文章编号:1672-058X(2013)08-0067-05

一种改进的 BP 算法在消费水平中的应用*

宋　峰
(重庆师范大学 数学学院,重庆 401331)

摘　要:通过对标准 BP 算法的改进,提出了一种 L-M 贝叶斯正则化优化算法,并把它应用到成都市居民消费水平预测中。经试验验证,L-M 贝叶斯正则化的 BP 神经网络比相同条件下另外两种改进算法有更强的泛化能力,对居民消费水平有很好的预测效果。

关键词:BP 神经网络;L-M 优化算法;贝叶斯正则化算法;居民消费水平
中图分类号:TP183　　**文献标志码**:A

一个地区的消费水平反应了这个地区的的经济发展关系以及人民生活水平。而这一经济模型是一个复杂的非线性系统,针对这种解决机理尚不明确的非线性问题,可以采用人工神经网络来解决。而在人工神经网络中,BP 神经网络是目前研究比较成熟的网络模型,并在实践中得到了广泛的应用。

1　L-M 贝叶斯正则化 BP 神经网络

1.1　标准 BP 神经网络模型及算法原理

BP 神经网络是一种多层前馈神经网络,包括输入层、若干隐含层和输出层。根据 Kplmogorov 定理,只要一个隐层的三层 BP 网络就可以实现对任意精度的逼近。BP 算法是采用反向传播的权值调整的一种算法。算法步骤:对一个输入样本,从输入层经隐含层逐层正向计算,得到输出层的输出。若得到期望的输出,学习算法结束;否则,转至反向传播。反向传播就是将误差信号按原连接通路反向计算,由梯度下降法调整各层神经元的权值和阈值,从而使网络输出逼近期望输出,通常使之达到误差均方值取最小为止。

1.2　BP 神经网络的缺陷与改进

BP 神经网络学习收敛速度慢,容易出现平坦区陷入局部极小点而无法得到全局最优解。针对这些缺陷,人们在标准 BP 算法的基础上进行了许多的改进,如加入动量法[1]、自适应优化学习速率[2]等。这些方法在不同程度上改善了网络的性能,提高了收敛速度,但是它们的神经网络泛化能力都不甚理想。此处把 L-M 优化算法与贝叶斯正则化相结合,提出了一种 L-M 贝叶斯正则化 BP 神经网络,使 BP 网络的泛化能力有了很大的提高。

1.2.1　Levenberg-Marquardr 优化算法

在标准 BP 神经网络中采用梯度下降法来调整权值与阈值,但梯度下降法在达到误差最小的附近会出现精度低和收敛慢的缺陷。由文献[3]知,将高斯-牛顿法与梯度下降法相结合,会得出一种 L-M 优化算法。

具体算法如下:假设神经网络训练样本为:$D=(X^{(i)},t^{(i)})$,$i=1,2,\cdots,n$。其中,n 为训练样本数,$X^{(i)}$ 为

收稿日期:2013-02-03;**修回日期**:2013-03-13.
* **基金项目**:重庆市自然科学基金计划项目(CSTC2011BB2116).
作者简介:宋峰(1988-),男,山东淄博人,硕士研究生,从事人工神经网络与智能计算研究.

输入样本数据,$t^{(i)}=(t_1^{(i)},t_2^{(i)},\cdots,t_j^{(i)},\cdots,t_q^{(i)})$,$t^{(i)}$为期望输出。将BP神经网络各层权值和阈值用向量$W$表示。总误差为:

$$E=\frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{q}(t_j^{(i)}-O_j^{(i)})^2\right)^2=\frac{1}{2}\sum_{i=1}^{n}(e^{(i)})^2=\frac{1}{2}e^2 \tag{1}$$

其中,$e^{(i)}=\sum_{j=1}^{q}(t_j^{(i)}-O_j^{(i)})^2$,$O_j^{(i)}$表示第$i$个样本在输出层第$j$个神经元的实际输出,$t_j^{(i)}$为相应期望输出,$e=(e^{(1)},e^{(2)},\cdots,e^{(n)})$,并且$e$是$W$的函数。

