SVM脑电波分类实验报告

李浩然

2017210914

1. 实验目的

给出二十个人的脑电波采样数据以及对应的睡眠阶段。利用前19个人的脑电波-睡眠阶段作为训练数据，训练多类支持向量机。再利用该支持向量机预测第二十个人的脑电波采样数据对应的睡眠阶段。将预测结果和真实结果作比较，并进行分析。

1. 数据描述

本次实验数据与之前贝叶斯判决实验使用同一数据集。

1. 实验环境

本次实验在Windows 10系统中借助Octave 4.2.1软件完成。

处于编程方便考虑，本次实验将“数据”和“标签”文件夹重命名为“data”和“label”。

1. 实验过程
   1. 读入数据

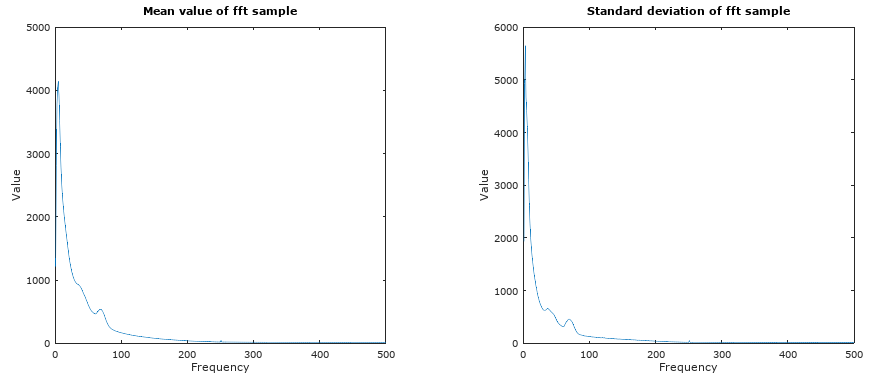
本次实验使用与之前贝叶斯判决实验使用相同的读入数据方法，得到仅包含1~3睡眠阶段的训练数据(sample, label)和测试数据(test\_sample, test\_label)。

* 1. FFT转换

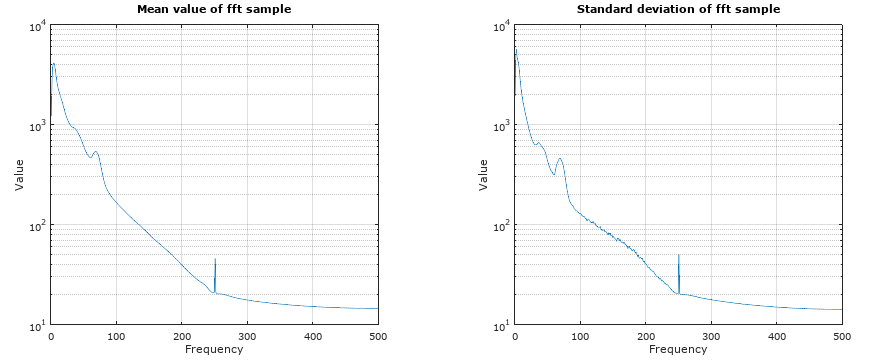
有上次实验报告中的分析可知，各个样本的特征在时域上没有明显的对应关系。所以需要将数据转换到频域。

