An Optimal Feature Subset Selection Using GA for Leaf Classification

Valliammal Narayan and Geethalakshmi Subbarayan
Department of Computer Science, Women Deemed University, India

Abstract: *This paper describes an optimal approach for feature extraction and selection for classification of leaves based on Genetic Algorithm (GA). The selection of the optimal features subset and the classification has become an important methodology in the field of Leaf classification. The deterministic feature sequence is extracted from the leaf images using GA technique, and these extracted features are further used to train the Support Vector Machine (SVM). GA is applied to optimize the features of color and boundary sequences, and to improve the overall generalization performance based on the matching accuracy. SVM is applied to produce the false positive and false negative features. Our experimental results indicate that the application of GA for feature subset selection using SVM as a classifier proves computationally effective and improves the accuracy compared to KNN to classify the leaf patterns.*

Keywords: *Feature extraction, feature selection, classification, GA, SVM, geometric, color, boundary and ripple features.*

Received January 5, 2012; accepted March 21, 2013; published online February 26, 2014

1. Introduction

Although one of the upcoming research areas is the need for the development of automatic plant recognition system such as Computer Aided Plant Leaf Recognition (CAP-LR) [11, 12]. Botanists need a computer-aided tool without human interaction to study and identify leaves instead of holding a plant encyclopedia. Huge volumes of biological information are now providing online access to hundreds and thousands of images of specimens, helping to digitize the complete specimen collection of the leaf images [3]. Such a system will return within seconds the top matching species, along with supporting data that describes about textual descriptions and high resolution type specimen images just by feeding into the computer the photograph of a leaf specimen [5]. By using our system, a botanist in this field can quickly search the entire collections of plant species within seconds which earlier took hours together.

The classification problem involves multi-dimensional information that is used to determine which data belongs to what class out of a set of possible classes. The variables stored in the multi-dimensional data sets are referred as features. Regrettably numerous of potential features have considerable impact on the efficiency of the classifiers such as SVM, KNN etc., Most of these features are either partially or completely irrelevant or redundant to the classified target [8]. In advance, to discriminate among the classes, these features will not provide sufficient information. It is also, infeasible to include all possible features in the processes of classifying the patterns and objects [1, 2, 3]. Feature selection is one of the major tasks in classification problems. The discriminate features have to be carefully extracted from the image and the extracted features are used to train the classifiers. The optimal features subset is selected to increase the matching accuracy based on the performance of the classifiers [4, 13, 15]. Reducing the dimensions of the feature space not only reduces the computational complexity, but also, increases estimated performance of the classifiers. Genetic Algorithm (GA) based approach which is a powerful feature selection tool is used to select a feature subset that can describe the classification performance by using the SVM classifier [9, 14]. In our approach an optimized method using GA is brought forward for plant and tree classification.

The paper is organized as follows: section 2 describes the overview of optimal approach. Experimentation and their performance measurement results are discussed in section 3. Finally, the conclusion is summarized in section 4 with references.

2. Overview of Optimal Approach Using GA Based Feature Subset Selection System for Leaf Recognition

There are many approaches available for leaf recognition such as Classifiers, Hidden Markov Models (HMM), Bayesian networks, SVM and Dynamic Time Warping etc.

Among these approaches SVM classifier has proven to be a powerful tool for solving problems of prediction, classification and pattern recognition. Such systems can achieve greater accuracy than HMM based systems and can handle low quality and noisy data. Our optimal approach is proved efficiently in experimental results. Figure 1 explains the framework for the classification system.

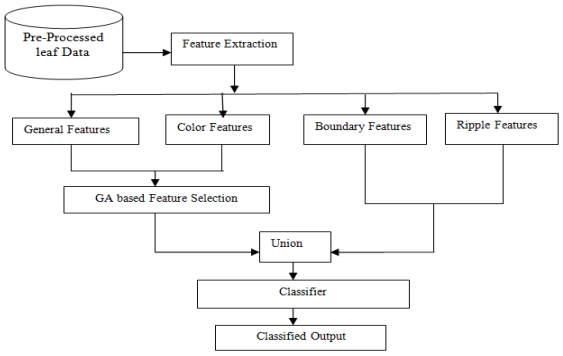

Figure 1. Framework for GA based feature subset selection for leaf classification.

