University of Cape Town

Multivariate Statistics

Multivariate Techniques for Leaf Species Classification: A Study on Feature-Extracted Image Data

Author:
Andomei Smit

Student Number:
SMTAND051

March 12, 2025

Contents

1 Introduction

Automatic plant identification from images has greatly benefited many industries including Forestry and Environmental Monitoring [2], Medicinal Plant Research [6] and Agriculture and Crop Management [9]. The sheer number of plant species (estimated around 374 000 in 2016 [3]) and the complexity of accurate identification blend to create the perfect problem for advanced computer vision, machine learning and classification applications. Fortunately, much work has been done in this area of research.

Tan et al.[8] explored various combinations of feature extraction and traditional computer vision techniques (such as Convolutional and Artificial Neural Networks) in order to classify species using Leaf Vein Morphometric (the study of leaf vein structures). Kumar et al. [5] took it even further by developing an app that can identify any of the 184 trees found in the Northeastern United States using images taken by users' mobile phones. Using a nearest-neighbor approach with histogram intersection as the distance metric, the user is presented with a few best matches and various other characteristics of each tree and is left to make the final classification.

Mallah et.al [7] attempted to find a classification method that could cope well with a small training set, a comparatively large number of species and the possibility of incomplete feature extraction. Using a sample of 1600 images of plant leaves corresponding to 100 unique species each 16 different images, they extracted shape, margin and texture features.

Using two density estimation methods (one by Fukunaga [4] that uses a standard K-NN density estimation and one by Atiya [1] that uses a weighted K-NN density estimation) to estimate posterior probabilities for each feature, the performance of all combinations of the features were compared by taking the product of the posterior probability vectors as a single input for the classification task. The best species classification accuracy was obtained by using all three features, giving a mean classification accuracy of more than 96%.

This project will explore if the results of Mallah et al. could be improved upon by using various dimension reduction techniques using their extracted features. It will compare the efficacy of the K-Nearest Neighbors (K-NN) algorithm using Euclidean distance metrics when applied to the various lower dimensional representations of the features to the results from Mallah et. al. Since the original results of the density-based K-NN algorithm was very high (more than 96%), it is not expected that the dimension reduction would improve the results, but it is expected that the clustering accuracy based on the lower dimensional spaces would at least be able to match the results closely.

2 Data Description

The data for this project is from Mallah et. al [7]. It is made up of 1600 images in total from 100 unique species (each with 16 different images) taken of leaves of plants from the Royal Botanic Gardens, Kew, in the United Kingdom. Three features were extracted from these images: shape (representing the leaf contour), margin (representing the leaf edge) and texture (representing the internal patters of the leaf). A black-and-white version of the original images is shown in Figure 1.

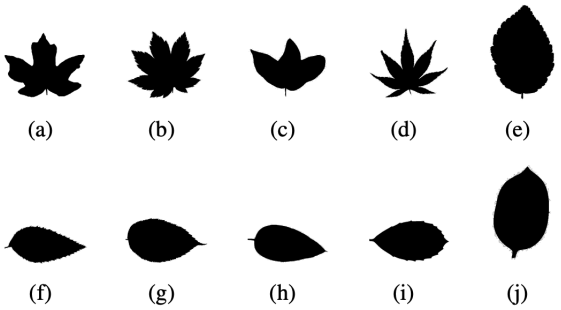

Figure 1: An example of 10 leaf images included in the data after being converted to be binary (black-and-white) images. The binary version of the images was used to extract the shape and margin features, whereas a greyscale version was used to extract the texture feature.

The texture and margin features used histogram accumulation to extract the features, while a normalized description of contour (thus independent of size or orientation) was used for the shape feature.

These features will serve as the independent variables with the species label (a unique label for each species) as the dependent categorical variable. Each feature is made up of 64 $1 \times n$ vectors, where n is the number of images. In total there are 192 feature vectors (64×3) across the three features.

Note that of the 1600 images from 100 species, two of the species had incomplete data in the feature set (i.e. were missing one of the shape, margin or texture features for at least one of the samples of a species). These were originally left out by the authors to test the ability of their probability based estimators when dealing with missing features.

However, since the images do not have a unique sample identifier across the extracted features, it is impossible to know which sample of that species was missing that feature vector. Thus, the entire species that is missing a feature for one of the

samples needs to be removed. It is further assume that the order of samples in the species is consistent across the three different feature sets (i.e that the first sample for the species is the same first sample across all the features).

This project will thus only use the features extracted from the remaining 1568 images of 98 unique species. The final dimensions of the data is dimensions 1568 (images, or n) by 192 (features or p). Each of the feature vectors are continuous vectors on a scale roughly between 0 and 0.2. The exact summary statistics for these features was left out due to little information being shown in those tables given the size of the feature space, but Table 1 shows an example of the first two feature vectors for each feature for two samples of two different species as an example.

