计算机科学 1994Vol. 21 №. 2

解释与智能系统*)

杨 莉 葛建新

(浙江大学计算机系 杭州 310027)

摘 要 In this paper, a brief overview of explanation in symbolic artificial intelligence is given firstly, and then various proposals for explanation in feedforward perceptrons and relaxation-type neural networks (NN) are discussed, and several successful implementations of explanation components in NN systems are presented. It is shown that the introduction of structure mechanism (e. g. the explicit encoding of relations and modular network architectures) to a network significantly facilitates the generation of explanation structures in NN.

1. 前言

解释在人工智能（AI）系统中是一个关键的功能。在基于事例的推理中，当一个预期的断言失败时，解释用于调整知识的结构，即基于错误驱动的学习；在某些情况下，用户不是领域专家，但用户有接受或拒绝由 AI 系统所产生结果的权力，这时解释用于使推理过程的结果对用户明朗化；当一个完整的领域理论（知识）被给予时，解释能用于知识密集型的学习。

解释在非结构化的神经网络（NN）里是一个难以实现的功能，NN 没有明显的、陈述式知识结构来允许解释结构的表示和产生，象推理路径、预期失败的解释等等。在 NN 中，知识被编码在网络的权值和门阈值里且分布在整个神经网络中，因而在 NN 系统中要实现解释部件必须做出几点假设（约定）。

本文重点讨论 NN 产生解释的能力。我们将会看到，为了能够容纳解释和解释部件（EC），神经网络系统趋于使用显式的关系编码和高度结构化的网络。

2. 符号人工智能里的解释

2.1 用户解释

在符号 AI 里，术语解释指一个显式结构。对于推理和学习，这种结构能在系统内部使用，对用户，用于结果的解释。例如，在基于规则的系统里，解释包括推理过程的中间步骤，即规则使用的次序，一个证明结构等等。这个结构能回答“how”问题，例如，结论“w”如何由推理系统得出？

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（IS）却是绝对关键的。专家系统（ES）的应用表明，用户需要 ES 对所产生的结果有一个解释，如果没有任何解释的话，用户不会接受结论[1]。因此，在 ES 领域里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有意义和合情合理的。

虽然，符号 AI 已经成功地实现了几种不同形式的用户解释，但无法保证一个解释是否透彻。例如，在一个组织上层次性很差的规则库里，每个规则有成百的前提，基于规则的跟踪将完全损坏解释的透彻性（合理性）。因此，基于规则跟踪的解释已普遍地被认为太严格、刻板和不灵活。基于规则跟踪的解释总是反映知识库（KB）的当前结构，经常指示到内部过程（例如计算），可能包括很多重复（例如，一个推理执行多次），且解释的粒度常常是错误的[2]。Moore 和 Swartout[3]指出，早期千篇一律的版本或模板作为用户解释部分的使用太刻板，以致系统总是以相同的方法解释问题，且缺乏回答（反应）策略。尽管在利用具有混合（多个）输入端的自然语言对话、用户模型和明显计划好的解释策略方面作出了很多努力，但人们似乎还是认为，当前的解释系统仍然不够灵活，不能应答，解释不连贯、不灵敏且太刻

*) 国家自然科学基金和浙江省自然科学基金资助项目，并得到浙江大学 CAD&CG 重点实验室的支持。

杨莉 工学博士，现为浙江大学计算机系博士后，主要从事人工智能、神经网络、智能 CAD、CAPP 等领域的研究工作。
葛建新 工学博士，主要从事人工智能、CAD、图形学、CIMS、用户界面等领域的研究工作。

板。

在基于事例的推理系统里，解释的另一种可能形式是：如果用户不能接受或理解一个结果，系统能选择一个相似的事例，并呈现给用户作为回答。

2.2 上下文处理里的解释

解释在上下文和故事情节理解里也起非常重要的作用，根据他或她的行动，解释能定义一种方法将一个刺激赋予一个字母，解释的这种形式已借助于广播刺激和标志传播方法来实现，即在语义网络中一个广度优先搜索用于发现概念间的联接，一个部件用于评估概念间的关系。在标记传播系统里，"is-a"层次结构里的对象由一个广播刺激过程索引且用于建立解释。

2.3 机器学习里的解释

学习是指调整一个或更多的知识结构，在学习里解释也起一个关键的作用，这时解释本身是一个显式结构。例如，在基于事例的推理中，一个关键功能是去解释预期的失败。总会发生这样的事情：当一个情形和以前的事例不吻合时，必须分类（划分）新的情形，预期结果和实际事件间的不同性将触发学习过程。

