基于DNS流量分析识别加密货币矿工的研究和实现

Research and Implementation of Cryptocurrency Miner Detection Based on DNS Traffic Analysis

周婧莹，黎　宇，黄　坤，梁洪智（中国联通广东分公司，广东 广州 510000）
Zhou Jingying，Li Yu，Huang Kun，Liang Hongzhi（China Unicom Guangdong Branch，Guangzhou 510000，China）

摘　要：
随着加密货币的兴起，恶意挖矿在全球肆虐，成为国家整治的重点，而挖矿木马不断进化，以各种方式规避当前主流的挖矿监测手段。为解决这一问题，提出一种AI算法，基于运营商DNS流量、AAA日志和挖矿威胁情报数据，自动识别挖矿行为和矿工。通过模型训练和结果分析，选择使用GMM和Bisecting K-means混合模型对DNS查询和响应中的模式进行识别，实时准确地检测矿工及其行为规律。

关键词：
加密货币；挖矿；DNS流量分析；机器学习
doi：10.12045/j.issn.1007-3043.2023.08.011
文章编号：1007-3043（2023）08-0048-05
中图分类号：TN915.08
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
With the rise of cryptocurrencies，malicious mining activities have spread globally and become a key concern for governments. Mining Trojans continue to evolve，and employ various methods to evade current mainstream mining detection techniques. To address this issue，it proposes an AI algorithm that can automatically identify mining behaviors and miners based on DNS traffic of operators，AAA logs，and mining threat intelligence data. Through model training and result analysis，it selects the Gaussian Mixture Model（GMM）and Bisecting K-means clustering algorithm as a hybrid model to recognize patterns within DNS queries and responses，enabling real-time and accurate detection of miners and their behavioral patterns.

Keywords：
Cryptocurrency；Mining；DNS traffic analysis；Machine learning

引用格式：周婧莹，黎宇，黄坤，等. 基于DNS流量分析识别加密货币矿工的研究和实现[J]. 邮电设计技术，2023（8）：48-52.

0 引言

加密货币挖矿是一项热门商业活动，但未经所有者同意使用资源挖矿会带来负面影响，同时对推动高质量发展和节能减排带来不利影响。挖矿会产生大量网络流量，包括与矿池通信的数据包、挖矿算法的计算结果等，因此当前基本是通过监测网络流量的数量、频率和目的地等指标特征对挖矿行为进行检测。而挖矿行为具有隐蔽性，其流量特征经常变化，需要人工长期追踪挖矿特征。另外，监测流量的方式成本高，如需全面、准确地发现挖矿行为，需实现全覆盖，不适用于中小型企业和公众用户。

对此，中国联通某分公司提出一种基于运营商DNS数据的AI算法，自动发现加密货币矿工。通过从DNS日志提取有关域名、IP地址和时间戳等信息来构建特征向量，将特征向量作为输入来训练模型，使模型对用户访问矿池域名的数据特征进行分类，从而识别出矿工，模型采用非监督学习算法。运营商DNS数

收稿日期：2023-06-02

据覆盖大中小企业用户和公众用户，可识别隐藏在企业和公众用户中的恶意矿工，及时预警处理，提升网络净化的能力，助力实现“双碳”目标。

1 数据采集和预处理

基于DNS数据的矿工识别AI算法的关键在于如何构建特征向量，而DNS日志包含大量冗余信息，如DNS响应码、查询类型和回答数量等，因此需对日志预处理，减少特征向量的维度，还需考虑如何选择合适的特征使模型可以准确识别矿工。矿池特征每天自动更新，通过训练好的模型，输入特征后自动判定是否为矿工，相比传统方式，人工介入更少，准确性更高。

1.1 数据集的构建

数据集分为矿池数据、DNS日志数据、AAA日志数据三类数据，其中矿池数据用于监控矿池的活跃度并反向监控矿机，从而检测出相关联的矿工；DNS日志数据用于分析访问矿池的情况和矿工的行为特征；AAA日志数据用于溯源宽带用户以便提供提醒服务。

1.1.1 矿池数据的获取

矿池活跃排名数据从互联网获取，通过提供全球多个数字货币矿池的实时数据和统计信息的网站，获取包括各个矿池的挖矿效率、算力分布、最近的块奖励等数据。这些数据可以帮助货币矿工了解市场的实时情况和趋势，以便做出更好的决策。采集数据后，对矿池进行标签打标，格式为（矿池域名，矿池ID），如（矿池1，1），（矿池2，2）。采集出的热点域名如图1所示。

