
華中科技大學

課程實驗報告

題目： 基於朴素貝葉斯分類器的語音性別識別

課程名稱： 機器學習

專業班級： CS1701

學 號： U201714501

姓 名： 熊逸欽

指導教師： 李玉華

報告日期： 2020 年 7 月 26 日

計算機科學與技術學院

目 录

| | | |
|----------|-------------------|-----------|
| 1 | 项目目的 | 1 |
| 2 | 问题分析 | 2 |
| 2.1 | 项目选题及背景..... | 2 |
| 2.2 | 数据集特点 | 2 |
| 2.3 | 任务要求 | 3 |
| 3 | 设计与分析..... | 5 |
| 3.1 | 朴素贝叶斯分类原理..... | 5 |
| 3.2 | 项目建模 | 6 |
| 3.3 | 设计思路与流程图..... | 7 |
| 3.4 | 系统改进和优化..... | 9 |
| 4 | 结果分析 | 12 |
| 4.1 | 实际运行结果..... | 12 |
| 4.2 | 优化前后性能对比..... | 13 |
| 4.3 | 分析影响性能的因素 | 15 |
| 5 | 思考与总结..... | 17 |
| 5.1 | 项目小结 | 17 |
| 5.2 | 心得收获 | 17 |

1 项目目的

通过本次项目开发，希望能够达到以下目的：

- （1）掌握建立机器学习模型的方法，掌握朴素贝叶斯算法的推导和使用，掌握使用 Python 语言实现算法的能力。
- （2）理解贝叶斯方法的基本原理和基本概念，理解机器学习的技术路线，理解算法优化的意义和重要性。
- （3）提高使用机器学习经典算法和思想解决实际问题的能力。
- （4）锻炼和培养独立思考的精神以及创新意识。

机器学习课程是人工智能方向最核心的课程，学习这门课程的重要性不言而喻。作为结课项目，希望在这次项目开发的实践过程中能够锻炼自己，提升自己，努力达成和实现上述目的。

2 问题分析

2.1 项目选题及背景

本次项目选择的题目是基于朴素贝叶斯分类器的语音性别识别。本题要求对 Kaggle 平台上提供的语音数据集（项目名称为 Voice Gender）进行性别识别，区分声音的来源为男性还是女性。

在这一背景下，设计一个基于朴素贝叶斯的分类器，使得能够根据声音的特征对声音来源的性别进行分类。

2.2 数据集特点

数据集以 csv 格式文件的形式呈现，文件名为“voice.csv”，大小为 1.02MB，其中包含从男性和女性说话者那里收集的 3168 个录制的语音样本信息。根据该题目在 Kaggle 上的介绍可知，数据集中的语音样本经过预处理，分析的频率范围为 0hz 至 280hz。

