**基于最小错误率Bayes分类器和Fisher分类器**

**实现手写数字识别**

xxxxx 控制科学与工程

**一、实验目的**

本次实验旨在学习模式识别中分类器的设计概念，利用Matlab设计基于最小错误率Bayes的手写数字识别平台，通过动手实践来对分类器有更深刻地认识，理解通过最小错误率Bayes的判别函数进行决策分类的原理。本文采用了基于主成分分析的模板匹配（PCA），以及基于最小错误率Bayes的方法进行手写数字的识别。

**二、编程环境**

本实验以windows 10为运行环境，以Matlab R2017a为平台进行设计。

**三、Bayes分类器实验原理**

1、设计思想

最小错误率贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。

2、实验原理

（1）贝叶斯公式

假设类别，模式样本，则其贝叶斯公式为



其中，表示后验概率，表示条件概率，表示先验概率。

（2）判别函数

对于二类分类的决策规则如下：



可见，贝叶斯最小错误率分类器设计的决策规则就相当于后验概率最大的决策规则，因此，多类贝叶斯的判别函数可写为



其中，为类的先验概率，为类的类概率密度函数。

经过代数运算和化简后，可得



其中d代表了样本个数，代表了数字i的样本特征向量的均值，代表数字i的样本特征矩阵的协方差矩阵。

**四、Fisher分类器实验原理**

1、设计思想

基于主成分分析的模板匹配（PCA）和Fisher线性判别（FLD）的方法对手写数字进行识别。利用每一类对应的特征线性向量唯一，且不同类之间对应的特征线性向量不同的两个原理，通过Fisher准则找出这些特征向量的最优投影方向，使之在投影方向上能够有最大间距。因此，可以先通过PCA降维，然后通过模板匹配法提取特征，再由FLD将降维后样本空间映射到一个一维判别特征空间进行分类。

2、实验原理

（1）线性判别函数

假设样本集为，属于类的样本子集为，属于类的样本子集为，则线性判别函数的一般表达形式为：



其中，为维特征向量，为权向量，为阈权值。

线性判别规则为



（2）基本参量

假设投影后的样本为，且满足，则在原来的样本空间中，类别均值为



其中是对应类中的样本个数，为样本特征向量。

各类内离散度矩阵和总类内离散度矩阵为





样本类间复杂度矩阵为



（3）Fisher准则函数

Fisher决策的出发点是：把所有的样本都投影到一维空间，使得在投影线上最易于分类 。而选择投影方向的准则是：使投影后两类相隔尽可能远，而对同一类的样本又尽可能聚集。

用以评价投影方向的函数为



将代入Fisher准则函数，得



（4）最佳投影方向和阈权值的求解

为基于Fisher准则将维空间投影到一维空间的最佳投影方向，满足



利用拉格朗日乘数法定理，有：



解得为一个维的向量，其分量值是对原维特征向量求加权和的权值。

而阈权值选用较常见的一种计算方法：



当和都确定后，即可在所属的一维空间内用线性判别规则进行分类。

**五、实验内容**

1、搭建平台

首先，利用Matlab中的GUI设计工具，创建可视化图形用户界面如图1，GUI设计文件为ui.m和ui.fig。实现通过滑动鼠标以拖动屏幕光标进行手写数字输入、训练样本归类、保存、训练样本集数量展示、样本识别和结果显示等功能。具体使用说明如下：

运行ui.m程序，进入GUI界面。

手写数字作为训练样本保存：在手写板区域拖动鼠标进行手写数字的书写，写完之后，数字图片会自动保存在当前目录下，在“归类选择”下拉菜单选择对应的数字标签，再点击其下方“保存”按钮即可将样本存入训练样本集；另外一种保存训练样本的放法是在数字识别之后，若识别结果正确，可以点击识别结果右侧的“保存”按钮，同样可以将样本存入训练样本集。

手写数字识别：手写板书写数字之后，在“分类算法选择”下拉菜单中选择分类算法（目前有贝叶斯二值分类、贝叶斯最小错误率分类、Fisher分类三种分类器），点击“识别”按钮，识别结果即会出现在下方结果展示区域。

