****

****

**本 科 生 毕 业 论 文**

**题目:基于机器学习的网络态势感知模型的研究**

|  |  |
| --- | --- |
| 姓 名： | 范早阳 |
| 学 号： | 201424010111 |
| 专 业： | |  | | --- | | 计算机科学与技术 | |
| 年 级： | 2016级 |
| 学 院： | 信息科学技术学院 |
| 完成日期： | 2020年4月 |
| 指导教师： | 付海艳 |

**本科生毕业论文独创性声明**

本人声明所呈交的毕业论文是本人在导师指导下进行的研究工作及取得的研究成果，除了文中特别加以标注和致谢的地方外，本论文中没有抄袭他人研究成果和伪造数据等行为。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

论文作者签名： 日期： 年 月 日

**本科生毕业论文使用授权声明**

海南师范大学有权保留并向国家有关部门或机构送交毕业论文的复印件和磁盘，允许毕业论文被查阅和借阅。本人授权海南师范大学可以将本毕业论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复印手段保存、汇编毕业论文。

论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

目 录

**[1网络态势感知研究概述 1](#_Toc7330)**

**[1.1网络态势感知的研究意义 1](#_Toc11134)**

**[1.2网络态势感知的研究现状 2](#_Toc6911)**

**[2 数据集的介绍 3](#_Toc22512)**

**[2.1数据说明 3](#_Toc8115)**

**[2.2数据分析 5](#_Toc10712)**

**[3 数据预处理 6](#_Toc3979)**

**[3.1 增加目标特征 6](#_Toc6332)**

**[3.2 导入需要用到的库 6](#_Toc14907)**

**[3.3 导入数据集 7](#_Toc6841)**

**[3.4特征的探索性分析 8](#_Toc30140)**

**[3.4.1 取值分布情况 8](#_Toc19697)**

**[3.4.2 特征的分布情况 9](#_Toc13706)**

**[3.4.3 特征的选择 11](#_Toc31811)**

**[3.5数据标准化 12](#_Toc1273)**

**[4 网络态势感知模型研究 13](#_Toc30946)**

**[4.1分类建模 13](#_Toc17652)**

**[4.2支持向量机 13](#_Toc13193)**

**[4.3 Xgboost算法 15](#_Toc20701)**

**[4.4 逻辑回归 17](#_Toc22871)**

**[4.5 随机森林 18](#_Toc2935)**

**[4.6 分类算法比较 19](#_Toc2977)**

**[5模型的参数优化 19](#_Toc7155)**

**[5.1 随机森林参数优化 20](#_Toc31546)**

**[5.2 逻辑回归优化 22](#_Toc9374)**

**[6随机森林模型优化 23](#_Toc11003)**

**[6.1 可视化单个决策树 23](#_Toc18366)**

**[6.2 核心特征 25](#_Toc31635)**

**[6.3 参考模型 27](#_Toc13476)**

**[7结束语 27](#_Toc29424)**

**[参考文献 27](#_Toc1533)**

**[致谢 29](#_Toc951)**

**基于机器学习的网络安全态势感知模型的研究**

作者：范早阳 指导教师：付海艳 教授

（海南师范大学信息科学技术学院，海口，571158)

**摘 要：**网络安全态势感知融合了大数据、机器学习等技术，能够更加全面、及时的对网络安全问题作出评估和预测。网络态势感知的准确性和机器学习算法的选择息息相关，算法选择的正确性直接影响着威胁情报的准确性以及应急措施的有效性。本文利用kdd99数据集，对4种分类算法进行比较研究，然后优化算法的参数，降低算法的维度。实验结果说明改进的模型的性能得到了很大提升，随机森林模型的f1-score达到0.93。

**关键词:**态势感知；机器学习；GridSearchCV；特征提取；

**Research on network security situational awareness model based on machine learning**

Author： FanZaoYang Tutor： FuHaiYan

(Department of Computer Science and Technology, School of Information Science and Technology,HaiNan Normal University, HaiKou,571158)

**Abstract：**Cyber security situational awareness integrates big data and machine learning technology to make a more comprehensive and timely assessment and prediction of cyber security issues. The accuracy of network situational awareness is closely related to the selection of machine learning algorithms. The accuracy of algorithm selection directly affects the accuracy of threat intelligence and the effectiveness of emergency measures. In this paper, we use KDD99 data set to compare four classification algorithms, and then optimize the parameters of the algorithm to reduce the dimension of the algorithm. The experimental results show that the performance of the improved model has been greatly improved, and the f1-score of the random forest model reaches 0.93.

**Keywords**: Situation awareness; machine learning; GridSearchCV; feature extraction;

# 1网络态势感知研究概述

## **1.1网络态势感知的研究意义**

随着网络的快速发展，我国的信息系统的攻击也逐渐增加。从个人来讲：造成自己的信息被泄露。对于大型电商平台而言，则会造成大量的用户流失。对于国家而言，网络安全则是关系到国家科技、人民的资产安全。目前网络安全成为了每个国家关心的问题。与此同时，网络安全态势感知也应运而生。

网络态势感知的首次提出是在1999年，网络态势是指由各种网络设备运行状况、网络行为以及用户行为等因素所构成的整个网络的当前状态和变化趋势。态势强调环境、动态性以及实体间的关系，是一种状态、一种趋势、一个整体和宏观的概念，任何单一的情况或状态都不能称其为态势。[1]

网络态势感知技术实现了入侵行为检测、入侵率计算、入侵者身份和入侵者行为的识别、态势评估、威胁评估等功能。[2]对预防网络安全问题提供了强有力的保障。

## 1.2网络态势感知的研究现状

网络安全态势感知的关键技术有：网络安全要素提取、网络态势安全理解、网络安全态势预测；在数据来源方面由于数据的产生平台、格式不一样等等给数据融合带来了很大的困难，目前融合技术仍然存在不足。

文献[17]指出现在的网络安全态势感知主要关注安全态势，并没有集成现有的各单元网络管理技术，无法实现对全局态势的综合评估和展现；对于态势因子的综合，以层次结构、加权函数为主流的评估方法，尽管简单方便直观、便于于分析，但是仍然缺少科学理论支持，不是很客观。

