

研究 综述

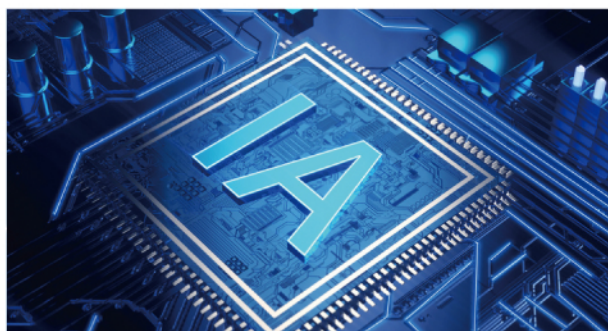
AI - 自动化机器学习

文——索信达苏钰

自动化机器学习 研究综述

CONTENTS

1. 介绍
2. 研究意义
3. 应用
 - 3.1 数据预处理
 - 3.2 特征工程
 - 3.3 模型选择
 - 3.4 算法选择
 - 3.5 深度学习
4. 技术
 - 4.1 调优器技术
 - 4.2 评估器技术
 - 4.3 经验方法
5. autoML 库
6. 总结
7. 引用



1. 介绍

近年来，机器学习在各个领域都有着巨大的贡献，然而一个高效的机器学习算法往往依赖于人类指导，例如数据处理，特征工程，算法选择以及超参选定等。随着机器学习算法越来越复杂，可选择的算法及过程数量越来越多，为了节约时间及人力成本以及提高精度，自动化机器学习（autoML）孕育而生。本文将分别从以下四方面进行介绍。



现今流行的
自动化机器学习库



为什么需要
自动化机器学习

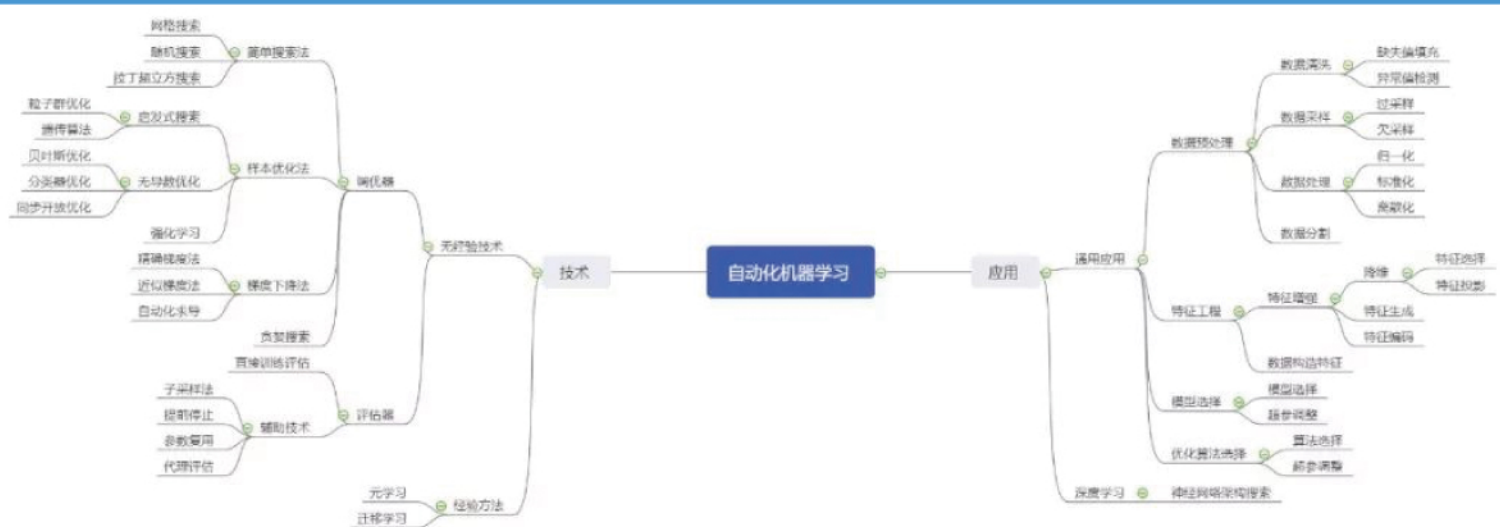


如何实现
自动化机器学习



为什么需要
自动化机器学习

下图为自动化机器学习的知识架构图



自动化机器学习本质是一个空间搜索优化问题，即从有限空间范围内以最短时间寻找最优解的过程。具体来说，自动化机器学习对机器学习每个过程产生一个备选空间，通过从每个空间中通过最优搜索选出一个最佳选项，并将其连接成流程化管道的过程。通过上述方法，在保证高精度的同时可以最小化时间及人力成本。

