Project 2: Dry Beans Classification

Data Science and Machine Learning for Engineering Applications

Politecnico di Torino

Introduction

We want to train an automatic classification model that predicts the bean category given its input features. Seven different types of dry beans were used in this research, considering the characteristics such as shape, form, type, and texture. Since they are very similar to each other, it is necessary to develop a system to distinguish them more effectively and quickly than the human eye to be able to carry out storage.

Objective

Classify the type of bean given its characteristics.

Dataset and Task

Task

The task is a multi-class classification problem. The dataset consists of two files in CSV format: a *Dry_Bean_Dataset_train.csv* file that contains the data to be used for training (and possibly validation) and a file *Dry_Bean_Dataset_test.csv* that should be used for testing. Remember that the test data should not be used in the training phase under any circumstances. Once the model has been trained, the final evaluation should be done on the test data.

Dataset Attributes

The following attributes are present in the dataset:

- Area (A): The area of a bean zone and the number of pixels within its boundaries.
- Perimeter (P): Bean circumference is defined as the length of its border.
- Major axis length (L): The distance between the ends of the longest line that can be drawn from a bean.
- Minor axis length (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.
- Aspect ratio (K): Defines the relationship between L and l.
- Eccentricity (Ec): Eccentricity of the ellipse having the same moments as the region.
- Convex area (C): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
- Equivalent diameter (Ed): The diameter of a circle having the same area as a bean seed area.
- Extent (Ex): The ratio of the pixels in the bounding box to the bean area.

- Solidity (S): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
- Roundness (R): Calculated with the following formula: $(4piA)/(P^2)$
- Compactness (CO): Measures the roundness of an object: Ed/L
- ShapeFactor1 (SF1)
- ShapeFactor2 (SF2)
- ShapeFactor3 (SF3)
- ShapeFactor4 (SF4)
- Class (Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira)

Reference

The original dataset can be found in the following URL: https://www.kaggle.com/datasets/sansuthi/dry-bean-dataset

KOKLU, M. and OZKAN, I.A., (2020), Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques. Computers and Electronics in Agriculture, 174, 105507. DOI: https://doi.org/10.1016/j.compag.2020.105507

Note

The dataset used in the project is a re-sampling of the original dataset (i.e., it is a modified version of the original dataset). The data collection is the result of manual collection, so it is necessary to pay attention to possible errors, missing values, etc.

For the multi-class classification task, we are interested in predicting the **Class** column (i.e., Seker, Barbunya, Bombay, Cali, Dermosan, Horoz, and Sira).

Important: The dataset is rather **imbalanced**. Pay attention because the system must be effective for all classes.