**中国民航大学**

**硕士研究生论文开题报告评审表**

学 号 2017122063

姓 名 吴优

申请学位 硕士学位

专业方向 航空工程

导师姓名 顾兆军

所在院系 中欧航空工程师学院

报告日期：2018 年 9 月 29 日

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 论文题目 | | 基于机器学习的僵尸网络检测技术研究 | | | | | | |
| 支持论文研究的科研项目 | | | | |  | | | |
| 论文选题依据（包括论文选题的意义、国内外研究现状分析等）   1. 选题的背景与意义   由于互联网技术发展迅速，同时也为国家发展带来了巨大的经济效益，我国的互联网规模在一段时期内还将不断扩大。而随着互联网的不断普及，用户人数不断增加，同时产生了多种类的新型网络应用，网络已经覆盖了生产生活的各个方面，社会发展对网络的依赖性也逐渐增强。而网络规模的告诉发展也带来一定问题，首先一些使用新型的网络应用如基于P2P或VoIP的软件对于网络的占用率较高，影响其他网络应用的运行;其次随着电子商务的兴起，网络中信息价值不断提高，导致网络安全问题日益突出；最后，互联网信息扩散十分迅速，为不良信息的传播提供了条件。近年来，航空业越来越依赖于互联网，网络环境的监管工作亟待加强。[1]  当前网络环境有流量大、种类多、发展快的特点，无论是网络的监督管理，还是安全防护，在全流量的基础上进行无疑是事倍功半，迫切需要对流量进行分类以进行有针对性的工作，因此网络流量的分类技术需要不断升级以应对日益复杂的网络环境。流量分类识别是进行各项网络分析管理工作的第一步，同时，在该方向上的研究成果也可以部分迁移至后续的分析步骤中，故流量分类技术的研究对于当前网络安全具有重大意义。[2]  流量分类的传统手段有基于端口的分析方法以及深度包分析，端口分析直接根据流量的端口信息进行分类，然而在当前网络环境下，存在大量可变端口的协议及应用，再加上端口隐匿、跳变技术的广泛使用，端口信息已经不能作为独立的分类依据，只能作为分类的辅助信息。深度包分析通过对采集的数据包内容进行直接分析，并从中提取一定模式特征，再利用这些特征去分类流量，这种方法可以得到较高的准确率，但其缺陷决定了它不适用于当前网络环境。首先，当前网络中，加密技术的使用愈加广泛，而由于加密数据包内容的随机性，该方法对加密流量几乎无效；其次，当前网络发展快速，流量类型日渐繁多，负载逐渐加大，而该方法需要维护庞大的特征库，并且需要进行频繁的更新，同时又过于依赖人工，运行代价未免过大。 | | | | | | | | |
| 机器学习算法在各个领域有广泛应用，在流量分类中应用机器学习算法，不仅可以得到较好的分类效果，分类速度也可以得到保障，同时，这种方法可以不依赖于数据内容进行分类，更加切合当前的实际需求，因此研究机器学习在网络流量分类中的应用可以有效助力网络的监管工作。[3]   1. 国内外研究现状   Roughan等人最早开始基于机器学习的流量分类研究，其中使用了k-NN算法和线性分类，并对个别数据特征的有效性进行了分析。[4]2004年，怀卡托大学的Antony等人尝试用无监督聚类算法进行分类。[5]剑桥大学的Moore等人曾于2005年对进行了流量的采集工作，并提出了可以应用于流量分类的248种流量特征，最后，利用所得数据提取相应特征，并使用改进核函数的朴素贝叶斯算法，结合特征优化算法进行了流量分类。[6]之后，各类算法都被应用到流量分类中，并且研究问题也变得更加有针对性。[7]2009年，Giuseppe等人采用深度学习结合特征优化，对加密的移动设备流量分类进行了研究。[8]2010年，Yuan等人使用支持向量机（Support Vector Machine, SVM）进行了较为准确的流量分类。[9]Tomasz等于2012年提出了一种基于C5.0决策树算法的流量分类方法。[10]2014年，Peng等人采用RBF网络研究了流量发生初期的分类技术。[11]Arash等人在2016年对 VPN流量的辨别方法进行了研究，[12]并在2017年对Tor流量的分类进行了研究。[13]近年来，基于以往的基础，也出现了更多针对网络威胁检测的研究工作。Qiu等人于2017年研究了半监督学习算法在僵尸网络检测中的应用。Wang等人结合了卷积神经网络，研究了恶意流量的检测。[14]  由于国内互联网的发展和机器学习研究的起步较晚，因此在流量分类工作上的研究成果稍微落后。2009年，徐鹏等人对基于支持向量机的流量分类技术进行了研究。[15]2012年张震等人研究了基于传播学习和半监督学习的流量分类方法。[16]潘吴斌等人于2014年研究了流量分类中的特征选择技术。[17]2017年，曹杰等人进一步研究了基于SVM的网络特征方法，同时研究了特征降维的方法。[18] | | | | | | | | |
| 论文研究方案（包括研究目标、研究内容和拟研究解决的关键问题、拟采取的研究方法、技术路线、实验方案及可行性、可能的创新之处）  1.研究目标  互联网环境日趋复杂，给网络监管带来了许多挑战，为了解决监管难题，首先需要对网络流量进行区分，从而降低监管难度。本课题的目标是利用python实现网络流量数据的预处理和特征优化，提出一种合理的方案来达成较好的流量分类效果，并实现对僵尸网络的识别。  2.研究内容  a.当前网络环境中存在的安全问题  b.各类机器学习算法的原理和特性，以及数据分析理论  c.网络协议工作原理及流量数据包的组成结构  d.比较与分析不同算法在数据上的分类效果，提出一种方案可以较好地对网络流量进行分类，并实现对僵尸网络流量的识别  3.拟研究解决的关键问题  a. 网络流量数据的特征提取和优化  b. 机器学习算法的优化改进  c. 僵尸网络的有效识别  4.拟采取的研究方案（或技术路线、实验方案）  本课题拟采取的研究课题如流程图4-1：  a.第一步是对流量数据进行预处理，包括去除无效数据、冗余数据以及数据缺失值的处理工作等。 | | | | | | | | |
