**摘 要**

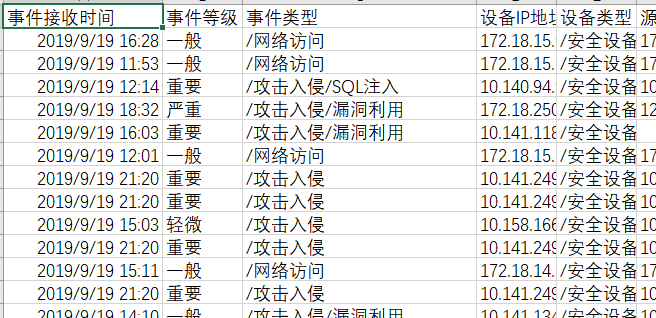
此次实验内容为对给定的数据集进行分类。分类的整体流程为数据预处理、特征选择、建立模型、训练、模型评价。实验过程中分别使用朴素贝叶斯、逻辑回归、SVM（LinearSVC）、神经网络（MLPClassifier）四种方法进行分类，评估各自的分类效果，对各方法的优缺点进行比较分析。最后对实验过程和结果进行总结分析，并提出不足与可改进之处。

**关键词：**模式识别；朴素贝叶斯；SVM;逻辑回归；神经网络

**第1章 实验环境与数据集**

1.1 实验环境、数据集获取与分析

实验环境为：Windows 10 RAM:8GB CPU:I5-7300。Python 3.7 Jupyter Notebook。使用给定的event.csv做数据集，编码格式为 gb18030.通过分析可知为分类问题，标签为 “事件类型”列。其余列为特征。



**第2章 数据预处理**

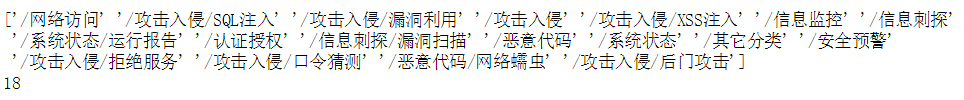
2.1 加载数据，查看总体情况

可知数据共 99809（行）\*66（列），包括标签列，也就是99809个样本，65个特征，1个标签。其中各列数据的数量，均值，最大最小值等见下图。

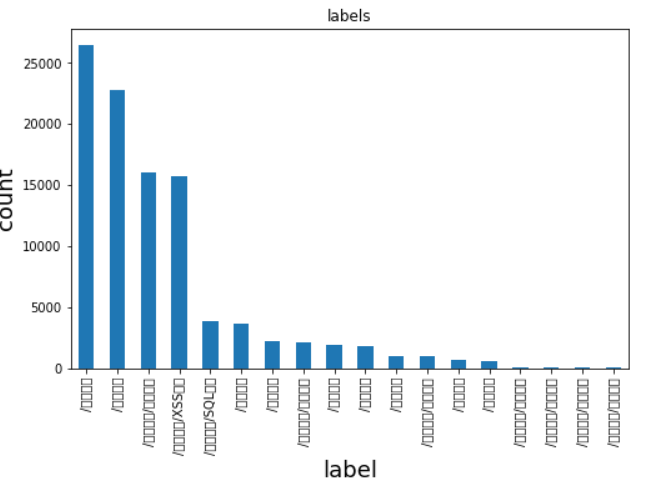
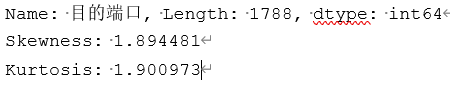


2.2 统计分析

查看label情况：label数量，共18种



Label分布。分布如图，图中因中文编码出现乱码。可见各类分布并不均匀。

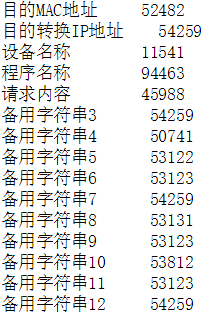
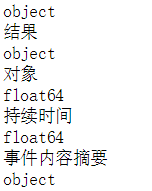
 

