计算机科学 1996 Vol. 23 №. 3

联接机制与符号机制相结合的机器学习系统：发展与展望

孟祥武　程　虎

（中国科学院软件所　北京 100080）

摘　要　This paper reviews some of the research on machine learning that combines the symbolic and connectionist learning. In conclusion the paper demonstrates that combining symbolism and connectionism methods is a promising approach to machine learning, and points out open problems.

关键词　Machine learning, Symbolic, Connectionist, Artificial intelligence, Hybrid algorithms

机器学习是人工智能研究的一个重点，并已出现了多种学习方法，如：机械学习、归纳学习、演绎学习、类比学习、基于解释学习、遗传算法和联接学习等。机器学习的发展大致经历了三个阶段[Carbonell et al. 83]：神经系统模型和决策理论技术；面向符号概念的学习；知识密集型系统。

近年来，联接机制与符号机制相结合的机器学习方法越来越受到重视，是机器学习领域中的一个新的研究方向。综合现有的多种学习方法也是机器学习领域的一个研究重点[Saita et al. 93, Tecuci 93]。

一、发展历史

1988 年 Utgoff 将决策树与感知器结合起来，提出了联接机制与符号机制相结合的学习方法的一个算法，该算法首先尝试使用感知器学习，如果失败了，用 ID3 的信息理论方法将数据集合分成子集，然后对每个子集递归地应用 ID3 算法。

Gallant 提出了一个松耦合混合系统，高层决策是由符号处理的，低层决策是由神经网络处理。

Oliver 和 Schneider 用规则把问题分解成子问题，结果表明：分解后的问题显著地减少了神经网络的学习时间。

1988 年 Suddarth 和 1989 年 Le Cun 提出把规则加入神经网络，使网络以一种特定的方式进行学习，提高了学习效率。这里规则主要用于减少训练时间，使得网络在训练过程中通过提出的约束条件控制训练样本集合。

1989 年，Shavlik 和 Towell 从机器学习观点出发，分析了神经网络的一些不足之处，也指出了解释学习的不足，提出将解释学习(EBL)与人工神经网络(ANN)相结合，形成混合 EBL-ANN 系统，用 EBL 指导初始神经网络的构造，混合系统的 EBL 部分产生一个粗略的解释，然后用标准的解释概括(generalization)算法概括，再将概括结果转换成神经网络。分类规则(领域理论)的前件决定网络输入层到隐藏层的权值，规则的结论决定隐藏层到输出层的权值，构造好人工神经网络后，利用标准神经网络学习技术和附加例子求精网络，最后形成一些输入量的函数，这样的函数用领域理论是不能表示的，因为混合系统(EBL-ANN)可以逐步增加这种表达函数，这同时增强了表达的精确性，还适应于所学概念是逐步变化的。该混合系统可以用不完全领域理论进行解释学习，并能选择一个好的初始神经网络结构。与单纯的神经网络相比，该混合系统需要较少的训练例子；与解释学习相比，可以用近似(不完全)领域理论。

Katz 描述了一个类似于 EBL-ANN 的系统，用一个知识库映射成神经网络。缺点是：仅提高分类速度，但不修改生成解释的规则，开始神经网络与解释

孟祥武　博士生，主要研究领域为神经网络和机器学习。程　虎　研究员，主要研究领域为语言编译，软件工程，人工智能。

结构相似，但在学习过程中，这种相似性很快就改变了。

Mooney 等通过试验比较了联接机制与符号机制学习，他们用 ID3 算法作为符号机制学习的代表，因为 ID3 算法简单而且广泛使用，属于一种归纳学习算法。另外试验表明[Rendell 89]：在符号学习算法中，ID3 一般完成得较好。感知器是较早的联接机制学习算法之一，尽管它有局限性，但仍不失简单实用。BP 算法是证明有效且流行的，代表了一类联接机制学习算法。这类算法克服了感知器的不足。试验结果表明：ID3 和感知器的学习和分类新例子时间明显比 BP 算法快一至二个数量级，正确分类结果也相似。对于有噪音的数据集合，BP 算法分类更精确。对于学习结果，符号机制容易解释，但联接机制却很难。对于 BP 算法，经验表明：若初始点（参数或权值）不合理，会不收敛。如何选择初始点？没有较好的办法。但许多符号学习算法却能选取适当的参数，保证较好的性能。

