(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108446741 B
(45)授权公告日 2020.01.07

(21)申请号 201810270934.5

(22)申请日 2018.03.29

(65)同一申请的已公布的文献号
申请公布号 CN 108446741 A

(43)申请公布日 2018.08.24

(73)专利权人 中国石油大学(华东)
地址 266580 山东省青岛市经济技术开发区长江西路66号

(72)发明人 孙运雷 魏倩 孔言

(74)专利代理机构 济南圣达知识产权代理有限公司 37221
代理人 黄海丽

(51)Int.Cl.
G06K 9/62(2006.01)

(56)对比文件
CN 105701509 A,2016.06.22,全文.
CN 106295682 A,2017.01.04,全文.
CN 105531725 A,2016.04.27,全文.

审查员 陈国灿

权利要求书5页 说明书6页 附图1页

(54)发明名称

机器学习超参数重要性评估方法、系统及存储介质

(57)摘要

本发明公开了机器学习超参数重要性评估方法、系统及存储介质，获取OpenML中不同的数据集，并提取元特征来表示每个数据集，同时收集待评估分类算法在不同超参配置下性能的数据；提取元特征来表示使用的目标数据集，并通过计算元特征之间的距离获得目标数据集与历史数据集之间距离的递增序列；使用待评估分类算法不同超参的性能数据来评估超参重要性，根据历史数据集与目标数据集距离递增的有序序列，对距离目标数据集较近的前m个历史数据集依次执行提出的Relief和聚类算法，最终获得待评估分类算法的超参重要性排序并指导的自动化调参过程。本发明对于分类算法黑盒的超参调整给予一定的指导，从而达到节省时间，提高效率的目的。

CN 108446741 B

1.基于机器学习超参数重要性评估的待分类数据分类系统,其特征是,包括:

历史数据集获取模块,其被配置为:从开放式机器学习环境OpenML中获取与目标数据集类型相似的若干新数据集,并对每个新数据集提取元特征,使得每个新数据集都用元特征向量来表示;

从开放式机器学习环境OpenML中收集待评估分类算法在不同超参数配置下性能的数据;

将每个新数据集的元特征向量以及不同超参数配置对应的性能数据存储于对应的历史数据集中;

距离序列获取模块,其被配置为:提取目标数据集的元特征向量来表示目标数据集,计算目标数据集元特征向量与历史数据集元特征向量之间的距离,获得目标数据集与每个历史数据集之间距离由近至远的距离序列;

输出模块,其被配置为:对距离目标数据集最近的前f个历史数据集依次执行Relief-Cluster算法:通过Relief算法得到的每类超参数的权重,进一步计算每类超参数的平均权重,利用每类超参数的平均权重初步得到每类超参数重要性权重排序;利用聚类算法进一步验证超参数重要性评估的准确性;最后,得到待评估分类算法的超参数重要性排序;

分类模块,其被配置为:根据得到的待评估分类算法的超参数重要性排序,对重要性排序靠前的若干个参数进行设置,然后,利用设置好参数的分类算法对待分类数据进行分类。

2.如权利要求1所述的系统,其特征是,所述历史数据集获取模块中,每个数据集D_i被描述为由F个元特征表示的向量$V^i=(V_1^i,...,V_F^i)$; i=1,2,...N。

3.如权利要求1所述的系统,其特征是,所述历史数据集获取模块中,元特征,包括:简单的元特征、数据集的统计元特征和重要性元特征;

所述简单的元特征,包括:数据集样本数量、特征数量、类别数量或缺失值数量;

所述数据集的统计元特征,包括:平均值、方差或距离向量的峰度;

所述重要性元特征,包括:在数据集上运行机器学习算法获得的性能。

4.如权利要求1所述的系统,其特征是,所述历史数据集获取模块中待评估分类算法在不同超参数配置下的性能,包括:错误分类率或者RMSE。

5.如权利要求1所述的系统,其特征是,利用元特征向量之间的距离来衡量目标数据集D_{N+1}与历史数据集D_i之间的距离$d_{pn}(D_{N'},D_i)$:

$d_{pn}(D_{N'},D_i)=||V_{N'}-V_i||_{pn}$

其中,$V_{N'}$表示目标数据集$D_{N'}$的元特征向量,V_i表示历史数据集D_i的元特征向量,pn表示p范数;

通过目标数据集与历史数据集元特征向量之间的距离比较,得到历史数据集与目标数据集距离由近至远的排序序列π(1),...,π(N)。

6.如权利要求1所述的系统,其特征是,

所述通过Relief算法得到的每类超参数的权重包括:

根据不同超参数配置下的性能数据大小设置阈值,将历史数据集中不同超参数配置对应的性能数据分为高性能样本和低性能样本,Relief算法首先从性能数据中随机选择一个样本s_i,然后从性能高样本和性能差样本中各选择一个距离s_i最近的样本;