上次实验中利用DTFT算法得到频谱的轴对称特性，将特征的维度减小了一半，即最终得到500维特征的数据进行后续处理。

本次实验尝试更进一步发掘实验数据的傅里叶频谱性质。观察发现，四分之一周期后的分量相比于其他频率分量已经非常小，接近于噪声了（见下图）。



将利用semilogy函数将y轴转换为对数坐标形式（见下图），可以看出，函数图象在x坐标为250附近有一个较明显的斜率变化。

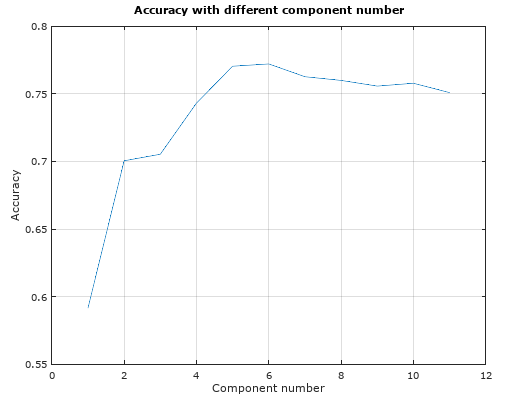


因此对数据进一步降维到250维。

* 1. PCA变换

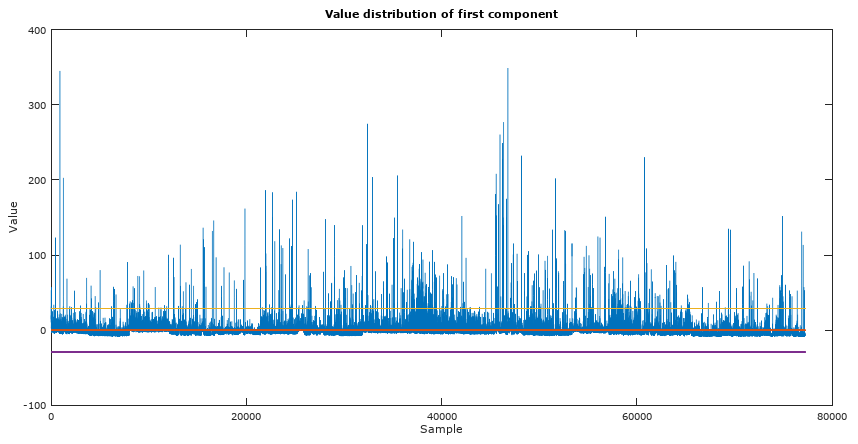
经过上述处理的250维数据维度仍然太高，所以考虑用PCA进行降维。不同于之前贝叶斯判决实验中简单地规整化（减去均值）后直接做PCA，本次实验考虑先对各个特征做标准正态化(normalize)。因为PCA算法选择每个分量的时候，都尽量使得投影其上的数据尽量分散，所以如果有个维度的数据过于分散，PCA在选取投影向量的时候就会偏向于选取该特征所在的分量，从而丢失了较多其他维度的信息。

对保留不同维数的数据集重复之前的贝叶斯判决（如下图），可以看出当选取6个成分的时候，预测的准确率达到最高的77.204%。



* 1. SVM数据预处理

为了防止某个特征的分布相比于其他特征太过分散，在SVM之前需要对数据做相应的缩放(Scaling)。以PCA降维后的第一个分量为例（见下图），图中红线是分量均值，上下两条黄线和紫线分别是均值加减三个标准差。可以看出，有少部分样本的分布超出了均值三个标准差，甚至更多。这些少量数据对数据缩放(Scaling)会产生巨大影响，所以考虑移除这些异常样本，即移除所有分布在远离均值三个标准差以外的样本。移除异常值的操作由filter.m脚本完成。



本次实验采用线性缩放，即利用下式将样本特征映射到[-1, 1]空间中。其中*Min*和*Max*分别是特征的最小值和最大值。再利用这两个参数，带入下式对测试数据的相应特征进行缩放。

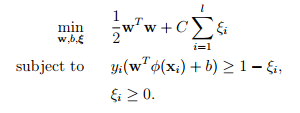
* 1. SVM

本次实验中使用[LIBSVM](https://www.csie.ntu.edu.tw/~cjlin/libsvm/)库完成支持向量机相关操作。

* + 1. LIBSVM原理

已知一组样本和每个样本对应的分类，这些样本可以用若干特征来描述。将样本集可以表示为，分类可以表示为。多分类SVM对每个类训练一个模型用于判断一个数据是否属于某类，预测时候综合各个模型的预测概率，选取概率最高的类。

