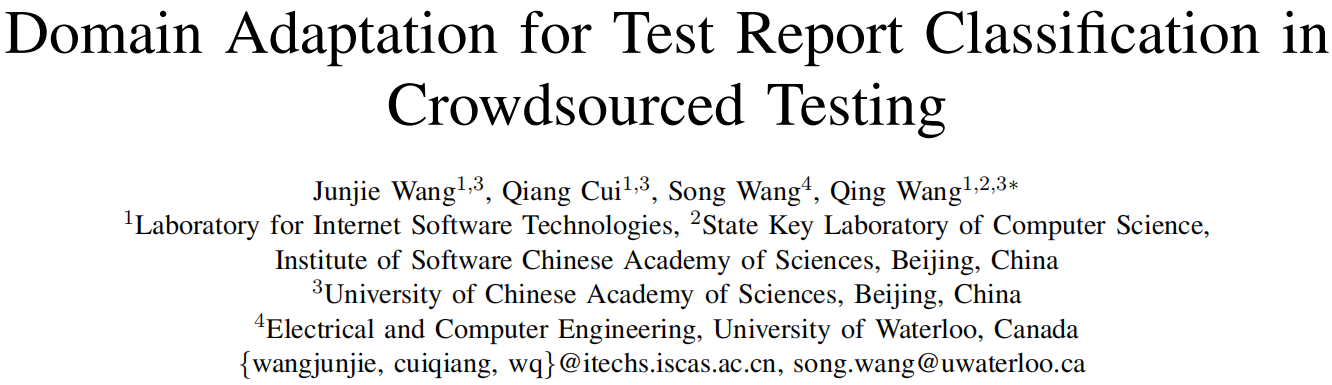
众包测试中测试报告分类的领域适应



1. 摘要

在众包测试中，从众包人员提交的大量测试报告中，将实际揭示故障的测试报告自动分类为真实故障是有益的。大多数现有的方法都是简单地利用历史数据来训练机器学习分类器，并对新传入的报告进行分类。然而，我们对真实工业数据的观察表明，众包测试下的项目来自不同的领域，提交的报告通常包含不同的技术术语来描述每个领域的软件行为。当使用不同的数据分布进行跨域报表分类时，分类模型的性能会显著降低。

为了建立一个有效的跨领域分类模型，我们利用深度学习来发现跨领域共享的中间表示，通过领域特定术语和领域未知术语之间的共现。具体来说，我们使用堆叠去噪自动编码器从原始文本术语中自动学习高级特征，并利用这些特征进行分类。我们对来自中国最大的众包测试平台之一的10个领域的58个商业项目进行了评估，结果表明，与三个常用的、最先进的基准相比，我们的方法可以产生令人满意的结果。此外，我们也使用真实案例研究来评估其有效性。来自真实世界测试人员的反馈证明了它的实用价值。

1. 介绍

众包测试是软件工程界和工业实践的一个新兴趋势。在众包平台执行测试任务后，人群工作者必须提交测试报告。然而，这些测试报告往往有许多假阳性，即测试报告被认为是失败的，实际上描述了一个正确的行为。项目经理或测试人员需要手动检查这些失败的测试报告，以验证它们是否实际显示了故障，只有不到50%的人最终被确定为真正的错误。因此，这个过程是耗时、乏味和低效的。因此，从人群工作者提交的大量测试报告中自动分类真正的错误将是有益的。

我们对真实行业数据的观察显示，众包测试的项目来自各行各业，从旅游、音乐到安全、摄影。不同领域的测试报告使用不同的技术术语来描述软件行为。例如，旅行领域的报告将包含“位置”、“导航”和“地点”等术语，而音乐领域的报告将包含“播放”、“歌词”和“歌曲”等术语。因此，来自这两个域的文本特征在它们的分布上有很大的不同。跨域的不同特征分布会降低机器学习分类器在跨域分类中的性能。这是因为大多数机器学习模型是在假设训练集和测试集是从相同的数据分布中提取的情况下设计的。大多数现有的报告分类方法直接利用文本特性来构建模型，而没有考虑跨领域的不同数据分布问题。这将导致众包报告的跨域分类性能较差。我们之前的工作提出从训练集中选择相似的数据实例来构建分类器。虽然它可以在一定程度上缓解分布差异问题，但在相似实例数量较少的情况下不起作用。