Species	Margin Feature 1	Margin Feature 2	Shape Feature 1	Shape Feature 2	Texture Feature 1	Texture Feature 2
Acer Circinatum	0.00000	0.00000	0.00070	0.00072	0.02441	0.00098
Acer Rubrum	0.00195	0.00000	0.00000	0.00099	0.00101	0.00000

Table 1: The first two Feature Vectors for each Feature for two different Species. In total each Feature (Margin, Shape and Texture) had 64 of these vectors, thus each sample has 192 feature measurements each.

3 Analysis Approach

3.1 Independent and dependent variables

Let I_{si} represent the i^{th} image from species s. This project will use the features extracted from images of the 98 complete species, which can be written as

$$\sum_{s=1}^{98}\sum_{i=1}^{16} I_{si} = 1568$$

Thus there were 16 unique images of each species s.

Each I_{si} had 192 features extracted, of which shape, margin and texture had 64 features each. These features will serve as the independent variables in this project. Since this is a classification problem of plant species, the response is a categorical variable with values $s = 1, ..., 98$ representing the index of the unique species.

3.2 Dimension Reduction Techniques

Three different clustering algorithms will be applied and compared, namely Principal Component Analysis (PCA), Isomap and Locally Linear Embedding (LLE).

Each of these techniques has its own set of parameters that need to be specified and/ or tuned. For PCA the number of principal components to use will be determined by finding an elbow in the Scree plot of the cumulative variance explained by the principal components, past which the marginal gain in variance explained is reduced significantly.

For Isomap, there are three different parameters that need to be tuned, namely number of nearest neighbors, the dimension of the lower dimensional space and the metric that will be used to estimate the lower dimensional space (i.e. either Manhattan, Euclidean or Minkowski distances). Using different combinations of all three parameters, a grid search will be performed, and dimensionality reduction will be applied for each combination of parameter sets. Similarly for LLE, a grid search will be applied to find the number of nearest neighbors and number of dimensions of the lower dimensional space.

3.3 Clustering Algorithm

Each of the above dimension reduction configurations will be evaluated by the K-Nearest Neighbors (K-NN) algorithm using Euclidean distance and calculating the effectiveness of the clustering outcome in terms of a Silhouette score (to measure how well the clusters are separated) and Inertia (to measure compactness of the clusters).

Using the best clustering configuration (in terms of the Silhouette score and Inertia) for each dimension reduction technique, the clustering accuracy will be calculated for each as the percentage of correctly clustered samples. This percentage will be compared to the mean accuracy score from Mallah et. al.[7] to see if it can be improved upon.

If time allows, the procedure will be repeated using K-NN with a density estimator on the lower dimensional spaces of each of the dimension reduction techniques to see if the results improve with the same clustering algorithm as in the original paper ([7]).

References

[1] Amir F. Atiya. Estimating the posterior probabilities using the k-nearest neighbor rule. *Neural Computation*, 17(3):731–740, 2005.

[2] M. A. Beck, C. Y. Liu, C. P. Bidinosti, C. J. Henry, C. M. Godee, and M. Ajmani. An embedded system for the automated generation of labeled plant images to enable machine learning applications in agriculture. *PLoS One*, 15(12):e0243923, 2020.

[3] Maarten J.M. Christenhusz and James W. Byng. The number of known plant species in the world and its annual increase. *Phytotaxa*, 261(3):201–217, 2016.

[4] Keinosuke Fukunaga and David M. Hummels. Bayes error estimation using parzen and k -nn procedures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):121–125, 1983.

[5] Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida C. Lopez, and João V. B. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, volume 7573 of *Lecture Notes in Computer Science*, pages 502–516, 2012.

[6] Owais A. Malik, Nazrul Ismail, Burhan R. Hussein, and Umar Yahya. Automated real-time identification of medicinal plants species in natural environment using deep learning models—a case study from borneo region. *Plants*, 11(15):1952, 2022.

[7] Charles Mallah, James Cope, and James Orwell. Plant leaf classification using probabilistic integration of shape, texture and margin features. In *Proceedings of the IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA)*, pages 105–110, 2013.

[8] Jing Wei Tan, Siow-Wee Chang, Sameem Abdul-Kareem, Hwa Jen Yap, and Kien-Thai Yong. Deep learning for plant species classification using leaf vein morphometric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(1):82–90, 2020.

[9] Girma Tariku, Isabella Ghiglieno, Gianni Gilioli, Fulvio Gentilin, Stefano Armiraglio, and Ivan Serina. Automated identification and classification of plant species in heterogeneous plant areas using unmanned aerial vehicle-collected rgb images and transfer learning. *Drones*, 7(10):599, 2023.