Report on Big Homework
Ordered Sets in Data Analysis

Aleksandr Atlasov

December 2023

1 Introduction

That is my report for Big Homework for Neural FCA.
Since we have 3 datasets, I will not repeat same explanations for data preprocessing, experiments setups etc. They are described in corresponding sections. For each dataset there is a separate section, which contains results with comments, if they were necessary. Everything is implemented in ipynb from the same repository.

To-do list

- ☑ Choose 3 datasets with binary classification
- ☑ Improve binarization of data
- ☑ Choose quality metrics
- ☑ Select best features
- ☑ Select best concepts using different approaches
- ☑ Compare with classical machine learning models
- ☑ Try various non-linearities
- ☑ Vizualize best ConceptNetworks

Contents

2 Setup

2.1 Data Preprocessing

First of all, it is necessary to import each dataset and do some classic moves. Specifically to recode, if required, the value of the target variable, handle missing values and remove identifying data.
Now let us describe binarization process, which is the same for all datasets. While evaluating models for the first time I dropped numeric features and apply OneHotEncoding for categorical ones. Such approach is fast to implement, but in leads to significant loss of information. So, it was necessary to binarize numeric features as well.
For each numeric feature, we create 10 equal segments starting from the minimum of this feature and ending with its maximum. After that create columns that indicate whether an observation belongs to one of the obtained segments.
Finally, we set bool type for all features.

2.2 Quality metrics

As a main metric I will use F1-score, which is harmonic mean between recall and precision:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

But also to compare ConceptNetwork with other models I used Accuracy and AUC-ROC.

2.3 Feature selection method

After feature preprocessing we get a lot of binary columns, and obviously not all of them contain useful information for classification. That is why we need to select features somehow. As it was taught on Machine Learning classes, there are many different ways to do that. I chose filter method, which nicely suits for binary classification. It is necessary to calculate the following t-statistic for feature j:

$$t(j) = \frac{|\mu_+ - \mu_-|}{\sqrt{\frac{n_+\sigma_+^2 + n_-\sigma_-^2}{n_+ + n_-}}},$$

where μ, σ, n mean, standard deviation and number of objects in classes correspondingly. The higher the statistic's value, the more important the feature is. Basically, this value reflects the difference by-class for some attribute.

2.4 Experiments design

Step 1 Data binarization as it was described. Plots for class labels.

Step 2 Selecting best number of features and concepts using custom grid-search function.

Step 3 Comparing ConceptNetwork with finetuned number of features and concepts with classical ML model. I chose Gradient Boosting, kNN, Random Forest, Logistic Regression and Naive Bayes Classifier. Each ML model was also finedtuned using grid-search (with the same amount of features as for ConceptNetwork).

Step 4 Selecting best activation function using number of features and concepts, obtained before.

Step 5 Vizualization of the best ConceptNetwork.

3 Heart Disease Dataset

3.1 Description and main info

This database contains 14 attributes, such as patient's age, gender, cholestrol level, resting ECG levels etc. The "goal" field refers to the presence of heart disease in the patient. There are 303 observations. As we can see on **??**, classes are more or less balanced.

Figure 1: Class balance plot

3.2 Choosing best number of features and concepts

Hereinafter, a heat map will be presented for various combinations of the number of features and concepts depending on the obtained F1-score.

