ALiPy：Python 中的主动学习

唐英鹏 tangyp@nuaa.edu.cn
李国祥 guoxiangli@nuaa.edu.cn
黄胜军* huangsj@nuaa.edu.cn
南京航空航天大学计算机科学与技术学院 模式分析与机器智能工信部重点实验室

中国南京 211106

抽象的

监督机器学习方法通 常需要大量标记示例来进行模型训练。然而，在许多实际应用中，无标签数据较多，但有标签数据有限；而且购买标签的成本很高。主动学习（AL）通过迭代地选择最有价值的数据来从注释器查询其标签，从而降低了标签成本。本文介绍了一个Python工具箱ALiPy[1]以便主动学习。ALiPy提供了基于模块的主动学习框架实现，允许用户方便地评估、比较和分析主动学习方法的性能。在工具箱中，学习框架的每个组件都有多种选项，包括数据处理、主动选择、标签查询、结果可视化等。此外还实现了 20 多种最先进的主动学习算法，ALiPy还支持用户在不同的主动学习设置下轻松配置和实现自己的方法，例如针对多标签数据的AL、带有噪声注释器的AL、具有不同成本的AL等等。该工具箱在 Github 上有详细的文档记录和开源[2]，并且可以通过 PyPI 轻松安装。

关键词： 主动学习、Python、工具箱、机器学习、半监督学习

一、简介

主动学习是利用有限的标记数据进行学习的主要方法。它试图通过主动查询最重要的示例来减少人类在数据注释上的工作（Settles（2009））。

ALiPy是一个用于主动学习的Python工具箱，适合各种用户。一方面，全过程主动学习得到了很好的落实。用户可以通过几行代码轻松进行实验，完成从数据预处理到结果可视化的整个过程。此外，工具箱中还实现了20多种常用的主动学习方法，为用户提供了多种选择。表 1 总结了 ALiPy 中实现的主要方法。另一方面，ALiPy支持用户高度自由地实现自己关于主动学习的想法。通过分解

*。通讯作者
1. http://parnec.nuaa.edu.cn/huangsj/alipy
2. https://github.com/NUAA-AL/ALiPy

ALiPy以低耦合的方式设计，将主动学习过程分解为多个组件，并相应地用不同的模块来实现，从而让用户可以自由配置和修改主动学习的任何部分。此外，除了传统的主动学习设置之外，ALiPy还支持其他新颖的设置。例如，数据示例可能是多标签的，预言机可能是嘈杂的，并且注释可能是成本敏感的。

表 1：在不同环境中实施主动学习策略。

具有实例选择的 AL	不确定性（Lewis 和 Gale (1994)）、委员会查询（Abe 和 Mamitsuka (1998)）、预期误差减少（Roy 和 McCallum (2001)）、随机、图密度（Ebert 等人（2012））、BMDR（ Wang 和 Ye (2013)))、QUIRE (Huang 等人 (2010))、LAL (Konyushkova 等人 (2017))、SPAL (Tang 和 Huang (2019))
多标签数据的 AL	奥迪（Huang 和 Zhou (2013)）、QUIRE（Huang 等人（2014））、MMC（Yang 等人（2009））、自适应（Li 和Guo（2013））、随机
通过查询特征进行AL	AFASMC（黄 等 等人。(2018)），稳定（Chakraborty 等人（2013）），随机
不同成本的 AL	HALC（Yan 和 Huang（2018）），随机，性价比
AL 与嘈杂的预言机	CEAL（Huang 等人（2017））、IEthresh（Donmez 等人（2009））、重复（Sheng 等人（2008））、随机
具有新颖查询类型的 AL	AURO（Huang 等人（2015））
适用于大规模任务的 AL	二次采样

2. ALiPy 中的模块

如图 1 所示，我们将主动学习实现分解为多个组件。为了便于在不同设置下实施不同的主动学习方法，我们基于多个模块开发了ALiPy，每个模块对应于主动学习过程的一个组成部分。

