International Journal of Computer Applications Technology and Research
Volume 2– Issue 4, 449 - 454, 2013, ISSN: 2319–8656

Shape Oriented Feature Selection for Tomato Plant Identification

A. Hazra
Department of Computer Science & Engineering, Jadavpur University, Kolkata, India

K. Deb
Department of Computer Science & Engineering, Jadavpur University, Kolkata, India

S. Kundu
Department of Computer Science & Engineering, Jadavpur University, Kolkata, India

P. Hazra
Faculty of Horticulture, Bidhan Chandra Krishi Viswavidyalaya, Kalyani, Nadia, India

Abstract: Selection of relevant features for classification from a high dimensional data set by keeping their class discriminatory information intact is a classical problem in Machine Learning. The classification power of the features can be measured from the point of view of redundant information and correlations among them. Choosing minimal set of features optimizes time, space complexity related cost and simplifies the classifier design, resulting in better classification accuracy. In this paper, tomato (Solanum Lycopersicum L) leaves and fruiting habits were chosen with a futuristic goal to build a prototype model of leaf & fruit classification. By applying digital image processing techniques, tomato leaf and fruit images were pre-processed and morphological shape based features were computed. Next, supervised filter and wrapper based feature selection techniques were adopted to choose the optimal feature set leading to small within-class variance and large among-class distance which may be of utter importance in building the model for recognition system of the tomato leaf and fruiting habit genre.

Keywords: Tomato Leaf, Tomato Fruit, Morphology, Feature Selection, Filter based Feature Selection, Wrapper based Feature Selection

1. INTRODUCTION

Being one of the most consumed vegetables worldwide and cultivating in almost every corner of the world, Tomato (Solanum Lycopersicum L) gained economic importance by the beginning of twentieth century. Wild Tomatoes are native of western South America, distributed from Ecuador to northern Chile, and with two endemic species in the Galápagos Islands (Peralta and Spooner, 2005). It is one of the most investigated species, both in genetic and genomic studies (Foolad 2007). So it becomes a necessity to classify and recognize the large number of cultivars present in Tomato for farmers, seed producing agencies, botanists, Agricultural R&D labs and others. Plant Identification is important in GIS based remote sensing and in national parks, where a botanist manually identifies the species through time-consuming, tedious experiments. Thus digital image processing techniques can be applied to the verification process to increase speed up, accuracy and fully automation. The following table (Table 1) shows the Species for tomatoes and their wild relatives along with their fruit colour.

Leaf/fruit classification is a tough task as inter-class similarities or intra-class variations are quite natural in mathematical modelling of biological samples (colours and textures are quite similar in different tomato species). Also colour depth, variation along with textures of the leaves and fruits usually change with plant age so colour based recognition of the samples is not realistic. The variations in the gray scale histogram images (Figure 1) belonging to same cultivar/species were shown which gives enough evidence about the limitations of the colour image processing techniques in this case. Here we have chosen shape oriented feature extraction strategies, after which best features were selected based upon filter and wrapper based techniques. A detailed flow diagram is given (Figure 2).

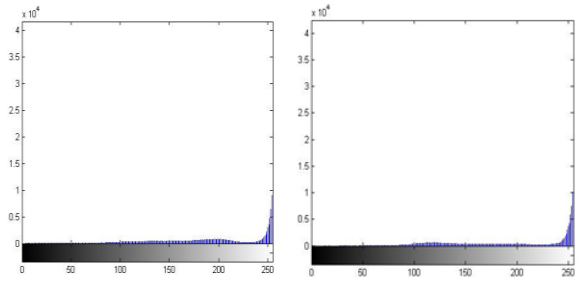

Figure 1 – Histogram showing Intra class variations in same tomato leaf/fruit species

