

神经网络解释机制*)

Explanation Mechanism in Neural Networks

贵忠华 刘振凯 严新民

(西北工业大学 西安710072)

摘 要 Neural Networks(NNs) have no explicit, declarative knowledge representation and therefore have considerable difficulties in generating explanation. We think that additional structures and algorithms are needed for generating explanation in NNs. This paper discusses the necessity of explanation in NNs, analyses the explanation methods which have been published, and proposed a case/rule-based explanation method for NNs.

关键词 Neural networks, Explanation, Inference, Knowledge representation, Case

解释就是人工智能系统对用户提问给出一个清晰的、完全的和易于理解的回答,即对其行为作出合理说明、使系统具有透明性。解释的类型虽然很有限,但对于用户是否接受一个推理系统却是绝对关键的。专家系统的应用表明,如果没有任何解释的话,势必影响用户对结论的可信度,降低系统的可接受性。因此,在专家系统研究中,人们作出了很多努力来允许产生解释,且这些解释至少在一定程度上是合情合理的。

神经网络可用于知识表示和推理,但是在本质上神经网络需要解释吗?一个极端的情况是完全信任的训练。例如,在神经网络的学习过程中,当能得到很大的一个训练样本集合,有一个非常有效的训练方法就足够了,可以不需要解释。这时网络能够容易地识别以前训练过的模式和不在训练样本集合里的一些新的实例(即泛化)。然而,对现实世界的应用而言,自然环境是不断变化的,推理者必须能处理不完全和不一致的信息;学习可能被中断;在一个适当的时间间隔里必须完成训练;甚至一个完美的训练过程也不影响用户对解释的需要,由于解释已成为任何推理系统中的一个关键功能。因此,解释机制在

神经网络中同样也是必备的。

人们已尝试性地提出了一些实现神经网络解释机制的方法,这些方法可分为下面三种:(1)采用基于规则的专家系统进行解释;(2)从神经网络中提取用于解释的规则;(3)基于实例的解释方法,即“举例说明”。下面将对这些方法的优缺点进行分析,并提出了一种基于实例/规则的解释方法。

1 采用基于规则的专家系统进行解释

这是 Caudill 于1991年提出的一种方法^[2],其基本思想是构造一个附加的、基于规则的专家系统用于辅助神经网络的解释,如图1所示。当用户要求解

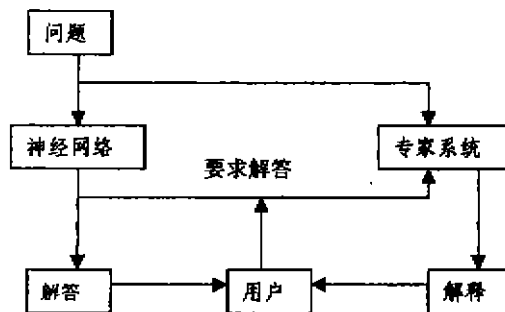


图1 用专家系统辅助神经网络的解释

*)国家自然科学基金资助。

该树可从B处分成一棵因果树(右半部分)和一棵倒因果树(左半部分)。

因果树中的传播体制与因果链中的传播体制是相似的,但它又具有树自己本身的特点:每个结点与以它为根的子树中的每个子结点有因果关系。象图1中的交叉因果链被分成两棵因果树后,应注意考虑它们的联系。

参 考 文 献

- [1]涂象初等,集合搜索的一些命题,科学通报,1988,33(12),957
- [2]涂象初等,集合搜索作为具有布尔代数元的 IF-THEN 结构,科学通报,34(9)1989
- [3]张尧庭等,专家系统中不确定性方法学,广西师范大学出版社,1994,6版

释时,将神经网络的输出和输入同时传给一个简单的专家系统,这个专家系统不具备解决问题的能力,但它能从神经网络的解答和输入进行反向推理,提供如何得到解答的解释。

这种方法的难点是如何开发这个简单的、不具备解决问题能力而又能进行解释的专家系统。因此,这种方法在原理上是可行的,而实际上却很少使用。

2 从神经网络中提取用于解释的规则

Gallant 的联接专家系统(Connectonist expert system)^[3]及 Satio 和 Nakano 的基于 PDP 模型的医疗诊断专家系统^[4]都采用从神经网络中提取规则知识用于解释。这种方法是思想是通过算法将神经网络的隐式、分布式的知识,转化为显式的规则知识,从而对结论进行解释。因此它可以采用现有的一些规则提取算法^[5],不过用于解释时,只需对特定的结论提取相应的规则知识,而无须提取神经网络所包含的全部规则知识。下面介绍 Gallant 在联接专家系统中所使用的解释方法。

假设用户要求解释某一结点为何取“真”或“假”值。系统则用 IF-THEN 规则来回答这一问题,但这些规则不是在知识库中显式表示的,而是当需要解释时由特定的算法产生。