文献[19]指出网络安全态势感知的研究是一个正处于发展中的课题,大部分研究都集中在重构攻击活动方面,基本 都是网络入侵检测领域研究的延伸,已有很好的基础但也有很多问题需要研究和解决，但是网络安全态势感知的研究仍处于初级阶段。

文献[20]指出网络安全态势感知缺少合理的算法和有效的技术平台。网络态势感知的准确性和大数据分析中的算法的选择息息相关，算法选择的正确性直接影响着威胁情报的准确性以及应急措施的有效性。

文献[21]指出网络态势感知技术能够综合网络、安全和应用系统等各方面因素,从整体上动态反映网络的安全状况和运行状况,并对其发展趋势进行关联分析和评估预测。结合大数据技术特有的海量存储、并行计算、高效查询等特点,为大规模网络安全态势感知技术的突破创造了机遇。

文献[22]指出目前我国科技发展使得网络安全态势感知技术也愈加丰富，技术人员能够通过对大数据的综合分析和挖掘，对网络安全态势进行科学的感知和分析，促进网络技术在新时期的健康发展。

伴随人工智能的快速发展，机器学习对未来的态势感知预测也会越来越准确。基于人工神经网络的安全态势预测技术因为全局优化、收敛速度等特点，在未来的研究领域也是非常热门的。

# 2 数据集的介绍

## 2.1数据说明

Kdd99数据集总共有500万条数据，研究选择其中的20%数据分别作为训练集和测试集。其中训练集有49万条数据，测试集有30多万条数据。其数据的每个特征及其含义见表1

**表1数据集41个特征名称和含义**

|  |  |  |  |
| --- | --- | --- | --- |
| 特征名称 | 含义 | 特征名称 | 含义 |
| duration | 连接持续时间 | is\_host\_login | 登录是否属于“hot”列表 |
| protocol\_type | 协议类型 | is\_guest\_login | 是否guest登录 |
| service | 目标主机网络服务类型 | count | \*目标主机的连接数 |
| flag | 连接正常或错误的状态 | srv\_count | \*服务的连接数 |
| src\_bytes | 源主机到目标主机的数据的字节数 | serror\_rate | \*目标主机的连接中出现“SYN” 错误 |
| dsr\_bytes | 目标主机到源主机的数据的字节数 | srv\_serror\_rate | \*相同服务的连接中，出现“SYN” 错误的 |
| land | 连接来自/送达同一个主机/端口 | rerror\_rate | 目标主机的连接中，出现“REJ” 错误的 |
| wrong\_fragment | 错误分段的数量 | srv\_rerror\_rate | \*服务的连接中，出现“REJ” 错误的 |
| urgent | 加急包的个数 | same\_srv\_rate | \*目标主机的连接中，与当前连接具有相同服务的 |
| hot | 访问系统敏感文件和目录的次数 | diff\_srv\_rate | \*目标主机的连接中，与当前连接具有不同服务的连接 |
| num\_failed\_logins | 登录尝试失败的次数 | srv\_dif\_host\_rate | \*服务的连接中，与当前连接具有不同目标主机的连接 |
| logged\_in | 是否登录 | dst\_host\_count | 前100个连接中，与当前连接具有相同目标主机的连接数 |
| num\_compromised | compromised条件（\*\*）出现的次数 | dst\_host\_srv\_count | 前100个连接中，与当前连接具有相同目标主机相同服务的连接数 |
| root\_shell | 是否获得root shell | dst\_host\_same\_srv\_rate | 前100个连接中与当前连接具有相同目标主机相同服务的连接所占的百分比 |
| su\_attempted | 是否出现’su root’ | dst\_host\_diff\_srv\_rate | 前100个连接中，与当前连接具有相同目标主机不同服务的连接所占的百分比 |
| num\_root | root用户访问次数 | dst\_host\_same\_src\_port\_rate | 前100个连接中，与当前连接具有相同目标主机相同源端口的连接所占的百分比 |
| num\_file\_creations | 文件创建操作的次数 | dst\_host\_srv\_diff\_host\_rate | 前100个连接中，与当前连接具有相同目标主机相同服务的连接中，与当前连接具有不同源主机的连接所占的百分比 |
| num\_shells | 使用shell命令的次数 | dst\_host\_serror\_rate | 前100个连接中，与当前连接具有相同目标主机的连接中，出现SYN错误的连接所占的百分比 |
| num\_access\_files\_ | 访问控制文件的次数 | dst\_host\_srv\_serror\_rate | 前100个连接中，与当前连接具有相同目标主机相同服务的连接中，出现SYN错误的连接所占的百分比 |
| num\_outbound\_cmd | 一个FTP会话中出站连接的次数 | dst\_host\_rerror\_rate | 前100个连接中，与当前连接具有相同目标主机的连接中，出现REJ错误的连接所占的百分比 |
| dst\_host\_srv\_rerror\_rate | 前100个连接中，与当前连接具有相同目标主机相同服务的连接中，出现REJ错误的连接所占的百分比 | | |

注：表格\*字符代表过去2秒内，与当前连接具有相同的

## 2.2数据分析

目标特征总共有5大类，细分有39小类见表2

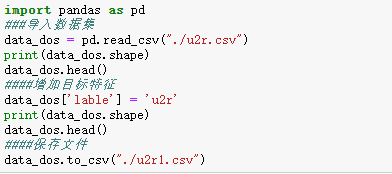
**表2目标特征分析**

|  |  |  |
| --- | --- | --- |
| 大类 | 含义 | 小类 |
| dos | 拒绝服务攻击 | back、land、neptune、pod、smurf、teardrop |
| probe | 监视和探测其他活动 | ipsweep、nmap、portsweep、satan |
| u2r | 普通用户对本地超级用户特权的非法访问 | bufffer、overflow、loadmodule、perl、rootkit |
| r2l | 来自远程机器的非法访问 | tp\_write、guess\_passwd、imap、multihop、phf、spy、warezclient、warezmaster |
| normal | 正常数据 |  |

