



**课 程 报 告**

|  |  |
| --- | --- |
| 课程名称 | 模式识别 |
| 题目名称 | 基于人脸图像的性别与年龄识别 |
| 专业班级 | 2017级自动化创新班 |
| 学号姓名 | 江熠铧 3117000995 |
|  | 蔡瑞涵 3117000985 |
|  | 谭美健 3217000 |
| 指导教师 | 邢延 |

2020年06月 11日

**目录**

[**1 模式识别系统设计 1**](#_Toc41594369)

[**2 数据预处理 1**](#_Toc41594370)

[**2.1 标签转化和删去无效值 1**](#_Toc41594371)

[**2.2 数据标准化和数据平衡 1**](#_Toc41594372)

[**2.3 获取原始图像数据 2**](#_Toc41594373)

[**3 支持向量机算法 2**](#_Toc41594374)

[**3.1 经典算法部分 2**](#_Toc41594375)

[**3.11 支持向量机算法原理与相关概念 2**](#_Toc41594376)

[**3.12 参数设定与调整 4**](#_Toc41594377)

[**3.13 实验结果分析与比较 5**](#_Toc41594378)

[**3.14 经典算法结果总结 5**](#_Toc41594379)

[**3.2 改进算法部分 6**](#_Toc41594380)

[**3.21 卷积神经网络算法原理与相关概念 6**](#_Toc41594381)

[**3.22 CNN提取特征SVM训练 6**](#_Toc41594382)

[**3.23 参数设定与调整 6**](#_Toc41594383)

[**3.24 实验结果分析与比较 6**](#_Toc41594384)

[**3.25 改进算法总结 6**](#_Toc41594385)

[**3.3 传统算法与改进算法比较 6**](#_Toc41594386)

[**3 B算法 7**](#_Toc41594387)

[**3.1 基本原理 7**](#_Toc41594388)

[***3.1.1 子标题* 7**](#_Toc41594389)

[**4 C算法 7**](#_Toc41594390)

[**5 结论 7**](#_Toc41594391)

[**参考文献 7**](#_Toc41594392)

[**附录 9**](#_Toc41594393)

# 1 模式识别系统设计

内容要求：

1. 说明识别的目标和类别，例如：识别目标：性别，类别：男和女两类、
2. 数据来源及数据特点（数据量、特征维度、类别是否均衡、数据量是否足够等）
3. 采用的开发工具（编程语言、集成开发环境等、要求用Python）
4. 采用的模式识别方法（注意要与数据特点相匹配）
5. 是否做特征提取与选取，若是，采用什么算法（经典、改进）
6. 采用的分类算法（含经典算法、智能算法/改进算法）
7. 分类器性能评价方法（K折交叉验证，分类准确率/错分率/AUC,是否考虑算法的时间复杂度等）

