Big Homework

Vatslav Sokolovskii

December 15, 2023

Contents

1 Project Part №1

In this paper will be introduced Lazy FCA classification algorithm based on pattern structures, binary Lazy FCA algorithm and popular models: xGboost, CatBoost, Logistic Regression and k-NN. For comparison I used three popular datasets:

- Water potability dataset
  `https://www.kaggle.com/datasets/adityakadiwal/water-potability`;
- Wine quality dataset
  `https://www.kaggle.com/datasets/subhajournal/wine-quality-data-combined`;
- Stroke Prediction dataset
  `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`.

You can find code for all stages of the project in my GitHub repository `https://github.com/RinokuS/HSE-OSDA-Project`. Code base structure:

1. Project_Task2.ipynb – notebook with performing classification on chosen datasets with xGboost, CatBoost, Logistic Regression and k-NN;
2. Project_Task3.ipynb – notebook with performing classification on chosen datasets with binary Lazy FCA;
3. Project_Task4.ipynb – notebook with performing classification on chosen datasets with Lazy FCA based on pattern structures;
4. datasets – folder with datasets.

All other files and folders are used to complete common Homeworks.

2 Project Part №2

First of all, as part of the second common Homework I performed classification on my datasets, using popular classification algorithms: xGboost, CatBoost, Logistic Regression and k-NN.

As for the result, we got following F1-scores table:

	xGboost	catboost	k-NN	LogisticRegression
Stroke Prediction Dataset	0.97	0.97	0.97	0.97
Water Quality	0.69	0.76	0.71	0.74
Wine Quality Data	1.00	0.71	0.94	0.59

And highlighted the least useful features for our three datasets (that were dropped on next two project parts, so we can save time on training FCALC model):

1. Stroke Prediction Dataset: id, work_type, Residence_type, smoking_status, gender
2. Water Quality: Conductivity, Organic_carbon, Trihalomethanes, Turbidity
3. Wine Quality Data: Unnamed: 0, citric acid, residual sugar, free sulfur dioxide, chlorides, total sulfur dioxide, pH

And for conclusion of this project part, we saw that the most stable model from the list of chosen are: xGboost and k-NN (And it was kind of unexpected, cause I thought that catboost will be the second one it the list of 'the best'. So maybe I used bad hyperparameters).

3 Project Part №3

Here were our first acquaintance with the algorithm of Lazy FCA. As the task of the part, we performed classification with the binary version of Lazy FCA.

And firstly we had to binarize all the data we got, and hat was hard enough cause almost all data from my three datasets were either categorical or numeric. So to binarize all the data I got (even after dropping all useless columns from the previous part of the Project) I had to search information about all possible thresholds on the Internet and this part was not so fun.

As for the result, we got following two tables (F1-score and accuracy):

Accuracy:

	xGboost	catboost	k-NN	LogisticRegression	FCALC
Stroke Prediction Dataset	0.95	0.95	0.95	0.95	0.62
Water Quality	0.64	0.66	0.58	0.59	0.39
Wine Quality Data	1.00	0.54	0.96	0.72	0.57

F1-Score:

	xGboost	catboost	k-NN	LogisticRegression	FCALC
Stroke Prediction Dataset	0.97	0.97	0.97	0.97	0.33
Water Quality	0.69	0.76	0.71	0.74	0.30
Wine Quality Data	1.00	0.71	0.94	0.59	0.57

As conclusion, binary FCALC was not good enough to beat popular algorithms, but still showed quite satisfactory results even after my rough data binarization. And we still can assume that data was just unsuitable for binarization and subsequent classification.

4 Project Part №4

For the 4th part of our Project (Big Homework) we performed classification on our datasets using Lazy FCA classification algorithm based on pattern structures. For this part we were able to get rid of data binarization, and all of the preprocessing was tied only to dropping least useful columns.

An unexpected difficulty of this stage was the model training time. For example, on the Wine Quality dataset training just one model took about 15 minutes on 15% of the dataset (And I got $3 * 3 = 9$ models with 2 folds cross validation each).

