高技术通讯 2019 年 第 29 卷 第 12 期:1165 ~ 1174

doi:10.3772/j.issn.1002-0470.2019.12.001

基于超限学习机与随机响应面方法的深度学习超参数优化算法①

孙永泽②*** 陆忠华③*

(*中国科学院计算机网络信息中心 北京 100190)

(**中国科学院大学 北京 100049)

摘 要 恰当的超参数设置是决定深度模型性能的关键因素,实现优秀高效的超参数优化算法能够提高深度学习模型的效果,提升模型超参数搜索调优的效率和速度,降低深度学习模型的应用门槛。超参数优化算法的典型代表是贝叶斯优化算法(BOA),此类基于代理模型的全局优化算法,相对随机搜索、网格搜索等简单算法理论上具备更好的优化效率。本文提出基于超限学习机(ELM)对超参数空间建立确定性代理模型,并改进随机响应面方法,实现了一种针对深度学习模型的超参数优化算法 SurroOpt1。实验表明,本文提出的算法,在深度卷积网络模型超参数优化任务中,相对贝叶斯优化和 TPE 算法这 2 种最先进的已知算法,在函数求解次数相同的情况下,具备更好的模型优化效果。

关键词 超参数优化;代理模型;超限学习机(ELM);随机响应面法;深度学习模型

0 引 言

各类深度学习模型如深度神经网络、深度卷积网络、长短期记忆网络等,近期被广泛用于图像分类、人脸识别、文本分类等领域。深度学习模型一般包含多个需要预先选择、设置的超参数,如网络结构、激活函数和学习速率等,超参数的设置将直接决定深度学习模型的性能与实际应用效果。然而,人为的手动设置、调整超参数是一项极为困难的任务,为获得相对较优的超参数设置,一般需要丰富的经验并进行大量的反复尝试,这成为影响深度学习模型在各类不同问题中能否成功应用的主要因素。因此,实现自动化、高效的超参数优化算法对深度学习模型的研究、应用有重要的现实意义,此研究领域吸引了相当的关注[1-3]。

深度学习模型的超参数优化问题可以定义为,对于模型需要设置的 d 个超参数 $\theta \in R^d$,找到最优的超参数设置 θ^*,使得基于此超参数设置训练得到的深度学习模型,有最优的性能评价指标 λ,常用的性能评价指标包括验证集误差等。一般认为 d 个超参数 θ 和性能评价指标之间存在未知函数关系 $\lambda = f(\theta)$,其中函数的解析形式未知,必须通过实际的训练来获得特定对应的值,而深度模型的训练一般需要较长的时间,当网络结构复杂、训练数据集规模较大时,单次的训练时长可能达到数天。因此获取一组超参数 θ 的性能评价值 λ,即求解函数值 $f(\theta)$,需要很大的计算开销。综上所述,可以将深度学习模型的超参数优化问题定义为一个全局高开销黑盒函数优化问题(expensive black-box global function optimization)。

由于超参数优化的高开销特性,导致解决相应的优化问题极为困难。一些简单的方法如随机搜索、网格搜索,在中小规模机器学习模型超参数优化问题中能够获得较好的效果,但在高开销场景中必须实现更高的效率以较少的函数求解次数尽快获得

① 国家自然科学基金(61873254)资助项目。

② 男,1988 年生,博士生;研究方向:高性能计算与计算金融;E-mail:sunyongze@outlook.com

③ 通信作者,E-mail:zhlu@sccas.cn

(收稿日期:2019-01-21)

相对较优解,简单的随机搜索和网格搜索算法在这一点上缺少理论基础和相关机制;而被广泛应用于全局黑盒函数优化问题的各种基于种群的方法,如遗传算法、差分进化、协方差进化算法等,每一步都需要求解种群中各个个体的函数值,导致单步较高的计算代价,此类算法在类似于深度模型超参数调优这种高代价函数优化场景中应用较少。为提高优化效率,基于代理模型的全局黑盒函数优化算法在研究和实践中得到了重点的关注,此类算法通过代理模型利用有限的已获取的参数值和其真实函数值,对函数关系 $f(\theta)$ 建立近似模型,来推测参数空间中的更优点位置,提高算法的效率,降低优化过程中所需要的函数求解次数。近期受到广泛关注的代表为贝叶斯优化算法(Bayesian optimization algorithm, BOA),其使用概率模型高斯过程对参数空间进行建模,并能够获得预测的不确定性度量,以此为基础建立优化策略。贝叶斯优化算法被广泛地应用于深度学习模型的超参数优化任务中[2,4]。其他被广泛应用的基于代理模型的超参数优化算法还有 TPE(tree-structured Parzen estimator)等[1]。

本文提出了一种采用超限学习机作为确定性代理模型的超参数优化算法,主要贡献如下:

(1) 提出使用超限学习机作为代理模型用于超参数优化任务中;

(2) 改进了随机响应面方法中的优化策略,基于超限学习机代理模型,建立了用于深度学习模型的超参数优化算法 SurroOpt1;

