计算机科学
COMPUTER SCIENCE

面向超参数估计的贝叶斯优化方法综述

李亚茹，张宇来，王佳晨

引用本文

李亚茹，张宇来，王佳晨. 面向超参数估计的贝叶斯优化方法综述[J]. 计算机科学, 2022, 49(6A): 86-92.

LI Ya-ru, ZHANG Yu-lai, WANG Jia-chen. Survey on Bayesian Optimization Methods for Hyper-parameter Tuning[J]. Computer Science, 2022, 49(6A): 86-92.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

多示例学习算法综述

Review of Multi-instance Learning Algorithms

计算机科学, 2022, 49(6A): 93-99. https://doi.org/10.11896/jsjkx.210500047

基于多源数据和逻辑推理的行为识别技术研究

Study on Activity Recognition Based on Multi-source Data and Logical Reasoning

计算机科学, 2022, 49(6A): 397-406. https://doi.org/10.11896/jsjkx.210300270

一种基于异质模型融合的 Android 终端恶意软件检测方法

Android Malware Detection Method Based on Heterogeneous Model Fusion

计算机科学, 2022, 49(6A): 508-515. https://doi.org/10.11896/jsjkx.210700103

基于 Stacking 多模型融合的 IGBT 器件寿命的机器学习预测算法研究

Study on Machine Learning Algorithms for Life Prediction of IGBT Devices Based on Stacking Multi-model Fusion

计算机科学, 2022, 49(6A): 784-789. https://doi.org/10.11896/jsjkx.210400030

机器学习在金融资产定价中的应用研究综述

Application of Machine Learning in Financial Asset Pricing:A Review

计算机科学, 2022, 49(6): 276-286. https://doi.org/10.11896/jsjkx.210900127

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300208

面向超参数估计的贝叶斯优化方法综述

李亚茹　张宇来　王佳晨
浙江科技学院信息与电子工程学院　杭州 310023
(yrlizust@foxmail.com)

摘　要　对绝大部分机器学习模型而言,超参数选择对模型的最终效果起到了至关重要的作用,所以超参数的选择与估计是机器学习理论与实践中的重要问题。从超参数空间中的点到模型泛化性能的映射可以看作一个具有高评估代价的复杂黑箱函数,一般的最优化方法难以适用。贝叶斯优化是一种非常有效的全局优化算法,适合求解具有解析式不明确、非凸、评估成本高等特点的优化问题,只需较少的目标函数评估就可以获得理想解。总结了贝叶斯优化在超参数估计问题上的基本理论和方法,综述了近年来该方向的研究热点和最新进展,包括代理模型、采集函数、算法实施等方面的研究,总结了现有的研究中尚待解决的问题,期望帮助初学者快速了解贝叶斯优化算法并理解典型的算法思想,为其之后的研究起到一定的指导作用。

关键词:超参数;贝叶斯优化;黑箱优化;概率代理模型;机器学习
中图法分类号　TP181;O212

Survey on Bayesian Optimization Methods for Hyper-parameter Tuning

LI Ya-ru,ZHANG Yu-lai and WANG Jia-chen
School of Information and Electronic Engineering,Zhejiang University of Science and Technology,Hangzhou 310023,China

Abstract　For most machine learning models,hyper-parameter selection plays an important role in obtaining high quality models. In the current practice,most of the hyper-parameters are given manually. So the selection or estimation of hyper-parameters is an key issue in machine learning. The mapping from hyper-parameter set to the model's generalization can be regarded as a complex black box function. The general optimization method is difficult to apply. Bayesian optimization is a very effective global optimization algorithm,which is suitable for solving optimization problems in which their objective functions could not be expressed,or the functions are non-convex,computational expensive. The ideal solution can be obtained with a few function evaluations. This paper summarizes the basics of the Bayesian optimization based on hyper-parameter estimation methods,and summarizes the research hot spots and the latest developments in the recent years,including the researches in agent model,acquisition function,algorithm implementation and so on. And the problems to be solved in existing research are summarized. It is expected to help beginners quickly understand Bayesian optimization algorithms,understand typical algorithm ideas,and play a guiding role in future researches.

Keywords　Hyper-parameters,Bayesian optimization,Black box optimization,Probabilistic surrogate model,Machine learning

1　引言

超参数(Hyper-parameter)的选择与估计一直是机器学习模型在实际应用中的关键问题,模型泛化性能的优劣依赖于对超参数的合理选择。当前,包括深度神经网络在内的机器学习模型在图像识别、自然语言处理等领域取得了举世瞩目的成就,但在实践中这些模型的超参数很大程度上仍然依靠人工经验或者反复试验来确定[1]。随着模型的日趋复杂,其超参数的数量也在增长,人工选择合适的超参数变得更为困难,同时也缺乏有效的理论保证与指导。

如果把机器学习模型的评价看作一个从超参数到泛化性能的映射,则超参数的估计问题可以看作对这个具有高评估代价的黑箱函数的优化问题。该映射函数往往具有非凸、多模态、含噪等复杂特征[2]。局部优化方法需要优化的目标有梯度,很容易陷入局部最优,不能很好地处理噪声;更重要的是,当前以深度学习为代表的机器学习模型训练时间非常长,获得一组超参数下的模型泛化误差代价较高,局部优化方法缺乏样本效率。而贝叶斯优化(Bayesian Optimization)正是这样一种能够利用有限的函数采样值,在较少的评估次数下获得复杂目标函数最优值的方法[3]。近年来,贝叶斯优化由于其较高的样本效率和较完善的理论性能保证,成为了超参数估计的一种重要方法[4]。

本文将主要综述面向超参数估计的贝叶斯优化方法及其研究现状。第2节简述了基本的超参数优化方法;第3节介绍了面向超参数估计的贝叶斯优化算法;第4节对近年来该领域的研究发展进行了综述;最后讨论了当前该领域面临的

