**2023年数据科学导论期末课程项目**

|  |  |  |  |
| --- | --- | --- | --- |
| **负责人:** | 许廷阳 | **学 号:** | 922127940155 |
| **电 话:** | 19951196072 | | |
| **选 题:** | 题目二: 分类、集成模型及其应用 | | |
| **专 业:** | 数据科学与大数据技术 | | |
| **学 院:** | 网络空间安全学院 | | |

**南京理工大学基础前沿交叉中心**

**2023年 12 月 28 日**

**期末课程项目注意事项及要求**

* **本课程期末考核包含三个项目, 需完成其中任意一个, 请谨慎选择. 选定一个项目后, 请删除其余不相干内容. *请务必按照格式规范撰写大作业.***
* **考核内容: 侧重数据科学导论课程中算法及模型的理解与应用.**
* **数值试验题应同时提交纸质和电子版报告(包含程序并打包), 其中书面报告有详细的推导和数值结果及分析. 说明: 尽管完成的方式、使用的工具、编程的技巧都很重要, 但本项目强调理论分析、看重方法的使用、重视结果的分析…**
* **成绩评定: 平时作业约占考核总体的60%, 项目大作业占比40%. 迟交一天(24小时)打折10%, 不接受晚交4天的作业和项目, 任何时候理由都不接受.**
* **评分细则: (1)完成要求的任务. (2)鼓励提出原创性的方法并予以实现. (3)提交一份完整的项目报告(包括纸质版、电子版源文件及程序). (4)2024年1月11日18:00点前发(递交)至课代表处.**
* **期末课程项目可分组完成(自由组合, 不超过3人一组，可以1个人), 可以同学间相互讨论或者找老师答疑. 如果是多人组队, 请填写附表1, 明确说明每人负责的部分和内容.**
* **鼓励开源交流, 禁止直接抄袭. 有讨论或从其它任何途径取得帮助, 请列出来源. 格式请参考本科生毕业论文写作模板(http://bysj.njust.edu.cn/Index.aspx).**
* **可选加分项目: (1)代码部分所使用的函数自己编写, 非调包. (2)将项目报告以slides形式展示给他人, 包括老师与同学. (3)在GitHub中将代码分享给需要使用的人, 非上传至盈利组织(如某文库、某丁网等).**
* **挑战项目: DataHub成员类项目(包含但不限于：泰迪杯、全国大学生数学建模大赛、美国大学生数学建模竞赛、Kaggle竞赛…), 具体咨询组内成员或相关老师.**
* **请使用公式编辑器对项目内包含的公式进行编辑……**
* **本项目及其相关内容在未经授课教师准许, 请勿随意上传至网络, 仅限于校内交流合作使用.**

**目 录**

[课程项目二 分类、集成模型及其应用 - 1 -](#_Toc154705405)

[【1】项目背景 - 1 -](#_Toc154705406)

[【2】方法陈述 - 1 -](#_Toc154705407)

[【3】案例实战 - 1 -](#_Toc154705408)

[【4】案例代码 - 19 -](#_Toc154705409)

[附表1 参与人员 - 1 -](#_Toc154705410)

# 课程项目二 分类、集成模型及其应用

## 【1】项目背景

在回归模型中假设响应变量是定量的, 但很多情况下响应变量却是定性的. 例如, 眼睛的颜色是定性变量, 取值蓝色、棕色或绿色. 定性变量也称为分类变量, 两者的统计含义是一样的. 预测一个观测的定性响应值也指对观测分类, 因为它涉及将观测分配到一个类别中. 一些分类方法先从预测定性变量不同类别的概率开始, 将分类问题作为概率估计的一个结果. 从这个角度上看, 分类与回归方法有许多类似之处.

《数据科学导论》课程中已习得几种分类方法, 如逻辑回归、近邻、决策树、支持向量机、随机森林和AdaBoost等. 现有分类器中, 支持向量机(support vector machine, SVM)是一种起源于机器学习社区的流行方法, 它是使用单一分类函数对二分类问题进行最大间隔分类的典型例子, 在实际应用中性能良好. 在很多问题中, 数据通常包含两个以上的类别. 尽管SVM在二分类上取得了成功, 但如何将其应用于多类别分类问题仍然是一个挑战. 文献中使用SVM进行多分类的方法可分为两大类: 第一类方法先训练一系列的二值支持向量机, 再将结果组合起来进行多类别分类. 例子包括一对一(one-versus-one)和一对多(one-versus-rest)方法. 尽管这种方法在概念和实现上都很简单, 但它也有自身的缺点. 第二类方法是在一个优化问题中同时考虑所有个类. 常用的方法是用个分类函数来表示对应优化中的个类, 并根据哪个分类函数最大给出最终的预测结果. 目前, 在第二类框架下提出了许多同步多分类支持向量机分类器. 本项目重点关注支持向量机与其他方法的比较, 以及如何使用SVM处理多分类问题.

## 【2】方法陈述

**问题一:** 请对现有的种分类方法进行总结, 试搜寻每类方法的优缺点(SVM需要着重介绍).

**问题二:** 尝试借助课程及网络, 检索SVM及其变体模型, 并对其相关机理进行论述.

## 【3】案例实战

**1. 教材案例回顾:** **乳腺癌诊断**

早期的乳腺癌检测主要检查乳腺组织的异常肿块. 如果能利用机器学习算法, 通过乳腺肿块的检测数据自动进行诊断, 将会给医疗系统带来很大的益处: 不仅能够大大提高检测效率, 还可以降低误判的风险.

数据集共有569个样本, 每个样本包含乳腺细胞的30个特征, 这30个特征都是由数字化细胞核的10个基础特征(包括半径、质地、周长、面积、对称性等) 的均值、标准差及最大值构成. 请使用分类方法(包含SVM), 对乳腺癌数据进行分类, 并选择适当的指标对最终的分类结果进行评价.

注: 评价指标可参考教材附录E.