2.1. Feature Extraction Methodology

The main work of a leaf recognition system is to extract common features among the images belonging to the same class of the data set and consequently indexing them. This method is applied to capture visual content of images for indexing and retrieval. The great variability of shapes and size of the leaves for plants and trees makes the recognition task difficult. By extracting features from image processing sequence, classification can be done by a discriminative classifier. They should be easy to compute in order for the approach to be feasible for a large image collection and rapid retrieval. Feature extraction methodologies analyse leaf images to extract the most prominent features that are representative of the various classes of objects [6, 7]. The below stated features are considered for extraction.

2.1.1. General Features

- *Leaf Area*: The value of leaf area is easy to evaluate, just counting the number of pixels of binary value on smoothed leaf image with in ROI. It is calculated as:

$$E = \iint dxdy \tag{1}$$

- *Circularity (C)*: Circularity is based on the bounding points of the ROI and is the ratio of the mean distance between the center of the ROI and all of the bounding points (μ_R) and the quadratic mean deviation of the mean distance (σ_R):

$$C = \mu_R / \sigma_R \tag{2}$$

- *Convex Perimeter Ratio (CPR)*: The convex perimeter ratio is the ratio of the ROI perimeter (P_{ROI}) and the convex hull perimeter (P_C).

$$CPR = P_{ROI} / P_C \tag{3}$$

2.1.2. Color Features

For RGB color space, the three features are extracted from each plane R, G, and B. The following statics are used to capture those moments. It is calculated using M and N represents the dimension and total number of pixels in the image. P_{ij} is values of color on i_{th} column and j_{th} row. The other moment used is termed Standard Deviation. The standard deviation is the square root of the variance of the distribution. Another moment used for extracting features is known as Skewness. It is the measure of the degree of asymmetry in the distribution:

$$Mean\ \mu = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} P_{ij} \tag{4}$$

$$STD\ \sigma = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(P_{ij} - \mu\right)^2} \tag{5}$$

The above table 1 shows the different features for plant and tree leaf images and the corresponding graphical representation is displayed in Figure 2.

Table 1. Features for plant and tree leaf images.

Features	No.	Feature List	Feature Values (Plant Leaves)	Feature Values (Tree Leaves)
General Features	1	Leaf Area	23166	10950
	2	Circularity	1.6795	1.0858
	3	Convex Perimeter Ratio	1388.035	885.193
Color Features	4	Mean	1.48+00	1.17E+00
	5	Standard Deviation	2.50E-01	1.39E-01
	6	Skewness	6.20E-02	1.78E+00
	7	Kurtosis	1.00E+00	4.19E+00
Boundary Features	8	Boundary	25124	12456
Ripple Features	9	Ripple Counting	10	8
	10	Ripple Pixel Counting	232	185

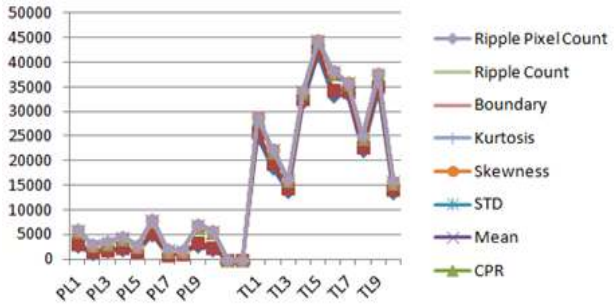

Figure 2. Graphical representation of plant and tree leaf features.

$$Skewness\ \theta = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \left(P_{ij} - \mu\right)^3}{MN\sigma^3} \tag{6}$$

$$Kurtosis\ \gamma = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \left(P_{ij} - \mu\right)^4}{MN\sigma^4} \tag{7}$$

2.1.3. Boundary Features

The system applies the contour algorithm, by applying threshold values 0.1 and 0.5 and calculates the pixel value within the boundary to extract the leaves boundary from the image.

2.1.4. Ripple Features

The ripple describes the fluctuation of the leaf boundary and can be obtained by finding the

differences between the leaf image and the average boundary of the leaf image. The average leaf boundary is calculated by R = LB/10 where R = range and LB is the length of boundary. The final image is called the ripples image. The ripple features are divided into two sub features:

- Ripples counting the ripples are the remaining objects in the ripple image.
- Ripples pixels counting-this process counts all the white pixels in all ripples.

