广东工业大学华立学院

机器学习课程论文

题目: 客户信用分类预测贝叶斯优化

学　　院　国际学院

专　　业　计算机科学与技术

班　　级　22 计算机 G1 班

组　　员　蒋国威 5112221011512

单嘉乐 5112221011503

李俊成 5112221011504

陈泽豪 5112221011510

吕承骏 5112221011501

指导教师　林睿翀

2024 年 6 月 25 日

报告：客户信用评分数据集分类预测与算法组合

摘要：

本项目旨在金融领域中提高客户信用评分的准确性，以辅助金融机构做出更明智的贷款批准决策。研究的核心是通过贝叶斯算法对客户信用评分进行预测，并探索将多个贝叶斯分类器与随机森林算法结合的方法来提升预测性能。项目首先进行了数据预处理和探索性分析，包括数据清洗、特征选择、缺失值处理，并使用可视化手段分析了关键特征与信用风险的关系。随后，项目采用了贝叶斯分类器进行初步预测，并进一步通过优化，将贝叶斯分类器的预测结果作为新特征输入到随机森林模型中，实现模型的集成学习。这种方法不仅利用了贝叶斯分类器的特征提取能力，也借助随机森林的泛化能力，有效提升了模型的抗过拟合能力和分类精度。研究发现，通过算法组合，模型在准确度、精确度和召回率等评价指标上均有显著提升，为金融领域的信用风险管理提供了有力的技术支持和策略指导。

1. 背景和目标

在金融领域，客户信用评分是一项重要的任务。准确地预测客户是否会违约可以帮助金融机构做出明智的贷款批准决策，从而降低信用风险。本报告旨在探讨如何利用贝叶斯算法进行客户信用评分预测，并通过将多个贝叶斯分类器与随机森林结合起来，提高整体性能。

客户信用评分数据集是一个经典的数据集，用于研究信用评分和风险分析。该数据集包含了申请贷款的个人客户的相关信息，以及一个二元分类标签，用于表示该客户是否违约（"+"表示违约，"-"表示没有违约）。

数据集中的特征包括了各种个人和贷款相关的信息，例如年龄、性别、婚姻状况、贷款金额、贷款期限、收入情况等。这些特征可以用来预测客户的信用违约风险，以帮助银行或金融机构做出贷款批准或拒绝的决策。

2. 方法学

2.1 数据预处理与探索分析：使用客户信用评分数据集，包含个人客户的相关信息和二元分类标签。进行数据清洗、特征选择和缺失值处理。

（1）加载数据集：从指定的 URL 加载数据集，该数据集是关于信用评分的数据集，具有 16 个特征（A1 到 A15）和一个目标变量（class）。

（2）数据探索：

使用 sns.countplot()绘制了目标变量（class）的分布情况，展示了各个类别的样本数量。

在客户评分预测数据集中，A2 为客户年龄，A3 和 A8 代表了与客户评分相关的关键信息或特征，A3 是客户的工作年龄，而 A8 是客户的收入水平，这些因素通常与客户的信用风险或者购买能力密切相关，因此对于评估客户的信用分数

可能具有重要影响。相比之下，其他特征可能并不像 A2，A3 和 A8 那样直接与客户评分相关，或者它们的影响相对较小。因此，使用了 Matplotlib 和 Seaborn 库对 A3 和 A8 进行可视化和分析有助于更全面地理解客户评分预测模型中重要的影响因素。绘制了特征 A2 的直方图和特征 A3 与 A8 的散点图，通过可视化展示了这些特征的分布和它们之间的关系。

