基于迁移学习的异构磁盘系统大型数据中心少数磁盘故障预测

会议文件 • 2019年8月

DOI：10.1145/3337821.3337881

引文
2

阅读
429

8位作者，包括：

张吉
华中科技大学
3份出版物，30份引文
见简介

克周
华中科技大学
146份出版物990份引文
见简介

黄平
华中科技大学
51份出版物，269份引文
见简介

基于迁移学习的异构磁盘系统大型数据中心少数磁盘故障预测

张吉[§]，周可[§]黄平[§§]，何旭斌[$]，Zhili Xiao[ζ]，程斌[ζ]，吉永光[ζ]，王银虎[ζ]

[§]武汉光电国家实验室和计算机科学与技术学院
[§]信息存储系统重点实验室，科大和腾讯智能云存储联合研究中心
[$]坦普尔大学，[ζ]腾讯公司。
{jizhang，k.zhou}@hust.edu.cn，{templestorager，xubin.he}@temple.edu
{tomxiao，bencheng，raidmanji，yhwang}@tencent.com

摘要

大型数据中心的存储系统通常建立在数千甚至数百万个磁盘上，其中磁盘故障不断发生。 如果无法恢复丢失的数据，磁盘故障可能导致严重的数据丢失，从而导致系统不可用，甚至造成灾难性后果。 虽然复制和擦除编码技术已被广泛应用以保证存储的可用性和可靠性，但磁盘故障预测正日益流行，因为它有可能防止磁盘故障首先发生。 最近的趋势转向应用基于磁盘SMART属性的机器学习方法来预测磁盘故障。 然而，传统的机器学习(ML)方法需要大量的训练数据才能提供良好的预测性能。 在大规模存储系统中，新磁盘逐渐进入以增加存储容量或替换失败的磁盘，导致存储系统由来自不同供应商的少量新磁盘和/或来自同一供应商的不同型号组成，随着时间的推移。 我们将这个相对较少的磁盘称为少数磁盘。 由于缺乏足够的训练数据，传统的ML方法无法在由异构少数磁盘组成的进化存储系统中提供令人满意的预测性能。 为了应对这一挑战，提高大数据中心少数磁盘的预测性能，我们提出了一种基于传输学习方法的少数磁盘故障预测模型TLDFP。 我们在两个现实数据集上的评估结果表明，与基于传统ML算法和两种最先进的迁移学习方法的四种流行预测模型相比，TLDFP可以提供更精确的结果。

1 导言

硬盘被广泛用作现代数据中心大规模存储系统的常用和主要存储设备。 在这样的数据中心，它一直是一个极具挑战性的承诺，以确保高可用性和可靠性的IT管理，因为各种磁盘故障不断发生在现场，无论是硬盘[，，]还是基于闪存的SSD[，]。112146719 如果现有的数据保护方案无法恢复丢失的数据，例如，由于磁盘故障超过设计的校正能力[，]复制和擦除代码，磁盘故障可能导致临时数据丢失，从而导致系统无法使用，甚至导致永久数据丢失。39 硬盘是一个相当复杂的系统，由各种磁性、机械和电子部件组成，每个部件都可能失效。 因此，由于许多原因，硬盘故障表现出不同的表现形式和程度[]，这在主要IT公司[，]的数据中心已经观察到。 341329与传统的无源容错技术如EC（擦除代码）和RAID（独立磁盘冗余阵列）[]相比，主动磁盘故障预测倾向于提前确保大规模存储系统的可靠性和可用性。41 因此，成功的磁盘故障预测不仅降低了丢失数据的风险，而且降低了与恢复驻留在失败磁盘上的数据相关的数据恢复成本(即网络带宽。

1硬盘制造商在磁盘固件中实现了自我监视、分析和报告技术(SMART)技术。 大多数SMART属性包含有关磁盘逐渐退化和可能缺陷的信息。 在内部，磁盘使用基于SMART值的所谓“阈值方法”[]来声明其故障状态，这意味着如果SMART属性的值超过相应的预定义阈值，硬盘将引起警报。39 *然而，这种“阈值法”只实现了3%-10%的故障检测率(FDR)和0.1%的虚警率(FAR)[]*。39 换句话说，这些数字突出了这种方法的保守性质，即它宁愿错过检测更多磁盘故障的机会，也不愿以更高的速度报告错误警报。

82628为了提高预测性能，[，，]提出了几种基于机器学习(ML)的磁盘故障预测模型，利用训练SMART数据来预测磁盘故障。 不幸的是，这些工作集中在大量具有足够训练数据的同质磁盘上。在大规模存储系统场景中，成群的新磁盘逐渐进入以取代失败的磁盘，导致存储系统由来自不同供应商的异构磁盘和来自同一供应商的不同模型组成，随着时间的推移。 异构磁盘

ICPP2019，2019年8月5-8日，日本京都

https://doi.org/10.1145/3337821.3337881

2227许多磁盘模型在数据中心[，]中是常见的。 此外，在不断发展的存储系统中，一些磁盘模型比其他模型要少得多，我们将这种相对少量的磁盘少数磁盘（相反，大量磁盘作为多数磁盘）称为异构磁盘系统的大数据中心。3.1. 我们发现，在两个真实世界的数据中心中，大约25%的具有许多模型（超过50个）的磁盘是少数数盘，如第2节所详述的那样，由于少数数盘的样本小，训练数据不足，传统的ML算法使用少数数盘的训练数据会大大增加过度拟合的风险(Section或泛化不良[]，这将削弱预测模型的性能，严重影响存储系统的可靠性。3.1) 48 因此，我们准备开发一个磁盘故障预测模型TLDFP，在具有丰富的异构磁盘数据集的情况下预测少数磁盘的故障。 我们的基本思想是从可用的多数磁盘数据集预测少数磁盘故障，这是传输学习的一种应用。

在本文中，我们旨在寻求以下问题的答案：（1）就故障预测而言，少数磁盘数据集的定义是什么？（2）为什么要使用迁移学习来预测少数磁盘故障？ （3）如何运用迁移学习方法预测少数磁盘故障？ （4）何时使用转移学习进行少数磁盘故障预测？ 此外，当应用于来自世界上最大的社交网络公司之一的公共Backblaze和腾讯的两个真实世界数据集时，我们的方法TLDFP在处理现实的系统挑战时，在进行跨盘模型故障预测时，平均达到96%的故障检测率和0.5%的虚警率。

2 背景和相关工作

几乎所有硬盘驱动器和基于闪存的SSD都带有内置的自我监控、分析和报告技术(SMART)，这是磁盘健康状况的指标。 智能技术的规范包含多达30个属性，报告各种磁盘操作条件。 智能数据直接或间接反映磁盘的健康状况，甚至包含一些统计信息。 智能数据可以通过指定的磁盘协议获得，磁盘制造商对此达成协议。如果SMART属性的值超过相应的预定义阈值，硬盘将引发警报。每个SMART属性条目由五个元素组成，描述为元组(ID、归一化、Raw、Threshold、最坏)。

- ID：SMART属性的指定序列号。
- 归一化：当前或最后一次归一化值（大多数归一化为由制造商特定算法使用其原始值计算的最佳值253和最差值1之间的值）。
- 原始：与传感器和特定供应商提供的计数或物理状态相对应的原始值。
- **阈值：磁盘报警失败的阈值。**
- 最差：给定属性的最低值或最差值。

元组中并非所有的五个元素都被使用。 在我们的论文中，我们重点讨论了我们收集的前三个元素(ID、归一化和原始数据集。 **为了方便起见，我们使用“smart_ID_Raw：V”来表示ID为V的SMART属性的原始值。 例如，smart_1_Raw：10表示读取错误率属性(ID：1)的原始值为10，smart_5_Normalized：56表示重新分配的扇区计数属性(ID：5)的归一化值为**56. 关于我们在评估中使用的SMART属性的更具体信息见表6.