图2说明规则是如何产生的,其基本思想是:设由推理得知 u_7 的值为“真”,则应设法找到一个获得此结果的已知信息的最小子集。算法的过程如下:

第1步,将所有对 u_7 最终获得真值的输入结点列表,在图2的例子中,它们是 u_1, u_2, u_5 。因为 $u_4 < 0, w_{57} > 0$, 而 u_7 已被定为“真”,故 u_4 被删去。

第2步,按权值绝对值的递减顺序排序; u_2, u_5, u_1 。

第3步,由这一排序表不断产生 IF-THEN 规则的子句,直到满足

$$\left\| \sum_{u_i \text{ 已用于子句中}} w_{i7} \right\| > \left\| \sum_{u_i \text{ 未用于子句中}} w_{i7} \right\| + 0$$

这样就产生规则: IF u_2 为“假”且 u_5 为“真” THEN u_7 为“真”。

应该指出的是,对于结点 u_i 的任何规则仅涉及

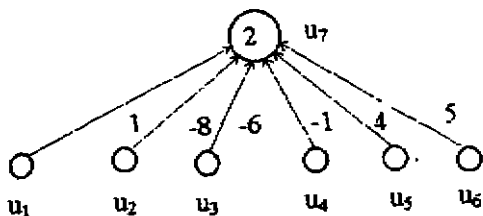


图2 提取 IF-THEN 规则的一个例子

与 u_i 相连接的结点。这种方法主要适用于定性的问题(即神经元的状态值“0”对应于“假”,而“1”对应于“真”),而不适应于定量的问题(即神经元的状态值为连续值)。对于这个局限性,Narazaki 针对数据分类问题提出了一种解决方法^[6],他的方法是对神经网络学习到的映射函数进行分析从而提取相应的规则,而不是直接分析权值矩阵,但是这种方法的有效性尚待进一步研究。

3 基于实例的解释方法

这是我国学者王珏研究员和戴汝为院士于1990年提出的一种方法^[7]。他们认为,对于神经网络推理结果的解释,使用传统的“基于踪迹”的技术所作出的解释是不能令用户满意的,而“举例说明”是人对问题解释的一个重要方法,因此他们提出了“基于实例”的解释方法。下面针对分类问题说明该方法的思路。

3.1 神经网络被用于分类问题时的几何解释

如果神经网络的输入模式有 n 个特征(features),而输出模式有 k 个类别,那么可将其理解为:神经网络算法将一个 n 维欧氏空间(记为 I'),映射到一个 k 维欧氏空间(记为 O')。

令 $S = \{s_i\}$ 表示具有 q 个样本的训练集, $i=1, 2, \dots, q$, I 表示与样本 s_i 对应的一个 n 维特征矢量(即输入矢量), O 表示与样本 s_i 对应的一个 k 维特征矢量(即输出矢量),则

$$S = \{(I_i, O_i)\} \quad (1)$$

经过学习之后, q 个在训练集中的样本根据它们的输入矢量被分布在上述 n 维欧氏空间 I' 中,而这些样本的输出矢量将分布在 k 维欧氏空间 O' 中。

在推理过程中,一个新的实例 s^* 将被确定在上述 n 维空间中,然后,它将被公式(1)映射在 O' 空间中,根据导出规则的计算可知它应属于哪个类。这就导致一个问题:是否可以利用在 I' 及 O' 空间中的距离找出一个在训练集中与 s^* 距离最近的样本作为解释呢?显然这种方法是行得通的。

3.2 “基于实例”的解释方法

令 s^* 是一个被求解的实例,它的输入矢量 I_s^* ,经神经网络计算后的输出矢量记为 O_s^* , S 是一个训练集(见式(1))。

再令 $\{O\}$ 是 S 中样本的输出矢量集合, $\{s_{min}\} \subset S$ 是在欧氏空间 O' 中 O_s^* 与 $O_i \in \{O\}$ 距离最小的样本集, $j=1, 2, \dots, q$, 则:

$$D_s^*(O_s^*) = \min_{j=1, \dots, q} \|O_s^* - O_j\|^{1/2} \quad (2)$$

其中 O_i 是 $\{s_{min}\}$ 中的一个样本的输出矢量(满足式(2)的 O_i 可有多)。

令 $\{I_{min}\}$ 是 $\{s_{min}\}$ 中所有输入矢量的集合,在 I' 中 I_s^* 与 $I_j \in \{I_{min}\}$ 距离最小的样本由下式决定:

$$D^*(I^*) = \min_i \|I_i^* - I_j\|^{1/2} \quad (3)$$

其中 I^* 是 $\{s_{mn}\}$ 中某个样本的输入矢量, 即 $(I^*, O^*) \in \{s_{mn}\}$ 。

令 P 是一个被表达成模式的训练集, p_i 是一个样本, $p_i \in P, i=1, 2, \dots, q$, 即:

$$P = \{p_i\} \quad (4)$$

则 p_i 可被表示成:

$$p_i = (I_i^1, O_i^1) \quad (5)$$

其中 I_i^1 与 O_i^1 为两个表, 分别表示了一个样本的输入模式及解。经导入规则与导出规则, p_i 被变换成矢量表达 $s_i \in S$ 。

$p^* \in P$ 是被求解的实例 s^* 的一个解释, 当且仅当与 p^* 对应的矢量表达式 (I^*, O^*) 满足式(2)和式(3)。

在实际系统中, “基于实例”的解释的基本规则可描述如下:

