Assignment_Q2_MAT3373

John Bute

2024-09-21

Question 2

In order to answer the question we need to declare some variables.

Firstly, W is the logical representation we use to represent the following phrase: An individual who owns a classical music album.

Then, Z represents People who own a hiphop album

Y equals the people who own the Beatles' Seargant Pepper's Lonely Hearts Club Band album.

Finally, X represents people who are born before 1976.

Now we are trying to determine which of the following rules are more useful, and which ones are more surprising:

W -> Z (Y AND W) -> X X -> Y

To determine a rule's strength we compute rule metrics such as frequency, confidence, interest, lift, and conviction of each rule:

```
#say W = individual owns a classical music album
#Say Z = they own a hip-hop album

library(knitr)
library(kableExtra)


Freq_W <- 2010

Freq_Z <- 6855

Freq_W_n_Z <- 132

#own both the seargent peppers lonely hearts club and a classical music album, then they were born befo
#Y = own beatles record....
#W = classical music record
#X = born before 1976

Freq_Y_n_W <- 1852

Freq_Y_n_W_n_X <- 1778
#FROM COURSE NOTES:
```

```r
N <- 15356
Freq_X <- 3888
Freq_Y <- 9092

Freq_X_n_Y <- 2720

#rules to evaluate:

#X -> Y
Support_XY <- Freq_X_n_Y/N
Confidence_XY <- Freq_X_n_Y/Freq_X
Interest_XY <- Confidence_XY - Freq_Y/N
Lift_XY <- (N^2 * Support_XY)/(Freq_X * Freq_Y)
Conviction_XY <- (1 - Freq_Y/N)/ (1 - Confidence_XY)


#W -> Z
Support_WZ <- Freq_W_n_Z/N
Confidence_WZ <- Freq_W_n_Z/Freq_W
Interest_WZ <- Confidence_WZ - Freq_W/N
Lift_WZ <- (N^2 * Support_WZ)/(Freq_W * Freq_Z)
Conviction_WZ <- (1 - Freq_Z/N)/(1 - Confidence_WZ)


# (Y AND W) -> X
Support_YWX <- Freq_Y_n_W_n_X/N
Confidence_YWX <- Freq_Y_n_W_n_X/Freq_Y_n_W
Interest_YWX <- Confidence_YWX - Freq_Y_n_W/ N
Lift_YWX <- (N^2 * Support_YWX)/(Freq_Y_n_W * Freq_X)
Conviction_YWX <- (1 - Freq_X/N)/(1 - Confidence_YWX)

#Table built with the calculated values:
metrics_table <- data.frame(
  Metric = c("Support", "Confidence", "Interest", "Lift", "Conviction"),
  `W -> Z` = c(Support_WZ, Confidence_WZ, Interest_WZ, Lift_WZ, Conviction_WZ),
  `(Y AND W) -> X` = c(Support_YWX, Confidence_YWX, Interest_YWX, Lift_YWX, Conviction_YWX),
  `X -> Y` = c(Support_XY, Confidence_XY, Interest_XY, Lift_XY, Conviction_XY)
)

# Display the table
kable(metrics_table, caption = "Association Rules Metrics") %>%
  kable_styling(full_width = FALSE, position = "left")
```

Table 1: Association Rules Metrics

Metric	W....Z	X.Y.AND.W.....X	X....Y
Support	0.0085960	0.1157854	0.1771295
Confidence	0.0656716	0.9600432	0.6995885
Interest	-0.0652218	0.8394389	0.1075072
Lift	0.1471121	3.7917755	1.1815751
Conviction	0.5925055	18.6904107	1.3578665

Based on this table, we can analyze the strength of each rule. When it comes to W -> Z, the support stands out as it is 0.008%, meaning that this rule appears so rarely that it would be foolish to assume any utility for the rule, as the rule is not significant within the dataset.

On the contrary, (Y and W) -> X and X->Y posted support metrics of 11.5% and 17.7% respectively, meaning that they do somewhat appear within the dataset.

However, (Y and W) -> X stands out when it comes to its confidence metric as it has posted a 96% confidence rate, meaning that we have a 96% probability of finding X if (Y AND W) is true. Therefore, we can assume there is a very strong predictive pattern where if Y AND W, then we will always see X being true. For W->Z, a lowly confidence rate of of 0.06% suggests that there seems to be no link of causality between Z being present and the presence of W. For X-> Y, a confidence rate of 70% (rounded) is not terrible at all, considering it being the highest frequency rule out of the three.

When it comes to the interest, once again (Y AND W) -> X is king, as its high interest metric suggests that this causality relation is not random chance, but rather a strong relationship between the the terms (Y AND W) and X. However, for X -> Y, an interest metric of 0.1 is closest to 0, which generally means that the rule is neither surprising nor strong. Finally, for W -> Z, a negative lift means that the presence of W actually makes Z less likely to appear than random chance. Therefore, maybe a more suitable rule would be W -> NOT Z

The strength of (Y AND W) -> X is demonstrated in its lift rate of 3.79, as it shows that they are positevly correlated. Meanwhile, W->Z is negatively correlated while X->Y 's lift rate of 1.18 is close to 1, which means that the pair are closer to being independent (X and Y are very slightly correlated positively).

Finally the conviction of (Y AND W) -> X of 18 indicates that this is a strong rule than what would be expected by random chance.

Overally, in terms of usefulness, (Y AND W) -> X is more useful as it is a strong relationship that happens frequently in the dataset. Therefore, utilizing this rule means that you can be very certain that if you had someone with qualities Y AND X, then they would are most defenitely born before 1976. Meanwhile, X -> Y is somewhat useful, as its slightly positive metrics in interest, lift, and conviction make it slightly useful, but not as strong and reliable as (Y AND W) -> X. Finally, W->Z is the least useful, however W->Z is surprising as its negative correlation of owning a classical album actually makes you less likely of owning a hip-hop album, contrary to what we believe.