Journal of System Simulation

Volume 32 | Issue 1 Article 1

1-17-2020

Curvature-based BP Algorithm Optimization and Its Application in FNN

Weili Xiong
1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;;2. Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China;

Wenxin Sun
1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;;

Xudong Shi
1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;;

Curvature-based BP Algorithm Optimization and Its Application in FNN

Abstract

Abstract: In order to improve the optimization efficiency of BP algorithm affected by the selection of step size, a *step size optimization BP algorithm based on curvature information is proposed* and applied to the training process of FNN (Fuzzy Neural Network). *Reference to Newton's method, The gradient of the cost function and the curvature information in the direction are calculated to determine the direction and magnitude of the parameter adjustment in each iteration.* This method only needs to consider the two order information of the gradient direction, so it does not need the storage and processing of Hessian matrix. The effectiveness and efficiency of the proposed method are verified by a numerical simulation and data simulation of blast furnace ironmaking process.

Keywords

Step size optimization, BP algorithm, Fuzzy neural network, Learning speed

Recommended Citation

Xiong Weili, Sun Wenxin, Shi Xudong. Curvature-based BP Algorithm Optimization and Its Application in FNN[J]. Journal of System Simulation, 2020, 32(1): 1-8.

第 32 卷第 1 期 系统仿真学报© Vol. 32 No. 1
2020 年 1 月 Journal of System Simulation Jan., 2020

基于曲率信息的改进 BP 算法及其在 FNN 中的应用

熊伟丽[1,2]，孙文心[1]，史旭东[1]
(1. 江南大学 物联网工程学院，江苏 无锡 214122；2. 江南大学 轻工过程先进控制教育部重点实验室，江苏 无锡 214122)

摘要：针对步长选取影响误差反向传播(BP，Back Propagation)算法优化效率问题，*提出一种基于曲率信息的步长优化 BP 算法*，并将其应用到了模糊神经网络(FNN)的训练过程中。*参考牛顿法的思想，根据代价函数的梯度及梯度方向上的曲率信息来确定模型参数调整的方向和幅度。*仅需考虑梯度方向上的二阶信息，因此不需要存储和处理 Hessian 矩阵。通过一个数值仿真和高炉炼铁过程数据建模实验，验证了方法的有效性及训练效率。

关键词：步长优化；BP 算法；模糊神经网络；学习速度
中图分类号：TP183 文献标识码：A 文章编号：1004-731X (2020) 01-0001-08
DOI: 10.16182/j.issn1004731x.joss.17-0461

Curvature-based BP Algorithm Optimization and Its Application in FNN

Xiong Weili[1,2], *Sun Wenxin*[1], *Shi Xudong*[1]
(1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;
2. Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China)

Abstract: In order to improve the optimization efficiency of BP algorithm affected by the selection of step size, a *step size optimization BP algorithm based on curvature information is proposed* and applied to the training process of FNN (Fuzzy Neural Network). *Reference to Newton's method, The gradient of the cost function and the curvature information in the direction are calculated to determine the direction and magnitude of the parameter adjustment in each iteration.* This method only needs to consider the two order information of the gradient direction, so it does not need the storage and processing of Hessian matrix. The effectiveness and efficiency of the proposed method are verified by a numerical simulation and data simulation of blast furnace ironmaking process.

Keywords: Step size optimization; BP algorithm; Fuzzy neural network; Learning speed

引言

误差反向传播(BP，Back Porpagation)算法是神经网络训练中常用的算法，其本质是一种梯度下降法，而“误差反向传播”是对神经网络中梯度求解步骤的生动描述。标准的 BP 算法具有运算代价低、简单易实现、适用范围广等优点，但是也存在收敛速度慢、优化过程不稳定[1]等不足，会导致模型训练耗时过长或甚至不收敛。文献[2]对这些不足进行了分析。当代价函数的 Hessian 矩阵存在“Big Killers”[2] 时，即存在较大的特征值时，代价函数对该特征向量上参数的变化相当敏感，优化过程中该现象容易使参数的调整在敏感方向上发生震荡从而使优化过程止步不前甚至无法收敛至局部最优。一些简单的改进方法能够缓解此种现象并提高 BP 算法的性能。如：动量法和 NAG[3]

收稿日期：2017-11-08 修回日期：2018-01-18；
基金项目：国家自然科学基金(61773182)；
作者简介：熊伟丽(1978-)，女，河南洛阳，博士，教授，硕导，研究方向为复杂工业过程建模及优化，智能优化算法及应用；孙文心(1993-)，男，江苏无锡，硕士生，研究方向为复杂工业过程建模及优化。

(Nesterov's Accelerated Gradient)在标准的 BP 算法上加了一个阻尼项，从而稳定了优化过程；变步长法[4]在优化过程发生震荡时自动减小步此外，基于二阶导数信息的优化方法能进一步提升 BP 算法的收敛速度。如：牛顿迭代法、高斯-牛顿法能利用二阶导数信息合理控制各方向上的学习率从而能提高优化过程的稳定性和收敛速度，但也需要付出较大的计算代价[5]；BFGS (拟牛顿法)通过迭代的方法来近似求解二阶导数信息从而缓解了牛顿迭代法中计算代价过大的问题。Adagrad[6]和 RMSProp[7]为每一个待优化参数设置了各自的步长，并在优化过程中不断调整来近似高斯-牛顿法的优化过程。上述近似二阶方法已用于多种神经网络模型的训练过程并展现出出色的优化效果[8]。

