**《数据挖掘导论》课程大作业**

**题目：数据挖掘在网络安全领域的应用**

**姓名：黄辉坤**

**学号：220810100104**

**班级：22级计算机科学与技术4班**

**日期：2024年12月26日**

目 录

[一、背景意义................................................................................................................1](#_Toc90576207)

[二、数据集介绍............................................................................................................1](#_Toc90576208)

[2.1 数据集的选取.................................................................................................1](#_Toc90576213)

[2.2 数据集的下载.................................................................................................2](#_Toc90576214)

[2.3 数据集的说明.................................................................................................3](#_Toc90576214)

[三、数据探索和预处理..................................................................................................3](#_Toc90576209)

[3.1 算法介绍.........................................................................................................3](#_Toc90576213)

[3.2 数据集及预处理.............................................................................................5](#_Toc90576213)

[四、算法实现................................................................................................................7](#_Toc90576209)

[4.1 平衡数据.........................................................................................................7](#_Toc90576213)

[4.2 标签归类.........................................................................................................7](#_Toc90576213)

[4.3 归一化.............................................................................................................8](#_Toc90576213)

[4.4 PCA降维.........................................................................................................8](#_Toc90576213)

[4.5 实验环境与模型调参.....................................................................................9](#_Toc90576213)

[4.6 结果分析.......................................................................................................1](#_Toc90576213)0

[五、实验总结..............................................................................................................1](#_Toc90576217)2

[六、心得体会..............................................................................................................1](#_Toc90576217)3

[参考文献......................................................................................................................1](#_Toc90576231)5

**一 、背景意义**

如今的网络攻击变得更加频繁和先进，具有规避和持续行为的高级网络威胁也在增加，网络安全问题变得尤为突出。传统的网络安全和情报技术在获取威胁情报、分析能力和利用能力方面存在局限性**[1]**。而数据挖掘在网络安全领域正发挥着越来越重要的作用，帮助我们更有效地发现和预测网络安全事件，提高对网络安全威胁的应对能力。

数据挖掘技术的发展为网络攻防领域提供了新的分析工具。数据挖掘技术可以处理和分析大规模数据集，通过机器学习、统计模型和其他分析技术，找出潜在安全威胁的蛛丝马迹，例如模式、异常和趋势等。由于传统的基于规则的网络防火墙无法适应复杂多变的攻击手段，而数据挖掘技术可以对网络流量进行分析和建模，从而识别出异常行为，提高网络入侵检测的准确性和效率。

不仅如此，数据挖掘技术可以整合外部威胁情报信息源，运用已知威胁、漏洞和攻击技术信息拓展分析内容，还可以帮助我们更好地了解网络安全事件的特征、发现潜在的安全风险，预测未来的安全威胁。

总而言之，数据挖掘和分析技术在网络攻防领域可以提高对网络安全威胁的识别、预测和响应能力，以及通过大数据技术提升网络安全防御的智能化和自动化水平**[2]**。因此数据挖掘与分析技术在网络安全领域具有广阔的应用前景和重要的研究意义。

**二 、数据集介绍**

真实的数据能够更有效训练出准确的模型，使得模型的实用意义和可解释性更强**[3]**。建立模型之前选择合适的数据集进行训练测试也同样重要。本节将基于数据的选取和下载以及对数据的说明进行描述。

### 2.1数据集的选取

CIC-IDS2017是一个公开的网络流量数据集，于2017年由加拿大网络安全研究所（CIC）所创建，旨在模拟真实的网络流量以评估和改进网络异常流量检测系统的性能。该数据集由一系列精心设计的实验生成，包含了良性和多种网络攻击类型。

数据的采集期共5天，从周一开始，周一只包括正常的流量。其余四天实施的攻击包括但不限于蛮力FTP、蛮力SSH、DoS、Web攻击、渗透、Botnet和DDoS等网络攻击。每天的网络攻击类型详见表1。

