计算机科学1996 Vol. 23№. 1

76-79

前馈神经网络中的反向传播算法及其改进:进展与展望

刘曙光 郑崇勋 刘明远
(西安交通大学 西安710049)

TP18

A

摘 要 BP网络和算法是使用最广泛的神经网络模型之一,但由于它使用梯度算法,因而存在固有的局部极小及收敛速度慢等问题。本文首先回顾了BP算法的产生和和发展过程,之后对BP算法固有的特点进行了阐述,最后针对原基本BP算法的缺陷对各种改进方法进行了全面综述,并指出了这一研究中的有关问题。

关键词 人工神经网络 反向传播算法 收敛 极值

前馈神经网络

1 BP算法的产生和发展

1958年,心理学家Rosenblatt提出了最早的前馈层次网络模型,并称为感知器(Perceptron)。在这种模型中,输入图形 $x=(x_1,x_2,\cdots,x_n)$ 通过各输入结点分配给下一层的各结点,这下一层就是所谓中间层,中间层可以是一层也可以是多层,最后通过输出层结点得到输出图形 $y=(y_1,y_2,\cdots,y_c)$。在这类前馈网络中没有反馈连接,没有层内连接,也没有隔层的前馈连接,每一结点只能前馈连到其下一层的所有结点。然而,对于含有隐蔽层的多层感知器当时没有可行的训练办法,所以初期研究的感知器为一层感知器。1969年,Minskey和Papert对Rosenblatt提出的简单感知器进行了详细的分析。他们引用的一个典型例子是所谓XOR(exclusive-or)问题。Minskey和Papert指出没有隐层的简单感知器在许多像XOR问题的情形下显得无能为力,并证明了简单感知器只能解决线性分类问题和一阶谓词问题。对于非线性分类问题和高阶谓词问题,必须引用隐单元层。隐单元可以在某一权值下对输入模式进行再编码,使得在新编码中模式的相似性能支持任何需要的输入输出映射,而不再像简单感知器那样使映射难以实现。

隐层的引入使网络具有很大的潜力。但正像Minskey和Papert当时所指出的,虽然对所有那些能用简单(无隐层)网络解决的问题有非常简单的学习规则,即简单感知器的收敛程序(主要归功于Widrow和Hoff于1960年提出的Delta规则),但当时并没有找到同样有效的含隐层的网络的学习规则。对此问题的研究有三个基本的结果。一种是使用简单无监督学习规则的竞争学习方法,但它缺乏外部信息,难以确定适合映射的隐层结构。第二条途径是假设一个内部(隐层)的表示方法,这在一些先约条件下是合理的。另一种方法是利用统计手段设计一个学习过程使之能有效地实现适当的内部表示法,Hinton等人(1984年)提出的Bolzmann机是这种方法的典型例子,它要求网络在两个不同的状态下达到平衡,并且只局限于对称网络。Barto和他的同事(1985年)提出了另一条利用统计手段的学习方法。但迄今为止最有效和最实用的方法是Rumelhart、Hinton和Williams(1986年)提出的一般Delta法则,即反向传播(BP)算法。Parter(1985年)也独立地得出过相似的算法,他称之为学习逻辑。此外,Lecun(1985年)也研究出大致相似的学习法则。

2 BP网络特性

BP网络的学习过程是一种误差修正型学习算法,由正向传播和反向传播组成。在正向传播过程中,输入信号从输入层通过作用函数后,逐层向隐含层,输出层传播,每一层神经元状态只影响下一层神经元状态。如果在输出层得不到期望的输出,则转入反向传播,将误差信号沿原来的连接通路返回,通过修改各层神经元的连接权值,使得输出误差信号最小。BP网络的连接结构和映射过程与多层Perceptron相同,只是后者阶跃的单元激发函数被换作Sigmoid函数。Rumelhart等人在1985年重新发现用

(19)

