Report on Assignment 1

Vaibhav Mishra

02/18/2018

1 Abstract

In the era of information explosion, enormous amounts of data have become available on hand to decision makers. Big data refers to data sets that grow so huge that they become difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these data sets. Such value can only be provided by using big data analytic, which is the application of advanced analytics techniques on big data. This paper aims to analyze the different data set.

2 Data sets

2.1 Letter recognition

Following two data sets are selected for classification problem

- Letter recognition `https://archive.ics.uci.edu/ml/machine-learning-databases/letter-recognition`
  - Data Exploration
    In this part we will try to explore the data set and reveal some interesting facts about the Letter recognition.
    1. Total no of class : 26
    2. Total number of observation : 20000

    The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20.000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15
    because of mentioned reason makes it hard to be analyzed using current technologies and techniques.

510/LetterRecognisition/DataDistribution.png

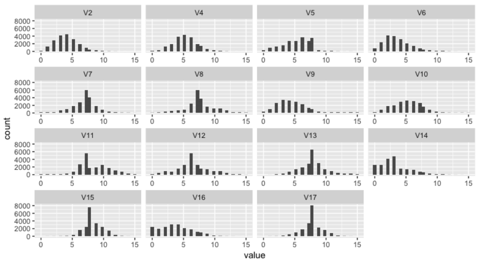
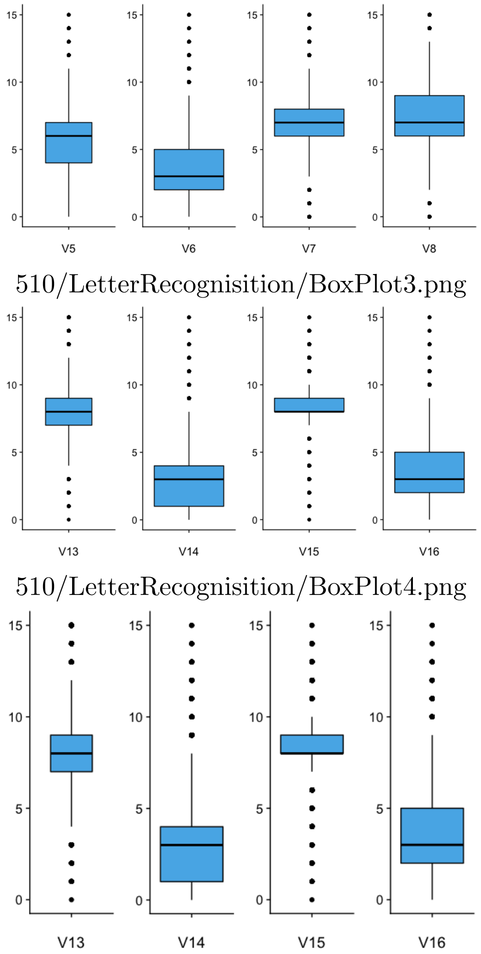


Figure 1: Histogram

2.2 EDA on Letter recognition

Histograms are graphical representation of distribution of the numerical data. The purpose of histogram is to graphically summarize the distribution of a data. We have used histograms for studying the spread and skewness of the data(Figure 1).

510/LetterRecognisition/BoxPlot2.png



510/LetterRecognisition/BoxPlot3.png

510/LetterRecognisition/BoxPlot4.png

Figure 2: BoxPlots

2.3 Boxplots

Boxplots are excellent tools for understanding the range of the data and detecting the outliers. The boxplots also explain about the quartiles, mean and median of the data. We used boxplots for a for a mentioned purpose. Boxplots for all the variables are given below,

From the plots (figure 1) it can be observed that the count of outliers is less. It is also observed that for attributes V5, V8, V9, V13 and V15 outliers are present on the higher end of the plot. For other variables except V2 outliers are present on both sides of the plot and for V2 no outliers are observed. From the above plots, we can say that the outliers are less and thus, we decided to use the data without removal of outliers.

510/LetterRecognisition/CorrelationMatrixPlot.png

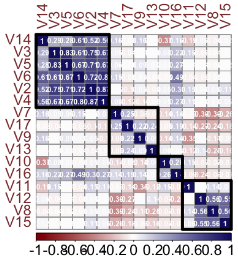


Figure 3: correlation

2.4 correlation

The main purpose of the correlation is to get a brief idea about the relationship of one variable with other variables. The correlation plot for all the variables is given in figure 2. The plot will help us to understand the dependency between the variables. In the above plot red color indicates negative correlation and blue color indicates positive correlation.

3 Robot Wall Navigation

Dataset can be found at:
`https://archive.ics.uci.edu/ml/datasets/Wall-Following+Robot+Navigation+Data` The data were collected as the SCITOS G5 robot navigates through the room following the wall in a clockwise direction, for 4 rounds, using 24 ultrasound sensors arranged circularly around its 'waist'.