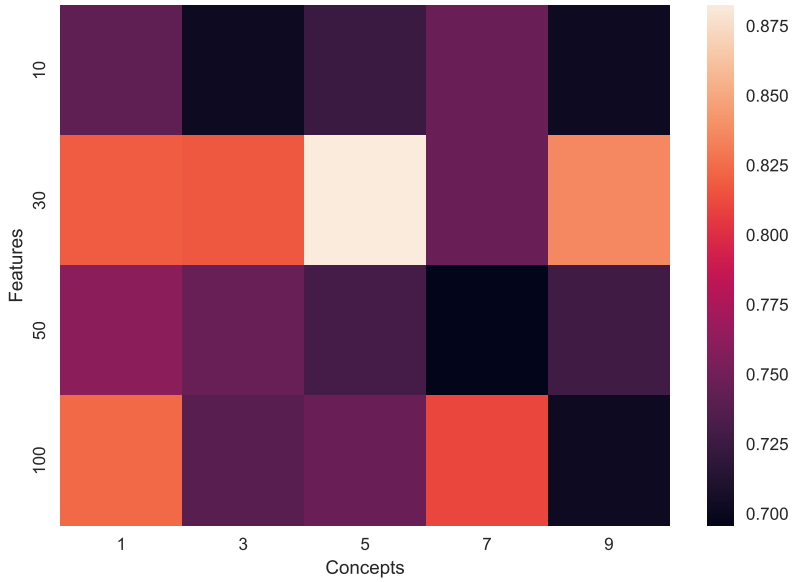


Figure 2: Heatmap for combinations of the number of features and concepts

It's easy to notice that the best combination is 30 features and 5 concept.

3.3 Comparison with classical machine learning models

Table 1: ConceptNetwork vs ML model

Model	Accuracy	F1-score	AUC-ROC
ConceptNetwork	0.836	0.844	0.838
CatBoost	0.803	0.812	0.805
NaiveBayes	0.639	0.738	0.613
LogisticRegression	0.82	0.831	0.82
RandomForest	0.77	0.788	0.769
kNN	0.77	0.794	0.766

As we can see ConceptNetwork superior all other models for all metrics. I guess it happens, because there are only 30 features, which is comfortable only for CN. Also there are not much observations.

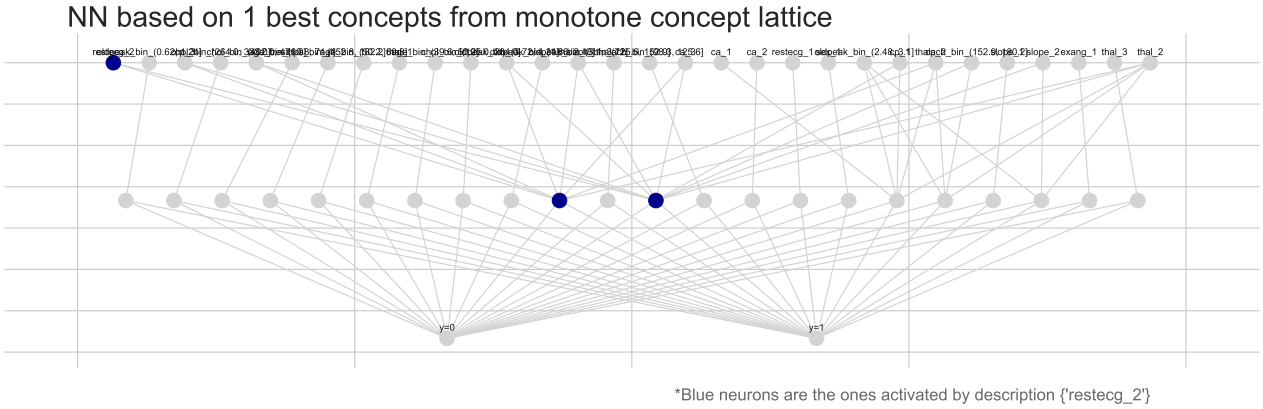
3.4 Selecting best non-linearity

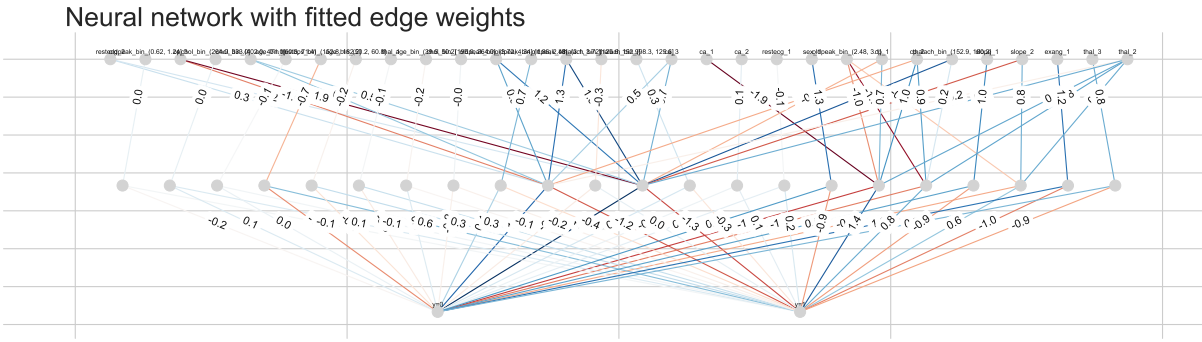
Table 2: Results for different activation functions

