RAW DATA CLASSIFICTION USING PCA

BHUVANA KORRAPATI
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, USA
bkorrapati@gmail.com

***Abstract*—The main aim of this assignment is to create a feature space with the pixel data. The images are collected and divided into the pixels into a particular range and, image processing is used to extract feature space.**

I. Introduction

This document is a report of all the activities that are done as part of assignment-1.The raw data is collected as images and the feature space is created with lables and merge to get the final dataset. It includes all the steps from setting up the environment to final solution.

II. Setting up Environment

A. Choosing Environment

Mostly python or R language is used for the image processing and machine learning. Here for this assignment Python is the programming language used to implement the concept.It is a simple and consistent programming language. Python provides access to large number of libraries and frameworks. Anaconda is a free open-source environment which consists of over 250 automatically installed packages. It is available in all the versions for windows and mac OS as well. In addition to this over 7,500 open-source packages can be installed with "conda install" command. After the installation of the anaconda, we can access Jupiter notebook through terminal as well to execute the code. Spyder is another IDE that is used for the same python code to execute and can be easily accessed with the anaconda navigator. For latest python libraries, the pip should be upgraded with the command pip3 install –upgrade python. OpenCV which stands for Open Source Computer Vision Library is an open-source computer vision and machine learning software library available in languages C++, Python, MATLAB, and Java. It is the most popular library used by Data scientists for image processing.

B. Installation

Anaconda environment was downloaded from the official Anaconda website[3] and for proper installation followed the steps mentioned in the website[4]. OpenCV is used as computer vision tool in this assignment which was downloaded and installed by running "conda install -c conda-forge opencv" common in Anaconda Prompt as shown in Fig.1(left).Jupyter Notebook can be accessed by the command jupyter notebook on terminal as shown. Spyder IDE is integrated with the Anaconda which can be accessed from Anaconda by running "spyder" command.

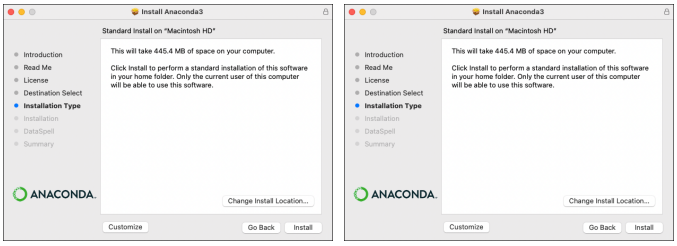

Fig. 1: Left: anaconda for mac. Right:jupyter notebook.

III. Collecting Raw Data(Images)

In this assignment, three different fruits are used to extract features namely LIME, ORANGE, APPLE. All the images that are used in this assignment are downloaded from the folder provided. The exact images used are in Fig.2 left, middle and right respectively.

Fig. 2: Left: lime Raw Image[6]. Middle: orange Raw Image[7]. Right: apple Raw Image[8]

IV. Reading and displaying Images

The input images are given using the 'imread' command and stored with unique variable names. These variables are used to access a specific image throughout the code. Each image will have 3 colour channels namely Red, Green and Blue whose numeric values are 0, 1 and 2 respectively. Colour channels of an image is displayed used the code "plt.imshow(image[:,:,0])" where the third value of the array represents the colour code. Fig.3 left, middle and right are the colour channels of the image that is used in this assignment.
The images used in this assignment are converted to Gray scale before extracting Features. An image can be converted to Gray Scale by using an OpenCV library. Below is the code that is used in the assignment to convert an image to Gray Scale and find its shape.

Grayscaling is the process of converting an image from other color spaces e.g. RGB, CMYK, HSV, etc. to shades

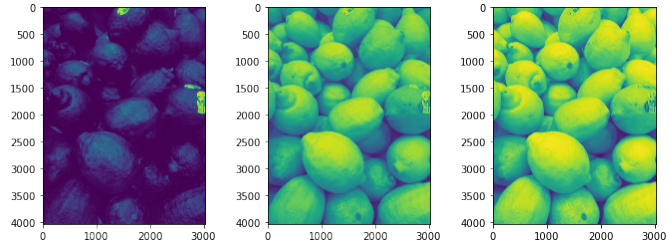

Fig. 3: Left: Red Colour Channel of lime. Middle: Green Colour of lime. Right: Blue Colour Channel of lime.

of gray. It varies between complete black and complete white.The three input images are considered with different colour channels and changed to the grayscale images. The main importance of the grayscaling is 'dimension reduction', 'Reduces model complexity' and 'For other algorithms to work,. The images are provided for all the three fruits.