2. 研究意义

机器学习近些年已经进入到生活中的各个领域，包括科技，金融，医疗等行业，然而虽然机器学习名字中是“机器”在学习，它却依旧需要大量人工的指导，其中包括选择合适的数据训练，对数据及其中特征进行处理，选择合适的模型并进行超参调整等等。如此一个复杂的工作流程，需要大量的专业人员对其进行深入的学习，这无疑增大的机器学习的使用成本。更糟糕的是，随着机器学习的不断发展，新的算法如雨后春笋，技术人员越难以掌握所有技术并能选择最优的技术以解决特定的问题，这意味着技术人员往往难以探触到模型的天花板。

因此我们需要一个方法可以使得机器学习的使用更为简单，快速，效果好，可拓展，以此可以有助于推动机器学习本身在各个行业中的推广和发展。因此，自动化机器学习孕育而成。

3. 应用

自动化机器学习可以用于数据预处理，特征工程，模型选择以及算法选择之中，接下来我将分别介绍。

3.1 数据预处理



传统数据预处理包括数据清洗，数据采样，数据处理，数据分割等等过程，其中每一步都有多种备选方法，选择方法前，往往需要对数据进行先一步的数据分析，整个数据预处理过程往往重复而耗时，于是我们可以将该过程加入搜索空间中，由机器一并完成。

3.2 特征工程



特征工程自动化具体可以分成从数据创造特征和增强特征区分能力两个子问题，从数据创造特征十分依赖人为的专业知识，较难自动化。所以我们重点介绍自动化特征增强，特征增强的方法主要包括降维，特征生成和特征编码。此处降维与传统特征工程相同，不做过多解释。特征生成则是发现原特征间的交互项，通过两变量相乘和标准化的形式引入新变量。特征编码则是通过基于数据所得的字典重新解释原始特征，流行方法有离散编码，局部线性编码，以及核方法等。

3.3 模型选择



主要包括两部分，分别是选择模型和设置超参数。解决同一个问题可能有多个备选模型，且每个模型的超参数也是未知，传统方法多是通过人力专业知识以及反复测试得到“最优”结果，自动化机器学习则可以解决这个繁琐的问题。其搜索空间由预选模型和他们的超参数组成，但是超参数的选择需要优先确定模型。

3.4 算法选择



对于经典学习模型来说，优化并不困难，因为大多有凸损失函数，困难点在于如何优化效率。然而随着模型的复杂度越来越高，调优的代价越来越高，以至于会对学习效果造成影响。所以算法选择的目的是自动找到一个优化算法可以平衡其效率和表现。例如一些流行方法的目标是最小化平滑目标函数，需要在梯度下降，随机梯度下降和L-BFGS算法中进行选择。梯度下降法有少的超参数，但是收敛较慢，每个迭代复杂度低。L-BFGS消耗更高但是收敛更快，SGD每个迭代消耗很小但是需要的迭代次数较多。此时自动化机器学习就能够在他们之中权衡效率和表现以选择最优算法。

