北京大学

北京大学软件与微电子学院

2020 级研究生组会

会

议

记

录

目录

北京大学软件与微电子学院

2020 级组会会议记录
（19-20 级交接组会）

参会人员：

19 级：贺韬 李子超 雷建辉 彭俊 戴启宇 游禹韩 符豫 黄宇 袁正光

20 级：阚婉玲 王煜 杨海兰 曾晓雨 林天睿 林权 季晓东 林坤

主持人：贺韬

2020 年 9 月 19 日

会议主题：19-20 级交接会

会议时间：2020 年 09 月 19 日（星期六）

会议地点：线上（企业微信）

主持人员：贺韬

参会人员：

19 级：贺韬 李子超 雷建辉 彭俊 戴启宇 游禹韩 符豫 黄宇 袁正光

20 级：阚婉玲 王煜 杨海兰 曾晓雨 林天睿 林权 季晓东 林坤

汇报人员：戴启宇 雷建辉

记录人员：阚婉玲

会议记录：

一、汇报内容

1.1 戴启宇

详细介绍了基于行为序列的用户异常行为分析实验过程，包括数据来源、数据处理、计算用户行为相似度和相关系数以及如何判断用户行为是否异常，并对实验结果进行分析。

简要介绍了基于深度强化学习的用户异常行为分析实验原理，包括强化学习模型、深度强化学习模型以及深度强化算法中的训练策略，并对单一异常检测和多类异常检测的实验结果进行分析。

1.2 雷建辉

详细介绍了 pcap2020-conn.log 数据以及实验构造的 9 个基于时间的网络流量统计特征和 10 个基于主机的网络流量统计特征，并用一

分类支持向量机算法进行处理，然后用 pca 算法降维数据，画散点图。最后详细地分析了实验中各个字段的正异常数据分布情况。

二、老师指导意见

1、探究 proto 字段为 udp 的数据被判断为异常的原因。

2、19 级的同学对之前所做的研究做一个阶段性的梳理与总结。

3、20 级的同学收集和汇总一些日志作为公开数据集。

4、20 级的同学阅读某一领域、方向的综述、文献，找到自己感兴趣的小方向，了解这个方向目前是否有人做、做到何种程度以及有哪些经典的方法。

5、希望 20 级的同学毕业前发表一篇文章，锻炼论文写作能力。

三、确定 20 级主持人名单

阚婉玲（本周） 林天睿 王煜 曾晓雨 杨海兰 林权 林坤 季晓东

四、确定最佳汇报人

戴启宇 雷建辉

北京大学软件与微电子学院

2020 级组会会议记录
（19-20 级交接组会 2）

参会人员:

19 级：贺韬 李子超 雷建辉 彭俊 戴启宇 游禹韩 符豫 黄宇 袁正光

20 级：阚婉玲 王煜 杨海兰 曾晓雨 林天睿 林权 费思源 林坤

主持人：彭俊

2020 年 9 月 26 日

会议主题：19-20 级交接会 2

会议时间：2020 年 09 月 19 日（星期六）

会议地点：线上（企业微信）

主持人员：彭俊

参会人员：

19 级：贺韬 李子超 雷建辉 彭俊 戴启宇 游禹韩 符豫 黄宇 袁正光

20 级：阚婉玲 王煜 杨海兰 曾晓雨 林天睿 林权 费思源 林坤

汇报人员：彭俊 符豫 李子超 游禹韩 袁正光

记录人员：林天睿

会议记录：

一、汇报内容

1.1 彭俊

工作总结：分析 Zeek IDS 日志数据和 Pcap 网络流量抓包数据，从中找出异常流量。基于 Zeek IDS 部分特征对数据聚类，根据 PyOD 框架计算结果，对异常数据标注，将无监督学习转化为监督学习。特征工程：GBDT、ANOVA 方法完成特征选择，合并日志来扩充数据。无监督学习部分：尝试 K-means/LOF/ABOD 算法等，最后选择孤立森林算法来建立无监督学习模型。（准确率 60%）

