第二次竞赛

河北师范大学2017级软件学院Kagfle第2次竞赛

实验步骤

1. 使用pandas加载数据集
2. 训练集测试集用君之代替
3. 进行数据降维
4. 进行数据升维
5. 构建SVM算法进行决策（一开始是用的全连接神经网络发现怎么调参结果都不好）

函数

分为open\_file文件和SVM处理文件

Open\_file文件中

文件开启

**def** open\_csv(CSV\_train\_path, CSV\_test\_path):

训练集数据与类别划分

**def** Separation\_train(train\_pd):

数据的归一化

**def** frequency(data):

获取数据为？的下标

**def** findindex(org, x):

缺失值填补

**def** Missing\_value\_processing(data):

文件保存

**def** Document\_Storage(test\_arr):

箱式图

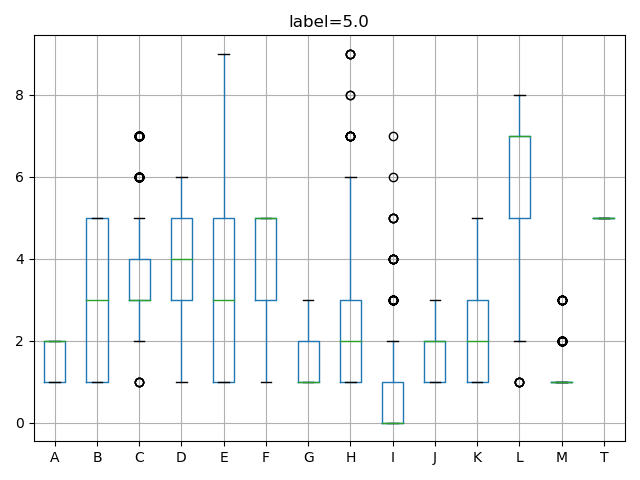
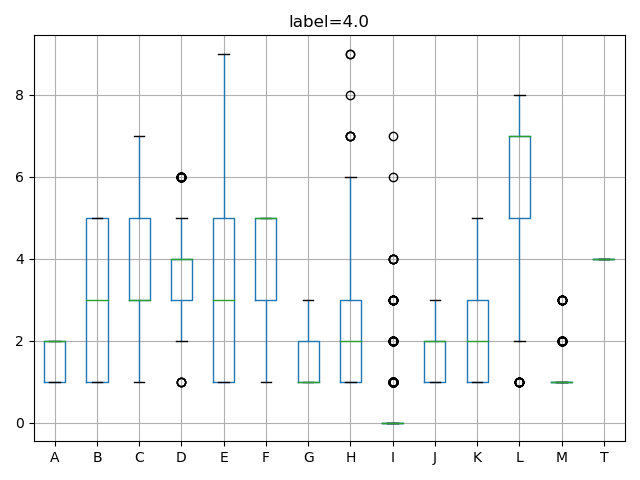
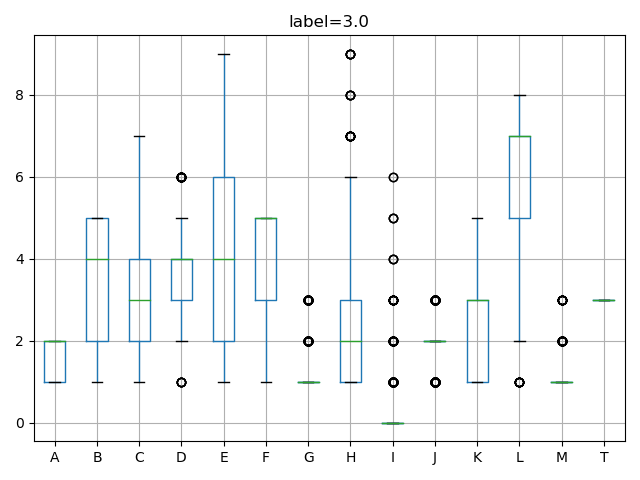
