Applying Machine Learning Clustering and Classification to Predict Banana Ripeness States and Shelf Life

Nandan Thor

Master of Science, Industrial Engineering Candidate, California Polytechnic State University, San Luis Obispo

Publication Date: 20 February 2017


Abstract Food waste accounts for over $15 billion annually. Not only is food waste a financial problem, it is also an ethical problem. The use of machine learning algorithms for clustering and classification provide an opportunity to help reduce food waste. K-Means clustering is proposed to determine banana ripeness states and the Decision Tree Classifier algorithm is proposed to classify banana shelf-life. An experiment is undertaken to provide data by imaging bananas and extracting color features using computer vision. The resultant data is then clustered to determine banana ripeness states. The states are used to determine the end-point of data collection. Seven different machine learning classification algorithms are tested to classify fruit shelf-life. The most accurate classifier is the Decision Tree Classifier which has an accuracy around 52%. The combination of machine learning algorithms and big data analysis becomes a powerful tool in working to reduce food waste.

Keywords agriculture; food waste; machine learning classifier; machine learning clustering

1. Introduction

Food waste is a pressing problem globally that will continue to get worse. In 2008, fruit food waste accounted for $15.1 billion of discarded food (Buzby, 2011). Further, 50% of all fruit produced is discarded (Iverson, 2015). Part of the problem is stringent acceptance guidelines for sellable fruit by distribution centers. The other part of the problem is that determining fruit shelf-life is still, at best, an experienced guess. There is waste throughout the agriculture supply chain but the most wasteful steps are at the retail and consumer level.

Banana ripeness is fundamentally subjective. There are loose classifications by the USDA based on ethylene content but the cost of measuring ethylene content is enormous (Kader, 1998). By using machine learning unsupervised clustering techniques, a method for objectively and inexpensively determining ripeness states is introduced.

Many agriculture retail distribution centers employ computer vision systems for quality classification. As the fruit enters the system, it is photographed, compared to existing standards for size, shape, color, etc. and then a binary decision is made to accept or reject the fruit. By using computer vision
feature extraction and machine learning classification algorithms, a method to predict shelf-life using presently-collected data is presented.

The clustering and classification algorithms presented will reduce food waste by allowing riper fruit to have their shelf-life accurately estimated, which will extend the boundaries for acceptable fruit for agriculture distribution centers. By accurately determining fruit shelf-life, fruit that may have previously been rejected as too ripe to sell may still be accepted and sold locally. Using machine learning to classify fruit shelf-life may, then, extend the boundaries for acceptable produce. By determining ripeness specifically, retailers will ensure that that fruit is sold at the exact time it is meant to be sold. Further, standards for ripeness states may be determined and shared. The two techniques will simultaneously reduce food waste and increase profits throughout the agriculture supply chain.

2. Materials and Methods

2.1. Data Collection

A sample size of n=46 bananas are purchased from a retail distribution center. Every day, the bananas are photographed by a Canon EOS 20-D camera using a Macro Ring illumination source (Figure 1). The temperature, humidity, lighting conditions, time of day, background paper, and distance from camera to fruit are kept constant between days.

Figure 1: Canon EOS 20-D camera with attached Macro Ring
After all of the images have been acquired, the raw image Red, Green, Blue (RGB) color coordinates of the banana images are extracted using computer vision algorithms written in LabVIEW and OpenCV. Then, the RGB coordinates are transformed to Hue, Saturation, Intensity (HSI), L,a*,b*, and Percent Green, Percent Yellow, and Percent Brown color coordinates. The percentage color coordinates are determined using the standard color wheel and the HSI color representation (Welch, 1991). The reason to convert from the RGB color representation is because RGB is very sensitive to lighting variations and RGB coordinates may vary between devices (Vidal, 2013). Further, HSI consolidates all color features into one dimension (Hue) and the a* and b* dimensions of L,a*,b* measure color on a green to yellow spectrum. This color coordinate transformation is performed for every banana image every day for fourteen days.

The result of the image capturing, feature extraction using computer vision in LabVIEW and OpenCV, and color representation transformations is a spreadsheet that contains H,S,I, L,a*,b*, percent yellow, percent brown, percent green, and (post-processed) shelf-life data for every banana on every day.

3. Results and Discussion

3.1. State Clustering

Ripeness is a subjective term that is very time-consuming and expensive to measure. Agriculture distribution centers will err on the side of caution when accepting or rejecting a piece of fruit. Considering storage and transportation costs, a distribution center would rather reject an acceptable piece of fruit than accept an unacceptable piece of fruit.

