Training Neural Networks Based on Formal Concepts

Sergei O. Kuznetsov[1,*,†], Mariia Zueva[2,†]

[1]HSE University, 20 Myasnitskaya St, Moscow, 101000, Russian Federation

Abstract

This paper presents a modernization of the neural network architecture based on concept lattices, *FCA-CLNet*, utilizing pre-clustering of data based on groups of attributes, unified by a shared interpretable meaning. This approach aims to create a compact model for data classification, with the added benefit of enabling subsequent interpretation of results in scenarios involving a large number of data features.

Keywords

Neural Networks, Clustering, Formal Concept Analysis

1. Introduction

Interpretability in the context of neural networks is an important aspect of research, as it allows us to understand how and why the model makes certain decisions. In recent years, interpretable neural networks have been actively researched and developed in order to overcome the problem of the "black box" and ensure the clarity and explainability of the decision-making process. This is especially important in areas where the decisions made by the model have a significant impact on people's lives and well-being, such as medicine, finance and justice. Finding a balance between the high performance of the model and its interpretability is a key factor for creating reliable and transparent systems capable of interacting with people in confidence.

With the growing demand for AI explainability, many papers addressed the problem of explaining «black box» systems and simultaneously tried to formulate the criteria and measures for evaluating explainability of the model design. In [1] the authors suggested using three core criteria for evaluating machine learning models, namely, interpretability, transparency and explainability. In [2] it was proposed to use expert opinions combined with statistical methods to measure the effectiveness of machine learning models. A first attempt in making a theory of interpretable neural networks (INNs) seems to be made in [3]. The authors managed to align the sparse coding method with existing neural network's architecture, so that the system had the interpretability of the model-based method and the efficiency of the learning-based one.

A series of works have intended to review and classify all existing interpretable methods. In [4] the authors have classified existing interpretable approaches by problem addressed, black-box type and explanation provided, with the purpose to help researchers solve the needed tasks. In [5] the authors suggested to divide interpretable neural network approaches into two types, model decomposition neural networks and semantic interpretable neural networks (INNs). The first one unites methods which inherit domain theoretical knowledge and implement it in the neural network architecture. The decomposition alternative INN starts by taking a complicated mathematical or physical model and breaking it down into smaller, manageable modules. After that it maps the computing of the obtained modules in accordance with the prior knowledge with hyper-parameters of neural network or its hidden layers, thus enhancing their interpretability [6, 7]. The idea can be described as using controllable artificial parameters and structures of neural network instead of the weights without mathematical and physical meaning. This

FCA4AI 2024: The 12th International Workshop "What can FCA do for Artificial Intelligence?", October 19 2024, Santiago de Compostela, Spain.

*Corresponding author.
†These authors contributed equally.
skuznetsov@hse.ru (S. O. Kuznetsov); m.zueva@hse.ru (M. Zueva)
0000-0003-3284-9001 (S. O. Kuznetsov); 0009-0000-6332-9936 (M. Zueva)

approach requires a theoretical model of the domain. An illustration of this concept is the utilization of convex or non-convex optimization algorithms to address mathematical modeling challenges, providing a framework for shaping the objective function. This method is applicable to such tasks as solving partial differential equations (PDE) [8], image deblurring, super-resolution, and other problems [9, 10, 11]. The second approach is semantic INNs [12], which is meant to explain the model's decision afterwards, with the process close to human semantic interpretation. The authors highlighted three different branches in this approach, namely, convolution neural network (CNN) visualization [6], decision tree regularization [13], and semantic knowledge graph [14]. In [15] the taxonomy of interpretable methods was proposed. This paper categorized existing architecture designs by three criteria, namely, the type of engagement (passive or active), the type of explanation and the focus, varying from local to global interpretability and provided the way how to order them in subcategories. The first architecture of neural networks based on the formal concept analysis (FCA) approach was proposed in [16]. In this work the authors propose building neural networks based on concept lattices and on lattices coming from monotone Galois connections. Later, in 2022 in [17] the authors integrated conceptual information into the message passing through graph neural networks (GNNs). The authors of [18] proposed an approach using BERT, which can learn more information from the maximal bi-cliques, which correspond to formal concepts, and use them to make link prediction.

