OSDA BIG Home Assignment Report

Mikhail Shkliar

December 15, 2023

Contents

1 Project Description

This paper describes research on Lazy FCA classification algorithm. It includes not only binary Lazy FCA classification algorithm and Lazy FCA algorithm based on pattern structures, but also such well-known models as xGboost, Logistic Regression, KNN means and Catboost. To provide valuable results and compare the result of different models' work I have taken such datasets:

- Heart Attack Analysis and Prediction dataset
  (`https://www.kaggle.com/adityakadiwal/water-potability`)
- Quality of Wine dataset
  (`https://www.kaggle.com/subhajournal/wine-quality-data-combined`)
- HR Analytics: Job Change of Data Scientists dataset
  (`https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists`)

1.1 Project files structure:

1. All the datasets are located in the datasets_bigHA folder.
2. OSDA_BHA2.ipynb: In this notebook you can find the datasets classification with the help of such standard models as xGboost, Logistic Regression, KNN means and CatBoost. It will be used to compare the results of popular models with the results of Lazy FCA classifier.
3. OSDA_BHA3.ipynb: In this notebook you can find the datasets classification with the help of binary Lazy FCA classifier.
4. OSDA_BHA4.ipynb: In this notebook you can find the datasets calssification with the pattern structures classifier.
   We will take a particularly close look at the last 2 models to draw conclusions about Lazy FCA classifier.

2 OSDA_BHA2 Project

This project is related to Homework №2.

In this projects I've performed classification of the taken datasets with the help of standard Classification algorithms. Such as xGboost, Logistic Regression KNN means and CatBoost.
As a result I have got such Accuracy:

	xGboost	**Log. Regression**	**KNN**	**Catboost**
Heart Attacks Prediction	0.90	0.87	0.69	0.85
Quality of Wine	1.00	0.74	0.96	0.54
Job Change of D.Scientists	0.76	0.77	0.76	0.78

Also we've got the following macro average F1-scores:

	xGboost	**Log. Regression**	**KNN**	**Catboost**
Heart Attacks Prediction	0.90	0.86	0.68	0.84
Quality of Wine	1.00	0.71	0.96	0.77
Job Change of D.Scientists	0.66	0.61	0.60	0.69

Here we performed the base analyses of the datasets which helped us in following projects. We've got rid of all missing values and prepocessed data. For instance, we had a lot of numeric and ordinal data in the data scientists job change datasets. Therefore, we had to analyse this data, divide it and transform all of it to the appropriate format.

What is more, we can conclude that xGboost is by far the best stable model. The second place goes to Catboost, the third to KNN means and the last to Logistic Regression. However, the difference in last 3 models performans is not so great.
I assume, the results are expected enough.

3 OSDA_BHA3 Project

This project is related to Homework №3.

In this projects I've performed classification of the taken datasets with the help of binary Lazy FCA algorithms.

First of all we had to preprocess the data but is wasn't a stumbling block thanks to the project 2.
The second step was to reduce the time of training FCALC model. To tackle this task I've decided to drop the less necessary features from all the datasets. I used xGboost classifier to find out and visualize the importance of the features. As a result I have removed such of the colomns from the corresponding datasets:

1. Heart Attack Analysis and Prediction dataset: trtbps;
2. Quality of Wine dataset: Unnamed: 0, citric acid, residual sugar, free sulfur dioxide, pH, chlorides, total sulfur dioxide;
3. HR Analytics: Job Change of Data Scientists dataset: nothing - every feature importance rate was higher than 0.01.

The further step was to define fcalc performer function.

After that it was necessary to binarize all the data, otherwise it was impossible to use the model. Practically each and every colomn of my model was numeric or categorial.
Therefore, it took significant amount of time to search for tresholds. Unfortunately, I haven't found valuable information for HR Analytics: Job Change of Data Scientists dataset. So, I've decided to take the mean value as a treshold and after some analysis found out that the most successful strategy is to reduce the median value slightly (e.g., by 0.07 or 7%).
After that I noticed an interesting fact, that FCALC showed better results with the median value treshold reduced by 7% for each column than with the tresholds from the article)

The final step was to start fclal_performer and to draw conclusions.