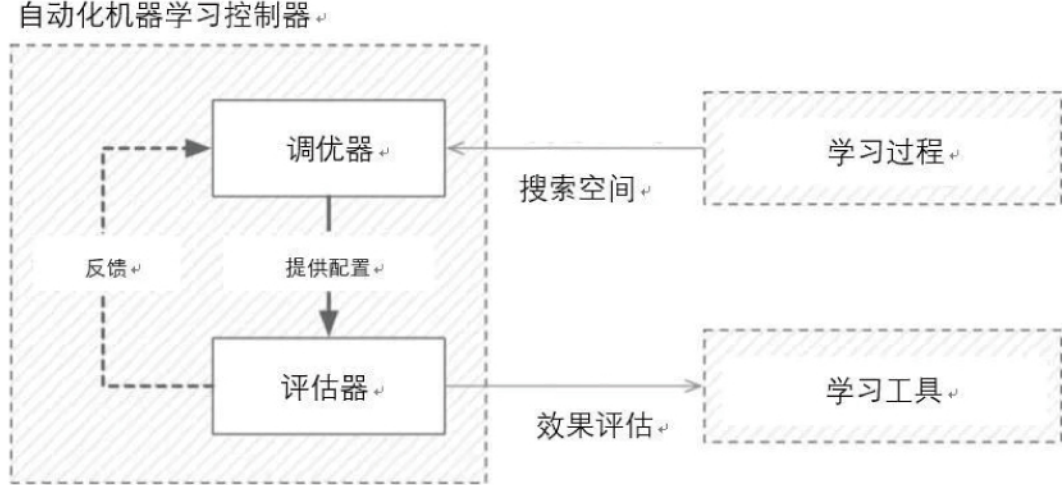
3.5 深度学习



对于经典学习模型来说，优化并不困难，因为大多有凸损失函数，困难点在于如何优化效率。然而随着模型的复杂度越来越高，调优的代价越来越高，以至于会对学习效果造成影响。所以算法选择的目的是自动找到一个优化算法可以平衡其效率和表现。例如一些流行方法的目标是最小化平滑目标函数，需要在梯度下降，随机梯度下降和L-BFGS算法中进行选择。梯度下降法有少的超参数，但是收敛较慢，每个迭代复杂度低。L-BFGS消耗更高但是收敛更快，SGD每个迭代消耗很小但是需要的迭代次数较多。此时自动化机器学习就能够在他们之中权衡效率和表现以选择最优算法。

4.技术

一般来说，自动化机器学习主要包括评估器和调优器两部分。评估器负责测量学习工具在调优器提供的参数条件下的表现，之后将结果反馈给调优器。调优器则负责通过反馈信息为学习工具更新或生成配置信息。



由上述应用可以看出，自动化机器学习的最大难点在于其庞大的搜索空间，所以找到一种可以快速高效的搜索算法则是整个自动化机器学习研究的主要目标。

4.1

调优器技术

对于调优器技术来说，主要有三种方法，分别是简单搜索方法，根据样本优化方法和梯度下降方法。

简单搜索方法

搜索空间不需要假设，每个配置之间相互独立。主要有网格搜索和随机搜索两种方法。简单但是不高效。为了解决这个问题，后来推出了拉丁超立方算法能够在减少取样点的同时保证空间的大概率被探索。简单搜索算法的主要问题在于，搜索方法不能自动化最优适应不同的问题场景。

样本优化法

根据先前的评估样本设定新的配置。不需要特定的假设但是更加高效。主要有三类，分别是启发式搜索，基于模型的非导优化和强化学习。

对于调优器技术来说，主要有三种方法，分别是简单搜索方法，根据样本优化方法和梯度下降方法。

启发式搜索：取灵感于生物行为和现象，广泛用于非凸，不平滑，不连续的调优问题。基础思想是初始化一个群体，通过调优器和原种群获得新群体，然后评估新种群，重复迭代上述过程。核心是如何更新群体。

粒子群优化 (particle swarm optimization)：粒子群优化灵感取源于群体迁徙的过程。群体之间通过共享位置，每轮迭代向最优点靠近。优点是整个过程可并行执行。

遗传算法 (evolutionary algorithm)：遗传算法灵感取源于基因杂交的过程。通过将各个个体进行杂交和变异，为了尽可能的保留优秀的基因组合，在接下来的杂交过程中，优秀基因的祖先越有可能被选中。

基于模型的非导优化：通过样本构建模型，之后根据评估生成新的样本，然后重复迭代上述过程，以达到有目标的空间搜索的目的，可以用于不可导的空间优化，其中主要包括贝叶斯优化，基于分类的优化，以及同步乐观优化。

贝叶斯优化：通过构建一个概率模型（例如高斯过程，树模型，深度网络），之后基于概率模型定义一个采集函数（例如期望提升，上置信界限），每次迭代由采集函数获得新样本，用于更新概率函数。该方法是现今最主流的方法之一，贝叶斯优化有收敛速度快的优点，缺点在于当空间不平滑时效果不佳，