提出了一种基于SMART数据的磁盘故障预测模型ML算法。32 哈默利和埃尔坎[]采用两种贝叶斯方法，基于量子公司的SMART数据对磁盘故障进行建模，该数据由1,927个良好驱动器和9个失败驱动器组成。 他们将问题归类为异常检测，并建立了一个称为NBEM的混合模型，简称用期望最大化训练的朴素贝叶斯聚类和另一种称为朴素贝叶斯分类器的方法。 它们对NBEM的故障检出率为35-40%，朴素贝叶斯分类器的故障检出率为55%，约为1。 Hughes等人。35 []探索两种统计方法来提高预测性能。 他们探索统计测试的能力，如秩和测试和OP-ed单变量测试，并测试这两种方法与7，744驱动器数据（其中36是故障）从两个不同的磁盘模型跨越3个月。 它们的故障检出率分别为60%和0.5%的误报率)。 Murray等人。15 []比较SVM、秩和检验、无监督聚类和反向安排检验的预测性能。

朱等人。8 []探讨了反向传播(BP)神经网络和改进的SVM模型建立基于SMART数据的预测模型的能力。[4]) 许多研究人员使用支持向量机(SVM)，因为他们声称SVM可以有效地使用核技巧执行非线性分类，隐式地将输入映射到高维特征空间[][]。 4431为了提高磁盘故障预测模型的稳定性和可解释性，Li等人。16 []提出了一种基于分类回归树(CART)的硬盘驱动器故障预测模型)。 回归树可以给磁盘一个健康评估，而不是一个简单的分类结果。301718 提出了梯度Boosted Regression Tree(GBRT[])对磁盘失效[，]进行建模，其中GBRT是一种基于树平均的梯度下降Boosting技术，是一种精确有效的ML方法，既可用于回归问题，也可用于分类问题。 为了避免过度拟合，GBRT算法训练了许多树桩作为周学习者，而不是完整的，高方差树。 此外，正则化贪婪森林(RGF[])方法是一种强大的非线性分类方法。36 它是GBRT的一个变体，其中结构搜索和优化是解耦的，它利用结构稀疏的概念直接对基于森林结构的森林节点进行贪婪搜索。 Mirela Madalina Botezatu等人。 采用这种方法对磁盘故障进行建模，取得了良好的效果[]。2 徐等人。102425 []提出了一种递归神经网络(RNN[，])方法，利用顺序信息预测硬盘故障。 他们使用从一个实字数据中心收集的数据集，其中包含3个不同的磁盘模型，分别表示为W、S和M，并建立这些磁盘模型的预测模型。 他们对长期依赖的顺序SMART数据进行建模，并展示了他们的预测模型的能力。 最近，Mahdisoltani等人。12 []建议使用传统的ML算法来预测磁盘

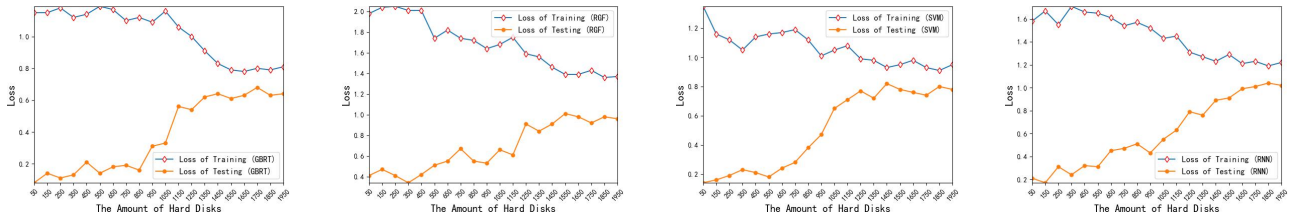


图1：四种流行的传统ML算法的训练损失和测试损失。 请注意，y轴表示随着数据集大小的增加而丢失的值。

使用SMART数据集的扇区错误。 我们在本文中的目标是作出整个磁盘故障预测，这需要更高的准确性，因为成本考虑。

如前所述，当新磁盘模型的培训数据有限时，就需要进行转移学习，这种情况经常发生在不断发展的存储系统中。 随着大数据存储库变得越来越普遍，使用与焦点或兴趣的目标领域相关但不完全相同的现有数据集使转移学习解决方案成为一种有吸引力的方法。2043633 迁移学习已经成功地应用于各种应用，包括多语言文本分类[，]，图像分类[]，人类活动分类[]，文本情感分类[]，Web文档分类[]等。1445 不足为奇的是，近年来，研究人员开始使用转移学习方法[，]解决少数磁盘故障预测问题。222 Mirela Madalina Botezatu等人。 提出了样品选择去偏置方法[]，我们将其表示为SSDB。2 它的主要思想是训练一个分类器，它可以根据与目标磁盘模型相关的样本的相似性对链接到特定磁盘模型的观测值进行排序。 该方法也是一种类似于TLDFP算法的单一源域转移学习方法。 FLF Pereira等人。22 提出了贝叶斯网络的多源域转移学习[]，本文将其表示为TLBN。 提出了一种新的基于聚类的信息源构建方法，并根据多个HDD的相似性对其进行分组，以构建一种新的信息源进行迁移学习。 虽然这些方法也为少数磁盘故障预测提供了解决方案，但我们将TLDFP与它们进行了比较，并表明我们的方法提供了更好的预测性能。此外，请注意，我们的工作是第一个系统地（什么，为什么，如何和什么时候）提出使用转移学习方法来解决基于SMART属性的大型、活动的、进化的存储系统的少数磁盘故障预测。

3 初步研究和动机

在本节中，我们通过实验检查定义了少数磁盘数据集，研究了SMART数据的分布，并证明了为什么我们使用传输学习来预测少数磁盘故障。

3.1 少数民族磁盘数据集

如前所述，我们的目标是提高磁盘数据集的磁盘故障预测性能，该数据集没有足够的训练数据，并且传统的ML算法提供了次优性能。 在这一部分中，我们给出了一个少数磁盘数据集的定义，并通过实验来定量地评估它们，实验显示了四种流行的ML算法(GBRT、RGF、SVM和RNN)在磁盘故障预测中的训练损失和测试损失23 2101731[，，，]. 损失是一个数字，表明模型的预测有多糟糕。请注意，我们在所有四种方法中都增加了一个正则化项来构造损失函数，这可以有效地防止由于模型具有非常多的参数而导致的过拟合。1 图说明了结果。 从图中可以看出，随着数据集的增加，训练集的损失在一定程度上增加，而测试集的损失减少，因为数据集的增加导致了需要拟合训练模型的更复杂的情况。 更具体地说，当磁盘数量小于1500时，随着数据集的扩大，训练和测试损失之间的差距减小，这被称为由少数磁盘引起的过度拟合。 当磁盘的数量超过1500，间隙变小并稳定。 因此，我们可以得出以下结论：（1)包含少于1500个磁盘的磁盘数据集可能导致过度拟合，我们将其命名为少数磁盘数据集；(2）当数据集包含少于1500个磁盘时，四种流行的传统ML算法无法提供令人满意的性能。 据我们所知，我们是第一个定义少数磁盘数据集并通过广泛的数据分析和实验定量评估它们的人。 我们研究了两个真实的数据中心，并按1500的阈值对磁盘数量进行了分类。1， 如数据中心BackBaze中的表所示，91种不同的磁盘模型只占所有磁盘的24%以下，12种模型占76%以上%。 我们称这91种型号为少数磁盘。 腾讯数据中心也发现了类似的观察。

