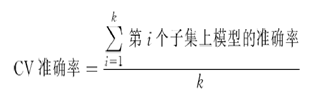
数据处理思路和所用的具体方法

根据本次实验的要求，本次实验采用的降维方法有T检验提取特征基因、基因芯片数据的PCA和基于SVM的rfe方法，分类方法采用的是SVM和K近邻方法，将这些方法组合，依据实验基本顺序：

**筛选特征表达基因（部分方法需样本划分）—K折交叉验证—构建分类器—验证准确率—对比评价**

实验步骤（包括下载数据）及其结果（要写清楚试验步骤及其结果和其它自己要补充的方面，步骤要求按照所写的步骤的指导能够对实验过程及结果重现）

1.实验原始数据为前期预处理完毕的基因芯片的表达数据。本次实验中，为了避免训练中可能出现的过拟合现象，每次迭代中采用K倍交叉验证法进行试验，得到K个模型，最后得到交叉验证的准确率其计算的公式如下：



2.基于PCA的分类。

首先对原始数据进行主成分分析，使用软件为cluster3.0，得到一个主成分文件，将其导入matlab工作空间，作为样本特征。

1. 用K近邻分类器。

将不同权重的主成分依次循环读入，每一次循环用k折交叉验证法进行验证，防止过拟合问题，分类时使用函数knnclassify，使用未参与分类的数据进行验证，计算正确率。试验中，经过多次的对比验证，K近邻分类器选择“夹角余弦距离法“和”最近邻原则“配合有最佳的实验准确率，之后的K近邻分类器亦是使用此种参数搭配。

（2）用SVM分类器

将不同权重的主成分依次循环读入，每一次循环用k折交叉验证法进行验证，libsvm和matlab的fitcsvm分类器都是基于SVM方法的分类器，经过对比，fitcsvm准确率相对较高，故使用其作为分类函数。并且使用未参与分类的数据进行交叉验证，计算正确率。

3.基于T检验的分类。

留出部分作为样本划分集，对余下的数据进行T检验，取p值小鱼0.001的部分。

（1）用K近邻分类器。

将差异基因依次循环读入，每一次循环用k折交叉验证法进行验证，防止过拟合问题，分类时使用函数knnclassify，使用未参与分类的数据进行验证，计算正确率。试验中，经过多次的对比验证，K近邻分类器选择“夹角余弦距离法“和”最近邻原则“配合有最佳的实验准确率。

（2）用SVM分类器

将不同权重的主成分依次循环读入，每一次循环用k折交叉验证法进行验证，libsvm和matlab的fitcsvm分类器都是基于SVM方法的分类器，经过对比，fitcsvm准确率相对较高，故使用其作为分类函数。并且使用未参与分类的数据进行交叉验证，计算正确率。

注：由于T检验的基因数据较多，为了使acc数据曲线变得比较平滑，使用简单移动平均法剔除掉随机波动，以反映出其基本轨迹。

4.基于SVM-RFE的分类。

在安装有libsvm的支持下，运行svm-rfe特征基因筛选器，从所有基因中选择前200个作为特征基因。

（1）用K近邻分类器。

将差异基因依次循环读入，每一次循环用k折交叉验证法进行验证，防止过拟合问题，分类时使用函数knnclassify，使用未参与分类的数据进行验证，计算正确率。试验中，经过多次的对比验证，K近邻分类器选择“夹角余弦距离法“和”最近邻原则“配合有最佳的实验准确率。

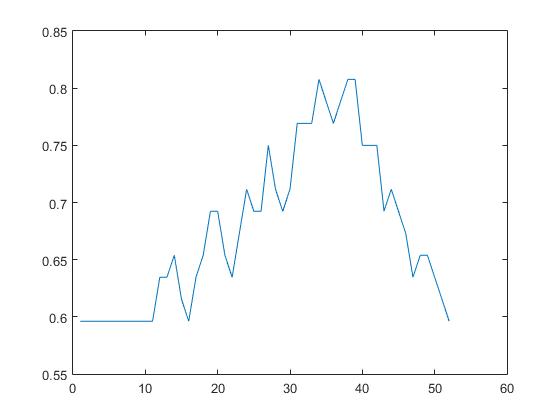
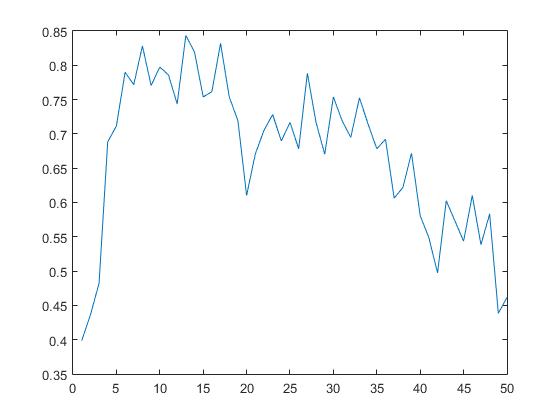
（2）用SVM分类器

