周来平

zlp@sjtu.edu.cn

摘要

本次大作业使用PCA+SVM，Reduced Rank LDA，QDA，PCA+KNN  
等方法对心电数据进行分析。

统计学习与推理大作业

心电图ECG数据分析报告

# 第一部分 心电数据预处理

## 数据对齐

经过观察，心电图的开始位置（P波开始处）与尖峰R的距离普遍约为200个单位长度。因此首先我通过极大值获得了尖峰的位置，然后将整张图移动，使得R的位置刚好在200处。用这种方法可以使心电开始处刚好在零点。除此以外，将所有心电图后面补零至相同长度，方便后续使用各种算法。

## 心电图绘制与查看

为方便观察，我编写了心电图查看程序，可直接查看某个病人某次心跳的12导心电图数据。如图1-1。此病人的数据已经经过1.1的数据对齐处理。

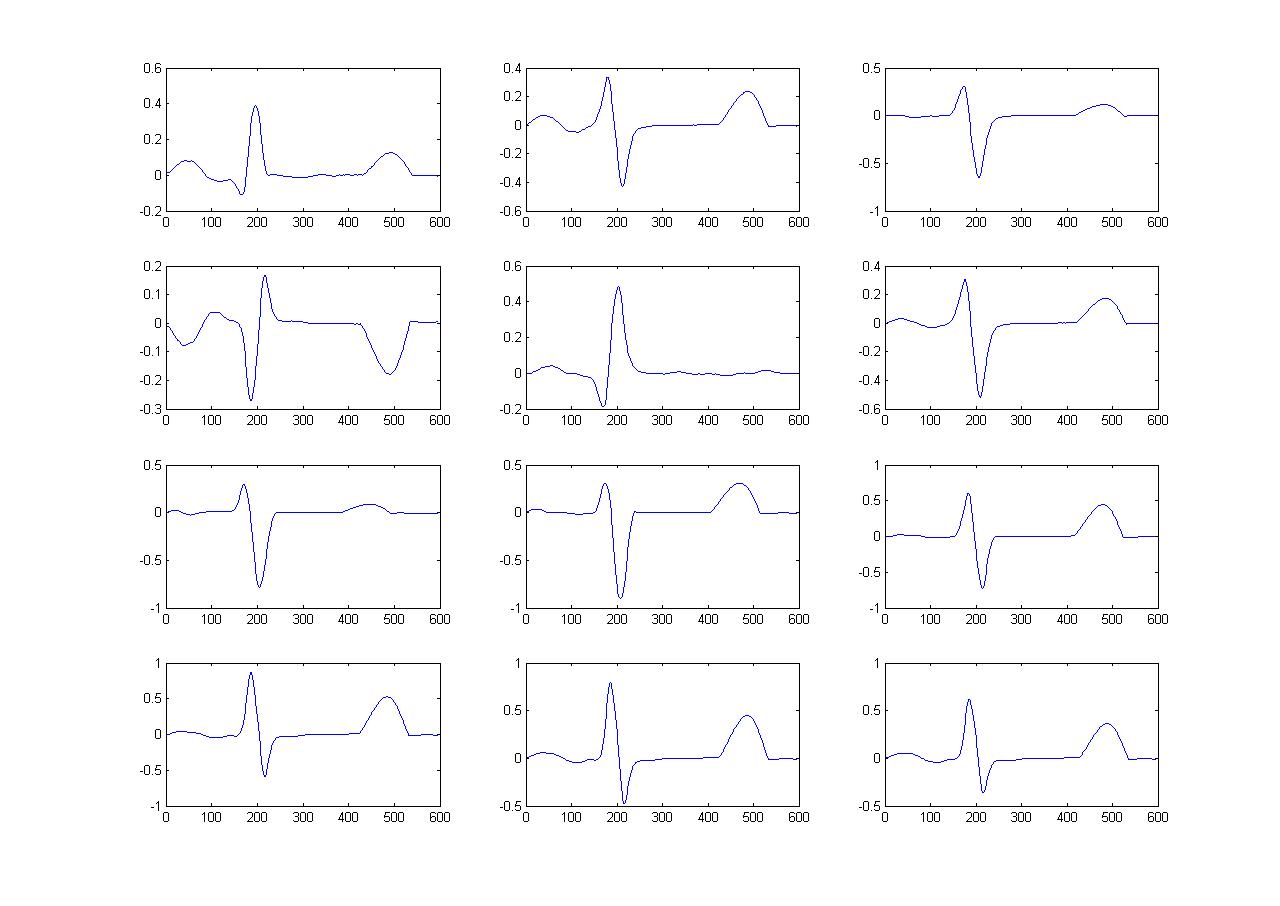


图1-1 标注为”窦性心律，电轴左偏“的一位病人的心电图，经过数据对齐处理

## 分割数据

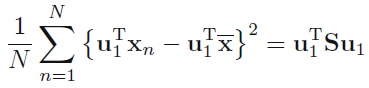
将数据分成训练数据和测试数据(9:1)，以方便后续进行交叉校验。这里疾病共有5类，每类10个病人。正常心电共50个人的样本。因而每个测试数据中每种疾病各有一个病人的数据，并有5个正常人的数据。其他数据归为训练数据。这样就可以做10折的交叉校验。

# 第二部分 PCA降维处理及SVM分类

## 2.1 Principal Component Analysis

PCA的主要用途是降维，有损数据压缩，特征提取，数据可视化等。

此次心电数据最长序列约为850维，因此需要降维操作，将误差减小，并暴露出主要影响因素。故此优化问题是使得投影到的维度上方差尽可能大，也就是说使得下式尽可能大。



其中S是数据的协方差矩阵。所以降维至N维空间等价于求S的前P大特征值，等价于求中心化的X数据矩阵前N大的奇异值。而这个空间则是这些特征值对应的特征向量张成的。如图2-1，是PCA的图形化解释。

