

Ridge Polynomial 神经网络带动量项异步梯度算法的收敛性

喻 昕 唐利霞 于 琰

(广西大学计算机与电子信息学院 南宁 530004)

摘 要 将动量项引入到 Ridge Polynomial 神经网络异步梯度训练算法的误差函数中,有效地改善了算法的收敛效率,并从理论上分析了 Ridge Polynomial 神经网络的带动量项的异步梯度算法的收敛性,给出了算法的单调性和收敛性(包括强收敛性和弱收敛性)。算法的这些收敛性质对于如何选取学习率和初始权值来进行高效的网络训练是非常重要的。最后通过计算机仿真实验验证了带动量项的异步梯度算法的高效性和理论分析的正确性。

关键词 Ridge Polynomial 神经网络,异步梯度算法,动量项,单调性,收敛性

中图法分类号 TP183 **文献标识码** A

Convergence of Asynchronous Gradient Method with Momentum for Ridge Polynomial Neural Networks

YU Xin TANG Li-xia YU Yan

(School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China)

Abstract The momentum was introduced into the conventional error function of asynchronous gradient method to improve the convergence efficiency of Ridge Polynomial neural network. This paper studied the convergence of the asynchronous gradient method with momentum for training Ridge Polynomial neural network, and a monotonicity theorem and two convergence theorems were proved, which are important for choosing appropriate learning rate and initial weights to perform an effective training. To illustrate above theoretical finding, a simulation experiment was presented.

Keywords Ridge polynomial neural networks, Asynchronous gradient algorithm, Momentum, Monotonicity, Convergence

1 引言

很多前馈神经网络具有非线性映射能力, Ridge Polynomial 神经网络(Ridge Polynomial Neural Network, 简称 RPNN)^[1,2]就是这样的一种前馈神经网络。RPNN 是一种以 Pi-Sigma 神经网络(Pi-Sigma Neural Network, 简称 PSNN)^[3]为模块,具有多项式乘积构造的求积神经元高阶神经网络。它克服了一般的前馈神经网络学习速度慢、泛化能力差的缺点,既保留了高度的非线性映射能力,又具有广泛的应用范围。目前,RPNN 已经广泛应用到很多的领域中^[4-7],而对于 RPNN 的理论研究较少,有必要对其进行深入的理论分析,以提高 RPNN 的应用效率。

梯度算法是高阶神经网络应用最广泛的训练算法,标准的梯度算法在调整权值时,只是按当前时刻误差的梯度下降方向调整,而没有考虑前一时刻的梯度方向,从而使训练过程发生震荡,收敛缓慢。针对以上问题的出现,本文提出一种将动量项加入到误差函数中的方法,并研究了加入动量项后的 Ridge Polynomial 神经网络异步梯度算法误差的单调性和权值的收敛性问题。在权值调整公式中加入动量项,不仅考虑了当前梯度方向,同时也兼顾了以前积累的调整经验,使得权

值的调节不产生大的摆动,从而达到跳出局部极小区域和加快迭代收敛速度的目的。与不带动量项的异步梯度算法的实验结果相比,本方法具有更快的收敛速度。

2 相关研究

早期的前馈神经网络只含有求和神经元,在处理复杂非线性问题时效率很低,因此引入了具有求积神经元的高阶神经网络,其中包括具有高阶处理单元的神经网络(High Process Unit Neural Network, 简称 HPUNN)^[8]、Sigma-Pi 神经网络(Sigma-Pi Neural Network, 简称 SPNN)^[9]、积单元神经网络(Product Unit Neural Network, 简称 PUNN)^[10]等。RPNN 是一种以 PSNN 为模块的新型高阶神经网络且得到了广泛的应用。梯度算法是高阶前馈神经网络应用最广泛的训练算法,按样本的输入方式可将其分为批处理梯度算法和在线梯度算法;按训练过程权值的更新方式可将其分为异步梯度算法和同步梯度算法。目前学者们已经将很多种梯度训练算法应用到高阶神经网络训练中,得出了相关的收敛性理论^[11-16],在文献[11]中,批处理梯度算法被应用在三层的 SPNN 训练中,并给出了相应的收敛性结论;文献[12]中使用在线梯度算法来训练 PUNN,并对算法的单调性和收敛性进

到稿日期:2013-05-21 返修日期:2013-07-09 本文受国家自然科学基金项目(61063045),广西科技攻关项目(桂科攻 11107006-1),广西教育厅项目(TLZ100715)资助。

喻 昕(1973—),男,博士,副教授,主要研究方向为人工神经网络、互连网络等,E-mail:yuxin21@126.com;唐利霞(1987—),女,硕士,主要研究方向为人工神经网络,E-mail:tanlixia23@163.com(通信作者);于 琰(1986—),女,硕士,主要研究方向为智能计算。

行了分析,为全局搜索算法和局部优化算法的组合算法的局部优化提供了理论支持;可变学习率是解决神经网络收敛速度慢的一个有效方法,文献[13]中分析了前馈神经网络可变学习率的梯度算法的收敛性;惩罚项不仅可以提高神经网络的收敛速度,同时也可以提高神经网络的通用性,文献[14]中分析了带惩罚项的梯度算法的收敛性;文献[15]中提出了 PSNN 乘法随机单点在线梯度算法,利用乘法来克服因初始权值选取不当而导致的收敛速度过慢的问题;文献[16]给出了 PSNN 带动量项的异步梯度算法的收敛性,然而带动量项的异步梯度算法应用于 RPNN 的收敛性问题还没有得到讨论。

3 Ridge Polynomial 神经网络及其带动量项的异步梯度算法

Ridge Polynomial 神经网络是由 Pi-Sigma 神经网络组成的,接下来首先介绍 Pi-Sigma 神经网络。

3.1 Pi-Sigma 神经网络

Pi-Sigma 神经网络是一种具有单隐含层和单输出的前馈神经网络。图 1 展示了 k 阶单输出的 Pi-Sigma 神经网络的拓扑图,从图中可以看出隐含层到输出层之间的权值固定为 1,该性质减少了网络的训练时间。

