在网络安全中实施主动学习：检测删节邮件中的异常

Mu-Huan (Miles) Chung[a]、Lu Wang[a]、Sharon (Siyuan) Li[a]
Yuhong (Alisha) Yang[c]、Calvin Giang[c]、Khilan Jerath[c]、Abhay Raman[c]David Lie[b]、Mark Chignell[aa]机械与工业工程系，多伦多大学[b]电气与计算机工程系，多伦多大学[c] 企业服务部，永明金融集团

arXiv:2303.00870v2 [cs.HC] 3 Mar 2023

摘要—**电子邮件异常检测研究通常依赖专门构建的数据集，而这些数据集可能无法充分反映行业环境中的实际数据类型。在我们的研究中，由于一家大型金融服务公司存在隐私问题，我们无法查看电子邮件正文和附件详情（但可以查看主题和附件文件名）。这使得在由此产生的删节邮件中标注可能存在的异常更加困难。另一个困难来源是电子邮件数量庞大，而资源又十分有限，因此机器学习（ML）成为必要之选；与此同时，也需要更高效地对人员进行培训，使其能够训练机器学习模型。主动学习（AL）被提出作为提高人员训练机器学习模型效率的一种方法。然而，由于可能存在人工分析员不确定性，主动学习方法的实施是一项以人为中心的人工智能挑战；此外，在网络安全领域（或医疗保健、航空等领域），标注任务可能更加复杂，因为标注错误可能造成极其严重的后果。本文介绍将主动学习应用于删节邮件异常检测的研究结果，并比较了在这一情境下实施主动学习的不同方法的效用。我们评估了不同的主动学习策略及其对最终模型性能的影响。我们还探讨了专家对其标签的置信度评分如何为主动学习提供参考。最后，我们结合这些结果对主动学习方法及专家在模型辅助邮件异常筛查中的作用所具有的意义进行了讨论。**

索引词—《**机器学习》、主动学习、人机交互、邮件外泄、内部威胁。**

I. 引言

大型组织面临着来自各种渠道、日益严峻的网络安全威胁。对手借助日益复杂的工具，伺机寻找可供利用的薄弱环节。由于恶意代理足智多谋，攻击手段也层出不穷，要防范所有可能的威胁十分困难。仅仅强化组织边界并不足够，因为许多威胁来自内部人员，或通过社会工程学等各种方法获取有效凭据的恶意外部人员（即所谓的伪装者）。粗心大意的内部普通用户也可能泄露敏感信息，这会构成重大威胁，尤其是在员工居家办公时。因此，组织需要在有意或无意的数据外泄流程造成重大损害之前，及时发现并加以阻止。

当前网络安全领域的最新技术水平采用《机器学习》（ML）方法来检测电子邮件异常。异常检测模型可以支持检测不当行为，例如包含敏感信息的外发邮件。然而，对于机器学习模型而言，始终存在如何获取或生成已标记实例训练集的问题。在网络安全领域，安全与隐私问题是首要考虑因素，因此很难获取高质量的已标记数据。通常，大量传感器/嗅探器会生成噪声数据集，这些数据集难以进行预处理和标注。没有适当的标注，就不可能训练监督模型并准确检测异常活动。

此外，在网络安全领域训练和部署机器学习模型还受到臭名昭著的“不平衡数据”问题的阻碍。由于恶意活动往往很少见或隐蔽，因此很难构建可用于训练机器学习模型的已标记实例集合。使用不平衡实例集训练模型可能会产生大量误报。因此，分析人员可能需要每天处理大量警报，以确定被标记的实例中哪些属于真实异常。

主动学习（AL）是一种交互式学习方法，旨在改进数据标注和模型训练过程。主动学习过程通常只需少量已标记数据来训练初始学习器模型，之后便可开始从人工反馈中学习。学习器模型会向人工分析人员查询那些导致模型不确定性最大的案例，以提升预测/分类性能。在实践中，主动学习对位于决策边界上或距其稍远的案例可能效果更好，具体取决于实例数据的分布特性等因素。

主动学习已广泛应用于数据标注，并被证明能有效降低标注成本 [1,2]。它可支持机器学习数据准备，并帮助机器学习模型克服棘手挑战。然而，尽管主动学习已在各种非专家标注任务中经过测试并得到应用，但对于需要专业知识的标注任务，其表现尚未得到充分研究。

供盲审使用的资金来源信息已匿名化

专家标注任务可以视为一种通用的标注任务，但标注人员未必总能确定自己赋予的标签。例如，在某些图像识别任务中，如果待标注图像中的物体分辨率相对较低，标注人员可能难以赋予标签。以往研究指出，在这种情况下，适当的做法是收集“软”标签 [3],，例如汇总一名或多名标注人员给出的常规标签，并将其转换为概率分布，从而根据似然最高的类别确定实际标签[4,5]。

专家对自己给出的标签可能并不完全有把握，原因可能是他们在相关领域的知识有限，也可能是数据隐私不允许他们获取所有相关信息（电子邮件异常检测中经常会出现这种情况）。当标注不确定性较高时，主动学习是否仍然有效？这种不确定性不仅存在于机器学习模型一侧，也存在于负责标注的人类一侧。为研究如何应用主动学习，同时考虑专家和机器学习模型的标注不确定性，我们提出以下研究问题。本文余下部分将通过一项电子邮件异常检测案例研究来评估这些问题：

- **研究问题1**：专家给标注决策赋予自信度评分时，校准得如何？（也就是说，当专家对自己的标注决策更有信心时，模型预测是否会有所改善？）
- **研究问题2**：与个人组相比，使用群体训练机器学习模型是否能带来更好的模型性能？
- **研究问题3**：专家彼此之间的一致程度如何？不同专家在训练机器学习模型方面是否存在个体差异（训练质量以其后续模型的性能来衡量）？

研究这些研究问题有助于我们了解主动学习的有效性（假设无法获得预言机的完整知识），以及分析人员与机器学习协作的效率。在下文报告的研究中，我们开展了一项关于出站电子邮件异常检测的案例研究。在本文余下部分，我们将首先回顾主动学习领域。随后，我们将概述一项案例研究的方法、过程和结果。该研究考察了在具有挑战性的电子邮件异常检测流程中应用主动学习的情况，同时还收集了专家对其所分配标签的置信度评分。

II. 背景

标注成本一直是监督式机器学习模型训练中的一个主要问题。在许多领域，众包 [6,7]通过招募参与者为数字、图像或视频字幕进行标注来帮助解决这一问题。验证码 [8] 是一种广为人知的标注应用案例，最初被实施为一种类似图灵测试的安全检查，用于识别请求资源的用户是人类还是机器。随后，ReCaptcha 应用程序利用验证码登录挑战来支持大规模图书数字化 [9]。然而，尚不清楚如何在网络安全这类数据集通常私有且敏感的领域复制众包标注的成功。

众包专有标签会导致数据泄露给不明外部人员。此外，网络安全领域高度依赖专业知识，非专家参与者无法提供高质量标签；而出于成本考虑，招募大量分析人员也不现实。虽然也可以使用外部承包商，但这同样成本高昂，而且组织还需要将敏感数据托付给承包商。因此，主动学习（AL）颇具吸引力，因为它不需要很多参与者，而且组织可以利用现有人员开展主动学习。