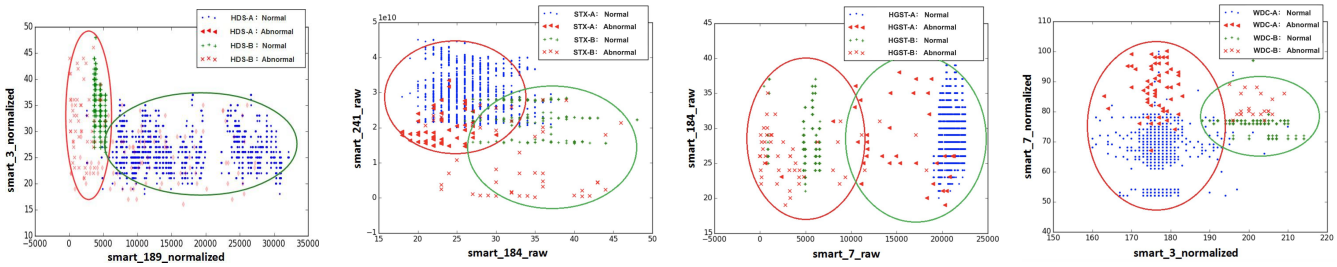
上述描述和分析表明，对少数磁盘进行磁盘故障预测是一个需要解决的现实问题。

表1：磁盘种群的特征

数据中心	磁盘号码	磁盘模型	总数	百分比
呆头呆脑	≥ 1500	12	114,570	76.61%
	< 1500	**91**	34,978	**23.39%**
腾讯	≥ 1500	8	18,996	73.32%
	< 1500	**52**	52,235	**26.67%**

3.2 使用传统ML的基线结果仅限于少数磁盘数据集

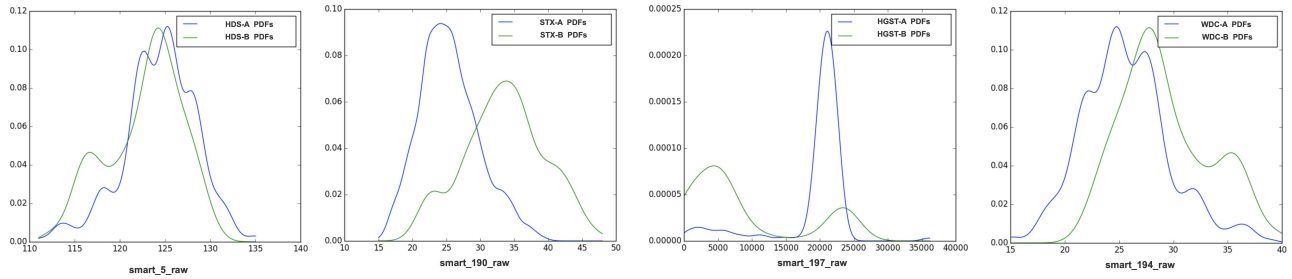
为了研究仅在少数磁盘数据集上训练的传统ML方法的预测结果，我们使用来自四个磁盘制造商的四个少数磁盘模型，基于四种流行的传统ML方法：GBRT、RGF、SVM和RNN进行本实验，这些方法包括流行的树结构算法和深度学习算法，并已广泛应用于磁盘故障预测。 请注意，我们只使用70%的少数磁盘数据集作为训练集，其余30%作为测试集，没有任何其他数据集。 此外，我们始终使用以下缩略语为这四个供应商在整个



(a) HDS STX HGST (d)妇女发展中心

图2：来自四个制造商的两个磁盘模型的两个SMART属性的分布，即日立、希捷、HGST、WDC。 每个子图显示由随机选择的两个磁盘模型的SMART属性对指示的异常和正常状态。这四个子图表明，每个制造商的两个磁盘模型表现出由两个SMART属性分布表示的相似故障模式，SMART数据分布在不同的值范围内，这促使我们应用转移学习来进行跨模型磁盘故障预测。

图3：来自四家制造商的两种不同磁盘模型的SMART属性值的PDF。



在其磁盘模型中也使用的纸张：日立(HDS)和希捷(ST)来自Backblaze，日立全球存储技术(HGST)和西方数字(WDC)来自腾讯。 2，从表中可以看出，四种传统的ML方法都不能提供高FDR和低FAR(FDR和FAR见节)。5.1.2 我们知道，由于使用基于传统ML的小齐次数据集引起的过度拟合，我们的预测性能很差。

表2：通过传统ML预测少数磁盘故障的预测结果

方法	制造商	罗斯福	很远
gbrt	hds/stx/hgst/wdc	27.3%/37.5%/31.6%/38.5%	29.0%/19.4%/17.6%/21.8%
RGF	hds/stx/hgst/wdc	36.4%/50.0%/47.4%/53.8%	44.3%/22.4%/ 53.4%/36.6%
svm	hds/stx/hgst/wdc	50.0%/41.7%/57.9%/30.8%	20.7%/47.8%/ 24.0%/43.7%
RNN	hds/stx/hgst/wdc	40.9%/33.3%/36.8%/30.8%	28.3%/31.3%/ 39.7%/38.7%

5 5).我们还进行了实验，直接使用现有多数磁盘的大型数据集来预测基于四种流行的传统ML技术的少数磁盘故障，性能也不令人满意(详情如图所示，为了了解原因，我们分析了SMART数据分布如下。

3.3 智能数据分布

有趣的是，观察到指示来自同一制造商的不同磁盘模型的磁盘健康条件的SMART属性的值表现出相似的分布模式。 我们分析了两个公开可用的SMART数据集从Backblaze[1] 以及来自腾讯数据中心的数据集。2 图显示了显示的SMART数据分布模式。 每个子图显示来自同一制造商和每个圆圈的两个不同磁盘模型的一对SMART属性值的分布模式

用于突出显示不同的磁盘模型。 显然，由两个磁盘模型的两个SMART属性指示的异常状态和正常状态之间的关系显示出类似的模式，只有SMART值的差异在不同的范围内。2(a)，2(b)2(c)2(d) 数字，并分别显示来自同一制造商的两个磁盘模型的异常状态在正常状态的上方、下方和左侧。 此外，SMART值分布在不同的光谱中。2(b) 以图希捷为例，模型STX-B的分布区域比模型STX-A的分布区域低。 传统的ML算法只有在从相同的分布中提取训练和测试数据时才能提供良好的预测性能[]。40 因此，当涉及到跨盘模型故障预测时，由于不同的分布谱，它们无法令人满意地执行，如图所示2.

为了深入研究来自同一制造商的不同磁盘模型的SMART值的分布，进一步激励从一个磁盘模型到不同模型的传输学习，并解释为什么我们使用传输学习来进行少数磁盘故障预测，我们研究了SMART数据分布的差异，并对来自同一制造商的不同磁盘模型进行了直观的分析。 概率密度函数统计量(PDF)常用于描述连续随机变量的强度。 为了便于观测，我们采用高斯核密度估计(GKDE)作为核函数，得到平滑的曲线。

3 图显示了来自四家制造商的两种型号的SMART属性的值的PDF。 两个磁盘模型的SMART数据的分布是不同的，但相似的，因为它们在不同的点和大小上表现出相似的尖峰。 我们将这种现象称为相关预测因子之间的协变量转移47