训练某类模型的时候，整数空间上的分类标签会根据是否属于某类重新映射成1和-1。此时SVM要处理的问题就可以如下表达：

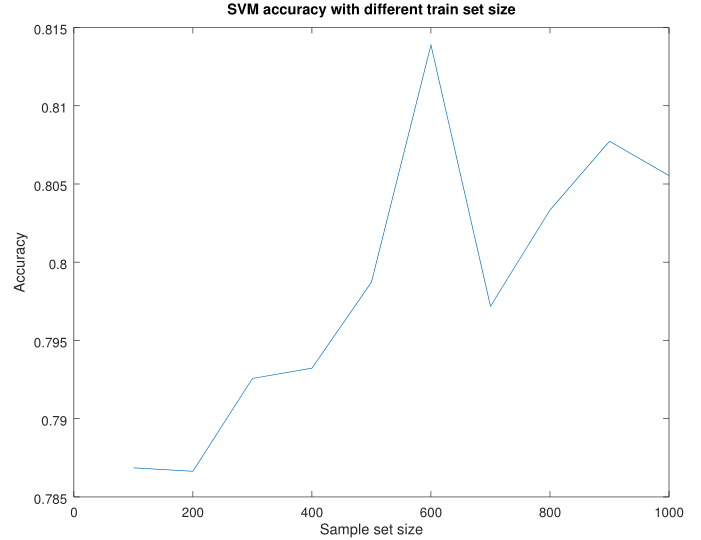


其中是样本根据核函数在高维空间中的映射。C是惩罚参数，即可以容忍多小的区分度，C也是在实验中那个需要重点调整的参数。

该约束式可以运用拉格朗日乘子进一步转化为优化问题，利用梯度下降等方法求解。

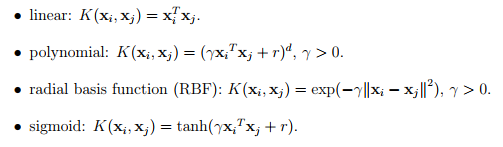
* + 1. SVM预测

由于样本数量较大，考虑每一类数据各抽取一部分进行训练，为了确定抽取样本的数量，以100为步长尝试100到1000规模的样本数量，利用缺省参数进行训练，观察预测准确率，得到下图。