打开“voice.csv”文件后其内容如下图所示：

| # | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U |
|----|----------|------|--------|------|------|------|-------|-------|--------|-------|-------|----------|----------|---------|----------|----------|---------|----------|---------|---------|-------|
| 1 | meanfreq | sd | median | Q25 | Q75 | IQR | skew | kurt | sp.ent | stm | mode | centroid | meanfun | minfun | maxfun | meandom | mindom | maxdom | dfrange | modindx | label |
| 2 | 0.059781 | 0.06 | 0.032 | 0.02 | 0.09 | 0.08 | 12.86 | 274.4 | 0.8934 | 0.492 | 0 | 0.059781 | 0.084279 | 0.0157 | 0.275862 | 0.007813 | 0.00781 | 0.007813 | 0 | 0 | male |
| 3 | 0.066009 | 0.07 | 0.0402 | 0.02 | 0.09 | 0.07 | 22.42 | 634.6 | 0.8922 | 0.514 | 0 | 0.066009 | 0.107937 | 0.01583 | 0.25 | 0.009014 | 0.00781 | 0.054688 | 0.0469 | 0.05263 | male |
| 4 | 0.077316 | 0.08 | 0.0367 | 0.01 | 0.13 | 0.12 | 30.76 | 1025 | 0.8464 | 0.479 | 0 | 0.077316 | 0.098706 | 0.01566 | 0.271186 | 0.00799 | 0.00781 | 0.015625 | 0.0078 | 0.04651 | male |
| 5 | 0.151228 | 0.07 | 0.158 | 0.1 | 0.21 | 0.11 | 1.233 | 4.177 | 0.9633 | 0.727 | 0.084 | 0.151228 | 0.088965 | 0.0178 | 0.25 | 0.201497 | 0.00781 | 0.5625 | 0.5547 | 0.24712 | male |
| 6 | 0.13512 | 0.08 | 0.1247 | 0.08 | 0.21 | 0.13 | 1.101 | 4.334 | 0.972 | 0.784 | 0.104 | 0.13512 | 0.106398 | 0.01693 | 0.266667 | 0.712813 | 0.00781 | 5.484375 | 5.4766 | 0.20827 | male |
| 7 | 0.132786 | 0.08 | 0.1191 | 0.07 | 0.21 | 0.14 | 1.933 | 8.309 | 0.9632 | 0.738 | 0.113 | 0.132786 | 0.110132 | 0.01711 | 0.253968 | 0.298222 | 0.00781 | 2.726563 | 2.7188 | 0.12516 | male |
| 8 | 0.150762 | 0.07 | 0.1601 | 0.09 | 0.21 | 0.11 | 1.531 | 5.987 | 0.9676 | 0.763 | 0.086 | 0.150762 | 0.105945 | 0.02623 | 0.266667 | 0.47962 | 0.00781 | 5.3125 | 5.3047 | 0.12399 | male |
| 9 | 0.160514 | 0.08 | 0.1443 | 0.11 | 0.23 | 0.12 | 1.397 | 4.767 | 0.9593 | 0.72 | 0.128 | 0.160514 | 0.093052 | 0.01776 | 0.144144 | 0.301339 | 0.00781 | 0.539063 | 0.5313 | 0.28394 | male |
| 10 | 0.142239 | 0.08 | 0.1386 | 0.09 | 0.21 | 0.12 | 1.1 | 4.07 | 0.9707 | 0.771 | 0.219 | 0.142239 | 0.096729 | 0.01796 | 0.25 | 0.336476 | 0.00781 | 2.164063 | 2.1563 | 0.14827 | male |
| 11 | 0.134329 | 0.08 | 0.1215 | 0.08 | 0.2 | 0.13 | 1.19 | 4.787 | 0.9752 | 0.805 | 0.012 | 0.134329 | 0.105881 | 0.0193 | 0.262295 | 0.340365 | 0.01563 | 4.695313 | 4.6797 | 0.08992 | male |
| 12 | 0.157021 | 0.07 | 0.1682 | 0.1 | 0.22 | 0.12 | 0.979 | 3.974 | 0.9652 | 0.734 | 0.096 | 0.157021 | 0.088894 | 0.02207 | 0.117647 | 0.460227 | 0.00781 | 2.8125 | 2.8047 | 0.2 | male |
| 13 | 0.138551 | 0.08 | 0.1275 | 0.09 | 0.2 | 0.12 | 1.627 | 6.291 | 0.966 | 0.752 | 0.012 | 0.138551 | 0.104199 | 0.01914 | 0.262295 | 0.246094 | 0.00781 | 2.71875 | 2.7109 | 0.13235 | male |
| 14 | 0.137343 | 0.08 | 0.1243 | 0.08 | 0.21 | 0.13 | 1.379 | 5.009 | 0.9635 | 0.736 | 0.108 | 0.137343 | 0.092644 | 0.01679 | 0.213333 | 0.481671 | 0.01563 | 5.015625 | 5 | 0.0885 | male |
| 15 | 0.181225 | 0.06 | 0.191 | 0.13 | 0.23 | 0.1 | 1.369 | 5.476 | 0.9374 | 0.537 | 0.22 | 0.181225 | 0.131504 | 0.025 | 0.275862 | 1.277114 | 0.00781 | 2.804688 | 2.7969 | 0.41655 | male |
| 16 | 0.183115 | 0.07 | 0.1912 | 0.13 | 0.24 | 0.11 | 3.568 | 35.38 | 0.9403 | 0.571 | 0.05 | 0.183115 | 0.102799 | 0.02083 | 0.275862 | 1.245739 | 0.20313 | 6.742188 | 6.5391 | 0.13933 | male |
| 17 | 0.174272 | 0.07 | 0.1909 | 0.12 | 0.23 | 0.11 | 4.485 | 61.76 | 0.951 | 0.635 | 0.05 | 0.174272 | 0.102046 | 0.01833 | 0.246154 | 1.621299 | 0.00781 | 7 | 6.9922 | 0.20931 | male |
| 18 | 0.190846 | 0.07 | 0.208 | 0.13 | 0.24 | 0.11 | 1.562 | 7.834 | 0.9385 | 0.539 | 0.05 | 0.190846 | 0.113323 | 0.01754 | 0.275862 | 1.434115 | 0.00781 | 6.320313 | 6.3125 | 0.25478 | male |
| 19 | 0.171247 | 0.07 | 0.1528 | 0.12 | 0.24 | 0.12 | 3.207 | 25.77 | 0.937 | 0.586 | 0.06 | 0.171247 | 0.079718 | 0.01567 | 0.262295 | 0.106279 | 0.00781 | 0.570313 | 0.5625 | 0.13835 | male |
| 20 | 0.168346 | 0.07 | 0.1456 | 0.12 | 0.24 | 0.12 | 2.704 | 18.48 | 0.9345 | 0.56 | 0.06 | 0.168346 | 0.083484 | 0.01572 | 0.231884 | 0.146563 | 0.00781 | 3.125 | 3.1172 | 0.05954 | male |
| 21 | 0.173631 | 0.07 | 0.1536 | 0.12 | 0.24 | 0.12 | 2.805 | 20.86 | 0.9309 | 0.518 | 0.06 | 0.173631 | 0.09013 | 0.0157 | 0.210526 | 0.193044 | 0.00781 | 2.820313 | 2.8125 | 0.06812 | male |
| 22 | 0.172754 | 0.08 | 0.1777 | 0.12 | 0.25 | 0.13 | 2.968 | 20.08 | 0.9255 | 0.523 | 0.06 | 0.172754 | 0.093574 | 0.01576 | 0.2 | 0.235877 | 0.00781 | 0.71875 | 0.7109 | 0.23507 | male |
| 23 | 0.181015 | 0.07 | 0.1693 | 0.13 | 0.25 | 0.13 | 2.587 | 12.28 | 0.9153 | 0.475 | 0.06 | 0.181015 | 0.098643 | 0.01615 | 0.275862 | 0.209844 | 0.00781 | 3.695313 | 3.6875 | 0.05994 | male |
| 24 | 0.163536 | 0.07 | 0.1455 | 0.11 | 0.23 | 0.11 | 3.588 | 28.65 | 0.927 | 0.542 | 0.06 | 0.163536 | 0.062542 | 0.01569 | 0.197531 | 0.059622 | 0.00781 | 0.445313 | 0.4375 | 0.0917 | male |
| 25 | 0.170213 | 0.08 | 0.1461 | 0.12 | 0.25 | 0.13 | 2.817 | 13.76 | 0.9138 | 0.488 | 0.06 | 0.170213 | 0.077698 | 0.0157 | 0.192771 | 0.101563 | 0.00781 | 0.5625 | 0.5547 | 0.16179 | male |

