Development of an Intelligent Decision Support System to Classify Unsafe Driving

Polytechnic University of Catalonia

MAI-IDSS

Practical Work 3

June 9, 2016

**Abstract**

Contents

[1.0 Introduction 4](#_Toc451872579)

[1.1 Kaggle 4](#_Toc451872580)

[1.2 State Farm Distracted Driver Detection Competition 4](#_Toc451872581)

[1.3 Dataset 4](#_Toc451872582)

[1.4 Problem Architecture 5](#_Toc451872583)

[1.5 Tools Used 5](#_Toc451872584)

[2.0 Data Pre-Processing 5](#_Toc451872585)

[3.0 Exploratory Data Analysis 5](#_Toc451872586)

[3.1 Distribution of Class Training Image Counts 5](#_Toc451872587)

[3.2 Distribution of Class Training Subject Counts 6](#_Toc451872588)

[4.0 Data Driven Models 7](#_Toc451872589)

[5.0 Model Validation 7](#_Toc451872590)

[6.0 Final Results and Conclusions 7](#_Toc451872591)

[7.0 Future Work 7](#_Toc451872592)

**Table of Figures**

[Figure 1: Example Training Image 4](#_Toc451872593)

[Figure 2 - Distributions of Class Training Subject Counts 6](#_Toc451872594)

**Table of Tables**

[Table 1- Distribution of Class Training Image Counts 5](#_Toc451872595)

# 1.0 Introduction

This report outlines the development of an Intelligent Decision Support System (IDSS) to identify from images if the driver of a vehicle is performing an action that may be considered unsafe. An overview of the problem is presented, with details about the data structure and architecture, along with the exploratory, modelling and validation analysis required in building the IDSS. The final results are discussed and future work with potential improvements are recommended to further develop the model.

## 1.1 Kaggle

Kaggle is an online platform that connects companies with difficult Data Science problems with Data Science practitioners, in the form of competitions. Companies provide details, data and a scoring methodology that competitors use to build the best IDSS models to solve the specific data driven problem.

This project attempts to build an IDSS that is competitive with other models in the [State Farm Distracted Driver Detection](https://www.kaggle.com/c/state-farm-distracted-driver-detection) Kaggle competition.

## 1.2 [State Farm Distracted Driver Detection](https://www.kaggle.com/c/state-farm-distracted-driver-detection) Competition

This competition aims to accurately classify the behavior of vehicle drivers to help determine if their action might be considered unsafe. With the advent of in-car radios, fast-food and cellphones, distracted driving has become a common cause of car accidents and fatalities. This goal of this competition is to build a system that can highlight these distracted driving activities.

## 1.3 Dataset

The State Farm Distracted Driver Detection Kaggle competition consists of 22,424 training images that have been sorted into 10 classes, and 79,726 unclassified images for testing. An example image is presented in Figure 1.



Figure : Example Training Image

The 10 classes to predict are:

c0: normal driving

c1: texting - right

c2: talking on the phone - right

c3: texting - left

c4: talking on the phone - left

c5: operating the radio

c6: drinking

c7: reaching behind

c8: hair and makeup

c9: talking to passenger

## 1.4 Problem Architecture

## 1.5 Tools Used

# 2.0 Data Pre-Processing

# 3.0 Exploratory Data Analysis

The high dimensionality of image processing and Computer Vision problems make the application of traditional Exploratory Data Analysis (EDA) and Feature Engineer (FE) difficult, but there are still important attributes of the dataset to explore.

## 3.1 Distribution of Class Training Image Counts

Table 1 outlines the distribution of counts of available training images, to ensure that an even distribution of each class exists within the training set.

Table - Distribution of Class Training Image Counts

|  |  |  |
| --- | --- | --- |
| Class ID | Class | Training Image Count |
| C0 | normal driving | 2489 |
| C1 | texting - right | 2267 |
| C2 | talking on the phone - right | 2317 |
| C3 | texting - left | 2346 |
| C4 | talking on the phone - left | 2326 |
| C5 | operating the radio | 2312 |
| C6 | drinking | 2325 |
| C7 | reaching behind | 2002 |
| C8 | hair and makeup | 1911 |
| C9 | talking to passenger | 2129 |

The classes seem to be fairly consistently represented, with class 0 (Safe Driving) being the most represented and class 8 (Hair and makeup) being the least represented. This makes safe driving 30% more represented than Hair and Makeup, but most classes have close to 2000 sample images. The effects of these different training sample sizes should be considered when developing the IDSS model.

## 3.2 Distribution of Class Training Subject Counts

Each image in the training set has been labelled with the "subject" or individual person observed in the image. Figure 2 Visualizes below a stacked barchart for the class representation of each individual.

