DOI:10.19651/j.cnki.emt.2108046

一种基于 QBC 不一致性的恶意加密流量识别方法*

张荣华 刘 智 罗 琴
(西南石油大学计算机科学学院 成都 610500)

摘 要: 当前基于机器学习的恶意加密流量识别主要采用有监督学习,依赖大量标注样本,但在真实环境中恶意流量不仅稀缺而且标注依赖专家经验,标注成本较高,而主动学习通过迭代训练选择困难样本(hardsample)进行训练,一定程度上减少了训练样本量,但当前基于委员会投票的 hardsample 选择策略粒度较粗,所选样本质量不佳。针对该问题,提出一种改进委员会投票(QBC)的恶意加密流量识别方法 CBU,设计了委员会对样本不一致性的计算方法,并结合已标注与未标注样本相似性分析,有效度量样本不确定性,从而选择高质量 hardsample,以减少样本标记和训练量。实验使用业界标准数据集 CTU 以及真实恶意数据集进行测试,结果表明,相比传统委员会投票策略,CBU 样本标记量减少 1 倍,只采用 15%数据的情况下识别准确率达到 96%,有效减少样本标注和训练量,具有较强实用性。

关键词: 加密流量;主动学习;样本选择;恶意识别
中图分类号: TP393.08 **文献标识码:** A **国家标准学科分类代码:** 520.1060

Method for identifying malicious encrypted traffic based on QBC inconsistency

Zhang Ronghua Liu Zhi Luo Qin
(School of Computer Science, Southwest Petroleum University, Chengdu 610500, China)

Abstract: At present, the identification of malicious encrypted traffic based on machine learning mainly uses supervised learning and relies on a large number of labeled samples. However, in the real environment, malicious traffic is not only scarce but also depends on expert experience, and the labeling cost is high. Active learning selects difficult samples through iterative for training, which reduces the amount of training samples to a certain extent, but the current hardsample selection strategy based on committee votes has a coarser granularity, and the quality of the selected samples is not good. In response to this problem, a CBU is proposed to improve the query by committee (QBC) method for identifying malicious encrypted traffic. Labeling sample similarity analysis, effectively measuring sample uncertainty, and selecting high-quality hardsamples to reduce sample labeling and training volume. The experiment uses the industry standard data set CTU and real malicious data sets for testing. The results show that compared with the traditional committee voting strategy, the amount of CBU sample labeling is doubled, and the recognition accuracy rate of only 15% of the data amount is 96%, which effectively reduces the sample labeling. And training volume, and it has strong practicability.

Keywords: encrypted traffic;active learning;sample selection;malicious identification

0 引 言

近年来,全球加密流量占比逐年增高,据 Google 透明度最新报告,从 2018 年 6 月至今,Chrome 产品和服务中加密流量的比例一直高达 95%。加密通信广泛应用于各大领域,能有效保护隐私安全,但加密流量不断增长也带来了各种网络恶意行为。僵尸网络[1]、APT 攻击[2]等网络攻击层出不穷。据最新一期的国家互联网应急中心网络安全信息与动态周报显示,较之于上周境内感染计算机恶意程序主机数量又提升了 1.1%。对于这些攻击的检测,传统的流量分析检测方法比如深度数据包解析[3]已经无法实现,而基于机器学习的恶意加密流量检测,一直是网络安全领

收稿日期:2021-10-10
* 基金项目:国家自然科学基金(61902328)项目资助

域的研究热点。

基于机器学习和深度学习的恶意加密流量识别研究很有成效[4-7]，但当前研究存在样本标注成本较高的问题。而在减少样本标注问题上主动学习[8](active learning)方法运用广泛。徐海龙等[9]利用 SVM 模型将样本到分类超平面的距离作为选点依据，利用置信度减少相似样本选择。胡峰等[10]根据模型预测结果，以模型预测概率的差值为不确定度量方法，根据不确定性计算的概率差值划分区域并进行样本标注。毛蔚轩等[11]通过计算样本相似性，再结合最小化估计风险实现对恶意代码的检测，在小样本情况下降低了错误率，但是最小化估计风险策略需要从已标记数据中构建数据依赖网络找出重要资源对象，不具备通用性。李翼宏等[12]在最小化风险上进行了改进，选出样本间相距较远数据与已标记样本计算相似度，再结合模型对未标记样本的预测结果估算风险，挑选出风险最小的数据进行标注。

