Acta Biologica Turcica 32 (2) 90-95 2019 Received 16 Jun. 2018; Accepted 26 Jan. 2019

ACTA BIOLOGICA TURCICA

© 1950-1978 Biologi, Türk Biologi Dergisi, Türk Biyoloji Dergisi, Acta Biologica

E-ISSN: 2458-7893, http://www.actabiologicaturcica.com

Original research

An unsupervised classification and feature selection approach for discrimination of *Ailanthus altissima* (Mill.) Swingle tree leaves

Çağlar CENGIZLER

Çukurova University, Faculty of Engineering and Architecture, Department of Biomedical Engineering, Adana, Turkey

e-mail: ccengizler@cu.edu.tr

Abstract: An exhaustive search approach is presented for automated discrimination of *Ailanthus altissima* (Mill.) Swingle tree leaves in this study. Experimental setup is consisting of a total of 20 different tree specimens and 735 leaf images taken from Leafsnap Dataset where a binary mask is extracted from each of the leaves. 10 salient features defining shape and morphology are extracted from each of these masks. In this study, it was aimed to evaluate all different combinations of these features as subsets to find an optimal feature set for clustering of *A. altissima* tree leaves. Accordingly, two widely known unsupervised data clustering methods, Fuzzy C-means and K-Means are implemented as classifier. Multiclass and two class discrimination experiments are achieved via these methods and F-Score is utilized for objective evaluation of the performance. Performed exhaustive search revealed the best combination of extracted features for unsupervised clustering based classification of the leaves. Additionally, experiments show that, evaluated clustering methods are functioning promisingly and they may discriminate *A. altissima* tree leaves with high accuracy and sensitivity.

Keywords: *Ailanthus altissima* (Mill.) Swingle, Classification, Unsupervised, Clustering, Feature Selection

Citing: Cengizler, Ç. 2019. An unsupervised classification and feature selection approach for discrimination of Ailanthus altissima (Mill.) Swingle tree leaves. *Acta Biologica Turcica, 32*(2): 90-95.

Introduction

Ailanthus altissima (Mill.) Swingle tree widely known as tree of heaven from the family Simaroubaceae Genus *Ailanthus* is a highly invasive and fast growing species which is originating from china (Heisey, 1996). Automated classification of that tree would be a necessity due to its highly invasive character and applications in Chinese traditional medicine (Zhao et al., 2005).

Classification based on examination of leaf images is one of the low-cost solutions (Sladojevic et al., 2016). Accordingly imaging leaves and discriminate them according to extracted measures would be an optimized solution to automated classification of tree species (Fu et al., 2004).

It was aimed to classify *A. altissima* according to leaf characteristics in this study. Automated discrimination of extracted features defining characteristics of leaves would be accomplished by both supervised and unsupervised machine learning approaches. For example, probabilistic neural networks are utilized as a supervised approach in the previous literature (Kadir et al., 2013). Also generalized softmax perceptron model is previously utilized for classification leaves of the sunflowers (Arribas et al., 2011). Most of the supervised methods require a training stage with an extra teaching dataset (Kotsiantis et al., 2007). An unsupervised approach which is based on clustering extracted features is utilized in this study. Therefore, proposed system completes classification of shape and morphology based features without any training. Ten features are extracted from each of the leaves. An exhaustive search is performed to determine best subset of these features for classification of *A. altissima.* Accordingly, all combinations of these features are examined with implemented classifiers. Two of the well-known clustering mechanisms, K-means and Fuzzy C-means are utilized to examine and compare the performance of clustering approaches. Both of them are organizing numerical data into clusters which consist of relatively similar elements of features (Bezdek at al., 1984) (Jain, 2010). Performance of the implemented algorithm is judged by F-score as some objective criteria.

The rest of this paper is organized as follows. Section 2 introduces the utilized data set, extracted features, feature selection and discrimination process. Section 3 presents the results of the experimental setup and Section 4 is consist of discussions of presented results and performance evaluation. Finally, Section 5 presents the conclusion.

