数据挖掘互评作业4：离群点分析与异常检测

姓名：曹健 学号：3120190978

1. **问题**

本次作业将从Anomaly Detection Meta-Analysis Benchmarks提供的benchmark数据集中任选两个进行分析。

可以使用[Python Outlier Detection (PyOD)](https://github.com/yzhao062/pyod)或其他已知的工具包来完成分析工作。

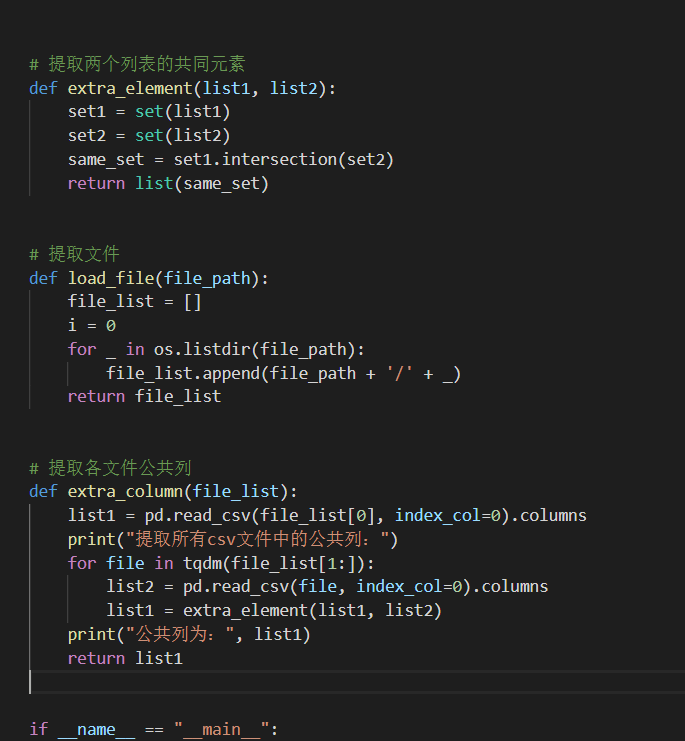
1. **数据**
2. 数据集选择：wine\_benchmarks数据集、abalone\_benchmarks数据集
3. 数据集描述：

两个数据集都分为两部分数据，分别为meta\_data文件夹中的原始整体数据以及benchmarks文件夹中进行划分后的多个子数据集，子数据集之间相互独立。

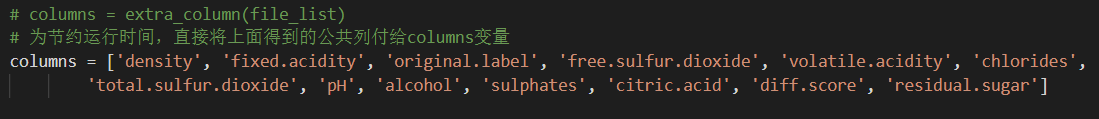
**三、数据集预处理（以数据集wine\_benchmarks举例）**

**1. 找出公共列**

观察发现每个数据集中的各csv文件大都具有不同的列，为了更有效地进行数据挖掘，我们提取所有文件中共有的列名。



提取结果即公共列为：



**2. 数据划分**

将数据集按照7:3的比例划分为训练集与测试集，在训练集上对各种检测离群点的方法进行训练，再在测试集上查看各种方法的效果，并进行对比。

**3. 算法选择**

实验中通过PyOD工具，使用CBLOF、KNN、PCA、IForest四种算法来检测异常，并计算每种算法在各个csv文件中的ROC值和PRN值。

**CBLOF算法**

CBLOF算法是基于聚类和邻近性的异常检测方法，它将数据分为香型集群和大型集群，然后根据点所属的簇的大小以及到最近的大簇的距离来计算异常分数。

**KNN算法**

KNN算法是典型的基于邻近性的异常检测算法，对于任何数据点，到第k个最近邻居的距离可以被视为远离分数，PyOD支持三个KNN探测器，分别为最大KNN、均值KNN以及中位数KNN，区别在于作为离群值得分的距离的计算方法。在这里我们用默认的最大KNN算法，也就是使用最大距离作为离群值。