基金项目:国家自然科学基金青年科学基金(61803337)
This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China(61803337).
通信作者:张宇来(zhangyulai@zust.edu.cn)

尚待解决的研究问题。

2 超参数估计的基本方法

超参数通常指模型本身或相应的训练算法中的配置变量，与参数不同，它不能通过对训练数据进行学习来得到。超参数广泛存在于各种机器学习模型中，如深度神经网络的学习率和批量大小，支持向量机中的惩罚系数，以及 k 最近邻算法中的 k 值等等，这些定义模型属性或者训练算法过程的参数都是模型的超参数。超参数的选择对模型的最终预测效果有很大的影响，如神经网络可能因学习率过大而收敛效果差，过小时收敛速度又过慢。

实践当中我们常常参考其他问题的超参数取值，然后通过反复尝试人工寻找最佳值。但这样做费时费力，并且缺乏理论指导与保证。随着计算机算力的增长，人们希望能够自动化地获得最优的模型超参数选择。现有的超参数估计方法主要包括了网格搜索、随机搜索以及本文将要介绍的贝叶斯优化方法。

2.1 网格搜索

网格搜索是应用较广泛的一种朴素的超参数估计方法[5]，通过查找超参数空间中网格上的所有点来确定最优值。网格搜索可以看作一种近似方法，通过给出较大的搜索范围以及较小的搜索步长，以较高的精度逼近全局最大值或最小值。但是，网格搜索显然十分消耗计算资源，特别是网格设置较密且需要调优的超参数比较多的时候。因此，实际使用网格搜索时，会先使用较广的搜索范围和较大的步长找到全局最优的可能位置，再缩小搜索范围以及步长，从而找到精确的最值，这样可以有效缩短搜索时间。但是，由于在超参数估计中，目标函数一般是非凸的，在使用较广的搜索范围和较大的步长时，常常会因为找到一个局部最优而错过全局最优。

2.2 随机搜索

随机搜索也是一种朴素的超参数估计方法，它通过在搜索范围中随机取样本点来寻找全局最优点。随机搜索有几个比较重要的优点：1）算法结构简单，易于在计算机上并行实现；2）不需要考虑目标函数的非凸性等即可以正常工作，理论上只要随机样本点集足够多，就可以找到全局的最大或最小值或它们的近似值；3）可以方便地根据不同的启发式思想进行修改，提高搜索效率。实践中，随机搜索一般会快于网格搜索[6]，然而应该在超参数空间的什么区域以何种方式引入随机方法，很多时候还缺乏有效的理论保证。

2.3 贝叶斯优化

贝叶斯优化（Bayesian Optimization）首先由美国伊利诺伊大学厄巴纳-香槟分校（UIUC）的 Pelikan 等学者于 1998 年提出[7]，它在已知有限样本点的情况下，通过构造黑箱函数输出的后验概率来寻找函数的最优值[8-9]。目前贝叶斯优化已经成为超参数估计领域的重要方法[10]。与网格搜索和随机搜索不同，贝叶斯优化的算法框架是序贯的，即当前的最优值搜索是在之前搜索结果的基础上，充分利用已知数据点的信息来进行的[11]，而其他搜索方法大多忽略了这个信息。所以贝叶斯优化方法中主要用到了两个核心部件：1）概率代理模型（Probabilistic Surrogate Model），用于近似表示当前的黑箱目标函数；2）采集函数（Acquisition Function），用于估计在当前已知数据条件下，最优点最有可能出现的位置。同时为了避免陷入局部极值，贝叶斯优化算法还通常会加入一定的随机性，在随机探索和根据后验分布取值之间做出权衡。贝叶斯优化是当前为数不多的，具有较好的收敛性理论保证的超参数估计方法[12]。本文的第 3 部分将从概率代理模型、采集函数、算法框架 3 个方面介绍贝叶斯优化方法的基本内容。

3 超参数优化中的贝叶斯优化方法

记 $f(\boldsymbol{x})$ 为从超参数向量 $\boldsymbol{x}$ 到模型泛化性能的映射，其中 $\boldsymbol{x}\in X$，$X\subseteq R^d$，X 为 d 维的超参数空间。超参数优化的目标是在超参数空间内寻找使得模型泛化性能最优的 d 维超参数 x^*。本文以寻找最大值为例：

$$x^* = \arg\max f(\boldsymbol{x}) \tag{1}$$

由于 $f(\boldsymbol{x})$ 表达了泛化精度等模型泛化指标关于模型超参数的度量，而在当前根据一组超参数进行一次模型训练和评估需要较大的计算量和较长的时间，所以 $f(\boldsymbol{x})$ 是一个具有高评估代价的黑箱目标函数。假设已知数据为 $D_{1,t}=(x_i, y_i)$，$i=1,2,\cdots,t$，其中 y_i 是在超参数 x_i 下训练得到的模型的测试集精度，下文中将当前已有的评估数据简记为 D。我们希望能够在已知较少数据，即较小的 t 的情况下，估算出超参数的最优值。因此，需要首先由 3.1 节中介绍的概率代理模型来拟合黑箱函数，而后由 3.2 节中介绍的采集函数来根据已知数据递推估算最优值，整个计算框架将在 3.3 节中介绍。