This paper explores the potential of incorporating clustering methods into a compact neural network architecture. Specifically, it introduces a modernization of the neural network framework, *FCA-CLNet*, which leverages concept lattices in conjunction with pre-clustering of data based on semantic attribute groups. The proposed approach is particularly suited for scenarios involving a large number of data features and aims to improve the interpretability of the model's performance.

2. Clustering

Clustering is a widely used useful tool for working with big data and data mining. A large number of clustering algorithms have been developed, each of which has its own area of application. A number of works have been devoted to creating a taxonomy of clustering methods. In [19] authors proposed a categorization framework to classify existing clustering algorithms into groups. They divided all algorithms into partitioning-based, hierarchical-based, density-based, grid-based and model-based:

Partitioning-based algorithms [20] first set the clusters from initial data and then redistribute the data points towards better group organization. Widely used K-means algorithm belongs to this category.

Hierarchical-based methods are intended to organize data hierarchically, based on the medium of proximity. Using these methods datasets can be represented by dendrograms, where each leaf node corresponds to an individual data point. Hierarchical-based methods are divided into agglomerative (bottom-up) and divisive (top-down) approaches [21]. In the former the process starts with clusters containing one object, and then they are united together towards more suited. In the latter the whole dataset is one cluster at the beginning and then it is recursively split into smaller ones till reaching the stopping criterion.

Grid-based methods are based on splitting the data space on grids and accumulating grid-data. The advantages of this approach are fast processing time and independence of the number of data objects.

In model-based methods [22] it is supposed that there is a mixture of probability distributions that generate the given data, so these approaches try to accommodate the data to the predefined mathematical model. These approaches are divided into statistical and neural network approaches.

In this paper, we chose four well-known clustering algorithms for preclustering the data features: K-means, Mean-Shift, DBSCAN and HDBSCAN.

3. Formal Concept Analysis

In FCA-CLNet architecture we operate with the terms related to formal concepts analysis (FCA). Let us recall some basic definitions of FCA [23]. The basic FCA structure is a binary datatable, called formal context, where rows stay for the set of objects, denoted by G, the columns stay for the set of attributes, denoted M and binary relation $I \subseteq G \times M$ is defined in the way so that $(g, m) \in I$ if the object g possesses the attribute m. The triple $K = (G, M, I)$ is called a *formal context*. Derivation operators $(\cdot)'$ for $A \subseteq G$, $B \subseteq M$ are defined as follows:

$$A' = \{m \in M \mid gIm \ for \ all \ g \in A\}, \tag{1}$$

$$B' = \{g \in G \mid gIm \ for \ all \ m \in B\}, \tag{2}$$

These derivation operators form *(antimonotone) Galois connection* on the ordered powersets $(2^G, \subseteq)$ and $(2^M, \subseteq)$.

we define a *classical formal concept of a formal context* K as a pair (A, B) such that $A \in G$, $B \in M$, $A' = B$, $B' = A$. Here A is called an *extent* and B is called an *intent* of the formal concept (A, B).Classical formal concepts are ordered by the relation $\geq$:

$$(A_1, B_1) \leq (A_2, B_2) \iff A_1 \subseteq A_2, \tag{3}$$

which defines a complete (algebraic) lattice on the set of concepts called *concept lattice* $L = (G, M, I)$. The covering relation corresponding to the partial order $\leq$, (if it exists) is defined as $\prec$:

$$(A_1, B_1) \prec (A_2, B_2) \iff (A_1, B_1) \leq (A_2, B_2) \tag{4}$$

and there is no concept (A_3, B_3) such that $(A_1, B_1) < (A_3, B_3) < (A_2, B_2)$.

Classical formal concepts are also called antimonotone formal concepts or formal concepts based on antimonotone Galois connection.

