Survey on Techniques for Plant Leaf Classification

Prof. Meeta Kumar[1], Mrunali Kamble[2], Shubhada Pawar[3],
Prajakta Patil[4],Neha Bonde[5]
* (Department of Computer Engineering, MIT College of Engineering/ Pune University, India)
** (Department of Computer Engineering, MIT College of Engineering/ Pune University, India)
*** (Department of Computer Engineering, MIT College of Engineering/ Pune University, India)
**** (Department of Computer Engineering, MIT College of Engineering/ Pune University, India)
***** (Department of Computer Engineering, MIT College of Engineering/ Pune University, India

ABSTRACT

In this paper we present survey on various classification techniques which can be used for plant leaf classification. A classification problem deals with associating a given input pattern with one of the distinct classes. Plant leaf classification is a technique where leaf is classified based on its different morphological features. There are various successful classification techniques like k-Nearest Neighbor Classifier, Probabilistic Neural Network, Genetic Algorithm, Support Vector Machine, and Principal Component Analalysis. Deciding on the method for classification is often a difficult task because the quality of the results can be different for different input data. Plant leaf classifications has wide applications in various fields such as botany, Ayurveda, Agriculture etc. The goal of this survey is to provide an overview of different classification techniques for plant leaf classification.

Keywords - **Leaf classification, image preprocessing, classifier, k-Nearest Neighbor, SVM**

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 characterising various plants visually. Plant leaves have two-dimensional nature and thus they are most suitable for machine processing.

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, Neural Network, Genetic Algorithm, Support Vector Machine, and Principal Component Analysis. In section II we will discuss preprocessing to be performed on the acquired image .In section III we have discussed overview of classification techniques and how they can be used for recognition of a species of a plant leaf. Finally in section V we conclude and discuss the future work that can be done. Table 1 shows comparative study for classification techniques we studied through this survey.

II. LEAF IMAGE ACQUISITION AND PREPROCESSING

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.

III. CLASSIFICATION TECHNIQUES

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

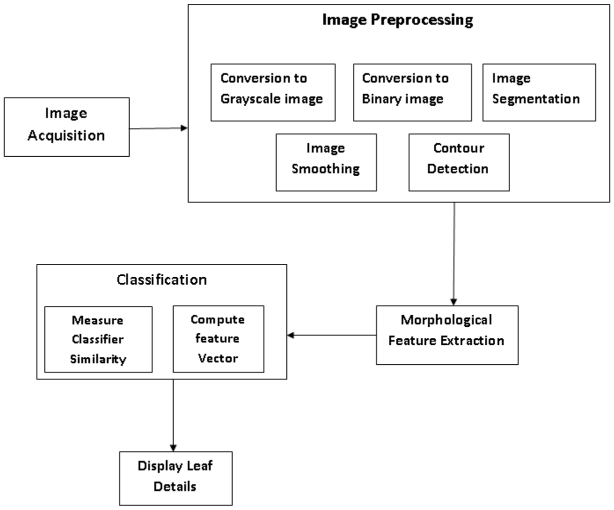

Figure 1 Block diagram for Plant Leaf classification

2.1 k-NEAREST NEIGHBOUR CLASSIFIERS

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.

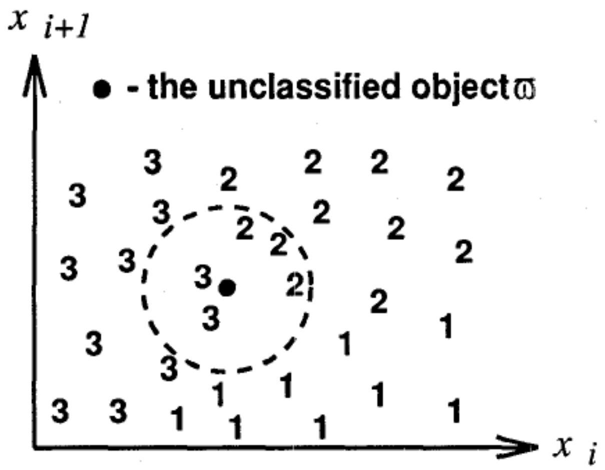

Figure 2 Example for classification using k-NN rule

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.