(3) 通过实验证明,本文提出的算法相比经典的 BOA 算法、TPE 算法,在深度学习优化任务中具备性能优势,特别是在超参数个数较多、超参数空间维度较高的情况下,性能和效率优势更为显著。

1 相关工作

文献[5]总结了基于代理模型进行全局优化的一般性框架,基于模型的贯序优化(sequential model-based optimization, SMBO),符合 SMBO 的全局优化算法一般可以用算法 1 表示。

算法 1 基于模型的贯序优化(SMBO)

(1) $S \leftarrow \varnothing$; /* S 表示已求解的参数点和对应的值 */;

(2) for $t \leftarrow 1$ to T;

(3) $x^* \leftarrow \text{argmin}_x A(x, M_{t-1})$;/* 基于代理模型 M 定义采集函数 A,并基于采集函数决定下一个函数求解点 x^* */;

(4) 求解函数值 $f(x^*)$;

(5) $S \leftarrow S \cup (x^*, f(x^*))$;/* 更新已知点集 */;

(6) $M_t \leftarrow fit(S)$;/* 基于已知点集 S 更新代理模型 M */;

(7) end for;

(8) return S。

符合 SMBO 框架的典型算法包括贝叶斯优化算法(BOA)和 TPE 算法。BOA 算法使用概率模型高斯过程作为代理模型,并基于代理模型预测值和不确定性度量建立采集函数,常用的采集函数包括提升期望、提升概率等。TPE 算法是当前在深度学习模型超参数调优中广泛采用的另一种算法,与贝叶斯优化算法不同,TPE 算法不直接对参数空间的函数值分布建立拟合模型,而是将参数空间划分为高函数值和低函数值 2 部分,并使用派尔森树(Parzen tree)模型对参数在 2 个部分的概率密度分布进行建模,在此基础上构建采集函数。

BOA 算法和 TPE 算法都被广泛应用于超参数优化任务中[6,7],是当前最具代表性的 2 种深度学习模型调参算法。文献[7]表明,贝叶斯优化算法和 TPE 算法分别在参数数量较少(低维超参数空间)和数量较多时(高维超参数空间)取得最优的超参数优化效果和效率,因此本文提出的算法将与贝叶斯优化算法和 TPE 算法在典型深度学习超参数优化任务上进行对比。

确定性模型同样能够作为代理模型用于全局优化问题。文献[8]使用神经网络作为代理模型,提高了遗传算法在空气动力学问题中的优化效率。文献[9]使用支持向量机代理模型提高了 CMA-ES 算法(covariance matrix adaptation evolution strategy)的

优化效率,并建立了自适应的代理模型调用策略。文献[10]使用深度神经网络代理模型用于多目标优化算法 NSGA-ii,在不同尺度上建立参数空间的预测模型。

文献[11]提出了基于确定性代理模型和随机采样方法进行全局优化的通用性框架随机响应面方法。通过随机采样在参数空间内采样生成大量候选点,并基于代理模型和距离尺度对候选点进行打分,选择得分最高的候选点作为下一个函数值求解点。可以看出,随机响应面方法仍然符合 SMBO 框架。文献[11]基于随机响应面法和径向基函数模型(radis basis function, RBF)代理模型建立了2种优化算法,全局随机响应面算法和局部随机法,其中局部随机法通过在当前最优点附近进行随机扰动产生候选点,是侧重于局部搜索的算法,而全局随机响应面法在整个参数空间采样产生候选点,是偏向于全局搜索的方法。文献[12]进一步为局部随机响应面法引入了动态坐标搜索机制,在产生候选点时随机固定当前最优点的不同坐标轴,提高了随机响应面方法在高维全局优化问题中的性能。文献[13]证明了基于候选点采样与代理模型的全局优化方法与遗传算法具备同样的性能。文献[14]将随机响应面方法应用到深度学习模型的超参数优化中,获得了与贝叶斯优化等方法同等或更优的性能。

超限学习机(extreme learning machine, ELM)是文献[15]提出的一种单隐层结构神经网络。与一般的神经网络使用误差梯度反向传播方法训练更新网络权重不同,超限学习机的输入层权重通过随机采样生成,输出层权重通过最小二乘法直接确定,不需要迭代求解。超限学习机训练速度快、使用简单,特别是在中小规模数据场景下,超限学习机表现出良好的泛化性能。文献[16]从理论上证明了超限学习机和最小二乘支持向量机(least square support vector machine, LS-SVM)、近似支持向量机(proximal support vector machine, PSVM)模型在理论上具有等价形式,通过实验表明超限学习机在多个数据集的回归与分类任务中具备同等或优于支持向量机算法的性能,且其泛化性能不随输入维度的增加而退化。因此相对一般的确定性模型如神经网络、支持向量机,超限学习机具备更快的训练与更新速度,且在回归任务中具备良好的泛化性能[17]。本文研究并证明了基于超限学习机作为参数空间的代理模型用于优化算法的可行性。