2. DATA COLLECTION, PRE-PROCESSING

Images were collected at dawn in Bidhan Chandra Krishi Viswavidyalaya to avoid the sunlight noises and each image (2816 X 2112 pixels) was captured from an equidistant point to avoid the overhead related to image normalization. The image samples were categorized based on their genetic characters i) Top Leaflet and ii) Fruit Bearing Habit (Figure 3). Then the noisy backgrounds were replaced by a white uniform background (Figure 4). After that, the pre-processing steps were carried out in which the images were complemented, converted to HSV Colour space, grey scale and finally to binarized images. Next the morphological operations (thinning, small component removal, morphological dilation, noise prone zone removal) were performed to get the enhanced binarized images (Figure 5). Once we get the enhanced binarized images, morphological shape features were calculated by approximating the leaf boundary with an ellipse and fruiting habit as an irregular shape. The longest length (Major Axis/Branch Length) and breadth (Minor Axis/Branch Width) of leaves and fruits

(Figure 5) were calculated from the developed user interface. All shape based features are described in the next section.

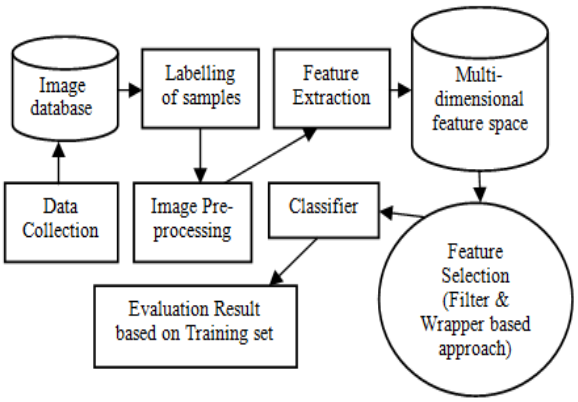

Figure 2 – Proposed feature selection and evaluation approach

Figure 3 - Top Leaflet and Fruiting Habit

Figure 4 - Processed Images of Top Leaflet and Fruiting Habit

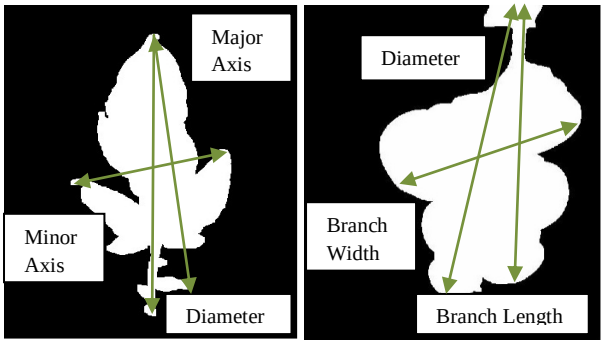

Figure 5 – Choosing Main features from the GUI end

Table 1 – Different tomato cultivars all around the world

New nomenclature	Fruit color
Solanum lycopersicoides	Green-yellow when maturing, black when ripe
Solanum sitiens	Green-yellow when maturing, black when ripe
Solanum juglandifolium	Green to Yellow green
Solanum ochranthum	Green to Yellow green
Solanum pennellii	Green
Solanum habrochaites	Green with darker green stripes
Solanum chilense	Green to whitish green with purple stripes
Solanum huaylasense	Typically green with dark green stripes
Solanum peruvianum L.	Typically green to greenish-white, sometimes flushed with purple
Solanum corneliomuelleri	Typically green with dark green or purple stripes, sometimes flushed with purple
Solanum arcanum	Typically green with dark green stripes
Solanum chmeilewskii	Typically green with dark green stripes
Solanum neorickii	Typically green with dark green stripes
Solanum pimpinellifolium L.	Red
Solanum lycopersicum L.	Red
Solanum cheesmaniae	Yellow, orange
Solanum galapagense	Yellow, orange

