

CAMELLIA (THEACEAE) CLASSIFICATION WITH SUPPORT VECTOR MACHINES BASED ON FRACTAL PARAMETERS AND RED, GREEN, AND BLUE INTENSITY OF LEAVES

W. JIANG¹, Z.M. TAO¹, Z.G. WU¹, N. MANTRI², H.F. LU* AND Z.S. LIANG

College of Life science, Zhejiang Sci-Tech University, Hangzhou 310018, China

Keywords: Camellia; Classification; Fractal analysis; RGB; SVM.

Abstract

Leaf traits are commonly used in plant taxonomic applications. The aim of this study was to test the utility of fractal leaf parameters analysis (FA) and leaf red, green, and blue (RGB) intensity values based on support vector machines as a method for accurately discriminating *Camellia* (68 species from five sections, 11 from sect. *Furfuracea*, 13 from sect. *Paracamellia*, 15 from sect. *Tuberculata*, 24 from sect. *Theopsis* and 5 from sect. *Camellia*). The results showed that the best classification accuracy was up to 96.88% using the RBF SVM classifier ($C = 16$, $g = 0.5$). The linear kernel overall accuracy was 90.63%, and the correct classification rates of 40.63% and 93.75% were achieved for the sigmoid SVM classifier ($C = 16$, $g = 0.5$) and the polynomial SVM classifier ($C = 16$, $g = 0.5$, $d = 2$), respectively. A hierarchical dendrogram based on leaf FA and RGB intensity values was mostly on agreement with the generally accepted classification of the *Camellia* species. SVM combined with FA and RGB may be used for rapidly and accurately classifying *Camellia* species and identifying unknown genotypes.

Introduction

Camellia L. is a commercially important genus of family Theaceae. It is cultivated globally, particularly in tropical and subtropical regions of East and Southeastern Asia (Ming, 2000; Gao *et al.*, 2005; Lu *et al.*, 2012). Some *Camellia* species are used to produce tea, others are cultivated as ornamental plants, and the seeds of some species are used for making edible oils (Chen *et al.*, 2005; Vijayan *et al.*, 2009; Jiang *et al.*, 2012). Currently, there are number of discrepancies in relation to classification of species from this important genus. There are three popular *Camellia* monographs developed by Sealy (1958), Chang (1998) and Ming (2000) that differ significantly in species, section and subgenus arrangement. All these taxonomic classifications are based on the morphology. Many studies have shown that classifications purely based on the traditional morphological characteristics are insufficient for closely related species because low divergence prevents having reasonable qualitative features to support the taxonomic systems (Bari *et al.*, 2003; Lu *et al.*, 2008a,b; Pandolfi *et al.*, 2009; Jiang *et al.*, 2010). As a result, there is no concordance in the method for classification of *Camellia* and further taxonomic research is necessary (Pi *et al.*, 2009).

Leaf characters have been successfully exploited to solve plant taxonomy problems (Plotze *et al.*, 2005, Lin *et al.*, 2008; Ye and Weng, 2011). Traditionally, leaf traits such as shape (Ming, 2000), morphology research (Barthlott *et al.*, 2009), and leaf anatomy (Pi *et al.*, 2009; Jiang *et al.*, 2010) have been used for classification. Recently, several researchers have used fractal parameters

* Corresponding author. Email: luhongfei0164@163.com

¹ Zhejiang Institute of Subtropical Crops, Zhejiang Academy of Agricultural Sciences Wenzhou 325005, China.

² School of Applied Sciences, Health Innovations Research Institute, RMIT University, Melbourne 3000, Victoria, Australia.

DOI: <http://dx.doi.org/10.3329/bjpt.v24i1.33034>

for plant identification (Mancuso *et al.*, 2003; Azzarello *et al.*, 2009; Pandolfi *et al.*, 2009). Mancuso (1999) highlighted the importance of the leaf fractal geometry for fingerprinting plants. In addition, leaf colour information provides useful data for judging maturity of agricultural products (Gunasekaran *et al.*, 1985), detecting diseases (Howaith *et al.*, 1990), and fruit sorting (Harrell *et al.*, 1989). Thus, the leaves really provide plenty of characteristics that can be used as a source of data for plant taxonomy (Yang and Lin, 2005).

Supervised techniques are one of the most effective analysis tools in classification field currently (Lu *et al.*, 2012). These tools apply available information about a category membership of samples to developed model for classification of the genus. Support vector machines (SVM) is a supervised pattern recognition technology which has the algorithm developed in the machine learning community and is capable of learning in high-dimensional feature spaces (Cortes and Vapnik, 1995; Lu *et al.*, 2011). The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of, which makes the SVM is a non-probabilistic binary linear classifier. Recently, SVM has been used in a variety of areas like information retrieval (Jain *et al.*, 1999), object recognition (Pontil and Verri, 1998), food bruise detection (Lu *et al.*, 2011), qualitative assessment of tea (Chen *et al.*, 2008), and fruit classification (Zheng *et al.*, 2010). Chen *et al.* (2007) demonstrated that SVM fixes the classification decision function based on structural risk minimum mistakes instead of the minimum mistake of the misclassification on the training set to avoid over-fitting problem. Compared to other pattern recognition tools such as artificial neural networks (ANNs), SVM is a powerful method with a higher training speed and can avoid overtraining (Jack and Nandi, 2002; Kumar *et al.*, 2011). In addition, Burges (1998) suggested that SVM could get the best solution of data set with better ability of generalization.

So far there is no knowledge about the utility of leaf image analysis and machine learning as a taxonomic toolkit for classification of genus *Camellia*. In this study, we combine the fractal leaf parameters and leaf red, green, and blue intensity values with SVM to analyze the taxonomical classification of *Camellia* plants. The main objective of this work was to (a) develop and evaluate the effectiveness of SVM for identifying 68 species in genus *Camellia*, and (b) confirming these relationships based on fractal parameters and red, green, and blue (RGB) intensity values of leaves. Our purpose is to provide a potential tool for accurate classification of *Camellia* species.

Material and Methods

Materials

All plant materials were collected from the International *Camellia* Garden in Jinhua, Zhejiang Province (29°07' N, 119°35' E, 40 m in altitude) in July 2011. All plants share the same environment in this garden which reduces the major effect of geographical distribution on leaf development. Healthy leaf samples following Chang's taxonomic treatment (1998), 11 species from sect. *Furfuracea*, 13 species from sect. *Paracamellia*, 15 species from sect. *Tuberculata*, 24 species from sect. *Theopsis*, and five species from sect. *Camellia*, for a total of 68 species were examined, and split into two groups: 36 for training phase of SVM model construction and the other 32 for the validation phase (Table 1). All samples were taken from the third mature leaves that was fully exposed to sunlight and horizontally arranged on the two-year-old branches of the plants. At least three plants per species were selected. Means of data were obtained using SAS version 9.0 (SAS Institute, Cary, NC, USA). Voucher specimens for all species were deposited in the Chemistry and Life Science College of Zhejiang Normal University (ZJNU) (see Appendix 1 for voucher details).

Image acquisition and fractal parameters

A Canon EOS 50D camera with a Canon EF-S 18-55 mm f/3.5-5.6 IS lens at 50 mm, was used to acquire leaf images. All image acquisition was carried out at least in five and the lighting for images was entirely from natural light on a sunny summer morning. Leaf fractal parameters were calculated using fractal image analysis software (HarFA, Harmonic and Fractal Image Analyzer 5.4) as previously described by Mancuso (2002), Pandolfi *et al.* (2009) and Zheng *et al.* (2011). Briefly, Figure 1 shows schematic diagram of HarFA output and five parameters in detail. The basic procedure was as follows: (1) each *Camellia* leaf image was split into the constituent color channels (red, green, blue); (2) each channel was set for a threshold color value between 0 and 255; (3) the fractal dimension (D) for red, green, and blue channel was calculated by box counting method; (4) then the D which is presented as a function of thresholding condition in fractal spectrum was plotted against the colour intensity to obtain the fractal spectra of the three channels; (5) determining the baseline ($D = 1$) that separates the fractal ($D > 1$) from the non-fractal ($D < 1$) zone of the spectrum. For this study, we selected $D = 1.2$ as the baseline. (6) Finally, the five fractal parameters (X_1 , X_2 , X , Y , and S) were determined by Origin Lab (version 8.0). Additionally, average RGB intensity values from *Camellia* images were assessed using the colour histogram tool of Image J (National Institutes of Health, Bethesda, MD).