主动学习通过反馈回路克服机器学习训练中的“瓶颈” [10]。对于许多本质上更适合人类大脑处理的任务，例如对象分割和视频标注，主动学习可以通过融入人类知识来提升性能。现在，我们来探讨可以采用的不同类型的主动学习场景。

A. 主动学习场景

主动学习可根据数据流的类型和目标任务采用不同的场景，主要有两种：流式场景和池式场景。流式场景最初被提出，作为某些任务中替代随机采样的一种采样策略，其前提是假设获取未标记数据的成本很低（或无需成本）[11,12]。主动学习器可以根据实例数据的分布特性，决定是否为即将到来的实例请求标签，或跳过该实例。由于需要针对数据流中的每个实例（或序列／行）作出决定，因此这种场景称为流式场景。

池式场景关注的是可以一次性获取大量未标记数据的情况 [10]。池式主动学习通常从一个使用少量已标记数据训练的初始学习器模型（初始化器）开始。随后，主动学习器会为大量未标记数据（即数据池）中的每个实例请求标签 [11,12]。实例一旦获得标签，就会从数据池中移除，主动学习过程随即继续查询下一个实例。

在流式场景和池式场景中，每次只请求一个标签有时都会带来问题。例如，对于电子邮件异常检测任务，为单封电子邮件请求标签可能会在两个方面耗费时间。首先，分析人员必须调查这一单个实例。调查过程可能需要大量支持，包括但不限于跨部门协作。其次，如果主动学习器背后的机器学习模型很复杂，那么每增加一个带标签的实例后都重新训练该模型，可能会造成较高的计算开销。因此，可以采用批量模式作为替代方案，一次查询多个实例 [13,14]。

批处理的概念使主动学习器能够同时抽取一批不确定样本（如果采用不确定性抽样）。这可能会显著降低两种传统场景的时间成本。然而，

基于批次的场景有时会导致批次中出现过多冗余，以至于某次迭代中几乎无法获得信息增益 [15]。一种可行的解决方案是选择涉及不同特征组合的多样化实例，以提高标注效率。这些实例通常与决策边界有一定距离 [16]。因此，“多样性查询”可以确保纳入多样化实例，从而减少冗余并提高查询效率 [17]。

B. 主动学习查询/采样策略

主动学习需要一种采样策略，而最常见的策略之一，是根据机器学习模型的不确定性 [18,19]，优先选择实例进行标注。采用这种不确定性最小化策略时，待标注实例队列会根据其不确定性排名进行排序，以便模型在每次标注判断后最大化信息增益。例如，如图 1 所示（二分类），靠近决策边界的橙色气泡更有可能在主动学习中被查询，因为它们的模型不确定性较高（这是由于它们接近决策边界）。

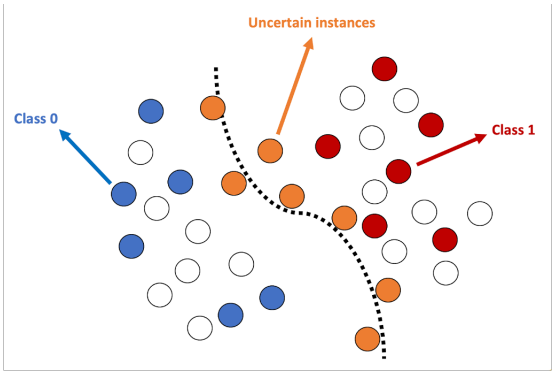

图 1. 二分类任务中的主动学习实现

有几种方法可以根据预测概率计算不确定性 [20]。例如，在三分类任务中，可以使用 Least Confident（公式 1）计算不确定性，其中 y' 是三个类别中最有可能的预测结果。如果各类别的概率分别为 [0.15、0.65、0.20]，则不确定性可计算为 0.35。

$$L(x) = 1 - P(y'|x) \qquad \text{(eq. 1)}$$

最不自信度的计算忽略了其他两个类别的信息，因此在许多情况下并非最优。因此，人们提出了另一种使用分类间隔（公式 2）的方式。该策略关注概率最高的类别y'与概率第二高的类别$y"$之间的差异。根据该策略，对于各类别预测概率分别为 [0.15、0.65、0.20] 的一行，其不确定性将计算为 0.45。

$$M(x) = P(y'|x) - P(y"|x) \qquad \text{(eq. 2)}$$

另一种常用的不确定性计算策略是计算分类熵（公式 3）。分类熵对应于期望对数损失 [20]，，它与找到真实类别 [21] 所需猜测次数的平均值成正比。

$$H(x) = -\sum_i P(x_i|x) log P(x_i|x) \qquad \text{(eq. 3)}$$

除了不确定性抽样 [10] 之外，还有其他查询策略。可以将多个模型组织成流水线，并组合使用不同模型（例如，委员会查询、分歧采样、期望模型变化或估计误差降低等），以提升每个反馈回路中的模型性能。因此，查询策略的选择以任务为导向，并会根据主动学习器所依赖的算法而变化。

虽然已有其他查询策略（例如， [22]-[25]）被提出，但不确定性驱动的采样策略仍然简单直接。在本文余下部分，我们将探讨如何在应用场景中改进主动学习的不确定性驱动采样，同时考虑人工标注者的不确定性。

III. 案例研究

本案例研究涉及一家大型金融服务公司。我们针对一项出站电子邮件异常检测任务，考察了主动学习的有效性和效率，以及专家与模型交互中的人为因素。我们重点关注违反公司政策的电子邮件行为（尤其是员工发送到自己电子邮件地址的邮件）。这类行为被认为存在问题，因为公司要求员工使用受限的公司电脑工作。员工可能会通过向自己发送电子邮件，秘密地将数据传输到自己使用的其他（不安全）电脑上，从而规避这项政策。

在研究开始之前，公司每周由一名安全分析师手动执行这项检测任务。该分析师会查看仪表板上的电子邮件，并将每封邮件/每个案例标记为异常或正常。仪表板提供筛选功能，分析师可以筛选出包含特定敏感词组合的电子邮件。筛选后，系统会通过两张散点图展示所有电子邮件的大小和所有用户的邮件数量，通过两张趋势图展示所选用户过去的电子邮件数量和大小变化记录，并通过一张详细表格展示所选用户当天的电子邮件活动。整个流程通常需要约两个工作日（每周）才能完成，其中包括从约 10,000 封电子邮件中进行筛选，以及筛查和审阅来自约 25 至 50 名不同用户的 1,000 多封筛选后邮件；这些用户发送的邮件数量众多和/或体积较大，且主题行或附件名称中包含某些敏感词。

此外，另一个拥有更高级别安全许可的团队还开展了第二阶段调查。后续调查包括审查详细的电子邮件属性、电子邮件正文、附件详情，有时还会与发件人的经理交谈。这一流程完全没有自动化

因此，审查每封电子邮件都很耗时。

整个手动检测过程效率低下，而且很可能容易漏检或产生假阳性。因此，需要通过机器学习干预来提高检测速度和准确性。此前，一家外部供应商曾尝试使用自动化机器学习（以往的手动检测结果作为训练/测试数据集）对公司电子邮件进行异常检测。然而，误报率高达 96%；经过进一步调查，该供应商标记的异常中只有 4% 被证实为真实异常。该外部供应商依赖公司提供的人工分配标签，我们怀疑标注质量不高可能是此次自动化机器学习预测性能不佳的原因之一。这促使我们考虑采用主动学习，更有效地利用人类知识来标注数据。

