Fuzzy Rules Extraction from Deep Neural Networks*

Alexey Averkin[1, 2][0000-0003-1571-3583] and Sergey Yarushev[3] [0000-0003-1352-9301]

[1] Federal Research Centre of Informatics and Computer Science of RAS, Moscow, Vavilova, 42, Moscow, Russia
[2] Educational and Scientific Laboratory of Artificial Intelligence, Neuro-technologies and Business Analytics, Plekhanov Russian University of Economics, Moscow, Russia
averkin2003@inbox.ru
[3] Plekhanov Russian University of Economics, Moscow, Russia
yarushev.sa@rea.ru

Abstract. This article presents the basic methods of machine learning and explanational artificial intelligence that can help in the issue of extracting rules and other models of knowledge representation not only from data, but from the artificial neural networks themselves. The paper discusses classification methods for rule-based learning methods for neural networks and the current state of technologies for extracting rules from neural networks. Next, we formulate the main problems that arise when extracting rules from neural networks, as well as the main methods for solving them.

Keywords: Deep learning, Neural networks, rule extraction, Convolutional neural network, Machine learning, Artificial intelligence, explanational artificial intelligence.

1 Introduction

This paper introduces the main concepts of machine learning that are relevant to the context of extracting rules from classical and deep neural networks. It includes the problem of classification of rule-based teaching methods and neural networks. Then, we will look at the current state of rule extraction from neural networks. Here we define the problem as well as the main approaches to its solution and present some of the existing rules extraction algorithms. The last part discusses specific problems when working with deep neural networks and neuro-fuzzy systems. At this stage, we also propose some algorithms that can successfully extract rules from these more complex neural networks.

Artificial Neural Networks (ANN) are widely known parallel computing models that exhibit excellent behavior in solving complex problems of artificial intelligence. However, many researchers refuse to use them due to their being a “black box”. This

* The reported study was funded by RFBR according to the research project № RFBR 19-07-01208 and the reported study was funded by RFBR and BRFBR, project number20-57-00024

Proceedings of the of the XXIII International Conference "Enterprise Engineering and Knowledge Management" (EEKM 2020), Moscow, Russia, December 8-9, 2020.

means that determining why a neural network makes a specific decision is a difficult task.

This is a significant drawback, since it is difficult to trust the reliability of the network that solves real problems without the ability to make acceptable decisions. For example, this is the case in critical, in terms of safety, applications where hidden failure can lead to life-threatening actions or huge economic losses

In addition, studying how neural networks extract, store and transform knowledge can be useful for future machine learning methods. For example, increasing the transarency of neural networks can help detect the so-called hidden dependencies that are not present in the input data, but appear as a result of their integration into the neural network. To overcome this lack of neural networks, researchers came up with the idea of extracting rules from neural networks, which can became a bridge between symbolic and connectionist models of knowledge representation in artificial intelligence.

Most authors focus on extracting the most understandable rules, and at the same time they should mimic the behavior of the neural network as precisely as possible, right up to an isomorphic representation of fuzzy rules in the form of a neuro-fuzzy system. Since 1992, since Chang's doctoral thesis on neuro-fuzzy networks, much work has been done in this area, which ended with the creation of the direction of soft computing. Since then, many methods for extracting rules from neural networks have been developed and evaluated, and excellent results have been obtained for many approaches.

However, despite the fact that there are quite a few available algorithms, none of them has ever been explicitly tested in deep neural networks. In addition, most authors focus on networks with only a small number of hidden layers. Only in the last few years pioneering work has appeared on the analysis of specific methods for extracting rules from deep-seated networks and new algorithms are presented that are capable of performing this task.

2 Methods for Extracting Rules from the Neural Network

In artificial intelligence, neural networks and rule-based learning methods are two approaches to solving classification problems. Both methods are known variants of studying models that predict classes for new data. For many tasks, NN rules-based teaching methods excel in accuracy.

However, neural networks have one major drawback: the ability to understand what a trained concept models, NN is not as strong as for rule based approaches. The concepts learned by neural networks are difficult to understand because they are represented using a large set of parameters [1].

