## 一、使用的数据集，算法和结果

使用了skin和wave两个数据集，最终结果用所有csvs数据集AUC值的平均值来体现;

对于每个数据集分别采用了pyod中的PCA、KNN、IForest、OCSVM和MCD算法计算了他们的AUC值，每个算法的代码分别在**数据集名\_算法名**中，如skin数据集用PCA实现的代码文件命名为skin\_PCA，最终结果如下：

对于skin数据集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | PCA | KNN | IFrest | OCSVM | MCD |
| 训练集AUC值 | 0.637383 | 0.817711 | 0.774665 | 0.633942 | 0.801208 |
| 测试集AUC值 | 0.623441 | 0.798716 | 0.762229 | 0.621339 | 0.791498 |

对于wave数据集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | PCA | KNN | IFrest | OCSVM | MCD |
| 训练集AUC值 | 0.594552 | 0.753252 | 0.763253 | 0.642452 | 0.796343 |
| 测试集AUC值 | 0.586366 | 0.742432 | 0.756574 | 0.634536 | 0.786363 |

## 二、代码分析思路和结果分析

分析思路：

由于数据集有不符合实际的值和多余的属性，需要先对数据集进行处理，只提取出对分类结果有明显影响的属性，因此对于像id，motherset，origin与分类明显无关的属性可以删除。

可以把label当作分类结果，将该问题看成有监督问题，也可以不加入label属性，当成无监督问题看待。

接着把数据集划分为训练集和测试集，选择任意一种pyod提供的分类算法，对训练集数据进行拟合，训练分类器，使用decision\_scores\_方法预测训练集是否为离群点。

最后用sklearn中的roc\_auc\_score方法得到auc值，即roc曲线下面的面积，对该值求平均值。

结果分析：

从结果可以看出，预测结果基本位于0.6到0.8之间，效果不是很好，后面可能需要选择更好的分类算法，或者对一些参数进行调整以及引入神经网络等改进结果。

除此之外，测试集的AUC值比训练集稍差，符合一般情况，说明算法实现的结果本身还是可靠的。