In our study we use another type of formal concepts called formal concepts based on *monotone Galois connection* or monotone formal concepts [24]. They are defined as pairs (A, B), which satisfy monotone Galois connection, that is

$$A^{\vee} = \{b \mid \nexists a \in G \backslash A \text{ such that } aIb\}, \tag{5}$$

$$B^{\wedge} = \{a \mid \exists b \in B \text{ such that } aIb\}, \tag{6}$$

where $A \subseteq G, B \subseteq M$ and $A = B^{\vee}, B = A^{\wedge}$.

In other words, for each set of objects A, we match all the attributes belonging only to objects from A'. On the other hand, the set of attributes B corresponds to the set of all objects B' satisfying at least one attribute from B. A and B are also called an *extent* and an *intent* of the formal concept.

A partial order on the set of all monotone formal concepts is defined as:

$$(A_1, B_1) \leq (A_2, B_2) \iff A_1 \subset A_2 \leftrightarrow B_1 \subset B_2. \tag{7}$$

.

We also can define *monotone concept lattice* based on this partial order.

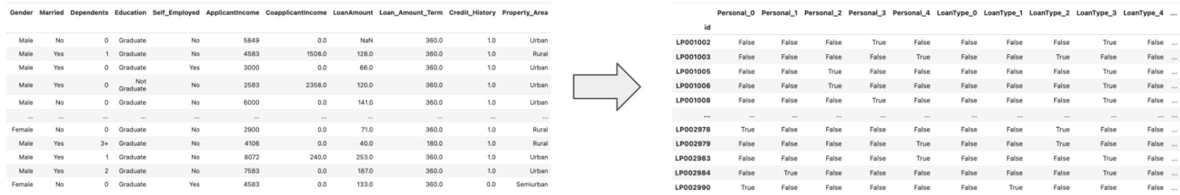
All monotone formal concepts can be obtained from the given formal context $K = (G, M, I)$ by finding its complement context $\bar{K} = (G, M, \bar{I})$ and then finding all its classical formal concepts.

4. FCA-CLNet

The proposed method utilizes a neural network architecture based on concept lattices. The idea of this neural network was proposed in [16]. This article extends the approach by incorporating an additional step, namely data pre-clustering, to derive novel features.

The method description is as follows:

Suppose $K = (G, M, I)$ is a *formal context*, where G is the set of objects, M is a set of attributes and I is a binary relation.



Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban
Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural
Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban
Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban
Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban
...	...	...	...	...	...	...	...	...	...	...
Female	No	0	Graduate	No	2900	0.0	71.0	360.0	1.0	Rural
Male	Yes	3+	Graduate	No	4106	0.0	40.0	180.0	1.0	Rural
Male	Yes	1	Graduate	No	8072	240.0	253.0	360.0	1.0	Urban
Male	Yes	2	Graduate	No	7583	0.0	187.0	360.0	1.0	Urban
Female	No	0	Graduate	Yes	4583	0.0	133.0	360.0	0.0	Semiurban

	Personal_0	Personal_1	Personal_2	Personal_3	Personal_4	LoanType_0	LoanType_1	LoanType_2	LoanType_3	LoanType_4	...
id											
LP001002	False	False	False	True	False	False	False	False	True	False	...
LP001003	False	False	False	False	True	False	False	True	False	False	...
LP001005	False	False	True	False	False	False	False	False	True	False	...
LP001006	False	False	True	False	False	False	False	False	True	False	...
LP001008	False	False	False	True	False	False	False	False	True	False	...
...	...	...	...	...	...	...	...	...	...	...	...
LP002978	True	False	False	False	False	False	False	True	False	False	...
LP002979	False	False	False	False	True	False	False	True	False	False	...
LP002983	False	False	False	False	True	False	False	False	True	False	...
LP002984	False	True	False	False	False	False	False	False	True	False	...
LP002990	True	False	False	False	False	False	True	False	False	False	...