与s_i同类的样本s_j用M表示，与s_i不同类的样本s_j用Q表示，每类超参数h的权重w_h根据公式(1)更新：

$w_h = w_h - diff(h,s_i,M)/rt + diff(h,s_i,Q)/rt$ (1)

$diff(h,s_i,M)$表示两个样本s_i与M在超参数h上的差异；

$diff(h,s_i,Q)$表示两个样本s_i与Q在超参数h上的差异；

两个样本s_i与s_j在超参数h上的差异$diff(h,s_i,s_j)$定义为：

若超参数h为标量型超参数，

$$diff(h,s_i,s_j)=\begin{cases}0, s_{ih}=s_{jh}\\ 1, s_{ih}\neq s_{jh}\end{cases} \quad (2)$$

若超参数h为数值型超参数，

$$diff(h,s_i,s_j)=\left|\frac{s_{ih}-s_{jh}}{max_h-min_h}\right| \quad (3)$$

其中，$1\leqslant i\neq j\leqslant m$，$1\leqslant h\leqslant ph$，$max_h$为超参数h在样本集中的最大值，$min_h$为超参数h在样本集中的最小值，m表示样本数，每个样本包含ph个超参数，rt表示迭代次数，$rt>1$，s_{ih}表示在样本s_i上超参h的值，s_{jh}表示在样本s_j上超参h的值。

7.基于机器学习超参数重要性评估的待分类数据分类系统，其特征是，包括：存储器、处理器以及存储在存储器上并在处理器上运行的计算机指令，所述计算机指令被处理器运行时，完成以下步骤：

步骤(1)：从开放式机器学习环境OpenML中获取与目标数据集类型相似的若干新数据集，并对每个新数据集提取元特征，使得每个新数据集都用元特征向量来表示；

从开放式机器学习环境OpenML中收集待评估分类算法在不同超参数配置下性能的数据；

将每个新数据集的元特征向量以及不同超参数配置对应的性能数据存储于对应的历史数据集中；

步骤(2)：提取目标数据集的元特征向量来表示目标数据集，计算目标数据集元特征向量与历史数据集元特征向量之间的距离，获得目标数据集与每个历史数据集之间距离由近至远的距离序列；

步骤(3)：对距离目标数据集最近的前f个历史数据集依次执行Relief-Cluster算法：通过Relief算法得到的每类超参数的权重，进一步计算每类超参数的平均权重，利用每类超参数的平均权重初步得到每类超参数重要性权重排序；利用聚类算法进一步验证超参数重要性评估的准确性；最后，得到待评估分类算法的超参数重要性排序；

步骤(4)：根据得到的待评估分类算法的超参数重要性排序，对重要性排序靠前的若干个参数进行设置，然后，利用设置好参数的分类算法对待分类数据进行分类。

8.如权利要求7所述的系统，其特征是，所述步骤(1)中，每个数据集D_i被描述为由F个元特征表示的向量$V^i=(V^i_1,...,V^i_F)$；i=1,2,...N。

9.如权利要求7所述的系统，其特征是，所述步骤(1)中，元特征，包括：简单的元特征、数据集的统计元特征和重要性元特征；

所述简单的元特征，包括：数据集样本数量、特征数量、类别数量或缺失值数量；

所述数据集的统计元特征，包括：平均值、方差或距离向量的峰度；

所述重要性元特征，包括：在数据集上运行机器学习算法获得的性能。

10.如权利要求7所述的系统，其特征是，所述步骤(1)中待评估分类算法在不同超参数配置下的性能，包括：错误分类率或者RMSE。

11.如权利要求7所述的系统，其特征是，利用元特征向量之间的距离来衡量目标数据集D_{N+1}与历史数据集D_i之间的距离$d_{pn}(D_{N'},D_i)$：

$d_{pn}(D_{N'},D_i)=||V_{N'}-V_i||_{pn}$；

其中，$V_{N'}$表示目标数据集$D_{N'}$的元特征向量，V_i表示历史数据集D_i的元特征向量，pn表示p范数；

通过目标数据集与历史数据集元特征之间的距离比较，得到历史数据集与目标数据集距离由近至远的排序序列π(1),...,π(N)。

12.如权利要求7所述的系统，其特征是，

所述通过Relief算法得到的每类超参数的权重包括：

根据不同超参数配置下的性能数据大小设置阈值，将历史数据集中不同超参数配置对应的性能数据分为高性能样本和低性能样本，Relief算法首先从性能数据中随机选择一个样本s_i，然后从性能高样本和性能差样本中各选择一个距离s_i最近的样本；