The provided file contain the raw values of the measurements of all 24 ultrasound sensors and the corresponding class label. Sensor readings are sampled at a rate of 9 samples per second. It is worth mentioning that the 24 ultrasound readings and the simplified distances were collected at the same time step, so each file has the same number of rows (one for each sampling time step). The wall-following task and data gathering were designed to test the hypothesis that this apparently simple navigation task is indeed a non-linearly separable classification task. Thus, linear classifiers, such as the Perceptron network, are not able to learn the task and command the robot around the room without collisions.

3.1 Boxplots

Boxplots are excellent tools for understanding the range of the data and detecting the outliers. The boxplots also explain about the quartiles, mean and median of the data. We used boxplots for aforementioned purpose. Boxplots for all the variables are given below,

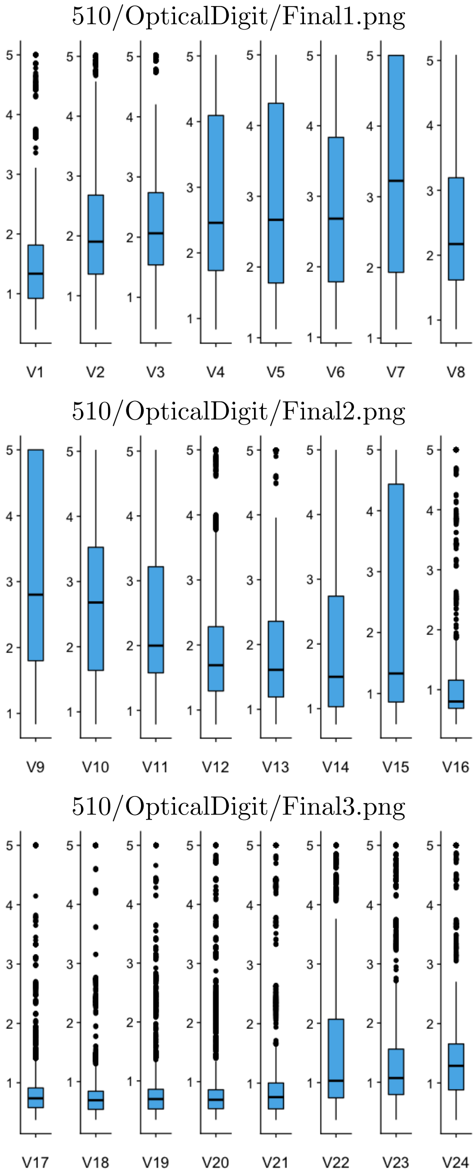
From the above plots (figure 4) it can be observed that the count of outliers is less. It is also observed that for attributes V1,V2,V3 V12, V16, V17, V18 , V19,V20,V21,V22,V23 and V24 outliers are present on the higher end of the plot. For other variables except V2 outliers are present on both sides of the plot and for V7,V9 and V15 no outliers are observed.

3.2 correlation

The main purpose of the correlation is to get a brief idea about the relationship of one variable with other variables. The correlation plot for all the variables is given in figure 2. The plot will help us to understand the dependency between the variables. In the above plot red color indicates negative correlation and blue color indicates positive correlation.(Figure 5)

3.3 Histogram

Histograms are graphical representation of distribution of the numerical data. The purpose of histogram is to graphically summarize the distribution of a data. We have used histograms for studying the spread and skewness of the data(Figure 6).


510/OpticalDigit/Final1.png

510/OpticalDigit/Final2.png

510/OpticalDigit/Final3.png

Figure 4: BoxPlots

3.4 Density

univariate data density provide important insight into the nature and usability of said data(Figure 7). From figure 7 we can visualize distribution of data