2.2 PROBABILISTIC NEURAL NETWORK

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 [3]. 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 [3]. 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. Fig.3 shows architecture of PNN.

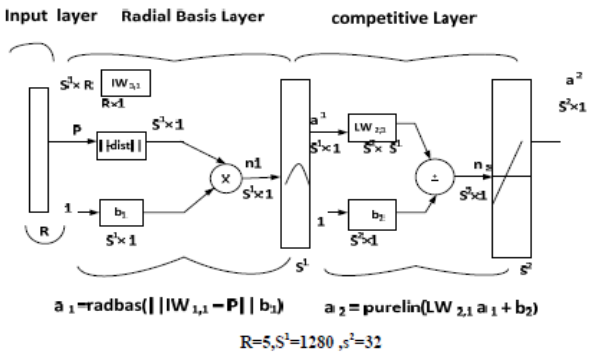

Figure 3 Architecture of PNN

2.3 GENETIC ALGORITHM

Genetic Algorithms are mainly used for feature classification and feature selection. The basic purpose of genetic algorithms (GAs) is optimization .GAs give a heuristic way of searching the input space for optimal x that approximates brute force without enumerating all the elements and therefore bypasses performance issues specific to exhaustive search. Genetic algorithm is used effectively in the evolution to find a near-optimal set of connection weights globally without computing gradient information and without weight connections initialization [1]. Though solution found by genetic algorithms is not always best solution. It finds "good" solution always. Main advantage of GA is that is adaptable and it possess inherent parallelism. Genetic Algorithms handle large, complex, non differentiable and multi model spaces for image classification and many other real world applications.

2.4 SUPPORT VECTOR MACHINE

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 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 [3]. 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 (1)

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

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 [4]. 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 4 shows the linear separating hyper plane where support vector are encircled.

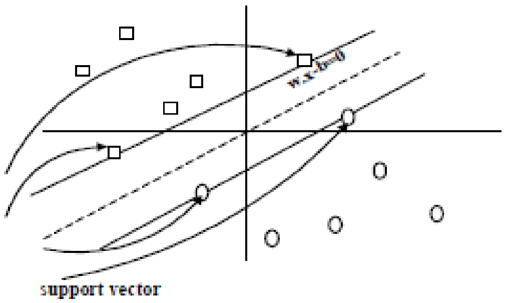

Figure 4 Linear separating hyper planes, the support vectors are circled

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.

2.5 PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is a variable reduction procedure. It is useful when you have obtained data on a number of variables (possibly a large number of variables), and believe that there is some redundancy in those variables. Intuitively, Principal components analysis is a method of extracting information from a higher dimensional data by projecting it to a lower dimension.

Principal component analysis is a basically used because it reduces the dimension of input vector of neural network. This method generates a new set of variables, called principal components. Each principal component is a linear combination of the optimally-weighted observed variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data. Mathematically, PCA transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on [2]. Each coordinate is called a principal component.

Often the variability of the data can be captured by a relatively small number of principal components, and, as a result, PCA can achieve high dimensionality reduction with usually lower noise than the original patterns. The objective of PCA is to perform dimensionality reduction while preserving as much of the randomness in the high-dimensional space as possible But the limitation with PCA is it depends on scaling of variables and it is not always easy to interpret principal components. The main limitation of PCA is that it does not consider class separability since it does not take into account the class label of the feature vector.

IV. Conclusion & Future Work

From study of above classification techniques we come up with following conclusion. 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.

PCA is used because it has advantage of reduced vector. The main limitation of PCA is that it does not consider class separability since it does not take into account the class label of the feature vector.

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.