* 1. **数据预处理**
  2. **SVM**
     1. Normal SVM
        1. 模型初步分析

SVM对于任意维特征值均可以找到一个超平面来对其实现分类，在本数据集中属于一个二分类模型，我们故为防止与后续GA-SVM重复，选择linear线性核作为我们的核函数，而由于我们在后续使用了GA算法对参数进行筛选，故没有对其超参数调整。

* + - 1. 模型结果分析

表（1.2.1.2-1）为模型结果参数表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.98 | 1 | 0.99 | 108 |
| 1 | 1 | 0.97 | 0.98 | 63 |
| accuracy |  |  | 0.99 | 171 |
| macro avg | 0.99 | 0.98 | 0.99 | 171 |
| weighted avg | 0.99 | 0.99 | 0.99 | 171 |

表（1.2.1.2-1）

图表, 树状图

描述已自动生成 图（1.2.1.2-2）为混淆矩阵

图（1.2.1.2-2）

根据表（1.2.1.2-1）所示的结果，我们可以看到模型的阳性样本查准率为100%查全率为97%最后精确度为0.98830，处于一个较为良好的程度，不过任然会出现误诊的情况，即将阳性样本误判为阴性，该问题我们试图在下面GA算法调参的SVM中（1.2.2章节）进一步优化和阐述。

* + 1. GA-SVM
       1. 模型初步分析

由于SVM含有C和gamma两个超参数，且这两个超参数对模型性能影响较大，光是使用传统的交叉验证调整参数对于参数选择数量上有一定的局限性，且很难将模型优化到一定程度，也就是说模型预测正确率提高是有一定限度的，且gamma参数通常精度会处于10-2以上，利用交叉验证很难调整到合适的参数，使得模型的预测效果不能达到最佳。

这里我们使用遗传算法（GA）对C和gamma两个参数进行选择（GA和SVM结合的具体描述已在方法陈述中详细说明），原因在于在数据大小可以接受的范围内，对于当前种群所随机的超参数在迭代过程中可以自行淘汰和选择，又由于根据种群大小，参数可以有很多种选择，相比传统交叉验证来说更容易找最为适合的参数。

GA和SVM结合的根本逻辑在于将SVM的在训练集上的测试准确度作为GA算法的目标函数，以在一次次迭代中优化选择参数。

* + - 1. 模型结果分析

图表, 折线图

描述已自动生成在经过12次迭代后模型目标函数值变化如图（1.2.2.2-1）所示

图（1.2.2.2-1）

可以看到在迭代到5代时，种群的目标函数的平均值已经到达了0.95是个相当可观的是数字，而这个种群最佳的目标函数值在模型迭代初期随机产生时就已经接近0.97左右，证明了模型在训练集上的优异表现

表（1.2.2.2-2）为模型迭代次数中的目标函数值的变化情况

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| gen | eval | f\_opt | f\_max | f\_avg | f\_min | f\_std |
| 0 | 50 | 0.974542 | 0.974542 | 0.669073 | 0.625275 | 0.105095 |
| 1 | 100 | 0.974542 | 0.96978 | 0.699857 | 0.625275 | 0.130396 |
| 2 | 150 | 0.974725 | 0.974725 | 0.753538 | 0.625275 | 0.156165 |
| 3 | 200 | 0.974725 | 0.974725 | 0.834319 | 0.625275 | 0.152626 |
| 4 | 250 | 0.974725 | 0.974725 | 0.945319 | 0.796154 | 0.0318183 |
| 5 | 300 | 0.974725 | 0.974725 | 0.951084 | 0.873993 | 0.0222437 |
| 6 | 350 | 0.974725 | 0.974725 | 0.942344 | 0.625275 | 0.0678664 |
| 7 | 400 | 0.974725 | 0.974725 | 0.928963 | 0.625275 | 0.0912569 |
| 8 | 450 | 0.974725 | 0.974725 | 0.947396 | 0.625275 | 0.0668796 |
| 9 | 500 | 0.974725 | 0.974725 | 0.944465 | 0.625275 | 0.0675292 |
| 10 | 550 | 0.974725 | 0.974725 | 0.944429 | 0.625275 | 0.0662672 |
| 11 | 600 | 0.974725 | 0.974725 | 0.936443 | 0.625275 | 0.0796977 |
| 12 | 650 | 0.974725 | 0.974725 | 0.95193 | 0.625275 | 0.0482395 |

表（1.2.2.2-2）

从表中可以看到种群迭代到后期的最佳目标参数选择已经到达一个稳定的数值，而由于种群差异性，平均值处于一个波动的范围当中，由此，我们选择的C和gamma参数如表（1.2.2.2-3）所示

|  |  |
| --- | --- |
| 最优参数/结果 | 数值 |
| 目标函数值 | 0.974725275 |
| C | 256 |
| gamma | 0.00390625 |

表（1.2.2.2-3）

从表（1.2.2.2-3）中可以看出，该参数在训练集中的准确率已达到97.47%，该准确率在所有用到的模型中已经处于一个比较优异的水平。在测试集的结果如表（1.2.2.2-4）所示。

|  |  |
| --- | --- |
| 结果参数 | 值 |
| 验证集查准率 | 100.00% |
| 验证集查全率 | 96.83% |
| 验证集F1值 | 98.39% |

表（1.2.2.2-4）

从表（1.2.2.2-4）中可以得到模型的查准率为100%，说明没有乳腺癌阴性患者被错分类为阳性，而查全率为96.83%则说明存在个例阳性患者被错分类为阴性，在实际医疗诊断中，阳性被错分类为阴性可能导致病人的治疗时间的拖延，会造成严重后果，这样说明了模型若用于实际诊断还有优化的空间，且只能作为一个辅助诊断手段。F1值98.39%则说明两者平衡下来的效果是不错的。同时，我们针对SVM超参数选择对模型性能提升方面可以从其防止过拟合角度考虑，SVM通过增加正则化项，如软间隔来避免过拟合，使得SVM在训练集上可以有较好的性能，并且能在未见过的数据集上保持较好的泛化能力。这里我们选择超参数C的值为256该惩罚程度一定上保证了在训练集上分类的准确性，在测试集上优异的测试性能也证明了超参数C并没有导致过拟合的问题发生。gamma值选择0.0039在该数据集下也使得每个样本点互相之间的影响在一个合理范围内。