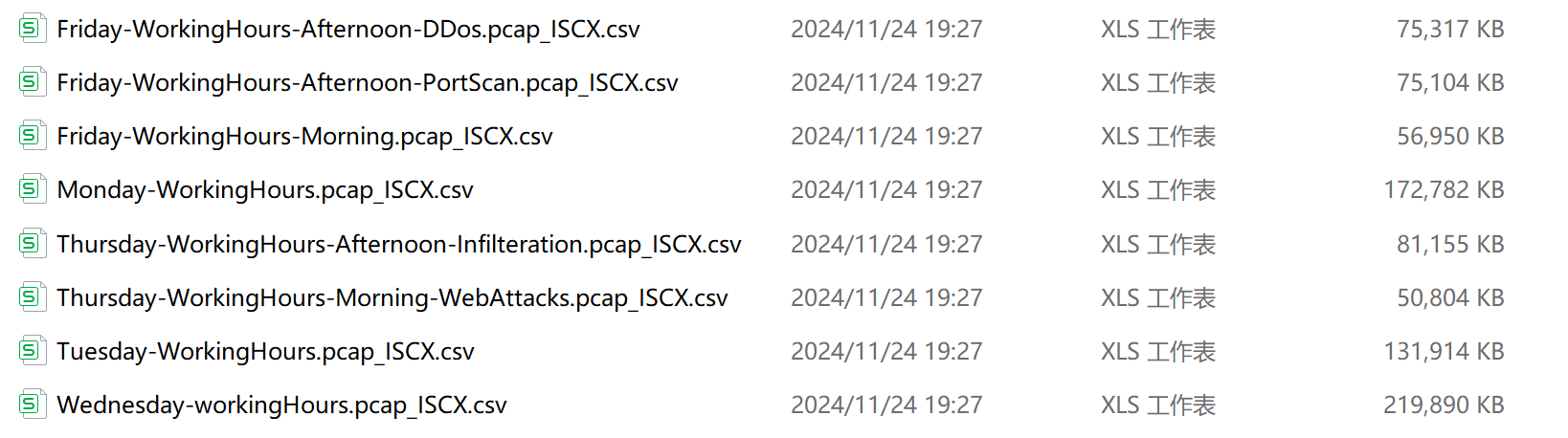
**表1 CIC-IDS2017网络流量数据类型表**

|  |  |
| --- | --- |
| Days | Labels |
| Monday | Benign |
| Tuesday | BForce,SFTP and SSH |
| Wednesday | DoS and Heartbleed Attacks  Slowloris,Slowhttptest,  Hulk and GoldenEye |
| Thursday | Web and Infiltration Attacks  Web BForce,XSS and Sql Inject.  and Cool disk |
| Friday | DD0S LOIT,Botnet ARES,  PortScans |

CICIDS2017数据集也因其真实性、多样性、丰富的标注信息、庞大的数据规模以及对技术创新的推动作用，在网络入侵检测系统测试中占有重要地位。

### 2.2数据集的下载

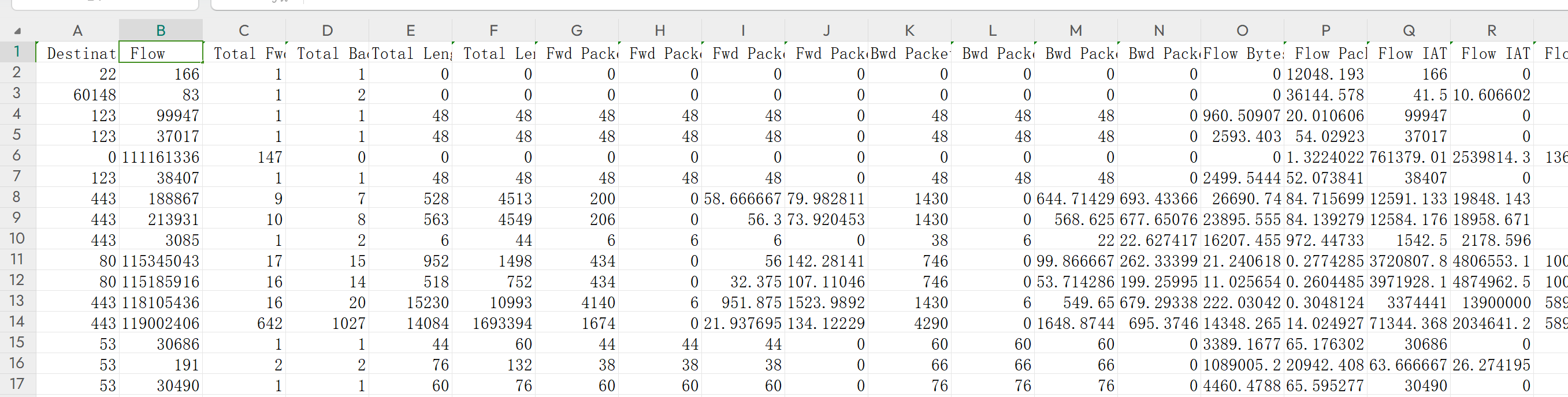
对于CIC-IDS2017数据集的下载，可以通过在其官方网站下载或者通过上传到的网盘或代码仓库上下载。以访问[官网](https://www.unb.ca/cic/datasets/ids-2017.html" \t "https://kimi.moonshot.cn/chat/_blank)为例，在下载目录中，选择下载目录名为MachineLearningCSV的数据压缩包，这是进行修剪过后的版本，剔除了IP和时间戳等不适合机器学习的属性列，适合进行机器学习训练使用。而在MachineLearningCSV目录下则是多个分散的子数据集，可见下图1，对应了表1每天的的攻击类型，本次大作业选用的就是这些数据集作为原数据集。



**图1 分散的子数据集**

### 2.2数据集的说明

CIC-IDS2017数据集的每一行数据都包含79列，其中78列为数值特征，最后1列为分类标签，可见下图2。特征列包括网络流特性，如流持续时间、数据包长度、端口、标志等，标签列的属性值则是代表这行数据属于正常流量或是哪一种网络攻击类型。



**图2 数据集部分数据实况**

对于数据集的大小和组成，数据集在5天内捕获了超过280万条实例，包括正常流量和各种网络攻击类型，如暴力FTP、Heartbleed、Botnet、DoS、DDoS、Web攻击和渗透等。