# 2 数据预处理

1. **标签转化和删去无效值**
2. **数据标准化和数据平衡**

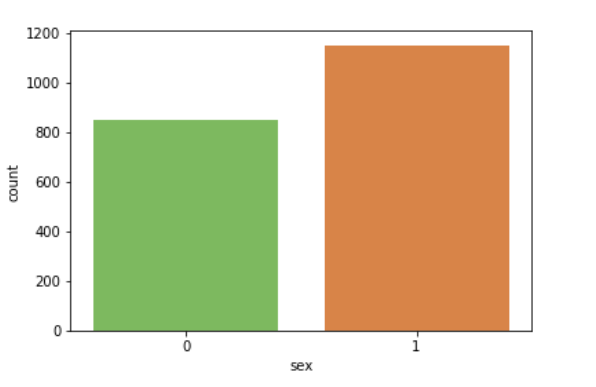


图2-3 数据性别分布

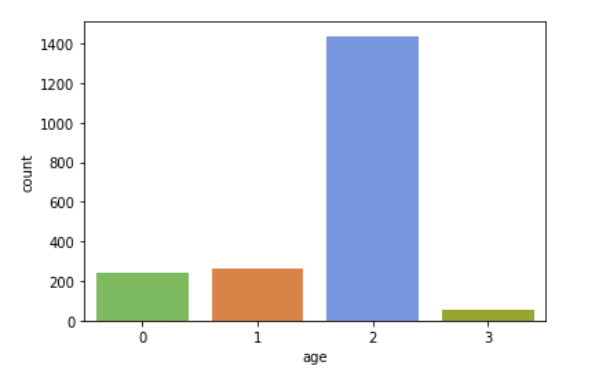


图2-3 数据年龄分布

1. **获取原始图像数据**

# 3 支持向量机算法

1. **经典算法部分**
2. **支持向量机算法原理与相关概念**

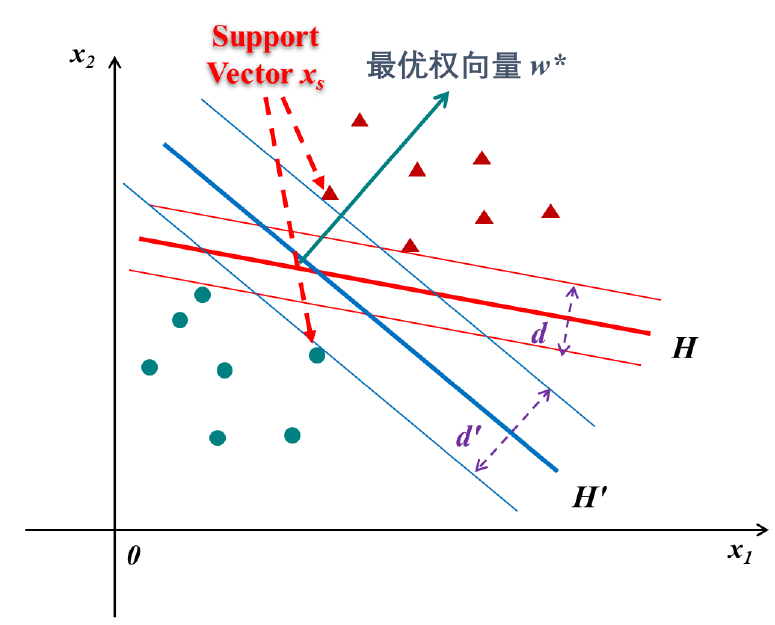


图3-1 SVM基本原理

* **基本原理**

支持向量机是一种二分类模型，基本模型（如上图）是定义在特征空间上的间隔最大的线性分类器。因此，SVM的学习策略就是**间隔最大化**，学习算法就是求解**凸二次规划**的最优化算法。

* **线性可分支持向量机与硬间隔最大化**

1. 线性可分支持向量机：给定线性可分训练数据集，通过间隔最大化或等价地求解相应的凸二次规划问题学习得到的**分离超平面**为

以及相应的**分类决策函数**

称为线性可分支持向量机。

1. 函数间隔和几何间隔

函数间隔即训练集T中所有样本点距离超平面的**距离**

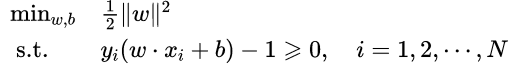
几何间隔，在对于超平面发现量加上某些约束时，如规范化后，，使得间隔是确定的，这时的函数间隔就是几何间隔，常表示为：

1. 最大间隔法

输入：线性可分训练数据集

输出：最大间隔分离超平面和分类决策函数

1. 构造并求解约束最优化问题：



求解最优

1. 得到超平面和决策函数：

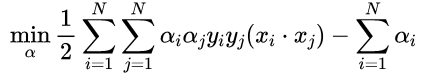
由于最大间隔分离超平面存在**唯一性**，因而能够找到使得两类点完正确分开的分离超平面。

1. 拉格朗日乘子法

对优化问题中，每一个不等式约束引入**拉格朗日乘子**得到拉格朗日函数及拉格朗日乘子向量。进而根据其对偶性，**求解极大极小问题**：



得到



因而可以利用上述约束函数求解最优值，进而求得分离平面和决策函数

* **线性不可分问题—软间隔支持向量机**

对于线性不可分训练数据，显然上述的线性可分支持向量机是不适用的。为了解决这一类情况，将硬间隔最大化改为**软间隔最大化**。

线性不可分意味着某些样本点不满足函数间隔大于等于1的约束条件，修改约束条件引入松弛变量，得到**约束条件**为：

**目标函数**变为：(C>0惩罚参数)

* **线性不可分问题—非线性支持向量机**

对于解线性分类问题，线性分类支持向量机有很好的性能，但是实际情况中许多分类问题都是非线性的。这时可以利用核技巧，实现低维空间到高维空间的映射计算，进而得到非线性支持向量机。