3.1 概率代理模型

由于原始待评估的目标黑箱函数评估代价高昂且复杂，所以用概率代理模型近似表示当前的目标函数。如果将得到的测试集精度数据 y 看作泛化精度的一次随机观测实现，则有 $y=f(\boldsymbol{x})+\varepsilon$，其中噪声 ε 满足 $p(\varepsilon)=N(0,\sigma_\varepsilon^2)$ 且独立同分布。我们希望概率代理模型可以从最初的假设先验出发，通过不断增加数据信息来完善模型。而高斯过程（Gaussian Processes，GP）由于可递推进行建模过程，成为了代理模型的重要选择。高斯过程模型是非参数模型[13]，在形式上它的每个有限子集都服从多元正态分布。不失一般性，假设模型输出期望为 0，则已知数据 D 和新的观测点 (x_{t+1}, y_{t+1}) 的联合分布可以表达为：

$$[y_{1:t+1}]\sim N\left(0,\begin{bmatrix}\boldsymbol{K}+\sigma_\varepsilon^2\boldsymbol{I} & \boldsymbol{k}\\ \boldsymbol{k}^{\mathrm{T}} & k(x_{t+1},x_{t+1})\end{bmatrix}\right) \tag{2}$$

其中，$k:\boldsymbol{x}*\boldsymbol{x}$ 为协方差函数，$\boldsymbol{k}=[k(x_1,x_{t+1}),\cdots,k(x_t,x_{t+1})]^{\mathrm{T}}$，Gram 矩阵 $\boldsymbol{K}$ 为：

$$\boldsymbol{K}=\begin{bmatrix}k(x_1,x_1) & \cdots & k(x_1,x_t)\\ \vdots & \ddots & \vdots\\ k(x_t,x_1) & \cdots & k(x_t,x_t)\end{bmatrix} \tag{3}$$

其中，$\boldsymbol{I}$ 是单位阵，σ_ε^2 是噪声方差，此时可以通过考虑原来的观测数据以及新的 $\boldsymbol{x}$ 来做出预测。由于 y_{t+1} 的后验分布为：

$$p(y_{t+1}\mid y_{1:t},x_{1:t+1})=N(\mu_t(x_{t+1}),\sigma_t^2(x_{t+1})) \tag{4}$$

可以得到 y_{t+1} 的数学期望和方差为：

$$\mu_t(x_{t+1})=\boldsymbol{k}^{\mathrm{T}}(\boldsymbol{K}+\sigma_\varepsilon^2\boldsymbol{I})^{-1}y_{1:t} \tag{5}$$

$$\sigma_{t+1}=k(x_{t+1},x_{t+1})-\boldsymbol{k}^{\mathrm{T}}(\boldsymbol{K}+\sigma_\varepsilon^2\boldsymbol{I})^{-1}\boldsymbol{k} \tag{6}$$

高斯过程表达函数丰富分布的能力仅依赖于协方差函数，自动相关性确定（Automatic Relevance Determination，

ARD)平方指数函数常常是高斯回归的默认选择：

$$K_{SE}(x,x')=\theta_0\exp\left\{-\frac{1}{2}r^2(x,x')\right\} \tag{7}$$

$$r^2(x,x')=\sum_{d=1}^{D}(x_d,x_d')^2/\theta_d^2 \tag{8}$$

ARD 平方指数核中，$f(x)$与 $f(x')$的协方差仅依赖于 x 与 x'的欧几里得距离。Matern-52 也是常用协方差函数之一：

$$K_{M52}(x,x')=$$
$$\theta_0\left(1+\sqrt{5r^2(x,x')}+\frac{5}{3}r^2(x,x')\right)\exp\{-\sqrt{5r^2(x,x')}\} \tag{9}$$

从上面可以看出，高斯过程根据核函数和一些观测值，可以精确地计算出闭合形式的后验分布；另外，随机森林模型和贝塔-伯努利模型[9]也在一些应用中作为概率代理模型来使用。

3.2 采集函数

直接从代理模型中选择一个最优点 x_{t+1}进行评估，通常代价高昂。在贝叶斯优化的每次迭代循环中，通常使用采集函数 α 来获得下一个用于评估的样本点。采集函数由已经观测到的数据的后验分布构成，通过对其最大化选择下一个需要评估的样本点，即：

$$x_{t+1}=\arg\max\alpha_t(x;D) \tag{10}$$

常见的采集函数有概率提升函数 PI、期望提升函数 EI、置信上界函数 UCB 3 种。

(1)PI(Probability of Improvement)采集函数的意义是极大化新样本相对于当前目标函数最大值有提升的概率。记 x^+为当前最优值，则更优的超参数可以通过极大化下列采集函数得到：

$$\alpha_{PI}(x)=P(f(x)\geqslant f(x^+)-e)=\Phi\left(\frac{f(x^+)-e-\mu(x)}{\sigma(x)}\right) \tag{11}$$

这里假设超参数在 x^+附近服从正态分布，故 $\Phi(\cdot)$ 表示的是正态累积分布函数；e 是用来平衡探索和利用的随机参数，e 较大时，PI 倾向于探索未知信息，e 较小时，PI 倾向于利用已知信息。PI 考虑了未知点比已知最大值提升的概率，但是没有考虑未知点比已知最大值提升的量。

(2)EI(Expected Improvement)采集函数也可以看作一种提升策略，它考虑了未知点比已知最大值提升的量，其采集函数为：

$$\alpha_{EI}(x)=\begin{cases}(\mu(x)-f(x^+))\Phi(z)+\sigma(x)\Phi(z), & \sigma(x)>0\\ 0, & \sigma(x)=0\end{cases} \tag{12}$$

与上面一样，$f(x^+)$表示现有的最大值。EI 函数求的是未知点函数值比 $f(x^+)$大的期望。$z=\dfrac{\mu(x)-f(x^+)}{\sigma(x)}$，是对当前最优点取值的归一化。

(3)UCB(Upper Confidence Bound)采集函数考虑的是尽可能地提高代理模型上置信边界的值，它是 Srinivas 等于 2010 年提出的一种针对高斯过程置信边界的策略：GP-UCB[14]。其采集函数为：

$$\alpha_{UCB}(x)=\mu(x)+\gamma\sigma(x) \tag{13}$$

其中，γ 为调节参数，同时 UCB 采集函数通过控制这个调节参数控制对探索未知和已知数据信息之间的平衡。UCB 在许多应用上取得了较好的效果。

3.3 贝叶斯优化算法框架

综上所述，面向超参数估计的贝叶斯优化算法框架如算法 1 所示，由于在实际中时间以及计算资源是有限制的，所以对代价昂贵的评估设置次数上限 T。

算法 1 BO

输入：代理模型 f，采集函数 α

输出：超参数向量 x^*

1. 初始化超参数向量 x_0
2. for t=1,2,…,T do
3. 最大化采集函数，得到下一个评估点：$x_{t+1}=\arg\max_{x\in X}\alpha(x|D)$；
4. 评估目标函数值 $y_{t+1}=f(x_{t+1})+\varepsilon_{t+1}$；
5. 整合数据：$D_{t+1}=D\cup(x_{t+1},y_{t+1})$，并且更新概率代理模型；
6. end for

4 研究进展

近年来贝叶斯优化被大量用于机器学习模型的超参数估计和模型的自动选择中[15-17]，从而带动了面向超参数估计的贝叶斯优化方法在理论研究、算法设计等方面的研究，并取得了一定的成果。本节首先综述了近年来在贝叶斯优化方法的两大组成部分即概率代理模型和采集函数的研究进展；随后介绍了在停止条件、最优初始化选择和整数约束优化等几个实际应用问题上的研究进展。

4.1 代理模型研究进展

4.1.1 具有复合函数结构的代理模型

2019 年，Astudillo 等提出部分超参数估计问题的目标函数具有复合函数 $f(x)=g(h(x))$的形式[18]，并在该假设下对贝叶斯优化算法进行了改进。复合函数中 h 是具有高评估代价的黑箱函数，且输出为向量；g 是已知解析式的简单函数，输出为标量。该方法在 h 上建立一个多输出高斯过程模型，所以当 g 是非线性函数时，整个复合函数的概率代理模型则呈现出非高斯的性质，这样就扩展了原有的基于高斯过程的代理模型的表达能力。同时对整个代理模型进行拆解，其中 $h(x)$函数的输出值提供了原先直接从 $f(x)$的输出中无法得到的信息。所以具有复合函数形式的代理模型的贝叶斯优化方法在符合该假设的应用中得到了显著优于标准方法的结果。

4.1.2 具有线性可加结构的代理模型

标准贝叶斯优化方法大多在低维的输入空间中运行良好，然而在诸如计算机视觉等应用领域，所使用模型的超参数数量庞大，通常需要在高维输入空间上优化目标函数 f，此时标准方法会呈现出相当高的复杂度。Kandasamy 等于 2015 年提出了将高维代理模型函数分解为多个不相交的变量子集上的低维函数之和的方法[19]，即：

$$f(x)=f_1(x^{(1)})+f_2(x^{(2)})+\cdots+f_d(x^{(M)}) \tag{14}$$

其中，M 为超参数的分组数，$M<d$。然而很多时候对代理模型直接进行分解是困难的。Li 等于 2016 年扩展了上述方法[20]，假设超参数分组在某种待定投影形式下可加。2018 年 Rolland 等再次推广了 Kandasamy 等的方法[21]，改变了超参数子集不相交的假设，考虑子集交集可为非空的加性模型，并通过图表示变量之间依赖关系，从而推导出一种有效的消息传递算法来优化采集函数。值得一提的是，上述方法都通过最大化高斯过程模型的边际似然函数来得到近似最优的代理模型分解，但在全特征空间进行这种最大化在计算上是困难的。Wang 等于 2017 年给出了一种新的高维贝叶斯优化

方法 HDBO[22]，该方法假设目标函数存在一个潜在可加结构，并且提出通过 Gibbs 抽样学习潜在结构，并行执行多个评估，从而减少优化过程所需的迭代次数。

另外，许多实际应用需要对多个模型的泛化性能指标进行优化，此类多任务问题的代理模型具有天然的可加性。Swersky 等将多任务高斯过程[23]扩展到贝叶斯优化框架，实现了多任务贝叶斯优化[24]，同时将已有的优化任务中获得的信息运用到新任务中，可以更有效地找到全局最优值，明显加快了优化进程。

4.1.3 具有子空间结构的代理模型

具有可加结构的代理模型虽然可以有效降低在具有高维超参数向量时的算法的复杂度，但对代理模型的限制过严。在更一般的情形下，可以先在一个低维空间计算得到超参数向量的低维投影，再在此基础上计算整个超参数向量。从这个角度看，前面的线性可加情形可以看作一种特殊的子空间结构。Djolonga 等认为超参数空间中存在对模型泛化性能影响较大的低维子空间，可以通过矩阵分解学习得到该子空间[25]。Munteanu 等也于 2019 年提出了一种基于低维子空间嵌入的贝叶斯优化方法 HeSBO[26]。该方法通过特征变换得到一个超参数空间的低维嵌入，并证明了在低维嵌入空间中得到的超参数最优值估计的误差是严格有界的。Kirschner 等于 2019 年提出了 LINEBO 算法[27]，通过将全局问题分解为一系列可以有效求解的一维子问题来解决高维超参数问题。

4.2 采集函数研究进展

4.2.1 基于信息熵搜索的采集函数

Ho 等于 2012 年提出了基于信息熵来构造采集函数的基本想法[28]；2014 年 Hoffman 提出了基于信息熵搜索(Entropy Search)的采集函数[29]，这种采集函数的基本想法是新选择的点能够使得关于 x 的分布函数的信息熵增益极大。ES 的采集函数为：

$$\alpha_{\mathrm{ES}}(x;D)=H(P(x_*\mid D))-E_{P(y\mid x,D)}[H(P(x_*\mid D\cup\{x,y\}))] \tag{15}$$

其中，$H(P(x))=-\int p(x)\log p(x)\mathrm{d}x$ 为 $p(x)$ 的熵。由于 $P(x_*\mid D)$ 本身比较复杂，熵的计算会很困难，所以 ES 在计算采集函数时，需要采用近似技术。

Wang 等于 2017 年提出了一种新的基于互信息的采集函数[22]：最大值熵搜索(Max-value Entropy Search，MES)。不同于之前的熵搜索策略使用有关 x^* 的信息，该策略使用有关最大估计值 $y^*=f(x^*)$ 的信息。采集函数是最大估计值 y^* 和下一个可能最优点之间的互信息增益，可以通过评估预测分布的熵来近似：

$$\alpha_{\mathrm{MES}}(x)=H(P(y\mid D,x))-E[H(P(y\mid D,x,y'))] \tag{16}$$

其中，$y=f(x)$，$P(y\mid x,D)=\mathrm{N}(\mu_t(x),\sigma_t^2(x)+\sigma^2)$，与以前关于 x 的分布不同，$P(y\mid D)$ 是一维的，因此计算变得容易了很多。Hernández-Lobato 等 2018 年提出 PESMO 方法[30]。PESMO 的采集函数为：

$$\alpha(x)=H(x^*\mid D)-E_y[H(x^*\mid D\cup\{(x,y)\})] \tag{17}$$

其中，x^* 是根据 Pareto 原则得到的一个超参数的有效子集。由于互信息是对称的，因此作者也给出了这个采集函数的基于互信息的等价表达式：

$$\alpha(x)=H(y\mid D,x)-E_{\chi^*}[H(y\mid D,x,\chi^*)] \tag{18}$$

在多优化任务中，采集函数不仅需要选择下一个可能最优点，还要确定评估任务。Henry 等提出一种适用于多优化任务的熵搜索[31]，用扩展的采集函数和成本函数平衡优化效用和成本：

$$(x_{t+1},z_{t+1})=\arg\max_{(x,z)\in\chi\times Z}\frac{\alpha_n(x,z)}{c(x,z)} \tag{19}$$

其中，$z\in Z$，是任务空间；$\alpha_n:\chi\times Z\to\mathrm{R}$ 为采集函数；$c:\chi\times Z\to\mathrm{R}^+$ 为成本函数。作者通过超参数调优等实验验证了该方法的性能，计算效率高，具有可伸缩性和高效性，可以有效应用于具有多优化任务的问题。

4.2.2 多采集函数的并行与集成估计

由于不同超参数之间的效果评估计算本身是完全独立的，评估函数的计算可以高效并行，所以朴素的并行运行多组贝叶斯优化算法已被证明是一种较好的降低时间复杂度的方法[32]。Contal 等于 2013 年提出了 GP-UCB-PE 算法[33]，通过在不同的线程中使用不同的采集函数来获得更多样性的超参数估计。该算法在并行的同一时间步上，同时使用了 UCB 采集函数和纯随机探索策略的采集函数，纯随机探索策略获得的信息用来辅助 UCB 策略寻找下一个可能最优点。从累积误差的角度分析，作者证明了使用该方法，泛化误差的理论上界可以得到优化。Lyu 等于 2018 年提出了 MACE 方法[34]。该方法在每次优化迭代中对多个采集函数进行优化，每个采集函数代表一种独特的选择策略，不同的采集函数对于下一个点的采样位置可能不一致。此时根据上文中提到的 Pareto 原则解 max mize $[UCB(x),PI(x),EI(x)]$，即在不降低其他采集函数的输出，无法得到某个采集函数的更优点时，当前解即为最优。这种批量贝叶斯优化策略可以显著提高优化的效率。

在具有多个代理模型或者多任务问题中，天然存在多个不同的采集函数。例如 Kandasamy 等提出的 HBO 算法[19]假设未知函数可以分解为多个独立的符合高斯分布的函数的加和，这使得可以将采集函数类似地分解成独立的采集函数的和，即 $\alpha=\sum\alpha^i(x^{(i)})$，每个采集函数可以独立地最大化，从而将总计算成本降低到输入维度数目的线性级。而多任务问题中的多采集函数也可以合并进行最优化计算。Paria 等于 2019 年提出一种基于随机标量化的多任务贝叶斯优化方法[35]，假设 k 个优化任务 $(y_1,\cdots,y_k)$ 均服从高斯分布，通过参数 λ 将多任务标量化 $s_\lambda(y):R^K\to R$，优化问题转化为 $x^*=\arg\max s_\lambda(f(x))$，采集函数为：

$$s_\lambda(\mu^t(x)+\sqrt{\beta_t}\sigma^{(t)}(x)) \tag{20}$$

其中，λ 表示对优化任务的偏好，标量化处理使得该优化方法的计算复杂度与任务数量呈线性关系。

4.2.3 采集函数最优值的近似估计

随着机器学习模型中超参数规模的增长，最优化采集函数问题的搜索空间也在不断增长，蒙特卡洛方法和变分法等近似方法越来越受到重视。Snoek 等于 2012 年提出使用蒙特卡洛采样实现并行贝叶斯优化算法[1]，此方法在传统的采集函数上做了扩展：

$$\hat{\alpha}(x;D,\theta,\{x_j\})=\int\alpha(x;D,\theta,\{x_j,y_j\})\,p(\{y_j\}_{j=1}^J\mid\{x_j\}_{j=1}^J,D)\mathrm{d}y_1,\cdots,y_J \tag{21}$$

其中，θ 是高斯过程协方差函数(6)中的参数，$\theta=[\theta_0,\theta_d]$，$D$ 中的是已经完成评估的点，通过随机采样 GP 模型得到 J 个

待评估的点。该策略利用高斯过程代理模型的蒙特卡洛采样来计算采集函数的蒙特卡罗估计。这样的蒙特卡罗近似计算还被用在了其他改进形式的算法中[3]。Gong 等于 2019 年提出 QSBO 算法[36],该算法使用变分方法[37]对采集函数进行最优化,同时将基于分位数的风险度量和基于熵的正则化项引入采集函数,分别约束了采集函数可能导致的最差情形,增强了最优化结果的多样性,避免了过拟合。

4.3 算法实施中问题的改进

4.3.1 停止条件的确定问题

传统的贝叶斯优化需要迭代地运行训练过程,那么可以考虑利用训练过程中可用的信息,将最终表现不佳的超参数提前停止模型训练,从而提高 BO 算法的效率。Swersketal 提出的 Freeze Thaw BO[38]用小的步长训练模型,在初始阶段探索了不同的超参数集,然后逐渐专注于少量表现好的超参数。2016 年 Kandasametal 提出的多精度 BO[39],通过利用低精度函数来减少 BO 的资源消耗,所述低精度函数可以通过使用训练数据的子集或通过针对小的步长训练 ML 模型来获得。Dai 等于 2019 年提出 BO-BOS 方法[40],BOS 在提供原则性的最优封顶机制方面的能力使其成为在理论上合理地将提前停止引入 BO 的首选方案。在贝叶斯最优停止中,停止和继续之间的决策是为了最大化预期效用或等效地最小化预期损失。

4.3.2 初始化超参数的选择问题

虽然贝叶斯优化已经成为机器学习算法的超参数优化的成功工具,但对于大型数据集,训练和验证单个超参数通常需要数小时、数天甚至数周的时间。Klein 等提出了将验证误差的生成模型建模为关于训练集大小的函数[41],该模型在优化过程中学习,并通过探索较小子集上的初始构型外推到完整的数据集,将损失和训练时间用有关数据集大小的函数表示,并自动权衡关于全局最优的高信息增益与计算成本。

在许多情况下,贝叶斯优化存在冷启动问题,即在找到一个好的搜索区域之前,可能会找到大量函数值较低的点,这使得优化过程非常漫长。由于源数据与目标任务的关联性是未知的,Ramachandran 等提出了一种贝叶斯优化中迁移学习的最优点选择方法[42]。

4.3.3 带约束的超参数空间

许多模型的超参数选择是有约束的,例如深度神经网络中的批量大小就要求是整数,并且在具有 2 的幂次形式时计算效率较高。而大多数贝叶斯优化方法都假设了连续的超参数空间,其中一个原因是许多方法是建立在高斯过程(Gaussian Process,GP)之上的,而 GP 模型最初是为连续输入空间提出的,并且主要用于连续输入空间。所以当一些输入变量必须是一个集合中的离散值或整数值时,就必须引入额外的近似。当考虑搜索多维离散空间时,对于 k 个类别的 M 个分类变量,可能的组合的数量规模与 $O(M^k)$ 成正比,这样的问题并不比连续空间更容易解决。Oh 等于 2019 年设计了一种新的组合贝叶斯算法 COMBO[43],通过利用图模型在组合结构上引入函数的光滑性。Merchán 等最近提出在对目标进行评估之前,应该对离散变量使用一个独热编码近似,或者在整数值变量的情况下四舍五入到最接近的整数[16]。然而,当在下一次贝叶斯优化迭代计算协方差时,这种舍入不被并入。

还有许多现实优化问题具有其他形式的约束条件,如果这些约束是已知的,则可以将它们直接合并到采集函数的优化中。例如 González 等基于函数的 Lipschitz 常数的估计提出了一种高效的启发式方法[44],该方法通过在采集函数上增加惩罚项的方式选择批量的待评估点,总的采集函数可以表达为:

$$x_{t,k}=\arg\max\{g[\alpha(x;I_{t,0})]\prod_{k-1}^{j=1}\varphi(x,x_{t,j})\} \tag{22}$$

其中,$\varphi(x,x_{t,j})$ 是采集函数在 $x_{t,j}$ 处的惩罚因子,$0\leqslant\varphi(x,x_{t,j})\leqslant1$。但是很多情况下,约束的形式无法显式表达。例如在多优化任务中,如果两项优化任务密切相关,那么就可以通过优化较便宜的任务来降低较昂贵任务的优化成本。Swersky 等基于这种动态的多任务策略提出了一种采集函数[24],该函数考虑了基于熵搜索策略的噪声的不确定性造成的成本。

5 前沿研究问题

随着研究的深入,面向超参数估计的贝叶斯优化方法与理论已经日趋完善。但机器学习模型的发展也对超参数估计与选择问题不断地提出新的挑战。下面罗列了部分近年来超参数估计实践中提出的热点研究问题。

5.1 高维度超参数估计问题

随着机器学习模型的日趋复杂,特别是当前深度神经网络在计算机视觉与自然语言处理问题中的广泛使用,模型的超参数数量在不断增长,高维的超参数空间导致了更大的搜索空间,同时每一次模型的训练与测试集误差数据获取也变得更为困难。贝叶斯优化方法相对于超参数向量长度的计算,复杂度是指数级的,所以由高维超参数带来的计算代价在很多问题中是不可承受的。在上文中提到的很多方法均通过将超参数向量分解成低维子向量的方式解决问题,例如文献[26-27]等;MCMC 采样和变分法等近似计算方法也在高维问题中大量被采用,如文献[36-37]等。如何在高维超参数假设下进一步降低贝叶斯优化方法的计算复杂度是当前的一个重要研究问题。