# 3 数据预处理

## 3.1 增加目标特征

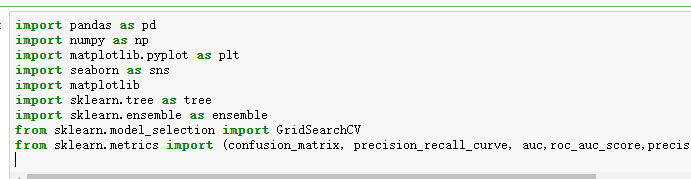
每个文件进行如下操作，并保存文件。为了便于数据处理，训练集和测试集的文件分别放到一个总的表格里面。



**图3.1 增加目标特征**

## 3.2 导入需要用到的库

导入处理数据集涉及的函数库：pandas、numpy、以及可视化的seaborn、还有网格搜索、随机森林等等。



**图3.2 导入函数库**

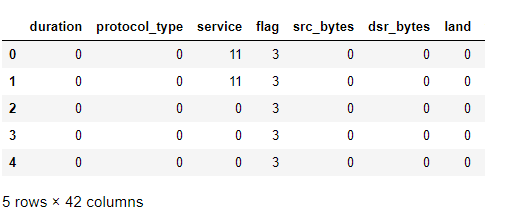
## 3.3 导入数据集

导入数据集train2.csv和test2.csv；并命名为data\_train、data\_test，代码和数据集文件放到同一个目录下面，或者写绝对路径，但是会造成代码的可移植性会很差。



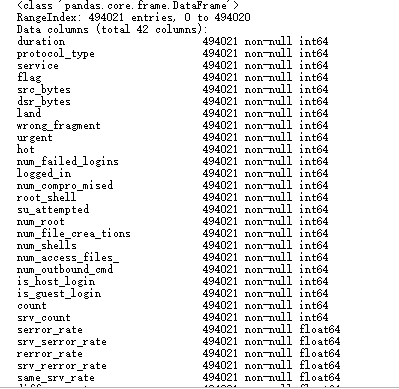
**图3.3 导入数据集**

查看前五行的数据样本（部分）见图3.4：此时数据集有42个特征，其中有很多特征（duration、src\_bytes、land）的值都是0，后续处理很大概率不给予考虑这些特征。



**图3.4 展示前五行数据集**

使用函数：data\_train.info()观察数据集是否有缺失值，观察每个特征的数据。



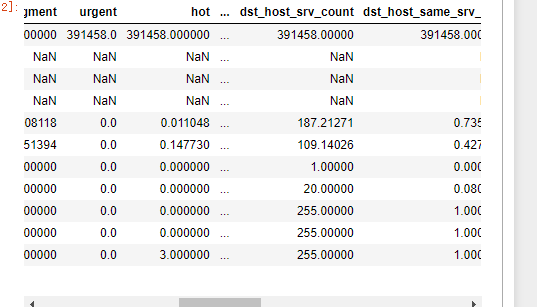
**图3.5 特征样本数据统计**

很清楚地发现训练数据集并无缺失值。测试集也用同样的方式进行处理，没有缺失值。

## 3.4特征的探索性分析

### 3.4.1 取值分布情况

通过统计函数查看个特征取值分布情况：data\_train\_dos.describe(include = "all")

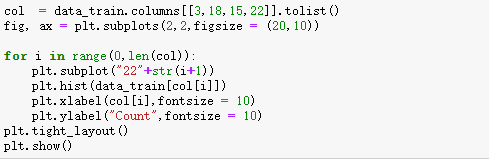


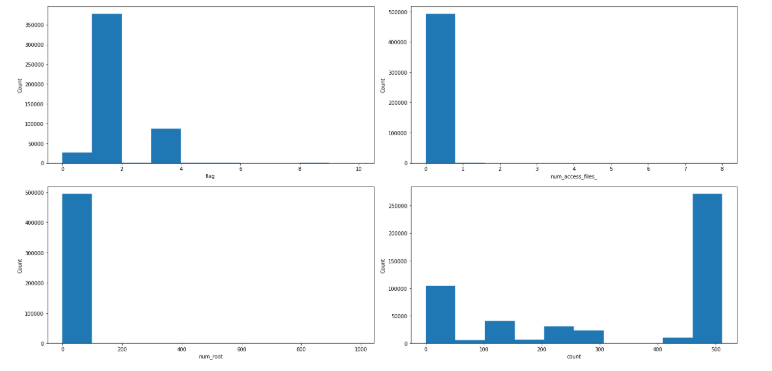
**图3.6 样本数据范围**

观察图3.6发现urgent所有数据都是0，选择特征值的时候就可以不需要考虑urgent。对我们的研究问题影响不大，在已知数据的范围之内，并不存在异常的数值。

### 3.4.2 特征的分布情况

用直方图观察离散型取值特征（ flag、 num\_root、num\_access\_files\_、 srv\_count ）分布情况：col是返回的一个列表，包含了需要观察的四列数据，fig,ax是需要画的一个2\*2坐标系，通过for函数依次把每一个特征的情况描绘出来，并计算每个值的数据总和。