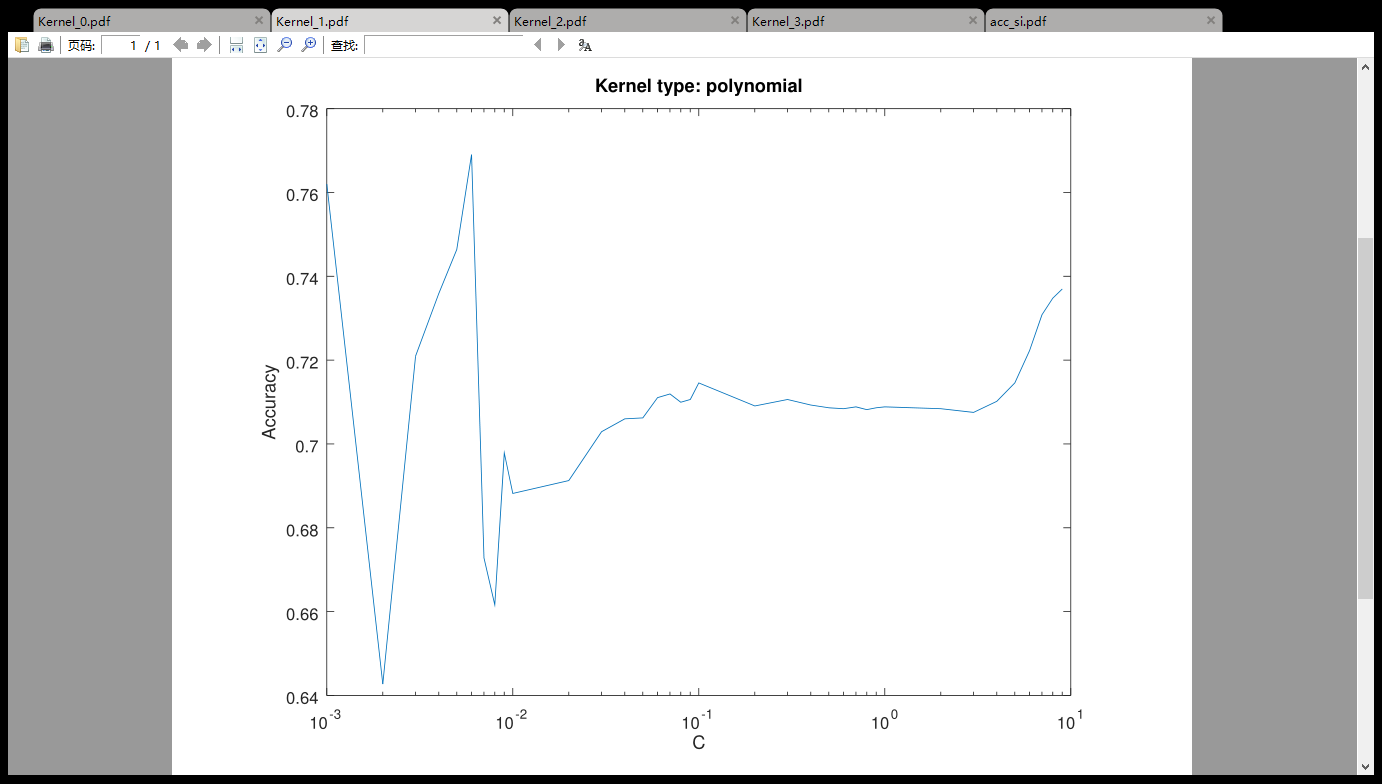
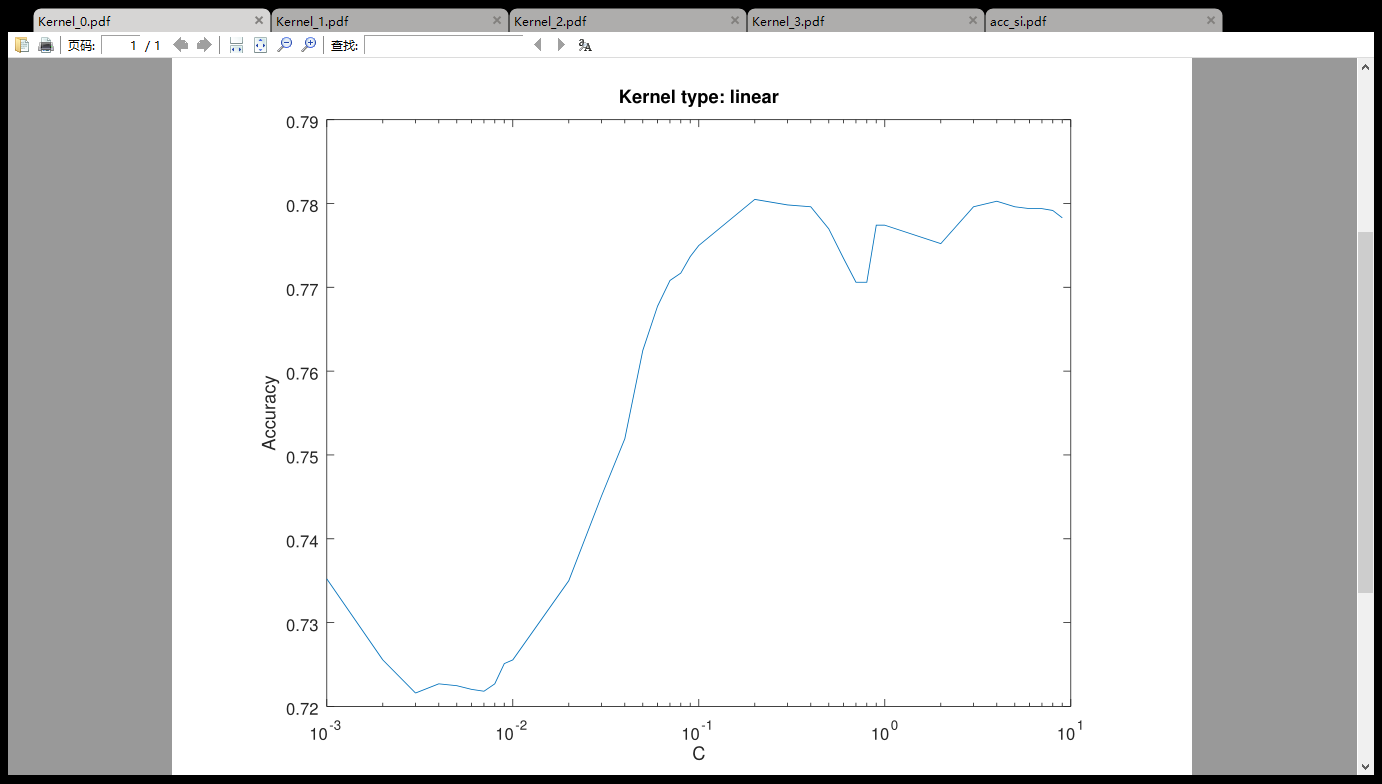


