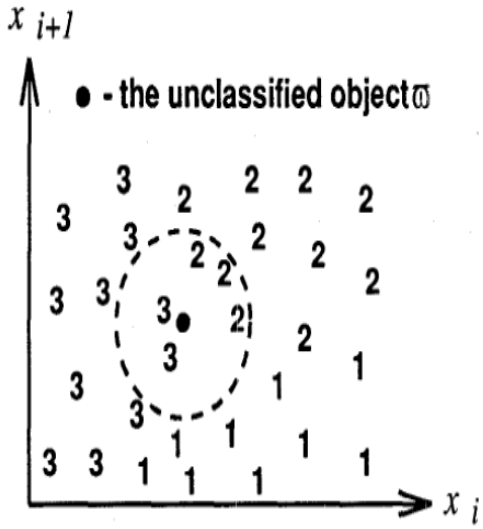

Figure 2: Example for classification using k-NN rule

.

For plant leaf classification, we first find out feature vector of test sample and then calculate Euclidean distance between test sample and training sample. This way it finds out similarity measures and accordingly finds out class for test sample. The k-nearest neighbor's algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. k is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose k to be an odd number as this avoids tied votes. It is intuitive that the k-NN rule doesn't take the fact that different neighbors may give different evidences into consideration. Actually, it is reasonable to assume that objects which are close together (according to some appropriate metric) will belong to the same category.

The well-known k-NN approach to classification has proven successful in many applications. In this method, we measure the distance from a test set item to each of the training set items, noting the k training set items that are nearest. We then classify the test set item by whichever class is most common among those k "nearest neighbors," letting each neighbor "vote." (In case of ties, we have chosen to include all training set items no farther away than the kth nearest neighbor, so in this case there will be more than k voters.) A number of investigators have considered the question of how best to measure distance: approaches have included global metrics (Fukunaga and Flick, 1984), local metrics (Short andFukunaga, 1980), metrics that are specific to the problem (Simardet al., 1993) andso on. By far the most common metric, though, has been Euclidean distance, under which the distance between two points Xr and Xs, say, is given by the square root of the (possibly weighted) sum of the squared distances over each co-ordinate.

In our approach, we attack the choices of k and of which variables to include by using a stepwise approach. Our implementation permits either forward or backward selection; for reasons of speed and parsimony we usually use the former. In this scheme, we start with every variable out of the model, plus a vector of possible values of k. Currently, this set is not chosen by reference to the data; we merely use the set 1, 3, 5,….., 31 since this has a "reasonably large" range. We use leave-one-out cross-validation to estimate the misclassification rate of the classifier for each choice of k. That is, each element of the training set is classified by all the others, using the current set of variables in the model and the entire vector of k's. Of course, at the very beginning of this process when every variable is "out" of the model, every training set item is equidistant from every test set item, and regardless of k, every training item gets to vote. Here, every test set item is simply given the most frequent training set classification.

ISSN 2348 – 8034
Impact Factor- 4.022

Then one of the variables is added to the model and a new set of misclassification rates, one for each k, is computed. This is done for each variable in turn. At the end of this process we choose the combination of k and added variable that produces the lowest misclassification rate. If no addition produces an improvement then the process is finished; the current set of variables and the best k are used. If the addition of a variable produces a misclassification rate strictly better than that of the current set, then that variable is added to the current set and the process continues. Our approach, as with other "greedy" algorithms, is reasonable but not guaranteed to produce an optimal set of variables. Since we require strict improvement at each stage we expect our routine to be resistant to the presence of "noise" variables.

- ***Disadvantages***
  - Expensive for testing each and every instance sensitive to noise
  - Gives irrelevant inputs

B. Probabilistic neural network classifier

Probabilistic neural networks can be used for classification problems. It has parallel distributed processor that has a natural tendency for storing experiential knowledge. PNN is derived from Radial Basis Function (RBF) Network. PNN basically works with 3 layers. First layer is input layer. The input layer accepts an input vector. When an input is presented, first layer computes distances from the input vector to the training input vectors and produces a vector whose elements indicate how close the input is to a training input. The second layer sums these contributions for each class of inputs to produce as its net output a vector of probabilities. Radial Basis Layer evaluates vector distances between input vector and row weight vectors in weight matrix. These distances are scaled by Radial Basis Function nonlinearly. The last layer i.e. competitive layer in PNN structure produces a classification decision, in which a class with maximum probabilities will be assigned by 1 and other classes will be assigned by 0.A key benefit of neural networks is that a model of the system can be built from the available data.