Figure 1: Dataset preprocessing using clustering

1. From the set of attributes M choose disjoint sets of attributes $M_1, M_2, \ldots, M_k$, such that $M_1 \cup M_2 \cup \ \ldots \ \cup M_k = M$ and elements of each set can be unified by a shared interpretable meaning. For example, for the formal context related to banking data, such attributes as “gender”, “marital status”, “number of dependents” can be unified as “client personal information”, and “education”, “self-employment”, “income”, “co-applicant’s income” as “the client’s ability to repay the loan”.
2. Separately apply a chosen clustering method to the attribute sets $M_1, M_2, \ldots, M_k$ and obtain clustering results as sets of clusters $C_1 = \{c_{11}, ..., c_{1t}\},\ C_2 = \{c_{21}, ..., c_{2t}\}, \ldots, C_k = \{c_{k1}, ..., c_{kt}\}$.
3. Create a new formal context $K_{cl} = (G, M_{cl}, I_{cl})$, where G is the initial set of objects, $M_{cl} = \{C_1 \ \cup C_2 \ \cup \ \ldots \ \cup \ C_k\}$ is a new attribute set, where each attribute stands for a cluster, I_{cl} - a binary membership relation to a given cluster. The example of dataset transformation is shown at Figure 1.
4. Find the most stable concepts based on monotone Galois connection [24] according to $\Delta - stability$ index [25]. Algorithm Sofia [26] can be used for this purpose.
5. Choose the “most interesting” concepts based on interestingness indices [27] to reduce the size of concept lattice (F1-score, accuracy, etc.) The example of concept lattice size reduction is shown at Figure 2.
6. Build neural network based on the reduced concept lattice. The architecture of the neural network is given as follows (Figure 3):
   - ***Input layer*** is created by the obtained attributes from dataset pre-clustering. Each attribute represents one of the clusters.
   - ***Hidden layers*** consisting of neurons corresponding to the resulting clusters.
   - ***Last hidden layer*** is connected to an ***output layer*** in which the number of neurons corresponds to the number of classes.

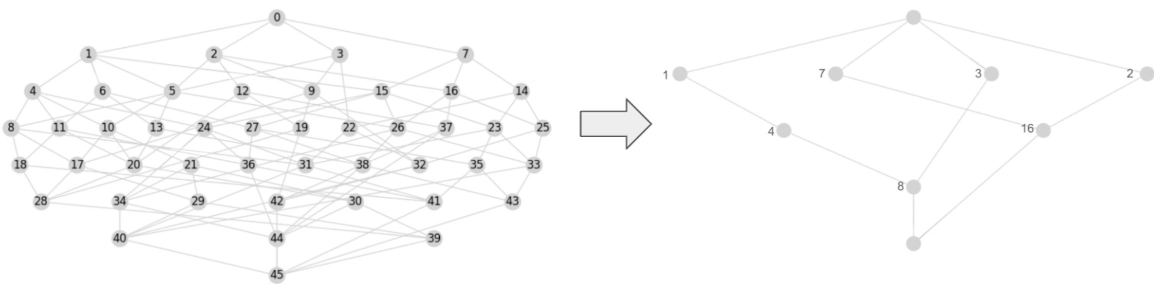


Figure 2: Concept lattice size reduction using "most interesting" concepts

In the current study, two approaches for choosing “the best” concepts were tested: based on F1-score and based on the accuracy metrics. For a single concept (A, B), the metrics was calculated with the following method:

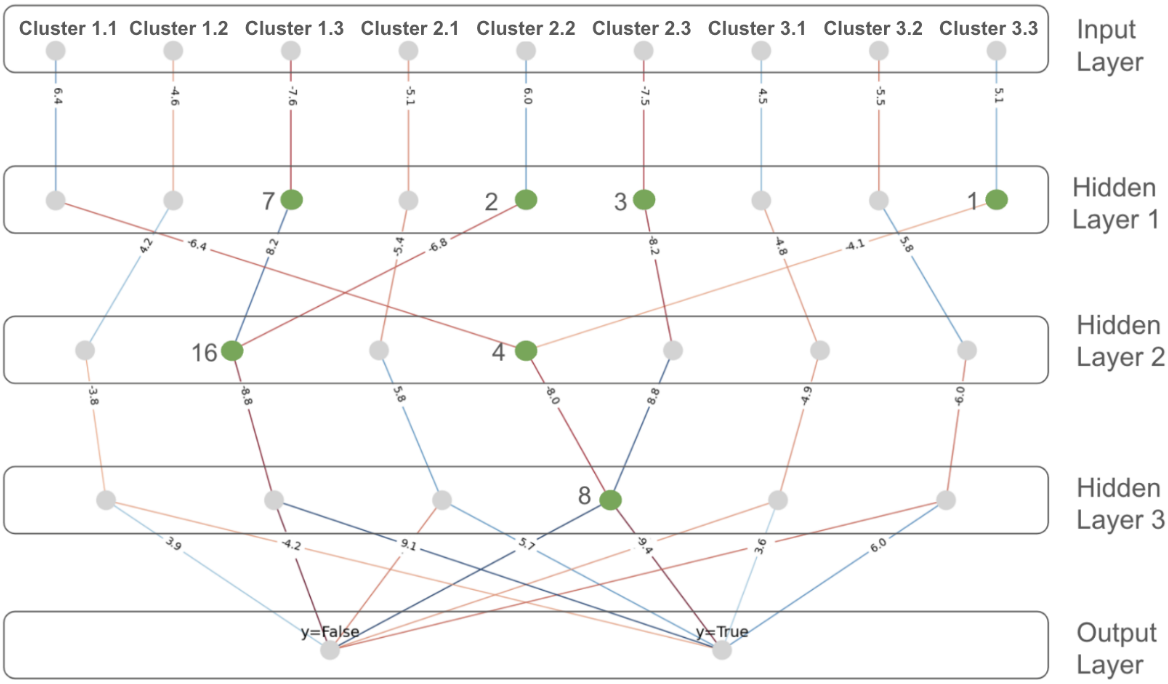


Figure 3: Neural network architecture based on concept lattice

- Assume that:
  $y_{pred}[g_i] = True, \ if \ g_i \in A$,
  $y_{pred}[g_i] = False, \ if \ g_i \notin A$; - an object is predicted True if it is in the extent of the concept and False otherwise;
- F1-score
  F1-score = F1-score(y, y_{preds}), where y_{preds} - predicted target values, y - real target values;
- accuracy = accuracy(y, y_{preds});
- Sort the concepts by the metrics value and choose 10 top concepts for building the neural network.

5. Experimental Part

To automatically find concepts for the *FCA-CLNet* architecture, build and train a neural network, this study uses the *FCApy* library (https://pypi.org/project/fcapy). This library provides the necessary tools for working with formal concepts and allows to automate the process of building and training a neural network based on these concepts.

Also, for a general understanding of the neural network, it is worth noting that sigmoid activation function is used for hidden layers. The value of softmax function is used for the output layer. When learning, binary cross-entropy is used as a loss function, and the Adam algorithm with the learning rate = 0.01 is used as an optimizer.

In this study, the performance of the model was compared with the following basic methods: *KNeighborsClassifier*, *LogisticRegression*, *RandomForestClassifier*, *CatBoostClassifier*, *XGBClassifier* and *TabNetClassifier*. Each of these methods was tested both on the initial dataset and on the dataset after clustering.

6. Data Description

For the purpose of our study we have chosen three datasets for binary classification from UCI Machine Learning Repository (https://archive.ics.uci.edu/) (Table 1):

Table 1
Dataset characteristics

Dataset	Number of objects	Number of attributes	Number of classes in target attribute
Credit Approval	690	15	2
Wine Quality	4898	11	2
Mammographic Mass	961	5	2

Each dataset represents a separate task and has its own unique characteristics, such as feature types, data size, class distribution, and noise presence. This approach allows one to consider different scenarios and evaluate the performance of models on different types of data.

