NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science

PROJECT REPORT
on the course of Ordered Sets in Data Analysis

FCA toolbox

Student:
Angelina Parfenova

Moscow, 2020

I decided to use the data on Heart disease UCI as the first dataset. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The "goal" field refers to the presence of heart disease in the patient.

Attribute information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

First of all, the data was preprocessed. Scale features were converted to categorical by taking several intervals. Categorical features then were transformed into binary. Then data was split for creating the plus and minus contexts.

FCA algorithms:

The first implementation. It is a simple implementation of FCA logic based on plus and minus contexts. For any object, if the number of intersections with our object with plus context is more than with minus one, then the object is classified as positive, otherwise as negative.

$$h_+(g) = \sum_{g+\in G+} g' \cap g_+$$

$$h_-(g) = \sum_{g-\in G-} g' \cap g_-$$

G – the set of all objects

G_+ - set of plus context (where the target variable is 1)

G_- - set of minus context (where the target variable is 0)

Algorithm:

1) For each object from plus context, look for similar descriptions of (g_+) and descriptions of (g'), and check whether there are also similar descriptions of this object with minus examples (g_-). If for (g_+) in G_+ we find more similar descriptions in G_- we assign the answer to false positive.

2) For each object from minus context, look for similar descriptions of (g_-) and descriptions of (g'), and check whether there are also similar descriptions of this object with plus examples (g_+). If for (g_-) in G_- we find more similar descriptions in G_+ we assign the answer to false negative.

The second implementation. Another implementation of FCA follows the logic of dividing the number of intersections with one or another context by the number of all elements in the context. Thus, we can evade the problem of differences of contexts' sizes and get more correct results.

$$h_+(g) = \frac{\sum_{g+\in G+} g' \cap g_+}{G_+}$$

$$h_-(g) = \frac{\sum_{g-\in G-} g' \cap g_-}{G_-}$$

The third implementation. Next, let's make the algorithm stricter introducing the rule for determining the positive or negative class. If the proportion of plus intersections over minus intersections exceeds some value/proportion (P), then it will be classified as positive. Now being just less or more is not enough. The same logic will be applied to the minus context with the value (M):

$$h_+(g) \,/\, h_-(g) > P$$

$$h_-(g) \,/\, h_+(g) > M$$

Three state of art algorithms were used: Random Forest, KNN and Logistic regression. The comparison by main metrics is below. An additional metrics that combines precision and recall was used based on their harmonic mean:

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Table 1

Heart UCI

	FCA_1	**FCA_2**	**Random Forest**	**KNN**	**Logistic Regression**	**FCA_3 (1.1, 1.1)**	**FCA_3 (1.25, 1.25)**	**FCA_3 (1.25, 1.1)**	**FCA_3 (0.8, 1)**
True positive	157	143	37	36	25	129	106	102	158
True Negative	83	113	25	25	12	98	76	97	113

False Positive	6	22	7	11	9	35	58	61	6
False Negative	54	22	6	3	12	38	60	40	23
True Positive Rate	0.744	0.86	0.86	0.92	0.67	0.77	0.64	0.72	0.87
True Negative Rate	0.93	0.84	0.78	0.69	0.57	0.74	0.57	0.61	0.95
Negative Predictive Value	0.6	0.84	0.81	0.89	0.5	0.72	0.56	0.70	0.83
False Positive Rate	0.06	0.16	0.22	0.30	0.43	0.26	0.43	0.38	0.05
False Discovery Rate	0.036	0.13	0.16	0.23	0.26	0.21	0.35	0.37	0.04
Accuracy	0.8	0.85	0.83	0.81	0.64	0.76	0.60	0.66	0.90
Precision	0.96	0.86	0.84	0.76	0.74	0.79	0.64	0.63	0.96
Recall	0.74	0.86	0.86	0.92	0.67	0.77	0.64	0.72	0.87
F	0,84	0,86	0,85	0,83	0,70	0,78	0,64	0,67	0,91

Then the data from my own research[1] was used:

Features are:

1. Believe in covid-19
2. Frequency of meetings with friends
3. Frequency of hands washing
4. Believe that covid-19 will affect the person
5. Probability that friend will be offended by lack of physical contact
6. Frequency of making an official pass
7. Frequency if joking about COVID-19
8. Frequency of going to public places

Target is the frequency of going out, I recoded this variable into a dichotomic one, 1-5 frequency was transformed into 0 and 6-10 - into 1.

Table 2

COVID-19 behavioral patterns

	FCA_1	**FCA_2**	**Random Forest**	**KNN**	**Logistic Regression**	**FCA_3 (0.7, 1)**
True positive	226	136	51	47	48	288
True Negative	130	218	45	48	27	221
False Positive	62	152	23	22	19	0
False Negative	122	34	16	18	27	31

Parfenova, A. (2020), "Will you shake my hand? Factors of noncompliance with COVID-19 behavioral rules in the framework of enforced social isolation in Russia", International Journal of Sociology and Social Policy, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/IJSSP-06-2020-0246

True Positive Rate	0.65	0.8	0.76	0.72	0.64	0.9
True Negative Rate	0.68	0.59	0.66	0.68	0.58	1
Negative Predictive Value	0.52	0.86	0.74	0.72	0.5	0.88
False Positive Rate	0.32	0.41	0.34	0.31	0.41	0.0
False Discovery Rate	0.22	0.53	0.31	0.32	0.28	0.0
Accuracy	0.66	0.65	0.71	0.70	0.62	0.94
Precision	0.78	0.47	0.69	0.68	0.71	1.0
Recall	0.65	0.8	0.76	0.72	0.64	0.90
F	0,70	0,59	0,72	0,70	0,67	0,94

Conclusion

Analyzing the results of metrics for the first dataset, we can see that the most effective algorithm is FCA_3 with coefficients 0,8 and 1. Nevertheless, we can’t say that this algorithm is good for predicting new data, as these coefficients mostly describe only the current data distribution. Among state of art algorithms, the most successful one is Random Forest.

As for the second dataset the results are similar. If we adjust the coefficients for FCA_3 making it more specific to the current data, then it performs very well with F-metrics = 0,94. Among state of art algorithms the most successful one is again Random Forest. However, its quality is much lower here. We can make a conclusion that to build better models we need to take into account the non-interval nature of variables in the dataset. Moreover, the initial targe variable was recoded from 1-10 scale to binary, so the results are expectedly lower. And classification task is probably not applicable for this data (better to build a linear regression for example).