训练样本集数量展示：GUI下方表格会展示当前手写数字自建样本库中各类数字样本集中样本的数量。每次点击“保存”按钮（两个都可以），只要保存成功，都会实时在表格中更新样本数量。



Figure 5-1手写数字识别图形化界面

2、特征提取

特征提取阶段是将数字字符形状转换成一组特征值的过程，其好坏决定着最终的识别效果。本实验选用基于PCA的模板匹配法。

（1）模板匹配法与数字图片特征提取

将手写的图像读入后，首先将其二值化并取反。随后为了提取特征，先将包含信息的部分提取出来，即找到每个手写样本的起始位置，以此为出发点在附近搜索该样本的宽度和高度。找到该图像矩阵中所有为0的点的坐标，取其最大最小值，截取两个最值之间的区域即是所需的有效图像矩阵。

将搜索到的每个有效图像矩阵的长度和宽度分为N等份，构成一个N×N的均匀小区域，即得模板，对每一小区域内的黑像素个数进行统计，再除以该区域的总面积，即得特征值。这样处理后，形状相同但大小不同的样本得到的特征值相差不大，因此能够将其视为同类。

可以推出，N值越大，模板越大，特征越多，区分不同物体的能力也越强，但同时计算量增加，运行等候的时间增长，所需的样品数也要成倍增加。而N值过小，不利于不同物体间的区分。因此本设计取N=5，即采用5\*5的模板，将手写数字的特征提取出来。具体的函数见程序GetFeature.m。

（2）PCA主成分分析

本设计对提取出的特征进行主成分分析，按照一定的贡献值，提取前m个主分量，将原空间变为m维空间，再用m维特征对样本进行分类。

选用Matlab软件中的自带函数：[PC, latent, tsquare]=pcacov(X) 通过协方差矩阵X进行主成分分析。其中，PC为主成分，latent为协方差矩阵X的特征值（是一个按从大到小顺序排列的列向量），而tsquare为每个特征值的方差贡献率。

经过测算，本文选用的累计比例为90%，即只要latent累积的值大于90%（对应的为前m个主成分），就认为降维后的m维空间样本能够很好地还原并表示原空间样本。本实验进行主成分分析处理后，得到的m=13，因而最后新的模式矢量pc为25\*13的数组。在选择了需要保留的特征矢量并组成了模式矢量后，进行转置，并将转置后的模式矢量左乘原始数据，得到新的数据。具体的函数见程序pcapro.m。

3、训练样本集的构建

本设计选用0~9九个数字的样本数为200左右不等，通过上述PCA+模板匹配的特征提取方法，将得到的新样本特征向量读入并保存，定义一个元胞数组template来保存0~9九个数字类的样本特征向量。这个元胞数组中包含一个1\*10维的结构数组，命名为pattern，结构数组中分为每个数字类的样本数num和每个数字类的特征向量的集合，即一个每一列都是一个25维特征向量的矩阵。将该元胞数组存储于一个mat中，每次需要更新或者使用时就将该元胞数组从template.mat中读出，完成处理后再将其重新存入template.mat中。

