基于贝叶斯神经网络的非参数回归*

杨 斌**[1] 聂在平[1] 夏耀先[2] 蒋荣生[2]

(1.电子科技大学电子工程学院 成都 610054; 2.中海油田服务有限公司 北京 101149)

【摘要】提高神经网络模型推广能力的关键是控制模型的复杂度。该文探索了贝叶斯神经网络的非参数回归的建模方法，通过融入模型参数的先验知识，在给定数据样本及模型假设下进行后验概率的贝叶斯推理，使用马尔可夫链蒙特卡罗算法来优化模型控制参数，实现了对神经网络模型中不同部分复杂度的控制，获得了模型参数的后验分布及预测分布。在5个含噪二维函数回归问题上的应用显示了模型的复杂度能根据数据的复杂度而自适应调整，并给出了较好的预测结果。

关 键 词 贝叶斯神经网络; 非参数回归; 正则化器; 马尔可夫链蒙特卡罗模拟

中图分类号 TP183

Bayesian Neural Network for Nonparametric Regression

Yang Bin[1] Nie Zaiping[1] Xia Yaoxian[2] Jiang Rongsheng[2]

(1.College of Electronic Engineering, UEST of China Chengdu 610054; 2.China National Offshore Oil Corporation Service Beijing 101149)

Abstract With neural networks, the main difficult in improving the model generalization capability is controlling the complexity of the model. This paper investigates a Bayesian neural network learning for nonparametric regression. Prior knowledge about the model parameters can be incorporated within Bayesian inference and combined with training data to control complexity of different parts of the model. A Markov chain Monte Carlo algorithm is used to optimize model control parameters and obtain the predictive distribution. We show that the complexity of the models adapts to the complexity of the data and produces good results on five noisy test functions in two dimension. The performance and advantage of this approach are compared with conventional neural network methods.

Key words neural networks; nonparametric regression; regularizer; Markov chain Monte Carlo

越来越多的研究表明，若不对神经网络的复杂度进行控制，结果将不可避免地出现过拟合而降低网络的推广(泛化)能力。目前有很多克服神经网络训练过拟合的方法，有效地通过加入一个或多个惩罚复杂函数的正则化器来改变目标函数的正则化，如权衰减法[1]，正则化器为 aE_w ，a 为权衰减率或正则化常数，E_w 为模型权向量内各分量的平方和。当使用这样一个正则化器时，克服过拟合问题就转化为对该函数复杂性起控制作用的超参数 a 的设定问题。若 a 值太大，对复杂函数惩罚过度而使内插太平滑导致忽略数据内的本质结构；若 a 值太小，则失去对模型参数的控制作用而自然就会出现过拟合，两者结果都将给出较差的网络推广能力。在训练过程中如何合理而有效地控制超参数 a 值是亟待解决的难题。

2001年11月13日收稿

* 国家自然科学基金资助项目，编号：69871004；油气藏地质及开发工程国家重点实验室开放基金资助项目，编号：PLC9913

** 男 34岁 博士 副教授

贝叶斯神经网络为解决 α 值问题提供了新思路和新方法。在贝叶斯分析框架下将模型参数视为不确定性量，使用显式的概率分布假设进入到模型中并加以分析和推断，对未知变量的先验知识通过先验分布来定量表述。它将数据的误差解释为一个似然函数定义，而正则化器可对应于在网络权上的先验概率分布，贝叶斯神经网络是通过融入先验分布的假设由给定的观测数据来调整寻找出权变量后验概率分布，网络预测就是基于后验分布的贝叶斯推理。由于神经网络的模型复杂度在贝叶斯框架中可以很自然地加以显式表达和控制，通过定义一些超参数的模糊先验来控制模型参数复杂性的未知程度，再进一步通过对模型参数的分组，及在各权组上使用不同的公共超参数，允许模型在不同的部分具有不同的复杂度。目前对参数、超参数控制和贝叶斯推理近似方法主要有两种，一种是MacKay的基于经验贝叶斯的高斯近似法[2]，另一种是Neal的基于马尔可夫链蒙特卡罗MCMC的全贝叶斯方法[3]。后者用蒙特卡罗模拟来从参数和超参数的联合概率分布中采样，马尔可夫链可被看成是后验平衡分布中的样本。本文探索了贝叶斯神经网络用于非参数回归上的实现方法，得到了模拟实验结果。