But despite all the difficulties, we got a good result. Lazy FCA classification algorithm based on pattern structures were much more stable, comparing to binary one, giving us much higher F1-score values. But on the other hand, unfortunately accuracy of the model was not high enough to amuse me. As for concrete results, we got these final tables:

Accuracy:

	xGboost	catboost	k-NN	LogisticRegression	FCALC	PS FCALC
Stroke Prediction Dataset	0.95	0.95	0.95	0.95	0.62	0.74
Water Quality	0.64	0.66	0.58	0.59	0.39	0.51
Wine Quality Data	1.00	0.54	0.96	0.72	0.57	0.62

F1-Score:

	xGboost	catboost	k-NN	LogisticRegression	FCALC	PS FCALC
Stroke Prediction Dataset	0.97	0.97	0.97	0.97	0.33	0.56
Water Quality	0.69	0.76	0.71	0.74	0.30	0.49
Wine Quality Data	1.00	0.71	0.94	0.59	0.57	0.62

And these graphics:

4.1 Water Potability

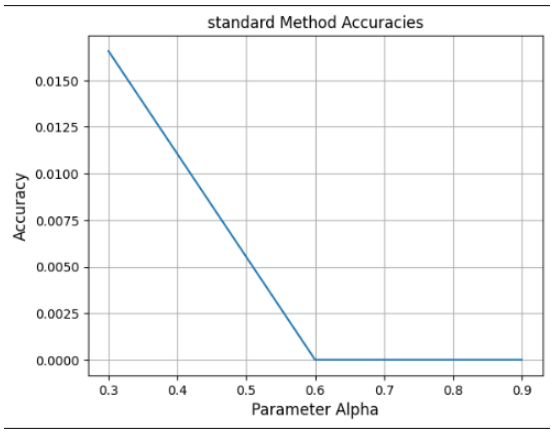

Figure 1: WP: Accuracies for different alpha on 'standard' method

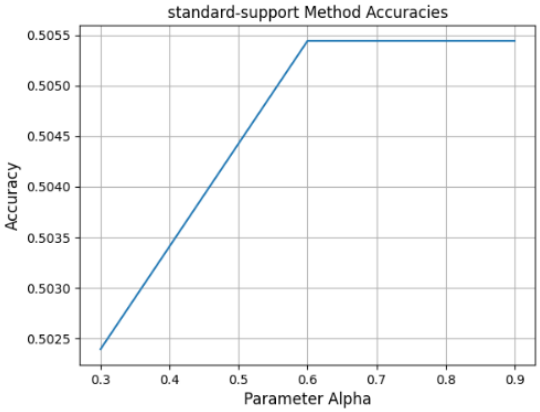

Figure 2: WP: Accuracies for different alpha on 'standard-support' method

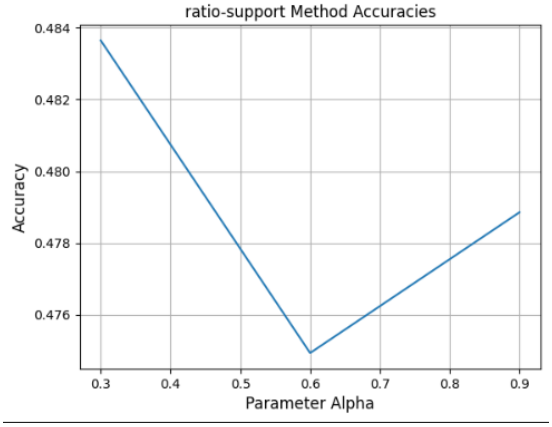

Figure 3: WP: Accuracies for different alpha on 'ratio-support' method

As we can see, 'standard-support' method was the best for this dataset.

