Ministry of Science and Higher Education of the Russian Federation

National Research University Higher School of Economics

Faculty of Computer Science

School of Data Analysis and Artificial Intelligence

HOMEWORK REPORT

NEURAL FCA

Subject: *Ordered Sets for Data Analysis*

Student	*M.D. Kirdin*
Teaching Assistant	*A. Tomat*
Teaching Assistant	*M. Zueva*
Professor	*S.O. Kuznetsov*

Moscow, 2024

CONTENTS

INTRODUCTION

The overarching task of this homework is to implement a merger of machine learning and formal concept analysis. This is done by basing the neural network(NN) architecture on the covering relation (graph of the diagram) of a lattice coming from monotone Galois connections as proposed by Kuznetsov and his colleges [1]. The vertices of such NNs are related to sets of similar objects with similarity given by their common attributes, thus they are easily interpretable. The edges between vertices can be interpreted in terms of concept generality (bottom-up) or conditional probability (top-bottom).

You can find the source code for this homework at https://github.com/succSeeded/NeuralFCA.

1 MODEL EVALUATION

1.1 Problem statement

For this, a dataset has to be chosen, its data binarized using scaling (binarization) strategy of choice, and finally the target attribute defined. Then, a comparison between several standard classification methods and NN should be made by calculating performance metrics best suited for the dataset.

1.2 Model evaluation on the Employee Attrition dataset

The Employee Attrition dataset has 15K entries after dropping of all rows containing at least one empty value, hence it was decided to work only with a selection of 2400 randomly selected elements of this set.

The NNs based on 4(minimal concept amount that covers all the training data) and 20 best-performing formal concepts with all the numerical features binarized using ordinal encoding with 3 nodes will serve as our baseline. Here the performance of each concept was measured as f_1-score of its extent used as a prediction. This produces architectures that can be seen on fig. 1 and fig. 2. The results of baseline classification can be seen in the table A.1.

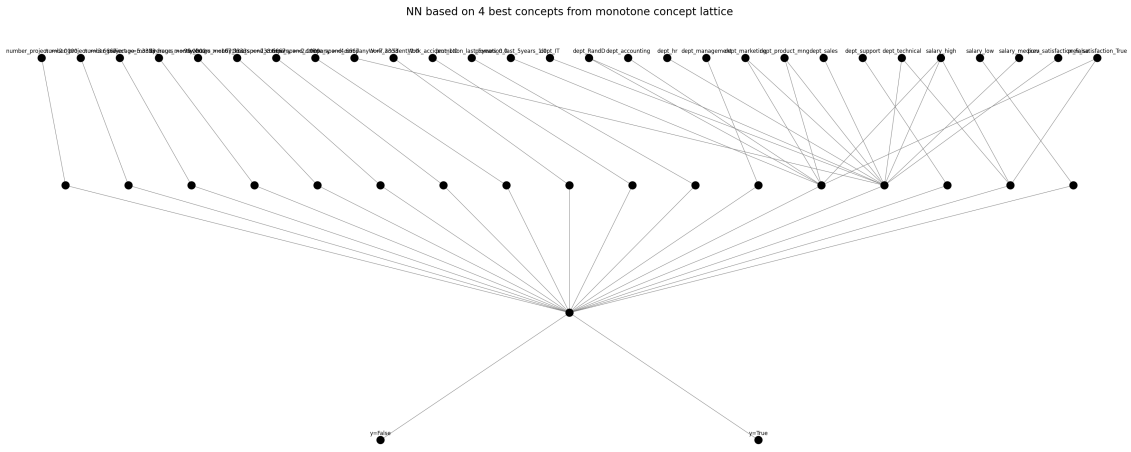

Figure 1. NN architecture produced with the 4 best formal concepts(FCs)