图 2.1 数据集内容概览

数据集共 3168 行，每一行表示一个声音样本，包含 20 个属性和 1 个标签。其中 20 个属性的说明如下：

- meanfreq: 平均频率（以 kHz 为单位）
- sd: 频率标准偏差
- median: 中位数频率（以 kHz 为单位）

- Q25: 第一个分位数 (以 kHz 为单位)
- Q75: 第三分位数 (以 kHz 为单位)
- IQR: 分位数范围 (以 kHz 为单位)
- skew: 偏斜
- kurt: 峰度
- sp.ent: 频谱熵
- sfm: 光谱平坦度
- mode: 模式频率
- centroid: 频率质心
- peakf: 峰值频率 (能量最高的频率)
- meanfun: 跨声信号测得的基频平均值
- minfun: 跨声学信号测得的最小基频
- maxfun: 跨声学信号测得的最大基频
- meandom: 整个声信号测得的主频的平均值
- mindom: 跨声学信号测得的主频率的最小值
- maxdom: 整个声信号测得的主频率最大值
- dfrange: 跨声信号测得的主频范围
- modindx: 调制指数。计算为相邻基频测量之间的累计绝对差除以频率范围

最后一个属性列为该声音样本的标签: 'male'表示男性, 'female'表示女性。

从图 2.1 中不难发现, 数据集中存在空缺数据的情况, 缺省以 0 值进行填补, 后续可以对空缺项进行处理。

2.3 任务要求

2.3.1 算法要求

要求使用朴素贝叶斯方法。

对于参数的获取, 使用连续性的参数估计方法得到, 即将属性看作连续的, 并假设其服从属性列均值 μ 和属性列标准差 σ 的高斯分布。

2.3.2 数据集划分要求

要求对数据集以 7:3 的比例划分训练集和测试集，抽取过程要求随机化。

2.3.3 结果展示要求

最终要求以 2*2 表格的形式呈现预测情况，表格四部分内容如下表 2.1 所示：

表 2.1 结果展示例表

| | |
|-------|-------|
| 男声正确率 | 男声错误率 |
| 女声正确率 | 女声错误率 |

3 设计与分析

3.1 朴素贝叶斯分类原理

3.1.1 算法简介

贝叶斯方法是以贝叶斯决策论作为理论基础，使用概率统计的知识对样本数据集进行分类。贝叶斯决策论简单来说就是在所有相关概率都已知时，根据已知概率结合误判损失来综合给出最优的样本类别判断结果。贝叶斯方法结合了先验概率和后验概率，算法简单但却能够保持较高的准确率。

朴素贝叶斯算法是在贝叶斯方法基础上的简化，它具有一个非常重要的假设：对于一个目标值的各个属性分量之间都相互条件独立。这个条件非常苛刻，实际生活中基本不可能满足，但实际使用时正确率依然有所保证，而且大大简化了算法，因此使用非常广泛。

我认为三者的关系可以这样看待：贝叶斯决策论是数学基础，贝叶斯方法是抽象模型，而朴素贝叶斯算法则是解决实际问题的具体方法。

3.1.2 算法步骤

首先假设数据集 $D = \{d_1, d_2, \dots, d_n\}$ ，数据分类变量 $Y = \{y_1, y_2, \dots, y_m\}$ ，每个样本的特征属性集 $X = \{x_1, x_2, \dots, x_p\}$ 。通俗来说就是数据集的 n 个样本可以被分为 m 类，每个样本具有 p 个特征属性。且假设 p 个特征属性 x_1, x_2, \dots, x_p 是相互独立且随机的。

由上述假设条件可知， Y 的先验概率 $P_{pri} = P(Y)$ ，后验概率 $P_{pos} = P(Y|X)$ ，分类条件概率 $P(X|Y)$ ，特征出现概率 $P(X)$ ，由概率论知识可知：

$$P(Y|X) = \frac{P(Y) P(X|Y)}{P(X)}$$

在给定类别为 y 的情况下，由于各个特征属性之间相互独立且随机，因此有：

$$P(X|Y = y) = \prod_{i=1}^p P(x_i|Y = y)$$

因此后验概率可以表示为：

$$P_{pos} = P(Y|X) = \frac{P(Y) \prod_{i=1}^p P(x_i|Y = y)}{P(X)}$$

对于不同的类别 y_i 而言， $P(X)$ 是恒定不变的，因此只需要比较上式中分子部分的 $P(Y) \prod_{i=1}^p P(x_i|Y = y)$ 的大小即可。

其中 $P(Y)$ 是先验概率，可以根据历史数据经验求得， $\prod_{i=1}^p P(x_i|Y = y)$ 的表示的是分类条件概率，也就是各个特征属性对于类型 y 的条件概率的连乘积。

对于 $\prod_{i=1}^p P(x_i|Y = y)$ 的求解方式有两种：

(1) x_i 对应属性为离散属性

对于离散属性，由于属性的取值是有限的，不妨设在 x_i 对应属性下有 m 个相异属性值，而属性值为 x_i 的样本共有 n 个，则 $P(x_i|Y = y)$ 就可以使用频率表示概率：

$$P(x_i|Y = y) = \frac{n}{m}$$

(2) x_i 对应属性为连续属性

对于连续属性，可以假设各特征属性的值都服从正态分布（高斯分布），根据数据集可以计算出各属性列在类别为 y 的条件下的平均值 μ_y 和方差 σ_y ，从而得到正态分布的密度函数，再将样本的属性值代入即可得到 $P(x_i|Y = y)$ 的概率：

$$P(x_i|Y = y) = G(x_i, \mu_y, \sigma_y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}}$$