Model	Accuracy	F1-score	AUC-ROC
ReLU	0.869	0.879	0.868
LeakyReLU	0.836	0.853	0.832
GELU	0.754	0.789	0.746
ELU	0.82	0.836	0.817
Tanh	0.803	0.829	0.797

The best activation function is ReLU.

3.5 Vizualiztion of best ConceptNetwork





4 Airline Passenger Satisfaction

4.1 Description and main info

This dataset contains an airline passenger satisfaction survey. It includes factors that are highly correlated to a satisfied (or dissatisfied) passenger. The main aim is to predict passenger satisfaction. There are 103594 observations.

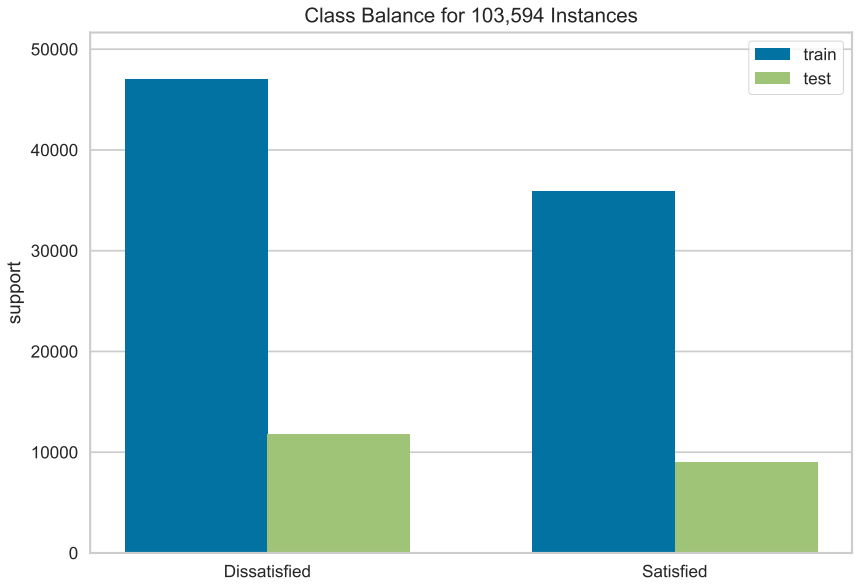


Figure 3: Class balance plot

4.2 Choosing best number of features and concepts

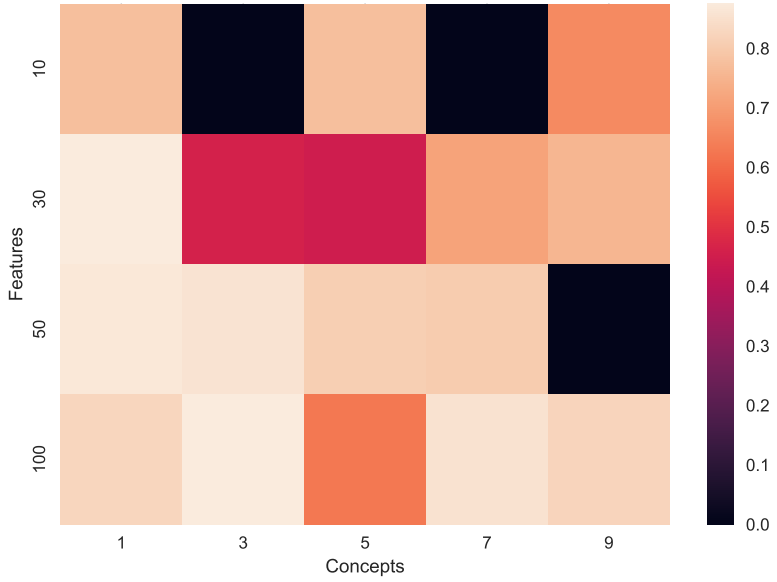


Figure 4: Heatmap for combinations of the number of features and concepts

Best number of features is again 30, and number of concepts is 1.

