南京理工大学

NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

数据科学与大数据技术课程设计

基于 KDD 数据集的网络攻击检测方法研究

学　　号：921127940122

姓　　名：顾翔

专　　业：数据科学与大数据技术

南京理工大学网络空间安全学院

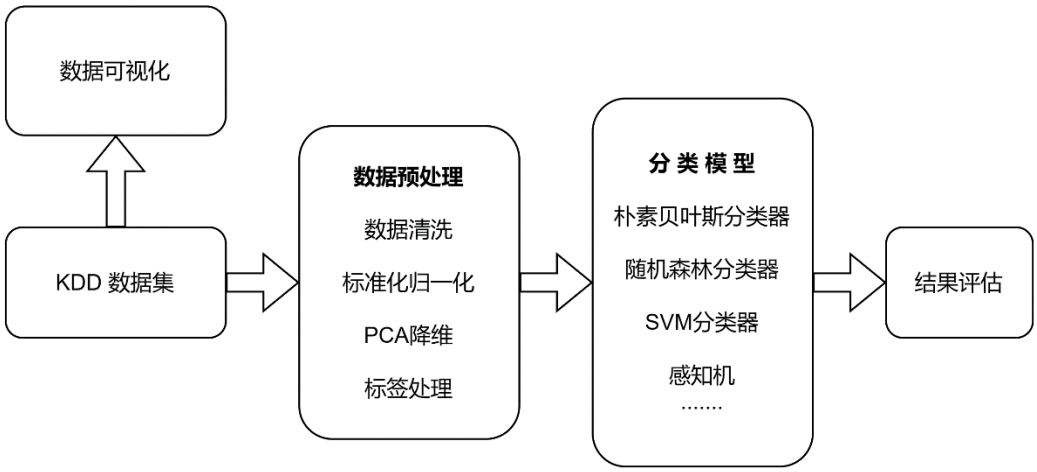
2023 年 9 月 28 日

目录

一、 概述

在日常的互联网中，每时每刻都发生着大量不同种类的网络攻击，对网络空间安全造成巨大的危害。因此，如何有效的从海量的连接中，识别出潜在的网络攻击，从而构建智能入侵检测系统，成为了一个重要的研究方向。本报告围绕 NSL-KDD 数据集，分析研究网络攻击的特征，采用多种机器学习方法探索数据集，创建分类模型来识别是否出现网络攻击及网络攻击的种类(图一)，取得了一定的成果。

关键词：网络入侵检测系统、NSL-KDD 数据集、机器学习



图一 流程图

二、数据统计描述

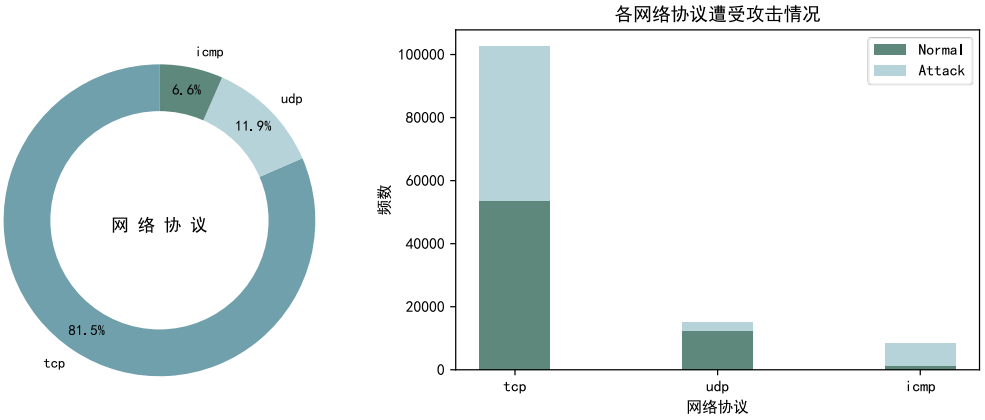
NSL-KDD 数据集是一个在 KDD99 [1]的基础上优化构建的入侵检测数据集，不包含冗余数据。在本次任务中，提供的输入数据为完整的 KDD 训练集，KDDTrain+.TXT。任务目的是通过训练，对测试集 KDDTest+.TXT 进行分类并评估。