与s_i同类的样本s_j用M表示，与s_i不同类的样本s_j用Q表示，每类超参数h的权重w_h根据公式(1)更新：

$$w_h=w_h-diff(h,s_i,M)/rt+diff(h,s_i,Q)/rt \quad (1)$$

$diff(h,s_i,M)$表示两个样本s_i与M在超参数h上的差异；

$diff(h,s_i,Q)$表示两个样本s_i与Q在超参数h上的差异；

两个样本s_i与s_j在超参数h上的差异$diff(h,s_i,s_j)$定义为：

若超参数h为标量型超参数，

$$diff(h,s_i,s_j)=\begin{cases}0, s_{ih}=s_{jh}\\ 1, s_{ih}\neq s_{jh}\end{cases} \quad (2)$$

若超参数h为数值型超参数，

$$diff(h,s_i,s_j)=\left|\frac{s_{ih}-s_{jh}}{max_h-min_h}\right| \quad (3)$$

其中，$1\leqslant i\neq j\leqslant m$，$1\leqslant h\leqslant ph$，$max_h$为超参数h在样本集中的最大值，$min_h$为超参数h在样本集中的最小值，m表示样本数，每个样本包含ph个超参数，rt表示迭代次数，$rt>1$，s_{ih}表示在样本s_i上超参h的值，s_{jh}表示在样本s_j上超参h的值。

13.一种计算机可读存储介质，其特征是，其上运行有计算机指令，所述计算机指令被处理器运行时，完成以下步骤：

步骤(1)：从开放式机器学习环境OpenML中获取与目标数据集类型相似的若干新数据集，并对每个新数据集提取元特征，使得每个新数据集都用元特征向量来表示；

从开放式机器学习环境OpenML中收集待评估分类算法在不同超参数配置下性能的数据；

将每个新数据集的元特征向量以及不同超参数配置对应的性能数据存储于对应的历史数据集中；

步骤(2)：提取目标数据集的元特征向量来表示目标数据集，计算目标数据集元特征向

量与历史数据集元特征向量之间的距离，获得目标数据集与每个历史数据集之间距离由近至远的距离序列；

步骤(3)：对距离目标数据集最近的前f个历史数据集依次执行Relief-Cluster算法：通过Relief算法得到的每类超参数的权重，进一步计算每类超参数的平均权重，利用每类超参数的平均权重初步得到每类超参数重要性权重排序；利用聚类算法进一步验证超参数重要性评估的准确性；最后，得到待评估分类算法的超参数重要性排序；

步骤(4)：根据得到的待评估分类算法的超参数重要性排序，对重要性排序靠前的若干个参数进行设置，然后，利用设置好参数的分类算法对待分类数据进行分类。

14.如权利要求13所述的介质，其特征是，所述步骤(1)中，每个数据集D_i被描述为由F个元特征表示的向量$V^i=(V_1^i,...,V_F^i)$；i=1,2,...N。

15.如权利要求13所述的介质，其特征是，所述步骤(1)中，元特征，包括：简单的元特征、数据集的统计元特征和重要性元特征；

所述简单的元特征，包括：数据集样本数量、特征数量、类别数量或缺失值数量；

所述数据集的统计元特征，包括：平均值、方差或距离向量的峰度；

所述重要性元特征，包括：在数据集上运行机器学习算法获得的性能。

16.如权利要求13所述的介质，其特征是，所述步骤(1)中待评估分类算法在不同超参数配置下的性能，包括：错误分类率或者RMSE。

17.如权利要求13所述的介质，其特征是，利用元特征向量之间的距离来衡量目标数据集D_{N+1}与历史数据集D_i之间的距离$d_{pn}(D_{N'},D_i)$：

$d_{pn}(D_{N'},D_i)=||V_{N'}-V_i||_{pn}$；

其中，$V_{N'}$表示目标数据集$D_{N'}$的元特征向量，V_i表示历史数据集D_i的元特征向量，pn表示p范数；

通过目标数据集与历史数据集元特征向量之间的距离比较，得到历史数据集与目标数据集距离由近至远的排序序列π(1),...,π(N)。

18.如权利要求13所述的介质，其特征是，所述通过Relief算法得到的每类超参数的权重包括：

根据不同超参数配置下的性能数据大小设置阈值，将历史数据集中不同超参数配置对应的性能数据分为高性能样本和低性能样本，Relief算法首先从性能数据中随机选择一个样本s_i，然后从性能高样本和性能差样本中各选择一个距离s_i最近的样本；