求解得到 $\prod_{i=1}^p P(x_i|Y = y)$ 之后代入即可求得后验概率 P_{pos} ，比较后验概率的大小，取后验概率最大时对应的类型 y 作为分类器最终输出的类型。

3.2 项目建模

对本次具体的语音性别识别项目进行建模如下：

设数据集 $D = \{d_1, d_2, \dots, d_{3168}\}$ ，数据分类变量 $Y = \{'male', 'female'\}$ ，每个样本的特征属性集 $X = \{x_1, x_2, \dots, x_{20}\}$ ，每个属性 x_i 的属性值设为 v_i 。也就是数据集的 3168 个样本可以被分为男女两类，每个样本具有 20 个特征属性。且假设这 20 个特征属性是相互独立且随机的。

根据 3.1 描述的算法，只需要关心分子 $P(Y) \prod_{i=1}^{20} P(x_i|Y = y)$ 的大小，其中先验概率 $P_{pri} = P(Y)$ 可以根据历史情况求得，因此计算的两项如下：

$$P(Y = 'male'|X) = P(Y = 'male') \prod_{i=1}^{20} P(v_i|Y = 'male')$$
$$P(Y = 'female'|X) = P(Y = 'female') \prod_{i=1}^{20} P(v_i|Y = 'female')$$

其中 $P(v_i|Y = 'male')$ 和 $P(v_i|Y = 'female')$ 都可以使用 3.1.2 节中提到的 x_i 对应属性为连续属性的情况，使用高斯分布得到参数。

计算完毕后，比较 $P(Y = 'male'|X)$ 和 $P(Y = 'female'|X)$ 的值，选取较大值对应的类型作为分类器的输出。

3.3 设计思路与流程图

3.3.1 设计思路

(1) 对数据集进行预先处理

可以发现数据集中存在一些值为 0 的项，对于这种空缺项，一种简单的处理方法就是取得其他值的平均数，再用这个平均值填补空缺项。

(2) 拆分训练集和测试集

项目要求使用 7:3 的比例拆分训练集和测试集，观察图 2.1 所示的数据集可知，数据集的排列是按照先'*male*'后'*female*'的顺序进行组织的，若直接截取前 70%作为训练集，后 30%作为测试集，显然是不合理的。故需要对数据集做如下处理：

1. 将数据集按行随机混洗，打乱原有行的顺序；
2. 从混洗后的数据集中随机抽取 70%作为训练集；
3. 把剩余的 30%数据集作为测试集。

(3) 计算先验概率 p_{male} 和 p_{female}

统计训练集中 label 列为'*male*'的行数 row_{male} ，再统计总行数 row_{all} ，则计算步骤如下：

$$p_{male} = P(Y = 'male') = \frac{row_{male}}{row_{all}}$$

$$p_{female} = P(Y = 'female') = 1 - P(Y = 'male')$$

(4) 计算平均值和标准差

计算训练集中各属性列（除标签之外）的平均值 μ 和标准差 σ ，便于计算高斯分布概率。

(5) 计算各特征属性的权重*

这一步作为可选的优化项，先计算训练集中男性样本的各属性值的平均数 $\bar{x}_{l_{男}}$ 和女性样本各属性值的平均数 $\bar{x}_{l_{女}}$ ，然后计算两者的相对偏差，把各属性值相对偏差的大小关系映射为各属性的权重。

(6) 遍历测试集的每个样本，对其使用贝叶斯分类器进行分类

将测试集的每个样本送入贝叶斯分类器。

贝叶斯分类器的设计流程如下：

1. 设置 p_male_cond 和 p_female_cond 初始值。若有对数优化，则设为 $\log(p_male)$ 和 $\log(p_female)$ ，否则设为 p_male 和 p_female 。
2. 对于每个属性列做如下操作：
 - a) 设置属性权重 $weight$ 。若有属性权重优化，则设为对应属性权重，否则设为 1.0;
 - b) 通过高斯分布函数计算当前属性的条件概率 g_male 和 g_female ;
 - c) 若有对数优化，则对 p_male_cond 和 p_female_cond 分别累加上 $weight * \log(g_male)$ 和 $weight * \log(g_female)$ 。否则对 p_male_cond 和 p_female_cond 分别累乘上 $(g_male)^{weight}$ 和 $(g_female)^{weight}$ 。
3. 比较 p_male_cond 和 p_female_cond 的大小，若 $p_male_cond > p_female_cond$ ，则分类器输出'male'，否则分类器输出'female'。

