1. **绪论**
   1. **研究背景和意义**

A Ranking’s background(Introduction)

近年来，市面上的各种各样的软件程序以及应用软件系统极大地丰富了我们的生活，在不断满足人们日益增长的物质精神需求的同时，也带来了诸多不可避免的问题。几十年来，软件行业得到了迅猛的发展。无论是从软件的技术还是设计方面来看，软件都已经有了长足的进步和发展。在未来的生活里，软件行业依然会是各行各业中不断壮大的一支队伍，软件行业也会更好地服务于经济、军事、政治以及人民的娱乐等多个方面。一旦软件发生错误，将会造成非常严重的后果。因此，在软件的开发过程中，保证软件的质量就需要采取一系列的措施。

为了保证软件的质量，软件缺陷预测就成为软件开发过程中必不可少的一个环节。软件缺陷预测首先通过挖掘和分析软件的历史仓库，从中提取程序块并且进行某种类标记。之后，通过分析软件源代码的内在复杂度特征和开发过程的联系，设计出和软件缺陷存在强相关性的度量元，并且对这些程序块进行度量。软件缺陷预测技术仅仅凭借模块的属性值就可以对软件模块是否存在缺陷进行预测。如果目标任务中仅仅含有少量高质量的历史数据，那对于缺陷的预测将造成极大的麻烦。当前项目的软件缺陷预测也不能顺利进行。

由于软件缺陷预测在软件质量保证以及维护领域的重要性，缺陷预测在软件工程的研究领域渐渐成为最活跃的话题。根据我们的了解，目前存在主流的两种技术来应对软件缺陷预测：项目内缺陷预测和跨项目缺陷预测。

一般来说，项目内缺陷预测模型有一个致命的弱点就是该方法往往需要足够的高质量的历史数据，并且这些数据需要来自同一类项目。因此，应用项目内缺陷预测模型来解决那些全新的不活跃的项目，将发生冷启动现象，导致该模型根本无法使用。随着互联网上可获得的标签化的缺陷数据的数量不断上升，到现在，跨项目缺陷预测已经成为了该领域最热门的技术。即便该技术与项目内缺陷预测技术的比较中，经常因为低下的预测水平而遭到抨击。

先前有大量的针对与跨项目缺陷预测的研究都是将软件缺陷预测当成二分类问题看待[1-8]，它们用于提高预测准确度的主要方法是使用大量的机器学习技术[4,6-8],例如特征选择，降维分析以及数据采样。然而，估计给定类别或者软件模块的缺陷倾向性对软件测试和软件维护中的实际活动的影响有限，特别是当缺乏人类的资源。从软件开发者的角度来看，缺陷倾向软件实体的排名次序肯定要比关于有多少软件实体可能存在缺陷的信息更有价值。

假设有如下的一个典型的应用场景：有两个软件开发人员使用Java语言共同开发一个全新的工程项目，其中一个开发人员(称为A)负责软件的测试。由于新版本(由1000多个Java类组成)严格的截止期限，A的一个健全的技术解决方案是在执行单元测试之前确定最有可能出现缺陷的版本类。因此，A根据常用的软件度量标准，使用朴素贝叶斯分类算法创建了一个跨项目缺陷预测模型。根据其他类似的成熟项目培训的跨项目缺陷预测模型的预测结果，类别存在问题的可能性非常低，大约只有7%，但是整个测试过程中的效率仍然相对较低，因为A必须执行一个随机的测试。在这种情况下，A实际上更愿意获得一个类别的排名表，来确定每个有缺陷类的优先级，从而制定出一个具有成本效益的计划。

数据搜集 训练模型

可用仓库 带标记缺陷数据 缺陷预测模型

(a) 传统跨项目缺陷预测模型的结构化流程

I1 预测 I1 I2 ... In

的I

的I8 I

I8

的I

缺陷预测模型

I2 YES YES NO

二分类预测结果

I3

.