7. Results

For the experimental evaluation, four clustering methods were applied for feature pre-clustering: K-means, Mean-Shift, DBSCAN, and HDBSCAN. 10 "most interesting" concepts were selected as neurons for the neural network architecture using two distinct concept selection methods. The results obtained for these two methods are presented in Table 2.

Table 2
FCA-CLNet concept selection method results (10 concepts). Weighted F1-score.

Clustering method	Best concepts selection	Loan Approval	Wine Quality	Mammographic
K-means	F1-score	0.79	**0.76**	**0.79**
	Accuracy	**0.83**	0.7	0.76
Mean-Shift	F1-score	0.84	0.72	0.74
	Accuracy	0.84	0.74	**0.79**
DBScan	F1-score	0.79	0.7	0.75
	Accuracy	0.81	0.69	0.76
HDBScan	F1-score	0.79	0.71	**0.77**
	Accuracy	**0.86**	0.7	0.73

The table shows that there is no significant difference in performance among the concept selection methods across all three datasets. For the K-means clustering approach, the F1-score-based concept selection method demonstrates better results in two out of the three datasets. The higher performance observed in the Loan Approval dataset may be attributed to the fact that the grouped features for clustering are more semantically similar than in the other datasets. Conversely, the method performs worst on the Wine Quality dataset, potentially indicating that this method is more effective when features can be easily divided into interpretable groups.

Subsequently, the performance of the proposed model was compared with that of classical machine learning methods on the same datasets. The performance of the FCA-CLNet model is very close to that of classical machine learning models, see Table 3.

8. Conclusion

In this paper, we investigated the application of feature pre-clustering for computing neural network architecture based on concept lattices, The proposed *FCA-CLNet* method demonstrated performance comparable to that of classical machine learning models, suggesting the potential for successfully integrating clustering methods into FCA-based approaches. While the results are promising, further development of the model is necessary to enhance its performance.

Table 3
Model Comparison. Weighted F1-score.

ML method	Clustering Method	Loan Approval	Wine quality	Mammographic
K-Neighbors	Without clustering	0.64	0.71	0.80
	K-Means	0.70	0.69	0.82
	Mean-Shift	0.71	0.66	0.81
	DBScan	0.70	0.67	0.81
	HBDScan	0.69	0.66	0.80
Logistic Regression	Without clustering	0.72	0.74	0.81
	K-Means	0.72	0.71	0.84
	Mean-Shift	0.72	0.68	0.84
	DBScan	0.74	0.69	0.81
	HBDScan	0.72	0.69	0.81
Naive Bayes	Without clustering	0.19	0.73	0.81
	K-Means	0.72	0.68	0.80
	Mean-Shift	0.22	0.38	0.84
	DBScan	0.68	0.54	0.64
	HBDScan	0.35	0.60	0.62
Random Forest	Without clustering	0.72	0.80	0.79
	K-Means	0.70	0.72	0.82
	Mean-Shift	0.73	0.68	0.84
	DBScan	0.72	0.69	0.81
	HBDScan	0.72	0.74	0.82
XGBoost	Without clustering	0.66	**0.81**	0.80
	K-Means	0.72	0.73	0.82
	Mean-Shift	0.73	0.68	0.84
	DBScan	0.72	0.70	0.82
	HBDScan	0.72	0.68	0.82
FCA - CLNet	K-Means	0.79	0.76	0.79
	Mean-Shift	**0.84**	0.72	0.74
	DBScan	0.79	0.7	0.75
	HBDScan	0.79	0.71	0.77

9. Acknowledgements

The work by S.O. Kuznetsov and M.M. Zueva in preparing this article was supported by grant 22-11-00323 of the Russian Science Foundation and carried out at the National Research University Higher School of Economics, Moscow.