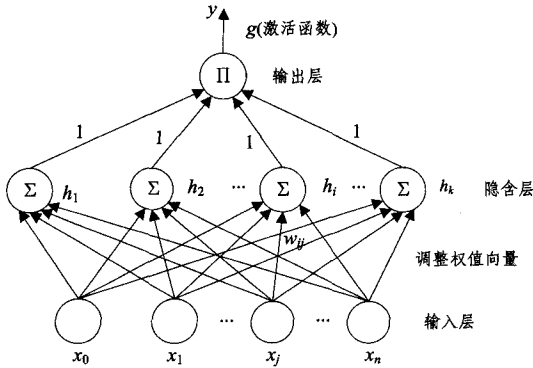


图 1 k 阶 Pi-Sigma 神经网络

设输入样本为 $x = (x_0, x_1, \dots, x_j, \dots, x_n) \in R^{n+1}$, 其中 $x_n = -1$ 为其阈值, $w_i = (w_{i0}, w_{i1}, \dots, w_{ij}, \dots, w_{in})^T$ 为连接输入层各结点与隐含层第 i 个结点的权值向量, 其中 w_{ij} 为连接隐含层第 i 个结点与输入层第 j 个结点的连接权值, 则 h_i 单元的 outputs 为:

$$h_i = w_i \cdot x = \sum_{j=0}^n w_{ij} x_j, i=1, 2, \dots, k \quad (1)$$

PSNN 最后的输出为:

$$y_1 = g(\sum_{i=1}^k h_i) \quad (2)$$

式中, $g(\cdot)$ 为激活函数。

3.2 Ridge Polynomial 神经网络带动量项的异步梯度算法

Ridge Polynomial 神经网络是由 Pi-Sigma 神经网络组成的高阶神经网络。图 2 展示的是以 k 个 PSNN 为模板组成的 Ridge Polynomial 神经网络, 它的隐含层的每个神经元即为一个 PSNN, 记 $PSNN_i$ 为第 i 个 PSNN, 其相应地含有 i 个隐含求和结点。

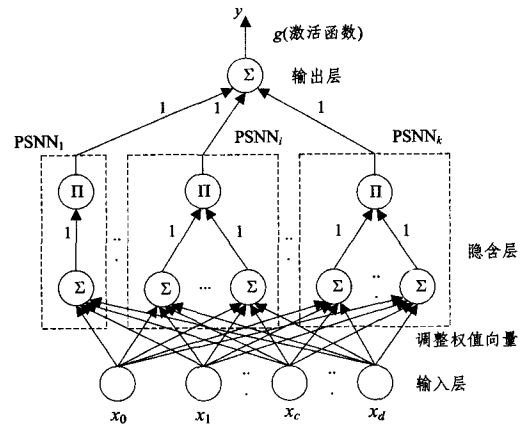


图 2 k 阶 Ridge Polynomial 神经网络

假定 RPNN 的输入层、隐含层和输出层的结点数分别为: $d+1, k$ 和 1。记 $w_{ij} = (w_{ij0}, w_{ij1}, \dots, w_{ijc}, \dots, w_{ijd})^T \in R^{d+1}$ 为连接输入层各结点与 $PSNN_i$ 中的第 $j - \frac{i(i-1)}{2}$ 个求和结点的权值向量, 则有:

$$\begin{aligned} w^m &= ((w_{11}^m)^T, (w_{21}^m)^T, \dots, (w_{k1}^m)^T)^T \\ &= ((w_{11}^m)^T, ((w_{22}^m)^T, (w_{23}^m)^T)^T, \dots, ((w_{k(k-\frac{1}{2}+1)}^m)^T, \dots, \\ &\quad (w_{k(k-\frac{1}{2}+1)}^m)^T)^T)^T \end{aligned} \quad (3)$$

记 $P_i(x)$ 为 $PSNN_i$ 的输出, 则 RPNN 的输出为:

$$y = g(\sum_{i=1}^k P_i(x)) \quad (4)$$

其中

$$P_i(x) = \prod_{j=\frac{i(i-1)}{2}+1}^{\frac{i(i+1)}{2}} h_{ij}, i=1, 2, \dots, k \quad (5)$$

$$h_{ij} = \sum_{c=0}^d w_{ijc} x_c \quad (6)$$

式中, h_{ij} 为 $PSNN_i$ 的第 $j - \frac{i(i-1)}{2}$ 个求和结点的输出, w_{ijc} 为第 c 个输入结点与 $PSNN_i$ 的第 $j - \frac{i(i-1)}{2}$ 个隐含求和结点之间的连接权值, $g(\cdot)$ 为激活函数, $P_i(x)$ 作为 $PSNN_i$ 的线性输出, 适用于 PSNN 的异步梯度算法, 也将适用于 RPNN。

设样本集为 $(x^a, O^a)_{a=1}^A \in R^{n+1} \times R$, 其中 x^a 为样本输入, O^a 为期望的样本输出, 定义 RPNN 输出的平方误差函数为:

$$E(w) = \frac{1}{2} \sum_{a=1}^A (O^a - y^a)^2 = \frac{1}{2} \sum_{a=1}^A (O^a - g(\sum_{i=1}^k P_i(x^a)))^2 \quad (7)$$

记

$$g_a(\sum_{i=1}^k P_i(x^a)) = \frac{1}{2} (O^a - g(\sum_{i=1}^k P_i(x^a)))^2 \quad (8)$$

由式(7)和式(8)得:

$$E(w) = \sum_{a=1}^A g_a(\sum_{i=1}^k P_i(x^a)) \quad (9)$$

对于任意取定的初始权值 w^0 , 假设异步迭代过程的第 m ($m=1, 2, \dots$) 轮迭代中, 隐层神经元的求和层的前 n ($n=0, \dots, \frac{k(k+1)}{2}-1$) 个隐含求和结点对应的权值已经调整更新, 将此时的网络误差函数值称为瞬时网络误差, 定义为:

$$E^n(w^m) = \sum_{a=1}^A g_a(\sum_{i=1}^k P_i^m(x^a)) \quad (10)$$

其中

$$\sum_{i=1}^k P_i^{n,m}(x^a) = \sum_{i=1}^{q-1} P_i^{m+1}(x^a) + \sum_{i=q+1}^k P_i^m(x^a) + \prod_{j=\frac{q(q-1)}{2}+1}^{\frac{q(q+1)}{2}} P_i^m(x^a) \quad (11)$$

$$(w_{q,n+1}^{m+1} \cdot x^a) \prod_{j=n+1}^{\frac{q(q+1)}{2}} (w_{q,n+1}^m \cdot x^a) \quad (11)$$

特别地,定义 $\prod_{j=\frac{q(q+1)}{2}+1}^{\frac{q(q+1)}{2}} (w_{q,n+1}^m \cdot x^a) = 1, \sum_{i=1}^0 P_i^{m+1}(x^a) = 0, q = 1, 2, \dots, k, a = 1, 2, \dots, A$.

