Q3

Rahul Atre

2023-09-27

Q3 [Association Rules Mining]

A store that sells accessories for smart phones runs a promotion on faceplates. Customers who purchase multiple faceplates from a choice of 6 different colours get a discount. The store managers, who would like to know what colours of faceplates are likely to be purchased together, collected past transactions in the file Transactions.csv.

1. For each rule, compute the support, confidence, interest, lift, and conviction.

Recall: The above five rule metrics are stated as follows:

$$
\begin{aligned}
Support(X \to Y) &= \frac{Freq(X \cap Y)}{N} \in [0,1] \\
Confidence(X \to Y) &= \frac{Freq(X \cap Y)}{Freq(X)} \in [0,1] \\
Interest(X \to Y) &= Confidence(X \to Y) - \frac{Freq(Y)}{N} \in [-1,1] \\
Lift(X \to Y) &= \frac{N^2 \cdot Support(X \to Y)}{Freq(X) \cdot Freq(Y)} \in (0, N^2) \\
Conviction(X \to Y) &= \frac{1 - Freq(Y)/N}{1 - Confidence(X \to Y)} \geq 0
\end{aligned}
$$

Loading the data from Transactions.csv:

```r
transData = read.csv("Transactions.csv")
```

First, we can create a function that calculates all the rule metrics, given the frequency of x, y, XAndY, and N.

```r
rule_metric_func<- function(freqX, freqY, freqXandY, N) {
  support <- paste("Support: ", freqXAndY/N * 100, "%")
  confidence <- paste("Confidence: ", freqXAndY/freqX * 100, "%")
  interest <- paste("Interest: " , freqXAndY/freqX - freqY/N)
  lift <- paste("Lift: ", (N^2 * freqXAndY/N)/(freqX * freqY))
  conviction <- paste("Conviction: ", (1 - freqY/N)/(1 - freqXAndY/freqX))

  output <- cat(support, confidence, interest, lift, conviction, sep = "\n")
}
```

i) {red, white} → {green}

```r
freqX <- nrow(transData[transData$Red == 1 & transData$White == 1,])
freqY <- nrow(transData[transData$Green == 1,])
freqXAndY <- nrow(transData[transData$Red == 1 & transData$White == 1 & transData$Green == 1,])
N <- nrow(transData)

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  1.12044817927171 %
## Confidence:  38.0952380952381 %
## Interest:  0.0448179271708683
## Lift:  1.13333333333333
## Conviction:  1.07239819004525
```

ii) {green} → {white}

```r
freqX <- nrow(transData[transData$Green == 1,])
freqY <- nrow(transData[transData$White == 1,])
freqXAndY <- nrow(transData[transData$Green == 1 & transData$White == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  1.5406162464986 %
## Confidence:  4.58333333333333 %
## Interest:  -0.0129901960784314
## Lift:  0.779166666666667
## Conviction:  0.986385820703827
```

iii) {red, green} → {white}

```r
freqX <- nrow(transData[transData$Red == 1 & transData$Green == 1,])
freqY <- nrow(transData[transData$White == 1,])
freqXAndY <- nrow(transData[transData$Red == 1 & transData$Green == 1 & transData$White == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  1.12044817927171 %
## Confidence:  25.8064516129032 %
## Interest:  0.199240986717268
## Lift:  4.38709677419355
## Conviction:  1.26854219948849
```

iv) {green} → {red}

```r
freqX <- nrow(transData[transData$Green == 1,])
freqY <- nrow(transData[transData$Red == 1,])
freqXAndY <- nrow(transData[transData$Green == 1 & transData$Red == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  4.34173669467787 %
## Confidence:  12.9166666666667 %
## Interest:  0.0115196078431373
## Lift:  1.09791666666667
## Conviction:  1.0132282578103
```

v) {orange} → {red}

```r
freqX <- nrow(transData[transData$Orange == 1,])
freqY <- nrow(transData[transData$Red == 1,])
freqXAndY <- nrow(transData[transData$Orange == 1 & transData$Red == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  0.280112044817927 %
## Confidence:  3.38983050847458 %
## Interest:  -0.0837487537387837
## Lift:  0.288135593220339
## Conviction:  0.913312693498452
```

