3. Association Mining

3.1 Apriori Algorithm

In this part, we will investigate the associations in our data set using the Apriori Algorithm. In order to using the algorithm, first, we converted our data set into a format suitable for association mining by using the one-out-of-k encoding technique because there are some continual and categorical attributes in our data set. The result is as follows:

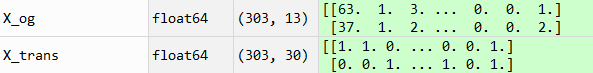


Fig. Comparing the size of the data set

From this figure, we can find that the size of the data set has been transformed from 30313 to 30330, and all the attributes are binary values. Then, we generate many transactions based on the observations and attributes name.

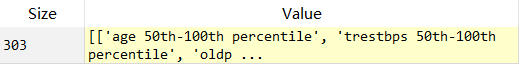


Fig. Transactions

Finally, we can apply the Apriori Algorithm on our data set. For easier analyzing, we set a threshold for support and confidence, the algorithm will find association rules with ≥20% and ≥80%. Because there are too many associations so we just show part of results below:

|  |  |  |
| --- | --- | --- |
| Association | Support | Confidence |
| {cp\_0, thal\_3} -> {target\_0} | 0.234 | 0.910 |
| {sex\_0, target\_1} -> {thal\_2} | 0.228 | 0.958 |
| {target\_0, trestbps 0th-50th percentile} -> {sex\_1} | 0.211 | 0.901 |
| {thal\_2, oldpeak 0th-50th percentile, age 0th-50th percentile} -> {target\_1} | 0.221 | 0.918 |
| {thal\_2, age 0th-50th percentile, slope\_2} -> {target\_1} | 0.201 | 0.924 |

3.2 Analysis

The reason why we set support ≥20% is we set the maximal k is 4 that when we applied the one-out-of-k encoding techniques on part of the attributes, so we cannot have support ≥25%, that’s one of the limitations of Apriror algorithm which mentioned in the textbook [1].

Our data set is for predicting the heart disease, so all attributes except target are the health conditions of a person, i.e. the values of an observation. The ‘sex\_0’ means male, ‘sex\_1’ means female, the ‘oldpeak’ means ST depression induced by exercise relative to rest, the ‘trestbps means resting blood pressure, the ‘slope’ means slope of the peak exercise ST segment, the ‘cp’ means chest pain type and the ‘thalach’ means maximum heart rate achieved.

So, according to the results table, we can say that people who have a chest pain type 3 and thal type 3 are more likely don’t have a heart disease; people who are male and have heart disease are more likely have a thal type 2;people who don’t have heart disease and the resting blood pressure lower than 130 are more likely a female; people who have a thal type 2, the ST depression induced by exercise relative to rest lower than 0.8 and the age lower than 55 are more likely to have heart disease; people who have a thal type 2, the age lower than 55 and the slope of the peak exercise ST segment is 2 are more likely to have a heart disease.