5.2 分布式算法实现问题

当前很多问题中计算效率的提高主要依靠并行与分布式计算的使用,可以说算法的并行与分布式实现是当前大规模机器学习问题实践中的主流方案。如前所述,基本的贝叶斯优化方法需要利用已有的数据来推断新的最优超参数,这种具有时序关系的算法框架本身并不能有效地并行实现。在已有的研究中,有的工作是利用多个相对独立地代理模型进行并行计算[21-22],有的则是在同一个代理模型下利用多个不同的采集函数进行并行计算[32-33]。如果各进程独立的进行计算,事实上是与贝叶斯方法的出发点相违背的。所以如何充分地利用不同并行进程中所得到的超参数与测试集误差数据的信息,更有效地将现有的算法并行化,是当前的重要研究问题。

5.3 整数约束优化问题

实际的超参数选择问题面临许多方面的约束,其中最重要的一种类型的约束是很多超参数的取值必须是整数。众所周知,整数优化是最优化问题中最困难的一种,并且具有指数级的时间复杂度。与此同时,部分超参数的选择还影响了除模型泛化性能之外的度量,如损失函数的稳定性,GPU 的并行优化效率等,这使得整个问题变得更为复杂。现有的方法如文献[16,44]等基本是将问题先在实数域中解决,然后将结果映射到整数域上。所以,整数约束问题的算法还有相当大

的优化空间。同时，与该问题紧密联系的多任务优化问题与多目标优化问题也是当前的一个研究热点[24,30,34]。

5.4 探索利用的权衡

事实上，正如文献[19]所指出的那样，超参数选择问题和具有未知环境的强化学习问题，以及系统辨识中的输入设计问题，具有相同的问题结构。而存在探索与利用的权衡正是这些问题的共同特点与难点所在。在强化学习方法中，对探索与利用的权衡有较多的讨论，而在超参数选择问题中，这样的讨论还相对较少。缺乏这种讨论的原因在于超参数选择问题所面临的环境探索的计算代价远大于其他两个问题，这使得在当前的贝叶斯方法当中探索仅限于在采集函数的选择上展开，如文献[14,35]等在采集函数上添加了与噪声方差相关的正则项。对于复杂函数而言，足够力度的探索在能够避免局部极值问题上的收益与其付出的计算代价之间的权衡问题，还有待在算法整体上做进一步的讨论。

结束语 在超参数估计问题中，贝叶斯优化方法利用已知的超参数及其相应的模型测试集误差作为数据来构造概率代理模型，并通过最优化采集函数来推断更优的超参数配置。这种做法充分利用了现有的实验数据，极大地提高了超参数估计的效率与效果，并且具有较为完善的理论依据与收敛性保证，在实践中获得了较好的效果。由于近年来机器学习模型被广泛应用，超参数估计方法也得到了重视，贝叶斯优化在理论研究与算法实践等多方面都取得了长足的进步。该算法虽然依旧面临众多问题与挑战，但很明显会在未来成为一种重要的通用超参数选择与估计方法，应该受到本领域广大学者与研究人员的重视。

参 考 文 献

[1] SNOEK J, LAROCHELLE H, Adams R P. Practical bayesian optimization of machine learning algorithms[J]. arXiv: 1206.2944, 2012.

[2] BROCHU E, CORA V M, DE FREITAS N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning[J]. arXiv: 1012.2599, 2010.

[3] LETHAM B, KARRER B, OTTONI G, et al. Constrained Bayesian optimization with noisy experiments[J]. Bayesian Analysis, 2019, 14(2): 495-519.

[4] BERGSTRA J, BARDENET R, BENGIO Y, et al. Algorithms for hyper-parameter optimization[C]// 25th Annual Conference on Neural Information Processing Systems (NIPS 2011). Neural Information Processing Systems Foundation, 2011.

[5] BAO Y, LIU Z. A fast grid search method in support vector regression forecasting time series[C]// International Conference on Intelligent Data Engineering and Automated Learning. Berlin: Springer, 2006: 504-511.

[6] BERGSTRA J, BENGIO Y. Random search for hyper-parameter optimization[J]. Journal of Machine Learning Research, 2012, 13(1): 281-305.

[7] PELIKAN M, GOLDBERG D E, CANTÚ-PAZ E. BOA: The Bayesian optimization algorithm[C]// Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-99). 1999: 525-532.

[8] FRAZIER P I. A tutorial on Bayesian optimization [J]. arXiv: 1807.02811, 2018.

[9] SHAHRIARI B, SWERSKY K, WANG Z, et al. Taking the human out of the loop: A review of Bayesian optimization[C]// Proceedings of the IEEE. 2015: 148-175.

[10] MAHENDRAN N, WANG Z, HAMZE F, et al. Adaptive MCMC with Bayesian optimization[C]// Artificial Intelligence and Statistics. PMLR, 2012: 751-760.

[11] JONES D R, SCHONLAU M, WELCH W J. Efficient global optimization of expensive black-box functions[J]. Journal of Global Optimization, 1998, 13(4): 455-492.

[12] JIANG M. Research and Application of Bayesian Optimization algorithm[D]. Shanghai: Shanghai University, 2012.

[13] RASMUSSEN C E. Gaussian processes in machine learning [C]// Summer School on Machine Learning. Berlin: Springer, 2003: 63-71.

[14] SRINIVAS N, KRAUSE A, KAKADE S M, et al. Gaussian process optimization in the bandit setting: No regret and experimental design[J]. arXiv: 0912.3995, 2009.

[15] THORNTON C, HUTTER F, HOOS H H, et al. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms[C]// Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 847-855.