4.3 Comparison with classical machine learning models

Table 3: ConceptNetwork vs ML model

Model	Accuracy	F1-score	AUC-ROC
ConceptNetwork	0.861	0.83	0.852
CatBoost	0.939	0.928	0.936
NaiveBayes	0.875	0.846	0.865
LogisticRegression	0.913	0.898	0.91
RandomForest	0.912	0.893	0.904
kNN	0.925	0.912	0.922

With large amount of data CN starts to lose its superiority, but still it show results not much worse.

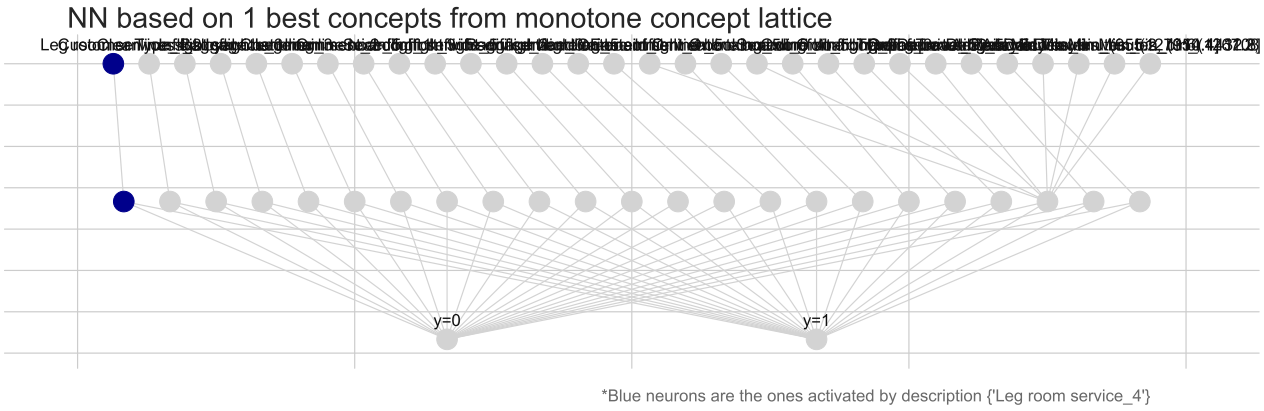
4.4 Selecting best non-linearity

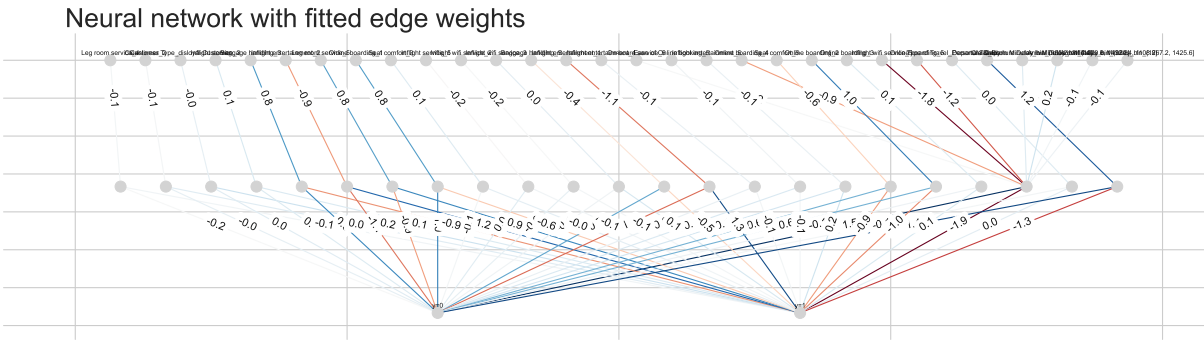
Table 4: Results for different activation functions

Model	Accuracy	F1-score	AUC-ROC
ReLU	0.906	0.886	0.899
LeakyReLU	0.837	0.807	0.831
GELU	0.882	0.859	0.876
ELU	0.881	0.859	0.876
Tanh	0.838	0.803	0.829

The best is obviuosly ReLU.