上述方法都在单一方向对主动学习方法进行调整和研究，没有结合考虑主动学习方法、模型以及样本间的关系，简单以减少相似样本为选点原则。自动标记方法虽然在其他领域很有成效[13-15]，但加密流量只有统计特征，自动标记会存在严重误差，进而影响最后的识别准确率。为了进一步降低传统委员会投票(query by committee，QBC)的样本标记量，本文针对传统委员会投票粒度较粗问题，设计了一种委员会投票不一致性的计算方法，增强委员会投票结果的可信性，再结合样本间的相似性有效度量模型对样本预测结果的不确定性，选择高质量 hardsample 进行实验，提高模型泛化性，减少样本标记量。为了验证本方案可行性，本文使用业界公认的标准数据集 CTU 以及真实恶意数据集进行实验，结果表明本方案相比于传统委员会投票效果更好，进一步降低了样本标记量。

1　基于委员会不一致性的方案设计

1.1　整体流程

主动学习方法能通过找出困难样本进行人工标记，再结合机器学习模型的反馈完成对数据的迭代标记，从而有效减少数据标注量。CBU(committee-based uncertainty)核心流程如图 1 所示。

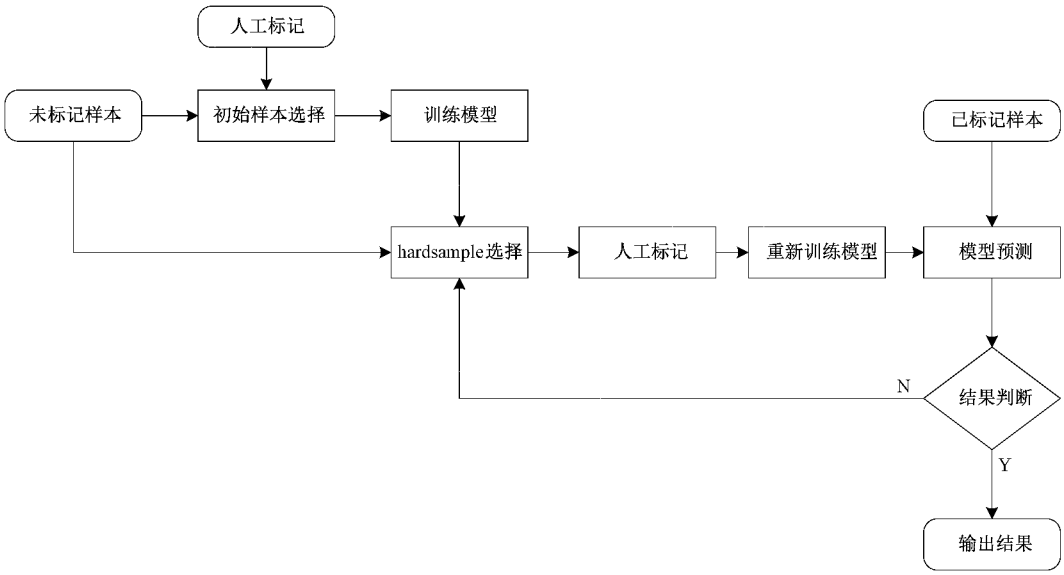

图 1　CBU 方案流程

1)将全部样本随机分为 10 份，取 1 份进行人工标记得到已标记样本 L_1，作为后续模型的测试集使用，剩余未标记样本 U_1 用于主动学习模型迭代。

2)对未标记样本 U_1 进行初始样本选择，将选择出来的样本进行人工标记得到已标记样本 L_2，剩余未标记样本为 U_2。

3)将已标记数据 L_2 作为输入，训练多个机器学习模型以备后续使用。

4)将未标记样本 U_2 作为输入，利用基于 hardsample 选择策略对 U_2 进行选择，挑选出困难样本进行人工标记并更新已标记样本 L_2。

5)将更新后的 L_2 作为输入重新训练模型，然后用这些模型分别对步骤 1)中得到的测试集 L_1 进行预测，若 3 个模型的测试结果中准确率最高的模型在继续迭代后准确率相对稳定，则实验结束，反之则继续迭代，重复步骤 4)，直至最后所有样本迭代完毕，最后输出实验结果。