在数据统计分析阶段，需要将其中明显出错的数据筛出改正或清除，其他数据在本阶段无需进一步处理，最后对数据进行可视化。操作步骤如下：

1. 清除出现数据缺失、错位的数据
2. 数字特征中存在以字符串形式存在的数字，将其转换成纯数字
3. 可视化并分析数据集。

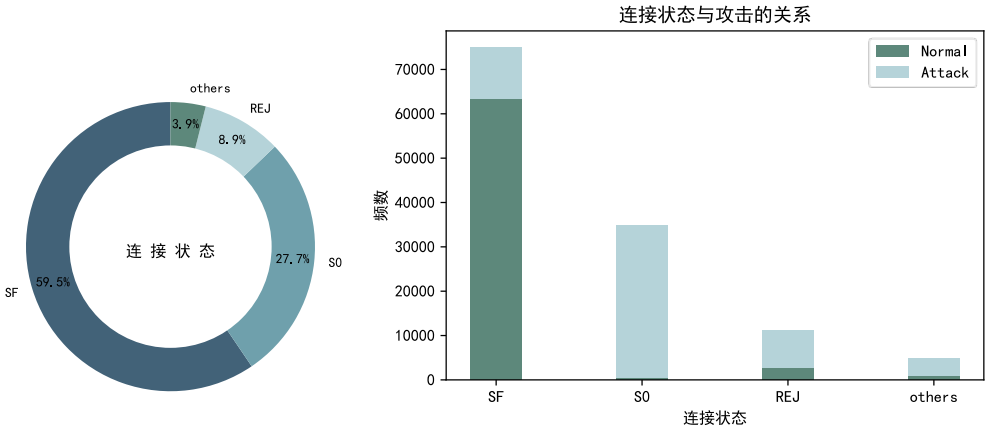
对于数据集中的文字属性，即网络协议分布、连接状态、服务内容，绘制了多个环图来直观体现数据的分布，总体而言，不同类别的数据在数据量上有较大的差异，是否发生攻击也与一些特征有较强的相关性，具体如下。

对于网络协议（图二），出现的网络协议总共有三类，其中，tcp 协议占比最高，icmp 占比最少。统计各个协议中正常访问和攻击行为，可以发现，在 icmp 协议中发生攻击的可能性更高，udp 最低。



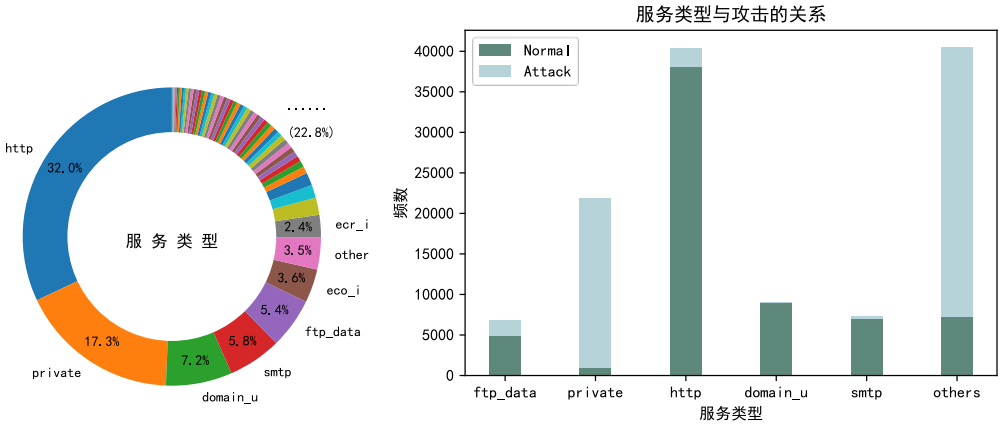
图二 网络协议与攻击的关系

连接状态与是否遭受攻击之间也有较为明显的联系（图三），通过对连接状态的可视化并分析，发现在表示正常建立连接并终止的 SF 状态，存在攻击的可能性较低，而其他的异常状态，有较高的可能性发生了网络入侵。



