# 基于正态分布参数估计的贝叶斯决策算法实现手写数字识别

## 实验目的

使用正态分布参数估计的方法设计一贝叶斯决策分类器，实现手写数字的识别。

## 实验方法

### 搭建平台

使用Windows系统自带的画图工具分别构建训练样本集和测试样本集，使用MATLAB GUI编程环境对它们进行处理、识别。

### 特征描述

将手写部分从图像中分割，并转化为灰度图像，而后将其分割为6\*6=36个小图片，用每个小图片中黑色像素占总像素的比重作为该样本的特征，也就是一个36维的向量。

子函数get\_feature(image)代码如下：

function[feature] = get\_feature(image)

%从图像中分割出手写部分并转为灰度

image2 = pretreat(image);

A = mat2cell(image2,[10,10,10,10,10,10],[7,7,7,7,7,7]);

B = reshape(A,1,36);

for i=1:length(B)

temp = cell2mat(B(i));

%feature(i)表示黑像素点占一个小块的比例

feature(i) = sum(temp(:) ~= 255) / numel(temp);

end

%feature

end

### 关键算法

本次实验假定各个同类样本满足正态分布，然后通过最大似然估计的方法估算出正态分布概率密度函数的两个重要参数和。

#### 最大似然估计

下面对最大似然估计的方法进行简要概括：

最大似然原理假定样本数据是群体的一个代表，一旦观测数据给定，就仅仅是的函数，使得概率密度函数最大的是最可能的取值，它就是最大似然估计，记作。

由独立同分布假设，样本集D出现的概率，定义似然函数。

接下来，用最大似然法估计的最大似然估计，使似然函数最大。为了得到最大值，必须满足的条件是似然函数对的梯度为0，即：



根据该条件求出的就是。

#### 正态分布下的似然估计

以单变量正态分布情况下估计其均值和方差为例子说明最大似然估计的用法。

单变量正态分布的形式为



其中未知的参数即是我们要估计的参数，用于估计的样本是，由以上讨论，最大似然估计应该是以下方程的解：



其中

。

因此，最大似然估计是以下方程的解：



。

容易解得：



。

对于多元正态分布，分析原理与上面相同，结论也和单变量的情况很相似，即：



。

#### 代码实现

%正态分布参数估计求类条件概率密度函数pXwi

function[pXwi] = get\_pXwi(new\_feature)

load mean\_feature;

%mean\_feature放到主函数里执行

%到这里mean\_feature的每一行都是一个类别的mean\_feature

%也就是要估计的参数miu

%下面估计sigma

pXwi = [];

for i =0:9

miu = mean\_feature(i+1,:);

image\_list = dir(strcat('./样本',num2str(i),'/\*.png'));

sigma = zeros(length(miu),length(miu));

for j=1:length(image\_list)

feature = get\_feature(imread(strcat('./样本',num2str(i),'/',image\_list(j).name)));

sigma = sigma+(feature-miu)'\*(feature-miu);

end

%无法保证sigma非奇异，所以在对角线加一个数字，几乎不对结果产生影响

sigma = (sigma+10\*diag(ones(1,length(new\_feature))))/length(image\_list);

%代公式求类条件概率密度函数

p = ((2\*pi)^(-length(new\_feature)/2)) \* ((det(sigma))^(-1/2)) \* exp(-0.5\*((new\_feature-miu)\*inv(sigma)\*(new\_feature-miu)' ));

pXwi = [pXwi,p];