Mitchell 等使用了术语——基于解释的一般化（EBG），在符号人工智能里指示知识密集型的学习方法。EBG 是一个两步学习：首先，构造为什么一个例子符合一个特殊目标概念的解释，并寻找训练实例里能贴切地刻画这个目标概念的特性（属性）。第二步，决定例子的一般化，它对于目标概念来说是一个充分的概念定义且满足一个可操作性准则。然而，EBG 需要一个完备的领域理论，但一个先决条件对于现实世界应用是不切实际的，因此在使用更弱的领域模型方面人们作出了很多的努力。

3. 神经网络中的解释机制

神经网络在模式识别、自适应行为（包括机器学习和一般化）、数据合成，概率和似然推理等领域中已被证明是有用的计算方法，而且 NN 为开发大规模并行性和容错行为的处理提供了有效的处理方法。

然而，在本质上 NN 需要解释吗？一个极端情形是完全信任的训练。例如，在 NN 的学习过程中，特别是当能得到很大的一个训练事例集且网络是完备训练时，有一个非常有效的训练技术就足够了，不需要解释。这时网络能够容易地识别以前已经训练过的模式和不在训练集里的一些新的实例（一般化）。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，因为解释已成为任何推理系统的一个关键功能，解释机制在 NN 中同样也是必备的。

前言中已简述了解释机制在非结构化的 NN 里很难实现，人工神经系统没有显式、符号化和陈述式知识表示，知识用数值化的权矩阵进行表示且知识分布表示在整个网络中。当必须对用户解释一个结果时，上述 NN 的知识表示方法限制了 NN 中可以使用的解释方法的范围。Charniak[4]借助于描述所谓"推出每个事件"的问题，展现了在非结构化 NN 里解释能力的不足。在 charniak 的术语里，解释意指去产生一个情形的画面，这个画面包括行走的人和他们可能的走向。为了保证能产生这样一个情形的画面，NN 必须具有表达所有可能事实合取的节点（因为这对解释很重要）。如果在一个松散型系统里后者用模式完备方法来实现，那么将产生一个情形的所有特征，而不仅仅是构造一个好的解释所需的一小部分特征。

至今为止，人们对于在非结构化 NN 里如何引进解释已经作了几点建议。一个思想起源于：在问题空间里基于规则系统的搜索与状态空间里基于松散型 NN 的搜索这两个不同的概念之间建立一个假想的相似性，上述两个"问题解决过程"基于一个特殊的开始点（输入）和一个结束状态，结束状态在第一种情形里是满足目标条件，在第二种情形里是一个局部极小点或一个全局极小值。两者的主要不同之处是：在基于规则的系统里通过问题空间很容易得到执行轨迹的跟踪，但在松散型 NN 系统中，很难跟踪所有状态的变化和保证所有中间状态在语义上均有意义。

模拟退火（例如在随机型 Boltzmann 机中）产生解释甚至更为困难，系统的行为将取决于最优方法的性能和随机因素。系统改变它的状态并最终到达一稳定点的这一过程可能包括能量的爬山行为和一些明显的随机行为，因此解释必须建立在系统的"理想化"行为上，而不能是在一特殊路径上。解释在前馈感知机（无反馈网）中也很难实现，分类是在一种简单的刺激向前传播的过程中完成的，如果隐含节点上特征的分布表示已被学习，那么通常在分类过程中没有能用于解释的中间步骤，因此用户只好信任系统工程师，认为学习被正确地执行，也认为系统拥有一个充分大的训练实例集，且系统是可靠的。

Mozer 和 Smolensky[6]指出，用简单的规则比用

很多的权值和激活值更能容易地理解前馈感知机的行为。他们的"缩减"技术有助于决定前馈网络中个体单元的合适性和从网络中移去冗余无用的单元。这个过程可得到一最小网络，它仅由确实对结果有所贡献的单元组成（即表示合适特性）。因此，一般化被限制，且根据简单的规则应能更容易地理解网络的性能和行为。然而，这一方法仅在输入单元和合适特征的数目相当少的情况下才具有一定优越性；规则的数目和复杂性不会促进解释，相反将造成权的分解问题。另外，Mozer 和 Smolensky 没能给出他们算法的终止性条件，指出什么时候停止从网络中移去单元。