3. FEATURE SELECTION

3.1 List of Morphological Features

- **Major Axis/Branch Length (L):** The length between the top and bottom end of the leaf/ fruit branch.(Averaged) (Figure 5)
- **Minor Axis/Branch Width (W):** The breadth between the two distant ends of the leaf/ fruit branch.(averaged) (Figure 5)
- **Aspect Ratio/Length Width Ratio (A_R):** Ratio between the Major Axis (Branch Length) and Minor Axis (Branch Width). [$A_R = L / W$]
- **Orientation/Branch Angle (α):** Angle between the Major Axis (Branch Length) with X axis; representing leaf/fruiting branch bending nature).
- **Eccentricity (€):** Ratio of the distances between two foci of an ellipse. (Applicable for leaf feature only). [€ = $\sqrt{1 - b^2/a^2}$, 2a,2b being the lengths of Major and Minor Axis respectively,$0 < € < 1$ in case of ellipse]
- **Area (A):** Total amount of space inside the two dimensional leaf/fruiting branch surface.
- **Perimeter (P):** Two dimensional 8 connectivity based neighborhood boundary of the closed geometric leaf/fruit surface.
- **Equivalent Diameter(E_D):** The diameter of the circle with the same area with the two dimensional leaf/fruit surface.
- **Number of On Pixels (O_P):** The total count of white pixels inside the leaf/fruit surface.
- **Form Factor(F_F):** The "roundness" of the leaf/fruiting branch.[$F_F = 4\pi A / P^2$]
- **Rectangularity(R):** Measures how rectangular the leaf/fruiting branch is. [R = L.W/A]
- **Solidity(S):** Ratio between the Area (A) of the binarized image and the Area of its Convex Hull ($A_{convex\ hull}$). [$S = A / A_{convex\ hull}$]

- **Concavity (C_a):** Difference between the Convex Hull Area ($A_{convex\ hull}$) and Area (A) of the binarized image. [$C_a = A_{convex\ hull} - A$]
- **Perimeter Ratio/Major-Minor Axis(P_R):** The ratio between the Perimeter(P) and the summation of its Major Axis(Branch Length) and Minor Axis(Branch Width). [$P_R = P/ L+W$]
- **Convexity(C):** Ratio between the Perimeter of the Convex Hull ($P_{convex\ hull}$) and the actual Perimeter (P) of the binarized image. [$C= P_{convex\ hull} / P$]
- **Smooth Factor(S_F):** Ratio of the image area smoothened by a 5×5 median filter to that smoothened by a 2×2 median filter.
- **Diameter (D):** The longest distance between any two points on the closed geometric surface of the leaf/fruiting branch.(Fig. 5)
- **Narrow Factor (N_F):** The "narrowness" of the leaf/fruiting branch. [$N_F = D/L$]
- **Perimeter Ratio of Diameter (P_{RD}):** Ratio of the Perimeter (P) and Diameter (D) of the Leaf/fruiting branch. [$P_{RD} = P/D$]
- **Compactness (C_A):** Associating the Area (A) of the image sample over its Diameter (D).[$C_A = \sqrt{4A/\pi}/D$]
- **R-Factor (R_F):** Ratio of the Perimeter of the Convex Hull ($P_{convex\ hull}$) and Diameter (D). [$R_F = P_{convex\ hull} / D$]
- **Euler Number (E_N):** The topology of a binarized image measured as the total number of objects in the image minus the total number of holes in the image.

3.2 Feature Selection

There are several approaches available for feature subset selection in machine learning. Selection of most effective features for classification from a large data set can be obtained by three basic selection strategies i) filter method ii) wrapper method and iii) embedded method. In filter based method, an attribute evaluator is used to evaluate the attributes/features and a ranker to rank all the features present in the feature data set. Next, the lower ranked features are omitted one by one and predictive accuracy is checked each time through a classifier (by measuring Mean Absolute Error, RMS Error, Relative Absolute Error, Root relative Squared Error etc.). One problem is that there is a possibility of being over fit of the model in filter based methods because the weights put by the ranker algorithm in order to rank the features can be very different than the weights put by the classification algorithm. The wrapper based approach uses a subset evaluator which creates all possible subsets from the multidimensional feature data set by using a search technique (Best First Search, Linear Floating forward Selection). Then a classifier is used to evaluate each feature subset to consider the subset with which the classifier performs the best recognition result. In embedded method, classifier dependent feature selection is done. In this work, filter and wrapper based techniques were adopted to observe the outcomes.

Feature selection and classification tasks were performed with the open-source WEKA Machine Learning workbench. In case of filter based technique, both univariate (Information Gain, Ben-Bassat, 1982) and multivariate (Correlation based feature selection, Hall, 1999) approaches were used. In case of wrapper based approach, classifier subset evaluation using Naive Bayes classifier was utilized to test the instances along with different search strategies. Knowledge of the dependencies among the features can be gained by finding out the internal relationships among the features and the quantification of their descriptive powers.