Cluster analysis

As a method of grouping data based on attributes of given population into similar and dissimilar groups, we conducted clustering analysis to classify 68 species in genus *Camellia* based on 15 fractal parameters and average RGB intensity values of leaf and compared it to Chang's (1998) results. A hierarchical dendrogram was constructed using Unweighted Pair-Group Method with Arithmetic Mean analysis (UPGMA). The Gower General Similarity Coefficient was applied to address multi-dimensional scaling. The multivariate statistical package (Version 3.13n, Kovach Computing Services) was used to conduct the cluster analysis.

SVM analysis

Support vector machine (SVM) was first proposed for pattern recognition applications by Vapnik (1995) based on statistical learning theory. The classification mechanism of SVM can be described as simple as: SVM tries to create an appropriate boundary (hyperplane) that meets the requirements of classification, the distance between the boundary and the nearest data points (support vectors) are maximal while the classification precision is also guaranteed. Theoretically, SVM can realize the optimal classification of linearly separable data. In order to solve non-linear problem, SVM converts the data from a low dimension input space to a high dimension feature space through a transformation function (kernel function).

All SVM algorithms are implemented with LIBSVM (Version 3.0) under MATLAB software (The Mathworks, Inc., Natick, MA, USA, version 7.9 R2009b). The LIBSVM is a library for support vector machines (2001).

Results

The fractal dimension and RGB intensity values of species

As shown in the flow chart (Fig. 1), for each species, the five fractal parameters (X_1 , X_2 , X , Y , S) were derived from the fractal spectra of each (red, green, and blue) colour channels (15 variables). The fractal values obtained for different *Camellia* species belonging to sections *Furfuracea*, *Paracamellia*, *Tuberculata*, *Theopsis*, and *Camellia* are shown in Figs. 2-4). These RGB intensity values were shown in Table 2. Thus, 18 input variables were obtained for modeling.

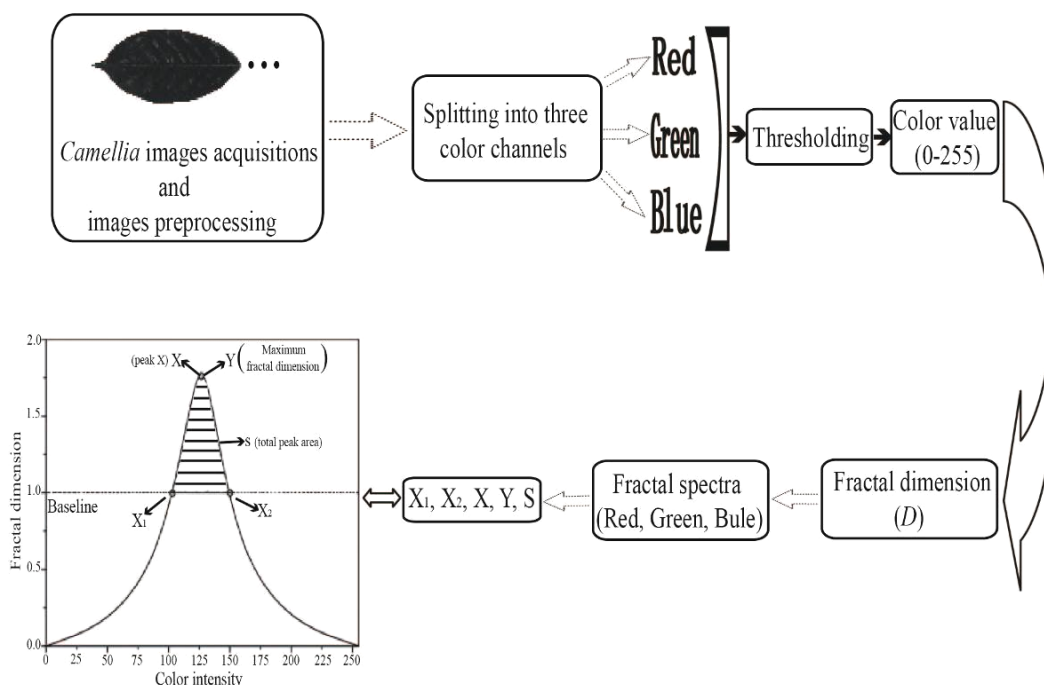


Fig. 1. Schematic diagram of the experimental protocol used to get fractal parameters and RGB intensity values from the image analysis of *Camellia* leaves.

Unsupervised cluster analysis

The relationship between the 68 *Camellia* species was examined by constructing a dissimilarity dendrogram using the 18 variables described above (Fig. 5). The species classified under sect. *Theopsis* by Chang (1998) clustered together (number 40 to 63) in the current study. Further, species number 12 to 24 and number 28 grouped together as an independent branch, which is also mostly congruous with Chang's treatment of sect. *Paracamellia*. Species number 1 to 11 belonging to sect. *Furfuracea* according to Chang's taxonomy also clustered together. However, two species, viz. *C. tuberculata* and *C. obovatifolia* from sect. *Tuberculata* also clustered with them. The other sect. *Tuberculata* species clustered together apart from *C. rhytidophylla* that clustered with sect. *Paracamellia*. Finally, species from sect. *Camellia* clustered together apart from *C. xiafongensis* that clustered with sect. *Theopsis*.

Support vector machine (SVM) classification accuracy

The training set and test set of SVM model is presented in Table 1. The class designation is important for training of SVM algorithms. The 68 species analyzed in the current study were divided into five categories, so the class designation followed the predefined Chang's (1998) taxonomy. Two SVM parameters namely regularization parameter (C) and kernel parameter (g), which are the keys to obtain good model performance, are optimized by cross validation. In current work, $\log_2 C$ and $\log_2 g$ were distributed from -5 to 5 with increments of 0.5. As seen in Fig. 6, the highest average accuracy of 83.33% was achieved when $C = 16$ and $g = 0.5$ for the training data set. The parameter of polynomial SVM were the combinations of another polynomial degree (d) with $d \in \{2, 3, 4, 5, 6, 7, 8, 9\}$. The classification results of linear, radial basis function (RBF), and

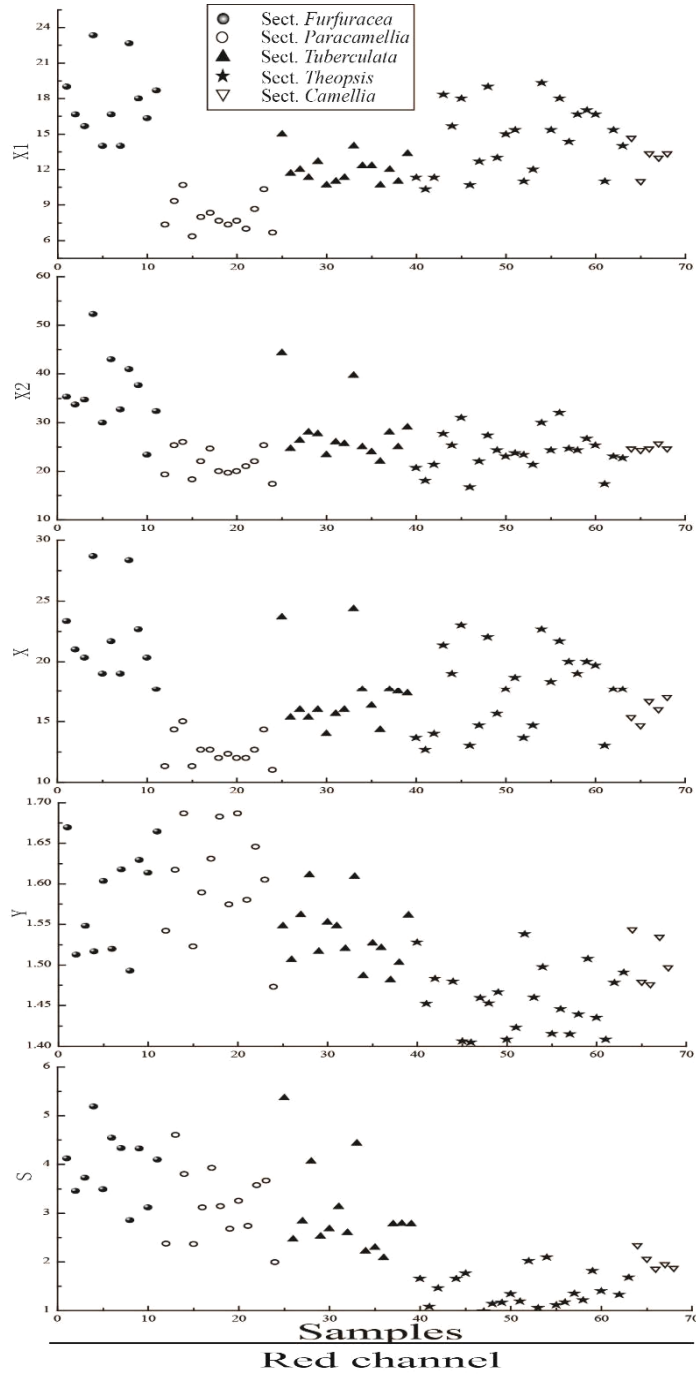


Fig. 2. Scatter plot of fractal parameters used for SVM models in this study. The five fractal parameters (X1, X2, X, Y, S) derived from the samples using **red** channel are shown. Numbers in the figure correspond to the species numbers in Table 1.