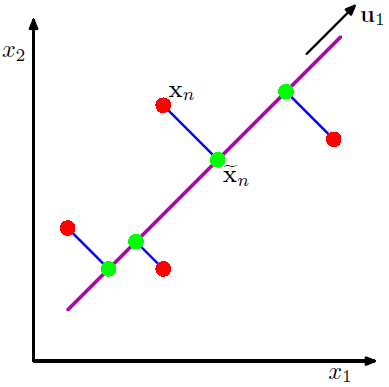


图2-1 PCA降维，数据被投射到方差最大的方向上

## 2.2 PCA降维结果

下面例举训练出来的PCA主方向的示意图，参见图2-2。

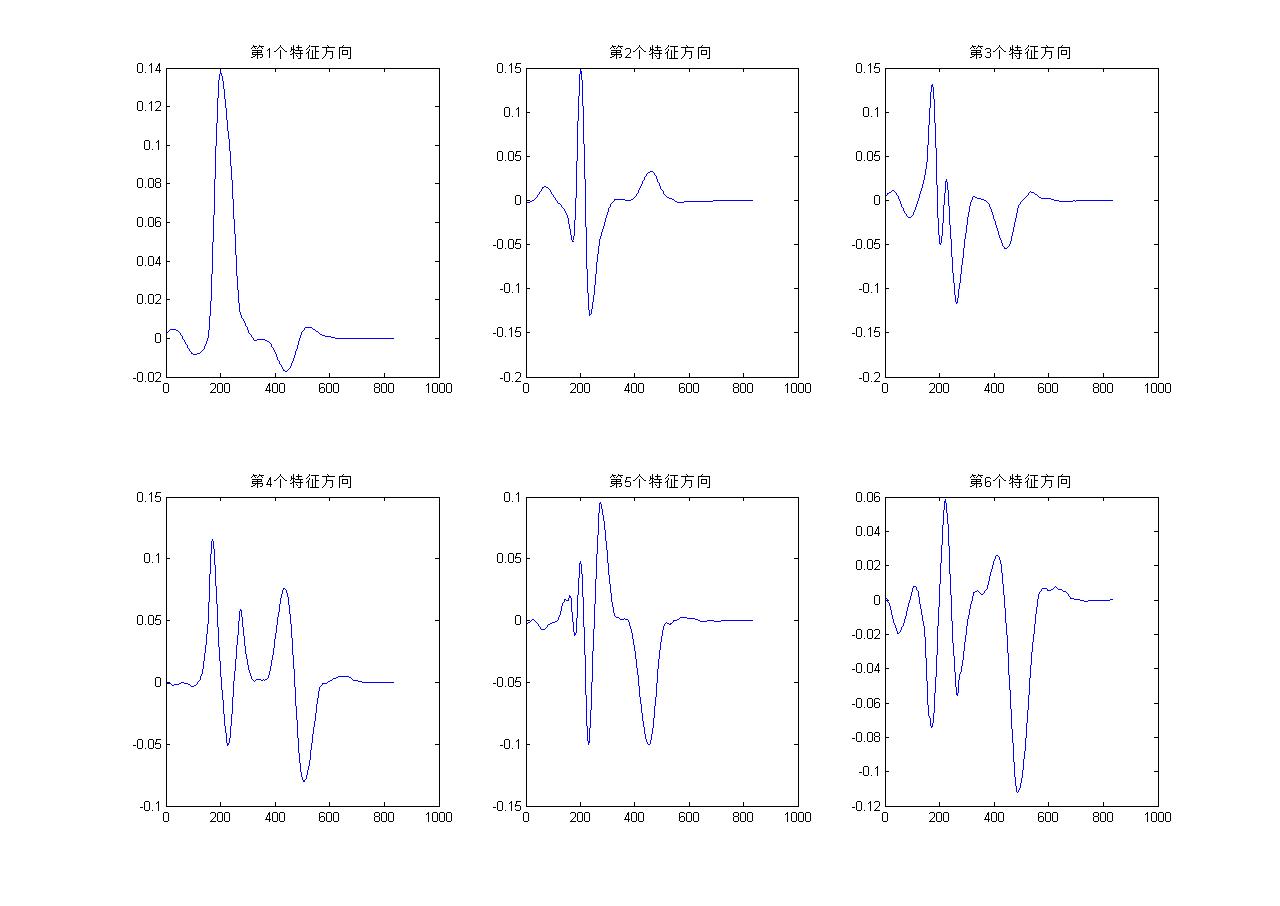
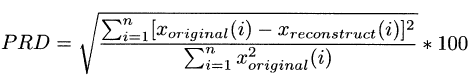


图2-2 各特征向量示意图

越靠前的特征向量（特征值越大），对应的系数越能体现出一张图的特点。通过查阅相关文献，我在[1]中找到了一种特征方向的评价方式：PRD



这个量主要用于衡量特征向量对原始数据描述的完整程度。PRD值越高，说明还原度越高。本实验中PRD值如表2-1。

表2-1 特征向量数的PRD值

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 特征向量数 | 1 | 2 | 5 | 10 | 20 | 50 | 100 |
| PRD | 0.71002 | 0.5256 | 0.36891 | 0.22628 | 0.12305 | 0.05107 | 0.02329 |

我们可以看到达到k = 50时，特征向量几乎可以表示心电图的全部特征，这比800维小多了。不过并不是k越大，降维效果越好，这需要后续的进一步评估。

接下来我们看投影到新的低维空间上的向量分布。如图2-3，是五种疾病的病人以及正常人的10维向量示意图。每幅图将他们每个人约10次心跳绘制出来，以体现降维结果的稳定性。