核函数包括：

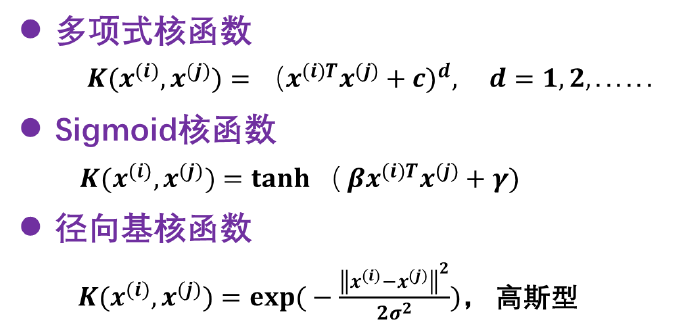


图3-2 SVM核函数

1. **参数设定与调整**
2. 训练器参数说明及选择

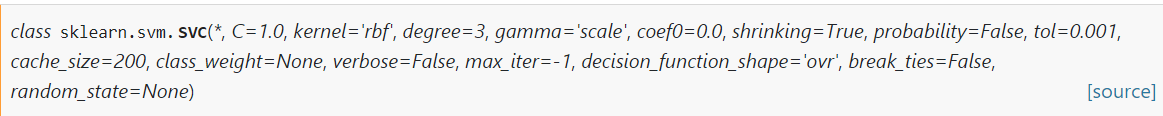
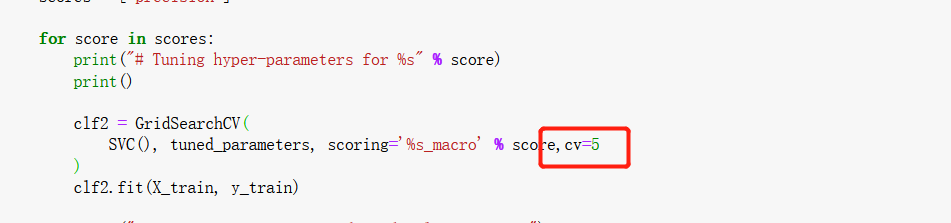


图3-3 SVC函数参数

|  |  |
| --- | --- |
| 参数 | 说明及参数的选择 |
| C | 惩罚因子，采用L2正则化 |
| 选用0.1，1，10三个不同量级的参数去训练，以观察不同正则化程度对分类器的影响 |
| kernel | 指定算法的内核类型，可选择‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’} |
| 结合本次的数据要求，选择linear’, ‘rbf’进行训练 |
| Probability | 是否启用概率估计，这个在内部会采用5折交叉验证 |
| 为更好的寻求一个最优分类器，此次选择False和True |
| verbose | 启用详细输出 |
| 选择true，观察训练过程的输出 |
| class\_weight | 调整类别权重 |
| 可选用，但因为本次数据已经经过SMOTE算法进行平衡这里就保持默认值 |
| 其他未调整的参数保持默认值 | |