end

## 实验结果

### 训练样本集构建

使用Windows系统自带的画图软件创建训练样本集，使用10个文件夹分别保存10个类的训练集，如下图。



图 三-1 训练样本集

每一个样本都有一个固定的模板图片，用画图工具在该图片的指定位置手写数字，模板图片如下图。

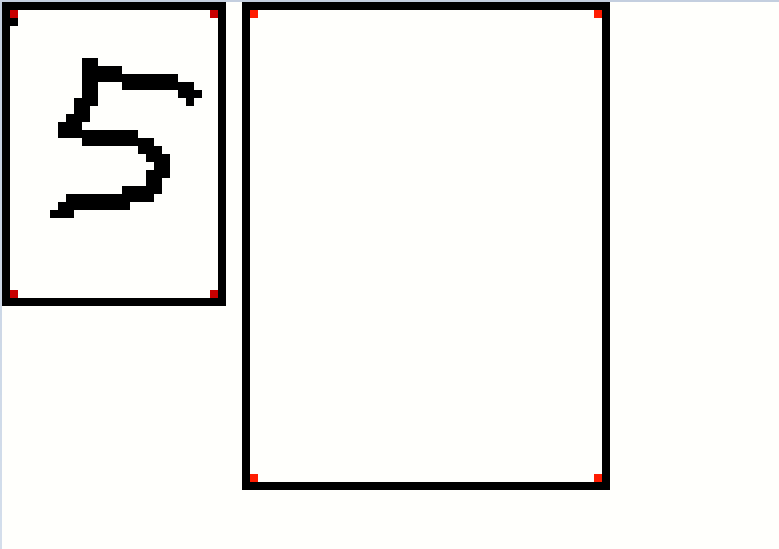


图 三-2 样本构建的模板图片

某一手写完成的样本如下图。

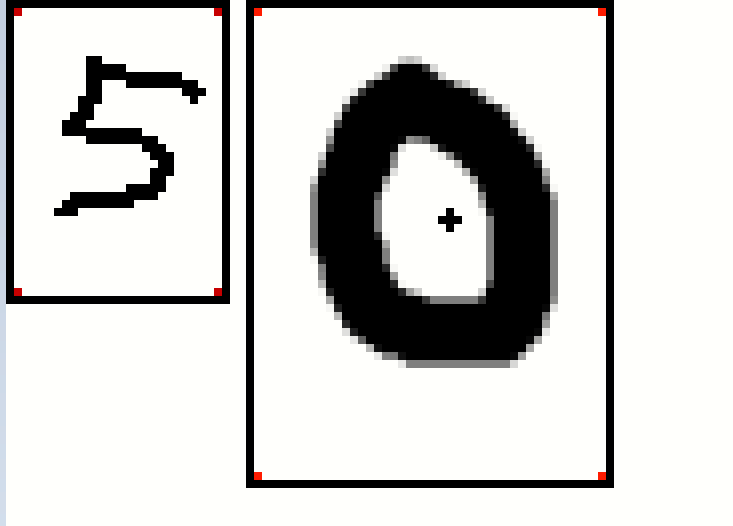


图 三-3 某一手写完成的样本

### 测试集构建

测试集样本的制作与训练集基本相同，不同的是，将所有图片都保存至同一文件夹“测试样本集”中，以供程序读取识别。

本次实验测试样本集如下图。