Increasing the transparency of neural networks by extracting rules has two main advantages. First, it gives the user some insight into how the neural network uses input variables to make a decision — and can even reveal hidden functions in NN when the rules are used to explain individual neurons. Identification of particularly important attributes or identification of the causes of neural network errors can be part of the understanding. Trying to make opaque neural networks more understandable, methods for extracting rules eliminate the gap between accuracy and clarity [2-4].

A more comprehensible form is required if, for example, a neural network is to be used in safety-critical applications, such as aircraft and power plants. In these cases, it is extremely important that the system user have the opportunity to check the output of the artificial neural network under all possible input conditions [5].

To formalize the task of extracting rules from a neural network, we give Craven's definition: “Given the trained neural network and the data on which it was trained, create a description of the network hypothesis that is understandable, but comes close to the network prediction behavior" [6].

To distinguish between different approaches for extracting rules from neural networks, Andrews introduced the widely used multi-dimensional taxonomy [5]. The first dimension they describe is the expressive power of the extracted rules (for example, IF-THEN rules or fuzzy production rule).

The second dimension is called translucency and describes the strategy followed by the algorithm for extracting rules. If the method uses a neural network only as a black box, regardless of the architecture of NN, we call it the pedagogical approach. If, instead, the algorithm takes into account the internal structure of the neural network, we call this approach decompositional. If the algorithm uses components of both pedagogical and decomposition methods, then this approach is called eclectic.

The third dimension is the quality of the rules extracted. Since quality is a broad term, it is divided into several criteria, namely, accuracy, fidelity, consistency, and comprehensibility. While accuracy measures the ability to correctly classify previously unseen examples, fidelity measures the degree to which rules can imitate the behavior of a neural network well [2].

Fidelity can be considered as accuracy relative to the output of NN. Consistency can only be measured when the rule extraction algorithm involves learning the neural network instead of processing the already trained NN: The extracted rule set is considered consistent when the neural network generates rule sets that correctly classify test data for different training sessions. Comprehensibility is considered here as a measure of the size of the rules, that is, short and few rules are considered more understandable [5]

In this paper we will focus only on the three criteria described. In accordance with [7], we focus on methods that do not impose special requirements on how the neural network was trained before the rules were extracted. In addition, only algorithms that are capable of extracting rules from direct propagation neural networks, despite any other characteristics of the architecture, are analyzed. According to [3] we want the algorithm to offer a high level of generality.

Let us analyze some methods for extracting rules that meet the above characteristics. We start with decomposition approaches. As mentioned earlier, decomposition approaches for extracting rules from neural networks operate at the neuron level. Usually, the decomposition method analyzes each neuron, and forms rules that imitate the behavior of this neuron. For various reasons, we do not take into account all available decomposition approaches in the subsequent review. We consider here the KT algorithm, Tsukimoto's polynomial algorithm and rule extractor via decision tree induction.

The KT algorithm was one of the first decomposition approaches for extracting rules from neural networks was presented in [8]. The KT algorithm describes each neuron (layer by layer) with the IF-THEN rules by heuristically searching for combinations of input attributes that exceed the threshold of the neuron. The rewrite module is used to obtain rules that refer to the original input attributes, and not to the outputs of the previous level. To find suitable combinations, the KT method applies a search on the tree, that is, a rule (represented as a node in the tree) at this level generates its child nodes by adding an additional available attribute [8]. In addition, the algorithm uses a number of heuristics to stop the growth of a tree in situations where further improvement is impossible.

Tsukimoto's polynomial algorithm to extracting rules from a neural network is very similar to the KT method. It also uses a layered decomposition algorithm to extract the IF-THEN rules for each neuron, and also monitors the strategy for finding input configurations that exceed the threshold of the neuron. The main advantage of the Tsukimoto method is its computational complexity, which is polynomial, while the KT method is exponential [9]. The algorithm achieves polynomial complexity by searching for relevant terms using the space of multilinear functions. In the second stage, these terms are used to create IF-THEN rules. Subsequently, if any, training data is used to improve the accuracy of the rules. In the last step, the Tsukimoto algorithm attempts to optimize comprehensibility c by removing non-essential attributes from the rules.