以下是 ALiPy 中的模块列表。

- alipy.data 操作：它提供了数据预处理和分区的基本功能。支持交叉验证或保留测试。
- alipy.查询策略：它由25种常用的查询策略组成。
- alipy.index.IndexCollection：它有助于管理标记和未标记示例的索引。
- alipy.metric:它提供了评估模型性能的多个标准。

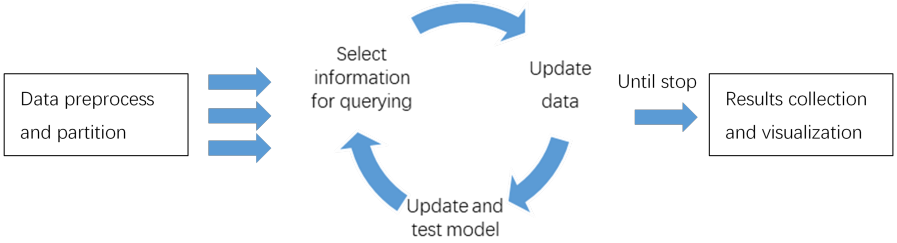

图 1：实施主动学习方法的总体框架。

- alipy.实验.状态和alipy.experiment.state io:它们有助于保存每次查询后的中间结果，并可以从断点恢复程序。
- alipy.experiment.stopping 标准它实现了一些常用的停止标准。
- alipy.oracle:它支持不同的oracle设置。人们可以设置多个具有嘈杂注释和不同成本的预言机。
- alipy.experiment.实验分析器：它提供了收集、处理和可视化实验结果的功能。
- alipy.utils.多线程：它提供了 k 次实验的并行实现。

以上模块均独立设计实现。不同部分之间可以不受限制地实施。这样，代码也就各自独立了模块可以替换为用户自己的实现（无需继承）。ALiPy中的模块不会相互影响，因此可以自由替换。

在每个模块中，我们还提供了高度的灵活性，使工具箱适应不同的设置。例如，在数据分割函数中，可以提供数据矩阵的形状或示例名称列表来进行分割。在oracle类中，可以进一步指定每个标签的成本，并在多标签设置中查询实例标签对。在分析器类中，对于成本敏感的设置，实验结果也可以不对齐，在绘制学习曲线时将自动执行插值。

更多详细信息请参考 http://parnec.nuaa.edu.cn/huangsj/alipy 的文档和 https://github.com/NUAA-AL/ALiPy 的 git 存储库。

3.ALiPy的使用

ALiPy为不同的用户提供了多种可选的用法。

对于不太熟悉主动学习并希望简单地将方法应用于数据集的用户，ALiPy 提供了一个类，封装了各种工具并实现了主动学习的主循环，即alipy.experiment.AlExperiment。

用户无需任何背景知识，只需通过该类几行代码即可运行实验。

对于想要通过实验评估现有主动学习方法性能的用户，ALiPy 提供了 20 多种最先进方法的实现，以及详细的说明和丰富的示例代码。

对于想要实现自己的想法并进行主动学习实验的用户，ALiPy提供基于模块的结构来支持用户修改主动学习的任何部分。更重要的是，还支持一些新颖的设置，使实现更加方便。我们还为每个模块和设置提供了详细的 api 参考和使用示例，以帮助用户快速入门。需要注意的是，ALiPy 并不强制用户使用任何工具类，它们是以独立的方式设计的，可以用用户自己的实现来替换，而无需继承任何东西。

有关详细信息，请参阅 ALiPy 主页和 github 上提供的文档和代码示例。

参考

阿部直树和间冢宏。使用 boosting 和 bagging 查询学习策略。在*第 15 届国际机器学习会议论文集*，第 1-9 页，1998 年。

Shayok Chakraborty、Jiayu Zhou、Vineeth Nallure Balasubramanian、Sethuraman Panchanathan、伊恩・戴维森和叶介平。主动矩阵补全。在*IEEE 第 13 届国际数据挖掘会议*，第 81-90 页，2013 年。