4.5 Vizualiztion of best ConceptNetwork





5 HR Analytics: Job Change of Data Scientists

5.1 Description and main info

This dataset designed to understand the factors that lead a person to leave current job for HR researches too. It is necessary to predict the probability of a candidate to look for a new job or will work for the company. There are 8955 observations. And as we can see classes are imbalanced.

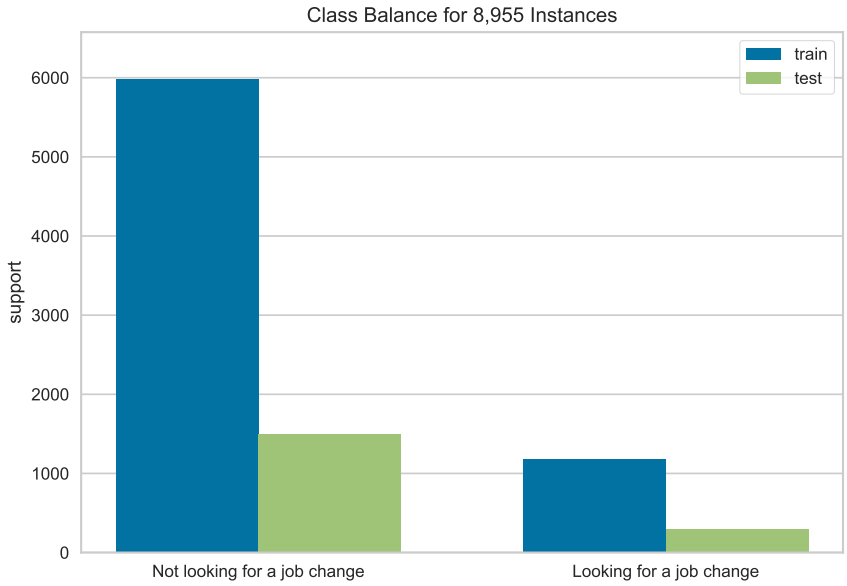


Figure 5: Class balance plot

5.2 Choosing best number of features and concepts

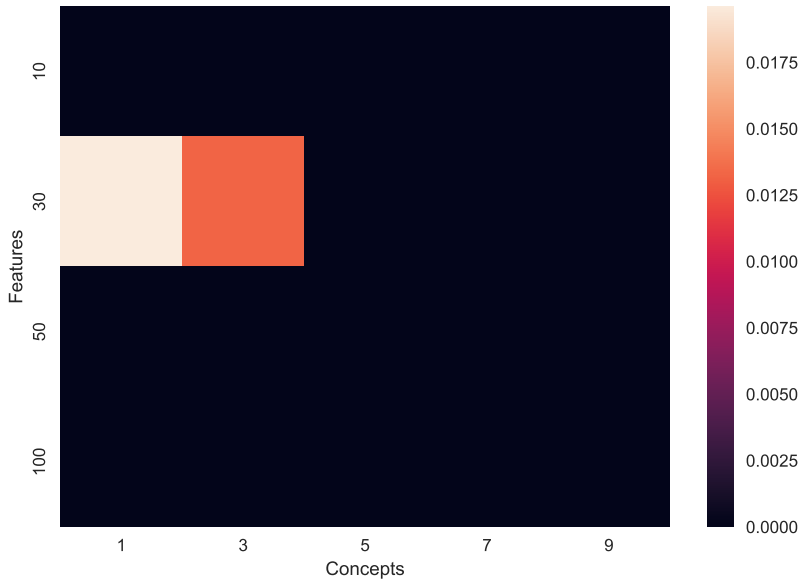


Figure 6: Heatmap for combinations of the number of features and concepts

Best number of features is another time 30, and number of concepts is 1. Actually, for all combinations F1-score is approximately 0. It means that CN is bad for imbalanced datsets.