1.2　初始样本选择

初始样本选择模块是针对所有未标记样本展开。该模块具有两个功能，得到初始已标记样本集和待标记样本集。已标记样本集用于训练初始模型以备后续委员会投票使用，待标记样本集用于后续模块迭代挑选 hardsample。

1)初始已标记样本选择策略

初始已标记样本决定初始机器学习模型预测效果，方

案采用多子类抽样策略从无标记样本中选择数据。多子类抽样策略的目的是使抽样的数据样本尽可能扩大初始模型的训练集覆盖范围，既能降低已标记样本集的冗余度，又能使训练得到的初始模型更具泛化性。

本文采用聚类选择样本，利用 K-means 聚类将无标记样本集聚为若干个子类，然后抽取每个子类中心点进行人工标记形成初始已标记数据集 L_2。经过实验分析，在 k 值范围 100～1 000 中，以 100 为叠加条件进行多次对比，结果显示子类个数 $k=500$ 时，样本量适中且模型泛化性较强。K-means 聚类算法高效且易于实现，算法基于欧几里得度量也称为欧氏距离，可计算多维空间中的直线距离，再根据点与点之间距离不断调整质心位置，直至所有数据点都计算完毕。例如 $x=\{x_1\ x_2\ x_3\ \cdots\ x_n\}$ 和 $y=\{y_1\ y_2\ y_3\ \cdots\ y_n\}$ 之间的距离可用式(1)表示。

$$d(x,y):=\sqrt{\sum_{i}^{n}(x_i-y_i)^2} \tag{1}$$

2)待标记样本集选择策略

主动学习迭代初期已标记样本较少，用于迭代的待标记样本集代表性越强，越能挑选出模型更需要的样本提高模型泛化能力，减少实验迭代次数，进而减少总体样本标记量。

待标记样本集的样本选择依旧采用多子类抽样方法寻找有代表性的数据。而在子类中，子类中心点和子类边缘点能有效代表整个子类特点，所以抽取这两部分数据加入待标记样本集。本文利用聚类算法 K-means 实现，经过实验分析，在 k 值范围 2～10 中，以 1 为叠加条件进行多次对比，结果显示子类个数 $k=4$ 时效果最佳，具体过程如算法 1 所示，首先随机挑选部分数据点作为质心，再根据式(1)的欧氏距离计算各数据到质心之间的距离，把每个点分配给最近的质心，从而形成 k 个子类；然后根据式(2)重新计算此时的质心坐标；接着不断重复上述步骤，直至最后质心位置不再发生变化为止；最后计算各簇内样本距离质心的距离，按照 10%比例抽取距离最小和距离最大的样本放入待标记数据集 M 中，等待后续样本标注。其中 u_i 代表簇 C_i 的质心，x 代表该簇中的样本。

$$u_i=\frac{1}{|C_i|}\sum_{x\in C_i}x \tag{2}$$

算法 1　待标记样本集选择策略

输入：未标记数据集 U_2

输出：待标记数据集 M

1. 获取 U_2 行和列数 m，n

2. **for** $i<4$：//聚类个数 $k=4$

3. u_i=random(m)//在 U_2 中随机选择 4 个样本作为聚类中心

4. **end for**

5. **while** clusterChange：

6. clusterChange=False//循环停止条件，质心不再变化时 clusterChange=True，停止循环

7. **for** $i<m$：

8. $d(x,\ u_m)$=distance()//计算每个点到聚类中心的距离

9. 根据 $d(x,\ u_m)$ 把每个样本分配给最近的聚类中心，形成 4 个簇

10. u_n=mean(x)//计算每个簇内所有样本数据的平均值并更新质心坐标 u_n

11. **end for**

12. **if** $u_m=u_n$：//质心不在发生变

13. clusterChange=True

14. **end if**

15. **end while**

16. 再次计算每个簇中 $d(x,\ u_m)$

17. 抽取 10%比例的 d_{min} 和 d_{max} 的数据放入 M

18. **end**

1.3　harsample 选择策略

传统的委员会投票利用预测类别是否一致进行样本标记，该策略是放弃预测类别一致的样本，选择预测结果不一致的样本进行标记。虽然该方法也有一定的效果，但是对投票结果不一致性判断粒度较粗，无法有效选择 hardsample，导致样本质量不佳，样本标记量较大。hardsample 属于委员会模型预测困难样本，具有委员会预测结果高度不一致性的特征，hardsample 更有利于提高模型泛化性。