假设在当前位于W_k(第k次迭代时网络的权值和阈值向量),并向W_{k+1}移动。如果移动量很小,则可将e展成二阶Taylor展开式:

$$e(W_{k+1})=e(W_k)+Z^{\mathrm{T}}(W_{k+1}-W_k)+\frac{1}{2}(W_{k+1}-W_k)^{\mathrm{T}}D(W_{k+1}-W_k) \tag{2}$$

其中,$Z^{\mathrm{T}}=2J^{\mathrm{T}}(W_k)e(W_k)$,$J(W_k)$为$e(W_k)$的雅克比矩阵;$D$为$e(W_k)$的Hessian矩阵,$D=2J^{\mathrm{T}}(W_k)J(W_k)$。要使$E$最小,即求$e$的最小值,显然可以对$W_{k+1}$求导,并使其导数为零。因此,会得到高斯-牛顿法的向量表达式:

$$W_{k+1}=W_k-(J^{\mathrm{T}}(W_k)J(W_k))^{-1}J^{\mathrm{T}}(W_k)e(W_k) \tag{3}$$

但是,高斯-牛顿公式经常会出现奇异现象,即$J^{\mathrm{T}}(W_k)J(W_k)$可能不是可逆矩阵。因此,记$H=J^{\mathrm{T}}(W_k)J(W_k)$,设

$$G=H+\mu I \tag{4}$$

其中I为单位矩阵。

假定H的特征根和特征向量分别为:$\{\lambda_1,\lambda_2,\cdots,\lambda_l,\cdots,\lambda_L\}$和$\{u_1,u_2,\cdots,u_l,\cdots,u_L\}$,则有:

$$Gu_l=(H+\mu I)u_l=Hu_l+\mu u_l=\lambda_l u_l+\mu u_l=(\lambda_l+\mu)u_l \tag{5}$$

因此,G的特征向量和H的特征向量相同,而特征值为$\lambda_l+\mu$。若增加μ使$\lambda_l+\mu>0$,可使G变为正定矩阵,而正定矩阵是可逆的。因此,把式(3)和式(4)结合,会得到L-M算法向量表达式,即:

$$W_{k+1}=W_k-(J^{\mathrm{T}}(W_k)J(W_k)+\mu I)^{-1}J^{\mathrm{T}}(W_k)e(W_k) \tag{6}$$

其中,μ为迭代变量。随着μ的增加,它接近于梯度下降法向量表达式。即当μ很大时,有:

$$W_{k+1}=W_k-\frac{1}{\mu}J^{\mathrm{T}}(W_k)e(W_k) \tag{7}$$

而当μ减小到0时,它变成式(3),即高斯牛顿法向量表达式[4]。因此,把式(6)带入式(1),最终会得到总误差:

$$E_{k+1}=\frac{1}{2}\|W_k-(J^{\mathrm{T}}(W_k)J(W_k)+\mu I)^{-1}J^{\mathrm{T}}(W_k)e(W_k)\|^2 \tag{8}$$

在计算过程中,依据误差E大小来决定μ的取值大小,即:

当$E_{k+1}<E_k$时,$\mu_{k+1}=\alpha_1\mu_k(\alpha_1<1)$;当$E_{k+1}>E_k$时,$\mu_{k+1}=\alpha_2\mu_k(\alpha_2>1)$。

L-M算法既可以避免高斯-牛顿法中容易出现的Jacobian矩阵病态和假收敛的现象,又可以避开梯度下降法中在精确解附近逼近精度低和假收敛的缺点。L-M算法能够保证权值和阈值的每次调整都使E减小,避免出现网络的震荡[5]。

1.2.2 贝叶斯正则化方法

(1) 正则化方法。在式(8)基础上,增加了一个限制逼近函数复杂性的一项E_ω,使得网络的性能函数(误差函数)改进为:

$$F(W)=\alpha E_\omega+\beta E_D \tag{9}$$

其中,

$$\begin{cases}E_\omega=\dfrac{1}{M}\sum_{m=1}^{M}\omega_m^2=\dfrac{1}{M}\|W\|^2\\ E_D=\dfrac{1}{2}\|W_k-(J^{\mathrm{T}}(W_k)J(W_k)+\mu I)^{-1}J^{\mathrm{T}}(W_k)e(W_k)\|^2\end{cases} \tag{10}$$