**三、数据探索和预处理**

### 3.1 算法介绍

对于CIC-IDS2017数据集进行数据挖掘，可以采用多种数据挖掘的算法和模型，但不同算法之间的原理和检测效果也有所差异。本节将对几类常见的、应用广泛的数据挖掘分类算法进行介绍研究。

**1）朴素贝叶斯**

朴素贝叶斯算法的核心思想是通过考虑特征概率来预测分类，即对于给出的待分类样本，求解在此样本出现的条件下各个类别出现的概率，哪个最大，就认为此待分类样本属于哪个类别。朴素贝叶斯算法是在贝叶斯算法的基础上进行了相应的简化，即假定给定目标值时属性之间相互条件独立**[4]**。

贝叶斯公式如下：

|  |  |
| --- | --- |
|  | (1) |

其中 P(A|B)是后验概率，P(B|A)是似然度，P(A)是类别的先验概率，P(B) 是证据。朴素贝叶斯是贝叶斯分类算法中最简单的一个，一般用于处理二分类或多分类任务。“朴素”二字体现的是特征与特征相互独立，即根据给定特征的条件概率来计算后验概率，从而进行分类。

**2）决策树**

决策树是一种树形结构，用于分类和回归问题。它的构建主要基于“信息增益”或“基尼不纯度”等指标。决策树的核心思想是通过递归地将数据集划分为多个子集，形成树状结构。每个内部节点代表一个特征的决策，每个叶子节点则是决策结果（分类或回归值）。

基尼系数是一种衡量数据不纯度的指标。在决策树的上下文中，不纯度是指数据集中不同类别（对于分类问题）或不同数值（对于回归问题）的混合程度。基尼系数的值介于0到1之间，值越小表示数据越纯。

|  |  |
| --- | --- |
|  | (2) |

其中，是类别i在数据集中的比例。

信息增益率是决策树算法中用于选择特征的一种方法，特别是在C4.5算法中。它在信息增益的基础上进一步考虑了特征的分裂能力，以避免选择那些具有许多值的特征，因为这些特征可能会产生许多分支，从而导致过拟合。

|  |  |
| --- | --- |
|  | (3) |

其中 H(T)是父节点的熵，是子节点，V是子节点的数量。

**3）KNN**

KNN是一种基于距离的分类算法，通过测量不同特征值之间的距离来进行分类。原理就是当预测一个新样本的类别时，根据它距离最近的 K 个样本点是什么类别来判断该新样本属于哪个类别，即多数投票原则。

计算测试样本与训练样本之间的距离，如欧氏距离公式：

|  |  |
| --- | --- |
|  | (4) |

**4）随机森林**

由于不同算法具有差异性，为避免算法在应对不同数据集时表现偏好性，可以把不同弱监督学习算法组成一个强监督学习算法。随机森林作为一种集成学习方法，随机森林算法的核心思想是“集成学习”，即将固定数量和不同特征的决策树集成起来**[5]**，其预测是所有树预测结果的平均（回归）或多数投票（分类）。随机森林算法的优点在于它对数据集的扰动具有强大的适应性，在许多方式中对不良事实具有较高的鲁棒性。此外，它也是一种可以处理高维特征的算法，因为随机森林算法可以自适应地减少对不重要特征的依赖。

**5）多层感知器（MLP）神经网络**

MLP（Multilayer Perceptron）是一种前馈神经网络，由输入层、隐藏层和输出层组成。它通过调整层与层之间的权重和偏置来学习输入和输出之间的映射关系。是以网络神经元算法为基础发展而来。

神经元的加权输入和激活函数如下：

|  |  |
| --- | --- |
|  | (5) |

其中  是激活函数，如Sigmoid或ReLU， 是权重，b 是偏置。

3.2 数据集及预处理

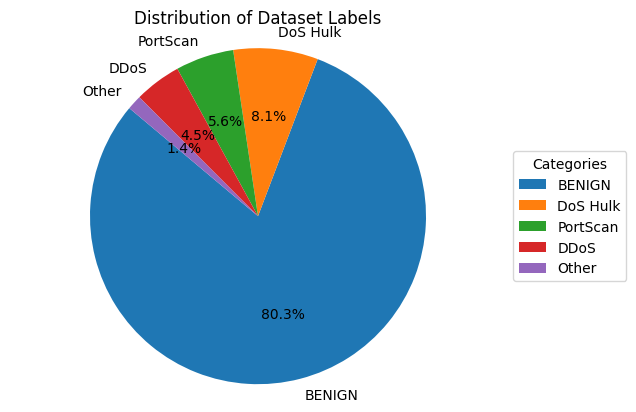
建立模型之前将选择好的数据进行预处理也同样重要**[6]**。常见的数据预处理的步骤有数据清洗、特征提取、标准化和归一化等，由于CIC-IDS2017数据集具有高维度和类别不平衡的特点，因此可能还需要采用一些特定的技术，如特征数值化、数据平衡等方法，以提高模型的性能和泛化能力。

1）合并数据集

由于原始数据集是由分散的数据集组成的，所以首先需要将这些子数据集合并在一起。合并后的数据集按标签列值分类的情况见表2。可以看到未加处理的初始数据集共有15种攻击类型，其中正常数据（BENIGN）占据绝大多数。下图3以可视化的方式直观体现了数据的不平衡性。