对于各列元素出现次数，偏斜度，峰度分析：

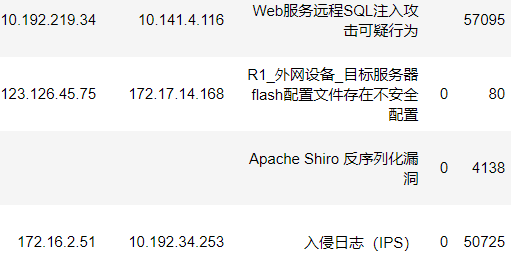
以目的端口为例：偏度1.894481，大于0，说明端口数据相比正态分布右偏，峰度1.900973>0，说明端口比正态分布要陡峭。

2.3 缺失数据分析与处理（填充）

统计哪些列存在缺失情况，分别缺失多少数据，以及数据类型。可见存在大量缺失，且缺失数据类型包括object，float两种。对象和持续时间完全缺失。

经过考虑，删除完全缺失的对象，持续时间列，其他列对缺失数据进行填充处理。填充方法为对object类型列填充字符’0’，其他列填充数据0.依据为：此数据集并不适合并不适合采用均值填充和上下文填充方法，因为可能导致分类的重要特征被破坏。填充结果如图：图中没有显示的部分经检验均为字符’’.



2.4 数据编码

经过填充后的数据集仍然含有很多字符，不利于模型的训练。需编码为数字。编码方法有：onehot，词向量模型，TF-IDF,顺序编码等。由于数据99809\*66列，内存可用8gb，若采用直接onehot编码，考虑内存占用需要保证列数控制在8gb/8byte/99809 约等于160\*66-66列，但原数据过于复杂，最后列数会超过上述上限，超出内存限制。若采用词袋法，进行词向量模型或TF-IDF编码需要先分词，filter过滤等工作，操作上比较复杂，这里选择顺序编码（sklearn的factorize方法）。编码后的数据如图：



2.5 数据标准化

常用的数据标准化方法有：缩放到0-1之间，正则化，先计算均值方差然后Zscore标准化等等。本次实验分别采用这些标准化方法，然而均导致最终的模型分类效果很差（比如采用Z分数标准化后使用逻辑回归模型的准确率召回率均低于10%），故最终不采用数据标准化方法。

**第3章 特征选择与特征抽取(PCA)**

3.1 特征选择方法概述

首先排除低方差的特征，然后对剩余特征进行pearson相关系数计算，选择绝对值较高的特征。同时，由于pearson相关系数只对于线性关系敏感，可能具有误导性，故在排除低方差特征后也进行卡方检验，计算特征和label的相关性，选择最优的几个特征，然后与根据相关系数选择出的特征进行对比。

除数学计算外，也根据人为经验选择一些对分类重要影响的特征。

对上述得到的三组特征分别进行模型训练，根据预测效果评分，选择合适的特征。

最后进行特征的组合，筛选（单独增减某些特征，对比预测效果）。

此过程对于下面的四种模型（SVM,朴素贝叶斯，逻辑回归，神经网络）分别独立，也就是进行四轮，每轮中从原始的三组特征开始不断调整，得到多组特征，对比预测效果，从而针对每个模型得到最合适的特征。

3.2 通过Pearson相关系数选择特征

由于特征已经编码，缺失数据也已经填充，计算过程比较轻松，值得注意的点是需要全部转换为float类型，否则会导致计算出错。结果如下：

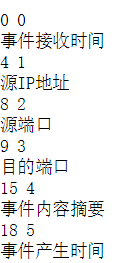
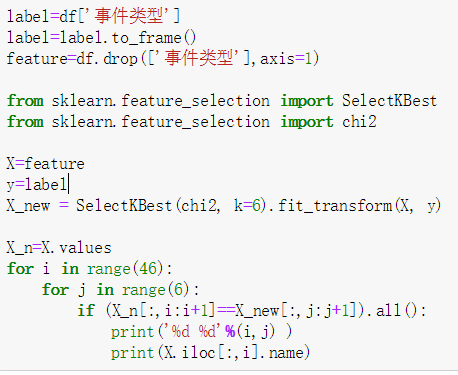