我们为了验证超参数是否导致在该数据集上过拟合，我们后续将训练集和测试集五五分成、三七分成，（超参数C训练后分别选择25，179）F1值同样可以分别达到96.45%、95.33%，处于一个较高的水平，两个测试结果为我们证明了该模型并没有存在过拟合的问题。

* 1. **朴素贝叶斯**
     1. 模型初步分析

这是一个二分类问题，对于数据集中的噪声，朴素贝叶斯分类器可以很好的避免其造成的影响，属于模型在这个数据集中的优势之一，而由于该数据集很小，在各个模型下的训练，预测下都有着较快的速度，无法体现朴素贝叶斯分类器在快速训练和预测速度上的优势，当然由于数据集很小，朴素贝叶斯使用概率模型，更好利用了有限的数据，因此可以给出较好的结果。该数据集共有30个维度，朴素贝叶斯在此高维数据集下由于假设各个特征维度之间相互独立，可以降低维度灾难的影响。

* + 1. 模型结果分析

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.94 | 0.95 | 0.95 | 108 |
| 1 | 0.92 | 0.9 | 0.91 | 63 |
| accuracy |  |  | 0.94 | 171 |
| macro avg | 0.93 | 0.93 | 0.93 | 171 |
| weighted avg | 0.94 | 0.94 | 0.94 | 171 |

图表, 树状图

描述已自动生成表（1.3.2-1）

图（1.3.2-2）

|  |  |  |
| --- | --- | --- |
| type | train | test |
| 0 | 0.625628 | 0.631579 |
| 1 | 0.374372 | 0.368421 |

表（1.3.2-3）

表（1.3.2-1）为朴素贝叶斯分类器在此模型下的表现，其最终准确率为0.935672514619883，从图（1.3.2-2）混淆矩阵也能得到结论模型本身来说是一个较为良好的水平，但相比之前使用的SVM分类器性能要差上一些这里我们对其原因进行一些分析。

首先，朴素贝叶斯分类器假设特征之间相互独立，在预处理的特征向量相关图中我们可以发现一些特征向量存在相关性，也就是说相互独立这个假设实际上是不成立的，这个方面可能会导致朴素贝叶斯分类器的性能的下降。其次，我们考虑到验证集测试集数据不平衡可能会对朴素贝叶斯性能产生影响，但在预处理过程中，我们验证了数据集和测试集之间是平衡的如表（1.3.2-3），故不存在该情况产生的影响。最后，比较SVM分类器，朴素贝叶斯分类器相较来说对噪声更为敏感一些，使得其性能不如SVM分类器。

* 1. 逻辑回归
  2. 决策树
  3. 随机森林

**2. 请将问题1中的分类器应用至下述数据集(二选一)**

(1) 教材第四、五章中其余6个数据任选三个

(2) Gas、Optical、Vertebral数据集(数据集见附件二)

强调分析不同分类器在不同数据集下的分类结果.

·数据预处理

* 1. SVM
     1. 图表, 折线图

        描述已自动生成数据集Gas表现

图（2.1.1-1）

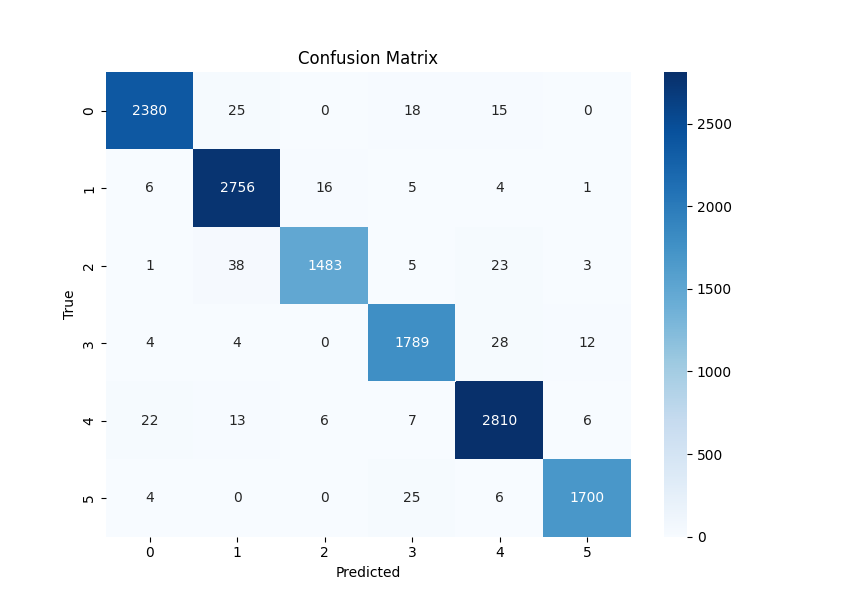
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| gen | eval | f\_opt | f\_max | f\_avg | f\_min | f\_std |
| 0 | 50 | 0.974215 | 0.974215 | 0.765808 | 0.661775 | 0.084724 |
| 1 | 100 | 0.974215 | 0.974215 | 0.782943 | 0.208635 | 0.121709 |
| 2 | 150 | 0.974215 | 0.974215 | 0.764341 | 0.208635 | 0.142324 |
| 3 | 200 | 0.974275 | 0.974275 | 0.813611 | 0.660326 | 0.089269 |
| 4 | 250 | 0.974275 | 0.974275 | 0.883663 | 0.732548 | 0.074192 |
| 5 | 300 | 0.977174 | 0.977174 | 0.896091 | 0.208635 | 0.115274 |
| 6 | 350 | 0.977174 | 0.975725 | 0.905279 | 0.586836 | 0.072746 |
| 7 | 400 | 0.977174 | 0.975725 | 0.915667 | 0.705012 | 0.055133 |
| 8 | 450 | 0.977174 | 0.977114 | 0.930570 | 0.821981 | 0.037391 |
| 9 | 500 | 0.978563 | 0.978563 | 0.932746 | 0.709360 | 0.041439 |
| 10 | 550 | 0.978563 | 0.978563 | 0.925806 | 0.735386 | 0.048587 |
| 11 | 600 | 0.978563 | 0.974275 | 0.934233 | 0.821981 | 0.028511 |
| 12 | 650 | 0.978563 | 0.974275 | 0.931943 | 0.682065 | 0.043917 |
| 13 | 700 | 0.978563 | 0.974275 | 0.933545 | 0.818659 | 0.039668 |
| 14 | 750 | 0.978563 | 0.974275 | 0.937030 | 0.739734 | 0.044462 |
| 15 | 800 | 0.978563 | 0.975725 | 0.902665 | 0.208635 | 0.119352 |
| 16 | 850 | 0.978563 | 0.975725 | 0.887860 | 0.208635 | 0.151423 |
| 17 | 900 | 0.978563 | 0.977174 | 0.887418 | 0.208635 | 0.152284 |
| 18 | 950 | 0.978563 | 0.977174 | 0.906812 | 0.739734 | 0.065377 |
| 19 | 1000 | 0.978563 | 0.974215 | 0.899761 | 0.208635 | 0.120832 |
| 20 | 1050 | 0.978563 | 0.971377 | 0.902320 | 0.208635 | 0.118109 |
| 21 | 1100 | 0.978563 | 0.971377 | 0.862443 | 0.208635 | 0.176612 |
| 22 | 1150 | 0.978563 | 0.972766 | 0.906797 | 0.702597 | 0.069797 |
| 23 | 1200 | 0.980012 | 0.980012 | 0.922505 | 0.729650 | 0.052874 |
| 24 | 1250 | 0.980012 | 0.980012 | 0.924052 | 0.806099 | 0.049105 |
| 25 | 1300 | 0.980012 | 0.979952 | 0.920430 | 0.818659 | 0.047841 |
| 26 | 1350 | 0.980012 | 0.979952 | 0.903758 | 0.208635 | 0.150861 |
| 27 | 1400 | 0.980012 | 0.978563 | 0.909437 | 0.208635 | 0.148833 |
| 28 | 1450 | 0.980012 | 0.980012 | 0.909771 | 0.443599 | 0.092281 |
| 29 | 1500 | 0.980012 | 0.980012 | 0.917089 | 0.729650 | 0.070470 |