5.3 Comparison with classical machine learning models

Table 5: ConceptNetwork vs ML model

Model	Accuracy	F1-score	AUC-ROC
ConceptNetwork	0.834	0.0	0.5
CatBoost	0.859	0.529	0.706
NaiveBayes	0.839	0.368	0.616
LogisticRegression	0.864	0.567	0.734
RandomForest	0.855	0.487	0.679
kNN	0.846	0.433	0.649

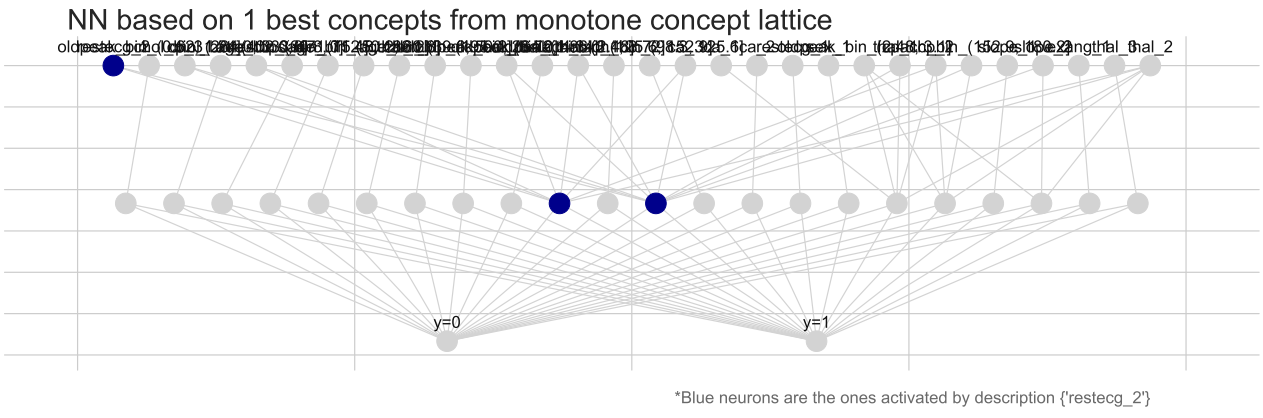
5.4 Selecting best non-linearity

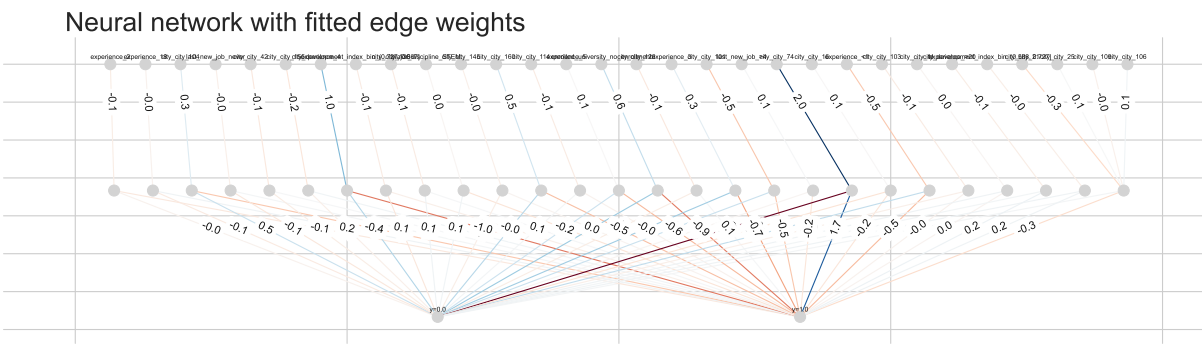
Table 6: Results for different activation functions

Model	Accuracy	F1-score	AUC-ROC
ReLU	0.834	0.0	0.5
LeakyReLU	0.834	0.0	0.5
GELU	0.832	0.02	0.503
ELU	0.834	0.0	0.5
Tanh	0.834	0.0	0.5

The best is GELU, because it gives at least non-zero results.

5.5 Vizualiztion of best ConceptNetwork





6 Conclusion

1. ConceptNetwork is good for small number of observations.
2. It does not work with imbalanced data
3. Does not require tricky non-linearity functions (ReLU is basically enough)