

**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**

[2020年9月19日 1](#_Toc27846)

[2020年9月26日 4](#_Toc28586)

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：

1. **汇报内容**
   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. **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。

1. **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。

1. **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

1. **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