TO-DO：无监督学习打标结果只能快速筛选离群点，需要领域知识来对离群点打标。比如模仿 NSL-KDD 数据集打标签。

1.2 符豫

首先介绍了 github 的使用。

工作总结：Zeek 日志获取的数据（包括 conn,dns,http,file 日志）原始数据集处理、日志字段含义表、日志处理思路、HTTP 日志相关分析、History 字段理解 TCP 连接情况、介绍其他数据集、数据处理工具 ET Inteligence Replist。

TO-DO：验证 History 字段的含义。

1.3 李子超

工作总结：常见无监督异常检测算法的复现、TEXTCNN 模型、RCNN 模型、DPCNN 模型、XGBoost 模型、BIGRU 模型、BAT 模型。不平衡数据集处理。工作流程：数据预处理（是否不平衡、类别标签）、阅读 paper 复现模型、提升模型、迁移到西门子数据集。

TO-DO：一些无监督异常检测算法的复现、生成对抗网络、生成更好的标签、探索更好的监督学习模型。

1.4 游禹韩

工作总结：数据字段解析、DNS 解析、尝试随机森林算法、LightGBM 算法、CatBoost 算法构建 Demo。

TO-DO:尝试将这些算法应用到西门子数据集上。

1.5 袁正光

工作总结：NSL-KDD 数据集介绍、尝试进行多表融合、测试集含有训练集未出现过的小攻击类型、对 object 特征进行离散化、特征重要性分析、使用不同方法给西门子数据集打标签、缺失值补充、基于高斯混合模型聚类、自编码器做聚类。