[1] https://wbackblaze.com/b2/harWWD-Drive-test-data.html

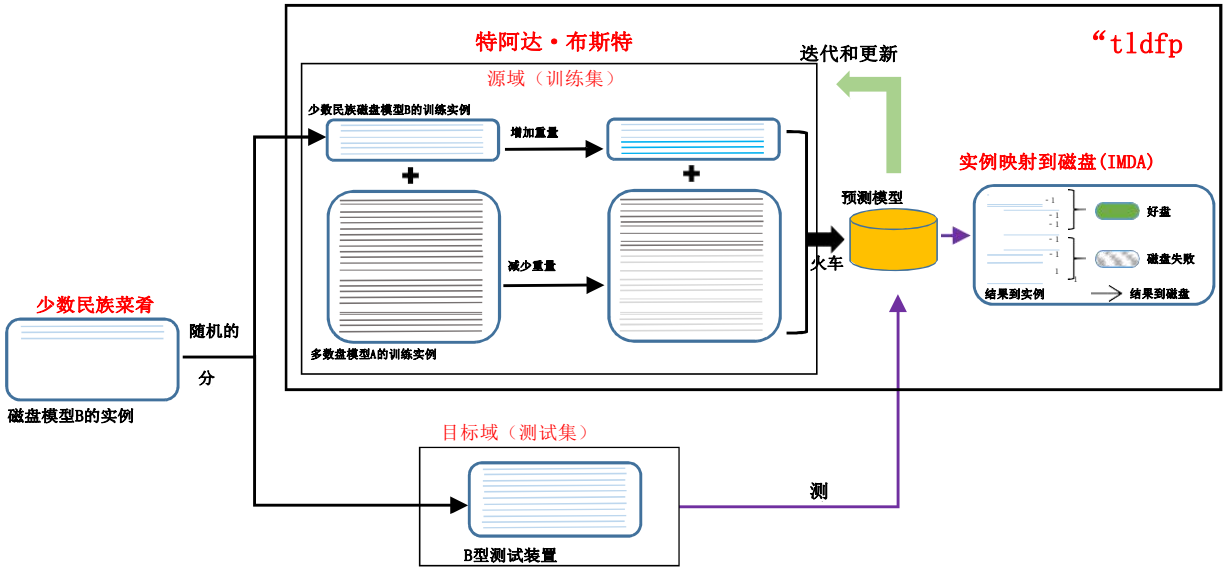


图4：TLDFP的总体结构，其中包含传输学习算法TrAda Boost和实例映射到磁盘算法(IMDA)。 **源域包含大多数磁盘模型A的完全标记数据集和少数磁盘模型B的一小部分标记数据集。 目标域中的测试数据是少数磁盘模型B的剩余未标记数据集。**

来自同一制造商的不同型号之间。 因此，我们得出结论，来自同一制造商的不同磁盘模型表现出不同的SMART值分布。 对于少数磁盘故障预测问题，其含义是，使用来自一个磁盘模型的训练数据建立在传统ML算法基础上的预测模型不适用于其他不同的模型，即使来自同一制造商。 因此，为了利用建立在足够的SMART训练数据基础上的磁盘模型的预测模型来建立一个仅有有限训练数据的不同模型的预测模型，我们可以采用传输学习算法，该算法本质上适合于将健康状态信息从一个磁盘模型转移到同一制造商的另一个磁盘模型。

考虑到上述异常和正常磁盘状态之间关系的规律性，以及不同磁盘模型之间不同的SMART数据分布谱的变化，我们被激励应用传输学习来预测少数磁盘的磁盘故障，使用来自大多数磁盘的知识，我们称之为TLDFP。

4 少数磁盘故障预测

在本节中，我们将回答如何和何时使用转移学习来预测少数磁盘故障的问题。 具体来说，我们详细介绍了少数磁盘故障预测TLDFP的转移学习方法，然后是基于KLD的源域选择方法。

4.1 TLDFP：用于少数磁盘故障预测的传输学习

在本节中，我们详细介绍了如何使用转移学习来预测少数磁盘故障。4 图说明了我们提出的预测方法TLDFP的总体结构。 它由两个主要组件组成：传输学习算法TrAdaBoost[]和实例映射到磁盘算法IMDA。5 注意，我们将少数磁盘模型B的SMART数据随机分为两部分。 第一部分包括标记目标域的一小部分(例如，10

数据，然后将其与大多数磁盘模型A的数据作为一个组合源域，建立两个不同磁盘模型之间的关系，以减少它们分布之间的变化。 另一部分包含剩余未标记数据作为测试数据。 然后，我们使用我们的TLDFP方法建立一个预测模型，并对少数磁盘模型B训练数据进行故障预测。 通过上述描述，我们在本文中要解决的问题可以被正式定义为：给定足够的标记训练数据S_a少量标记训练S_b 和未标记的测试数据T_b 主要目的是利用S的有用部分$_a$和S_b并训练分类器C，实现了对未标记训练集T的分类性能$_b$. TrAda Boost是传统ML方法Ada Boost的扩展。 Ada Boost是一种迭代算法，其关键过程包括训练几个不同权重的弱分类器，然后将这些弱分类器合并到强分类器中，以提高预测性能。 根据AdaBoost算法，它首先给所有训练实例一个初始权重。 当发现源域中的实例被错误分类时，我们认为这个实例很难分类，从而增加了它的权重。 这样，在下一次迭代中，这个实例的意义将变得更大。 然而，AdaBoost是一种传统的ML方法，它只能为与训练数据具有相同分布的测试数据建立有效的预测模型。 在转移学习算法TrAdaBoost中，当来自组合源域的磁盘模型B的实例被错误分类时，我们在下一次迭代中增加了这个实例的权重，这类似于AdaBoost。 然而，当磁盘模型A的实例被错误预测时，假设该实例与磁盘模型B不同。 因此，与Adaboost不同，它在下一次迭代中降低了该实例的权重，以减少其对目标域的影响。

TrAda Boost算法的输入包括两个磁盘模型'
训练和测试数据，以及最大迭代次数T。

它初始化训练数据的权重并执行迭代过程。在每个周期中，我们使用基本学习者，如GBRT、RGF、SVM、RNN和权重分布I^t构建分类器h_b在测试数据和计算标记源域磁盘模型B数据集上的错误率$_b$。 最后，我们根据先前的迭代结果和错误率来设置新的权重。 注意，如果源域的多数磁盘模型A实例被错误分类，则它们被认为与少数磁盘模型不同

b. 因此，我们减少了这些实例的权重，以减少它们在下一次迭代中对预测模型的影响。

具体来说，我们将实例乘以w ϕ 在哪里

Φ范围从0到1和c（•） 是SMART属性的真实标签。

另一方面，如果组合源域中的磁盘模型B实例被错误分类，我们将增加这些实例的权重，以便在下一次迭代中通过mul-获得更多的注意-

用w提示这些实例 ϕ 在哪里Φ$_t$是更大的

比1。 经过几次迭代(我们将最大迭代次数设置为22次，尽管我们将研究这个值如何影响Section中的预测性能，但适合少数磁盘模型B的源域中大多数磁盘模型A的实例将获得更大的权重，而与磁盘B不同的实例将具有更小的权重。6.4)，

由于故障预测模型的输入包括许多不同时刻来自许多磁盘的SMART实例，因此每个输出结果都指示特定实例的预测结果，而不是磁盘健康状态。 因此，我们需要将多个SMART实例的结果映射到最终的磁盘状态。 *为此，我们提出了一种实例映射到磁盘算法(IMDA)。* IMDA通过考虑磁盘的所有SMART实例来确定磁盘的最终健康状态。 具体来说，如果任何实例被归类为失败，则相应的磁盘被视为失败，其他替代选项将在本节中讨论6.5.

4.2 基于KLD的源域选择何时可以使用TLDFP进行少数磁盘故障预测？

为了回答这个问题，我们使用KullbackLeibler发散(KLD)，它是一个度量一个概率分布从另一个预期概率分布的发散程度的度量[]。38 KLD值表示两个随机变量分布之间的差异。 零KLD值意味着两个随机分布是相同的，而KLD值随着两个随机分布之间的差异扩大而增加。 一般来说，KLD值越大，两个分布之间的差异就越大，两个分布之间的知识转移就越困难。3 表给出了对应于图中所示PDF的KLD值，显示所有KLD值都不等于零，证实了各自的SMART数据分布确实不相同。3. 3 3, 3. 请注意，如表所示，KLD值趋势与PDF差异从HDS增加到STX，到HGST，再到图中所示的WDC是一种减少源域和目标域之间数据分布差异的方法，因此我们推断，一个磁盘模型与另一个磁盘模型之间的KLD值越大，TLDFP就越难传输经验。 5.2 节中的预测结果证明了我们的猜想，我们还在节中进行了详细的实验和讨论6.1.

