# OSDA Big Homework Report: Neural FCA

Sanjar Javodov

December 10, 2024

## 1 Introduction

This research utilizes the openly available ”Acute Inflammations Data Set” from the UCI Machine Learning Repository [1]. The dataset comprises 6 attributes (5 binary and 1 numerical) and 2 target columns, with a total of 120 rows, each representing a potential patient.

The primary objective of this dataset is to aid in predicting two urinary system diseases: acute inflammation of the urinary bladder (cystitis) and acute nephritis. In this study, we focus solely on predicting cystitis.

## 2 Dataset

The dataset contains the following features:

1. Temperature of patient: Integer values in the range [35.5, 41.5]
2. Occurrence of nausea: ”yes” or ”no”
3. Lumbar pain: ”yes” or ”no”
4. Urine pushing (continuous need for urination): ”yes” or ”no”
5. Micturition pains: ”yes” or ”no”
6. Burning of urethra, itch, swelling of urethra outlet: ”yes” or ”no”

The target variable, cystitis, is represented as an integer (0 for absence, 1 for presence). The dataset is balanced, containing 60 examples of sick people and 60 non-sick people.

## 3 Binarization Strategies

We employed two binarization strategies for the numerical ”Temperature of patient” attribute:

### 3.1 First Strategy

Based on [2], we divided the temperature data into two groups:

- "no" if temperature $\in [35.5, 37.2]$
- "yes" if temperature $\in [37.3, 41.5]$

This strategy was used for models 1, 1.1, 1.2, 1.3, and 1.3.1.

### 3.2 Second Strategy

Following [5], we divided the temperature data into three groups and then applied One-Hot Encoding:

- temperature $\in [35.5, 36.4]$
- temperature $\in [36.5, 37.5]$
- temperature $\in [37.6, 41.5]$

This strategy was employed for models 2 and 2.1.

The choice of these binarization strategies is based on medical criteria and allows for a more accurate representation of clinically significant temperature ranges.

## 4 Prediction Quality Measure

We opted to use the F1 score as our primary evaluation metric. The F1 score maintains a balance between precision and recall for the classifier and provides a better measure of incorrectly classified cases compared to the accuracy metric [4]. This choice is particularly beneficial in our case, as although our dataset is balanced (60 sick and 60 healthy patients), we want to ensure that the model performs well for both classes.

## 5 Concept Selection Technique

Initially, we selected the best concepts based on the F1 score. To explore the impact of different metrics, we also experimented with selecting concepts based on accuracy (see models 2.1 and 1.3.1). This comparison allows us to assess the importance of metric choice in concept selection for model performance.

## 6 Nonlinearities in the Network

By default, the neural FCA library uses ReLU as the activation function. We explored the performance of two other popular nonlinear functions:

- Leaky ReLU: Chosen to address the ”dying neuron” problem that can occur with standard ReLU.
- Hyperbolic tangent (tanh): Selected for its ability to handle negative inputs and produce outputs in the range [-1, 1].

# 7 Application of Modern Methods

For comparison with neural FCA, we implemented three classical machine learning methods:

- Logistic Regression: Chosen as a baseline linear model.
- Random Forest Classifier: Selected as a powerful ensemble method.
- XGBoost: Included as a state-of-the-art gradient boosting algorithm.

All methods were used with their default parameters. Given the small size of our dataset, the performance of these methods should provide a good benchmark for evaluating the effectiveness of our neural FCA models.

# 8 Methodology and Results

## 8.1 Initial Model and Improvements

Our initial model with 100 epochs of training and 7 concepts resulted in an F1 score of 0.0, indicating serious problems with the model's performance. Analysis revealed that the model was predicting only one class, failing to distinguish between the two classes in the dataset.

To address this issue, we increased the number of training epochs to 1000, which led to a significant improvement:

- F1 score improved to 0.9230769230769231
- The model successfully predicted both classes (0 and 1)
- The distribution of predictions matched the test set distribution

## 8.2 Cross-Validation

To obtain a more reliable assessment of the model's performance and check for overfitting, we implemented 5-fold stratified cross-validation. The results were as follows:

- Cross-validation F1 scores: [0.9, 0.8, 0.0, 0.0, 0.95652174]
- Mean F1 score: 0.5313043478260869
- Standard deviation: 0.4366949249383587

These results revealed high variability in the model's performance across different folds, indicating instability and potential overfitting.