为了克服不同的数据分布问题，更有效地进行跨领域的报告分类，我们提出了领域适应报告分类(Domain Adaptation report classification， DARS)。利用堆叠去噪自编码器(SDA)，一个强大的表示学习算法(也称为深度学习)，学习高级特征，然后利用这些特征进行分类。为了抽象高级特性，SDA从跨域共享的原始文本术语中发现中间表示。直观地说，中间表示是通过上述提到的领域特定术语和领域未知术语之间的共现来学习的(例如，跨领域出现的术语，如“按钮”、“打开”和“错误”)。

我们在中国最大的众包测试平台之一的25564份测试报告上，对DARS的有效性和优势进行了实验研究，这些测试报告涉及10个领域的58个商业项目。结果表明，DARS的中位数可以达到0。77 F1和0。84 AUC。它显著优于三种常用和最先进的基线报告分类方法。此外，我们还与真实世界的测试人员进行了案例研究和调查，以进一步评估DARS的有效性。反馈显示，76%的测试人员同意DARS的有效性，并愿意将其用于实际的众包测试。

本文的主要贡献如下：

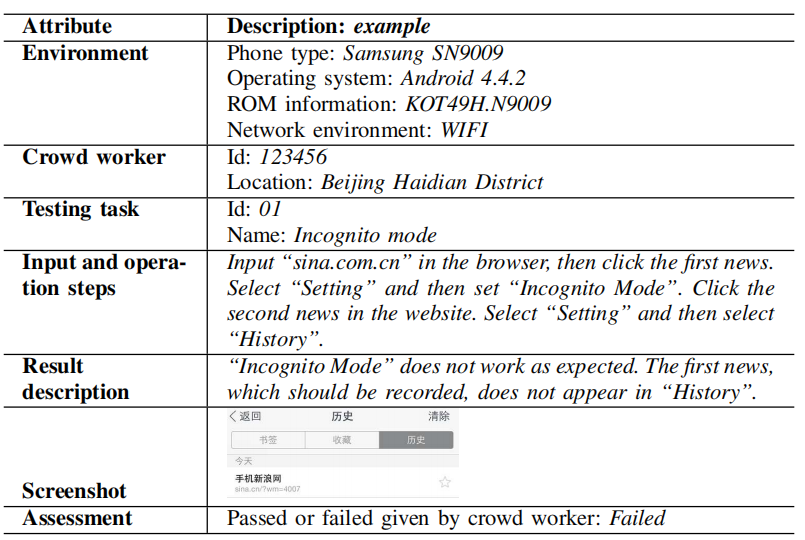
* 提出了面向跨领域众包报告的领域适应报告分类(Domain Adaptation Report claSsification， DARS)方法，克服了跨领域的数据分布差异。
* 我们从中国最大的众包测试平台之一收集了来自10个领域58个商业项目的25564份测试报告，评估了我们的方法，结果很有希望。
* 我们通过实际案例研究评估DARS的有效性，并讨论在实践中采用我们的方法时所吸取的教训和遇到的挑战。

1. 背景知识

3。1众包测试

我们的实验使用百度众包测试平台。一般来说，测试人员准备测试任务，并将它们分发到众包测试平台上。然后众包工作者可以签到进行任务，完成任务后提交众包测试报告。表1展示了一个典型的众包报告的属性。它包含操作步骤、结果描述、屏幕截图等，以及评估工人是否认为软件的行为是正确的(即通过)，还是错误的(即失败)。对于每个测试任务，平台还包含相关项目所属域的名称。