如果 $D^*(I^*) < h$

那么 p^* 为推理解释的实例

其中 h 是一个阈值。

“基于实例”的解释方法存在的问题在于, 欧氏距离的方法仍是假设已知的实例与被求解的问题需要使用同一空间中的同一测量标准, 这对很多情况是不适用的。在实际情况中, 举例说明往往是与问题情景有关。这意味着选择例子的时候, 不仅要考虑问题与某个例子距离的远近, 而且还要考虑是否与某些具体特征有关, 换句话说, 还要对某些因素有所侧重。

4 基于实例/规则的解释方法

基于实例的解释方法也可以从分类问题推广到一般的非线性映射问题。但是, 由于训练样本数往往只占输入/输出模式空间的很小一部分, 因此可能会出现 $D^*(I^*) \geq h$ 的情况。这时有两种选择, 一种是不把 p^* 作为解释的实例, 即不提供解释; 另一种是仍把 p^* 作为解释的实例, 但是由于 p^* 和 p_i 的欧氏距离超过了阈值, 能否被用户接受就成了一个问题。

针对这个问题, 本文提出了一种基于实例/规则的解释方法, 该方法将基于实例的解释与基于规则的解释相结合, 具体描述如下:

如果 $D^*(I^*) < h$

那么 p^* 为推理解释的实例

否则 搜索解释规则集, 使用 p^* 和匹配的规则作为解释

其中“解释规则集”是一个事先建立好的、用于对实例 p^* 进一步解释的规则集, 利用它可以将 p^* 和 p_i 作比较, 激活相应的解释规则。

下面以一个具体实例说明该方法。在电机设计中, 电磁负荷选择的经验知识很难用规则形式表示,

作者在文[8]中提出了用神经网络表示这类经验知识。但是如何解释由神经网络得到的电磁负荷呢? 作者认为, 可以采用这种基于实例/规则的解释方法。

假设要确定一台功率为90KW、同步转速为3000转/分的小型三相异步电动机(IP44)的电磁负荷, 通过神经网络计算得到线负荷为 $313 \times 10^2 \text{ A/m}$, 气隙磁密为6600高斯, 记为:

$$P_i = (90, 3000, 313, 6600)$$

对于该问题的解释, 首先找出欧氏距离最小的实例为:

$$p^* = (75, 3000, 296, 6420)$$

此时有 $D^*(I^*) \geq h$, 那么搜索解释规则集, 匹配了下面的规则:

IF 同步转速相等 AND 功率较大 THEN 线负荷应较大

IF 同步转速相等 AND 功率较大 THEN 气隙磁密应较大

那么结合实例 P^* 和上面的规则, 就可以产生如下的解释:

功率为90KW, 同步转速为3000转/分的电机, 线负荷取 $313 \times 10^2 \text{ A/m}$, 气隙磁密取6600高斯, 是因为功率为75KW、同步转速为3000转/分的电机, 线负荷取 $296 \times 10^2 \text{ A/m}$, 气隙磁密取6420高斯, 同时由于同步转速相等且功率较大的电机, 线负荷和气隙磁密应较大。

这样的解释显然更有说服力, 这种方法也与人解释问题的方法相似, 即, 人在采用举例说明的同时, 有时也采用一些合情合理的规则解释。

解释规则集可根据具体问题构造, 其功能主要用于实例比较和对实例作进一步的解释。

参考文献

- [1] J. Diederich Explanation and artificial neural networks, Int. J. Man-Machine Studies, 37(5)1992
- [2] M. Caudill Expert networks, Byte 16(10)1991
- [3] S. I. Gallant, Connectionist expert systems, CACM, 31(2)1988
- [4] K. Satio and R. Nakano, Medical diagnostic expert system based on PDP model, Proc. of the IEEE Intl. Conf. on Neural Network, San Diego, 1988
- [5] L. M. Fu, Rule generation from neural networks, IEEE Trans. on Systems, Man, and Cybernetics, 24(8)1994
- [6] H. Narazaki et al., Reorganizing knowledge in neural networks: An explanation mechanism for neural networks in data classification problems, IEEE Trans. on Systems, Man, and Cybernetics, Part B, Cybernetics, 26(1)1996
- [7] 王珏、戴汝为, 建造带有人工神经网络知识系统的一种方法, 计算机学报, 13(5)1990
- [8] 刘振凯、蔡青, 基于人工神经网络的电机设计知识表示方法研究, 电工技术学报, 待发表