By using unsupervised K-Means clustering on the collected data, clear boundaries for “unripe” (green), “ripe” (yellow), and “spoiled” (brown) are shown in Figure 2.

![Figure 2: K-Means Clustering of banana ripeness states](image)

Using the Extra Trees Classifier in the scikit-learn module in Python, the most important features in each color representation are found. Average hue (of HSI), Percent Green (of relative color percentages), and average a* of (L, A*, b*) are found to be the most indicative of banana shelf-life. These three features are then used to cluster the data, which is projected from the three aforementioned dimensions to two dimensions using Principal Component Analysis (PCA).
By using unsupervised clustering (dropping the post-processed “Shelf-Life” feature), the subjectivity and cost are taken out of determining banana ripeness. Any banana image can be plotted (using the features enumerated above) and its ripeness state can be determined quickly, cheaply, and objectively.

### 3.2. Shelf-Life Classification

Using the banana ripeness state clusters, it is possible to determine the end of shelf-life for the experimental bananas. The first day that the banana enters the “spoiled” state (as defined by the clustering above) indicates the end of shelf-life for that banana. It is, then, possible to determine the shelf-life of the banana at every day and at its corresponding color coordinates. Because data is collected once per day, the resolution of shelf-life is in discrete days and, thus, banana shelf-life is a classification prediction problem as opposed to a regression prediction problem. The goal is to classify the shelf-life category (i.e. four days left), not predict exactly how many days of shelf-life remain (i.e. 4.786 days). Because we are predicting discrete data instead of continuous data, we use classifiers instead of regressors. All of the color representation features (HSI, color percentages, and L,a*,b*) are used as classification features.

Seven different machine learning classification algorithms are tested on the data: Gaussian Naïve Bayes, Ada Boost Classifier, Random Forest Classifier, Decision Tree Classifier, Gaussian Process Classifier, Support Vector Classifier, and K-Neighbors Classifier. All of the algorithms are implemented using scikit-learn in Python. The data is split into a training set and a testing set (80/20) using k-fold cross validation with k=5. The accuracy score for each classifier is computed and is shown in Figure 3. The accuracy scores shown are averages accuracy scores for each fold of the k-fold cross validation in order to utilize all of the collected data.

The Decision Tree Classifier performs the best at about 52% accuracy, closely followed by the K-Neighbors Classifier at around 51% accuracy. The reason for the high performance of the Decision Tree Classifier is because of its unique node splitting criteria (Safavian, 1990). Also, the Decision Tree Classifier over fits the training data which ends up working best because of the similarity between the training and testing data sets (the training set is a representative sample of the population). The results may also be explained in the natural clustering of the data in n-dimensional space (where n is
the number of features) which allows for axial splitting of the data in the decision trees. The worst performing classifier is the Support Vector Classifier at about 20%. This may be explained by the lack of linearity in the data (Lee, 2007) and the fact that each banana begins at a unique color feature set (high covariance among individual bananas).

### 3.3. Conclusion and Future Work

Machine learning is becoming ubiquitous. Clustering and classification algorithms have been applied to banana features extracted using computer vision. The clustering analysis provides a cheap, objective way of determining fruit ripeness. The classification analysis provides a way to determine shelf-life of bananas with accuracy around 50%. Future work may collect different features to improve on accuracy. A sample size of n=46 bananas may be too small to achieve significant results in machine learning, so future work should focus on expanding data collection and on applying different classifiers to the data.

Both the clustering and classification machine learning analyses presented may lead to reduce food waste and may increase profits for agriculture retailers. By accurately determining banana ripeness states, the guesswork is taken out of banana ripeness state prediction and boundaries are established and extended for each state. Using color features to classify banana shelf-life allows for informed distribution and consumption decisions which will decrease food waste at each point in the agricultural supply chain.

Future work may focus on different fruits other than bananas and on different clustering and classification algorithms. The proposed machine learning techniques focus on the distribution level of the agriculture supply chain but it may be valuable to look upstream at the farm level and downstream at the retail and consumer level.

Food waste is a multi-billion-dollar loss as well as an ethical dilemma. By applying cutting-edge machine learning algorithms and big data analysis, steps may be taken to reduce food waste.

### Acknowledgements

The data collection experiments and computer vision performed were part of Nandan Thor’s M.S. Industrial Engineering master’s thesis defended on May, 2016.

Master’s thesis committee members: Dr. Jose Macedo (committee chair), Dr. Reza Pouraghhabagher, and Dr. Daniel Waldorf (all are professors in the Industrial Engineering department at California Polytechnic State University: San Luis Obispo).

### References