COMPARATIVE STUDY

TABLE 1 Comparative Study of classification techniques for plant leaf classification

Classification Techniques	Pros	Cons
1. kNN Classifier	1. Simplest 2.Robust with regard to search space 3.No training is required, confidence level can be obtained	1. Expensive testing of each instance 2. Sensitiveness to noisy or irrelevant inputs 3.Lazy Learning
2. Probabilistic Neural Network	1. Tolerant of noisy inputs 2. Instances can be classified by more than one output 3. Adaptive to changing data	1. Long training time 2. Large complexity of network structure 3. too many attributes can result in over fitting
3.Genetic Algorithm	1. Handle large, complex, non differentiable and multi model spaces 2. Refining irrelevant and noise genes 3. Efficient search method for a complex problem space	1. Computation or development of scoring function is nontrivial 2. Not the most efficient method to find some optima, rather than global 3. Complications involved in the representation of training/output data
4. Support Vector Machine	1. Good generalization capability 2. Sparseness of the solution and the capacity control obtained by optimizing the margin 3. SVMs can be robust, even when the training sample has some bias	1. Slow training 2. Difficult to understand structure of algorithm 3. limitation is speed and size, both in training and testing
5.Principal Component Analysis	1. Used for variable reductions 2. Choose weights depending on the frequency in frequency domain. 3. Extract the maximum information in the data by maximizing the variance of the principal components.	1. Does not perform linear separation of classes 2. Scaling of variables 3. The largest variances do not correspond to the meaningful axes

REFERENCES

Journal Papers:

[1] M.Seetha, I.V.muralikrishna, B.L. Deekshatulu, B.L.malleswari, Nagaratna, P.Hegde a *Artificial neural networks and other methods of image classification, Journal of Theoretical and Applied Information Technology*, © 2005 - 2008 JATIT. All rights reserved.

[2] Krishna Singh, Indra Gupta, Sangeeta Gupta, *SVM-BDT PNN and Fourier Moment Technique for classification of Leaf, International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 3, No.* 4, December, 2010.

[3] Krishna Singh, Dr. Indra Gupta and Dr Sangeeta Gupta , *Retrieval and classification of leaf shape by support vector machine using binary decision tree, probabilistic neural network and generic Fourier moment technique: a comparative study, IADIS International Conferences Computer Graphics, Visualization, Computer Vision and Image Processing 2010*

[4] J.-X. Du, X.-F. Wang, and G.-J. Zhang, *Leaf shape based plant species recognition,Applied Mathematics and Computation, vol. 185*, 2007.

[5] A. Kadir, L. E. Nugroho, A. Susanto, P. Insap Santosa, "*Leaf Classification Using Shape, Color, and Texture Features*", *International Journal of Computer Trends and Technology- July to Aug Issue 2011*

[6] Hongjun Lum, Rudy Setiono, *Effective Data Mining using Neural Network, IEEE transactions on knowledge and data engineering, vol. 8, no. 6*, december 1996

[7] J. M. Zurada, *Introduction to Artificial Neural Networks System*. Jaico Publishing House.

[8] J.-X. Du, X.-F. Wang, and G.-J. Zhang, *Leaf shape based plant species recognition*,Applied Mathematics and Computation, vol. 185, 2007.

[9] D.E Goldberg, *Genetic Algorithms in Search, Optimization* and Machine Learning. Addison-Wesley, New York, 1989.

[10] M. Z. Rashad, B.S.el-Desouky, Manal S .Khawasik, *Plants Images Classification Based on Textural Features using Combined Classifier, International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011*

[11] D S Guru, Y. H. Sharath, S. Manjunath, *Texture Features and kNN in classification of Flower Images, IJCA special issue on" Recent Trends in Image Processing and Pattern Recognition" RTIPPR, 2010*

[12] Mahmood R Golzarian, Ross A Frick, *Classification of images of wheat, ryegrass and brome grass species at early growth stages using principal component analysis,* Golzarian and Frick Plant Methods 2011, 7:28 http://www.plantmethods.com/content/7/1/28

Proceedings Papers:

[13] Thair Nu Phyu, Survey of Classification Techniques in Data Mining, *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009*, Mar 18 - 20, 2009, Hong Kong.

[14] S. Wu, F. Bao, E. Xu, Y. Wang, Y. Chang, and Q. Xiang, A leaf recognition algorithm for plant classification using probabilistic neural network, in *Proceedings of 2007 IEEE International Symposium on Signal Processing and Information Technology*, Giza, Dec 2007.