1 贝叶斯神经网络

考虑一个回归问题，由数据 $D=\{\boldsymbol{x}_i,t_i\}_{i=1}^N$ 进行 x~t 的映射和回归。x_i 为第 i 个样本向量，t_i 为第 i 个学习样本期望输出值。使用带有加性噪声 e 的确定性函数将目标与输入联系起来：$t=f(\boldsymbol{x})+e$，其中 e 为高斯分布 $N(0,\sigma_t^2)$ 的噪声。给定新输入样本 x 的期望 t 的概率密度为

$$P(t\mid \boldsymbol{x})\propto \exp[-\frac{1}{2\sigma_t^2}(f(\boldsymbol{x})-t)^2] \tag{1}$$

当给定了依赖于新的输入 x 及模型权 W 集有输出 $y(x;W)$ 值的神经网络模型来进行回归时，其参数积分为

$$P(y\mid \boldsymbol{x},D)=\int P(y\mid \boldsymbol{x},\boldsymbol{W})P(\boldsymbol{W}\mid D)\mathrm{d}\boldsymbol{W} \tag{2}$$

式中 $P(W\mid D)$ 为权参数后验分布。令 $P(W)$ 为在获得任何训练数据之前对网络权参数 W 的先验分布，则定义一个正则化器为

$$R(\boldsymbol{W})=-\ln P(\boldsymbol{W}) \tag{3}$$

应用贝叶斯规则及式(2)、(3)得

$$\ln P(\boldsymbol{W}\mid D)=\ln P(D\mid \boldsymbol{W})+\ln P(\boldsymbol{W})+\mathrm{const}=-E(\boldsymbol{W})-R(\boldsymbol{W})+\mathrm{const} \tag{4}$$

式中const代表常数。定义贝叶斯神经网络回归模型的总误差函数为 $U(\boldsymbol{W})=E(\boldsymbol{W})+R(\boldsymbol{W})$，由式(4)得

$$P(\boldsymbol{W}\mid D)=(1/Z)\exp[-U(\boldsymbol{W})] \tag{5}$$

式中 Z是一个归一化常数。将式(5)代入式(2)得

$$P(y\mid \boldsymbol{x},D)=\int P(y\mid \boldsymbol{x},\boldsymbol{W})\frac{1}{Z}\exp[-U(\boldsymbol{W})]\mathrm{d}\boldsymbol{W} \tag{6}$$

由此，网络预测就变成了对式(6)的计算，但式(6)无法直接解析求解，这里求助于一些数值积分近似法，如马尔可夫链蒙特卡罗法(MCMC)。假设马尔可夫链满足遍历性，则由该链上的大量采样所达到的稳定分布就可代表式(5)的后验概率分布。用结果马尔可夫链上的样本序列 $\{W_t\}$，对式(6)积分的近似计算式为

$$P(y\mid \boldsymbol{x},D)=\frac{1}{n_s}\sum_{t=n_0+1}^{n_0+n_s}P(y\mid \boldsymbol{x},\boldsymbol{W}_t) \tag{7}$$

式中 n_0 为被舍弃的一些初始马尔可夫链以保证收敛性；n_s 为来自平衡后验分布的权向量样本数。

对模型自由度控制的先验知识定量地体现在权参数分布 $P(W)$ 中。将 W 参数放置一个模糊收缩均

值为0的高斯先验：$P(W)\sim N(0,\alpha)$。这里α是正态分布中的精度(方差的倒数)，称为超参数。较大的α值导致整个或权组内权幅度的较小变化，其作用类似于权衰减法中的权值惩罚系数。超参数也存在一些不可忽视的不确定性，可使用更高一级的先验分布(超先验)来控制超参数α的取值。设超参数服从的共轭超先验为倒Gamma分布：$P(\alpha)\sim Inv\_gamma(\alpha_0,n_\alpha)$，$\alpha_0$和$n_\alpha$为给定的参数。由于引入了未知超参数$\alpha$，因此需在式(5)、(6)中将$\alpha$显式地表示出来。

