DBDM Assignment 3

Mining association rules on traffic accidents in the UK

Eva van Weenen (s1376969), Paul Couzy (s1174347),
Alex van Vorstenbosch (s1526146), Irene Haasnoot (s1258745)

December 2018

Abstract

This project tests the performance of the Apriori algorithm in finding association rules in a data set of 1.6 million traffic accidents in the United Kingdom. Three experiments are done to try and find association rules that might be useful to anticipate and prevent accidents. We report on several of the association rules found for each experiment and conclude that the performance of the algorithm on the data set is sub optimal. The algorithm appears to be biased towards association rules that offer little insight and therefore a lot of pre processing is required.

1 Introduction

Traffic incidents remain the most important death causes of young adults of 15-29 years of age worldwide and it is the ninth most important cause of death of all ages world wide (World Health Organization, 2016). Therefore, road safety is a major concern to governments, but also to citizens. Deriving relations between the variables involved in traffic incidents is therefore of importance if we want to prevent these incidents.

This project analyses traffic data from the United Kingdom between 2005 and 2014, excluding 2008. The main goal is to use algorithms such as the Apriori algorithm to mine interesting association rules from the data set. The Apriori algorithm was first proposed by Agrawal and Srikant (1994) as a way to mine association rules in transactional databases. It is commonly used as a method to perform market basket analyses.

In the following sections we will discuss the hardware and software used to mine the data set, as well as the methods used to analyze the performance of the algorithm.

2 Methods

In this section the system-parameters, the code, the data-set, and the methods and processes used to analyze the data are discussed.

2.1 System Parameters

The following systems parameters are used to conduct the experiments:

Setup – STRW University

- CPU: 3.20GHz Intel(R) Core(TM) i5-3470, 32K L1d cache, 32K L1i cache, 256K L2 cache, 6144K L3 cache
- Main memory: 8GB
- Disk: 500GB HDD
- OS: Fedora 21.0

The data used in this setup was stored on a network disk on a different STRW server that was mounted on the computer used. This disk has a storage capacity of 1.5 TB.

2.2 Code

This project was carried out using `Python 3.6` (Python Core Team, 2018). The pre-processing was based on a project carried out by Landenberger (2018) and the accident data is treated as if being a large item set. In order to do this the data was converted and analyzed using `mlxtend`, as described by Raschka (2018).

2.3 Data set

The traffic accidents data set was taken from Kaggle.com (2018) and represents an extended version of the data-set provided by U.K. Department of Transport (2017). The data-set consist of several `.csv` files containing the following data:

- UK road accidents from 2005 to 2007
- UK road accidents from 2009 to 2011.
- UK road accidents from 2012 to 2014

Our data contains the following variables. Abbreviations of the headers are displayed between brackets and might be used in the tables further in this report.

- **Accident ID** - *Accident index*
- **Location parameters of the accidents** - *Location Easting OSGR (LE), Location Northing OSGR (LN), Longitude (Long), Latitude (Lat), LSOA of accident location (LSOA), Urban or Rural area (UoR)*
- **Information about the police force** - *Police force, Did police officer attend scene of accident (DPOA-SOA), Local Authority Highway (LAH), Local Authority District (LAD)*
- **Information about the accident** - *Accident severity, Number of vehicles (# vehicles), Number of casualties (# casualties)*
- **Time of the accident** - *Date, Day of Week, Time, Year*

- **Information about the road of the accident** - *1st Road class (1RC), 1st Road number (1RN), Road type, Speed limit*
- **Information about the junction of the accident** - *2nd Road class (2RC), 2nd Road number (2RN), Junction detail, Junction control*
- **Pedestrian crossing information** - *Pedestrian crossing human control (PCHC), Pedestrian crossing physical facilities (PCPF)*
- **Conditions at the site** - *Light conditions (Light), Weather conditions (Weather), Road surface conditions (Road surface), Special conditions at site (Special), Carriageway Hazards*

2.4 Pre-processing

The dataset as presented by Kaggle.com (2018) is not suitable for mining association rules. Pre-processing was required to create a suitable item set format for the association rule mining.

First of all identifiers such as the LSOA, Accident Index, 1st Road Number and 2nd Road Number were removed. These variables would never enter a frequent item set as they are case specific identifiers, keeping these would only slow down the algorithm without adding value. Next, values that were unknown, either labelled as 'unknown' or without a value, were set to NaN. All accidents with at least 1 missing value were dropped from the table to ensure proper processing of the data.

