Matthew Ferguson

Review: *Dry Beans Classification Using Machine Learning*

ECE 5984

Summary:

Slowinski created a set of machine learning classifiers to predict dry bean species. Computer vision was applied to 13 thousand dry bean samples to extract geometric features. Features and target data was leveraged to train several classification models, a multinomial Bayes model, a support vector machines model, a decision tree, a random forest model, a voting classifier, and an artificial neural network. Overall accuracy for Slowinksi's models ranges between 88-94%.

Process and Methods

Slowinski accessed a dataset of 13,000 photographs of 7 dry bean species from the University of California, Irvine machine learning repository. Slowinski conducts the analysis using Python. The libraries that he leverages are NumPy, pandas, Matplotlib, seaborn, scikit-learn, and Keras. Models were computed using 'Google Colab' computers. 16 Geometric features were extracted from the photographs using computer vision methods.

The 7 species of bean in the sample set are Barbunya, Bombay, California, Dermason, Horoz, Seker, and Sira. The 16 geometric features are area, perimeter, major axis length, minor axis length, aspect ratio, eccentricity, convex area, equivalent diameter, extent, solidity, roundness, compactness, and four shape factors. Notably color is not used, a feature that likely highly correlates with bean species. Color could certainly be extracted as RGB or otherwise in future agricultural classification work.

After creating the bean features dataset, Slowinski proceeds to correlation analysis and feature reduction. Since all these features are geometric some highly correlate to one another. Compute time and the risk of overfitting can be lowered by reducing features to the vital few. Basically, several features may closely represent the same dimension and the duplication of information can be reduced to yield a better classification model. To this end, Slowinski excludes area, perimeter, eccentricity, convex area, equivalent diameter, compactness, shape factor 1, and shape factor 3. The remaining features that are used for training are major axis length, minor axis length, aspect ratio, extent, solidity, roundness, shape factor 2, and shape factor 4.

Finally, the dry bean extracted features were used to train several machine learning model types. Slowinski creates sets of multinomial Bayes models, support vector machine models, decision trees, random forest models, voting classifiers, and an artificial neural network. Overall accuracy for Slowinksi's models is 88-94%. These models are discussed in further detail in the results and conclusion section.

Results and Conclusions

Slowinski succeeded in accurately classifying bean species using machine learning techniques. Slowinski concludes that bean classification is a simple task in terms of computation and time. Time to train models is under 1 minute. We will give an overview of the model types and their resulting performance in this section.

The worst performing model by classification accuracy was the Naïve Bayes model. Naïve Bayes is a fast and simple model to compute, with lower performance typically. Slowinski uses the Naïve Bayes as a baseline for classification performance. This baseline classification accuracy is 64%.

Support vector machines are flexible algorithms that can be used for either regression or classification models. They do however require significantly more computing power than Naïve Bayes. Support vector machines draw lines between classes in a higher dimensional feature space. Basis functions of linear, 3-degree polynomial, and radial produced accuracies between 91-92%.

A decision tree classification model is a schema of yes/no questions about feature values that once answered prescribe a class to an event. Decision trees use information gain to figure out which questions reduce entropy best. However, care is required when developing decision trees to prevent overfitting. Note that decision trees make splits on one feature at a time. Slowinksi's creates three decision trees. The first has a depth of 4 questions with 9 leaves, the second has a depth of 5 and 16 leaves, and the final has a depth of 10 with 30 leaves. Classification accuracy ranged between 88-92% with the largest tree performing best. My personal interpretation of Slowinski's largest decision tree is that there will be a degree of overfitting due to the large depth.

Random forest decision trees were used to create a bean classification model. Random forest employs many decision trees with subsets of events or features to create a 'forest of decision trees' that are not over correlated to the training data. The decision trees vote for a class, and the most voted for class is output. Generally speaking, random forest models are more accurate, less prone to over fitting and generalize better however they behave like a black box making interpretation of results more challenging. Slowinski's random forest contains 150 decision trees with no restrictions on depth or features. The classification accuracy of this random forest classifier is 94%, the best model performance overall. In addition, training time was only 2 seconds, which was 10-20 quicker than the support vector machine models.

