Jacob Redden ([jjr0017@auburn.edu](mailto:jjr0017@auburn.edu), Coordinator), Yifeng Wu ([yzw0134@auburn.edu](mailto:yzw0134@auburn.edu)), Robert Smith (res0059@auburn.edu)

COMP 6600 Artificial Intelligence

Group 17

03/28/2024

Project Proposal Report

Group 17 has been tasked with implementing 5 different Artificial Intelligence classifiers on a dataset obtained from huggingface.co. The restrictions on this dataset provided are 100k training samples, 10k testing samples, 2 classes being classified, and at least 20 attributes. The group decided to do an image classification problem so the CelebA-attrs dataset was chosen since it contained enough data to meet the criteria. This dataset has 40 different attributes identified for each picture in the dataset. We will use the 128x128 resolution pictures as an input to the models and we will focus on classifying these pictures based on the ‘Smiling’ attribute and different hair attributes to try to determine the color of hair of the individual in the picture. The hair attributes are ‘Blonde Hair,’ ‘Gray Hair,’ ‘Brown Hair,’ and ‘Black Hair.’

Image Classification is a well studied topic in the Artificial Intelligence industry. Many different Artificial Intelligence techniques have been used for image classification including Random Forests, Support Vector Machines, K-Nearest Neighbor, and Convolutional Neural Networks. For this assignment, we will be implementing Decision Tree, Naive Bayes Classifier, Logistic Regression, Support Vector Machine, and Multi-Layered Perceptron. Because the input to this problem is a picture represented by a 128x128 grid of RGB pixels, there are no human distinguishable patterns in the data that can allow for the classifying of pictures by hair color or determining if the subject is smiling. Machine Learning can find patterns in large amounts of data that humans cannot perceive which makes this problem suitable for Machine Learning to solve.

We expect to have the challenge of tuning all the hyper-parameters for each classification method to obtain an optimal solution. Image classification can take time and the dataset is fairly large so the tuning of hyper-parameters to yield optimal results may take a substantial amount of effort and time.

The dataset we will be using is found on huggingface.co at <https://huggingface.co/datasets/tpremoli/CelebA-attrs>. This dataset has 163k training samples, 19.9k testing samples, and 20k validation samples. Huggingface has a python module we will employ called ‘datasets’ that will download and store the dataset in a cache directory for future use.

We plan to use python as our coding language and the sklearn library for the implementation of the different Artificial Intelligence models. sklearn is a large python library that can be used for data preprocessing, classification, and metrics, plus many other applications we will not use. This will allow us to focus on tuning the hyper-parameters rather than focusing on a correct, efficient implementation of each model.

We anticipate a lot of time being spent tuning hyper-parameters. For the Decision Tree Classifier, we will determine the best results by tuning the decision criterion function, the max depth of the tree, the minimum number of samples required to split, and the minimum number of samples required to create a leaf node. For the Naive Bayes Classifier, we will tune the variance smoothing and potentially look at using different probability distributions. For the Logistic Regression Classifier, we will experiment with the tolerance for the stopping criteria, the norm for the penalty assessed in the regression, and the solver used for the logistic regression. For the Support Vector Machine Classifier, we will try different kernels to find the optimal one for this problem and tune the regularization parameter. For the Multi-Layered Perceptron Classifier, we will tune the hidden layer sizes and the activation function used to introduce non-linearity.

We will use various metrics to determine how well our models do at classifying the images. Some metrics we will use are precision, accuracy, sensitivity, and specificity. The 5 different models will be compared with each other using the metrics obtained by classifying the Validation samples of the dataset. A conclusion will be drawn determining which model works best for this image classification problem.

The due date of the project is April 30th with the Progress Report due April 16th. We expect to have the initial supporting code done to implement argument parsing, download the dataset, do any preprocessing necessary on the data, and put the data in a format the classifiers can understand by April 1st. The first couple models will be implemented and tuned by April 12th which will allow for a few days to write the progress report. The final models will be implemented and tuned by April 22nd to allow for a week to write the Final Report.