Materials and Methods

It was aimed to present a feature selection and evaluation methodology in this study. Accordingly, experimental classifications are performed on previously extracted features where feature extraction stage is followed by data clustering based exhaustive search process. It should be noted that evaluation of each feature subset is achieved by test classifications. Fundamental stages of the approach are shown in Figure 1.

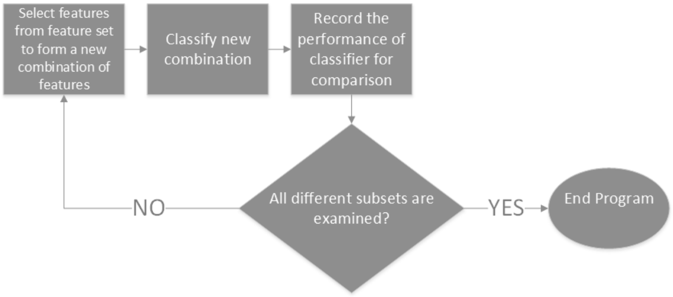
Select features from feature set to form a new combination of features
Classify new combination
Record the performance of classifier for comparison
All different subsets are examined?
NO
YES
End Program

Figure 1. Fundamental stages of the proposed exhaustive methodology are given in block diagram form.

Utilized Data

Utilized data set consist of 735 leaf images and their binary masks. All data is taken from leafsnap database which covers tree species from the Northeastern United States (Kumar et al., 2012). Twenty species including *A. altissima* are collected in our experimental dataset. Names of the species classified in this study is given in Table 1 with total leaf image counts.

All leaf images have a binary mask in the data set defining its area and outline. A sample image and its mask is shown in Figure 2.

Table 1. Classified species in the study.

Specimen Name	Leaf Count
Abies nordmanniana	35
Acer rubrum	45
Ailanthus altissima	16
Betula lenta	11
Broussonettia papyrifera	54
Catalpa speciosa	72
Cornus kousa	15
Fagus grandifolia	31
Ficus carica	45
Ginkgo biloba	21
Liriodendron tulipifera	51
Magnolia denudata	60
Malus floribunda	31
Oxydendrum arboreum	47
Picea pungens	49
Prunus subhirtella	47
Quercus bicolor	21
Sassafras albidum	20
Tilia americana	15
Ulmus rubra	49

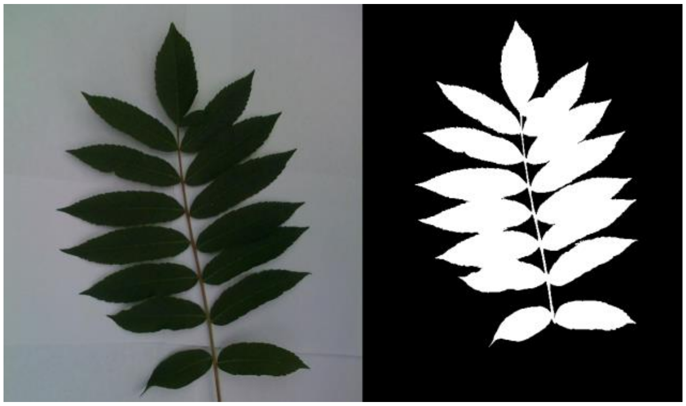

Figure 2. *Ailanthus altissima* leaf image from data set (left) and its binary mask (right).

Extracted Features

Ten features are extracted from each of the binary masks. These features are shape and morphology based numerical measurements which are defining the shape character of the specimen. A list of the extracted features and their cluster centroids with standard deviations for *A. altissima* tree is given in Table 2.

A brief explanation of each feature introduced in Table 2 are given below.