**表2 合并后各攻击类型情况表**

|  |  |  |
| --- | --- | --- |
| 攻击类型 | 含义 | count |
| BENIGN | 正常数据 | 2260360 |
| DoS Hulk | 一种拒绝服务攻击 | 229198 |
| PortScan | 端口扫描攻击 | 157703 |
| DDoS | 分布式拒绝服务攻击 | 127082 |
| DoS GoldenEye | 一种拒绝服务攻击 | 10289 |
| FTP-Pataor | 针对FTP服务的暴力破解攻击 | 7894 |
| SSH-Pataor | 针对SSH服务的暴力破解攻击 | 5861 |
| DoS slowloris | 一种拒绝服务攻击 | 5771 |
| DoS Slowhttptest | 一种拒绝服务攻击 | 5485 |
| Bot | 僵尸网络攻击 | 1943 |
| Web Attack | Brute Force | Web暴力破解攻击 | 1497 |
| Web Attack | XSS | 跨站脚本攻击 | 648 |
| Infiltration | 渗透攻击 | 34 |
| Web Attack | Sql Injection | SQL注入攻击 | 21 |
| Heartbleed | 安全漏洞攻击 | 11 |



**图3 数据集类别数量饼状图**

2）数据清洗与数值化

数据清洗可以去除错误、重复和不一致的数据，提高数据集的整体质量。而对数据进行数值化操作就是将一些数据集中难以识别或处理的非数值型数据转化为易于识别的数值型数据。由于CICIDS2017数据集本身已做了一些数据清洗与数值化处理，于是我们就简单对数据集进行是否有异常值和空缺值的检测，若有异常和空缺的数据则删去。并对每个属性列从0开始编号，共有78个特征列，最后一列即第79列为标签列。

1. **算法实现**

接下来对数据集进行数值化、标准化、数据平衡、PCA 降维处理等预处理操作，然后使用决策树、朴素贝叶斯、KNN、随机森林、MLP 模型进行研究实验，比较这些模型算法的检测精确率、F1、召回率等指标，对结果进行分析研究。

4.1 平衡数据

数据不平衡会导致模型偏向于多数类，导致过拟合的结果，数据平衡可以减少这种偏差，提高模型对少数类的预测能力，帮助模型学习到所有类别的特征**[7]**。

从表2我们可以看到数据总量很庞大，有近280多万条数据，其次数据在类别上分布很不均匀，正常数据（BENIGN）占了绝大多数，而有些类别却寥寥无几，这也是数据集本身的缺陷之一。

于是我们需要在缩小数据总量的同时进行过采样和欠采样，将正常流量数据与异常数据比列调至7：3左右的比例，本文就从正常数据中欠采样随机抽取10万条正常数据，而将其余类别为异常的类别调至3000条数据，共4.1万条异常数据，其中数量不足3000条的类别采用过采样的方法增至3000条数据。一方面一定程度上减轻了数据的不平衡性，另一方面减少了数据总量，降低了在训练模型时的空间和时间成本。

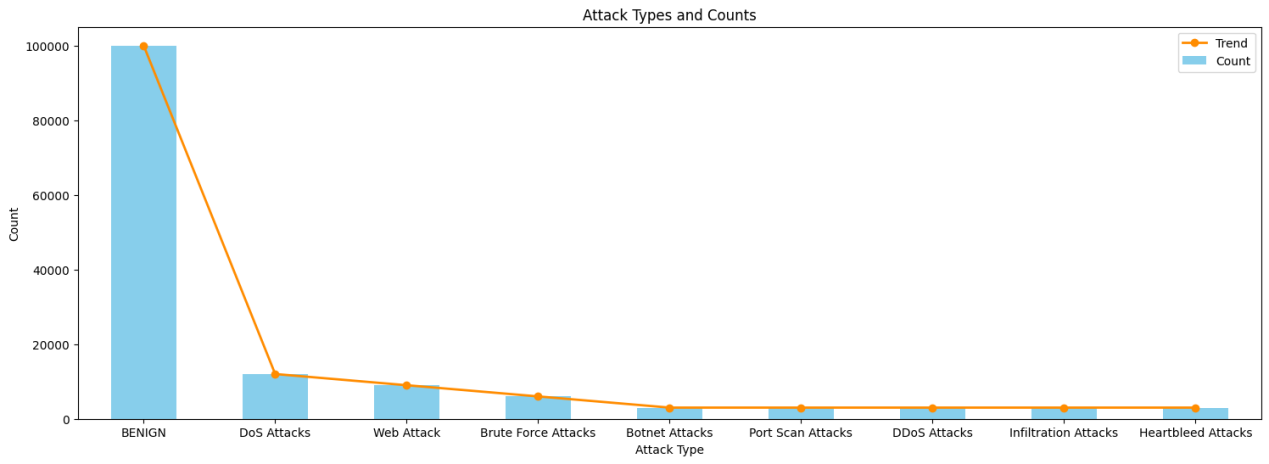
4.2 标签归类

在上一步数据平衡的基础上，又将原本标签属性值15个类别再进行归类，将攻击类型相近的归为同一类，类别减至9个，进一步降低了训练所需的时间成本，结果如下表3。

**表3 归类后的攻击类型数量表**

|  |  |
| --- | --- |
| 攻击类型 | count |
| BENIGN | 100000 |
| DoS Attacks | 12000 |
| Web Attacks | 9000 |
| Brute Force Attacks | 6000 |
| Botnet Attacks | 3000 |
| Port Scan Attacks | 3000 |
| DDos Attacks | 3000 |
| Infiltration Attacks | 3000 |
| Heartbleed Attacks | 3000 |