表3-1 参数选择与说明

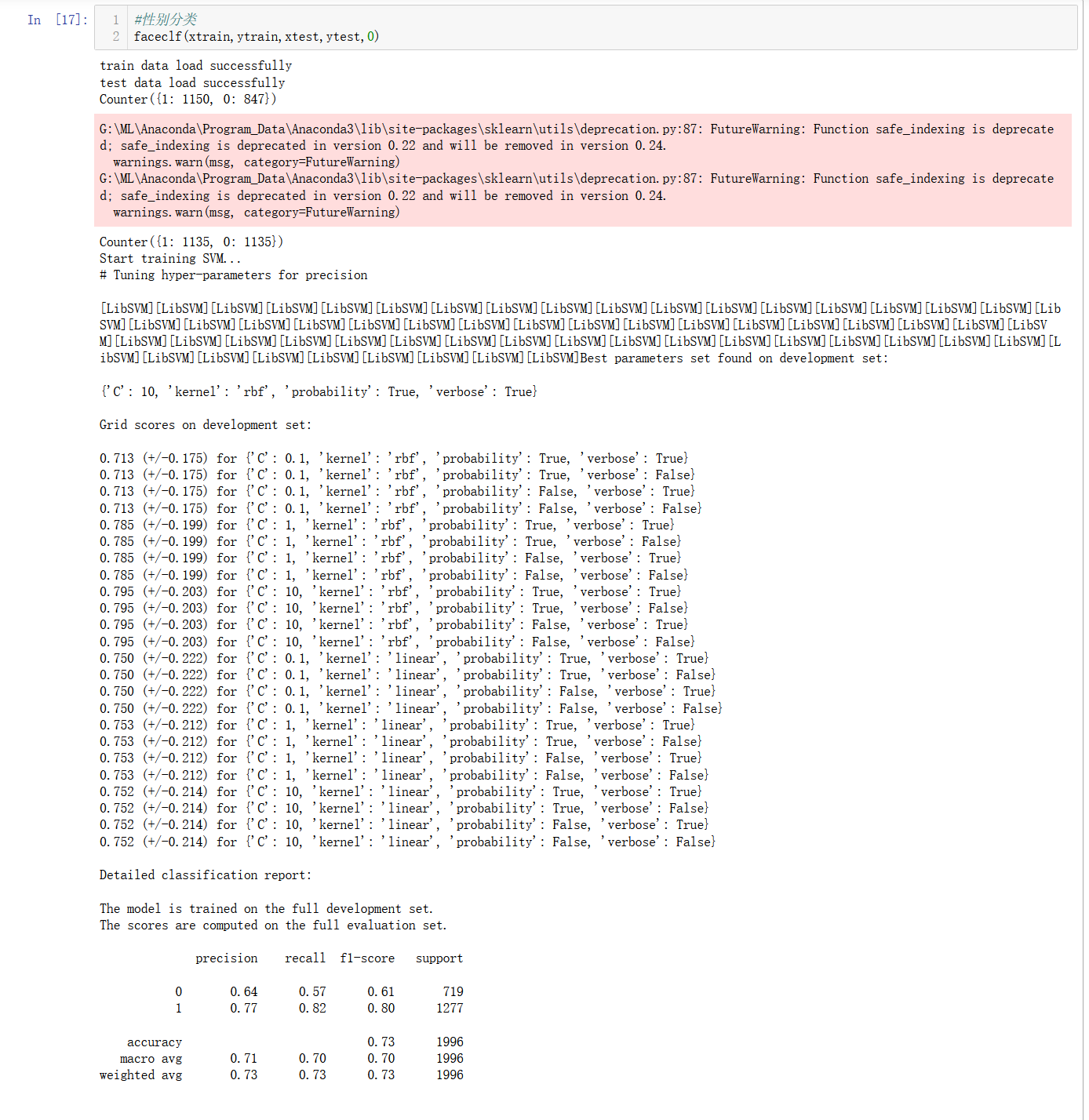
1. 交叉验证

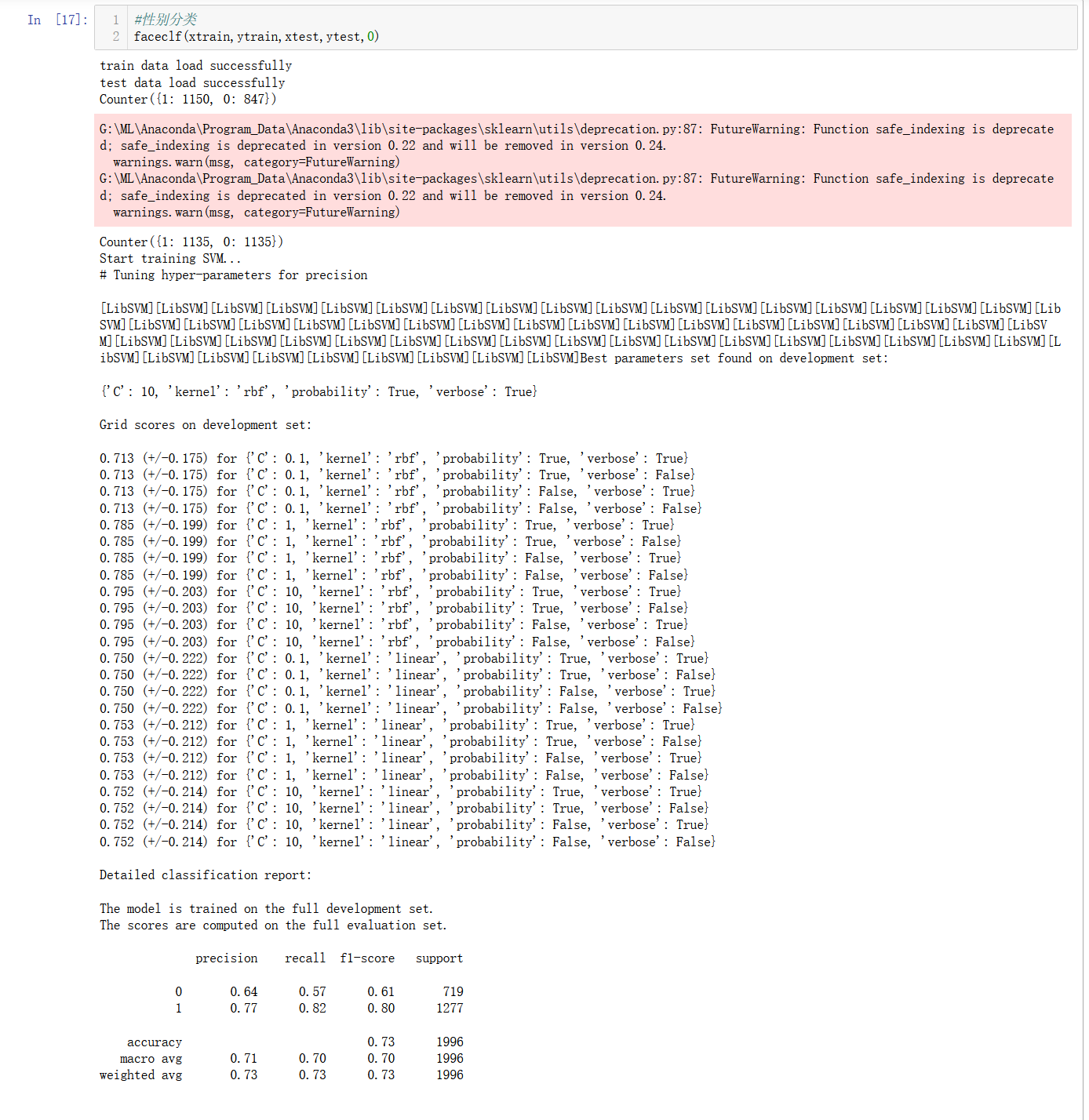


如图所示，本次实验采用5折交叉验证

1. **实验结果分析与比较**

1.性别结果：





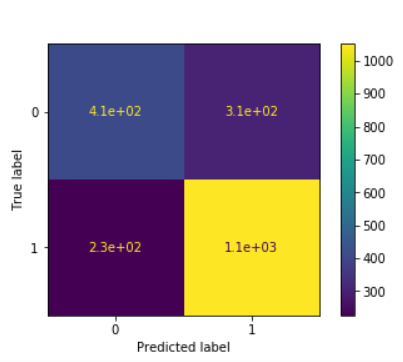


图3-4 Confusion matrix-sex

2.年龄结果：



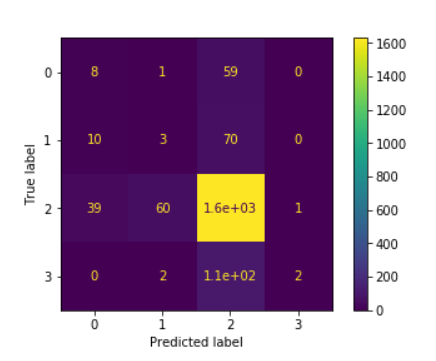


图3-4 Confusion matrix-age

3.结果分析与比较

从两次的寻找最优参数的过程中，可以发现实际上'probability'这个参数的变化对于训练集的准确率没有产生太大的影响，主要影响到准确率的是惩罚因子C和核函数kernel的选择。