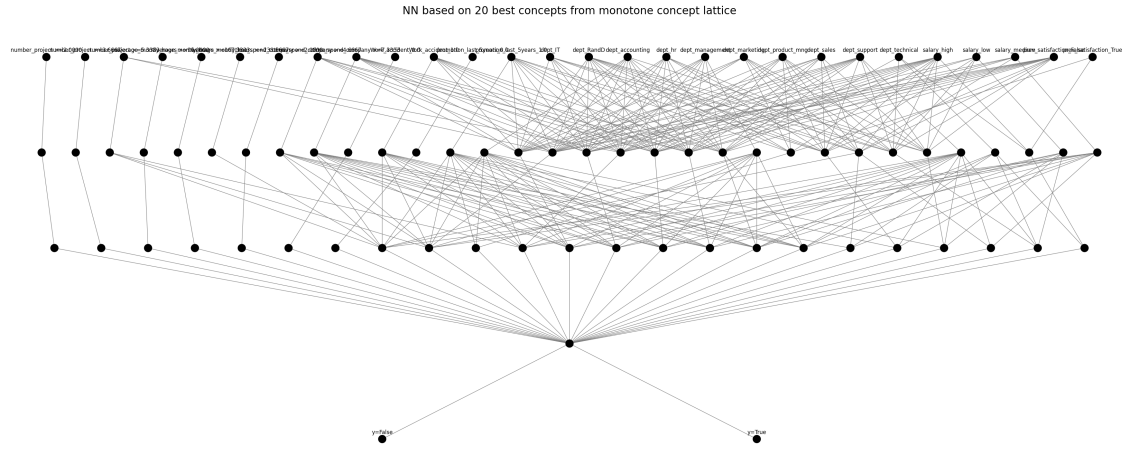

Figure 2. NN architecture produced with the 20 best formal concepts(FCs)

Another binarization type of interest is interordinal encoding. With double the amount of nodes it produces architectures seen on fig. 3 and fig. 4, its performance metrics are recorded in the table A.2.

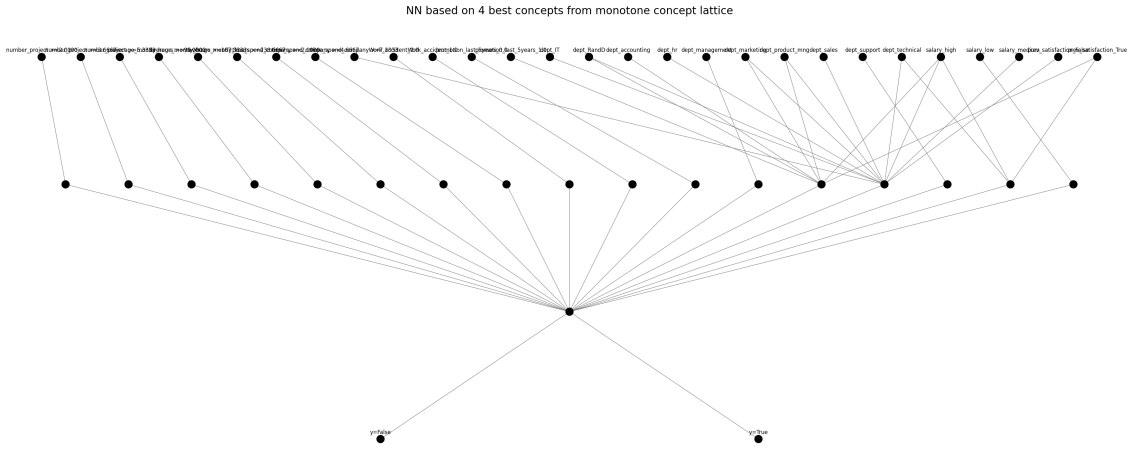

Figure 3. NN architecture produced with the 4 best formal concepts(FCs) with interordinal encoding

The interordinal encoding works by applying both ordinal and inverse ordinal encoding at the same time. It preserves connections between values near the nodes, thus giving a notable performance boost to model performances while working with specific feature types such as age.