2　马尔可夫链蒙特卡罗模拟

MCMC法可被用于对式(6)的积分近似。马尔可夫链模拟过程是通过产生一个样本$\{W_1,W_2,\cdots,W_n\}$链，而成员是来自与条件概率$P(W|\alpha,D)$成直接比例的采样，n就为链长($n=n_0+n_s$)。设给定了迭代步t的当前状态W_t，从一个关于W_t的对称分布中获得一个新的候选状态$\tilde{W}$,若$U(\tilde{W})<U(W_t)$，则接受该候选状态为新的状态，否则，仅以概率$P(W_{t+1}=\tilde{W}\,|W_t)=\exp[U(W_t)-U(\tilde{W})]$来接受。结果由选择的一系列$\{W_t\}$样本组成了一个马尔可夫链。Metropolis算法缺点是以缓慢的随机行走方式进行$\{W_t\}$空间的探索，连续的采样间存在高度的相关，使达到平衡分布的时间会很长，在此使用了Duane的混合蒙特卡罗(HMC)法来加快收敛速度[4]。它综合了Metropolis－Hastings算法与一个动力系统的模拟，利用了梯度信息而避免随机行走。从后验分布$P(W,\ \alpha\,|D)=P(\alpha|D)P(W|\alpha,D)$采样遵循二步法，第一步固定超参数$\alpha$，用混合蒙特卡罗法来对网络权的后验分布采样；第二步改变超参数α，从后验分布$P(\alpha\,|W,D)$中获得一个样本，这个二步法过程就执行了一个块吉布斯(Gibbs)采样[3,4]。

3　在二维函数回归上的应用

将MCMC的贝叶斯神经网络法应用于5个含噪函数的回归，其名称和形式如下[5]

1) 简单互作用函数(Sif)

$$f_1(x_1,x_2)=10.391((x_1-0.4)(x_2-0.6)+0.36) \tag{8}$$

2) 径向函数(Rad)

$$\begin{aligned} f_2(x_1,x_2)&=24.23(r^2(0.75-r^2))\\ r^2&=(x_1-0.5)^2+(x_2-0.5)^2\end{aligned} \tag{9}$$

3) 调和函数(Harm)

$$\begin{aligned} f_3(x_1,x_2)&=42.659(0.1+\tilde{x}_1(0.05+\tilde{x}_1^4-10\tilde{x}_1^2\tilde{x}_2^2+5\tilde{x}_2^4))\\ \tilde{x}_1&=(x_1-0.5),\tilde{x}_2=(x_2-0.5)\end{aligned} \tag{10}$$

4) 加性函数(Cadd)

$$f_4(x_1,x_2)=1.335\,6(1.5(1-x_1)+\mathrm{e}^{2x_1-1}\sin(3\pi(x_1-0.6)^2)+\mathrm{e}^{3(x_2-0.5)}\sin(4\pi(x_2-0.9)^2)) \tag{11}$$

5) 复杂互作用函数(Cif)

$$f_5(x_1,x_2)=1.9(1.35+\mathrm{e}^{x_1}\sin(13(x_1-0.6)^2)\mathrm{e}^{-x_2}\sin(7x_2)) \tag{12}$$

形成5个函数的225个含噪声训练数据及10 000个检验用数据的方法同文献[5]。预测误差测度时使用FVU(Fraction of Variance Unexplained)值定义为

$$FVU=\sum_{i=1}^{N}(t(\boldsymbol{x}_i)-y(\boldsymbol{x}_i))^2\Big/\sum_{i=1}^{N}(t(\boldsymbol{x}_i)-\bar{t})^2 \tag{13}$$