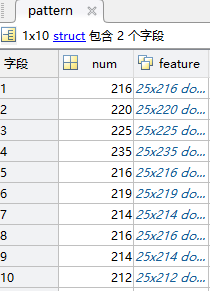


Figure 5-2 结构数组pattern

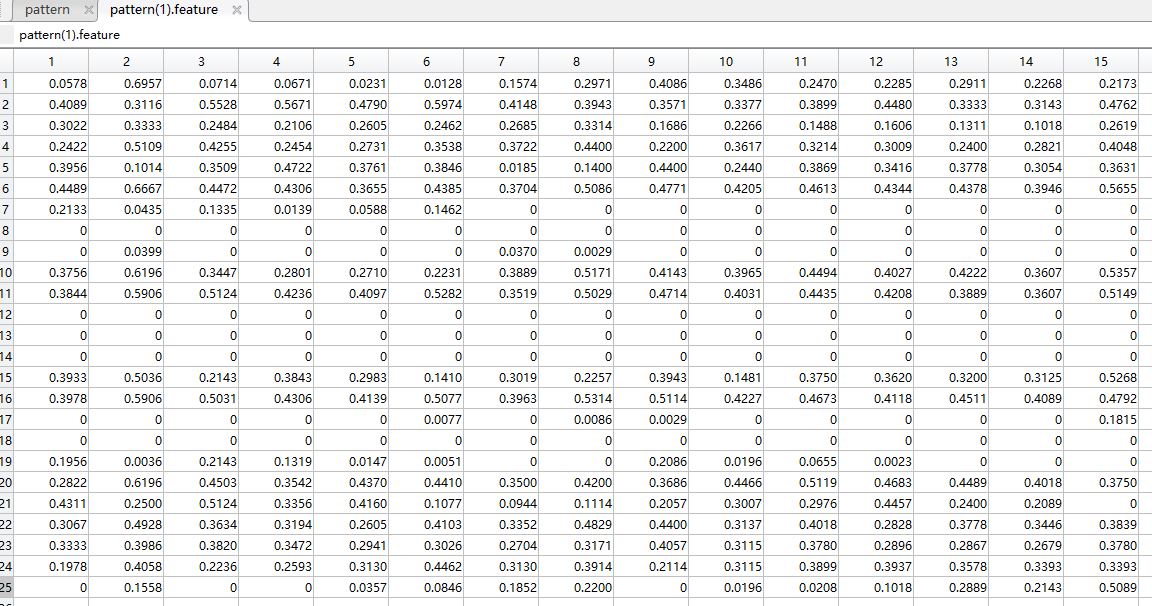


Figure 5-3 数字类1的像素向量集合

4、测试样本的构建

依据对手写数字进行特征提取的步骤，即经过主成分分析，并划分为5\*5的模板后，得到新的样本矩阵。然后统计每个区域内每一块黑像素块的数目，计算出每一个模板的黑像素率，从而得到25个处于[0，1]的值，存放于data矩阵中。该矩阵则作为测试样本输入给所设计的分类器。