的I

I8

I8

. 预测 I8 I2 ... I1

. 本文方法

In 1 2 n

排序预测结果

测试集合

(b)传统跨项目缺陷预测与本文方法比较

图1. 展示两种不同的缺陷预测方法

跨项目的软件缺陷预测是由于目标项目中往往缺少足够的高质量历史数据，在业界逐渐流行起来。从先前的研究来看，软件缺陷预测被看成是一个二分类问题或者回归问题。然而，对于那些有限的人力和时间的项目而言，使用现有的方法来解决这个问题可能并不是非常合适。同时，运用跨项目数据的所有样本来构建模型并非就可以得到一个具有足够高性能的模型，使用和目标域数据不相似的源域数据建立的模型通常具有很低的可用性。本文专注于面向排序的预测缺陷实体。图1.展示了他们之间的区别。简而言之，本文的主要目标是为跨项目缺陷预测提出一种基于排序的方法(称之为ROCPDP)，该方法可以根据优先级排列有问题的软件实体。**(Need revisit)**此外，我们希望通过基于公开数据集AEEEM[12]的五个开源项目的案例来研究验证ROCPDP的可行性。

* 1. **国内外研究现状**

本小节简要回顾现存的缺陷预测方法。这些方法大致上可以分为二类：1、通过分类技术预测软件模块是否缺陷，2、通过回归技术预测软件模块缺陷的数目。这里将以上二类以以下三种方式讨论。

* + 1. **本项目软件缺陷预测方法**

本项目软件缺陷预测是软件工程数据挖掘领域中一个研究的热点。[静态软件缺陷预测方法研究-陈翔]该方法通过分析软件代码或者开发过程设计出与软件缺陷相关的度量元。通过挖掘软件历史仓库(software historical repositories)来创建缺陷预测数据集。目前，可以挖掘与分析的软件历史仓库包括项目所处的版本控制系统(例如SVN，CVS，Git等)、缺陷跟踪系统(例如Bugzilla，Mantis，Jira或Trac)或相关开发人员的电子邮件等。最后，对搜索到的数据进行清洗得到可用缺陷预测数据集，并构建缺陷预测模型。应用模型对目标项目进行缺陷预测。

Akiyama等人[付-13]最早提出了软件缺陷预测方面的量化关系式，研究计算并给出了软件缺陷与代码行数之间的关系式：D = 4.86+0.018L，即代码行数与软件缺陷个数呈线性关系。这种量化关系式子最早是基于人们的经验估计得到。当时普遍认为平均每完成60行代码就会出现一个软件缺陷。Akiyama通过统计分析早期软件缺陷与代码行数的内在联系，提出了该量化关系式。在软件工程领域最早应用于估计软件缺陷数量。

支持向量机[yx-25-27]、神经网络[yx]、决策树以及贝叶斯等方法都是缺陷预测领域的分类方法。这些方法使用软件度量指标来预测软件模块是否存在缺陷。但是，数据集中缺陷倾向和无缺陷类别的高度不平衡性质降低了预测性能。这些问题也是本项目软件缺陷预测领域最常涉及到的问题。因此，现已提出特征选择和类别不平衡学习方法来解决以上问题。比如，Wang 等[improving cross-17]利用深度信赖网络从源代码中自动学习语义特征。

在缺陷预测过程中，对不同软件缺陷的误分类现象可以被分为两类：类1和类2。两类在误分类情况下的代价是不同的。因此，学者就提出一些成本敏感的学习方法，通过生成一个最有最小误分类代价的类型模型来处理这个问题。此外，另一些基于集成学习的方法[improving cross-20，21]也用于处理缺陷预测问题。然而，这些方法仅限于预测提供的软件模块是否存在缺陷。

传统的软件缺陷预测技术均是以本项目的历史数据作为训练基础，借助不同的方法模型来进行软件缺陷预测。但实际情况中，往往缺少足够的高质量的历史数据来构建模型，这种现象在新成立的项目中最为常见。为了解决这一问题，人们提出了跨项目软件缺陷预测，利用相似的有价值的源项目的历史数据来构建模型，预测目标项目的缺陷。

* + 1. **跨项目软件缺陷预测方法**

为了解决新的公司或者新的项目的历史数据太少而无法完成本项目缺陷预测的问题，跨项目缺陷预测(CPDP)应运而生。Zimmermann等[improving 25]在12个真实应用数据集上研究CPDP模型，由于源项目数据集和目标项目数据集中间存在分布上的差异，导致CPDP仍然存在严峻的挑战。为了减少源与目标项目之间的分布差异，这里提及三种主流的方法。