A. 范围与数据

本研究使用的数据集包含 27 列，用于描述每封电子邮件的信息，包括发件人、收件人、主题、附件、附件大小、发件人身份识别数据（角色、所在地、入职日期、状态等）、日期时间变量，以及主题行或附件名称中是否提及某些敏感词。此外，我们还统计并汇总了收件人数量。我们还加入了两个二元变量，分别表示电子邮件的主题行（第一个变量）和附件名称（第二个变量）中是否包含某些敏感词（这些词此前曾用于手动检测任务）。我们还计算了电子邮件地址与用户名之间的相似度，作为代理指标，用于衡量某人是否可能在向其其他电子邮件账户之一发送电子邮件（使用莱文斯坦距离 [35]）。随后，我们将这一相似度作为一项附加变量。

原始未标记数据集约有 320,000 行，其中包含两周的出站电子邮件。该原始数据集被拆分为八组未标记数据，并按照日期/时间顺序将每组分配至主动学习阶段的八个轮次，以模拟本案例研究中的真实检测任务。此外，还有一个由从原始数据中随机选取的 200 个实例组成的预标注数据集。后一个预标注实例数据集用于训练初始机器学习检测模型。

限制标注任务的一个因素是，提供给分析人员的数据不包含电子邮件正文和附件的详细信息。只有另一个团队（本研究未涉及该团队）获得了这些额外数据。该团队拥有更高级别的安全许可，负责在潜在异常被发现后进行调查。因此，在这一筛查/过滤阶段，人工判定人员只能根据主题行、文件名和一些可识别用户身份的信息做出标注决策。虽然这些变量足以检测疏忽导致的数据外泄，但也可能导致更多异常被标记出来（“以防万一”），进而造成更高的假阳性率。

B. 方法

本研究通过以下方式实施了主动学习（AL）过程：人工分析人员先自行标注数据，然后由模型预先标注数据。随后，模型根据分析人员提供的标签进行更新。该主动学习过程包含以下通常会采用的组成部分：

- 一小组预标注数据
- 使用预标注数据初始化的主动学习模型
- 选定的主动学习场景，以及与分析人员工作角色相适配的采样策略
- 终止准则（例如准确率阈值或准确率的渐近改进）
- 一种评估结果的方法

由200个案例组成的预标注数据子集用于初始化主动学习器（主动学习模型）。我们使用了由另一个团队准备的真实标签数据集，该团队可以访问电子邮件详情。我们采用了基于批次的场景，每轮由每位分析人员标注多个实例。本研究采用的整体主动学习过程如图2所示。

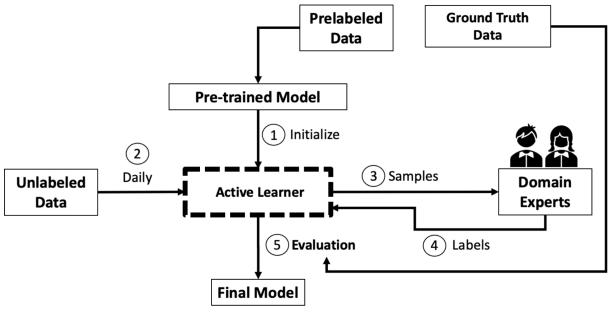

图2. 案例研究中采用的整体主动学习过程

主动学习面临的一个问题是， [26] 分析人员可能会因不得不标注算法无法确定的案例而感到沮丧，因为分析人员很可能也无法确定这些案例。因此，我们可以预期，当分析人员能够清晰地标注算法认为模棱两可的案例时，主动学习的效果最佳。可以想见，在实际情境中，分析人员是否愿意执行这项可能令人沮丧的任务，将限制他们对模棱两可案例的标注。

Carcillo 等人此前开展的研究 [27] 使用信用卡欺诈检测任务（这也是一项需要专业知识的主动学习异常检测任务）测试了多种主动学习采样策略。他们采用了最高风险查询（HRQ）策略，优先标注敏感度/风险最高的实例。

我们的方法以 Carcillo 等人描述的方法为基础，并根据案例研究的需求进行了调整。我们采用了高风险查询（HRQ）方法，让分析人员查看较高比例的预期异常。不过，为了探索性目的，我们还在每个反馈回路中加入了少量不确定和随机实例。采用 HRQ 方法的原因之一是

在人工分析人员审查大量异常时，这种方法可能更能激发他们的积极性。

1）参与者：实验共有 10 名参与者，他们代表了我们合作的工作组中所有具备所需专业知识的公司员工。参与者将标注任务作为工作职责的一部分来完成。实验产生的数据由同一家公司内的第三方收集，该第三方与参与者之间不存在利益关系，也没有利益冲突。第三方生成了 10 个匿名化身份标识，并分配给参与者。在整个实验过程中，研究人员只能访问这些身份标识和标签，因此无法将模型性能与特定参与者关联起来。相反，第三方可以将身份标识与参与者对应起来，但无法得知哪个模型是由哪位参与者训练的。

完成任务后，参与者可以选择是否愿意将其匿名化数据用于研究。该决定是匿名的；选择不完成标签，或不提供数据用于研究的人不会受到任何影响。此流程已获 X 大学（名称因盲审要求而隐去）伦理审查委员会批准（伦理审查协议编号 XXXXX——编号因盲审要求而隐去）。

标注任务作为参与者工作实践的一部分开展，共有 10 名参与者。对于每个需要标注的电子邮件实例，任务是审查所提供的数据（主题行、大小、文件名等），然后确定证据是否表明由于可能存在数据外泄活动（不当传输机密数据）而有必要进一步调查。随后，参与者为该实例提供标签（异常或非异常），并继续处理队列中的下一个电子邮件实例。在本案例研究中，我们使用 LightGBM [28]作为底层主动学习模型，因其在分类任务中的效率和性能。与 Carcillo 等人 [27] 成功采用的策略类似，我们使用了高风险查询（即机器学习预测模型相对确定为异常的实例）。

在研究的每一轮次（每天）中，每位参与者在 8 个工作日内每天标注 20 个实例，其中包括 14 个高风险查询（HRQ）、3 个不确定性查询（UQ）和 3 个随机查询（RQ）。可收集的数据量受限于公司主管认为应投入到这项工作中的时间量。原始数据被拆分为 8 个未标注数据集，以模拟现实中的日常标注/检测任务流程。本案例研究没有设置停止标准，因为我们希望更多地关注人类交互，而非寻求达到某一特定水平的模型性能。

2) 研究设计：我们的案例研究聚焦于以下构念/因素：

- 个体差异（在经验、专业知识等方面）
- 协作效应（多名分析人员训练同一个模型时）
- 人类不确定性（即分析人员在提供标签时自我报告的置信度，以及这是否会影响主动学习流程）
- 分析人员在与主动学习系统交互过程中发现的人为因素问题