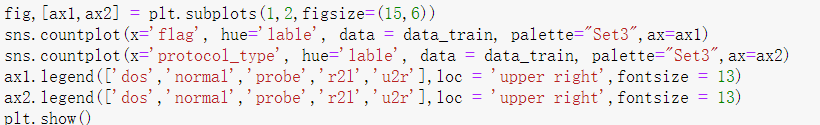
**图3.7** **取值特征分布**



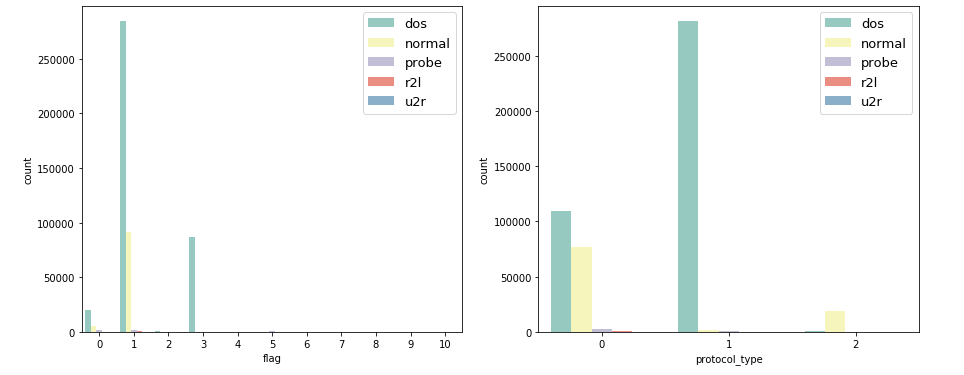
**图3.8 取值特征分布结果**

清楚的看到特征num\_root、num\_access\_files\_的数据几乎都是一致的，换而言之，这些特征对后续研究的影响是很小的，可以不用考虑。

将flag 和protocol\_type进行分组，查看不同分组的特征情况。观察这两个特征中不同目标特征的数据情况



**图3.9 特征分组**



**图3.10 flag、protocol\_type分组**

从分组的情况来看，protocol\_type 当值为icmp时候，有较好的分别，flag的值是1和3的情况，不同目标特征的也有很明显的差别。但是这两个特征对于r2l和u2r的攻击方式不是有很大明显区别。因此，我们可以选择这两个特征作为基础特征

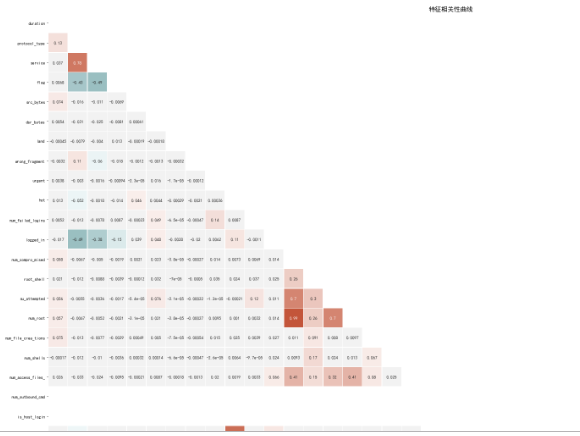
### 3.4.3 特征的选择

将flag和protocol\_type作为基础特征，通过计算特征之间的相关性，来选择其他的特征值。需要用到相关系数函数corr()来计算我们41个特征的相关性。用热力图画出先关系数图，如果热力图出现相关系数为1的情况，说明两个特征成线性相关，只考虑其中的一个特征。



**图3.11 计算相关性**

由于相关性热力图较大，选择其中的一部分展示：热力图纵横坐标分别是41个特征，通过开始设置基础特征，寻找其他相关的特征。

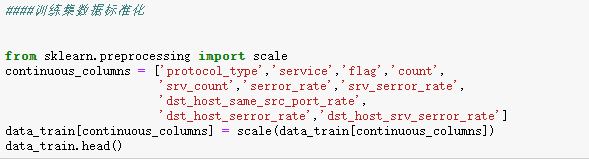


**图3.12 特征相关性热力图**

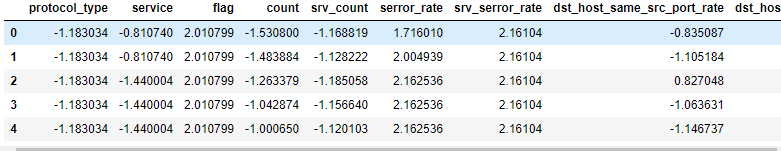
通过观察相关性热力图，将相关系数大于0.7并且不等于1的特征作为优先选择的特征：分别是（protocol\_type、service、flag、 count、 serror\_rate、 srv\_serror\_rate、 dst\_host\_same\_src\_port\_rate、dst\_host\_serror\_rate、 dst\_host\_srv\_serror\_rate）十个特征。

## 3.5数据标准化

在Sklearn中的preprocessing模块，存在scale()函数，用该函数对我们选择的十个特征进行标准化：数据集和测试集分别需要标准化



**图3.13 数据标准化**

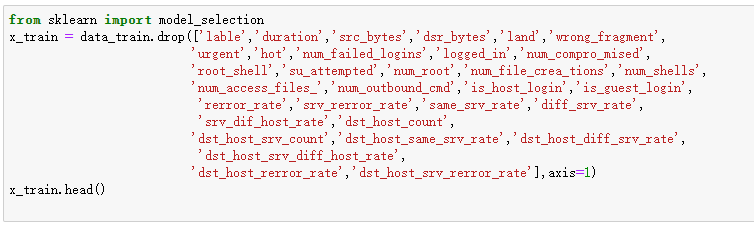


**图3.14 样本标准化后前5行数据**

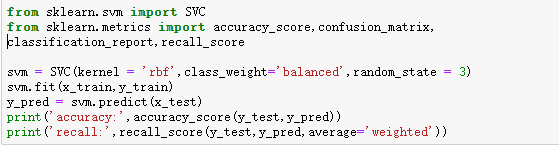
# 4 网络态势感知模型研究

## 4.1分类建模

去掉标签和没有选择的特征，同样也要对测试集数据进行相同的操作：

**图4.1 选择特征**

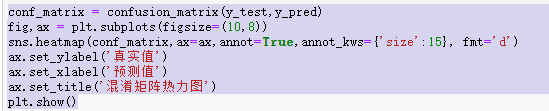
## 4.2支持向量机

支持向量机(SVM)是一种常见的基于监督的二分类模型。SVM已经广泛应用于统计分析，回归分析和机器学习中。SVC函数的参数主要有：kernel(’poly’,’rbf’)、class\_wegiht(类别的权重)、random\_state(数据洗牌时的种子数)；首先用该方法尝试看下准确率和召回率