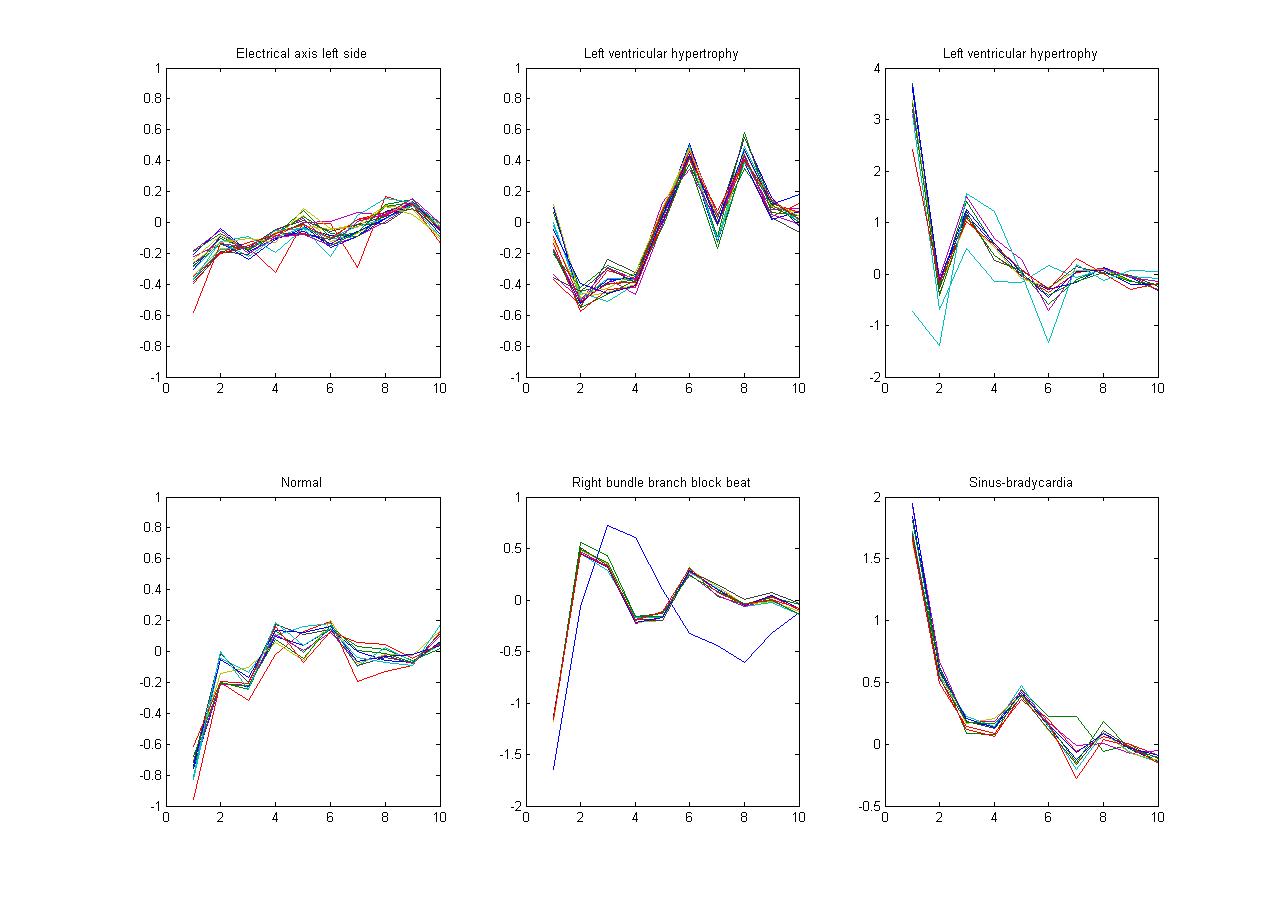


图2-3 不同病的病人前10个特征向量对应的系数，每幅图记录了10次心跳

这些新特征在空间中的分布如图2-4.这里仅仅使用了前两个特征向量，在平面图上表示出来。可以看出除了蓝色和红色的点比较显著外，其余的点比较勉强。其中原因可能是维数比较低。