Another method for extracting rules through decision tree induction was introduced in [10]. Their CRED algorithm converts each output unit of a neural network into a solution, where tree nodes are tested using nodes of a hidden layer, and leaves represent a class. After this, intermediate rules are extracted from this step. Then for each split point used in these rules, another decision tree is created using split points on the input layer of the neural network. In the new trees, the leaves do not directly choose the class. Extracting the rules from the second decision tree leads us to the description of the state of hidden neurons, consisting of input variables. As a final step, intermediate rules that describe the output layer through the hidden layer and those that describe the hidden layer based on the inputs of the neural network are replaced. Then they are combined into construction rules that describe the output of the neural network based on its input data.

The main group of pedagogical approaches of rule extraction consist of validity interval analysis , approaches for rule extraction using sampling and rule extraction by reverse engineering the neural network.

Pedagogical approaches do not take into account the internal structure of the neural network. The motive in pedagogical approaches is to treat trained NN as a single entity or alternatively as a black box [11]. The main idea is to extract rules by directly mapping inputs to outputs [12].

The pedagogical approaches usually have access only to the function of the neural network. This function returns the output of the neural network for random input, but offers no understanding of the internal structure of NN or any weights. Having NN, this class of algorithms tries to find coherence between possible input variations and outputs created by the neural network, while some of them use specified training data, and some do not.

Rule extraction based on interval analysis approach uses the interval confidence analysis (VIA), to extract rules that mimic the behavior of a neural network [13]. The main idea of this method is to find the input intervals in which the output signal NN is stable, that is, the predicted class is the same for slightly changing input configurations. As a result, VIA provides the basis for reliably correct rules.

Retrieving rules using sampling represents several methods that follow more or less the same strategy for extracting rules from a neural network using sampling, that is, they create an extensive set of data as a basis for extracting rules. After that, the selected data set is submitted to a standard learning algorithm for generating rules that simulate the behavior of a neural network. In [2] it is proved that the use of sample data exceeds the use of training data in the problems of extracting rules

One of the first methods that followed this strategy was the Trepan algorithm [14]. It works very much like «divide and conquer» algorithm C4.5 [23] by searching for split points on training data for individual instances of different classes. The main differences from divide and conquer are the best strategy for expanding the tree structure, additional split points and the ability to choose additional learning examples at deeper points of the tree. As a result, the algorithm also creates a decision tree, which, however, can be transformed into a set of rules, if necessary.

Another of these very general pedagogical approaches that use sampling to extract rules from the neural network is presented in [13]. The algorithm, called Binarized Input-Output Rule Extraction (BIO-RE), is capable of processing only NN with binary or binarized input attributes. BIO-RE creates all possible input combinations and requests them from the neural network. Using the NN output, a truth table is created for each example. From the truth table, it is just as easy to go to the rules, if necessary.

ANN-DT is another decision-based sampling method for describing the behavior of a neural network [14]. The overall algorithm is based on CART with some variations in the initial implementation. ANN-DT uses the sampling method to expand the training set so that most of the training sample is still representative. This is “achieved using the nearest neighbor method, in which the distance from the sample point to the nearest point in the training data set is calculated” [14] and compared with the reference value.

The idea of creating a large set of instances at the first stage is also implemented by the STARE algorithm [15]. Like BIO-RE, STARE also forms extensive truth tables for learning. The advantage of STARE is its ability not only to handle binary and discrete attributes, but also to work with continuous input data. For the formation of truth tables, the algorithm rearranges the input data, while for each continuous attribute "it is necessary to sample it over the entire range of values with a high frequency".

The example of pedagogical approach using a sample of educational data that we want to briefly present here is KDRuleEx [4]. Like Trepan, the algorithm also generates additional learning cases where the basis for the following separation points is too small. KDRuleEx uses a genetic algorithm to create new training examples. The technique leads to a decision table that can be converted, for example, into IF-THEN rules, if desired.

Eclectic approach are the methods for extracting rules include elements of both pedagogical and decompositional [3]. In particular, eclectic approaches use knowledge of the internal architecture and weight vectors in the neural network to complement the symbolic learning algorithm [5].