皮纳尔・唐梅兹、杰米・G・卡博内尔和杰夫・G・施奈德。高效学习准确率用于选择性抽样的标签来源。在*第15届ACM SIGKDD知识发现与数据挖掘国际会议论文集*，第 259-268 页，2009 年。

桑德拉・艾伯特、马里奥・弗里茨和伯恩特・席勒。RALF：强化主动学习公式-对象类别识别的灰化。在*IEEE 计算机视觉和模式识别会议*，第 3626-3633 页，2012 年。

黄胜军和周志华。用于增量多标签学习。*矿业*，第 1079-1084 页，2013 年。

不确定性和多样性驱动的主动查询*IEEE 第 13 届国际数据会议*

黄胜军、金荣、周志华。通过查询信息进行主动学习及代表性事例。在*神经信息处理系统的进展*，第 892-900 页，2010 年。

黄胜军、金荣、周志华。通过查询信息进行主动学习生动、有代表性的事例。*IEEE 模式分析和机器智能汇刊*，36（10）：1936-1949，2014。

黄胜军、陈松灿、周志华。多标签主动学习：查询类型很重要。在*第25届国际人工智能联合会议论文集*，第 946-952 页，2015 年。

黄胜军、陈家旅、穆欣和周志华。具有成本效益的主动学习-
来自不同的贴标商。在*第26届国际人工智能联合会议论文集*，第 1879-1885 页，2017 年。

黄胜军、徐苗、谢明坤、杉山正史、牛刚、松灿
陈。具有监督矩阵完成的主动特征获取。在*第24届ACM SIGKDD知识发现与数据挖掘国际会议论文集*，第 1571–1579 页，2018 年。

克谢尼娅・科纽什科娃、拉斐尔・斯尼特曼和帕斯卡・福阿。学习主动学习
数据。在*神经信息处理系统的进展*，第 4228–4238 页，2017 年。

大卫・D・刘易斯和威廉・A・盖尔。用于训练文本分类器的顺序算法。在
第 17 届国际 ACM-SIGIR 信息检索研究与开发年度会议论文集，第 3-12 页，1994 年。

李欣和郭雨红。具有多标签 SVM 分类的主动学习。在*会议记录*
第23届国际人工智能联合会议主旨报告，第 1479-1485 页，2013 年。

尼古拉斯・罗伊和安德鲁・麦卡勒姆。通过抽样实现最佳主动学习
误差减少的估计。在*第 18 届国际机器学习会议论文集*，第 441-448 页，2001 年。

B. 解决。主动学习文献调查。技术报告，威斯康星大学麦迪逊分校，2009 年。

Victor S. Shen、Foster J. Provost 和 Panagiotis G. Ipeirotis。拿另一个标签吗？即兴表演-
使用多个嘈杂的标签机来提高数据质量和数据挖掘。在*第14届ACM SIGKDD知识发现与数据挖掘国际会议论文集*，第 614-622 页，2008 年。

唐英鹏和黄胜军。自定进度的主动学习：询问正确的事情
在正确的时间。在*第33届AAAI人工智能会议论文集*，2019。

王政、叶介平. 查询批次的判别性样本和代表性样本
模式主动学习。在*第19届ACM SIGKDD知识发现与数据挖掘国际会议论文集*，第 158-166 页，2013 年。

严一凡和黄胜军。具有成本效益的分层多标签主动学习
分类。在*第27届国际人工智能联合会议论文集*，第 2962-2968 页，2018 年。

杨碧山、孙建涛、王腾蛟、陈正。有效的多标签活性
学习文本分类。在*第15届ACM SIGKDD知识发现与数据挖掘国际会议论文集*，第 917-926 页，2009 年。