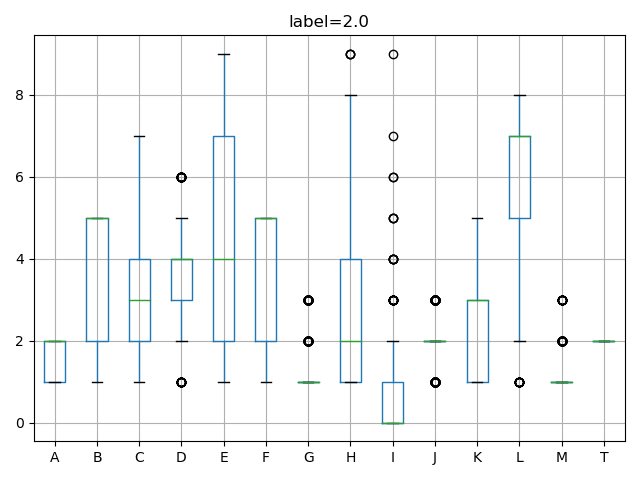
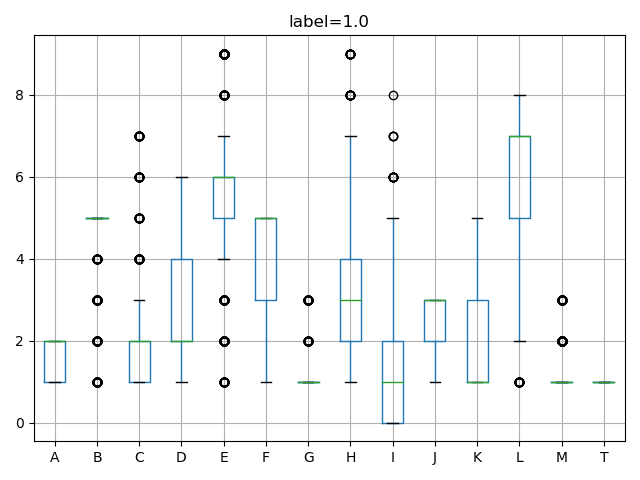
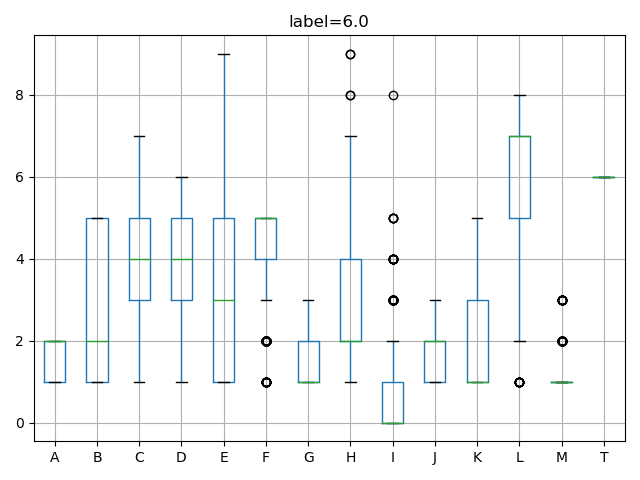
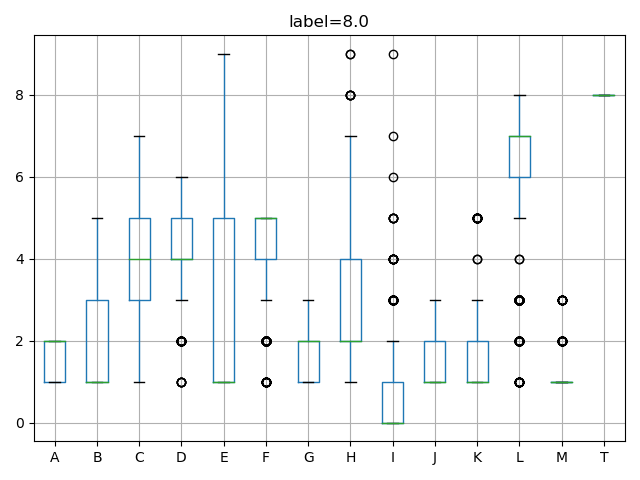
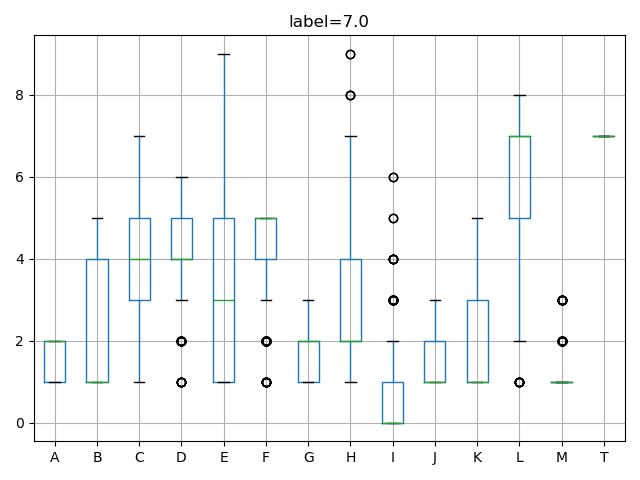
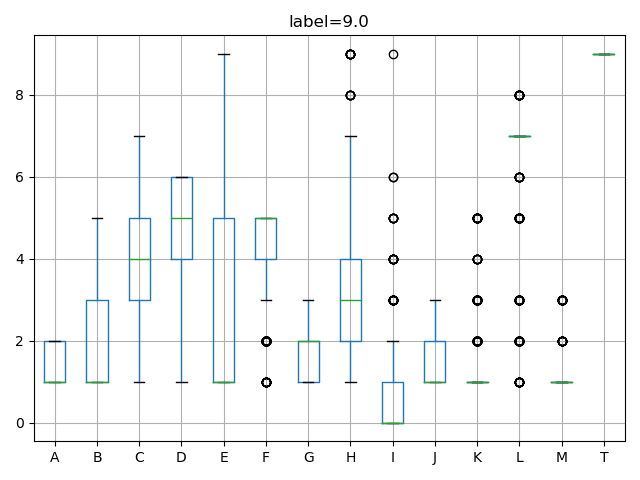
*def Box\_figure(dataset, label)这个存在于全连接神经网络阶段分析到后来我就删掉了*

数据处理类

**class** dataset():

SVM中是模型的构建和模型的选择和调参过程

一些数据分析



试验流程：

起初我用全连接神经网络进行了模型的构建和基本的数据处理，由于正确率只有百分之29我进行了一些盲目的数据降维工作，发现效果不是很好，我就计算了数据的相关性，找到了相关性较小的数据列进行了删除处理，然后效果值提升了百分之三我觉得这可能和各个类别数据重叠有关，我就进行了数据的升维，正确率确实提升了但是整体还是没有很大的改观，所以我进一步认为我的模型不是很符合，决定换一个模型，采用了SVM模型由于时间紧迫也没有达到理想值。