Apriori and FP-Growth

Amit Roy #1, Shekhor Chandra Saha #2, Fariha Ibnat #3, Sharika Tasnim Arshi#4

Computer Science and Engineering, East West University,  
Jahurul Islam City Gate, A/2 Jahurul Islam Ave, Dhaka 1212, Bangladesh

[12017-3-60-021@std.ewubd.edu](mailto:12017-3-60-021@std.ewubd.edu)

[22017-3-60-025@std.ewubd.edu](mailto:22017-3-60-025@std.ewubd.edu)

[32017-2-60-149@std.ewubd.edu](mailto:32017-3-60-025@std.ewubd.edu)

[42017-3-60-084@std.ewubd.edu](mailto:42017-3-60-084@std.ewubd.edu)

1. **Dataset**

**File Type:**

Here, Chess and Mushroom dataset are in dat file. Dat files store only specific information which are related to the program. More information is stored in a dat file then other file.

**Chess Dataset:**

Chess dataset is in dat file. There are 3195 rows and 37 columns. The total overview of chess dataset is given below.

Graphical user interface

Description automatically generated with low confidenceFrom the dataset we observe that there are 37 numbers of variable, 3195 numbers of observations, there are no missing cells and no duplicates rows. Total 923.7 KiB memory in size.

Few Column Explanations:

1st Column have two types of values. Those are 1 and 2. 1 is total 1668 times and 2 is total 1527 times. Frequency of 1 is 52.2% and 2 is 47.8%. These variables have high correlations with some fields. Those are 21,44 and 68.

Graphical user interface, application

Description automatically generated

19th columns have two types of values. Those are 19 and 20. 19 is total 1979 times and 20 is total 1216 times. Frequency of 19 is 61.9% and 20 is 38.1%. These variables have high correlations with some fields. Those are 15,17 and 46.

Graphical user interface, application

Description automatically generated

42th columns have two types of values. Those are 42 and 43. 42 is total 2713 times and 43 is total 482 times. Frequency of 42 is 84.9% and 43 is 15.1%. These variables have high correlations with some fields. Those are 23, 54, 64 and 74.

Graphical user interface, application

Description automatically generated

60th columns have two types of values. Those are 60 and 61. 60 is total 3148 times and 61 is total 47 times. Frequency of 60 is 98.5% and 43 is 1.5%. These variables have high correlations with some fields. Those are 66 and 72.

Graphical user interface, application

Description automatically generated

74th columns have two types of values. Those are 74 and 75. 74 is total 2406 times and 75 is total 789 times. Frequency of 74 is 75.3% and 75 is 24.7%. These variables have high correlations with some fields. Those are 23, 42, 54 and 64.

Graphical user interface, application

Description automatically generated

**Correlations:**

Pearson’s r:

Chart, scatter chart

Description automatically generated

The Pearson's correlation coefficient (r) is a measure of linear correlation between two variables. Its value lies between -1 and +1, -1 indicating total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation. Furthermore, r is invariant under separate changes in location and scale of the two variables, implying that for a linear function the angle to the x-axis does not affect r.

To calculate r for two variables X and Y, one divides the covariance of X and Y by the product of their standard deviations.

Spearman’s ρ:

Chart, scatter chart

Description automatically generated

The Spearman's rank correlation coefficient (ρ) is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's r. Its value lies between -1 and +1, -1 indicating total negative monotonic correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation.

To calculate ρ for two variables X and Y, one divides the covariance of the rank variables of X and Y by the product of their standard deviations.

Kendall’s T:

Chart, scatter chart

Description automatically generated

Similarly, to Spearman's rank correlation coefficient, the Kendall rank correlation coefficient (τ) measures ordinal association between two variables. Its value lies between -1 and +1, -1 indicating total negative correlation, 0 indicating no correlation and 1 indicating total positive correlation.

To calculate τ for two variables X and Y, one determines the number of concordant and discordant pairs of observations. τ is given by the number of concordant pairs minus the discordant pairs divided by the total number of pairs.

Phik (φ k):

Chart, scatter chart

Description automatically generated

Phik (φk) is a new and practical correlation coefficient that works consistently between categorical, ordinal and interval variables, captures non-linear dependency and reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution.

Cramer’s V (φ c):

Chart, scatter chart

Description automatically generated

Cramér's V is an association measure for nominal random variables. The coefficient ranges from 0 to 1, with 0 indicating independence and 1 indicating perfect association. The empirical estimators used for Cramér's V have been proved to be biased, even for large samples. We use a bias-corrected measure that has been proposed by Bergsma in 2013.

Missing values (count):

Chart

Description automatically generated

The above bar chart represents count of values per column ignoring missing values from the chess dataset. Each bar represents columns of the dataset. The height of the bar represents about the completeness of the column on the left, y-axis ranges from 0 to1. If the bar is less than 1.0 then we can say that there are missing values. On the right, index values are measured. In the above figure we can conclude that, this dataset has no missing values. Here bottom column represents sample of a column length and top column represents highest column length for that for that column.

Matrix:

Chart, bar chart

Description automatically generated

In the above figure we can see the plotting of matrix representing no missing values of chess dataset. Matrix plot is a useful tool that provides a color fill for each column. When values are missing it will be shaded in white. As we can see from the figure that there are no missing values so every column is shaded into blue.

**Mushroom Dataset:**

Mushroom dataset is in dat file. There are 8123 rows and 23 columns. The total overview of chess dataset is given below. From the dataset we observe that there are 23 numbers of variable, 8123 numbers of observations, there are no missing cells and no duplicates rows. Total 1.4MB memory in size. There are two types of carriable. One is categorical and another is Numerical. Categorical value is 14 and Numerical value is 9.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Few Column Explanations:

1st Column have two types of values. Those are 1 and 2. 1 is total 3915 times and 2 is total 4208 times. Frequency of 1 is 48.2% and 2 is 51.8%.