表1。 众包测试报告的一个例子



为了吸引更多的工人，测试任务经常得到经济补偿，特别是对于这些失败的报告。通常，该平台每月交付大约100个项目，平均每天接收超过1000个测试报告。然而，有些测试报告是假阳性的，也就是说，一个被标记为失败的测试报告实际上涉及到正确的行为或所研究的软件系统之外的行为。测试人员需要手动检查该平台上失败的测试报告，以判断它们是否真正揭示了一个错误。然而，对于一个测试人员来说，手工检查1000份报告可能需要半个星期的时间。此外，只有不到50%的错误最终被确定为真正的错误。显然，这样的过程是耗时和低效率的。

3。2堆叠去噪自动编码器(SDA)

堆叠去噪自编码器(Stacked Denoising Autoencoder， SDA)是一种无监督学习有效表示的人工神经网络。它由多层去噪自动编码器的组成。

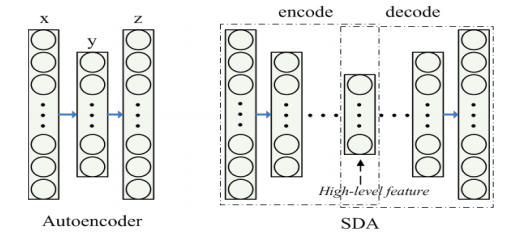


图1。 自动编码器和堆叠去噪自动编码器

一个自编码(如图1)需要一个输入向量x∈[0， 1]d，并编码到一个隐藏表示y∈ [0， 1]，通过一个确定的映射y = f0(x)=s(Wx+b)参数化，θ= W，b。W是d'×d向量权重矩阵和b是一个偏差矢量。然后将得到的潜在表示y解码为输入空间z = gθ'(y) = s(W'y+b')且θ'=W'，b'的重构向量z。反向映射的权矩阵W受W’=WT的约束。对模型参数进行优化，使平均重构误差最小。

去噪的自动编码器(DA)对自动编码器进行了轻微的修改，即在将输入编码为隐藏表示之前破坏输入。通过最小化L(x， g(h(x)))，训练它从已损坏的版本x中重建(或去噪)原始输入x。一种典型的破坏选择是二进制掩蔽噪声。它将每个输入的部分特征置零，其中噪声水平是一个输入参数。这对于众包测试报告的文本特征来说是很自然的，在这些测试报告中，个人特定的词汇偏好会影响单词的存在或缺失。

堆叠去噪自编码器(SDA)(如图1所示)将一系列DAs堆叠在一起，通过将第t个DA的隐藏表示作为(t+1)个DA的输入，构建了一个深入的体系结构。训练是贪婪的，一层一层地进行。最后一个DA中的潜在表示被视为高级特征。隐含层数和每层节点数为输入参数，可根据用户需求进行设置。

3。3动机

现有的测试报告分类方法大多假定训练集和测试集来自相同的数据分布。这意味着当使用这些方法来分类新的测试报告时，性能会迅速下降，这些测试报告与训练集有不同的分布。

我们为随机选择的四个领域，即旅行、音乐、安全和照片，展示了测试报告文本描述的术语云，以说明图2中的分布差异及其影响。我们可以很容易地观察到，技术术语在不同的领域中表现出显著的差异。例如，旅游领域(图2)中的报告包含诸如位置、导航和位置等术语，而音乐领域(图2)中的报告包含诸如播放、歌词和歌曲等术语。因此，从这两个领域衍生出来的文本特征在分布上存在着明显的差异。从这个意义上讲，基于旅游领域的报告构建的带有文本特征的模型可能无法有效地对音乐领域的报告进行分类。为了缓解这个问题，需要新的特性和方法。

在图2中，我们还注意到跨域的项目共享一定数量的公共术语。例如，许多项目包含诸如显示、设置和下载等行为术语，或者诸如无、错误和缺失等描述性术语。这些共享的术语可以帮助跨越领域的鸿沟。更具体地说，在旅游领域，有类似于展示建筑位置的描述，而在音乐领域，类似于展示歌曲歌词的描述是很常见的。在表示学习算法的帮助下，歌词、歌曲、术语定位等术语通过共享的术语显示建立关系。基于共生，表示学习算法将原始术语映射到高级特征。两个术语共享的上下文越相似，它们的高级特性就越相似。然后，这些高级功能可以用作构建预测模型和预测众包报告的输入。