The next model for bean classification is a voting classifier. Voting classifiers are similar to random forest classifiers in that multiple models vote for an aggregate result. The model classifications can be weighted by confidence, or every model output classification can be weighted equally. Slowinski implements a 'dumb' model with no weighting by confidence. The ensemble is composed of 3 previous models, the radial basis support vector machine model, the large decision tree, and the random forest. The voting classifier accuracy is 93%, inferior to the random forest alone, but still quite high.

The final model Slowinski creates is an artificial neural network. The artificial neural network has 3 hidden layers with 17, 12, and 3 neurons. The activation function for the model is rectified linear unit. The output layer has 7 neurons, one for each species. The output layer has a sigmoid activation function, and the loss function of the model is categorical cross entropy. Training occurred for 24 epochs, and 20% of the training dataset was used for validation. The artificial neural network classification accuracy is 93% Note that an additional artificial neural network was trained using a dropout layer to reduce the risk of overfitting. However, the dropout layer increased model accuracy an insignificant amount while being computationally expensive.

I affirm that Slowinski has successfully classified dry beans using machine learning techniques and I agree with Slowinski that bean classification is a relatively simple task in terms of computation. Slowinski is building on previous work for dry bean classification where 16 geometric features were used, and when we draw comparisons between this effort and the previous, we can see that all machine learning models have improved accuracy when using the reduced feature set except the decision tree models. This is likely due to decision trees operating on splits in a single feature and thus decision trees 'cannot combine features' as Slowinski puts it. Information for splits is lost when reducing features. Dermason and Sira were the most frequently mistaken beans likely due to their similarity, whereas Bombay beans were classified perfectly. Bombay beans are notably larger than the other species in the dataset. The author posits that future research should be done on elimination of highly correlated features.

Applications and Discussion:

Slowinski's value case for classification of dry bean species is that species identification is currently a highly manual task that can be automated using computer vision and machine learning classification. Many further applications of this technology beyond automation of bean species identification exist. Classification of dry beans or other agricultural products can be leveraged for many significant economic applications.

For perspective, the global agricultural market is worth over 10 trillion dollars a year with edible beans worth more than 17 billion dollars. About 2 million acres of edible dry beans are planted each year in the United States alone. 2 million acres is about 7% the size of Virginia. The United States is not even in the top five dry bean producers, ranked sixth after Brazil, India, China, Burma, and Mexico. A small automation of dry bean/general agriculture production can have a large economic impact by reducing labor requirements. Similarly, efficiency gains by means of machine learning tools in the agricultural sector can result in more crops using less land.

Bean classification models can hopefully generalize well to plant agriculture provided quality crop data exists or can be created. Using computer vision, I anticipate features can be extracted for any kind of crop. Intermixed agricultural products can be potentially separated autonomously using machine learning classification models that identify agricultural product type or species. Machine learning classification models can be trained using similar data to identify produce quality or detect foreign objects in a food packaging setting. Classification models can be trained using crop data which will potentially predict crop yields or crop quality before a harvest occurs.

Models trained on growing crops may potentially predict the levels of water and fertilizer required to meet a desired target. Pests, weeds, or diseases can be identified and remediated well in advance. Certainly, there are significant economic benefits in applying machine learning to agricultural classification tasks. Furthermore, general image classification can be used for many applications outside the agricultural sector. While an additional degree distanced from Slowinski's work, one can imagine the value of classifying aircraft parts by quality, or machine tools by degradation and nearness to failure.

I find Slowinski's work to be interesting and applicable to wider industry. The overall benefit of this work is that it demonstrates machine learning can be used successfully for classification of agricultural products. I can leverage the results of Slowinski's work to produce bean classification or other agricultural classification models by following his process. I think that the largest value to me personally is the computer vision techniques which extract geometric features. There is a disconnect between photographs of agricultural products, and trainable features which computer vision bridges. This has sparked a personal interest in computer vision and as a result I have enrolled in your class 'ECE 5554 Computer Vision'.

References

- Dry Beans Classification Using Machine Learning
- Edible Beans Market Size 2021 Global Future Growth, Share, Regional Trend, Leading Players Updates, Industry Demand, Current and Future Plans by Forecast to 2028 - MarketWatch
- Production Facts - US Dry Bean Council (usdrybeans.com)