As a result we have got such Accuracy in comparison with other models:

	xGboost	**Log. Regression**	**KNN**	**Catboost**	**FCALC**
Heart Attacks	0.90	0.87	0.69	0.85	0.73
Quality of Wine	1.00	0.74	0.96	0.54	0.59
DS-ts' Job Change	0.76	0.77	0.76	0.78	0.74

Also we've got the following macro average F1-scores:

	xGboost	**Log. Regression**	**KNN**	**Catboost**	**FCALC**
Heart Attacks	0.90	0.86	0.68	0.84	0.73
Quality of Wine	1.00	0.71	0.96	0.77	0.53
DS-ts' Job Change	0.66	0.61	0.60	0.69	0.65

Summing it all up, binary FCALC was not as good as the standart leading algorithms, however on Heart Attacks and Data Scientists job datasets the results are satisfying enough.
Maybe the poor results have been achieved on the wine dataset owing to the fact that we had to take just 15% of all the data to reduce the time of binary FCALC model studies. We can also suppose, not so good results are connected with the wrong tresholds binarization choice.

4 OSDA_BHA4 Project

Overview This project is related to Homework №4.

In this projects I've performed classification of the taken datasets with the help of pattern structures Lazy FCA classification algorithm.

Again first steps were related to preprocessing and feature selection. After project №3 this part was a great deal easier. However, this time feature importance analysis showed a bit different results.

The following step was to split the data into the train and test datasets. Surprisingly, the stumbling block here was the time of the performance. It was really difficult to choose such data size the algorithm to work an adequate amount of time. I had to take just 8% of the wine quality dataset and just 10% of the data scientists job change dataset.

The results of the FCALC model based on pattern structures performance compared to the results of binary FCALC model performance are shown in the tables below:

Accuracy of Pattern Structures FCALC algorithm in comparison with Binary FCALC one:

	Pattern Structures FCALC	**Binary FCALC**
Heart Attacks	0.73	0.85
Quality of Wine	0.59	0.42
Data Scientists' Job Change	0.74	0.74

Also we've got the following macro average F1-scores:

	Pattern Structures FCALC	**Binary FCALC**
Heart Attacks	0.73	0.85
Quality of Wine	0.53	0.33
Data Scientists' Job Change	0.65	0.71