4. EXPERIMENTAL RESULTS

4.1 Filter based Approaches (Information-Gain and Correlation measurement)

Information gain, being a goodness measurement criterion in machine learning and associating with entropy calculation, measures the purity of randomly drawn examples. Through Information Gain technique, all the features can be ranked using ranker search method from the feature set. The entropy of a random variable X is defined as

$$H(X) = -\sum_{i=0}^{m} p(x) log_2((p(x)) \quad \textbf{(1)}$$

Where m is the number of observed outcomes of the random variable X, p(x) being the probability density function for X. Now, if the observed value of random feature variable X is evenly distributed according to Y, then the entropy of X after observing Y is

$$H(X/Y) = -\sum_{i=0}^{m} p(y) \sum_{i=0}^{m} p\left(\frac{x}{y}\right) log_2((p(x/y)) \quad \textbf{(2)}$$

Hence the information gain depicts the extra information of X with respect to Y, saying the amount of entropy decreased for X, formulated as Information gain = H(X)−H(X/Y). According to the Info-gain phenomenon, the feature ranking of the 10 and 14 best features of tomato leaf and fruit feature data set from the 22 morphological features were shown (Figure 6). The classification result on the data set was given (Figure 7) which depicts that the percentage of the correctly classified leaf and fruit instances using Naive Bayes classifier are 82.2% and 97.62% respectively. The Cohen's kappa coefficient of tomato leaf and fruit classes is 0.8095 and 0.9744, describing the statistical measure of inter-rater agreement for tomato leaf and fruit feature variables along with mean absolute error, R.M.S. error, relative absolute error and root relative squared error measurements(Figure 7). The total numbers of instances used in training of tomato leaf and fruit classes are 45 and 42 respectively. The performance of the Naive Bayes classifier was measured by observing some crucial parameters e.g. True positive rate (TP rate), False positive rate (FP rate), Precision, Recall, F-score/F-measure, Receiver operating characteristics (ROC) area. It is observed that the average true class prediction (TP rate) rate being 82.22% and 97.6% for the tomato leaf and fruiting habit respectively. The class prediction rates (average) being positive in case of false samples (FP rate) are 0.013 and 0.002. The precision for the tomato leaf and fruiting habit indicates the ratio between the true positive over the total retrieved samples are 86.1% and 98.2% respectively and a recall indicates the fraction of relevant instances that are retrieved are 81.4% and 100%. F-score/F-measure is an accuracy measurement criterion of a classifier which is the harmonic mean of the precision and recall.

$$\text{F-Score/F-Measure} = \frac{2\times(precision\times recall)}{(precision\times recall)} \quad \textbf{(3)}$$

81.4% and 97.6% F-Measure are found in leaf and fruit feature data sets. The Receiver operating characteristics (ROC) curve is the plot between FP rate and TP rate showing

the behaviour of a classifier. The AUC (area under the curve) measure is noted as 0.981 and 1 which is almost ideal for a classifier (Figure 8). So, the result depicts that Naive Bayes classifier acts ideally on tomato leaf and fruiting habit data.

Ranking Value	Attribute Name
1.347	Minor Axis
1.209	Diameter
1.209	R-Factor
1.195	Perimeter Ratio of Diameter
1.155	Rectangularity
0.961	Compactness
0.909	Perimeter Ratio
0.787	Eccentricity
0.787	Aspect Ratio
0.562	Major Axis
0	Area
0	Orientation
0	Euler number
0	Convexity
0	Concavity
0	Narrow Factor
0	Smooth Factor
0	On Pixel
0	Equivalent Diameter
0	Solidity
0	Form Factor
Search Method	Attribute Ranking
Attribute Evaluator	Information Gain Ranking Filter
	(Supervised)