2.2. Feature Subset Selection Methodology

Feature selection is an important task that allows the determination of the most relevant features for pattern recognition. The extracted features are normalized or reduced by selecting appropriate features to improve the classification accuracy [15]. A good feature selection results in:

- Faster training and better generalization.
- Removes redundant leaf images.
- Focuses recognition to a small set of properties.
- Displays the final classified outcome.

In this paper, we describe about the selection of optimal set of features using GA, which provide the discriminating information to classify the leaf patterns and increases the matching accuracy. Feature subset selection is performed in two stages.

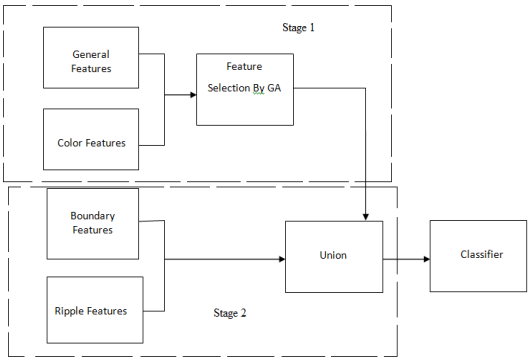

Figure 3. Frame work for Feature subset selection using GA.

Table 2. Features reduced by GA.

S.No.	Features	Features for Plant Leaves	Features for Tree Leaves
1	Leaf Area	23166	10950
2	Standard Deviation	2.50E-01	1.39E-01

The above table 3 shows the features reduced by our method and the graphical representation is shown in Figure 4.

Table 3. Features reduced by optimal approach.

S.No.	Features	Features for Plant Leaves	Features for Tree Leaves
1	Leaf Area	23166	10950
2	Standard Deviation	2.50E-01	1.39E-01
3	Boundary	25124	12456
4	Ripple Counting	10	8
5	Ripple Pixel Counting	232	185

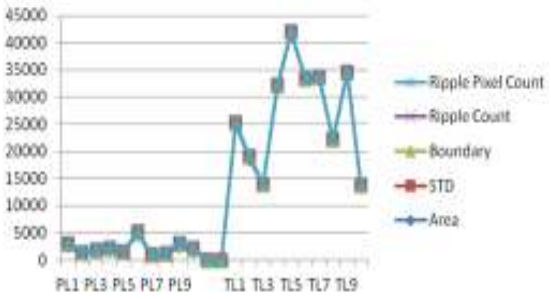

Figure 4. Graphical representation of features selected by optimal approach.

2.2.1. Feature Selection Based on GA

The normal process of searching for the features is computationally expensive; therefore GA is used as a search algorithm [9]. The GA sets up a group of imaginary lives having a string of codes for a chromosome on the computer. The GA evolves the group of imaginary lives (referred to as population), and gets almost optimum solution for the problem. The GA uses three basic operators over the population such as selection, crossover and mutation. For the first stage the parameters used for feature selection are leaf area, circularity, convex perimeter ratio and color features which are optimized through GA.

2.2.2. Evaluation Function

Selecting an appropriate evaluation function which produces the fitness of each individual in the population. GAs then use this feed back to bias the search process so, as to provide an improvement in the population's average fitness. The natural representation for the feature selection problem is the binary string of length N to indicate the presence or absence of each of the N possible features. The following Figure 3 shows the framework for optimal approach feature subset selection using GA.

The evaluation function selects the appropriate feature set which is solely based on the performance of the classification process .The calculated feature subset is recommended as the set of features to be used in the actual test data. In our system, the specified seven features in the first phase are reduced to two features. These selected features are forwarded to the additive or union operator path. The following Table 2 gives the feature value selected by GA method.

2.2.3. Feature Subset Selection

In second stage of feature selection path, the boundary, ripple features are taken for process. The combined two features are directly forwarded to the union operator path. The union of the optimal selected features from first stage and the combined features of the second stage are derived. Finally, the derived features are forwarded to SVM classifier.