综上，合理利用二阶导数信息能有效提高 BP 算法的收敛速度和稳定性。针对模糊神经网络的训练过程，提出一种步长优化方法来提高 BP 算法的收敛速度。在每次迭代中先使用反向传播算法计算代价函数的梯度，然后计算出代价函数在梯度方向上的曲率(二阶导数)，最后根据牛顿法的思想，利用曲率和一阶导数信息来估计最优步长并进行参数调整。将方法与变步长法和惯性法做比较，通过数值仿真和对高炉炼铁硅含量数据的拟合实验验证了方法的有效性和性能。

1 FNN 和 BP 算法

1.1 模糊神经网络简介

模糊神经网络(FNN)是模糊系统同神经网络相结合的产物，作为模糊系统，能够很好地描述不确定信息或定性的知识，具有很好的可解释性[9]；作为神经网络，具有非常优秀的非线性逼近能力[10]。因此被广泛应用于智能控制、复杂工业建模与软测量等多个方面[11-13]。

基于 Mamdani 模糊模型构造由单值模糊器、乘法推理机、中心平均解模糊器组成的 MISO 模糊神经网络[14-15]，其结构如图 1 所示。

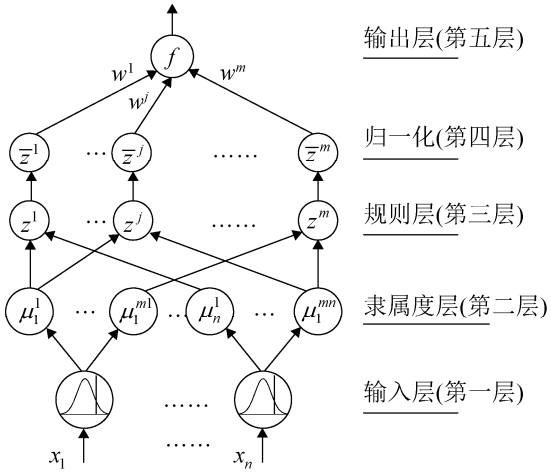


图 1 Mamdani 模型的模糊神经网络
Fig. 1 Fuzzy neural network based on Mamdani model

第一层：将精确集合上的输入变量 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$ 传入模糊神经网络。

该层节点数 $\mathrm{N}_1=n$ (输入变量的维数)

第二层：用 m_i 个语言变量 $A_i^{l_i}$ 来描述精确变量 x_i，并输出 x_i 隶属于语言变量 $A_i^{l_i}$ 的隶属度 $\mu_i^{l_i}$，其中 $i=1,2\cdots n$；$l_i=1,2\cdots m_i$。

$$\mu_i^{l_i}=\exp\left(-\left(\frac{x_i-c_i^{l_i}}{\sigma_i^{l_i}}\right)^2\right) \tag{1}$$

该层节点数 $\mathrm{N}_2=\sum_{i=1}^{n}m_i$ (语言变量总数量)，涉及模型参数变量 $c_i^{l_i}$ 和 $\sigma_i^{l_i}$ (语言变量 $A_i^{l_i}$ 的中心位置和宽度)。

第三层：计算模糊规则库中规则 R_j 对输入变量 $\boldsymbol{X}$ 的适用度 z^j ($j=1,2\cdots m$； $j_i=1,2\cdots m_i$)。规则 R_j 的形式如下：

$$R_j:\text{if } x_1 \text{ is } A_1^{j_1} \text{ and } x_2 \text{ is } A_2^{j_2}\cdots\text{and } x_n \text{ is } A_n^{j_n},$$
$$\text{then } f \text{ is } B^j$$

其中：f 为模型的精确输出量；B^j 为描述 f 的语言变量。用乘法推理机计算 R_j 的适用度

$$z^j=\prod_{1\le i\le n}\mu_i^{j_i} \tag{2}$$

该层节点数 $\mathrm{N}_3=m$ (规则库中的规则数量)。

第四层：归一化计算

$$\bar{z}^k=z^k\Big/\sum_{j=1}^{m}z^j\ \ (k=1,2\cdots m) \tag{3}$$

该层节点数 $N_4 = m$。

第五层：解模糊计算，利用中心平均解模糊器得出模型的精确输出

$$f(\boldsymbol{X}) = \sum_{k=1}^{m} \bar{z}^k w^k \tag{4}$$

由于设计的是 MISO 模糊神经网络，所以该层节点数 $N_5 = 1$。该层涉及模型参数 w^k，表示语言变量 B^k 的中心。

1.2 模糊神经网络的 BP 训练方法

标准 BP 算法实质上是一种梯度下降法，首先利用求导中的链式规则来求解代价函数关于模型参数的梯度，然后根据梯度来调整模型参数。

定义模型参数 θ、误差 $E(\theta)$和代价函数 $J(\theta)$为：

$$\theta \equiv \left[c_1^1 \cdots c_n^{m_n}, \sigma_1^1 \cdots \sigma_n^{m_n}, w^1 \cdots w^m \right]^{\mathrm{T}} \tag{5}$$

$$E(\theta) \equiv \left(f(\boldsymbol{X}) - y \right)^2 / 2 \tag{6}$$

$$J(\theta) \equiv \sum_{p=1}^{N} E(\theta)\big|_{\boldsymbol{X}=\boldsymbol{X}^p, y=y^p} \tag{7}$$

式中：$\boldsymbol{X}^p$ 和 y^p 为训练样本集中第 p 个样本输入变量 $\boldsymbol{X}$、输出变量 y 的观测值；N 为训练样本容量。