为了评估可能存在的人为因素问题，并回答我们提出的研究问题，我们将参与者分为3个团队（不同模型以字母A至F标记，如下表1所示）：

- **个人组**：个人组中的每位参与者在整个实验过程中训练（提供反馈给）自己的模型（B、C和D）
- **交换组**：交换团队中的每位参与者在实验前半段训练自己的模型（E和F），并在实验后半段改为训练另一个模型（F和E）
- **小组**：小组团队的参与者共同训练了一个模型 (A)——在这种方法中，如果五名分析人员中有两名认为某个实例属于 1（真），该实例就会被标记为 1（真）。

这种实验设计使我们能够比较个人组成员彼此之间的表现，以及他们与以小组形式工作的参与者之间的表现。

<table>
<tr><td rowspan="2">Day</td><td colspan="10">Analyst ID (10 participants in total)</td></tr>
<tr><td>Grp_1</td><td>Grp_2</td><td>Grp_3</td><td>Grp_4</td><td>Grp_5</td><td>Ind_1</td><td>Ind_2</td><td>Ind_3</td><td>Swp_1</td><td>Swp_2</td></tr>
<tr><td colspan="11">Pre-train Phase (Train a baseline model to initialize the AL process)</td></tr>
<tr><td>1</td><td colspan="5" rowspan="8">Model A</td><td rowspan="8">Model B</td><td rowspan="8">Model C</td><td rowspan="8">Model D</td><td rowspan="4">Model E</td><td rowspan="4">Model F</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td></tr>
<tr><td>5</td><td rowspan="4">Model F</td><td rowspan="4">Model E</td></tr>
<tr><td>6</td></tr>
<tr><td>7</td></tr>
<tr><td>8</td></tr>
<tr><td colspan="11">Evaluation Phase</td></tr>
</table>

表 I
主动学习实验的计划小组、交换组和个人组训练团队

在主动学习过程中，我们要求参与者报告自己对每个所提供标签的置信度。这不仅使我们能够考察标签不确定性的影响，还提供了一种训练额外（多分类）模型的方法。自我报告的置信度为 0 到 10 之间的整数（0 表示完全没有信心，10 表示完全有信心）。图 3 展示了各类别的平均置信度及其变化情况。

如图3所示，0（假）实例的置信度总体高于1（真）实例。参与者在赋予1（真）标签时，置信度低得多。因此，我们采用了如下公式4所示的转换，将置信度转换为该实例为1（真）的伪概率。

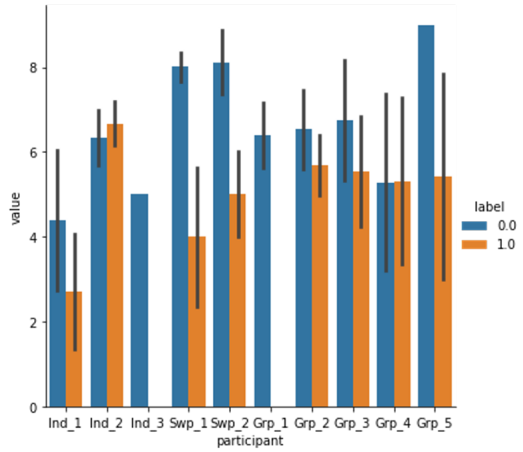

图3. 预标注数据集各标签类别参与者置信度值的均值和误差线

$$\begin{cases} 10 - Conf_i, \forall Label_i = 0 \\ ROUND((Conf_i + 10)/2), \forall Label_i = 1 \end{cases} \tag{4}$$

我们的数据收集和转换方法基于M´endez M´endez等人此前描述的方法。他们发现，收集参与者的置信度并将其转换为真正例的概率（在视频标注任务中），相比只收集二元标签 [3]，是更好的数据标注方式。

C. 评估方案与指标

在实验的最后阶段，我们需要一种评估主动学习结果以及分析师与模型交互的方法。我们采用两种不同的方式评估模型性能：

1) 典型的机器学习模型评估指标，例如精确率和召回率

由于该数据集高度不平衡，我们还报告了精确率-召回率曲线下面积（AUPRC）。相较于受试者工作特征曲线下面积（AUROC），AUPRC 更能体现模型在不平衡数据集上的性能，因此更受青睐 [29]。我们还纳入了 Fβ 分数，以突出模型性能偏向精确率或召回率时两者之间的权衡。测试数据集由另一支调查团队整理，该团队拥有完整权限，可查看电子邮件详情，包括电子邮件正文和附件文件内容。

2) 专家与模型的一致程度

通过评估每轮任务中检测为异常的 HRQ案例数量，确定模型的一致性。

此外，我们在实验结束时准备了一份问卷，访谈参与者了解他们对主动学习过程的看法和经验，以帮助我们更好地理解主动学习期间的人为因素和互动情况。

IV. 结果

A. 模型性能

本节展示主动学习与领域分析师合作的结果。我们首先考察模型性能，然后探讨研究问题。由于领域分析师的时间有限，本研究中的主动学习持续时间相对较短，因此我们的分析重点关注数据中观察到的趋势，以及在此类情境下实施未来主动学习系统的经验教训。

如图4所示，在整个研究期间，二分类模型性能（蓝线）始终低于多分类模型性能（橙线）。在其他研究中（例如 [3,30]开展的众包研究），二分类标注（相较于多分类标注）的表现也较差。在专家对判断感到高度不确定的机器学习任务中，期待他们给出明确的二分类判断可能并不合理。

在训练过程中，专家以低于 5 的置信度评分将真实异常标记为非异常的实例，被重新归类为真正例，其他研究也采用了这种做法（例如， [3]）。最终得到的多分类模型（图4中的橙色线所示）表现优于二分类模型，这可能是因为评估期间各类别更加均衡。在比较多分类模型时，交换组和小组参与者训练的模型并未优于个人参与者训练的模型。不同参与者组合的详细性能指标见表2。

Model Name	Multiclass Model Performance			
	AUPRC	Precision	Recall	F1
ind_1	0.81 (0.56)	0.62 (0.62)	0.99 (0.96)	0.76 (0.75)
ind_2	0.73 (0.64)	0.68 (0.72)	0.43 (0.09)	0.53 (0.15)
ind_3	0.74 (0.65)	0.67 (0.62)	0.48 (0.99)	0.56 (0.73)
swp_1	0.69 (0.64)	0.56 (0.67)	0.53 (0.45)	0.54 (0.54)
swp_2	0.65 (0.61)	0.64 (0.62)	0.11 (0.94)	0.19 (0.75)
grp	0.72 (0.61)	0.68 (0.56)	0.35 (0.02)	0.47 (0.04)

表 II 多分类模型与二分类模型的主动学习性能（二分类模型的性能列于括号内）

我们还采用第 3.3 节所述的第二种方法（评估方法 B）进行了评估。在后一种方法中，我们计算了整个案例研究期间被标记为 1（真实）的 HRQ所占百分比。之所以能够进行这种评估，是因为每轮的 14 个 HRQ实例都是主动学习模型预测出的最高风险异常。随后，我们评估这些预测是否与分析师决策/标签相符，以此评估模型性能。