梯度法修改权重(generalized delta rule)的样本学习算法可以有效地运用在多层网络上，使得过去在Perceptron模型中无能为力的XOR等学习问题获得解决。含有输入、输出和单层隐单元的三层BP网络富有的功能引起人们的注意。Lippmann(1987年)指出三层网络可以处理凸区域上模式识别问题。Wieland和Leighton(1987年)给出了一个例子，用三层网络将空间划分成凹的子空间。Huang和Lippmann(1987年)仿真演示了三层网络可以处理几种很复杂的模式辨识问题。这些研究促进了三层网络的广泛应用。Funashi和Hecht-Nielsen(1989年)分别证明了随着隐单元的增加，三层网络所实现的映射可以一致逼近紧集上的连续函数或按L^2范数逼近紧集上平方可积的函数，揭示了三层网络丰富的实现映射能力。Mitchison和Durbin(1989年)给出在一定条件下，三层网络学习容量的上、下限的估计。三层网络的输入和输出单元都由应用的问题所规定，只有隐单元的数目是可变的。应行仁(1990年)详细分析三层神经网络的记忆机制，指出具有足够多隐单元的三层神经网络可以记忆任给的样本集。采用渐近函数(非常一般的函数，包括阶跃函数、Sigmoid函数等)作为隐单元激发函数的三层神经网络，$k-1$个隐单元能够准确记忆k个实验值样本。采用阶跃激发函数时，$k+1$个随机给定的实数值样本能够被$k-1$个隐单元的网络记忆的概率为零。联想记忆在Sigmoid激发函数的网络中结果也是如此。BP网络除具有较强的对信息分布式记忆特点外，还具有一定的容错性和抗干扰性。孙德保、高超(1994年)对三层BP网络的容错性和抗干扰性进行了研究，得出了三层BP网络的容错能力取决于输入层到隐含层的连接权值矩阵与隐含层到输出层连接权值矩阵的乘积的结果。

3 基本BP算法存在问题

BP算法的基本形式为：

$$W^{(k)}=-\eta\nabla E(W^{(k)})+\alpha\triangle W^{(k-1)} \tag{1}$$

式中η是学习率或迭代步长，α是惯性因子。

用三层具有Sigmoid神经元非线性的网络可以以任意精度逼近任何连续函数，但是它主要存在如下缺点：①从数学上看它归结为一非线性的梯度优化问题，因此不可避免地存在局部极小问题；②学习算法的收敛速度很慢，通常需要上千次或更多；③网络结构为前向结构，没有反馈连接，因此它是一非线性映射系统。

基本的BP算法最大的问题是采取梯度法(LMS)时步长和势态项系数是由经验确定的。步长和势态项系数选取不好会使训练时间过长，甚至会引起完全不能训练，其原因：一是网络的麻痹现象，一是局部最小。

(1)网络的麻痹现象。在训练过程中，加权调得较大可能迫使所有的或大部分的节点的加权和输出s_j较大，从而工作在S型激发函数的饱和区。此时激发函数在其导数$F'(s)$非常小的区域。由于在计算加权修正量的公式中，各层误差正比于$F'(s)$，当$F'(s)\to 0$时各层误差趋于零，这使得$w_{ij}\to 0$，相当于调节过程几乎停顿下来。

(2)局部极小。采用梯度法的训练过程从某一起始点沿误差函数的斜面逐渐达到最小点$E\to 0$。对于复杂的网络，其误差函数面在多维空间，就像个碗，碗底是最小点，但这个碗的表面凹凸不平，因而在训练过程中可能陷入一小谷区，称小谷区为最小点。由此点向各方向变化均使E增加，以至无法逃脱局部最小点。

4 各种改进算法

由于在ANN中BP占据了非常重要的位置，所以近几年许多研究人员对BP作了深入的研究，提出了很多BP的改进方案，其主要目的是为了加快训练速度，避免陷入局部极小和改善概括能力。

模式I：改进误差函数。误差函数表示为：

$$E(w)=\frac{1}{2}\sum_j(\hat{y}_j-y_j)^2 \tag{2}$$

误差函数的定义不是唯一的，可以选用别的函数$f(\hat{y}_j,y_j)$代替$(\hat{y}_j-y_j))^2$，只要f函数在$\hat{y}_j=y_j$时达到最小就可以。这样导出的BP算法除输出层的δ_j不同以外，其它各层的方程与基本BP法没有什么差别。Baum和Wilczek等人(1988年)提出了一种误差函数

$$E=\sum_{k,j}\left[\frac{1}{2}(1+\hat{y}_j^k)\log\frac{1+\hat{y}_j^k}{1+y_j^k}+\frac{1}{2}(1-\hat{y}_j^k)\log\frac{1-\hat{y}_j^k}{1-y_j^k}\right] \tag{3}$$

该式同样满足当$y_j^k=\hat{y}_j^k$时，$E=0$，但是，当$y_j\to\pm 1$时，式(3)发散，而式(2)则趋于常数，即处于E的平坦区，从而长时间离不开，这就是所谓麻痹现象。在这种情况下，转移函数采用$F(s)=\mathrm{th}(s)$，将式(3)微分并代入$F'(s)=1-(y_j^k)^2$，可得

$$\delta_j^k=\hat{y}_j^k-y_j^k \tag{4}$$

显然，一般式中的$F'(s)$项消失了，事实上，如$F'(s)$

存在，则当|s|增大时，它进入转移函数的平坦区，因为F'(s)→0，即正平面的平坦部分。s→0，相当于E平面的突变区，F'(s)很大，可能导致过调或振荡。于是，Fahlman(1989年)提出了一种折衷方案，即