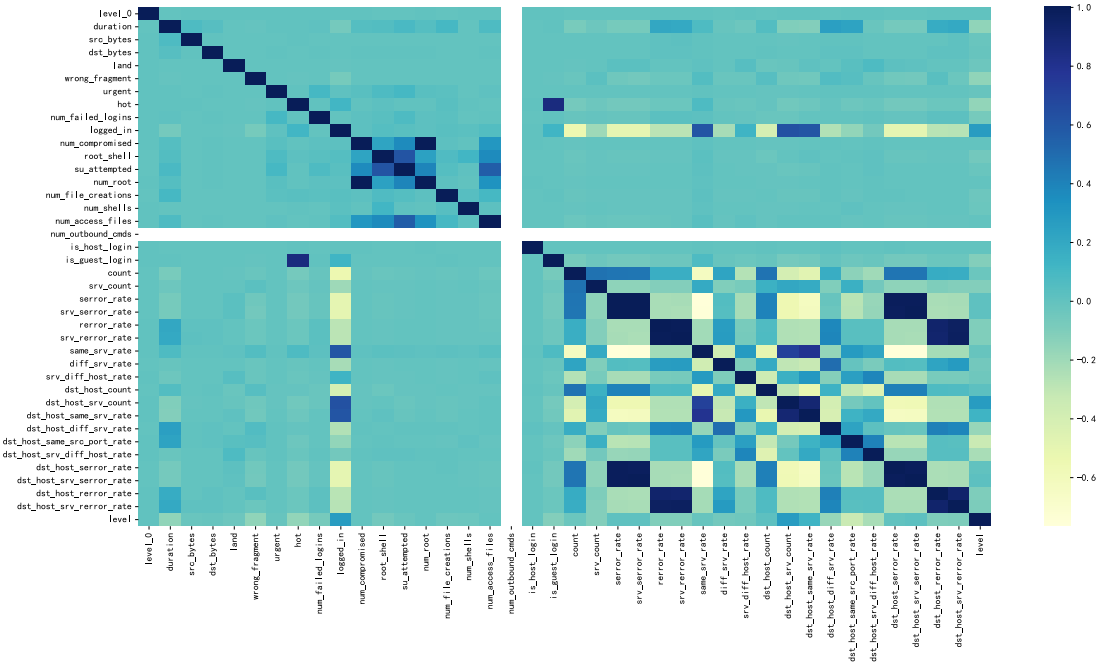
图三 连接状态与攻击的关系

服务类型的种类很多，所以很难直接找出特征与攻击间显著的关系，但是通过对服务类型的可视化（图四），可以发现不同服务类型之间发生攻击的频率也是不尽相同的，因此，可以推断，服务类型对判断是否发生网络攻击也是有一定的帮助的。