The fast retrieval of rules from a neural network approach includes the FERNN approach, which first tries to identify the corresponding hidden neurons, as well as the corresponding inputs to the network. For this step, a decision tree is constructed using the well-known algorithm C4.5. The rule extraction process leads to the generation of M-of-N and IF-THEN rules. Having a set of properly classified teaching examples, FERNN analyzes the activation values of each hidden unit. For each hidden unit, activation values are sorted in ascending order. Then use the C4.5 algorithm to find the best split point to form the decision tree.

3 Extracting rules from deep neural networks and neuro-fuzzy networks

The most interesting from the point of view of this study is the extraction of rules using neuro-fuzzy models. Systems based on fuzzy rules (FRBS), developed using fuzzy logic, have become a field of active research over the past few years. These algorithms have proven their strengths in tasks such as managing complex systems, creating fuzzy controls. The relationship between both worlds (ANN and FRBS) has been carefully studied and equivalence results were obtained [17]. This fact gives two immediate and important conclusions. First, we can apply what was discovered for one of the models to the other. Secondly, we can translate the knowledge embedded in the neural network into a more cognitively acceptable language - fuzzy rules. In other words, we get a clear interpretation of neural networks [18-20].

Since 2012, the revolution of deep learning networks began . Consider one of the first and probably the most cited works in convolutional NN- Alexnet - it has 7 hidden layers, 650,000 neurons, 60,000,000 parameters. She was the first champion in pattern recognition and studied at 2 GPU for 1 week. Where did we get enough images to train her?

In 2010 dataset Imagenet with 15000000 images has appeared. The emergence of Imagenet brought the learning of neural networks to a whole new level. Parallel rapidly developed computing power, which led computer vision to the kind that we know and love it now. Since 2010, the annual Imagenet competition has also been held, where for the first time in 2012, the Alexnet convolutional neural network won and, since then, the deep networks has not lost its positions. The last winner, the National Assembly presented by scientists from China, contained 269 layers.

In order to get   the semantic interpretation of deep learning blackbox neuro-fuzzy networks can be used instead of the last complete connection layer. For example, ANFIS (adaptive neuro-fuzzy system) is a multilayer network of forward propagation. This architecture has five layers, such as a fuzzy layer, a product layer, a normalized layer, a defuzzification layer, and a common output. Fixed nodes are represented by a circle, and nodes represented by a square are adapted nodes. ANFIS brings the benefits of a mix network and fuzzy logic.

The aim of mixing fuzzy logic and neural networks is to design an architecture, which uses a fuzzy logic to show knowledge in fantastic way, while the learning nature of neural network to maximize its parameters. ANFIS put forward by Jang in 1993 integrate the advantages of both neural network and fuzzy systems, which not only have good learning capability, but can be interpreted easily als. ANFIS has been used in many applications in many areas, such as function approximation, intelligent control and time series prediction [26].

Deep learning networks and neuro-fuzzy networks can be mixed by different ways A hypothetical system can be created using two components [24]. The first is deep learning feature generation which can be used to create representative features from text directly. The deep learning system would initially be trained on unlabeled data. Once these features are extracted from the deep learning system, they will be integrated into fuzzy-inference systems. These systems can incorporate both the features detected from the deep learning as well as subjective information from an analyst. These two pieces together can be used for classification purposes. The final system would therefore be able to report both classification results and the specific features and rules that were activated for the system to arrive at its conclusion. Additionally, the final system could be further biased by an analyst as a form of feedback.

Very interesting approach is suggested in [22-23], where the author established a fundamental connection between two important fields in artificial intelligence i.e. deep learning and fuzzy logic. He shows, how deep learning could benefit from the comparative research by re-examining many trail-and-error heuristics in the lens of fuzzy logic, and consequently, distilling the essential ingredients with rigorous foundations. The author proposed deep generalized hamming network (GHN) as such not only lends itself to rigorous analysis and interpretation within the fuzzy logic theory but also demonstrates fast learning speed, well-controlled behavior and state-of-the-art performances on a variety of learning tasks. In [24] it is presented another approach for incorporating such rule-based methodology into neural networks by embedding fuzzy inference systems into deep learning networks.

Thanks to the theory of fuzzy sets, using fuzzy relationships and rules, you can create an effective model for predicting time series with a large number of inputs and one output (forecast). Such an approach allows us to make a kind of justification for the operation of an artificial neural network using neural-fuzzy models on the one hand and fuzzy cognitive maps on the other. We have developed a hybrid modular forecasting model that combines the theory of fuzzy logic, cognitive maps and artificial neural networks.