Fig. 4: Left: Grayscale for lime. Middle: Grayscale for orange. Right: grayscale for apple.

V. Image Resizing and Binarization

The raw images we collect can be of any size, which creates abnormalities while extracting features. A function imresize(int height, int width) is created resize all the images without changing their aspect ratio by keeping height as 264 pixels. To avoid the data loose in the next steps the width of the images is taken in such away that their value is divisible by 12. A OpenCV library is used to resize an image.The input for the above code will be the height and width of a Gray scale image. The output will be a new image with resized Gray scale. Binarization is of the image is converting a grayscale image into a black and white image (i.e. 0 and 255 pixels respectively). This can be achieved with the help of a process called thresholding. So, with the help of thresholding, we can get binary images. Now, thresholding can be defined as a process in which each pixel is converted to either 0 or 255 depending on whether its value is greater than or less than a threshold value. If the value of the pixel is greater than the threshold value, then it is converted to 255 otherwise it is converted to 1.

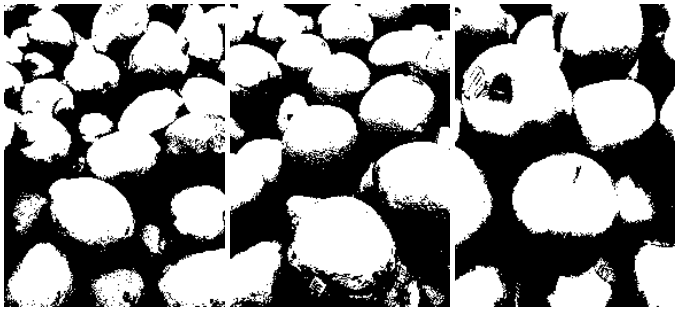

Fig. 5: Left: Grayscale for lime. Middle: Grayscale for orange. Right: grayscale for apple.

VI. 12X12 block-feature vectors

A. Generating 12X12 Block-feature vectors

To find the best feature pair of a fruit an 12X12 pixel block is considered and each pixel in that block is considered as a feature. A pixel value is uniquely considered in this method. The blocks move 12 steps at a time to avoid repetition. N number of feature pairs are created for a Gray scale image, where N is number of 12X12 pixel blocks. Values of these feature pairs are used to identify an image and perform various machine learning techniques. Below is the code used to covert the image to 12X12 pixel blocks and convert them to vectors of 144 size.The input for the above code will be a Gray scale image and it's size. The code will compute the generate an output of csv file. This csv file will contain value of each pixel of an image with 145 columns, where 145th column is used to label the fruit type, where lime is labelled for 0, 1 for orange, 2 for apple.

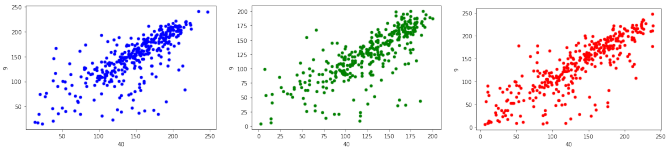

Fig. 6: Left:Scatter plot for lime. Middle: Scatter plot for orange. Right: Scatter plot for apple.

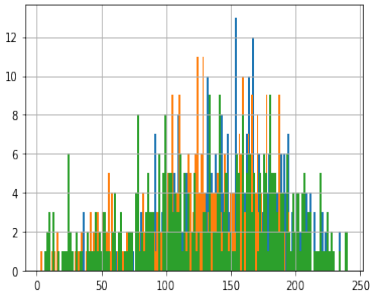

Fig. 7: histogram for feature vectors

VII. Sliding feature vector for 12X12

A. Generating 12X12 Block-feature vectors

The overlapping of the vectors is considered and sliding causes the increase in number of observations as well. A pixel value is uniquely considered in this method. The blocks move 1 step further at a time to avoid repetition. N number of feature pairs are created for a Gray scale image, where N is number of 12X12 pixel blocks. The number of observations increases by sliding features. The input for the above code will be a Gray scale image and it's size. The code will compute the generate an output of csv file. This csv file will contain value of each pixel of an image with 145 columns, where 145th column is used to label the fruit type.