The array was turned into a sparse Boolean array indicating which conditions applied to a specific accident, in a way creating a table with 'items' that either were or weren't there in the transaction. Finally this Boolean array was used to create an array of frequent item sets present in the data.

Some numeric data was binned into categories; a day is divided into 'Rush Hour','Midday' and 'Night' and the week is binned into 'weekend' and 'weekday'.

2.5 Apriori algorithm

The Apriori algorithm works by identifying frequent item sets in the database and extending these frequent item sets until a threshold value for the support of these item sets is reached. The **support** of an item set is defined as:

$$\text{Support(X)} = \frac{\text{Item sets containing X}}{\text{Total Item sets}} \tag{1}$$

From these frequent item sets association rules are generated. In order to compare the quality of the mined association rules, a set of quality measures is introduced.
The **confidence** is similar to the conditional probability of Y given X, and tells us how likely we are to see item set Y given item set X.

$$\text{Confidence(X} \Rightarrow \text{Y)} = \frac{\text{Support(X} \cup \text{Y)}}{\text{Support(X)}} \tag{2}$$

The **lift** measures how much more often Y is observed given X, than if X and Y were independent. A score of 1 implies X and Y are independent.

$$\text{Lift(X} \Rightarrow \text{Y)} = \frac{\text{Confidence(X} \Rightarrow \text{Y)}}{\text{Support(Y)}} \tag{3}$$

The **leverage** computes the difference between the observed frequency of X ∪ Y compared to the assumption that X and Y are independent. A score of 0 implies X and Y are independent.

$$\begin{aligned}\text{Leverage(X} \Rightarrow \text{Y)} = \text{Support(X} \Rightarrow \text{Y)} \\ -\text{Support(X)} \times \text{Support(Y)}\end{aligned} \tag{4}$$

The **conviction** calculates how much more often the association rule would be wrong if X and Y were completely independent.

$$\text{Conviction(X} \Rightarrow \text{Y)} = \frac{1 - \text{Support(Y)}}{1 - \text{Confidence(X} \Rightarrow \text{Y)}} \tag{5}$$

If **Kulczynski** is near 0 or 1, then we have an interesting rule that is negatively or positively associated, respectively.

$$\text{Kulczynski(X} \Rightarrow \text{Y)} = \frac{1}{2}\left(P(X|Y) + P(Y|X)\right) \tag{6}$$

The **Imbalance Ratio** tells us how the data is balanced, where 0 is perfectly balanced and 1 is very skewed.

$$\text{IR(X} \Rightarrow \text{Y)} = \frac{|\text{Support(X)} - \text{Support(Y)}|}{\text{Support(X)} + \text{Support(Y)} - \text{Support(X} \cup \text{Y)}} \tag{7}$$