Fisher 等也通过试验比较了 ID3 和 BP 算法，他们认为两个算法各有特点，应该结合起来，发展符号与联结机制相结合的混合系统。

1990 年，Shavlik 把神经网络与符号逻辑结合起来研究知识求精。Hall 等提出一种 SCnet 模型可以完成有限的模糊推理，可以通过增量式学习加以扩充，但联接权不能学习。SCnet 被用于诊断和分类。

Fu，Murphy，Towell 和 Shavlik 等分别提出的方法属于转换模型，这类方法把专家系统的知识转换成神经网络，或把神经网络转换成专家系统知识。转换前的系统成为源系统，转换后的系统成为目标系统。在目标系统工作时，源系统保持静态。目标系统工作结束后，可以根据需要把目标系统再转换到源系统，或不再转换。本质思想是：用优先知识决定如何初始化神经网络，即网络选择一个好的起始点（初始的权值设置和网络拓扑结构）。转换模型可以把源系统的优点结合到目标系统中。如果源系统是专家系统，目标系统是神经网络，则可获得学习能力及自适应、速度（并行）等。反过来则可获得单步推理能力、解释能力及知识的显式表示等。缺点是：缺乏精确的转换方法。

这里 Fu 提出的模型称为 KBCNN，其学习包括用 BP 和爬山法相结合的权修正过程，规则的删除等，但不能进行增量学习，即不能增加新结点。

Benachenhou 等提出一种松耦合模型，主要是专家系统与神经网络通过一个中间媒介（数据文件）进行通讯、交换数据。

1991 年，Hendler 等提出了一种紧耦合模型，两者之间的数据交换是直接通过内部数据完成的，而不是通过外部数据进行，效率很高。如在专家系统中，用神经网络作为知识获取模块，用符号机制作为解释模块，对网络的行为进行解释。

Pomerleau 等描述了基于规则技术，与 Gallant 思想相似，将多层神经网络与符号知识相结合。由于让神经网络实现符号任务，如关于图的规划和推理是困难的，故该技术用标记图（Annotated maps）提供网络缺乏的高层决策能力，这种标记图也称环境图，可决定特定情况下哪个网络适合于目前的环境，标记图可被看成是高层符号知识。

Cios 等研究将连续 ID3 算法和机器学习算法 CLILP2 结合起来，产生一神经网络结构。1992 年，Cios 等提出一连续 ID3 算法，转换决策树为神经网络的隐层，帮助决定反馈神经网络的结构（结点数和隐层）。该算法可以解释嵌入连接和权上的知识，即这些知识可被转换成决策规则，揭示了归纳学习和反馈神经网络之间的关系。

Lacher 等提出了一种根据知识库构造相应的神经网络的方法（对应的神经网络为专家系统）。这种方法与 Fu 的方法较类似。网络可以通过 BP 算法修正其联接权，但是只对网络的终端结点进行学习。由于联接权与规则强度有一定的对应关系，故通过这种方法可以修正规则强度，但网络的拓扑结构不能通过学习加以改变。

Towell 证明用基于知识的人工神经网络的领域理论求精比纯符号理论求精系统要好。Towell 和 Shavlik 还提出一种技术，用符号的归纳学习表示好的输入特性，即它们权值大。结果发现这种经过预处理的神经网络有较好的归约（generalization）能力。

1992 年 Samad，1993 年 Sun 提出的系统属于一种完全集成模型，将两种机制完全融合在一起，共享数据结构和知识表示。这种模型主要在联接机制的专家系统中可以看到。

神经网络的学习包括两方面，拓扑结构的变化及联接权的变化。大部分神经网络学习算法都是基于联接权变化的学习。

1993 年，Optiz 和 Shavlik 提出了一个算法，在

训练期间，利用符号机制指导，解释和说明神经网络中何处动态地增加新结点，这样学习不仅仅是神经网络中权的变化，而且拓扑结构也可以变化，同时可以向规则库增加新规则，对领域理论进行求精。但问题是：究竟网络中何处增加新结点最好？为什么？还没有很好的解决，需要进一步的研究。