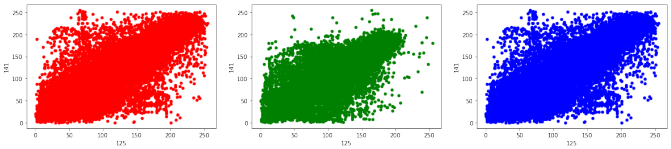

Fig. 8: Left:Scatter plot for lime. Middle: Scatter plot for orange. Right: Scatter plot for apple.

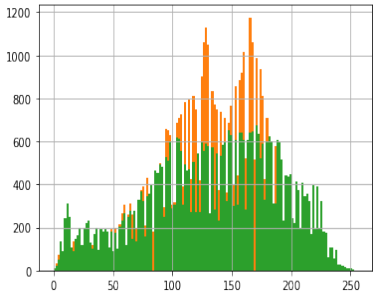

Fig. 9: Histogram for sliding vectors.

VIII. Statistical Description

Number of observations, the dimentions of the dataset that is generated through both the feature vectors and sliding feature vectors are considered. The number of observations for sliding feature vectors is more than general feature vectors. The dimentions that is number of columns remains same.the mean and standard deviation is calculated for the feature vectors as we can easily get to know the data and the labels as well. this can be even helpful to check the correlation between the features in the dataset accordingly. The mean and standard deviations are mentioned in the table.

Number of Observations in lime	352
Number of Observations in orange	352
Number of Observations in apple	352
Dimension of Data in lime	144
Dimension of Data in orange	144
Dimension of Data in apple	144

TABLE I: Statistical Information of Sliding window feature

Number of Observations in lime	45360
Number of Observations in orange	45360
Number of Observations in apple	45360
Dimension of Data in lime	144
Dimension of Data in orange	144
Dimension of Data in apple	144

TABLE II: Statistical Information of Sliding window feature

	Mean	Standard deviation
Mean and SD for lime	146.73	47.48
Mean and SD for orange	124.80	42.47
Mean and SD for apple	133.80	57.56

TABLE III: Mean and SD for feature vectors

IX. Merge the feature vectors

The generated block feature vectors for all the three images are merged to get the datasets with lables. The datasets are merged and converted to the single dataset. We merge to train and test the data for algorithms. the scatter plots and histograms are generated through the merged datasets. for the generated block features if two fruits are merged we get three merged datasets and all the three fruits are merged to get one dataset as well with three different lables.

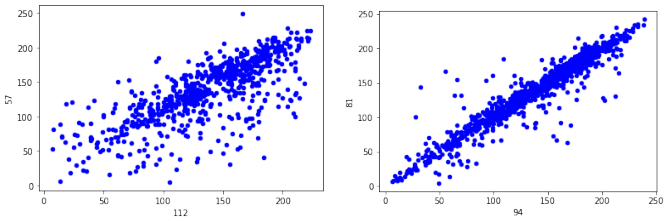

Fig. 10: Left:Two class lables. Right: Three class lables

X. Effects of block size and code changes

The block size of the image is considered for the detail feature extraction as we get more features to explore from the same image. It is highly important to consider the size of the pixels as we can have sparsity, high dimentionality or low dimentionality problems as well.As the dimensionality increases, the distance between objects may be heavily dominated by noise. That is, the distance and similarity between two points in a high-dimensional space may not reflect the real relationship between the points. Globe is the library that is used to read multiple files from the folder for which the path is mentioned and then the feature space is generated as well with the same library in opencv.

XI. Divide the Datasets

In this assignmen,t 3 different fruits are used to extract features namely LIME, ORANGE, APPLE. The merged datasets that is two dimentional feature space we get six datasets that is three for general and other for sliding feature vectors. The dataset is divided by 80:20 rule that is 80 percent of training data and 20 percent of testing data. It is a proven rule to split the dataset for better accuracy as well and to reduce the

	Mean	Standard deviation
Mean and SD for lime	133.66	57.53
Mean and SD for orange	125.57	42.35
Mean and SD for apple	133.98	57.53

TABLE IV: Statistical Information of Sliding window feature

problem of non seen data. The training and testing datasets are saved as a csv files in the output folder.