熵是一种度量信息混乱程度的方法，CBU 采用一种基于熵的委员会不一致性计算方法并结合样本相似性挑选 hardsample 进行人工标记。

CBU 不确定性计算策略(如算法 2 所示)分为两个部分：基于熵的委员会不一致性计算和度量样本间的相似性。基于熵的委员会不一致性计算先用 L_2 训练 K 个模型组成委员会，再用委员会对待标记样本集 M 进行预测。由于是二分类问题所以每个委员针对 M 中的样本 x_i 会得到 $T=2$ 个概率预测值，即可求得该委员投票结果的熵值 H，进而求出整个委员会投票的熵值总和 H_{sum}，熵值总和越大代表委员会投票的不一致性也越大。委员会不一致熵的计算公式如式(3)所示，其中 $m\in T$，$n\in K$，$x\in M$。

$$H_{sum}(x)=-\sum_{i=0}^{n}\sum_{j=0}^{m}p_j(x_i)\log_2 p_j(x_i) \tag{3}$$

相似度计算采用余弦相似度。如式(4)所示，待标记样本集 M 中的样本 x 与已标记样本集 L_2 中的每个样本 y 都有一个相似值，由于两个样本越相似时，其样本数据构成的余弦夹角越小，余弦值越大，因此以最大值 $S_{\max}$ 度量样本与已标记样本集间的相似性。

$$sim_{(x,y)}=\frac{\sum_{i=1}^{n}x_i\cdot y_i}{\sqrt{\sum_{i}^{n}(x_i)^2}\cdot\sqrt{\sum_{i}^{n}(y_i)^2}} \tag{4}$$

hardsample 选择策略是选择不一致性较大并且与已标记样本集尽可能不相似的样本，所以最终以 V 衡量投票样本的不确定性程度，计算方式如式(5)所示。

$$V=\frac{H_{sum}(x)}{S_{\max}(x)}(x\in M) \tag{5}$$

算法 2 hardsample 选择策略

输入：待标记数据集 M，已标记样本集 L_2

输出：样本的不确定性程度 V

1. train=L_2[1…h]//已标记样本集 L_2 作为训练集
2. test=M [1…q]//待标记集 M 作为测试集
3. model ()//用 train 训练委员会组
4. model. predict ()//对 test 进行预测，并输出预测概率
5. **for** $i<q$：
6. H_{sum}=Entropy (i)//求委员会组总的不一致熵
7. $S_{\max}=sim$ (i)//计算 M 与 L_2 中每个样本的余弦相似度，取最大值衡量样本与已标记样本集间的相似性
8. 计算 V//计算样本的不确定性程度
9. **end for**
10. **end**

2 实 验

2.1 实验环境

实验环境为 AMD R5 3600X 的 CPU、32 G 内存，操作系统为 Ubuntu16.04。

2.2 数据集

实验数据使用网络开源 CTU 数据集[16]和恶意软件分析网站数据集，两个开源数据集都包含真实恶意加密流量以及相关的恶意 IP 信息，便于做数据标注处理。

本实验目前主要对 TLS 加密流量进行研究，所以需要对两个数据集都含有的大量杂流量进行数据包过滤操作。过滤采用网络流量分析系统进行过滤，最后筛选出的恶意加密流量数据包总量约为 75.6 G。良性流量部分使用 CTU 良性流量并在个人主机上利用 Wireshark 收集访问 Alexa 排名前列网站产生的流量。最后用于实验的正负样本数据总共有 371.6 万条流(样本数据分布如表 1 所示)。在表 1 中，正负样本大小以及流数量差异很大，是由 CTU 的流量收集模式造成的，CTU 样本收集会以月为单位收集，使得数据量和流数目都比较大，但有效数据很少。