图2。 众包报告不同数据分布的说明性例子

1. 研究方法

图3展示了领域适应报告分类(DARS)方法的概述。一般来说，DARS首先训练SDA来有效地编码输入特征。然后将训练集和测试集输入到训练的SDA中，生成高级特征。然后根据高级特征进行分类。DARS包括四个主要步骤：1)提取文本特征，2)训练SDA， 3)利用SDA生成基于文本特征的高级特征，4)构建分类器并使用学习到的高级特征进行分类。

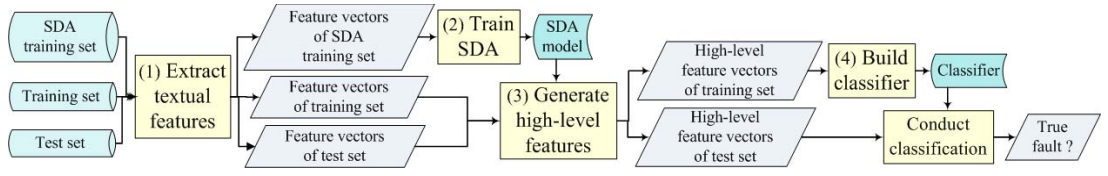


图3。 DARS概览

4。1提取文本特征

特征提取的目的是从众包报告中获得特征，这些报告可以用作训练SDA的输入。我们首先收集不同来源的文本描述(输入和操作步骤，结果描述)。然后我们进行分词，因为我们实验中的众包报告是用中文写的。我们采用ICTCLAS2进行分词，并将描述分词成词。然后我们移除停止字(即on、the等)以减少噪音。由于工作者经常使用不同的词来表达相同的概念，我们引入同义词替换技术来缓解这个问题。采用LTP3的同义词库。剩下的每一个术语都对应一个特征。对于每个特征，我们取其在描述中出现的频率作为其值。我们将这些特征组织成一个特征向量，每个特征值作为相应的词频。

4。2训练SDA

为了生成用于众包报告分类的高级特征，我们首先需要使用SDA训练集训练SDA。已有研究表明，SDA训练集的规模越大，的性能越好。此外，表示学习技术不涉及类信息(即，报告是否是真实错误)。因此，通常的做法是在培训过程中利用所有可用的数据实例。训练一个SDA就是确定权值w和偏差b，使训练好的SDA能够有效地对输入特征进行编码。为了训练一个有效的SDA来生成高级特征，我们需要调整四个参数：1)隐藏层数，2)每个隐藏层的节点数，3)噪声水平，4)训练迭代次数。现有的利用SDA为自然语言处理和图像识别生成特征的工作报告称，SDA生成的特征的性能对这些参数很敏感。输入和输出层中的节点等于特征向量的大小。对于其他层，我们将节点数量设置为相同的，以简化我们的模型。另外，SDA要求输入向量的长度相同。为了利用SDA生成高级特征，我们首先收集SDA训练集中出现的所有特征，并将它们组织成一个联合特征向量。然后将原始特征向量转化为联合特征向量。具体来说，对于原始特征向量中包含的特征，我们使用其原始值作为联合特征向量中的值。否则，我们将其值设置为0。添加0并不会影响结果，因为它仅仅意味着特定的术语没有出现在报告中。

注意，SDA要求输入数据的值在0到1之间。由于输入向量表示每个特征的词频，其值可以大于1。为了满足这一要求，我们使用最小—最大归一化对SDA训练集的特征向量值进行归一化。

4。3生成高级特征

在我们训练了一个SDA之后，权重w和偏差b都是固定的。对于训练集和测试集的特征向量，我们首先将其映射到联合特征向量中。然后我们使用min-max归一化将它们归一化到0到1的范围内。将训练集和测试集的归一化联合特征向量分别输入SDA。SDA的最后一个隐藏层中的表示随后被视为高级特性。

