Association Rules Mining - Accidents

V2 Maestros

Contents

Problem Statement

The input dataset contains information about 1000 fatal accidents. It has different feature variables associated with the accident. The goal is to find patterns in the variables - which accident conditions frequently occur together.

Techniques Used

1. Association Rules Mining
2. Converting Feature data into Basket Data

Data Engineering & Analysis

```
setwd("C:/Personal/V2Maestros/Modules/Machine Learning Algorithms/Association Rules  Mining")

accident_data <- read.csv("accidents.csv")

str(accident_data)
```

Loading and understanding the dataset

```
## 'data.frame':    1000 obs. of  16 variables:
##  $ Accident_Index                            : Factor w/ 620 levels "2.01E+12","2.01E+264",..: 10 1
##  $ Police_Force                              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Accident_Severity                         : int  3 3 3 3 3 3 3 3 2 3 ...
##  $ Number_of_Vehicles                        : int  3 1 2 2 1 2 2 1 3 2 ...
```

```
##  $ Number_of_Casualties                       : int  2 1 1 1 1 1 2 1 2 1 ...
##  $ Day_of_Week                                : int  2 3 3 7 6 5 7 7 5 3 ...
##  $ Local_Authority_.District.                 : int  1 2 2 3 11 4 4 13 14 14 ...
##  $ Road_Type                                  : int  3 2 6 6 6 3 6 3 6 6 ...
##  $ Speed_limit                                : int  30 30 30 30 30 50 30 50 30 30 ...
##  $ Junction_Detail                            : int  3 3 8 6 0 6 3 0 3 6 ...
##  $ Pedestrian_Crossing.Physical_Facilities    : int  5 5 0 1 1 5 0 0 0 0 ...
##  $ Light_Conditions                           : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ Weather_Conditions                         : int  1 1 1 1 1 2 1 1 1 1 ...
##  $ Road_Surface_Conditions                    : int  1 1 1 1 1 2 2 1 1 1 ...
##  $ Urban_or_Rural_Area                        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Did_Police_Officer_Attend_Scene_of_Accident: int  1 1 1 2 1 1 1 1 1 1 ...
```

```
summary(accident_data)
```

```
##        Accident_Index  Police_Force  Accident_Severity Number_of_Vehicles
##  2.01E+12     :381    Min.   : 1.0   Min.   :1.00      Min.   :1.00
##  2.01E+264    :  1    1st Qu.:10.0   1st Qu.:3.00      1st Qu.:1.00
##  200501BS70049:  1    Median :31.0   Median :3.00      Median :1.00
##  200501BS70468:  1    Mean   :32.4   Mean   :2.76      Mean   :1.58
##  200501CP00113:  1    3rd Qu.:46.0   3rd Qu.:3.00      3rd Qu.:2.00
##  200501CP00275:  1    Max.   :98.0   Max.   :3.00      Max.   :5.00
##  (Other)      :614
##  Number_of_Casualties  Day_of_Week   Local_Authority_.District.
##  Min.   :1.00         Min.   :1.00   Min.   :  1
##  1st Qu.:1.00         1st Qu.:2.00   1st Qu.:146
##  Median :1.00         Median :4.00   Median :346
##  Mean   :1.49         Mean   :4.09   Mean   :374
##  3rd Qu.:2.00         3rd Qu.:6.00   3rd Qu.:544
##  Max.   :8.00         Max.   :7.00   Max.   :939
##
##    Road_Type     Speed_limit   Junction_Detail
##  Min.   :1.00   Min.   :20.0   Min.   :0.00
##  1st Qu.:6.00   1st Qu.:30.0   1st Qu.:0.00
##  Median :6.00   Median :30.0   Median :1.00
##  Mean   :5.24   Mean   :40.8   Mean   :1.97
##  3rd Qu.:6.00   3rd Qu.:60.0   3rd Qu.:3.00
##  Max.   :9.00   Max.   :70.0   Max.   :9.00
##
##  Pedestrian_Crossing.Physical_Facilities Light_Conditions
##  Min.   :0.000                           Min.   :1.00
##  1st Qu.:0.000                           1st Qu.:4.00
##  Median :0.000                           Median :4.00
##  Mean   :0.671                           Mean   :4.29
##  3rd Qu.:0.000                           3rd Qu.:6.00
##  Max.   :8.000                           Max.   :7.00
##
##  Weather_Conditions Road_Surface_Conditions Urban_or_Rural_Area
##  Min.   :1.00       Min.   :1.00            Min.   :1.00
##  1st Qu.:1.00       1st Qu.:1.00            1st Qu.:1.00
##  Median :1.00       Median :1.00            Median :1.00
##  Mean   :1.64       Mean   :1.43            Mean   :1.39
##  3rd Qu.:1.00       3rd Qu.:2.00            3rd Qu.:2.00
##  Max.   :9.00       Max.   :4.00            Max.   :2.00
```

```
##
##  Did_Police_Officer_Attend_Scene_of_Accident
##  Min.   :-1.00
##  1st Qu.: 1.00
##  Median : 1.00
##  Mean   : 1.09
##  3rd Qu.: 1.00
##  Max.   : 2.00
##
```

```
head(accident_data)
```

```
##   Accident_Index Police_Force Accident_Severity Number_of_Vehicles
## 1  200501CW10664            1                 3                  3
## 2  200501CW11407            1                 3                  1
## 3  200501EO40954            1                 3                  2
## 4  200501EO41326            1                 3                  2
## 5  200501FH10618            1                 3                  1
## 6  200501GD10263            1                 3                  2
##   Number_of_Casualties Day_of_Week Local_Authority_.District. Road_Type
## 1                    2           2                          1         3
## 2                    1           3                          2         2
## 3                    1           3                          2         6
## 4                    1           7                          3         6
## 5                    1           6                         11         6
## 6                    1           5                          4         3
##   Speed_limit Junction_Detail Pedestrian_Crossing.Physical_Facilities
## 1          30               3                                       5
## 2          30               3                                       5
## 3          30               8                                       0
## 4          30               6                                       1
## 5          30               0                                       1
## 6          50               6                                       5
##   Light_Conditions Weather_Conditions Road_Surface_Conditions
## 1                4                  1                       1
## 2                4                  1                       1
## 3                4                  1                       1
## 4                4                  1                       1
## 5                4                  1                       1
## 6                4                  2                       2
##   Urban_or_Rural_Area Did_Police_Officer_Attend_Scene_of_Accident
## 1                   1                                         1
## 2                   1                                         1
## 3                   1                                         1
## 4                   1                                         2
## 5                   1                                         1
## 6                   1                                         1
```

