DOI:10.19609/j.cnki.cn10-1339/tn.2018.05.014

开源软件缺陷预测技术与迁移学习

文/任俊桦 刘峰

开源软件对于商业来说有许多强有力的竞争优势和价值，随之而来，开源软件的缺陷预测也成为当下软件工程领域中的一个研究热点。

开源软件的低成本性和灵活性，使得更多的IT开发企业和行业领域越来越倾向于使用开源软件构建产品或系统，随之带来了软件质量保障问题。采用机器学习技术预测软件缺陷已成为提高软件质量的重要途径，但仅局限于同项目历史训练数据完整的情况下。

面对一个全新的或历史数据稀缺的项目，迁移学习方法可用源项目的相关知识来为目标项目构建预测模型，有效利用其他项目或领域已有的训练数据来构建缺陷预测模型，并迁移和应用到另一个项目中，其技术挑战是，由于不同项目之间的应用领域、开发流程、编程语言、开发人员经验等并不相同，如何在数据集间存在较大的分布差异性的条件下提升缺陷预测方案的实际性能。本文重点考察了面向开源软件的缺陷预测技术以及正在兴起的迁移学习方法。

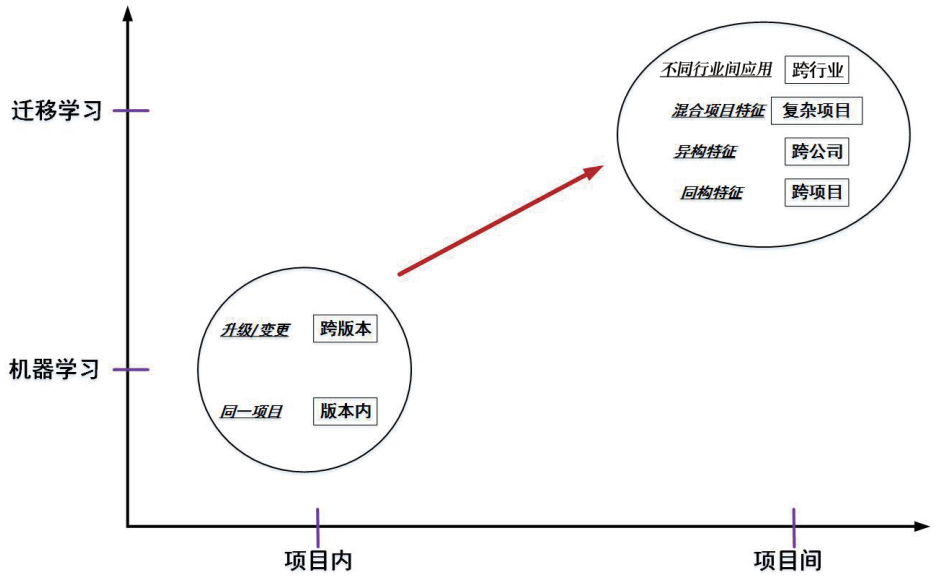
北京交通大学软件评测实验室刘峰

由于开源软件的开放性，客户可以在开源软件上开发定制来满足自己的要求。并且使用者和开发者能共同参与使用和测试，发现其中的缺陷并及时修正，从而使软件更加稳定与可用，所以开源软件对于商业来说有许多强有力的竞争优势和价值，随之而来，开源软件的缺陷预测也成为当下软件工程领域中的一个研究热点。

软件缺陷是影响软件质量的首要因素，人们从事各项活动的效率和安全在很大程度上依赖于软件的质量和系统的可靠性，而隐含缺陷的软件在部署后可能会产生意料之外的结果或行为，严重的时候会给企业带来巨额的经济损失，甚至有时候会引发人员伤亡。

