基于 XGBoost 与 SVM 的网络入侵检测系统：以 UNSW-NB15 数据集为例

王宇飞 (22122850)

摘要：200 字左右，应包括目的、方法、结果和结论等要素。目的。方法。结果。结论。

1 引言

1.1 提出问题（300-500 字）

1.1.1 三级标题

正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文。正文。[1]

1.1.2 三级标题

正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文。正文。正文。

1.2 求解方案分析（300-500 字）

正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文。正文。正文。

1.3 论文概述（200 字）

正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正

文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。

2 数据处理与模型构建

2.1 UNSW-NB15 数据集

UNSW-NB15 数据集是由澳大利亚新南威尔士大学堪培拉分校（UNSW Canberra）的网络靶场实验室（Cyber Range Lab）使用 IXIA PerfectStorm 工具生成的。该数据集旨在模拟现代网络环境中的正常活动和合成攻击行为，从而为网络入侵检测研究提供高质量的实验数据。通过 Tcpdump 工具捕获了 100 GB 的原始网络流量数据（如 Pcap 文件），并进一步处理生成了包含多种攻击类型和正常流量的结构化数据集。在这项工作中，UNSW-NB15 数据集中共有 257,673 个数据实例，其中包括 175,341 个训练数据实例和 82,332 个测试数据实例。

2.1.1 数据集分类情况

UNSW-NB15 数据集涵盖了一种正常流量类型和九种主要的攻击类型，包括 Fuzzers、Analysis、Backdoors、DoS（拒绝服务攻击）、Exploits、Generic、Reconnaissance（侦察攻击）、Shellcode 和 Worms。这些攻击类型覆盖了广泛的网络威胁，能够有效支持入侵检测系统的训练与评估。

表 1: UNSW-NB15 数据集中的标签及其含义

ID	Type	Description
1	Normal	Natural transaction data.
2	Analysis	An attack to invade web applications through emails ports or web scripts.
3	Backdoor	A covert attempt to circumvent normal authentication measures or other processes by allowing for secure remote access.
4	DoS	A malicious attempt to disrupt the computer resources by attacking memory.
5	Exploits	An instruction to take advantage of bugs or errors caused by unintentional behavior on the network.
6	Fuzzers	An attack to crash the system by inputting a lot of random data.
7	Generic	A technique to clash the block-cipher configuration by using hash functions.
8	Reconnaissance	A probe to evade network security controls by collecting relevant information.
9	Shellcode	A piece of code that is executed to exploit software vulnerabilities.
10	Worms	A set of virus code which can add itself to computer system or other programs.

2.1.2 数据集特征情况

本数据集的 257,673 项数据中，共含有 49 个特征（包含类别标签），其中所有特征被分为 6 组，包括流量、基础、内容、时间、附加生成和标记特征。特别地，特征 1 至 35 代表从数据包中集成收集的信息。附加生成的特征进一步分为通用特征和连接特征两部分。特征 36 至 39 被视为连接特征，而特征 40 至 47 被视为通用特征。特征 48 至 49 为标记特征。

表 2: UNSW-NB15 数据集中的特征

No.	Name	No.	Name	No.	Name	No.	Name
1	srcip	14	service	27	Sjit	40	ct_ftp_cmd
2	sport	15	Sload	28	Djit	41	ct_srv_src
3	dstip	16	Dload	29	Stime	42	ct_srv_dst
4	dsport	17	Spkts	30	Ltime	43	ct_dst_ltm
5	proto	18	Dpkts	31	Sintpkt	44	ct_src_ ltm
6	state	19	swin	32	Dintpkt	45	ct_src_dport_ltm
7	dur	20	dwin	33	tcprtt	46	ct_dst_sport_ltm
8	sbytes	21	stcpb	34	synack	47	ct_dst_src_ltm
9	dbytes	22	dtcpb	35	ackdat	48	attack_cat
10	sttl	23	smeansz	36	is_sm_ips_ports	49	Label
11	dttl	24	dmeansz	37	ct_state_ttl		
12	sloss	25	trans_depth	38	ct_flw_http_mthd		
13	dloss	26	res_bdy_len	39	is_ftp_login		

2.1.3 数据分布情况

由图1可以发现，数据集中各类别的条目数量分布并不均衡，这有可能导致模型在训练和测试过程中对于少数类别的识别效果不佳。因此，在训练与评估模型时，需要采取一些策略来处理数据集中的类别不平衡问题。