(7) 统计结果

对每个样本的分类结果，与测试集标签进行对比，分性别统计预测正确的个数和正确率。

3.3.2 程序流程图

程序流程图如下图 3.1 所示：

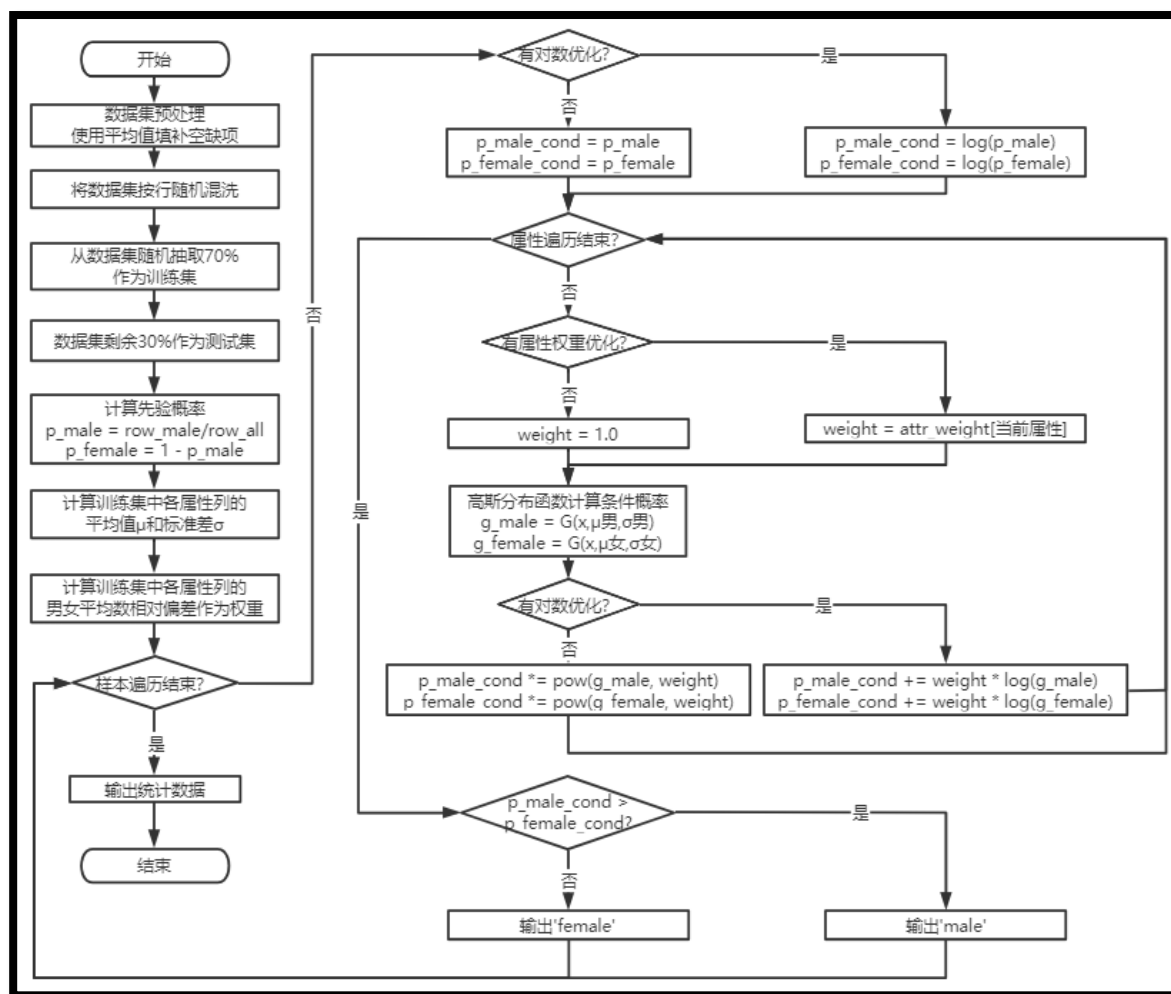


图 3.1 程序流程图

3.4 系统改进和优化

3.4.1 填补数据集空缺时与性别相关联

在 3.3.1 节的第（1）步中对数据集进行预处理，使用列平均值填补数据集空缺的方法存在以下两个方面的漏洞：一方面，求列平均值时包含了值为 0 的项，会使得平均值偏小；另一方面，直接以平均值填补空缺忽略了性别因素，对分类器分类会产生不利影响。

综上，可以对填补数据集空缺时的方法作出如下改进：

（1）仅对属性列中的非 0 值求平均值。

（2）求平均值时区分男性和女性，分别求男性样本的列属性平均值 $\bar{x}_{i男}$ 和女性样本的列属性平均值 $\bar{x}_{i女}$ 。在遇到空缺项时，若待填补样本的标签为男性，则使用 $\bar{x}_{i男}$ 填补空缺，否则使用 $\bar{x}_{i女}$ 填补空缺。

3.4.2 使用取对数累加法代替连乘

根据朴素贝叶斯算法的计算公式，计算出各条件概率之后要进行累乘。由于本题中有 20 个特征属性，再加上先验概率，需要计算 21 个概率的连乘积。由于概率 P 总是小于 1 的，因此这个连乘积的值会非常小。即使 Python 计算时使用的是 64 位浮点数 float 64，但也是有精度限制的，太小的值会产生截断误差，对结果造成影响。

因此，可以对连乘积取对数，转变为各个概率取对数之后的累加操作，公式为：

$$\log(P(Y) \prod_{i=1}^{20} P(v_i|Y)) = \log P(Y) + \sum_{i=1}^{20} \log P(v_i|Y)$$

取对数之后，对于接近 0 的值，也能够非常显著地体现出来，解决了连乘积的值过小带来的精度问题，一定程度上降低了误判的概率。

3.4.3 引入属性权重特征

在朴素贝叶斯算法中，假设各属性相互独立且对分类判断的影响程度相同。但实际上各个属性对于最终判断的关联程度是不一样的。

举例而言，比较本项目中的属性 Q75 和属性 IQR，如图 3.2 所示：

| | | |
|-----|----------|----------|
| Q75 | 0.226346 | 0.223184 |
| IQR | 0.110784 | 0.057834 |

图 3.2 Q75 和 IQR 的平均值

(1) 数据集的属性 Q75 对应的男性平均值为 0.226346，女性平均值为 0.223184，计算男女平均值之间的相对偏差为 1.4%。

(2) 数据集的属性 IQR 对应的男性平均值为 0.110784，女性平均值为 0.057834，计算男女平均值之间的相对偏差为 62.8%。

根据上述对比，显然属性 IQR 比属性 Q75 更适合作为区分男性和女性的依据，因为男性和女性在属性 IQR 上体现的差距更明显。

因此可以做出如下优化：

(1) 计算训练集中男性样本各属性的平均值 $\bar{x}_{i男}$ 和女性样本各属性值平均值 $\bar{x}_{i女}$

(2) 计算两者的相对偏差 $relative_gap$ ，公式如下：

$$relative_gap_i = \frac{|\bar{x}_{i男} - \bar{x}_{i女}|}{(\bar{x}_{i男} + \bar{x}_{i女}) / 2}$$

