Feature Scale Shapes Nearest Neighbor Classification of Dry Beans

Abstract

Image-derived seed measurements mix quantities reported on different numerical scales. We tested feature standardization in a fixed 7-nearest-neighbor classifier using 13,543 records from seven Dry Bean varieties. Across five paired stratified folds, standardization fitted only on training rows increased mean held-out balanced accuracy from 71.62% to 93.46%, a gain of 21.84 percentage points. Barbunya-to-Cali errors fell from 460 to 86 of 1,322 specimens. Dermason-to-Sira errors rose from 187 to 213 of 3,546 specimens, even as overall Dermason recall improved.

Measurement Scale and Neighborhoods

Dry Bean measurements include area in pixels alongside scale-free shape ratios [1, 2]. In Euclidean nearest-neighbor classification, multiplying a feature changes its contribution to squared distance and can reorder the neighbors. A linear transform provides a general way to change the distance geometry [3]. Earlier work evaluated feature scaling on Dry Bean for several other classifiers [4]. We compared two input representations for the same 7-neighbor rule to measure sensitivity to that choice and see which variety errors changed.

Paired Evaluation

The public UCI Dry Bean dataset supplies 13,611 rows with 16 geometric measurements and seven variety labels [1]. We removed 68 exact duplicates of the complete feature-and-label rows, preserving the order of first occurrences. We used five shuffled stratified folds with seed 42. For each feature, we subtracted the training-fold mean and divided by the training-fold population standard deviation in both the training and test rows. The raw arm used the original measurements. Both arms used exact Euclidean search, seven equally weighted neighbors, and the same folds, with no parameter tuning. For each held-out fold we computed balanced accuracy as the unweighted mean recall over seven classes. We then averaged the five fold scores equally. Recall by variety pooled each row's one held-out prediction.

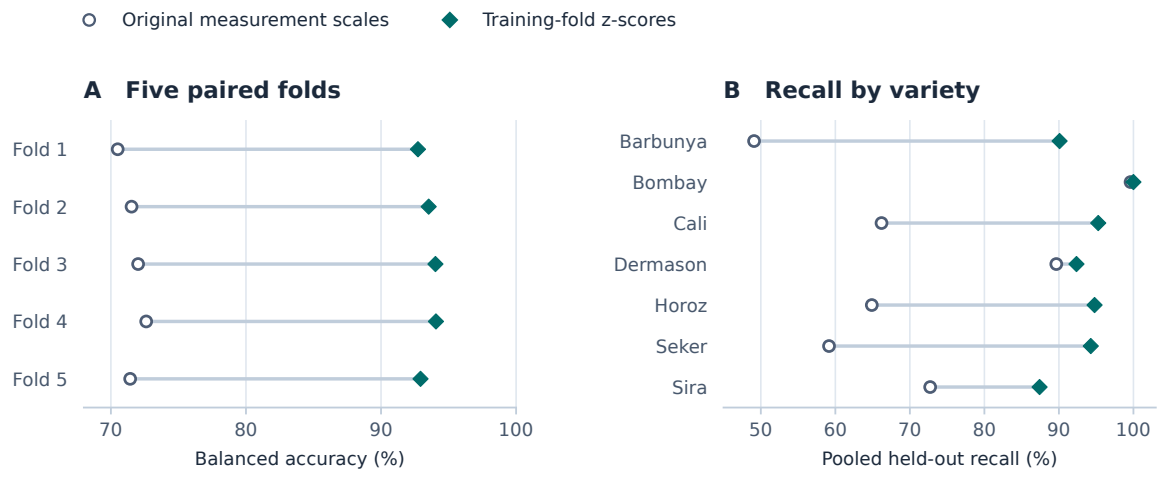


Figure 1. Standardization improved balanced accuracy on each paired fold. A Scores paired by held-out fold. **B** Recall by variety pooled over all held-out predictions. Open circles mark original scales and filled diamonds mark training-fold z-scores.

Results

Balanced accuracy increased from 71.62% to 93.46% on average. Each paired gain fell between 21.45 and 22.23 percentage points (Fig. 1A). The training sets overlapped across folds, so the range describes variation across the chosen splits.

Scaling improved the pooled recall of all seven classes (Fig. 1B). The largest raw error cell was Barbunya classified as Cali. It fell from 460 to 86 of 1,322 Barbunya specimens. Dermason classified as Sira increased from 187 to 213 of 3,546 Dermason specimens, even as overall Dermason recall rose. After scaling, Sira still had the lowest recall among the varieties at 87.41%.

Discussion

Dividing each coordinate by its training standard deviation changes its weight in the Euclidean distance. The gain shows that the original numerical scales materially affected this fixed classifier's measured performance. The experiment changes all scales together, leaving open which features account for the gain. It also uses one seed collection, so the magnitude cannot be generalized to a different acquisition environment without external validation.

Data and Analysis

The source measurements are available from UCI under CC BY 4.0 [1]. The accompanying analysis script and fold-level results reproduce the comparison. Predictions used scikit-learn 1.8.0 with an exhaustive neighbor search and tied class votes resolved by the first label in alphabetical order.

References

[1] *Dry Bean* [dataset]. UCI Machine Learning Repository (2020). doi:10.24432/C50S4B.

[2] Köklü M, Ozkan IA. Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture* 174, 105507 (2020). doi:10.1016/j.compag.2020.105507.

[3] Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10, 207–244 (2009). Journal full text.

[4] Słowiński G. Influence of data dimension reduction, feature scaling and activation function on machine learning performance. *Concurrency, Specification and Programming*. CEUR Workshop Proceedings 2951 (2021). Proceedings full text.