4.1 Heart Attacks prediction analysis

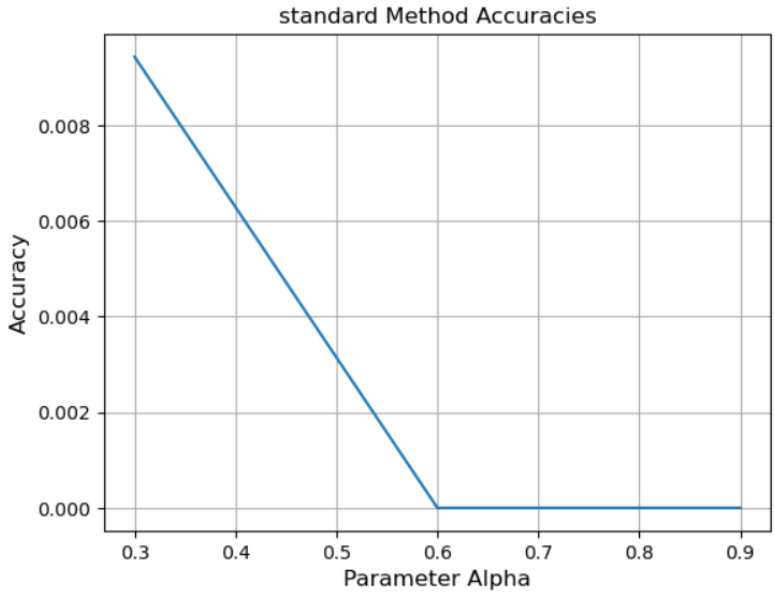

Figure 1: Accuracy for various alpha parameters on standard method

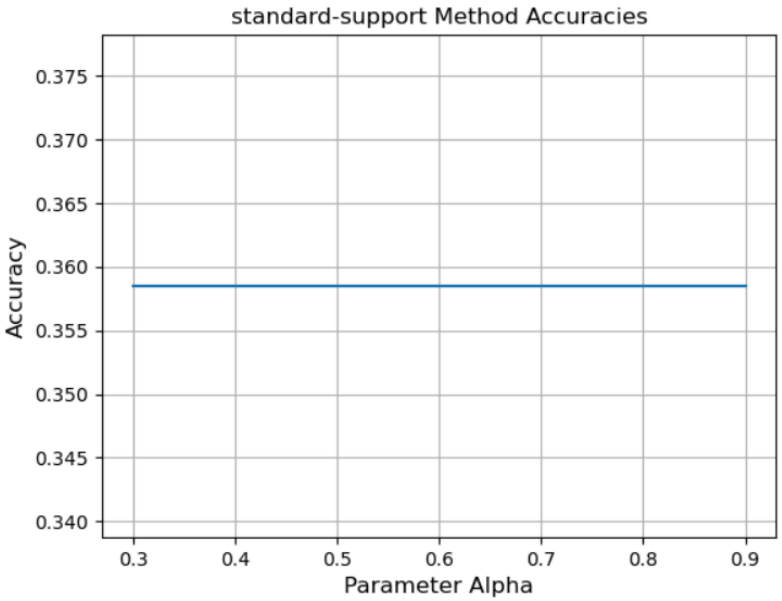

Figure 2: Accuracy for various alpha parameters on standard-support method

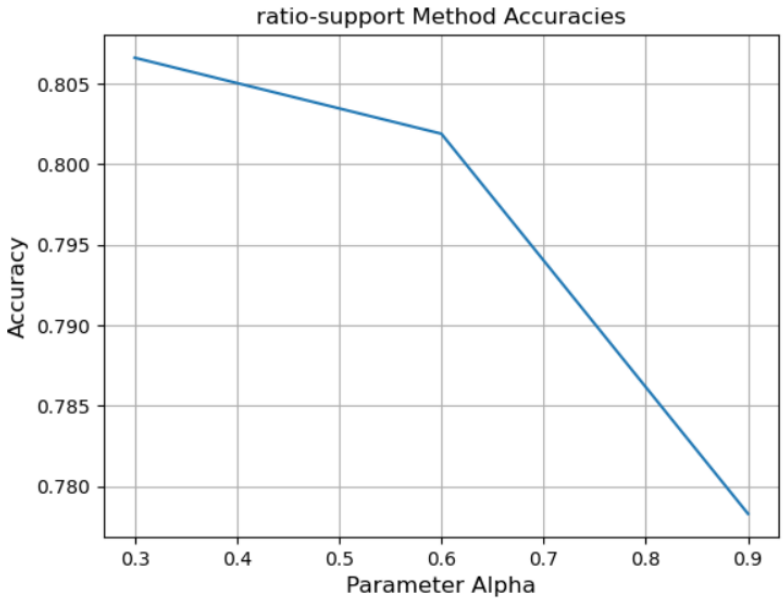

Figure 3: Accuracy for various alpha parameters on ratio-support method

We have achieved really good results. Our macro f1-score and accuracy are really good. Results even grater than some popular models have. Moreover, we can claim that this template based FCALC is much more stable on classification, as our F1 score is higher than during the binary classification analysis.