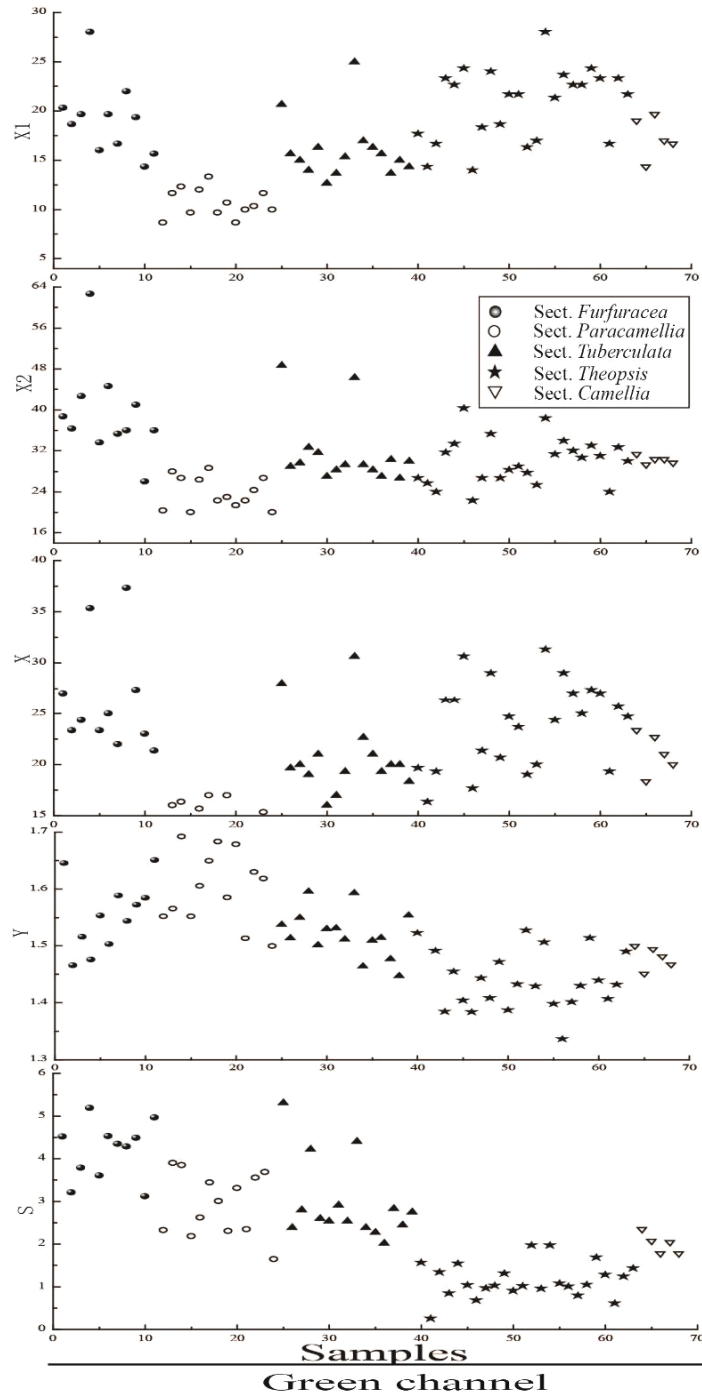


Fig. 3. Scatter plot of fractal parameters used for SVM models in this study. The five fractal parameters (X1, X2, X, Y, S) derived from the samples using **green** channel are shown. Numbers in the figure correspond to the species numbers in Table 1.

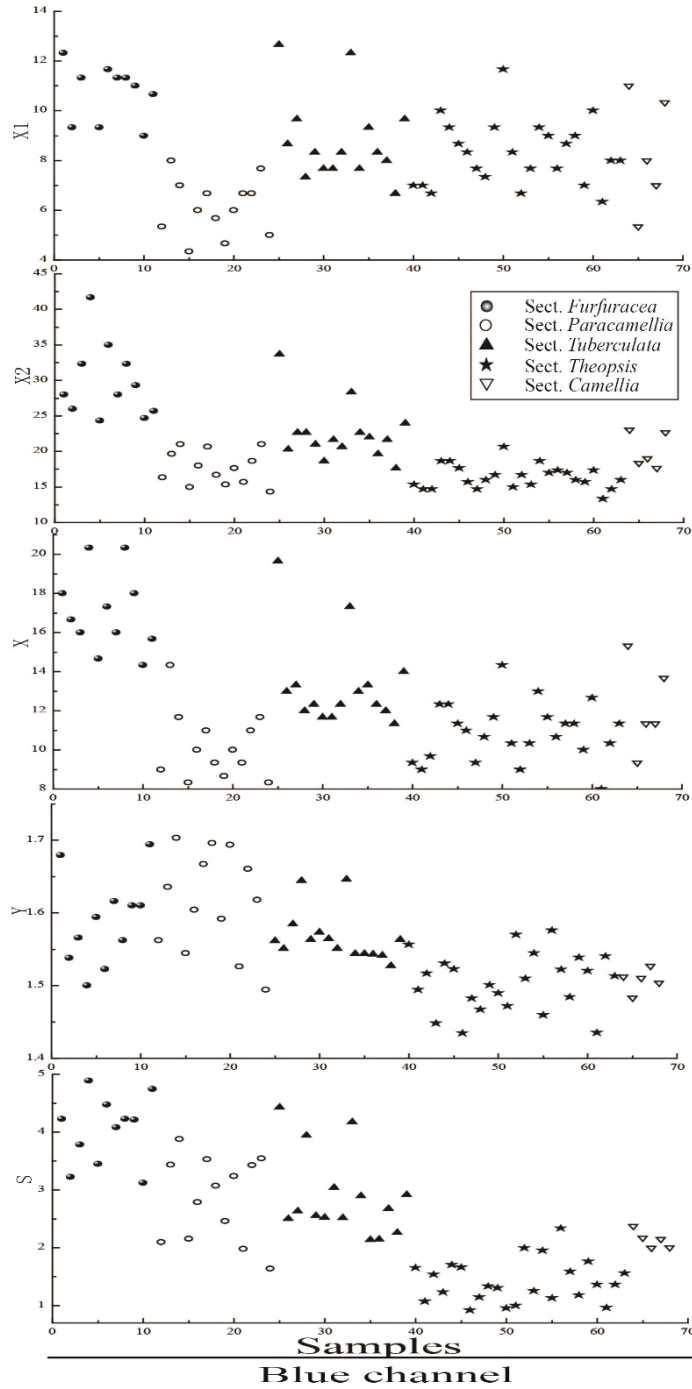


Fig. 4. Scatter plot of fractal parameters used for SVM models in this study. The five fractal parameters (X1, X2, X, Y, S) derived from the samples using **blue** channel are shown. Numbers in the figure correspond to the species numbers in Table 1.