M 为网络参数总和,式(9)中 α,β 是超参数,控制着其他参数(包括权值和阈值)的分布形式。对于正则化方法而言,难点在于超参数的确定。而贝叶斯正则化方法则可以在网络训练过程中自适应地调整超参数大小,并使其达到最优。

(2) 贝叶斯学习。David Mackay[6]将贝叶斯方法用于神经网络建模过程中,之后又对网络训练过程中后验概率计算问题提出了运用高斯逼近方法来计算 Hessian 矩阵。在贝叶斯理论的框架下,网络的参数被认为是随机变量,给定样本数据,由贝叶斯规则,参数分布函数为:

$$P(W\mid D,\alpha,\beta)=\frac{p(D\mid W,\beta)P(W\mid\alpha)}{p(D\mid\alpha,\beta)} \tag{11}$$

式(11)中,$p(D|W,\beta)$ 为似然函数,$p(D|\alpha,\beta)$ 为归一化因子,$P(W|\alpha)$ 为先验概率,表示在没有数据样本下的参数 W 的先验知识。

假设训练样本的总体分布是正态分布,权参数的先验分布也是正态的。从而,似然函数和先验分布函数分别为:

$$p(D\mid W,\beta)=\frac{1}{Z_D(\beta)}\exp(-\beta E_D) \tag{12}$$

$$P(W\mid\alpha)=\frac{1}{Z_W(\alpha)}\exp(-\alpha E_W) \tag{13}$$

式(12)、(13)中,$Z_W(\alpha)=(\frac{\pi}{\alpha})^{\frac{M}{2}}$,$Z_D(\beta)=(\frac{\pi}{\beta})^{\frac{n}{2}}$,把式(12)(13)带入式(11),得到:

$$P(W\mid D,\alpha,\beta)=\frac{\frac{1}{Z_W(\alpha)}\frac{1}{Z_D(\beta)}\exp(-\alpha E_W-\beta E_D)}{\text{归一化因子}}=\frac{1}{Z_F(\alpha,\beta)}\exp(-F(W)) \tag{14}$$

贝叶斯学派认为,最优的权参数应该极大化后验概率 $P(W|D,\alpha,\beta)$。对 W 来说,$Z_F(\alpha,\beta)$ 可以看做常数。因此,从式(14)可以得出,极大化后验概率等价于极小化总误差函数 $F(W)$。

下面来优化超参数,应用贝叶斯规则求得后验分布:

$$P(\alpha,\beta\mid D)=\frac{P(D\mid\alpha,\beta)p(\alpha,\beta)}{p(D)} \tag{15}$$

假设先验分布是一个很宽的分布函数,也即均匀分布,因为式(15)中的归一化因子与 α,β 无关,因此求取最大后验概率的问题就转化为求解最大似然函数 $P(D|\alpha,\beta)$。由于似然函数是式(11)的归一化因子,因此可得到:

$$P(D\mid\alpha,\beta)=\frac{Z_F(\alpha,\beta)}{Z_D(\beta)Z_W(\alpha)} \tag{16}$$

当样本数据较多时,后验分布趋于正态分布,若后验概率曲线足够窄,峰值足够尖锐,则可利用泰勒展开式对问题作进一步的简化,以求得 $Z_F(\alpha,\beta)$。

设 $F(W)$ 取最小值时所对的参数为 $W_{优}$,将 $F(W)$ 在 $W_{优}$ 附近泰勒展开,忽略高次项,得到:

$$Z_F=(2\pi)^{\frac{M}{2}}(\det((\nabla^2F(W_{优}))^{-1}))^{\frac{1}{2}}\exp(-F(W_{优})) \tag{17}$$

式(17)中,$\nabla^2F(W_{优})=\beta\nabla^2E_D+\alpha\nabla^2E_W$ 是 $F(W)$ 在 $W_{优}$ 点的 Hessian 矩阵。将式(17)带入式(16),并利用极大似然原理,求出满足似然函数 $P(D|\alpha,\beta)$ 的最大的 α 和 β,即得到最优超参数:

$$\begin{cases}\alpha_{优}=\dfrac{\gamma}{2E_W(W_{优})}\\ \beta_{优}=\dfrac{n-\gamma}{2E_D(W_{优})}\end{cases} \tag{18}$$