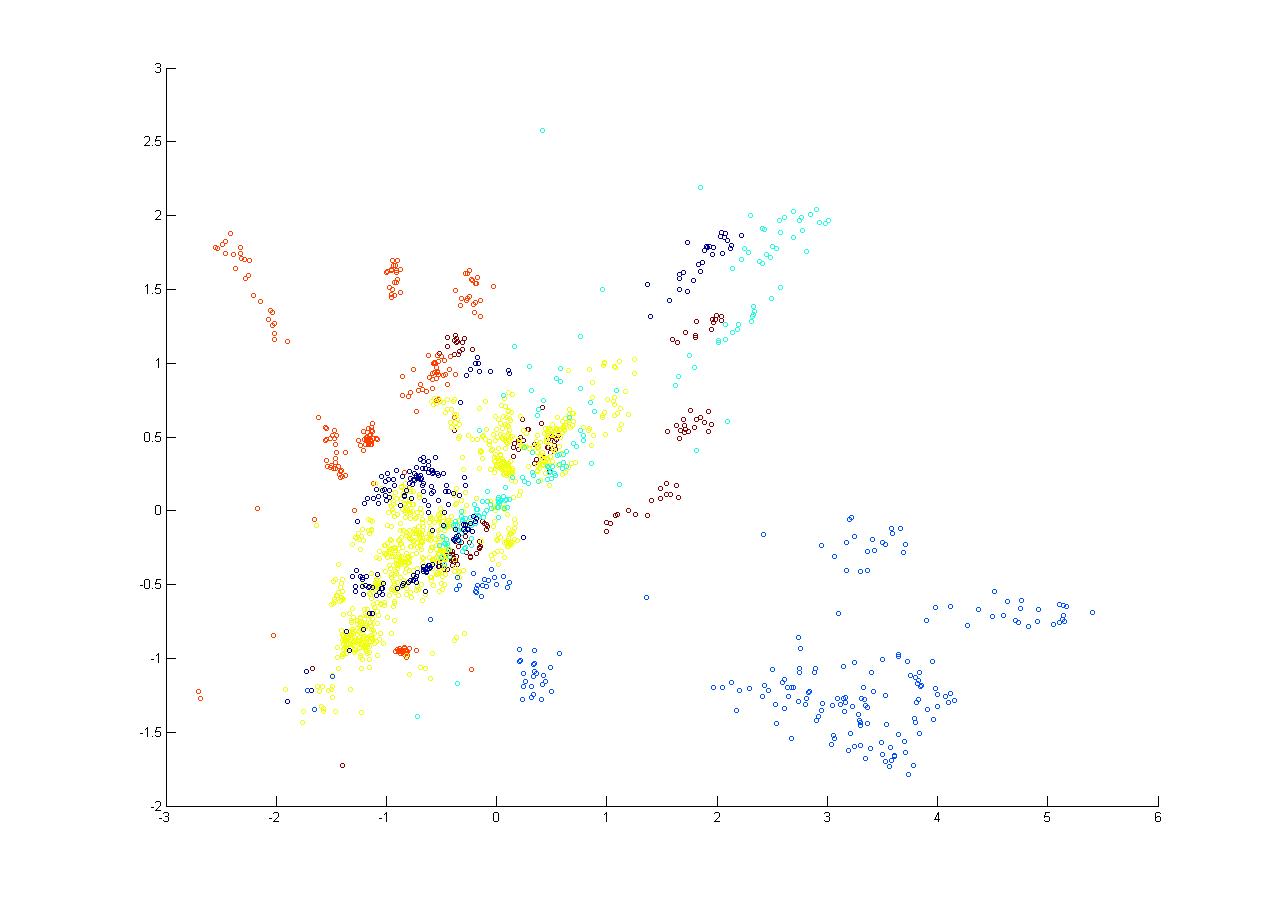


图2-4 仅使用前两个特征向量绘制的散点图

## 2.3 支持向量机(Support Vector Machine)

支持向量机是一种非常常用有效的分类器。这里我们采取这种方法。由于这里是6类分类问题，我训练了6个分类器，用来判断第k个类是否正确。这种方法比起一对一的方法有更加充足的训练数据，所以我采用了它。