使用梯度算法来训练网络的最终目的,是找到一个 w^* , 使 $E(w^*) = \min E(w)$, 此时,瞬时误差函数 $E^n(w^m)$ 对权向量 $w_{q(n+1)}^m$ 求梯度:

$$\nabla_{q(n+1)} E^n(w^m) = \sum_{a=1}^A g_a' \left(\sum_{i=1}^k P_i^{n,m}(x^a) \right) \prod_{j=\frac{q(q-1)}{2}+1}^{\frac{q(q+1)}{2}} (w_{q,n+1}^{m+1} \cdot x^a) \prod_{j=n+2}^{\frac{q(q+1)}{2}} (w_{q,n+1}^m \cdot x^a) \quad (12)$$

在训练过程中,第 m 次迭代的权值更新为:

$$w_{q(n+1)}^{m+1} = w_{q(n+1)}^m + \Delta w_{q(n+1)}^m \quad (13)$$

$$m = 0, 1, 2, \dots, n = \frac{q(q-1)}{2}, \dots, \frac{q(q+1)}{2} - 1$$

式中, $\Delta w_{q(n+1)}^m$ 是权值改变量。常用的异步梯度算法的权值改变量取为:

$$\Delta w_{q(n+1)}^m = -\eta \nabla_{q(n+1)} E^n(w^m)$$

式中, η 是学习速率,带动量项异步梯度算法的权值改变量取为

$$\Delta w_{q(n+1)}^m = -\eta \nabla_{q(n+1)} E^n(w^m) + \tau_{q(n+1)}^m \Delta w_{q(n+1)}^{m-1} \quad (14)$$

式中, $\tau_{q(n+1)}^m$ 是动量项系数,满足

$$\tau_{q(n+1)}^m = \begin{cases} \min(\tau, \frac{\mu \|\nabla_{q(n+1)} E^n(w^m)\|}{\|\Delta w_{q(n+1)}^{m-1}\|}), & \Delta w_{q(n+1)}^{m-1} \neq 0 \\ 0, & \text{其他} \end{cases} \quad (15)$$

式中, μ 和 τ 是常数,满足 $0 < \tau < 1, \mu > 0$ 。在带动量项的异步梯度算法的迭代过程中,起主要作用的仍是梯度项,动量项的作用是加速收敛,所以动量因子一般不宜过大。

4 相关的引理及其证明

首先给出论述中所需的假设:

假设 1 对于 $\forall t \in R$, 都存在一个常数 $C_1 > 0$, 使得

$$\max\{|g_a'(t)|, |g_a''(t)|\} \leq C_1$$

假设 2 存在常数 $C_2, C_3 > 0$ 满足

$$|(w_{ij}, x^a)| \leq C_2, |x^a| \leq C_3, 1 \leq i \leq k, \frac{i(i-1)}{2} + 1 \leq j \leq \frac{i(i+1)}{2}, 1 \leq a \leq A$$

$$\frac{i(i+1)}{2}, 1 \leq a \leq A$$

假设 3 学习速率 η 和参数 μ 足够小,且满足

$$0 < \eta < \frac{2}{AC_1 C_2^{(k-1)} C_3^2}$$

$$0 < \mu < \frac{-1 - AC_1 C_2^{(k-1)} C_3^2 \eta + \sqrt{1 + 4AC_1 C_2^{(k-1)} C_3^2 \eta}}{AC_1 C_2^{(k-1)} C_3^2 \eta}$$

假设 4 序列 $\{w^m\}_{m=0}^\infty$ 包含在紧集 D 中,且 D 中的临界点集是有限的。

引理 1 若假设 1—假设 3 满足,序列 $\{w^m\}_{m=0}^\infty$ 由异步梯度算法式(13)和式(14)生成,则有

$$E(w^{m+1}) \leq E(w^m) - (\eta - \mu - \frac{1}{2} AC_1 C_2^{(k-1)} C_3^2 (\eta + \mu)^2)$$

$$\sum_{q=1}^k \sum_{n=\frac{q(q-1)}{2}}^{\frac{q(q+1)}{2}-1} \|\nabla_{q(n+1)} E^n(w^m)\|^2 \quad (16)$$

成立。

证明:为了简便,记

$$\varphi_{a,k}^{n,m} = \sum_{i=1}^k P_i^{n,m}(x^a) \quad (17)$$

$$\varphi_{a,k}^{n+1,m} = \varphi_{a,k}^{n+1,m} - \varphi_{a,k}^{n,m}$$

$$T_{q,a}^{n,m+1} = \prod_{j=\frac{q(q-1)}{2}+1}^{\frac{q(q+1)}{2}} (w_{q,n+1}^{m+1} \cdot x^a) \quad (18)$$

$$\bar{T}_{q,a}^{n,m} = \prod_{j=n}^{\frac{q(q+1)}{2}} (w_{q,n}^m \cdot x^a)$$

替换式(12)和式(14)得

$$\nabla_{q(n+1)} E^n(w^m) = \sum_{a=1}^A g_a'(\varphi_{a,k}^{n,m}) T_{q,a}^{n,m+1} \bar{T}_{q,a}^{n+2,m} x^a \quad (19)$$

和

$$\begin{aligned} \Delta w_{q(n+1)}^m &= -\eta \nabla_{q(n+1)} E^n(w^m) + \tau_{q(n+1)}^m \Delta w_{q(n+1)}^{m-1} \\ &= -\eta \sum_{a=1}^A g_a'(\varphi_{a,k}^{n,m}) T_{q,a}^{n,m+1} \bar{T}_{q,a}^{n+2,m} x^a + \tau_{q(n+1)}^m \Delta w_{q(n+1)}^{m-1} \end{aligned} \quad (20)$$

由式(17)、式(11)和式(18)可得

$$\text{当 } \frac{q(q-1)}{2} + 1 \leq n < \frac{q(q+1)}{2} \text{ 时}$$

$$\begin{aligned} \varphi_{a,k}^{n+1,m} - \varphi_{a,k}^{n,m} &= \sum_{i=1}^k P_i^{n+1,m}(x^a) - \sum_{i=1}^k P_i^{n,m}(x^a) \\ &= \prod_{j=\frac{q(q-1)}{2}+1}^{n+1} (w_{q,n+1}^{m+1} \cdot x^a) \prod_{j=n+2}^{\frac{q(q+1)}{2}} (w_{q,n+1}^m \cdot x^a) - \prod_{j=\frac{q(q-1)}{2}+1}^n (w_{q,n}^{m+1} \cdot x^a) \prod_{j=n+1}^{\frac{q(q+1)}{2}} (w_{q,n}^m \cdot x^a) \\ &= (T_{q,a}^{n+1,m+1} \bar{T}_{q,a}^{n+2,m} - T_{q,a}^{n,m+1} \bar{T}_{q,a}^{n+1,m}) \\ &= T_{q,a}^{n,m+1} \bar{T}_{q,a}^{n+2,m} (w_{q(n+1)}^{m+1} - w_{q(n+1)}^m) \cdot x^a \end{aligned} \quad (21)$$

$$\text{当 } n = \frac{q(q+1)}{2} \text{ 时}$$

$$\begin{aligned} \varphi_{a,k}^{n+1,m} - \varphi_{a,k}^{n,m} &= \sum_{i=1}^k P_i^{n+1,m}(x^a) - \sum_{i=1}^k P_i^{n,m}(x^a) \\ &= \sum_{i=1}^q P_i^{m+1}(x^a) + \prod_{j=\frac{q(q+1)}{2}+1}^{n+1} (w_{q(n+1)}^{m+1} \cdot x^a) \prod_{j=n+2}^{\frac{(q+2)(q+1)}{2}} (w_{q(n+1)}^m \cdot x^a) \\ &\quad + \prod_{j=\frac{q(q+1)}{2}+1}^n (w_{q,n+1}^{m+1} \cdot x^a) \prod_{j=n+1}^{\frac{q(q+1)}{2}} (w_{q,n+1}^m \cdot x^a) + \sum_{i=q+1}^k P_i^m(x^a) \\ &= (T_{q+1,a}^{n+1,m+1} \bar{T}_{q+1,a}^{n+2,m} - P_{q+1}^m(x^a)) \\ &= (T_{q+1,a}^{n+1,m+1} \bar{T}_{q+1,a}^{n+2,m} - T_{q+1,a}^{n+1,m} \bar{T}_{q+1,a}^{n+2,m}) \\ &= \bar{T}_{q+1,a}^{n+2,m} (w_{q(n+1)}^{m+1})_{(n+1)} - w_{q(n+1)}^{m+1} \cdot x^a \\ &= T_{q+1,a}^{n+1,m+1} \bar{T}_{q+1,a}^{n+2,m} (w_{q(n+1)}^{m+1})_{(n+1)} - w_{q(n+1)}^{m+1} \cdot x^a \end{aligned} \quad (22)$$

对每个 $1 \leq a \leq A$ 和 $0 \leq n \leq \frac{k(k+1)}{2} - 1$, 利用泰勒定理,

在点 $\varphi_{a,k}^{n+1,m}$ 处对函数 $g_a(\varphi_{a,k}^{n+1,m})$ 展开,有:

$$g_a(\varphi_{a,k}^{n+1,m}) = g_a(\varphi_{a,k}^{n,m}) + g_a'(\varphi_{a,k}^{n,m})(\varphi_{a,k}^{n+1,m} - \varphi_{a,k}^{n,m}) + \frac{1}{2} g_a''(\varphi_{a,k}^{n,m})(\varphi_{a,k}^{n+1,m} - \varphi_{a,k}^{n,m})^2 \quad (23)$$

式中, $\varphi_{a,k}^{n+1,m}$ 介于 $\varphi_{a,k}^{n+1,m}$ 与 $\varphi_{a,k}^{n,m}$ 之间。

取 $a=1, 2, \dots, A$, 对式(23)累加, 由假设 1、假设 2、式(22)、式(21)、式(19)和式(20)得

当 $\frac{q(q-1)}{2} + 1 \leq n < \frac{q(q+1)}{2}$ 时

$$\sum_{a=1}^A g_a(\varphi_{a,k}^{n+1,m}) = \sum_{a=1}^A g_a(\varphi_{a,k}^{n,m}) - \eta \|\nabla_{q(n+1)} E^n(w^m)\|^2 + \delta_1 + \delta_2 \quad (24)$$

其中

$$\delta_1 = \tau_{q(n+1)}^m \Delta w_{q(n+1)}^{m-1} \cdot \nabla_{q(n+1)} E^n(w^m)$$

$$\delta_2 = \sum_{a=1}^A \frac{1}{2} g_a''(\varphi_{a,k}^{n,m})(\varphi_{a,k}^{n,m})^2$$

由假设 1、假设 2、式(15)、式(20)和 Cauchy-Schwartz 不等式, 得

$$\begin{aligned} \|\delta_1\| &\leq \mu \|\nabla_{q(n+1)} E^n(w^m)\|^2 \\ \|\delta_2\| &\leq \frac{1}{2} AC_1 C_2^{q-1} C_3^q (\eta \|\nabla_{q(n+1)} E^n(w^m)\| + \tau_{q(n+1)}^m \|\Delta w_{q(n+1)}^{m-1}\|)^2 \\ &\leq \frac{1}{2} AC_1 C_2^{q-1} C_3^q (\eta + \mu)^2 \|\nabla_{q(n+1)} E^n(w^m)\|^2 \end{aligned}$$