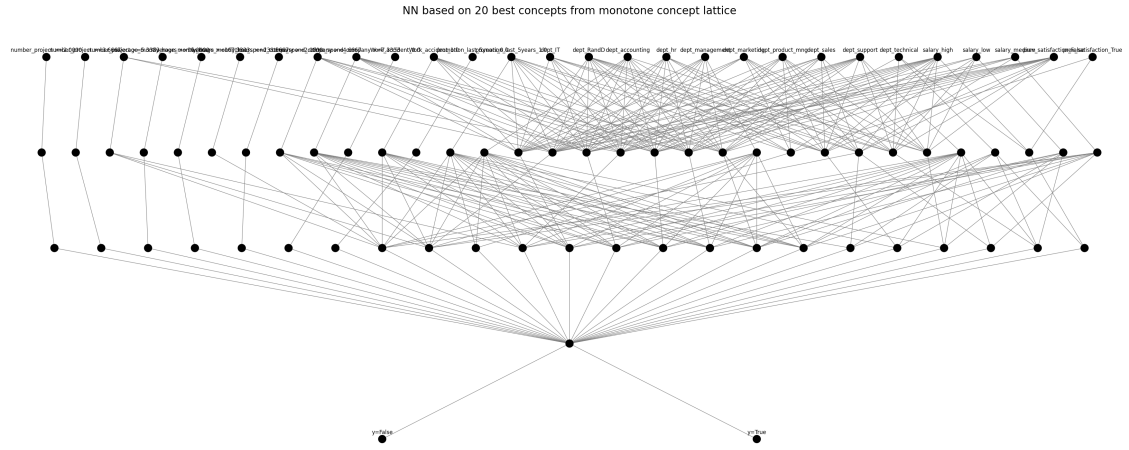

Figure 4. NN architecture produced with the 20 best formal concepts(FCs) with interordinal encoding

A finer numerical feature binarization should be considered in terms of its influence on the quality of predictions. With the increase in quantity of features, new dependencies in numerical data could be found. Table A.3 shows the performance metrics for this case; fig. 5 and fig. 6 demonstrate the NN architectures with this type of encoding.

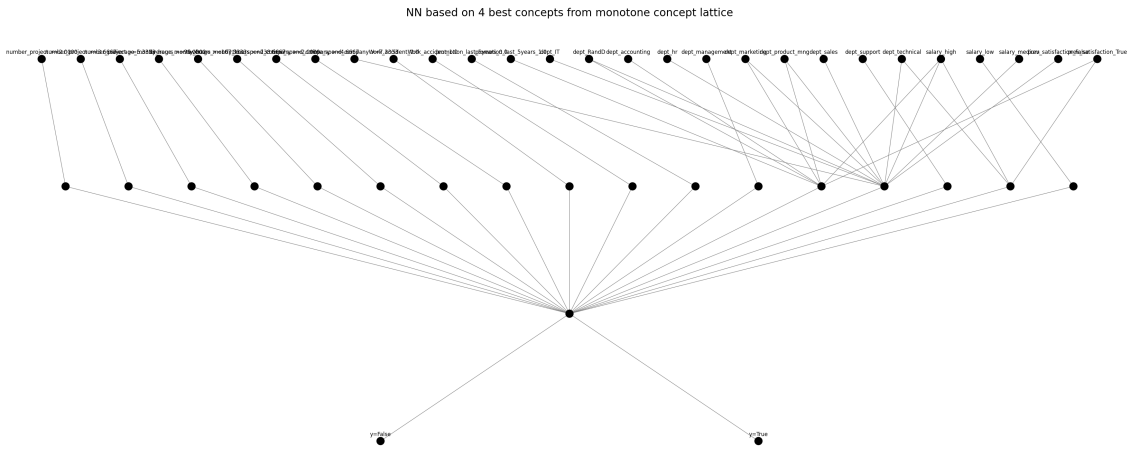

Figure 5. NN architecture produced with the 4 best formal concepts(FCs) with finer ordinal encoding

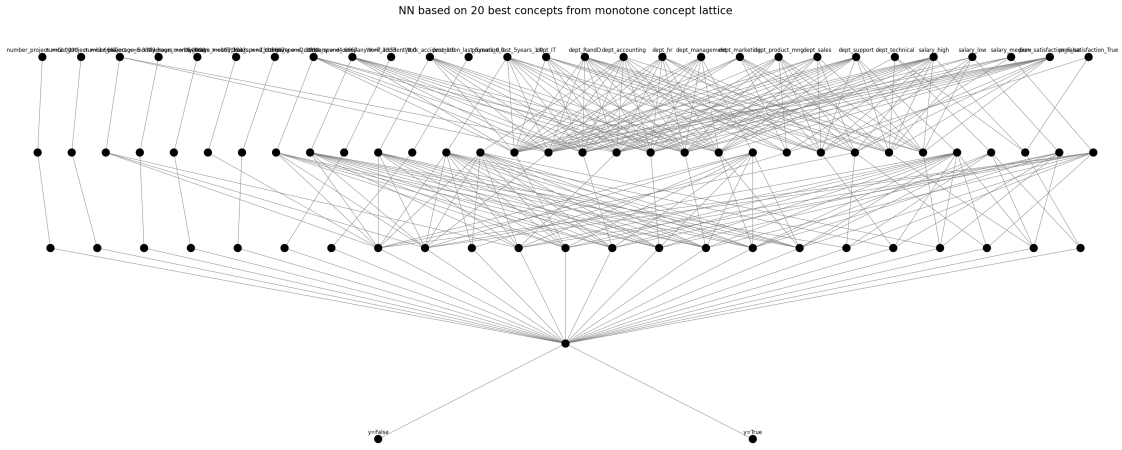

Figure 6. NN architecture produced with the 20 best formal concepts(FCs) with finer ordinal encoding