Table 2. Extracted features with cluster centroids and standard deviations for *Ailanthus altissima* (SD: Standard Deviation)

Feature Name	Centroid	SD
Area	65449,75	18432,44
Convex Area	110566,75	27091,21
Eccentricity	0,79	0,07
Major Axis Length	451,44	68,04
Minor Axis Length	264,3	36,41
Major Axis Length, Minor Axis Length Ratio	1,72	0,27
Perimeter	3523,38	895,76
Equivalent Diameter	285,73	42,41
Extent	0,435	0,06

Area: Number of pixels inside the leaf region.

Convex Area: Number of pixels inside the smallest convex polygon that can contain the leaf region.

Eccentricity: It is a scalar value within a range of 0 and 1. It is measured by division of the distance between the foci of the ellipse (an ellipse which has the same second-moments as the region) and its major axis length.

Major Axis Length: Magnitude of major axis of the ellipse in pixels that has the same second central moments as the leaf.

Minor Axis Length: Magnitude of minor axis of the ellipse in pixels that has the same second central moments as the leaf.

Major Axis Length, Minor Axis Length Ratio: Ratio of minor and major axes in pixels.

Perimeter: Number of pixels that forms the boundary of the leaf.

Equivalent Diameter: Diameter of a circle which has the same area as the leaf.

Extent: Ratio of number of pixels inside the leaf area and bounding mask.

Solidity: Ratio of area of the leaf region and convex area.

Classification Stage

Two experimental classification setups are utilized in the study. First setup involves a multiclass classification where *A. altissima* leaves are accepted as positive samples and features of all other specimen as negatives (Platt et al., 2000). Accordingly, algorithm tried to discriminate leaves of *A. altissima* tree from all other leaves.

Second setup involves one versus one classification experiments where algorithm discriminates leaves of *A. altissima* tree from only one specimen at once (Khan and Madden, 2009).

All of the classification tasks are completed with 1023 different feature subsets which are derived by combining 10 previously extracted features. It was aimed to find best feature subsets with both of the setups while evaluating the clustering performance.

Two well-known clustering algorithms are implemented for all classification tasks and each method is experimented with same data for comparison.

One of the classifiers implemented for examining the effectiveness of each feature subset is K-means. It is operating for partition n observations into certain number of clusters where observations belong to the cluster with the nearest mean (Kanungo et al., 2002). It is possible to formulate all observations by:

$$X = \{x_1,x_2,x_3,........,x_n\} \quad (1)$$

and centers by:

$$V = \{v_1,v^2,.......,v_c\} \quad (2)$$

With respect to (1) and (2) cluster centers are calculated by:

$$v_i = (\frac{1}{c_i})\sum_{j=1}^{c_i} x_i \quad (3)$$

where ‘c_i’ stands for the number of observations in i_{th} cluster.

Clustering process starts with random cluster centers then distance between each observation and cluster centers is calculated. Following by, each observation is assigned a cluster center according to its distance. Nearest cluster center should be chosen at that point. Then cluster centers are recalculated. K-means clustering algorithm repeats the calculating distances until minimizing the within-cluster sum of squares. which is formulated by:

$$j(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(\|x_i - v_j\|)^2 \quad (4)$$

In addition to K-means another well-known clustering approach Fuzzy C-means (FCM) is implemented in the study. FCM algorithm is based on similar principles. However, in difference to K-means, FCM operates on membership degrees of observations (Cannon et al., 1986). Objective function of FCM with membership degree is given below:

$$j_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2 ; 1 < m < \infty \quad (5)$$

where, u_{ij} is the degree of membership, x_i is i_{th} observation of data, and c_j is center of the j_{th} cluster. Accordingly, u_{ij} is calculated by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad (6)$$

Where center of the clusters is calculated by:

$$c_j = \frac{\sum_{j=1}^{N} u_{ij}^m . x_i}{\sum_{j=1}^{N} u_{ij}^m} \quad (7)$$

Results

Multiclass and two class discrimination setups are utilized in the study. Clustering performance is measured by several parameters in both of these setups.