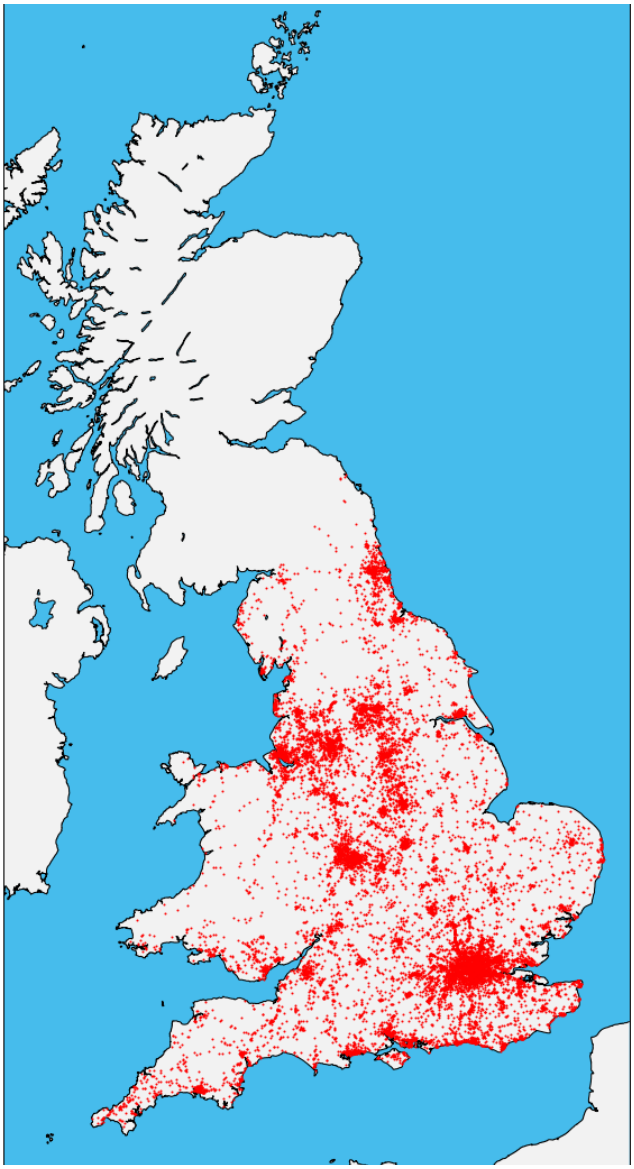

Figure 1: Fatal accidents in the UK in the period 2005-2014, excluding 2008.

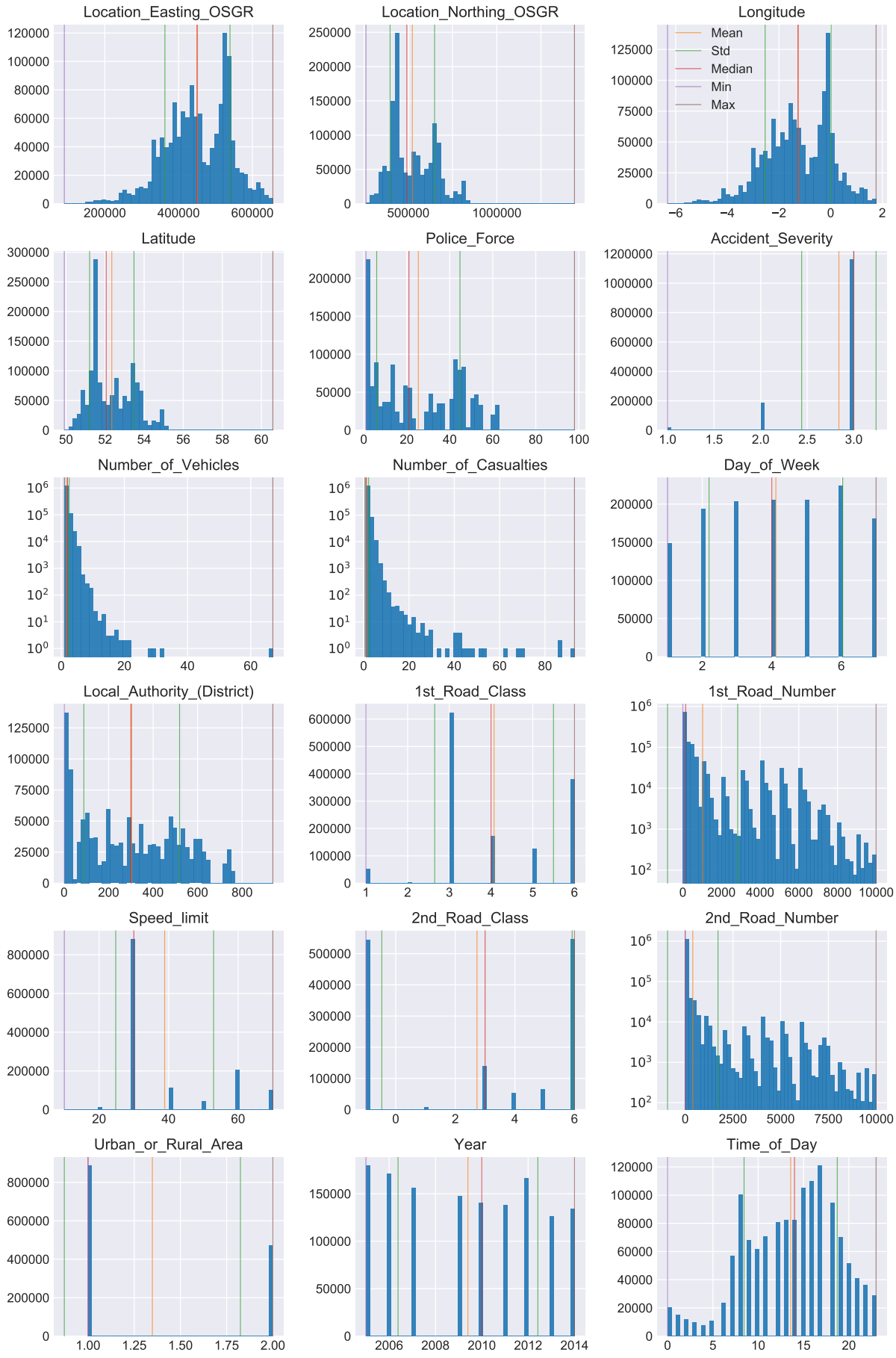

Figure 2: This figure shows the distribution, Mean, Median, Standard Deviation, Minimum and Maximum of each numerical variable in the dataset.