Finally, the impact of an increase in number formal concepts used to construct a concept lattice needs to be considered. Results of evaluating performance metrics on data with ordinal encoding of numerical features can be seen on fig. 7.

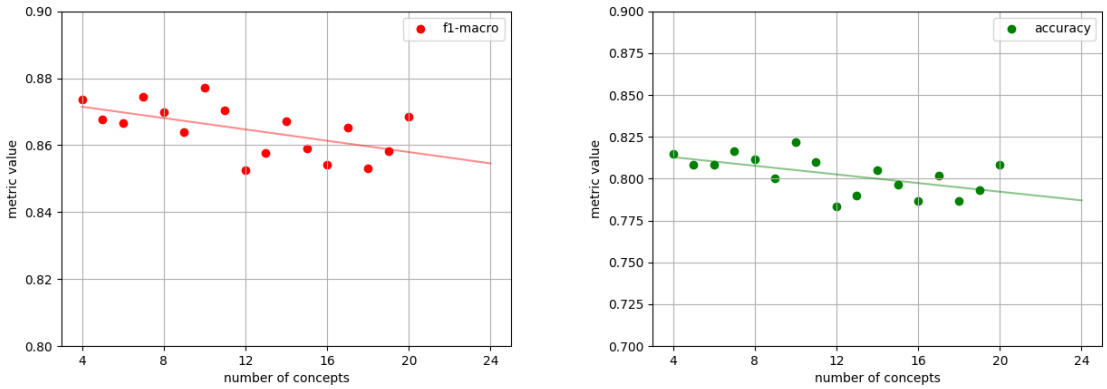

Figure 7. Model performance metrics as functions of number of concepts

1.3 Model evaluation on the Estonia Disaster Passenger List dataset

This dataset small both in terms of number of features and objects and its only numerical feature is age. Therefore, interordinal encoding is the only logical choice for this case since it conserves closeness relation between people of similar age in different age groups. Here lattices from 21 concepts, which is minimal possible amount, and 25 concepts are considered, as shown in fig. 8 and fig. 9, performance measures of those NNs can be seen in table A.4.

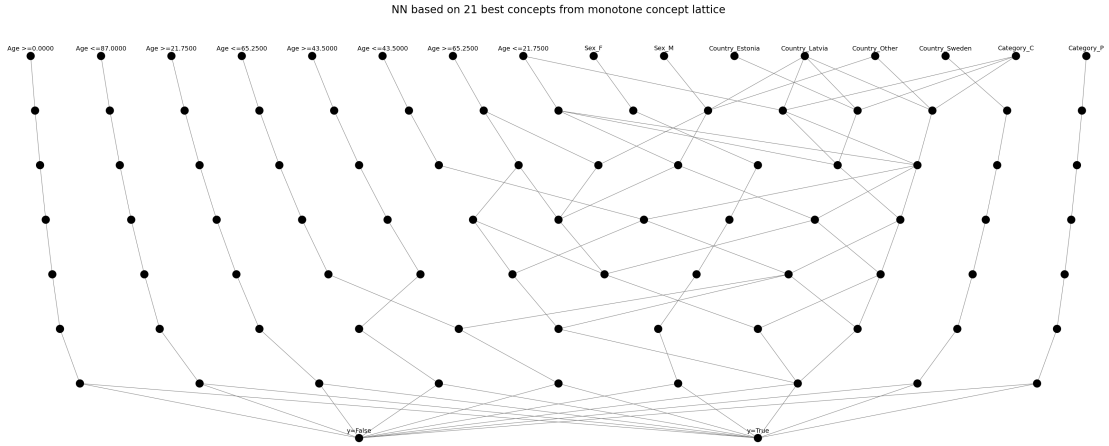

Figure 8. NN architecture produced with the 21 best formal concepts(FCs) on Estonia Disaster dataset