软件缺陷预测是软件测试中的一项重要工作，本文首先对软件缺陷预测技术现状进行了概述，其中包括基于机器学习的项目内软件缺陷预测和跨项目软件缺陷预测，其次面临跨项

图1 软件缺陷预测技术演进



目预测遇到的度量元差异问题，重点阐述了迁移学习用于软件缺陷预测中相应的解决方法。

软件缺陷预测概述

软件缺陷预测主要是根据历史数据来预测软件中潜在的缺陷，具体流程是对项目的历史数据进行统计分析，从中抽取程序模块并进行类型标记，随后通过分析软件代码的内在逻辑或开发过程特征，设计出与软件缺陷相关的度量元，并借助这些度量元可以构建出用于模型训练的缺陷预测数据集，最后基于特定的研究方法构建出缺陷预测模型，并用于对项目中的新程序模块进行预测，挖掘软件缺陷的分布规律，并应用于实际的软件缺陷预测中。

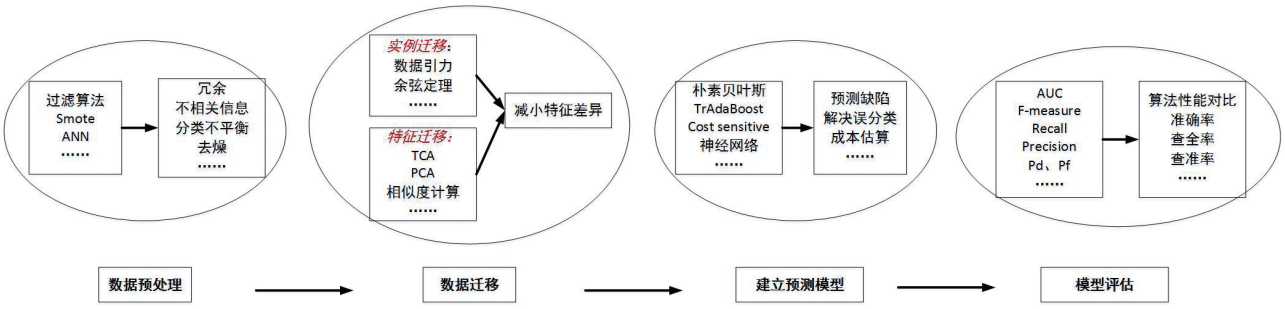
软件缺陷预测希望能够在项目开发的早期阶段，预先识别出项目内的潜在缺陷程序模块，并对这类程序模块分配足够的测试资源以确保可以进行充分的代码审查或单元测试，最终达到提高软件产品质量的目的。

另外，软件缺陷与软件演化过程是紧密相关的，同一软件的不同版本间的缺陷也会存在一定关联，在软件投入使用后，由于缺陷修复、需求增加和环境变化等，维护人员需要对软件进行变更以满足新的需求，或者为了提高软件的性能和可靠性进行改进等，这时候需要对项目进行跨版本预测。

作为人工智能和数据科学的核心，机器学习已成为当今发展最快的技术领域之一，致力于研究如何利用经验来提高系统自身的性能，采用机器学习的基础理论、核心算法和关键技术，解决软件缺陷预测中的实际问题，不仅能够提高软件缺陷预测能力，更重要的是能够有效提高软件质量和软件可靠性。

传统机器学习预测软件缺陷的大部分研究工作都集中关注同项目缺陷预测问题，均是假设训练数据以及测试数据来自于同一个项目，二者的特征和样本分布特性是一致的，即同项目的部分数据集作为训练集来构建模型，

图2 迁移学习预测软件缺陷流程



并用剩余的数据作为测试集来获得模型的预测性能，但在实际的软件开发场景中，需要进行缺陷预测的目标项目可能是一个新启动项目，或这个项目已有的训练数据较为稀缺。

开源软件缺陷的评测数据集

目前，大部分的开源软件缺陷预测研究的数据集已公开共享，分别来自于航天航空软件、电器行业软件、开源项目（例如Ant、Eclipse等）等累积的软件缺陷历史数据，这些数据集均可以在PROMISE（http://openscience.us/repo）中进行下载。

跨项目缺陷预测

目前，如果要对一个全新的项目进行缺陷预测，但是此项目中的数据不充足，缺陷信息缺乏，训练集不足而无法构建预测，然而却有大量的其他相关的项目缺陷信息数据，但是此训练数据与所需进行的分类任务中的测试数据特征分布不同，在这种情况下，研究者提出了跨项目缺陷预测方法，即将不同项目间的缺陷信息共享，通过在一个有足够历史缺陷信息的项目（源项目）上进行训练，并将得到的预测模型应用到另一个项目（目标项目）进行测试。

