Data Mining Project Report

Priyank Prajapati
Id-201501138

Analysis of road accidents data

There are a lot of vehicles driving on the roadway every day, and traffic accidents could happen at any time anywhere. Some accident involves fatality, means people die in that accident. As human being, we all want to avoid accident and stay safe. To find out how to drive safer, data mining technique could be applied on the traffic accident dataset to find out some valuable information, thus give driving suggestion.

Road and traffic accidents are uncertain and unpredictable incidents and their analysis requires the knowledge of the factors affecting them. Road and traffic accidents are defined by a set of variables which are mostly of discrete nature. The major problem in the analysis of accident data is its heterogeneous nature. Thus heterogeneity must be considered during analysis of the data otherwise, some relationship between the data may remain hidden. Although, researchers used segmentation of the data to reduce this heterogeneity using some measures such as expert knowledge, but there is no guarantee that this will lead to an optimal segmentation which consists of homogeneous groups of road accidents. Therefore, cluster analysis can assist the segmentation of road accidents.

Data preprocessing:

Data preprocessing mainly deals with removing noise, handle missing values, removing irrelevant attributes in order to make the data ready for the analysis. In this step, our aim is to preprocess the accident data in order to make it appropriate for the analysis.

Accident data for this project were obtained from data.gov.uk web-site. The data set consists of 18k road accidents for 7 years period from 2009 to 2015 in different csv files. Firstly I combined all files in one csv files using pandas library in python and perform data cleaning and data transformation on it. Some of the features are deleted from the dataset. Selected features are converted to given below form:

Dataset Description and transformation:

Number Of vehicles:	**Season:(converted from date)**	**Time :**
1 => 1 2 => 2 3+ => 3	{JAN, FEB, DEC} => {WNT} {MAR, APR, MAY} => {SPR} {JUN, JULY, AUG} => {SMR} {SEP, OCT, NOV} => {FALL}	0 to 4 am => T1 4 to 8 am => T2 8 am to 12 pm => T3 12 to 4 pm => T4 4 to 8 pm => T5 8 to 12 pm => T6

Surface:	**Light:**	**Causality class:**
Dry => DRY Wet / Damp => WET Snow =>SNOW Frost / Ice => ICE Flood => FLD	Daylight: street lights present =>DLT Daylight: no street lighting => DLT Daylight: street lighting unknown => DUS Darkness: street lights present and lit => RLT Darkness: street lights present but unlit => RLT Darkness: no street lighting => NLT Darkness: street lighting unknown => DUS	Driver or rider => Driver Vehicle or pillion passenger=> passenger Pedestrian => Pedestrian
Gender:	**Weather:**	**Casualty Severity:**
Male => M Female => F	Fine without high winds = > FINE Raining without high winds => RAIN Snowing without high winds => SNOW Fine with high winds => FINE Raining with high winds=> RAIN Snowing with high winds => SNOW Fog or mist – if hazard => FOG Other => OTH Unknown => OTH	Fatal => Fatal Serious => Serious Slight => Slight
Age Group:	**Type of Vehicle:**	
Age < 20 => CHD Age between 20 to 30 => YNU Age between 30 to 60 => ADL Age > 60 => SNR	Pedal cycle, motor cycle => TWH Taxi/Private hire car => CAR Car => CAR Minibus (8 – 16 passenger seats) => BUS Bus or coach (17 or more passenger seats) => BUS Ridden horse => ANI	

Clustering algorithm

The objective of clustering algorithm is to divide the data into different clusters or groups such that the objects within a group are similar to each other whereas objects in other clusters are different from each other. I have used K-modes clustering which is an enhanced version of K means algorithm.

Distance measure

Given a data set D, the distance between two objects X and Y, where X and Y are described by N categorical variables, can be computed as follows:

$$d(x, y) = \sum_{i=1}^{n} \delta(\mathbf{Xi}, Yi)$$ **Where δ(Xi, Yi) = 0 if Xi=Yi**

=1 if Xi != Yi

In the above equations, Xi and Yi are the attribute i values in object X and Y. This distance measure is often referred as simple matching dissimilarity measure.

In order to cluster the data set D into k cluster, K-modes clustering algorithm perform the following steps:

1. Initially select k random objects as cluster centers or modes.

2. Find the distance between every object and the cluster center using distance measure.

3. Assign each object to the cluster whose distance with the object is minimum.
So suppose one cluster consists three nodes A = [1 1 0] , B=[0 1 1], C=[0 0 1] then the new updated mode will be M=[0 1 1] as first column has highest '0' frequency and other 2 have highest '1' frequency.