表（2.1.1-2）

|  |  |
| --- | --- |
| 最优参数/结果 | 数值 |
| 目标函数值 | 0.980012077294686 |
| C | 214.37392618879676 |
| gamma | 0.2543907091021538 |

表（2.1.1-3）

|  |  |
| --- | --- |
| 结果参数 | 值 |
| 验证集查准率 | 97.84% |
| 验证集查全率 | 97.58% |
| 验证集F1值 | 97.70% |

表（2.1.1-4）

图（2.1.1-5）

在数据集Gas当中，因为数据量在14000左右，若使用GA算法帮助SVM调参的化会出现因时间复杂度过高而导致的计算时间过长，无法在有限的时间内获得我们需要的结果，故我们把总数据集5%划分为训练集，其余95%划分为数据集。最后得到的结果：图（2.1.1-1）为GA调参且以SVM在训练集上准确率为目标函数值时选取的参数最大正确率和评价正确率随迭代次数的变化折线图。表（2.1.1-2）则为迭代过程中目标函数值的详细记录，表（2.1.1-3）为经过迭代后选取的SVM的C和gamma值，表（2.1.1-4）为该参数下rbf核的SVM在测试集下的测试效果。图（2.1.1-5）为混淆矩阵。

* + 1. 图表, 折线图

       描述已自动生成数据集Optical表现

图（2.1.2-1）

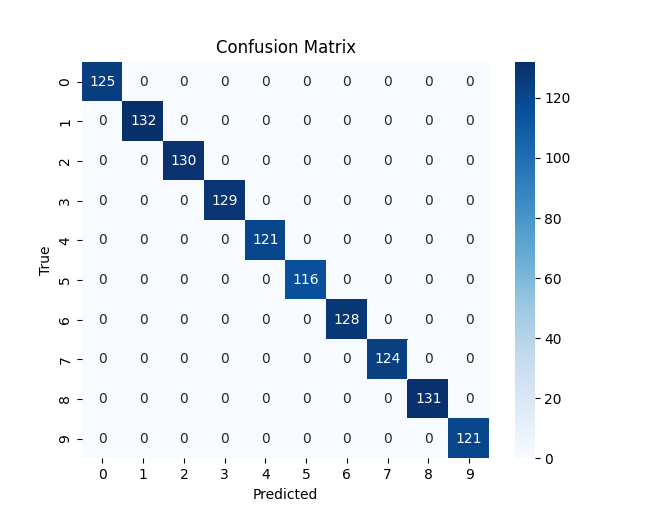
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| gen | eval | f\_opt | f\_max | f\_avg | f\_min | f\_std |
| 0 | 50 | 0.106581 | 0.106581 | 0.095740 | 0.095470 | 0.001560 |
| 1 | 100 | 0.983314 | 0.983314 | 0.114005 | 0.095470 | 0.124221 |
| 2 | 150 | 0.983314 | 0.983314 | 0.148963 | 0.095470 | 0.210801 |
| 3 | 200 | 0.983314 | 0.983314 | 0.237445 | 0.095470 | 0.325263 |
| 4 | 250 | 0.983314 | 0.983314 | 0.351703 | 0.095470 | 0.396468 |
| 5 | 300 | 0.983314 | 0.983314 | 0.387943 | 0.095470 | 0.409675 |
| 6 | 350 | 0.983314 | 0.983314 | 0.524902 | 0.095470 | 0.436816 |
| 7 | 400 | 0.983314 | 0.983314 | 0.892352 | 0.095470 | 0.252824 |
| 8 | 450 | 0.983314 | 0.983314 | 0.900566 | 0.096264 | 0.241272 |
| 9 | 500 | 0.983314 | 0.983314 | 0.951082 | 0.095470 | 0.131613 |
| 10 | 550 | 0.983314 | 0.983314 | 0.860785 | 0.095470 | 0.285155 |
| 11 | 600 | 0.983314 | 0.983314 | 0.834142 | 0.095470 | 0.321899 |