从kernel的调节来看，对于两组数据样本的进行训练，当kernel=’linear’时，改变惩罚因子C并没有使得分类器对于训练样本有比较大程度的性能变化。反观，当kernel=’linear’时，改变惩罚因子C，可以明显的看到分类器的性能有明显的变化。

从惩罚因子C来看，有着越大准确率越高的趋势，也符合正则化减小过拟合的理论依据。但考虑到正则化的性质，当太大的时候可能会导致出现欠拟合，此次实验采用了三个量级的参数较为合理。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Prediction | | | | |
|  | Precision | Recall | F1-score | support |
| 0-male | 0.64 | 0.57 | 0.61 | 719 |
| 1-female | 0.77 | 0.82 | 0.80 | 1277 |
| accuracy |  |  | 0.73 | 1996 |
| Marco avg | 0.71 | 0.71 | 0.70 | 1996 |
| Best parameters | | | | |
| {'C': 10, 'kernel': 'rbf', 'probability': True, 'verbose': True} | | | | 0.795 (+/-0.203) |

表3-2 性别分类结果

从上表数据我们可以看到分类器对于测试集样本的准确率虽相比于训练时有一定程度的降低，但也比较符合实际情况。从具体的数据来看，两个类别的查准率，查全率和F1值存在一定的差距。但是结合我们实际用到的数据来看，原来的数据样本的分布存在一定的差异（0:847,1:1150），在采用SMOTE算法对数据进行平衡后的数据在一定程度上和实际数据还是有差异的，所以导致了这两部分

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Prediction | | | | |
|  | Precision | Recall | F1-score | support |
| 0 | 0.14 | 0.12 | 0.13 | 68 |
| 1 | 0.05 | 0.04 | 0.04 | 83 |
| 2 | 0.87 | 0.94 | 0.91 | 1730 |
| 3 | 0.67 | 0.02 | 0.03 | 115 |
| accuracy |  |  | 0.82 | 1996 |
| Marco avg | 0.43 | 0.28 | 0.28 | 1996 |
| weighted avg | 0.82 | 0.82 | 0.79 | 1996 |
| Best parameters | | | | |
| {'C': 10, 'kernel': 'rbf', 'probability': True, 'verbose': True} | | | | 0.976 (+/-0.071) |

表3-3 年龄分类结果

1. **经典算法结果总结**
2. **改进算法部分**
3. **卷积神经网络算法原理与相关概念**

卷积层

1. 局部感知
2. 空间排列
3. 参数共享

池化层

全连接层

Dropout层

1. **CNN提取特征SVM训练**
2. **参数设定与调整**

**CNN参数选择**

#### SVM参数选择

1. **实验结果分析与比较**

1.性别结果：

2.年龄结果：

3.结果分析与比较

1. **改进算法总结**
2. **传统算法与改进算法比较**
3. 经典算法的原理、参数设定与调整、实验结果的分析比较、结论
4. 智能算法/改进算法的原理、参数设定与调整、实验结果的分析比较（必须要有与经典算法的结果对比）、结论

eijkel2

图 2-1 曲线图.

# 3 B算法

另一类算法[[1](#_ENREF_1)].

## 3.1 基本原理

原理

表 3.1-1 衡量指标

| 名称 | 含义 |
| --- | --- |
| F1 |  |
| F2 |  |

继续。

### *3.1.1 子标题*

公式

(3.1.1-1)

继续。

# 4 C算法

# 5 结论

内容要求：

1. 做了什么，取得了什么结论（同种算法之间？不同种算法之间？）
2. 需要改进的方面

# 参考文献

1. Cunha, J.C., O.F. Rana, and P.D. Medeiros, *Future Trends in Distributed Applications and Problem-solving Environments.* 2005.

# 附录

表 A1 成员分工及贡献度自评

| 姓名 | 个人分工 | 贡献度 |
| --- | --- | --- |
| 张三 | 具体内容，课程报告第1、2节 | 60% |
| 李四 | 具体内容，课程报告第3节 | 40% |
|  |  |  |
|  |  |  |

课程体会与建议

具体内容