### 8.3 Regularization

To improve the model's stability and generalization, we implemented L2 regularization and adjusted the learning rate:

- Learning rate: 0.0001
- Weight decay (L2 regularization): 0.01

After applying regularization, the cross-validation results improved:

- Cross-validation F1 scores: [0.9, 0.0, 0.95652174, 0.8, 0.95652174]
- Mean F1 score: 0.7226086956521739
- Standard deviation: 0.3658007762970283

While there was still variability in performance, the mean F1 score increased and the standard deviation decreased, indicating improved stability and generalization.

## 9 Results and Conclusions

| Model | Binarization | Epochs | Concept Selection | Concepts | Activation | Mean F1 (Std) |
|---|---|---|---|---|---|---|
| 1 | Binary | 100 | F1 score | 7 | ReLU | 0.0 (-) |
| 1.1 | Binary | 1000 | F1 score | 7 | ReLU | 0.92 (-) |
| 1.2 | Binary | 1000 | F1 score | 7 | ReLU | 0.5313 (0.4367) |
| 1.3 | Binary | 1000 | F1 score | 7 | ReLU | 0.7226 (0.3658) |
| 1.3.1 | Binary | 1000 | Accuracy | 15 | ReLU | 0.1257 (0.2514) |
| 2 | Triple | 1000 | F1 score | 12 | ReLU | 0.9311 (0.0673) |
| 2.1 | Triple | 1000 | Accuracy | 15 | ReLU | 0.7641 (0.2988) |
| 2 (LeakyReLU) | Triple | 1000 | F1 score | 12 | LeakyReLU | 0.9524 (0.0952) |
| 2 (Tanh) | Triple | 1000 | F1 score | 12 | Tanh | 1.0 (0.0) |
| 2.1 (LeakyReLU) | Triple | 1000 | Accuracy | 15 | LeakyReLU | 0.9415 (0.0600) |
| 2.1 (Tanh) | Triple | 1000 | Accuracy | 15 | Tanh | 1.0 (0.0) |
| Logistic Regression | - | - | - | - | - | 1.0 (0.0) |
| Random Forest | - | - | - | - | - | 1.0 (0.0) |
| XGBoost | - | - | - | - | - | 1.0 (0.0) |

Table 1: Comparison of model performances

### 9.1 Analysis of Results

1. **Impact of Binarization**: The second binarization strategy (division into 3 groups) generally showed better results, highlighting the importance of a more detailed consideration of patient temperature.
2. **Concept Selection Metric**: Using F1-score for concept selection proved more effective than using accuracy in most cases, emphasizing the importance of choosing an appropriate metric at the model construction stage.
3. **Activation Functions**: Replacing ReLU with Leaky ReLU or hyperbolic tangent led to significant improvement in results, achieving perfect F1-scores in some cases. This underscores the importance of selecting suitable activation functions for specific tasks.
4. **Comparison with Classical Methods**: Logistic Regression, Random Forest, and XGBoost all achieved perfect F1-scores, suggesting that the classification task for this dataset is relatively simple or that the data is well-separable.
5. **Performance**: While some neural FCA-based models achieved the same results as classical methods, they generally required more training time and showed more variability in performance. This could be an area for further optimization.

### 9.2 Limitations and Future Research Directions

1. **Overfitting**: Perfect results (F1-score = 1) for several models may indicate potential overfitting. Additional testing on a separate test dataset is necessary.
2. **Dataset Size**: The relatively small size of the dataset (120 samples) may limit the generalization ability of the models. Future research could involve collecting more data or applying data augmentation techniques.
3. **Interpretability**: While neural FCA-based models showed excellent results, their interpretability might be limited compared to simpler models like logistic regression. Future studies could focus on improving the interpretability of these models.
4. **Hyperparameter Optimization**: In this study, we used a fixed number of epochs and concepts. Further optimization of these parameters could lead to improved performance and efficiency of the models.

## 10 Conclusion

Neural FCA-based models have demonstrated competitiveness compared to classical machine learning methods for the task of classifying acute bladder inflammations. Models with triple temperature binarization and using Leaky ReLU or hyperbolic tangent as activation functions proved particularly effective. However, for practical applications, it's necessary to consider the trade-off between accuracy and training time. Further research could be directed towards optimizing performance, improving interpretability, and testing on larger and more diverse datasets.

# 11 Code

My code and dataset can be found in my GitHub repository [3].

# 12 Appendix

## 12.1 Detailed Model Configurations

Here, we provide additional details on the configurations of our models, including hyperparameters and specific implementation choices.