1. 利用数据过滤方法寻找最合适的训练数据(improving 23 24 26)。例如Turhan等[23]提出NN过滤技术来筛选训练数据集。Peter等[24]提出Peters过滤器，通过其他项目的结构来选择训练数据。他们将过滤器与其他两种质量预测方法进行比较，以评估Peters过滤器的性能。结果显示①本项目缺陷预测对于小数据集的表现较弱。②Peters过滤器配合CPDP构建了更好和更有用的预测器。Kawata等[26]提出并检验了基于其中一种高级聚类算法的数据过滤方法---基于密度的空间聚类(DBSCAN)。我们提出的方法也属于这一种。
2. 第二个主流的方法是设计基于迁移学习技术的高效的缺陷预测器([22,27,28])。例如，Ma等[27]提出了一种新的被称为迁移朴素贝叶斯(TNB)的新算法，将跨项目数据信息转化为训练数据的权重，然后基于重新加权的跨项目数据构建新的预测器。结果表明TNB在在工作特征曲线下的面积(AUC)更准确，运行的时间比最先进的方法还少，并且可以有效地实现跨项目缺陷预测任务。Chen等人[28]提出了双重迁移提升模型(DTB)。另一个CPDP的挑战在于源项目数据和目标项目数据之间的度量集通常是异构的。异构跨项目缺陷预测(HCPDP)任务是指源数据和目标项目数据之间是异构的。Jing等人[22]为HCPDP提供了有效的解决方案。他们提出了源和目标项目数据的统一度量表示(UMR)，并将典型相关分析(CCA)（一种有效的迁移学习方法）引入到CPDP中，以使源项目和目标项目的数据分布相似。结果表明他们的方法明显优于部分不同指标的HCPDP和具有完全不同指标的HCPDP和现代CPDP方法，其方法是有效的。
3. 应用无监督分类器，不需要任何的训练数据来执行CPDP(参见 [29 30])，因此训练项目数据与目标项目数据之间的分布差异不再成为问题。例如，Zhang等人[30]提出应用基于连接的无监督分类器，该分类器基于谱聚类来执行CPDP。
   * 1. **面向软件排序的软件缺陷预测方法**

近年来，一些研究人员试图从新的角度研究软件软件缺陷预测。他们将这个问题作为排序问题而不是二元分类问题或者回归问题。[A ranking-O9,21]。因此，最近在信息检索领域中引入了学习排序(LTR)[26]的方法来研究软件缺陷预测。但是，上述研究是基于这样的一个假设：缺陷倾向软件实体的等级与其包含缺陷的实际数量成正比。此外，他们也使用常规的验证(将项目中的数据集以8/2划分成训练集和测试集)或10折交叉验证在WPDP项目中进行。因此，在以前的研究中提出的这些方法是否可以用于实际的CPDP任务仍然是一个未知数[9,21]。

提出以排序方式应对CPDP的目的是预测哪些模块最有可能存在大部分缺陷以便分配软件质量改进工作[A-learning-to-rank7,10,11]。也就是说CPDP排序任务的目标是预测相对缺陷数量，尽管估算模块缺陷的准确数量要好于估算模块的排序，因为缺陷的准确数量能提供更多有价值的信息。无论如何，由于实践中总是缺乏高质量的数据，要预测模块缺陷的准确数量是有困难的，甚至是不可能的。实际上，对于哪些试图明确预测软件模块中缺陷数量的现有的方法，他们总是使用这些预测数字对模块进行排序，以指导软件质量负责的团队针对最有可能产生错误的模块[3,10,11]。因此，Ohlsson[13]提出了Alberg图[13]（也称为累积提升图CLG[6]）作为性能指标，Weyukers等人[10]应用错误百分比平均值(FPA)和前20%模块中缺陷的百分比来评估预测模型。

换句话说，软件模块的顺序是根据它们的缺陷，而不是用于指导CPDP测试资源的分配任务的特定缺陷编号。在这种情况下，直接了解软件模块的排序更为自然，不是预测每个模块中缺陷的准确数量，而是使用这些数字对模块进行排序。实际上，对于排序任务的CPDP，具有较高预测准确度（较小的平均绝对误差AAE或平均相对误差ARE）的模型可能会给出较差的排序。

相较于CPDP中的二分类器而言，CPDP对排序任务的研究是很有限的[A-learning ranking11]。Ohlsson和Alberg[13]使用线性回归方法构建CPDP模型。他们使用这些模型来预测编码开始之前软件模块中的缺陷数量。 Ohlsson等人[13]应用Alberg图（与CLG图相同）和顶层模块中缺陷的百分比来评估CPDP模型，并证明Alberg图的有效性。

Ostrand等人[A-learning-ranking-3]应用负二项回归NBR和基于LOC非常简单的模型来构建模型，以预测下一版大型商业系统中每个模块的预期缺陷数量，并利用前20%模块中缺陷的百分比来评估CPDP模型。NBR取得比简单模式更好的结果。然后，从他们的实际缺陷模型的图表来看，改进CPDP模型仍然存在很大的进步空间。Ostrand等人还使用预测模型根据缺陷密度给出模块的顺序。