Weigend 等人[6]提出了一种更为复杂的权值删除过程，他们的方法包含对传统 B-P 学习算法的扩充以引入一个更复杂的代价函数。该过程从某一给定问题的庞大前馈网络开始，把代价函数和网络中的每一联接相连。如果在训练集上，能通过少数几个权值来获得一个给定的性能标准，那么代价函数将推进删减过程，并最终删去尽可能多的联接。进一步，权的删除扩充到单元节点的删除且能从网络中移去最不重要的隐含单元。在某些情况下，当仅有少数几个隐含节点被保留下来时，就可能去识别出重要的输入特征并解释哪些特征对一个分类或预测有所贡献。

在 NN 中更易于实现第二种可能的解释方法——呈现相似情形，这在理论上也是可行的。因为图解和事件在分布式系统中的表示本质上就是基于相似性的。

综上所述，我们可得到如下结论：解释在非结构化的 NN 中是难于实现的。

4. 结构化神经网络系统中的解释机制

目前已有几种不同类型解释功能的 NN 系统，这里介绍的系统使用了结构化程度较高的知识表示，例如用连接线表示引起因果的联系。应注意到，标记传播系统已成功地用作解释部件的一部分。标记传播模型是大规则并行推理系统，它使用一个启发式推理部件（路径评估器）来对被激活概念的路径进行分析。下面将不讨论这类模型，而只着重介绍自由解释的、有带权联接和位或值传播的 NN 模型。

4.1 诊断型问题求解神经网络中的解释问题

Peng 和 Reggia[7]与 Wald 等人[8]分别描述了用于诊断型问题求解的一种神经网络模型。这种模型包括双向推理，模型的方法是基于竞争的。使用"multiple WTA"方法，它允许相互有抑制联接的单元间用"获胜单元"作为竞争的结果。Goel 等人[9]在 Hopfield 网络的基础上，也提出了一个类似的方法。

Peng 和 Reggia 的方法在概念上是基于最小覆盖理论的，这个理论中，有一无序集 D，一个现象（特征）集 M，一个关系 $D\times M\supseteq C$ 表示"无序集 D"和"现象集 M"间的因果联系，$(d_i, m_j)\in C$ 当且仅当"无序 d_i 是现象 m_j 出现的原因"。基于关系 C，对任一 $d_i\in D$，$m_j\in M$，定义两个集合"结果"和"原因"。一个假设 D_i 表达了对一个现象集 M^+ 的可能解释（M^+ 为所有现存的现象）。这里，对 M^+ 而言，D_i 是一无序的集合，表示当哪一个无序出现时，能造成或解释 M^+。

这个模型可用两层神经网络实现：D 和 M 是两个节点集且 C 是连接线集，每个连接 $(d_i, m_j)\in C$ 有一恒定的权值用于表示连接强度。当在时刻 t_s 松驰地到达平衡状态时，模型收敛到一个"胜者"集合，换言之，对任一 d_i，如果 $d_i(t_s)$ 近似等于 1 或 0，那么无序集 $D_i=\{d_i | d_i(t_s)\approx 1\}$ 被看作是 NN 模型的问题结果。

这里讨论的重点是 Peng 和 Reggia[7]网络产生解释的能力，一个解释是一个无序集，是已观察现象（特征）集的原因所在。然而现象集和无序集之间的因果强度在 Peng 和 Reggia 文章中是随机产生的，不能被学习（常数），而在 Wald 等人文章中由领域专家提供因果强度。

Peng 和 Reggia 的方法对本文的讨论很重要，因为它允许系统大体上去回答"why"问题。给定很多现象，系统用很多"无序"来解释"why"这些现象被产生。不幸的是，胜者"无序"节点的激活程度近似为 1（最大值），因而不可能把输出处理为概率或可能性的形式。如果能够解释现象对结果（即产生的无序）的贡献程度，那么对结果的可能解释将更为详细和有意义。

4.2 NN 中"解释合适性"的问题

下面是一个结构化 NN 系统[10]用于模型化"解释合适性"的一个例子，即选择一个假设集，它能够最好地被证据数据、直接观察和其它假设所解释。在这些命题间产生内部"合适性"能用一个 NN 系统来模拟。这里给出 Thagard 方法的讨论，因为对于解释的透彻性而言，逐字"合适性"是重要的。

