Open-environment Machine Learning 问题综述

by zzh

概要

传统机器学习：假设在封闭环境场景中，学习过程中的重要因素保持不变。

越来越多涉及开放环境的实际任务（开放环境机器学习，Open ML），开始受到关注。在开放环境任务中，数据随着时间积累，类似于数据流，而像传统研究那样在收集所有数据后再训练机器学习模型变得十分困难。 本文简要介绍了该研究领域的一些进展，重点关注新兴类别、递减/增量特征、数据分布变化、学习目标多样化等技术，并讨论了一些理论问题。

1.简介

典型的机器学习任务，通过优化特定目标从训练数据集中学习一个预测模型，训练数据集由训练示例组成。一个训练示例包含两个部分：描述对象外观的特征向量（实例）和指示相应真实输出的标签。在分类中，标签表示训练实例所属的类别；在回归中，标签是对应实例的实数响应。本文主要集中讨论分类问题。

formally，考虑学习任务 f:X→Y,从训练数据集 D={(x1,y1),...,(xm,ym)}中学习，其中 x_i∈X 是特征空间 X中的特征向量，y_i∈Y 是给定标签集 Y 中的真实标签。

值得注意的是，当前机器学习大多基于假设在封闭环境场景下的任务，即学习过程中的重要因素保持不变。例如，所有要预测的类别标签都是已知的，描述训练/测试数据的特征从未改变，所有数据来自相同的分布，并且学习过程针对不变的唯一目标进行优化。图1展示了封闭环境机器学习研究中假设的典型不变因素。

特征不变 (feature invariant)：

用于描述事件/对象的特征向量不会改变，无论是训练阶段还是测试阶段，特征的结构和类型是固定的。

学习目标不变 (learning objective invariant)：

图的中心部分显示了从训练数据中学习一个模型，目的是学习特定的预测目标。这个目标在整个过程中保持不变。无论数据的大小或性质如何，模型的优化方向是固定的。

数据分布不变 (data distribution invariant)：

假设训练数据和测试数据来自相同的分布，模型的泛化能力可以直接应用于新数据。

类别标签不变 (class label invariant)：

在训练和测试数据中，类别标签 (✓ 和 ×) 不变。所有的类别在训练阶段是已知的，测试阶段使用的类别标签与训练时相同，不会发生新的类别。

封闭环境的假设提供了一种简化的抽象，使得复杂任务能够更容易地处理，但如今要求新一代能够处理学习过程中重要因素变化的机器学习技术，称为开放环境机器学习（Open-environment Machine Learning），或简称为开放学习（Open Learning）或 Open ML。

解决方案1：提前模拟可能的变化来人工生成大量训练样本，然后将这些数据输入到一个强大的机器学习模型中，例如深度神经网络。

问题：仅在用户对变化有所了解或至少能估计出变化的内容及其发生方式时适用。实际大数据任务中的数据通常是随着时间积累的，例如实例一个接一个地接收，像数据流一样，这使问题变得更加复杂。

解决方案2：仅用新数据对已训练的深度神经网络进行优化

问题：灾难性遗忘

解决方案3：频繁地基于存储的所有接收到的数据进行重新训练

问题：巨大计算和存储成本

解决方案4：持续学习（Continual Learning）,帮助深度神经网络抵御遗忘

问题：需多次扫描大批量训练数据并进行离线训练，面对大数据流的计算和存储问题仍然存在严重的顾虑。

最近关于开放机器学习的研究取得了相当大的进展。本文将简要介绍该研究方向的一些进展，重点关注新兴类别、递减/增量特征、数据分布变化和多样化学习目标等技术，同时也将讨论一些理论问题

2.新兴类别 Emerging New Classes

封闭环境的机器学习研究通常假设任何未见实例$\hat{x}$的类别标签必须是提前已知的给定标签集中的成员，即$\hat{y}$∈Y。

然而，现实中并非总是如此。例如，考虑一个机器学习模型辅助的森林疾病监控系统。显然，很难提前列举出所有可能的类别标签，因为有些森林疾病可能是全新的，比如由以前从未在该地区遇到的入侵害虫引发的疾病。能够处理$\hat{y} \notin$Y是开放环境机器学习（Open ML）的基本要求。