为保证恶意加密流量识别有效性，本文数据收集时间跨度为 2011 年～2021 年，降低数据流量随时间变化对恶意行为检测的影响。CTU 数据集是从 2011 年开始收集，缺少近几年的数据流量，所以恶意流量使用恶意分析网站的数据集进行补充，良性流量自我收集进行补充，保证所有流量分布均匀，数据集的正负样本分布比例为 1∶1。

表 1 数据集分布

类型	数据大小 /G	TLS 流数 /K	TLS 占比 /%
恶意流量	75.6	3 687	94.89
正常流量	4.1	29	99.18
总量	79.7	3 716	99.94

2.3 特征选择

通过对恶意软件的行为分析和研究采用 3 个类别总共 24 个特征进行实验研究，TLS 特征 5 个、传输层特征 13 个、X509 证书特征 6 个(部分特征如表 2 所示)。表 2 中除版本号、服务器选择的密码套件外，其余特征包括存活时间、包大小、证书长度以及证书有效时间等均为统计型特征。

表 2 部分特征

类型	部分特征
TLS	ssl_version
	cipher_suite_serve
	resumed
Conn	max_duration
	ssl_flow_ratio
	avg_size
	recv_sent_size_ratio
	percent_of_established_state
X509	avg_key_length
	avg_cert_valid_day
	std_cert_valid_day
	percent_of_valid_cert

2.4 数据预处理

数据采用 DPI 软件 Zeek 进行提取。如图 2 所示，首先通过 Zeek 分析得到数据包的日志文件，然后通过 Zeek 设置的唯一识别 ID 将日志文件进行整合，在整合过程中会根据选择的特征进行相应的统计特征计算，而且将四元组信息以及协议类型相同的数据流合并为一个流条目数据。最后总的数据流条目有 10 090 条。

由于样本特征中 TLS 版本号以及服务器选择的密码套件为字符串格式，模型无法训练和预测。本实验利用 One-hot 独热编码对字符串进行转化，从而进行模型训练与预测。样本数据条目是由多个流合并而成，所以某些特征内容比如 TLS 流数、最大持续时间等内容特别大，因此为了提高模型效果将数据进行归一化处理。

2.5 模型选择

CBU 选择逻辑回归、决策树和随机森林 3 个经典模型

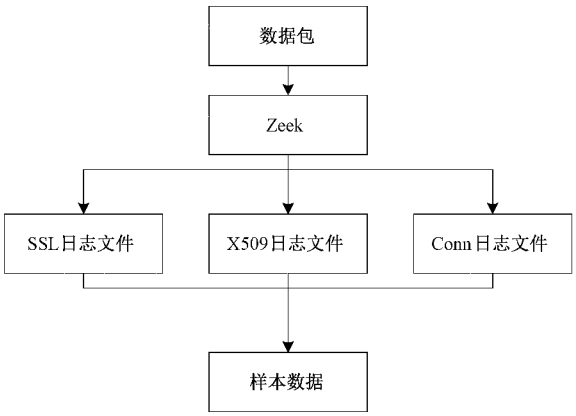

图 2　数据预处理流程

组成委员进行实验，聚类算法选择 K-means。各模型关键参数均在全量数据下多次对比实验取得。对于 K-means 算法，在初始样本选择阶段存在两次聚类，聚类个数分别为 500 和 4。对于委员会模型组，逻辑回归模型的最大迭代次数为 100，决策树模型的最大深度为 11，随机森林模型树的个数为 200。

2.6　实验与分析

1)实验设计

为测试 CBU 有效性，本文设计如下 3 组实验：

(1)不同初始数据选择策略比较。设计了样本随机选择和具有初始样本选择策略的委员会投票方法比较。

(2)hardsample 选择策略比较。设计了 CBU、传统委员会投票、只采用不确定信息熵以及 Baseline 进行比较。Baseline 为常用的有监督学习方法，使用全部数据进行标记和训练，模型采用检测效果较好的随机森林。

(3)对比迭代过程中抽样比例对各实验方法在检测效果和数据使用量的影响，样本标记的抽取比例包括 10%、20%以及 40%。

在模型指标方面采用准确率、精确率以及 F1 值对模型预测结果进行评估。为了保证结果的可信性，各方法均设置 10 次对比实验，使得每份数据都能作为测试集，提高实验结果有效性。模型结果预测也进行 10 次，结果取平均值，进一步降低偶然性，减少实验误差。