选择事件类型列，排序，选择绝对值在0.35以上的特征：目的端口，备用字符串1，操作，备用字符串2，事件名称，事件原始等级，事件原始类型。

3.3 通过卡方检验选择特征

将特征和label分离，分别计算特征和label两两间相关性，选择最优的特征。代码如下，其中K为选择的特征数量。最后选择了：事件接受时间，源IP地址，源端口，目的端口，事件内容摘要，事件产生时间。



3.4 通过经验选择特征

通过以往对网络安全知识的学习，选择：网络协议，网络应用协议，目的端口，事件内容摘要，操作。

3.5 特征抽取：PCA降维

PCA的作用是降低数据维度，压缩数据降低信息损失，减少计算力和内存消耗。每次得到一组特征都会进行PCA降维，下面给出一个例子。

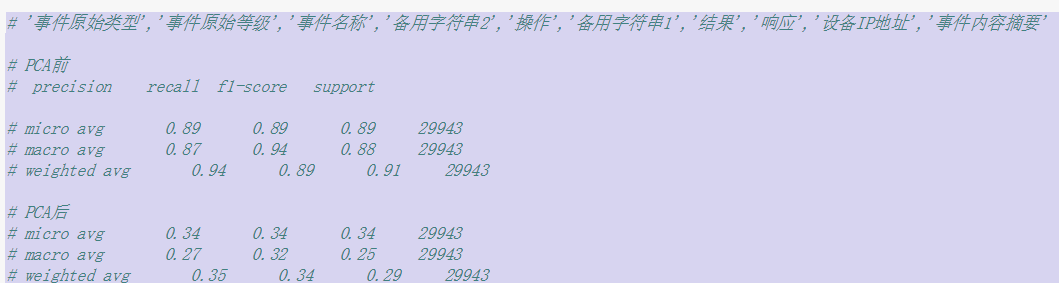
将原有13个特征根据降维后方差值占总方差值的比例，选择3个主成分。分别占比 99.9%，0.006%，0.002%。



[9.9990e-01 6.3469e-05 2.7110e-05]

[6.8925e+06 4.3750e+02 1.8688e+02]

采用朴素贝叶斯模型对比PCA降维前后的预测效果，可见PCA降维对模型预测准确性影响比较大，故后续实验未采用PCA降维。



**第4章 模型建立与训练，评估**

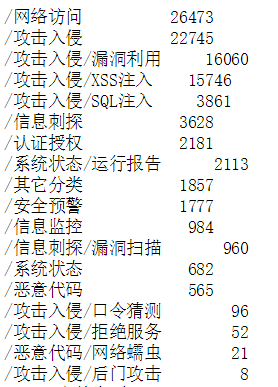
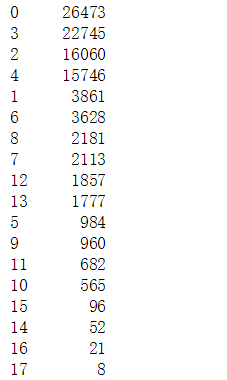
4.1 概述

使用sklearn提供的模型接口快速方便的对四种方法（朴素贝叶斯，SVM,逻辑回归，神经网络）建模，训练。由于篇幅所限，这里直接展示四种方法各自最优特征进行训练的结果。将数据集7：3分为训练集测试集。