$$\delta_i^k = (F'(s)+0.1)(y_i^k - y_i^k) \quad (5)$$

该式的效果比式(4)强。一方面恢复了F'(s)的某些影响，另一方面又消除了麻痹现象，即当|s|变化时，仍然保持$\delta \neq 0$。

模式Ⅰ：改进激发函数。Stornetta和Huberman(1987年)提出了一种改进激发函数的方法之一：双极性S激发函数BP算法。一般S型激发函数的输出动态范围为0→1，这不是最佳的。为了解决这一问题，可将输入范围变为±1/2，同时对S激发函数偏置，使结点的输出范围变为±1/2，即$y_{新}=(1+e^{-s新})^{-1}$改为$y_{新}=-1/2+(1+e^{-s新})^{-1}$。而在利用BP算法时，其中的一阶导数$F'(s)=1/4-y_{新}^2$。实验表明，收敛时间平均减小30～50%。姜天戟、袁曾任(1995)也对BP算法中使用的激发函数进行了详细分析，他们发现S型激发函数在BP算法中被多次使用，而其导数范围仅为[0，0.25]，这必将使得学习速率较慢。为了提高学习速率，可以构造一个组合函数，该函数满足S型激发函数的要求，但其导数在一些重要的点上取得较大(并非每点的导数在训练中同等重要)，从而加速学习的收敛。

模式Ⅱ：改进网络结构。前馈多层网络克服了单层网络的功能表示能力低的缺陷。但是，训练过程收敛太慢也是个大的缺点，Pao(1989年)提出了一个称为函数链路网络结构。该网络为单层网(没有隐蔽层)，在将输入样本加入到网络之前，先通过一函数链路进行某种非线性变换，将每一输入分量通过函数链路变换为一系列线性独立函数，从而将原样本的空间维数变为独立函数的高空间维数。这样，新的信息表达空间扩展了，使单层网络具有了分辨复杂图形的能力，同时，该网络收敛速度也加快了许多。在前向式网络拓扑结构中，输入节点与输出节点是由问题本身决定的，只有隐层的层数与隐节点的数目是可变的。对于隐层的层数，应行仁(1990)已有了详细的论述。相对来说，隐节点的选取较为困难，隐节点少，学习过程可能不收敛；隐节点多，网络性能下降，节点冗余。为了找到合适的节点，在学习过程中，根据环境要求，自组织和自学习自己的结构，这种网络学习方法称为自构性学习算法。自构学习(C. C. Lee 1991)分为两个阶段：预估阶段和自构阶段。在预估阶段，网络根据问题的大小定一个隐节点数较大的神经网络结构(预定的结构并不一定理想)。在自构阶段，网络根据学习情况合并无用的冗余节点，最后得到一个大小合适的自适应型神经网络。

模式Ⅳ：参数α和η自适应调整。对于一个特定问题，要选择一个适当的参数α和η值不是一件容易的事，通常凭经验和实验选取，然而在训练开始时较好的参数α和η不见得对后来的训练过程合适。为解决这一问题，Cater、Franzini、Vogl、Jacobs等人建议在训练过程中，自动调整这些参数。通常调节参数的准则是检查某特定加权的修正是否确实降低了误差函数，如果不是这样，就应该修改α和η。D. E. Rumelhart (1986)、R. A. Jocobs (1988)、T. Tollenare(1990)和F. M. Silva等(1990)就该问题研究了如何利用启发式信息使α和η得到自适应调整。

模式Ⅴ：提高收敛速度。为了提高收敛速度，研究人员对基于梯度法的BP算法进行了很多改进。谭永红(1994)将多层前向神经元分解成线性输入和非线性输出两部分，并注意到其输入到输出的映射是一对一的，而且其逆映射亦是一对一的特点，将递推最小二乘法估计技术用于估计神经元之间的连接系数。由于递推最小二乘法有较快的收敛速度，因而加速了神经网络的训练过程。徐嗣新等(1993)提出了前向神经网络的分段学习算法。该算法结合自适应BP算法与Newton算法，从而提高收敛速度。自适应BP学习方法在远离极小点时，学习速度快，接近极小点时，易产生振荡，使学习速度降低；而Newton法只在极小点附近才有效，但Newton法不易找到合适的初值。若将这两种方法结合起来，即先用自适应BP算法学习，接近极小点时再用Newton法学习，则可同时利用一、二阶导数提出的信息。邓志军等(1995)利用PID控制思想，提出了BP网络的一种二阶快速学习算法，给出了学习因子选择的必要条件与较佳区域，并结合一非线性正弦函数进行了仿真研究。结果表明，较之标准BP学习算法，利用此法可使学习收敛速度提高22倍左右。C. Charalambous (1992)提出了一种适合于BP学习的共轭梯度法，该方法与一种简单的不精确线性搜索相结合，极大地提高了BP学习速度，使收敛速度提高了两个数量级，同时不依赖于α和η的选择。林忠(1993)还从模式样本与收敛性的相关性的依从关系出发，提出用快速沃尔什-阿达玛变换(FWHT)对模式样本作正交化处理，从而大大加速其收敛过程。