2)初始数据选择策略比较

初始数据选择策略相比传统随机样本选择可以挑选有代表性的样本训练模型，让实验结果更快的贴近全量数据的结果，减少主动学习的迭代次数，进而减少数据的标注量。

通过图 3 委员会投票在选点方法为本文提出的选点策略和传统随机选择方法结果对比可以发现，准确率、精确率以及 F1 值均为 96%且具有初始选择策略的主动学习方法效果略好，恶意加密流量检测结果符合预期效果。

从图 4 和 5 初始数据选择策略与随机选择的方法在迭代次数和数据标注量对比可以发现，加入初始数据选择策略使得迭代次数平均在 7.2 次，而随机选择方法迭代次数平均为 9 次。经计算，初始数据选择策略的主动学习方法数据标注量为 25.06%，随机选择主动学习数据标注量为 29.44%，数据标注量减少 4%。

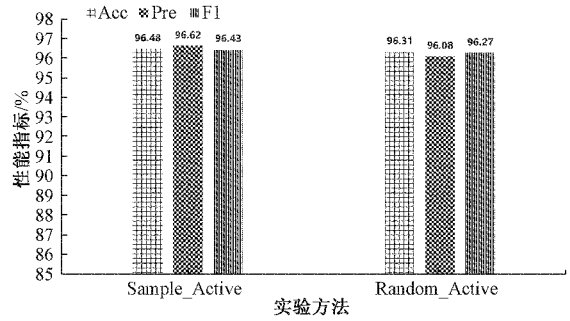

图 3　初始数据选择策略对比图

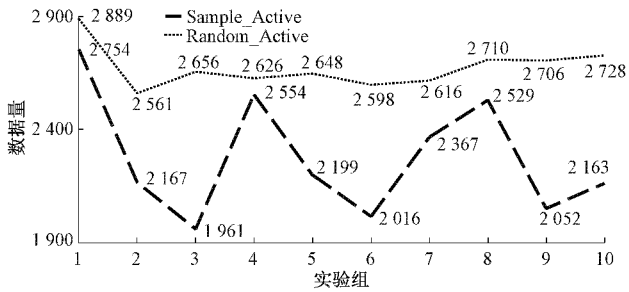

图 4　数据标注量对比图

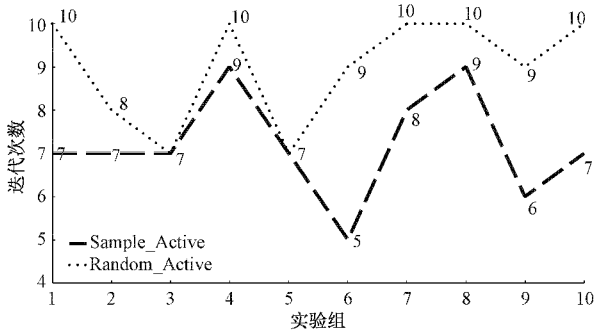

图 5　迭代次数效果对比图

实验结果表明方案的初始数据选择策略相比传统随机抽样方法减少了实验迭代次数，在一定程度上降低样本标记量。同时也说明传统委员会投票方法确实存在选点粒度较粗样本质量不佳从而导致样本标记量较大的问题。

3)hardsample 选择策略对比

CBU 方法通过挑选较高委员会不一致熵并且与已标记样本不相似的样本加入实验，在保证识别准确率的同时有效减少样本标记量，并且进一步降低已标记样本集冗余度。

通过实验(实验结果如图 6 所示)可以发现 CBU 在准确率、精确率和 F1 值上效果更好，并且 CBU 和常用的有监督学习方法在检测结果上几乎一致，展现了 CBU 方法在恶意加密流量上更具有效性。