## 2.4 交叉校验(Cross Validation)

如上文所说，我使用的是10折的交叉校验方法。对于精度与PCA降维维度的交叉校验如图2-5所示。

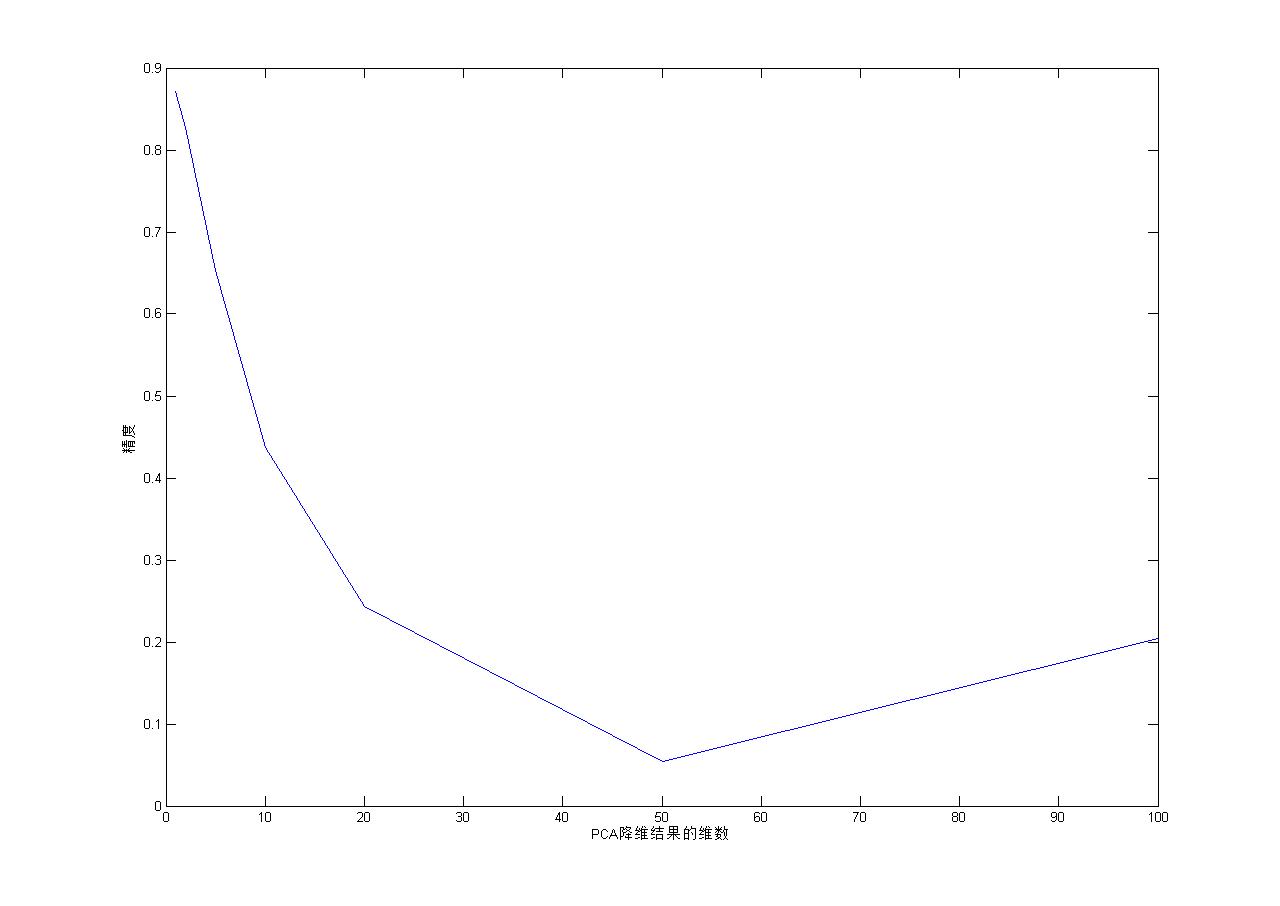


图2-5 精度与PCA维度的关系

出人意料的是，维数越低，交叉校验的结果越好！这里可能暴露一些问题，将在下面的章节详述。

不管怎样，交叉校验的结果告诉我们要取p = 1。此时精度为87.13%。

同时，为了提供对照，我也做了直接SVM训练原始生数据的对照组，交叉校验的精度为20.43%。可以看出PCA降维给SVM分类提供了十分正面的影响。

## 2.5 问题反思

与一般情况不同的是，这个PCA的结果是维数为1时效果最好。我对上面的很多图片做了一些分析研究，其中图2-3中对于得相同疾病的不同病人来说，我发现他们的10维特征还是有相当的差异的（尽管一个病人的所有心跳得到的特征都十分相似）。这可能说明心电图的个体存在差异，患同种疾病的病人的心电图可能有一些差别，而这些差别刚好被我的PCA捕捉到了，从而造成了结果不佳。另见图2-4，我发现同种疾病的数据点似乎有多个中心，这也支持了上述的说法。而PCA在1维时最佳这个情况表现出了过拟合的现象，过多的维数带来了过多噪声（比如说个体差异），影响了SVM的判断。解决过拟合的方法是增加数据，或降低维数（但这里降不下去了……），但本质的方法还是让一开始的特征更干净。

我感觉这种非高斯的同类多中心分类问题使用更加灵活的模型可能会更好，所以就有了下一节。

## 2.6 KNN尝试

KNN是非常灵活的模型，虽然我说过过拟合在项目中比较严重，但是KNN比较适合分类2-4这样的灵活的数据，所以我就应用了KNN方法来解决问题。

下图是KNN的交叉校验结果。这是使用PCA输出的10维特征向量完成的。

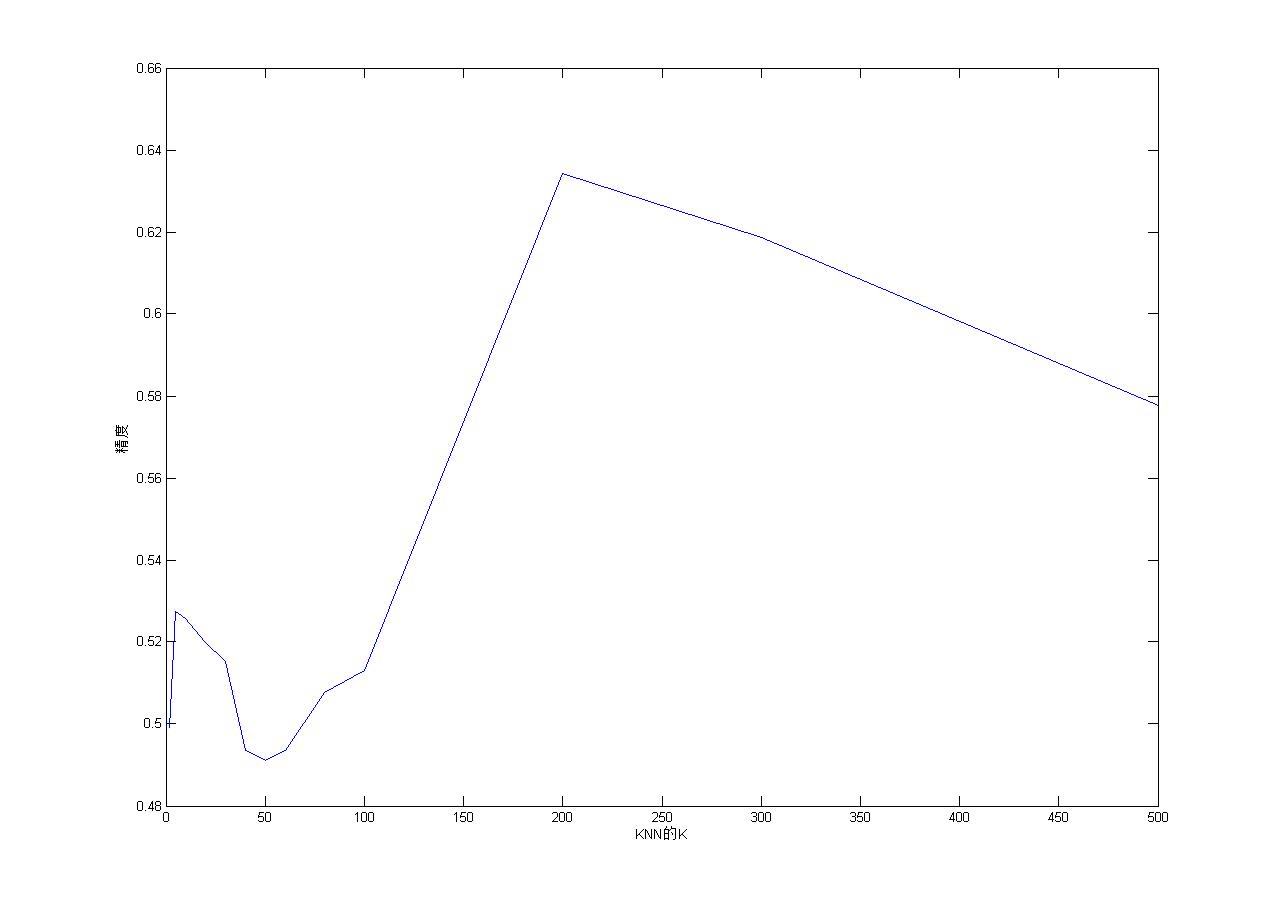


图2-6 KNN交叉校验结果

好消息是在K = 200时，结果很不错，为63.42%，比SVM的10维特征向量分类效果要好。然而坏消息是这仍然没有超过SVM处理1维特征向量的精度。而处理1维特征向量由于KNN的本质，效果会很差，无法与SVM相比，所以PCA解决这一问题的结论就是最佳方案是使用1维向量，且使用SVM分类器。