5、分类器流程

打开GUI，初始化

手写输入数字

选择分类器进行识别

Figure 5-4分类器流程图

本次实验所设计分类器程序总流程如上图。总共实现了贝叶斯二值分类和贝叶斯最小错误率分类，以及Fisher分类。

**六、实验结果**

1、结果分析

首先，运行ui.fig打开图形用户界面，滑动鼠标写入数字。选择分类器和训练样本，并点击Identify按钮可进行识别，部分识别结果如图所示。

Figure 6-1 Bayes二值分类器部分识别结果

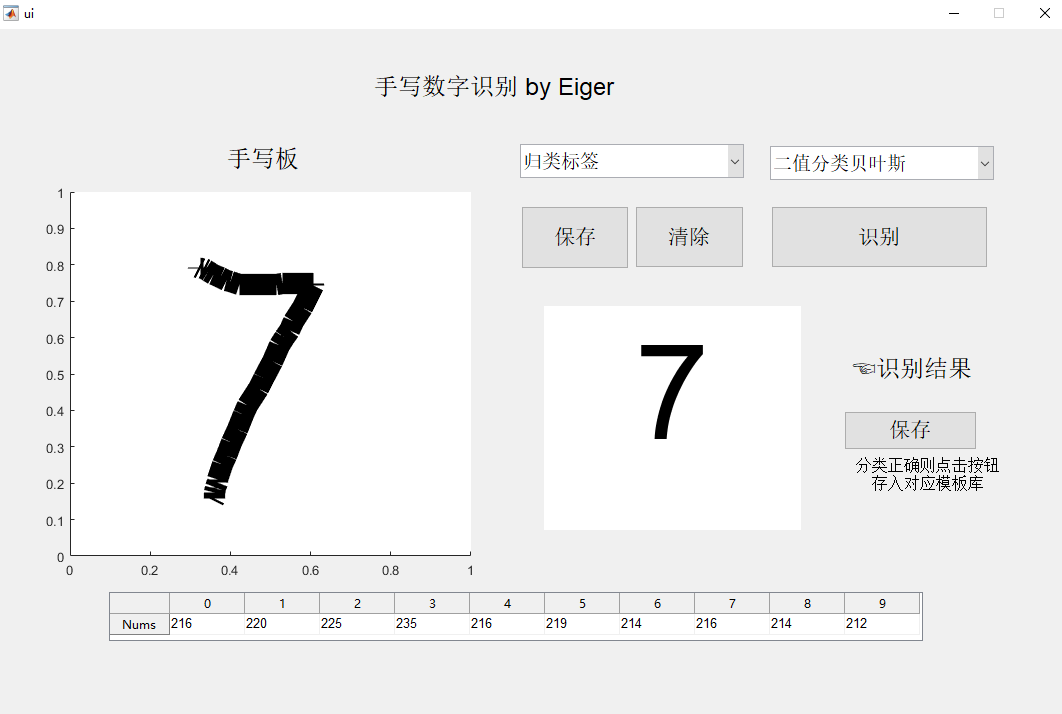
 

Figure 6-2 Bayes最小错误率分类器部分识别结果

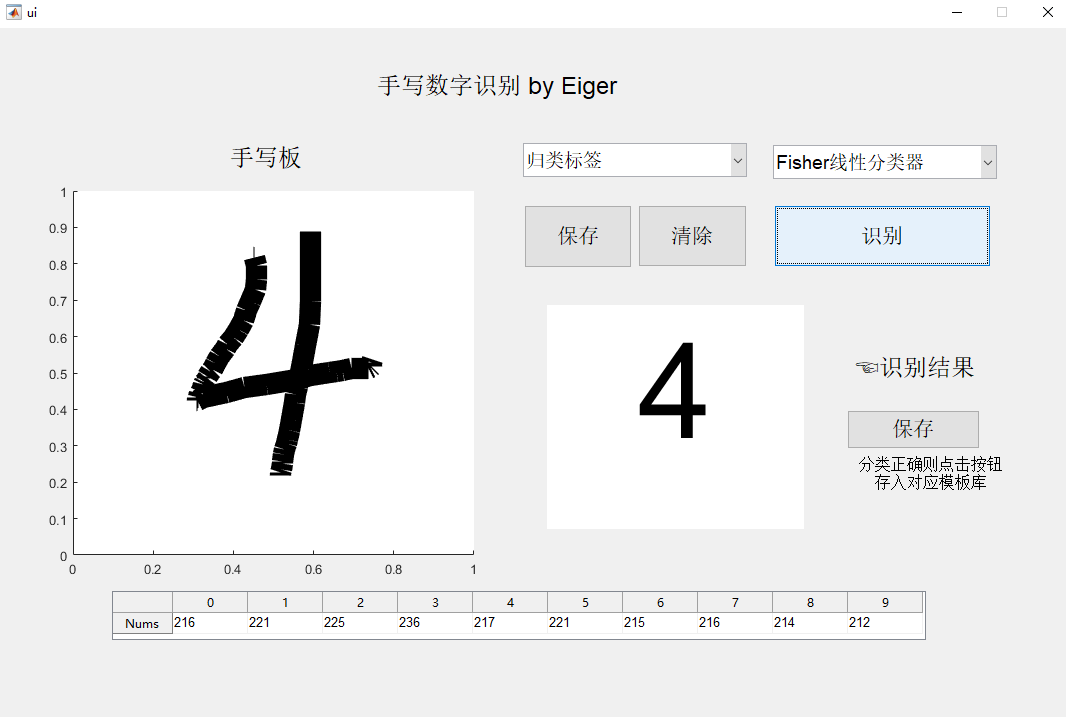
 

Figure 6-3 Fisher分类器部分识别结果

从使用体验来看，所设计手写数字平台操作简便，功能齐全。

从识别结果来看，在大量的测试下，分类器有较高的识别率，能正确识别数字0-9，具有一定的实用性。即使改变书写地位置和角度时，也能很好地识别。

三个分类相比较，贝叶斯最小错误率分类器和Fisher分类器准确率较高，但Fisher计算复杂度要远大于贝叶斯分类。因此认为贝叶斯分类器比较好用。

下面是三类分类器的分类准确性测试结果，测试程序依次为Accuracy1.m、Accuracy2.m、Accuracy3.m。

当样本总量为2192时，执行三个测试脚本，得到结果如下：

（1）贝叶斯二值分类算法为90.3285%，具体如下所示：