有人认为我们可以人为生成一些虚拟的训练样本用于新类别，就像在对抗深度神经网络中常用的训练技巧一样。这里的难点在于我们很难预见可能会出现什么未知类别（在后文中称为 NewClass），而训练一个可以处理所有可能类别的模型是不可能的，或者成本高得难以承受。

如果所有数据都在手中，那么处理 NewClass 可以被视为一种特殊的半监督学习任务。例如，可以为每个已知类别建立一个半监督的大间隔分隔器，对应于最紧密的轮廓线，然后将落在所有轮廓线之外的未标注实例视为 NewClass 实例。实际上，可以通过将未标注数据中的已知类别分布与未知类别分布分离，来近似估计 NewClass 的分布。然而，当数据是随着时间积累时，这类策略就不再直接适用了。(模型无法提前知道未来数据中可能会出现的新类别，因此无法在初始阶段建立准确的类别分隔。）

考虑如下的新兴类别学习场景：一个机器学习模型通过初始训练数据进行训练，然后被部署来处理数据流一样到来的未见实例。对于已知类别的输入实例，训练后的模型应该能够做出正确预测。对于未知类别的输入实例，模型应能够报告遇到一个 NewClass 实例；用户可以为 NewClass 创建一个新标签。在遇到几个 NewClass 实例后，训练后的模型应能够优化/更新，使得 NewClass 成为一个已知类别，并能够准确预测此类输入实例。理想情况下，整个过程不需要基于存储的所有已接收数据进行重新训练，因为这在实际的大数据任务中可能是极其昂贵的，甚至是不可行的。显然，上述描述的是一个有人工参与的无监督/监督混合任务。

新兴类别学习与已有学习方法的区别

零样本学习假设借助外部知识（如类别定义/描述/属性等侧信息）工作，这些信息可以帮助关联已见类别与未见类别，因此它可以被视为一种迁移学习；相比之下，新兴类别学习是一种不假设外部知识的一般机器学习设

定。换句话说，零样本学习假设未见类别是已知的，尽管它们没有出现在训练数据中，而学习新兴类别则应对的是未知类别这一重大挑战。因此，新兴类别学习的方法更为通用，且可以转换并应用于零样本学习。

开放集识别/分类扩展了拒绝选项（具有拒绝选项的分类旨在避免可能不正确的不确定预测，假设所有类别都已知。），考虑到在测试阶段可能会出现未知类别，其目标是识别已知类别并拒绝 NewClass。然而，它们并不尝试使训练后的模型能够处理 NewClass。一些广义开放集识别研究试图通过假设侧信息的可用性（如在零样本学习中提到的）来识别未知类别，而学习新兴类别则是一种不假设此类外部知识的通用机器学习设定。

新兴类别学习与增量学习的关系

新兴类别学习实际上是一种增量学习，强调对训练后的模型只需进行少量修改即可容纳新信息。增量学习的研究有着悠久的历史，大多数研究关注训练样本的增加，即 E-IL（样本增量学习）。增量学习的另外两种类型是 A-IL（属性增量学习）和 C-IL（类别增量学习）。A-IL 关注特征的增加，与本文第3节讨论的内容相关，尽管以往的研究通常致力于在给定所有数据/特征的前提下选择合适的特征空间。C-IL 关注类别的增加，与新兴类别学习相关，尽管以往的研究很少关注 NewClass（新类别）的识别，通常假设增量类别是已知的。

类别发现（Class Discovery）试图发现稀有类别，作为与类别预测分离的过程。如上所述，学习新兴类别是一个无监督/监督混合任务，而这些研究与其第一阶段（主要是无监督部分）有些相关。