模型评估方法：根据测试集预测效果，按照各类权重，计算整体准确率，召回率。同时查看18类各自的分类状况（准确率，召回率，准确数量，混淆矩阵）。

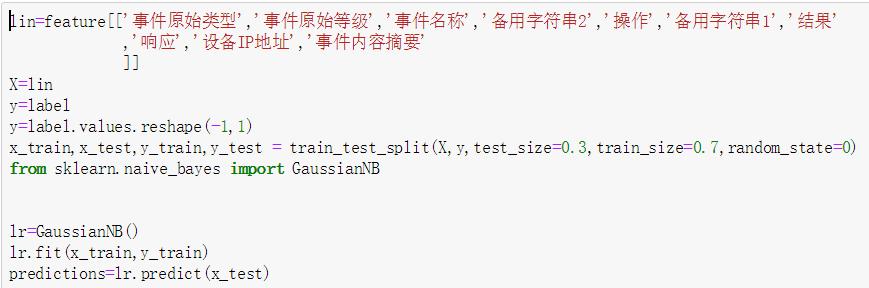
模型评估函数如下：

其中事件类型编码与原名称对应关系如下。

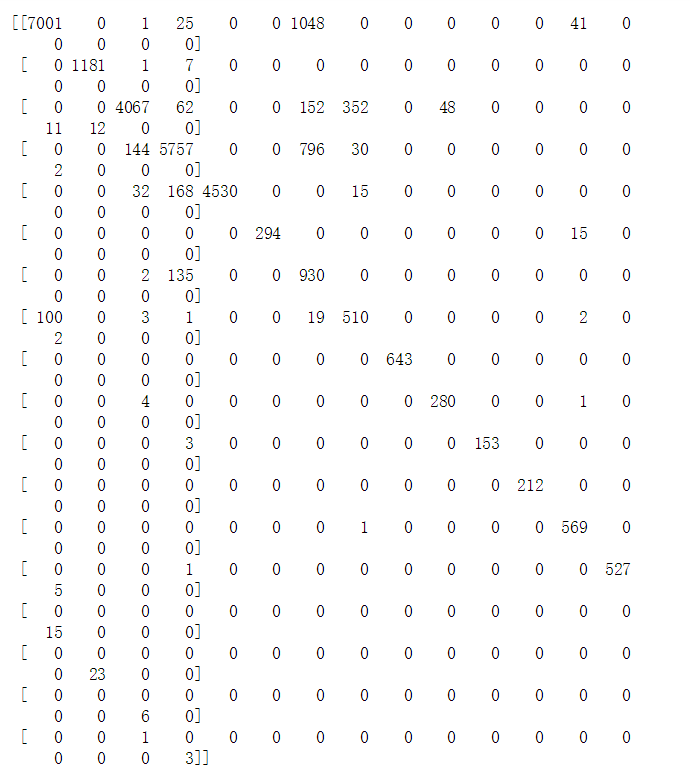
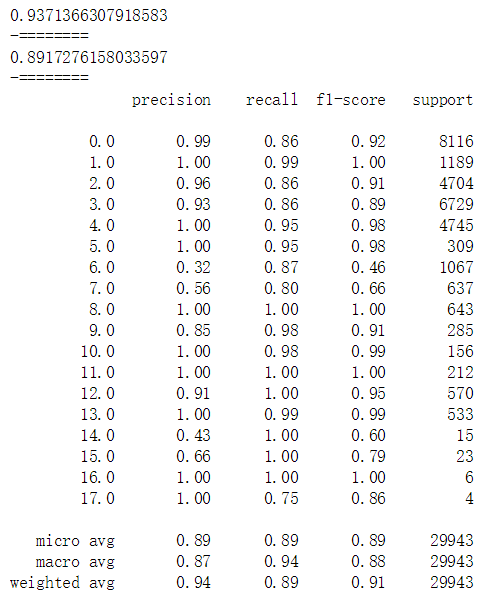
 

4.2 朴素贝叶斯

建模与训练：选择10个特征，如图：



评估：整体准确率93.7%，召回率89.1%。对于第6，14，15类（信息刺探，攻击入侵/拒绝服务，攻击入侵/口令猜测）分类效果比较差。对信息刺探识别准确率最低，仅32%，对攻击入侵/后门刺探召回率最低，75%。信息刺探共1067个测试样本，准确率仅32%，影响较大。总体而言分类效果比较一般。