2.3. Classification Through SVM

The recognition performance is efficiently processed through SVM. The parameters of the SVM are tuned to improve the overall generalization performance. SVM is primarily a two-class classifier that acts as an attractive and more systematic approach to learn linear or non-linear decision boundaries [8, 11]. Given a set of points, which belong to either of two classes, SVM finds the hyper-plane leaving the largest possible fraction of points of the same class on the same side, while maximizing the distance of either class from the hyper plane. The property of SVM is to simultaneously minimize the empirical classification error and maximize the geometric margin.

3. Experimental Results and Comparison

In our work the geometric feature, texture features and color features are extracted from different plant and tree leaves. These features are trained with SVM and K-Nearest classifiers. Large image data set is taken and 26 features are extracted for classification. 70% of images are used for training and 30% of images are used for testing purpose. For 50 leaf images, total 10 features are extracted and it is reduced to 5 features by using our method. Some sample dataset considered for classification is shown in the following figure 5.

Figure 5. Sample plant and tree leaf.

From the results obtained, the values of 1's and 0's in Table 4 corresponds to the classification of plant leaves features and misclassification whereas in Table these values represents tree leaves features and misclassification on the basis of SVM and KNN. The table 5 below shows the parametric results for classification. Accuracy is the proportion of correctly identified images from the total number of images. Sensitivity measures the ability of the proposed method to identify anomalous images. Specificity measures the ability of the method to identify normal images [8]. Our optimal method produces suitable results in terms of accuracy, sensitivity and specificity.

Table 4. Classifier results for plant leaf features.

S.No.	Feature Value for Plant Leaves	KNN	SVM
1	Leaf Area	0	1
2	Standard Deviation	1	1
3	Boundary	0	1
4	Ripple Counting	1	1
5	Ripple Pixel Counting	1	1

Table 5. Classifier results for tree leaf features.

S.No.	Feature Value for Tree Leaves	KNN	SVM
1	Leaf Area	1	1
2	Standard Deviation	1	0
3	Boundary	0	1
4	Ripple Counting	1	1
5	Ripple Pixel Counting	1	0

Table 6. Classifier results for plant leaf features.

S.No.	Classifiers	Accuracy	Sensitivity	Specificity
1	KNN	85%	75%	87%
2	SVM	88%	76%	89%

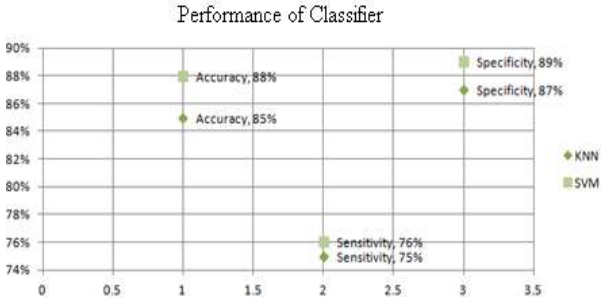

Figure 6. Sample plant and tree leaf.

Figure 6 shows the accuracy, sensitivity and specificity of KNN and SVM classifier performance. It is clearly proved that the SVM produces efficient and suitable results compared to the KNN classifier.

4. Conclusions and Future Work

This paper described about the optimal feature subset selection using GA for leaf classification. Two phases are adopted for the feature subset selection.

In the first stage the deterministic feature sequence is extracted from the leaf images using GA technique, and the extracted feature sequence is used to train the SVM. In the second stage combination of two features are extracted and are directly forwarded to SVM. Our approach based on GA for feature subset selection significantly improves the recognition accuracy of SVM to classify the leaf pattern .Future work involves research along the following directions: To consider large plant leaf dataset; feature selection and reduction method will be done by higher valid method for improving recognition accuracy and classification time.

References

[1] Chaki J. and Parekh R., "Plant Leaf Recognition Using Shape Based Features and Neural Network Classifiers," *the International Journal of*

Advanced Computer Science and Applications, vol. 2, no. 10, pp. 41 - 47, 2011.

[2] Guo G., Li S., and Chan K., "Support Vector Machines for Face Recognition," *Image and Vision Computing*, vol. 19, no. 9, pp. 631 - 638, 2001.

[3] Kadir A., Nugroho E., Susanto A., and Santosa P., "Leaf Classification Using Shape, Color, and Texture Features," *the International Journal of Computer Trends and Technology*, vol. 2, no. 10, pp. 224-230, 2011.