利用式(3)，(4)，(6)得到

$$\frac{\partial E(\theta)}{\partial w^k} = \bar{z}^k \left(f(\boldsymbol{X}) - y \right) \tag{8}$$

$$\frac{\partial E(\theta)}{\partial \bar{z}^k} = w^k \left(f(\boldsymbol{X}) - y \right) \tag{9}$$

$$\frac{\partial E(\theta)}{\partial z^j} = \left(f(\boldsymbol{X}) - y \right) \left(w^j - f(\boldsymbol{X}) \right) \bar{z}^j / z^j \tag{10}$$

由式(1)~(2)得到

$$\frac{\partial z^j}{\partial \mu_i^{l_i}} = z^j \delta_{l_i, j_i} / \mu_i^{l_i}$$

$$\frac{\partial z^j}{\partial c_i^{l_i}} = \frac{\partial z^j}{\partial \mu_i^{l_i}} \frac{\partial \mu_i^{l_i}}{\partial c_i^{l_i}} = 2 z^j \left(x_i - c_i^{l_i} \right) \delta_{l_i, j_i} / \left(\sigma_i^{l_i} \right)^2 \tag{11}$$

$$\frac{\partial z^j}{\partial \sigma_i^{l_i}} = \frac{\partial z^j}{\partial \mu_i^{l_i}} \frac{\partial \mu_i^{l_i}}{\partial \sigma_i^{l_i}} = 2 z^j \left(x_i - c_i^{l_i} \right)^2 \delta_{l_i, j_i} / \left(\sigma_i^{l_i} \right)^3 \tag{12}$$

式中：$\delta_{l_i, j_i} = \begin{cases} 1 & (l_i = j_i) \\ 0 & (l_i \neq j_i) \end{cases}$。

结合式(10)~(12)得到

$$\frac{\partial E(\theta)}{\partial c_i^{l_i}} = \frac{2 \left(x_i - c_i^{l_i} \right) \left(f(\boldsymbol{X}) - y \right)}{\left(\sigma_i^{l_i} \right)^2} \sum_{j=1}^{m} \delta_{l_i, j_i} (w^j - f(\boldsymbol{X})) \bar{z}^j \tag{13}$$

$$\frac{\partial E(\theta)}{\partial \sigma_i^{l_i}} = \frac{x_i - c_i^{l_i}}{\sigma_i^{l_i}} \frac{\partial E(\theta)}{\partial c_i^{l_i}} \tag{14}$$

最后，令 $\Delta\theta$ 为负梯度，根据式(7)，(8)，(13)，(14)得到代价函数的梯度并调整参数

$$\frac{\partial E(\theta)}{\partial \theta} = \left[\frac{\partial E}{\partial c_1^1} \cdots \frac{\partial E}{\partial c_n^{m_n}}, \frac{\partial E}{\partial \sigma_1^1} \cdots \frac{\partial E}{\partial \sigma_n^{m_n}}, \frac{\partial E}{\partial w^1} \cdots \frac{\partial E}{\partial w^m} \right]^{\mathrm{T}}$$

$$\nabla_\theta J(\theta) = \sum_{p=1}^{N} \frac{\partial E(\theta)}{\partial \theta}\bigg|_{\boldsymbol{X}=\boldsymbol{X}^p, y=y^p} \tag{15}$$

$$\Delta\theta \equiv -\nabla_\theta J(\theta)$$

$$\theta(q) = \theta(q-1) + \alpha \Delta\theta\big|_{\theta=\theta(q-1)} \tag{16}$$

式中：α 为训练步长；$\theta(q)$ 为参数变量经过 q 次迭代后的取值；q 为大于 1 的整数。

BP 算法在实际使用中通常可以利用变步长法或动量法来加快训练速度[4]，如表 1 所示。

表 1 BP 算法的优化方法

Tab. 1 Optimizing methods for BP algorithm

改进方法	公式描述
动量法	$\Delta\theta(q) = (\beta - 1)\nabla_\theta J(\theta) + \beta \Delta\theta(q-1)$ β 为动量因子 $(0 \le \beta < 1)$
变步长法	$\alpha(q) = \gamma^\tau \alpha(q-1); \gamma \geqslant 0$ $\tau = \mathrm{sgn}(\Delta\theta(q)\Delta\theta(q-1));$

2 基于曲率信息的 BP 算法

利用一阶导数信息可以准确地找到代价函数在小范围内的最速下降方向，但是仅根据一阶导数信息调整参数并不够合理。本文提出一种基于曲率信息的 BP 算法。基于代价函数的负梯度 $\Delta\theta(q-1)$，将步长的确定转化为以下优化问题：

$$\begin{aligned} &\min && J(\alpha) = J(\theta(q-1) + \alpha \Delta\theta(q-1)) \\ &\text{s.t.} && \alpha > 0 \end{aligned}$$

利用牛顿迭代法的思想，根据负梯度 $\Delta\theta(q-1)$ 方向上代价函数关于步长 α 的一阶导数和曲率信息来估计每次迭代的最佳步长 α，从而提高参数调整的效率。由于该方法仅需考虑

$\Delta\theta(q-1)$ 方向上的曲率信息，因此不需要存储和处理 Hessian 矩阵。

2.1 FNN 中曲率信息的求法

通过式(8)、(11)~(15)可以得到代价函数的梯度，在此基础上做进一步推导就能得到代价函数关于步长 α 的二阶导数。当负梯度 $\Delta\theta$ 确定后，参数调整步骤中变量 $c_i^{l_i}$，$\sigma_i^{l_i}$ 和 w^k 可以看作是关于步长 α 的因变量。