表（2.1.2-2）

|  |  |
| --- | --- |
| 最优参数/结果 | 数值 |
| 目标函数值 | 0.9833139759969028 |
| C | 247.41419351100922 |
| gamma | 0.00390625 |

表（2.1.2-3）

|  |  |
| --- | --- |
| 结果参数 | 值 |
| 验证集查准率 | 100.00% |
| 验证集查全率 | 100.00% |
| 验证集F1值 | 100.00% |

表（2.1.2-4）

图（2.1.2-5）

在数据集Optical中，由于每个数字的特征为一个特征矩阵，对于GA调参的SVM来说处理起来时间复杂度很高，故在每一轮迭代上都花费了近10分钟的时间，从图（2.1.2-1）中可以看出，起初几代的平均准确率并不理想，但在后面几代参数优化的很快。其中图（2.1.2-1）为GA调参且以SVM在训练集上准确率为目标函数值时选取的参数最大正确率和评价正确率随迭代次数的变化折线图。表（2.1.2-2）则为迭代过程中目标函数值的详细记录，表（2.1.2-3）为经过迭代后选取的SVM的C和gamma值，表（2.1.2-4）为该参数下rbf核的SVM在测试集下的测试效果。图（2.1.2-5）为混淆矩阵。

* + 1. 数据集Vertebral表现

图表, 折线图

描述已自动生成图表, 折线图

描述已自动生成图（2.1.3-1） 图（2.1.3-2）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| gen | eval | f\_opt | f\_max | f\_avg | f\_min | f\_std |
| 0 | 50 | 0.650595 | 0.650595 | 0.650595 | 0.650595 | 0.000000 |
| 1 | 100 | 1.000000 | 1.000000 | 0.657583 | 0.650595 | 0.048917 |
| 2 | 150 | 1.000000 | 1.000000 | 0.671464 | 0.650595 | 0.082604 |
| 3 | 200 | 1.000000 | 1.000000 | 0.685440 | 0.650595 | 0.104537 |
| 4 | 250 | 1.000000 | 1.000000 | 0.726917 | 0.650595 | 0.143756 |
| 5 | 300 | 1.000000 | 1.000000 | 0.767750 | 0.650595 | 0.163341 |
| 6 | 350 | 1.000000 | 1.000000 | 0.877917 | 0.650595 | 0.163446 |
| 7 | 400 | 1.000000 | 1.000000 | 0.975750 | 0.650595 | 0.082749 |
| 8 | 450 | 1.000000 | 1.000000 | 0.960679 | 0.650595 | 0.103903 |
| 9 | 500 | 1.000000 | 1.000000 | 0.983024 | 0.650595 | 0.068508 |
| 10 | 550 | 1.000000 | 1.000000 | 0.929393 | 0.650595 | 0.138186 |
| 11 | 600 | 1.000000 | 1.000000 | 0.946333 | 0.650595 | 0.118346 |

表（2.1.3-3）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| gen | eval | f\_opt | f\_max | f\_avg | f\_min | f\_std |
| 0 | 50 | 0.458929 | 0.458929 | 0.458929 | 0.458929 | 0.000000 |
| 1 | 100 | 0.458929 | 0.458929 | 0.458929 | 0.458929 | 0.000000 |
| 2 | 150 | 0.458929 | 0.458929 | 0.458929 | 0.458929 | 0.000000 |
| 3 | 200 | 0.995238 | 0.995238 | 0.469655 | 0.458929 | 0.075083 |
| 4 | 250 | 0.995238 | 0.995238 | 0.512560 | 0.458929 | 0.160893 |
| 5 | 300 | 0.995238 | 0.995238 | 0.587548 | 0.458929 | 0.228880 |
| 6 | 350 | 0.995238 | 0.995238 | 0.727083 | 0.458929 | 0.268155 |
| 7 | 400 | 0.995238 | 0.995238 | 0.970250 | 0.458929 | 0.100727 |
| 8 | 450 | 0.995238 | 0.995238 | 0.937976 | 0.458929 | 0.157534 |
| 9 | 500 | 0.995238 | 0.995238 | 0.929583 | 0.458929 | 0.173974 |
| 10 | 550 | 0.995238 | 0.995238 | 0.960476 | 0.458929 | 0.127142 |
| 11 | 600 | 0.995238 | 0.995238 | 0.990071 | 0.939881 | 0.014312 |

表（2.1.3-4）

|  |  |
| --- | --- |
| 最优参数/结果 | 数值 |
| 目标函数值 | 1.0 |
| C | 135.31434154510498 |
| gamma | 0.00390625 |

表（2.1.3-5）

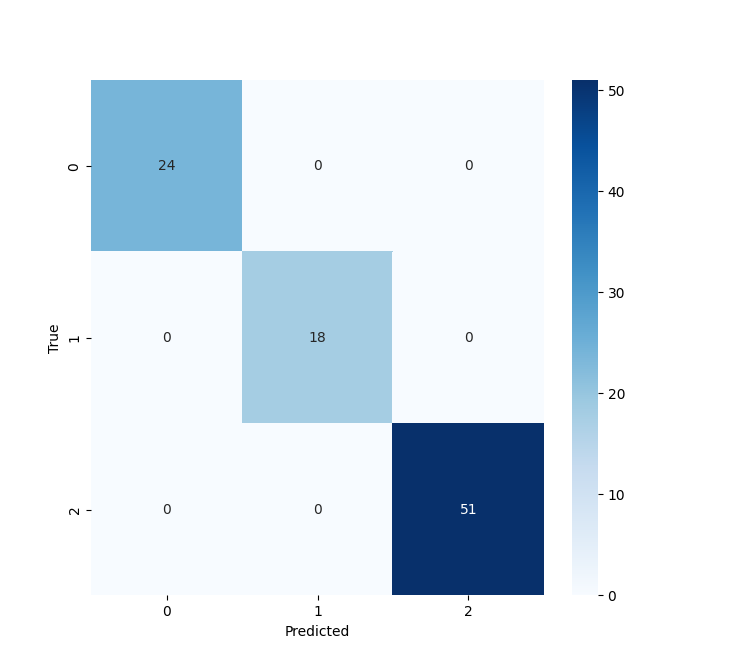
|  |  |
| --- | --- |
| 结果参数 | 值 |
| 验证集查准率 | 100.00% |
| 验证集查全率 | 100.00% |
| 验证集F1值 | 100.00% |