如图5所示，尽管参与者在早期轮次中标注为“是”的平均比例

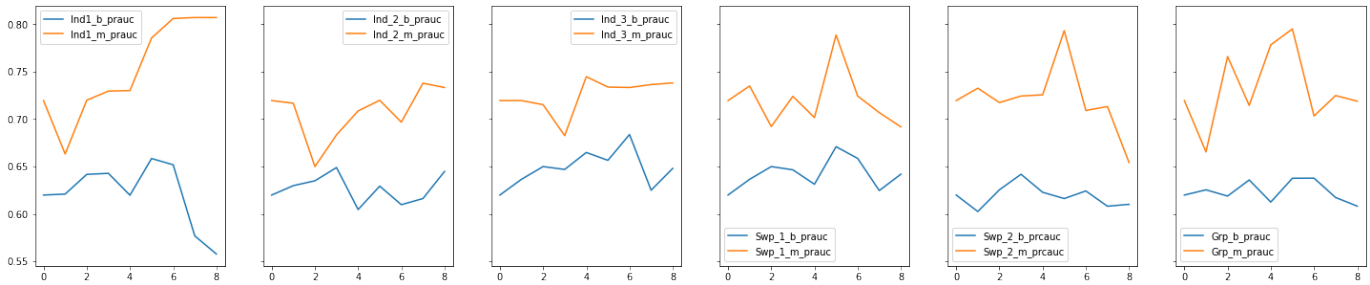

图4. 各团队二分类和多分类主动学习模型的精确率-召回率曲线下面积性能

较低，但可以观察到上升趋势（见图5）。使用（非参数）Jonckheere–Terpstra趋势检验确认了该趋势具有统计学显著性 [31,32],，如表3所示（p=0.02）。这一结果表明，使用专家提供的标签训练的主动学习模型，随着时间推移获得了越来越多的“是”标签。如前所述，这14个最高风险查询（HRQ）实例是模型预测的最异常实例。因此，这种正相关关系体现了模型准确性的提升，即其预测与专家判断的标注匹配程度有所提高。

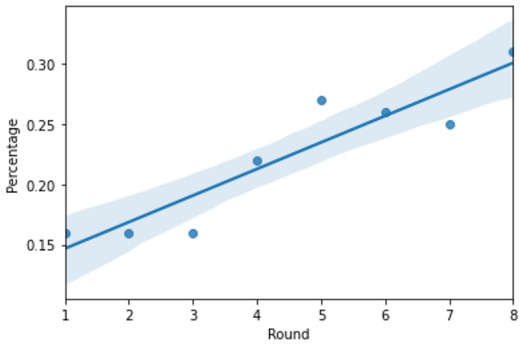

图5. 每轮14个HRQ实例的平均1（真）标签百分比趋势

Metric	Round							
	1	2	3	4	5	6	7	8
Mean	0.16	0.16	0.16	0.22	0.27	0.26	0.25	0.31

JT Score	1710.5
p-value	0.02

表III 每轮14个HRQ实例的平均1（真）标签百分比的Jonckheere趋势检验

B. 个人差异与标签可靠性

如图6所示（误差线表示95%置信区间），参与者的置信度评分分布各不相同：有些人的评分波动更大，或平均评分高于或低于其他人。这些差异从一开始就已显现（任务第一轮的评分差异见图6左侧面板），当时所有参与者都针对相同的20个实例进行了评估。在所有轮次中也观察到了明显的个体差异（如图6右侧面板所示）。

为了衡量小组内部的一致性，我们计算了群组训练团队（该团队的所有参与者在每轮都对相同的20个实例进行标签标注）每天所作决策的Krippendorff α系数 [33,34]（使用前文所述转换后的多类别标签）。图7左侧面板展示了置信度评分相似度这一指标的变化，可以观察到其数值总体呈下降趋势。这表明，随着时间推移，群组训练团队中的分析人员彼此间的分歧增多（而非减少）。这可能反映出研究期间参与者之间没有反馈或协商，而且部分参与者的动机可能比其他参与者变化更大，导致随着时间推移，一致性降低。

无论其原因如何，观察到的小组专家之间分歧的增加可能导致标签质量下降，并使小组培训模型的性能低于由个人参与者训练的模型。基于这些结果，我们建议应为组参与者安排一次或多次培训课程，以达成共识（在实际应用这种技术的情况下）。还可以根据参与者对案例进行初步标注的结果对其进行筛选，随后让主动学习专注于高性能个人或由高性能个人组成的小型子组。与当前的主动学习实践相比，这应能提升最终主动学习模型的性能；在当前实践中，所使用的评审小组在标注能力方面具有异质性。

C. 模型与人类不确定性

就标签可靠性而言，在本研究的设置下，有两种不确定性可能影响主动学习模型的性能：模型不确定性和人类不确定性。模型不确定性用于衡量模型对某些实例有多不确定；采用不确定性查询策略时，模型最不确定的实例会被提交查询。人类不确定性则是一种潜在属性，存在于每位分析人员的决策之中，会影响其置信水平以及所给出的标签。如表 4 所示，当人类对所判断的实例有把握，而模型对这些实例高度不确定时，能够获得最大的信息增益。在这种

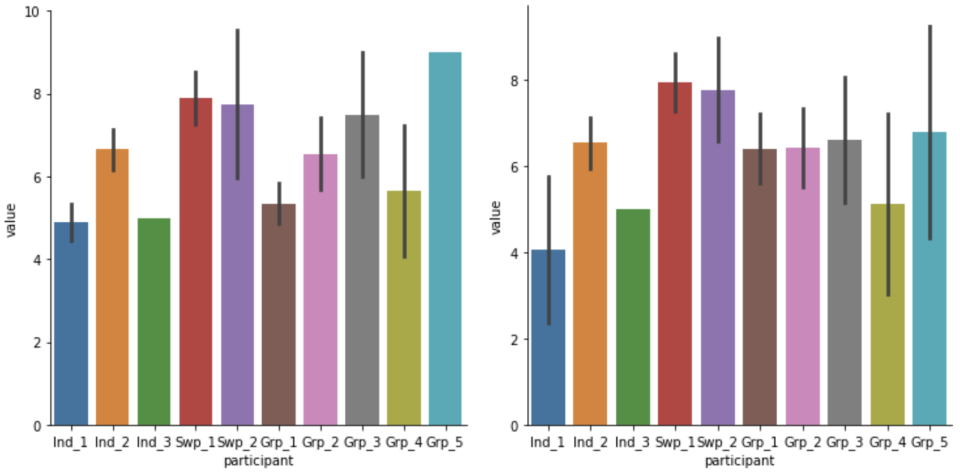

图6. 第1轮（左）和所有轮次（右）中平均置信度值及其变化方面的个体差异

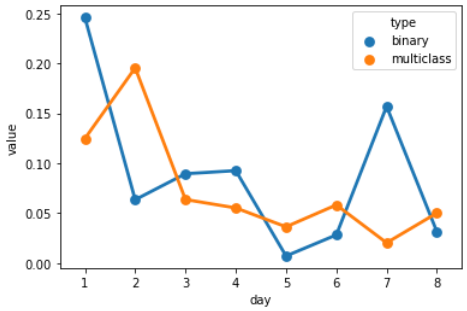

图7. 使用组参与者训练的二分类和多分类模型的 Krippendorff α系数值

情况下，人类可以通过提供正确的标签来引导模型，并通过消除模型的不确定性来带来高信息增益。

		Model: Low Certainty	Model: High Certainty
Human	Low Certainty	Noisy Labeling	Misleading Human Labeling
Human	High Certainty	**High Information gain**	Conflicting Knowledge (need for ground truth labeling)