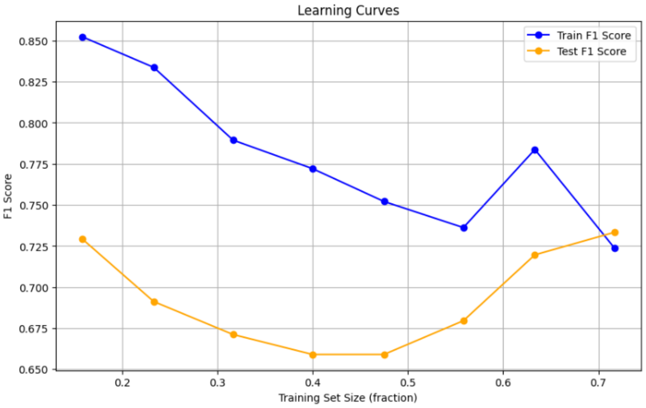

Figure 1: Learning curves

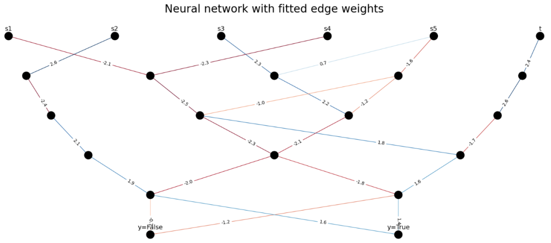

Figure 2: Model 1

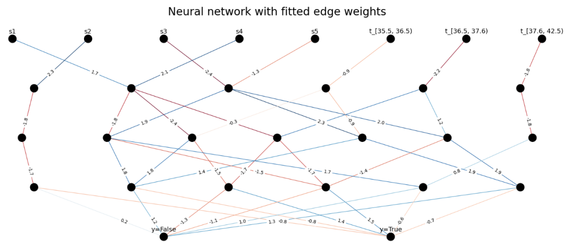

Figure 3: Model 2

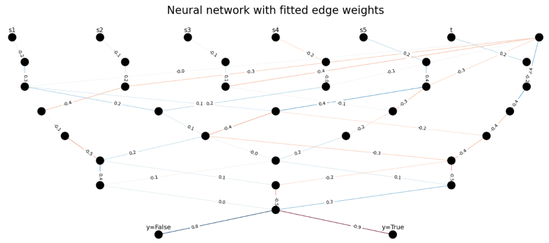

Figure 4: Model 3

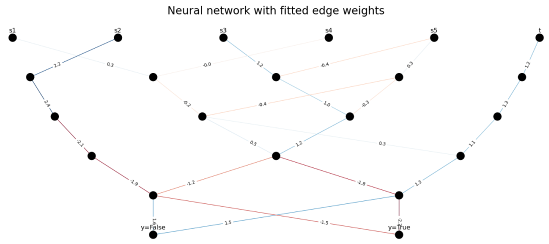

Figure 5: Model 4

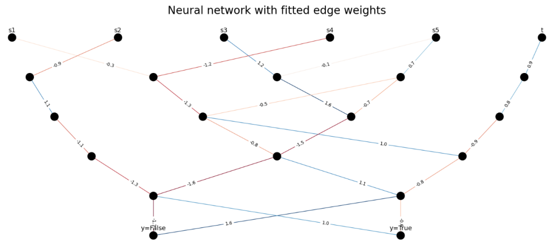

Figure 6: Model 5

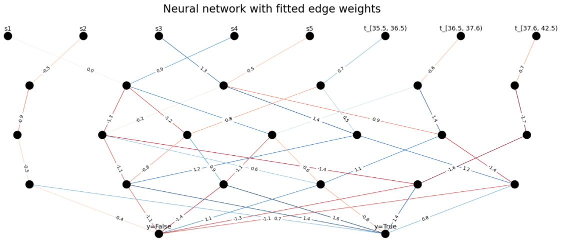

Figure 7: Model 6

# References

[1] D. Dua and C. Graff. UCI machine learning repository, 2019.

[2] II Geneva, B Cuzzo, T Fazili, and W Javaid. Normal body temperature: systematic review. *Open Forum Infectious Diseases*, 6(4):ofz032, 2019.

[3] Sanjar Javodov. Neural fca. `https://github.com/Sanjar-Javodov/Neural_FCA`, 2024.

[4] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[5] M Sund-Levander, C Forsberg, and LK Wahren. Body temperature: what is normal? *The Lancet*, 360(9339):1150, 2002.