These are accuracy:

$$\frac{Tp + Tn}{(N)} \quad (8)$$

sensitivity:

$$\frac{Tp}{(Tp + Fp)} \quad (9)$$

recall:

$$\frac{Tp}{(Tp + Fn)} \quad (10)$$

and F-Score:

$$F_{score} = 2 \frac{Sensitivity * Recall}{(Sensitivity + Recall)} \quad (11)$$

Where Tp, Tn, Fp, Fn and N are true positive, true negative, false positive false negative and number of observations respectively (Goutte and Gaussier, 2005).

All specimens are evaluated with multiclass setup which means each of them is classified from a pool of all other specimens by both of the clustering methods. A total of 1023 different feature subsets are evaluated for each of the specimens are results are given in Table 3.

Table 3. Multiclass classification results of both clustering methods are presented for each specimen

Specimen	Best K-means	Best Fuzzy C-means
Ailanthus altissima	0,983	0,945
Picea pungens	0,973	0,973
Betula lenta	0,965	0,93
Abies nordmanniana	0,963	0,963
Cornus kousa	0,962	0,927
Tilia americana	0,962	0,927
Sassafras albidum	0,958	0,924
Ulmus rubra	0,958	0,905
Ginkgo biloba	0,958	0,923
Quercus bicolor	0,958	0,923
Malus floribunda	0,952	0,917
Fagus grandifolia	0,95	0,915
Ficus carica	0,947	0,903
Acer rubrum	0,941	0,908
Oxydendrum arboreum	0,938	0,902
Prunus subhirtella	0,938	0,902
Liriodendron tulipifera	0,935	0,899
Broussonettia papyrifera	0,934	0,897
Magnolia denudata	0,928	0,891
Catalpa speciosa	0,919	0,881

With respect to Table 3, best classification performance for multiclass discrimination is achieved with K-means method for *A. altissima.* Also K-means is functioning better in the 90% of all of the multiclass tests. In addition to Table 3, best feature combinations graded with highest F-score are given for *A. altissima* in Table 4.

Moreover, leaves of *A. altissima* tree are classified from only one specimen at once. Results of clustering performance are given in Table 5 for each of the specimens.

Table 4. Best Combination of features classified with k-means are given with F-score, accuracy and sensitivity.

Feat.1	Feat.2	Feat.3	Feat.4	Feat.5	Feat.6	Feat.7	Feat.8	Feat.9	Feat.10	F-Score	Accuracy	Sensitivity
0	0	1	1	1	1	1	0	0	0	0.983	0,967	0,968
1	0	0	1	1	1	1	0	0	0	0.983	0,967	0,968
0	0	0	1	1	1	1	0	0	0	0.983	0,967	0,968
0	0	1	0	1	1	1	0	0	0	0.983	0,967	0,968
0	1	1	1	0	1	1	0	0	0	0.983	0,967	0,968
1	0	0	1	0	1	1	0	0	0	0.983	0,967	0,968
1	1	1	0	0	1	1	0	0	0	0.983	0,967	0,968
1	1	1	0	1	0	1	0	0	0	0.983	0,967	0,968
1	0	0	0	1	0	1	0	0	0	0.983	0,967	0,968
0	1	1	1	0	0	1	0	0	0	0.983	0,967	0,968
0	1	1	0	0	0	1	0	0	0	0.983	0,967	0,968
0	1	0	0	0	0	1	0	0	0	0.983	0,967	0,968
1	0	0	0	0	0	1	0	0	0	0.983	0,967	0,968

Table 5. Best two class classification performance scores of k-means and fuzzy c-means for all specimens are given.