Ranking Value	Attribute Name
2.335	Narrow Factor
1.933	Aspect Ratio
1.474	Perimeter Ratio
1.23	Branch Width
1.227	Rectangularity
0.827	Branch Length
0.738	Diameter
0.738	Perimeter Ratio of Diameter
0.736	On Pixel
0.736	Area
0.736	Equivalent Diameter
0.736	solidity
0.732	Perimeter
0.732	Convexity
0	R-Factor
0	Euler Number
0	Smooth Factor
0	Form Factor
0	Compactness
0	Orientation
Search Method	Attribute Ranking
Attribute Evaluator	Information Gain Ranking Filter
	(Supervised)

(Figure 6 - Ranking of the leaf and fruit attributes through Information Gain Ranking Filter)

Correctly Classified Instances	37	82.22%
Incorrectly Classified Instances	8	17.77%
kappa Statistic	0.8095	
Mean Absolute Error	0.0265	
Root mean Squared Error	0.1387	
Relative Absolute Error	21.31%	(Approx.)
Root relative Squared Error	55.61%	(Approx.)
Total Number of Instances	45	(Training)

Correctly Classified Instances	41	97.62%
Incorrectly Classified Instances	1	2.38%
kappa Statistic	0.9744	
Mean Absolute Error	0.0025	
Root mean Squared Error	0.0427	
Relative Absolute Error	1.88%	(Approx.)
Root relative Squared Error	16.57%	(Approx.)
Total Number of Instances	42	(Training)

Figure 7 - Evaluation result and accuracy measurement on Tomato Leaf/Fruiting habit data set using CFS and Info-gain techniques

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Tomato Leaf Class
1	0	1	1	1	1	Class 1
0.667	0	1	0.667	0.8	0.992	Class 2
1	0	1	1	1	1	Class 3
1	0	1	1	1	1	Class 4
1	0	1	1	1	1	Class 5
0.667	0	1	0.667	0.8	0.992	Class 6
1	0.024	0.75	1	0.857	0.976	Class 7
0.667	0.024	0.667	0.667	0.667	0.992	Class 8
0.333	0	1	0.333	0.5	0.968	Class 9
1	0.071	0.5	1	0.067	1	Class 10
1	0	1	1	1	1	Class 11
1	0.024	0.75	1	0.857	0.992	Class 12
0.333	0.024	0.5	0.333	0.4	0.865	Class 13
1	0.024	0.75	1	0.857	0.976	Class 14
0.667	0	1	0.667	0.8	0.968	Class 15
0.822	**0.013**	**0.861**	**0.822**	**0.814**	**0.981**	**(Weighted Avg.)**

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Tomato Fruit Class
1	0	1	1	1	1	Class 1
1	0	1	1	1	1	Class 2
1	0	1	1	1	1	Class 3
1	0	1	1	1	1	Class 4
1	0	1	1	1	1	Class 5
1	0	1	1	1	1	Class 6
1	0	1	1	1	1	Class 7
0.667	0	1	0.667	0.8	1	Class 8
1	0.026	0.75	1	0.857	1	Class 9
1	0	1	1	1	1	Class 10
1	0	1	1	1	1	Class 11
1	0	1	1	1	1	Class 12
1	0	1	1	1	1	Class 13
1	0	1	1	1	1	Class 14
0.976	**0.002**	**0.982**	**0.976**	**0.976**	**1**	**(Weighted Avg.)**

Figure 8 - Performance evaluation of Naive Bayes classifier for each class; feature selection based on Information gain technique

This section discusses on correlation based feature selection techniques. Correlation based Feature Selection (CFS) is a multivariate method used in feature selection with some popular search strategies like best first search (BFS) and linear floating forward selection (LFFS) etc. Naive Bayes classifier was used to validate the results. The correlations among attributes and category variable can be classified into weak correlation, strong correlation and without any correlation. An ideal feature vector is strongly correlated with category attribute and not correlated with any other feature vectors (otherwise they are redundant). In this work, linear correlation coefficient measurement with Pearson product moment criterion was computed to find the correlations among the features and category attribute. Generally, the correlation measurement is defined as

$$\text{Correlation } x, y = \frac{\sum_{i=1}^{n}(x_i - x')(y_i - y')}{(n-1)STD_x \; STD_y} \quad \textbf{(4)}$$

$$x' = \sum_{i=1}^{n} x_i \quad , \quad y' = \sum_{i=0}^{n} y_i \quad \textbf{(5)}$$

$$\mathbf{STD_x} = \frac{\sqrt{\sum_{i=1}^{n}(x_i - x')^2}}{n-1}, \; STD_y = \frac{\sqrt{\sum_{i=1}^{n}(y_i - y')^2}}{n-1} \quad \textbf{(6)}$$