4。4建立一个分类器

在训练集和测试集上获得每个众包报告生成的高级特征后，我们基于训练集构建机器学习分类器，然后使用测试集对构建的分类器的性能进行评估。为了更好地帮助手工检查，分类器将提供报告为真实故障的概率。概率值越大，报告包含真实错误的可能性越大。为了计算F1，我们使用0。5作为截止值，表示将概率值大于0。5的报告视为真故障，反之亦然。

1. 实验设置

5。1研究问题

RQ1：(有效性)：DARS对众包报告的分类有多有效?

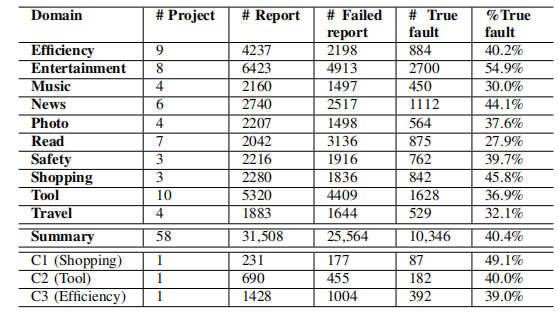
RQ2：(优势)：在分类众包报告方面，DARS能否超越现有的技术?

RQ3：(有用性)：DARS 对软件测试人员有用吗?

5。2数据收集

我们的实验基于来自百度众包测试平台存储库的众包报告。 我们收集2015年10月1日至2015年10月31日期间关闭的所有众包测试项目。 我们根据平台中记录的域名将这些项目分组为域。表二提供了这些领域的详细信息，包括项目的数量、提交的报告的数量、报告失败的报告的数量以及失败报告中真实故障的数量和比率。

表2。 调查中的项目



请注意，我们的分类是对失败的报告进行的，而不是完整的。由于以下原因，我们排除了通过的报告。正如我们提到的，失败的报告通常涉及正确的行为和真正的错误。然而，通过与公司的测试人员交谈，我们发现几乎没有通过的报告涉及真正的错误。这可能是因为补偿有利于错误，所以人群工作者不太可能错过错误。

每个报告的评估属性（见表一）可以作为分类的基本真相标签。对于具有评估属性的报告（只有2106份报告），我们只需使用存储的值作为分类的标签。对于其他报告，要求公司的两名测试人员为每个报告分配评估标签。在他们单独的标签后，我们分析了他们的标签之间的差异，并进行了后续访谈，直到达成共识。人们可能会认为，测试人员也可以在标记过程中产生假阳性，这将影响我们的实验结果。控制良好的标签过程，有两名测试人员和后续访谈，可以帮助缓解这一问题。更重要的是，公司的测试人员比人群工作者更有经验，他们这样做没有经济补偿的缺点。因此，我们认为基本真理标签是相对值得信赖的。

5。3实验设置和基线

正如我们在第三-B节中提到的，培训SDA不涉及班级信息。对于正在调查的10个域，我们使用一个域的所有众包报告作为测试集，并提取测试集的特征向量。我们从其他9个域中随机选择K个众包报告作为训练集，并提取训练集的特征向量。将K分别设置为100、200、500、1，000、1，500、2，000、3，000和5，000，以调查培训集大小对模型性能的影响。每个K的实验重复50次，以保证结果的稳定性。

为了进一步探索我们提出的方法的性能，我们将DARS与三种典型的基线方法进行了比较。

领域未知分类(DUC)：它是最直接的预测方法，不考虑不同领域之间的分布差异。利用训练集中所有众包报告构建机器学习分类器，并对测试集进行分类。

转移成分分析(TCA+)：它是领域适应的最新技术。TCA的目的是通过最小化数据分布之间的距离，在保留原始数据属性的同时，为训练集和测试集找到一个潜在的特征空间。TCA+通过自动规范化扩展了TCA，并且可以产生比TCA更好的性能。