[4] Kadir A., Nugroho E., Susanto A., and Santosa P., "A Comparative Experiment of Several Shape Methods in Recognizing Plants," *the International Journal of Computer Science & Information Technology*, vol. 3, no. 3, pp. 256 - 263, 2011.

[5] Krishna B. and Kaliaperumal B., "Efficient Genetic-Wrapper Algorithm Based Data Mining for Feature Subset Selection in a Power Quality Patter Recoginition Application," *the International Arab Journal of Information Technology*, vol. 8, no. 4, pp. 397 - 405, 2011.

[6] Kumar N., Pandey S., Bhattachary A., and Ahuja S., "Do Leaf Surface Characteristics Affect Agrobacterium," *Journal of Biosciences*, vol. 29, no. 3, pp. 309 - 317, 2004.

[7] Li Y., Chi Z., and David D., "Leaf Vein Extraction Using Independent Component Analysis," *in Proceedings of IEEE Conference on Systems, Man and Cybernetics*, Taipei, China, pp. 3890 - 3894, 2006.

[8] Michalak K. and Kwasnicka H., "Correlation-Based Feature Selection Strategy in Classification Problems," *International Journal of Applied Mathematics and Computer Science*, vol. 16, no. 4, pp. 503 - 511, 2006.

[9] Papadakis S., Tzionas P., Kaburlazos V., and Theocharis J., "A Genetic Based Approach to the Type I Structure Identification Problem," *Informatica*, vol. 5, no. 3, pp. 364 - 376, 2005.

[10] Pornpanomchai C., Supapattranon C., and Siriwisesokul N., "Leaf and Flower Recognition System (e-Botanist)," *the International Journal of Engineering and Technology*, vol. 3, no. 4, pp. 347 - 351, 2011.

[11] Pornpanomchai C., Rimdusit S., Tanasap P., and Chaiyod C., "Thai Herb Leaf Image Recognition System," *Kasetsart Journal-Natural Science*, vol. 45, no. 3, pp. 551 - 562, 2011.

[12] Shabanzade M., Zahedi M., and Aghvami S., "Combination of Local Descriptors and Global Features for Leaf Recognition, Signal & Image," *Signal & Image Processing: An International Journal*, vol. 2, no. 3, pp. 23 - 31, 2011.

[13] Sathya B., Valli S., Raju S., and Kumar V., "Content Based Leaf Image Retrieval (CBLIR) Using Shape, Color and Texture Features," *Indian Journal of Computer Science and Engineering*, vol. 2, no. 2, pp. 202-211, 2011.

[14] Sinan R., Emre M., and Metin T., "Feature Extraction and Classifier Combination for Image-based Sketch Recognition," *in Proceedings of the 7th Sketch-Based Interfaces and Modelling Symposium*, Aire-la-Ville, Switzerland, pp. 63-70, 2010.

[15] Tzionas P., Papadakis E., and Manolakis D., "Plant Leaves Classification Based on Morphological Features and a Fuzzy Surface Selection Technique," *in Proceedings of International Conference on Technology and Automation,* Thessaloniki, Greece, pp. 365-370, 2005.

[16] Wahyu W. and Hugh W., "Simple and Accurate Feature Selection for Hierarchical Categorisation," *in Proceedings of the ACM Symposium on Document Engineering*, pp. 111-118, 2002.

[17] Zheng X. and Wang X., "Leaf Vein Extraction Based on Gray-Scale Morphology," *the International Journal Image, Graphics and Signal Processing*, vol. 2, no. 2, pp. 25-31, 2010.

Valliammal Narayan is the Assistant Professor in the Department of Computer Science and pursuing her Ph.D in Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore. She has more than 15 Years of teaching Experience. Her research interests includes Image Processing, Pattern Recognition and Neural Networks. She has more than 15 publications at National and International Level. She is a Life member of the one of Professional organization in Indian Science Congress Association.

Geethalakshmi Subbarayan is working as Associate Professor in the Department of Computer Science in Avinashiligam Institute for Home Science and Higher Education for Women, Coimbatore. She has more than 20 years of teaching experience. Her areas of interest include Image Processing and Software Engineering. She has around 50 publications in her research area at National and International Level. Presently she is guiding M.Phil and Ph.D research scholars. She is currently the Principal Investigator of one of the Major research Project funded by NRB. She is a life member of one of the professional organization in Indian Science Congress Association.