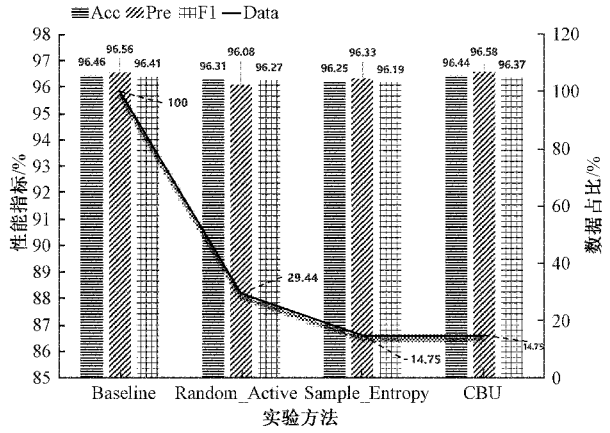

图 6 CBU 方法效果对比图

对比数据标记量，有监督学习方法使用了全部数据，传统委员会投票方法数据标注量为全部数据的 29.44%，而 CBU 方法只使用了全部数据的 14.75%，相比于传统委员会投票方法还降低了 15%，直接减少 1/2 数据标注量，效果显著。与 CBU 相比，不考虑样本相似性只使用投票不一致熵标记的数据量也是 14.75%，这是因为两个方法每次迭代选点都只标记总数据量 1%的样本，所以两个方法具有相同的迭代次数和数据量，但是恶意检测结果 CBU 略高。

综上，在恶意加密流量检测上，CBU 方法挑选样本质量更高训练出的模型泛化性更好，显著降低样本标记量，减少数据标注成本。

4)不同数据抽取比例对实验的影响

本文对上述方法在样本抽取比例上进行了实验对比，数据标记的抽取比例包括 10%、20%以及 40%，对比结果如表 3～5 所示。

表 3 比例 10%各方法对比 %

对比方法	准确率	精确率	F1 值	数据
CBU	96.44	96.58	96.37	14.75
Sample_Entropy	96.25	96.33	96.19	14.75
Sample_Active	96.48	96.62	96.43	25.06
Random_Active	96.31	96.08	96.27	29.44
Baseline	96.46	96.56	96.41	100.00

表 4 比例 20%各方法对比 %

对比方法	准确率	精确率	F1 值	数据
CBU	96.55	96.65	96.47	20.95
Sample_Entropy	96.39	96.60	96.33	23.40
Sample_Active	96.40	96.60	96.30	30.50
Random_Active	96.44	96.21	96.39	29.62
Baseline	96.46	96.56	96.41	100.00

表 5 比例 40%各方法对比 %

对比方法	准确率	精确率	F1 值	数据
CBU	96.61	96.96	96.54	37.99
Sample_Entropy	96.48	96.71	96.43	42.14
Sample_Active	96.38	96.49	96.31	30.70
Random_Active	96.25	96.06	96.20	32.66
Baseline	96.46	96.56	96.41	100.00

表中 Baseline 为有监督学习方法，利用全量数据进行实验，采用针对本数据集检测效果最好的随机森林模型进行恶意检测。表 3～5 结果显示其他 4 种主动学习方法在准确率、精确率以及 F1 值上均达到 Baseline 检测结果，达到了实验的检测效果。

通过表 3 在抽样比例为 10%时可以发现，由下往上随机选点的传统委员会投票学习方法标记数据占比为 29.44%，说明主动学习方法能有效降低样本标记量，减少样本标记成本；只加入本文抽样策略的传统委员会投票 Sample_Active 方法数据标记量只降低了不到 5%；但本文 CBU 方法以及具有抽样策略和委员会不一致熵的 Sample_Entropy 方法数据标注量占比只有 14.75%，显著降低样本标记量，相比于传统委员会投票和全量数据的有监督学习方法，极大的减少了样本标记成本，具有较强的实用性。但同时也说明在每次样本迭代选点量极小的情况下，样本相似度在较少数据标注量上没有展现明显的作用，只提高了样本质量，略微提高了检测准确率。

当抽样比例从 10%提升到 20%再到 40%时，CBU 检测效果依旧最好。从 10%到 20%后，CBU 方法数据标记量依旧远低于有监督学习和传统委员会投票方法。CBU 方法的相似性度量与只采用委员会不一致熵的差距拉大，但同时也表明单纯的抽样策略随着抽样比例的增大已经不再发挥作用。

如表 5 所示，当抽样比例上升到 40%时，传统的委员会投票方法样本标记量反而低于 CBU 方法，这是因为 CBU 随着抽样标记比例增大时，每次迭代后样本标记量成比例增加，而传统委员会投票由于投票粒度较粗，无法增大样本标记的比例，只能通过增大初始样本的选择比例，所以样本标记量变化小。通过对比 10%到 40%的传统委员会投票结果也可以发现，抽样比例的增大几乎不影响传统委员会投票方法的结果。