基于分类的优化 (Classification-based optimization)：通过旧样本训练一个分类器，将搜索空间分为正负区域，正区域中的样本更可能得到优秀结果。所以新样本从正区域获得，迭代上述步骤。该方法的优点是高效。

同步乐观优化 (Simultaneous Optimistic optimization)：是一种分支界限优化算法。在搜索空间内构造一个树结构，每个叶子节点为一个小区域。SOO 可以协调深度和广度，找到全局最优点。但是会由于数据维度上升而使树变得十分复杂。

强化学习 (reinforcement learning)：RL 是一种广泛并强大的优化框架，通过延迟反馈解决问题，与其他优化方法的区别在于它存在延迟反馈，以给学习增加时序的概念。其中主要有两个算法，分别是策略学习和 Q-learning。

策略学习 (Policy Learning)：通过把策略看作是函数，只有当前状态一个输入，根据事先的策略决定当前状态下该执行的动作。但是要事先知道策略并不是件容易事，其中需要深入理解这个把状态映射到目标的复杂函数。

Q-Learning：另一个指导代理的方式是给定框架后让代理根据当前环境独自做出动作，而不是明确地告诉它在每个状态下该执行的动作。与策略学习不同，Q-Learning 算法有两个输入，分别是状态和动作，并为每个状态动作对返回对应值。当你面临选择时，这个算法会计算出该代理采取不同动作时对应的期望值，以选出最优结果。

梯度下降

由于自动化机器学习的问题很复杂，很多情况不可导，甚至不连续，所以梯度下降不是很流行。但是对于一些可导的损失函数，梯度下降更加高效。

为了提高计算梯度的效率，有两种可以使用的方法，其一是可以不计准确梯度，而是计算非准确梯度，这样超参可以在模型收敛之前更新，来使梯度下降法更为高效。另一种计算梯度的方法是通过自动化求导，通过链规则和向后传播求导。

神经架构搜索

在神经架构搜索中，除了可以使用上述介绍的几种技术以外，由于深度学习本身独特的搜索空间，只有支持离散变量优化的算法才可以使用，这导致许多常用的优化算法无法用于神经架构搜索中，为了解决这个问题，在 NAS 中，可以通过将离散的配置空间改为连续空间，使其可以使用更强大的其他优化技术。

贪婪搜索

贪婪搜索方法，即每一步都会使用最优决定来解决问题，虽然贪婪搜索往往找不到全局最优，但是可以在较短的时间里找到近似全局最优的局部最优。

4.2

评估器技术

相对于调优器技术来说，评估器的整体消耗要大得多，而且评估器技术较少，现今的主要方法还是使用直接评估方法。

直接评估：最简单的方法，直接训练模型后评估，虽然准确但是消耗较大。随着数据量的增大和迭代次数的增多，直接评估无疑会给整个过程造成极大的负担，为了改善直接评估，研究人员们设计了以下方法以辅助直接评估方法，来减少其消耗。

子取样法：使用原样本或特征子集进行评估，训练数据越少，速度越快，噪音越多。

提前终止：不同于传统机器学习中，提前终止用于防止过拟合。这里是当遇到没有前途的配置信息，可以直接终止评估，免除不必要的浪费。缺点是可能会错过起初表现不佳，但最终经过训练后效果较好的配置。

参数复用：对于差异不大的配置信息，可以使用先前的参数作为起始信息，这样可以加快收敛速度，得到更好的表现。缺点在于，不同的起始点，可能造成不同的局部优化，为评估带来偏斜。参数复用可以看做迁移学习的一种。

代理评估：假如配置信息可以被量化，可以通过构建代理模型来预测所给配置的表现，其中后文介绍的元学习便可使用该技术。

4.3

经验方法

上述介绍的方法都是不需要凭借先前信息的方法，然而对于人来说，人往往不需要面对每个新问题都要重新进行一次学习，因为人们可以迁移之前的学习经验，或者套用之前问题的方法来解决新问题，因此研究人员提出了关于经验的机器学习算法。经验学习主要有两种方法，分别是元学习和迁移学习。由于经验学习发展还在初始阶段，两种方法并没有明确的界限，其中许多算法思想都是相通的，都是通过在配置生成和评估过程中减少消耗以提高自动化机器学习的效率。