与s_i同类的样本s_j用M表示，与s_i不同类的样本s_j用Q表示，每类超参数h的权重w_h根据公式(1)更新：

$$w_h=w_h-diff(h,s_i,M)/rt+diff(h,s_i,Q)/rt \quad (1)$$

$diff(h,s_i,M)$表示两个样本s_i与M在超参数h上的差异；

$diff(h,s_i,Q)$表示两个样本s_i与Q在超参数h上的差异；

两个样本s_i与s_j在超参数h上的差异$diff(h,s_i,s_j)$定义为：

若超参数h为标量型超参数，

$$diff(h,s_i,s_j)=\begin{cases}0, s_{ih}=s_{jh}\\ 1, s_{ih}\neq s_{jh}\end{cases} \quad (2)$$

若超参数h为数值型超参数，

$$\mathrm{diff}(h,s_i,s_j)=\left|\frac{s_{ih}-s_{jh}}{max_h-min_h}\right| \quad (3)$$

其中，$1\leqslant i\neq j\leqslant m$，$1\leqslant h\leqslant ph$，$max_h$为超参数h在样本集中的最大值，$min_h$为超参数h在样本集中的最小值，m表示样本数，每个样本包含ph个超参数，rt表示迭代次数，$rt>1$，s_{ih}表示在样本s_i上超参h的值，s_{jh}表示在样本s_j上超参h的值。

机器学习超参数重要性评估方法、系统及存储介质

技术领域

[0001] 本发明是机器学习超参数重要性评估方法、系统及存储介质。

背景技术

[0002] 机器学习为数据处理和数据分类提供了重要的技术支撑，然而模型选择和调参依然是困扰用户的两大难题，于是自动化机器学习系统应运而生。自动化机器学习系统利用自动化机器学习算法达到了自动化数据预处理，自动化选择算法，自动化调参的目的，提高了数据分类预测的准确性，同时将用户从选择算法和反复调参的繁重任务中解脱出来。

[0003] 由于自动化机器学习的核心是自动化算法选择及自动化超参配置，因此该系统将机器学习过程归约成了算法选择和超参优化(Combined Algorithm Selection and Hyper-parameter optimization，CASH)问题。CASH问题即把算法的选择当做根层次的新的超参数，从而将选择算法和超参数值的问题映射到选择超参值的问题。通过将数据预处理和特征选择技术作为超参数，系统可以自动选择数据预处理和特征选择技术。最终归结为的超参优化问题可以通过经典的贝叶斯优化算法找到最优解，从而达到提升数据分类预测精度的效果。

[0004] 然而目前的自动化机器学习系统的超参配置模块的配置过程全凭经验，或者通过反复迭代得到最后的结果来对若干个超参数的配置进行一一调整，这样存在的缺陷是：浪费机器学习的时间，而且反复迭代也浪费计算机资源，不分重要性地对所有超参数的配置进行调整会浪费用户的时间和精力。

发明内容

[0005] 本发明是机器学习超参数重要性评估方法、系统及存储介质，所要解决的技术问题是如何准确评估机器学习算法的超参重要性，并将其用于指导自动化超参配置以及增强超参配置的可解释性问题。

[0006] 作为本发明的第一方面：

[0007] 机器学习超参数重要性评估方法，包括：

[0008] 步骤(1)：从开放式机器学习环境OpenML中获取与目标数据集类型相似的若干新数据集，并对每个新数据集提取元特征向量，使得每个新数据集都用元特征向量来表示；

[0009] 从开放式机器学习环境OpenML中收集待评估分类算法在不同超参数配置下性能的数据；

[0010] 将每个新数据集的元特征向量以及不同超参数配置对应的性能数据存储于对应的历史数据集中；

[0011] 步骤(2)：提取目标数据集的元特征向量来表示目标数据集，计算目标数据集元特征向量与历史数据集元特征向量之间的距离，获得目标数据集与每个历史数据集之间距离由近至远的距离序列；

[0012] 步骤(3)：对距离目标数据集最近的前f个历史数据集依次执行Relief-Cluster算

法:通过Relief算法得到的每类超参数的权重,进一步计算每类超参数的平均权重,利用每类超参数的平均权重初步得到每类超参数重要性权重排序;利用聚类算法进一步验证超参数重要性评估的准确性;最后,得到待评估分类算法的超参数重要性排序。

[0013] 所述机器学习超参数重要性评估方法,包括以下步骤:

[0014] 步骤(4):根据得到的待评估分类算法的超参数重要性排序,对重要性排序靠前的若干个参数进行设置,然后,利用设置好参数的分类算法对待分类数据进行分类。

[0015] 所述步骤(1)中,每个数据集D_i被描述为由F个元特征表示的向量$V^i=(V_1^i,...,V_F^i)$;i=1,2,...N;

[0016] 所述步骤(1)中,元特征,包括:简单的元特征、数据集的统计元特征和重要性元特征;

[0017] 所述简单的元特征,包括:数据集样本数量、特征数量、类别数量或缺失值数量;

[0018] 所述数据集的统计元特征,包括:平均值、方差或距离向量的峰度;