**图4.2 支持向量机**

**图4.3 支持向量机结果**

运行的结果准确率和召回率都在0.69，效果不是很好，并且也很抽象。用混淆矩阵来可视化分类的情况：



**图4.4 支持向量机可视化**

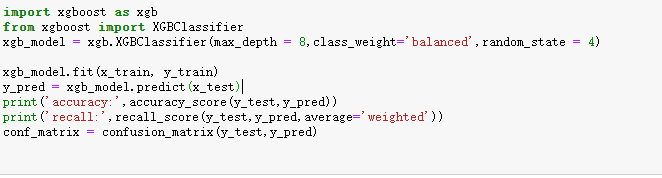


**图4.5 支持向量机混淆矩阵热力图**

很明显svm在解决多分类问题上面的效果不是很理想，dos类别近三分之一都被误分到其他的类别。其他的目标特征分类也很一般。

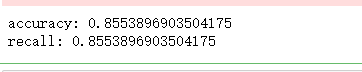
## 4.3 Xgboost算法

Xgboost（极端梯度提升）是Boosting算法的一种，在GBDT的基础上优化，它是由多个相关联的决策树联合决策的，使之更强大，适用于更大范围。改函数使用的参数主要有max\_depth(树的深度)、class\_weight = balanced（平衡输入样本不均衡带来的问题）、random\_state(随机数种子)



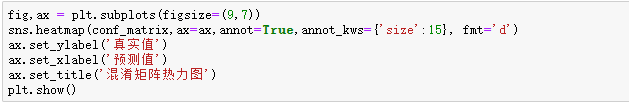
**图4.6 Xgboost**

Xgboost算法的准确率和召回率都达到了0.85,相对于svc而言，有很明显的提升；



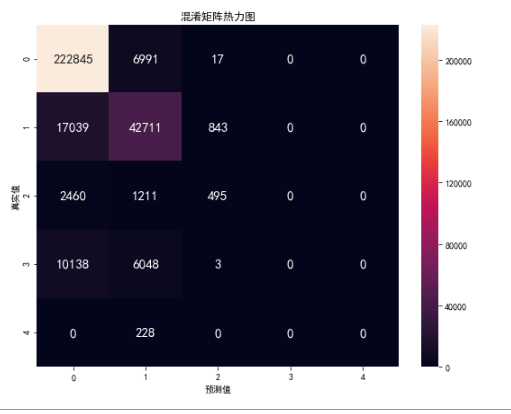
**图4.7 Xgboot结果**

用混淆矩阵热力图表示：



**图4.8 Xgboot可视化**

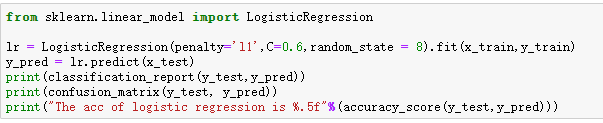
从图4.9可以看出：dos的分类效果还是不错的，23万条数据，只有7000被误认为正常数据，相对于支持向量机而言，分类效果还是可以接受的。正常数据也有近三分之一被当做dos攻击，对于其他的分类情况则是很一般。



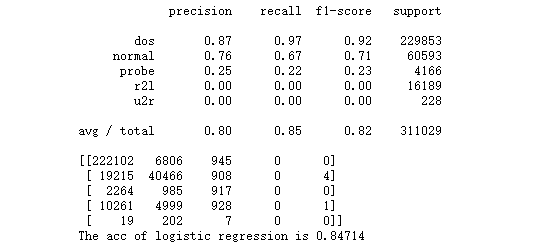
**图4.9 Xgboost混淆矩阵热力图**

## 4.4 逻辑回归

逻辑回归是用于解决二（多）分类问题的机器学习方法，这里我们只设置其中的三个参数：penalty(参数有l1和l2，使得使用的模型的泛化能力更强)、c(正则化系数λ的倒数)、random\_state(随机种子数)



**图4.10 逻辑回归**

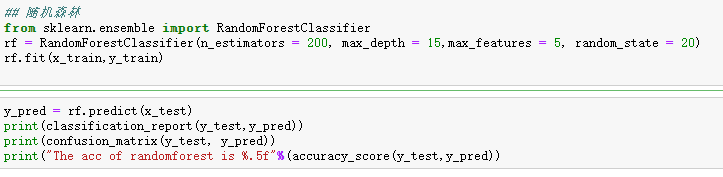


**图4.11 逻辑回归结果**

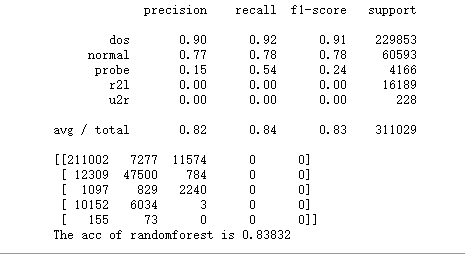
逻辑回归在对dos分类的recall达到了0.97，f1值也达到了0.92,分类的效果可以接受。正常数据数据和xgboost分类效果差不多，不是很理想。

## 4.5 随机森林

随机森林是机器学习中的一种常用的方法。其中使用到的参数有：n\_estimators（森林里面的决策树的数目）、max\_depth(最大深度)、max\_features(特征数目)、random\_state(随机种子数)



**图4.12 随机森林**



**图4.13 随机森林结果**

观察dos和normal的f1值分别是0.91和0.78。相对于Xgboost算法效果会差点，可能也因为参数设置的原因导致的。后续会针对这个参数利用gridSearchCV（网格搜索）方法进行优化

## 4.6 分类算法比较

**表3分类算法性能分析表**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision（Accuracy） | recall | f1 |
| SVM | 0.69 | 0.69 | 0.69 |
| Xgboost | 0.85 | 0.85 | 0.85 |
| 逻辑回归 | 0.87 | 0.97 | 0.92 |
| 随机森林 | 0.90 | 0.92 | 0.91 |

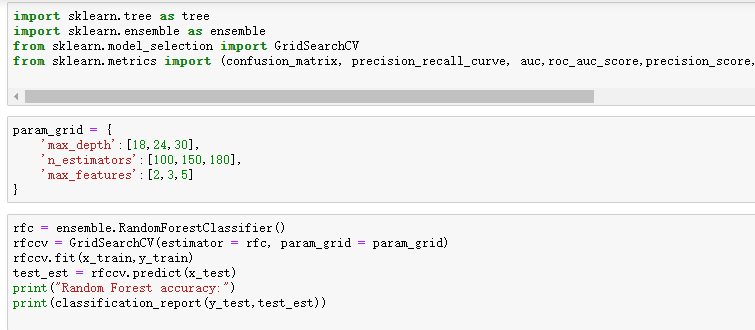
表格的数据是基于分类dos的效果作为参考，很清楚地看到SVM分类效果不是很理想，逻辑回归的f1值最高，其次是随机森林。