Antecedents			Consequents		Measure
PCPF	No physical crossing within 50 meters	→	**PCHC**	None within 50 metres	0.9982
DPOASOA	Yes				
Special	None	→	**PCHC**	None within 50 metres	0.9982
PCPF	No physical crossing within 50 meters				
DPOASOA	Yes				
Special	None	→	**PCHC**	None within 50 metres	0.9982
Weather	Fine without high winds				
PCPF	No physical crossing within 50 meters				
DPOASOA	Yes				

(a) Top 3 of measure confidence

Antecedents			Consequents		Measure
Special	None	→	**Carriageway Hazards**	None	1.3482
Speed limit	30		**UoR**	1	
Carriageway Hazards	None	→	**Special**	None	1.3482
UoR	1		**Speed limit**	30	
Carriageway Hazards	None	→	**Special**	None	1.3481
UoR	1		**Speed limit**	30	
PCHC	None within 50 metres				

(b) Top 3 of measure lift

Antecedents			Consequents		Measure
PCHC	None within 50 metres'	→	**Special**	None	-0.01622
DPOASOA	Yes		**Accident Severity**	3	
DPOASOA	Yes	→	**Special**	None	-0.01620
			Accident Severity	3	
PCHC	None within 50 metres	→	**Accident Severity**	3	-0.01604
DPOASOA	Yes		**Carriageway Hazards**	None	

(c) Top 3 of measure leaverage

Antecedents			Consequents		Measure
Special	None	→	**Weather**	Fine without high winds	10.938
Light	Daylight: Street light present				
Carriageway Hazards	None				
PCHC	None within 50 metres				
Road Surface	Dry				
Special	None	→	**Weather**	Fine without high winds	10.918
Light	Daylight: Street light present				
Road Surface	Dry				
Carriageway Hazards	None				
Special	None	→	**Weather**	Fine without high winds	10.840
Light	Daylight: Street light present				
Road Surface	Dry				
PCHC	None within 50 metres				

(d) Top 3 of measure conviction

Table 1: Top 3 association rules of Experiment 1 generated by the Apriori algorithm for different measures

Antecedents			Consequents	
DoW	Weekday	→	**ToD**	Rush Hour
Weather	Fog or mist			
Weather	Snowing without high winds	→	**DoW**	Weekday
PCPF	No physical crossing within 50 meters		**Road Type**	Single carriageway
			Accident Severity	3
Road Type	Roundabout	→	**Weather conditions**	Raining
ToD	Rush Hour		**Accident Severity**	3
			DoW	Weekday
DoW	weekday	→	**Accident Severity**	3
Road Types	roundabout		**ToD**	Rush hour

Table 2: Association Rules regarding special weather of Experiment 3, as produced by the Apriori algorithm

3 Results

3.1 Statistical Report

The data set contains a mix of data types consisting of floats, integers and strings. For the integers and floating point data statistics can be performed to clarify how the data is distributed and therefore what is to be expected when performing further analysis. In Table 4 we have made a table of these values. In Figure 2 the distributions are shown. Note that some of these variables identify categories and therefore the statistics lose part of their meaning. However, they are still revealing with regards to the skewness of the distribution.

3.2 Experiment 1: General rules

We test the Apriori algorithm by trying to find big general rules for the traffic accidents data set. For the support a threshold of 50% was used as the rules should apply often. Furthermore, a confidence threshold of 80% ensures a high chance of occurrence of the association rule. In order to ensure that the rule is of interest we use the filtering rules $0.15 <$ kulczynski or kulczynski > 0.85 and Imbalance > 0.5, see `measures.csv`.

As seen in table 3, the performance over all metrics is in general poor. However, the conviction manages to reach some higher values, but since the other metrics do not follow this indicates that the found associations are most likely highly independent. In Table 3 the best rules per metric have been listed.

3.3 Experiment 2: Things the Government can control

The government of the UK, or any of the local counties, may want to use the data of the accidents to reduce the amount of accidents. As we do not have the accidents that did not happen, we can focus on making the accidents less severe. The question we can ask ourselves is "How can we influence the accidents to be less severe, so that we save lives and reduce the damage caused by the accident?" We focus on the direct influence the authorities can make, i.e. condition of the roads, lights available, crossing facilities, speed limit etc. We do not use the conditions the government cannot control like the weather, date, number of vehicles or place where the accidents happen. One can say that the government can run a campaign for driving while it's dry, or have people move to other places, but it will be harder to set those things in motion then to improve the roads, so we focus on the latter.