根据式(16)可以得到

$$\Delta\theta=\left[\Delta c_1^1\cdots\Delta c_n^{m_n},\Delta\sigma_1^1\cdots\Delta\sigma_n^{m_n},\Delta w^1\cdots\Delta w^m\right]$$

$$\frac{\mathrm{d}c_i^{l_i}}{\mathrm{d}\alpha}=\Delta c_i^{l_i};\ \frac{\mathrm{d}\sigma_i^{l_i}}{\mathrm{d}\alpha}=\Delta\sigma_i^{l_i}\ ;\ \frac{\mathrm{d}w^k}{\mathrm{d}\alpha}=\Delta w^k$$

为了简化表述，令

$$h_i^{l_i}\equiv\frac{\left(x_i-c_i^{l_i}\right)\Delta\sigma_i^{l_i}}{\sigma_i^{l_i}}+\Delta c_i^{l_i}\ ;\ \ g^j\equiv w^j-f(\boldsymbol{X})$$

并可知：

$$\frac{\mathrm{d}h_i^{l_i}}{\mathrm{d}\alpha}=-\frac{h_i^{l_i}\Delta\sigma_i^{l_i}}{\sigma_i^{l_i}}\tag{17}$$

根据式(11)~(12)得到

$$\frac{\mathrm{d}z^j}{\mathrm{d}\alpha}=\sum_{\substack{1\leqslant i\leqslant n\\1\leqslant l\leqslant m}}\delta_{l_i,j_i}\left(\frac{\partial z^j}{\partial c_i^{l_i}}\frac{\partial c_i^{l_i}}{\partial\alpha}+\frac{\partial z^j}{\partial\sigma_i^{l_i}}\frac{\partial\sigma_i^{l_i}}{\partial\alpha}\right)=$$

$$2z^j\sum_{\substack{1\leqslant i\leqslant n\\1\leqslant l\leqslant m}}\delta_{l_i,j_i}\frac{\left(x_i-c_i^{l_i}\right)h_i^{l_i}}{\left(\sigma_i^{l_i}\right)^2}\tag{18}$$

对式(17)~(18)两端再进行求导得

$$\frac{\mathrm{d}^2z^j}{\mathrm{d}\alpha^2}=\frac{1}{z^j}\left(\frac{\mathrm{d}z^j}{\mathrm{d}\alpha}\right)^2-$$

$$2z^j\sum_{\substack{1\leqslant i\leqslant n\\1\leqslant l\leqslant m}}\delta_{l_i,j_i}\frac{h_i^{l_i}\left(3h_i^{l_i}-2\Delta c_i^{l_i}\right)}{\left(\sigma_i^{l_i}\right)^2}\tag{19}$$

同理，根据式(3)，(4)，(8)，(10)可以得到

$$\frac{\mathrm{d}f(\boldsymbol{X})}{\mathrm{d}\alpha}=\sum_{j=1}^{m}\left[\frac{\mathrm{d}z^j}{\mathrm{d}\alpha}g^j+z^j\Delta w^j\right]\Bigg/\sum_{j=1}^{m}z^j\tag{20}$$

$$\frac{\mathrm{d}^2f(x)}{\mathrm{d}\alpha^2}=\frac{\sum_{j=1}^{m}\left[\frac{\mathrm{d}^2z^j}{\mathrm{d}\alpha^2}g^j+2\frac{\mathrm{d}z^j}{\mathrm{d}\alpha}\left(\Delta w^j-\frac{\mathrm{d}f(\boldsymbol{X})}{\mathrm{d}\alpha}\right)\right]}{\sum_{j=1}^{m}z^j}\tag{21}$$

由式(6)可以得到

$$\frac{\mathrm{d}E(\theta)}{\mathrm{d}\alpha}=\left(f(\boldsymbol{X})-y\right)\frac{\mathrm{d}f(\boldsymbol{X})}{\mathrm{d}\alpha}\tag{22}$$

$$\frac{\mathrm{d}^2E(\theta)}{\mathrm{d}\alpha^2}=\left(\frac{\mathrm{d}f(\boldsymbol{X})}{\mathrm{d}\alpha}\right)^2+\left(f(\boldsymbol{X})-y\right)\frac{\mathrm{d}^2f(\boldsymbol{X})}{\mathrm{d}\alpha^2}\tag{23}$$

最后，根据式(7)得到代价函数关于步长 α 的一阶、二阶导数

$$\left.\frac{\mathrm{d}J(\theta)}{\mathrm{d}\alpha}\right|_{\theta=\theta(q-1)}=\sum_{p=1}^{N}\left.\frac{\mathrm{d}E(\theta)}{\mathrm{d}\alpha}\right|_{\boldsymbol{X}=\boldsymbol{X}^p,y=y^p}^{\theta=\theta(q-1)}\tag{24}$$

$$\left.\frac{\mathrm{d}^2J(\theta)}{\mathrm{d}\alpha^2}\right|_{\theta=\theta(q-1)}=\sum_{p=1}^{N}\left.\frac{\mathrm{d}^2E(\theta)}{\mathrm{d}\alpha^2}\right|_{\boldsymbol{X}=\boldsymbol{X}^p,y=y^p}^{\theta=\theta(q-1)}\tag{25}$$

至此，曲率信息求解完成。从式(18)~(25)依次求出 z，$f(x)$，$E(\theta)$ 和代价函数 $J(\theta)$ 关于步长的二阶导数。所以，整个求导数的过程可以看作为一次正向传播。