2 基于超限学习机的代理模型

本节主要介绍如何构建基于超限学习机的参数空间代理模型,其中 2.1 节介绍超限学习机的主要概念,2.2 节描述如何基于超限学习机构建参数空间代理模型。

2.1 超限学习机

超限学习机是一种单隐层前馈神经网络,由输入层、隐含层和输出层构成,超限学习机的基本网络结构如图 1 所示。

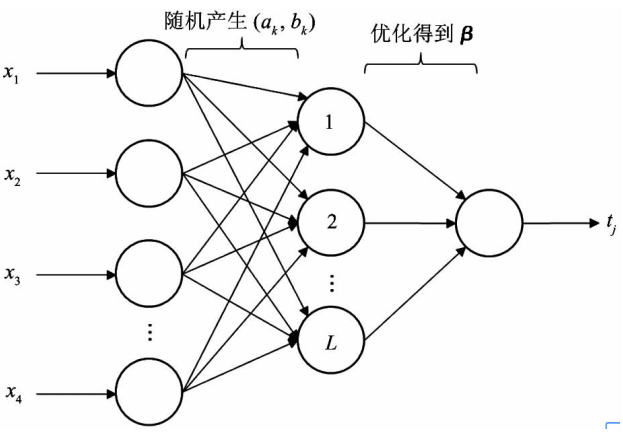

图 1 超限学习机的网络结构

对于 N 个训练样本 $(x_i, t_i) \in R^n \times R^m$,其中 x_i 为第 i 个 n 维输入样本,t_i 表示对应的目标输出,则 L 隐层节点的 ELM 模型输出可以表示为

$$f(x_i) = \sum_{k=1}^{L} \boldsymbol{\beta}_k G(a_k, b_k, x_i) = t_i \quad i = 1,2,\cdots,N \quad (1)$$

式中,α_k、b_k 是输入层权重和偏置,权值从均匀分布中随机采样生成;$\boldsymbol{\beta}_k$ 是连结第 k 个隐层单元和输出单元之间的权重矩阵;$G(a_k, b_k, x_i)$ 是第 k 个隐层节点的激活函数。将式(1)写成矩阵形式 $\boldsymbol{H\beta} = \boldsymbol{T}$,其中:

$$\boldsymbol{H} = \begin{bmatrix} G(a_1, b_1, x_1) & \cdots & G(a_L, b_L, x_1) \\ \vdots & \vdots & \vdots \\ G(a_1, b_1, x_N) & \cdots & G(a_L, b_L, x_N) \end{bmatrix}_{N\times L},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1^{\mathrm{T}} \\ \vdots \\ \beta_L^{\mathrm{T}} \end{bmatrix}, \boldsymbol{T} = \begin{bmatrix} t_1^{\mathrm{T}} \\ \vdots \\ t_N^{\mathrm{T}} \end{bmatrix} \quad (2)$$

则超限学习机输出层的权值由最小二乘法得到：

$$\boldsymbol{\beta} = \boldsymbol{H}^{\mathrm{T}}\left(\frac{1}{\gamma} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{T} \quad (3)$$

式中 γ 为正则化参数。

超限学习机在任意多数据样本上的训练能够通过式(3)一步完成。式(3)的求解，主要受 $M \times N$ 方阵 $\left(\frac{1}{\gamma} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)$ 求逆的性能决定，在不采用快速算法的情况下，其算法复杂度上限为 $O(N^3)$，在参数代理优化算法中，N 随着已知点集的扩大逐步增大，直到达到最大求解次数限制。而相对地，传统神经网络基于梯度反向传播进行训练的神经网络，需要使用梯度下降算法进行多步迭代训练模型，无法实现单步计算拟合模型。文献[16]证明超限学习机具备良好的预测性能和泛化能力，为此本文使用超限学习机作为代理模型构建优化算法。

2.2 基于超限学习机建立代理模型

对于全局优化过程中获取到的点集使用算法 2 构建或更新超限学习机代理模型。

算法 2 基于超限学习机的代理模型更新

输入：已求解点集 S，激活函数 G，隐层节点数量 L，参数空间上界 $highBound$，参数空间下界 $lowBound$，模型正则化系数 γ；

输出：超限学习机代理模型 M

(1) $\hat{x} \leftarrow \frac{x - lowBound}{highBound - lowBound}$；/* 对点集 S 中的参数点基于上下界进行最大最小归一化 */；

(2) a, $b \leftarrow$uniform(−1, 1)；/* 通过在区间(−1,1)均匀随机采样产生输入层的权值与偏置 */；

(3) 根据式(3)构建隐层输出矩阵 $\boldsymbol{H}$；

(4) $\boldsymbol{\beta} \leftarrow \boldsymbol{H}^{\mathrm{T}}\left(\frac{1}{\gamma} + \boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)^{-1}\boldsymbol{T}$；/* 基于式(4)得到输出层权重 */；

(5) return M。

得到输入层和输出层的权重后，即可根据式(1)获得对参数空间的各点预测。本文所用超限学习机模型的激活函数为三角激活函数。图 2 为超限学习机对 1 维函数 $y = x\sin(x) + x\cos(2x)$，$x \in [0,5]$ 基于采样点对一维参数空间的拟合，及其与 RBF 插值模型的对比。