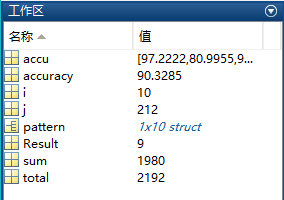


Figure 6-4 Bayes算法下总体准确率



Figure 6-5各类数字样本的准确率（1~10分别代表数字0~9）

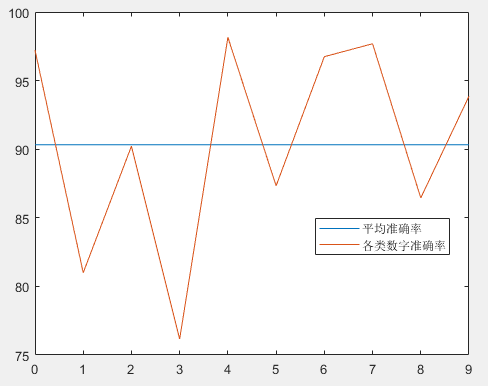


Figure 6-6 Bayes二值分类算法下的准确率折线图

（2）基于最小错误率的贝叶斯算法为97.1715%，具体如下所示：

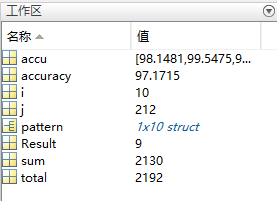


Figure 6-7 Bayes算法下总体准确率



Figure 6-8 各类数字样本的准确率（1~10分别代表数字0~9）

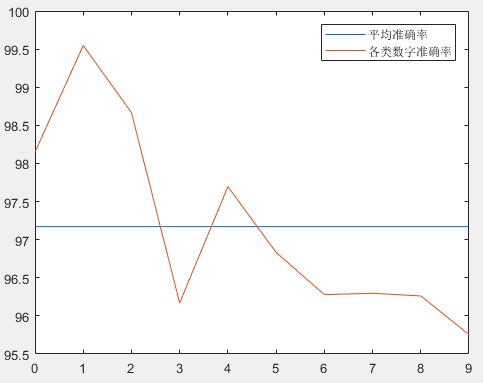


Figure 6-9 Bayes最小错误率分类算法下的准确率折线图

（3）Fisher算法为96.8124%(进行其准确率测试时，样本集已达到2196个)，具体如下所示：

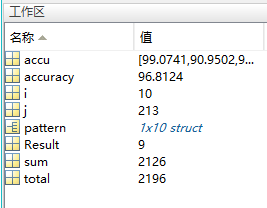


Figure 6-10总体准确率



Figure 6-11 各类数字样本的准确率（1~10分别代表数字0~9）

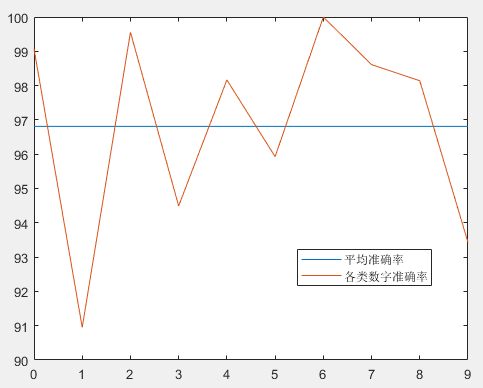


Figure 6-12准确率折线图

2、改进方案

为了提高手写数字识别的正确率，有如下几种改进方案：

1）增加模板样本的数目N

2）增加训练样本数量

3）改进分类器的算法或采用其他分类器，如非线性分类器