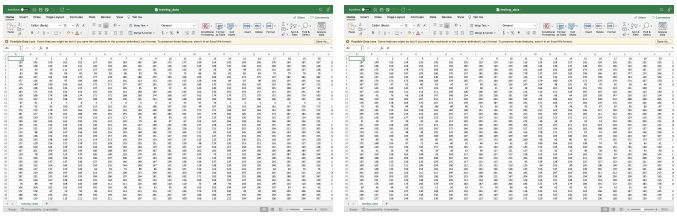

Fig. 11: Left: csv file of training data. Right: csv file of testing data.

XII. Feature selection and Linear Regression

From the training and testing dataset, two columns other than the label column is selected randomly. The two columns, statistical data is considered and check the correlation is checked. If both the features are correlated we can eliminate them to reduce the dimentionaly problem. It is the simple machine learning model that mostly uses the statistical information for the computation process. Regression is a supervised learning algorithm as the labels are considered. The histogram and scatter plot for non overlapping datasets.

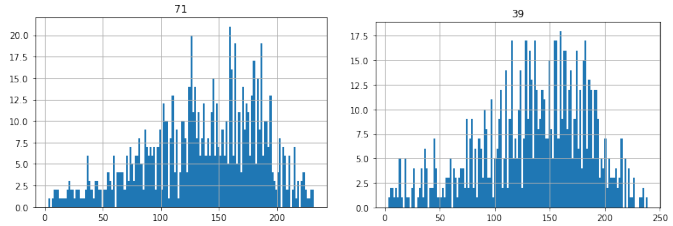

Fig. 12: Left: hist of training data. Right: hist of testing data.

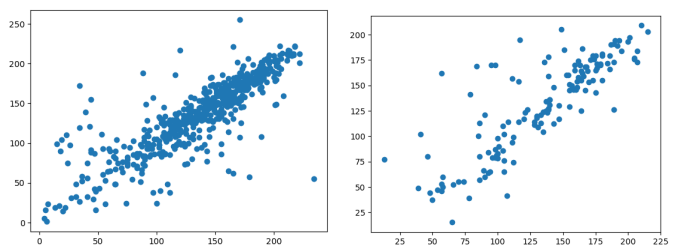

Fig. 13: Left: scatter plot of training data. Right: scatter plot of testing data.

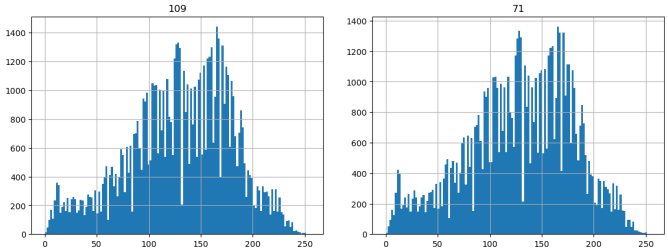

Fig. 14: Left: hist of training data. Right: hist of testing data.

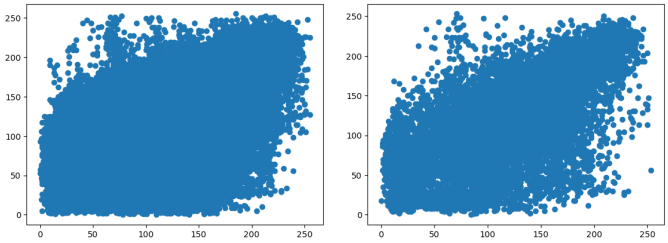

Fig. 15: Left: scatter plot of training data. Right: scatter plot of testing data.

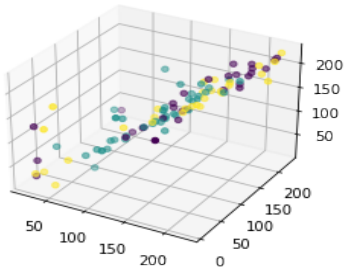

Fig. 16: Linear regression model.