* 1. **本文工作**

以下是文本工作的总结：**(Need revisit)**

1. 将跨项目缺陷预测看做一个排序问题，应用迁移学习中TCA/TCA+算法，同时结合经典排序算法Ranking SVM，完成了自动对原始数据的正规化、筛选出与目标项目最相似的源数据集并导入排序器中，在开源项目AEEEM基础上进行实验。查看不同正规化方法之间对于缺陷预测结果的表现力。
2. 文章基于AEEEM进行实证研究，结果表明，在不同的跨项目缺陷预测情境下，新的方法在两种常用评估指标方法的表现优于其他基准方法。例如，Spearman相关系数合Kendall等级相关系数。
   1. **本文组织结构**

本文以大量数据挖掘，机器学习文献为基础，结合必要的数学方法，完成了跨项目缺陷预测中原始数据的正规化，利用TCA/TCA+算法目标数据的筛选，利用现有的排序器对预处理后的数据进行实验，最后得到评估结果，以验证其可行性和有效性。

论文的具体安排如下：

1. 绪论。介绍了本文的研究意义以及国内外研究现状。简述研究方案以及论文组织结构。
2. 相关工作以及算法介绍。简单介绍了迁移学习的概念，详细介绍了TCA/TCA+算法以及排序器Ranking SVM。介绍基于以上运用到跨项目软件缺陷预测的可行性以及实验构建的流程。
3. 实验相关工作介绍。首先介绍了数据集的来源以及实验环境，然后运用正规化对原始数据进行预处理，再应用TCA/TCA+算法筛选出与目标数据集最相似的降维后的源数据，最后导入Ranking SVM排序器中进行实验。结果与不应用TCA/TCA+算法进行比较，验证其有效性，同时分析了基于不同训练集规模的实验，以分析运用多少比例的目标项目数据加入训练集能最大地提高模型的性能。
4. 总结全文，并介绍了对未来工作的期望及方向。
5. **相关理论及算法**
   1. **迁移学习理论**

传统的机器学习有两个基本的假设：①训练样本和测试样本的分布一致；②训练样本的数据量要足够多才能得到一个良好的模型。然而在多数情况下，这两个假设并不能同时成立。考虑推荐系统中经常遇到的冷启动问题，当一个新任务成立时，并没有足够的良好标记的数据提供给训练模型。有时会用人工手动标记数据的方式代替这一过程以完成训练的需要。但这一过程相当繁复，不仅会花费大量的人力和物力，同时效率底下，其科学性也根本得不到保证。

针对以上情况，人们根据日常生活经验，自然会有如下的问题。能不能利用从相关领域 学到的知识来帮助目标领域的学习，即使当前领域只有少数的训练数据？或者，怎样利用相关领域的知识来加速目标领域的学习过程？对以上问题的回答，产生了一个值得研究的领域-----迁移学习(Transfer Learning)。

传统机器学习从一个任务的训练数据中挖掘知识，并应用到该任务的测试数据中。而迁移学习不仅要从若干个源任务中学习知识，还需要通过某种机制，将这些知识应用到目标任务中去。

常见的迁移学习技术一般可以分为四类：

1. 基于实例。对来自源领域的标注数据进行加权后，再用于目标领域的学习任务中[陈德品-6,20,22,29,37,39,47]。
2. 基于特征表示。寻找隐含于源领域和目标领域之间的，分歧最小的一个特征表示[2,3,4,7,8,21,24,38]。
3. 基于模型。搜索源领域和目标领域的模型之间共享的参数或先验知识[9,27,33,44,69]。
4. 基于关系知识。建立源领域和目标领域之间关系知识的映射[25,51,52]。

本文所用到TCA/TCA+算法是基于特征表示的迁移学习算法。基于特征表示的迁移学习的直观想法是为了目标领域学习一个“好”的特征表示。源领域的知识通过特征表示迁移到目标领域，在该特征表示下，目标领域的学习性能能够得到明显的提升。

* 1. **排序学习算法及评价指标**
     1. **排序学习算法**

排序学习在学术界有一个专门的术语------learning to rank，它是一个监督和半监督的机器学习问题。排序是信息检索的核心问题，因此信息检索很自然地成为排序学习的主要应用领域之一。