Mitchell 等提出基于解释的神经网络学习(EBNN)，用于减少机器人需要学习的例子数。在EBNN 中，用一组神经网络表示领域理论。类似于符号 EBL，根据领域理论，解释和分析每一个目标函数训练例，利用领域理论指导了目标函数学习。EBNN 的优点是：1)在目标函数学习之前或当中，利用 BP 或其它神经网络学习过程，领域理论本身也可以从带噪音的数据中学习。2)基于训练例(归纳部分)和从解释中抽取的信息(分析部分)，EBNN 是一个自然的方法，逐步求精要学的目标概念。

1994 年，Soulie 提出两种提高神经网络性能，降低复杂度的策略。一是模块化，将问题分解成子问题，每个子问题对应于一个模块，先分开单独训练每个模块，然后再设法综合起来。训练后，一个模块代表一个局部最优，但又出现了如何使其全局最优的问题。为此提出了一个形式化框架，用以合作训练多模块系统结构的网络，该方法在许多应用中证明是有效的。二是变量选择，一些不重要的或几乎无用的变量，可以考虑当作噪音除去，而主要考虑最重要的变量，这样就减少了收集和处理的数据量。相应有多种变量选择方法。结果表明：可减少一半变量，并能提高性能。

Garvin 提出一个神经网络自构算法 IWA(Iterative Weighted Adaptive)，用优化子集阻止训练过度问题，产生好的归约结果。IWA 算法能决定正确的神经网络模型，模型的选择是基于训练数据的。

1995 年，Giraud-Carrier 等针对一类自组织动态网 ASOCS (Adaptive Self-Organising Concurrent Systems)，提出了一个动态增量式学习算法 AA1＊，该算法可改变整个网络拓扑结构，它用最小的网络增长达到最大预测精度。AA1＊能逐步处理例子和规则，结合了符号和联接主义表示。

传统神经网络模型分成学习和识别两个独立过程，学习过程不识别，识别过程不学习，Zochowski 等提出一种自控制学习神经网络模型 SMARTNET，设法将两个过程统一起来，在识别过程中学习，即改变权值，整个过程为：对输入模式在相对短的时间内判断其新颖程度，SMARTNET 主要学习新模式，识别已熟悉的模式。学习和识别过程是同时的，学习速度主要由新模式决定。

二、基于知识的人工神经网络

在联接机制与符号机制相结合的机器学习领域，较成功的系统是基于知识的人工神经网络(KBANN)。以美国威斯康星大学 Shavlik 为首的研究小组，从 1989 年起，对联接机制与符号机制相结合的机器学习领域做了深入的研究，在研究将解释学习与神经网络相结合的基础上，提出一种称为基于知识的人工神经网络(KBANN)的学习方法。它利用不完善领域理论和带有噪音的数据进行解释学习，产生一种近似正确的解释结构(规则树)，利用该解释结构构造初始神经网络，对该网络利用 BP 算法进行训练，形成正确的表示目标概念的网络，最后从训练好的网络中抽取出最终符号信息来。

KBANN 已被成功地应用到 DNA 序列分析上，并与六个实验性算法：BP、ID3、最近邻居、PEBLS、感知器和 Cobweb 等算法进行了比较，对试验结果进行了分析比较，最后得出 KBANN 是一种非常有效的方法，并优于以上算法。

1994 年，Shavlik 和 Towell 较完整地论述了 KBANN，但 KBANN 有以下缺点：

1)训练后的神经网络难以解释其决策，KBANN 学到的知识难以转换成有关问题的解。

2)不能处理带变量的规则，不能处理不确定性规则。

3)没有归纳新规则的能力，不能向过去不完全的规则集中增加新的符号规则。

4)不能改变网络的拓扑结构。

5)由于神经网络学习算法不能处理谓词演算变量，故 KBANN 只能适用于命题逻辑的领域理论，KBANN 是基于 BP 算法的，由于 BP 算法不能训练递归网，故 KBANN 只能适用于非递归的领域理论。

