

Associative rule mining Document

Vivek I(IMT2013017)

Aravind K(IMT2013019)

Nikhilesh P(IMT2013031)

Arun Prasad(IMT2013037)

As part of associative rule mining on our california crime data set, we have done the following:

- Discretized continuous attributes in our data, since attribute values must be discrete to apply associative rule mining algorithms.
- Applied apriori algorithm(by setting minsupport and minconfidence) to extract some rules i.e, doing **descriptive analysis** using these results
- Imposed some conditions in our apriori algorithm(on RHS of the implication) and arrived at certain rules i.e, **predictive analysis** using associative rule mining

Discretization of continuous variables :

Our data set has information about number of crimes in each category(like Murder, burglary etc...) for every place in california. So, as part of discretizing these continuous variables, we have converted them into three discrete variables namely L,M and H where L indicates that the number of crimes(of that category) in that place are low, M indicates that the number of crimes(of that category) in that place are medium and H indicates that the number of crimes(of that category) in that place are high.

Descriptive analysis using associative rule mining :

The rules obtaining after applying apriori algorithm(without imposing any conditions(on the data set(with minsupport=0.3 and minconfidence=0.7) are shown below:

The screenshot shows the RStudio interface. The script editor contains R code for discretizing the California crime data. The console displays the output of the `inspect(rules.sorted)` command, showing a list of associative rules with their support and confidence values. The Environment pane on the right shows the data objects created.

```
8 - for(i in c(1:11)){
9   data <- within(data, quartile <- as.integer(cut(data[,i], unique(quantile(data[,i], probs=0:3/3)), include.lowest=TRUE)))
10  data$quartile[data$quartile==1]<-'L'
11  data$quartile[data$quartile==2]<-'M'
12  data$quartile[data$quartile==3]<-'H'
13  data[i]<-data[,quartile]
14  data<-data[,1:11]
15 }
16 - for(i in c(1:11)){
17   data[,i] <- as.factor(data[,i])
18 }
25:1 (Top Level) ↓
```

Console Output:

lhs	rhs	support	confidence	lift
{Property.crime=L}	=> {Larceny..theft=L}	0.3217391	0.9610390	2.870636
{Larceny..theft=L}	=> {Property.crime=L}	0.3217391	0.9610390	2.870636
{Larceny..theft=L}	=> {Property.crime=H}	0.3108696	0.9346405	2.810030
{Property.crime=H}	=> {Larceny..theft=L}	0.3108696	0.9346405	2.810030
{Murder=L,Aggravated.assault=L}	=> {Violent.crime=L}	0.3152174	0.9359474	2.795005
{Violent.crime=L,Murder=L}	=> {Aggravated.assault=L}	0.3152174	0.9354839	2.794302
{Aggravated.assault=L}	=> {Violent.crime=L}	0.3173913	0.9480519	2.777732
{Violent.crime=L}	=> {Aggravated.assault=L}	0.3173913	0.9299363	2.777732
{Violent.crime=L,Aggravated.assault=L}	=> {Murder=L}	0.3152174	0.9931507	1.423207
{Violent.crime=L}	=> {Murder=L}	0.3369565	0.9872611	1.414411
{Aggravated.assault=L}	=> {Murder=L}	0.3304348	0.9870130	1.414411
{Motor.vehicle.theft=L}	=> {Murder=L}	0.3304348	0.9806452	1.405286
{Robbery=L}	=> {Murder=L}	0.3326087	0.9745223	1.396512
{Burglary=L}	=> {Murder=L}	0.3260870	0.9615385	1.377906
{Property.crime=L}	=> {Murder=L}	0.3173913	0.9480519	1.358579
{Larceny..theft=L}	=> {Murder=L}	0.3152174	0.9415584	1.349274
{Population=L}	=> {Murder=L}	0.3152174	0.9415584	1.349274
{Rape=L}	=> {Murder=L}	0.3869565	0.9368421	1.342515
{Murder=L}	=> {Rape=L}	0.3869565	0.5545171	1.342515
{Arson=L}	=> {Murder=L}	0.3956522	0.8834951	1.266068
{Murder=L}	=> {Arson=L}	0.3956522	0.5669782	1.266068

