FUNCTIONAL RULE EXTRACTION METHOD FOR ARTIFICIAL NEURAL NETWORKS

Caleb Princewill Nwokocha

Department of Computer Science, The University of Manitoba, Winnipeg, MB, Canada

nwokochc@myumanitoba.ca

ABSTRACT

The idea I propose in this paper is a method that is based on comprehensive functions for directed and undirected rule extraction from artificial neural network operations. Firstly, I defined comprehensive functions, then constructed a comprehensive multilayer network (denoted as $\mathbb{N}$*). Each activation function of* $\mathbb{N}$ *is parametrized to a comprehensive function. Following* $\mathbb{N}$ *construction, I extracted rules from the network by observing that the network output depends on probabilities of composite functions that are comprehensive functions. This functional rule extraction method applies to the perceptron and multilayer neural network. For any* $\mathbb{N}$ *model that is trained to predict some outcome given some event, that model behaviour can be expressed – using the functional rule extraction method – as a formal rule or informal rule obeyed by the network to predict that outcome. As example, figure 1 consist of a comprehensive physics function that is parameter for one of the network hidden activation functions. Using the functional rule extraction method, I deduced that the comprehensive multilayer network prediction depends on probability of that physics function and probabilities of other composite comprehensive functions in* $\mathbb{N}$*. Additionally, functional rule extraction method can aid in applied settings for generation of equations of observed phenomena. This generation can be achieved by first training an* $\mathbb{N}$ *model toward predicting outcome of a phenomenon, then extracting the rules and assuming that probability values of the network comprehensive functions are constants. Finally, to simplify the generated equation, comprehensive functions with probability* $p \sim 0$ *may be omitted.*

KEYWORDS

Rule extraction, comprehensive functions, neural network

1. INTRODUCTION

For decades, artificial neural network has been the most useful pattern recognition technique in various science and engineering areas like machine learning, computer vision, natural language processing, robotics, etcetera [3, 1, 2]. Real world applications of this technique span vast number of industries, with increasing utilization across domains [11, 10, 7]. This immense utilization of artificial neural network stem from explosion of data and predictive power of models.

A standard model consist of input, activation functions, hyperparameters, and output. The input include train set and testing set, activation functions may be sigmoid, rectified linear unit, or tanh, [4] hyperparameters are used to control activation functions, and output is the model prediction.

The model undergo a training procedure that involve optimizing hyperparameter values until the model learn to output accurate predictions when given further training set as input. After the training procedure, the model is tested with a testing set to measure generalization capacity of the model. If it past the test, the model is fully ready for use by its user.

Although, this model may serve its user adequately, the user lack concrete understanding of any rule obeyed by the model to predict an output. [9] This lack of concrete understanding of the model rules resulted to the term "black-box", meaning values of activation functions and hyperparameters are only numbers and they have no comprehensive relation to the probability distribution that the model learned. Without extracted rules, a user cannot prove that a model will make a right or wrong prediction at any given time. Before safely deploying a network model in a high-risk environment such as surgical robot in an operating room, autonomous vehicles on road highways, bots on trading system, etcetera, a user should be able to extract rules of the network model.

The functional rule extraction method explained in the following sections, is an efficient method for extracting rules from an N model consisting comprehensive functions. Rules extracted by this method can also be interpreted as equations of observed phenomena.

2. Related Work

3. Comprehensive Function

A solution to problem of rule extraction is forming a comprehensive relationship between the learned probability distribution and the input. I will use the sigmoid function $S(X)$ to form a comprehensive relationship, although other activation function can be used.

$$S(X) = \frac{1}{1 + e^{-X}} = (1 + \ e^{-X})^{-1}$$

A **comprehensive function** is a function that has explanatory relationship within a set of composite functions.

3.1. Univariate Comprehensive Function

A **univariate comprehensive function** f_c (for the purpose of this paper) is a comprehensive function with parameter wx, such that $f_c: wx \to \mathbb{R}$.

Let X be a univariate comprehensive function $f_c(\ wx)$, then

$$S(f_c(wx)) = \left(1 + \ e^{-f_c(wx)}\right)^{-1}$$

Some examples of univariate comprehensive function are:

1. Volume of a cube $V(s) = f_c(s) = s^3$
2. Trigonometry identity $T(a) = f_c(a) = \sin^2 a + \cos^2 a$

3.2. Multivariate Comprehensive Function

A **multivariate comprehensive function** f_c (for the purpose of this paper) is a comprehensive function with parameters $w_1x_1, \cdots, w_nx_n$, such that $f_c: w_1x_1, \cdots, w_nx_n \to \mathbb{R}$.