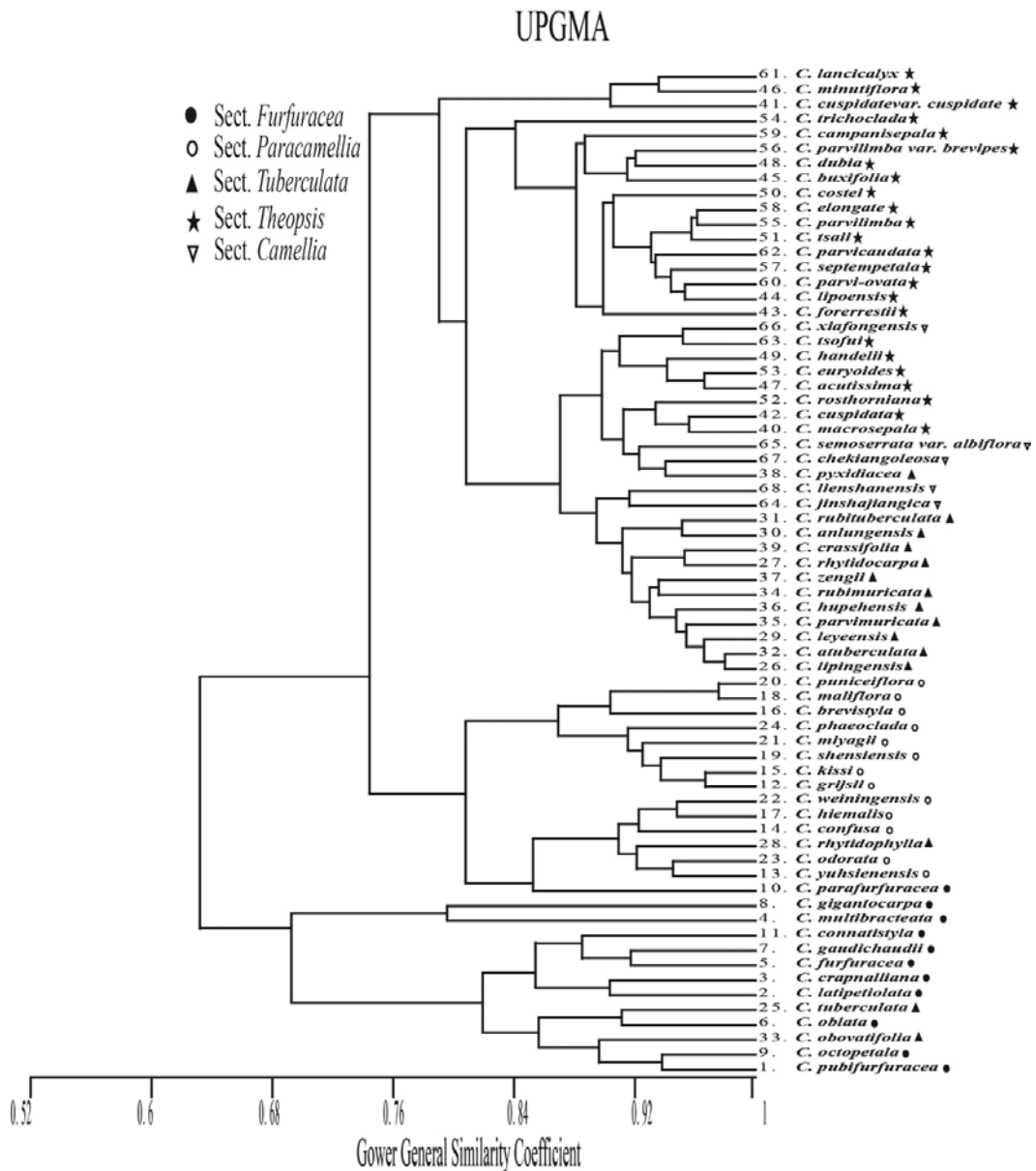


Fig. 5. UPGMA dendrogram of genus *Camellia* based on fractal parameters and RGB intensity values, sect. *Furfuracea* (●), sect. *Paracamellia* (○), sect. *Tuberculata* (▲), sect. *Theopsis* (□), sect. *Camellia* (▼).

sigmoid SVM models, with optimal parameters of C and g are presented in Fig. 7. The RBF SVM classifier offers the best conformance to Chang's classification with 96.88% accuracy rate (sect. *Furfuracea*-100%, sect. *Paracamellia*-100%, sect. *Tuberculata*-85.71%, sect. *Theopsis*-100%, sect. *Camellia*-100%). The only misclassification was in sect. *Tuberculata*, it suggested species number 12 (*C. grijsii*) belongs to sect. *Furfuracea*. Table 3 reveals that the classification results obtained by RBF SVM classifier approach in the training set is 100%, which highlights the good

Table 1. Species assessed, as classified by Chang (1998). Notes: ^{1, 2, 3, 4, 5} represent the sample labels (categories) used in SVM model. Species without parenthesis are training set, which are followed by parenthesis are test set. The numbers in parenthesis is the number of species in test set.

| Samples | | | | |
|--|---|---|---------------------------------|-----------------------------------|
| Sect. <i>Furfuracea</i> ¹ | 1. <i>C. pubifurfuracea</i> | 2. <i>C. latipetioata</i> | 3. <i>C. crapnalliana</i> | 4. <i>C. multibracteata</i> |
| | 5. <i>C. furfuracea</i> | 6. <i>C. oblata</i> | 7. <i>C. gaudichaudii</i> (1) | 8. <i>C. gigantocarpa</i> (2) |
| | 9. <i>C. octopetala</i> (3) | 10. <i>C. parafurfuracea</i> (4) | 11. <i>C. connatistyla</i> (5) | |
| Sect. <i>Paracamellia</i> ² | 12. <i>C. grijsii</i> | 13. <i>C. yuhsienensis</i> | 14. <i>C. confusa</i> | 15. <i>C. kissi</i> |
| | 16. <i>C. brevistyla</i> | 17. <i>C. hiemalis</i> | 18. <i>C. maliflora</i> | 19. <i>C. shensiensis</i> (6) |
| | 20. <i>C. puniceiflora</i> (7) | 21. <i>C. miyagii</i> (8) | 22. <i>C. weiningensis</i> (9) | 23. <i>C. odorata</i> (10) |
| | 24. <i>C. phaeoclada</i> (11) | | | |
| Sect. <i>Tuberculata</i> ³ | 25. <i>C. tuberculata</i> (12) | 26. <i>C. lipingensis</i> (13) | 27. <i>C. rhytidocarpa</i> | 28. <i>C. rhytidophylla</i> |
| | 29. <i>C. leyeensis</i> | 30. <i>C. anlungensis</i> | 31. <i>C. rubituberculata</i> | 32. <i>C. atuberculata</i> |
| | 33. <i>C. obovatifolia</i> | 34. <i>C. rubimuricata</i> | 35. <i>C. parvimuricata</i> | 36. <i>C. hupehensis</i> (15) |
| | | | (14) | |
| | 37. <i>C. zengii</i> (16) | 38. <i>C. pyxidiacea</i> (17) | 39. <i>C. crassifolia</i> (18) | |
| Sect. <i>Theopsis</i> ⁴ | 40. <i>C. macrosepala</i> (19) | 41. <i>C. cuspidatevar. synapidate</i> (20) | 42. <i>C. cuspidata</i> | 43. <i>C. forerrestii</i> |
| | 44. <i>C. lipoensis</i> | 45. <i>C. buxifolia</i> | 46. <i>C. minutiflora</i> | 47. <i>C. acutissima</i> |
| | 48. <i>C. dubia</i> | 49. <i>C. handelii</i> | 50. <i>C. costei</i> | 51. <i>C. tsaii</i> |
| | 52. <i>C. rosthorniana</i> | 53. <i>C. euryoides</i> | 54. <i>C. trichoclada</i> (21) | 55. <i>C. parvilimba</i> (22) |
| | 56. <i>C. parvilimba var. brevipes</i> (23) | 57. <i>C. septempetala</i> (24) | 58. <i>C. elongate</i> (25) | 59. <i>C. campanisepala</i> (26) |
| | 60. <i>C. parvi-ovata</i> (27) | 61. <i>C. lancicalyx</i> (28) | 62. <i>C. parvicaudata</i> (29) | 63. <i>C. tsofui</i> (30) |
| Sect. <i>Camellia</i> ⁵ | 64. <i>C. jinshajiangica</i> | 65. <i>C. semoserrata var. albiflora</i> | 66. <i>C. xiafongensis</i> | 67. <i>C. chekiangoleosa</i> (31) |
| | 68. <i>C. lienshanensis</i> (32) | | | |

Table 2. The RGB intensity values derived from samples used for SVM models in this study.