代入式(24)

$$\begin{aligned} \sum_{a=1}^A g_a(\varphi_{a,k}^{n+1,m}) - \sum_{a=1}^A g_a(\varphi_{a,k}^{n,m}) &\leq -\eta \|\nabla_{q(n+1)} E^n(w^m)\|^2 + \mu \|\nabla_{q(n+1)} E^n(w^m)\|^2 + \\ &\quad \frac{1}{2} AC_1 C_2^{q-1} C_3^q (\eta + \mu)^2 \|\nabla_{q(n+1)} E^n(w^m)\|^2 \\ &\leq -(\eta - \mu - \frac{1}{2} AC_1 C_2^{q-1} C_3^q (\eta + \mu)^2) \|\nabla_{q(n+1)} E^n(w^m)\|^2 \end{aligned} \quad (25)$$

当 $n = \frac{q(q+1)}{2}$ 时

$$\sum_{a=1}^A g_a(\varphi_{a,k}^{n+1,m}) = \sum_{a=1}^A g_a(\varphi_{a,k}^{n,m}) - \eta \|\nabla_{(q+1)(n+1)} E^n(w^m)\|^2 + \delta_3 + \delta_4 \quad (26)$$

其中

$$\delta_3 = \tau_{(q+1)(n+1)}^m \Delta w_{(q+1)(n+1)}^{m-1} \cdot \nabla_{(q+1)(n+1)} E^n(w^m)$$

$$\delta_4 = \sum_{a=1}^A \frac{1}{2} g_a''(\varphi_{a,k}^{n,m})(\varphi_{a,k}^{n,m})^2$$

由假设 1、假设 2、式(15)、式(20)和 Cauchy-Schwartz 不等式, 得

$$\begin{aligned} \|\delta_3\| &\leq \mu \|\nabla_{(q+1)(n+1)} E^n(w^m)\|^2 \\ \|\delta_4\| &\leq \frac{1}{2} AC_1 C_2^q C_3^q (\eta \|\nabla_{(q+1)(n+1)} E^n(w^m)\| + \tau_{(q+1)(n+1)}^m \|\Delta w_{(q+1)(n+1)}^{m-1}\|)^2 \\ &\leq \frac{1}{2} AC_1 C_2^q C_3^q (\eta + \mu)^2 \|\nabla_{(q+1)(n+1)} E^n(w^m)\|^2 \end{aligned}$$

代入式(26)得

$$\begin{aligned} \sum_{a=1}^A g_a(\varphi_{a,k}^{n+1,m}) - \sum_{a=1}^A g_a(\varphi_{a,k}^{n,m}) &\leq -\eta \|\nabla_{(q+1)(n+1)} E^n(w^m)\|^2 + \mu \|\nabla_{(q+1)(n+1)} E^n(w^m)\|^2 + \\ &\quad \frac{1}{2} AC_1 C_2^q C_3^q (\eta + \mu)^2 \|\nabla_{(q+1)(n+1)} E^n(w^m)\|^2 \\ &\leq -(\eta - \mu - \frac{1}{2} AC_1 C_2^q C_3^q (\eta + \mu)^2) \|\nabla_{(q+1)(n+1)} E^n(w^m)\|^2 \end{aligned} \quad (27)$$

由式(10)、式(25)和式(27)可推得

$$\begin{aligned} E(w^{m+1}) &= \sum_{a=1}^A g_a(\varphi_{a,k}^{\frac{k(k+1)}{2},m}) - \sum_{a=1}^A g_a(\varphi_{a,k}^{\frac{k(k+1)}{2}-1,m}) + \\ &\quad \sum_{a=1}^A g_a(\varphi_{a,k}^{\frac{k(k+1)}{2}-1,m}) - \sum_{a=1}^A g_a(\varphi_{a,k}^{\frac{k(k+1)}{2}-2,m}) + \dots + \\ &\quad \sum_{a=1}^A g_a(\varphi_{a,k}^{1,m}) - \sum_{a=1}^A g_a(\varphi_{a,k}^{0,m}) + \sum_{a=1}^A g_a(\varphi_{a,k}^{0,m}) \\ &\leq -(\eta - \mu - \frac{1}{2} AC_1 C_2^{q-1} C_3^q (\eta + \mu)^2) \sum_{n=\frac{k(k-1)}{2}}^{\frac{k(k+1)}{2}-1} \|\nabla_{k(n+1)} E^n(w^m)\|^2 - (\eta - \mu - \frac{1}{2} AC_1 C_2^{q-2} C_3^q (\eta + \mu)^2) \sum_{n=\frac{(k-1)(k-2)}{2}}^{\frac{k(k-1)}{2}-1} \|\nabla_{(k-1)(n+1)} E^n(w^m)\|^2 + \dots + \\ &\quad -(\eta - \mu - \frac{1}{2} AC_1 C_2^q C_3^q (\eta + \mu)^2) \sum_{q=1}^k \|\nabla_{1(n+1)} E^n(w^m)\|^2 + E(w^m) \\ &\leq E(w^m) - (\eta - \mu - \frac{1}{2} AC_1 C_2^{q-1} C_3^q (\eta + \mu)^2) \sum_{q=1}^k \sum_{n=\frac{q(q-1)}{2}}^{\frac{q(q+1)}{2}-1} \|\nabla_{q(n+1)} E^n(w^m)\|^2 \end{aligned} \quad (28)$$

引理 1 得证。

引理 2 设 $f: R^n \rightarrow R$ 在紧集 D 上连续可微, 假定 f 在 D 中的临界点集是有限的, 设序列 $\{x_k\} \subset D$ 满足

$$\lim_{k \rightarrow \infty} \|x_{k+1} - x_k\| = 0, \lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$$

那么, 存在一个 x^* ,

$$\lim_{k \rightarrow \infty} x_k = x^*$$

注: 此引理在文献[17]中已得到证明, 即文献[17]中的引理 3.5.10。

5 主要的收敛性定理及证明

定理 1(单调性) 若假设 1—假设 3 满足, 序列 $\{w^m\}_0^\infty$ 由异步梯度算法式(13)和式(14)生成, 则有

$$E(w^{m+1}) \leq E(w^m), m=0, 1, 2, \dots \quad (29)$$

证明:

由假设 3 知:

$$\eta - \mu - \frac{1}{2} AC_1 C_2^{q-1} C_3^q (\eta + \mu)^2 > 0 \quad (30)$$

由式(30)再结合引理 1 可得定理 1 成立。

定理 2(弱收敛性) 若假设 1—假设 3 满足, 序列 $\{w^m\}$ 由带动量项的异步梯度算法生成, 则有

$$\lim_{m \rightarrow \infty} \|\nabla E(w^m)\| = 0 \quad (31)$$

证明:

由引理 1 的式(16), 取 $C_0 = \eta - \mu - \frac{1}{2} AC_1 C_2^{q-1} C_3^q (\eta + \mu)^2 > 0$, 对于整数 $M > 0$, 有

$$\begin{aligned} E(w^M) &\leq E(w^{M-1}) - C_0 \sum_{q=1}^k \sum_{n=\frac{q(q-1)}{2}}^{\frac{q(q+1)}{2}-1} \|\nabla_{q(n+1)} E^n(w^{M-1})\|^2 \\ &\leq \dots \leq E(w^0) - C_0 \sum_{m=1}^M \sum_{q=1}^k \sum_{n=\frac{q(q-1)}{2}}^{\frac{q(q+1)}{2}-1} \|\nabla_{q(n+1)} E^n(w^{m-1})\|^2 \end{aligned} \quad (32)$$

对 $\forall M > 0$, 有

$$E(w^M) \geq 0 \quad (33)$$

由式(32)和式(33)得

$$C_0 \sum_{m=1}^M \sum_{q=1}^k \sum_{n=\frac{q(q-1)}{2}}^{\frac{q(q+1)}{2}-1} \|\nabla_{q(n+1)} E^n(w^{m-1})\|^2$$

$$\leq E(w^0) - E(w^M) \leq E(w^0) < \infty \quad (34)$$

令 $M \rightarrow \infty$, 则

$$\sum_{m=1}^{\infty} \sum_{q=1}^k \sum_{n=\frac{q(q-1)}{2}}^{\frac{q(q+1)}{2}-1} \|\nabla_{q(n+1)} E^n(w^{m-1})\|^2 < \infty \quad (35)$$

于是可得

$$\lim_{m \rightarrow \infty} \sum_{q=1}^k \sum_{n=\frac{q(q-1)}{2}}^{\frac{q(q+1)}{2}-1} \|\nabla_{q(n+1)} E^n(w^{m-1})\|^2 < \infty \quad (36)$$

对 $\frac{q(q-1)}{2} \leq n \leq \frac{q(q+1)}{2} - 1, q=1, 2, \dots, k$, 由式(19)和

式(20)可得

$$\begin{aligned} & \|w_{q(n+1)}^m - w_{q(n+1)}^{m-1}\| \\ &= \|\Delta w_{q(n+1)}^{m-1}\| \\ &\leq \eta \|\nabla_{q(n+1)} E^n(w^{m-1})\| + \tau_{q(n+1)}^{m-1} \|\Delta w_{q(n+1)}^{m-2}\| \\ &\leq \eta \|\nabla_{q(n+1)} E^n(w^{m-1})\| + \mu \|\nabla_{q(n+1)} E^n(w^{m-1})\| \\ &= (\eta + \mu) \|\nabla_{q(n+1)} E^n(w^{m-1})\| \end{aligned} \quad (37)$$

再由式(36)可得

$$\begin{aligned} \lim_{m \rightarrow \infty} \|w_{q(n+1)}^m - w_{q(n+1)}^{m-1}\| &= \lim_{m \rightarrow \infty} \|\Delta w_{q(n+1)}^{m-1}\| \\ &= \lim_{m \rightarrow \infty} \|\nabla_{q(n+1)} E^n(w^{m-1})\| = 0 \end{aligned} \quad (38)$$

因此

$$\begin{aligned} \lim_{m \rightarrow \infty} \|w^m - w^{m-1}\| &= \lim_{m \rightarrow \infty} \|((w_1^m - w_1^{m-1})^T, \dots, (w_k^m - w_k^{m-1})^T, \dots, (w_k^m - w_k^{m-1})^T)^T\| \\ &= \lim_{m \rightarrow \infty} \|((w_{11}^m - w_{11}^{m-1})^T, \dots, \\ &\quad ((w_{k(\frac{k-1}{2}+1)}^m - w_{k(\frac{k-1}{2}+1)}^{m-1})^T, \dots, \\ &\quad (w_{k(\frac{k+1}{2}}^m - w_{k(\frac{k+1}{2}}^{m-1})^T)^T)^T\| = 0 \end{aligned} \quad (39)$$

下面由假设 1、假设 2 和 Cauchy-Schwartz 不等式, 依据中值定理导出 3 个用于证明定理 2 的重要不等式。

$$\begin{aligned} & |\bar{T}_{q,a}^{n+1,m} - \bar{T}_{q,a}^{n+1,m-1}| \\ &= \left| \sum_{h=n+1}^{\frac{q(q+1)}{2}} \left(\prod_{j=n+1}^{\frac{q(q+1)}{2}} (w_{q,h}^m \cdot x^a) (\Delta w_{q,h}^{m-1} \cdot x^a) \right) \right| \\ &\leq C_2^{\frac{q(q+1)}{2}-n-1} C_3 \sum_{h=n+1}^{\frac{q(q+1)}{2}} \|\Delta w_{q,h}^{m-1}\| \\ &|\varphi_{a,k}^{0,m} - \varphi_{a,k}^{n,m-1}| \\ &= \left| \sum_{i=1}^k P_i^{0,m}(x^a) - \sum_{i=1}^k P_i^{n,m-1}(x^a) \right| \\ &= \left| T_{q,a}^{n,m} (\bar{T}_{q,a}^{n+1,m} - \bar{T}_{q,a}^{n+1,m-1}) + \sum_{i=q+1}^k (P_i^n(x^a) - P_i^{m-1}(x^a)) \right| \\ &= \left| T_{q,a}^{n,m} (\bar{T}_{q,a}^{n+1,m} - \bar{T}_{q,a}^{n+1,m-1}) + \sum_{i=q+1}^k (\bar{T}_{i,a}^{\frac{i(i-1)}{2}+1,m} - \bar{T}_{i,a}^{\frac{i(i-1)}{2}+1,m-1}) \right| \\ &\leq C_2^{q-1} C_3 \sum_{h=n+1}^{\frac{q(q+1)}{2}} \|\Delta w_{q,h}^{m-1}\| + C_2^{q-1} C_3 \sum_{i=q+1}^k \sum_{h=\frac{i(i-1)}{2}+1}^{\frac{i(i+1)}{2}} \|\Delta w_{q,h}^{m-1}\| \\ &|g_a'(\varphi_{a,k}^{0,m}) - g_a'(\varphi_{a,k}^{n,m-1})| \\ &= |g_a''(\tilde{\varphi}_{a,k}^{n,m})(\varphi_{a,k}^{0,m} - \varphi_{a,k}^{n,m-1})| \\ &\leq C_1 C_2^{q-1} C_3 \sum_{h=n+1}^{\frac{q(q+1)}{2}} \|\Delta w_{q,h}^{m-1}\| + C_1 C_2^{q-1} C_3 \sum_{i=q+1}^k \sum_{h=\frac{i(i-1)}{2}+1}^{\frac{i(i+1)}{2}} \|\Delta w_{q,h}^{m-1}\| \end{aligned}$$