510/OpticalDigit/FinalCor.png

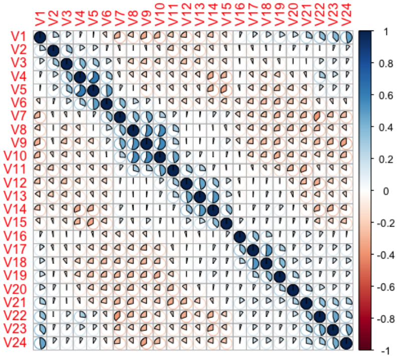


Figure 5: correlation

510/OpticalDigit/FinalDensity.png

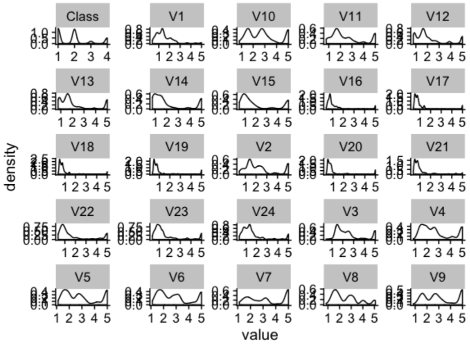


Figure 7: Data Density

510/OpticalDigit/FinalHist.png

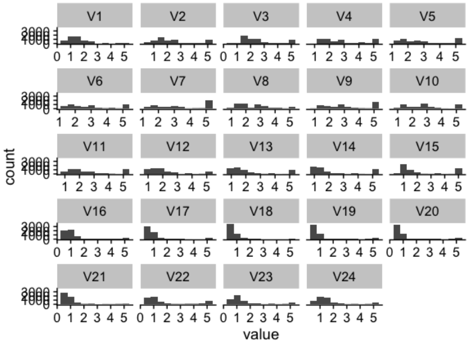


Figure 6: Histogram

510/OpticalDigit/MissingData.png

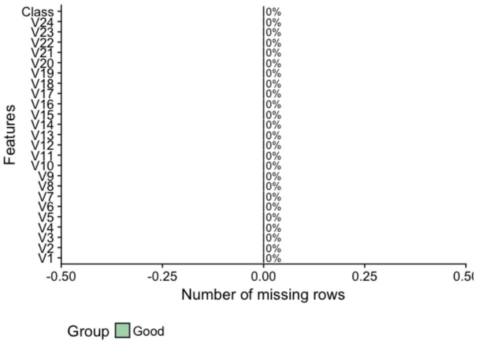


Figure 8: Missing data Plot

based on class for each feature.

3.5 Missing data analysis

Figure 8 provides the missing data analysis on each each of the feature

4 Data set creation

The Creating an Analytical Dataset provides foundational knowledge to input, clean, blend, and format data in preparation for analysis. As part of learning you should able to:

- The common sources and types of data
- To identify and correct common issues with data
- To format data in useful ways for analysis
- To blend data from multiple sources together

4.1 Image Classification

Image feature is a simple image pattern, based on which we can describe what we see on the image. For example human eye will be a feature on a image of a human. The main role of features in computer vision(and not only) is to transform visual information into the vector space. This give us possibility to perform mathematical operations on them, for example finding similar vector(which lead

510/HorseVsCar/ResizedImage/Scatterplot.png

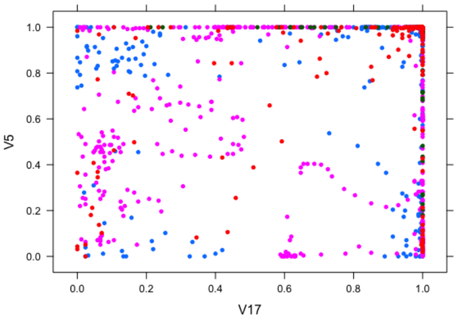


Figure 9: ScatterPlot

us to similar image or object on the image).

To make image classification as big data problem we have selected images(cat, dog, lion, panther) as set1 and other set includes images of horse, elephant , truck and car which makes it to 4 class classification problem. Steps performed to extract features from the images and transform to data file is as following:

- Per class one image is downloaded to local
- Each image is re sized to the 256*256
- From re sized image 8*8 block of images are processed at a time which results total 64 features and 4096 observation for each set
- Extracted image information is transformed to CSV file for further analysis

5 Data exploration results on dataset

5.1 ScatterPlot

Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another.(Figure 9) Provides result between between 2 features V5 and V17 .

510/HorseVsCar/ResizedImage/CorrelationMatrix.png

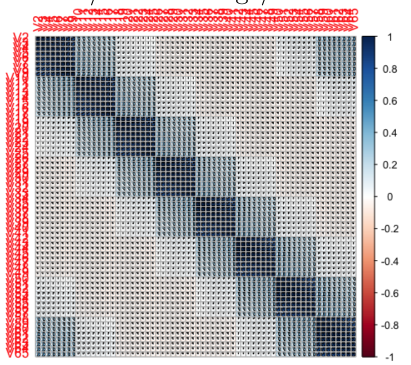


Figure 10: set1:correlation

510/Image classification/Set1Cor.png

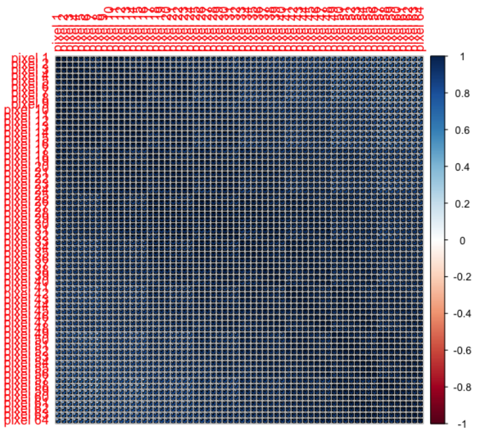


Figure 11: set2:correlation

Code and plots can be found in there respective folder

References

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/letter-recognition/letter-recognition.data

[2] Kotipalli, K., Suthaharan, S., 2014. Modeling of class imbalance using an empirical approach with

spambase dataset and random forest classification, in: Proceedings of the 3rd annual conference on Research in information technology, ACM. pp. 7580

[3] Suthaharan, S., 2014. Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM SIGMETRICS Performance Evaluation Review 41, 7073

[4] Suthaharan, S., 2015. Machine Learning Models and Algorithms for Big Data Classification: Think- ing with Examples for Effective Learning. volume 36. Springer