Hello and welcome to the presentation on Rapid Microbe Identification

Today we will be discussing the use of machine learning techniques to classify microorganisms, based on measurements generated automatically from their microscopic images

--

Millions of people worldwide are pet owners because we love pets and consider them part of the family. We enjoy their company and, like our families, we don’t ever want them to get sick. But sometimes they do.

In America, people spend hundreds of dollars or more on their pets each year and, while the majority of that is spent on pet food, the second highest spending category is veterinary care.

--

One of the ways our pets can get sick is through infections. When this happens, time is an important factor because microorganisms can replicate quickly. Many microbes have short doubling times, some measured in minutes.

Identification of these microbes can be time consuming. Culturing bacteria for identification methods takes time and so do genetic testing methods

Technicians can look at the microorganisms under a microscope and may be able to get a broad reading of the microbe type, but this can be subjective and can vary from person to person.

It is important to get an accurate identification because the ID can be used to determine precise treatment options. Without an ID, broad spectrum antibiotics are often used, which can further the already difficult problem of increasing antibiotic resistance.

--

To create a model for identifying microbes, I used a Microbes Dataset, available on Kaggle.

This dataset consists of 23 numeric features. These measurements were automatically generated from microscopic images of the microorganisms and include things like length, width, circumference and area as well as a host of other, more technical measurements. About half of these features had skewed distributions, but all were transformed using a Box-Cox transformation.

The target variable consists of 10, fairly broad microorganism types. The 10 classes were imbalanced, with the most common one making up about 1/4 of the dataset, while the least common was represented by only about 2% of the data. To help compensate for this imbalance, I first set the class\_weight parameter to ‘balanced’ for algorithms that offered this option. This led to a modest improvement. I later went back and rebalanced the training dataset by sampling the overrepresented classes and resampling the underrepresented classes, ultimately maintaining the same size dataset. This led to lowered performance across all metrics for ever algorithm. I suspect this is because approximately 25% of the information was lost from the original dataset.

One concern of mine is whether or not the samples in the dataset are representative of real-world samples. If they were grown in nutrient rich media under optimal conditions, their morphologies may vary slightly from what would be seen in the more competitive environment of the gut. Additionally, some of the classes listed represent a wild variety of shapes and sizes of organisms, like diatoms, so it may be beneficial to narrow the set down to particular species instead of the broad groupings.

Many of the measurements appeared to be redundant, with strong correlations and high Pearson Ranking. In an effort to reduce dimensionality, I used the ANOVA F-test to determine the optimal number of features and it came back with 23, so I kept all of them.

--

Of the 9 algorithms tested with the dataset, 2 performed extremely well: the Random Forest Classifier and the Decision Tree Classifier. Others performed quite poorly, but still better than random chance.

In addition to Accuracy, The Metrics considered for model evaluation include F1-score, Precision, and Recall. All three of these additional metrics scored comparably to Accuracy across all models.

Once I narrowed down the prospective models, I performed some hyperparameter tuning using GridSearchCV and was able to bump accuracy up a bit more, to 98.7%

--

Given the comparable performances between Random Forest and Decision Tree, I decided to proceed with the Decision Tree Classifier because it will be easier to interpret for troubleshooting or for further analysis and model improvements. As a bonus, it runs about ten times faster than the Random Forest Classifier.

Recommendations moving forward include generating more data as well as broadening the dataset to include more microbe taxons and increasing precision by including additional genera or species within the existing classes.

Additionally, it could prove useful to look for other measurements that could be made from the microbe images, for use in the model, or alternatively to train a neural network on the images themselves and let it extract the useful features directly.

Thank you for coming and have a great day!