可见，CBU 方法不适用于抽样比例过大的情况，当抽样比例较小时，CBU 效果显著，可以极大地减少数据标记量，显著降低样本标记成本。

3 结 论

本文针对真实环境中恶意加密流量不仅稀缺而且标注成本较高的情况，利用委员会投票迭代标记，实现少样本恶意流量识别。并且针对委员会选点粒度较粗设计了

委员会不一致性的计算方法，结合样本相似性有效度量样本不确定性，选择高质量 hardsample 进行实验，进一步减少样本标记量。同时实验对初始样本抽样比例进行对比，结果表明 CBU 方法在数据抽取比例为 10%时，样本标注量只需 14.75%，对于恶意加密流量识别效果显著。

但是整个实验也有不足，本方案只是在训练集选择策略上进行了学习和研究，对于模型的改进和算法的研究还不够，后续考虑加入对模型的改进优化，探索更有效的算法进行实验，也可以加入深度学习模型，对数据样本本身进行深入研究，不断改进 CBU 方法。

参考文献

[1] 李可，方滨兴，崔翔，等. 僵尸网络发展研究[J]. 计算机研究与发展，2016，53(10)：2189-2206.

[2] 贺诗洁，黄文培. APT 攻击详解与检测技术[J]. 计算机应用，2018，38(S2)：170-173，182.

[3] 饶亲苗，彭艳兵. 基于 DPI 的应用指纹自动提取方法研究[J]. 计算机应用与软件，2021，38(4)：328-333.

[4] ANDERSON B，PAUL S，MCGREW D. Deciphering malware's use of TLS (without decryption) [J]. Journal of Computer Virology and Hacking Techniques，2018，14(3)：195-211.

[5] 骆子铭，许书彬，刘晓东. 基于机器学习的 TLS 恶意加密流量检测方案[J]. 网络与信息安全学报，2020，6(1)：77-83.

[6] CHEN L，GAO S，LIU B，et al. THS-IDPC：A three-stage hierarchical sampling method based on improved density peaks clustering algorithm for encrypted malicious traffic detection[J]. The Journal of Supercomputing，2020，76(9)：7489-7518.

[7] KWON D，KIM H，KIM J，et al. A survey of deep learning-based network anomaly detection[J]. Cluster Computing，2019，22(1)：949-961.

[8] 张琳. 基于 a-BvSBEM 主动学习的高光谱图像分类[J]. 国外电子测量技术，2017，36(4)：17-20.

[9] 徐海龙，别晓峰，冯卉，等. 一种基于 QBC 的 SVM 主动学习算法[J]. 系统工程与电子技术，2015，37(12)：2865-2871.

[10] 胡峰，张苗，于洪. 基于三支决策的主动学习方法[J]. 控制与决策，2019，34(4)：718-726.

[11] 毛蔚轩，蔡忠闽，童力. 一种基于主动学习的恶意代码检测方法[J]. 软件学报，2017，28(2)：384-397.

[12] 李翼宏，刘方正，杜镇宇. 一种改进主动学习的恶意代码检测算法[J]. 计算机科学，2019，46(5)：92-99.

[13] 陈利琴，金聪. 基于异构描述子的新型高斯混合模型图像自动标注方法[J]. 电子测量技术，2015，38(11)：60-65.

[14] 张胜男，苑玮琦. 圆阵列平面靶标特征点的自动标记[J]. 计算机工程与应用，2016，52(5)：169-172.

[15] 何晓梅. 基于条件随机场的音乐共同语义标注[J]. 电子测量技术，2016，39(8)：70-74.

[16] GARCIA S，GRILL M，STIBOREK J，et al. An empirical comparison of botnet detection methods[J]. Computers & Security，2014，45：100-123.

作者简介

张荣华，硕士研究生，主要研究方向为网络空间安全与机器学习。

E-mail：1459697205@qq.com

刘智，博士，讲师，主要研究方向为网络空间安全与机器学习。

E-mail：zhi.liu@swpu.edu.cn

罗琴，博士，副教授，主要研究方向为网络空间安全。

E-mail：dorothy_lq@163.com