[0019] 重要性元特征,包括:在数据集上运行机器学习算法获得的性能。

[0020] 所述步骤(1)中待评估分类算法在不同超参数配置下的性能,包括:错误分类率或者RMSE;

[0021] 另外,对于许多常见算法,开放式机器学习环境OpenML已经包含了非常全面的性能数据,适用于各种数据集上的不同超参数配置,即收集数据集D_i在待评估分类算法下的超参配置θ_i及性能y_i数据$<\theta_i,y_i>_{k=1}^{K}$。

[0022] 对于目标数据集$D_{N'}$,提取元特征$V_{N'}$来表示目标数据集,并基于不相似的数据集其使用算法的超参数配置也具有差异这一原则,利用元特征向量之间的距离获得目标数据集与历史数据集之间的距离序列。对距离目标数据集近的前f个历史数据集,使用算法在不同超参数的性能数据来评估超参数重要性;

[0023] 利用元特征向量之间的距离来衡量目标数据集$D_{N'}$与历史数据集D_i之间的距离$d_{pn}(D_{N'},D_i)$:

[0024] $d_{pn}(D_{N'},D_i)=||V_{N'}-V_i||_{pn}$

[0025] 其中,$V_{N'}$表示数据集$D_{N'}$的元特征向量,V_i表示历史数据集D_i的元特征向量,pn表示p范数。

[0026] 通过目标数据集与历史数据集元特征向量之间的距离比较,得到历史数据集与目标数据集距离由近至远的排序序列π(1),...,π(N),其中$(\pi(i)\le\pi(j)) \Leftrightarrow (d(D_{N'},D_i)\le d(D_{N'},D_j))$。

[0027] 根据历史数据集与目标数据集距离由近至远的排序队列π(1),...,π(N),对距离目标数据集较近的前f个历史数据集依次执行Relief-Cluster算法。首先通过Relief算法得到的每类超参的平均权重来初步评估超参重要性,然后利用聚类算法的r(C)指标进一步验证超参重要性评估的准确性,重复以上两步m次,选择r(C)指标最大时对应的超参重要性评估结果,最后得到待评估分类算法的超参重要性排序,转而用于指导目标数据集在待评估分类算法的自动化调参过程。

[0028] 所述通过Relief算法得到的每类超参数的权重包括:

[0029] 根据不同超参数配置下的性能数据大小设置阈值,将历史数据集中不同超参数配置对应的性能数据分为高性能样本和低性能样本,Relief算法首先从性能数据中随机选择

一个样本s_i，然后从性能高样本和性能差样本中各选择一个距离s_i最近的样本；

[0030] 与s_i同类的样本s_j用M表示，与s_i不同类的样本s_j用Q表示，每类超参数h的权重w_h根据公式(1)更新：

[0031] $w_h = w_h - diff(h,s_i,M)/rt + diff(h,s_i,Q)/rt$ (1)

[0032] $diff(h,s_i,M)$表示两个样本s_i与M在超参数h上的差异；

[0033] $diff(h,s_i,Q)$表示两个样本s_i与Q在超参数h上的差异；

[0034] 两个样本s_i与s_j在超参数h上的差异$diff(h,s_i,s_j)$定义为：

[0035] 若超参数h为标量型超参数，

[0036] $$diff(h,s_i,s_j)=\begin{cases}0, s_{ih}=s_{jh}\\ 1, s_{ih}\neq s_{jh}\end{cases} \quad (2)$$

[0037] 若超参数h为数值型超参数，

[0038] $$diff(h,s_i,s_j)=\left|\frac{s_{ih}-s_{jh}}{max_h-min_h}\right| \quad (3)$$

[0039] 其中，$1\leqslant i\neq j\leqslant m$，$1\leqslant h\leqslant ph$，$max_h$为超参数h在样本集中的最大值，$min_h$为超参数h在样本集中的最小值，m表示样本数，每个样本包含ph个超参数，rt表示迭代次数，$rt>1$，为了避免一次抽样的随机性；s_{ih}表示在样本s_i上超参h的值，s_{jh}表示在样本s_j上超参h的值。

[0040] 由公式(1)可知，对于高性能贡献大的超参数表现为在异类间差异大而在同类间差异小，因此具有区分能力的超参数的权值为正值。

[0041] 为避免一次抽样的随机性，迭代进行$rt>1$次，得到每类超参的重要性权重排序。

[0042] 所述利用聚类算法进一步验证超参数重要性评估的准确性包括：

[0043] 根据得到的每类超参数的重要性权重排序，对位于前k类的超参数进行聚类，并计算超参数重要性，假设超参数样本集为S，T为超参数样本集合的大小，K为超参数样本所属类的个数，p_{ik}表示样本隶属于类k的概率，C_k表示超参数样本的实际类标签，C表示超参数集，则在C的重要性度量r(C)表示为：