Graphical user interface, application

Description automatically generated

52th Column have two types of values. Those are 52 and 53. 53 is total 4608 times and 52 is total 3515 times. Frequency of 53 is 56.7% and 52 is 43.3%.

Graphical user interface, application

Description automatically generated

54th Column have five types of values. Those are 54, 55, 56, 57 and 58. 54 is total 1119 times, 55 is total 556 times, 56 is total 3776 times, 57 is total 192 times and 58 is total 2480 times. Frequency of 54 is 13.8%, 55 is 6.8%,56 is 46.5%, 57 is 2.8% and 58 is 30.5%.

Graphical user interface, application

Description automatically generated

59th Column have four types of values. Those are 59, 60, 61 and 62. 59 is total 5175 times, 60 is total 552 times, 61 is total 2372 times and 62 is total 24 times. Frequency of 59 is 63.7%, 60 is 6.8%, 61 is 29.2% and 62 is 0.3%.

Graphical user interface, application

Description automatically generated

90th Column have three types of values. Those are 90, 91 and 92. 90 is total 7487 times, 91 is total 600 times and 92 is total 36 times. Frequency of 90 is 92.2%, 91 is 7.4% and 92 is 0.4%.

Graphical user interface, application

Description automatically generated

**Correlations:**

Pearson’s r:

Chart, treemap chart

Description automatically generated

The Pearson's correlation coefficient (r) is a measure of linear correlation between two variables. Its value lies between -1 and +1, -1 indicating total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation. Furthermore, r is invariant under separate changes in location and scale of the two variables, implying that for a linear function the angle to the x-axis does not affect r.

To calculate r for two variables X and Y, one divides the covariance of X and Y by the product of their standard deviations.

Spearman’s ρ:

Chart, timeline, treemap chart

Description automatically generated

The Spearman's rank correlation coefficient (ρ) is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's r. Its value lies between -1 and +1, -1 indicating total negative monotonic correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation.

To calculate ρ for two variables X and Y, one divides the covariance of the rank variables of X and Y by the product of their standard deviations.

Kendall’s T:

Chart, timeline

Description automatically generated

Similarly to Spearman's rank correlation coefficient, the Kendall rank correlation coefficient (τ) measures ordinal association between two variables. Its value lies between -1 and +1, -1 indicating total negative correlation, 0 indicating no correlation and 1 indicating total positive correlation. To calculate τ for two variables X and Y, one determines the number of concordant and discordant pairs of observations. τ is given by the number of concordant pairs minus the discordant pairs divided by the total number of pairs.

Phik (φ k):

A picture containing chart

Description automatically generated

Phik (φk) is a new and practical correlation coefficient that works consistently between categorical, ordinal and interval variables, captures non-linear dependency and reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution.

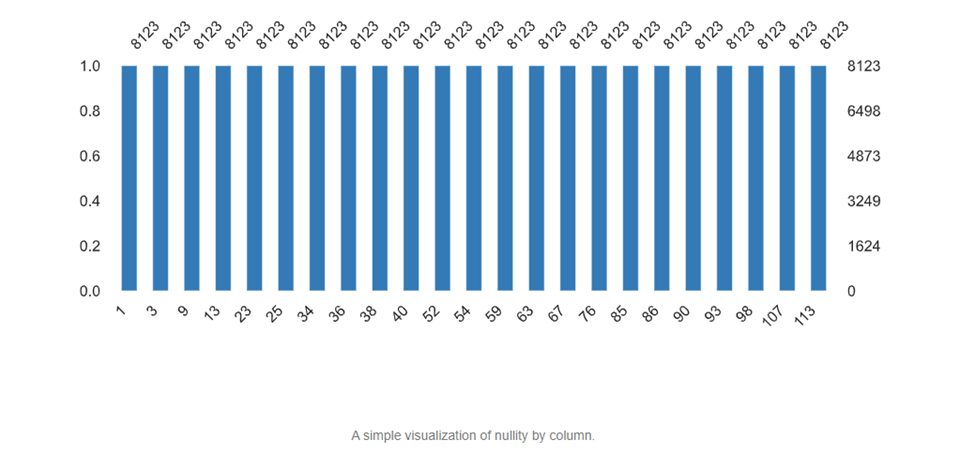
Cramer’s V (φ c):

A picture containing chart

Description automatically generated

Cramér's V is an association measure for nominal random variables. The coefficient ranges from 0 to 1, with 0 indicating independence and 1 indicating perfect association. The empirical estimators used for Cramér's V have been proved to be biased, even for large samples. We use a bias-corrected measure that has been proposed by Bergsma in 2013.

Missing values (counts):



The above bar chart represents count of values per column ignoring missing values from the mushroom dataset. Each bar represents columns of the dataset. The height of the bar represents about the completeness of the column on the left, y-axis ranges from 0 to1. If the bar is less than 1.0 then we can say that there are missing values. On the right, index values are measured. In the above figure we can conclude that, this dataset has no missing values. Here bottom column represents sample of a column length and top column represents highest column length for that for that column.

Matrix:

Chart

Description automatically generated

In the above figure we can see the plotting of matrix representing no missing values of mushroom dataset. Matrix plot is a useful tool that provides a color fill for each column. When values are missing it will be shaded in white. As we can see from the figure that there are no missing values so every column is shaded into blue.

1. **Result**

Advertisers,