DS5230 Project Pitch : Pattern Recognition in Accidents in the UK

Sahasrabhojanee, Adwait
sahasrabhojanee.a@husky.neu.edu

Sreekumar, Sreejith
sreekumar.s@husky.neu.edu

Yu, Xue
yu.xue1@husky.neu.edu

Li, Xuexian
li.xuex@husky.neu.edu

October 11, 2017

1 Abstract

In September 2017, Kaggle released a dataset that contains an aggregated collection of 1.6 million accident records compiled from a few UK government sources in the area during the time span of 2000 - 2016. For the course project of Unsupervised Machine Learning and Data Mining - DS 5230, we will be analyzing this data, working towards uncovering patterns in accidents.

Our final objective would be to derive insights and provide useful information to the managerial personnel for the betterment of traffic management plans.

2 The Dataset

The dataset contains two types of files:

- The average annual daily flow of traffic on major roads in the U.K
- Records on over 1.6 million road accidents, spread across three files

Attributes related to the accident are provided as columns in the records. These include the date and location of the accident, the types of the road and the vehicle involoved, the number of casualties, and the weather and geographic conditions. A complete description of the fields has been provided at : https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales

3 Questions we intend to answer

During the course of the analysis, we will be focusing on answering the following questions:

1. Does the season/time affect the number of accidents?
2. Does the type of the road significantly affect the chance of an accident occurring on it?
3. What determines whether or not a police officer will attend the scene of the accident?
4. Are there factors other than those recorded in the data that cause a higher number of accidents in a particular area? (Population, for instance)
5. Are any of the factors linearly related to each other?

4 Methods

We intend to use clustering methods to answer questions 1 through 3, holding out a potentially important variable (which takes one of K values) and group the data into K clusters. We will then compare the clusters formed with the held-out variable to see if they are roughly similar.
For analyzing the factors mentioned in question 4, we intend to cluster the data, and then try to explore external

factors which match the clusters formed.
Exploratory data analysis and statistical studies will be conducted on the variables to understand correlations and answer question five.

5 Results

1. Visual temporal and spatial exploratory data analysis of accident data
2. Answers to the aforementioned questions
3. Outputs from the statistical analysis of variables

6 Work Division

In a subgroup of two, we will be working towards answering questions one through four (two questions for each subgroup). Question five, being more generic, will be a split across everyone. This will proceed in a series of group discussions and experiments as the project progresses.

7 References

- Kaggle, 2017. 1.6M accidents & traffic flow over 16 years. Accessed at October 10, 2017. `https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales`