

学习率有限监督调整方法

李鸿雁¹, 刘宪亮², 鲍新华¹

(1. 吉林大学 环境与资源学院,长春 130026;2. 黄河水利职业技术学院,河南 开封 475003)

摘要:从学习率对BP网络误差下降曲线的影响机理入手,提出了能够使学习率具有自适应调整能力的有限监督调整方法,通过实证分析并结合实际工程算例阐明了该方法的有效性。

关键词:人工智能;学习率;有限监督调整;人工神经网络;BP网络

中图分类号:TP183 **文献标识码:**A **文章编号:**1671-5497(2007)04-0846-05

Limited supervising method for learning rate regulation

Li Hong-yan¹, Liu Xian-liang², Bao Xin-hua¹

(1. College of Environment and Resources, Jilin University, Changchun 130026, China; 2. Yellow River Conservancy Technical Institute, Kaifeng 475003, China)

Abstract: The mechanism of the impact of learning rate on the error-drop-curve of BP networks was studied. Then a limited supervising method for learning rate regulation was proposed, by which the learning rate of BP networks is capable for self-adaptive adjustment. This method was validated by demonstration study and engineering application.

Key words: artificial intelligence; learning rate; limited supervising to regulating; artificial neural network ;BP networks

0 引言

在人工神经网络训练中,学习率是影响训练速度和训练精度的重要因素之一。从网络权重的调整式(1)中可以看到,网络权重的调整量 $\Delta\mathbf{W}_j$,取决于网络的输入向量 \mathbf{X} 、网络的当前权重 \mathbf{W}_j 和学习率 l_r 三方面因素。下式中 Y_d 为网络的期望输出。

$$\Delta\mathbf{W}_j = l_r f(\mathbf{W}_j, \mathbf{X}, Y_d) \mathbf{X}^T \quad (1)$$

在网络训练的全过程中,初始权重和学习率共同决定了权重的修改路径和网络的最终收敛位置,但在训练过程中的某一迭代步,权重修改量只由学习率一个因素决定。因此,学习率不仅直接影响网络的训练速度,而且还会对网络的误差梯

度下降曲线产生影响。

对学习率进行自适应调整的研究一直是BP算法研究的重要内容,国内外很多专家学者就此问题展开了探索,结合各自的研究领域提出许多实用的方法^[1-6]。作者从学习率对BP网络误差下降曲线的影响机理入手,提出了能够使学习率具有自适应调整能力的有限监督调整方法,并通过实证分析和算例验证了该方法的有效性。

1 学习率对网络误差的影响

文献[7]采用实证分析的方法,较为系统地研究了学习率对网络误差下降曲线的影响。以“鸢尾花分类”问题为例,在相同的网络初始权重下研究了不同学习率对网络训练过程的影响。网络误

收稿日期:2006-01-12.

基金项目:国家自然科学基金资助项目(59809007).

作者简介:李鸿雁(1968-),女,副教授。研究方向:水文水资源,智能算法。E-mail:lihongyan@jlu.edu.cn.

差 E 的下降过程见图 1。 E 的定义如下:

$$E = \sum_{k=1}^n \|e\|^2 = \|\mathbf{Y}_{dk} - \mathbf{Y}_k\|^2 \quad (2)$$

式中: $\mathbf{Y} = [Y_1, Y_2, \dots, Y_i, \dots, Y_n]^T$ 表示网络输出的因变量序列; $\mathbf{Y}_d = [Y_{d1}, Y_{d2}, \dots, Y_{di}, \dots, Y_{dn}]^T$ 表示问题的真实因变量序列,即期望值。

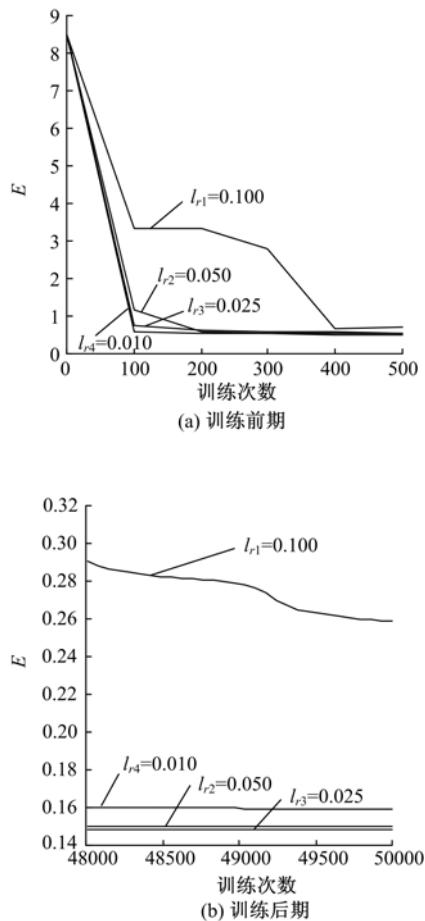


图 1 不同学习率网络误差的下降曲线

Fig. 1 Drop-curve of network errors under different original learning rates

从图 1 可知,当学习率取值在区间 $[0.01, 0.05]$ 时,网络的训练情况比较好;当学习率大于 0.1 时,网络训练情况不够理想。

具有相同初始权重的网络,由于学习率的不同,相应的网络误差下降路径亦不相同,这说明不同的学习率决定了不同的误差梯度下降曲线,最终会导致网络收敛于不同的位置,从而决定了网络不同的训练精度。如果不考虑误差下降路径,片面地讨论“如果学习率选得小,收敛速度会很慢”,这是不切实际的。但在极值点附近,较大的学习率确实不利于网络收敛。

由于 BP 算法是基于网络误差梯度下降的原

则对权重进行调整,所以,无论网络沿哪条误差曲线进行训练,在训练后期,逐渐减小的学习率都有利于网络更快地收敛。

2 学习率的有限监督调整

2.1 基本思想及实现方法

首先选定较大的初始学习率,开始进行网络训练。如果网络误差呈现明显的下降趋势,保持该学习率不变;如果网络误差呈现增大的趋势,按某比例减小学习率,同时将网络权重拉回网络误差未增加的位置,并继续进行网络训练。判断网络误差变化趋势的工作每隔一定训练次数进行一次,而不是每步都进行。这种学习率自适应调整的方法称为有限监督调整方法,其实现过程如下:

```

begin
   $e_0 \leftarrow \mathbf{Y}-\text{net}(\mathbf{W}_0, \mathbf{X})$ ; // 初始化网络权重并计算误差
   $SE_0 \leftarrow \text{SumSquare}(e_0)$ ; // 计算网络误差  $E$ 
   $Old\_W = W$ ; // 记录网络权重
   $Old\_SE = SE_0$ ; // 记录初始网络误差  $E$ 
  for Epochs = 1 : TerminalEpochs // 迭代训练开始
    if  $SE < se$ , Epochs = Epochs - 1; break, end
      // 判断是否停止训练
     $\Delta W \leftarrow f(l_r, SE)$ ; // 计算权重的调整量
     $e \leftarrow \mathbf{Y}-\text{net}(W + \Delta W, X)$ ; // 计算误差
     $SE \leftarrow \text{SumSquare}(e)$ ; // 计算网络误差  $E$ 
    if rem(Epochs, step) = 0 // 每隔 step 步判断
      一次网络误差  $E$  的变化趋势,以决定是否调整学习率
    if  $Old\_SE < SE$ , // 如果网络误差  $E$  增加
       $l_r = \alpha \cdot l_r$ ; // 学习率  $l_r$  以  $\alpha$  比例减小
       $W = Old\_W$ ; // 将网络权重拉回到迭代次数为
      Epochs-step 时的状态
    end
     $Old\_SE = SE$ ; // 记录网络误差
  end
end

```

网络误差变化趋势的判断步长 step 根据实际问题而定,一般取每 100 或 200 个迭代步进行一次判断为宜。

2.2 方法比较

常用学习率自适应调整的基本原则^[8,9]为:
①若网络误差 E 减小,则学习率增加;②若网络误差 E 增加到 k 倍,则学习率减小;③若网络误差处于其他状态,则学习率不变。上述规则可以

用条件方程来表示,见式(3)。

$$l_r(n) = \begin{cases} \alpha l_r(n-1), & E(n) < E(n-1) \\ \beta l_r(n-1), & E(n) \geq kE(n-1) \\ l_r(n-1), & \text{其他} \end{cases} \quad (3)$$

要求式中参数 $\alpha > 1$ 、 $\beta < 1$ 和 $k > 1$,典型的取值为 $\alpha=1.05$ 、 $\beta=0.7$ 和 $k=1.04$ 。

对于河道水位预报和流量预报这种复杂工程算例,存在网络节点和输入维数多、样本容量大的问题,如在文献[10]中,水文预报模型的结构为(4—40—20—1),流量预报模型的结构为(5—40—20—1),选取连续20年的代表性样本作为训练样本,水位模型和流量模型的样本容量分别为1501和1539。若采用式(3)的自适应调整方法来实时判断网络误差 E 的变化趋势,网络训练常常会出现不收敛的问题。

在网络初始权重不变的情况下,作者分别采用学习率有限监督调整方法和式(3)的自适应调整方法,对文献[10]的流量模型进行训练,网络的训练过程及训练结果见表1和表2。

表1 学习率有限监督调整的网络训练过程

Table 1 Network training process based on limited supervising to learning rate regulating

初始 l_r	Epoch	l_r	E
0.5	0	0.2	421.3
	100	0.2	212.8
	200	0.14	212.8
	300	0.098	212.8
	400	0.068	212.8
	500	0.068	212.8
	600	0.048	212.8
	700	0.048	212.8
	800	0.033	212.8
	900	0.033	29.45
0.2	1000	0.033	29.45
	2000	0.033	20.19
	3000	0.033	18.86
	0	0.1	421.3
	100	0.1	212.8
	300	0.049	212.8
	500	0.034	212.8
	1000	0.034	22.90
	2000	0.034	19.56
	3000	0.034	17.02
0.1	0	0.05	421.3
	100	0.05	212.8
	200	0.035	212.8
	300	0.035	62.85
	500	0.035	26.76
	1000	0.035	21.62
	2000	0.035	19.45
	3000	0.035	17.57

注:学习率缩减系数 $\alpha=0.7$ 。

表2 学习率基于式(3)调整的网络训练过程

Table 2 Network training process based on the rule of Eq. (3) for learning rate regulating

初始 l_r	Epoch	l_r	E
0.5	0	0.5000	421.359
	1	0.5250	212.807
	2	0.5250	212.807
	3	0.5250	212.807
0.2	⋮	⋮	⋮
	0	0.2	421.3
	1	0.21	212.8
	2	0.21	212.8
0.1	3	0.21	212.8
	⋮	⋮	⋮
	0	0.1	421.3218
	1	0.105	212.8
0.01	2	0.105	212.8
	3	0.105	212.8
	⋮	⋮	⋮
	0	0.0010	421.359
0.01	1	0.0011	224.496
	2	0.0011	90.0639
	3	0.0012	64.7992
	4	0.0008	94.3907
	5	0.0006	129.016
	6	0.0004	154.992
	7	0.0003	172.491
	8	0.0002	184.025
	9	0.0001	191.686
	10	0.0001	196.872
0.001	20	0.0000	209.808
	50	0.0000	211.531
	100	0.0000	211.578

注: $\alpha=1.05$, $\beta=0.7$, $k=1.04$ 。

2.3 功能分析

经多次试算,本算例适宜的初始学习率 l_r 接近于0.03。从表1来看,当取较大的初始学习率时,经过学习率的有限监督调整,最终都能稳定到0.03附近。

从表2来看,如果初始学习率取得较大,如取 $l_r=0.5$,从开始到第1步,网络误差 E 减小,满足式(3)的第一项,则学习率调整为 $l_r=0.5250$;从第1步到第3步,网络误差 E 不变,满足式(3)的第三项,则学习率保持不变,但是不变的学习率导致了网络的不收敛。为了避免这种现象发生,可以减小 k 值,如果采取较为极端的情况,即 $k=1$,则式(3)转化为式(4)。

$$l_r(n) = \begin{cases} \alpha l_r(n-1), & E(n) < E(n-1) \\ \beta l_r(n-1), & E(n) \geq E(n-1) \end{cases} \quad (4)$$

若采用式(4)作为学习率调整规则,当 $l_r=$

0.5 和 $l_r=0.001$ 时, 网络的训练过程见表 3。

表 3 学习率基于式(4)调整的网络训练过程

Table 3 Network training process based on the rule of Eq. (4) for learning rate regulating

初始 l_r	Epoch	l_r	E
0.5	0	0.5000	421.359
	1	0.5250	212.807
	2	0.3675	212.807
	3	0.2573	212.807
	4	0.1801	212.807
	5	0.1261	212.807
	10	0.0212	212.807
	20	0.0006	212.807
	50	0.0000	212.807
0.001	0	0.0010	421.359
	1	0.0011	224.496
	2	0.0011	90.0639
	3	0.0012	64.7992
	4	0.0008	94.3907
	5	0.0006	129.016
	6	0.0004	154.992
	7	0.0003	172.491
	8	0.0002	184.025
	9	0.0001	191.686
	10	0.0001	196.872
	20	0.0000	209.808
	50	0.0000	211.531
	100	0.0000	211.578

从表 3 来看, 如果初始学习率选取较大值时, 如 $l_r=0.5$, 从第 1 步到第 2 步, 网络误差减小, 学习率增加; 从第 2 步到第 3 步, 网络误差不变, 则学习率减小, 由于这里未采用将第 3 步权重拉回到第 2 步时的策略, 导致了学习率无限减小, 其结果是出现了网络不收敛的现象。如果初始学习率选取较小值时, 如 $l_r=0.001$, 从第 1 步至第 3 步, 学习率一直增加, 当第 4 步开始时, 由于网络误差出现增加现象, 则学习率减小, 最终导致学习率衰减为 0, 网络不收敛。

从表 2 来看, 如果初始学习率选取较小值, 如 $l_r=0.01$, 同样会出现学习率取值先是变大, 然后变小, 最后衰减为 0 的不合理情况。

2.4 讨论

与一般学习率自适应调整方法相比, 学习率有限监督调整方法主要存在以下区别。

(1) 只有当网络误差增加时, 才以某比例缩小学习率; 如果网络误差处于下降状态, 则保持原学习率不变。该策略认为, 如果网络误差处于下降状态, 尤其下降趋势明显时, 则说明此时的学习率是合适的, 不需调整, 特别不需作增加调整。

(2) 判断网络误差变化趋势的工作每隔一定迭代步数进行一次, 而非步步判断, 如可选步长为

100 或 200。这是因为: 一方面, 由于网络训练是一个连续的搜索过程, 只有经过一定的路径积累, 才能显示网络误差的变化趋势。另一方面, 如果步步都进行判断, 增加了计算量, 反而给网络训练增添了负担。图 2 和图 3 描述了不同步长时, 网络误差的下降趋势和学习率的变化过程。从图 2 来看, 当步长为 1 或 10 时, 学习率很快就衰减为 0, 此时的网络误差基本不会再下降, 如图 3 所示。当步长为 50, 学习率下降会很平稳, 但网络误差下降曲线出现了起伏振荡现象。当步长为 100 时, 网络训练情况较为理想。

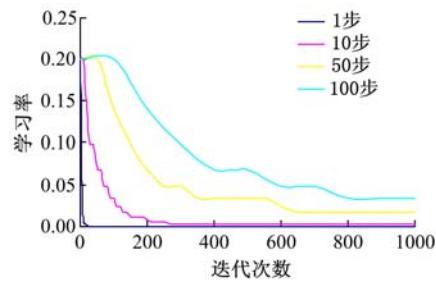


图 2 不同步长学习率的变化情况

Fig. 2 Variation of learning rates under different step sizes

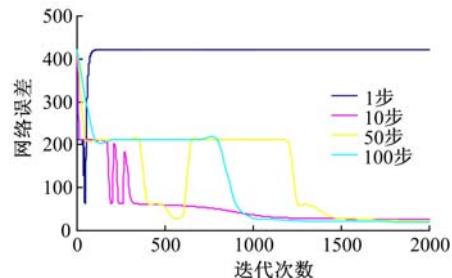


图 3 不同步长网络误差的变化情况

Fig. 3 Variation of network errors under different step sizes

(3) 当判断网络误差出现增大时, 有限监督调整的修改策略是在缩小学习率的同时, 将网络权重拉回到前一判断步所在的误差曲面上, 减小后的学习率在该节点开始训练。这相当于在该节点上, 重新选择误差下降路径。这也说明了当判断步长达到一定数值时, 能在很大程度上避免网络误差振荡问题的出现, 图 3 反映出的情况也证明了该观点。

3 结束语

作者从学习率对 BP 网络误差下降曲线的影响机理入手, 提出了能够使学习率具有自适应调

整能力的有限监督调整方法，并结合具体算例，阐明了该方法的有效性。人工神经网络模仿人脑的结构及功能，具有对信息进行并行处理、分布式存储以及自学习与推理的能力，表现出容错性、非线性、非局域性、非凸性等特点，适于对模糊信息或复杂的非线性关系进行识别与映射。结构的分布式和算法的并行性，决定了其非线性识别能力，同时也给采用数学方法对其功能和特性进行严格证明带来了困难，因此，实证分析和算例研究不失为一种有效验证的策略和途径。

参考文献：

- [1] Baldi P, Hornik K. Neural networks and principal component analysis: learning from examples without local minimum[J]. *Neural Networks*, 1989, 2(2): 53-58.
- [2] Yu Xiao-hu, Chen Guo-an. Efficient back propagation learning using optimal learning rate and momentum[J]. *Neural Networks*, 1997, 10(3): 517-527.
- [3] 王子才,施云惠,崔明根.一种具有动态最优学习率的BP算法[J].*系统仿真学报*, 2001, 13(6):775-776/815.
Wang Zi-cai, Shi Yun-hui, Cui Ming-gen. Back propagation algorithm using dynamic optimal learning rate[J]. *Journal of System Simulation*, 2001, 13(6):775-776/815.
- [4] 李勇,李洪源,叶荣学.自适应学习率的BP网络算法及其在汽轮发电机组故障模糊诊断中的应用[J].*热能动力工程*, 1997, 12 (6): 455-458.
Li Yong, Li Hong-yuan, Ye Rong-xue. BP network algorithm with self-adaptive learning rate and its use in the fuzzy diagnosis of turbogenerator failures[J]. *Journal of Engineering for Thermal Energy and Power*, 1997, 12 (6): 455-458.
- [5] 马正华,薛国新. BP神经网络训练算法的改进[J].*江苏理工大学学报:自然科学版*, 2000, 21(1): 79-83.
Ma Zheng-hua, Xue Guo-xin. Improvement in the training algorithm of BP neural networks[J]. *Journal of Jiangsu University of Science and Technology*, 2000, 21(1): 79-83.
- [6] 冯天瑾,陈哲,顾方方. BP网络学习参数模糊自适应算法的实现[J].*青岛海洋大学学报*, 2000, 30(1):137-141.
Feng Tian-jin, Chen Zhe, Gu Fang-fang. A fuzzy adaptive algorithm for learning parameters of BP networks[J]. *Journal of Ocean University of Qingdao*, 2000, 30(1):137-141.
- [7] 苑希民,李鸿雁,刘树坤,等.神经网络和遗传算法在水科学领域的应用[M].北京:中国水利水电出版社,2002.
- [8] 万书亭,李和明,李永刚.自适应神经网络在发电机组故障诊断中的应用[J].*华北电力大学学报*, 2002, 29(2):99-102.
Wan Shu-ting, Li He-ming, Li Yong-gang. Study of adaptive neural network for fault diagnosis of steam-turbine generator unit[J]. *Journal of North China Electric Power University*, 2002, 29(2):99-102.
- [9] 武美先,张学良,温淑花,等. BP神经网络的双学习率自适应学习算法[J].*现代制造工程*, 2005(10): 29-32.
Wu Mei-xian, Zhang Xue-liang, Wen Shu-hua, et al. Double self-adaptive learning rate algorithm for BPNN [J]. *Modern Manufacturing Engineering*, 2005(10):29-32.
- [10] 李鸿雁,刘寒冰,苑希民,等.人工神经网络峰值识别理论及其在洪水预报中的应用[J].*水利学报*, 2002(6):15-20.
Li Hong-yan, Liu Han-bing, Yuan Xi-min, et al. Peak value recognition theory of artificial neural network and its application to flood forecasting [J]. *Journal of Hydraulic Engineering*, 2002(6):15-20.