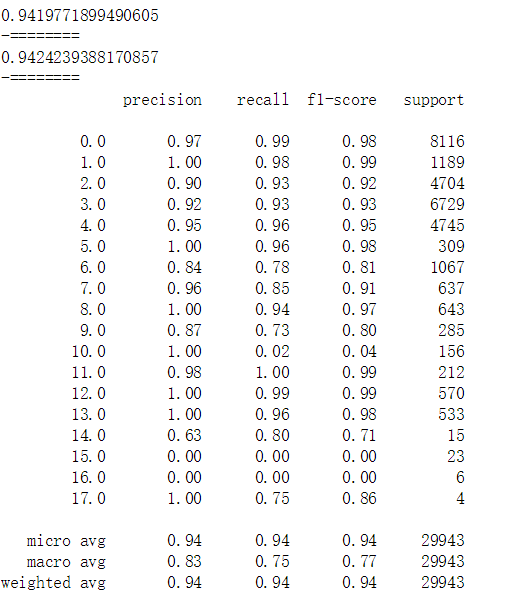
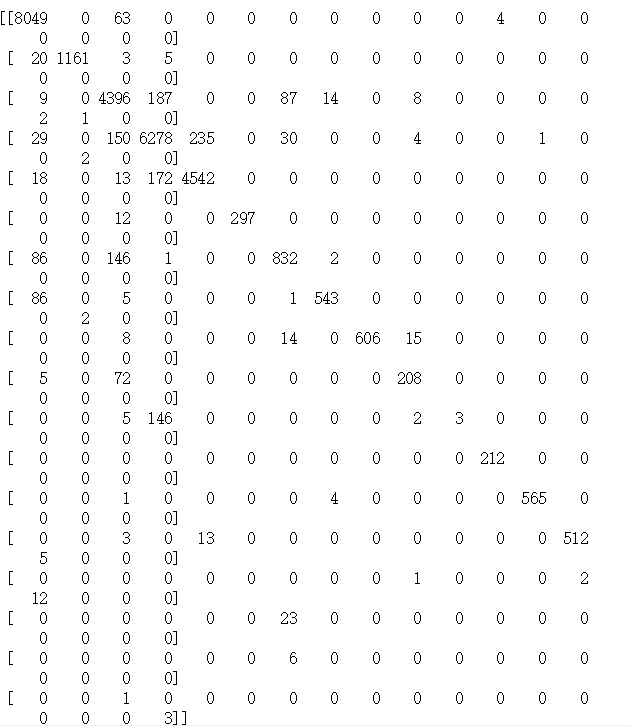


4.3 逻辑回归

建模与训练：采用13种特征，如下。penalty正则化损失函数为l2，减少过拟合，正则化系数C为1，逻辑回归损失函数优化方法为liblinear（采用坐标轴下降法迭代优化损失函数），class\_weight=None（各类权重相同，防止对于数量小的类识别效果差），最大迭次数iter为100，迭代终止误差范围tol为0.0001.

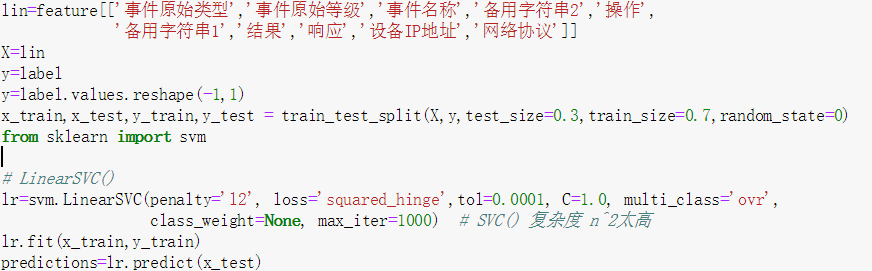


评估：整体准确率94.1%，召回率94.2%。对第10，14，15，16类（恶意代码，攻击入侵/拒绝服务，攻击入侵/口令猜测，恶意代码/网络蠕虫）识别较差。对攻击入侵/口令猜测，恶意代码/网络蠕虫识别准确率召回率均为0，对恶意代码识别召回率2%，很差。总体而言识别效果比朴素贝叶斯好，但是对于个别类（样本数量小的14，15，16）识别效果很差。