元学习

元学习顾名思义，即通过提取元信息来指导学习。元学习首先要将学习问题和学习工具特征化(例如数据的统计的特征，学习工具的超参)，然后从过去的经验的中提取出来的元特征，最后元学习器可以用元知识进行训练。

元学习在自动化机器学习有着重要意义，一方面，通过特征化学习问题和学习工具，可以发现重要的信息，比如数据中存在数据漂移（模型随时间推移不再准确），其次，伴随着特征化，容易发现相似的问题，从而能够知识可以在不同的问题间复用和转移。另一方面，元学习器编码过去的知识作为指导以解决未来的问题。

元学习可以运用于评估器中，以降低评估过程中训练所造成的的巨大消耗。通过把配置信息输入先前训练好的元学习器进行评估，以预测配置的表现或者适应度，最理想情况下，如果所有的配置已经被枚举出来，元学习器可以直接选出最优配置。

元学习也可以运用于调优器中，通过优化搜索空间，以减少调优过程的无意义消耗，在配置生成阶段，通过先前经验所得的元学习器，提取学习问题的特征作为输入来预测有前途的配置。同时可以应用于迁移学习，通过运用与之前任务元特征空间最相近的配置作为初始化数据来进行配置生成热身启动。

除此之外，元学习还可以应用于动态配置自适应，通过数据和特征的统计值，检测是否有概念漂移出现，一旦发现概念漂移，则重新预测有前景的配置，以保证模型可用性。

迁移学习

迁移学习通过使用先前的经验来指导学习，在机器学习中则是复用之前最优的训练过的代理模型或者是搜索策略，来节约消耗。在调优过程中，可以进行代理模型迁移，在网络结构问题中，由于网络的可转移性，迁移学习也广泛运用在神经架构搜索中。

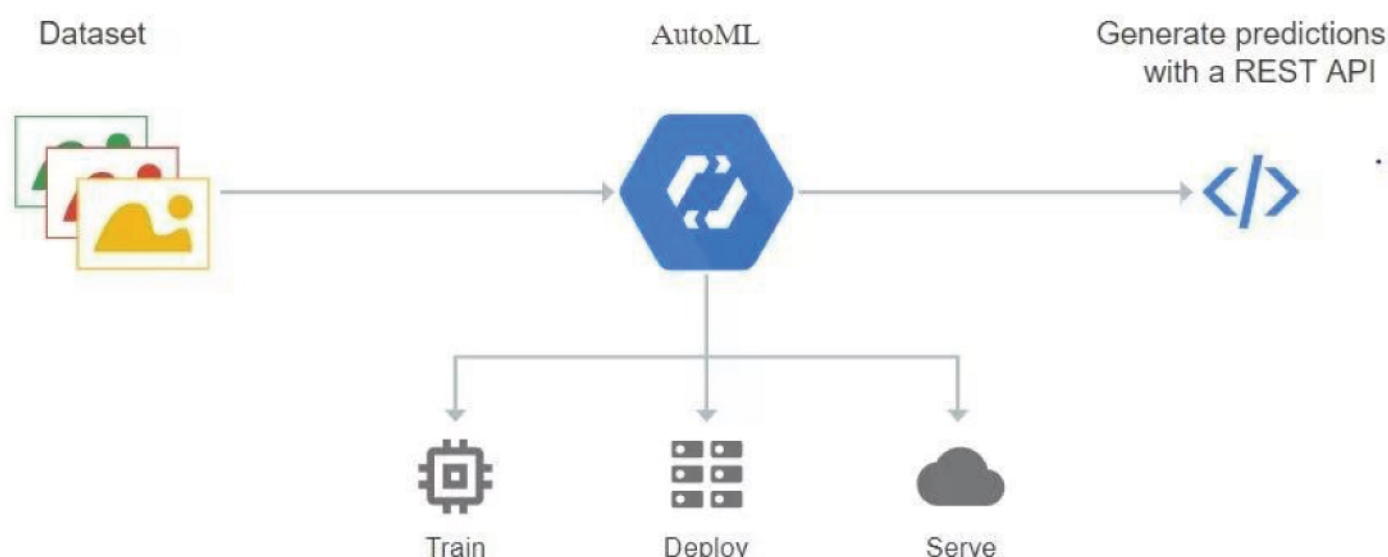
迁移学习运用在评估器中，可以加快对于预选配置的评估。对于一般的优化问题，迁移学习可以迁移模型参数，通过使用训练过的最优参数来进行初始化。另一种迁移学习的思路是通过功能保留转换，例如 Net2Net，通过要求和先前训练模型相同的功能来初始化新的网络，以加快网络结构的训练过程。