在经历数据平衡和标签归类后的数据集，就成为训练模型真正需要的数据集，此时的数据类别分布可见下图4。



**图4 预处理后的数据类别数量组合图**

4.3 归一化

数据归一化是数据预处理过程中常见且关键的步骤，其作用是统一量纲，避免一些小数据被吞噬**[8]**。数据归一化的方式有许多种，本文选取归一化操作中较为常用的方法，去均值和方差归一化，这是针对每一个特征维度来做的。公式如下：

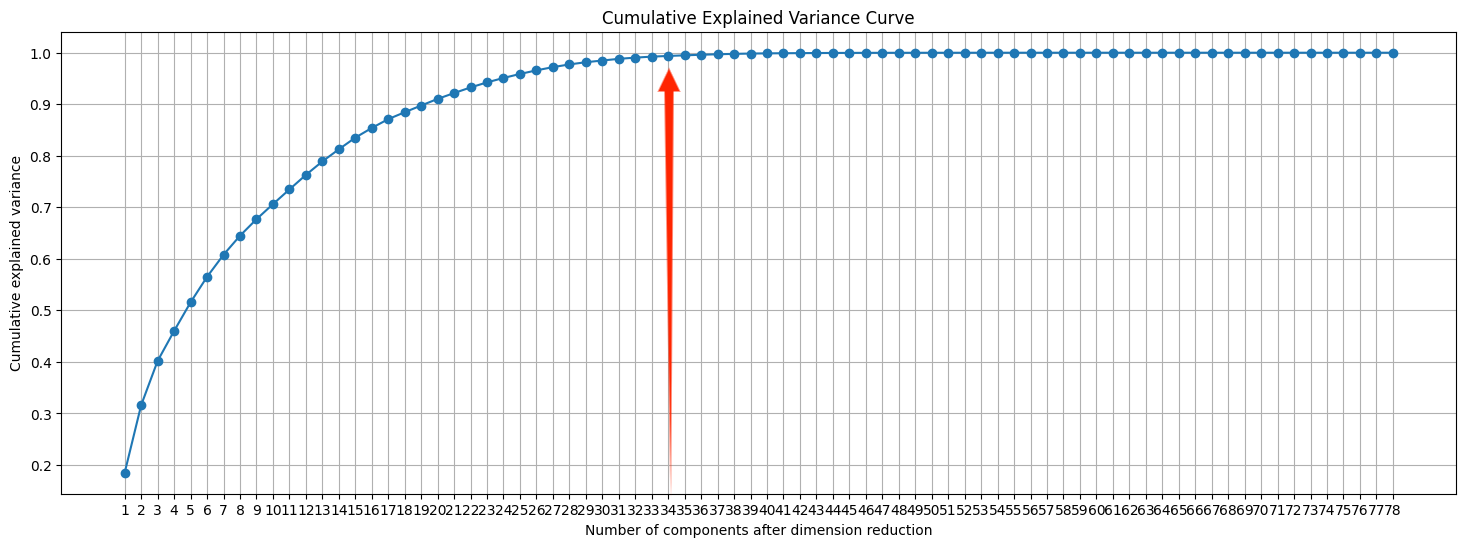
|  |  |
| --- | --- |
|  | (6) |

其中的是归一化后的数据，μ为所有样本数据的均值，σ为所有样本数据的标准差。

4.4 PCA降维

由于数据具有较高的维度，共有78个特征列，且数据样本数量在经过缩减后仍较大，需要较多的时间与空间进行实验，效率较为低下，因此，使用PCA算法进行降维。PCA算法的本质是找到数据方差最大的投影方向，方差越大，说明该投影方向包含的信息越多。数据协方差矩阵的特征值与方差对应，特征值越大，方差越大，代表着该投影方向信息含量高**[9]**。PCA通过删除特征值小的投影方向上的数据实现降维操作。

图5是经过PCA降维后的可解释方差贡献率曲线，可以看到当n\_components为34左右时，曲线开始趋于平缓。说明在34这个点之后，增加更多的主成分对解释数据方差的贡献很小，因此最好可以降维至34个特征。



**图5 可解释方差贡献率曲线图**

4.5 实验环境与模型调参

操作系统为Windows 11，CPU频率为2.50GHz，内存为16GB，512GB硬盘存储空间， 编程工具为Jupyter Notebook，编程语言Python3.9.10，选用sklearn库中的机器学习模型。