## 2.7 问题再反思

上面提到了同疾病的病人个体差异过于明显，而Reduced Rank LDA考虑了类标，最小化类内方差，最大化类间方差，十分适合解决上面的困境。因而我又继续采用Reduced Rank LDA方法进行数据分析。

# 第三部分 Reduced Rank LDA

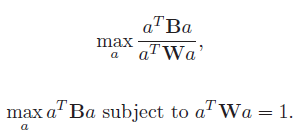
## 3.1 Reduced Rank LDA

1. 计算出K个中心点，维数为K \* p的中心点矩阵M和全部数据的协方差矩阵W（类内方差）。

2. 白化数据。M\* = MW-1/2

3. 计算M\*的协方差矩阵B\*，计算其谱分解，得到特征值最大的特征向量。将数据投射到这些特征向量张成的空间即可。

本质上是解决下面的问题：



## 3.2 Reduced Rank LDA实现结果

Reduced Rank LDA可将数据降至K – 1维。本实验是5维。

特征向量如下图所示。

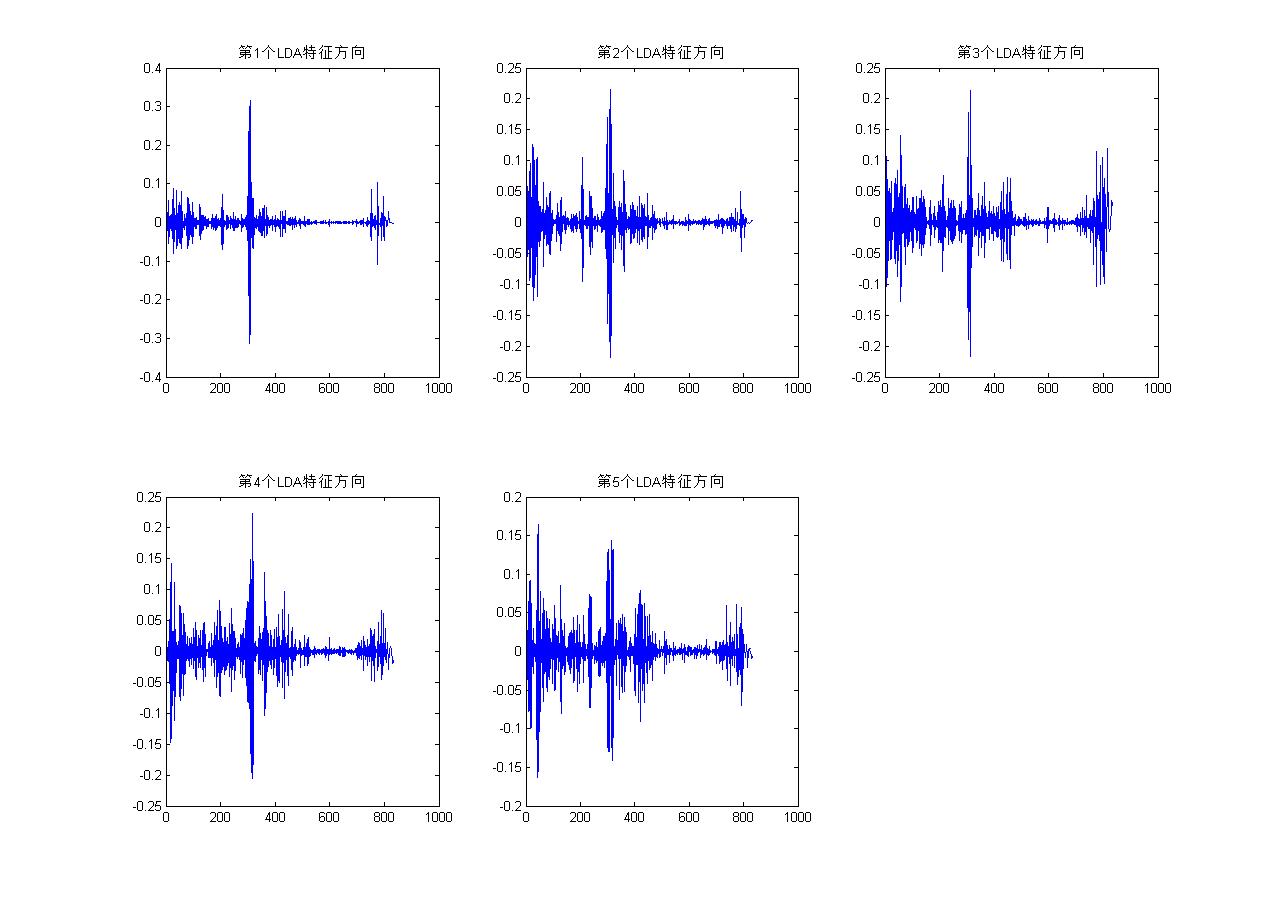


图3-1 Reduced Rank LDA全部5个特征向量的图示

如图3-2是前50个心跳的5维向量的示意图（这至少取自3个病人的心跳）。可以看出类内的差异不再那么明显。

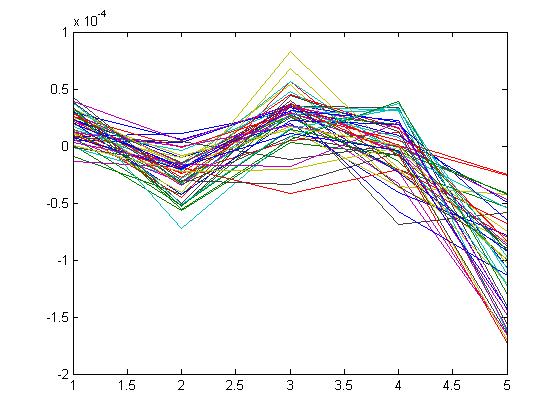