因此，KLD的值可以指导我们选择合适的多数磁盘数据集（源域）来训练少数磁盘故障预测模型。 据我们所知，我们第一次尝试提出一种基于KLD值的新方法，作为选择适当的多数磁盘模型和改进磁盘故障预测的有效指标。6.1 我们在章节中的评估结果表明，我们使用KLD的方法是非常有效和实用的。

表3：图中PDF的KLD值3

源域	目标域	智能属性	KLD
hds-a	hds-b	5_raw	0.61
STX-a	STX-b	190_raw	0.89
HGST-a	HGST-b	197_raw	1.35
WDC-a	WDC-b	194_raw	0.56

5 实验评价

在本节中，我们评估了TLDFP的预测性能。 我们首先描述方法，然后是实验

最先进的转移学习方法，根据评估指标。

5.1 方法

我们在实验和智能属性选择中描述了两个真实世界的SMART数据集的特征。 然后，我们介绍了ML中常用的四种评估指标和我们用于进行所有实验的一些测试方法。

5.1.1 ***数据集和属性选择。*** 数据集：我们使用来自实际数据中心的两个SMART数据集进行评估。4 表给出了这两个数据集的总体特征。 每个磁盘都被归类为“好”或“失败”。 “样本”表示SMART记录。 每个好的磁盘或失败的磁盘都有许多SMART记录。 由于原始数据集具有比故障磁盘更多的良好磁盘样本，因此我们使用大多数类下采样来改进在不平衡类的情况下的训练，以创建训练数据集。 我们选择了故障盘与良好盘的1：3的比率[]。12 在传统ML方法进行训练时，我们将数据分为70%的训练和30%的测试数据，这与现有的工作是一致的[]。22 **请注意，我们的实验的所有结果都是通过交叉验证[]获得的，以避免在ML中常见的偶然事故。**375 表列出了我们选择的磁盘进行评估。

表4：智能数据集

数据中心	期限	很好	很好的样本	失败了	故障样本
回到布拉兹	50个月	141,891	106,867,099	7,657	7,689
腾讯	26个月	68,436	774,994,430	2795	31,574,341

表5：用于评估的选定磁盘模型

数据中心	制造商	磁盘模型	很好	很糟糕
呆头呆脑	Hitachi	hds722020ala330	4774	225
		hds723030ala640	1048	72
	STX	标准4000dm000	37006	3157
		st4000dx000	222	81
腾讯	hgst	HGST-a	13367	451
		HGST-b	679	63
	WDC	WDC-a	6847	259
		WDC-b	472	42

智能属性选择：每个智能观察可以包含多达30个有意义的智能属性。 然而，有些属性与我们的磁盘故障预测模型无关，因为

它们是不可变的，或者没有经历过明显的异常变化。 因此，我们根据基于原理组件分析(PCA)的特征选择过程，选择性地保留与磁盘健康状态相关的属性，而忽略其他无关属性。 所选的智能属性列在表中，对于每个智能样本，我们使用归一化值和原始值。6. 归一化值通常表示属性的当前值。 然而，当从原始值转换时，某些归一化值失去了准确性，一些原始值对预测模型更敏感。 我们还使用其他的特征选择方法，如[，，]。21222 然而，对实验结果的影响并不显著，因此由于空间有限，我们没有进一步讨论。

此外，不同的SMART属性具有不同的输出范围，这将导致对预测模型的不同影响。 为了在我们的磁盘故障预测模型中对不同的SMART属性进行公平的比较，我们使用符合现有工作的最小-最大缩放来规范所有选定的SMART属性的范围[12]:

$$x_{规范} = \frac{x - x_{敏}}{x_{马克斯} - x_{敏}}$$

其中x是智能属性x的原始值$x_{马克斯}$和$x_{敏}$分别是训练集中属性的最大值和最小值。

表6：为我们的评估选择SMART属性

#id	智能属性名称	属性类型
001	原始读取错误率	标准化和生
003	旋转时间	正常化
005	重新分配的扇区计数	标准化和生
007	寻找错误率	标准化和生
009	供电时间	标准化和生
184	I/O错误检测和校正	标准化和生
187	报告的不正确的错误	标准化和生
188	命令暂停	生的
189	高飞写	标准化和生
190	气流温度	标准化和生
193	负载/卸载循环计数	标准化和生
194	温度	标准化和生
197	当前待定区计数	标准化和生
240	头部飞行时间	生的
198	离线不正确的扇区计数	标准化和生
241	总LBA写的	生的
242	总LBA阅读	生的

5.1.2 . 评估度量我们使用以下四个度量来报告我们的实验中的结果，这些结果通常用于评估ML中分类模型的能力[42].

故障检测率(FDR)也称为召回率。 它捕获了正确预测为失败的真实失败磁盘的比例。 FDR越高，模型越好。

FAR：错误警报率(FAR)，错误预测为失败的良好磁盘的比例。远越低，模型越好。

F-Score：F-Score是两个度量FDR和预测精度(PP)之间的平衡。 PP是被正确预测为失败的预测失败磁盘的比例。 F-Score越高，模型越好。

AUC-ROC曲线：曲线-接收器操作特性(AUC-ROC)曲线下的区域是在不同阈值设置下分类问题的性能测量。 ROC是一条概率曲线，AUC表示可分离程度或度量。 它是用FDR绘制的，其中FDR在y轴上，FAR在x轴上。 在磁盘故障预测中，AUC较高意味着模型更好地区分故障和良好的磁盘。

5.1.3 测试方法和配置。 为了验证我们提出的TLDFP的有效性，我们在三种情况下进行了实验：1)使用仅在少数磁盘数据集上训练的传统ML方法；2)将TLDFP与传统ML技术（基线)进行比较；3）将TLDFP与其他传输学习方法进行比较。 设置如下所述。

1) 传统的ML方法只在少数磁盘上训练：

详细说明见章节3.2

2) 传统ML技术的TLDFP：

在这种情况下，我们进行实验，研究基于四种传统ML算法的少数磁盘故障预测的性能，使用大型异构数据集进行来自不同磁盘模型和同一磁盘制造商的培训。 7 表显示了培训和测试数据集。请注意，我们随机选择10%的测试数据集进行所有训练数据集的训练。

3)与其他转移学习方法的TLDFP：

除了传统的ML技术外，我们还将我们的TLDFP与两种最先进的传输学习方法(SSDB和TLBN)进行了比较，以预测少数磁盘故障。请注意，我们在[]和[]中对这两种方法使用相同的数据集，以便在所有实验中进行公平的比较。222