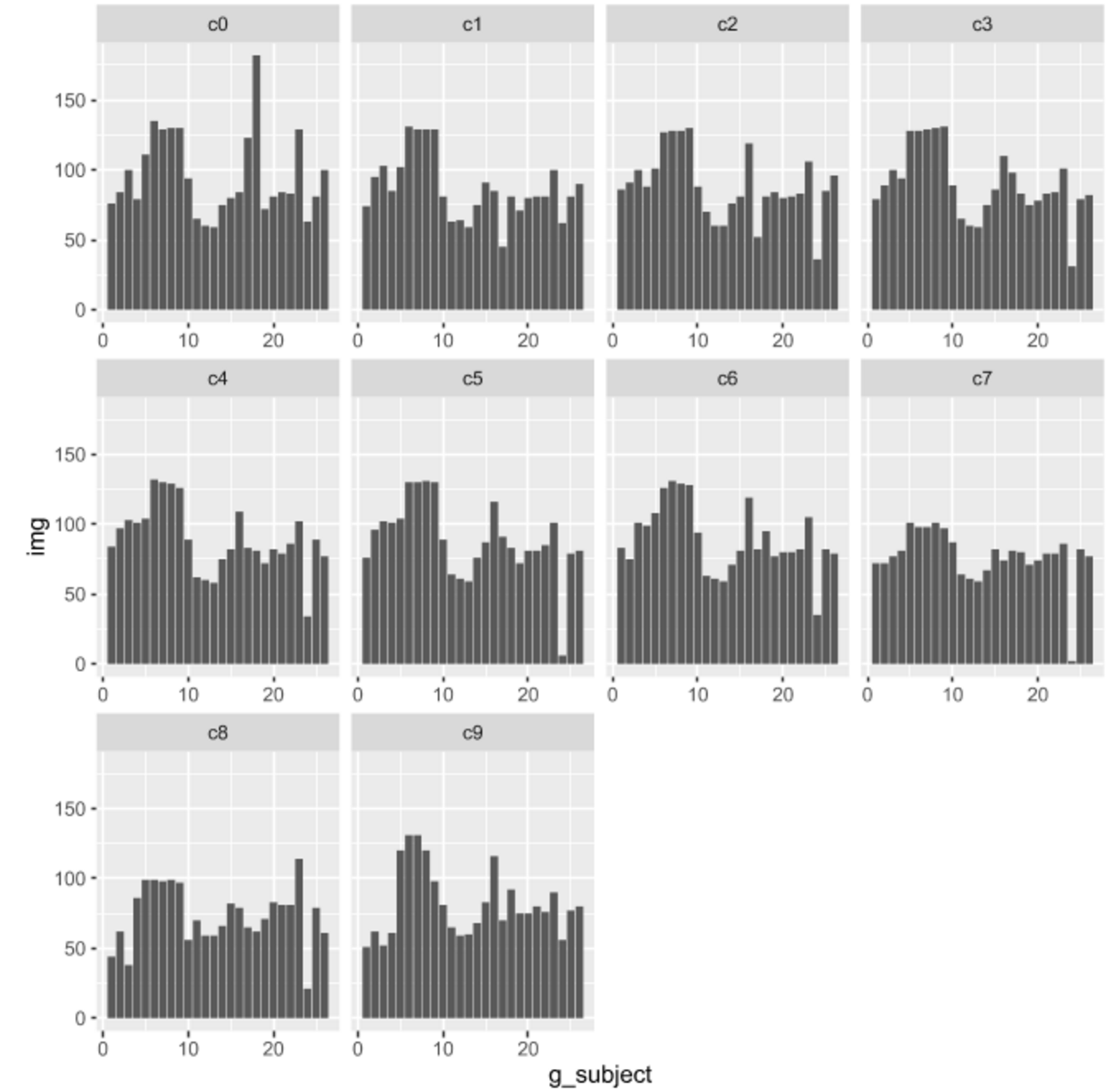


Figure - Distributions of Class Training Subject Counts

The distribution of individuals in each class are fairly similar, but there do exist obvious differences. Some easily identifiable outliers include subject 16 in class 0 (safe driving) being over represented, or subject 26 being largely underrepresented in many categories, such as C5 (operating the radio) and C7 (reaching behind). It is important to review these outliers in analyzing the performance of a model, as it may be possible to develop models that overfit to the above distributions, but do not generalize to real world distribution of these classes. For example, it could be possible to learn attributes about subject 26 and use those attributes as indication that the probability of being in classes 5 or 7 are very low. But because the purpose of our model is to generalize to the larger public, using these attributes specific to subject 26 may incorrectly categorize subjects with similar attributes, but do belong in classes 5 or 7.

# 4.0 Data Driven Models

# 5.0 Model Validation

# 6.0 Final Results and Conclusions

# 7.0 Future Work

Figure 3 shows the generic architecture for the IDSS prototype. Each data driven model is color coded to highlight which methods were utilized by that specific model in each step of the IDSS development.

**Legend**

\* All Models

\* L/A DAC

\* Decision Tree

\* Nearest Neighbor

\* Neural Network