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 based on comprehensive function for rule extraction from artificial neural network (ANN) operations. This method applies to the classical and multi-layer neural network – also known as deep learning – models.

KEYWORDS

Rule extraction, comprehensive functions, neural network

1. INTRODUCTION

In traditional ANN models, activation functions may be sigmoid, rectified linear unit, or tanh. With any of these activation functions for all units in the network, extracting rules of the resulting model representation is never easy. [1]This difficulty is due to the fact that sigmoid, rectified linear unit, and tanh has no comprehensive relation to the probability distribution that a model learns. Instead, these activation functions act as regularizes for desired units outputs.

2. COMPREHENSIVE FUNCTION

A solution to problem of rule extraction is forming a comprehensive relationship between the learned probability distribution and the dataset. 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.

2.1. Univariate Comprehensive Function

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

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 $T(a) = f_c(a) = \sin^2 a + \cos^2 a$

2.2. 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

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

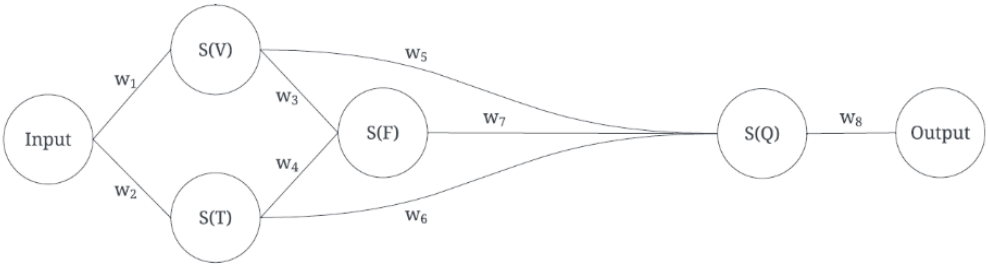

Figure 1: Comprehensive multi-layer artificial neural network

3. 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 multi-layer 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:

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

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

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."

[3]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.

REFERENCES

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

[2] 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.

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