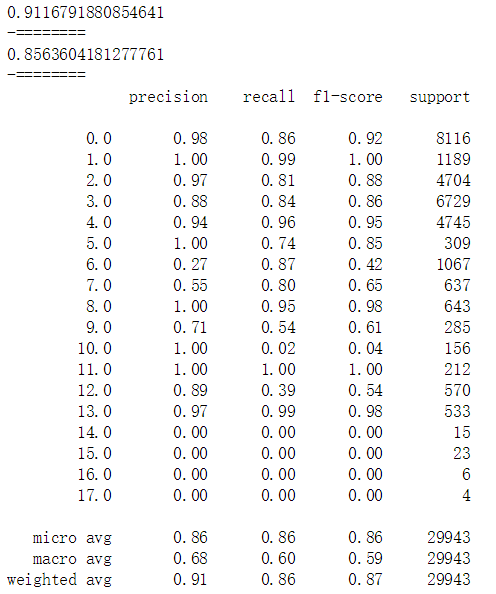
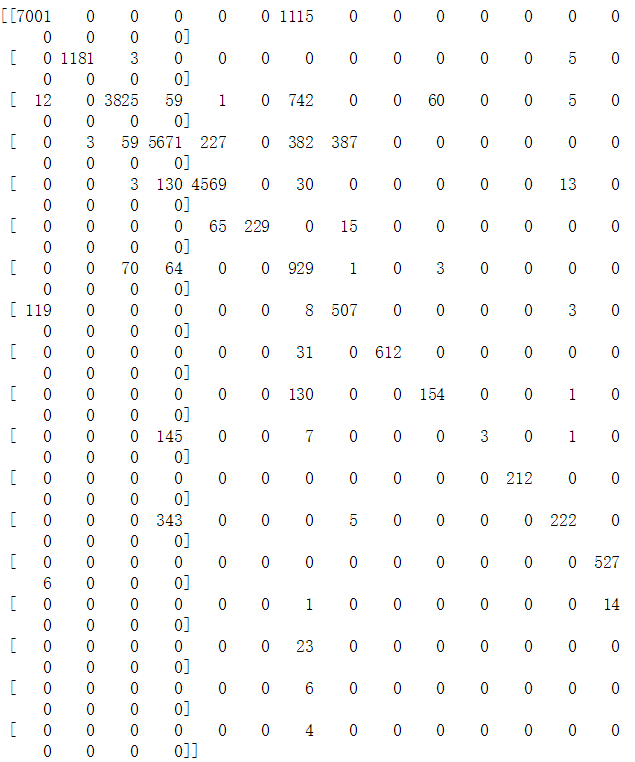
 

4.4 SVM

建模与训练：选择10个特征，如下。设置penalty（正则化损失函数）为l2，减少过拟合，正则化系数C为1，损失函数为squared\_hinge，最大迭次数max\_iter为1000，设置迭代终止误差范围tol为0.0001，lass\_weight=None（各类权重相同，防止对于数量小的类识别效果差），multi\_class为ovr（将分类中某一类当做正类，其他类做负类，求得每一类为正类时的准确率，取最高准确率的类为正类）。这里采用线性SVC，原因是普通的SVC复杂度为n^2,效率低。