Where x, y are feature vectors x', y' are their means, STD_x and STD_y are the standard deviations of feature vectors x and y. The range of Correlation varies between -1 to 1. The more the $|\text{Correlation}_{x, y}|$ approaches towards 1, correlation between x , y increases. If $\text{Correlation}_{x, y}$ is zero, the features are independent of each other, and if $\text{Correlation}_{x, y}$ approaches towards -1, the features are more negatively correlated. CFS technique evaluates the worth of a subset of features by considering their individual predictive ability along with the degree of redundancy among them with a searching strategy (Best First search, Exhaustive search, Linear Forward Selection etc.). The most relevant features found in CFS method are listed below. (Table 2) The performance of Naive Bayes classifier applied on the most relevant feature set found using CFS technique was given in the following figure. (Figure 9)

Table 2 - Top Leaflet and Fruiting Habit

Problem Set	Best Features
Tomato leaf feature data set	Minor Axis, Aspect Ratio, Rectangularity, Perimeter Ratio, Diameter, Perimeter Ratio of Diameter, R-Factor
Tomato fruiting habit feature data set	Branch Width, Aspect Ratio, Perimeter, Rectangularity, Concavity, Perimeter Ratio, Diameter, Narrow Factor, Perimeter Ratio of Diameter

TP Rate	FP rate	Precision	Recall	F-Measur	ROC Area	Class
1	0	1	1	1	1	a
0.667	0.024	0.667	0.667	0.667	0.992	b
1	0	1	1	1	1	c
1	0	1	1	1	1	d
1	0	1	1	1	1	e
0.667	0	1	0.667	0.8	1	f
1	0.024	0.75	1	0.857	1	g
0.667	0	1	0.667	0.8	1	h
0.667	0.024	0.667	0.667	0.667	0.992	i
1	0.048	0.6	1	0.75	1	k
1	0	1	1	1	1	l
0.667	0	1	0.667	0.8	1	m
0.333	0	1	0.333	0.5	0.992	n
1	0.048	0.6	1	0.75	1	o
0.822	0.013	0.863	0.822	0.817	0.998	Weighted Average

TP Rate	FP rate	Precision	Recall	F-Measur	ROC Area	Class
1	0	1	1	1	1	a
1	0	1	1	1	1	b
1	0	1	1	1	1	c
1	0	1	1	1	1	d
1	0	1	1	1	1	e
1	0	1	1	1	1	f
1	0	1	1	1	1	g
1	0	1	1	1	1	h
1	0	1	1	1	1	i
1	0	1	1	1	1	k
1	0	1	1	1	1	l
1	0	1	1	1	1	m
1	0	1	1	1	1	n
1	0	1	1	1	1	o
1	0	1	1	1	1	Weighted Average

Figure 9 - Performance evaluation of Naive Bayes classifier; features selected using CFS based technique

4.2 Wrapper based Approach

As a part of wrapper based feature selection, some search methods (best first search and linear floating forward selection) were used in combination with Naive Bayes classifier to find out the best subset of features with which the Naive Bayes classifier performs the best recognition result. A basic flow diagram was given. (Figure 10)

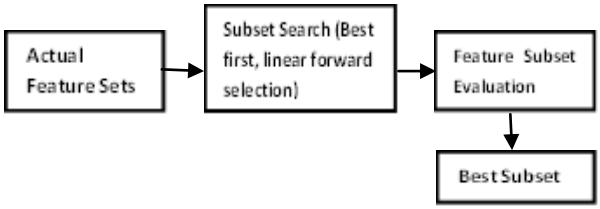

Figure 10 – Wrapper based classifier sub-set evaluation via Naive Bayes classifier

Best first (Greedy Hill climbing augmented with backtracking facility) search and Linear floating forward selection search strategies were taken individually to find out the best discriminating feature sub set for tomato leaf and fruiting habit recognition model. As a classifier, Naive Bayes algorithm was used as an inductive algorithm to estimate the merits of the feature sets. A stopping criterion for the selection of feature subset was specified as the number of subsets can be huge. So, a predefined number of iterations were specified for this purpose. After the selection of the best subset of features for tomato leaf and fruiting habit data set, a validation procedure was also adopted to check the predictive accuracy of the subset based on the Naive Bayes classifier.