5.2 实验结果

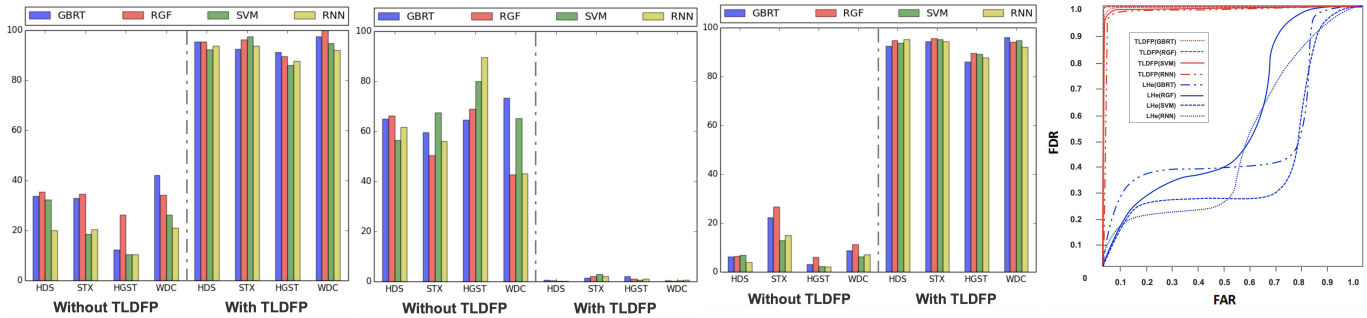
在这一部分中，我们展示了TLDFP与传统ML方法和其他转移学习方法的结果，其中分别提到了四个评价指标。 5.1.2 **请注意，我们已经显示了使用传统ML方法的不良基线结果，仅在节中的少数磁盘数据集上进行了训练。3.2**

5.2.1 与传统ML方法相比的评价。

- FDR/Recall Rate：我们利用两个真实数据中心的四个磁盘模型，对TLDFP和四种流行的传统ML方法的FDR进行了实验研究。5(a)， 从图中可以看出，四种传统的ML方法都不能使用大型异构数据集提供高FDR。 然而，TLDFP分别使用上述GBRT、RGF、SVM、RNN算法作为基本学习者，都获得了较高的FDR。
- FAR：请注意，我们的TLDFP的目标不仅是实现高FDR，而且对于少数磁盘故障预测也是低FA。5(b)。 图中显示了FAR的结果，所有其他方法都显示出更高的FAR，这在现实的数据中心是不可接受的。 此外，除了TLDFP之外，四种传统的ML方法都不能在少数磁盘上同时提供高FDR和低FAR。3.3， 通过分析，我们知道传统的ML方法造成的预测性能差，不具备减少少数群体之间分布差异的能力

表7：基于传统ML的大型异构数据集的少数磁盘故障预测数据集

数据中心	制造商	培训磁盘模型	测试磁盘模型	培训组	测试装置
呆头呆脑	Hitachi	hds-a	hds-b	良好的HDS-A和225失败的HDS-A 良好的HDS-B和7失败的HDS-B	良好的HDS-B和65失败的HDS-B
	STX	STX-a	STX-b	良好的STX-A和3157失败的STX-A 22个良好的STX-B和8个失败的STX-B	200个好的STX-B和73个失败的STX-B
腾讯	hgst	HGST-a	HGST-b	13367良好的HGST-A和451失效的HGST-A 良好的HGST-B和6失败的HGST-B	良好的HGST-B和57失败的HGST-B
	WDC	WDC-a	WDC-b	良好的WDC-A和259失败的WDC-A 良好的WDC-B和4失败的WDC-B	良好的WDC-B和38失败的WDC-B



FDR（越高越好） 远（越低越好） (c)F-Score（越高越好） AUC-ROC

图5：与四种传统的ML方法相比，基于TLDFP的四种磁盘模型的FDR、FAR、F-Score和AUC-ROC的结果。

目标域中的磁盘数据集和源域中的大多数磁盘数据集。

- 5(c)F-Score：图使用四个磁盘模型比较了两个数据集上不同预测模型的F-Score。 可以看出，TLDFP比其他不同的传统ML方法具有更高的F-Score。 例如，以RBF为基本学习者的TLDFP(RBF)的F核几乎是腾讯数据中心WDC磁盘算法RBF的9倍。4.2， 正如我们在章节中回顾的那样，HGST数据集的较大KLD值将导致更困难的知识转移。 这一结论也得到了F-Score观测的证实，因为HGST的TLDFP的F-Score通常比其他情况低。 我们将在本节中进一步详细讨论这个问题6.1.
- 5(d)AUC-ROC曲线：我们使用腾讯WDC磁盘模型绘制图中的AUC-ROC曲线。 如图所示，TLDFP的AUC-ROC曲线都接近左上角，与其他两种迁移学习方法相比，TLDFP(RGF)获得了更高的AUC值。 四种传统的ML方法获得了较低的AUC值，反映了它们在执行跨盘模型故障预测方面的分类能力差。

5.2.2 与其他迁移学习方法相比的评价。8， 如表TLDFP所示，FDR/FScore/和FAR比SSDB和TLBN高。 原因是，虽然SSDB将源域的分布与目标域匹配，但它只对源域中的观测值进行排序，而TLDFP对每一次观测值进行更有效的权重调整。 考虑到TLBN是一种多源域转移学习方法，TLDFP是一种单源域转移学习方法，我们分析了源域中每个磁盘模型与目标域中少数磁盘模型之间的KLD值。9. 结果如表所示，我们在磁盘模型中找到了源域中KLD值较大的数据（如ST320005XXXX和ST1500DL003）。 然而，TLDFP只使用KLD值最小的磁盘模型作为源域。

一个大的KLD值导致TLBN在减轻源域和目标域之间的分布差异方面遇到困难。 这一结果还表明，单源域转移学习比多源域转移学习性能更好，并且有一个很好的度量(例如KLD)来评估不同域之间的差异。 **请注意，由于空间限制，我们不包括所有方法和磁盘模型的结果。 从我们所有的测试中，TLDFP显示了最佳的性能。.**

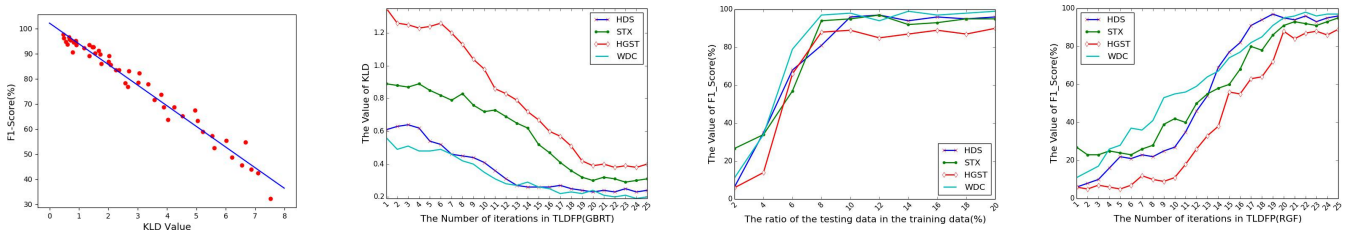
表8：TLDFP与SSDB、TLBN、FLBN、F-Score和AUC的比较

方法	制造商	罗斯福	很远	F-Score	奥克
tldfp(RGF)vsssdb	STX	94.9%/85.8%	1.6%/3.0%	92.6%/83.7%	0.93/0.86
	Hitachi	97.1%/70.8%	0.9%/5.4%	95.7%/69.0%	0.96/0.81
TLDFP(RGF)与TLBN	STX	91.3%/73.1%	0.6%/2.6%	91.3%/70.4%	0.91/0.83

表9：使用方法TLBN在训练集中的每个磁盘模型和测试集中的少数磁盘模型之间的KLD值

培训磁盘模型	测试磁盘模型	智能属性	KLD
圣320005xxxx	圣33000651as	5_raw	5.6
圣32000542as		190_raw	2.7
圣1500d1003		188_raw	7.1
圣31500341as		190_raw	1.5
圣31500541as		197_raw	0.83