vi) {white, black} → {yellow}

```r
freqX <- nrow(transData[transData$White == 1 & transData$Black == 1,])
freqY <- nrow(transData[transData$Yellow == 1,])
freqXAndY <- nrow(transData[transData$White == 1 & transData$Black == 1 & transData$Yellow == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  3.92156862745098 %
## Confidence:  90.3225806451613 %
## Interest:  0.852805638384386
## Lift:  17.9139784946237
## Conviction:  9.81232492997199
```

vii) {black} → {green}

```r
freqX <- nrow(transData[transData$Black == 1,])
freqY <- nrow(transData[transData$Green == 1,])
freqXAndY <- nrow(transData[transData$Black == 1 & transData$Green == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  28.8515406162465 %
## Confidence:  33.2794830371567 %
## Interest:  -0.00333962340994559
## Lift:  0.990064620355412
## Conviction:  0.994994608012697
```

2. Amongst the rules for which the support is positive (> 0), which one has the highest lift? Confidence? Interest? Conviction?

Among rules for which support is positive, the rule with highest lift, confidence, interest, and conviction is vi) {white, black} → {yellow}

3. Build an additional 5-10 candidate rules (randomly), and evaluate them. Which of the 5-10 candidate rules do you think would be most useful for the store managers?

- {red, yellow} -> {orange}

```
freqX <- nrow(transData[transData$Red == 1 & transData$Yellow == 1,])
freqY <- nrow(transData[transData$Orange == 1,])
freqXAndY <- nrow(transData[transData$Red == 1 & transData$Yellow == 1 & transData$Orange == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  0.140056022408964 %
## Confidence:  5.26315789473684 %
## Interest:  -0.0300014742739201
## Lift:  0.63693131132917
## Conviction:  0.968331777155306
```

- {white} -> {black}

```
freqX <- nrow(transData[transData$White == 1,])
freqY <- nrow(transData[transData$Black == 1,])
freqXAndY <- nrow(transData[transData$White == 1 & transData$Black == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  4.34173669467787 %
## Confidence:  73.8095238095238 %
## Interest:  -0.128851540616246
## Lift:  0.851373182552504
## Conviction:  0.508021390374331
```

- {green} -> {yellow}

```
freqX <- nrow(transData[transData$Green == 1,])
freqY <- nrow(transData[transData$Yellow == 1,])
freqXAndY <- nrow(transData[transData$Green == 1 & transData$Yellow == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  1.12044817927171 %
## Confidence:  3.33333333333333 %
## Interest:  -0.0170868347338936
## Lift:  0.661111111111111
## Conviction:  0.982323964068386
```

- {red} -> {white}

```r
freqX <- nrow(transData[transData$Red == 1,])
freqY <- nrow(transData[transData$White == 1,])
freqXAndY <- nrow(transData[transData$Red == 1 & transData$White == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  2.94117647058824 %
## Confidence:  25 %
## Interest:  0.191176470588235
## Lift:  4.25
## Conviction:  1.25490196078431
```

- {green, yellow} -> {red}

```r
freqX <- nrow(transData[transData$Green == 1 & transData$Yellow == 1,])
freqY <- nrow(transData[transData$Red == 1,])
freqXAndY <- nrow(transData[transData$Green == 1 & transData$Yellow == 1 & transData$Red == 1,])

rule_metric_func(freqX, freqY, freqXandY, N) #Calculating Rule Metrics
```

```
## Support:  0.980392156862745 %
## Confidence:  87.5 %
## Interest:  0.757352941176471
## Lift:  7.4375
## Conviction:  7.05882352941176
```

Out of the rules I created, I believe that the first two would be most useful for store managers. For the first rule, according to color science, if you mix red and yellow, the output is typically orange. It would be useful to know if customers are likely to purchase colors that mix together. For the second rule, white and black are opposites that contrast each other on the color spectrum. This is because black is made from "all colors" whereas white has absence of color. It would be insightful to know for store managers if such a concept has any basis in rule metrics.

4. How would one determine reasonable threshold values for the support, coverage, interest, and lift of rules derived from a given dataset?

In general, it is impractical to provide definitive or fixed threshold values for the following rules since it always depends on the specific context of the dataset. As data scientists, it is ideal to explore the association rules first before determining a reasonable threshold. Candidate rules afterwards, can be discarded or accepted in accordance to the metric thresholds determined. Specific sectors for data such as academic, professional, economic, government, military, commercial, etc., could have varying quality requirements, and therefore, seeking perfection in data through a strong threshold can be disruptive to data analysis.