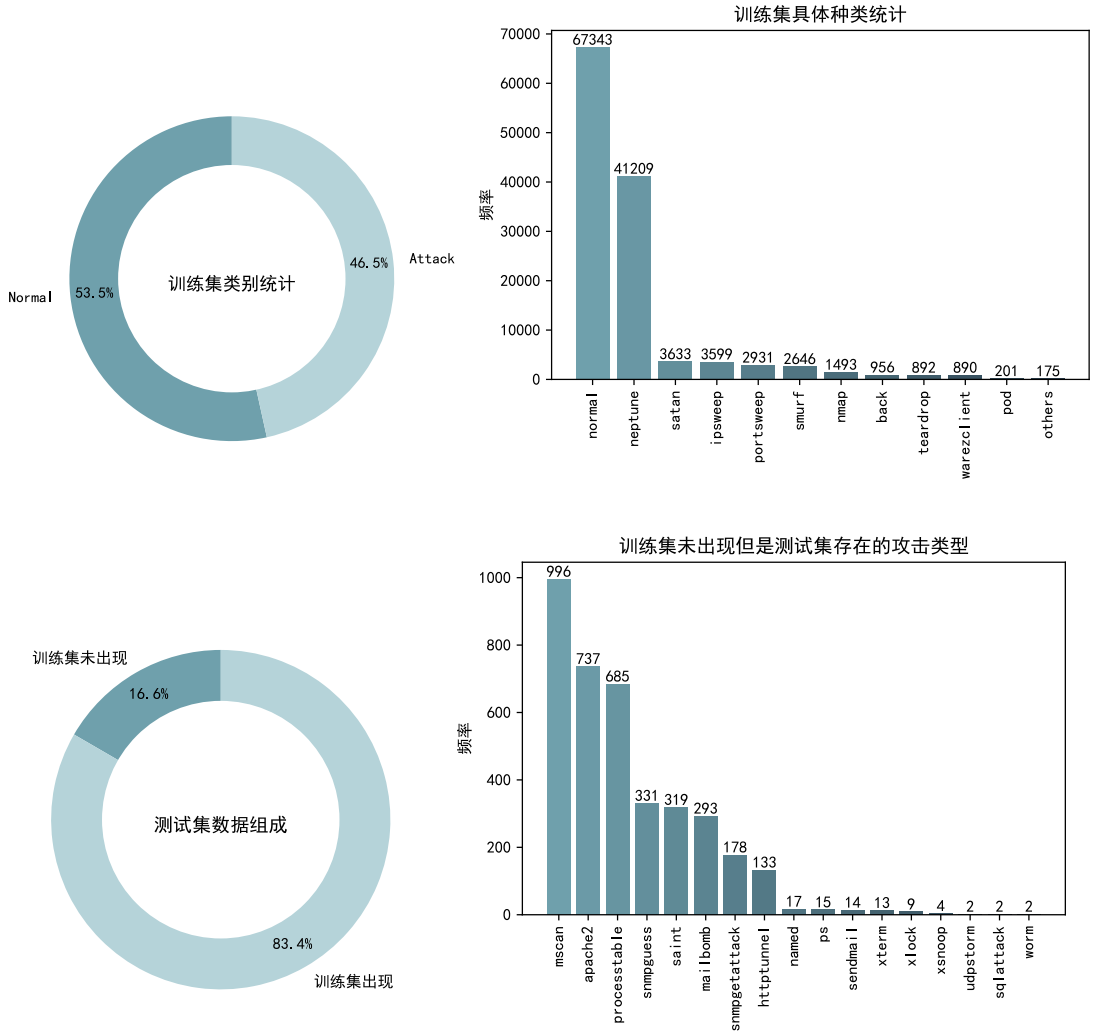
图四 服务类型与攻击的关系

对于数字特征，通过计算他们的关系矩阵可以绘制热力图（图五），通过热力图来判断数字特征之间的联系，从而对以后的特征选择可以起到一定的帮助。



图五 数字特征热力图

此外，针对需要预测的内容，即“是否存在攻击”和“攻击的种类”，在训练集可视化查看的同时，注意到，在测试集中，存在部分的攻击类别是未在训练集中出现的，将这些无法预测到的类也进行了统计并可视化结果，这些类占测试集总数据的 16.6%，因此可以预见到，在后续进行多分类的时候，在整体测试集上的分类效果将会很难超过 83.4%。(图六)



图六 攻击类型统计

三、数据预处理

目前的数据中，含有文本特征，不能直接作为输入用于分类任务，因此，需要对数据进行预处理，具体的处理步骤如下：

1. 将文本特征转化为零一向量，这样既可以将文本特征作为输入，同时，零一向量也避免了直接将其转换为连续数字造成的数据间的相互影响。
2. 将数字特征进行标准化和归一化操作
3. 对于需要分类的攻击标签，转化为“攻击(1)/正常(0)”的二分类标签和转化为数字的多分类标签，分别用于接下来的二分类和多分类任务
4. 由于输入的数据特征维数过多，因此尝试对数据进行降维处理，采用PCA[2]降维方法对数据进行特征提取。