总之，结果表明，TLDFP可以有效地解决少数磁盘故障预测问题，比传统的ML方法和其他两种相同数据集的传输学习方法具有更好的预测性能。 更具体地说，TLDFP不仅提供高的FDR、F-Score、AUC，而且同时显示出相当低的FAR。 比较结果的主要原因是我们的TLDFP算法能够利用少量的标记目标磁盘数据来建立源和目标磁盘模型之间的关系，这有助于大型异构磁盘模型对少数目标磁盘模型的特性进行良好的训练。6.2. 换句话说，TLDFP减少了源域和目标域之间的数据分布差异，我们将在第Note部分进一步讨论，我们已经验证了TLDFP在SSD和NVMe上的性能，并收集了一些有希望的结果，由于空间限制，这些结果不包括在内。



(a) (b) (c) (d)

图6：(A)KLD和F-Score的拟合曲线。 (b)培训中KLD值下降。 (c)FDR与百分比之间的关系绘制到源域的目标域数据。 (d)FDR与TLDFP中迭代次数之间的关系。

表10：F1-Score随KLD值的变化而变化

数据中心	方法	培训模式	测试模型	KLD	F1-Score
后面的火焰	*TLDFP(GBRT)*	hgst-k	HGST-1	2.67	76.8%
			HGST-m	1.76 ↓	85.9% ↑
			hgst-n	0.91 ↓	93.5% ↑
			HGST-O	0.69 ↓	95.7% ↑
			HGST-p	0.47 ↓	97.6% ↑
腾讯	*tldfp(svm)*	STX-k	STX-1	3.57	71.6%
			STX-m	2.58 ↓	78.2% ↑
			STX-n	2.26 ↓	83.4% ↑
			STX-O	1.35 ↓	93.1% ↑
			标准普尔	1.17 ↓	92.2% ↓
			STX-Q	0.71 ↓	95.3% ↑
			STX-r	0.66 ↓	96.6% ↑

6 意见和敏感性研究在本节中，我们从五个方面提供了几项额外的敏感性研究。

6.1 KLD在源域选择中的影响

4.2 5.2.1 5.2.2，为了验证我们在Section和Section中的猜想，并进一步解释Section和Section中的结果，我们分析了TLDFP中KLD和F-Score之间的关系，使用几个少数磁盘模型作为测试磁盘模型和一个磁盘模型的大型异构数据集，用于在两个真实数据中心从同一制造商进行培训。6(a)10. 结果如图和表所示，我们观察到F-Score的值随着KLD值的降低而不断上升。 换句话说，它表明SMART数据分布在源域和目标域之间的差异越小，TLDFP中的知识转移就越容易。 因此，我们使用KLD作为大型异构数据集源域选择的有效指标，通常选择KLD值最小的数据集。

6.2 TLDFP中KLD值的变化

6(b)。 为了更直观地观察TLDFP方法如何减少源域和目标域之间的数据分布差异，我们记录了TLDFP(GBRT)训练过程中模型每次迭代后，训练集和数据集之间的KLD值，如图所示，我们可以看到，随着模型迭代次数的增加，源域和目标域之间的KLD值不断减小，当迭代次数约为22次时，KLD值稳定在较小的值。 这表明，我们的TLDFP模型在训练过程中不断减少了两个域之间SMART数据分布的差异，使我们能够利用大型异构磁盘数据来预测少数磁盘数据，实现知识转移。

6.3 来自目标域的可变样本

正如前面所讨论的，我们的预测模型TLDFP使用目标域中标记数据集的一部分作为其源域的一部分数据集。 在本节中，我们研究了百分比数如何影响TLDFP的预测性能。 我们报告了以RGF为基本学习者的TLDFP的结果，并利用腾讯数据中心的WDC模型的磁盘数据，因为其他三个基本学习者显示了类似的结果。6(c) 图显示了FDR与将目标域数据放入源域的百分比之间的关系。 它清楚地表明，当百分比从2%增加到10%时，FDR急剧增加%。 当百分比超过10%时，FDR不会继续增加，而是保持一个相对较高的水平，这意味着将更多的目标域数据放入源域并不有助于进一步提高预测性能。 因此，我们在以前的实验中随机选择10%的目标域数据。

6.4 迭代的影响

Tr Ada Boost算法使用一个输入参数来限制算法执行的迭代。在这一部分中，我们研究迭代参数如何影响F-Score的预测性能。 我们报告了以RGF为基本学习者的TLDFP的结果。6(d) 图显示了F-Score如何随着不同数据集的迭代次数的变化而变化。从这个数字，我们可以作出类似的观察，如前一小节。 随着迭代次数的增加，F-Score快速增加，在22次迭代中达到稳定水平。 它还表明，随着算法在每次迭代中调整实例权重，TLDFP逐渐调整以收敛到测试数据。 由于22表示一个反射点，我们在前面的实验中采用了这个迭代数。

6.5 IMDA的敏感性研究

4.1，在部分中，我们介绍了IMDA算法。 特别是，如果任何实例被归类为失败，则相应的磁盘被视为失败。 在这里，我们进行实验来评估不同实例的影响，具体来说，1个实例，所有实例的1/3，所有实例的1/2，所有实例的2/3，以及所有被归类为失败的实例。

11. 本实验使用基于腾讯WDC磁盘模型数据的TLDFP(RNN)的结果如表所示，很明显，我们使用的选项（一个实例失败表示相应的磁盘故障）达到了最佳的性能。

表11：IMDA在TLDFP中的敏感性研究结果

方法	计量	1	1/3	1/2	2/3	全部
tldfp	罗斯福	95%	32%	24%	8%	3%
	很远	0.7%	0.7%	0.5%	0.2%	0.2%

7 结论

在本文中，我们开发了一个名为TLDFP的模型，以有效地预测少数磁盘故障，利用传输学习，其中

传统的ML方法表现不佳。 我们的主要贡献包括：（1)我们是第一个通过在异构磁盘模型的数据中心进行广泛的数据分析和实验来定义少数磁盘数据集并对其进行定量评估的人；(2）我们是第一个提出了一种基于KLD值的新方法来选择适当的多数磁盘模型；（3）我们开发了一种交叉磁盘模型故障预测方法，它具有重要的实际适用性，因为不同的磁盘模型逐渐被放置到现实的存储系统中来替换失败的磁盘。 我们对来自现实世界数据中心的两个数据集的实验表明，TLDFP在故障检测率和虚警率方面优于具有代表性的传统ML方法和现有的转移学习方法。 在进行交叉盘模型故障预测时，TLDFP平均达到96%的故障检测率，只有0.5%的虚警率。

感谢

这项工作得到了国家自然科学基金创新小组项目的支持。 61821003，国家重点研究开发项目批准号。 2016YFB0800402和美国。 国家科学基金资助CCF-1717660和CCF-1813081。 我们衷心感谢王元璋和王阳涛，他们给了我建议和时间听我说话。 感谢南京审计大学的胡立汉、孙周宝和徐振元的帮助。 黄平是本文的相应作者。

参考资料

[1] 布鲁斯·艾伦。 2004. 用SMART监视硬盘。 *Linux Journal117（2004），60-65。*

[2] Mirela Madalina Botezatu和Ioana Giurgiu等人。 2016. 预测磁盘替换到可靠的数据中心。 *在第22届ACM SIGKDD会议记录，旧金山，加利福尼亚州，美国，8月13日至17日。* 39－48.

[3] 布拉德·卡尔德和鞠王等人。 2011. Windows Azure Storage：一种高可用的云存储服务，具有很强的一致性。 *在第23届ACM SOSP会议记录，卡斯凯斯，葡萄牙，10月23日至26日。* 143－157.

[4] 科琳娜·科特斯和弗拉基米尔·瓦普尼克。 1995. 支持向量网络。 机器学习20，3（1995），273-297。

[5] 戴文元和杨强等人。 2009. 促进迁移学习。 在 *第24次ICML会议记录。 193－200.*

[6] Brian Kulis等人。 2011. 你看到的不是你得到的：使用非对称内核变换的域适应。 *在IEEE CVPR会议记录中。* 1785－1792.