从科学解释理论出发，Thagard[10]描述了"解释合适性"的一个计算理论，用以接受或拒绝科学假设集（或命题）。命题的相互竞争将建立稳定的结合，即

一致和非矛盾的能被证据数据等最好地解释的假设集。据 Thagard[10]所言："一个假设和命题集相合适（贴切）是指假设解释命题集或命题集解释假设，或两者共同解释其它的命题，或者命题集提供相似的解释"。用以描述直接观察到的事物的命题具有独立的可接受性，而一个解释假设被接受仅当它在总体上比竞争命题具有更好的"合适性"。

Thagard 理论本质上由建立命题间"合适性"的七个关系构成，这些关系为：对称性、解释、相似、数据优先权、矛盾、可接受性和系统合适性。下面对"解释"和"相似"加以更为详细的介绍以便给出"合适性关系"的一个简单实例，有关完整的描述请参见文 [10]。

Thagard 对"解释"理论描述如下：

如果 P_1，…，P_m 解释 Q，那么（1）对任一 P_i，$P_i=P_1$，…，P_m，P_i 和 Q 相合适；（2）对任一 P_i，P_j，$P_i=P_1$，…，P_m，$P_j=P_1$，…，P_m，P_i 和 P_j 相合适；（3）在（1）和（2）中，合适的程度反比于命题 P_1，…，P_m 的数目。

对"相似"理论描述如下：

（1）如果 P_1 解释 Q_1，P_2 和 Q_2，P_1 相似于 P_2，Q_1 相似于 Q_2，那么 P_1 和 P_2 合适，Q_1 和 Q_2 合适；

（2）如果 P_1 解释 Q_1，P_2 解释 Q_2，Q_1 相似于 Q_2，但 P_1 不相似于 P_2，那么 P_1 和 P_2 不合适。

上述的这类关系如何编码到一个 NN 模型中，这样的 NN 模型如何运行呢？首先，须有一个简单的高级描述语言，能允许诸如（EXPLAIN (H_1，H_2) E_1）和（CONTRADICT (H_1) (H_2)）形式的表达式（该表达式表示假设 H_1 与 H_2 共同解释证据 E_1，但 H_1 和 H_2 互相矛盾）。这些表达式被编辑在一个 NN 网络中，每一个命题用一个单一的节点表示。如果两个命题合适，那么在它们之间存在一个表示兴奋、具有正权值的对称连接。如果两个命题不合适，那么它们之间存在一个抑制连接。数据优先权由来自一个特殊数据节点的兴奋刺激链来实现。

在时刻 t，激活值大于零指示命题的接受，整个命题系统的合适性用下列函数表征（这个函数是 NN 系统中能量函数的负值）：

$$H(t)=\sum_i\sum_j W_{ij}a_i(t)a_j(t)$$

其中 W_{ij} 是从节点 i 到节点 j 的权值，$a_i(t)$ 是时刻 t 节点 i 的激活值。运行网络意指去产生命题的一个合适集（假如这个集合是可得到的），这个集合由命题节点上激活值的一个稳定模型来表示，这个稳定的结合体比其它具有更糟解释合适性的可能假设集拥有更多的控制权力。Thagard-[9]使用这种方法进行了科学解释和论证领域里的很多模拟实验。例如：Lavoisier 用氧气来反驳燃素理论的论点。至今为止，为了得到一稳定结果，最为复杂的应用包含 150 个节点和 210 次迭代过程。

（参考文献共 10 篇略）

（接第 22 页）

有 Mod (SETNAT) =I (SETNAT) =Z (SETNAT) = [PO]。

初始代数规范一般比终止代数规范和自由规范需要的公理数要少。初始代数规范的优越性在于它们为真（条件）等式规范，而终止代数规范和单构的自由规范则需要附加的非等式公理或附加的语义假设，这方面的探讨可参考有关文献。在以后的有关代数规范描述的应用中，我们将总是假定规范 SP 是 BOOLM 的完善，即 true≠false 是 SP 仅有的非等式，所有其它的公理都是（条件）等式公式。

参 考 文 献

[1] J. A. Goguen, J. Meseguer, 《Completeness of Many-Sorted Equational Logic》, 1981

[2] K. J. Barwise, 《Studies in Logic and the Foundation of Mathemaics》 Vol. 90, 1977

[3] P. Padawitz, M. Wirsing, 《Completeness of Many-Sorted Equational Logic Revisited》, 1984

[4] P. Padawitz, 《Computing in Horn Clause Theories》, Theoretical Computer Science 16, 1988

[5] M. Wirsing, 《Algebraic Specification》, MIP-8914, 1989

[6] 宋群，聂承启，杨茵，《抽象数据类型的代数方法研究》，《江西师大学报》3 期，1993