浙江大学毕业设计

本科学生中期报告

学生姓名	陈威志
实习单位/实验室	浙江大学
项目名称	基于 SMO 算法 SVM 的试卷信息统计
导师姓名	吴健
合作导师姓名	
项目开始日期	**2016.3.30**
项目结束日期	**2016.5.30**

填表日期：　2016 年　04 月　27 日

撰写提纲：（请以此结构，逐一论述）

一、 **项目概况**

1.1 项目简介

建立试卷信息统计系统：用户将学生完成答题，教师批阅后的试卷扫描或拍照图片传入系统中。系统对图像进行预处理，分块，归一化等操作，将每道题的老师判卷结果(手写对勾√和错误×)交给训练好的 SVM 分类器进行识别，然后将信息汇总给出试卷的统计信息。

1.2 理论意义与价值

试卷的数字化管理对提高教学质量有着十分重要的作用。借助数字化试卷管理系统,可以有效分析学生对知识点的掌握情况。因此,设计一个数字化的试卷管理系统具有重要的应用价值。而试卷信息统计又是其中必不可少的一个步，在教师完成了阅卷之后，如何快速，准确地统计出试卷中的信息就显得至关重要。并且统计出的信息（试卷的难易度，正确率等）也会为后续的出题工作提供反馈性的信息，是试题库系统中重要的环节。关于试卷的导入，试题库的建立，国内以有人在这方面做过相关的工作，但是关于学生完成考试和教师批阅后的试卷信息统计和反馈，相关的工作并不是很多。

1.3 工作与任务

1.3.1 机器学习部分

对处理好的符号信息进行识别。具体为：

对图像信息进行多种方法的特征向量的提取，并进行测试，最后选取一组，保证特征向量能很好的表现待识别字符的特征。

对特征向量进行归一化处理。

搜集不同人画出的不同样子对号和错号，对大规模数据进行上述工作，建立训练集和测试集。保证训练集的样本具有普遍性，能够得出健壮的模型。

尝试多种核函数，如线性、多项式、sigmod，rbf 等，最终确立 svm 内核。

对选取的 rbf 内核参数及惩罚项参数（γ，C）使用交叉验证和网格搜索，最终确定一个较优的参数。

1.3.2

基于 OCR 技术的文字信息提取，APP 应用开发，界面设计，用户个人数据库的建立。

二、 **工作成果及水平**

对于图像的获取、处理，以及机器学习的部分已经很好的完成，具体完成工作如下：

2.1 图像获取和识别

对不同光照下的图片都准确的提取出对号和错号的待识别部分。比较传统

的 RGB 颜色体系，我们采用 HSV 颜色体系将红色谱系范围提取出来，反复进行参数设置，确保提取信息的完整，以及不必要信息的过滤。通过灰度化，二值化，降噪，归一化等过程将图片信息转化为同一规格的二值矩阵。

2.2 图片特征向量的提取

因为手写字体的多样性，所以为了增加识别的精准度，需要强大的特征提取。模式特征对分类十分重要，手写字符识别的关键在于能否找到有效的特征。

2.2.1

在项目的初期，我采用的是现在识别手写数字中比较常用的 13 点网格特征提取：

a. 把字符水平分成四份,垂直分成两份,分别统计这 8 个区域的白像素的个数,得到 8 个特征。
b. 水平和垂直各划两条线把水平和垂直分割成三分, 统计这四条线穿过的白像素个数,得到 4 个特征。
c. 字符图像全部白像素数目作为 1 个特征, 总共得到 13 个特征。

最后发现对 200 个训练集的训练，交叉验证后的结果为 93%左右。我认可以对其进行改善。于是有了现在的新的特征提取方法：

a. 笔画密度特征：对 50*50 的样本每隔 10 行扫描一次，统计像素 个数提取一个特征，行 9 个，列 9 个，共 18 个。
b. 投影特征，将一个字符点阵划分为 4 个像素区域，共 12 个边界，对每个点向所在象限的 4 个边界投影，12 条边界上的投影长度作为 12 个特征。
c. 重心及重心距特征，首先求出重心的二维特征，再求 4 个象限关于重心偏离的重心距二维特征。
d. 傅里叶变换特征，首先对图像进行二维离散傅立叶变换，然后将直流分量移到频谱中心，取取傅立叶变换的实部于虚部，计算频谱幅值，归一化后频率矩阵的大幅值集中在图像的中心，能反应原图的轮廓信息，提取 48 个特征。
e. 字符轮廓特征，包括字符有效宽度，有效高度，宽高比等 14 个特征。