[7] Bianca Schroeder等人。 2016. *生产中的Flash可靠性：预期和意外。 在美国加利福尼亚州圣克拉拉市第十四届美国东部，2月22-2567-80。.*

[8] 朱炳鹏等。 2013. 大型存储系统的主动驱动故障预测。 *在IEEE第29届MSST研讨会上，5月6日至10日，长滩，加利福尼亚州，美国。* 1－5.

[9] 程黄等人。 2012. 在WindowsAzure存储中的擦除编码。 *在USENIXATC会议记录，波士顿，马里兰州，美国，6月13日至15日。* 15－26.

[10] 常旭等人。 2016. 递归神经网络硬驱动的健康状况评估和故障预测。 TOC65，11（2016），3502-3508。

[11] Eduardo Pinheiro等人。 2007. 大磁盘驱动器的故障趋势。 在 *第五届USENIXFAST，2月13日至16日，美国加利福尼亚州圣何塞。 17－28.*

[12] Farzaneh Mahdisoltani等人。 2017. 利用主动误差预测提高存储系统的可靠性。 *在USENIX ATC会议记录中。* 美国圣克拉拉协会，加利福尼亚州，391-402。

[13] Haryadi S。 Gunawi等人。 2018. 大规模故障：大型生产系统硬件性能缺陷的证据。 国家统计局14，3（2018），23：1-23：26。

[14] 王衡等。 2011. 密集贸易的行动承认。 *在IEEE CVPR会议上。* 3169——3176.

[15] Joseph F Murray等人。 2003. *利用非参数统计方法预测硬盘驱动器故障。在ICANN会议记录中。* 1－4.

[16] 李静等。 2014. 利用分类树和回归树进行硬盘驱动器失效预测。 *在IEEE关于DSN的国际会议上。* 383——394.

[17] 李静等。 2016. 准确是不够的：磁盘故障预测的新度量。 *在第35届IEEESRDS，布达佩斯，匈牙利，9月26日至29日。* 71－80.

[18] 李静等。 2017. 硬盘驱动器故障预测使用决策树。 *关系。 英格。 &西斯。安全164（2017），55-65。*

[19] 贾斯汀·梅扎等人。 2015. 大规模研究领域内的闪存故障。 *在ACM SIGMETRICS会议记录，波特兰，OR，美国，6月15日至19日。* 177－190.

[20] 周天一等。 2014. 混合异质迁移学习通过深度学习。 *在28AAAI人工智能会议记录，7月27日至31日，魁北克市，魁北克，加拿大。* 2213－2220.

[21] Lakshmi N。 Bairavasundaram等人。 2007. 磁盘驱动器中潜在扇区错误的分析。 *在ACM SIGMETRICS会议记录，圣地亚哥，加利福尼亚州，美国，6月12日至16日。* 289－300.

[22] Pereira等人。 2017. 应用于硬盘驱动器故障预测的贝叶斯网络传输学习。 *在巴西智能系统会议上。* 228－233.

[23] 范荣恩等人。 2008. LIBLINEAR：大型线性分类库。 *机器学习研究杂志9（2008），1871-1874。*

[24] Tomas Mikolov等人。 2009. 基于神经网络的高选择性语言语言模型。 *在IEEE ICASSP会议记录中，4月19日至24日，台北，台湾4725-4728。.*

[25] Tomas Mikolov等人。 2011. 递归神经网络语言模型的扩展。 *在IEEE ICASSP会议记录，5月22日至27日，布拉格会议中心，布拉格，捷克共和国。* 5528－5531.

[26] Teerat Pitakrat等人。 2013. 主动硬盘驱动器故障检测的机器学习算法比较。 *第13届国际ACM IARCS会议记录。* 17－21.

[27] 蒋伟航等。 2008.磁盘是存储故障的主要贡献者 -全面研究储存子系统的故障特点。 国家统计局4，3（2008），7：1-7：25。

[28] 杨文军等。 2015. 利用大数据预测硬盘驱动器故障。 *在IEEE34thSRDS中。* 13－18.

[29] 永旭等。 2018. *通过预测磁盘错误来提高云系统的服务可用性。 在USENIXATC，波士顿，马里兰州，美国，7月11日至13日。* 481－494.

[30] 杰罗姆·H。 弗里德曼。 2001. 贪婪函数逼近：一种梯度升压机。 *统计年鉴29，5（2001），1189-1232。*

[31] Sandipan等人。 很好。 2016. 云平台硬盘故障预测的实用方法：数据中心故障管理的大数据模型。 *在IEEE第二大数据服务中。 105－116.*

[32] 格雷格·哈默利和查尔斯·埃尔坎。 2001. 磁盘驱动器故障的贝叶斯方法。 *在第18次ICML会议记录中。* 202－209.

[33] Maayan Harel和Shie Mannor。 2011. 从多元展望中学习。 在 *第28次ICML会议记录。 401－408.*

[34] 宋煌和宋赋等。 2015. 用量化的磁盘退化符号描述磁盘故障：早期经验。 *在IEEEIISWC，亚特兰大，GA，美国，10月4-6日。* 150－159.

[35] 戈登·休斯和约瑟夫·F·默里等人。 2002. 改进磁盘驱动器故障警告。 IEEEtor51（2002），350-357。

[36] 李强森和张彤。 2014.利用正则化贪婪森林学习非线性函数。 *IEEE Trans。 模式肛门。 马赫。 直觉。* 36，5 (2014)，942－954.

[37] 罗恩·科哈维。 1995. 交叉验证和引导的准确性估计和模型选择研究。 *在IJCAI。* 1137－1143.

[38] s. 卡尔巴克和R。 a。 莱布勒。 79－86. 论信息与充足性。 *数学统计年鉴22（79-86），1951年。*

[39] Joseph F。 穆雷和戈登·F。 Hughes等人。 2005. 硬驱预测故障的机器学习方法：一种多因素应用。 *机器学习研究杂志6（2005），783-816。*

[40] 辛诺潘和强阳。 2009. 迁移学习调查。 IEEEtkde22（2009），1345-1359。

[41] 大卫·A。 帕特森和加思。 吉布森等人。 1988. *昂贵磁盘冗余阵列的案例)。 在1988年ACM SIGMOD会议记录，芝加哥，伊利诺伊州，美国，6月1-3日。* 109－116.

[42] 大卫·鲍尔斯。 2007. 评价：从精密度、召回率和F-度量到ROC、知情度、标记性和相关性。 *机器学习技术杂志2（012007），37-63。*

[43] 彼得·普雷滕霍夫和本诺·斯坦。 2010. 跨语言文本分类使用结构对应学习。 *在第48届ACL会议记录中。* 1118－ 1127.

[44] Felix Salfner和Maren Lenk等人。 2010. 在线故障预测方法的调查。 *ACM Comput。 幸存者。* 42，3 (2010)，10:1－10:42.

[45] Kanoksri Sarinnapakorn和Miroslav Kubat。 2007. 文本分类中的子量词组合：基于DST的解决方案和案例研究。 IEEEtkde19（2007），1638-1651。

[46] 比安卡·施罗德和加思·A。 吉布森。 2007. 现实世界中的磁盘故障：1000，000小时的MTTF对你意味着什么？。 *在美国加利福尼亚州圣何塞市，2月13日至16日，第5次USENIX飞行。* 1－16.

[47] h. Shimodaira。 2000. 通过加权对数似然函数来改进协变量移位下的预测推理。 *统计规划和参考杂志90，2（2000），227-244。*

[48] Igor V。 泰特科和大卫·J。 Livingstone等人。 1995. 神经网络研究，1过度拟合和过度训练的比较。 *化学信息和计算机科学杂志35，5（1995），826-833。*