# 5模型的参数优化

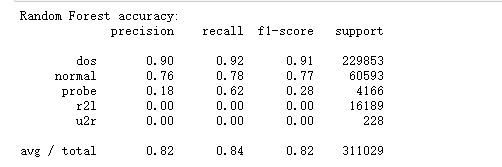
通过前面几种分类的方法比较，选择分类效果比较好的分类算法。下面利用网格搜索和交叉验证寻找最佳的参数

## 5.1 随机森林参数优化

导入相关函数，创建一个字典param\_grid保存传入的参数；函数GridSearchCV的参数：estimator（分类器）、param\_grid(字典值)



**图5.1 随机森林参数优化**

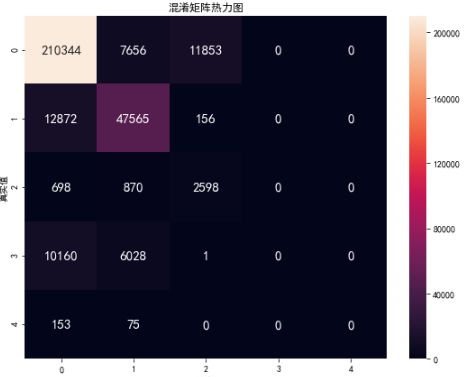


**图5.2随机森林参数优化结果**

经过遍历计算之后，树最大深度为为18，最大特征为3，决策树的个数为180的情况分类效果是可以接受的。

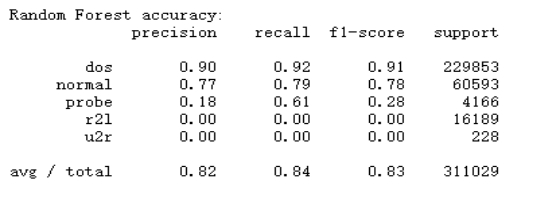


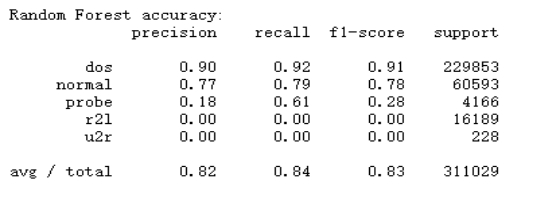
**图5.3 最佳参数**



**图5.4 随机森林混淆矩阵热力图**

网格搜索特别耗时，通过网格搜索和交叉验证多次进行操作，发现都会选择决策树个数都会选择偏大的数值，对于树的最大深度，则是选择数值偏小的。参数max\_depth 的范围大概是14-16，max\_features = 3，决策树的个数到500-800。



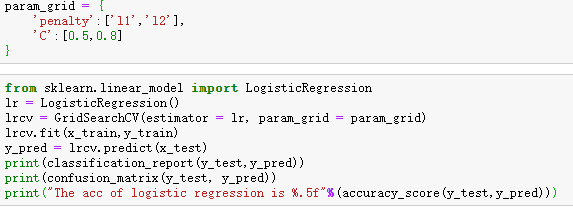


**图5.5优化随机森林结果**

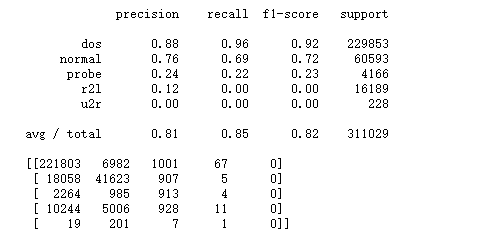
对于u2r的分类效果很差的原因，主要的原因是数据较少，然而r2l分类的效果较差原因是因为当初选择基础参数导致。

## 5.2 逻辑回归优化

优化逻辑回归时，参数Penalty值选择l1和l2，正则化系数的倒数C设置为0.5与0.8



**图5.6 逻辑回归优化**



**图5.7 逻辑回归优化结果**



**图5.8 逻辑回归最佳参数**

多次的参数调整，发现搜索参数c选择较小的值，惩罚项则是l2.最后当参数为c==0.4,penalty = ‘l2’的效果较好。

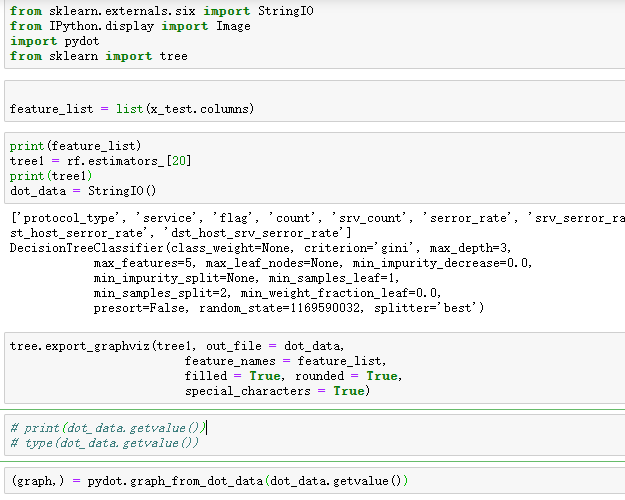
# 6随机森林模型优化

通过前面一系列的尝试之后，现在只是知道设计模型是好的，但是这个模型就只是一个黑盒子，所做的任务也仅仅只是提供了一些训练集让他进行训练，让它根据提供的测试集进行预测，并且评预测的一个结果，并且也看了这个模型是合理的。那么问题是：这个模型如何达到他的价值？下面就从两个方面来探讨：

