Q2

Rahul Atre

2023-09-26

Q2 [Association Rules Mining]

Evaluate the following candidate association rules for the British Musical Dataset introduced in the course notes.

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

$$Support(X \to Y) = \frac{Freq(X \cap Y)}{N} \in [0, 1]$$

$$Confidence(X \to Y) = \frac{Freq(X \cap Y)}{Freq(X)} \in [0, 1]$$

$$Interest(X \to Y) = Confidence(X \to Y) - \frac{Freq(Y)}{N} \in [-1, 1]$$

$$Lift(X \to Y) = \frac{N^2 \cdot Support(X \to Y)}{Freq(X) \cdot Freq(Y)} \in (0, N^2)$$

$$Conviction(X \to Y) = \frac{1 - Freq(Y)/N}{1 - Confidence(X \to Y)} \geq 0$$

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

```
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) $X \to Y$

```
#Data obtained from notes
freqX <- 3888
freqY <- 9092
freqXAndY <- 2720
```

```
N <- 15356

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

```
## Support:  17.7129460797083 %
## Confidence:  69.9588477366255 %
## Interest:  0.107507206201889
## Lift:  1.18157508341797
## Conviction:  1.35786645352136
```

ii) $W \rightarrow Z$

```
freqW <- 2010
freqZ <- 6855
freqWAndZ <- 132
N <- 15356

rule_metric_func(freqW, freqZ, freqWAndZ, N) #Calculating Rule Metrics
```

```
## Support:  17.7129460797083 %
## Confidence:  135.323383084577 %
## Interest:  0.906828516961947
## Lift:  3.03140170772687
## Conviction:  -1.56721875928664
```

iii) $(Y \cap W) \rightarrow X$

```
freqYAndW <- 1852
freqX <- 3888
freqYAndWAndX <- 1778
N <- 15356

rule_metric_func(freqYAndW, freqX, freqYAndWAndX, N) #Calculating Rule Metrics
```

```
## Support:  17.7129460797083 %
## Confidence:  146.868250539957 %
## Interest:  1.21549157026021
## Lift:  5.80069150023554
## Conviction:  -1.59342210613123
```

- If an individual owns a classical music album (W), then they also own a hip-hop album (Z), given that Freq(W) = 2010, Freq(Z) = 6855, and Freq(W ∩ Z) = 132.
- If an individual owns both the Beatles' Sergeant Peppers' Lonely Hearts Club Band and a classical music album, then they were born before 1976, given that Freq(Y ∩ W) = 1852 and Freq(Y ∩ W ∩ X) = 1778.
- Out of the 3 rules that have been established in the previous question (X → Y [course notes], W → Z, and (Y AND W) → X), which do you think is more useful? Which is more surprising?

Out of the 3 rules that were established, I believe that the second rule is the most useful one. Let's assume that an entertainment company is looking to expand their business into new genres of music, but are unsure

of which ones to explore. As a business, the next logical step would be to analyze data trends in the music landscape. If the company specializes in classical music, and their findings were similar to the second rule, then they can invest capital into hiring hip-hop artists to increase the overall revenue. The company could also bundle two albums (classical and hip-hop) together for a "discounted price" but silently raise the cost of one of the two albums for a higher profit margin. They could also allow for collaboration between hip-hop and classical artists to create innovative music.

The rule that I found most surprising was the third rule, stating that if an individual owns both the Beatles' Sergean Peppers' Lonely Hearts Club Band and a classical music album, they would be born before 1976. There seems to be a very strong correlation in the data, as less than 200 data points were lost when applying the intersection rule. Ideally, data scientists should not try to create artificial meaning from the data (i.e. Assume that the Beatles' were popular within the older generation's culture, therefore the total population liking Beatles' are overwhelmingly old). However, this is quite a surprising statistic as the number of data points lost after applying the intersection is very low. Ultimately, correlation does not equal causation, as a younger individual may find classical music pleasant or enjoy listening to the Beatles'.