基于聚类的分类方法(CURES)：这是众包报告分类的最先进技术，也可以缓解众包报告中数据分布的差异。首先对训练集的相似报告进行聚类，然后根据每个聚类的报告构建分类器。然后选择与测试集最相似的聚类，进行分类。

DARS和这些基线都涉及到利用不同的机器学习分类算法来构建分类器。在这项工作中，我们实验了被广泛报道为有效的许多不同的分类任务在软件工程——线性回归(LR)。

5。4评估指标

为了评估我们提出的方法，我们使用两个指标：F1和AUC。F1是一个广泛采用的指标来评估问题报告分类技术。它是对真实故障进行分类的准确率和召回率的调和手段。AUC是评价不平衡数据分类性能的最常用和最广泛的指标。由于众包报告的数据集通常是不平衡的(即较少的真实错误)，我们特别使用这个指标。它是ROC曲线4下的面积，用来衡量分类器的整体判别能力。完美模型的AUC为1，而预测所有实例为真或假的模型的AUC为0。

1. 实验以及分析

6。1RQ1（有效性）

图4展示了所有实验域的50个实验，在不同训练集大小下，DARS的F1和AUC。我们可以看到，随着训练集大小的增加，F1和AUC都会先提高，然后基本保持不变。

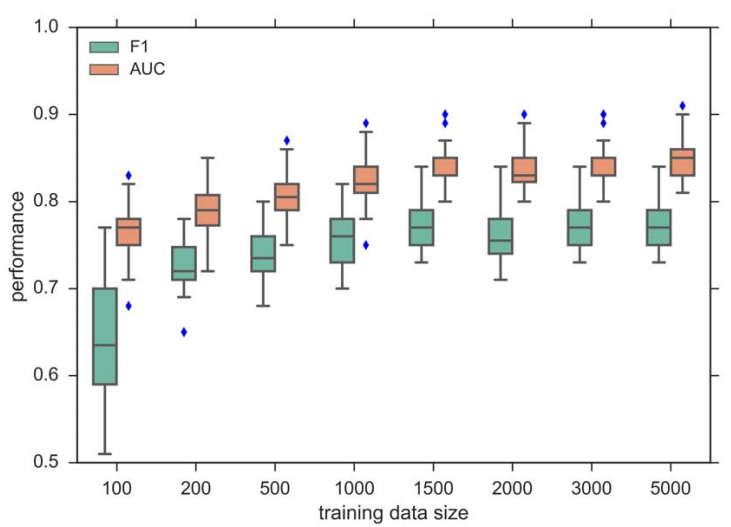


图4。 DARS的有效性

然后，我们对每个相邻训练集大小下的分类性能进行Mann-Whitney测试。结果表明，当训练集大小小于1，500时，相邻每个训练集F1和AUC的p值均小于0。05。当训练集尺寸大于1500时，每个相邻训练集F1和AUC的p值均大于0。05。这说明少于1500个实例的训练集会降低性能，超过1500个实例的训练集不会显著提高性能。因此，1，500为相对最优列车集尺寸。

我们关注训练集大小为1，500时的性能。F1范围为0。72 ~ 0。82，F1中位数为0。77。AUC范围为0。80 ~ 0。90，中位AUC为0。84。这意味着DARS只需要1500个标记数据实例就可以获得相对满意的性能。它受益于学习到的高级特性，这些特性可以基于未标记的数据学习。通常有大量的未标记数据，它们相对容易收集。我们的方法提供了一种有效利用这些未标记数据的方法。