Data Transformation The data frame needs to be converted into a Basket form to be loaded by the arules dataset. The following custom code does it.

```
#get column names of the data set
colnames <- names(accident_data)

#Start building a file in basket format - one row per transaction and each column value becoming
# a basket item in the format <column_name>=<column_value>

basket_str <- ""
for ( row in 1:nrow(accident_data))  {

  if ( row != 1) {
    basket_str <- paste0(basket_str, "\n")
  }
  basket_str <- paste0(basket_str, row,",")

  for (col in 2:length(colnames)) {
      if ( col != 2) {
        basket_str <- paste0(basket_str, ",")
      }
      basket_str <- paste0(basket_str, colnames[col],"=",accident_data[row,col])
  }

}
write(basket_str,"accidents_basket.csv")
```

Exploratory Data Analysis Typically, for Clustering problems, EDA is only required for finding out outliers and errors. If outliers are found, we would want to eliminate them since they might skew the clusters formed by moving the centeroids signficantly.

```
library(arules)
```

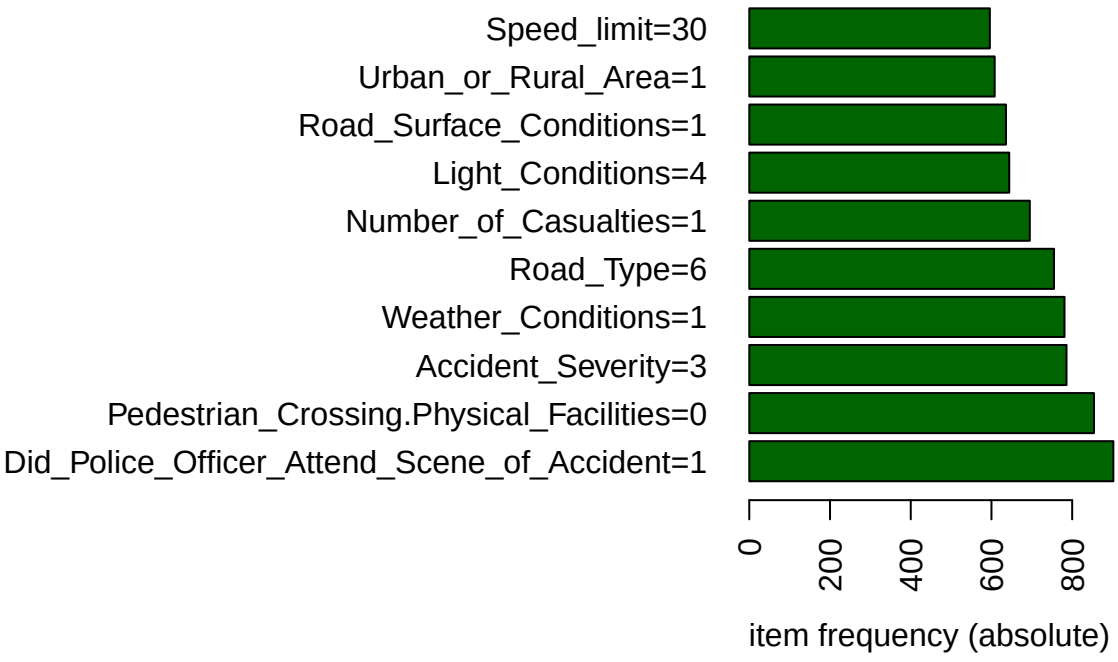
```
## Warning: package 'arules' was built under R version 3.1.1

## Loading required package: Matrix
##
## Attaching package: 'arules'
##
## The following objects are masked from 'package:base':
##
##     %in%, write
```

```
accidents <- read.transactions("accidents_basket.csv",sep=",")
summary(accidents)
```

```
## transactions as itemMatrix in sparse format with
##  1000 rows (elements/itemsets/transactions) and
##  1452 columns (items) and a density of 0.01102
##
## most frequent items:
## Did_Police_Officer_Attend_Scene_of_Accident=1
##                                          902
##     Pedestrian_Crossing.Physical_Facilities=0
```

```
##                                         854
##                             Accident_Severity=3
##                                         786
##                           Weather_Conditions=1
##                                         781
##                                  Road_Type=6
##                                         755
##                                    (Other)
##                                       11922
##
## element (itemset/transaction) length distribution:
## sizes
##   16
## 1000
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      16      16      16      16      16      16
##
## includes extended item information - examples:
##   labels
## 1      1
## 2     10
## 3    100
```

```r
itemFrequencyPlot(accidents,topN=10,type="absolute",
                  col="darkgreen", horiz=TRUE)
```