TO-DO：自编码器做打标工作。

二、老师指导意见

1、20 级同学学习 github 入门，一起合作完成一个项目，参考 git 手册。

2、20 级同学在下次组会前浏览相关技术文档，确定想做的方向。

3、分析西门子数据的含义，看出网络的基本情况、拓扑、服务器、协议，然后做异常相关检测，找到并使用公开数据集。

4、统计工作日的工作时间，确定组会时间。

三、主持人名单

阚婉玲 林天睿（本周）王煜 曾晓雨 杨海兰 林权 林坤 费思源

四、确定最佳汇报人

李子超

北京大学软件与微电子学院

2020 级组会会议记录

参会人员：阚婉玲 王煜 杨海兰 曾晓雨 林天睿

林权 费思源 林坤 丁志昊 付子裕

主持人：王煜

2020 年 10 月 17 日

会议主题：20 级第一次线下组会-相互了解与组内工作安排

会议时间：2020 年 10 月 17 日（星期六）

会议地点：2214 会议室

主持人员：王煜

参会人员：

阚婉玲 王煜 杨海兰 曾晓雨 林天睿 林权 费思源 林坤 丁志昊 付子裕

记录人员：王煜

会议记录：

一、成员个人情况介绍

1.1 付子裕

个人情况：本科信息安全。

职业规划：可能选择依次是攻读博士、选调生、央企国企、外资、民营。岗位选择算法岗优于开发岗。

综合实践：西门子数据分析处理。

工程实践成果：按照老师说的先以发论文为主。

1.2 王煜

个人情况：智能科技方向

职业规划：就业，目前更倾向于进入体制内。

综合实践：西门子数据分析处理。目前组内分析西门子数据工作有瓶颈，打算先从公共数据集着手。

工程实践成果：按照老师说的先以发论文为主。

1.3 曾晓雨

个人情况：智能科技方向

职业规划：不做学术，大概率想往运营或者产品经理方向走。

综合实践：西门子数据分析处理。目前思路是先在公开数据集上找好的算法用起来，再交叉看可用性，如果好的话尝试应用到西门子数据上。

工程实践成果：按照老师说的先以发论文为主。

1.4 杨海兰

个人情况：软件工程方向

职业规划：想做算法方面的岗位。之前一直做算法，想沿着本科方向继续做，现在想往推荐系统工程性方向发展。

综合实践：西门子数据分析处理。目前想把 state（主要是链接特征）作为目标，用决策树去跑，看显著影响它的特征。history 应该也有用。

工程实践成果：按照老师说的先以发论文为主。

其他：有意打 kaggle 比赛，对找算法岗求职比较有用。内部大家有兴趣的可以一起组队。

1.5 阚婉玲

个人情况：软件工程方向，本科信息安全。本科主要做的特征提取和操作系统安全（具体设计大标准要检测的小项，原标准没有）方向的项目。

职业规划：研二想开发和算法各实习一次，工作首选选调。

综合实践：西门子数据分析处理，也是做流量日志分析。

工程实践成果：按照老师说的先以发论文为主。

1.6 林权

个人情况：智能科技方向。

职业规划：试下算法，算法不行就开发，偏向体制内。

综合实践：开源方向。

工程实践成果：按照老师说的先以发论文为主。

1.7 林坤

个人情况：软件工程方向。

职业规划：想做开发，往推荐系统方向发展，最后选择是走选调生。

综合实践：西门子数据分析处理。

工程实践成果：按照老师说的先以发论文为主。

1.8 丁志昊

个人情况：智能科技方向，本科管理方向。

职业规划：想从发论文的过程中感受一下对学术有没有兴趣，有的话就往上读博，否则就找工作。老师建议先学一些基础开发知识。

综合实践：开源方向

工程实践成果：按照老师说的先以发论文为主。

1.9 林天睿

个人情况：智能科技方向。

职业规划：就业开发岗。

综合实践：西门子数据分析处理。

工程实践成果：按照老师说的先以发论文为主。

1.10 费思源

个人情况：智能科技方向。半科班，大一大二做偏硬件的项目，之后做过 cv 方向的项目，做有监督的分类。

职业规划：研二实习，互联网企业就业。

综合实践：开源方向。

工程实践成果：按照老师说的先以发论文为主。

二、老师指导意见

1. 开源社区方向介绍：做一个软件去进行评估的初筛，评价开源社区。即做一个 ranking，选个大体范围再去进一步评估。基于软工开源软件评价模型基础上，加上一些开源软件的评价，做综合评价。目前有社区评估和开源项目评估方向。

对于该项目的最初想法：先分职责，细分到小职责，再划分每个成员对应什么职责，做出一个对标。这样可以对成员有个层次的了解。比如在 git 里有哪些权限，基于这些权限做出成员活动范围。

想做该方面的同学可以先了解一下什么是开源社区，成员分布什么情况，成员（职业化）在社区里如何升层。可以看一些 paper，再接触上交博士师兄了解工程性的项目。

2. 工程实践成果建议：建议大家发论文，对个人来讲很有价值。在发之前需要和老师单独讨论，合格才能发。此论文最终也可以作为毕设，但是需要查重。英文的没问题，中文的需要写说明。想读博的同学一定要发 CCF A 类或 B 类论文，其他同学发 C 类也可以。如果发论文学院也会有相应的鼓励。物联网和开源方向一直有个固定会议不发表论文也可以参加。

3. **关于论文：**首先一定要看综述。推荐最好看的就是博士论文综述，中文的好些。可以先搜好学校的博士论文综述学习，下面看不懂很正常，但可以看综述，了解这个领域的基本的知识结构，学术现状和商业工程现状都要了解。可以先写一个综述发出去，但综述不好写，需要分类，总结，评测，对比，展望。期刊什么的一般不愿意接收综述，但软件学报一般会接收。最后写论文的时候，初稿出来之后和老师讨论修改，可以随时和老师约时间。提示大家要注意好目标会议和竞赛的时间点，每年定时定点千万不要错过。

4. **其他建议：**关于未来，希望大家有个坚定的目标去做，然后要能适应别的。关于工作推进，由于师兄已经有留存技术文档，因此这一届不需要特定阶段进行论文阅读。

5. **组会流程：**①列出上次会议留下的任务。②组内成员报告（5 个人为主作报告，5 个人简单汇报，在下次组会开始前由下次组会主持人确定主报告人名单）。③再产生问题与任务。

6. **组会时间：**暂定每周一晚 18:00。

7. **下次组会内容：**组内成员浏览对应方向的文档，根据自己学习的内容或者实验的内容做报告。阚婉玲和付子裕安全方面有一定基础，下次组会可以讲一下那些数据文件的意义，帮助大家理解数据和环境。