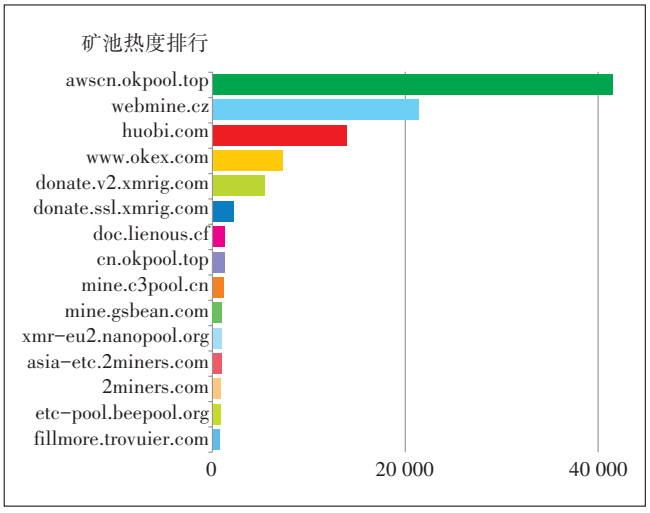

图1　矿池域名热度排行

对应图1中矿池的排行顺序，矿池打标后的数据标签如下：

（awscn.okpool.top，1）
（webmine.cz，2）
（huobi.com，3）
（www.okex.com，4）
（donate.v2.xmrig.com，5）
（donate.ssl.xmrig.com，6）
（doc.lienous.cf，7）
（cn.okpool.top，8）
（mine.gsbean.com，9）
（xmr-eu2.nanopool.org，10）
（2miners.com，11）
（etc-pool.beepool.org，12）
（fillmore.trovuier.com，13）

1.1.2 DNS数据的获取

采集运营商的DNS日志数据，格式如下：源IP地址|源端口|目的IP|目的端口|ID|域名|请求类型|解析结果|解析时间|状态码|请求或响应（q请求、r响应），例如：192.168.99.10|11764|192.168.99.1|53|11616|fatgb.88888g.cn|A|10.153.89.1;1.1.1.1|20221031141117.176|0|q。

状态码选项为：0-sucess，代表正确；1-format error，代表格式错误；2-server fail，代表服务器错误；3-nxdomain，代表无记录；4-not support type，代表不支持的类型；5-Edeny，代表错误。

1.1.3 AAA数据的获取

采集运营商AAA日志数据，AAA日志没有标准格式，但必须包含如下字段：请求IP、日志时间、公网IP、私网IP、NAT开始端口、NAT结束端口、状态。其中，状态有3种：上线、keepalive、下线。

1.2 数据预处理

挖矿行为持续且有规律，用户一般会采购专业挖矿设备或软件进行长时间挖矿，基于此特征建立用户行为轨迹分析，形成一套主动挖矿行为的规律模型，以“打点”等方式对挖矿行为长期标注，汇聚挖矿动态数据信息，识别主动挖矿行为。

步骤1：AAA账户信息和DNS日志回填。因DNS日志中只有请求IP，没有宽带信息，需根据AAA日志中用户上线信息进行匹配。匹配原则为：如果NAT的端口不为空或不为0，宽带用户为NAT转化分配的IP时，与DNS请求IP比对，若一致则判断请求IP的源端口是否在NAT端口的范围内，如果在且DNS请求时间

在用户上线的时间范围内，则匹配成功。如果NAT的端口为空或者为0，此时宽带用户为非NAT转化分配的IP，直接和DNS请求IP比对，若一致且DNS请求时间在用户上线的时间范围内则匹配成功。

步骤2：依据回填后的数据，对DNS日志的域名进行判断，形成一个序列。序列由用户名称、小时、矿池ID、访问次数4个部分组成，格式为：用户名称|小时|矿池ID1，访问次数，矿池ID2，访问次数…|小时|矿池ID1，访问次数，矿池ID2，访问次数…。举例如下：宽带用户名称|00|10，23，11，22…|01|10，22，11，19…|02|…|23|10，23，11，22。