## 6.1 可视化单个决策树

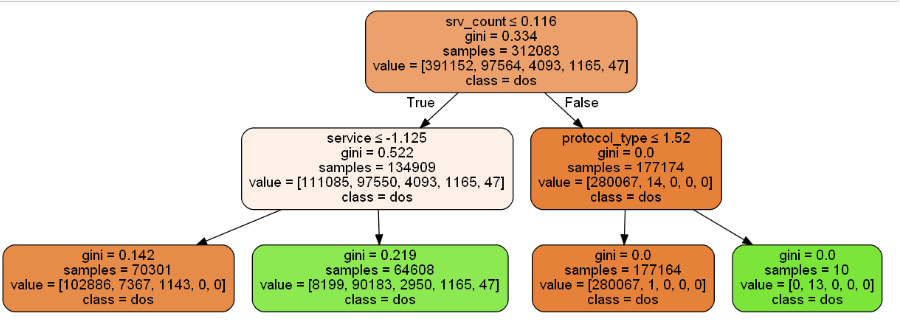
基于前面实验的数据，从随机森林中选择其中的一棵树，并且将这个树保存稳图像：

导入相关的库函数。函数export\_graphviz()的参数：tree 是我们将要可视化的一棵决策树，out\_file（保存的文件名字）、feature\_names（我们选择的特征名称）

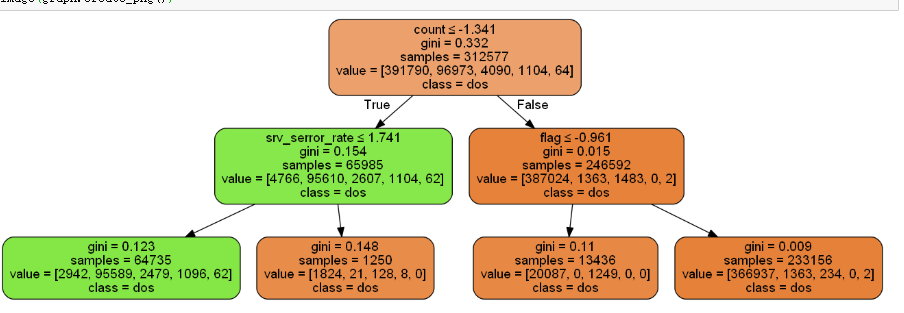


**图6.1 可视单棵化决策树**

为了能够清楚地看到输出图像，将树的深度设置为2：展示其中的两棵树



**图6.2 第二十一棵树**



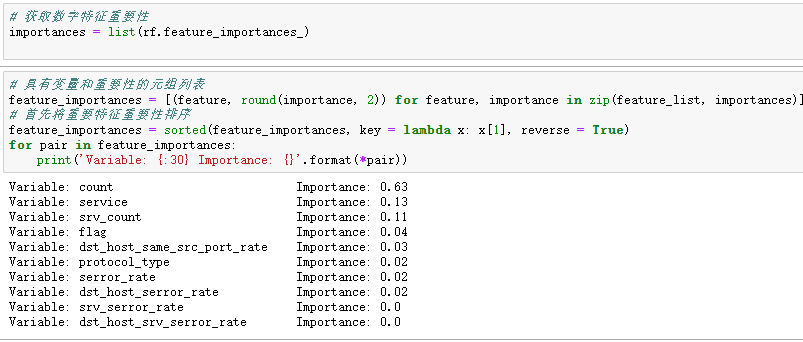
**图6.3 第三课树**

观察两棵树的根节点的特征，根节点是从所有特征中gini值最大的。基于其中的第21棵树就可以做一些预测：从测试集里面选择第四条数据，srv\_count = -0.976，service = -0.890，protocol\_type = -1.137，srv\_serror\_rate= -0.252从根节点开始，因为srv\_count = -0.976<0.116,值是true，左移，判断特征srv\_serror\_rate >-1.125 右移，最终判断这个数据被分类到dos的类里面。此时gini是0.219。

样本数据有49万条，但是我们根节点的样本数据大概只有31万条左右，这也是因为随进森林的每一棵树都是用稍微不同的训练集训练的，最终通过成百上千棵决策树一起决策。通过对树的可视化，清晰地发现对于单棵树只有一部分特征影响着决策树判断，那么找到这些特征就会大大提高模型的识别效率。

## 6.2 核心特征

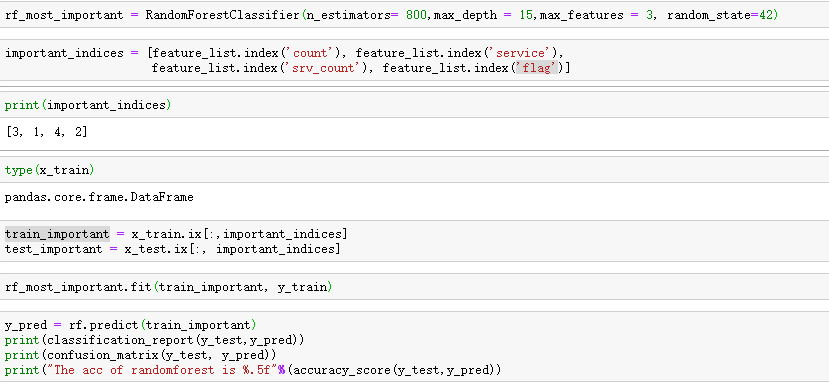
为了量化整个随机森林中所有变量的有用性，找到需要的核心特征，可以通过查看变量的相对重要性。使用这些数字进行变量之间的相对比较。从而进行优化设计的模型。



**图6.4 特征相对重要性**

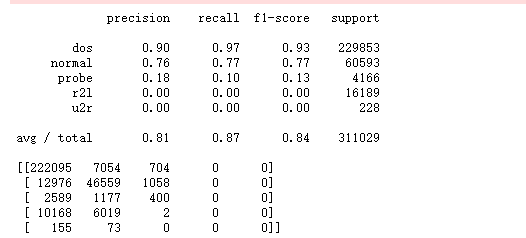
从运行的结果可以看到，在列表第一的是特征count、其次是service、srv\_count、flag。在接下来的训练过程中删除其他的特征，把这四个特征作为未来模型的核心特征。这样就大大减少模型的处理效率。