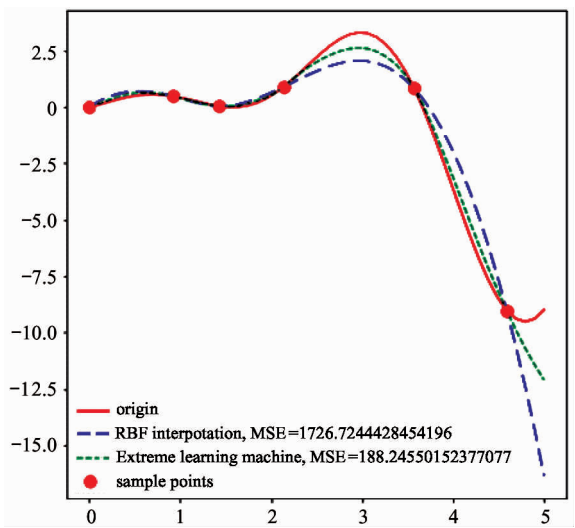

图 2 超限学习机和 RBF 作为代理模型对 1 维函数建模

3 基于随机响应面方法的超参数优化算法

随机响应面方法是文献[11]提出的一种基于随机采样和代理模型的全局优化算法。随机响应面方法的主要步骤如算法 3 所示。

算法 3 随机响应面方法

(1) 在参数空间采样，生成 N 个候选点；

(2) 使用代理模型计算各个候选点的预测函数值 S_i；

(3) 计算各个候选点的距离度量值 $D_i \leftarrow \min(\|x_i - x_{\mathrm{evaled}}\|)$；/* 距离度量为候选点到已知点集中各个点的最小欧式距离 */；

(4) $V_S \leftarrow (S_{\max} - S_i)/(S_{\max} - S_{\min})$；/* 计算各个候选点代理模型预函数值的相对得分，预测值较小的候选点具有更高的得分 */；

(5) $V_D \leftarrow (D_i - D_{\min})/(D_{\max} - D_{\min})$；/* 计算

各个候选点的距离度量相对得分,距离度量较大的候选点有更高的得分 */;

(6) $V \leftarrow w \times V_s + (1-w) \times V_D$, $w \in [0,1]$; /* 计算2个得分的加权平均作为各个候选点的最终得分,较大的 w 倾向于选择代理模型认为更优的点;较小的 w 倾向于探索不确定性较高的未知区域 */;

(7) 选择得分最高的候选点 x^* 作为下一个函数值求解点,得到 $f(x^*)$ 后更新已知点集,重复以上步骤,直至终止条件。

文献[11]基于以上步骤提出了2种算法,即全局随机响应面法和局部随机响应面法。主要区别是步骤(2)产生候选点和步骤(6)选择候选点的策略。为同时兼顾全局探索与局部搜索两者的优势,本文结合2种算法建立一种改进的随机响应面方法,候选点在当前最优点邻域内均匀采样生成,引入局部参数 ρ 来控制均匀采样候选点的分布范围,相对正态采样,保证候选点充分覆盖于 ρ 所限制的区域,同时在参数空间维度较高时,采用部分坐标扰动机制,提高高维情况下优化的效率,具体候选点生成如算法4所示。

算法4 候选点生成算法

输入:当前已获得最优参数坐标 x_{best},局部参数 ρ,最小扰动范围 $r1$,候选点个数 N,参数维度 D,参数空间上界 $highBound$,参数空间下界 $lowBound$;

输出:候选点集 P

(1) if $D \leqslant 8$;

(2) $p_{perturb} = 1$;

(3) else;

(4) $p_{perturb} \leftarrow (1-\rho(1-r1))$; /* $p_{perturb}$ 为坐标轴扰动的概率。当参数空间维度较高,通过 $p_{perturb}$ 能够实现随机固定当前最佳点某些维度坐标,在部分坐标轴扰动生成候选点的效果,当 ρ 为1时,坐标波动概率为最小值 $r1$ */;

(5) end if;

(6) for $i \leftarrow 1$ to N;

(7) $x_j \leftarrow x_{\text{best}}$;

(8) for $j \leftarrow 1$ to D /* 以概率 $p_{perturb}$ 对坐标轴加入随机扰动 */;

(9) $L_j \leftarrow lowBound[j]$, $H_j \leftarrow highBound[j]$; /* L_j 和 H_j 为参数空间在当前维度的上下界 */;

(10) $offset = \begin{cases} -uniform(0,\ x_j - low\_bound_j) & \text{if } rand < 0.5 \\ uniform(0,\ high\_bound_j - x_j) & \text{if } rand > 0.5 \end{cases}$; /* 以当前位置为起点,等概率的向上界或下界偏移,偏移终点在当前位置与边界之间等概率分布;

(11) $offset \leftarrow offset \times (1-\rho(1-r1))$; /* 通过局部参数 ρ 控制扰动偏移的大小,当 $\rho = 0$ 时,等价与在整个参数空间均匀采样产生候选点,当 $\rho = 1$ 时,偏移范围被限制在当前最佳点坐标的邻域,邻域大小由 $r1$ 系数决定 */;