其中 $\tilde{\varphi}_{a,k}^{n,m}$ 介于 $\varphi_{a,k}^{0,m}$ 和 $\varphi_{a,k}^{n,m-1}$ 之间, 故由以上 3 式推得

$$\begin{aligned} & \|\nabla_{q(n+1)} E(w^m) - \nabla_{q(n+1)} E^n(w^{m-1})\| \\ &= \sum_{a=1}^A \|g_a'(\varphi_{a,k}^{0,m}) T_{q,a}^{n,m} (\bar{T}_{q,a}^{n+2,m} - \bar{T}_{q,a}^{n+2,m-1}) x^a + (g_a'(\varphi_{a,k}^{0,m}) - g_a'(\varphi_{a,k}^{n,m-1})) T_{q,a}^{n,m} \bar{T}_{q,a}^{n+2,m-1} x^a\| \\ &\leq AC_1 C_2^{q-2} C_3 \sum_{h=n+1}^{\frac{q(q+1)}{2}} \|\Delta w_{q,h}^{m-1}\| + AC_1 C_2^{q-2} C_3 \sum_{h=n+1}^{\frac{q(q+1)}{2}} \|\Delta w_{q,h}^{m-1}\| \\ &\quad + AC_1 C_2^{q-2} C_3 \sum_{i=q+1}^k \sum_{h=\frac{i(i-1)}{2}+1}^{\frac{i(i+1)}{2}} \|\Delta w_{q,h}^{m-1}\| \end{aligned} \quad (40)$$

结合式(38)得

$$\lim_{m \rightarrow \infty} \|\nabla_{q(n+1)} E(w^m) - \nabla_{q(n+1)} E^n(w^{m-1})\| = 0$$

进而

$$\lim_{m \rightarrow \infty} \|\nabla_{q(n+1)} E(w^m)\| = \lim_{m \rightarrow \infty} \|\nabla_{q(n+1)} E^n(w^{m-1})\| = 0 \quad (41)$$

于是, 得

$$\begin{aligned} & \lim_{m \rightarrow \infty} \|\nabla E(w^m)\| \\ &= \lim_{m \rightarrow \infty} \|((\nabla_1 E(w^m))^T, (\nabla_2 E(w^m))^T, \dots, (\nabla_k E(w^m))^T)^T\| \\ &= \lim_{m \rightarrow \infty} \|((\nabla_{11} E(w^m))^T, \dots, (D_{k(\frac{k-1}{2}+1)} E(w^m))^T, \dots, (D_{k(\frac{k+1}{2}} E(w^m))^T)^T)^T\| = 0 \end{aligned} \quad (42)$$

定理 2 得证。

定理 3(强收敛性) 若定理 2 成立, 再若假设 4 满足, 则存在 $w^* \in D$, 使

$$\lim_{m \rightarrow \infty} w^m = w^* \quad (43)$$

证明: 令引理 2 中的 $x \approx w$, f 为梯度函数, $f(x) \approx \nabla E(w)$, 由假设 1 知 $\nabla E(w)$ 在 D 上连续可微, 结合式(39)、式(42)、假设 4 和引理 2, 定理 3 的强收敛性得证。

6 实验结果与分析

4 维奇偶问题是一类较难的分类问题, 该问题包含 2^4 个样本点, 每个样本点由取值为 -1 或 1 的 4 维向量构成。整个样本空间被划分为两类, 包含偶数个 1 的样本为一类, 包含奇数个 1 的样本为另一类。

为了证明带动量项的异步梯度算法的收敛性质和算法的高效性, 我们以求解 4 维奇偶分类问题为例, 分别用带动量项的异步梯度算法和异步梯度算法进行数值实验, 对比两种算法的学习效率。表 1 给出了 4 维奇偶问题的输入输出关系, 实验中将 4 维的输入数据通过输入结点输入网络获得网络的实际输出, 通过网络的自学习功能来调整网络的权值向量, 使得网络的实际输出与期望输出的误差逐渐变小, 直至收敛。

表 1 4 维奇偶问题

输入模式				输出分类			
1	1	1	1	0	-1	1	-1
-1	1	1	1	1	1	-1	-1
1	1	1	-1	1	1	-1	1
-1	1	1	-1	0	-1	-1	1
-1	-1	-1	1	1	1	-1	1
1	-1	1	-1	0	-1	-1	1
-1	-1	-1	-1	0	1	1	-1
1	-1	-1	-1	1	-1	1	-1

在该实验中, 取 Ridge Polynomial 神经网络的输入层、隐

层和输出层的结点数为 5,4,1。网络中输出层的激活函数取为 $g(x) = \frac{1}{1+e^{-x}}$, 随机生成如表 1 所列的 80 个输入样本, 初始权值 w^0 在 $[-0.2, 0.2]$ 内随机选取。当训练迭代次数达到 15000 或目标误差值小于 10^{-5} 时训练终止。选取学习速率 $\eta=0.001$, 动量参数 $\tau=0.9$, μ 分别选取为 0.001, 0.0015, 0.002, 可以验证实验设置的参数取值满足假设 1—假设 4。对于不同的 μ 的取值各做 50 次实验。实验结果如表 2 所列。

