Predicting Monarch Butterfly Sex

Stuart Sims

# Overview

The general goal of this project is to identify what features most reliably predict the sex class of the monarch butterfly, danaus plexippus. The aim is to narrow down the number of features to find those that are most relevant to classifying the sex given a wide spectrum of reflectivity data. There are approximately 1000 features in the dataset and a labeled feature for sex class.

## The Problem

There are two main problems given this data:

First, the spectral properties outside the visible spectrum of light is not known for Monarchs. Do they reflect the same, more, or less light in those areas as they do in the visible areas and do they display any sexual dimorphism?

Secondly, suppose that you wanted to reliably determine the sex class of a monarch butterfly from an image collected in the lab. Students may not be able to do so reliably and there are thousands of samples in cryogenic storage to process. Similarly, what if the most important features are not those that require extremely expensive and delicate imaging hardware and could be discerned from regular photographs? Identifying key features that may hold predictive power could help in feature selection among more commonly available and public data. A predictive model could aid in determining these types of characteristics. Similarly, it may be able to assist with wing fragments or partial specimens found in the field.

## Client Profile

Researchers studying Monarch populations would benefit from these findings, as they may have samples waiting to be processed for this very issue. The Altizer Lab, the Xerces Society for Invertebrate Conservation, North American Butterfly Association, USFS, University of Minnesota Monarch Lab, and several other entomological research labs come to mind as clients invested in similar research.

## Source Data

In remote sensing, reflectance represents the proportion of light that is returned from a target (i.e. reflects off of a surface). Certain wavelengths of light are associated with specific natural structures (like chlorophyll a and b concentrations in a leaf, leaf structure, amount of water, and so on). The reflectance at those wavelengths can quantify the presence, absence, or amount of a substance in an image. Remote sensing is often used after satellite, plane, or UAV collection methods. In this work, I set up an experiment to gauge reflectance of the dorsal view of monarch butterfly wings (outstretched) at a very close distance, comparatively. The intensity of the orangeness in monarch butterflies does vary between sex classes, on infectious status (whether the individual is parasitized by OE), and on life cycle stage (a summer-breeding individual vs a migrating individual). This data was all collected within the Deepak Mishra Lab at the University of Georgia School of Geography. It contains about 700 wavelength reflectance features and labeled sex data. The data is complete, so no imputation is needed for this specific data.

# Approach

## Data Collection

Sample specimens were provided by the Altizer Lab from The University of Georgia Odum School of Ecology. All Monarchs were deceased and some from cryogenic storage were provided as well. Sampling occurred in multiple sessions using a hyperspectral imaging camera known as the “SuperGIR.” The superGIR can collect reflectance of light at wavelengths ranging from 340 to 2500nm. Before each session, a calibration was performed against an industry-standard white test media. This calibration ensures reflectance values’ accuracy across sampling sessions. Due to the small imaging area, halogen lighting was utilized to boost the available light off the sample. Due to the warm-up time, the imaging area was left for 15 minutes for the lights to reach standard operating characteristics.

A small brass c shaped clip was used to ensure wings were fully outstretched for imaging. A baseline sample was also taken with the background and brass clip alone to ensure no additional noise was introduced to the reflectance data.

Each specimen was spread with a rear, dorsal view with the central axis of the specimen parallel and centered the central axis of the camera. Wings were kept outstretched using the small brass C shaped for each sample.

Each specimen was scanned three times. The scan number, specimen number, sex, and other categorical characteristics were recorded. Additionally, the camera captures a traditional image of the imaged area for each specimen. If any scan was disturbed or the individual was not centered properly in the imaging area, it was noted and discarded later. An additional sample was captured to ensure that each specimen had at least 3 samples.

Each set of scans was aggregated per specimen using an unweighted mean at each wavelength reflectance. The result is complete reflectance data across 96 Monarchs.