至此，从原始数据中构建了训练集和测试集，接下来进行分类任务。

四、模型选择与参数优化

模型选择：经过在多个模型下的尝试后，最终选择了效果最优的分类模型，并通过网格搜索进行调参

（一）、二分类任务：是否发生网络攻击

在二分类任务上，采用多层感知机模型（MLP）进行分类。对于此次二分类任务，设计了包含一个输入层、一个隐藏层、一个输出层的三层感知机模型，具体细节如下：

1. 模型初始化：

   输入：$X \in R^{m \times n}$ m 个样本 n 个特征

   隐藏层：$H = \sigma(XW_h + b_h)$　　$W_h \in R^{n \times h}$

   h 为隐藏层神经元个数　　σ 为激活函数

   输出层：$O = HW_o + b_o$

2. 训练：　迭代以下过程

   前向传播阶段：输入 x，得到神经网络各层输出，计算最后一层的误差项 δ^L

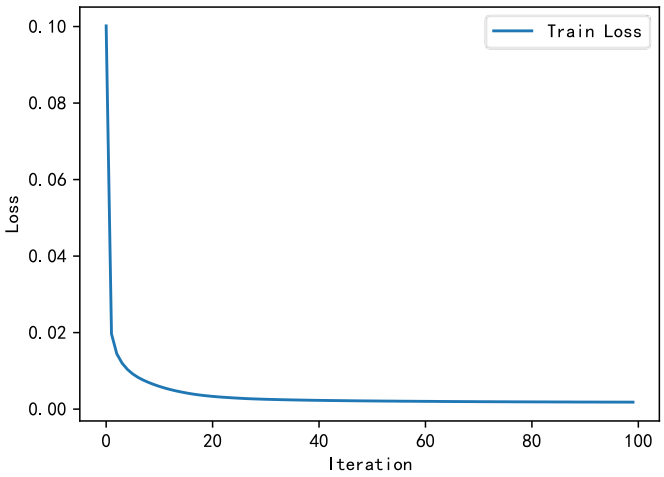
   反向传播阶段：根据链式法则，最后层的 δ^L 向前推导各层的δ^L，各层分别利用δ^L计算 $\frac{\delta L}{\delta w}$ 和 $\frac{\delta L}{\delta b}$

   更新各层的权重：$W \leftarrow W - \eta \frac{\delta L}{\delta w}$　　$b \leftarrow b - \eta \frac{\delta L}{\delta b}$

3. 超参数选取：使用经过 PCA 降维的数据作为输入；隐藏层通过网格搜索设置为：115 个神经元；损失函数采用了分类任务常规的交叉熵损失函数；激活函数为 ReLU 函数；学习率为 0.001，迭代次数为 100。

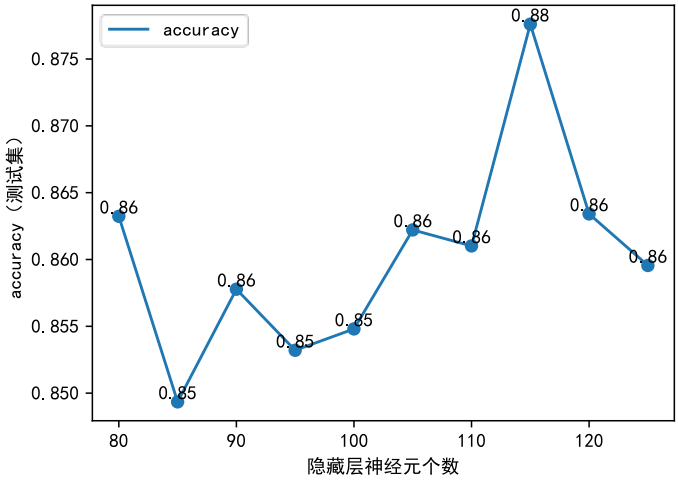
模型在全部测试集上的准确率为：88 %　（0.8776062461183568），在去除训练集未出现的攻击类别后的测试集上的准确率为：90%（0.9033631332481907）

其中，在训练时的损失变化如下（图七）：



图七 训练损失

对多层感知机，对超参数进行了交叉验证网格搜索，以下是对其神经元个数的超参数搜索结果，最终取 115 个神经元作为隐藏层神经元个数达到了较好的效果



图八 网格搜索超参数

除此之外，还对其他的分类模型进行了测试，测试结果如下（表一）：

模型类别	**准确率**	**运行时间**	**备注**
朴素贝叶斯	0.76	0.1s	未降维
决策树	0.86	11.6s	CART 决策树
随机森林	0.83	34s	50 棵树
多层感知机	**0.88**	**37s**	**网格搜索**