(12) if $rand < p_{perturb}$;

(13) $x_j \leftarrow x_j + offset$; /* 以概率 $p_{perturb}$ 为候选点当前坐标引入偏移 */;

(14) end if;

(15) end for;

(16) $P \leftarrow P \cup x_i$;

(17) end for;

(18) return P。

上述候选点生成算法中,当局部参数 ρ 为0时,等价于在整个参数空间均匀采样,而当 ρ 逐渐接近1,候选点生成的范围逐渐朝当前最优点的邻域收缩。进一步地,可以设置算法3步骤(6)中的 $w = \rho$,则候选点的选择策略也由 ρ 进行控制,当 ρ 较小时,倾向于选择候选点覆盖区域内不确定性较高的点;当 ρ 接近于1时,倾向于选择候选点覆盖区域内代理模型预测值较高的点。候选点坐标轴的扰动概率与 ρ 的关系如式(4)所示,局部坐标扰动机制的引入能够提高响应面方法在高维空间中的性能表现[12]。

$$p_{perturb} = \begin{cases} (1-\rho(1-r1)) & D > 8 \\ 1 & D \leqslant 8 \end{cases} \tag{4}$$

最后,为有效结合全局随机面法和局部随机面法,本文将每一轮的优化划分为2个阶段,探索阶段与全局随机面法类似,局部参数 ρ 由0经过 n_1 个探索步达到 $\rho_{\max}$;之后进入利用阶段,保持 $\rho = \rho_{\max}$,如果经过 n_2 步结果没有得到提高,则此轮结束,设置 $\rho = 0$,重新开始探索阶段。

综上所述,基于超限学习机和随机响应面方法的超参数优化算法 SurroOpt1 如算法 5 所示。

算法 5 基于超限学习机和随机响应面方法的超参数优化算法 SurroOpt1

输入:参数空间维度 D,参数空间上界 $highBound$,参数空间下界 $lowBound$,最大函数值求解次数 $N_{\max}$,探索阶段步数 n_1,利用阶段最大失败步数 n_2,最大利用参数 $\rho_{\max}$;

输出:最优参数 x_{best}

(1) $N_{intial} \leftarrow 2D+2$, $S \leftarrow \varnothing$, $\rho \leftarrow 0$; /* $N_{initial}$ 为初始随机采样点数量,S 为已评估点集;

(2) $n_{fail} \leftarrow 0$, $\rho_{\text{step}} \leftarrow (\rho_{\max} - \rho_0)/n_1$; /* n_{fail} 为利用阶段结果连续未获得提高的步数 */;

(3) for $i \leftarrow 1$ to $N_{initial}$ /* 随机均匀采样获得 $N_{initial}$ 个初始点 */;

(4) $x \leftarrow uniform(lowBound, highBound)$;

(5) 求解函数值 $f(x)$;

(6) $S \leftarrow S \cup (x, f(x))$, $x_{best} \leftarrow argmin_x f(x)$;

(7) end for;

(8) 使用算法 2 基于点集 S 建立 ELM 代理模型 M;

(9) for $i \leftarrow 1$ to $N_{\max}$/* 基于代理模型和随机响应面法的优化;

(10) 使用算法 4,基于当前 x_{best} 和 ρ,生成 $500D$ 个候选点;

(11) 使用代理模型 M 获得各候选点的预测值,并计算各个候选点的评分 V_i; /* 按照算法 3 中的步骤计算,权重 $w = \rho$ */;

(12) 选择评分最优的候选点作为下一个函数值求解点 x,并求解 $f(x)$;

(13) $S \leftarrow S \cup (x, f(x))$;

(14) 使用算法 2 基于点集 S 更新 ELM 代理模型 M;

(15) if $\rho < \rho_{\max}$/* 处于探索阶段,每一步控制 ρ 递增 ρ_{step} */;

(16) $\rho \leftarrow \rho + \rho_{step}$;

(17) end if;

(18) if $f(x) < f(x_{best})$;

(19) $x_{best} \leftarrow x$, $n_{fail} \leftarrow 0$;

(20) if $\rho \geqslant \rho_{\max}$ and $f(x) \geqslant f(best)$/* 处于利用阶段 */;

(21) $n_{fail} \leftarrow n_{fail} + 1$;

(22) end if;

(23) if $n_{fail} \geqslant n_2$;/* 如 n_{fail} 达到 n_2,则此轮结束,重新进入探索阶段 */;

(24) $\rho \leftarrow 0$;

(25) end if;

(26) end for;

(27) return x_{best}。

SurroOpt1 算法通过局部参数 ρ 在 2 个不同阶段的控制策略,结合了全局随机响应面法和局部随机响应面法,可以认为 SurroOpt1 算法是全局响应面方法和局部响应面算法的混合算法,周期内按顺序调用且在每个周期的起始实现重启机制,在充分利用最优点进行局部搜索的同时可以有效避免陷入局部最优。