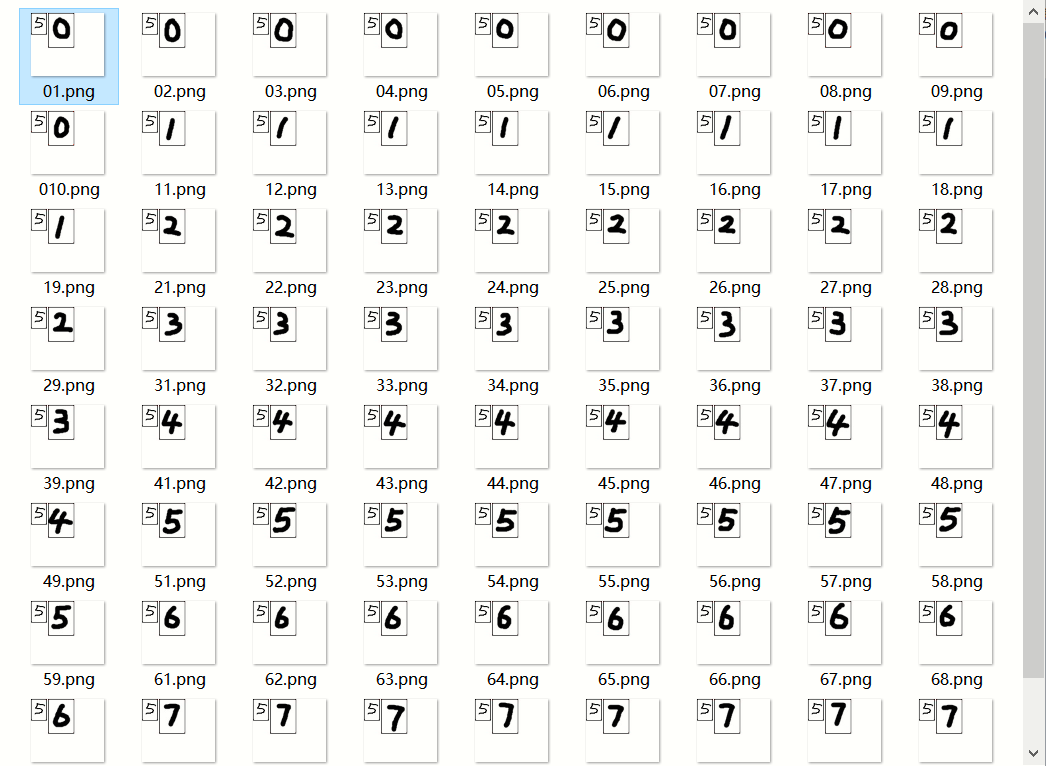


图 三-4 测试样本集

该样本集中有手写数字0-9各10个，它们的命名以各自的数字开头，用于实验中计算正确率。

### 实验结果

实验的运行环境是MATLAB GUI，本次实验各控件布局如下图。

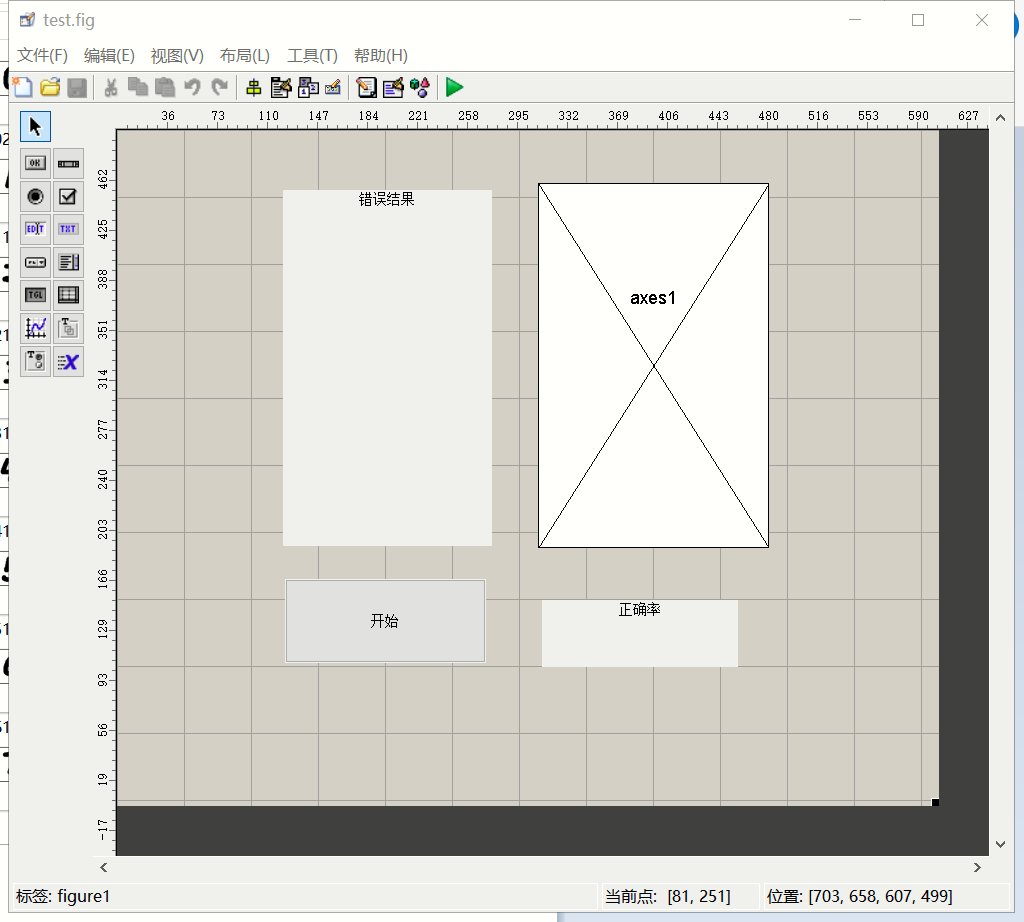


图 三-5 GUI控件的布局

图中已表明了各控件的作用，其中axes1用于运行时显示正在识别的样本。

程序运行效果如下图所示。



图 三-6 程序正在运行

程序运行结果如下图：

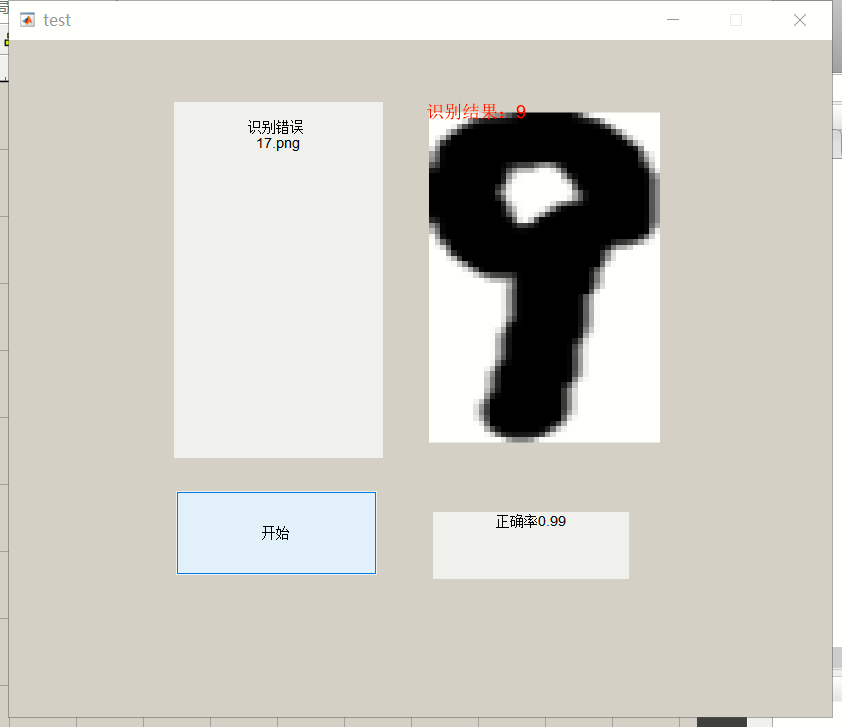


图 三-7 程序运行结束