XIII. Elastic-Net Regression

Elastic-net Regression is used as regression model to train the datasets. Elastic-net regression is trained on Training dataset and based on that trained model labels are predicted on the testing dataset. Those new labels are inserted into 65th column, these dataset is sent to csv. These 64th and 65th columns are compared to check the accuracy of the model. The accurary is comparatively low as the training is not done accurately, Confusion matrix concept is used to check the accuracy of the model. Though Elastic-net regression provides some accurate models for two-class classifier it fails to create an accurate model for three-class classifier. The confusion matrix for all the four data sets are

XIV. Evaluation of the learning models

Accuracy of all the models that are generated above are tested using Confusion Matrix. Confusion matrix is generated by the training and testing values of the algorithm. Now in-built measures are used to find the acccuracy of all the models. METRICS library is used from the sklearn is used to find the accuracy. ”metrics.accuracy-score” and even the precision,

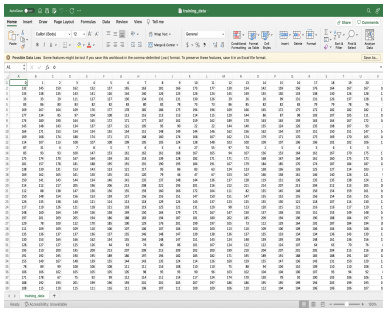

Fig. 17: screenshot of csv file with 65 col.

sensitivity and specificity is calculated by the confusion matrix. The below is the accuracy of every single model tested using built-in measures.

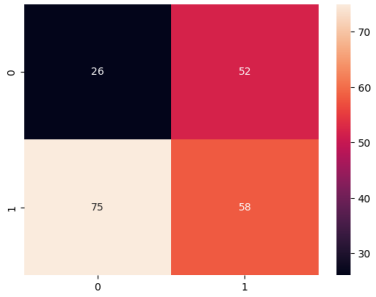

Fig. 18: Random-Forest graph.

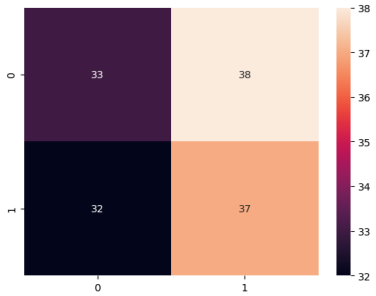

Fig. 19: Random-Forest graph.

Table.1.

Our Accuracy Score	0.44680851063829785
Our Precision Score	0.4782608695652174
Our Sensitivity Score	0.44000000000000006
Our Specificity Score	0.45454545454545453

TABLE V: Statistical Information of Sliding window feature

XV. Random-Forest Implementation

Random-Forest is the layered model of supervised learning algorithm that is mainly used for the feature generation. To overcome the draw back of Elastic-net regression for 3-class classifier Random Forest is used to train models for all the datasets. The features are considered and The training dataset in each category is trained using random forest and then using the trained model labels are predicted on the testing dataset. The new labels are inserted into the dataframe as the 145th column. The accuracy of this model is tested by using confusion matrix.

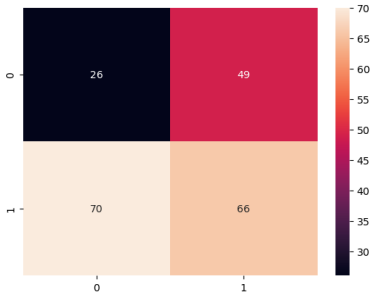

Fig. 20: Random-Forest graph.

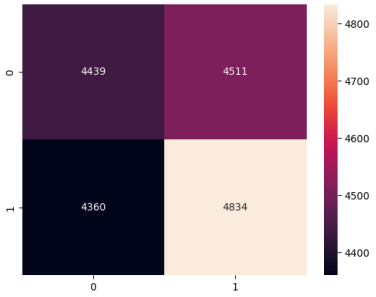

Fig. 21: overlapping graph.

XVI. Principle Component Analysis

It is a scalable machine learning model, used for the feature extraction, the number of features are reduced and only the first two are considered as the best features. In order to obtain an effective outcome, this comprehensive data analysis was carried out utilizing the Python programming language and the PCA algorithm of machine learning. we have created and applied PCA for all the four datasets that is two overlapping with two class classification and three class classification, and two with overlapping or sliding features.

XVII. References

[1] https://steelkiwi.com/blog/python-for-ai-and-machine-learning/
[2] https://docs.anaconda.com/anaconda/
[3] https://www.anaconda.com/products/individual
[4] https://docs.anaconda.com/anaconda/install/windows/
[5] https://www.analyticsvidhya.com/blog
[6] http://https://scikit-learn.org/stable
[7] https://dontrepeatyourself.org/post/how-to-resize-images-with-opencv-and-python/
[8] "Machine Learning Models and Algorithms for Big Data Classification Thinking with Examples for Eff effective Learning" – Shan Suthaharan