但跨项目缺陷预测所存在的问题是，不同项目的特征（例如所处的应用领域、采用的开发流程、使用的编程语言或开发人员的经验等）并不相同，所以源项目与目标项目的缺陷数据集的度量元存在很大的特征和分布差异，因此在缺陷预测模型构建时，如何从源项目中迁移出与目标项目相关的知识是其面临的研究挑战。

迁移学习与软件缺陷预测

针对新项目缺陷数据集的匮乏问题，大部分研究借助机器学习领域中的迁移学习方法来构建软件缺陷预测模型，迁移学习是运用已有知识，对具有一定相关性的领域的问题进行求解的一种机器学习方法，其目的是迁移已有知识来解决目标领域仅有少数已标记实例甚至没有的问题。

跨项目缺陷预测问题可以视为迁移学习在软件缺陷预测领域中的一个重要应用。采用合适的迁移学习方法则可以大大提高样本不充足任务的分类识别结果。

图1展示了软件缺陷预测技术中，从传统机器学习到迁移学习应用的方法演进过程。

迁移学习缺陷预测的流程主要分为四大部分：

①数据预处理：对项目数据（源项目或目标项目）进行筛选、去噪，去除无关样本，解决跨项目中出现的类不平衡、不相关信息、冗余信

息等问题；

②数据迁移：对筛选后的项目数据（源项目或目标项目）进行相应的特征或实例数据迁移，使目标项目数据更贴近源数据的分布。

③建立预测模型：基于机器学习的算法，对迁移后的数据建立缺陷预测模型。

④模型评估：利用相应的算法评估机制，对建立的缺陷预测模型进行评估，验证所使用的模型算法性能是否可用。

图2展示了迁移学习预测软件缺陷的流程。

面向复杂软件缺陷

随着人们对软件需求的日益增加，软件开发过程会越来越复杂，软件规模和逻辑程度也持续增长，比如新型的复杂系统中的大数据、云平台等，面对此种情况，单个源项目软件数据很难充分反映一个复杂目标项目的缺陷特征，可用多源异构迁移学习跨项目的方法来预测软件缺陷。

使用不同公司的数据作为源数据，来预测目标公司或目标项目未标记的软件模块，由于源数据和目标数据是从不同的公司和项目收集的，因此它们有不同的数据分布情况，所以后面的缺陷预测模型建立中，依然会用之前的特征迁移、实例迁移等方法。

HDP（异构迁移缺陷预测）方法包括特征选择和特征映射两个阶段，具体来说：首先借助特征选择方法从源项目中选出与类标强相关的特征. 随后借助特征映射方法将为源项目选出的特征与目标项目的特征进行映射，最后基于映射的特征构建缺陷预测模型。

借助典型相关分析（CCA）方法，来减少源项目与目标项目数据集分布间的差异程度，通过为源项目和目标项目寻找一个共有空间，使得投影到该共有空间的两个数据集间的相关性最大化，结果表明：若源项目与目标项目之间共有的度量元数越多，该方法的性能越好，同时他们也发现多对1项目的方式要优于1对1项目方式。

北京交通大学软件评测实验室任俊桦

面向跨行业迁移软件缺陷预测

目前关于迁移学习与跨行业软件缺陷预测的研究还不多，但是跨行业的迁移学习已有了大量研究，主要在语音识别、文本分类、图像识别等学科的研究较多。跨行业大致的思路是寻找提取源行业项目和目标行业项目的共性特征，尝试去建立一个在训练集和测试集都适用的模型。

关于未来的展望

对于一些新启动的软件项目或大部分中小规模企业来说，搜集充足的高质量缺陷预测数据集较为困难，研究人员已经通过挖掘开源项目，搜集了很多高质量的缺陷预测数据集，并共享到Promise库中。

针对跨项目软件缺陷预测问题的研究具有丰富的理论研究价值和工业界应用前景，虽然国内外研究人员已经取得了一定的研究进展，但我们认为，开源软件的迁移学习跨项目缺陷预测研究仍然是今后软件缺陷预测研究中值得关注的一个开放性研究课题。