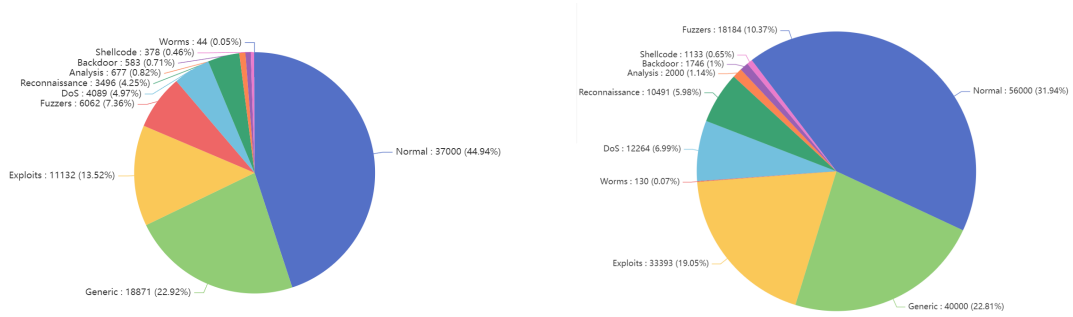

图 1: 训练集与测试集攻击分类占比情况

2.2 数据预处理

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文，如表3所示。

表 3: 表题标题

	速度/(m.s-1)	时间/s	频率/kHz
第一次			
第二次			
第三次			

2.3 算法三（300-500 字）

正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文，见公式1。

$$P(f) = \frac{1}{T} \mid 2\pi f A \exp\left[-\frac{(2\pi f\sigma)^2}{2}\right]^2 \tag{1}$$

3 算法实现描述

正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文。

3.1 算法总体框架（>500 字）

正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文

3.2 改进一及分析（>500 字）

正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文。

3.3 改进二及分析（>500 字）

正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文。正文正文正文正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文正文正文。

4 实验描述

4.1 实验数据和实验方案（>500 字）

正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文。正文。

4.2 实验一及结果分析（>500 字）

正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文。正文。

4.3 实验二及结果分析（>500 字）

正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文。正文。

5 结论（500 字）

正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文。正文。

6 学习体会和建议（300 字）

正文正文正文正文正文正文正文正文正文正文，正文正文正文正文正文正文正文正文正文正文正文。正文。正文。

参考文献

[1] Han K, Xiao A, Wu E, et al. Transformer in transformer[J]. Advances in neural information processing systems, 2021, 34: 15908-15919.

A 附录

1、图模板

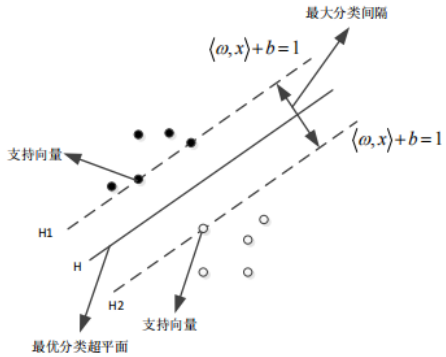

图 2: SVM 模型原理图

2、表模板

表 4: 最优算法的多指标分析

	精确率	召回率	F1 得分
石块	0.94	0.96	0.95
金属	0.92	0.97	0.95
塑料	0.96	0.89	0.93

3、公式模板

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t. } y_i\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\right) \geqslant 1, \quad i=1,2,\ldots,m \end{aligned} \tag{2}$$

4、伪代码模板

Algorithm 1 K 近邻算法步骤

Input: 训练数据集；待预测数据；

Output: 预测数据的类别;

1: 加载数据;
2: 初始化 K 值;
3: 计算预测样本与训练集中的每一个样本的距离;
4: 将距离和索引添加到有序集合中;
5: 对距离按从小到大排序方式对距离和索引的有序集合进行排序;
6: 从排序的集合中选择前 K 条数据;
7: 获得选的 K 条数据的标签;
8: 计算每一种标签的样本数量; **return** 数量最多的标签作为样本的预测值;

5、代码模板

```
#调整图片尺寸到统一大小，并扁平化为一维数据
def image_to_feature_vector(image, size=(128, 128)):

        return cv2.resize(image, size).flatten()

#提取图像在HSV颜色空间上的颜色直方图，将直方图扁平化，
#作为特征向量返回
def extract_color_histogram(image, bins=(32, 32, 32)):
        hsv = cv2.cvtColor(image, cv2.COLOR_BGR2HSV)
        hist = cv2.calcHist([hsv], [0, 1, 2], None, bins,
                [0, 180, 0, 256, 0, 256])
        if imutils.is_cv2():
                hist = cv2.normalize(hist)
        else:
                cv2.normalize(hist, hist)
        return hist.flatten()
```