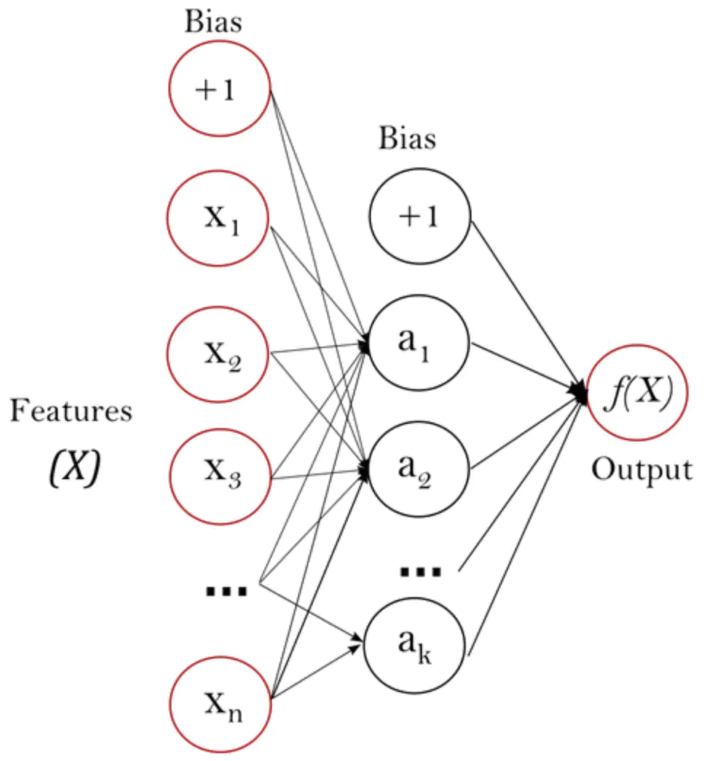


评估：整体预测准确率91.1%，召回率85.6%.对14，15，16，17类识别准确率均为0%，这是因为这四类样本数量都很少。分类效果比较差，不如逻辑回归和朴素贝叶斯。