The modular system as a whole consists of several specialized modules. In general, these modules have the following characteristics:

1. System modules are specific and have specialized computing architectures to recognize and respond to specific subtasks of a large common task.

2. Each module, as a rule, is independent of other modules in its functioning and does not affect other operation of other modules.

3. Modules have a simpler architecture compared to the system as a whole. Thus, the module is faster than a complex monolithic system.

4. The results of each module individually are combined using a special integration module (in our case, the forecast consensus module), due to which the highest forecast accuracy of the entire system is achieved.

The system has three main modules responsible for the forecasting task. The ANFIS neuro-fuzzy network performs a time series forecast based on numerical indicators and gives us the so-called quantitative forecast, the results of which pass through a verification system (assessment of the adequacy of the forecast), if the forecast corresponds to the necessary accuracy, then it is transmitted to the next module. In parallel with the neuro-fuzzy network, a module with a fuzzy cognitive map is working, which receives data on the event effect on the time series as an input, a cognitive map is constructed in which all factors of influence on a specific predicted indicator are taken into account. At the output, the cognitive map gives us a forecast with the probability of its fulfillment, that is, with the consonance of a factor that tells us whether the forecast will be fulfilled or not. Further, all the data received from these modules is sent to the third module, which operates on the basis of the ANFIS network, which aggregates the information received from the previous modules and gives the final consensus forecast. In Fig. 1 presents a model of a forecasting system.

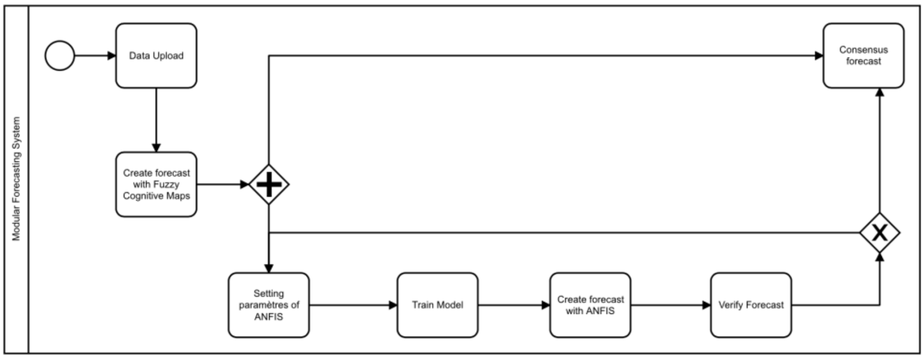
Modular Forecasting System
Data Upload
Consensus forecast
Create forecast with Fuzzy Cognitive Maps
Setting paramètres of ANFIS
Train Model
Create forecast with ANFIS
Verify Forecast

Fig.1. Modular forecasting system

4 Conclusion

This paper attempts to provide a review of several rule extraction algorithms from an artificial neural network. Some of the state-of-the-art algorithms are discussed from each category named as decompositional, pedagogical, and eclectics. Currently, deep learning provides an acceptable solution for lots of problems. It is a new machine learning area which is believed to move machine learning a step ahead. But it is still a black box system. Only several works try to established a fundamental connection between two important fields in artificial intelligence i.e. deep learning and fuzzy logic.

The study of fuzzy logic culminated in the end of the 20th century, and since then has begun to decrease [27]. This decrease can be partly attributed to the lack of results in machine learning. Extracting rules is one way to help understand neural networks. These studies will pave the way for fuzzy logic researchers to develop applications in artificial intelligence and solve complex problems that are also of interest to the machine learning community. Experience and knowledge in the field of fuzzy logic are well suited for modeling ambiguities in big data, modeling uncertainty in the representation of knowledge and providing transmission training with non-inductive inference, etc.