算法能够保证理论上的收敛,证明如下。

假设 $f(x)$ 存在最优点 x_{best},则当 $N_{\max} \to \infty$,参数空间中的每一个点都以概率 $p = 1$ 被采样到(步骤(24)等价于在整个参数空间均匀采样),即 x_{best} 以概率 1 被采样,算法以概率 1 收敛。

算法每一步的计算复杂度主要由 3 部分决定:

(1) 基于已知点集更新代理模型的矩阵求逆运算,如前所述,算法复杂度上限为 $O(N^3)$,N 为已知点集的个数;

(2) 基于代理模型,求 $500D$ 个候选点的代理模型预测值,主要包括式(2)的矩阵乘法运算和式(3)的矩阵-向量乘法运算,两者的算法复杂度分别为 $O(D^2L)$ 与 $O(DL)$,则求候选点代理模型预测值的算法复杂度为 $O(D^2L)$;

(3) 求候选点与已知点集间的距离,其算法复杂度为 $O(D^2N)$。

在算法初始阶段已知点集数量 N 较小时,第 2 部分的运算是主要的计算代价,而随着已知点集中点数的增多,第 1 部分和第 3 部分将成为主要的运算代价。

4 实验与结果

在本节中,使用本文提出的超参数优化算法,在3个典型卷积神经网络超参数优化问题上测试提出的算法。

4.1 算法相关参数设置

本文提出的超参数优化算法 SurroOpt1,本次实验中,设置的代理模型相关参数和算法参数取值如表1所示。

表1 算法相关参数取值

参数名称	参数取值
ELM 隐层单元数量 h_{num}	2 000
ELM 正则化系数 γ	2^{20}
ELM 激活函数	三角激活函数
最小扰动范围 v_1	0.12
最大控制参数 $\rho_{\max}$	0.9
探索步数 n_1	$\min(16, 2D)$
利用步数 n_2	$\min(8, D)$
最大评价函数求解次数 $N_{\max}$	200

4.2 卷积神经网络模型超参数优化实验与基准算法

本文设计3个不同的卷积神经网络超参数优化问题实验,验证本文算法在深度学习模型超参数调优问题中的有效性。卷积神经网络是当前最常用的深度网络结构之一,也是大量复杂网络结构的基础。在数据集的选择上,3个实验分别使用 MNIST 手写数据集和 Cifar10 数据集训练卷积神经网络,这2个数据集被广泛用于测试深度网络的基本性能,并在大量超参数调优算法相关研究中被作为测试超参数调优算法的基准训练数据集。

BOA 算法与 TPE 算法是当前最典型的2种超参数调优算法,在深度学习自动化参数调优领域收到广泛关注与研究,因此本文将算法的实验效果与最大提升期望 BOA 算法和 TPE 算法进行对比,分别使用2个基准算法和本文算法解决3个超参数调优问题,实验5次,比较3个算法的性能和效率表现。

在3个实验中,需要训练结构相似的 CNN 模型。模型使用2个卷积层,每个卷积层的输出使用批归一化(batch normalization)和最大化池化(max pooling)处理,在2个卷积层后使用2层全连接网络输出分类结果。前2个问题使用 MNIST 数据集训练网络,训练10轮;第3个问题使用 Cifar10 数据集训练网络,训练100轮。第3个问题在各层输出使用 Dropout 权值采样,提高泛化能力。这里采用随机梯度下降法(SGD)对网络进行训练,3个问题分别需要设置8、15、19个超参数。具体待优化超参数如表2所示,评价超参数的性能指标是训练结束后在模型验证集上的分类错误率。

表2 实验 CNN 模型超参数列表

	CNN8	CNN15	CNN19
超参数1	学习率	学习率	学习率
超参数2	动量	动量	动量
超参数3	全局权值正则化系数	全局权值正则化系数	全局权值正则化系数
超参数4	学习率衰减率	学习率衰减率	学习率衰减率
超参数5	全连接层1节点数	全连接层1LeakyRelu 函数斜率	全连接层1LeakyRelu 函数斜率
超参数6	全连接层2节点数	全连接层2LeakyRelu 函数斜率	全连接层2LeakyRelu 函数斜率
超参数7	卷积层1滤波器数	全连接层1正态分布初始化方差	全连接层1正态分布初始化方差
超参数8	卷积层2滤波器数	全连接层2正态分布初始化方差	全连接层2正态分布初始化方差
超参数9		卷积层1正态分布初始化方差	卷积层1正态分布初始化方差
超参数10		卷积层2正态分布初始化方差	卷积层2正态分布初始化方差
超参数11		批处理大小	全连接层1dropout 概率
超参数12		全连接层1节点数	全连接层2 dropout 概率

（表 2 续）

超参数 13	全连接层 2 节点数	卷积层 1 dropout 概率
超参数 14	卷积层 1 滤波器数	卷积层 2dropout 概率
超参数 15	卷积层 2 滤波器数	批处理样本数量
超参数 16		全连接层 1 节点数
超参数 17		全连接层 2 节点数
超参数 18		卷积层 1 滤波器数
超参数 19		卷积层 2 滤波器数