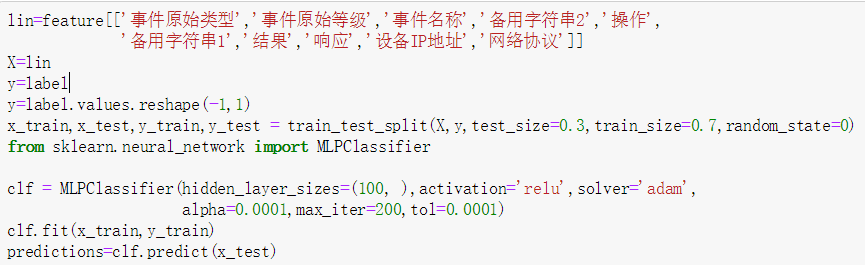
 

4.5 神经网络

建模与训练：采用sklearn提供的MLPClassifier监督学习算法。只使用一个隐藏层（共三层神经网络，如下图）。

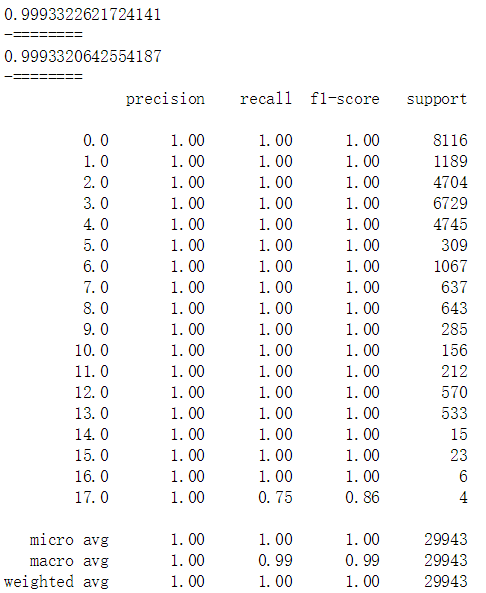
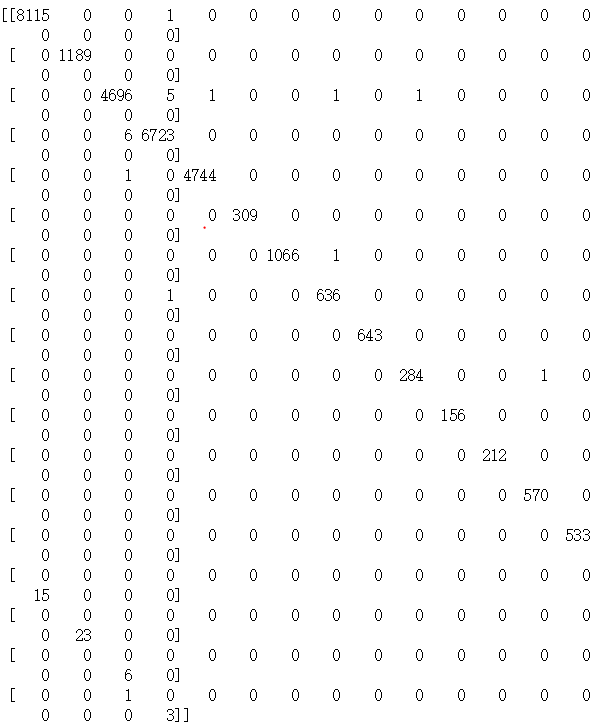


建模与训练代码如下：使用10个特征。设定hidden\_layer\_size=(100,)，也就是一个100个神经元的隐藏层，activate=’relu’（激活函数f(x)=max(x,0)），，lslover=’adam’(基于随机梯度的优化器，且适应几万的样本数量)，正则参数alpha为0.0001，最大迭代次数200，迭代终止误差范围tol为0.0001



评估：整体识别准确率99.9%，召回率99.9%。对0-16类识别准确率召回率都是100%，对只有4个样本的17类（攻击入侵/后门攻击）识别的准确率也有100%，召回率75%，仅错分类1个。

整体而言分类效果非常理想，远优于SVM,逻辑回归，朴素贝叶斯分类器。

**第5章 总结**

5.1 结果分析

从最后的识别效果来看，神经网络分类器是效果最好的，准确率达到99.9%，对于权重低的小样本的分类效果也很好，而朴素贝叶斯，逻辑回归，LinearSVC虽然整体识别效果都不错，准确率在90%-94%，但是对于个别类，尤其是样本数目少的类分类效果普遍比较差。

SVM分类器的效果比较差是出乎意料的，原因可能是特征选择的问题或者参数的问题，也可能和标签与样本的线性/非线性关系有关，有很大的改进空间。

神经网络分类器的分类效果很好则在预料之中，因为它具有出色的学习能力。

朴素贝叶斯为生成模型，逻辑回归为判别模型，朴素贝叶斯适应小数据集，逻辑回归在大数据集表现更好。但对于权重小的类别分类效果都一般。

5.2 可改进之处

有许多可改进之处，比如：

在编码时采用的顺序编码其实有许多问题，尤其是对字符串。可以先采用filter过滤掉无用的信息，字符，然后分词，然后采用onehot/词向量模型编码。

在特征选择时发现不同特征的顺序可能会对最后的预测效果产生影响，可以采用二进制选择，然后遍历排序，分别评估，最后确定最好的特征选择。

在模型训练评估时可以采用k折交叉验证，更有说服力。

文章中有很多数据，若转化成折线图，散点图等图像，会有更好的展示和分析效果。

**参考文献**

[1] Ng A Y, Jordan M I. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes[C]. neural information processing systems, 2001: 841-848.

[2] (美)Pang-Ning Tan等著.数据挖掘导论. 范明等译. 北京:人民邮电出版社,2011.

[3] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2017. Feature Selection: A Data Perspective. ACM Comput. Surv. 50, 6, Article 94 (December 2017), 45 pages.

[4] scikit-learn中文文档https://sklearn.apachecn.org/#/docs/14