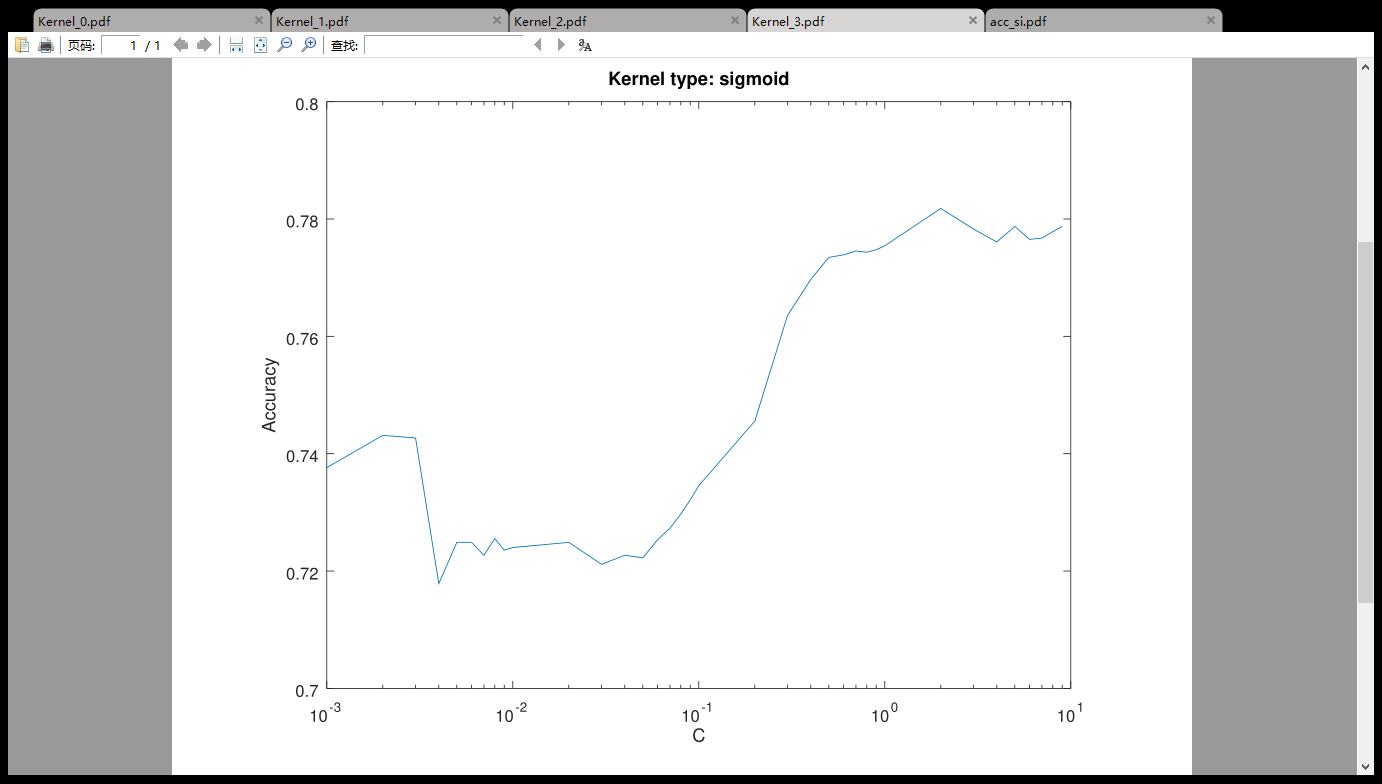
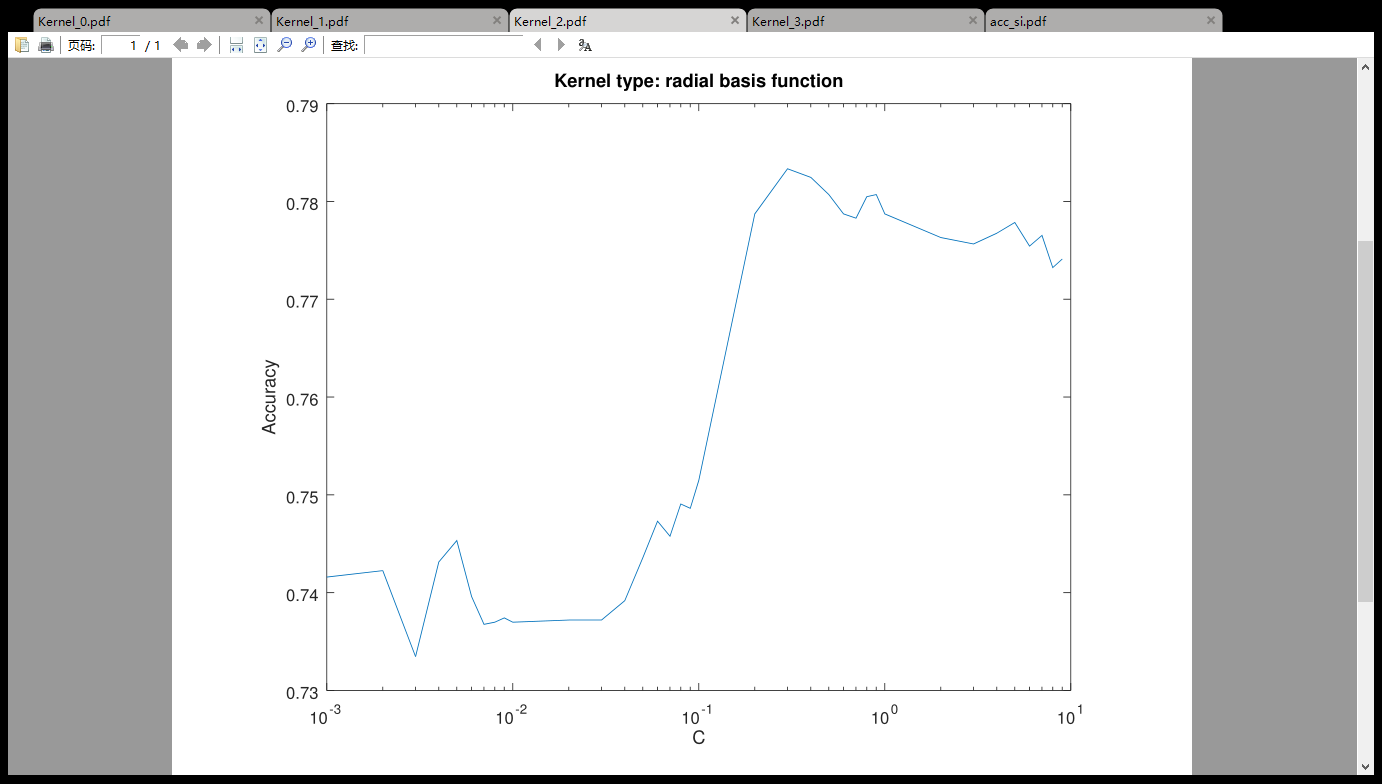
可以看出在样本数量为600时预测准确率最高，所以对总样本中每一类分别随机抽取600个作为最终的训练样本。