模式Ⅵ：全局优化。经典BP算法存在局部极小问题，而非所期望的全局极小点。局部极小问题使算

法甚至会在某些情况下失效。跳出局部极小区的一般做法是：增加隐单元数，或改变学习速率，或同时从多个初始点开始学习。这些做法需试凑多次才能成功，而每次学习时间又很长，致使训练很困难。王小同等(1994)提出了一种避免局部极小问题的方法。这一方法从寻找全局极小点的思路出发，将全局优化方法运用于前向网络学习算法，只需在原来学习算法中加入一个由全局优化方法形成的初值点选择模块，以选择好初始权值，从而自动地避免了局部极小问题的发生。他们分别采用了隧道效应法、填充函数法、测度论法三种全局优化方法，经实验验证，认为测度论法在求解 BP 算法全局极小值问题时更为有效。为了求解全局极小问题，可以采用一些优化理论的方法，如卡尔曼滤波、同伦优化等。S. Singhal 等(1989)利用最优估计理论中的卡尔曼方法，把 BP 算法的网络权值作为滤波的状态变量，从而利用推广卡尔曼滤波来实现非线性网络的学习，不仅避免了局部极值，而且大大提高了学习速度。J. Chow 等(1991)将一 BP 网络误差函数最小化问题转化为一非线性代数方程的求解问题，然后将连续同伦思想用于非线性代数方程的求解，建立了相应的同伦 BP 网络理论和学习算法。同伦 BP 算法不但是大范围收敛的，同时具有良好的收敛速度和可克服病态能力，在梯度法不收敛时，它仍能给出满意的结果。

5 BP 网络研究的有关问题

BP 网络理论已经引起了许多领域科学家的兴趣和关注，但它还处在迅速发展中，因此，还有许多工作要我们去做。概括起来，至少对下述问题的研究是有意义的：

——开发更为简便的具有全局最小和快速运算的 BP 算法。

——BP 网络中隐单元层数、隐单元个数的选取需要进一步严密的理论上的指导。

——自适应拓扑结构和 BP 算法更具有适应性。

——开发可用于 VISI 的 BP 网络结构与算法。

——BP 网络的研究不仅其本身正在向综合性发展，而且愈来愈与其它相关学科密切结合起来，发展出性能更强的结构。

人工神经网络的研究经历了七十年代低潮期以后，进入八十年代开始复苏，并掀起了第二次研究热潮。神经网络有两个与传统方法进行信息处理完全不同的性能：第一，神经网络是自适应和可以被训练的，它有自修改能力。第二，神经网络结构本身就决定了它是大规模并行机制，就是说神经网络从原理上就比传统方法快得多。因此，它的研究与应用已成为科学技术研究中的又一新的热点。BP 网络做为人工神经网络中最基本的和使用最广泛的网络，不仅在理论研究上日臻成熟，其应用也取得了令人鼓舞的进展。如家用电器、故障诊断、模式识别、图像处理、工业控制、专家系统、管理系统、运输系统、财政金融、VLSI 芯片等，已广泛采用了 BP 网络技术。(参考文献共15篇略)

(上接第81页)

五、结论

本文对 BMI 模型有关的两个问题进行了讨论，我们的结论是

(1)运用 PROLOG 实现基于 BMI 模型的专家系统时，形如 $E_1 \vee E_2 \xrightarrow{RS} H$ 的规则不能随意地分解为 $E_1 \xrightarrow{RS} H$ 和 $E_2 \xrightarrow{RS} H$。

(2)BMI 模型应用于分布式专家系统时，在某些情况下可能会导致系统运行结果不等价。

上面两个结论，对 MYCIN 的确定性因子模型仍成立。因此，我们在使用它们时必须谨慎小心。

参考文献

[1]罗旭东、蔡经球、邱玉辉，一种新的基于区间估计的不确定推理模型，西南师范大学学报(自然科学版)，1994

[2]张为群，基于区间估计的不确定推理模型 BMI 的性质分析，计算机科学，Vol. 22，№5，1995

[3] Shortliffe, E. H., Computer-Based Medical Consultations: MYCIN, American Elsevier Publishing Inc., New York, 1976

[4]Zadeh, L. A., The Concept of a Linguistic Variable and its Application to Approximate Reasoning, Part I, Inform. Sci., Vol. 8, 199-249; Part II, Inform. Sci., Vol. 8, 301-357; Part III, Inform. Sci. Vol. 9, 43-80, 1975

[5]Baldwin, J. F., Support Logic Programming, Int. J. Intell. Syst., 1, 73-104, 1986