三、组内工作

3.1 综合实践分组结果

西门子数据分析处理方向	付子裕、王煜、曾晓雨、杨海兰、阚婉玲、林权、林坤、林天睿
开源方向	丁志昊、费思源

3.2 组内资源情况

（1）实验室地点：研发楼四楼 1403。需要在进实验室前去财务处刷身份证，说明要进入导师实验室，之后才能用身份证刷开门禁进入实验室。

（2）服务器：共两台。

①显示器与服务器（windows server），开机登陆密码 abc@123。

②小服务器，一般用来存文件。开启后可以通过链接访问。

ftp server: http://pkulinux.quickconnect.to/

账号：wangzi1994328

密码：m6u3v7t4

（3）靶机：在实验室柜子里。密码与相关信息在柜子中的纸条上标明。

（4）书籍：报销费用的书籍最后需留在组内，存放于老师办公室。借还实验室书籍需在“实验室书籍借还登记.xlsx”中修改书籍持有人。

3.3 组费管理与报销事宜

目前由王煜接手管理。

可以购书报销，但需要汇总一起买，开具发票（发票抬头：北京大学、税号：12100000400002259P）之后交给负责人网上申请，签字，再去本部报销。报销过的书籍留在组内作为公共资源。

3.4 检索前沿信息任务分组

（1）大致分组：

国内组	付子裕、丁志昊、曾晓雨、林权
国外组	林天睿、阚婉玲、王煜、杨海兰、林坤、费思源

具体每人负责公司类或高校科研院所类（新闻媒体类辅助参考，因为是二手信息）由组内内部协商，最终每人写一条。

（2）收集：周五 12:00 前提交至下次组会主持人处，由组会主持人汇总发送到群里。

（3）要求：要保证来源切实，新（所以尽量搜集一手的信息）。内容上，以一条 AI 评论为例，具备评论时间、来源（论文、网站等）和原始新闻或论文的发表时间，其中英文的内容需要翻译一下。

（4）示例：

2020 年 4 月 17 日

论文作者：美国华盛顿大学(UW) 电子系统工程学院、生物工程学院

论文地址：

https://www.biorxiv.org/content/10.1101/523944v5.full.pdf

论文发表时间：2020 年 4 月 7 日

内容：美国华盛顿大学于 2020 年 4 月 7 日在 bioRxiv 平台上发表了一篇基于增长变换动力系统的尖峰神经元和种群模型的论文。论文指出，在神经形态工程学中，通常以自下而上的方式对神经种群进行建模，其中各个神经元模型通过突触相连以形成大规模的尖峰网络。然而，这些方法通常根据加标活动的某种统计量度（例如发射速率）来定义能量功能，不允许独立控制和优化神经动力学参数。该论文介绍了一种新的尖峰神经元和种群模型，其中神经元的动态和尖峰响应可以直接从网络目标或连续值神经变量（如膜电位）的能量函数中导出。该模型的主要优点在于，它可以独立控制三个神经动力学特性：（a）控制稳态种群动态，该种群动态编码了精确的网络能量功能的最小值；（b）控制网络中单个神经元产生的动作电位的形状，而不影响网络的最小值；（c）在不影响网络最小值或动作电位形状的情况下控制峰值统计和瞬态种群动态。所提出模型的核心是生长变换动力学系统的各种变体，无论网络规模和神经元连接性的类型（抑制性或兴奋性）如何，均可产生稳定且可解释的种群动态。研究者使用此网络构建了一个峰值关联存储器，与传统体系结构相比，它使用的尖峰更少，同时在高存储器负载下保持了较高的查全率。