也就是将男性平均值和女性平均值之差的绝对值除以两数的平均值。

(3) 将相对偏差进行规范化处理，成为属性的权重

由于在优化前，20 个属性的权重可以看作是 1，因此属性的总权重为 20。为了不改变这一数量关系，需要把相对偏差的数值整体调整，做法如下：

1. 求各属性相对偏差之和 `sum_weight`
2. 将各属性相对偏差的值乘以 20，再除以 `sum_weight`，得到各属性最终权重

这一步可以把各属性值相对偏差的大小关系映射为各属性的权重。

按照上述三个步骤计算完各个属性权重之后，在朴素贝叶斯分类器工作的过程中，就可以给各个属性分量结合权重进行计算。若是取对数的求和运算，则在取完对数之后乘以其权重，再相加；若是未取对数的连乘运算，则对每一项求其权重次幂，再相乘。

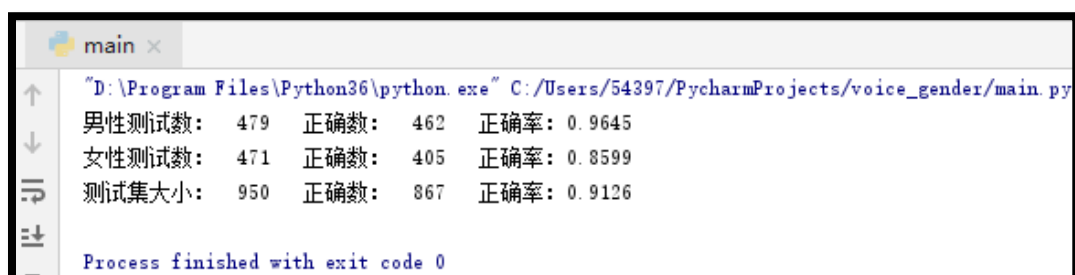
4 结果分析

4.1 实际运行结果

在 Python 3.6 的 64 位环境下运行编写好的程序代码，使用的模块如下：

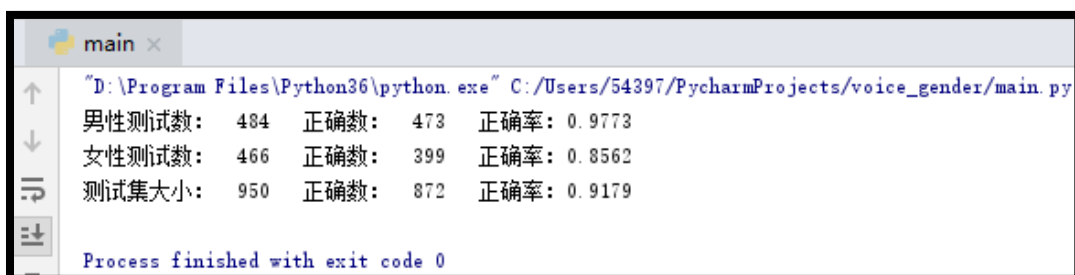
1. 使用了 pandas 模块用来读取 csv 文件为 DataFrame，并对 DataFrame 进行操作
2. 使用了 math 模块用来进行对数运算
3. 使用了 scipy 包里的 norm 模块用来计算高斯分布的函数值

实际运行结果展示时，使用对数优化加属性权重优化的最终版本，共运行四次，结果分别如下图 4.1 至 4.4 所示：



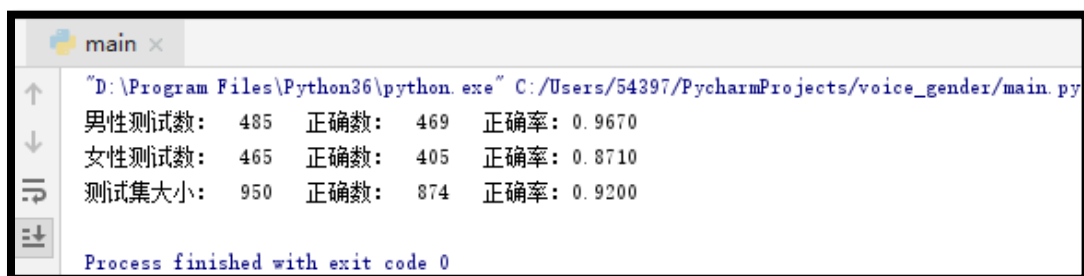
```
main x
"D:\Program Files\Python36\python.exe" C:/Users/54397/PycharmProjects/voice_gender/main.py
男性测试数: 479 正确数: 462 正确率: 0.9645
女性测试数: 471 正确数: 405 正确率: 0.8599
测试集大小: 950 正确数: 867 正确率: 0.9126
Process finished with exit code 0
```

图 4.1 测试点 1



```
main x
"D:\Program Files\Python36\python.exe" C:/Users/54397/PycharmProjects/voice_gender/main.py
男性测试数: 484 正确数: 473 正确率: 0.9773
女性测试数: 466 正确数: 399 正确率: 0.8562
测试集大小: 950 正确数: 872 正确率: 0.9179
Process finished with exit code 0
```