## Data Cleaning and Transformation

This foundational data contained reflectance values but also categorical data some categorical fields:

* ORIGINAL DATA SCAN #
* RECORDED SCAN #
* Monarch
* Group
* INFO
* INFECTED
* NOTES

These fields are not used in subsequent modeling at this time. The “Group” value “ORIENTATION TEST” was used to determine if the rotation of the specimen heavily influenced reflective characteristics. It contained multiple scans at multiple angles compared to the central axis of the camera and was removed first. Subsequently, these categorical fields were dropped from the dataset.

Some scan aggregations were missing specimen sex. These scan aggregations correspond to background samples of the area. Records with no “Sex” label were therefore dropped.

Finally, outlier detection was used to determine reflectance values outlier positions. Z scored outliers with a threshold of 3 standard deviations was investigated. Valid interpretation does require normality in each feature.

Inter-Quartile range was also used to determine outlier values in the dataset. Outlier thresholds were established per field at 1.5\* IQR on either side of the 25th and 75th percentile values. Ultimately, an ensemble method suited to handle outliers was used, so no steps were added to then remove the values identified as outliers at the cleaning stage.

## Exploratory Analysis and Inferential Statistics

The highest and lowest reflectance were identified among wavelengths. The standard deviation of each feature was explored to better understand where the features differed most. Subsequently, plotting each reflectance value across the spectrum of wavelength values (columns) allowed visual exploration of the reflectance “curves.” The mean reflectance at each feature for both sex classes as well as each individually was plotted to characterize where reflectance differed from the overall mean among male and female specimens. When coloring the line plots based on Sex, reflectance was visibly diverging starting at the orange/red wavelengths with males having higher reflectance values. Statistical testing was used to discern whether the mean reflectance in the blue, green, and red wavelengths differed significantly between sex classes. In all three tests, there was evidence to reject the null hypothesis at alpha = 0.1. Among visual portions of the spectrum, there are significant differences in reflectance. In this portion, dimensionality reduction was intentionally postponed. Because the primary goal is to identify features of interest, potentially reducing the feature space before modeling was counterproductive at this point.

## Machine Learning

In this portion of the analysis, some introductory visualizations were presented to allow readers to interact with plots and gain more context to the study system. Subsequently, principal Component Analysis was used to determine if only a few principal components could be used with the same success as modeling with all features. With the use of only three principal components, 94% of the variance could be explained. This indicated the data was a good candidate for dimensional reduction using this technique.

T-distributed Stochastic Neighbor embeddings was used as an exploration in the potential goal of dimensional reduction, but ultimately not adopted for use in classification. An ideal parameter combination in t-SNE to cleanly separate groups with a two-dimensional representation was not found despite testing a wide variety of perplexity and early exaggeration combinations.

In pursuit of both identifying important features and creating a predictive model, a random forest classifier was constructed to predict the binary variable for sex class using the entire dataset. Hyperparameter tuning methods were used to determine the best combination of parameters for use in the model. A Randomized search with 3 folds and cross-validation was initially implemented but proved highly variable between runs. Subsequently, a grid search with 3 folds and cross validation was used to identify the best parameter combination. Both methods ultimately arrived at same parameters with the training set. However, the best parameters identified resulted in overfitting to the training data and did not generalize as well to the testing data. Given this overfitting, default parameters for the random forest classifier from sci-kit learn were used and achieved better performance.

Gradient boosting was briefly explored only to see how it would perform since relatively few records were still being misclassified. However, it did not do any better than the implementation of the random forest. Hyperparameter tuning was not pursued with the gradient boosting approach mainly in the interest of minimizing model complexity given so few data points.

In discerning feature importances, importances from the random forest model were extracted and sorted. In addition to this, permutation importance was used to cross-compare. In both, multiple features in the orange and red wavelengths were identified as most important. These included 656.8nm, 658.2nm, and 693.3nm. Both identified one wavelength in the near infrared to be important as well: 822.2nm. In both methods, importance values were relatively small and no one feature reported an importance greater than 0.10 out of 1. Collinear features were likely given the visualizations of these curves.