Search Method for Leaf Data Set	Selected Attributes
	Minor Axis
Best First Search	Area
	Perimeter Ratio of Diameter
Linear Forward Floating Selection	Minor Axis
	Concavity
	Perimeter Ratio of Diameter

Search Method for Fruiting Habit Data Set	Selected Attributes
Best First Search	Aspect Ratio
	Perimeter Ratio
	Narrow Factor
Linear Forward Floating Selection	Aspect Ratio
	Perimeter Ratio
	Narrow Factor

Figure 11 – Most relevant features selected for Tomato Leaf/Fruiting Habit Data Set

Correctly Classified Instances	37	82.22%
Incorrectly Classified Instances	8	17.77%
Kappa Statistic	0.8095	
Mean Absolute Error	0.0191	
Root Mean Squared Error	0.1255	
Relative Absolute Error	15.3125	
Root Relative Squared Error	50.3026	
Total Number of Instances	45	

Correctly Classified Instances	42	100.00%
Incorrectly Classified Instances	0	0.00%
Kappa Statistic	1	
Mean Absolute Error	0	
Root Mean Squared Error	0	
Relative Absolute Error	0	
Root Relative Squared Error	0.00%	
Total Number of Instances	42	

Figure 12 – Evaluation through Naive Bayes classifier and accuracy measurement on Tomato Leaf/Fruiting Habit Data Set

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	a
0.667	0.024	0.667	0.667	0.667	0.992	b
1	0	1	1	1	1	c
1	0	1	1	1	1	d
1	0	1	1	1	1	e
0.667	0	1	0.667	0.8	1	f
1	0.024	0.75	1	0.857	1	g
0.667	0	1	0.667	0.8	1	h
0.667	0.024	0.667	0.667	0.667	0.992	i
1	0.048	0.6	1	0.75	1	j
1	0	1	1	1	1	k
0.667	0	1	0.333	0.5	0.992	l
0.333	0	1	0.333	0.5	0.992	m
0.667	0.024	0.667	0.667	0.667	0.992	n
1	0.048	0.6	1	0.75	1	0
0.822	0.013	0.863	0.822	0.817	0.998	Weighted Average

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	a
1	0	1	1	1	1	b
1	0	1	1	1	1	c
1	0	1	1	1	1	d
1	0	1	1	1	1	e
1	0	1	1	1	1	f
1	0	1	1	1	1	g
1	0	1	1	1	1	h
1	0	1	1	1	1	i
1	0	1	1	1	1	j
1	0	1	1	1	1	k
1	0	1	1	1	1	l
1	0	1	1	1	1	m
1	0	1	1	1	1	n
1	0	1	1	1	1	Weighted Average

Figure 13 – Performance evaluation of Naive Bayes classifier; features selected using Wrapper based technique

It is observed that the best discriminatory features recorded using best first search technique are minor axis, area, perimeter ratio of diameter, aspect ratio and perimeter ratio, narrow factor in case of tomato leaf and fruiting habit feature data set respectively. Similarly in case of linear floating forward selection strategy, the best relevant features are minor axis, concavity, perimeter ratio of diameter and aspect ratio, perimeter ratio, narrow factor for tomato leaf and fruiting habit feature data set. (Figure 11)

An overall 82.22% and 100% recognition accuracy using Naive Bayes algorithm with kappa statistic 0.8095 and 1 were noted with optimally chosen tomato leaf and fruiting habit features using wrapper based approach (Figure 12) and the predictive accuracies are almost ideal for each of the leaf and fruiting habit classes (Figure 13). An important observation was noted that in both filter and wrapper based approaches the recognition accuracy of tomato leaf data set was 82.22 % but in case of fruiting habit data set it became 97.6%(Info-gain), 100%(CFS) and 100%(wrapper based). So, better predictive accuracy for fruiting habit data set was observed in case of CFS and wrapper based techniques. These observation results may be of utter importance when building a leaf/fruiting habit recognition model.