其中,$\gamma=M-2\alpha_{优}\,\mathrm{tr}((\nabla^2F(W_{优}))^{-1})$,成为有效参数个数。$\gamma$ 表示有多少参数在减少总误差方面起作用,它的取值为[0,M]。

在进行优化求解时,需要计算 $F(W)$ 在其最小点 $W_{优}$ 处的 Hessian 矩阵也即需要计算 $\nabla^2 F(W_{优})$,计算量较大。为了提高速度,可以利用高斯-牛顿逼近法对 Hessian 矩阵做进一步简化,得:$\nabla^2 F(W_{优}) = 2\beta J^{\mathrm{T}} J + 2\alpha I_M$,其中,$J$ 是 E_D 在点 $W_{优}$ 处的雅克比矩阵[7]。

2 改进的 BP 神经网络在预测居民消费水平中的应用

2.1 数据的获取以及归一化

为了评估成都市居民消费水平,在此选取一些影响成都市消费水平的因素,建立它们与反映成都市消费水平的的成都市社会消费品零售总额之间的关系,并进行预测。选取 1997 年 1 月至 1998 年 8 月之间的数据,作为样本数据[2]。并选取 1997 年 1 月至 1998 年 3 月作为训练集,把 1998 年 4 月至 8 月作为样本测试集。

2.2 网络结构的确立

由于影响成都市社会消费品零售总额的变量主要有 8 个,因此,选择 BP 神经网络结构为输入节点的数 8 个,输出节点数为 1 个。对于隐含层节点的确定,采用试凑法。选取经验公式 $h<n-1$ 和 $h=\sqrt{pq}+\theta$(h 为隐层节点数,p 为输入层节点数,q 为输出层节点数,n 为样本数,θ 为 1-10 中的任意数),并综合考虑,得出隐节点的个数在 9-19 个[1,8]。通过 trainlm 训练函数进行训练网络得出:如图 1 隐节点为 15 时,只迭代 6 次,就达到精度[9]。也即网络结构取为 8-15-1。选取隐节点的激活函数是 logsig,输出节点的激活函数是 logsig,最大迭代次数为 1 000,误差上限为 0.000 01。

2.3 对比试验及结论

通过 MATLAB7.0 编程[9],分别用 3 种方法对 BP 网络进行训练,并以 1998 年 4 月到 1998 年 8 月为测试样本,对成都市消费水平做出预测。结果如表 1 所示。

表 1 各模型预测结果及误差分析

日期\类型	成都市社会消费品零售总额(亿元)	L-M 贝叶斯正则化方法		L-M 优化算法		自适应学习速率的梯度下降动量法	
		预测值(亿元)	相对误差	预测值(亿元)	相对误差	预测值(亿元)	相对误差
1998.4	32.70	33.451 8	0.023 0	37.939 5	0.160 2	38.555 9	0.179 1
1998.5	33.51	33.493 8	-0.000 48	38.776 6	0.157 2	37.816 5	0.128 5
1998.6	34.20	33.370 4	-0.024 3	36.205 4	0.058 6	37.173 7	0.087 0
1998.7	33.87	31.231 3	-0.007 5	34.123 6	-0.077 9	36.406 0	0.074 9
1998.8	34.38	31.450 5	-0.008 8	34.078 1	-0.085 2	36.373 4	0.058 0

从表 1 可以看出,L-M 贝叶斯正则化方法所得的模型精度高,性能最稳定,有较强的泛化能力。并且由以图 1、图 2、图 3 可知,L-M 算法收敛速度最快,但精度相对 L-M 贝叶斯正则化方法来说偏低。而对于自适应学习效率的梯度下降法,相对 L-M 贝叶斯正则化方法和 L-M 优化算法来说,收敛速度还需很大程度的提高,并且在精度、稳定性上也有一定差距。