2.2 算法步骤

当调参问题为一个凸优化问题时，牛顿迭代法能够表现出非常快的收敛速度，但在一般的非凸问题中，牛顿迭代法会发生异常，算法可能会最终收敛于鞍点甚至极大值(3.1 节的仿真给出了一个例子)。因此在利用曲率进行最优步长估计时必须保证结果大于零以保证代价函数递减；另外，还需对步长的大小做一定的限制，防止步长过大破坏训练过程的稳定。本方法分 2 种情况来决定参数调整步长，算法步骤为：

step 1：设置精度阈值和步长上限。

step 2：利用误差反向传播计算出代价函数的负梯度。

step 3：利用式(17)~(24)计算代价函数关于步长 α 的一阶导数和曲率，根据牛顿法的思想，分 2 种情况利用式(25)~(26)对最优步长进行估计。

情况 1：$\left.\frac{\mathrm{d}^2J(\theta)}{\mathrm{d}\alpha^2}\right|_{\theta=\theta(q-1)}>0$

$$\alpha(q)=0.6\left(\frac{\mathrm{d}J(\theta)}{\mathrm{d}\alpha}\Big/\frac{\mathrm{d}^2J(\theta)}{\mathrm{d}\alpha^2}\right)_{\theta=\theta(q-1)}\tag{26}$$

情况 2：$\left.\dfrac{\mathrm{d}^2J(\theta)}{\mathrm{d}\alpha^2}\right|_{\theta=\theta(q-1)}<0$

$$\alpha(q)=-1.8\left(\frac{\mathrm{d}J(\theta)}{\mathrm{d}\alpha}\Big/\frac{\mathrm{d}^2J(\theta)}{\mathrm{d}\alpha^2}\right)_{\theta=\theta(q-1)} \tag{27}$$

常数 0.6 和–1.8 属于参考值，可以根据不同情况替换。

step 4：由于梯度和曲率仅能描述代价函数的局部情况，因此，当 step 3 估计出的步长超出合理上限时，强制步长小于所设置的步长上限，利用线搜索方法[16]来确定一个相对合理的步长，再根据式(28)调整参数

$$\theta(q)=\theta(q-1)+\alpha(q)\Delta\theta(q) \tag{28}$$

step 5：进行准确度计算，如果满足精度要求则停止计算，否则返回 step 3。

在整个算法流程中需要经过两次运算才能进行一次参数调整，首先计算出 $\Delta\theta$，然后根据 $\Delta\theta$ 估计出最佳步长再调节参数。虽然需要付出一定计算代价，但是，由于利用了二阶导数信息，所以能取得理想的训练效率。

3 仿真实验

通过仿真实验来分析说明步本文改进 BP 算法的快速收敛性。首先，构造一个简单的优化问题来直观地分析该方法相比其他步长优化方法的优势；

3.2 节和 3.3 节利用数值仿真实验和高炉炼铁过程数据仿真实验来展现本文算法在实际使用中的优化效果，利用均方误差来描述训练过程中模型对数据的拟合程度，式(29)是均方误差的计算公式。

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-f(x_i))^2 \tag{29}$$

式中：n 为训练样本容量；y_i 和 x_i 分别为样本集中第 i 个样本的主导变量和辅助变量；$f(x_i)$为预测模型根据 x_i 对 y_i 做出的估计。

3.1 数值优化实验

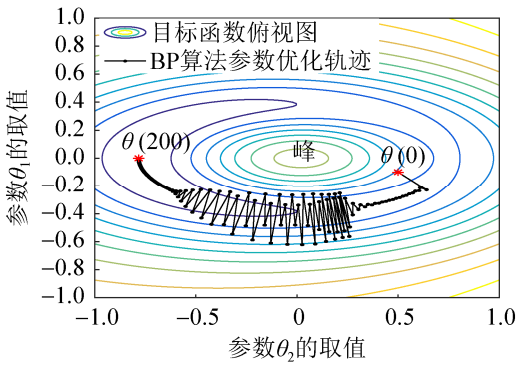
构造一个参数寻优问题来直观地说明普通 BP 算法收敛缓慢的原因及本文的步长优化方法相比其他方法的优势。构造目标函数如下

$$J(\theta_1,\theta_2)=\cos\left(\frac{5}{1+4\theta_1^2+\theta_2^2}\right)+\frac{\theta_2^2+4\theta_2}{20} \tag{30}$$

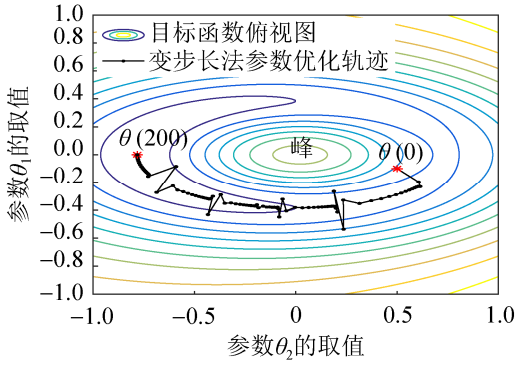
初始参数设为

$$\theta(0)=[\theta_1(0),\theta_2(0)]^{\mathrm{T}}=[-0.1,0.5]^{\mathrm{T}}$$

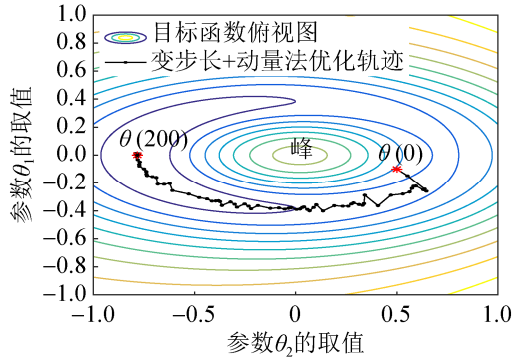
利用标准 BP 算法、多种 BP 改进方法及步长为 0.6 的牛顿迭代法来调节参数 θ_1 和 θ_2，使得式(30)尽可能达到最小，迭代次数设为 200 次，结果如图 2 所示。