An artificial neural network (ANN) is an interconnected group of artificial neurons simulating the thinking process of human brain. One can consider an ANN as a "magical" black box trained to achieve expected intelligent process, against the input and output information stream. Thus, there is no need for a specified algorithm on how to identify different plants. PNN is derived from Radial Basis Function (RBF) Network which is an ANN using RBF. RBF is a bell shape function that scales the variable nonlinearly. PNN is adopted for it has many advantages. Its training speed is many times faster than a BP network. PNN can approach a Bayes optimal result under certain easily met conditions. Additionally, it is robust to noise examples. We choose it also for its simple structure and training manner. The most important advantage of PNN is that training is easy and instantaneous. Weights are not "trained" but assigned. Existing weights will never be alternated but only new vectors are inserted into weight matrices when training. So it can be used in real-time. Since the training and running procedure can be implemented by matrix manipulation, the speed of PNN is very fast. The network classifies input vector into a specific class because that class has the maximum probability to be correct. In this paper, the PNN has three layers: the Input layer, Radial Basis Layer and the Competitive Layer. Radial Basis Layer evaluates vector distances between input vector and row weight vectors in weight matrix. These distances are scaled by Radial Basis Function nonlinearly. Then the Competitive Layer finds the shortest distance among them, and thus finds the training pattern closest to the input pattern based on their distance.

- **Disadvantages**
  - Large network structure
  - Too many attributes results in over fitting of the network

ISSN 2348 – 8034
Impact Factor- 4.022

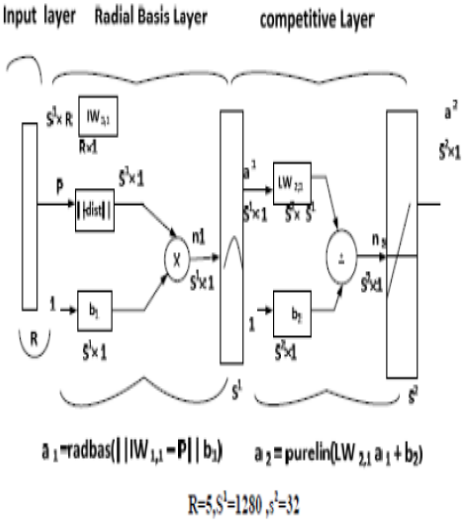

Fig 3: Architecture of PNN

We employ Probabilistic Neural Network (PNN) with image and data processing techniques to implement general purpose automated leaf recognition for plant classification. 12 leaf features are extracted and orthogonalized into 5 principal variables which consist the input vector of the PNN. The PNN is trained by 1800 leaves to classify 32 kinds of plants with accuracy greater than90%. c lassification system using PNN

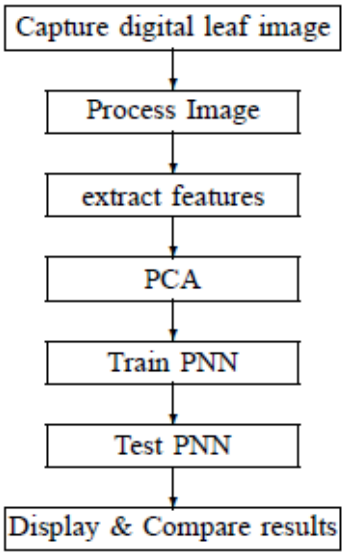

Fig 4: Flow diagram of

C. Support vector machine (SVM)

Support vector machine (SVM) is a non-linear classifier. The idea behind the method is to non-linearly map the input data to some high dimensional space, where the data can be linearly separated, thus providing great classification performance. Support Vector Machine is a machine learning tool and has emerged as a powerful technique for learning from data and in particular for solving binary classification problems [3]. The main

ISSN 2348 – 8034
Impact Factor- 4.022

concepts of SVM are to first transform input data into a higher dimensional space by means of a kernel function and then construct an OSH (Optimal Separating Hyper Plane) between the two classes in the transformed space. For plant leaf classification it will transform feature vector extracted from leaf's contour. SVM finds the OSH by maximizing the margin between the classes. Data vectors nearest to the constructed line in the transformed space are called the support vectors. The SVM estimates a function for classifying data into two classes. Using a nonlinear transformation that depends on a regularization parameter, the input vectors are placed into a high-dimensional feature space, where a linear separation is employed. To construct a nonlinear support vector classifier, the inner product (x, y) is replaced by a kernel function K (x, y), as in equation;

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i x) + b\right)$$

where f(x) determines the membership of x. We assume normal subjects were labeled as -1 and other subjects as +1.The SVM has two layers. During the learning process, the first layer selects the basis K (xi, x), i=1, 2….N from the given set of kernels, while the second layer constructs a linear function in the space. This is equivalent to finding the optimal hyper plane in the corresponding feature space. The SVM algorithm can construct a variety of learning machines using different kernel functions. Fig 5 shows the linear separating hyper plane where support vector are encircled. Main advantage of SVM is it has a simple geometric interpretation and gives a sparse solution. Unlike neural networks, the computational complexity of SVMs does not depend on the dimensionality of the input space. One of the bottlenecks of the SVM is the large number of support vectors used from the training set to perform classification tasks.