3 结束语

运用 L-M 贝叶斯正则化 BP 神经网络对成都市居民消费水平进行预测,通过对比试验,得知其具有较强

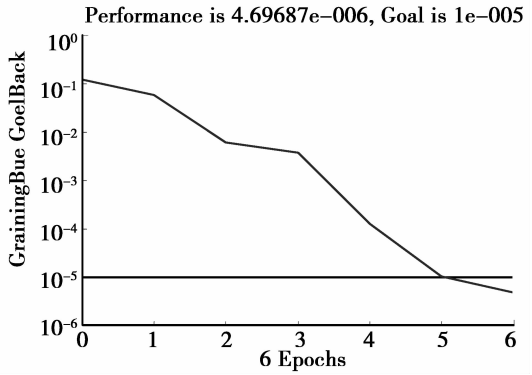

图 1　15 个隐节点的迭代过程

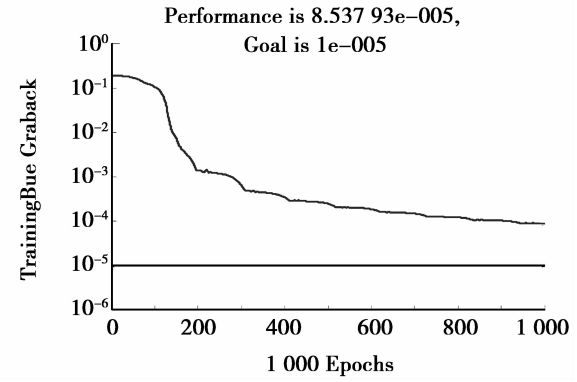

图 2　加入动量的 BP 神经网络

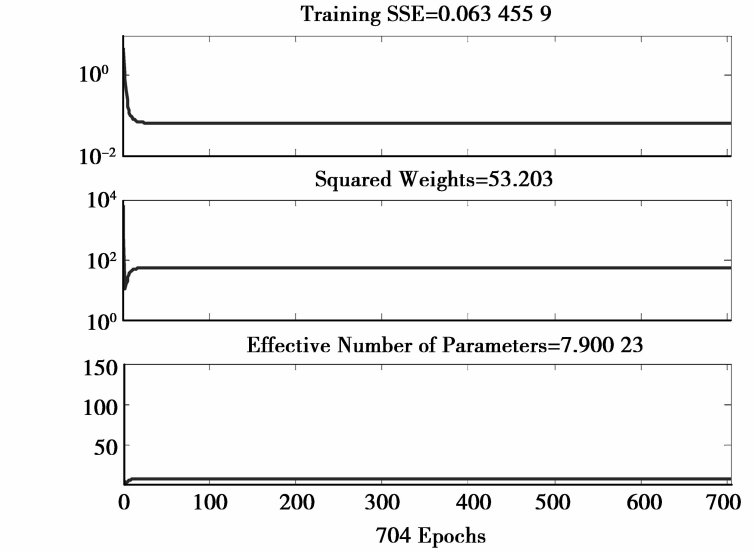

图 3　L-M 贝叶斯正则化的 BP 的神经网络

的泛化能力。但是,在收敛速度上,还需要更进一步提高,并且由于需要计算 J^TJ,所需计算机内存较大。但在有限的样本量时,使网络具有较强的泛化能力,是一种值得推广的方法。

参考文献:

[1] 韩立群.人工神经网络理论、设计及应用[M].2 版.北京:化学工业出版社,2007

[2] 李晓峰,徐玖平,王萌清,等.BP 人工神经网络自适应学习算法的建立及其应用[J].系统工程理论与实践,2004(5):3-5

[3] 田建平,曹东卫,李海楠.LM-BP 神经网络在于桥水库水质预测中的应用[J].水利信息化,2010(3):31-32

[4] 王赟松,许洪国.快速收敛的 BP 神经网络算法[J].吉林大学学报:工学版,2003,33(4): 80-81

[5] 吴方良,石仲堃,杨向辉,等.基于 L-M 贝叶斯正则化方法的 BP 神经网络在潜艇声纳部位自噪声预报中的应用[J].船舶力学,2007,11(1):136-142

[6] MACKAY D.Bayesian interpolation[J].Neural Computation,1992(4):415-447

[7] FORESEE F D,HAGAN M T.Gauss-Newton approxination to bayesian learning[A].In Proce-edings of International Conference on Neural Networks[C].Houston,Texas,1997

[8] 康迪,马寿峰,钟石泉.基于 BP 神经网络的微观交通安全预测方法[J].交通信息安全,2011,3(29):81

[9] 葛哲学,孙志强.神经网络理论与 MATLAB R2007 实现[M].北京:电子工业出版社,2007

(下转第 89 页)