（5）师兄整理的供参考信息源：

①谷歌学术（关键词 + 日期搜索）【前沿论文】

重点关注顶级期刊、会议论文

顶级期刊：《Nature》、《Science》

顶级会议：AAAI、ICML、ICLR、CVPR、ICCV、ACL 等

②知名院校评论、新闻 【院校进展】

MIT News: http://news.mit.edu/topic/artificial-intelligence

Stanford News: http://ai.stanford.edu/blog

CMU News: https://ai.cs.cmu.edu/newsroom

③知乎每日 ArXiv 论文专栏 【SOTA 论文】

https://zhuanlan.zhihu.com/arxivdaily

④知名公司 & 研究院 【业界动态】

Facebook AI: https://ai.facebook.com

Microsoft AI: https://www.microsoft.com/en-us/ai

Google AI: https://ai.google

IBM Research: https://www.research.ibm.com

达摩院: https://damo.alibaba.com

商汤科技、百度、旷视

⑤海外媒体 【外媒视角】

AI News 专题网站: https://artificialintelligence-news.com

谷歌新闻搜索

传统媒体：New York Times、Bloomberg、Wall Street Journal、BBC News、CNN、CNBC 等

⑥国内媒体 【中文视角】

新智源、cnBeta、凤凰新闻

四、主持人名单

阚婉玲 林天睿 王煜（本周） 曾晓雨 杨海兰 林权 林坤 费思源

付子裕 丁志昊

2020 级组会会议记录

参会人员：阚婉玲 王煜 杨海兰 曾晓雨 林天睿

林权 费思源 林坤 丁志昊 付子裕

主持人：曾晓雨

2020 年 10 月 26 日

汇报小组

第一组（本周汇报）：付子裕、阚婉玲、林天睿、王煜、曾晓雨

第二组：丁志昊、费思源、林坤、林权、杨海兰

汇报内容

网络安全基础知识

常用攻击类型（汇报人：付子裕）

异常类型	Dos	Prob	U2R	R2L
	Back	ipsweep	Buffer_overflow	ftp_write
	land	Msan	LoadModule	Guess_passwd
	Neptune	Nmap	perl	imap
	Pod	Portsweep	ps	multihop
	Smurf	Saint	rootkit	phf
	Teardrop	Satan	sqlattack	warezmaster
	Processtable		xterm	warezclient
	UdpStorm			spy
	MailBomb			sendmail
	apache2			Xlock
	Worm			Snmpguess

1. 端口扫描：端口扫描是指某些别有用心的人发送一组端口扫描消息，试图以此侵入某台计算机，并了解其提供的计算机网络服务类型(这些网络服务均与端口号相关)。

   #常用端口扫描工具：IPSweep、PortSweep、Nmap、Mscan

   #常用端口扫描类型：TCP connect()、TCP SYN、TCP FIN、IP 段扫描、TCP 反向扫描

2. DoS 攻击：DoS 是 Denial of Service 的简称，即拒绝服务攻击，是指故意的攻击网络协议实现的缺陷或直接通过野蛮手段残忍地耗尽被攻击对象的资源，目的是让目标计算机或网络无法提供正常的服务或资源访问，使目标系统服务系统停止响应甚至崩溃。