10. Declaration on Generative AI

During the preparation of this work, the authors used Chat-GPT-4 in order to: text translation, rephrasing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. Roscher, B. Bohn, M. F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries, Ieee Access 8 (2020) 42200–42216.
[2] O. Loyola-Gonzalez, Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view, IEEE access 7 (2019) 154096–154113.
[3] K. Gregor, Y. LeCun, Learning fast approximations of sparse coding, in: Proceedings of the 27th international conference on international conference on machine learning, 2010, pp. 399–406.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.
[5] Z. Liu, F. Xu, Interpretable neural networks: principles and applications, Frontiers in Artificial Intelligence 6 (2023).
[6] Q. Zhang, Y. N. Wu, S.-C. Zhu, Interpretable convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8827–8836.
[7] N. Shlezinger, J. Whang, Y. C. Eldar, A. G. Dimakis, Model-based deep learning, Proceedings of the IEEE (2023).
[8] S. H. Rudy, S. L. Brunton, J. L. Proctor, J. N. Kutz, Data-driven discovery of partial differential equations, Science advances 3 (2017) e1602614.
[9] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences 57 (2004) 1413–1457.
[10] Z. Wang, D. Liu, J. Yang, W. Han, T. Huang, Deep networks for image super-resolution with sparse prior, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 370–378.
[11] Y. Li, M. Tofighi, J. Geng, V. Monga, Y. C. Eldar, Efficient and interpretable deep blind image deblurring via algorithm unrolling, IEEE Transactions on Computational Imaging 6 (2020) 666–681.
[12] F.-L. Fan, J. Xiong, M. Li, G. Wang, On interpretability of artificial neural networks: A survey, IEEE Transactions on Radiation and Plasma Medical Sciences 5 (2021) 741–760.
[13] N. Frosst, G. Hinton, Distilling a neural network into a soft decision tree, arXiv preprint arXiv:1711.09784 (2017).
[14] Q. Zhang, R. Cao, S. Zhang, M. Redmonds, Y. N. Wu, S.-C. Zhu, Interactively transferring cnn patterns for part localization, arXiv preprint arXiv:1708.01783 (2017).
[15] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Transactions on Emerging Topics in Computational Intelligence 5 (2021) 726–742.
[16] S. O. Kuznetsov, N. Makhazhanov, M. Ushakov, On neural network architecture based on concept lattices, in: Foundations of Intelligent Systems: 23rd International Symposium, ISMIS 2017, Warsaw, Poland, June 26-29, 2017, Proceedings 23, Springer, 2017, pp. 653–663.
[17] M. Shao, Z. Hu, W. Wu, H. Liu, Graph neural networks induced by concept lattices for classification, International Journal of Approximate Reasoning 154 (2023) 262–276.
[18] S. Peng, H. Yang, A. Yamamoto, Bert4fca: A method for bipartite link prediction using formal concept analysis and bert, arXiv preprint arXiv:2402.08236 (2024).
[19] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, A. Bouras, A survey of clustering algorithms for big data: Taxonomy and empirical analysis, IEEE transactions on emerging topics in computing 2 (2014) 267–279.
[20] A. K. Jain, R. C. Dubes, Algorithms for clustering data, Prentice-Hall, Inc., 1988.
[21] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, C.-T. Lin, A review of clustering techniques and developments, Neurocomputing 267 (2017) 664–681.
[22] C. Fraley, A. E. Raftery, How many clusters? which clustering method? answers via model-based cluster analysis, The computer journal 41 (1998) 578–588.
[23] B. Ganter, R. Wille, Formal Concept Analysis. Mathematical Foundations, Springer, 2024.
[24] L. Zhao, M. J. Zaki, N. Ramakrishnan, Blosom: a framework for mining arbitrary boolean expressions, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 827–832.
[25] A. Buzmakov, S. O. Kuznetsov, A. Napoli, Scalable estimates of concept stability, in: Formal Concept Analysis: 12th International Conference, ICFCA 2014, Cluj-Napoca, Romania, June 10-13, 2014. Proceedings 12, Springer, 2014, pp. 157–172.
[26] A. Buzmakov, S. O. Kuznetsov, A. Napoli, Fast generation of best interval patterns for nonmonotonic constraints, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2015, pp. 157–172.
[27] S. O. Kuznetsov, T. Makhalova, On interestingness measures of formal concepts, Information Sciences 442 (2018) 202–219.