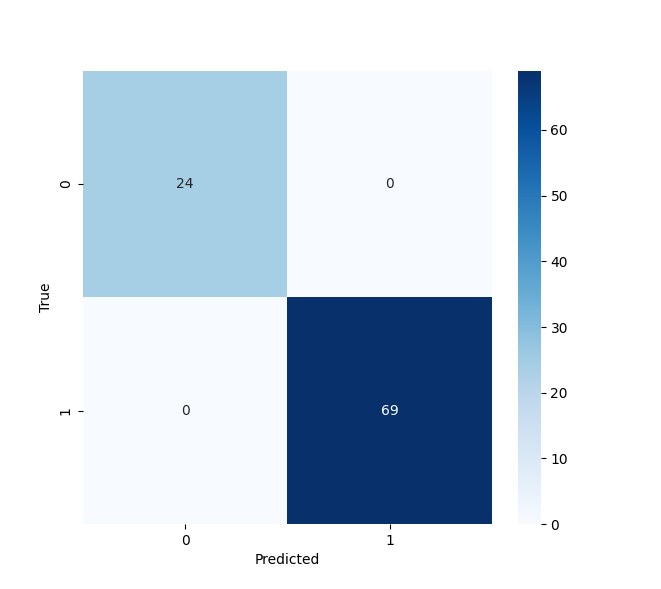
表（2.1.3-6）

|  |  |
| --- | --- |
| 最优参数/结果 | 数值 |
| 目标函数值 | 0.9952380952380953 |
| C | 204.21197187528014 |
| gamma | 0.00390625 |

表（2.1.3-7）

|  |  |
| --- | --- |
| 结果参数 | 值 |
| 验证集查准率 | 100.00% |
| 验证集查全率 | 100.00% |
| 验证集F1值 | 100.00% |

表（2.1.3-8）

图（2.1.3-9） 图（2.1.3-10）

在Vertebral数据集上，经过GA算法调参的SVM最后在测试集上的表现都达到了完美的程度，均分类正确。其中图（2.1.3-1）和图（2.1.3-2）分别为二分类数据集，三分类数据集下GA调参SVM返回训练集上正确率（目标函数值）的变化。表（2.1.3-3）和表（2.1.3-4）分别为二分类，三分类数据集迭代过程中目标函数值的详细记录，表（2.1.3-5）和表（2.1.3-6）为二分类，三分类数据集经过迭代后选取的SVM的C和gamma值，表（2.1.2-7）和表（2.1.3-8）为该参数下rbf核的SVM在二分类，三分类测试集下的测试效果。图（2.1.3-9）和图（2.1.3-10）分别为二分类和三分类的混淆矩阵

* + 1. 总结

总体上rbf核函数下的SVM经过GA遗传算法调参在各个数据集下都有着十分优异的表现，其在Vertebral和Optical数据集上的正确率全部都达到了100%，足以证明SVM分类器的泛化性，在合适的C超参数值和gamma值下，SVM分类器可以在当前数据集下发挥其极为优秀的性能。

在Gas数据集下，其F1值如表（2.1.1-4）所示高达97.70%，处于一个较为优异的范围，并且由于训练集相较测试集来说小很多，故不存在过拟合的可能性存在，由此可以验证该参数下，对于该数据集泛化性的优异，当然其正确率没有达到100%的一部分原因想必在于其训练集相比测试机来说小太多，可能存在测试集和训练集之间的差异性，这是可以通过调整训练集测试集划分优化的，但这必然要牺牲时间，对于再大一些的数据集，我们就更难通过此类方法找到SVM合适的参数了，这也体现了SVM的局限性，即C超参数和gamma超参数对于模型的性能影响很大，当前主流的调参方式如果为了节省时间，可供选择的参数列表是很小的，很难找到一个契合模型的超参数，而如果用GA之类的调参方法，时间复杂度是无法接受的，这是我们发现的SVM的一个短板，即在数据集很大的情况下，我们很难找到一个合适的参数使得SVM的性能优化。

针对Optical数据集，我们可以发现其在GA调参的初始阶段目标函数值（测试集预测正确率）处于一个较低的水平，可以见得对于矩阵这一类别的特征处理的时候参数调整的难度，一轮迭代接近10分钟的时间也可见得SVM在处理矩阵上时间方面的缺陷，但在迭代过程中，随着我们参数的调整，平均目标函数值（测试集上正确率）上升的速度很快，也进一步验证了一个好的超参数对于SVM性能提升方面的影响。

在小数据集上，也就是Vertebral数据集上，SVM在平均10轮不到的参数迭代中，其目标函数值即可达到一个很高的水平，可以见得SVM在小数据集上也有着不错的棒鲁性和泛用性，这里我们分析原因可能在于对于小数据集来说，SVM可更加容易地找到一个较小的超平面，从而更好地进行分类。同时，对于具有高维数据的小数据集来说，SVM可以用核函数将高维数据映射到一个更高维的特征空间，从而更容易找到一个较好的超平面来实现分类。且针对小数据集的非线性问题，核函数也可以将非线性问题转化为线性问题，从而更容易找到一个较好的超平面进行分类。同时，在小数据集上SVM可以有效避免过拟合，由于SVM通过最大化边界来寻找最优超平面，它对于噪声和异常值的容忍度相对较高，从而能够避免过拟合。