Quality metric	Maximum value
Confidence	0.998
Lift	1.348
Leverage	0.145
Conviction	10.937

Table 3: This table displays the maximum values encountered in the set of rules of Experiment 1, given the various quality metrics.

After pre-processing the following columns of our data remain:

- Accident Severity
- Road Type
- Special Conditions at Site
- Light Conditions
- Pedestrian-Crossing-Physical Facilities
- Junction Control
- Speed Limit
- Road Surface Conditions

There are some logical rules produced by the Apriori algorithm, some of the important ones are:

$$\text{Street lights present} \rightarrow \text{accident severity-3}$$
$$\text{Speed limit} = 30 \rightarrow \text{accident severity-3}$$
$$\text{Road surface conditions} = \text{dry} \rightarrow \text{accident severity-3}$$

This implies that street light should reduce the accident severity, which makes sense if a road can be seen properly there would be less accidents, and if one happens it would be less severe as the driver can anticipate on the accident as the braking distance is small with this speed limit and road condition.

The accidents which are not severe, namely accident severity-3, are dominant in our data-set, this is a good thing as most accidents are non-lethal. If we want to prevent deaths, we would like to analyze the most severe accidents which are sparse in our data-set. If we want to analyze it properly we need to set the condition for the support and the threshold for the confidence very low, namely, `sup=0.001` and `min_threshold = 0.0`. Some rules are filtered out by selecting only the rule with a Kulzcinsky measure smaller than 0.15 or bigger than 0.85 and with an Imbalance Ratio that is bigger than 0.5.

From the generated association rules we only look at frequent item sets, where the fatal accidents (accident severity-1) occur. The correlations found from these frequent item sets and the rules can be used by the government to prevent fatal accidents. These association rules can be found in the attached file `accident1.csv`.

Some interesting patterns arise from this; the counter-intuitive one is that a speed-limit of 60 is more correlated with a fatal accident than a speed-limit of 70. This would imply that driving faster would be more safe, which is most likely wrong. It can be that the roads where the speed-limit is 70 are designed to be more safe than the roads with a speed-limit of 60, which is the more standard speed limit in the UK. Accidents happen more frequently on roads which are not lighted at night in comparison to the roads which are lighted at night, this can be seen in the Lift, Confidence, Leverage and Conviction measures. There is also a negative correlation between the speed-limit=30 and the accident severity =1. This means that reducing the speed-limit lowers the chances at a fatal accident, which we would expect from common sense.

	Mean	Median	Min	Max	Standard Deviation
Location Easting OSGR (LE)	451063.63	449760.00	90180.00	655370.00	88238.72
Location Northing OSGR (LN)	272711.11	241723.00	10290.00	1189600.00	125809.47
Longitude (Long)	-1.26	-1.26	-6.32	1.76	1.28
Latitude (Lat)	52.34	52.06	49.91	60.59	1.14
Police Force	25.37	21.00	1.00	98	19.37
Accident Severity	2.84	3.00	1.00	3.00	0.40
Number of Vehicles (# vehicles)	1.84	2.00	1.00	67.00	0.71
Number of Casualties (# casualties)	1.36	1.00	1.00	93.00	0.83
Day of Week	4.12	4.00	1.000	7.00	1.92
Local Authority District (LAD)	303.85	300.00	1.00	938.00	215.10
1st Road Class (1RC)	4.07	4.00	1.00	6.00	1.42
1st Road Number (1RN)	1023.32	146.00	0.00	9999.00	1814.29
Speed limit	38.87	30.00	10.00	70.00	14.08
2nd Road Class (2RC)	2.73	3.00	-1.00	6.00	3.19
2nd Road Number (2RN)	396.08	0.00	-1.00	9999.00	1321.11
Urban or Rural Area (UoR)	1.35	1.00	1.00	2.00	0.48
Year	2009.40	2010.00	2005.00	2014.00	3.01
Time of Day	13.58	14.00	0.00	23.00	5.14

Table 4: A table containing the simple statistical analysis of the numerical variables

What we can do as a government:

- Make sure that all the roads are illuminated by lights during the night, so that there will be less frequent (fatal) accidents.
- Adjust the speed-limit in certain areas from 40 to 30, to make the roads more safe.
- Inspect the difference between the roads with a speed-limit of 60 and 70, to find a clarification for the counter-intuitive rule.