6)缺乏精确的转换方法，KBANN 要经过两次转换。

7)KBANN 要求领域知识必须是层次结构的，若规则不是层次结构的，就没有中间结论，规则转换成神经网络也没有隐藏层，结果 KBANN 就与感知器学习一样了。

8)KBANN 是根据由试验得出的经验性参数、计算、设置初始权值和阈值，缺乏理论上的分析和依据。

三、结束语

从以上的发展历史，我们发现有以下几个方面的工作还需进一步的研究：

1. 神经网络的学习包括两方面，权值的变化和拓朴结构的改变。目前的神经网络学习算法主要集中于权值的变化上，利用符号机器学习技术辅助神经网络学习，加快权值学习过程，另外使网络拓朴结构能合理地确定和变化，形成神经网络的一个统一学习算法(权值和拓朴结构统一变化)。目前还没有一个通用的、独立于问题的方法来选择一个好的网络拓朴结构，形成合理的网络拓朴结构是目前研究的一个热点。

2. 利用符号学习技术确定神经网络何时停止训练，防止训练过度或训练不足。

3. 如何利用符号学习技术减少学习例子，如用符号规则指导神经网络尽快结束训练，因为有时学习例子很难找到或寻找代价很大。

4. 训练例子的顺序与学习结果的关系问题。

5. 较好地处理带有噪音的数据以及个别有错误的数据。

6. 从神经网络中抽取符号规则是联接机制与符号机制学习相互转化的一种途径，也是两种机制结合的一个方法。同时使神经网络具有解释能力，便于更好地理解神经网络，目前也是一个研究热点。

7. 由于网络中联接权与知识库中规则强度有一定的对应关系，联接权的修正对应于规则强度的修正，从而为规则强度的学习提供了有效的手段。

8. 一般神经网络训练前，随机设置初始权值，然后开始训练，研究表明[Towell et al. 1994]：初始值影响神经网络学习，如何利用符号规则选择一个好的初始点(初始权值，初始网络拓朴结构，参数)是需要深入研究的。

9. 符号学习中归纳学习与神经网络学习相结合问题。这两者有共同之处，神经网络是从例子中学习，也是一种归纳学习。

10. 解释学习与神经网络相结合，发展基于解释的神经网络(EBL-ANN)。KBANN 就是 EBL-ANN 算法的发展。

11. 进一步提高神经网络的可靠性、鲁棒性。

总之，联接机制与符号机制相结合的机器学习是一个值得研究的方向，存在许多问题，研究下去，在机器学习或神经网络学习领域一定会有所突破。(参考文献共 38 篇略)

(上接第 62 页)

根据预备定理 4，第 i 小组的数据一致趋向于均匀分布($1\leqslant i\leqslant N$)。□

定理 2 递归分组排序的期望复杂度为 O(N)。

证明：由于递归分组排序算法第一次分组的期望时间为 O(N)；又 第一次分组后，再对每个小组分组排序，即用桶排序算法排序每个小组的数据，由定理 1，第一次分组后，每个小组内的数据服从均匀分布，故排序第 r 个小组数据的平均工作量小于 CN_r，其中 N_r 是第 r 个小组的数据个数，C 为一常数，$1\leqslant r\leqslant N+1$。于是，排序全部 N+1 个小组的数据的平均工作量小于：

$$\sum_{r=1}^{N+1} C\cdot N_r = CN$$

故 递归分组排序的期望时间为 O(N)。□

递归分组排序不要求数据服从某些特殊概率分布，其排序平均工作量为 O(N)。优于 Akl 等人出的桶排序算法。

4 实验结果

在 IBM-PC 机上，采用 BASIC 语言，对同一组随机数(由 RND 函数产生)，分别使用 Quicksort 和递归分组排序算法，所用的 CPU 时间(单位为秒)如表 1 所示。

表 1

数据个数	200	500	1000	1500
Quicksort(秒)	31	84	185	2 4
递归分组排序	18	44	87	130

实验结果表明递归分组排序算法的排序速度比快速排序算法的排序速度要快一倍，且数据量愈大，递归分组排序愈优。(参考文献共 5 篇略)