[0044] $$r(C)=1-\frac{1}{2}\left(\frac{1}{K}\left(\sum_{i=1}^{K}F_i(C)\right)+F(C)\right); \quad (4)$$

[0045] $$F(C)=\frac{\sum_{i\in S}\sqrt{\frac{1}{K}\sum_k(p_{jk}-c_{jk})^2}}{T} \quad ; \quad (5)$$

[0046] $$F_i(C)=\frac{\sum_{j\in S_i}\sqrt{\frac{1}{K}\sum_k(p_{jk}-c_{jk})^2}}{|X_i|} \quad ; \quad (6)$$

[0047] 其中，F(C)表示在超参数集C上聚类的结果与类标签在整个超参数样本集上的差异，C代表超参数集，$F_i(C)$表示在超参数集C上聚类的结果与类标签在各个类内的差异，X_i表示第i个类的超参数样本集合。

[0048] r(C)值越高，聚类结果与实际类标签之间的相关度越大，超参数集C对分类的影响越大。选择r(C)指标最大时对应的超参重要性评估结果。

[0049] 类标签是指性能高和性能低的标签。

[0050] 作为本发明的第二方面，

[0051] 机器学习超参数重要性评估系统，包括：存储器、处理器以及存储在存储器上并在处理器上运行的计算机指令，所述计算机指令被处理器运行时，完成上述任一方法所述的

步骤。

[0052] 作为本发明的第三方面，

[0053] 一种计算机可读存储介质，其上运行有计算机指令，所述计算机指令被处理器运行时，完成上述任一方法所述的步骤。

[0054] 本发明的有益效果：

[0055] 本发明可以准确评估机器学习算法的超参重要性，用于指导自动化超参配置以及增强超参配置的可解释性问题。用于描述机器学习算法本身的超参重要性，为超参配置过程提供有效借鉴和良好的可解释性。此模块着重解决的技术问题为如何准确评估机器学习算法的超参重要性，并将其用于指导自动化超参配置以及增强超参配置的可解释性问题。

[0056] (1)节约资源，节省时间，通过提供合适的先验知识，缩小搜索空间，使得超参配置过程具有一定的指导性，摆脱以往完全黑盒的状态。

[0057] (2)同时可以让用户直观的了解哪类超参数对算法性能影响更大。

附图说明

[0058] 构成本申请的一部分的说明书附图用来提供对本申请的进一步理解，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。

[0059] 图1为本发明提供的流程图；

具体实施方式

[0060] 应该指出，以下详细说明都是例示性的，旨在对本申请提供进一步的说明。除非另有指明，本文使用的所有技术和科学术语具有与本申请所属技术领域的普通技术人员通常理解的相同含义。

[0061] 本发明充分利用开放式机器学习环境OpenML中的多个数据集以及其每个数据集在多种算法下的性能数据，结合元学习方法计算目标数据集与历史数据集的距离，并利用Relief算法和聚类算法得到待评估分类算法每类超参数的重要性排序，排序结果转而用于指导目标数据集在待评估分类算法的自动化调参过程。本发明为提供合适的先验知识，缩小搜索空间，使得超参配置过程具有一定的指导性，摆脱以往完全黑盒的状态；同时可以让用户直观的了解哪类超参数对算法性能影响更大。

[0062] 如图1所示，本发明包括以下步骤：

[0063] 步骤A、获取OpenML中不同的数据集，并对每个数据集提取元特征，使得每个数据集都可以用元特征来表示，同时收集待评估分类算法在不同超参配置θ_i下性能y_i(例如，错误分类率或者RMSE)的数据$<\theta_i,y_i>_{k=1}^{K}$。并将每个数据集的元特征向量以及不同超参配置对应的性能数据存储于历史数据集样本库；

[0064] 在步骤A中提取的元特征主要包括:简单的元特征(例如，数据集样本数量，特征数量，类别数量，缺失值数量等)、数据集的统计元特征(例如，平均值，方差，距离向量的峰度等)、重要性元特征(例如在数据集上运行机器学习算法获得的性能等信息)这三大部分。

[0065] 步骤B、对于我们使用的目标数据集，我们也提取元特征来表示目标数据集，并基于不相似的数据集其使用算法的超参配置也具有差异这一原则，利用元特征向量之间的距离获得目标数据集与历史数据集之间的距离序列。对距离目标数据集较近的前f个历史数

据集,我们可以使用待评估分类算法不同超参的性能数据来评估超参重要性;