| | Sect. <i>Furfuracea</i> | | Sect. <i>Paracamellia</i> | | Sect. <i>Tuberculata</i> | | Sect. <i>Theopsis</i> | | Sect. <i>Camellia</i> | |
|---|-------------------------|------------|---------------------------|------------|--------------------------|------------|-----------------------|------------|-----------------------|------------|
| | Range (min-max) | Mean ± SD | Range (min-max) | Mean ± SD | Range (min-max) | Mean ± SD | Range (min-max) | Mean ± SD | Range (min-max) | Mean ± SD |
| R | 15.61-33.39 | 23.54±5.25 | 12.16-17.85 | 14.25±1.77 | 15.52-26.63 | 18.81±3.38 | 13.34-25.08 | 19.22±3.55 | 15.65-19.19 | 17.35±1.26 |
| G | 18.28-39.97 | 27.65±6.10 | 13.52-18.99 | 15.86±2.01 | 17.89-32.94 | 22.54±4.15 | 18.36-34.07 | 25.82±4.31 | 19.69-23.95 | 22.08±1.78 |
| B | 12.54-24.96 | 18.02±3.77 | 9.91-13.76 | 11.55±1.21 | 11.04-20.82 | 14.03±2.49 | 9.6-14.57 | 12.14±1.45 | 10.32-15.68 | 12.80±2.20 |

Table 3. The classification results in the training set of RBF SVM classifier.

| Samples | Sample number | Classification results | | | | | Total accuracy |
|---------------------------|---------------|-------------------------|---------------------------|--------------------------|-----------------------|-----------------------|----------------|
| | | Sect. <i>Furfuracea</i> | Sect. <i>Paracamellia</i> | Sect. <i>Tuberculata</i> | Sect. <i>Theopsis</i> | Sect. <i>Camellia</i> | |
| Sect. <i>Furfuracea</i> | 6 | 6 | 0 | 0 | 0 | 0 | 100% |
| Sect. <i>Paracamellia</i> | 7 | 0 | 7 | 0 | 0 | 0 | |
| Sect. <i>Tuberculata</i> | 8 | 0 | 0 | 8 | 0 | 0 | |
| Sect. <i>Theopsis</i> | 12 | 0 | 0 | 0 | 12 | 0 | |
| Sect. <i>Camellia</i> | 3 | 0 | 0 | 0 | 0 | 3 | |

Table 4. The predicted classification of the polynomial SVM under different degrees with the optimal parameters ($C=16, g=0.5$).

| Subset | Samples Number | Polynomial degree | | | | | | | |
|---------------------------|----------------|-------------------|--------|--------|--------|--------|--------|--------|--------|
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Sect. <i>Furfuracea</i> | 5 | 100% | 100% | 100% | 100% | 100% | 80% | 80% | 80% |
| Sect. <i>Paracamellia</i> | 6 | 100% | 83.33% | 66.67% | 66.67% | 66.67% | 66.67% | 66.67% | 66.67% |
| Sect. <i>Tuberculata</i> | 7 | 71.43% | 71.43% | 57.14% | 57.14% | 57.14% | 57.14% | 57.14% | 57.14% |
| Sect. <i>Theopsis</i> | 12 | 100% | 100% | 100% | 100% | 91.67% | 91.67% | 91.67% | 91.67% |
| Sect. <i>Camellia</i> | 2 | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Total accuracy (%) | | 93.75% | 90.63% | 84.38% | 84.38% | 81.25% | 78.13% | 78.13% | 78.13% |

performance of the RBF SVM classifier. The linear SVM classifier for five sections shows correct classification rate of 90.63% (sect. *Furfuracea*-100%, sect. *Paracamellia*-100%, sect. *Tuberculata*-57.14%, sect. *Theopsis*-100%, sect. *Camellia*-100%), but the sigmoid kernel overall accuracy for the test data set is worse than any other classifiers with only 40.63% (sect. *Furfuracea*-0%, sect. *Paracamellia*-33.33%, sect. *Tuberculata*-0%, sect. *Theopsis*-91.67%, sect. *Camellia*-0%). For polynomial classifiers, in fact, it is a linear classifier when polynomial degree $d = 1$. The classification results of polynomial SVM classifier with different degrees from 2 to 9 are shown in Table 4. The polynomial SVM classifiers with $d = 2$ achieved the best overall classification accuracies (93.75%) of the five sections (sect. *Furfuracea*-100%, sect. *Paracamellia*-100%, sect. *Tuberculata*-71.43%, sect. *Theopsis*-100%, sect. *Camellia*-100%). In addition, the active effect on the classification accuracies was very less when d was greater than 2, with increasing polynomial degree, the classification accuracies take on a descending trend (Table 4).

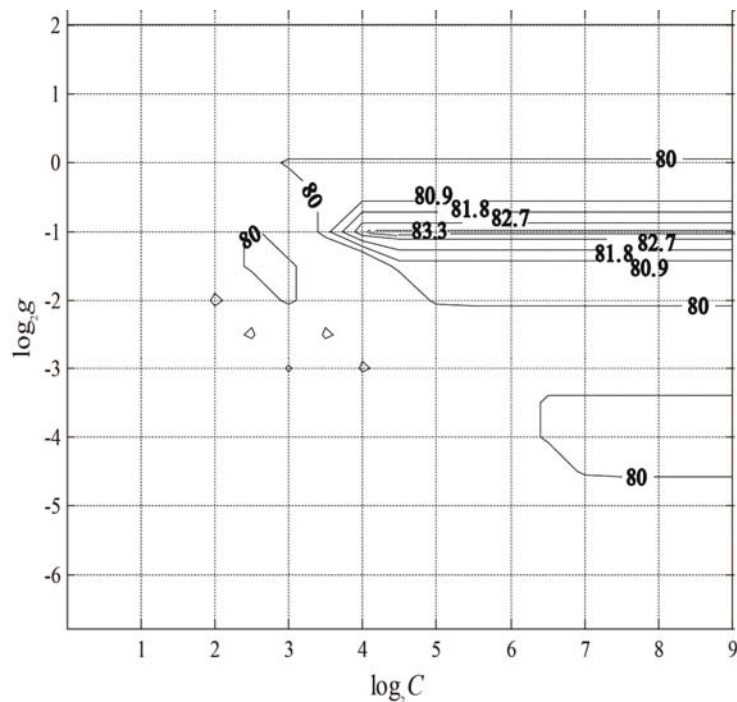


Fig. 6. Average classification accuracy in different kernel parameter (C) and regularization parameter (γ) by cross-validation.

Table 5 .Summary of the supervised techniques, materials, factors, and accuracies for Chang (1998)’s *Camellia* classification.

| Classification techniques | Materials | Factors | Accuracy | Reference | Demerit |
|--|--|---|--|-------------------|--|
| Cluster analysis | 63 species and 2 varieties in 4 sections | Fourier transform infrared data of leaves | 84.7% | Lu et al. 2008a | Expensive |
| Cluster analysis | 21 species from 4 sections | Fourier transform infrared data combined with leaf anatomy | 85.7% | Lu et al. 2008b | Wasting time, money and low-efficiency |
| Particle swarm optimization-aided fuzzy cloud classifier | 24 species from 3 sections | 23 quantitative features cover the characters of flower, fruit and leaf | 98.0% | Lu et al. 2009 | Laborious and time consuming |
| Pattern recognition techniques | 93 species from 5 sections | 31 variables from Leaf morphological and venation characters | LVQ1-ANN for 60% LVQ2-ANN for 91.11% DAN2 for 91.11% SVM for 97.78% | Lu et al. 2012 | Heavy workload |
| Cluster analysis and principal coordinate analysis | 19 species from 2 sections | 28 variables from floral morphology characters | 84.2% | Jiang et al. 2012 | Poor repeatability |
| Back-propagation neural networks | 47 species from 3 sections | 7 leaf anatomy attributes | 86.36% | Jiang et al. 2013 | Time consuming |