Given the frequency of correlation observed, a hierarchical clustering approach was used with the features to identify clusters of features and reduce the severity of those effects on the feature importance interpretation. Spearman R correlation coefficients were created and visualized. Ward Linkage was then calculated on this matrix of spearman R coefficients and subsequently clustered agglomeratively. The dendrogram from that clustering was visualized in order to select an appropriate threshold. A threshold of 20 was identified, leaving only 12 features representing the original 994.

These twelve features were used to pair down the columns in use for the feature training and testing datasets. A new random forest classifier was built and trained on this subset of features from the training data. The performance of this model was improved over the model using all features. Both overall accuracy and F1 scores for both labels were improved by using this reduction. The model feature importances were extracted again. They indicated that the most important feature, now represented by 667.1 accounted for just over 1/3 of the feature importance. The wavelength features in its cluster were printed for inspection and spanned from the red colors of light up into low values of the near infrared.

The second most important feature was 749.1nm. This area of the electromagnetic spectrum showed an important response outside the visible spectrum. The features clustered in this area were also printed for visual inspection. All of the features in this cluster were outside the visible spectrum and in the near infrared response range.

A third feature in the short wave infrared portion of light, 1902nm was also identified and the features clustered alongside it were printed for visual inspection as well.

The features in the UV spectrum had a feature importance of about one fourth or less than the strength of the NIR features.

LIME (locally interpretable model-agnostic explanation) was used to aid in interpretability for the random forest classifier, as . A prediction function was specified using the classifier trained on the reduced, 12 features. It was then fed a specified observation and the class probabilities along with a visualization of the decision probabilities at those 12 features was provided.

Finally, the simplified classifier model was pickled and stored. This feature reduction approach would allow someone to submit only 12 wavelength reflectance values to the model given the same experimental conditions and obtain a prediction. With PCA, it would not be nearly as simple, so the clustering approach was ultimately adopted.

## Deployment

Flask was used to build a simple application that created a port to which a user could submit a string containing reflectance values at these twelve wavelengths and have the stored model predict the label. The webpage used some basic html to create a form box and button and separate html was constructed for a page containing the predicted label. While this is not publicly available to the web, it provided valuable experience in making the results of modeling available in a more accessible way that could be integrated in a more robust stack of tools.

## Deliverables

Deliverables for this project include:

* A publicly available pickled classification model
* An overview of the analysis in laymen’s terms for use by other curious researchers in this area

## Conclusion

This data proved to be a good candidate for dimensional reduction techniques. Similarly, a classification model is moderately strong in predicting sex with the reduced feature sets. Finally, the most important features identified in those models do align with the current biological understanding of the Monarch Butterfly itself.

Some unexpected findings arose as well:

First, Monarch butterflies have lower reflectance in the visible portion of the spectrum as compared to areas such as the red edge, near infrared, and in portions of the short-wave infrared portions of the electromagnetic spectrum.

Secondly, the feature space can be reduced heavily while also increasing the performance of the model. The original 994 features, when clustered by the strength of their correlation, could be reduced to only 12. The same random forest classification model reached higher performance with the use of these 12 features

Regarding coloration, while red coloration is the most important feature by a factor of 2, features in the Near Infrared and short-wave infrared held medium importance as well. These NIR and SWIR areas should be investigated for potential biological importance in Monarch Butterflies. Subsequent analysis may include:

1. Any physiological ability to sense in these areas. For example, do Monarchs have any morphological structures that would allow visual sensing in this area? If so, does the sexually dimorphic nature of this response play a role in mate selection or success?
2. A potential predictor of health as opposed to observable redness, as that has shown to correlate with infectious status and migratory behavior. Can the effects of visible coloration and these other response areas be separated given enough samples? Do these reflectance responses outside the visible spectrum play any role in sexual selection or success or indicate infectious or other physiological health status?