References

1. Craven, M. and Shavlik, J. W. (1994). Using sampling and queries to extract rules from trained neural networks. In ICML, pages 37–45.
2. Johansson, U., Lofstrom, T., Konig, R., Sonstrod, C., and Niklasson, L. (2006).Rule extraction from opaque models–a slightly different perspective. In Machine Learning and Applications, 2006. ICMLA'06. 5th International Conference on, pages 22–27.
3. Craven, M. and Shavlik, J. (1999). Rule extraction: Where do we go from here. University of Wisconsin Machine Learning Research Group Working Paper, pages 99–108.
4. Sethi, K. K., Mishra, D. K., and Mishra, B. (2012b). KDRuleEx: A novel approach for enhancing user comprehensibility using rule extraction. In Intelligent Systems, Modelling and Simulation (ISMS), 2012 Third International Conference, pages 55–60.
5. Andrews, R., Diederich, J., and Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based systems, 8(6):373–389.
6. Craven, M. W. (1996).Extracting comprehensible models from trained neural networks. PhD thesis, University of Wisconsin-Madison.
7. Thrun, S. (1993). Extracting provably correct rules from artificial neural networks. Technical report, University of Bonn, Institut für Informatik III.
8. Fu, L. (1994). Rule generation from neural networks. Systems, Man and Cybernetics, IEEE Transactions on, 24(8):1114–1124.
9. Tsukimoto, H. (2000). Extracting rules from trained neural networks. Neural Net-works, IEEE Transactions on, 11(2):377–389.
10. Sato, M. and Tsukimoto, H. (2001). Rule extraction from neural networks via decision tree induction. In Neural Networks, 2001. Proceedings. IJCNN'01. Iternational Joint Conference on, volume 3, pages 1870–1875.
11. Tickle, A. B., Andrews, R., Golea, M., and Diederich, J. (1998). The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. IEEE Transactions on Neural Networks, 9(6):1057–1068.
12. Thrun, S. (1995). Extracting rules from artificial neural networks with distributed representations. Advances in neural information processing systems, pages 505–512.
13. Craven, M. W. and Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. Advances in neural information processing systems, pages 24–30.

14. Taha, I. A. and Ghosh, J. (1999). Symbolic interpretation of artificial neural networks. Knowledge and Data Engineering, IEEE Transactions on, 11(3):448–463.
15. Towell, G. G. and Shavlik, J. W. (1993). Extracting refined rules from knowledge-based neural networks. Machine learning, 13(1):71–101.
16. Setiono, R. and Leow, W. K. (2000). FERNN: An algorithm for fast extraction of rules from neural networks. Applied Intelligence, 12(1-2):15–25.
17. Averkin A., Yarushev S. Hybrid Neural Networks and Time Series Forecasting. Artificial Intelligence. Communication in Computer and Information Sciences 934 – Springer, 2018 – pp.230-239
18. Giovanni Pilato, Sergey A. Yarushev, and Alexey N. Averkin Prediction and Detection of User Emotions Based on Neuro-Fuzzy Neural Networks in Social Networks // Proceedings of the Third International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’18) Volume 2 - Springer, . Advances in Intelligent Systems and Computing, volume 875. pp. 118-26
19. Averkin, G Pilato and S A Yarushev An Approach for Prediction of User Emotions Based on ANFIS in Social Networks A N // Second International Scientific and Practical Conference Fuzzy Technologies in the Industry , FTI 2018– CEUR Workshop Proceedings, 2018) , pp. 126-134
20. Jan Ruben Zilke, Eneldo Loza Mencía and Frederik Janssen, DeepRED -- Rule Extraction from Deep Neural Networks, in: Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19--21, 2016, Proceedings, pages 457--473, Springer International Publishing, 2016
21. L.Fan. "Revisit Fuzzy Neural Network: Demystifying Batch Normalization and ReLU with Generalized Hamming Network", NIPS 2017.
22. Quinlan, J. R. (1993). C4.5: Programs for Machine Learning, volume 1. Morgan Kaufmann.
23. David Bonanno, Kristen Nock, Leslie Smith, Paul Elmore, Fred Petry, "An approach to explainable deep learning using fuzzy inference," Proc. SPIE 10207, Next-Generation Analyst V, 102070D , 2017.
24. S. R. Jang, “ANFIS: Adaptive-network-based fuzzy inference systems”. IEEE Trans. On Systems, Man, and Cybernetics. Vol. 23, 1992. pp . 665-685.