此外，由于模型的性能往往受到许多因素的影响，例如模型的[超参数](https://so.csdn.net/so/search?q=%E8%B6%85%E5%8F%82%E6%95%B0&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/m0_72410588/article/details/_blank)、数据的质量、特征选择等。其中，模型的超参数调节是模型优化中最重要的环节之一，因为正确的调参可以使模型的效果尽量最大化。在本次大作业的模型参数调优，使用sklearn的GridSearchCV，即使用了网格搜索（Grid Search）的方法来寻找所给范围内最优的超参数组合。具体模型与参数选择结构见表4。

**表4 模型参数表**

|  |  |
| --- | --- |
| 模型 | 参数/结构 |
| 决策树 | criterion= entropy, max\_depth=None, random\_state=42 |
| 朴素贝叶斯 | var\_smoothing=1e-09 |
| 随机森林 | criterion=gini, max\_depth=None, n\_estimators=20, random\_state=0 |
| MLP | activatio=tanh, hidden\_layer\_sizes=(50, 50), learning\_rate=constant, max\_iter=1000, solver=adam |
| KNN | metric=manhattan, n\_neighbors=3, weights=distance |

4.6 结果分析

在机器学习领域，评估模型的性能是至关重要的一步，因为它帮助我们理解模型在实际应用中的表现。精确度（Accuracy）、F1分数、召回率（Recall）等都是常用的评价指标，它们各自从不同的角度衡量模型的性能。表5是这些指标的介绍和定义。

**表5 模型评价指标表**

|  |  |  |
| --- | --- | --- |
| 评价指标 | 中文含义 | 定义 |
| Accuracy | 准确率 | 预测正确的样本数与总样本数的比例 |
| Precision | 精确率 | 对于每个类别，精确率是被正确预测为该类别的样本数占所有被预测为该类别的样本数的比例 |
| Recall | 召回率 | 召回率是被正确预测为该类别的样本数占所有实际为该类别的样本数的比例 |
| f1-score | F1分数 | 精确率和召回率的调和平均数，用于综合评估模型的性能 |

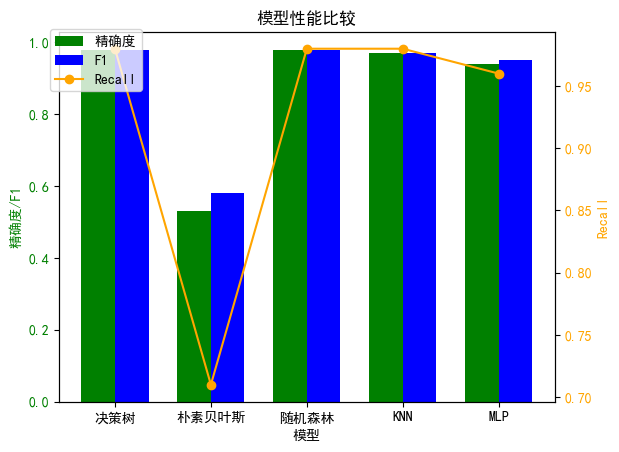
在经过上面一系列的数据预处理和模型调参等操作，我们将数据集划分成训练集：测试集=7：3的比例，并分别用到上文算法介绍中的5个模型算法进行模型训练，在设置完我们用网格搜素得到的最佳模型参数后，得到各模型的测试结果，并从准确率、宏平均精确率（每个类别精确率的算术平均值）、宏平均召回率（每个类别召回率的算术平均值）、宏平均F1分数（每个类别F1分数的算术平均值）以及性能评分（显示分类模型性能的综合评价）等多个指标进行对比分析评价，详见表6。

**表6 各模型评价指标对比表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 决策树 | 朴素贝叶斯 | 随机森林 | KNN | MLP神经网络 |
| 准确率 | 0.99 | 0.68 | 0.99 | 0.99 | 0.98 |
| 宏平均精确率 | 0.98 | 0.53 | 0.98 | 0.97 | 0.94 |
| 宏平均召回率 | 0.98 | 0.71 | 0.98 | 0.98 | 0.96 |
| 宏平均F1分数 | 0.98 | 0.58 | 0.98 | 0.97 | 0.95 |
| 性能评分 | 0.991 | 0.680 | 0.992 | 0.989 | 0.975 |

以精确度、F1、召回率为评价指标评估模型，实验结果可视化如图6所示。从整体角度分析，能明显看出朴素贝叶斯分类器的分类效果较之其它模型效果最差，各项指标均为最低。而集成学习的随机森林模型整体性能最佳，但是与其他机器学习算法相比优势并不明显。