4.3　实验结果

对于实验 1,8 超参数 CNN 网络优化问题,3 个算法最终得到的性能对比如表 3 所示。

表 3　8 超参数问题性能对比

	最优(%)	最差(%)	平均(%)
本文算法	0.78	0.82	0.8
最大提升 BOA 算法	0.84	0.91	0.87
TPE 算法	0.79	0.86	0.82

3 个算法对实验 1 问题的优化收敛曲线如图 3 所示。

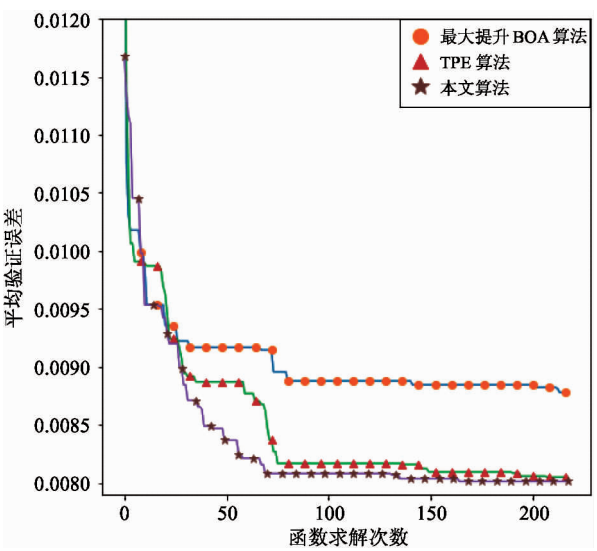

图 3　3 种算法在实验 1 上的效率比较

对于实验 2,15 超参数 CNN 网络优化问题,3 个算法得到的性能对比如表 4 所示。

3 个算法对实验 2 问题的优化收敛曲线如图 4 所示。

表 4　15 超参数问题性能对比

	最优(%)	最差(%)	平均(%)
本文算法	0.88	1.07	0.98
最大提升 BOA 算法	1.1	1.13	1.12
TPE 算法	1.01	1.04	1.02

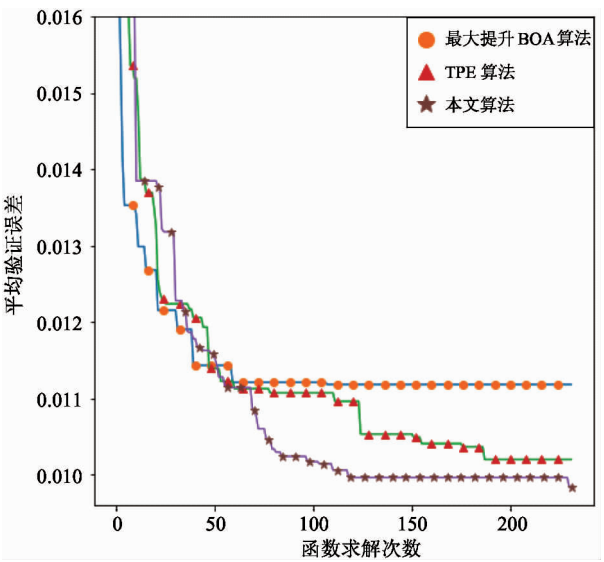

图 4　3 种算法在实验 2 上的性能比较

对于实验 3,19 超参数 CNN 超参数优化问题,3 个算法得到的性能对比如表 5 所示。

表 5　19 超参数问题性能对比

	最优(%)	最差(%)	平均(%)
本文算法	21.95	25.19	23.49
最大提升 BOA 算法	27.05	28.77	27.90
TPE 算法	24.43	27.05	25.56

3 个算法对实验 2 问题的优化收敛曲线如图 5 示。

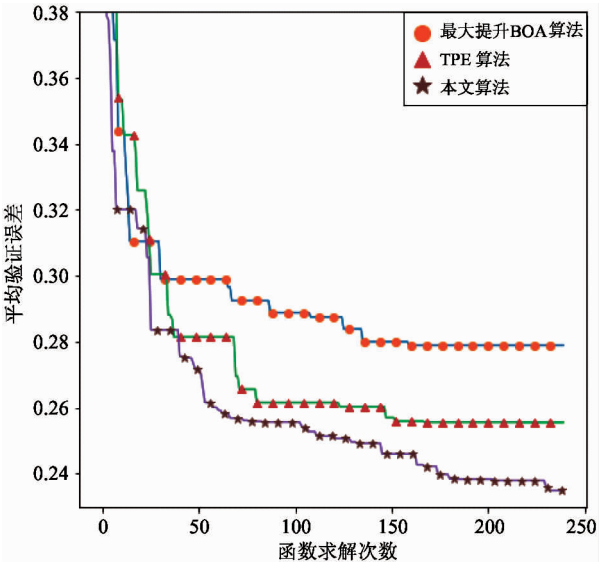

图 5　3 种算法在实验 3 的性能比较

以上结果显示,本文提出的算法 SurroOpt1,相对于基准算法,在 3 个卷积神经网络超参数优化实验中,均表现出更好的性能与优化效率,能够以更少的搜索次数,在较短时间内获得到相对更优的超参数设置。特别是实验 3,说明 SurroOpt1 算法在超参数数量较多、参数空间维度较高时,具备更显著的优势。

