**西安电子科技大学网络与信息安全学院**

**大数据分析与安全 实验报告**

**班 级： 2118021**

**学 号：**

**姓 名：**

**电子邮箱：**

**指导教师： 李 玥**

**2024年 10 月 28 日**

|  |
| --- |
| 实验题目：网络攻击分类器——GMM |
| 实验摘要：  用非监督学习算法GMM设计了一个通用的网络攻击分类器，对于KDD数据集的检测准确率达到了58.9%，F1 Score 达到了0.508。 |
| 题目描述：  用非监督学习算法设计一个通用的网络攻击分类器，将样本分为5类。 |
| 实验内容：   1. 实验目的   设计一个通用的网络攻击分类器，能够识别五种样本类型：benign（良性）、DoS类、r2l类、u2r类、probe类。探索SOM（自组织映射）、KMeans（K均值）、GMM（高斯混合模型）等不同的无监督学习算法在网络攻击分类中的表现。   1. 模型选择 我进行了SOM、KMeans、GMM三个模型的效果对比，具体原因：  * SOM：SOM具备较好的可视化特性，能够直观显示不同类型攻击样本的空间分布，使模型更便于解释和优化。 * KMeans：由于其计算复杂度较低，KMeans在大规模数据处理上具备较好的效率。对于网络流量数据，这一优势非常重要。 * GMM：GMM具有更灵活的簇形状（非线性），适用于复杂网络流量数据的分类，有助于捕捉攻击流量样本的多样性。  1. 实验步骤  1. 数据预处理  * 收集并加载数据，对缺失值和异常值进行处理。 * 进行特征选择和数据标准化，以保证特征的有效性和模型的收敛性  2. 模型训练  * 按照无监督学习方法的要求，选定不同的无监督模型（SOM、KMeans、GMM），并进行训练。 * 使用交叉验证或不同分割策略验证模型表现。  3. 结果可视化  * 使用混淆矩阵和降维（如PCA、t-SNE）等方法，将不同类样本的分类结果进行可视化，展示模型的分离效果。 * 对比各个模型的分类效果，总结分类效果最佳的模型及其参数配置。  1. 实验结果与分析  1. 数据预处理 对“KDDTrain+”数据进行了统计、清洗，其中统计结果如下图所示。对于数据中出现的错误数据（格式不正确、数据缺失），采用了中位数填补的方法。   2. 模型训练 使用“KDDTrain+”分别对三个模型进行训练，其中SOM、KMeans、GMM的训练可视化结果如下图所示。其中SOM的可视化结果是使用了数据样本在网格上形成拓扑结构，KMeans和GMM使用了训练完成后的，原数据使用PCA聚类成2维的散点图。     3. 结果可视化 最后使用三个模型进行异常检测，并给出结果散点可视化。其中SOM、KMeans和GMM的Accuracy分别可以达到33.2%、39.3%、58.9%，F1 Score分别达到了0.341、0.339、0.508。 |
| 实验总结：  本次实验通过SOM、KMeans和GMM三种算法设计了一个网络攻击分类器，并深入探讨了每种算法的分类效果、参数调优及可视化结果。但是，在五类攻击样本中，某些类别（如u2r和r2l）之间的特征相似度较高，导致模型难以区分这些类别，分类效果不理想。 |
| 参考文献：  [1] <https://www.kaggle.com/code/fazilbtopal/popular-unsupervised-clustering-algorithms>  [2] <https://www.kaggle.com/datasets/hassan06/nslkdd>  [3] <https://pandas.pydata.org/>  [4] <https://zhuanlan.zhihu.com/p/78798251> |