现有的排序算法可以分为pointwise、pairwise和listwise三大类。

pointwise方法假设训练数据中的每个查询-文档组合都有一个表示二者相关度的分数，将排序问题转化为一个回归问题，即给你一个查询以及一个文档，预测它的相关度分数。常见的pointwise排序学习算法包括McRank[陈德品46]，PRank[18]等，其中McRank将排序问题转化为传统的分类问题；而PRank将排序问题转化为顺序回归问题。

pointwise方法旨在精确预测每个文档的相关分数，但在实际应用中，我们更加关注文档之间的先后关系，即模型能将相关度高的文档排在相关度低的文档前面，而不关注预测的分数是不是足够精确。针对这一点，pairwise方法就诞生了，它将排序问题转化为一个二值分类问题，即训练一个二值分类器，给定一对文档，该分类器能回答两个文档中哪个相关度更高。优化的目标是尽量减少排序输出中倒置文档对的数量。代表性的pairwise算法有Ranking SVM，这也是本文将要用到的排序器。

listwise方法是最新提出来的一类排序学习算法，相对于前面两类，它直接使用训练数据中的整个搜索列表作为一个训练实例。根据K个训练实例（一个查询及其对应的所有搜索结果评分作为一个实例）训练得到最有评分函数F，对于一个新的用户查询，函数F对每一个文档打分，之后按照得分顺序由高到低排序，就是对应的搜索结果。

* + 1. **评价指标**

排序学习算法的研究离不开有效的评价体系。常用的有准确率(precision)、召回率(recall)、平均准确率的均值(Mean Average Precision, MAP)、平均排序倒数(Mean Peciprocal Rank, MRR)、NDCG(Normalized Discounted Cumulative Gain)。本文(revisit)使用了NDCG评价指标，它是对传统评价指标的一个改进。

* 1. **TCA/TCA+算法介绍**
     1. **TCA算法**

迁移成分分析Transfer Component Analysis(TCA)[1], 最早是由香港科技大学杨强教授团队提出。是最先进的特征提取迁移学习技术之一。TCA算法背后的动机是源领域和目标领域之间可能存在一些常见的潜在因素，即使这些领域的观察特征不同。为了揭示潜在因素，我们将这些域投影到一个新的空间，这个空间被称为潜在空间。以这种方式，可以减小域差异，同时可以保留原始数据结构（例如，数据方差和局部几何结构）因此，由潜在因素跨越的潜在空间可以用作跨域分类任务的桥梁。例如，Internet Explorer 8和FireFox是由同一组标准度量表示的两个项目，因为它们的开发过程显著不同[TCA-12]。但是，它们都是Web浏览器。这意味着它们在编码方面有一些共同点，即使共同性可能被隐藏。如果可以发现隐藏的共同性并用它来表示两个项目的数据，那么跨项目差异可能会大大降低。因此，训练一个项目数据的缺陷预测模型可以成为应用于另一个项目。

在数学上，TCA试图学习一种转换φ，它能将源域和目标域的原始数据映射到一个潜在的，两个域距离Dist(φ(),φ())较小的空间以及转换后的数据方差Var({φ(), φ()})较大。假设转换 φ ： 🡪 将原始特征向量映射到d维度子空间。那么，TCA的高级表示可以写成如下：

arg min Dist(φ(), φ()) + λR(φ), (1)

φ

s.t. constraints on φ() and φ()

其中 φ 是通过最小化域之间的差异来学习并满足一些约束来保存变换后的数据方差，R(φ)是对于φ的避免过拟合的正则化，λ ≥ 0是控制正则化影响的权衡参数。正则化是一种常规的用于机器学习和数据挖掘的技术。[Transfer-22,23]

* + 1. **TCA with Normalization**
    2. **TCA+算法**

1. **实验**
   1. **实验设置**
   2. **实验结果**
2. **总结与展望**
   1. **总结**
   2. **展望**

**跨项目软件缺陷预测研究**

**摘 要**

相关数据显示...

【关键词】 跨项目 软件缺陷预测 排序学习 ...

**Cross-Project Software Defect Predicton**

**Abstract**

According to the reports, the number and size of various types of ...

**【Key words】**cross-project software defect prediction ranking learning

**参考文献**

[1]. S. J. Pan, I. W. Tsang, J. T. Kwok and Q. Yang, "Domain Adaptation via Transfer Component Analysis," in IEEE Transactions on Neural Network, vol. 22, no. 2, pp. 199-210, Feb. 2011.doi: 10.1109/TNN.2010.2091281