图3-2 前50个心跳的5维向量示意图

最令人振奋的是图3-3的4幅图。从5个方向看去，似乎每个类都能在某一维被精确地分辨出来，而且蓝色，红色和青色三种类型能分得很清楚。不过不可否认的是数据之间还是没有较明显的Margin，这可能是后面分类效果不佳的原因之一。

与PCA的图2-4不同的是，这里同类的数据并没有被分为多个中心，看来Reduced Rank LDA的特性起作用了。

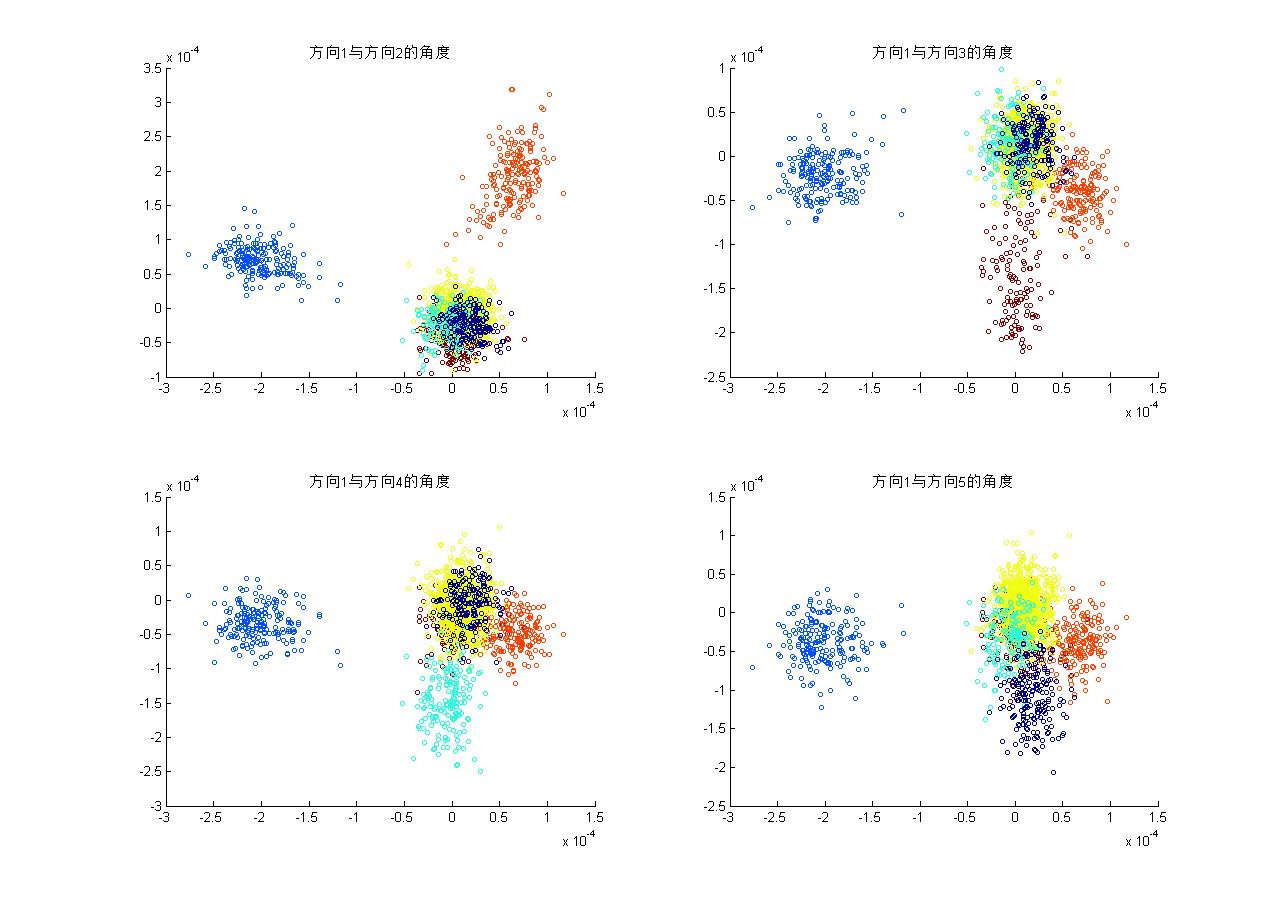
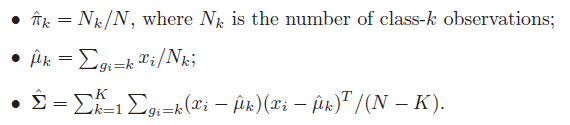


图3-3 5个维度上数据点的分布情况

## 3.3 使用LDA

降维后就可以使用LDA获得判别函数，进行判别了。

下面是极大似然估计法的结果，即训练过程:



判定函数是：



## 3.4 交叉校验(Cross Validation)

我们将12导的5维向量叠在一起进行训练，结果为69.05%。

使用subset方法，取5维向量的子集进行训练，共有五个CV的对象，分别是[1], [1, 2], [1, 2, 3], [1, 2, 3, 4], [1, 2, 3, 4, 5]，结果如下。可以看出没有明显的证据表明进一步降低维数对于精度的提升有帮助。（这里仅使用1导数据，所以效果不及上面12导数据好）。