新兴类别学习的一般解决方案

第一阶段——NewClass 的识别

通过异常检测实现。这里的挑战在于区分 NewClass 数据与已知类别的异常。通常情况下，这并不总是可行的。幸运的是，在许多实际任务中，可以合理假设 NewClass 实例比已知类别的异常"更异常"。如果在原始特征空间中这种假设不成立，我们可以通过核映射或表示学习识别合适的特征空间。之后，NewClass 实例的识别可以简化为流中的异常检测问题，这可以通过孤立森林(isolation forest)等方法解决。

第二阶段——在不牺牲已知类别的性能下优化训练好的模型以容纳 NewClass

对于深度神经网络，为避免灾难性遗忘，通常需要基于所有数据（或至少是精心选择的子样本）进行重新训练，这会带来巨大的计算和存储成本。理想情况下，只需对 NewClass 相关的局部区域进行微调，而不是进行可能严重影响已知类别的全局更改。一种解决方案是利用树/森林模型的优势，仅增量方式修改包含 NewClass 的树叶节点，这甚至不需要存储已知类别数据。其他替代方法包括可以局部化不同类别影响的技术，例如基于全局和局部概述的方法，以便 NewClass 的变化不会显著影响已知类别。

特殊情况处理

有多个新类别

可以利用 NewClass 数据的聚类结构。需要注意的是，NewClass 第一次被检测到的时间点与模型优化完成的时间点之间通常存在较大间隔。为了缩短这一间隔，已经有一些努力致力于基于较少的 NewClass 数据进行模型更新。对于包含新类别的多标签学习更具挑战性，因为在这种情况下，NewClass 实例可能也拥有已知类别标签，甚至可能出现在已知类别的密集区域中，这时关键在于检测特征组合和/或标签组合的显著变化。一个相关的研究方向是检查哪些已知类别与 NewClass 紧密相关，且已有一套关于从 NewClass 映射到已知类别的评估方法被开发出来。

训练数据中出现NewClass实例，但由于特征信息不足而被误认为是已知类别实例

这种情况更加复杂，只有一项非常初步的研究曾对此进行探讨。

3.递减/增量特征

封闭环境的机器学习研究通常假设所有可能的实例（包括未见过的实例）都位于相同的特征空间中，即$\hat{x} \in X$。然而，这并不总是成立。例如，在第2节提到的森林疾病监控中，由于某些传感器的电池耗尽，它们可能无法继续发送信号，导致特征减少；而新的传感器可能被部署，导致特征增加。能够处理$\hat{x} \in \hat{X} \not ={X}$是开放环境机器学习的另一个基本要求。需要注意的是，与新兴类别相比，只需要对新类别进行特殊处理，而消失的类别可以简单忽略；但对于递减和增量特征，必须同时给予关注，因为特征的递减可能会严重降低模型的性能。

考虑以下场景：一个通过初始训练数据训练出来的机器学习模型，被部署来处理像数据流一样到来的未见数据，其中可能包含递减和/或增量特征。对于输入的测试数据，模型应能够做出正确的预测；对于输入的新增训练数据，模型应能够相应地进行优化。理想情况下，整个过程不应需要基于已接收的所有数据进行重新训练。

通常情况下，构建一个能够从X中受益且适用于$\hat{x} \in \hat{X} \not ={X}$ 的机器学习模型并不总是可能的，因为机器学习是通过经验来提高性能的，而在大多数情况下，X中的学习经验对$\hat{X}$中的学习可能帮助甚微，尤其当 $X ∩\hat{X} =∅$ 时。例如，图3(a)展示了一个情景，如果第1阶段数据的特征空间（即 { (x 1 , y 1) , ⋯ , (x T 1 , y T 1) } {(x 1 ,y 1),⋯,(x T1 ,y T1)}）和第2阶段数据的特征空间（即 { (x T 1 + 1 , y T 1 + 1) , ⋯ , (x T 2 , y T 2) } {(x T1+1 ,y T1+1),⋯,(x T2 ,y T2)}）完全不同，那么在第1阶段训练的模型对第2阶段毫无帮助，必须基于第2阶段的特征集 S 2 S2 从头开始训练一个新模型。