表IV 人类与模型的确定性和不确定性的组合（突出显示人类确定而模型不确定的“最佳点”）

分析人员的人类不确定性对模型性能的影响见图8。图中的每个点代表平均原始置信度（Y轴）以及模型在下一轮中的精确率-召回率曲线下面积增减值（X轴）。由个人参与者1训练的模型（模型 Ind 1，以橙色标示）呈现出上升趋势（r=0.83；p=0.01）；然而，其他模型（Ind 2、Ind 3、Swp 1、Swp 2 和 Grp）合并后，置信水平与模型改进之间呈现不显著或临界关系（r=0.24；p=0.10）。基于这些初步结果，我们为未来研究提出两个假设：1）当分析人员具备足够的标注专业能力时，分析人员置信度与模型性能之间存在正相关关系；2）对于个人评分者而言，评分置信度与所得模型质量之间的关系强于评分者群体。

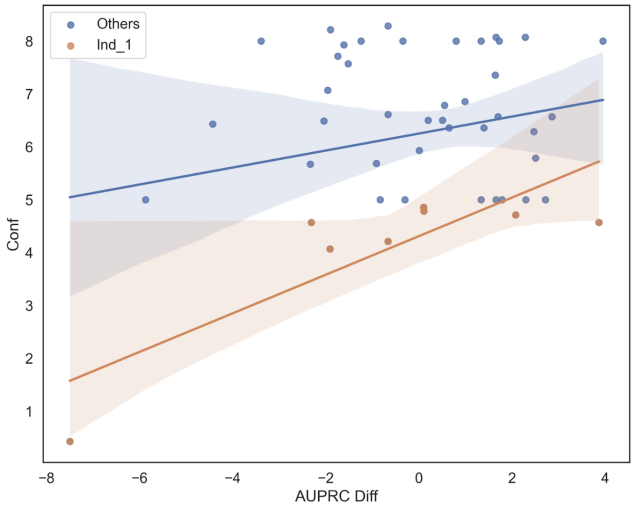

图8. Ind 1 与其他组（Ind 2、Ind 3、Swp 1、Swp 2 和 Grp）每天的平均原始置信度与模型精确率-召回率曲线下面积变化之间的趋势。

Test of Correlation				
	r	CI 95%	p-value	power
Others	0.24	[-0.05, 0.49]	0.10	0.39
Ind_1	0.83	[0.31, 0.97]	0.01**	0.80

表V 展示图8中趋势的统计数据（斜率是否不同于零）

此外，图9显示，高风险查询（HRQ）的平均置信度往往高于

不确定查询（UQs）。此外，从图9可以看出，总体而言，不确定查询（UQs）的置信度评分波动大得多。这表明，案例越不确定，分析人员对其评分的置信度就越容易出现波动。

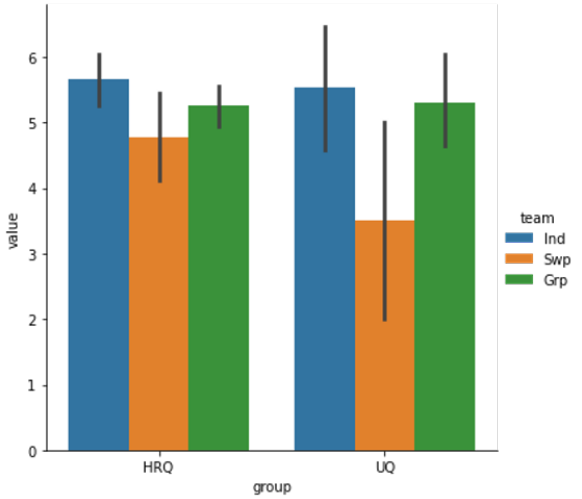

图9. 标签为1（真）时，群组训练团队中参与者的HRQ与UQ在置信水平的均值和变化方面的差异

D. 问卷答复

在案例研究结束时，我们使用问卷收集了有关分析人员使用主动学习模型的体验反馈，其中包括定量结果和定性结果。表6展示了每个问题的描述性结果，随后是对定性结果的总结和解读。接下来，我们将结合这些定性结果以及前文呈现的定量结果，评估本文先前提出的研究问题。

Question	Content	Mean	STDEV
1	How long did it generally take to label the 20 instances each round? (in minutes)	10.60	3.91
2	How confident were you in your ability to correctly label email anomalies before the experiment? (on a 10 point scale: 0 being least confident; 10 being completely confident)	5.70	1.25
3	After you finished the experiment, how confident were you that the labels you provided were generally accurate? (on a 10 point scale: 0 being least confident; 10 being completely confident)	5.90	1.20
4	When you did the labeling, the top 14 instances in every round were actually predicted as positives by the model. How well do you trust that the model is doing a good job of detecting true anomalies? (on a scale of 1 completely no trust to 10 completely trust)	4.20	1.40
5	How hard do you think it was to provide the labels in this task? (0 not hard at all; 5 moderately difficult; 10 extremely difficulty)	5.50	1.90
6	Please rate yourself in terms of how qualified or skilled you are at this email anomaly labeling task. (on a scale of 1 complete novice to 10 complete expert)	4.60	2.07
7	I had trouble maintaining my motivation to do the task and to provide the best labels I could as the experiment went on. (on a scale of 0 strongly disagree to 10 strongly agree).	4.70	2.79
8	I became better at doing the task and providing the best labels I could as the experiment went on. (on a scale of 0 strongly disagree to 10 strongly agree).	5.60	1.90

表VI描述性统计（均值和标准差）针对各项问题参与案例研究的分析人员报告称：

- 他们平均每轮花约10分钟为20个实例添加标签（这一过程从先前的标准（每周可能耗费专家 2 天时间）
- 他们通常对自己提供高质量标签的能力有信心（10名参与者中有9名报告的置信度高于平均水平，其中4名高于均值2个标准差以上）
- 他们通常认为维持动机并不困难（10名参与者中有4名报告维持动机有困难，其中只有1名高于均值2个标准差以上）
- 他们在整个实验过程中提供恰当标签的能力有所提升（10名参与者中有8名表示自己有所进步，其中4名的评分高于平均值2个标准差以上）

不过，他们担忧以下问题：

- 标注的难度（由于每个案例提供的信息有限）
- 他们是否能够信任模型输出。他们建议数据应包括：
- 活动历史的可视化，以及对相似角色发送者之间的比较
- 有关敏感/私人内容的更多信息（例如附件详情）

E. 研究结果总结

在展示模型结果和问卷反馈之后，我们重新审视第1节提出的研究问题，并解读在无法获得真实标签的现实情境下，主动学习的有效性。

研究问题1：专家在为自己的标注决策评定置信度时，校准得是否准确？（也就是说，专家对标注决策越有信心，模型预测效果是否就越好？）

分析人员对标注的不确定性似乎会显著影响主动学习提升预测效果的能力；在一位表现最佳的参与者标注员身上，这一点尤为明显（图8和表5）。因此，对于表现出色的专家标注员，这个研究问题或许可以得到肯定回答。理想情况下，如果我们能够信任人类对标注的置信度判断，那么这些判断就能为主动学习的预期改进幅度提供一个上限，因为只有当人工提供的标签确实可靠时，模型才能从这些标签的训练中获益。由于在无法获得真实标签时，很难确定人工标签的可靠性，因此，如果参与者具备足够的标注专业能力，置信度或许可以作为衡量标注确定性的有效代理指标。