2.3 基于 SMO 的 SVM 算法实现

实现了基于 SMO 算法(序列最小化优化算法)。先前的 SVM 算法必须求解复杂的二次规划问题，并且使用昂贵的二次规划问题工具。但是 SMO 算法较好地避免了这一问题。

在之前的工作中实现了 SMO 算法，使 SVM 训练的效率大幅提升。

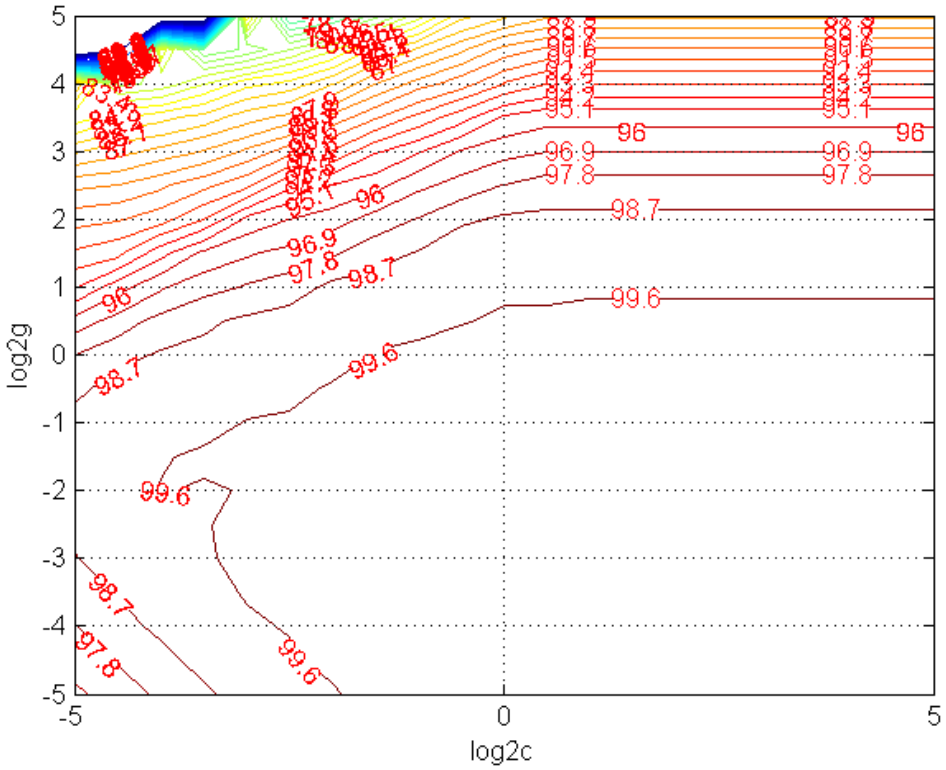
2.4 核函数的确定和最优参数选取

通过比较不同内核的训练效果，最终确定了使用 RBF 内核，$K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right), \gamma > 0$.RBF 内核非线性地映射样本进入一个更高维空间，因此不同于线性内核，可以处理那些分类标签和属性不是线性关系（线性不可分）的情形。

内核参数与惩罚项（γ，C）的确定，我采用的是网格搜索以及交叉验证的方法。思想是让（γ，C）在一定范围内跑，如 c = 2^(-5),2^(-4),...,2^(5),g =

2^(-5),2^(-4),...,2^(5)，然后用交叉验证的方法找到准确率最高的（γ，C）。然后不停细化范围，直至找到最佳的（γ，C）。

我收集了多个同学的手写对号和错号样本，保证样本具有普遍性。通过对1835 个样本进行测试，fold = 5 的交叉验证的结果为 99.9455%（即误分点只有一个）。对于错号和对号能够进行完美的分隔。最终确定的参数为 bestg = 0.3536，bestc = 0.7071.



网格搜索

应用新的特征提取方法，以及基于 RBF 内核优化参数后的 SVM 分类器可以做到，对于 104 个测试集进行测试的结果为，识别率 100%，据识率 0%，误识率 0%。这个结果是很优秀的，目前的研究以及相关的学术论文方面，在手写字符或数字的识别上鲜有做到如此准确的识别的。

三、 项目收获

在前一半的项目工作中，我的收获很大，对项目有了一定的理解和认识。对于监督学习的方法有了理解和掌握。在项目过程中也遇到了一些难题，锻炼了我独立思考解决问题的能力。在团队合作方面和同组的同学认真讨论，共同商量方案，锻炼了我的团队合作能力。

四、 对工作建议

对于个人今后的工作有以下几点方面：

1. 基于 ORC 的文字识别。
2. 试卷理解与试卷版面分隔。
3. APP 应用开发及界面设计。

4. 试卷信息统计，及个人数据库的建立。

五、 **其它**