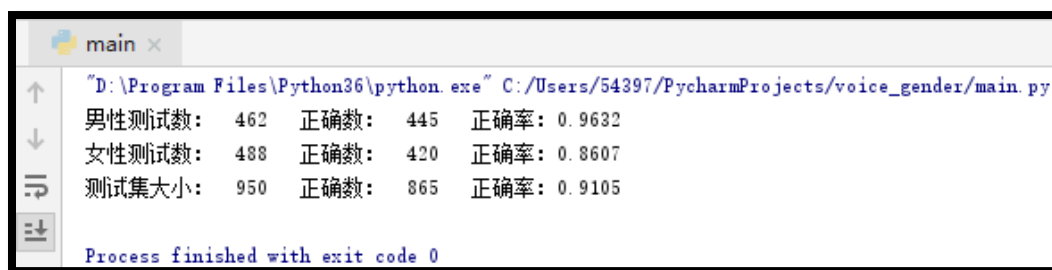
图 4.2 测试点 2



```
main x
"D:\Program Files\Python36\python.exe" C:/Users/54397/PycharmProjects/voice_gender/main.py
男性测试数: 485 正确数: 469 正确率: 0.9670
女性测试数: 465 正确数: 405 正确率: 0.8710
测试集大小: 950 正确数: 874 正确率: 0.9200
Process finished with exit code 0
```

图 4.3 测试点 3

华中科技大学项目开发报告



```
main x
"D:\Program Files\Python36\python.exe" C:/Users/54397/PycharmProjects/voice_gender/main.py
男性测试数: 462 正确数: 445 正确率: 0.9632
女性测试数: 488 正确数: 420 正确率: 0.8607
测试集大小: 950 正确数: 865 正确率: 0.9105

Process finished with exit code 0
```

图 4.4 测试点 4

对上述结果求平均值，可计算得到男性的平均正确率为 0.9680，女性的平均正确率为 0.8620，总体的平均正确率为 0.9153。

根据上述结果，作出表 4.1 的结果：

表 4.1 实际运行结果表

| 男声正确率 | 男声错误率 |
|--------|--------|
| 0.9680 | 0.0320 |
| 女声正确率 | 女声错误率 |
| 0.8620 | 0.1380 |
| 总体正确率 | 总体错误率 |
| 0.9153 | 0.0847 |

4.2 优化前后性能对比

根据 3.4 节系统改进和优化的描述，本次尝试了三种优化方法。经过实践检测，发现填补数据集空缺对性能影响不显著，因此下面分析对数优化、属性权重优化对性能的影响。对四种情况分别运行 4 次，求 4 次的平均值并作出下面的表格：

(1) 无对数优化、无属性权重优化

表 4.2 无对数优化、无属性权重优化的运行结果表

| 男声正确率 | 男声错误率 |
|--------|--------|
| 0.9752 | 0.0248 |
| 女声正确率 | 女声错误率 |
| 0.7763 | 0.2237 |
| 总体正确率 | 总体错误率 |
| 0.8758 | 0.1242 |

(2) 有对数优化、无属性权重优化

表 4.3 有对数优化、无属性权重优化的运行结果表

| 男声正确率 | 男声错误率 |
|--------|--------|
| 0.9807 | 0.0193 |
| 女声正确率 | 女声错误率 |
| 0.7872 | 0.2128 |
| 总体正确率 | 总体错误率 |
| 0.8821 | 0.1179 |

(3) 无对数优化、有属性权重优化

表 4.4 无对数优化、有属性权重优化的运行结果表

| 男声正确率 | 男声错误率 |
|--------|--------|
| 0.9637 | 0.0363 |
| 女声正确率 | 女声错误率 |
| 0.8527 | 0.1473 |
| 总体正确率 | 总体错误率 |
| 0.9074 | 0.0926 |

(4) 有对数优化、有属性权重优化

表 4.5 有对数优化、有属性权重优化的运行结果表

| 男声正确率 | 男声错误率 |
|--------|--------|
| 0.9680 | 0.0320 |
| 女声正确率 | 女声错误率 |
| 0.8620 | 0.1380 |
| 总体正确率 | 总体错误率 |
| 0.9153 | 0.0847 |