2 算法模型

2.1 输入/输出数据特征

挖矿行为通常会产生相对规律性，如数量、频率、大小等规律；同时目标地址固定，如报文都指向矿池域名或挖矿机域名；挖矿协议不断升级的特征也需考虑。

输入数据集从DNS日志中提取有关域名、IP地址和时间戳等信息构建特征向量，通过特定筛查条件如DNS响应码、查询类型和回答数量等，简化出数据向量。输入数据可抽象为数值型数据列表，如某用户小时内的矿池访问次数[10，10，11]等。由于数据规模大，生成的模型需相对简洁，复杂度较低，并且还能保存聚类中心和标签数据。

而输出结果由于是用于判断是否为挖矿行为，只需输入特征数据，使用非监督算法，返回挖矿行为百分比，故选用的算法只需要输出内容是行为比率即可。

2.2 算法选择

基于输入数据特征和输出数据的要求，对现有算法进行对比选择，具体聚焦以下3种算法。

2.2.1 K-means算法

K-means是一种基本的聚类方法，该方法将数据点分成K个簇。K-means算法需先给定簇的数量K，然后随机初始化K个簇中心（cluster center），将每个数据点分配到最近的簇中心，并重新计算每个簇的中心。K-means算法可有效处理大规模数据集，适用于数值型数据。K-means算法基于特征空间的欧氏距离来度量数据点之间的相似性，使用场景为凸形状聚类、非凸形状聚类和确定聚类数目。

2.2.2 Bisecting K-means算法

Bisecting K-means是对K-means的改进，算法的基本思想是将所有数据点看作一个大簇，再分成为2个子簇，选择其中一个子簇重新划分，直到簇的数量达到预定的K值。每次划分，算法会选择最优的划分方式，即选择SSE（Sum of Squared Errors）最大的子簇进行划分，以保证划分的质量。Bisecting K-means比K-means算法更适用不同形状和密度的数据集。

2.2.3 Gaussian Mixture Model（GMM）算法

GMM是一种概率生成模型，与K-means不同，GMM的每个数据点不是硬分配到某个簇，而是以一定的概率属于每个簇。GMM可用于解决复杂聚类问题，特别是K-means算法解决不了的数据集中簇具有重叠或不同方向方差的问题。

基于DNS数据集的构建和预处理将数据集构造为数值型数据，挖矿行为有一定的时效特征性，适合使用以上3个算法模型。由于K-means和Bisecting K-means算法在输入/输出数据上基本一致，需针对实验结果进行选择。在算法选择上，基于机器学习的常用方法，使用混合模型Bisecting K-means和GMM同步验证，可更准确地找出矿工数据。

3 实验及结果

3.1 数据特征

以一天的数据作为测试数据，因大部分用户属于非矿工用户，使用全天所有用户的数据进行分类可能会出现较大偏差，因此需根据以下规则删除无特征的用户数据。

a）连接域名为非矿池域名的数据。本算法仅需针对直连矿池的报文进行分析，故去除非矿池连接报文可减少噪声。

b）一天内连接矿池域名10次以下的用户数据。挖矿需不断接收/发送报文，若24 h内连接次数少于10次，可认为是不小心连接上矿池或只是浏览，选择10次是因为有的币会以小时为连接次数级。

c）白名单地址相关数据。

通过以上数据集构建和数据预处理，最终形成的特征数据如下：

$$[a1,a2,a3,a4,a5,a6\cdots\cdots a24]$$

其中$a1$代表00:00—01:00访问矿池的次数，后面以此类推。

3.2 寻找最佳聚类数

K-means的惯性（inertia）是一种衡量聚类算法性

能的指标，是每个数据点与其所属簇质心距离平方和的总和。计算惯性的目的是希望让簇内的数据点越接近彼此，而簇与簇之间的距离越远，以便更好地将数据点分配到不同的簇。

通常惯性越小的模型越好，但伴随K值的增加，惯性下降的速度变慢，所以“肘部”的K值为最优选择。选择不同的K值计算惯性，这里选择簇数为3代入3个模型。K-means惯性如图2所示。

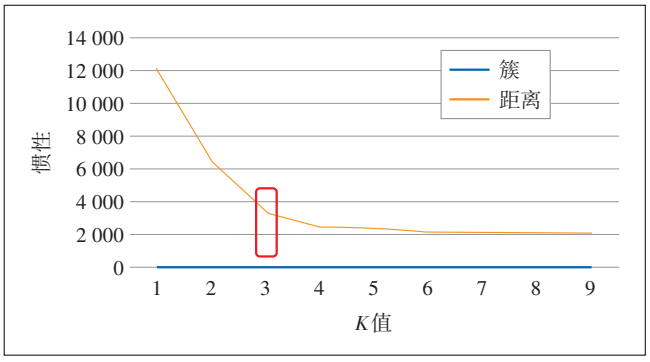

图2　K-means惯性

3.3 模型训练和结果分析

首先对K-means、Bisecting K-means训练结果进行分析，采样3 000条数并进行分类，执行后的簇中心如表1所示，标蓝部分为疑似矿工。

K-means的惯性值为3 400、Bisecting K-means的惯性值为3 140，惯性值越小说明聚集度越高，故选择Bisecting K-means模型更适合。