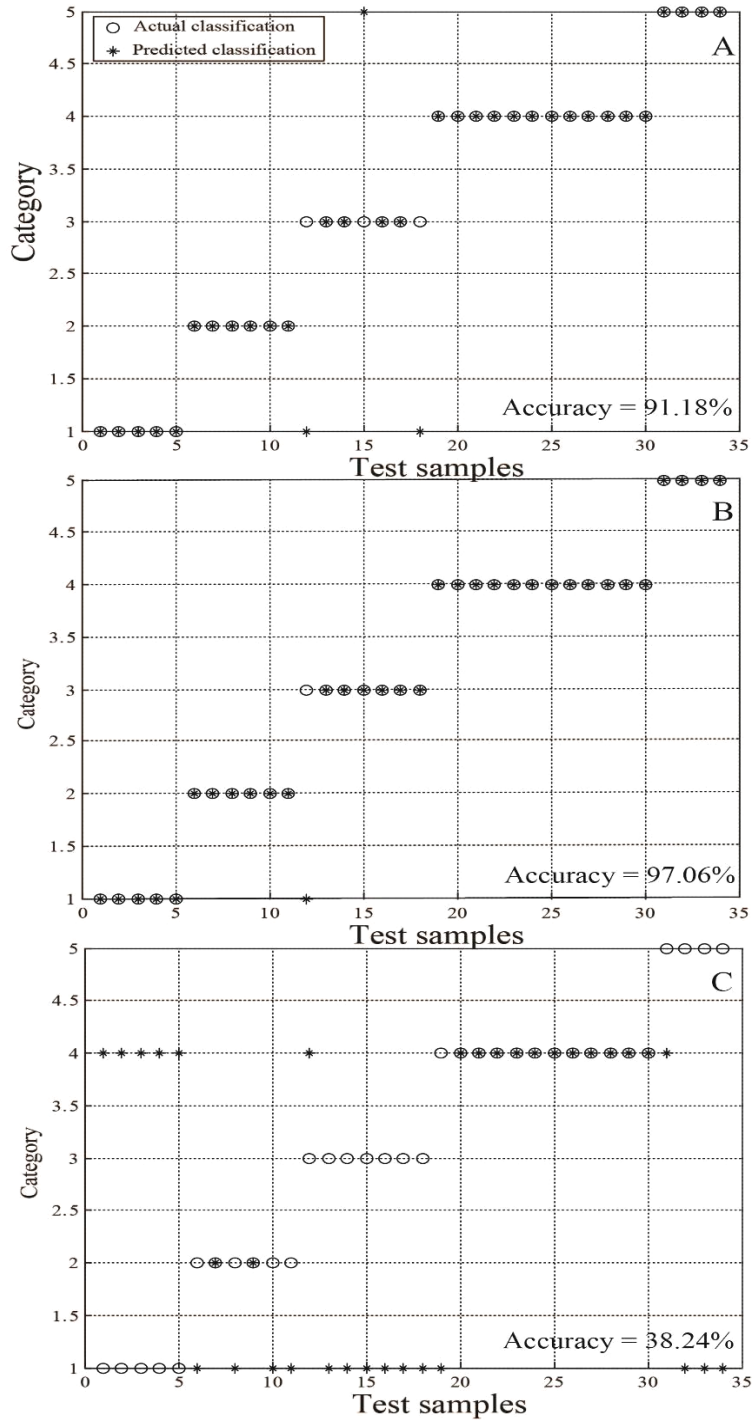


Fig. 7. The classification results of linear (A), RBF (B) and sigmoid (C) SVMs with the optimal parameters.

Discussion

Plant numerical taxonomy applies numerical methods or supervised techniques like SVM in the classification of taxonomic units. It converts the information content of taxa to numerical quantitative and its aim is in its objectivity. Thus, developing a taxonomic toolkit is becoming an indispensable aid in modern systematics. Traditionally, leaf characters have been used as a basis for plant taxonomy and they have been successfully used to solve plant classification problems (Linnaeus, 1753). Contemporary classification especially for genus *Camellia*, have involved use of advanced technology tools. Some examples are, classification within genus level based on simulated annealing aided cloud classifier (Pi *et al.*, 2011); use of genetic information with molecular biotechnology tools; fourier transform infrared spectroscopy (FTIR) combined with shape and anatomy analysis of *Camellia* leaves (Lu *et al.*, 2008b; Shen *et al.*, 2008), which suggested that the chemical method also had important taxonomic significance. However, as shown in Table 5, some of these methods are laborious and expensive, and do not always guarantee satisfactory results. Moreover, a defect common to all the approaches (Table 5) is that they get quantitative features of plant is based on damaging leaves. However, fractal analysis and RGB intensity values combined with support vector machine (SVM) used in our study are not only non-destructive, but are simple, and easily performed. The fractal spectrum was introduced as a botanical identification key by Mugnai *et al.* (2008). Actually, leaf colour is a very special characteristic but often ignored by taxonomists. *Camellia* species are both trees and shrubs, and plant height and leaf feature may interfere with plant photosynthesis. The chlorophyll content in turn is correlated to the leaf colour (Du *et al.*, 2009). Moreover, the long-term evolution of *Camellia* species have made them a stable system, therefore they can be classified based on leaf traits like chlorophyll content.

Chang (1998) and Ming (2000) are two comprehensive floras prominently used by *Camellia* researchers. People often turn to flora to identify a new species; however, traditional information retrieval processes is frequently cumbersome. Further, some basic characteristics can only be manually identified which needs experience and is often subjective. These limitations can be overcome by developing an automated method of plant identification which is rapid and efficient. We have developed an automated method using leaf fractal parameters in SVM model to classify 68 *Camellia* species. The taxonomic results are very encouraging allowing us to achieve accuracy of up to 96.88% using the RBF fractal values. As a modern pattern recognition tool, the SVM is advantageous over other methods like back-propagation artificial neural network (BP-ANN). The common problem with neural networks is the networks structure; BP-ANN may suffer from the over-fitting problem because its approaches are based on the empirical risk minimization principles. Comparatively, the over-fitting can be easily controlled in SVM by choosing a suitable margin to get the best resolution of entire data set (Burges, 1998). In addition, SVM does not need a great quantity of training sets for developing model.

Our results were mostly congruent with Chang's (1998) classification of *Camellia* species with some differences. However, it should be noted that other researchers have also reported deviations from Chang's classification. For example, when our results are compared to *Camellia* classification by Vijayan *et al.* (2009), the general agreement in classification of the 68 *Camellia* species indicates the usefulness of fractal parameters and RGB intensity in detecting phylogenetic relationships. For the plants from sect. *Furfuracea* and sect. *Theopsis*, all collected species from two sections were joined and intermixed respectively (Fig. 5), which is in agreement with the classification by Vijayan *et al.* (2009). In addition, our results support the grouping of *C. yuhsienensis* (No. 13, from sect. *Paracamellia*) and *C. rhytidophylla* (No. 28, from sect. *rhytidophylla*) together as reported by Vijayan *et al.* (2009). This is however different from Chang's (1998) treatment of these two species. Further, as shown in Fig. 5, species from sect.

Paracamellia grouped together, whilst Vijayan *et al.* (2009) taxonomic treatment advocates these species as three clades. In analyzing results from the SVM classifiers, we found that the species number 12 (*C. grijsii*) from sect. *Paracamellia* was incorrectly classified as a species from sect. *Furfuracea* by all SVM classifiers [linear, RBF, sigmoid, and polynomial ($d = 2$) classifiers]. The deviation from this classification needs further investigation to see if this misclassification is due to the underlying algorithm's fitting of the data, or *C. grijsii* really has a close relationship with sect. *Furfuracea*.

In addition, high quality seeds are the key to develop the modern agriculture, it is necessary to select good seed varieties for improvement of crops yield. An elite variety with greater benefits should replace the variety with inferior quality seeds. Bacchetta *et al.* (2011) identified Sardinian species of *Astragalus* section *Melanocercis* by seed image analysis. Developing countries are still using traditional manual seed separation method. In this context, the application of SVM based on fractal leaf parameters analysis (FA) and leaf red, green, and blue (RGB) intensity values used in the present study is not only proposed as a complementary method for botanical identification, but also proposed as a modern method of good seed selection. The SVM-FA-RGB system is very simple to establish and requires only a personal computer and an optical scanner. Therefore it could potentially replace old methods that are complicated, labour-intensive and expensive.

Conclusion

We have developed a system for automatic binary classification of 68 *Camellia* species into five sections based on SVM and discussed the important features of this classification. The hierarchical dendrogram based on fractal parameters and RGB intensity values confirms the morphological classification of the five sections proposed by Chang's (1998) research. The linear, polynomial ($d = 2$), RBF SVM classifier with $C = 16$, $g = 0.5$ work well in the classification of the genus *Camellia*. Especially RBF SVM classifier showed encouraging results that obtaining a correct classification rate of 96.88%. The above results indicate that fractal parameters and RGB intensity values analysis using SVM, particularly RBF kernel, can be effectively used to distinguish the *Camellia* at genus level, or even at higher taxa level. In addition, the SVM-FA-RGB system could be used to select high quality seeds in agriculture breeding programs.