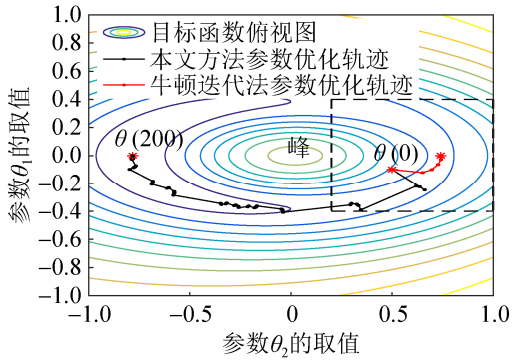
(a) 固定步长的 BP 算法



(b) 变步长法



(c) 变步长法+动量法



(d) 改进 BP 算法与牛顿迭代法

图 2 参数优化过程俯视图

Fig. 2 Contouring of parameter optimization process

如图 2(a)所示，由于在一些“狭长”区域内目标函数对参数 θ_1 的调整过于敏感，容易使参数的调整不均匀(θ_1 的调整过大以致发生振荡， θ_2 的调整则过缓)，所以固定步长的 BP 算法对参数的调节往往比较低效并且不够稳定。变步长法能够使步长自动地适应参数优化过程，能够使优化过程在敏感的参数调整方向上减小振荡，并在正确的调整方向上尽可能地提高步长以加快收敛速度(如图 2(b)所示)，加上动量法后能进一步提高优化过程的稳定性和收敛速度(如图 2(c)所示)。但是变步长法对步长的调节存在滞后性，当参数调整发生振荡时才会减小步长，调整过缓时才会增加步长，因此参数优化过程会呈现出阶段性。本文的算法在调整步长时利用了曲率信息，因此具有以下 2 个特点：(1) 由于参数选择合理，经过“狭长”区域时不至于发生严重的振荡(如图 2(d)所示)；(2) 每次的参数调整方向差别较大，经过几次迭代后就能找到合适的下降方向，并以准确的步长进行参数调整。

最后，图 2(d)展示了牛顿迭代法出现奇异的情况，图 3 是图 2(d)虚线框内容的展示。由于所定义的优化问题非凸，牛顿法在迭代中 Hessian 矩阵出现了负特征值，从图 3 可以看出，牛顿迭代法最终收敛到了鞍点处。本文方法在迭代中同样也出现了曲率为负的现象，但由于所选下降方向为梯度方向，且限制了步长大于零，因此能保证代价函数随迭代逐步递减，不会被鞍点、极大值点所吸引。

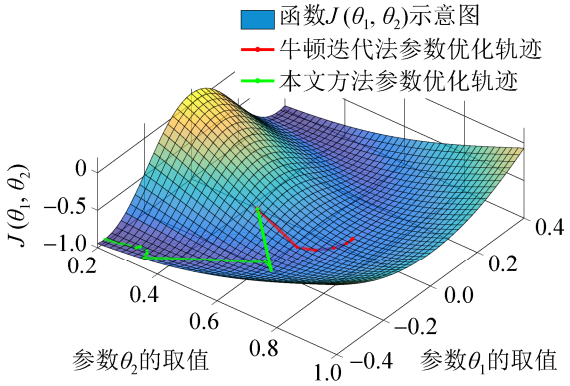


图 3 优化轨迹示意图

Fig. 3 Schematic of trajectory in optimization

3.2 非线性系统逼近实验

构造非线模型为：

$$y(t)=\frac{x_1(t)}{1+0.5\sin\left(x_2(t)\right)}+\frac{x_2(t)}{1+0.5\sin\left(x_1(t)\right)}+\varepsilon(t)$$

$$x_1(t+1)=\left(\frac{x_1(t)}{1+x_1^2(t)}+1\right)\sin\left(x_2(t)\right)$$

$$x_2(t+1)=x_2(t)\cos\left(x_2(t)\right)+\exp\left(-\frac{x_1^2(t)+x_2^2(t)}{8}\right)x_1(t)+$$

$$u^3(t)/[1+u^2(t)+0.5\cos\left(x_1(t)+x_2(t)\right)]$$

式中：$u(t)$和$y(t)$为模型的输入输出变量；$x_1(t)$和$x_2(t)$为模型的状态变量； $\varepsilon(t)$ 是方差为 0.08 的高斯白噪声，初始状态设为 0。令模型输入变量 $\boldsymbol{X}(t)$为 $[y(t-1)\cdots y(t-3),u(t)\cdots u(t-2)]$，输出变量 $f(t)$为 $y(t)$的估计。令输入信号为 $u(t)=\sin(0.2\pi t)+\sin(0.8\pi t)$。