研究问题2：将个人组分配给主动学习过程与将小组分配给主动学习过程，会如何影响后续的模型预测准确率？

由一群人进行标注，效果不如由个人进行标注。这一结果与众包（常见的主动学习）文献中的典型发现相矛盾，因为相关研究中的参与者掌握完整的知识。在这些无需专业知识的常见主动学习任务中，由多名标注人员产生的冗余有助于提高标签质量和可靠性。

不过，在本案例研究中，参与者仅获得了有关每封电子邮件的部分信息。小组团队（以及交换团队）中参与者之间不同的置信度／不确定性水平，可能削弱了模型改进的整体效果。根据我们获得的结果，我们认为，在需要专业知识且隐私问题至关重要的企业环境中，经过精心挑选和培训的个人，其表现可能优于经验／知识水平参差不齐的群体。

研究问题3：专家之间的意见一致程度如何？不同分析人员所给标签的可靠性如何（即，以后续模型的预测指标来定义性能）？

通过 Krippendorff's Alpha 评估的参与者间一致性相对较低。考虑到在一项关键安全任务中作出标注决策的难度，这并不令人意外，因为此类任务只提供部分信息。鉴于观察到的显著个体差异，显然需要通过培训课程，就如何“正确”标注达成共识，并根据已确定的共识方法，教会参与者采用最佳标注实践。也可能需要进行人员选拔，以招募合格的个人执行标注任务。

五、讨论：意义、局限性与未来工作

此处获得的结果和发现揭示了在主动学习中运用专业知识所面临的挑战，也提出了在缺乏真实标签时应对相关情形的可能策略和补救措施。一些发现与先前研究一致，这些研究考察了主动学习任务中人类知识存在错误或不完整的情形。先前研究者建议，在征求标签的同时收集置信度测量值。将置信度与二分标签结合，可以生成多类别标签，从而可能提供信息量更丰富、更加均衡的数据，并提升模型性能。

目前的一些研究结果与早期研究有所不同（其中许多研究在标注任务中使用了非专家）。网络安全领域的特性使得采用新颖研究策略成为必要。在潜在专家众多的领域中，通过小组培训产生冗余，并利用这些冗余预先筛除错误标签（在模型训练之前），或许是可行的；但在大多数网络安全场景中，这并非可行策略。由于网络安全专家往往稀缺且成本高昂，因此不适合通过大规模群体收集数据。因此，要解决专家之间可能出现的意见分歧，或许无法通过在较大的群体中寻找彼此意见一致的可靠参与者子集来实现。

本研究在大型组织中研究网络安全相关主动学习，这一点相对独特。我们希望本研究能够推动这一重要领域开展更多“现实”的研究。在行业环境中开展研究的一个缺点是（正如本案例研究所体现的那样），研究是在受到多重限制条件/局限性的环境中进行的。例如，由于隐私问题和有限的安全许可，本案例中一些采集的变量无法访问，从而使标注任务变得更加不确定。可获得信息的有限性反过来限制了领域专家充分运用其知识来提供标签的能力。此前提到的另一个局限性是参与的人类参与者数量及其参与多轮标注的可用性。在本研究中，即使每轮只需分配 20 个标签，也花费了数周时间才获得 8 轮标注数据。我们认为，在这类环境中收集数据存在这样的困难并不罕见。然而，这些限制条件对迭代次数（轮次）、每轮查询集的规模以及专家参与者数量所产生的最终影响，使得在回答某些研究问题时难以获得具有统计显著性的结果。

本文报告的研究对20个案例进行了8轮标注，每轮有10名参与者。尽管对于行业环境而言，这个样本量相对较大，但其普遍适用性仍有限。不过，也可以认为，主动学习在不同环境中的表现很可能有所不同。因此，与其试图在少数环境中获得大样本，不如在多种环境中使用小样本，考察置信度评分以及不同参与者分组的影响，这样更有益处。当前结果还有助于说明，在隐私问题导致可用数据特征受到限制的情况下，主动学习的表现如何。

尽管存在上述局限性，所得发现仍为在真实场景中使用主动学习提供了新的见解，也可能为如何在组织内实施网络安全应用中的主动学习提供初步指导。我们预计，本研究的研究设计和所得发现将为该领域未来的研究者奠定良好基础。交互式机器学习的问题，即如何让人类与人工智能交互，是一个极其重要的问题，而直到过去几年才开始得到应有的关注。未来研究的一个富有潜力的方向，是探究人类与模型不确定性之间的关系。需要开展研究以：A）检验是否存在更好的方法来引出专家知识，从而提高标签可靠性；B）检验是否存在更好的方法来询问专家，例如只询问那些他们更有可能能够回答或更有把握回答的实例。

在本研究中，我们无法获取分析人员在主动学习过程中判断每个案例所花费的时间。不过，正如前文所述，主动学习的效率最终应根据模型改进与分析人员所付出的努力之间的权衡来评估。本文界定了这一问题，并证明专家能够赋予置信度标签，而这些标签可以预测后续的机器学习预测准确率。既然置信度标注似乎有助于指导主动学习，未来研究就应探讨

由人工分析人员的置信度程度指导的主动学习方法的相对效率。

VI. 结论

网络安全领域的特点使得在采用《机器学习》（ML）技术时，难以不考虑领域专家的人为因素。领域专家会对组织的整体安全以及相关用户和客户作出关键决策。他们的时间极为宝贵，因此他们提供的标签获取成本也很高。在本案例研究中，我们将主动学习（AL）作为一种干预手段，应用于一项由人工执行的电子邮件异常检测任务；该任务被认为难以通过自动化机器学习完成。这项干预旨在提高标注、训练和检测的效率。我们还提出了三个研究问题，以研究领域分析师的人为因素及其与机器学习模型交互时的行为。

我们采用基于批次的场景，结合高风险查询（HRQ）、不确定性查询（UQ）和随机查询（RQ）策略，使标注任务贴近专家日常工作负荷，同时满足密度和多样性要求，从而获得更好的主动学习查询。为了研究分析人员在不同团队中使用主动学习时的行为，我们将参与者分为三个团队，每个团队采用不同的训练方案：个人训练一个模型；在案例研究中途交换个人训练的模型；以小组形式训练一个模型。

结果揭示了专家使用机器学习模型时的行为特点。根据研究发现提出的建议，以及对研究问题的回答，可总结如下：

- 参与者的置信度至关重要，因为这有助于训练出性能更好的多分类模型
- 将分析人员分组以增加冗余，以及交换分析人员以减轻工作负荷，可能无助于提升模型性能
- 可能需要制定培训／筛选指南，以找到最适合执行标注任务的人员
- 更好的模型改进可能与较高的平均置信度相关；应根据标注专业能力筛选标注人员，从而使模型得到更好的训练并提升性能
- 在设计标注流程时，组织应考虑其中可能存在的权衡：一方面，查询模型不确定的样本以获得更多信息增益；另一方面，查询人工分析人员有把握的样本，以获得质量更高的标签