| 图4-1 拟研究方案  b.数据处理过后，首先对数据样本的组成和分布进行初步分析，并确定合理的评价指标，用于评价后续的实验结果，之后对各类机器学习算法的特点进行分析，并进行实验对比，选择合适的算法。  c.确定使用的算法后，进行数据的特征选择工作，利用python实现不同的特征选择方法，并分别进行特征筛选，对不同方法的筛选结果进行对比，并根据先前确定的指标，选择最优的特征集合。  d.确定使用的数据特征后，进行算法模型的调优工作，首先在较大的参数范围内进行粗略搜索，之后再缩小搜索范围，并缩小搜索步长，最后基于交叉验证的方法，得到最优的算法模型。  e.通过对数据的观察，并结合算法自身特点，提出一种改进的方式，从而达成更良好的数据分类效果，可以从数据处理、混合方法、训练机制或者算法本身拉进行改进。  机器学习算法的选择和改进是本课题的核心。现阶段利用机器学习进行分类的方式主要为有监督学习，之后将进行其他方法的研究。有监督学习是指数据的类别已经预先进行了标定，算法模型将以数据标签作为基本事实，与分类结果进行对比，并设定一个最优化的目标，通过不断迭代更新，最终得到一个满足要求的模型。  每当F中的一个样本输入C，C会根据现有参数计算输出H，并将该结果与 | | | | | | | | |
| 样本的目标值O进行对比，并根据预测结果的偏差修正参数，得到新的分类器C’，并用C’进行下一次训练，直到训练结束，即可将得到的分类器用于新样本的分类。  其中，集成机器学习的方法是一类有效的有监督学习算法。Adaboost算法和随机森林算法是比较经典的集成机器学习算法。    图4-2 有监督学习原理图  5.可能的创新之处  一种改进的基于机器学习的僵尸网络流量分类算法 | | | | | | | | |
| 论文工作计划：  （1）2018年9月-11月  查找并阅读与课题相关的文献与资料，学习如何使用Python的各类API，实现机器学习算法以及进行数据分析。  （2）2018年12月-2018年2月  研究僵尸网络的主要运作方式，并在实验室中搭建模拟环境，用于模仿僵尸网络的行为，并收集网络流量，获取可用数据。  （3）2019年3月-4月  完成数据的预处理工作，并对数据进行初步分析，确定算法的评价指标，并对比不同算法的效果。进行特征优化工作，应用不同的特征选择策略，对选择结果进行分析，并确定最终使用的特征集合，之后利用该特征集合进行算法模型的调优工作。  （4）2019年5月-7月  算法优化，综合考虑数据和算法的特点，进行有效的理论分析，并进行现有分类方法的改进，并通过实验对比改进方法和原方法的分类效果，以验证方法的 | | | | | | | | |
| 有效性。  （5）2020年8月-12月  进行工作总结，撰写学位论文。 | | | | | | | | |
| 预期达到的目标、预期的研究成果  预期目的：  实现僵尸网络流量的准确识别，以及一般应用流量的分类  预期研究成果：  完成研究生论文的撰写，并发表一篇中文核心期刊论文 | | | | | | | | |
| 参考国内外文献情况：  国内文献 6 篇，国外文献 12 篇。  主要文献目录：（不少于10篇）  [1].柏骏, 夏靖波, 吴吉祥,等. 实时网络流量分类研究综述[J]. 计算机科学, 2013, 40(9):8-15.  [2].潘吴斌, 程光, 郭晓军,等. 网络加密流量识别研究综述及展望[J]. 通信学报, 2016, 37(9):154-167.  [3].Boutaba, Raouf, et al. "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities." Journal of Internet Services and Applications 9.1 (2018): 16.  [4].Roughan M, Sen S, Spatscheck O, et al. Class-of-service mapping for QoS:a statistical signature-based approach to IP traffic classification[C]// ACM SIGCOMM Conference on Internet Measurement. ACM, 2004:135-148.  [5].Mcgregor A, Hall M, Lorier P, et al. Flow Clustering Using Machine Learning Techniques[C]// International Passive and Active Network Measurement International Workshop, PAM 2004, April 19-20, 2004, Proceedings. DBLP, 2004:205-214.  [6].Moore, Andrew, Denis Zuev, and Michael Crogan. Discriminators for use in flow-based classification. 2005.  [7].ACETO, Giuseppe, CIUONZO, Domenico, MONTIERI, Antonio, et al. Mobile Encrypted Traffic Classification Using Deep Learning.  [8].Yuan, Ruixi, et al. "An SVM-based machine learning method for accurate internet traffic classification." Information Systems Frontiers 12.2 (2010): 149-156.  [9].Bujlow, Tomasz, Tahir Riaz, and Jens Myrup Pedersen. "A method for classification of network traffic based on C5. 0 Machine Learning Algorithm." Computing, Networking and Communications (ICNC), 2012 International Conference on. IEEE, 2012.  [10.]Peng, Lizhi, Bo Yang, and Yuehui Chen. "Hierarchical RBF neural network using for early stage internet traffic identification." Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on. IEEE, 2014. | | | | | | | | |
| [11.]Draper-Gil, Gerard, et al. "Characterization of Encrypted and VPN Traffic using Time-related." Proceedings of the 2nd international conference on information systems security and privacy (ICISSP). 2016.  [12.]Lashkari, Arash Habibi, et al. "Characterization of Tor Traffic using Time based Features." ICISSP. 2017.  [13.]Qiu, Zhicong, David J. Miller, and George Kesidis. "Flow based botnet detection through semi-supervised active learning." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.  [14.]Wang, Wei, et al. "Malware traffic classification using convolutional neural network for representation learning." Information Networking (ICOIN), 2017 International Conference on. IEEE, 2017.  [15.]徐鹏, 刘琼, 林森. 基于支持向量机的Internet流量分类研究[J]. 计算机研究与发展, 2009, 46(3):407-414.  [16.]张震, 汪斌强, 李向涛,等. 基于近邻传播学习的半监督流量分类方法[J]. 自动化学报, 2013, 39(7):1100-1109.  [17.]潘吴斌,程光,郭晓军,王艳.基于选择性集成策略的嵌入式网络流特征选择[J].计算机学报,2014,37(10):2128-2138.  [18.]曹杰. 基于SVM的网络流量特征降维与分类方法研究[D]. 吉林大学, 2017. | | | | | | | | |
| 实验设备条件  实验室计算机 | | | | | | | | |
| 指导教师意见：  本课题主要研究基于机器学习的僵尸网络流量分类识别技术。吴优同学经查阅相关国内外文献，了解国内外流量分类技术研究现状，拟收集网络流量数据，通过机器学习等算法对网络流量进行分析与分类工作，并进行僵尸网络流量的识别。主要研究内容包括：流量数据收集、数据分析和特征提取、机器学习算法的对比、优化与改进。已经完成机器学习算法的对比研究，为后续工作打下基础。  该课题研究目标明确，思路清晰，拟定方案可行。同意研究生开题。  导师签字：  年 月 日 | | | | | | | | |
| 评审小组成员 |  | | 姓 名 | | | 职 称 | 研究方向 | 签字 |
| 组长 | | 苏志刚 | | | 教授 | 航空工程 |  |
| 成员 | | 陈亚军 | | | 副教授 | 航空工程 |  |
| 成员 | | 胡雪兰 | | | 副教授 | 航空工程 |  |
| 成员 | | 徐登明 | | | 副教授 | 数学 |  |
| 成员 | | 林洁 | | | 副教授 | 数学 |  |
| 评审小组意见： | | | | | | | | |
| 开题报告成绩  （**合格或不合格**） | | | |  | | 组长签字：  年 月 日 | | |
| 备 注 | | | | | | | | |

说明：

1. 若开题评审小组对该研究生的论文题目有不同看法，请详细填写在“备注”栏内，并对是否重新选题提出明确的意见。
2. 该表于每年第三学期末完成，按规定时间交研究生部。