生成 1 000 个样本。利用多种算法对模糊神经网络进行优化，优化过程的误差曲线如图 4 所示，图 5 表示本文方法在每次参数调整中的步长取值示。

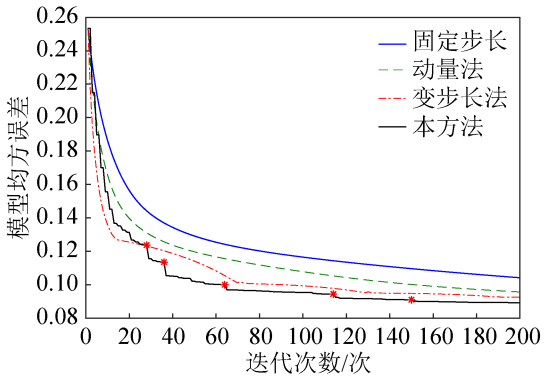


图 4 数值仿真效果对比

Fig. 4 Comparison of numerical simulation results

https://dc-china-simulation.researchcommons.org/journal/vol32/iss1/1
DOI: 10.16182/j.issn1004731x.joss.17-0461

从图 5 可以看出，在训练过程中有 5 次参数调整步长取到了较大的值，这 5 次参数调整都使图 4 中的训练曲线发生了显著下降，这说明本文提出的方法经过多次迭代后找到了合适的参数调整方向，并以准确的步长对参数进行了调整，因此能够加快训练过程的收敛速度。

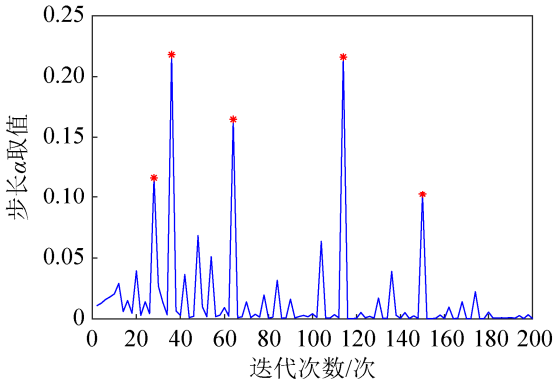


图 5 训练过程中本方法的步长选择
Fig. 5 Step size selection of our method during training

3.3 高炉炼铁硅含量数据仿真

高炉炼铁过程中铁水的硅含量是表征炉温状态的重要参考变量[17]。实际生产中，影响铁水硅含量的因素众多，如矿石种类、炉温、风量等，并且高炉炼铁过程的机理复杂，难以建立精确的机理模型。所以有必要利用现场采集的数据建立数据驱动模型来估计铁水中的硅含量。利用模糊神经网络对高炉炼铁硅含量数据进行建模，模型的输入变量有 2 h 和 3 h 前铁水硅含量的测量值、喷煤、风量、透气性、风温、料速、压差和富氧量，输出量为此时刻硅含量的估计，输入量描述如表 2 所示。

表 2 模型的输入变量
Tab. 2 Input variables of the model

物理量	单位	物理量	单位
Si (t-1)	wt%	风温	℃
Si (t-2)	wt%	料速	t/h
喷煤	t	压差	kPa
风量	M^3/min	富氧率	wt%
透气性	M^3/min•kPa		

数据的采样周期为 1.5~2 h。样本的输入量经过了 PCA[18]降噪处理，样本大小 N=500。利用多种训练方法对模糊神经网络进行训练，比较结果如图 6 所示。

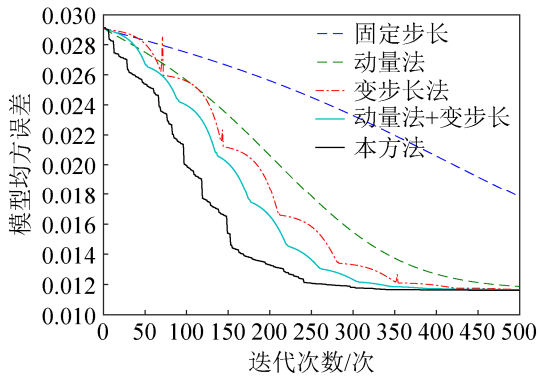


图 6 各方法的训练误差曲线
Fig. 6 Deviation curves of each methods

为了显示各方法的步长选择合理性，表 3 对参数训练过程中相邻 2 次迭代的参数调整方向夹角 $\langle \Delta\theta(q), \Delta\theta(q-1) \rangle$ 进行了统计，如果夹角出现在 $(30,50]$°范围内就可近似说明步长选择合理，出现在 $(150,180]$°范围内则说明训练过程发生了振荡。

表 3 训练中夹角出现在各区间的统计频率
Tab. 3 Statistical frequency of angles that appeared in each ranges during training process %

方法名称	$(30,50]^\circ$	$(30,150]^\circ$	$(150,180]^\circ$
固定步长	99.8	0.2	0.0
动量法	99.0	1.0	0.0
变步长法	94.8	4.2	1.0
动量法+变步长	91.4	8.4	0.2
本方法	12.9	87.1	0.0

从表 3 可以看出，在训练过程中，固定步长的 BP 算法往往会在同一个方向上对模型参数进行多次调整，因此训练比较低效。变步长法由于能利用前一步迭代的经验自动调节步长，所以能大幅度提高训练效率，但是这种调节不够合理，从表 3 中可以看出真正高效的参数调整次数小于总迭代次数的 10%，并且使训练发生了振荡。动量法相当于给负梯度 $\Delta\theta$ 加上了一个阻尼项，能够有效地减小训练过程中的振荡趋势，从而加快训练速度(如图 6 所示)。本文方法由于利用了曲率信息，步长选择更加合理，所以能得到更快的收敛速度。