[16] GARRIDO-MERCHÁN E C, HERNÁNDEZ-LOBATO D. Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes [J]. Neurocomputing, 2020, 380: 20-35.

[17] TOSCANO-PALMERIN S, FRAZIER P I. Bayesian optimization with expensive integrands[J]. arXiv: 1803.08661, 2018.

[18] ASTUDILLO R, FRAZIER P. Bayesian optimization of composite functions[C]// International Conference on Machine Learning. PMLR, 2019: 354-363.

[19] KANDASAMY K, SCHNEIDER J, PÓCZOS B. High dimensional Bayesian optimisation and bandits via additive models [C]// International Conference on Machine Learning. PMLR, 2015: 295-304.

[20] LI C L, KANDASAMY K, PÓCZOS B, et al. High dimensional Bayesian optimization via restricted projection pursuit models [C]// Artificial Intelligence and Statistics. PMLR, 2016: 884-892.

[21] ROLLAND P, SCARLETT J, BOGUNOVIC I, et al. High-dimensional Bayesian optimization via additive models with overlapping groups[C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2018: 298-307.

[22] WANG Z, LI C, JEGELKA S, et al. Batched high-dimensional Bayesian optimization via structural kernel learning[C]// International Conference on Machine Learning. PMLR, 2017: 3656-3664.

[23] WILLIAMS C, BONILLA E V, CHAI K M. Multi-task Gaussian process prediction[C]// Advances in Neural Information Processing Systems. 2007: 153-160.

[24] SWERSKY K, SNOEK J, ADAMS R P. Multi-task bayesian optimization[C]// Advances in Neural Information Processing Systems. 2013.

[25] DJOLONGA J, KRAUSE A, CEVHER V. High-dimensional gaussian process bandits[C]// Neural Information Processing Systems. 2013.
[26] NAYEBI A, MUNTEANU A, POLOCZEK M. A framework for Bayesian optimization in embedded subspaces[C]// International Conference on Machine Learning. PMLR, 2019: 4752-4761.
[27] KIRSCHNER J, MUTNY M, HILLER N, et al. Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces[C]// International Conference on Machine Learning. PMLR, 2019: 3429-3438.
[28] HENNIG P, SCHULER C J. Entropy Search for Information-Efficient Global Optimization[J]. arXiv: 1112. 1217, 2012.
[29] HERNÁNDEZ-LOBATO J M, HOFFMAN M W, GHAHRAMANI Z. Predictive entropy search for efficient global optimization of black-box functions[J]. arXiv: 1406. 2541, 2014.
[30] HERNÁNDEZ-LOBATO D, HERNANDEZ-LOBATO J, SHAH A, et al. Predictive entropy search for multi-objective bayesian optimization[C]// International Conference on Machine Learning. PMLR, 2016: 1492-1501.
[31] MOSS H B, LESLIE D S, RAYSON P. Mumbo: Multi-task max-value bayesian optimization[J]. arXiv: 2006. 12093, 2020.
[32] WANG Z, GEHRING C, KOHLI P, et al. Batched large-scale bayesian optimization in high-dimensional spaces[C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2018: 745-754.
[33] CONTAL E, BUFFONI D, ROBICQUET A, et al. Parallel Gaussian process optimization with upper confidence bound and pure exploration[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2013: 225-240.
[34] LYU W, YANG F, YAN C, et al. Batch bayesian optimization via multi-objective acquisition ensemble for automated analog circuit design[C]// International Conference on Machine Learning. PMLR, 2018: 3306-3314.
[35] PARIA B, KANDASAMY K, PÓCZOS B. A flexible framework for multi-objective Bayesian optimization using random scalarizations[C]// Uncertainty in Artificial Intelligence. PMLR, 2020: 766-776.
[36] GONG C, PENG J, LIU Q. Quantile stein variational gradient descent for batch bayesian optimization[C]// International Conference on Machine Learning. PMLR, 2019: 2347-2356.
[37] LIU Q, WANG D. Stein variational gradient descent: A general purpose bayesian inference algorithm[J]. arXiv: 1608. 04471, 2016.
[38] SWERSKY K, SNOEK J, ADAMS R P. Freeze-thaw bayesian optimization[J]. arXiv: 1406. 3896, 2014.
[39] PERDIKARIS P, KARNIADAKIS G E. Model inversion via multi-fidelity Bayesian optimization: a new paradigm for parameter estimation in haemodynamics, and beyond[J]. Journal of The Royal Society Interface, 2016, 13(118): 20151107.
[40] DAI Z, YU H, LOW B K H, et al. Bayesian optimization meets Bayesian optimal stopping[C]// International Conference on Machine Learning. PMLR, 2019: 1496-1506.
[41] KLEIN A, FALKNER S, BARTELS S, et al. Fast bayesian optimization of machine learning hyperparameters on large datasets[C]// Artificial Intelligence and Statistics. PMLR, 2017: 528-536.
[42] RAMACHANDRAN A, GUPTA S, RANA S, et al. Selecting optimal source for transfer learning in Bayesian optimisation[C]// Pacific Rim International Conference on Artificial Intelligence. Cham: Springer, 2018: 42-56.
[43] OH C, TOMCZAK J M, GAVVES E, et al. Combinatorial bayesian optimization using the graph cartesian product[J]. arXiv: 1902. 00448, 2019.
[44] GONZÁLEZ J, DAI Z, HENNIG P, et al. Batch Bayesian optimization via local penalization[C]// Artificial Intelligence and Statistics. PMLR, 2016: 648-657.

LI Ya-ru, born in 1990, postgraduate. Her main research interests include machine learning and parameter tuning.

ZHANG Yu-lai, born in 1983, Ph.D, professor. His main research interests include parameter tuning theory and method and application of data mining.