4. Select a new center or mode for every cluster and compare it with the previous value of center or mode; if the values are different, continue with step 2.

After dividing all data cluster wise, It will directly go to it's cluster files in a clusters folders. All cluster wise divided data are used for Association rule mining and analysis.

Analysis of cluster 1: It consists two wheeler accidents that occurred in summer season. In this cluster above 70% of accidents involved two vehicles and sex of causality of all people is male who are young and adult. The surface of the road is Dry. All accident are occur in daylight and between 12 pm to 8 pm. Around 25% of accident occur in rainy weather. The class of casualty is passenger.

Analysis of cluster 2: It consists Car accidents that involved passenger as a class of causality. All the passenger as a class of causality are In this cluster above 80% of accidents involved two vehicles. All accident are occur in darkness when street light is present and between 4 pm to 8 pm in rainy weather. Other cluster analysis and graphs are submitted with the project.

Association rules

Association rule mining is a very popular data mining technique that extracts interesting and hidden relations between various attributes in a large data set. Association rule mining produces a set of rules that define the underlying patterns in the data set. The associativity of two characteristics of accident is determined by the frequency of their occurrence together in the data set. A rule A → B indicates that if A occurs then B Will also occur. I used Apriori algorithm for analysis the association rules on each clusters. The output of the code will show frequent item sets, association rules, support, confidence and lift. The result of cluster 5 is:

Min support : 65% Min confidence: 65%
----------FREQUENT ITEMSET----------
['CAR'], ['Slight'], ['F'], ['WET'], ['ADL'], ['CAR', 'Slight'], ['CAR', 'WET'], ['Slight', 'WET'], ['ADL', 'CAR'], ['ADL', 'Slight'], ['ADL', 'WET']

----------ASSOCIATION RULES----------

RULES	SUPPORT	CONFIDENCE	LIFT
Rule# 1 : ['Slight'] ==> ['CAR']	70	81	0.939733
Rule# 2 : ['CAR'] ==> ['Slight']	70	88	1.112680
Rule# 3 : ['WET'] ==> ['CAR']	68	80	0.943699
Rule# 4 : ['CAR'] ==> ['WET']	68	86	1.082838
Rule# 5 : ['Slight'] ==> ['WET']	72	84	0.972137
Rule# 6 : ['WET'] ==> ['Slight']	72	85	1.003145
Rule# 7 : ['ADL'] ==> ['CAR']	69	78	0.877037
Rule# 8 : ['CAR'] ==> ['ADL']	69	87	1.104154
Rule# 9 : ['Slight'] ==> ['ADL']	77	89	1.033346
Rule# 10 : ['ADL'] ==> ['Slight']	77	86	0.971851
Rule# 11 : ['WET'] ==> ['ADL']	74	87	1.025437
Rule# 12 : ['ADL'] ==> ['WET']	74	83	0.934603

Other cluster's association rules are submitted with project.

Conclusion:

In this project, I proposed a framework for analyzing accident patterns for different types of accidents on the road which makes use of K modes clustering and association rule mining algorithm. The study uses 18,776 accidents that have occurred on during 2009 to 2015 in UK. K modes clustering find six cluster (C1–C6) based on attributes number of vehicles, road type, lightning on road and road feature. Association rule mining have been applied on each cluster as well as on main data to generate rules. Strong rules with high lift values are taken for the analysis. Rules for every cluster reveal the circumstances associated with the accidents within that cluster. These rules are compared with the rules generated for the main data file and comparison shows that association rules for main data file does not reveal appropriate information that can be associated with an accident. More information can be identified if more feature are available that is associated with an accident. To strengthen our methodology, I also performed trend analysis of all data on monthly and hourly basis. The results of trend analysis also supports our methodology that performing clustering prior to analysis helps in identify better and useful results that we cannot obtained without using cluster analysis.

References:

- https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-015-0035-y
- https://analyticsdefined.com/using-k-modes-clustering-categorical-data/
- https://www.kdnuggets.com/2016/03/doing-data-science-kaggle-walkthrough-cleaning-data.html/2
- https://data.gov.uk/dataset/road-traffic-accidents
- https://pypi.python.org/pypi/kmodes/