Modeling & Prediction

We discover the frequently occuring patterns with arules.

```
rules <- apriori(accidents, parameter=list(supp=0.1, conf=0.3))
```

```
##
## parameter specification:
##  confidence minval smax arem  aval originalSupport support minlen maxlen
##         0.3    0.1    1 none FALSE            TRUE     0.1      1     10
##  target   ext
##   rules FALSE
##
## algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## apriori - find association rules with the apriori algorithm
## version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[1452 item(s), 1000 transaction(s)] done [0.00s].
## sorting and recoding items ... [29 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 7 8 9 done [0.01s].
## writing ... [21772 rule(s)] done [0.01s].
## creating S4 object  ... done [0.01s].
```

```
inspect(rules[1:40])
```

```
##    lhs                            rhs                                                 support confidence
## 1  {}                          => {Road_Surface_Conditions=2}                           0.330     0.3300
## 2  {}                          => {Number_of_Vehicles=2}                                0.391     0.3910
## 3  {}                          => {Urban_or_Rural_Area=2}                               0.392     0.3920
## 4  {}                          => {Junction_Detail=0}                                   0.493     0.4930
## 5  {}                          => {Number_of_Vehicles=1}                                0.525     0.5250
## 6  {}                          => {Speed_limit=30}                                      0.596     0.5960
## 7  {}                          => {Urban_or_Rural_Area=1}                               0.608     0.6080
## 8  {}                          => {Road_Surface_Conditions=1}                           0.636     0.6360
## 9  {}                          => {Light_Conditions=4}                                  0.644     0.6440
## 10 {}                          => {Number_of_Casualties=1}                              0.695     0.6950
## 11 {}                          => {Road_Type=6}                                         0.755     0.7550
## 12 {}                          => {Weather_Conditions=1}                                0.781     0.7810
## 13 {}                          => {Accident_Severity=3}                                 0.786     0.7860
## 14 {}                          => {Pedestrian_Crossing.Physical_Facilities=0}           0.854     0.8540
## 15 {}                          => {Did_Police_Officer_Attend_Scene_of_Accident=1}        0.902     0.9020
## 16 {Day_of_Week=5}             => {Did_Police_Officer_Attend_Scene_of_Accident=1}        0.103     0.9196
## 17 {Police_Force=1}            => {Speed_limit=30}                                      0.111     0.9174
## 18 {Police_Force=1}            => {Urban_or_Rural_Area=1}                               0.118     0.9752
## 19 {Police_Force=1}            => {Light_Conditions=4}                                  0.118     0.9752
## 20 {Police_Force=1}            => {Did_Police_Officer_Attend_Scene_of_Accident=1}        0.110     0.9091
## 21 {Weather_Conditions=2}      => {Road_Surface_Conditions=2}                           0.124     0.9841
## 22 {Road_Surface_Conditions=2} => {Weather_Conditions=2}                                0.124     0.3758
## 23 {Weather_Conditions=2}      => {Accident_Severity=3}                                 0.103     0.8175
```

```
## 24 {Weather_Conditions=2}      => {Pedestrian_Crossing.Physical_Facilities=0}       0.102     0.8095
## 25 {Weather_Conditions=2}      => {Did_Police_Officer_Attend_Scene_of_Accident=1}   0.123     0.9762
## 26 {Day_of_Week=6}             => {Road_Type=6}                                    0.106     0.7794
## 27 {Day_of_Week=6}             => {Accident_Severity=3}                            0.104     0.7647
## 28 {Day_of_Week=6}             => {Pedestrian_Crossing.Physical_Facilities=0}      0.110     0.8088
## 29 {Day_of_Week=6}             => {Did_Police_Officer_Attend_Scene_of_Accident=1}   0.122     0.8971
## 30 {Road_Type=3}               => {Road_Surface_Conditions=1}                      0.100     0.6757
## 31 {Road_Type=3}               => {Light_Conditions=4}                             0.100     0.6757
## 32 {Road_Type=3}               => {Weather_Conditions=1}                           0.115     0.7770
## 33 {Road_Type=3}               => {Accident_Severity=3}                            0.122     0.8243
## 34 {Road_Type=3}               => {Pedestrian_Crossing.Physical_Facilities=0}      0.114     0.7703
## 35 {Road_Type=3}               => {Did_Police_Officer_Attend_Scene_of_Accident=1}   0.138     0.9324
## 36 {Accident_Severity=2}       => {Number_of_Vehicles=1}                           0.120     0.6522
## 37 {Accident_Severity=2}       => {Speed_limit=30}                                 0.106     0.5761
## 38 {Accident_Severity=2}       => {Urban_or_Rural_Area=1}                          0.103     0.5598
## 39 {Accident_Severity=2}       => {Road_Surface_Conditions=1}                      0.114     0.6196
## 40 {Accident_Severity=2}       => {Light_Conditions=4}                             0.116     0.6304
```

Looking at the output of the top 40 rules, we see patterns of conditions that frequently occur together. For example " {Weather_Conditions=2} => {Accident_Severity=3} " is an interesting pattern since it tells us that when a given weather condition exists, there is a specific severity of accident that occurs

Conclusions

Association Rules Mining is a powerful and flexible machine learning technique that can be used to find hidden patterns and relationships in large datasets.