3.4 Experiment 3: Weather conditions

For this experiment the goal is to see if bad weather has a bigger probability to cause accidents on particular types of roads and at particular times of the day. If this is the case the government can take precautions when this type of weather arrives to make sure there are as little accidents as possible.
During pre-processing the following variables are selected:

- Weather Conditions
- Road Type
- Accident Severity
- Day of Week
- Time of day

The data is filtered to exclude cases with 'Fine without high winds', as we are only interested in cases where special weather conditions apply. For the support threshold 1% was used as many weather conditions occur sparsely in the data set. Association rules were selected with a confidence threshold of at least 50% to make sure that the chance of occurring would be high given the antecedents. A government would want to be sure that that there will be a high chance of preventing some occurrence. Finally, the rules are filtered using a Kulczynski measure smaller than 0.15 or higher than 0.85 and an Imbalance Ratio higher than 0.5. The final result is ordered by the Lift. Similar rules are removed and the final result is shown in Table 2

We notice a couple of interesting association rules. The first one tells us that if it's foggy outside on working days, accidents are most likely to occur during Rush Hour. This second one states that snow during rush hour is particularly dangerous on single carriageways on weekdays, and will likely cause a light accidents. The third one says that an accident on a weekday on a roundabout will probably happen during rush hour and have severity 3. Note that this is not related to the weather and demonstrates how difficult it can be to generate rules related to a specific category/topic.

From all these rules it is clear that rush hour is a dangerous period to be on the road regardless of the type of weather, most likely because any difficult weather hinders the driving ability of people during the already stressful conditions of rush Hour.

4 Discussion

During the experiments for this data set it seems that this type of mining doesn't work for this kind of data set. The Apriori algorithm tends to return related conditions. Such as the type of road and the speed limit, or the time of day and if the road lights were turned on. This is inherent to the Apriori algorithm as it searches for these types of occurring conditions, but this makes it difficult to find more interesting patterns that might occur a lot less often. As also shown in Experiment 3 it can be difficult to focus the algorithm on a specific topic. Even with very few variables present, the main topic of interest was not present in the Top 3 association rules. This demonstrates that the data requires a lot of pre-processing for each research question one might ask, enhancing the risk of biasing the data towards the expected answers. Furthermore filtering the data set using the Kulczynski metric is a computationally expensive operation which made running the code costly time wise. The majority of the generated rules did not offer new insights into the problems at hand, but rather followed common sense conditions that would also make sense without the use of the algorithm. It would therefore make much more sense for this data set to use different algorithms that can handle continuous variables such as time, and to make for example heat maps to deduce correlations.

During the writing of the report we noticed an error in the code filtering the rules based on the Imbalance and the Kulczynski. In order to apply this condition we used $0.15 <$ Kulczynski < 0.85 AND Imbalance < 0.5, but this should have been OR. Therefore not all rules were properly filtered and some were kept which should have been discarded. This error in the code might partially explain why a lot of non-insight association rules are present. However, this does not fully explain the problem as often the amount of rules left after the incomplete filtering was already small.

5 Conclusion

In this project the associating rules in the traffic accidents data set were mined. The Apriori algorithm was used in three separate experiments to gauche its performance. Rules were found connecting variables the government can control to the severity of the accident, making it easier for the government to take action preventing severe accidents. Rules were also found connecting bad weather and the time of day and type of road to anticipate where and when bad weather might cause the most problems. While some meaningful association rules were found and reported, it is also concluded that the data set did not have the right format to be mined using the Apriori algorithm, making it difficult to mine new insights, if at all present.

Bibliography

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1-55860-153-8. URL http://dl.acm.org/citation.cfm?id=645920.672836.

Kaggle.com. 1.6 million uk traffic accidents, 2018. URL https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales/.

Heidi Landenberger. Trafikipy, March 2018. URL https://github.com/hmlanden/TrafikiPy.

Python Core Team. *Python: A dynamic, open source programming language, version 3.6.6*. Python Software Foundation, 2018. URL https://www.python.org/.

Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018. doi: 10.21105/joss.00638. URL http://joss.theoj.org/papers/10.21105/joss.00638.

U.K. Department of Transport. 1.6 million traffic accidents in the u.k., 2017. URL https://www.dft.gov.uk/traffic-counts/download.php.

World Health Organization. Number of road traffic deaths, July 2016. URL http://www.who.int/gho/road_safety/mortality/traffic_deaths_number/en/.