### 性能量化评估

测试结果显示，对于测试样本集中的100个样本，识别的正确率达到了0.99，说明算法比较可靠。

## 结果讨论

运行结果显示，对数字1的样本“17.png”识别错误，在命令窗口中对其进行测试。

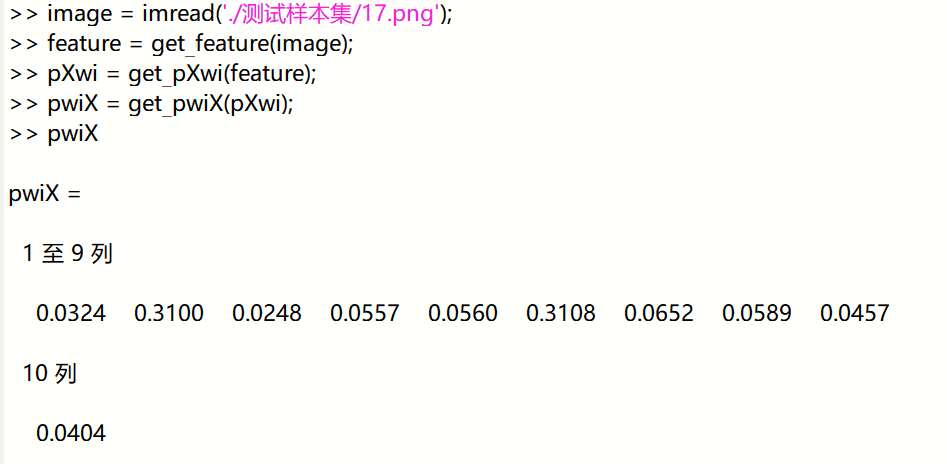


图 四-8 对测试样本集中的“17.png”测试

说明程序认为该样本是1的概率为0.3100，是5的概率为0.3108，差距很小，通过增加样本数可以提高正确率。

## 结论

本次实验使用MATLAB编程环境完成了基于正态分布参数估计的贝叶斯决策算法的编写，实现了手写数字的识别，通过对测试样本集（100个样本）的识别，结果显示正确率达99%，表示实验过程中各个样本集的建立以及所用算法较为可靠。