根据上述结果，绘制四种情况的性能对比图如下：

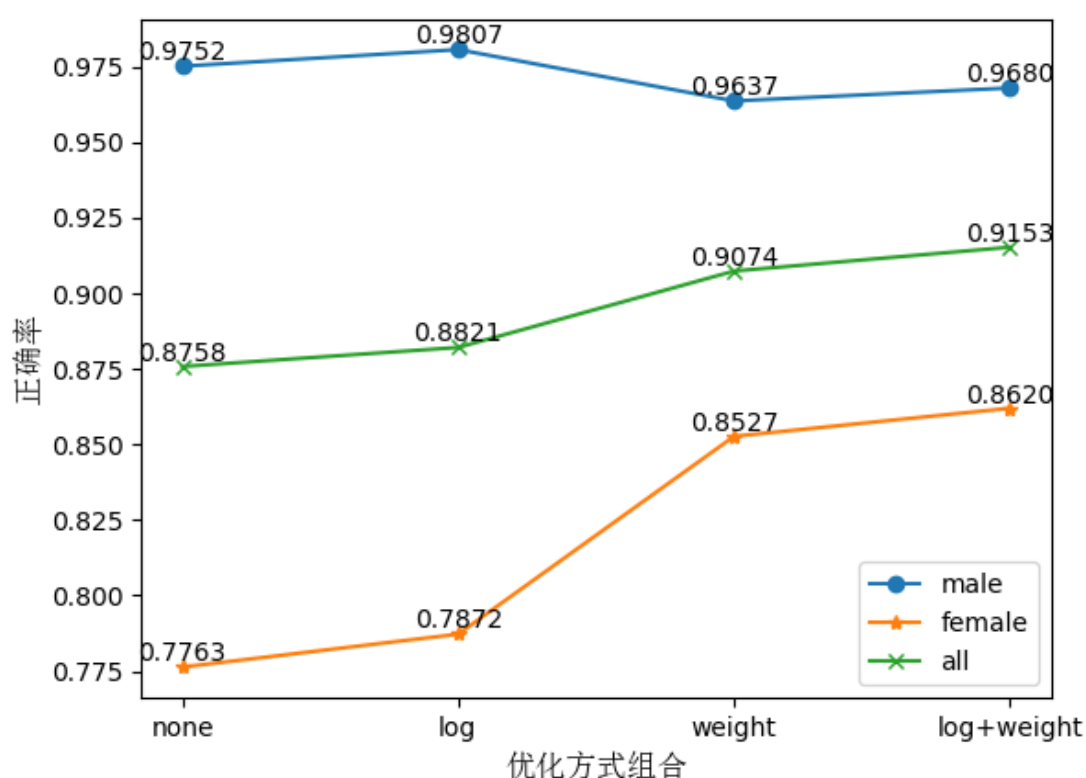


图 4.5 优化项目性能对比图

由上图 4.5 可以看出，对数优化对于系统性能的提升较小但较稳定，而引入属性权重之后，对系统综合性能的提升是比较明显的。

此外还能看到，对数优化对男女声音的识别率都有所帮助，而引入属性权重的优化方法则会略微降低男性声音的识别正确率，而显著提高女性声音的识别正确率。

4.3 分析影响性能的因素

不难看出，识别正确率在男女之间存在较大差异，对男性声音的识别正确率显著高于女性。出现这种情况的原因可能是朴素贝叶斯要求相互独立，但实际上属性之间不可避免存在关联，且女性对关联的敏感度更高。还有一种可能的原因是女性声音的频率较高，且变化范围大，数据集的大小有限，限定的声音频率范围也有限，无法准确描述女性声音的动态变化。

此外，两种优化方法也对分类器性能产生影响：

(1) 引入对数运算进行优化

减少了计算机精度限制引起的误差，一定程度上可以提高分类器的性能，减少误判。

通过图 4.5 的分析，在计算时引入对数运算进行优化，能够使得正确率提高 0.7% 至 0.9%。

（2）引入属性权重进行优化

能够根据属性与最终类型的相关程度，按权重进行计算，使得相关度低的属性对结果产生的影响下降，而使得与类型相关度高的属性能够尽可能发挥分类的作用。

通过图 4.5 的分析，在计算时引入属性权重进行优化，能够使得正确率提高 3.6% 至 3.8%。

5 思考与总结

5.1 项目小结

5.1.1 所做的工作

本次项目开发过程中，我完成了以下工作：

- (1) 分析题目，并以朴素贝叶斯算法为基础建立了解体模型。
- (2) 处理数据集，分性别使用非空数据的平均值填补数据集中的空缺值。
- (3) 将数据集按照 7:3 的比例划分为训练集和测试集。
- (4) 使用 Python 语言实现了使用朴素贝叶斯算法解决该问题的程序。
- (5) 思考了程序的优化方法，并在条件概率的计算过程中通过引入对数计算和属性特征权重这两种方法进行优化，分别使得正确率提升约 0.8%和约 3.7%。
- (6) 对比分析了不同优化方法组合的实际效果。
- (7) 分析了影响系统性能（正确率）的因素。

5.1.2 未来改进的方向

未来对系统的改进方向包括：

- (1) 调整权重计算方式

目前的权重计算方式只和属性特征对于男女性别的相对偏差有关。后续可以通过分析与结果关联度较大的属性，加大其权重，使得计算出来的属性特征权重更贴合实际。

- (2) 尝试使用离散化模型

本次实验使用的是高斯分布对连续的属性值做估计，另一种方法是将数据集中的属性值进行量化，使得各属性值离散分布，然后再使用离散的条件概率计算公式，用频率表示概率。随着量化阶数的提高，分类器的正确率应该也会有所提高。

5.2 心得收获

通过机器学习这门课程的学习，我学到了很多非常经典的机器学习算法，如 KNN、SVM、Logistic、朴素贝叶斯等。很多算法的实现过程并不难，却能取得非常好的效果，体现了算法的巧妙。

华中科技大学项目开发报告

本次项目开发主要使用的是朴素贝叶斯算法，这个算法本身比较简单，但一开始得到的结果并不令我满意，女声的识别准确度只有 80% 不到。后来我把程序核心部分对于条件概率的计算过程逐一打印出来分析，发现了有些属性对结果影响很大，有些却影响不大。根据这一发现，启发我引入了属性权重的概念。使用各属性的男女平均值的相对偏差作为该属性的权重，能够很好的发挥部分重要属性对识别结果的影响力。在这个过程中，我思考了对朴素贝叶斯算法的优化方法，并跟着自己的想法去实践，并且最终取得了一定的效果，这个过程使我非常难忘。

此前的我从未接触过机器学习，这门课程作为我的启蒙课，通过老师讲解和项目的练习，不仅让我掌握了机器学习的基本知识，还让我对它产生了浓厚的兴趣。很感谢老师的教学和指导，即使疫情期间只能上网课，我也感觉确确实实学到了东西，收获颇丰。