训练模型的参数选择是的决策树的个数是800，最大深度是15，max\_feature = 3,这些数据也是通过优化，得到较好的参数。从以前的训练集和测试集选择需要的四个特征的所有数据集分别保存在train\_important 和test\_important 。



**图6.5 降维后的随机森林**

运行后的结果见图6.6，发现dos的分类效果又提升了两个百分点，对于正常数据的分类效果也可以接受，整体的效果满足一开始的期望。



**图6.6 随机森林结果**

从上面的结果可以看出，针对dos攻击，并不需要收集其他更多的信息，只需要收集核心特征的四个变量，这样一来减少了收集数据的样本的时间成本。通过优化后的模型就可以很大概率检测到某次网络连接记录是否是dos攻击。

## 6.3 参考模型

**表4 模型优化前后分析**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | Precision | recall | F1-score | 特征个数 |
| 原始 | 0.90 | 0.92 | 0.91 | 10 |
| 优化 | 0.90 | 0.97 | 0.93 | 4 |

从表4很清楚地观察到，优化后的模型特征个数减少了一半，间接地使得模型在实际运用的过程中，提高了分类效率。

# 7结束语

本文是根据kdd99数据集，结合机器学习的多种分类算法，根据分类的效果，选择较为合适的随机森林分类算法，然后逐步对改算法参数的优化，找到合适的参数。通过对单个决策树的可视化，发现选择的十个基本特征中，存在部分特征对决策树并没有很大的影响，通过变量之间的相对比较，最后寻找到随机森林的四个核心特征，让随机深林模型的分类效果又提升了很多，并且也大大减少来了收集和处理数据的时间，然而由于数据集的特征比较多，最后设计的模型，对其他3种攻击方式的分类效果较差，但是模型对于dos、normal、攻击的分类效果较为良好，最终的模型也是可以接受的。

# 参考文献

1. 常镒恒,马照瑞,李霞,巩道福.网络安全态势感知综述[J].网络空间安全,2019,10(12):88-93.
2. 龚正虎,卓莹.网络态势感知研究[J].软件学报,2010,21(07):1605-1619.
3. 叶欣. 基于多分类安全事件的网络安全风险评估技术研究[D].天津理工大学,2019.
4. 朱博文. 基于大数据的网络安全态势感知模型研究[D].华侨大学,2018.
5. 王卫华.网络安全态势评估模型研究[J].青岛远洋船员职业学院学报,2017,38(01):28-31.
6. 管磊,胡光俊,王专.基于大数据技术的网络安全态势感知平台研究[J].保密科学技术,2016(05):13-19.
7. 徐振华.网络安全态势感知模型设计和实现[J].电子测试,2016(03):23-25.
8. Yuanquan Shi, Tao Li, Renfa Li, et al. An Immunity-Based IOT Environment Security Situation Awareness Model. 2017, 5(7):182-197.
9. GUO Fangfang,HU Yibing,XIU Longting,FENG Guangsheng,WANG Shuaishuai.A Hierarchical P2P Model and a Data Fusion Method for Network Security Situation Awareness System[J].Wuhan University Journal of Natural Sciences,2016,21(02):126-132.

[10] Abasi. The Analysis on the Network Security Situation Assessment and Situational Awareness. 2014, 2987:2752-2755

[11] 冯波. 网络安全态势评估模型研究[D].电子科技大学,2014.

[12] Liu, Xiaowu, Wang, Huiqiang, Yu, Jiguo, et al. Network Security Situation Awareness Model Based on Multi-Source Fusion. 2012, 5(2):775-779(5).

[13] Yixiang Luo. Network Security Situation Awareness Model-Inspired by Immune[C]. Intelligent Information Technology Application Association.Manufacturing Systems and Industry Application（ICMEAT 2011 MSIA）.Intelligent Information Technology Application Association:智能信息技术应用学会,2011:648-651.

[14] 鲁智敏. 基于态势感知的网络安全体系部署与价值分析[C]. 公安部第三研究所、江苏省公安厅、无锡市公安局.2019中国网络安全等级保护和关键信息基础设施保护大会论文集.公安部第三研究所、江苏省公安厅、无锡市公安局:《信息网络安全》北京编辑部,2019:112-115.

[15]李全华.基于安全态势感知在网络攻击防御中的具体运用[J].信息通信,2018(04):191-192.

[16]贾焰, 韩伟红, 杨行. 网络安全态势感知研究现状与发展趋势[J]. 广州大学学报(自然科学版), 2019(3).

[17]卓莹. 基于拓扑流量挖掘的网络态势感知模型[C]. 中国指挥与控制学会.第三届中国指挥控制大会论文集（上册）.中国指挥与控制学会:中国指挥与控制学会,2015:261-266.

[18]钱真坤.网络安全态势感知框架及随机森林评估模型[J].西南师范大学学报(自然科学版),2019,44(11):118-123.

[19]龚俭,臧小东,苏琪,胡晓艳,徐杰.网络安全态势感知综述[J].软件学报,2017,28(04):1010-1026.

[20]胡志军.基于大数据的网络安全态势感知平台的应用思考[J].金融科技时代,2019(10):44-46.

[21]毛军礼,汲锡林. 基于大数据的网络态势感知体系架构[ J] . 无线电通信技术,2018,44(3) :217-223.

[22]张媛.网络安全态势感知防御技术[J].信息技术与信息化,2019(08):121-123.

# 致谢

本研究及学位论文是在我的导师付海燕老师的亲切关怀和悉心指导下完成的。他严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。付海艳老师不仅在学业上给我以精心指导，同时还在思想、生活上给我以无微不至的关怀，在此谨向付老师致以诚挚的谢意和崇高的敬意。我还要感谢在一起愉快的度过毕业论文小组的同学们，正是由于你们的帮助和支持，我才能克服一个一个的困难和疑惑，直至本文的顺利完成。

在论文即将完成之际，我的心情无法平静，从开始进入课题到论文的顺利完成，有多少可敬的师长、同学、朋友给了我无言的帮助，在这里请接受我诚挚的谢意!最后我还要感谢培养我长大含辛茹苦的父母，谢谢你们!

最后，再次对关心、帮助我的老师和同学表示衷心地感谢!