表3。 每个测试域的性能

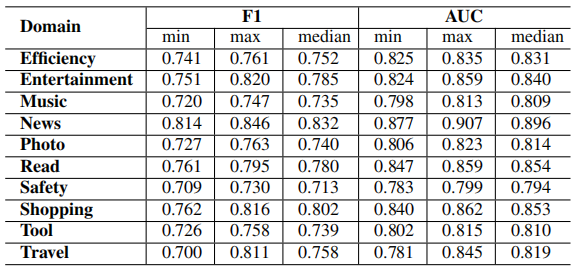
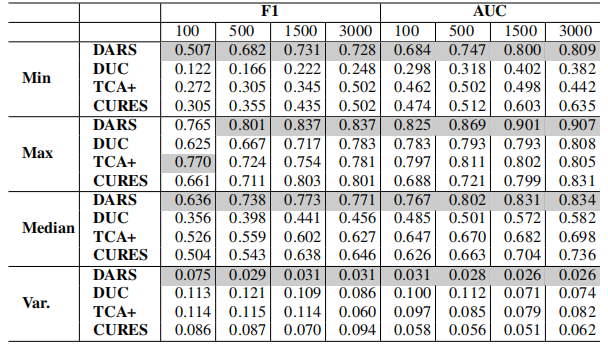


表3展示了作为测试集的每个域的分类性能，训练集大小为1500。我们可以很容易地观察到，不同领域的性能可能在某种程度上有所不同。不同区域F1中位数为0。71 ~ 0。83，AUC中位数为0。79 ~ 0。89。最差的表现出现在诸如安全和音乐等领域。我们进一步分析了其背后的原因。在这两个领域的众包报告中出现了大量的新词，例如歌曲名称。在建立中间表示时，这些术语会带来噪音。

6。2RQ2（优势）

表4说明了DARS和三条基线的性能。我们随机选择众包报告作为训练集，每个训练集大小重复50次。对于这些方法，我们给出了随机实验的最小值、最大值和中值性能。背景颜色为黑色的值表示每个训练集大小的最佳F1或AUC。

表4。 性能与基线的比较

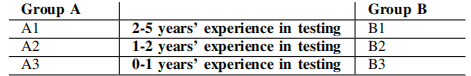


对于每一个训练集大小，DARS都能以最小的方差达到最高的中位数F1和AUC。我们还在DARS和每个基线之间对F1和AUC进行了Mann-Whitney检验。所有测试的p值都小于0。05。这进一步说明了我们方法的有效性和优点。在三个基线中，DUC的性能最差，这说明如果不考虑数据在不同领域的分布差异，分类器的性能会很差。

对于所有训练集大小，TCA+在F1和AUC中的性能都比DARS差。这是合理的，因为现有的领域适应方法(即TCA+)是基于原始特征的线性投影来生成新的特征。然而，深度学习技术(如SDA)可以从原始特征的非线性映射中学习高级特征，从而编码复杂的数据变化。我们的方法也优于CURES，这是最先进的技术分类众包报告。这是因为CURES依赖于历史上类似的报告来构造分类器。当训练集不够大时，CURES系统性能会下降。

6。3RQ3（有用性）

表5。 案例研究参与者

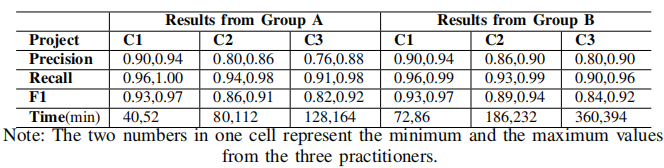


为了进一步评估DARS的有效性，我们在百度进行了个案研究和调查。我们随机选择了三个项目作为案例研究(详见表二)，其中包括来自众包测试组的六名测试人员。根据他们的经验，我们将他们分为两组，具体情况见表五。

本案例研究的目标是评估DARS在从众包测试报告中分类真实故障方面的有效性。首先，我们利用训练好的SDA为三个实验项目的所有报告生成高级特征。然后，我们使用最初标记的报告来构建一个分类器。通过对三个项目进行分类，我们得到每个报告为真故障的概率。

对于这两个组，我们要求从业人员用yes或no标记每个报告，表示该报告是否涉及真正的错误。对于B组的从业者，我们只向他们展示分类的众包测试报告。对于A组的从业者，除了众包报告外，我们还为他们提供了每个报告为真实错误的预测概率。此外，报告以概率值降序显示。