表一 多个模型二分类结果

（二）、多分类任务

多分类任务中，选择了随机森林[3]作为多分类的模型。随机森林是多棵决策树的集成，随机森林模型的具体细节如下（参考《数据科学导论》）：

输入：训练集 D，特征维度为 d，随机选取特征数量 m，决策树学习算法 h

输出：集成模型 H(x)

for t = 1 to T do

使用 bootstrap 采样，从训练集 D 获得大小为 n 的抽样数据 D_t

从 d 个特征中随机选取 m 个特征，基于D_t中随机选取的 m 个特征，

使用决策树模型学习得到一颗决策树$h_t(x)$

输出集成模型（分类）：$H(x) = majority\_vote(\{h_t(x)\}_{t=1}^{T}$

模型在全部的测试集上的准确率为：72%，在去除训练集未出现的攻击类别后的测试集上的准确率为：86%

除了随机森林，还对其他的分类模型进行了测试，测试结果如下（表二）：

模型类别	准确率	运行时间	备注
朴素贝叶斯	0.55	0.1s	未降维
决策树	0.69	13.6s	CART 决策树
随机森林	**0.72**	**39.8s**	**网格搜索**
SVM	0.69	29.5s	RBF 核
多层感知机	0.70	53.3s	三层

表二 多个分类模型在完整测试集多分类结果

五、结果分析

对于二分类模型，模型只需要识别到是否存在攻击，依靠训练集中的学习到的攻击发生特征，可以将正常与攻击行为进行区分，因此即使在测试数据集包含 16.6%的训练集未出现攻击类型，模型依然能准确的进行分类是否发生攻击，达到 88%的准确率，具体评估结果如下（表三）：

	Precision	Recall	F1-Score	Support
Class 0(normal)	0.81	0.93	0.87	9710
Class 1(attack)	0.94	0.84	0.89	12832
Accuracy	0.88	-	-	22542

表三 感知机二分类结果分析

对于多分类模型，模型在随机森林和多层感知机模型下的表现接近，综合考虑模型性能和训练时间，选择了随机森林模型。分析发现，模型在少样本类别上的表现不佳，当训练集的某个攻击类别样本较少时，模型在测试集无法正确分类，如 guess_passwd 攻击方法只在训练集中出现了 53 次，在测试集并没有被准确分类，导致模型测试集上的表现下降，具体统计如下表（表四）。

少样本类别名称	precision	recall	f1-score	训练集数量	测试集数量
guess_passwd	0.0	0.0	0.0	53	1231
buffer_overflow	0.0	0.0	0.0	30	20
imap	0.0	0.0	0.0	11	1
rootkit	0.1250	0.3077	0.1778	10	13

少样本类别名称	precision	recall	f1-score	训练集数量	测试集数量
loadmodule	0.0	0.0	0.0	9	2
ftp_write	0.0	0.0	0.0	8	3
multihop	0.0	0.0	0.0	7	18
phf	0.3333	0.5	0.4	4	2
perl	0.0	0.0	0.0	3	2

表四 模型在少样本类别的表现

此外，16.6%的测试集攻击类型未在训练集出现，这对分类的准确率也造成了一定的影响。这对于只需要区分是否发生攻击的二分类任务而言影响并不显著，但对于多分类任务而言，需要分类出训练集未出现的类别是相当困难的，多分类模型在整个测试集上的测试结果为 72%，而在去除这些未出现的攻击类型后，在测试集上的测试结果为 86%（+14%），与 16.6%的训练集未出现数据量吻合，这也证明了采用随机森林作为多分类模型是有效的，其性能下降主要是受到其中的少样本类别和训练集未出现的类别的影响。

六、总结

本报告围绕 NSL-KDD 数据集，首先分析研究网络攻击的特征，可视化结果，再采用多种机器学习方法探索数据集，结合效率、准确率等多个角度综合评估，并通过参数搜索，最终构建了较为完善的网络攻击检测模型。

未来，可以尝试选用一些深度学习网络来尝试进一步优化分类模型的效果。而针对目前存在的：数据集类别比例失衡导致的少样本类别表现不佳问题，可以尝试通过少样本学习[4]来改善，针对测试集中出现而训练集中未出现的类别，可以通过零次学习、重采样等方法进行尝试。

七、参考文献

[1] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*, 2009, pp. 1-6: Ieee.

[2] A. Daffertshofer, C. J. Lamoth, O. G. Meijer, and P. J. J. C. b. Beek, "PCA in studying coordination and variability: a tutorial," vol. 19, no. 4, pp. 415-428, 2004.

[3] G. Biau and E. J. T. Scornet, "A random forest guided tour," vol. 25, pp. 197-227, 2016.

[4] Y. Yu and N. J. I. A. Bian, "An intrusion detection method using few-shot learning," vol. 8, pp. 49730-49740, 2020.