5. autoML库

5.1

Cloud AutoML

Cloud AutoML 是由 google 开发的现今集成程度最高的自动化机器学习平台，其中集合了包括自然语言处理，图像处理，表格处理等当今几乎所有机器学习领域。使用者几乎不需要任何专业技能，只要将数据上传后根据引导点击一些设定，整个机器学习过程就可以高效完成。



然而缺点在于该平台是闭源的，换句话说，Cloud AutoML 更像一个近似 AWS 的服务，因此它是收费的。更多使用细则见官方文档 <https://cloud.google.com/automl/>

5.2

MLBox

MLBox 是一个强大的自动化机器学习库，它包含了快速读取和分布式数据预处理，有稳定的特征工程，并且能在高维情况下，准确自动化调参，库中包含最新的预测模型，并能产生满足模型解释性的预测结果。



MLBox, Machine Learning Box

其在 kaggle 比赛中都有着较为出色的排名，更多的使用方法和官方文档，可见 <https://mlbox.readthedocs.io/en/latest/>

5.3

Auto-Sklearn

Auto-Sklearn 是一个自动化机器学习工具组，它可以用于替代 sklearn 的评估器。

scikit-learn 作为最火的 python 机器学习包，已经得到了广泛的运用。Auto-Sklearn 让使用者不再需要关心算法选择和超参调整。其中运用了贝叶斯调优，元学习以及集成构造等方法，更多具体技术内容可见论文 <http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf> 或者官方文档 <https://automl.github.io/auto-sklearn/master/>

5.4

TPOT



TPOT 是基于 Scikit-Learn 的一个自动化机器学习工具，不同于 Auto-Sklearn 的是，TPOT 使用遗传算法作为空间搜索算法，在特定任务下，可以得到比无导数优化方法更好的结果。更多信息可见官方文档 <http://epistasislab.github.io/tpot/>

5.5

AutoKeras



AutoKeras

TPOT 是基于 Scikit-Learn 的一个自动化机器学习工具，不同于 Auto-Sklearn 的是，TPOT 使用遗传算法作为空间搜索算法，在特定任务下，可以得到比无导数优化方法更好的结果。更多信息可见官方文档 <http://epistasislab.github.io/tpot/>

6. 总结

本文主要总结了现阶段自动化机器学习的发展情况，简单介绍了自动化机器学习的应用和内部的技术，以及当今一些较为成熟的库。现阶段的自动化机器学习仍然存在许多的问题，一方面是，经验学习方法包括元学习和迁移学习才刚刚起步，无论是理论还是实践都有着很多的空白。另一方面，即使是较为成熟的传统机器学习调参部分，也有着评估器效率低，调优器空间搜索迭代次数多等问题。这些方向都等待着研究人员的解决，因此，研究自动化机器学习是一个十分有前途的方向，在接下来的文章中，我将具体介绍自动化机器学习中的各个部分。

7. 引用

【1】 Quanming Yao, Mengshuo Wang, Yuqiang Chen, Wenyuan Dai, Yi-Qi Hu, Yu-Feng Li, Wei-Wei Tu, Qiang Yang, Yang Yu (2019), Taking the Human out of Learning Applications: A Survey on Automated Machine Learning, arXiv:1810.13306 [cs.AI]

【2】 Gang Luo (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values, Network Modeling Analysis in Health Informatics and Bioinformatics, 2016, Volume 5, Number 1, Page 1

【3】 Parul Pandey, AutoML: The Next Wave of Machine Learning, <https://heartbeat.fritz.ai/automl-the-next-wave-of-machine-learning-5494baac615f>