代入GMM模型，3类数据的分布情况和Bisecting K-means的对比如表2所示，无挖矿行为是指该用户没有访问过任何矿池；偶发挖矿行为是偶尔出现访问矿池的行为，但不频繁无规律，不足以判定为矿工；疑似挖矿行为是指有规律地访问矿池，大概率为矿工。

因2个模型在分布上几乎一致，故使用机器学习常用的方式，即使用混合模型发现挖矿行为，故模型选择为Bisecting K-means和GMM，若同时支持为挖矿行为则判断为矿工。特征数据举例如下：

[11,11,11,11,11,10,8,6,12,11,13,11,
8,9,10,10,11,16,12,14,12,11,10,9]

其中第1位数字11代表该矿工在00:00—01:00时访问了11次矿池。

3.4 模型准确度阐述

模拟挖矿场景验证模型分析结果的准确性，使用多台机器分别进行主动挖矿和使用蜜罐机被动挖矿，

表1　簇中心

算法	簇中心1	簇中心2	簇中心3	算法	簇中心1	簇中心2	簇中心3
Bisecting K-means	0.79	6.33	7.67	K-means	11.50	1.47	2.67
	0.64	1.33	11.33		12.00	0.60	4.67
	0.50	1.33	7.33		11.00	0.47	1.33
	0.36	3.33	10.33		11.00	0.73	4.33
	0.00	5.67	7.33		11.00	0.00	5.67
	0.21	1.67	5.33		8.00	0.20	1.67
	0.43	0.67	8.33		7.00	0.40	4.33
	0.79	1.33	4.67		7.00	0.73	1.33
	0.36	3.33	6.67		10.00	0.73	1.33
	0.29	5.33	10.33		9.50	0.27	9.33
	0.00	1.00	7.67		11.50	0.00	1.00
	0.21	1.33	10.33		10.50	0.20	4.67
	0.07	1.33	6.00		9.00	0.07	1.33
	0.29	5.33	6.33		9.50	0.27	5.33
	0.14	1.33	11.00		10.00	0.13	5.67
	0.21	1.33	6.67		10.00	0.20	1.33
	0.36	5.00	12.33		13.00	0.33	8.67
	0.00	4.67	10.00		15.00	0.60	1.67
	0.43	0.67	8.67		13.00	0.40	0.67
	1.50	5.00	12.67		12.00	1.40	9.67
	0.57	1.33	7.33		11.00	0.53	1.33
	0.07	1.33	11.00		10.50	0.07	5.33
	0.21	1.00	7.33		11.00	0.20	1.00
	0.07	1.33	10.67		10.00	0.07	5.33

表2　特征分布情况

算法	无挖矿行为	偶发挖矿行为	疑似挖矿行为
Bisecting K-means	0.64	0.35	0.05
GMM	0.72	0.25	0.03

对算法分析结果进行统计。

用12台机器（10台主动，2台被动）开展10轮测试，模型识别结果准确度均在80%以上。图3为一次验证结果，其中第1列为实际行为操作，第3列为模型识别结果，可看到有3台部分算法结果有误，在实际应用中可通过人工判别弥补。

4 结论

某省联通提出基于运营商DNS日志的AI算法自动发现加密货币矿工。算法采用挖矿威胁情报作为数据分析的触发点，锁定特定数据，并利用运营商DNS日志数据和AAA日志数据，结合AI算法混合模型对