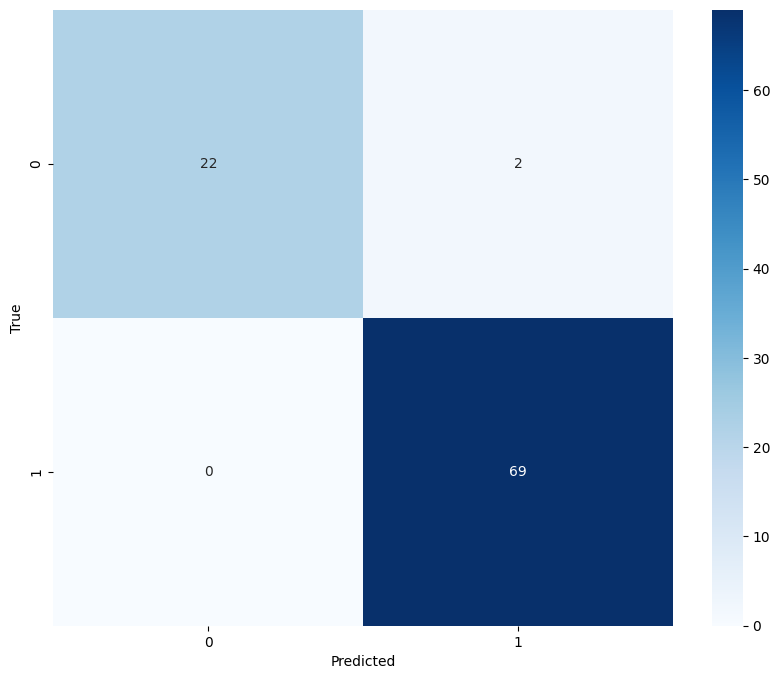
* 1. 朴素贝叶斯
     1. 数据集Vertebral表现

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 1 | 0.92 | 0.96 | 24 |
| 1 | 0.97 | 1 | 0.99 | 69 |
| accuracy |  |  | 0.98 | 93 |
| macro avg | 0.99 | 0.96 | 0.97 | 93 |
| weighted avg | 0.98 | 0.98 | 0.98 | 93 |

表（2.2.1-1）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| DH | 0.67 | 0.56 | 0.61 | 18 |
| NO | 0.7 | 0.79 | 0.75 | 24 |
| SL | 0.98 | 0.98 | 0.98 | 51 |
| accuracy |  |  | 0.85 | 93 |
| macro avg | 0.78 | 0.78 | 0.78 | 93 |
| weighted avg | 0.85 | 0.85 | 0.85 | 93 |

图表

描述已自动生成表（2.2.1-2）

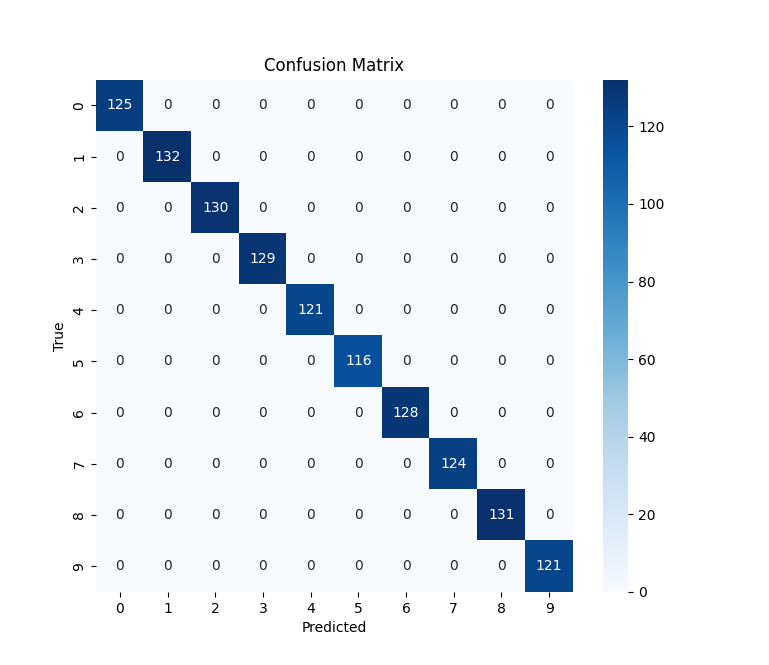
图（2.2.1-3） 图（2.2.1-4）

其中表（2.2.1-1）和表（2.2.1-2）分别为脊椎病数据集下的二分类，三分类对应的分类报告，图（2.2.1-3）和图（2.2.1-4）为其混淆矩阵，可以从中见得其在二分类下的正确率有接近98%的水平，而在三分类下的分类正确率只有85%的水平，具体原因我们在2.2.4中进一步阐述。

* + 1. 数据集Optical表现

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.97 | 1 | 0.98 | 32 |
| 1 | 0.83 | 0.67 | 0.74 | 30 |
| 2 | 0.86 | 0.58 | 0.69 | 31 |
| 3 | 0.94 | 0.71 | 0.81 | 42 |
| 4 | 0.94 | 0.86 | 0.9 | 37 |
| 5 | 0.89 | 0.87 | 0.88 | 38 |
| 6 | 0.98 | 0.98 | 0.98 | 44 |
| 7 | 0.7 | 0.97 | 0.81 | 34 |
| 8 | 0.51 | 0.9 | 0.65 | 40 |
| 9 | 1 | 0.56 | 0.72 | 32 |
| accuracy |  |  | 0.82 | 360 |
| macro avg | 0.86 | 0.81 | 0.82 | 360 |
| weighted avg | 0.86 | 0.82 | 0.82 | 360 |