**PCA算法**

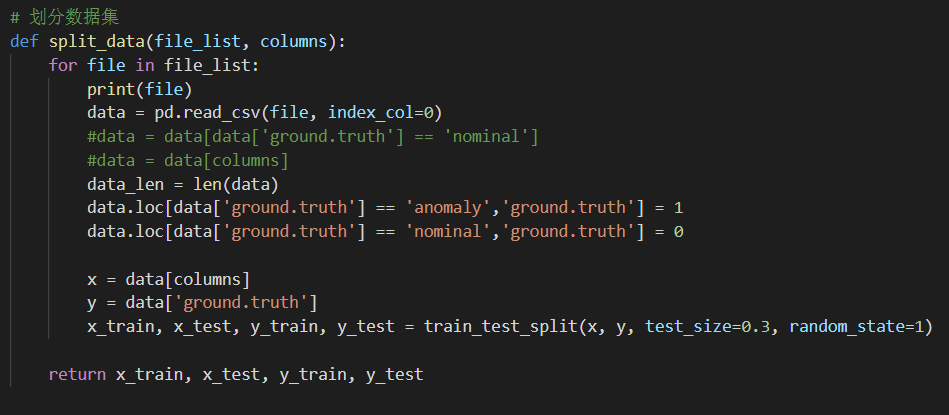
PCA算法是一种最常见的数据降维的方法，它可以将原数据进行线性变换，并找出数据中信息含量最大的主要成分，去除信息含量较低的成分，从而减少冗余，降低噪音。通常在异常检测的语境里，噪音（noise）、离群点（outlier）和 异常值（anomaly）是同一件事情的不同表述。所以，PCA既然可以识别噪音，自然也可以检测异常。

**Isolation Forest算法**

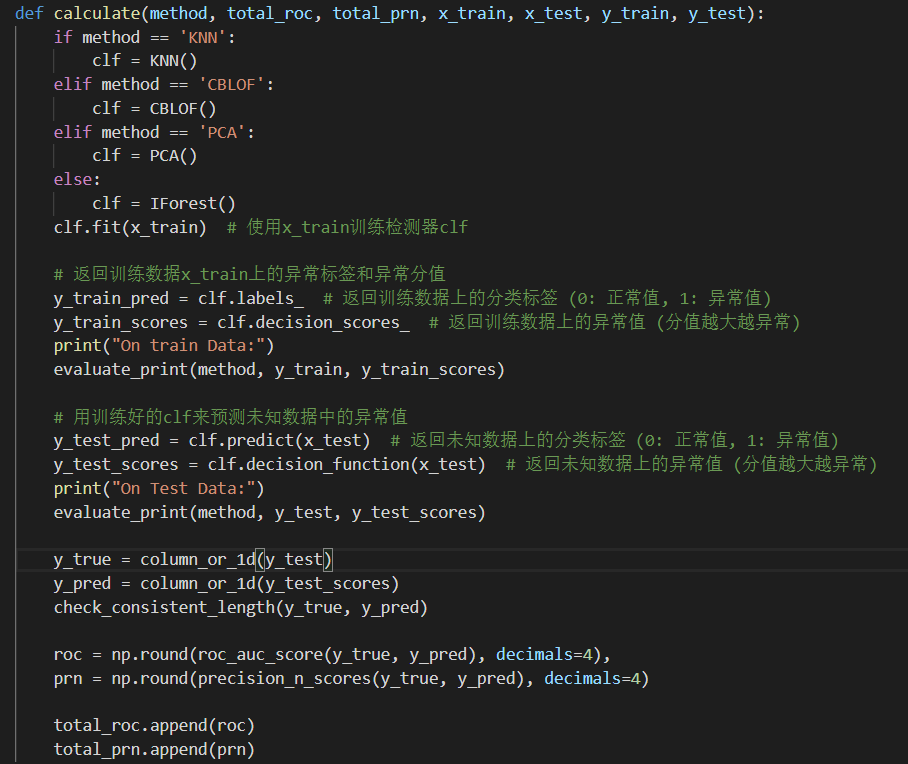
该方法是基于聚类的方法，使用一组树完成数据分区。隔离森林提供了一个异常分数用来查看结构中点的隔离程度。