Acknowledgements

The study was partially supported by grants from the Science and Technology research Plan of Jinhua city, China (No. 2009-2-020). The authors thank Mr. Bin Wang, Ms. Jingjing Lou and Ms. Zhihui Zhu (Zhejiang Normal University) for collecting living species and assistance with the experiments.

References

- Azzarello, E., Mugnai, S., Pandolfi, C., Masi, E., Marone, E. and Mancuso, S. 2009. Comparing image (fractal analysis) and electrochemical (impedance spectroscopy and electrolyte leakage) techniques for the assessment of the freezing tolerance in olive. *Tress-Structure Function* **23**: 159–167.
- Bacchetta, G., Fenu, G., Grillo, O., Mattana, E. and Venora, G. 2011. Identification of Sardinian species of *Astragalus* section *Melanocercis* (Fabaceae) by seed image analysis. *Ann. Bot. Fenn.* **48**: 449–454.
- Barthlott, W., Wiersch, S., Colic, Z. and Koch, K. 2009. Classification of trichome types within species of the water fern *Salvinia*, and ontogeny of the eggbeater trichomes. *Botany* **87**: 830–836.
- Burges, C.J.C. 1998. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.* **2**: 121–167.

- Chang, H.T. 1998. Theaceae. In: Flora of China Editorial Committee (Ed.), Vol. **49**. Flora of China. Science Press, Beijing.
- Chen, L., Wang, S. and Nelson, M. 2005. Genetic diversities within *Camellia* species confirmed by random amplified polymorphic DNA (RAPD) markers. *J. Am. Soc. Hortic. Sci.* **40**: 993–1147.
- Chen, Q.S., Guo, Z.M. and Zhao, J.W. 2008. Identification of green tea's (*Camellia sinensis* L.) quality level according to measurement of main catechins and caffeine contents by HPLC and support vector classification pattern recognition. *J. Pharmaceut. Biomed.* **48**: 1321–1325.
- Chen, Q.S., Zhao, J.W., Fang, C.H. and Wang, D.M. 2007. Feasibility study on identification of green, black and Oolong teas using near-infrared reflectance spectroscopy based on support vector machine (SVM). *Spectrochim. Acta. A* **66**: 568–574.
- Cortes, C. and Vapnik, V. 1995. Support-vector networks. *Mach. Learn.* **20**: 273–297.
- Du, P., Ling, Y.H., Sang, X.C., Zhao, F.M. and Xie, R. 2009. Gene mapping related to yellow green leaf in a mutant line in rice (*Oryza sativa* L.). *Genes Genom.* **31**: 165–171.
- Gao, J.Y., Parks, C.R. and Du, Y.Q. 2005. Collected species of the genus *Camellia*, an illustrated outline. Zhejiang Science and Technology Press, Hangzhou.
- Gunasekaran, S., Paulsen, M.R. and Shove, G.C. 1985. Optical methods for nondestructive quality evaluation of agricultural and biological materials. *J. Agri. Eng. Res.* **32**: 209–241.
- Harrell, R.C., Slaughter, D.C. and Adsit, P.D. 1989. A fruit-tracking system for robotic harvesting. *Mach. Vision Appl.* **2**: 69–80.
- Howaith, M.S., Searcy, S.W. and Birth, G.S. 1990. Reflectance characteristics of fresh market carrots. *T. ASABE* **33**: 961–964.
- Jack, L.B. and Nandi, A.K. 2002. Fault detection using support vector machines and artificial neural networks: augmented by genetic algorithms. *Mech. Syst. Signal Pr.* **16**: 373–390.
- Jain, A.K., Murty, M.N. and Flynn, P.J. 1999. Data Clustering: A Review. *ACM Comput. Surv.* **31**: 264–323.
- Jiang, B., Peng, Q.F., Shen, Z.G., Moller, M., Pi, E.X. and Lu, H.F. 2010. Taxonomic treatments of *Camellia* (Theaceae) species with secretory structures based on integrated leaf characters. *Plant Syst. Evol.* **290**: 1–20.
- Jiang, W., Nitin, M., Jiang, B., Zheng, Y.P., Hong, S.S. and Lu, H.F. 2012. Floral morphology resolves the taxonomy of *Camellia* L. (Theaceae) Sect. *Oleifera* and Sect. *Paracamellia*. *Bangladesh J. Plant Taxon.* **19**: 155–165.
- Kumar, K.S., Jayabarathi, T. and Naveen, S. 2011. Fault identification and location in distribution systems using support vector machines. *Eur. J. Sci. Res.* **51**: 53–60.
- Lin, X.Y., Peng, Q.F., Lu, H.F., Du, Y.Q. and Tang, B.Y. 2008. Leaf anatomy of *Camellia* Sect. *Oleifera* and Sect. *Paracamellia* (Theaceae) with reference to their taxonomic significance. *J. Syst. Evol.* **46**: 183–193.
- Linnaeus, C. 1753. *Species Plantarum*. Stockholm: Salvias.
- Lu, H.F., Jiang, B., Shen, Z.G., Shen, J.B., Peng, Q.F. and Cheng, C.G. 2008a. Comparative leaf anatomy, FTIR discrimination and biogeographical analysis of *Camellia* section *Tuberculata* (Theaceae) with a discussion of its taxonomic treatments. *Plant Syst. Evol.* **274**: 223–235.
- Lu, H.F., Shen, J.B., Lin, X.Y. and Fu, J.L. 2008b. Relevance of Fourier transform infrared spectroscopy and leaf anatomy for species classification in *Camellia* (Theaceae). *Taxon* **57**: 1274–1288.
- Lu, H.F., Pi, E.X., Peng, Q.F., Wang, L.L. and Zhang, C.J. 2009. A particle swarm optimization-aided fuzzy cloud classifier applied for plant numerical taxonomy based on attribute similarity. *Expert Syst. Appl.* **36**: 9388–9397.
- Lu, H.F., Zheng, H., Hu, Y., Lou, H.Q. and Kong, X.C. 2011. Bruise detection on red bayberry (*Myrica rubra* Sieb. & Zucc.) using fractal analysis and support vector machine. *J. Food Eng.* **104**: 149–153.
- Lu, H., Jiang, W., Ghiassi, M., Lee, S. and Nitin, M. 2012. Classification of *Camellia* (Theaceae) species using leaf architecture variations and pattern recognition techniques. *PLoS ONE* **7**(1): e29704.