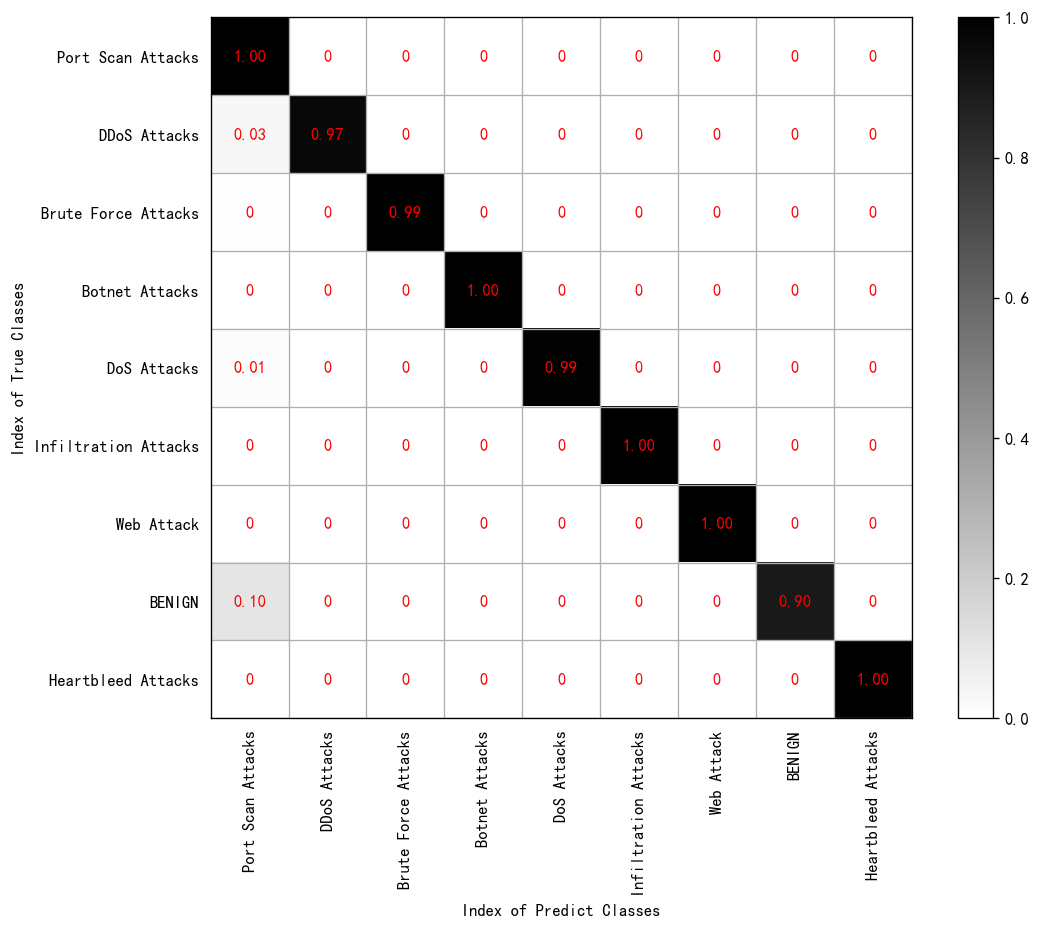
首先在准确性上，可以看到除朴素贝叶斯外的模型在准确性等指标上表现都很好。决策树和随机森林这样的结果主要是因为依赖于参数调整和数据预处理，由于在采取简单的过采样的方法后，并没有增加噪声，致使数据会出现较多重复冗余的情况，可能导致训练的模型会出现过拟合的情形；KNN算法则是依赖于多个近邻的投票结果，更方便地进行多分类任务，且与朴素贝叶斯不同，KNN对数据的分布没有特定的假设，这使得它在某些情况下更加准确**[10]**；MLP模型则适合处理大量特征的高维数据集，并且能通过调整隐藏层的数量和每层的神经元数量，可以适应各种复杂性的问题；朴素贝叶斯表现的效果不佳，这可能是因为朴素贝叶斯算法的核心假设是特征之间相互独立，但在实际数据中，这种假设往往很难完全成立，当特征之间存在较强的相关性时，会影响模型的分类精度**[11]**。



**图6 各模型评价指标对比组合图**

另外在训练时间上，决策树和朴素贝叶斯的训练时间较短，而MLP神经网络和随机森林由于是深度学习和集成学习，所需的训练时间较长。

为了进一步可视化每种分类的效果如何，可以使用用混淆矩阵来实现每种类别分类效果的可视化。混淆矩阵的作用是以表格形式展示了模型的预测结果与实际标签之间的关系，使得分类结果一目了然。便于识别模型在哪些类别上容易出错，从而有针对性地改进模型。通过比较不同模型的混淆矩阵，也可以直观地看出哪个模型的分类效果更好。如图6是随机森林的混淆矩阵图，可以看到它的总体分类效果比较好。



**图6 随机森林模型混淆矩阵图**

1. **实验总结**

在本次对CICIDS2017数据集进行分类模型的构建和评价指标的实验中，采用了五种不同的机器学习模型：决策树、随机森林、K最近邻（KNN）、朴素贝叶斯以及多层感知器（MLP）。通过这些模型，我们能够对网络流量中的正常行为和各种攻击行为进行有效分类。

决策树模型以其直观易懂的树状结构，提供了清晰的分类规则。它在处理非线性数据和特征交互时表现出色，但容易过拟合；随机森林通过集成多个决策树，显著提高了模型的泛化能力和稳定性。对CICIDS2017数据集中的复杂模式有良好的识别能力；KNN模型简单易实现，对数据的分布没有假设，能够处理多分类问题；朴素贝叶斯以其概率论为基础，适合大规模数据集。但是，它假设特征之间相互独立，这在实际应用中可能不总是成立；MLP，作为一种神经网络模型，具有很好的非线性拟合能力。通过适当的训练，MLP能够学习到复杂的数据模式。