基于这项研究，我们建议主动学习算法能够采取以下任一方式：a）预测专家置信度，并且只请求专家能够有把握地作出判断的实例的标签；或 b）接受多标签或多分类标签，以便筛除人类不确定如何标注的实例，同时不损害模型性能。

总之，当前迫切需要提高防御网络安全威胁的有效性。尽管在边界防御等领域以及利用机器学习检测异常方面已经取得了长足进展，但人类专业知识在提升机器学习性能方面所发挥的作用却常常被忽视。虽然主动学习等技术为将人类专业知识融入机器学习中的关键标签标注/训练任务提供了途径，但本研究结果也表明，纳入人类专业知识会带来复杂性。本研究所强调的三个主要问题是：1）如何纳入专家对其所分配标签所持有的不确定性；2）如何应对标注专业能力方面的个体差异；3）在主动学习过程中，如何最好地利用多位专家（包括对其判断进行分组和优化）。

参考文献

[1] Tomanek, K. 和 Olsson, F.（2009）。关于使用主动学习支持文本数据标注的网络调查。NAACL HLT自然语言处理主动学习研讨会论文集，45–48。[2] Settles, B.（2011）。从理论到查询：实践中的主动学习。机器学习研究杂志：研讨会与会议论文集，16，1–18。[3] M´endez M´endez, A. E.、Cartwright, M.、Bello, J. P. 和 Nov, O.（2022）。通过引导置信度评估改进众包音频标注。《美国计算机协会人机交互汇刊》，6(CSCW1)，1-25。[4] Augustin, A.、Venanzi, M.、Hare, J.、Rogers, A. 和 Jennings, N. R.（2017年7月）。分类分布的贝叶斯聚合及其在众包中的应用。AAAI出版社/人工智能国际联合会议。[5] Chung, J. J. Y.、Song, J. Y.、Kutty, S.、Hong, S.、Kim, J. 和 Lasecki, W. S.（2019）。高效引导方法：估计群体答案。《美国计算机协会人机交互汇刊》，3(CSCW)，1-25。[6] Schenk, E. 和 Guittard, C.（2009年12月）。众包：哪些工作可以外包给群体，为什么？载于开放源代码创新研讨会，法国斯特拉斯堡（第72卷，第3页）。[7] Estell´es-Arolas, E. 和 Gonz´alez-Ladr´on-de-Guevara, F.（2012）。迈向一体化的众包定义。《信息科学杂志》，38(2)，189-200。[8] Von Ahn, L.、Blum, M.、Hopper, N. J. 和 Langford, J.（2003年5月）。验证码：利用困难的人工智能问题保障安全。载于密码技术理论与应用国际会议（第294-311页）。施普林格，柏林，海德堡。[9] Von Ahn, L.、Maurer, B.、McMillen, C.、Abraham, D. 和 Blum, M.（2008）。ReCaptcha：通过网络安全措施实现基于人类的字符识别。《科学》，321(5895)，1465-1468。[10] Settles, B.（2009）。主动学习文献综述。技术报告。[11] Atlas, L.、Cohn, D. 和 Ladner, R.（1989）。利用查询和选择性采样训练连接主义网络。神经信息处理系统进展2（NIPS）。[12] Cohn, D.、Atlas, L.、Ladner, R. 和 Waibel, A.（1994）。利用主动学习改进泛化。《机器学习》，15(2)，201–221。[13] Hoi, S. C. H.、Jin, R. 和 Lyu, M. R.（2006）。通过批量模式主动学习进行大规模文本分类。第15届万维网国际会议论文集。[14] Demir, B.、Persello, C. 和 Bruzzone, L.（2011）。用于遥感图像交互式分类的批量模式主动学习方法。《IEEE地球科学与遥感汇刊》，49(3)，1014–1031。[15] Guo, Y. 和 Schuurmans, D.（2007）。判别式批量模式主动学习。神经信息处理系统进展20（NIPS）。[16] Li, X. 和 Guo, Y.（2013）。用于图像分类的自适应主动学习。IEEE计算机视觉与模式识别会议论文集。[17] Xu, Z.、Akella, R. 和 Zhang, Y.（2007）。在相关性反馈的主动学习中纳入多样性和密度。欧洲信息检索会议：信息检索进展，246–257。

[18] Holub, A.、Perona, P. 和 Burl, M. C.（2008）。用于目标识别的基于熵的主动学习。2008年IEEE计算机学会计算机视觉与模式识别研讨会, CVPR研讨会。[19] Joshi, A. J.、Porikli, F. 和 Papanikolopoulos, N.（2010）。用于图像分类的多类别主动学习。2372–2379。[20] Settles, B.（2012）。不确定性抽样。载于《主动学习》（第14–16页）。Morgan and Claypool出版社。[21] Danka, T. 和 Horvath, P.（2018）。modAL：用于Python的模块化主动学习框架。[22] Zhu, J.、Wang, H.、Tsou, B. K. 和 Ma, M.（2010）。通过基于不确定性和密度的采样进行数据标注的主动学习。《IEEE音频、语音与语言处理汇刊》，18(6), 1323–1331。[23] Settles, B. 和 Craven, M.（2008）。序列标注任务主动学习策略分析。1070–1079。[24] Huang, S.、Jin, R. 和 Zhou, Z.（2010）。通过查询信息丰富且具有代表性的样本进行主动学习。神经信息处理系统进展，23。[25] He, T.、Zhang, S.、Xin, J.、Zhao, P.、Wu, J.、Xian, X.、Li, C. 和 Cui, Z.（2014）。一种兼顾不确定性、代表性和多样性的主动学习方法。《科学世界杂志》。[26] Lee, J. U.、Meyer, C. M. 和 Gurevych, I.（2020）。赋能主动学习，联合优化系统和用户需求。arXiv预印本，arXiv:2005.04470。[27] Carcillo, F.、le Borgne, Y. A.、Caelen, O. 和 Bontempi, G.（2018）。面向真实场景信用卡欺诈检测的流式主动学习策略：评估与可视化。《国际数据科学与分析杂志》，5(4)，285–300。[28] Ke, G.、Meng, Q.、Finley, T.、Wang, T.、Chen, W.、Ma, W.、Ye, Q. 和 Liu, T.-Y.（2017）。Lightgbm：一种高效的梯度提升决策树。神经信息处理系统进展30（NIPS 2017）。[29] Saito, T. 和 Rehmsmeier, M.（2015）。在不平衡数据集上评估二分类器时，精确率-召回率图比ROC图更具信息量。《公共科学图书馆·综合》，10(3)，e0118432。[30] Dumitrache, A.、Aroyo, L. 和 Welty, C.（2018年6月）。在众包框架消歧中捕捉歧义。载于第六届AAAI人类计算与众包会议。[31] Jonckheere, A. R.（1954）。针对有序备择假设的无分布假设的k样本检验。《生物计量学》41：133-145。[32] Terpstra, T. J.（1952）。当一个排序中存在并列值时，Kendall趋势检验的渐近正态性与一致性。《数学研究》14：327-333。[33] Krippendorff, K.（2011）。计算Krippendorff α信度。[34] Krippendorff, K.（2004）。内容分析：方法导论。Sage出版社。[35] Levenshtein, V. I.（1966年2月）。能够纠正删除、插入和逆转的二进制码。载于《苏联物理学文摘》（第10卷，第8期，第707-710页）。