Environment pane:

- data: 460 obs. of 11 variab..
- rules: Formal class rules
- rules.so: Formal class rules

Files pane:

- abind: Combine Multidimensional Arrays 1.4-5
- arules: Mining Association Rules and Frequent Itemsets 1.5-0
- asse...: Easy pre and post assertions. 0.1-0
- base...: Tools for base64 encoding 0.1-3
- BH: Boost C++ Header Files 1.62-1
- bitops: Bitwise Operations 1.0-6
- caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc. 1.17-0

Note: In our data set property crimes are divided into Burglary, Larceny thefts and motor vehicle thefts and violent crimes are divided into Murder, Rape, Robbery and aggravated assault.

Some useful and interesting rules(from the above set of 21 rules) and their interpretation are the following:

[1], [3] – It says that, in a city(in california state) if the property crimes are low (or) high, then the larceny thefts are also low (or) high respectively. Another way of interpreting this is that, among all the property crimes larceny thefts mostly determine the number of property crimes. So, to reduce the number of property crimes in a city, more care can be taken towards reducing the larceny thefts.

[5]- Similar interpretation as above i.e, In a city, among all the violent crimes, murder and aggravated assault mostly determine the number of violent crimes.

[7],[10]-Among aggravated assault and murder, most contributing factor to violent crimes are aggravated assaults since rule [7] has more lift value than rule [10]

[13]- In a city if robberies are less, then murders are less. Another interpretation of this can be that people doing robbery are more prone to kill the people.

[15]-In a city, if the property crimes are low, then the number of murders are also low. So, we can think that people who are involved in burglary, larceny thefts are more prone to kill the people.

[17]-This is a straight forward implication which says that, in a city if the population is low then the number of murders are also low.

[18]-In a city if number of rapes are low then the number of murders are also low. Another interpretation of the same thing is that: people who commit rapes also kill them.

Predictive analysis using associative rule mining :

We want to determine the factors which account for murders in a city. So, we imposed this rule on the RHS in our apriori algorithm and the result we got is the following:

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for loading data, discretizing it into quartiles, and applying the `apriori` function to generate association rules.
- Console:** Displays the output of the `inspect(rules.sorted)` command, showing a list of rules with their left-hand side (lhs), right-hand side (rhs), support, confidence, and lift values.
- Environment:** Shows the loaded data object 'data' with 460 observations and 11 variables, and the generated rules object 'rules'.
- User Library:** Lists installed packages such as 'arules', 'arulesL', 'base64enc', 'BH', 'bitops', and 'caTools'.

Console Output (Rules):

lhs	rhs	support	confidence	lift
{Violent.crime=L,Aggravated.assault=L}	=> {Murder=L}	0.3152174	0.9931507	1.423207
{Violent.crime=L}	=> {Murder=L}	0.3369565	0.9872611	1.414767
{Aggravated.assault=L}	=> {Murder=L}	0.3304348	0.9870130	1.414411
{Violent.crime=L,Rape=L}	=> {Murder=L}	0.3000000	0.9857143	1.412550
{Motor.vehicle.theft=L}	=> {Murder=L}	0.3304348	0.9806452	1.405286
{Robbery=L}	=> {Murder=L}	0.3326087	0.9745223	1.396512
{Burglary=L}	=> {Murder=L}	0.3260870	0.9615385	1.377906
{Property.crime=L}	=> {Murder=L}	0.3173913	0.9480519	1.358579
{Property.crime=L,Larceny..theft=L}	=> {Murder=L}	0.3043478	0.9459459	1.355561
{Population=L}	=> {Murder=L}	0.3152174	0.9415584	1.349274
{Larceny..theft=L}	=> {Murder=L}	0.3152174	0.9415584	1.349274
{Rape=L}	=> {Murder=L}	0.3869565	0.9368421	1.342515
{Arson=L}	=> {Murder=L}	0.3956522	0.8834951	1.266068