Some examples of multivariate comprehensive functions are:

1. Force $F(m, a) = f_c(m, a) = m \cdot a$
2. Quadratic formula $Q(a, b, c) = f_c(a, b, c) = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \left(-b \pm \sqrt{b^2 - 4ac}\right)(2a)^{-1}$

By definitions of comprehensive function, I deduce that from example 1 of the univariate case, $f_c(s) = f_c(wx)$ if and only if $s = wx$. This means that the activation function

$$S(f_c(wx)) = \left(1 + e^{-f_c(wx)}\right)^{-1}$$

can be treated as

$$S\big(V(wx)\big) = \left(1 + e^{-[(wx)^3]}\right)^{-1} iff\ V(s) = V(wx)$$

Likewise, the second univariate example is treated as

$$S\big(T(wx)\big) = \left(1 + e^{-\sin^2 wx - \cos^2 wx}\right)^{-1} iff\ T(a) = T(wx)$$

For the multivariate case, the comprehensive force sigmoid activation function is written as

$$S\big(F(w_1x_1,\ w_2x_2)\big) = \left(1 + e^{-(w_1x_1w_2x_2)}\right)^{-1}$$
$$iff\ F(m, a) = F(w_1x_1,\ w_2x_2)$$

and the comprehensive quadratic sigmoid activation function is written as

$$S\big(Q(w_1x_1,\ w_2x_2,\ w_3x_3)\big) = \left(1 + e^{\left(w_2x_2 \pm \sqrt{(w_2x_2)^2 - 4w_1x_1w_3x_3}\right)(2w_1x_1)^{-1}}\right)^{-1}$$
$$iff\ Q(a, b, c) = Q(w_1x_1,\ w_2x_2,\ w_3x_3)$$

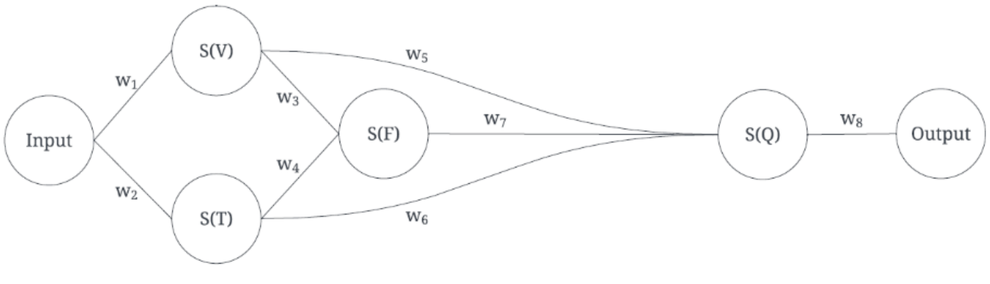

Figure 1: Comprehensive multilayer artificial neural network

4. RULE EXTRACTION

[2]According to M.Sato and H. Tsukimoto, rule extraction from neural networks is the task for obtaining comprehensible descriptions that approximate the predictive behavior of neural networks. I will extract the rule from the network model in figure 1. Firstly, I describe the model as a comprehensive multilayer artificial neural network because its hidden unit are activated by comprehensive sigmoid functions introduced in earlier examples. The comprehensive model is made up of six units: one input unit i, one output unit o, and four hidden units. The output unit has a standard sigmoid function. Algebraically, this model output can be expressed as the following:

$$o\left(S\left(Q\left(S(V(iw_1))w_5, S\left(F(S(V(iw_1))w_3, S(T(iw_2))w_4)\right)w_7, S(T(iw_2))w_6\right)\right)w_8\right)$$

$$= \left(1+e^{-w_8\left(1+e^{\left(w_7\left(1+e^{-w_3w_4\left(1+e^{-iw_1}\right)^{-1}\left(1+e^{-iw_2}\right)^{-1}}\right)^{-1} \pm \sqrt{w_7\left(1+e^{-\left(w_3w_4\left(1+e^{-iw_1}\right)^{-1}\left(1+e^{-iw_2}\right)^{-1}\right)^2}\right)^{-1} - 4w_5\left(1+e^{-iw_1}\right)^{-1}w_6\left(1+e^{-iw_2}\right)^{-1}}\right)\left(w_5\left(2+2e^{-iw_1}\right)^{-1}\right)^{-1}}\right)^{-1}}\right)^{-1}$$

To extract rule from the model, I extract the composites $r_i \in r, \forall i : 1 \leq i \leq 4$ of comprehensive functions (together with their weights) from the above output expression:

$$r_1 = w_8\left(-i^2w_1w_2w_3w_4w_7 \pm \sqrt{(i^2w_1w_2w_3w_4)^2w_7 - 4i^2w_1w_2w_5w_6}\right)(iw_1w_5)^{-1}$$

$$r_2 = w_8\left(-VTw_3w_4w_7 \pm \sqrt{(VTw_3w_4)^2w_7 - 4VTw_5w_6}\right)(Vw_5)^{-1}$$

$$r_3 = w_8\left(-Fw_7 \pm \sqrt{F^2w_7 - 4VTw_5w_6}\right)(Vw_5)^{-1}$$

$$r_4 = Qw_8$$

The above compositions are **formal rules** of the model in figure 1. An **informal rule** is worded statement of a formal rule. As shown, r_1 is a much complex formal rule than r_2, r_3, and r_4. This complexity in r_1 result from explicitly stating complete weight distribution across the input and comprehensive functions. Even probabilistic $w_1, \cdots, w_8$ may not sum to 1 for simple explanation of the r_1. Hence, I normalize $w_1, \cdots, w_8$ sum to 1 by using softmax σ equation on all weights and mapping $\sigma(w_i)$ to p_i. Here p_i represent associated probabilit(ies) to an input or comprehensive function.