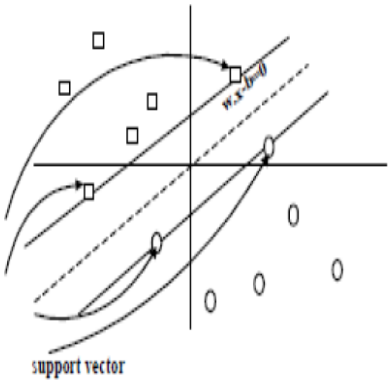

Fig 5: Linear separating hyper planes, the support vectors are circled.

The SVM classifier is widely used in bioinformatics (and other disciplines) due to its high accuracy, ability to deal with high-dimensional data such as gene expression, and ability in modeling diverse sources of data. SVMs belong to the general category of kernel methods. A kernel method is an algorithm that depends on the data only through dot-products. When this is the case, the dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional feature space. This has two advantages: First, the ability to generate non-linear decision boundaries using methods designed for linear classifiers. Second, the use of kernel functions allows the user to apply a classifier to data that have no obvious x dimensional vector space representation.

ISSN 2348 – 8034
Impact Factor- 4.022

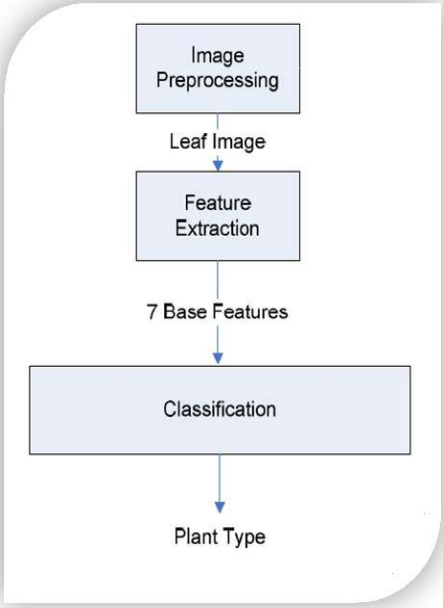

Fig 6: Classification using SVM classifier

- **Disadvantages**
  - Limitations in speed and size, both in training and testing data
  - Difficult to understand the structure of the algorithm

The nearest-neighbor method is perhaps the simplest of all algorithms for predicting the class of a test example. An obvious disadvantage of the kNN method is the time complexity of making predictions. Considerable amount of work has been done for recognizing plant species using k Nearest Neighbor technique. Classifying using PNN and SVM can further be explored by researchers, SVM being relatively a new machine learning tool. The most important advantage of PNN is that training is easy and instantaneous. Additionally, neural networks are tolerant to noisy inputs. But in neural network it's difficult to understand structure of algorithm. SVM was found competitive with the best available machine learning algorithms in classifying high-dimensional data sets. In SVM computational complexity is reduced to quadratic optimization problem and it's easy to control complexity of decision rule and frequency of error. Drawback of SVM is it's difficult to determine optimal parameters when training data is not linearly separable. Also SVM is more complex to understand and implement. Another technique we studying is genetic algorithm. Genetic algorithms are good at refining irrelevant and noisy features selected for classification. But representation of training/output data in genetic programming is complicated. Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. Future direction for researchers can be to explore more robust techniques for recognition of plant leaves using a combination of classifying techniques like SVM, kNN, PNN.

Mobile applications for plant leaf classification can be created which can be best learning tool for botany students. Also this application can be used in agricultural field for weed identification which in turn will help for proper determination of pesticides and fertilizers.

ISSN 2348 – 8034
Impact Factor- 4.022

III. CONCLUSION

In the identification of a plant, the leaves play a paramount role due to its proximity throughout the year and it is easier to access through a computer. The shape of the leaf is one of the most significant features to characterize several plants visually. The categorization of leaves is a crucial process in some agricultural-based industries. There are various successful classification techniques like k-Nearest Neighbor Classifier, Probabilistic Neural Network, Genetic Algorithm, Support Vector Machine, and Principal Component Analysis. Considerable amount of work has been done for recognizing plant species using k Nearest Neighbor technique. Classifying using PNN and SVM can further be explored by researchers, SVM being relatively a new machine learning tool. In order to overcome the drawbacks of the above said classification methods, a novel technique is used for the effective classification of the leaf. We can classify the disease of the plant leaf by determining the effects of adding multiple leaf features and also extend the proposed work to classify the medicinal leaf images, which is used in the Ayurvedic medicines for curing the human diseases

REFERENCES

[1] *Trishen Munisami, "Plant leaf recognition using shape features and colour histogram with k-nearest neighbour classifier", Second international symposium on computer vision and the internet (VisionNet'15)*
[2] *Stephen Gang Wu, "A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network", Dept. of Electronic Engineering, National Taiwan Univ. of Science & Technology, Taiwan, R. O. China*
[3] *Sushma S. Patil, "Identification and Classification of Cotton Leaf Spot Diseases using SVM Classifier", International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 4, April - 2014*