**四、训练与测试（以数据集wine\_benchmarks举例）**

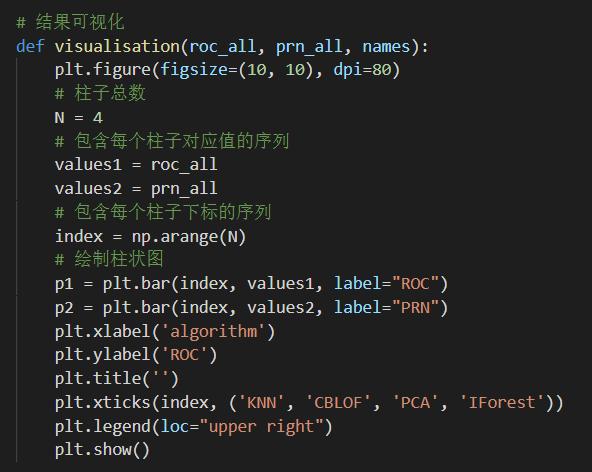
**1. 对每个csv文件，分别将其换分为训练集和测试集。**



**2. 在训练集和测试集上分别计算ROC值和PRN值。**



**3. 结果可视化。**



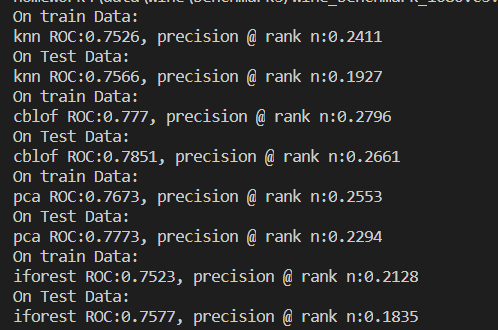
**五、实验结果分析**

**1. wine\_benchmarks数据集**

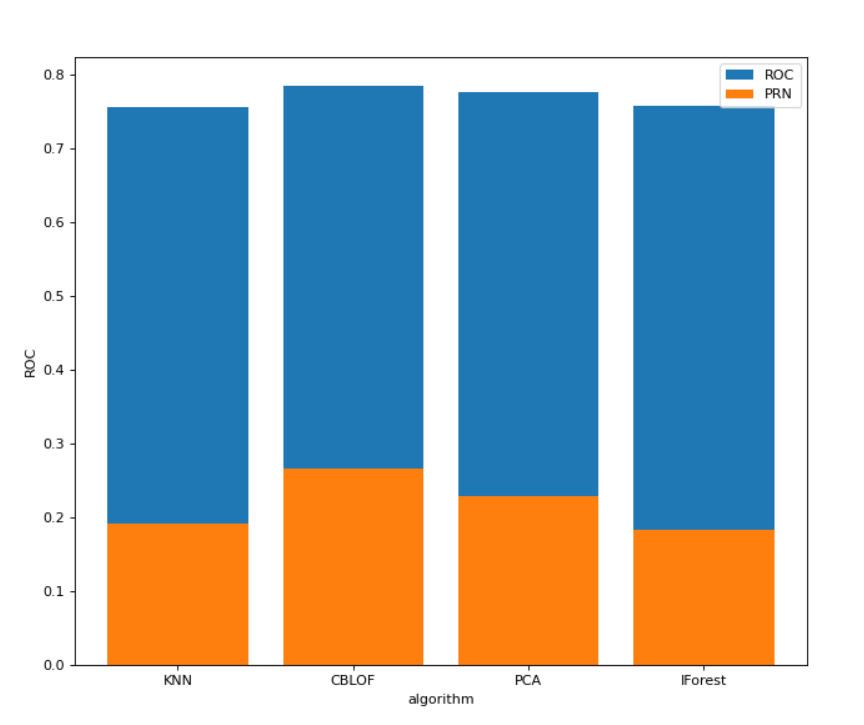
该数据集共包括1210个.csv数据文件（注：文件的最大标号为1680，但实际只包括了1210个文件），其中各个数据文件中共有的公共列有13个，分别为：

columns = ['density', 'fixed.acidity', 'original.label', 'free.sulfur.dioxide', 'volatile.acidity', 'chlorides', 'total.sulfur.dioxide', 'pH', 'alcohol', 'sulphates', 'citric.acid', 'diff.score', 'residual.sugar']