$$p_i \leftarrow \sigma(w_i) = \frac{e^{w_i}}{\sum_{i=1}^{n} e^{w_i}}, \forall i: 1 \leq i \leq 8$$

Then the formal rules $r_1, \cdots, r_4$ can be noted as **normal formal rules** below.

$$r_1 = p_8\left(-i^2p_1p_2p_3p_4p_7 \pm \sqrt{(i^2p_1p_2p_3p_4)^2w_7 - 4i^2p_1p_2p_5p_6}\right)(ip_1p_5)^{-1}$$

$$r_2 = p_8 \left(-VTp_3p_4p_7 \pm \sqrt{(VTp_3p_4)^2 w_7 - 4VTp_5p_6} \right) (Vp_5)^{-1}$$

$$r_3 = p_8 \left(-Fp_7 \pm \sqrt{F^2 p_7 - 4VTp_5p_6} \right) (Vp_5)^{-1}$$

$$r_4 = Qp_8$$

And a **normal informal rule** of r_3 can be stated as “The output of the network is based on rule that the probability p_8 of quadratic relationship between negative force at probability p_7, square root of squared positive force at probability p_7, square root of four negative cubic volume and trigonometry identity products at probability p_7p_6, and single cubic volume at probability p_5, is true.”

[8] Furthermore, rules can either be directed or undirected. A **directed rule** is a rule that is obtained from a directed neural graph model, while an **undirected rule** is a rule obtained from undirected neural graph model. Figure 1 is an example of a directed neural graph model.

3. CONCLUSIONS

Papers in this format must not exceed twenty (20) pages in length. Papers should be submitted to the secretary AIRCC. Papers for initial consideration may be submitted in either .doc or .pdf format. Final, camera-ready versions should take into account referees’ suggested amendments.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] P. Li, J. Li and G. Wang, "Application of Convolutional Neural Network in Natural Language Processing," 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2018, pp. 120-122, doi: 10.1109/ICCWAMTIP.2018.8632576.

[2] B. Raducanu and M. Grana, "Morphological neural networks for robust visual processing in mobile robotics," Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, pp. 140-143 vol.6, doi: 10.1109/IJCNN.2000.859387.

[3] E. Nishani and B. Çiço, "Computer vision approaches based on deep learning and neural networks: Deep neural networks for video analysis of human pose estimation," 2017 6th Mediterranean Conference on Embedded Computing (MECO), 2017, pp. 1-4, doi: 10.1109/MECO.2017.7977207.

[4] Wikipedia contributors. (2022, July 27). Hyperparameter (machine learning). In *Wikipedia, The Free Encyclopedia*. Retrieved 02:31, July 28, 2022, from https://en.wikipedia.org/w/index.php?title=Hyperparameter_(machine_learning)&oldid=1100733664

[5] Goodfellow, I., Bengio, Y.,, Courville, A. (2016). Machine Learning Basics. In *Deep Learning*. MIT Press. ISBN: 9780262035613.

[6] M. Sato and H. Tsukimoto, "Rule extraction from neural networks via decision tree induction," IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), 2001, pp. 1870-1875 vol.3, doi: 10.1109/IJCNN.2001.938448.

[7] Chao Deng, Fanlun Xiong, Ying Tan and Zhenya He, "Sequential learning neural network and its application in agriculture," 1998 IEEE International Joint Conference on Neural Networks

Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227), 1998, pp. 221-225 vol.1, doi: 10.1109/IJCNN.1998.682266.

[8] Grimaldi, R. P. (2004). An Introduction to Graph Theory. In *Discrete and Combinatorial Mathematics: an Applied Introduction*. Boston: Pearson Addison Wesley. ISBN: 0201726343 9780201726343.

[9] Z. F. Wu, J. Li, M. Y. Cai, Y. Lin and W. J. Zhang, "On membership of black-box or white-box of artificial neural network models," 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), 2016, pp. 1400-1404, doi: 10.1109/ICIEA.2016.7603804.

[10] Hong Zhang, Zhen Zhang and A. W. Partin, "Neural network based systems for prostate cancer stage prediction," Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, pp. 659-662 vol.3, doi: 10.1109/IJCNN.2000.861399.

[11] L. Zhong, C. Qi and Y. Gao, "Construction of Continuing Education Teaching Quality Evaluation System Based on BP Neural Network," 2021 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE), 2021, pp. 348-351, doi: 10.1109/ISAIEE55071.2021.00091.

Author

Short Biography

Photo