5. CONCLUSION

An elaborative description and observation results are produced in this work where the concentration was given to supervised learning based feature subset selection problems using both filter and wrapper based approaches. Here, we have investigated both univariate (Information gain) and multivariate (Correlation based) feature selection strategies with horticultural feature data set. But embedded feature selection strategies like Simulated Annealing, Decision Tree or Random Multinomial Logit (RMNL) also demand a lot of attention which need to be explored. The relevant attributes found in separate feature selection approaches were evaluated and compared using Naive Bayes algorithm. In Info-gain method, reduced number of features were 10, 14 and in CFS technique it is 7, 9 whereas in wrapper based approach it is 3, 3 respectively from 22 number of features. When these different techniques can be combined a better classification result may be obtained for tomato leaf and fruiting habit. Also scope for automating the whole process and investigating other methods for feature selection requires further studies.

6. REFERENCES

[1] S. Kundu, A. Hazra, K.Deb, P. Hazra, “Dimensionality Reduction of Morphological features of Tomato Leaves and Fruiting Habits”, IEEE International Conference on Communications, Devices and Intelligent Systems, CODIS-2012, pp. 608-611.

[2] Pavan Kumar Mishra, Sanjay Kumar Maurya, Ravindra Kumar Singh, Arun Kumar Misra, “A semi-automatic plant identification based on digital leaf and flower Images”, IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM - 2012).

[3] M. Dash, H. Liu, “Feature Selection for Classification”, Intelligent Data Analysis 1 (1997) 131-156, Elsevier.

[4] Hossain. J. and M. A. Amin, “Leaf Shape Identification Based Plant Biometrics”, Proceedings of 13th International Conference on Computer and Information Technology (ICCIT 2010).

[5] J. Huang, N. Huang, L. Zhang, H. Xu, “A method for feature selection based on the correlation analysis”, IEEE International Conference on Measurement, Information and Control (MIC 2012).

[6] George H. John, Ron kohavi, Karl Pfleger, “Irrelevant Features and the Subset Selection Problem”, Machine Learning: Proceedings of the Eleventh International Conference, 1994Morgan Kauffmann Publishers, San Francisco, CA.

[7] Pat Langley, “Selection of Relevant Features in Machine Learning”, Proceedings of the AAAI Fall Symposium on Relevance (1994), New Orleans, LA.

[8] Yvan Saeys, Inaki Inza, Pedro Larranaga, “A review of feature selection techniques in bioinformatics”, BIOINFORMATICS, vol. 23 no. 19 2007, pages 2507-2517.

[9] Dionysios Lefkaditis, Georgios Tsirigotis, “Morphological feature selection and neural classification for electronic components”,Journal of Engineering Science and Technology Review 2(1), 2009, 151-156.

[10] Jinsong Leng, Craig valli, Leisa Armstrong,“A wrapper-based Feature Selection for Analysis of Large Data Sets”, 3rd International Conference on Computer and Electrical Engineering,(ICCEE 2010),pp. 166-170.

[11] Yvan Saeys, Inaki Inza, Pedro larranaga, “ A review of feature selection techniques in bioinformatics”,BIOINFORMATICS,vol. 23 no. 19 2007. Pages 2507-2517.

[12] Ron Kohavi, George H. John, “ Wrappers for feature subset selection”, Artificial Intelligence 97(1997) 273-324, ELSEVIER.

[13] Hai Nguyen, Katrin Franke, Slobodan Petrovic, “Improving Effectiveness of Intrusion Detection by Correlation Feature Selection”, IEEE International Conference on Availability, Reliability And Security, pp. 17-24.

[14] M.Hall,“Correlation Based Feature Selection for Machine Learning”, Doctoral Dissertation, University of Waikato, Department of Computer Sceince ,1999.

[15] Ben-Bassat, M., “Pattern Recognition and Reduction of Dimensionality”, In P.R. krishnaiah and L.N. Kanal,editors, Handbook of statistics-II, North Holland,1982, 773-791.