3. 僵尸网络：指采用一种或多种传播手段，将大量主机感染 bot 程序（僵尸程序）病毒，从而在控制者和被感染主机之间所形成的一个可一对多控制的网络。

4. 缓冲区溢出攻击：缓冲区溢出攻击通过往程序的缓冲区写超出其长度的内容，造成缓冲区的溢出，从而破坏程序的堆栈，使程序转而执行其它指令，以达到攻击的目的。

#可以使用此方法获得管理员权限

数据库分析

西门子数据库

PCAP 数据包理解（汇报人：阚婉玲）

数据包结构：Frame(物理层数据帧信息)、Ethernet II(数据链路层以太网帧头部信息)、Internet Protocol Version 4(网络层 IP 包头部信息)、Transmission Control Protocol(传输层报文头部信息，此处是 TCP 协议)

```
▶ Frame 875: 60 bytes on wire (480 bits), 60 bytes captured (480 bits)
▶ Ethernet II, Src: Cisco_2f:00:ef (c8:00:84:2f:00:ef), Dst: Siemens_6e:67:14 (20:87:56:6e:67:14)
▶ Internet Protocol Version 4, Src: 156.96.155.228, Dst: 10.10.10.11
▶ Transmission Control Protocol, Src Port: 3229, Dst Port: 3389, Seq: 1773, Ack: 2090, Len: 0
```

```
▼ Ethernet II, Src: Cisco_2f:00:ef (c8:00:84:2f:00:ef), Dst: Siemens_6e:67:14 (20:87:56:6e:67:14)
  ▶ Destination: Siemens_6e:67:14 (20:87:56:6e:67:14)    目的MAC地址
  ▶ Source: Cisco_2f:00:ef (c8:00:84:2f:00:ef)           源MAC地址
    Type: IPv4 (0x0800)
    Padding: 000000000000
▼ Internet Protocol Version 4, Src: 156.96.155.228, Dst: 10.10.10.11
    0100 .... = Version: 4                               IP协议版本
    .... 0101 = Header Length: 20 bytes (5)              首部长度
  ▶ Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
    Total Length: 40
    Identification: 0x3943 (14659)
  ▶ Flags: 0x4000, Don't fragment                        不支持分组
    Fragment offset: 0                                   分组偏移量为0
    Time to live: 113                                    生存时间
    Protocol: TCP (6)                                    此包内封装的上层协议为TCP
    Header checksum: 0x8433 [validation disabled]        头部数据的校验和
    [Header checksum status: Unverified]
    Source: 156.96.155.228                               源IP地址
    Destination: 10.10.10.11                             目的IP地址
▼ Transmission Control Protocol, Src Port: 3229, Dst Port: 3389, Seq: 1773, Ack: 2090, Len: 0
    Source Port: 3229                                    源端口号
    Destination Port: 3389                               目的端口号
    [Stream index: 4]
    [TCP Segment Len: 0]
    Sequence number: 1773    (relative sequence number)  序列号
    Sequence number (raw): 2013726870
    [Next sequence number: 1773    (relative sequence number)]下一个序列号
    Acknowledgment number: 2090    (relative ack number) 确认序列号
    Acknowledgment number (raw): 1859959692
    0101 .... = Header Length: 20 bytes (5)              头部长度
  ▶ Flags: 0x010 (ACK)                                   TCP标志字段
    Window size value: 65535                             流量控制的窗口大小
    [Calculated window size: 65535]
    [Window size scaling factor: -2 (no window scaling used)]
    Checksum: 0x9e9a [unverified]                        TCP数据字段的校验和
    [Checksum Status: Unverified]
    Urgent pointer: 0
  ▶ [SEQ/ACK analysis]
  ▶ [Timestamps]
```

服务器功能分析：根据端口号推测服务器功能

IP	端口号	说明
192.168.16.10	137、138	
192.168.16.11	80、8080、443	HTTPS 服务器

192.168.16.12	138	
192.168.16.20	80、8089	HTTP 服务器
192.168.16.255	137、138	NetBIOS 服务（在局域网中提供计算机的名字或 IP 地址查询服务）
192.168.17.4	61616	ActiveMQ（开放源代码消息中间件）
192.168.17.6	514/TCP	不必登录的远程 shell

CONN.LOG 和 PCAP 文件的关系（汇报人：阚婉玲）

统计结论：统计 conn.log 和 pcap 文件里重合的 IP

#对应关系有待进一步研究确认，以下结论假设对应关系已经确认

（1） 在 conn.log 文件里连接的目的 IP 均为 192.168.16.11，而在 pcap 文件里数据包的目的 IP 均为 10.10.10.11。

（2） conn.log 文件里连接记录的源端口号、目的端口号与 pcap 文件里数据包的源端口号、目的端口号不一致。

（3） conn.log 文件里连接记录的连接状态为 duser=RSTO（表示发送方发送 RST，终止连接）或者 duser=RSTOS0（表示发送方发送 SYN，然后又发送了 RST）

- RST(Reset the connection)：用于复位因某种原因引起出现的错误连接，也用来拒绝非法数据和请求。
- 有三个条件可以产生 RST 包：①建立连接的 SYN 到达某端口，但是该端口上没有正在监听的服务；②TCP 想取消一个已有连接；③TCP 接收到了一个根本不存在的连接上的分节。
- RST 攻击：假设有一个合法用户(1.1.1.1)已经同服务器建立了正常的连接，攻击者构造攻击的 TCP 数据，伪装自己的 IP 为 1.1.1.1，并向服务器发送一个带有 RST 位的 TCP 数据段。服务器接收到这样的数据后，认为从 1.1.1.1 发送的连接有错误，就会清空缓冲区中建立好的连接。这时，如果合法用户 1.1.1.1 再发送合法数据，服务器就已经没有这样的连接了，该用户就必须重新开始建立连接。