LIBSVM中，训练SVM的函数提供了4种核函数，分别为：



尝试不同核函数以及代价参数的取值组合，得到如下结果：

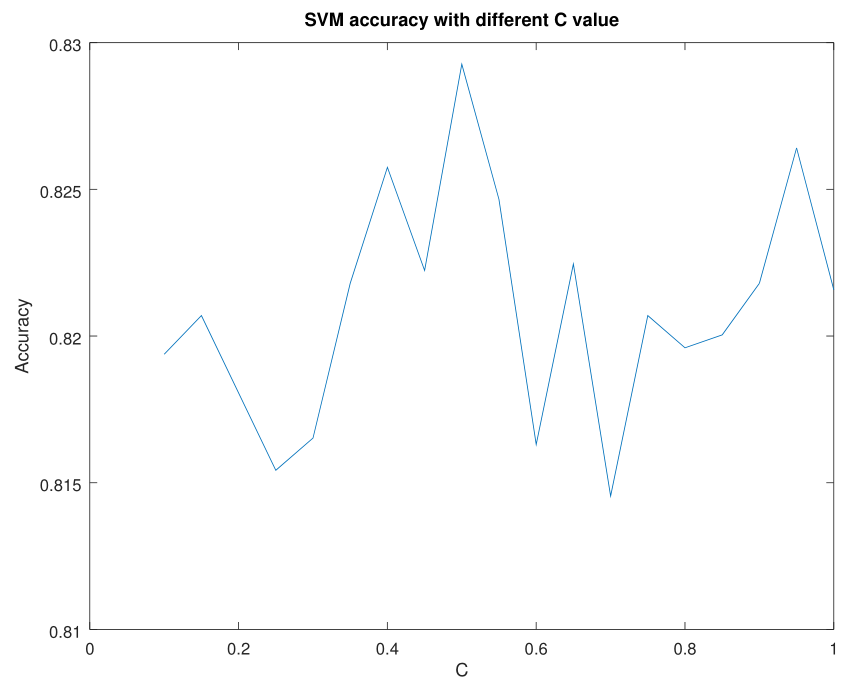




由上图可知RBF核函数的预测结果最优。

1. 实验结果

每类取出600个样本训练SVM，以0.05为步长尝试不同的C参数，对于每个参数尝试多次选取最好的预测结果。



当C参数为0.5的时候，预测准确率达到最高值82.158%。

1. 结果分析

观察实验数据可以发现，准确率的变化没有呈现出平稳的变化，而且选取不同的样本子集做训练的到的预测准确率也不尽相同，最多的时候相差可以达到5%。可能的原因有：

1. 数据集本身的区分度有限，导致预测结果不够准确。
2. 选取的样本太少。随机的数据在分布全体数据外围和内部的情况预测准确率会相差很多。
3. LIBSVM算法实现的稳定性。手册上提到，标签的格式等因素会影响计算结果的稳定性。具体原因有待深入了解LIBSVM实现细节。