在指标评价方面，我们使用了准确率、精确度、召回率、F1分数等指标。实验结果表明，随机森林在多个指标上表现最佳，而朴素贝叶斯在处理数据集时效果不理想。

另外，通过对整个实验过程进行分析，也能发现实验中会有以下这些不足。首先，在开始的数据预处理，平衡数据时所用的只是简单的随机过采样和欠采样，这就会有一定的随机性导致抽取出的样本不足以涵盖多数类的多样性和通过复制生成的样本使模型过拟合，可以考虑用SMOTE过采样和Tomek Links欠采样代替随机采样的方式。其次，在降低数据集维数时，除了使用PCA降维的方式，还可以通过对特征列进行相关性分析，找出与标签列相关性最高的前几个特征列，这样能将维数降到更低。最后在参数调优方面，由于出于对训练时间的考量，可能没有进行充分的网格搜索来找到最优的超参数组合。

总的来说，这次实验不仅让我深入了解了不同模型的特性和适用场景，还锻炼了我在数据预处理、特征选择、模型训练和参数调优等方面的实践能力。在未来的研究中，可以考虑上述不足，进一步优化模型，提高分类的准确性和可解释性。同时，探索更多的算法和集成学习技术，以应对复杂的网络入侵检测任务。

**六、心得收获**

通过一学期的数据挖掘课程，对我来说，不仅仅是学习一门课程完成每次的实验和作业任务那样简单，更多的是让我开始学习用各种各样的算法、模型和原理来解决我们生活实际或相关的问题，而不是简单的纸上的理论学习。

我们都知道，数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息或知识的非平凡过程。正是通过这门课程，我能学习和了解多种数据挖掘技术，包括分类、聚类、关联规则学习、异常检测等，这些技术对于解决实际问题各有用途。同时我也清楚认识到数据预处理的重要性，包括数据清洗、特征选择、数据转换等，这些步骤对于提高模型性能和准确性都至关重要。

此外，课程中的实验作业也让我有机会将理论知识应用到真实的数据集上，间接扩展了我在看待问题时的理解力，也帮助我加强了用Python编程建模解决问题的能力，相信这一定会对我未来的学习也会有所帮助。其次，一学期以来的课程学习，让我在看待解决问题需用何种方法时认识到，所谓算法，并不是说那些复杂的数学模型才是算法，哪怕是一个简单的计算公式，只要能够解决现有问题的痛点，有了自己的模型思路，它就是一个算法，只是它可能不够通用，只能解决特定需求而已。

最后是我个人认为在数据挖掘这门课的改进建议，我觉得在以后的课程学习中可以引入更多关于数据挖掘在不同领域的应用案例，如金融、医疗、社交媒体等，简单来说，就是同生活实际联系更多一点，这也侧面帮助学生更好地理解数据挖掘的实际应用和行业需求。另外，我建议可以适当增加课程的实践性，比如让学生自己参与数据的收集、整理与处理，提高参与感。

**参考文献：**

[1]2020年网络安全态势报告[J].信息安全研究,2021,7(03):198-206.

[2]许宁林.大数据时代人工智能在计算机网络技术中的应用[J].造纸装备及材料,2024,53(10):109-111.

[3]曹卿,靳荣.数据挖掘算法在大数据网络安全防御中的应用研究[J].黑龙江工业学院学报(综合版),2024,24(05):88-94.DOI:10.16792/j.cnki.1672-6758.2024.05.026.

[4]刘慧慧,岳丽颖.数据挖掘技术在网络入侵检测中的应用[J].软件, 2023, 44(11): 112-114.

[5]要丽娟.基于数据挖掘技术的入侵检测模型设计[J].电子技术, 2024, 53(10): 302-303.

[6]陈汗.基于网络流量的攻击特征分析与检测方法研究[D].东北石油大学, 2023. DOI:10.26995/d.cnki.gdqsc.2023.001243.

[7]江轲.一种改进随机森林算法及在入侵检测中的应用[D].武汉邮电科学研究院, 2021.DOI:10.27386/d.cnki.gwyky.2021.000003.

[8]卢卓尔.基于自编码器与深度学习的异常流量入侵检测技术的研究与实现[D].北京邮电大学,2024.DOI:10.26969/d.cnki.gbydu.2024.000320.

[9]谭淞.基于特征选择的入侵检测方法研究[D].西华大学, 2021. DOI: 10. 27411 /d. cnki.gscgc.2021.000115.

[10] 朱琨, 张琪. 机器学习在网络入侵检测中的应用[J]. 数据采集与处理, 2017, 32(03):479-488.

[11]贾栋豪.基于机器学习的网络入侵检测系统研究[D].华北理工大学, 2023. DOI: 10.27108/d.cnki.ghelu.2023.000652.

[12]Bilal H A .一种基于特征选择和机器学习算法的入侵检测方法研究[D].中北大学,2022.DOI:10.27470/d.cnki.ghbgc.2022.000292.