式中　N为样本个数；x_i为输入的样本向量，$t(x_i)$为期望函数值，$y(x_i)$为网络实际输出值，$\bar{t}$是期望输出的均值。所有的MCMC模拟都是先在准平衡阶段运行50～100个迭代步，接着进行400～800个迭代步较长时间的采样阶段，取最后200个迭代步$n_s=200$上的链样本作为最终预测式(7)的模型。网络隐层单元为tanh传输函数，输出单元为线性函数。为了控制模型不同部分的复杂度，将模型参数分为4个权组，分别为输入－隐层、隐层单元阈值、隐层－输出层、输出层单元阈值。每个权组服

从均值为0、精度为a_g (g=1,…,4)的高斯先验分布，而超参数a_g由一个模糊Gamma先验分布控制，取该模糊先验分布的形状参数a_0=0.5，均值n_a=0.05。对225个有噪声样本的训练及预测结果如表1所示。FVU训练误差上5个函数具有相似的均值和较小的标准偏差说明各自模拟的马尔可夫链达到了稳定状态，得到的预测分布是收敛的平衡分布。

表1 5个函数的含噪声训练样本的学习和预测结果

函数	网络结构	FVU训练误差均值	FVU训练误差标准偏差	FVU预测误差均值	FVU预测误差标准偏差
Sif	2-10-1	0.062 6	0.005 6	0.001 96	0.000 07
Cif	2-10-1	0.056 7	0.005 7	0.025 58	0.000 61
Rad	2-10-1	0.050 9	0.005 1	0.009 06	0.000 39
Cadd	2-10-1	0.053 5	0.006 7	0.017 08	0.000 24
Harm	2-10-1	0.050 5	0.004 8	0.013 09	0.000 34

为了显示预测分布的预测推广能力，本文与用其他算法的预测结果进行了对比，其比值如表2所示。表中GNBP为批训练高斯-牛顿算法[5]，PPL为投影寻踪学习算法[5]，SMART为自动平滑样条投影寻踪回归算法，Reg.JNN为正则化Jacobian网络学习算法。可见在预测误差上，对于Sif、Cadd、Cif三个非线性函数，本文方法已获得了最小的预测误差。对于另两个函数，本文的结果亦有小的预测误差，体现出贝叶斯神经网络较好的预测推广性能。

表2 在独立检验集上5个函数的检验误差FVU值表

算法	Sif	Rad	Harm	Cadd	Cif
GNBP	0.017 00	0.026 0	0.210 0	0.019	0.070
PPL	0.007 67	0.032 7	0.091 0	0.007	0.031
SMART	0.018 00	0.016 0	0.160 0	0.008	0.049
Reg.JNN	0.011 00	0.008 0	0.024 0	0.053	0.061
本文方法	0.001 96	0.009 1	0.013 1	0.017	0.026

4 结 论

贝叶斯推理的神经网络方法以一种完全不同于常规神经网络学习和预测方式，利用网络模型参数的先验知识和所获得的样本数据信息，得到网络参数的后验分布，并对后验分布进行预测。将网络模型复杂度的不确定性以超参数变量的形式和显式的概率分布融入到贝叶斯模型中，且能根据实际样本数据的特性自动加以调整模型参数的分布，达到了控制模型复杂度的效果，有关函数回归问题有待作进一步研究与探讨。

参 考 文 献

1 阎平凡，张长水. 人工神经网络与模拟进化计算. 北京：清华大学出版社，2000

2 MacKay D C. A practical Bayesian framework for backpropagation networks. Neural Computation, 1992, 4(3): 415-447

3 Neal R M. Bayesian learning for neural networks. Berlin:Springer-Verlag, 1996

4 Duane S, Kennedy A D. Hybrid Monte Carlo. Physics Letters B, 1987, 195(2): 216-222

5 Hwang J, Lay S, Maechler M. Regression modeling in back-propagation and projection pursuit learning . IEEE Trans Neural Networks, 1994, 5(3): 342-353