[0066] 在步骤B中,利用元特征向量之间的距离来衡量目标数据集$D_{N'}$与历史数据集D_i(i=1,2,…N)之间的距离,其中的距离公式我们使用的是衡量数据集元特征向量之间差异的常用p-范数:$d_{pn}(D_{N'},D_i)=||V_{N'}-V_i||_{pn}$。通过目标数据集与历史数据集元特征向量之间的距离比较,我们可以得到历史数据集与目标数据集距离由近至远的排序序列π(1),...,π(N),其中$(\pi(i)\leq\pi(j)) \Leftrightarrow (d(D_{N'},D_i)\leq d(D_{N'},D_j))$。

[0067] 步骤C、根据历史数据集与目标数据集距离由近至远的有序序列,对距离目标数据集较近的前f个历史数据集依次执行我们提出的Relief-Cluster算法。首先通过Relief算法得到的每类超参的平均权重来初步评估超参重要性,然后利用聚类算法的r(C)指标进一步验证超参重要性评估的准确性,重复以上两步m次,选择r(C)指标最大时对应的超参重要性评估结果,最后得到待评估分类算法的超参重要性排序转而用于指导目标数据集在待评估分类算法的自动化调参过程。

[0068] 在本发明中,步骤C具体包括以下步骤:

[0069] 步骤C1、我们根据不同超参配置下的性能数据大小设置阈值将数据分为性能高的一类和性能差的一类,Relief算法首先从超参样本集合中随机选择一个样本s_i,然后从两类样本中各选择一个距离s_i最近的样本。与s_i同类的样本用M表示,与s_i不同类的样本用Q表示,每类超参h的权重w_h根据公式(1)更新:

[0070] $w_h=w_h-diff(h,s_i,M)/rt+diff(h,s_i,Q)/rt$ (1)

[0071] 上述公式中,两个样本s_i与s_j(1≤i≠j≤m)在超参h(1≤h≤ph)上的差定义为:

[0072] 若超参h为标量型超参,

[0073] $$diff(h,s_i,s_j)=\begin{cases}0, s_{ih}=s_{jh}\\ 1, s_{ih}\neq s_{jh}\end{cases} \quad (2)$$

[0074] 若超参h为数值型超参,

[0075] $$diff(h,s_i,s_j)=\left|\frac{s_{ih}-s_{jh}}{max_h-min_h}\right| \quad (3)$$

[0076] 其中,max_h和min_h分别为超参h在样本集中的最大值和最小值。

[0077] 由公式(1)可知,对于高性能贡献较大的超参应该表现为在异类间差异较大而在同类间差异较小,因此具有区分能力的超参的权值应为正值。为避免一次抽样的随机性,上述过程迭代进行rt>1次。

[0078] 步骤C2、根据上步得到的每类超参的重要性权重排序,我们对位于前k类的超参进行聚类,并计算特征重要性,假设超参样本集为S,T为超参样本集合的大小,K为超参样本所属类的个数,p_{ik}表示样本隶属于类k的概率,C_k表示超参样本的实际类标号,C表示超参子集,则在C的重要性度量r(C)可以表示为:

[0079] $$r(C)=1-\frac{1}{2}\left(\frac{1}{K}\left(\sum_{i=1}^{K}F_i(C)\right)+F(C)\right) \quad (4)$$

[0080] $$F(C)=\frac{\sum_{i\in S}\sqrt{\frac{1}{K}\sum_k(p_{jk}-c_{jk})^2}}{T} \quad (5)$$

[0081] $$F_i(C)=\frac{\sum_{j\in S_i}\sqrt{\frac{1}{K}\sum_k (p_{jk}-c_{jk})^2}}{|X_i|} \quad (6)$$

[0082] 其中F(C)表示在超参集C上聚类的结果与类标签在整个超参样本集上的差异,C代表超参子集,$F_i(C)$表示各个类内的差异,X_i表示第i个类的超参样本集合。r(C)值越高,聚类结果与实际类标签之间的相关度越大,超参集C对分类的影响越大。

[0083] 对以上两步迭代m次,选取r(C)最大时对应的超参重要性排序,最后将得到的超参重要性排序结果转而用于指导目标数据集在待评估分类算法的自动化调参过程。

[0084] 本发明中Relief-Cluster算法的流程图:

[0085] 输入:超参数样本集S,超参数类别数hc,取样/迭代次数rt

[0086] 输出:聚类评价指标r(C),超参数重要性权重矩阵W

Begin:

For i←1 to m do

[0087] For i←1 to hc do

w_i=0;

For i←1 to rt do

[0088] 从S中随机选择一个样本s_i;

[0089] 从与s_i同类的样本中选择与s_i最近的一个近邻,记为M;

[0090] 从与s_i异类的样本中选择与s_i最近的一个近邻,记为N;

[0091] 采用公式(1)更新超参重要性权重向量W;

[0092] 选取大小为X的超参子集;

[0093] 在超参子集上对样本聚类;

[0094] 计算聚类结果与实际结果的相关度r(C)