4.2 Wine quality

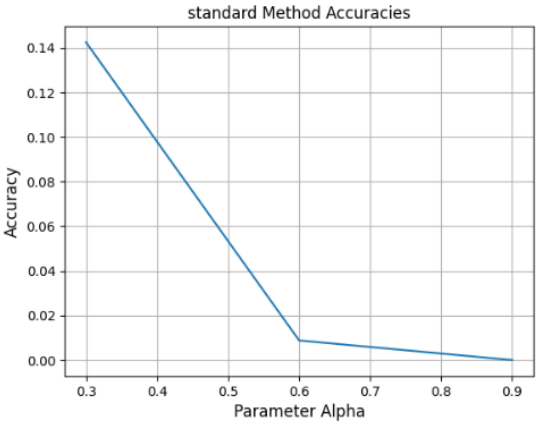

Figure 4: WQ: Accuracies for different alpha on 'standard' method

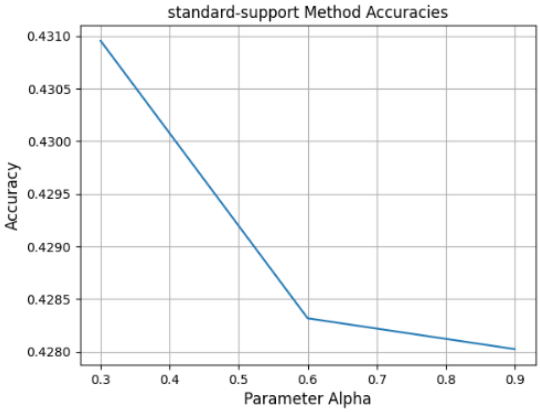

Figure 5: WQ: Accuracies for different alpha on 'standard-support' method

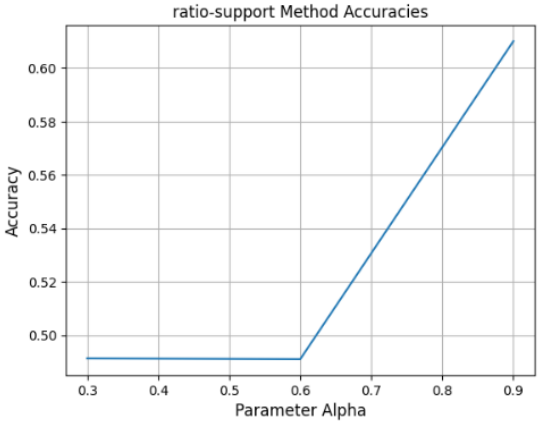

Figure 6: WQ: Accuracies for different alpha on 'ratio-support' method

As for this dataset, the 'ratio-support' method was the best one

4.3 Stroke prediction

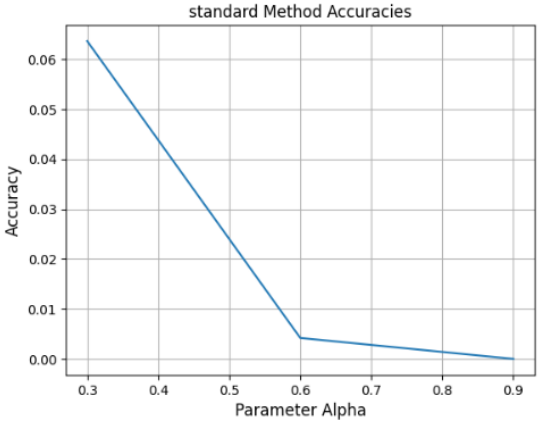

Figure 7: SP: Accuracies for different alpha on 'standard' method

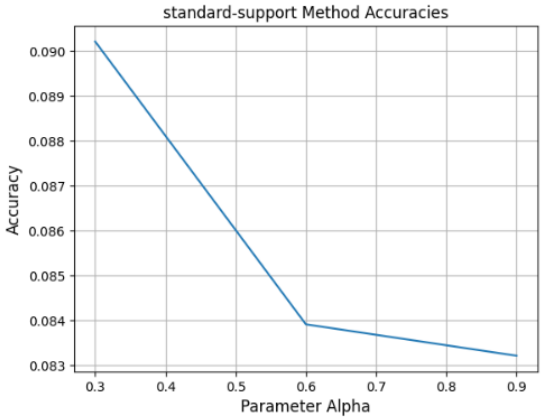

Figure 8: SP: Accuracies for different alpha on 'standard-support' method

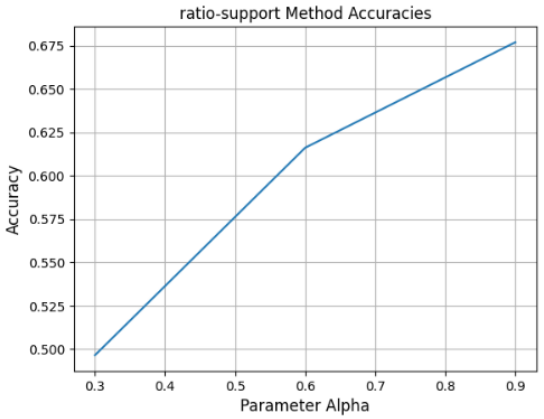

Figure 9: SP: Accuracies for different alpha on 'ratio-support' method

5 Conclusion

In conclusion, I can mention that algorithm based on pattern structures has greater results than binary one (especially for F1-score) and has comparable results with popular classification models. And as for the methods, we can assume that 'ratio-support' method is the best almost on every chosen dataset.

However there is some problems with performance and over fitting, so there are a huge field for research and testing different approaches.