（4） pcap 文件里数据包出现[TCP Out-of-Order]错误。

- [TCP Out-of-Order]的原因一般是网络拥塞，导致顺序包抵达时间不同，延时太长或者包丢失，需要重新组合数据单元。

公开数据库

NSL-KDD（汇报人：王煜）

数据库文件构成：KDDTrain+.ARFF（完整 NSL-KDD 训练集，带有 ARFF 格式的二进制标签）、KDDTrain+.TXT（完整的 NSL-KDD 训练集，包括 attack 的标签以及严重性）、KDDTrain+_20Percent.ARFF（ KDDTrain+.ARFF 文件的 20%的子集）、KDDTrain+ _20Percent.TXT（KDDTrain+ _20Percent.TXT 文件的 20%的子集）、KDDTest+.ARFF（完整的 NSL-KDD 测试集，带有 ARFF 格式的二进制标签）、KDDTest+.TXT（完整的 NSL-KDD 测试集，包括 attack 的标签以及严重性）、KDDTest-21.ARFF（不带有严重性得分为 21 的

KDDTest+.ARFF 文件）、KDDTest-21.TXT（不带有严重性得分为 21 的 KDDTest+.TXT 文件）

网络记录向量的维度：一条记录 43 个字段，1-9 为 TCP 连接基本特征、10-22 为 TCP 连接的内容特征、23-31 为基于时间的网络流量统计特征、32-41 为基于主机的网络流量统计特征、42 为标签，43 为严重性得分

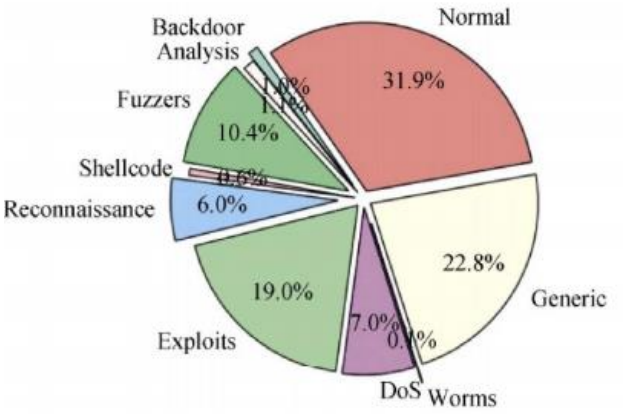
攻击类型：DoS、Smurf、Probe、R2L、U2R

UNSW-NB15（汇报人：曾晓雨）

数据库发布时间：2015 年

网络记录向量的维度：42 维特征+1 维标记

数据集正异常数据成分：9 大类攻击流量、一类正常流量，比例如下



攻击类型：Generic、Worms、DoS、Exploits、Reconnaissance、Shellcode、Fuzzers、Analysis、Backdoor

数据集缺陷：攻击类型分布极度不均衡、，不利于神经网络的训练

CICIDS-2017（汇报人：林天睿）

数据集文件构成：原始流量（Pcap 文件）、提取后的特征集（GeneratedLabeledFlows.csv 文件）、进一步预处理（MachineLearning.csv 文件）#可直接用于机器学习#

数据集正异常数据成分：共 5 天流量数据、周一的数据全部为正常流量、周二到周五混杂有各种攻击的流量(最常见的 7 种攻击类型)