4.2 Quality of wine analysis

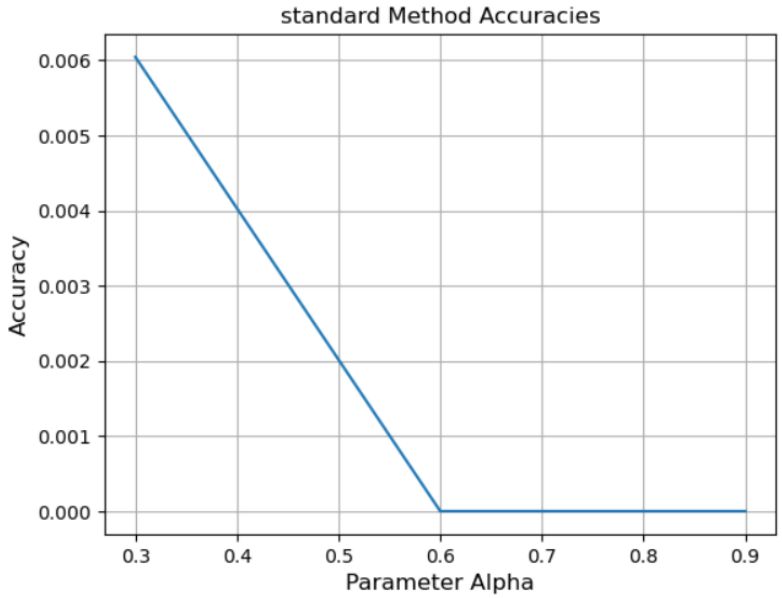

Figure 4: Accuracy for various alpha parameters on standard method

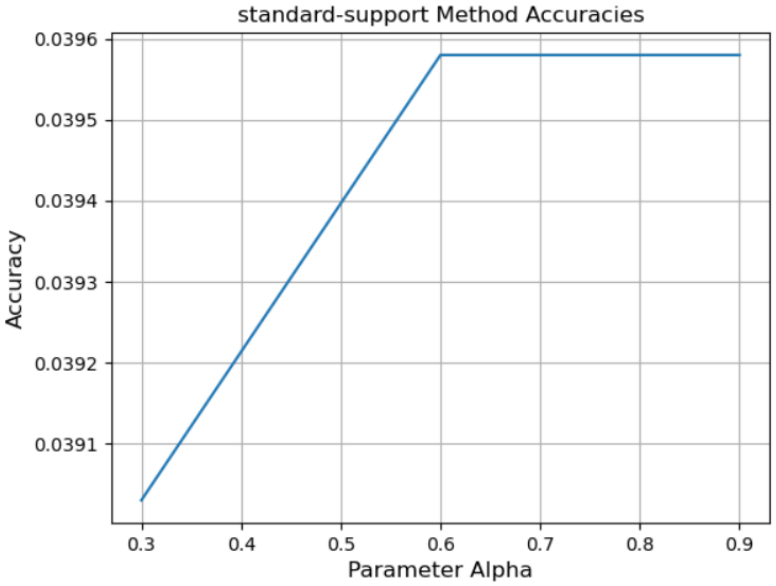

Figure 5: Accuracy for various alpha parameters on standard-support method

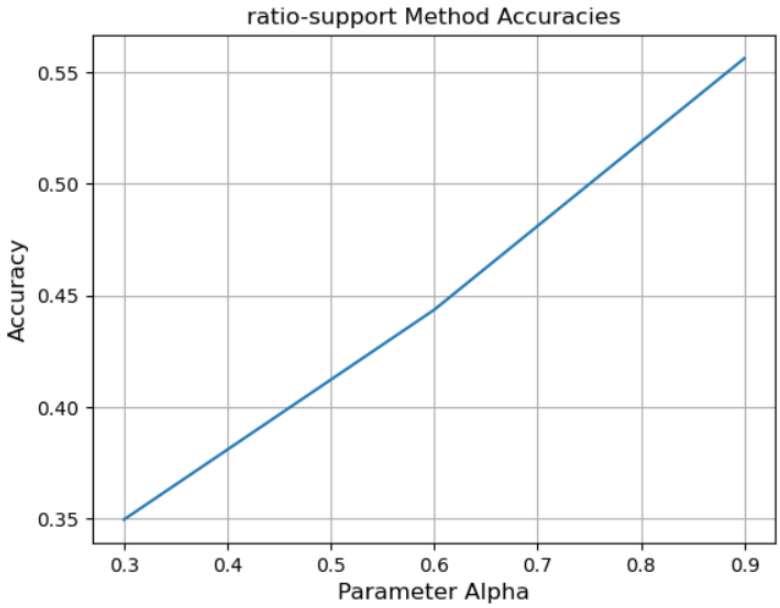

Figure 6: Accuracy for various alpha parameters on ratio-support method

We have achieved poor enough results. Both of our macro f1-score and accuracy aren't good. Maybe the quality of our results is due to the fact that the dataset was trimmed. We have taken just 8% of the data but even with such amount the process of studies took too long time. It was a bit frustrating to wait for more than 6 hours while the classification process is over and to get such unappealing results.

