"leaf" dataset

Pedro F. B. Silva André R. S. Marçal Rubim Almeida da Silva

February 2014

1 Data Description

The present database comprises 40 different plant species. Table 1 details each plant's scientific name and the number of leaf specimens available by species. Species numbered from 1 to 15 and from 22 to 36 exhibit simple leaves and species numbered from 16 to 21 and from 37 to 40 have complex leaves.

Class	Scientific Name	#	Class	Scientific Name	#
1	Quercus suber	12	21	Fraxinus sp.	10
2	Salix atrocinera	10	22	Primula vulgaris	12
3	Populus nigra	10	23	Erodium sp.	11
4	Alnus sp.	8	24	Bougainvillea sp.	13
5	Quercus robur	12	25	Arisarum vulgare	9
6	Crataegus monogyna	8	26	Euonymus japonicus	12
7	Ilex aquifolium	10	27	Ilex perado ssp. azorica	11
8	Nerium oleander	11	28	Magnolia soulangeana	12
9	Betula pubescens	14	29	Buxus sempervirens	12
10	Tilia tomentosa	13	30	Urtica dioica	12
11	Acer palmatum	16	31	Podocarpus sp.	11
12	Celtis sp.	12	32	Acca sellowiana	11
13	Corylus avellana	13	33	Hydrangea sp.	11
14	Castanea sativa	12	34	Pseudosasa japonica	11
15	Populus alba	10	35	Magnolia grandiflora	11
16	Acer negundo	10	36	Geranium sp.	10
17	Taxus bacatta	5	37	Aesculus californica	10
18	Papaver sp.	12	38	Chelidonium majus	10
19	Polypolium vulgare	13	39	Schinus terebinthifolius	10
20	Pinus sp.	12	40	Fragaria vesca	11

Table 1: Leaf database: plant species (class) and number of specimens available (#).

Each leaf specimen was photographed over a coloured background using an Apple iPAD 2 device. The 24-bit RGB images recorded have a resolution of 720×920 pixels. Binary versions are provided for simple leaves. Figure 1 provides an overview of the general aspect of the typical leaves of each plant.

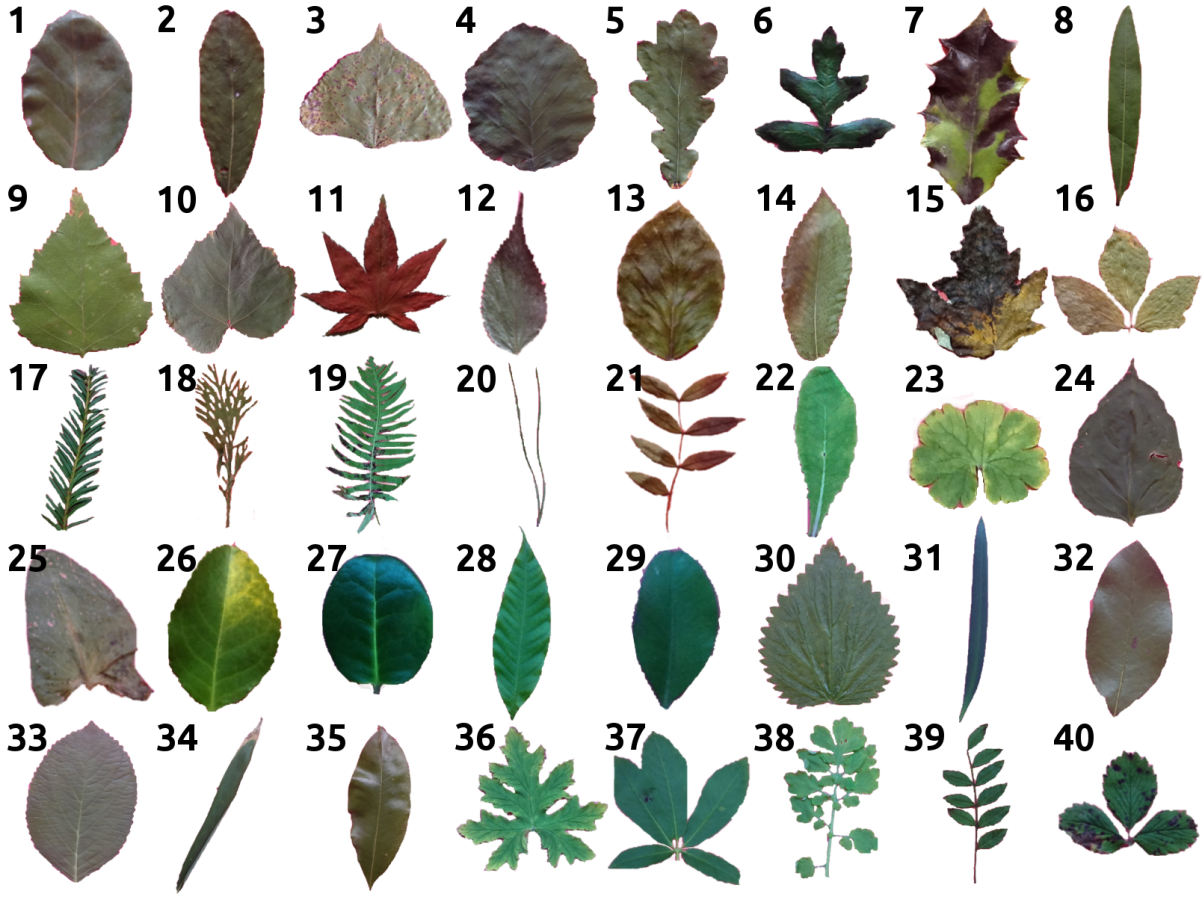

Figure 1: Leaf database overview

2 Attributes

The provided data comprises the following shape (attributes 3 to 9) and texture (attributes 10 to 16) features:

1. Class (Species)
2. Specimen Number
3. Eccentricity
4. Aspect Ratio
5. Elongation
6. Solidity
7. Stochastic Convexity
8. Isoperimetric Factor
9. Maximal Indentation Depth
10. Lobedness
11. Average Intensity
12. Average Contrast
13. Smoothness
14. Third moment
15. Uniformity
16. Entropy

Let I denote an object of interest in a binary image, ∂I its border, $D(I)$ its diameter, i.e., the maximum distance between any two points in ∂I and $A(I)$ its area. Let $A(H(I))$ denote the area of the object's convex hull (i.e. any 'optimal' inscribing convex polygon) and $L(\partial I)$ the object's contour length. Assume that the operator $d(.)$ stands for the Euclidean distance. Table 2 details the definition of attributes 3 to 10.

Shape feature	**Description**
Eccentricity	Eccentricity of the ellipse with identical second moments to I. This value ranges from 0 to 1.
Aspect Ratio	Consider any $X, Y \in \partial I$. Choose X and Y such that $d(X,Y) = D(I)$. Find $Z, W \in \partial I$ maximizing $D^{\perp} = d(Z,W)$ on the set of all pairs of ∂I that define a segment orthogonal to $[XY]$. The aspect ratio is defined as the quotient $D(I)/D^{\perp}$. Values close to 0 indicate an elongated shape.
Elongation	Compute the maximum escape distance $d_{\max} = \max_{X \in I} d(X, \partial I)$. Elongation is obtained as $1 - 2d_{\max}/D(I)$ and ranges from 0 to 1. The minimum is achieved for a circular region. Note that the ratio $2d_{\max}/D(I)$ is the quotient between the diameter of the largest inscribed circle and the diameter of the smallest circumscribed circle.
Solidity	The ratio $A(I)/A(H(I))$ is computed, which can be understood as a certain measure of convexity. It measures how well I fits a convex shape.
Stochastic Convexity	This variable extends the usual notion of convexity in topological sense, using sampling to perform the calculation. The aim is to estimate the probability of a random segment $[XY]$, $X, Y \in I$, to be fully contained in I.
Isoperimetric Factor	The ratio $4\pi A(I)/L(\partial I)^2$ is calculated. The maximum value of 1 is reached for a circular region. Curvy intertwined contours yield low values.
Maximal Indentation Depth	Let $C_{H(I)}$ and $L(H(I))$ denote the centroid and arclength of $H(I)$. The distances $d(X, C_{H(I)})$ and $d(Y, C_{H(I)})$ are computed $\forall X \in H(I)$ and $\forall Y \in \partial I$. The indentation function can then be defined as $[d(X, C_{H(I)}) - d(Y, C_{H(I)})]/L(H(I))$, which is sampled at one degree intervals. The maximal indentation depth $\mathfrak{D}$ is the maximum of this function.
Lobedness	The Fourier Transform of the indentation function above is computed after mean removal. The resulting spectrum is normalized by the total energy. Calculate lobedness as $F \times \mathfrak{D}^2$, where F stands for the smallest frequency at which the cumulated energy exceeds 80%. This feature characterizes how lobed a leaf is.

Table 2: Shape features (attributes 3 to 10)

Attributes 11 to 16 are based on statistical properties of the intensity histograms of grayscale transformations of the original RGB images. The definition of each attribute is given in Table 3.

If Z is a random variable indicating image intensity, its nth moment around the mean is $\mu_n = \sum_{i=0}^{L-1} (z_i - m)^n p(z_i)$, where m is the mean of Z, $p(.)$ its histogram and L is the number of intensity levels.

Texture feature	Description
Average Intensity	Average intensity is defined as the mean of the intensity image, m.
Average Contrast	Average contrast is the the standard deviation of the intensity image, $\sigma = \sqrt{\mu_2(z)}$.
Smoothness	Smoothness is defined as $R = 1 - 1/(1 + \sigma^2)$ and measures the relative smoothness of the intensities in a given region. For a region of constant intensity, R takes the value 0 and R approaches 1 as regions exhibit larger disparities in intensity values. σ^2 is generally normalized by $(L-1)^2$ to ensure that $R \in [0, 1]$.
Third moment	μ_3 is a measure of the intensity histogram's skewness. This measure is generally normalized by $(L-1)^2$ like smoothness.
Uniformity	Defined as $U = \sum_{i=0}^{L-1} p^2(z_i)$, uniformity's maximum value is reached when all intensity levels are equal.
Entropy	A measure of intensity randomness.

Table 3: Texture features (attributes 11 to 16)

3 References

1. "Evaluation of Features for Leaf Discrimination", Pedro F. B. Silva, Andre R.S. Marcal, Rubim M. Almeida da Silva (2013), Springer Lecture Notes in Computer Science, Vol. 7950, 197-204.

2. "Development of a System for Automatic Plant Species Recognition", Pedro Filipe Silva, Dissertação de Mestrado (Master's Thesis), Faculdade de Ciências da Universidade do Porto. Available for download or online reading at `http://hdl.handle.net/10216/67734`