训练数据集、测试数据集在不同算法下的ROC值和PRN值如下所示：



我们将各算法在测试集上的实验结果进行可视化对比，得到如下对比图：



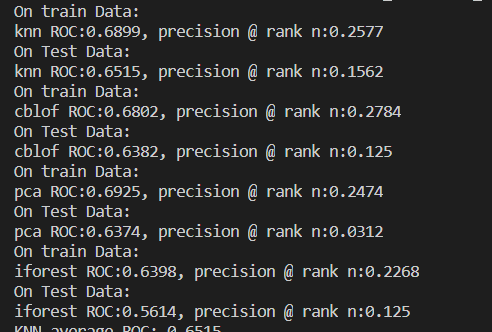
从图中我们可以发现，在ROC值方面，CBLOF算法的实验结果最佳；在PRN值上，同样也是CBLOF算法的实验结果最佳。也可以大致看出，在两个评测值上，四种算法的优劣状态基本一致。

**2. abalone\_benchmarks数据集**

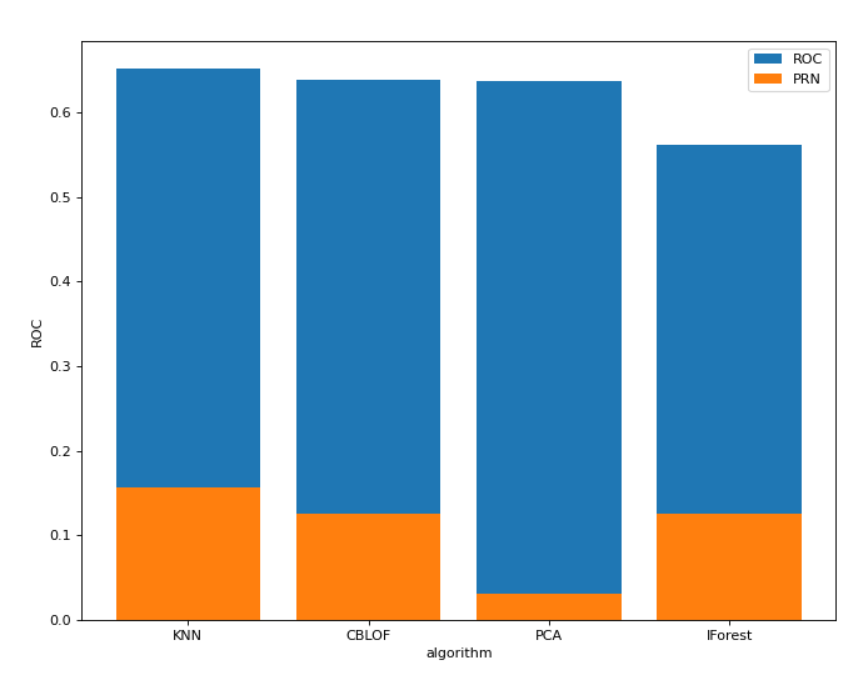
该数据集共包括1725个.csv数据文件（注：文件的最大标号为1800，但实际只包括了1725个文件），其中各个数据文件中共有的公共列有9个，分别为：

columns = ['V2', 'diff.score', 'original.label', 'V1', 'V4', 'V3', 'V7', 'V5', 'V6']

训练数据集、测试数据集在不同算法下的ROC值和PRN值如下所示：



我们将各算法在测试集上的实验结果进行可视化对比，得到如下对比图：



从图中我们可以发现，在ROC值方面，KNN算法的实验结果最佳，IForest算法结果最差；在PRN值上，同样是KNN算法的实验结果最佳，但结果最差的却是PCA算法。