4 结论

针对传统 BP 算法存在的不足，设计一种基于曲率的 BP 优化算法，并推导得到了该方法在 FNN 训练中的实施步骤。该方法在不需存储 Hessian 矩阵的条件下，利用上二阶导数信息来有效加速普通的一阶方法，虽然并没有求出完整的二阶信息，当问题相对简单时无法达到牛顿法的二阶收敛速度，但减小了内存负担且能有效适用于更加复杂的优化问题。由于仅需计算梯度方向上的曲率，求解难度小于其它依赖二阶导数的优化方法，因此，能为更多代价函数二阶可导的优化问题提供新的思路。

参考文献：

[1] 王海军. 带自适应步长神经网络学习算法的收敛性分析[D]. 大连: 大连理工大学, 2016.
Wang Haijun. Convergence Analysis of Neural Network Learning Algorithm with Adaptive Learning Rate[D]. Dalian: Dalian University of Technology, 2016.

[2] LeCun Y A, Bottou L, Orr G B, et al. Efficient backprop[M]. Neural networks: Tricks of the trade. Springer Berlin Heidelberg, 2012: 9-48.

[3] Nesterov Y. Gradient methods for minimizing composite functions[J]. Mathematical Programming (S0025-5610), 2013, 140(1): 125-161

[4] 房振勇, 游文虎, 冯汝鹏. 改进 BP 算法在模糊神经网络中的应用[J]. 北京航空航天大学学报, 2007, 33(11): 1321-1324.
Fang Zhenyong, You Wenhu, Feng Rupeng. Application of improved BP algorithm in fuzzy neural networks[J]. Journal of Beijing University of Aeronautics and Astronautics, 2007, 33(11): 1321-1324.

[5] 高雪鹏, 丛爽. BP 网络改进算法的性能对比研究[J]. 控制与决策, 2001, 16(2): 167-171.
Gao Xuepeng, Cong Shuang. Comparative Study on Fast Learning Algorithms of BP Networks[J]. Control and Decision, 2001, 16(2): 167-171.

[6] Hadgu A T, Nigam A, Diaz-Aviles E. Large-scale learning with AdaGrad on Spark[C]// IEEE International Conference on Big Data. Beijing, China: IEEE, 2015: 2828-2830.

[7] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. Massachusetts, USA: MIT Press, 2016.

[8] Mukkamala M C, Hein M. Variants of RMSProp and Adagrad with logarithmic regret bounds[C]// Proceedings of the 34th International Conference on Machine Learning, PMLR 70. Sydney, Australia: ICML, 2017: 2545-2553.

[9] Wang L X. A course in fuzzy systems and control[M]. Englewood, USA: Prentice-Hall, Inc. 1996.

[10] Hecht-Nielsen. Theory of the backpropagation neural network[C]// International Joint Conference on Neural Networks. Honolulu, HI, USA: IEEE, 2002(1): 593-605.

[11] 董海鹰, 魏占宏, 杨玺, 等. 基于自适应模糊神经网络的风速软测量[J]. 电力系统及其自动化学报, 2013, 25(1): 60-65.
Dong Haiying, Wei Zhanhong, Yang Xi, et al. Wind Speed Soft Sensor Based on Adaptive Fuzzy Neural Network[J]. Proceedings of the CSU-EPSA, 2013, 25(1): 60-65.

[12] Pratama M, Er M J, Li X, et al. Data driven modeling based on dynamic parsimonious fuzzy neural network[J]. Neurocomputing (S0925-2312), 2013, 110: 18-28.

[13] Zhang M, Liu X. A soft sensor based on adaptive fuzzy neural network and support vector regression for industrial melt index prediction[J]. Chemometrics & Intelligent Laboratory Systems (S0169-7439), 2013, 126: 83-90.

[14] 李国勇, 杨丽娟. 神经·模糊·预测控制及其 MATLAB 实现. [M]. 3 版. 北京: 电子工业出版社, 2013: 257-260.
Li Guoyong, Yang Lijuan. Neural Netowrk. Fuzzy. Predictive Control and its MATLAB implementation [M]. 3 Edition. Beijing: Publishing House of Electronics Industry, 2013: 257-260.

[15] 丛爽. 几种模糊神经网络系统关系的对比研究[J]. 信息与控制, 2001, 30(6): 486-491.
Cong Shuang. Comparative Research on Relationships between Several Fuzzy-Neural Network Systems[J]. Information and Control, 2001, 30(6): 486-491.

[16] 李炯城, 肖恒辉, 李桂愉. 高效的线搜索寻优方法[J]. 计算机工程与应用, 2013, 49(3): 53-56.
Li Jiongcheng, Xiao Henghui, Li Guiyu. Efficient line search optimization method[J]. Computer Engineering and Applications, 2013, 49(3): 53-56.

[17] 蒋朝辉, 董梦林, 桂卫华, 等. 基于 Bootstrap 的高炉铁水硅含量二维预报[J]. 自动化学报, 2015, 42(5): 715-723.
Jiang Zhaohui, Dong Menglin, Gui Weihua, et al. Two-dimensional Prediction for Silicon Content of Hot Metal of Blast Furnace Based on Bootstrap[J]. Acta Automatica Sinica, 2015, 42(5): 715-723.

[18] Mika S, Lkopf B, Smola A, et al. Kernel PCA and de-noising in feature spaces[C]// Conference on Advances in Neural Information Processing Systems. Marriott, Denver, USA: MIT Press, 1999: 536-542.

https://dc-china-simulation.researchcommons.org/journal/vol32/iss1/1
DOI: 10.16182/j.issn1004731x.joss.17-0461