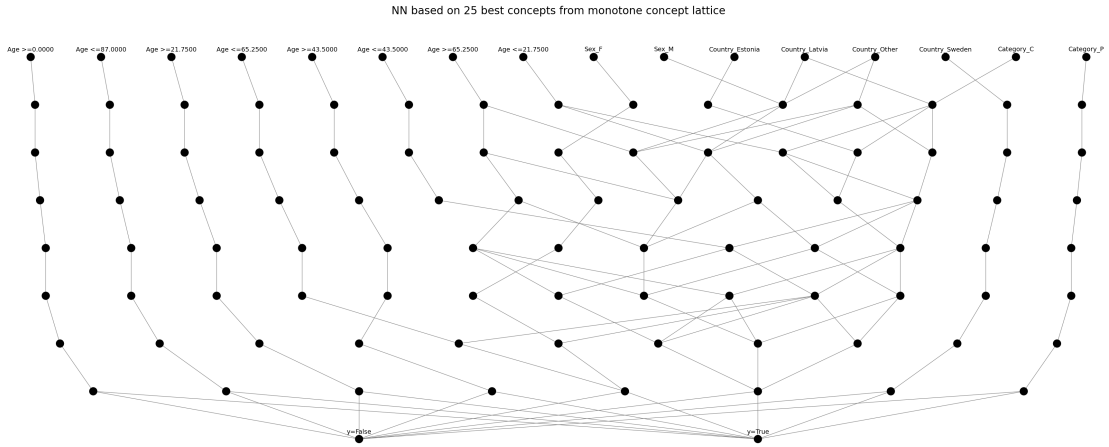

Figure 9. NN architecture produced with the 25 best formal concepts(FCs) on Estonia Disaster dataset

RESULTS

Finer binarization produces better results in terms of model metrics than both the baseline and binarization using interordinal encoding with similar number of nodes, however this comes at the cost of interpretability. Increasing the number of formal concepts used to construct the concept lattice does not improve the model's performance; on the contrary, it seems to correlate with decrease in classification performance. Furthermore, the NN architecture used in this work struggles with overfitting on skewed data as exemplified by the Estonia Disaster dataset.

CONCLUSION

Neural Networks based on concept lattices perform on par with several ensemble models and show promise in some machine learning applications. As exemplified by the Estonia Disaster dataset, these NNs can show connections between various features and their impact on its final decision, hence improving human interpretability of their inner workings.

APPENDIX A

Table A.1 Baseline NN performance on the Employee Attrition dataset

Classifier	f_1**-score**	**accuracy**
GaussianNB	0.831797	0.756667
RandomForestClassifier	0.888889	0.838333
HistGradientBoostingClassifier	0.879147	0.830000
NeuralFCA(4 concepts)	0.864310	0.801667
NeuralFCA(20 concepts)	0.864253	0.800000

Table A.2 NN performance on the Employee Attrition dataset with interordinal encoding

Classifier	f_1**-score**	**accuracy**
GaussianNB	0.831797	0.756667
RandomForestClassifier	0.888889	0.838333
HistGradientBoostingClassifier	0.879147	0.830000
NeuralFCA(4 concepts)	0.870748	0.810000
NeuralFCA(20 concepts)	0.856487	0.791667

Table A.3 NN performance on the Employee Attrition dataset with finer ordinal encoding

Classifier	f_1**-score**	**accuracy**
GaussianNB	0.831797	0.756667
RandomForestClassifier	0.888889	0.838333
HistGradientBoostingClassifier	0.879147	0.830000
NeuralFCA(4 concepts)	0.888889	0.838333
NeuralFCA(20 concepts)	0.900115	0.855000

Table A.4 NN performance on the Estonia disaster dataset

Classifier	f_1**-score**	**accuracy**
GaussianNB	0.831797	0.756667
RandomForestClassifier	0.888889	0.838333
HistGradientBoostingClassifier	0.879147	0.830000
NeuralFCA(21 concepts)	0.00000	0.862903
NeuralFCA(25 concepts)	0.000000	0.862903

REFERENCES

1. Kuznetsov, S.O., Makhazhanov, N., Ushakov, M. On Neural Network Architecture Based on Concept Lattices // Foundations of Intelligent Systems. ISMIS 2017. Lecture Notes in Computer Science(), vol 10352, pp. 653-663. Springer, Cham. https://doi.org/10.1007/978-3-319-60438-1_64