表6。 案例研究结果



我们提到，我们只要求从业者分配“是”或“否”的标签，因为真正的错误的概率是复杂的，需要手工测量。因此，我们不提出需要计算概率的AUC。相反，除了F1，我们在表六中给出了分类的精度、召回率和所需时间。用DARS(A组)提供的概率辅助的分类可以找到与没有辅助的分类(B组)一样多的真正错误，时间要少得多。特别是，随着项目规模的增加，没有援助的人工分类所消耗的时间可以急剧增加，而其准确性(F1)并没有显示出根本的差异。

1. 总结讨论

我们已经提到，众包报告分类的主要挑战是跨域报告的不同数据分布。简单地利用其他领域的报告来构建分类器很容易导致低精度。我们假设表示学习算法可以使跨域的特征分布更加相似，从而减少特定于域的信息引入的噪声。为了直观地证明这一点，我们研究了每对域的特征分布之间的A距离。距离是两个数据集的概率分布之间相似性的度量。我们假设，在表示学习之后，区分不同的领域应该更加困难，这意味着领域适应的有效性。这可以通过更相似的特征分布来说明。

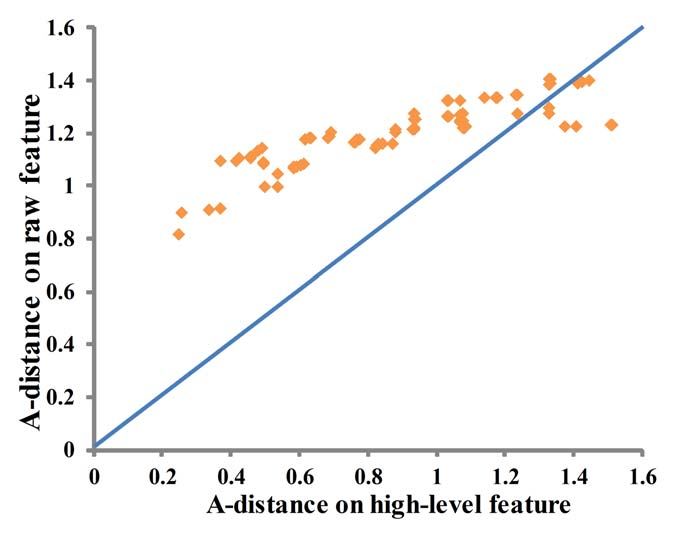


图5。每个域对的距离

图5报告了每个域对的原始特性和高级特性的距离。正如预期的那样，对于高级特征，A-距离减小，这意味着在表示学习之后，域之间的特征分布变得更加相似。这就是为什么基于高级特征的分类可以获得更高的性能。

也存在一些威胁，外部威胁胁关系到本研究的一般性。首先，我们的数据集包括58个项目，涵盖10个领域，从中国最大的众包之一收集测试平台。项目的各个领域和数据的规模相对减少了这种威胁。其次，本研究中调查的所有众包报告都是用中文写的，我们不能保证在其他语言的众包项目中也能观察到类似的结果。但这一点得到了缓解，因为我们没有进 行语义理解，而是简单地标记句子，并使用单词作为表示学习的标记。关于内部威胁，而不是实验研究SDA训练集对分类性能的影响，我们只是将所有可用的数据实例视为SDA训练 集。这得到了先前几项研究的结果的支持。此外，实验结果证明了其有效性。无论如何，我们将进 行精心设计的实验，以进一步研究其影响。

本文提出了领域适应报告分类(DARS)方法，以克服众包报告分类中跨域的数据分布差异。我们从中国最大的众包测试平台之一的有效性、优势和有用性的角度来评价DARS，结果是有希望的。应当指出，所提交的材料只是正在进行的工作的起点。我们正在与百度众包平台密切合作，并计划在网上部署该方法。返回的结果将进一步验证有效性，并指导我们改 进我们的方法。未来的工作还将包括探索其他特性和技术，以进一步提高模型的性能和稳定性。

本文由南京大学软件学院2021级硕士研究生石孟雨翻译转述。