Specimen	Max. FCM	Max. KM
Abies nordmanniana	1	1
Betula lenta	1	1
Catalpa speciosa	1	1
Cornus kousa	1	1
Fagus grandifolia	1	1
Ficus carica	1	1
Ginkgo biloba	1	1
Liriodendron tulipifera	1	1
Magnolia denudata	1	1
Oxydendrum arboreum	1	1
Picea pungens	1	1
Prunus subhirtella	1	1
Quercus bicolor	1	1
Sassafras albidum	1	1
Tilia americana	1	1
Broussonettia papyrifera	0,99	0,99
Ulmus rubra	0,989	0,989
Acer rubrum	0,988	0,988
Malus floribunda	0,966	0,966

Discussion

Results presented in Table 3, indicate that the specimen with the highest multiclass classification success is *A. altissima.* According to Table 4, both of the methods would be effective. However, K-means functions better at 90% of the classifications.

All subsets of extracted features are evaluated with both of the methods. Table 4. is presenting best resulting subsets with K-means. It would be possible to interpret from the table that, three of the most distinctive features for *A. altissima* may be, perimeter, major axis length and Eccentricity.

In addition to multiclass classification tests also one versus one discrimination is achieved in the study. Results given in Table 5. are indicating that both of the FCM and K-means are functioning well for two class discrimination of *A. altissima.* They have classified the leaves with 100% success from 78.94% of the other specimens.

Conclusion

This paper introduces a novel utilization of basic unsupervised classification approaches for discrimination of *A. altissima* leaves. All combinations of extracted features are examined with presented methodology. Results are promising. *Ailanthus altissima* is classified with the success of 0.983% F-Score, 0,967% accuracy and 0,968% sensitivity. Accordingly, it would be possible to conclude that basic unsupervised clustering methods would be efficient classifiers for shape and morphology based features of *A. altissima* tree leaves.

References

Arribas, J.I., Sánchez-Ferrero, G.V., Ruiz-Ruiz, G. and Gómez-Gil, J., 2011. Leaf classification in sunflower crops by computer vision and neural networks. Computers and Electronics in Agriculture, 78(1), pp.9-18.

Bezdek, J.C., Ehrlich, R. and Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), pp.191-203.

Cannon, R.L., Dave, J.V. and Bezdek, J.C., 1986. Efficient implementation of the fuzzy c-means clustering algorithms.

IEEE transactions on pattern analysis and machine intelligence, (2), pp.248-255.

Fu, H., Chi, Z., Feng, D. and Song, J., 2004, December. Machine learning techniques for ontology-based leaf classification. In Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th (Vol. 1, pp. 681-686). IEEE.

Goutte, C. and Gaussier, E., 2005, March. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European Conference on Information Retrieval (pp. 345-359). Springer, Berlin, Heidelberg.

Heisey, R.M., 1996. Identification of an allelopathic compound from *Ailanthus altissima* (Simaroubaceae) and characterization of its herbicidal activity. American Journal of Botany, pp.192-200.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), pp.651-666.

Kadir, A., Nugroho, L.E., Susanto, A. and Santosa, P.I., 2013. Leaf classification using shape, color, and texture features. arXiv preprint arXiv:1401.4447.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), pp.881-892.

Khan, S.S. and Madden, M.G., 2009, August. A survey of recent trends in one class classification. In Irish Conference on Artificial Intelligence and Cognitive Science (pp. 188-197). Springer, Berlin, Heidelberg.

Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, pp.3-24.

Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C. and Soares, J.V., 2012. Leafsnap: A computer vision system for automatic plant species identification. In Computer vision–ECCV 2012 (pp. 502-516). Springer, Berlin, Heidelberg.

Platt, J.C., Cristianini, N. and Shawe-Taylor, J., 2000. Large margin DAGs for multiclass classification. In Advances in neural information processing systems (pp. 547-553).

Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D. and Stefanovic, D., 2016. Deep neural networks based recognition of plant diseases by leaf image classification. Computational intelligence and neuroscience, 2016.

Zhao, C.C., Shao, J.H., Li, X., Xu, J. and Zhang, P., 2005. Antimicrobial constituents from fruits of *Ailanthus altissima* SWINGLE. Archives of pharmacal research, 28(10), pp.1147-1151.