[0095] 从m个r(C)中选取值最大时对应的超参重要性排序;

[0096] End

[0097] 以上所述仅为本申请的优选实施例而已,并不用于限制本申请,对于本领域的技术人员来说,本申请可以有各种更改和变化。凡在本申请的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本申请的保护范围之内。

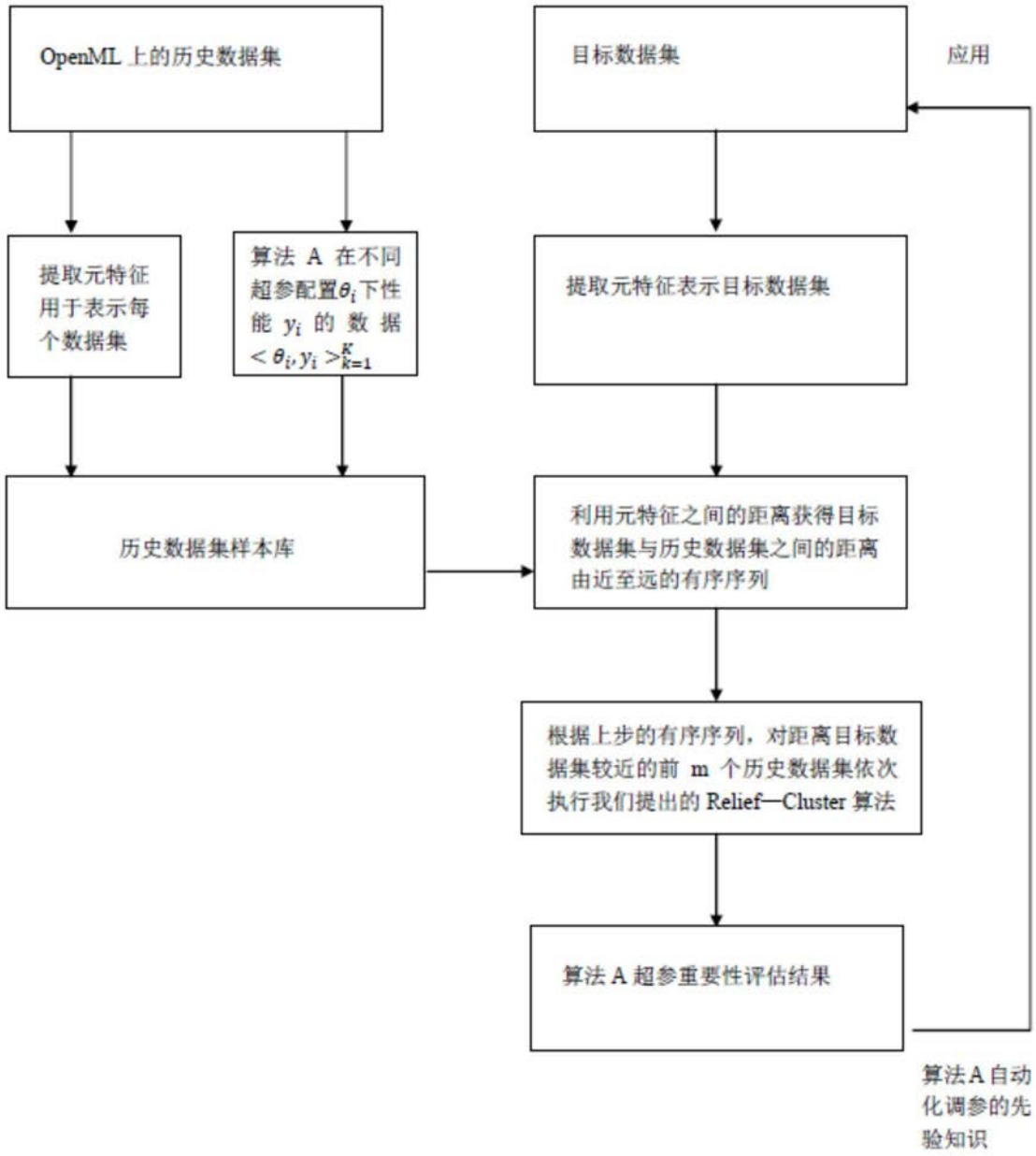
OpenML 上的历史数据集
目标数据集
应用
提取元特征用于表示每个数据集
算法 A 在不同超参配置θ_i下性能 y_i 的数据 $<\theta_i, y_i>_{k=1}^{K}$
提取元特征表示目标数据集
历史数据集样本库
利用元特征之间的距离获得目标数据集与历史数据集之间的距离由近至远的有序序列
根据上步的有序序列，对距离目标数据集较近的前 m 个历史数据集依次执行我们提出的 Relief—Cluster 算法
算法 A 超参重要性评估结果
算法 A 自动化调参的先验知识

图1