攻击类型：蛮力攻击、心脏流血攻击、僵尸网络、拒绝服务攻击、分布式拒绝服务攻击、网络攻击、内网渗透

流量分析器 CICIFlowMeter：输入 pcap 文件，以 csv 文件形式输出 pcap 文件中包含的 80 多维数据包特征信息。

#配置步骤：下载安装 IDEA、安装 winpcap 包、IDEA 打开项目后配置编译器、项目、模块、为 jnetpcap 添加依赖

#代码:https://github.com/manjusakalin/CICFLOWMETER

#配置参考:https://blog.csdn.net/u010916338/article/details/84397495

异常流量分类算法

有监督算法

《基于深度学习的网络流量入侵检测研究》、《基于深度学习的网络流量分类及异常检测方法研究》、《大规模网络流量异常检测方法研究》综述阅读（汇报人：曾晓雨）

基于分类的算法：有监督的算法

#工业界主流算法：云端签名匹配法

常用的研究方法及标志性论文：

SVM：Method and system for confident anomaly detection in computernetwork traffic

NB：Early detection of network element outages based on customer trouble calls

神经网络：Applying convoluntional neural network for network intrusion detection

A deep learning approach for network intrusion detection system

算法弊端：网络攻击方式和恶意流量的模式不断增加的，训练数据集内部信息陈旧，其上训练出的算法很难投入现实应用。

无监督算法

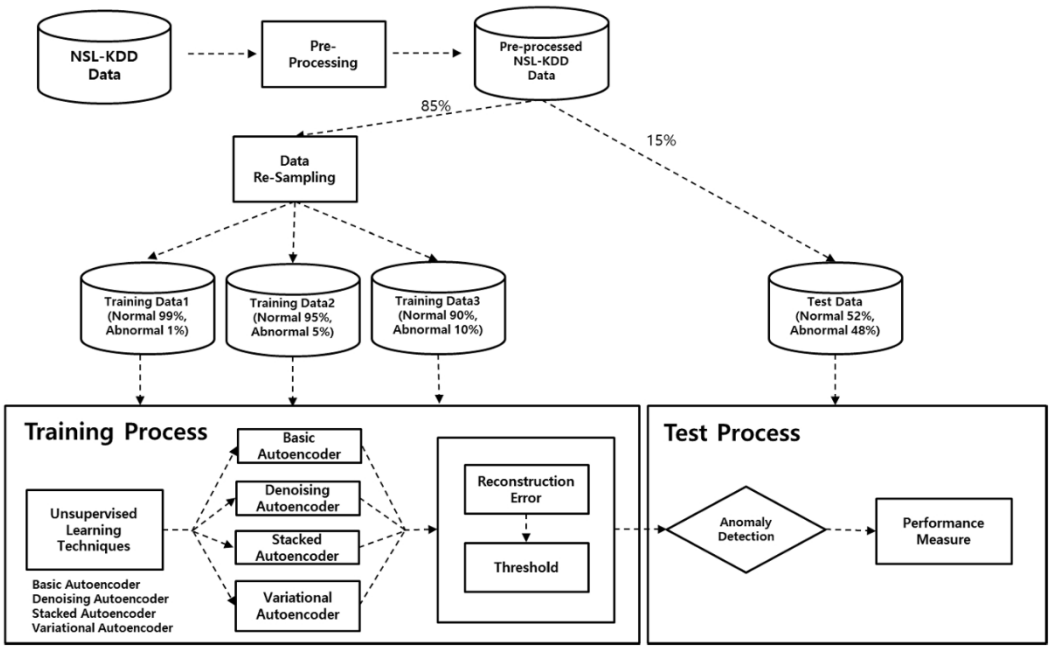
UNSUPERVISED LEARNING APPROACH FOR NETWORK INTRUSION DETECTION SYSTEM USING AUTOENCODERS 论文汇报（汇报人：林天睿）

论文使用数据集：NSL-KDD

论文创新点：提出一种基于训练数据中异常数据的百分比设置重建损失阈值的启发式方法。将训练数据中的异常数据百分比 α%当作阈值，将重建误差超过阈值的数据作为异常数据。

论文成果：本文无监督学习模型实现了 **91.7%**的精度，优于聚类算法（80%）。

论文研究框架：



论文核心公式：确定重建误差θα的值，高于此值的样本定义为异常。公式 θα=Mα+Zα*Σα，（Mα：所有样本的均值、Zα：正态分布 α 分位点、Σα：标准差）

数据分析结果：随着训练数据异常数据比例的增加，基本自动编码器、去噪自动编码器异常检测结果显示性能较低。（可能原因:随着异常数据的增加，生成的数据变得更加异构，导致重建过程更加困难。在混合正常和异常数据并存的实际情况下，具有较高复杂性的模型比基本自动编码器、去噪自动编码器性能更好）

Metric	Basic autoencoder	Denoising autoencoder	Stacked autoencoder	Variational autoencoder
Accuracy(1%)	0.9170	0.8802	0.8782	0.8766
Accuracy(5%)	0.6971	0.6845	0.9012	0.8468
Accuracy(10%)	0.6220	0.6689	0.9038	0.8626

主持人名单

阚婉玲 林天睿 王煜 曾晓雨(本周) 杨海兰 林权 林坤 费思源

付子裕 丁志昊

最佳汇报人

阚婉玲