表 2 对比实验结果

μ 的取值	不带动量项的平均迭代次数	带动量项的平均迭代次数
$\mu=0.001$	14715	7233
$\mu=0.0015$	14668	5834
$\mu=0.002$	14413	4562

从表 2 可以看出对于 μ 取不同的值, 不带动量项的异步梯度算法的平均迭代次数远远大于带动量项的异步梯度算法的平均迭代次数, 由此可说明带动量项的异步梯度算法比不带动量项的异步梯度算法更高效。同时可以看出 μ 取值较大时, 收敛速度较快, 但 μ 的取值不应超过假设 3 给定的界限。

为了说明带动量项的异步梯度算法的收敛性质, 给出了当 $\eta=0.001$, $\tau=0.9$, $\mu=0.002$ 时误差函数变化和权值调整的曲线图, 如图 3 和图 4 所示。

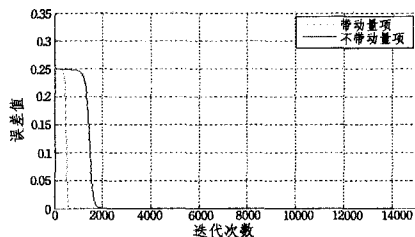


图 3 带动量项和不带动量项算法的误差值变化曲线

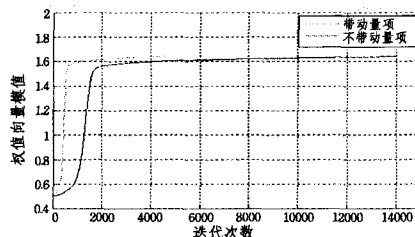


图 4 带动量项和不带动量项算法的权值向量模值变化曲线

图 3 展示了使用带动量项的异步梯度算法和异步梯度算法的误差变化, 可以看出随着训练次数的增长, 两种算法的误差函数值都单调下降, 且误差函数的切线最后是趋于 0, 实验结果与单调性与弱收敛性理论是一致的。但是对比可以看出加入动量项后的异步梯度算法在迭代 4500 多次后达到收敛, 而不加动量项的异步梯度算法需要迭代 14000 多次, 可见改进后算法的收敛速度更快, 效率更高。图 4 展示了两种算法的权值向量模值的变化趋势, 对比可以看出带动量项的异步梯度算法比不带动量项的异步梯度算法的权值收敛速度更快, 因此可说明改进后的算法更高效。从图中还可看到在迭代次数达到 2000 之前, 两种算法的权值矩阵的模值都随着训练次数的增长而迅速调整增大, 这点变化与图 3 是吻合的, 因为在这个阶段误差函数梯度值较大, 使得权值向量调整也比较大。权值矩阵的模值从迭代次数 2000 之后随着训练次数的增长变化较小并逐渐趋于稳定, 表明权值向量达到了误差

函数的极小点, 与定理 3 的强收敛性的结果一致。

结束语 本文提出了将动量项加入到神经网络异步梯度训练算法的方法, 分析了 RPNN 带动量项的异步梯度算法的收敛性。从仿真实验分析可以看出, 加入动量项后的异步梯度算法比不加动量项的异步梯度算法的收敛速度快, 实验结果也验证了文中得出的收敛性结论。

参考文献

- [1] Shin Y, Ghosh J. Approximation of Multivariate Functions Using Ridge Polynomial Networks [J]. Neural Networks, 1992, 2:380-385
- [2] Shin Y, Ghosh J. Ridge Polynomial Networks [J]. IEEE Transactions on Neural Networks, 1995, 6(3):610-622
- [3] Shin Y, Ghosh J. The pi-sigma network; an efficient higher-order neural network for pattern classification and function approximation [J]. IEEE Transactions on Neural Networks, 1991, 1: 13-18
- [4] Christodoulos V, Yiannis S B, Basil G M. Ridge polynomial networks in pattern recognition [C]// 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications, 2003, 519-524
- [5] Ghazali R, Hussian A J, Liatsis P. Dynamic ridge polynomial neural network; forecasting the univariate non-stationary and stationary trading signal [J]. Expert Systems with Applications, 2011, 38:3765-3776
- [6] Ghazali R, Hussian A J, Nawi N M, et al. Non-stationary and stationary prediction of financial time series using dynamic ridge polynomial neural network [J]. Neurocomputing, 2009, 72(10-12):2359-2367
- [7] Hacib T, Bihan Y L, Mekideche M R, et al. Ridge polynomial neural network for non-destructive eddy current evaluation [J]. Studies in Computational Intelligence, 2011, 327:185-199
- [8] Giles C L, Maxwell T. Learning, invariance, and generalization in a high-order neural network [J]. Applied Optics, 1987, 26(23):4972-4978
- [9] Rumelhart D E, McClelland J L. Parallel Distributed Processing Explorations in the microstructure of cognition [M]. Cambridge: MIT Press, 1986
- [10] Durbin R, Rumelhart D E. Product units: A computationally powerful and biologically plausible extension to backpropagation networks [J]. Neural Computation, 1989, 1:133-142
- [11] Zhang Chao, Wu Wei, Xiong Yan. Convergence Analysis of Batch Gradient Algorithm for Three Classes of Sigma-Pi Neural Networks [J]. Neural Processing Letters, 2007, 26(3):177-189
- [12] 张超, 李正学, 陈先华, 等. 用在线梯度法训练积单元神经网络的收敛性分析 [J]. 高等学校计算数学, 2010, 32(3):261-274
- [13] 邵红梅, 安凤仙. 一类训练前馈神经网络的梯度算法及收敛性 [J]. 中国石油大学学报, 2010, 34(4):176-180
- [14] Shao Hong-mei, Wu Wei, Liu Li-jun. Convergence of Online Gradient Method with Penalty for BP Neural Networks [J]. Communications in Mathematical Research, 2010, 26(1):67-75
- [15] 喻昕, 邓飞, 唐利霞. Pi-Sigma 神经网络的乘子法随机单点在线梯度算法 [J]. 计算机应用研究, 2011, 28(11):4074-4077
- [16] 熊焱, 张超. Pi-Sigma 神经网络的带动量项的异步批处理梯度算法收敛性 [J]. 应用数学, 2008, 21(1):217-212
- [17] 袁亚湘, 孙文瑜. 最优化理论与方法 [M]. 北京: 科学出版社, 2001