主被动挖矿清单

实际行为	IP地址	算法判断	地市	矿池名称	域名
主动挖矿		Bisecting K-means：主动挖矿 GMM：主动挖矿		matpool.io	matpool.io
主动挖矿		Bisecting K-means：主动挖矿 GMM：主动挖矿		k8s.entrypoint，ops.sparkpool.com	k8s.entrypoint，ops.sparkpool.com
主动挖矿		Bisecting K-means：主动挖矿 GMM：主动挖矿		cybtc.info	cybtc.info
主动挖矿		Bisecting K-means：被动挖矿 GMM：主动挖矿		cryptonotepool.org.uk	cryptonotepool.org.uk
主动挖矿		Bisecting K-means：主动挖矿 GMM：主动挖矿		donate，v2xmrig.com	donate，v2xmrig.com
主动挖矿		Bisecting K-means：主动挖矿 GMM：主动挖矿		bitminter.com	bitminter.com
主动挖矿		Bisecting K-means：主动挖矿 GMM：主动挖矿		donate.v2xmrig.com	donate.v2xmrig.com
主动挖矿		Bisecting K-means：被动挖矿 GMM：主动挖矿		poolgpu.com	poolgpu.com
主动挖矿		Bisecting K-means：主动挖矿 GMM：主动挖矿		gbminers.com	gbminers.com
主动挖矿		Bisecting K-means：主动挖矿 GMM：主动挖矿		pool.supportxmr.com	pool.supportxmr.com
被动挖矿（病毒）		Bisecting K-means：被动挖矿 GMM：主动挖矿		donate，sslxmrig.com	donate，sslxmrig.com
被动挖矿（病毒）		Bisecting K-means：被动挖矿 GMM：被动挖矿		viabtc.com	viabtc.com

图3　算法结果验证情况

挖矿主被行为和矿工进行识别。该模型可应用在配合上级监管部门整治虚拟货币挖矿高压态势，也可为中小型企业和公众用户提供安全服务。未来将不断完善算法，提高其准确率和效率，并应用于更广泛的网络环境。

参考文献：

[1] 余文珣，余斯聪，钟英南，等. 一种基于流量特征识别挖矿程序的方法和系统：CN202010123819.2[P]. 2020-06-19.
[2] 王伟兵，孙秀兰. 数字货币洗钱黑色产业研究与打击难点分析[J]. 网络空间安全，2021，12(2)：1-7.
[3] 张琦，宫忱. 利用DNS日志数据检测僵尸网络技术探讨[J]. 现代电信科技，2016，46(4)：39-43.
[4] ZHAUNIAROVICH Y，KHALIL I，YU T，et al. A survey on malicious domains detection through DNS data analysis[J]. ACM Computing Surveys，2019，51(4)：1-36.
[5] 辛毅，高泽霖，黄伟强. 挖矿木马的检测与防护技术分析[J]. 网络空间安全，2022，13(1)：41-46.
[6] 黄子依，秦玉海. 基于多特征识别的恶意挖矿网页检测及其取证研究[J]. 信息网络安全，2021，21(7)：87-94.
[7] 宋文纳，彭国军，傅建明，等. 恶意代码演化与溯源技术研究[J]. 软件学报，2019，30(8)：2229-2267.
[8] 杨钰杰，霍云龙. 基于Cyber Kill Chain的铁路信息网络安全防御研究[J]. 铁路计算机应用，2021，30(11)：64-67.
[9] 王伟，罗鹏宇. 基于机器学习建模的DGA恶意域名检测[J]. 通信技术，2022，55(6)：753-761.
[10] 任益辰. 基于程序类基因的恶意程序相似性分析技术研究[D]. 北京：北京邮电大学，2021.
[11] 汤飞. 恶意加密货币挖矿软件的检测与防御[D]. 西安：西安电子科技大学，2020.
[12] 李建. 基于流量的P2P僵尸网络检测[J]. 计算机时代，2016(5)：45-48.
[13] 程叶霞，付俊，彭晋，等. 区块链安全分析及针对强制挖矿的安全防护建议[J]. 信息通信技术与政策，2019(2)：45-51.
[14] 赵文军. 挖矿病毒处理案例分析及思考[J]. 现代信息科技，2020，4(12)：145-147.
[15] 史博轩，林绅文，毛洪亮. 基于网络流量的挖矿行为检测识别技术研究[J]. 计算机应用研究，2022，39(7)：1956-1960.

作者简介：
周婧莹，毕业于中南大学，高级工程师，硕士，主要从事运营商云网安全防护及安全产品能力孵化工作；黎宇，毕业于华南理工大学，正高级工程师，硕士，主要从事IP网、网络安全技术和云计算技术研究和应用工作；黄坤，毕业于西安邮电大学，工程师，学士，主要从事运营商数据互联网维护及公众用户安全防护工作；梁洪智，毕业于吉林大学，工程师，学士，主要从事云资源网络信息安全合规、网络安全防护能力的建设、维护及运营等工作。