- Mancuso, S. 1999. Fractal geometry-based image analysis of grapevine leaves using the box counting algorithm. *Vitis* **38**: 97–100.
- Mancuso, S. 2002. Discrimination of grapevine (*Vitis vinifera* L.) leaf shape by fractal spectrum. *Vitis* **41**: 137–142.
- Mancuso, S., Nicese, F.P. and Azzarello, E. 2003. The fractal spectrum of the leaves as a tool for measuring frost hardness in plants. *J. Hortic. Sci. Biotech.* **78**: 610–616.
- Ming, T.L. 2000. Monograph of the Genus *Camellia*. Yunnan Science and Technology Press, Kunming.
- Mugnai, S., Pandolfi, C., Azzarello, E., Masi, E. and Mancuso, S. 2008. *Camellia japonica* L. genotypes identified by an artificial neural network based on phyllometric and fractal parameters. *Plant Syst. Evol.* **270**: 95–108.
- Pandolfi, C., Messina, G., Mugnai, S., Azzarello, E. and Masi, E. 2009. Discrimination and identification of morphotypes of *Banksia integrifolia* (Proteaceae) by an Artificial Neural Network (ANN), based on morphological and fractal parameters of leaves and flowers. *Taxon* **58**: 925–933.
- Pi, E.X., Lu, H.F., Jiang, B., Huang, J., Peng, Q.F. and Lin, X.Y. 2011. Precise plant classification within genus level based on simulated annealing aided cloud classifier. *Expert Syst. Appl.* **38**: 3009–3014.
- Pi, E.X., Peng, Q.F., Lu, H.F., Shen, J.B., Du, Y.Q., Huang, F.L. and Hu, H. 2009. Leaf morphology and anatomy of *Camellia* section *Camellia* (Theaceae). *Bot. J. Linn. Soc.* **159**: 456–476.
- Plotze, R.D.O., Falvo, M., Pádua, J.G., Bernacci, C. and Vieira, M. 2005. Leaf shape analysis using the mutiscale Minkowski fractal dimension, a new morphometric method: A study with *Passiflora* (Passifloraceae). *Can. J. Bot.* **83**: 287–301.
- Pontil, M. and Verri, A. 1998. Support vector machines for 3-D object recognition. *IEEE T Pattern Anal.* **20**: 637–646.
- Sealy, J.R. 1958. A Revision of the Genus *Camellia*. The Royal Horticultural Society, London.
- Shen, J.B., Lu, H.F., Peng, Q.F., Zheng, J.F. and Tian, Y.M. 2008. FTIR spectra of *Camellia* sect. *Oleifera*, sect. *Paracamellia*, and sect. *Camellia* (Theaceae) with reference to their taxonomic significance. *J. Syst. Evol.* **46**: 194–204.
- Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. NY, New York, USA: Springer-Verlag.
- Vijayan, K., Zhang, W.J. and Tsou, C.H. 2009. Molecular taxonomy of *Camellia* (Theaceae) inferred from nrITS sequences. *Am. J. Bot.* **96**: 1348–1360.
- Yang, Z.R. and Lin, Q. 2005. Comparative morphology of the leaf epidermis in *Schisandra* (Schisandraceae). *Bot. J. Linn. Soc.* **148**: 39–56.
- Ye, P. and Weng, G.R. 2011. Classification and recognition of plant leaf based on neural networks. *Key Eng. Mater.* **464**: 38–42.
- Zheng, H., Lu, H.F., Zheng, Y.P., Lou, H.Q. and Chen, C.Q. 2010. Automatic sorting of Chinese jujube (*Zizyphus jujuba* Mill. cv. 'hongxing') using chlorophyll fluorescence and support vector machine. *J. Food Eng.* **101**: 402–408.

(Manuscript received on 7 September 2016; revised on 10 October 2016)

Appendix 1

Table A1. Collection localities and vouchers of studied specimens

| Taxon; vouchers; accession number (all specimens from China, Zhejiang, Jinhua International Camellia Species Garden) |
|--|
| <p><i>C. pubifurfuracea</i>; H.E. Tian; 200610170901 (ZJNU). <i>C. latipetiolata</i>; H.E. Tian; 200610171001 (ZJNU). <i>C. crapnalliana</i>; H.E. Tian; 200610171101 (ZJNU). <i>C. multibracteata</i>; H.E. Tian; 200610171201 (ZJNU). <i>C. furfuracea</i>; H.E. Tian; 200610171301 (ZJNU). <i>C. oblate</i>; Q.F. Peng; 200610221401 (ZJNU). <i>C. gaudichaudii</i>; H.E. Tian; 200610171501 (ZJNU). <i>C. gigantocarpa</i>; Q.F. Peng; 200610221601 (ZJNU). <i>C. octopetala</i>; Q.F. Peng; 200610221701 (ZJNU). <i>C. parafurfuracea</i>; H.E. Tian; 200610171801 (ZJNU). <i>C. connatistyla</i>; H.E. Tian; 200610171901 (ZJNU). <i>C. grijsii</i>; J.B. Shen; 200701122001 (ZJNU). <i>C. yuhsienensis</i>; J.B. Shen; 200701122102 (ZJNU). <i>C. confusa</i>; J.B. Shen; 200612232201 (ZJNU). <i>C. kissi</i>; J.B. Shen; 200701122302 (ZJNU). <i>C. brevistyla</i>; J.B. Shen; 200612232501 (ZJNU). <i>C. hiemalis</i>; J.B. Shen; 200612232602 (ZJNU). <i>C. maliflora</i>; J.B. Shen; 200701122802 (ZJNU). <i>C. shensiensis</i>; J.B. Shen; 200701122901 (ZJNU). <i>C. puniceiflora</i>; J.B. Shen; 200701123001 (ZJNU). <i>C. miyagii</i>; J.B. Shen; 200701123001 (ZJNU). <i>C. weiningensis</i>; S.S. Hong; 2011040139 (ZJNU). <i>C. odorata</i>; X.Y. Lin; 20070128350 (ZJNU). <i>C. phaeoclada</i>; J.B. Shen; 200701123602 (ZJNU). <i>C. tuberculata</i>; Q.F. Peng; 05112301 (ZJNU). <i>C. lipingensis</i>; B. Wang & W. Jiang; 2011070101 (ZJNU). <i>C. rhytidocarpa</i>; B. Jiang; 06111201 (ZJNU). <i>C. rhytidophylla</i>; B. Wang & W. Jiang; 2011070102 (ZJNU). <i>C. leyeensis</i>; B. Wang & W. Jiang; 2011070103 (ZJNU). <i>C. anlungensis</i>; Q.F. Peng; 05112302 (ZJNU). <i>C. rubituberculata</i>; B. Jiang & Q.F. Peng; 06062302 (ZJNU). <i>C. parvimuricata</i>; B. Jiang & Q.F. Peng; 06090404 (ZJNU). <i>C. hupehensis</i>; Q.F. Peng; 05112303 (ZJNU). <i>C. zengii</i>; B. Jiang & Q.F. Peng; 06061201 (ZJNU). <i>C. pyxidiacea</i>; B. Jiang & Q.F. Peng; 05112304 (ZJNU). <i>C. crassifolia</i>; S.S. Hong; 2011040104 (ZJNU). <i>C. macrosepala</i>; S.S. Hong; 2011040107 (ZJNU). <i>C. cuspidate</i> var. <i>synapidate</i>; B. Wang & W. Jiang; 2011070104 (ZJNU). <i>C. cuspidate</i>; B. Wang & W. Jiang; 2011070105 (ZJNU). <i>C. forerrestii</i>; S.S. Hong; 2011040109 (ZJNU). <i>C. lipoensis</i>; S.S. Hong; 2011040110 (ZJNU). <i>C. buxifolia</i>; S.S. Hong; 2011040111 (ZJNU). <i>C. minutiflora</i>; S.S. Hong; 2011040112 (ZJNU). <i>C. acutissima</i>; S.S. Hong; 2011040114 (ZJNU). <i>C. dubia</i>; B. Wang & W. Jiang; 2011070106 (ZJNU). <i>C. handelii</i>; S.S. Hong; 2011040115 (ZJNU). <i>C. costei</i>; S.S. Hong; 2011040116 (ZJNU). <i>C. tsaii</i>; B. Wang & W. Jiang; 2011070107 (ZJNU). <i>rosthorniana</i>; S.S. Hong; 2011040117 (ZJNU). <i>C. euryoides</i>; S.S. Hong; 2011040118 (ZJNU). <i>C. trichoclada</i>; B. Wang & W. Jiang; 2011070108 (ZJNU). <i>C. parvilimba</i>; S.S. Hong; 2011040119 (ZJNU). <i>C. parvilimba</i> var. <i>brevipes</i>; S.S. Hong; 2011040120 (ZJNU). <i>C. septempetala</i>; S.S. Hong; 2011040121 (ZJNU). <i>C. elongate</i>; S.S. Hong; 2011040122 (ZJNU). <i>C. campanisepala</i>; S.S. Hong; 2011040123 (ZJNU). <i>C. parvi-ovata</i>; S.S. Hong; 2011040124 (ZJNU). <i>C. lancicalyx</i>; S.S. Hong; 2011040125 (ZJNU). <i>C. parvicaudata</i>; B. Wang & W. Jiang; 2011070109 (ZJNU). <i>C. tsofui</i>; S.S. Hong; 2011040126 (ZJNU). <i>C. jinshajiangica</i>; B. Wang & W. Jiang; 2011070110 (ZJNU). <i>C. semoserrata</i> var. <i>albiflora</i>; B. Wang & W. Jiang; 2011070111 (ZJNU). <i>C. xiafongensis</i>; B. Wang & W. Jiang; 2011070111 (ZJNU). <i>C. chekiangoleosa</i>; Q.F. Peng & B. Jiang; 2006101111 (ZJNU). <i>C. lienshanensis</i>; B. Wang & W. Jiang; 2011070112 (ZJNU)</p> |