将不同权重的主成分依次循环读入，每一次循环用k折交叉验证法进行验证，libsvm和matlab的fitcsvm分类器都是基于SVM方法的分类器，经过对比，fitcsvm准确率相对较高，故使用其作为分类函数。并且使用未参与分类的数据进行交叉验证，计算正确率。

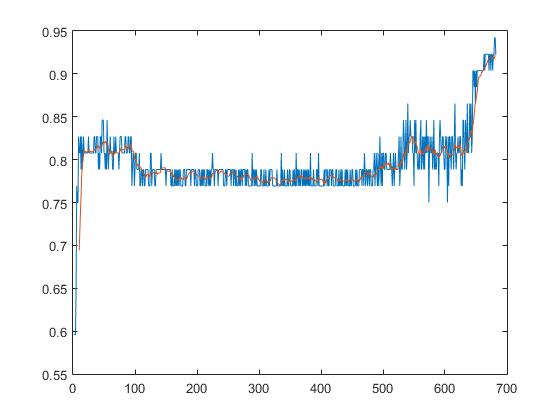
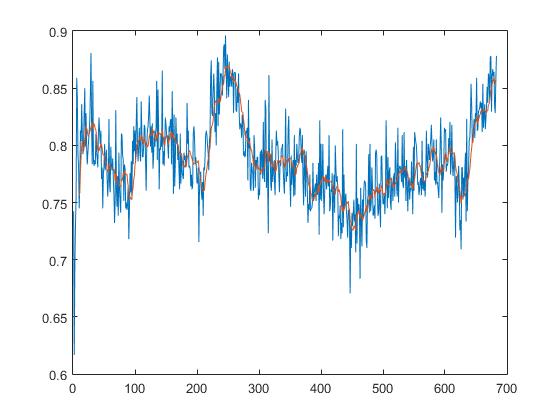
5.实验结果分析。

KNN分类 SVM分类

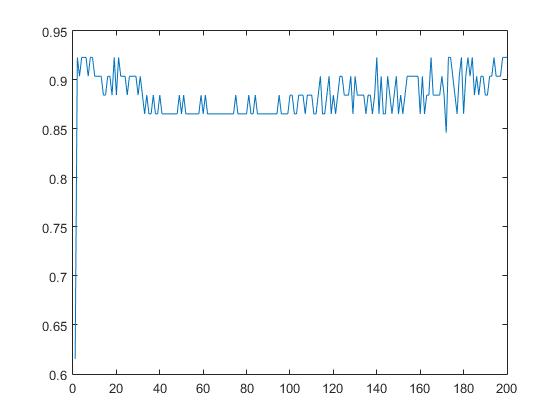
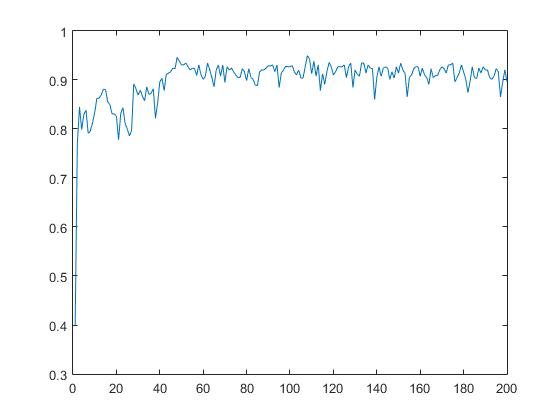
**PCA**



**Ttest**



**Svm-rfe**



（1）不同分类方法下ACC的对比

总体来看，k近邻方法下的准确率略高于svm方法，但差别很小，k近邻方法的波动较大，稳定性较差。

（2）不同差异基因筛选方法下的ACC对比

a.PCA方法下，都是先升高然后降低，可能原因为PCA之后的数据权重不同，越向后越权重越小，对基因的差异度贡献不大。K近邻方法得到的ACC较高些。

b.T检验方法下筛选，knn结果波动较大，svm的结果波动较小。两者的ACC都在0.75-0.85之间。

c.svm-rfe方法下的准确率都较高，在0.8-0.9之间波动，并且波动幅度较小，我们可以认为在所有差异基因筛选的过程中，此种方法具有最佳效果。