5　结　论

本文基于随机响应面法,使用超限学习机作为代理模型,提出了一种新的用于深度学习模型超参数优化的全局优化算法 SurroOpt1。算法利用了超限学习机训练快速、泛化性能好的优势,并进一步改进了随机响应面法的优化策略。本文提出的算法在性能和效率方面优于已有的经典超参数优化算法,在超参数优化场景具备良好的应用前景。未来将继续进一步研究 SurroOpt1 算法的并行实现,提高解决大规模超参数优化问题的效率。

参考文献

[1] Bergstra J, Bardent R, Bengio Y, et al. Algorithms for hyper-parameter optimization [C] // Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 2011: 2546-2554

[2] Snoek J, Larochelle H, Adams R P. Practical bayesian optimization of machine learning algorithms [C] // Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, 2012: 2951-2959

[3] 王凯, 王珏翎, 刘文革. 滚子包络端面啮合蜗杆传动参数优化[J]. 高技术通讯, 2018, 28(7): 651-656

[4] Golovin D, Solnik B, Moitra S. Google vizier: a service for black-box optimization [C] // Proceedings of the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Canada, 2017: 1487-1495

[5] Hutter F, Hoos H H, Leyton-Brown K. Sequential model-based optimization for general algorithm configuration [C] // Proceedings of the International Conference on Learning and Intelligent Optimization, Berlin, Germany, 2011: 507-523

[6] Klein A, Falkner S, Bartels S. Fast Bayesian optimization of machine learning hyperparameters on large datasets [C] // Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Florida, USA, 2017: 528-536

[7] Zhao M, Li J. Tuning the hyper-parameters of CMA-ES with tree-structured Parzen estimators [C] // Proceedings of the Advanced Computational Intelligence, Xiamen, China, 2018: 613-618

[8] Mengistu T, Ghaly W. Aerodynamic optimization of turbomachinery blades using evolutionary methods and ANN-based surrogate models [J]. *Optimization and Engineering*, 2008, 9(3): 239-255

[9] Poloczek J, Kramer O. Local SVM constraint surrogate models for self-adaptive evolution strategies [C] // Annual Conference on Artificial Intelligence, Berlin, Heidelberg, 2013: 164-175

[10] Peter T. Using deep learning as a surrogate model in multi-objective evolutionary algorithms [R]. Magdeburg: Otto von Guericke University, 2018

[11] Regis R G, Shoemaker C A. A stochastic radial basis function method for the global optimization of expensive functions [J]. *Informs Journal on Computing*, 2007, 19(4): 497-509

[12] Regis R G, Shoemaker C A. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization [J]. *Engi-*

neering Optimization, 2013, 45(5): 529-555

[13] Müller J, Shoemaker C A. Influence of ensemble surrogate models and sampling strategy on the solution quality of algorithms for computationally expensive black-box global optimization problems[J]. *Journal of Global Optimization*, 2014, 60(2): 123-144

[14] Ilievski I, Akhtar T, Feng J. Efficient hyperparameter optimization for deep learning algorithms using deterministic RBF surrogates[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, USA, 2017: 822-829

[15] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: theory and applications[J]. *Neurocomputing*, 2006, 70(1-3): 489-501

[16] Huang G B, Zhou H, Ding X. Extreme learning machine for regression and multiclass classification[J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2012, 42(2): 513-529

[17] 朱敏, 许爱强, 陈强强. 一种基于改进 KELM 的在线状态预测方法[J]. 北京航空航天大学学报, 2019, 45(7): 1370-1379

A hyperparameter tuning algorithm based on extreme learning machine and stochastic response surface method for deep learning

Sun Yongze[*,**], Lu Zhonghua[*]

([*] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190)

([**] University of Chinese Academy of Sciences, Beijing 100049)

Abstract

Appropriate setting of hyperparameters is a critical factor that determines the performance of a deep learning model. Realization of highly efficient hyperparameter tuning algorithm contributes to improvement of the speed and efficiency of deep learning application, and reducing the difficulty of applying deep learning model. One of the state-of-art hyperparameter tuning algorithms is Bayesian optimization algorithm (BOA) based on surrogating model. Theoretically the performance and efficiency of algorithm based on surrogating model can be superior to several simple hyperparameter tuning algorithms such as grid search and random search. This article presents a high performance hyperparameter tuning algorithm SurroOpt1 that adopts extreme learning machine (ELM) as deterministic surrogating model and an improved random respond surface method as optimization strategy. Experiments prove that the proposed algorithm can achieve superior performance and efficiency in hyperparameter tuning task for deep convolutional neural nets to Bayesian optimization algorithm and tree-structured Parzen estimator (TPE) algorithm, which are two of the state-of-art algorithms.

Key words: hyperparameter tuning, surrogating model, extreme learning machine (ELM), random respond surface, deep learning model