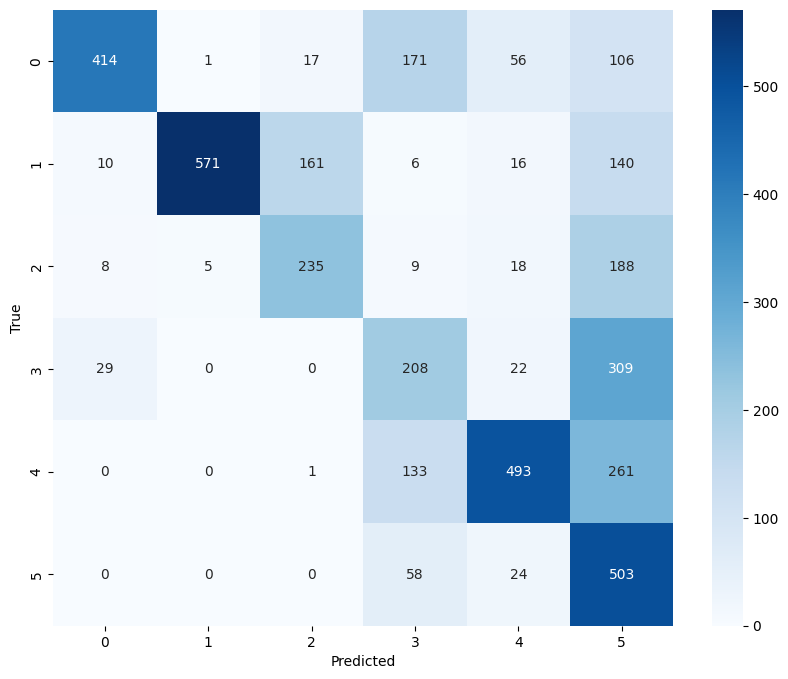
表（2.2.2-1）

图（2.2.2-2）

其中表（2.2.2-1）分别为光学数字识别数据集下对应的分类报告，图（2.2.2-2）为其混淆矩阵，可以从中见得其正确率只有接近82%的水平，具体原因们在2.2.4中进一步阐述。

* + 1. 数据集Gas表现

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 1 | 0.9 | 0.54 | 0.68 | 765 |
| 2 | 0.99 | 0.63 | 0.77 | 904 |
| 3 | 0.57 | 0.51 | 0.54 | 463 |
| 4 | 0.36 | 0.37 | 0.36 | 568 |
| 5 | 0.78 | 0.56 | 0.65 | 888 |
| 6 | 0.33 | 0.86 | 0.48 | 585 |
| accuracy |  |  | 0.58 | 4173 |
| macro avg | 0.65 | 0.58 | 0.58 | 4173 |
| weighted avg | 0.7 | 0.58 | 0.61 | 4173 |

表（2.2.3-1）

图（2.2.3-2）

其中表（2.2.3-1）分别为气体识别数据集下对应的分类报告，图（2.2.3-2）为其混淆矩阵，可以从中见得其正确率很低只有接近58%的水平，算是朴素贝叶斯分类器的一个短板具体原因们在2.2.4中进一步阐述。

* + 1. 总结

由最后朴素贝叶斯模型在三个数据集下的表现结果可以发现，相较调整好参数的SVM模型来说，正确率均没有达到一个理想的水平，其只有在Vertebral二分类数据集下准确率（98%）可以达到接近SVM的水平。而在多分类问题中，SVM正确率均在85%以下。

在拥有128维特征向量的Gas数据集其正确率出现了明显下滑只有58%，可以发现朴素贝叶斯在高维数据下出现了很明显的缺陷。具体原因我们经过分析，由预处理得到的特征值关联矩阵的展示下我们可以很快否定掉朴素贝叶斯的先验性假设，即各个特征值之间是相互独立的，这个缺陷在高维数据下展现的更加明显，故出现了准确率只有58%的场面，当然在128维特征向量下，可能存在许多无关的或冗余的特征。这些特征可能会干扰模型的学习和预测能力，导致预测准确率降低。这里经过检验也确定了不存在训练集和测试集之间不平衡的情况，在这个大数据集下，我们朴素贝叶斯利用了70%的数据作为训练集，也不存在训练集不足的情况，从中也能看出朴素贝叶斯的一个有点，即在大数据集下的计算速度相较SVM还是有着很大的优势的，在满足朴素贝叶斯假设的前提下，朴素贝叶斯分类器是个不错的选择。

在Optical数据集下，我们发现朴素贝叶斯分类器的准确率有82%，根据混淆矩阵可以得知，我们在测试集下出现了很多将3误判成8的例子，针对这一现象，我们一是可以归结为3和8两个数字的特征矩阵有很多相似的地方，这个对于朴素贝叶斯来说难以区分，也符合实际情况下3误认为是8。在其他数字上，朴素贝叶斯还是保持了一个不错的区分度，同时，朴素贝叶斯运行时间较SVM短很多，属于其模型上的优势，但缺陷也很明显，对于相似的两类，朴素贝叶斯的区分性能并不是很好。

在Vertebral数据集下，二分类准确率威98%而三分类只有82%这个原因可能在于在二分类任务中，两个类别之间的区分度可能相对较高，导致模型可以较好地进行分类。但在三分类任务中，不同类别之间的区分度可能存在差异，其中一些类别之间的区分度较低，导致模型在这些类别上的预测准确率下降。同时朴素贝叶斯算法基于特征之间的条件独立性假设，但在三分类任务中，特征之间的相关性可能更复杂，无法满足条件独立性假设。这可能导致模型在三分类任务中的预测准确率下降。

* 1. 逻辑回归
     1. 数据集Gas表现
     2. 数据集Optical表现
     3. 数据集Vertebral表现
     4. 总结
  2. 决策树
     1. 数据集Gas表现
     2. 数据集Optical表现
     3. 数据集Vertebral表现
     4. 总结
  3. 随机森林
     1. 数据集Gas表现
     2. 数据集Optical表现
     3. 数据集Vertebral表现
     4. 总结

## 【4】案例代码

（请将案例代码粘贴在此处, 并给出必要注释）

**案例一代码**

**案例二代码**

附表1 参与人员

项目负责人及参与人员情况表

| 序号 | 姓名 | 性别 | 学号 | 专业 | 投入时间占比 | 任务分工 | 项目中职务 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 许廷阳 | 男 | 922127940155 | 数据科学与大数据技术 |  |  | 项目负责人 |
| 2 | 谭清海 | 男 |  | 数据科学与大数据技术 |  |  | 骨干 |
| 3 |  | 男 |  | 数据科学与大数据技术 |  |  | 骨干 |

备注：本表需如实填写, 可注明共同完成项目内容, 将作为成绩评判的重要依据.