4.3 Data scientists job change data analysis

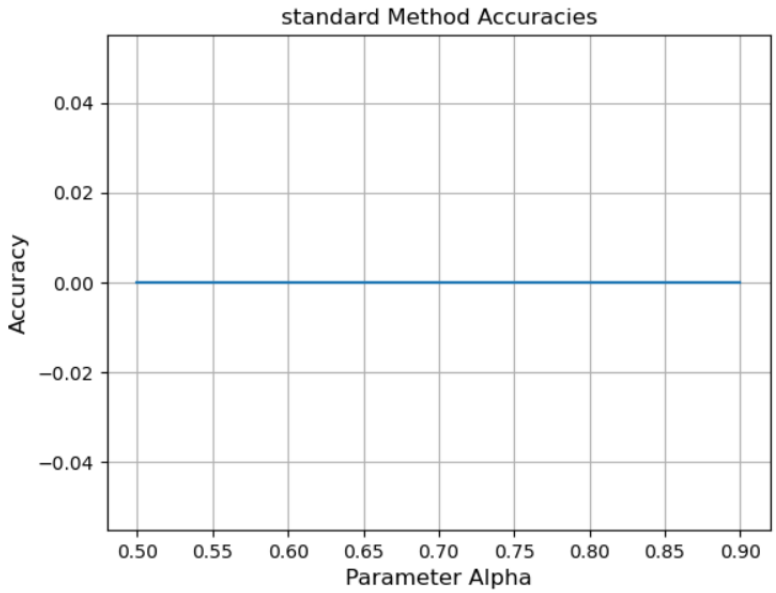

Figure 7: Accuracy for various alpha parameters on standard method

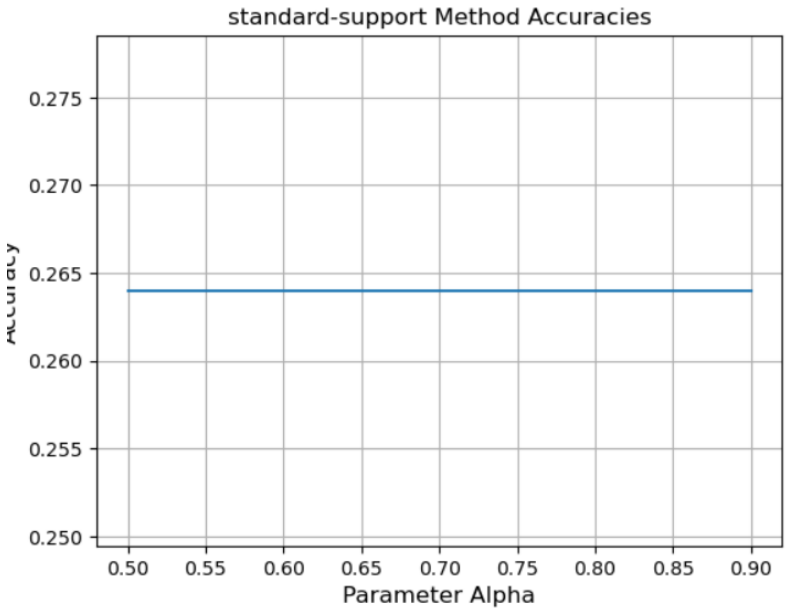

Figure 8: Accuracy for various alpha parameters on standard-support method

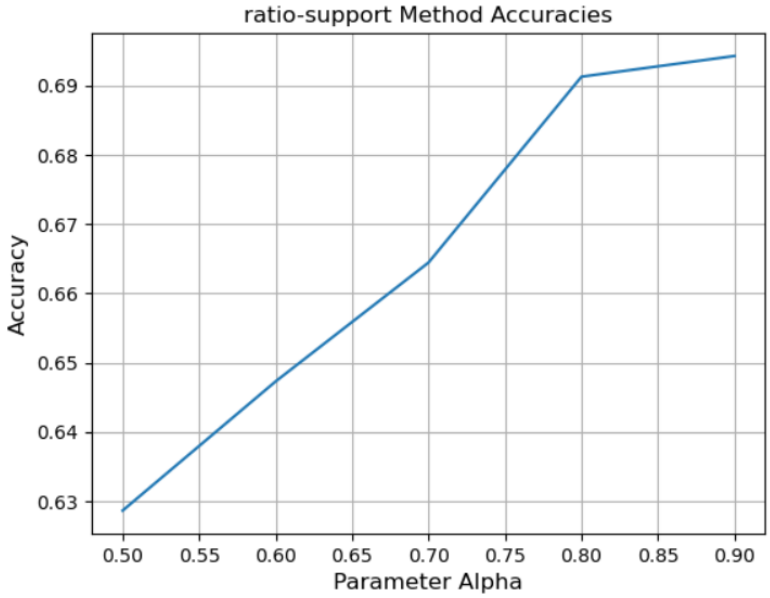

Figure 9: Accuracy for various alpha parameters on ratio-support method

We have achieved rather good results. Our macro f1-score and accuracy are note so bad) F1-score is higher than binary FCALC achieved, which means our new model is more stable. Maybe in case of taking at least 12% of the dataset as it was before, we could achieve even better results. It must be mentioned that it took 5 hours to get the results of the model performance.

5 Conclusion

Drawing a conclusion I should say that the Pattern Structures Lazy FCA algorithm generally shows better results than Binary Lazy FCA algorithm. In some cases it outperformes even popular models. However, practically in each case standard models like xGboost, Logistic Regression, KNN means and Catboost show better results. When it comes to the preferable methods, I would rather opt for ratio-support one as it shows the best result practically with all datasets ans alpha parameters. What is more, Pattern Structures Lazy FCA algorithm is generally more stable as it shows better F1 score than Binary one even having less train & test data. It must be mentioned that too much time for performance can become a stumbling block for the FCALC based on Pattern Structures. It is necessary to reduce your dataset to about 2000, otherwise the process of training the model will last for days. So, in case of a small dataset FCALC based on pattern structures can become a good choice for achieving good results.