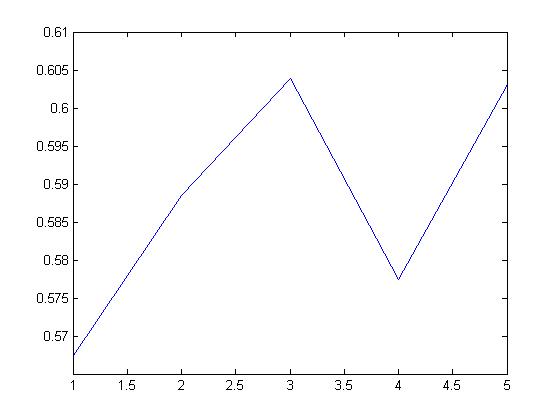


图3-4 k个维度(k <= 5)的交叉校验

## 3.5 使用QDA

QDA与LDA唯一的差别就是不同类的协方差可以不同，判别线是二次曲线。训练与判别函数构造是十分相似的。

QDA交叉校验的精度为65.06%。看来不需要使用这一模型来优化判别函数。

## 3.6 再次使用KNN

对Reduced Rank LDA降维得到的5维向量使用KNN进行分类，结果的交叉校验如下（这也是用1导数据做的）：

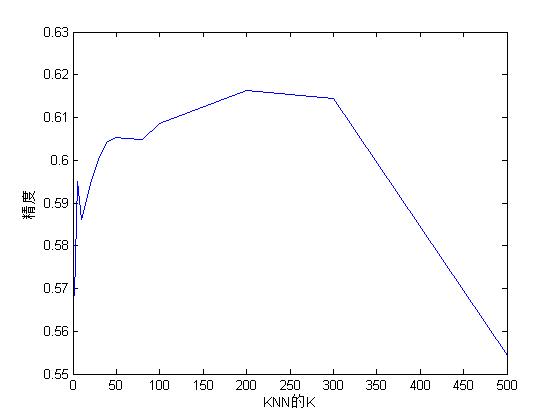


图3-5 KNN的交叉校验

可以看出KNN对于Reduced Rank LDA得到的结果的分类能力略优于LDA的结果。但此优势很小。且在K = 200时效果最佳。

## 3.7 问题反思

Reduced Rank LDA的分类结果相比PCA差，这是意料之外的事情。虽然图3-3中类别分得如此之好，但是在我调试的时候有很多测试点在2-3维上准确地落在了区域内，却在其他几维上跑出很远的距离（然而也没有跑到别的中心上），最终以微弱劣势输给了先验最大“无病组”。还有很多点落在了边界上，也是以微弱劣势选择了错误的类别。因而我引入了KNN方法，因为从直观上说它更容易将边界点判断地正确一些，可惜效果不佳。Reduced Rank LDA的不良表现让我确信预处理可能还不够好，这可能是与PCA的通病。

# 第四部分 总结

## 4.1 结果

如上所说，PCA+SVM效果最好，然而却在PCA降到1维的前提下，可见过拟合很严重。RRLDA次之，尽管它的类别分得很清楚，但是仍然效果不尽如人意。可能是因为噪声干扰。QDA和KNN要弱于LDA，这说明线性高斯模型足以解释心电模型。而PCA+KNN比同维度的PCA+SVM要更好，这说明PCA得到的模型比高斯模型要复杂，需要KNN来解释。

表4-1 交叉校验精度结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方法 | PCA+SVM | PCA+KNN | RRLDA | RRQDA | RRLDA+KNN |
| 精度 | 0.8713 | 0.6342 | 0.6905 | 0.6506 | 0.6164 |

## 4.2 代码结构

common 公用代码部分

load\_from\_file.m 从文件夹中读入全部数据

pre\_process.m 预处理，数据对齐

data\_plot.m 绘制心电图

data\_load.m 加长向量，放入数据矩阵中等琐碎操作

knn.m 三行KNN实现代码

data\_division.m 分割数据为训练和测试两部分

train\_svm.m 如正文所说地训练K个SVM分类器

data 数据(放在这里就好，我不会提交数据的)

data\_tmp Cross validation等的暂存结果

fda Reduced Rank LDA

fda.m, lda.m, lda\_classify.m, qda.m, qda\_classify.m 训练和分类函数

test\_fda.m, test\_fda\_all.m 测试和12导测试函数

fda\_cv.m, fda\_qda\_test.m, fda\_subset\_cv.m, fda\_test.m, knn\_cv.m 测试，交叉校验的脚本

images 本文的截图

pca PCA方法

direct\_cv.m, direct\_test.m 作为对照的直接SVM训练的脚本

pca\_cont.m, pca\_test.m, prd.m PCA程序

knn\_cv.m, pca\_cv.m, test\_pca.m 测试和交叉校验的脚本

trial 实验过程中的一些测试函数或脚本

## 4.3 致谢

十分感谢张老师这门课程的谆谆教诲，这门课虽然很难，花费了我很多课余时间却仍然不能很好地掌握很多内容，但是重要的是为我开拓了眼界，让我能够更加游刃有余地进行接下来的科研和学习工作。同时也感谢助教和周正中博士在大作业过程中给我的悉心帮助和指导！

# 参考文献

[1] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction[J]. University of Chicago, 2001.

[2] Bishop C. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer[J]. Journal of Electronic Imaging, 2006, 16(4):140-155.

[3] Zhang H, Zhang L Q. ECG analysis based on PCA and Support Vector Machines[C]// Neural Networks and Brain, 2005. ICNN&B '05. International Conference on. IEEE, 2005:743-747.