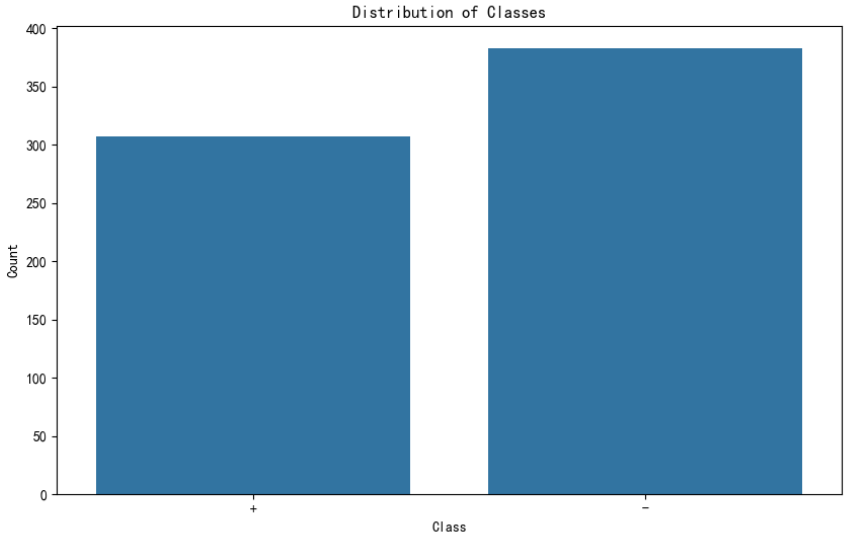

图 1 目标变量（class）的分布情况

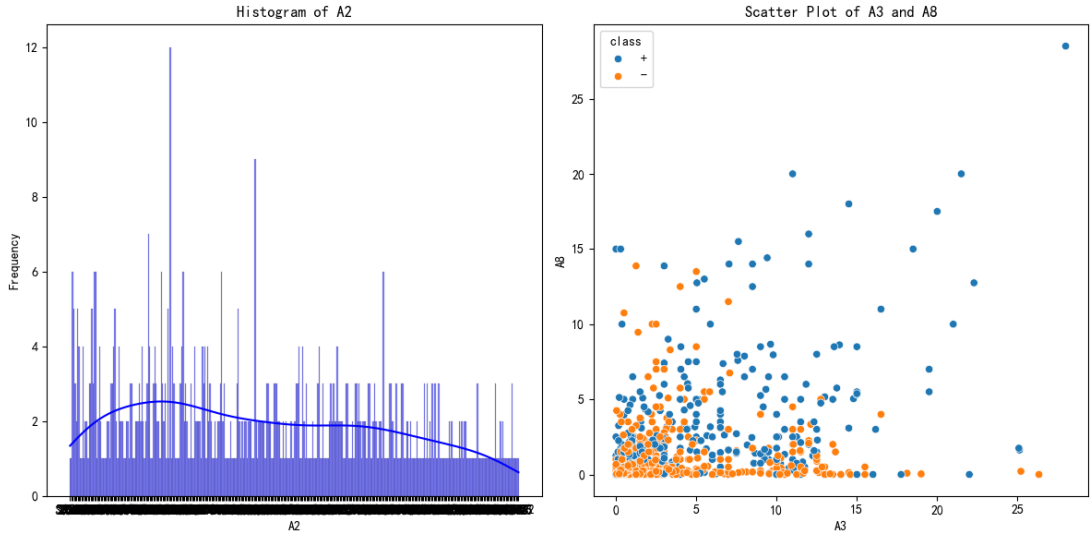

图 2 特征 A2 的直方图和特征 A3 与 A8 的散点图

（3）数据预处理：

使用 Python 中 pandas 库以及使用 replace()函数将数据中的'?'替换为 NaN，以便后续处理缺失值，使用 dropna()删除含有缺失值的样本，使用 pd.get_dummies()将分类变量转换为哑变量（One-Hot 编码），将特征数据（X）和目标变量（y）分开。

（4）划分数据

使用 train_test_split()将数据集划分为训练集和测试集，其中测试集占 20%。

2.2 贝叶斯分类器预测及优化

（1）单个贝叶斯分类器预测：

单独使用贝叶斯分类器对数据集进行分类预测，并评估准确率、精确率和召回率等标准

（2）优化：

将多个贝叶斯分类器与随机森林结合起来，以提高整体性能。可以使用随机森林作为元分类器，而贝叶斯分类器作为基分类器。这样，随机森林将综合多个贝叶斯分类器的预测结果来进行最终的分类。

随机森林被用作元分类器，而多个贝叶斯分类器被用作基分类器，它们的预测结果会被整合在一起。这种组合的方法可以利用随机森林的强大泛化能力和贝叶斯分类器的特征提取能力来提高整体性能。

将多个贝叶斯分类器与随机森林结合起来的关键在于如何将贝叶斯分类器的预测结果整合到随机森林中。下面是具体的步骤：

借助 Python 中的 scikit-learn 库进行建模和分析。

训练基分类器： 首先，对多个不同类型的贝叶斯分类器进行训练，例如高斯朴素贝叶斯、多项式朴素贝叶斯和伯努利朴素贝叶斯。每个基分类器都将独立地对数据进行建模，并生成预测结果。

生成基分类器预测： 使用训练好的贝叶斯分类器对测试数据进行预测，得到每个基分类器的预测结果。

组合预测结果： 将基分类器的预测结果作为新的特征输入到随机森林中。每个基分类器的预测结果都是一个新的特征，随机森林将学习如何有效地组合这些特征以获得最终的分类结果。

训练随机森林： 使用包含基分类器预测结果的新特征的数据训练随机森林模型。随机森林将学习如何从多个基分类器的预测中提取信息，并做出准确的分类决策。

整体预测：使用训练好的随机森林模型对新的未见过的数据进行分类预测。随机森林将综合多个贝叶斯分类器的预测结果，并基于这些结果做出最终的分类决策。

优势和工作原理：

结合多个贝叶斯分类器和随机森林的方法具有以下优势：

利用多样性： 随机森林通过训练多个不同的基分类器来引入多样性，从而提高了模型的泛化能力。贝叶斯分类器的引入可以进一步增加多样性，从而进一步提高整体性能。

充分利用特征信息： 贝叶斯分类器具有良好的特征提取能力，可以有效地利用特征之间的相关性和条件概率信息。结合随机森林的决策树模型，可以充分利用这些特征信息来进行分类。

抗过拟合能力强： 随机森林通过随机抽样和特征子集选择来减少过拟合的风险。贝叶斯分类器通常具有较好的鲁棒性和泛化能力，可以进一步增强模型的抗过拟合能力。

综上所述，将多个贝叶斯分类器与随机森林结合起来可以充分利用两种算法的优势，从而提高整体性能。这种组合方法不仅可以提高分类精度，还可以增加模型的稳定性和鲁棒性，可适用于处理复杂的数据集和分类问题。

3. 发现及数据分析可视化展示

可视化展示：绘制了单独贝叶斯分类器和算法组合的预测结果的混淆矩阵、ROC 曲线等，以直观展示性能提升。

贝叶斯分类器性能：单独使用贝叶斯分类器时，获得了一定的准确率、精确率和召回率，并且生成了混淆矩阵，但单一分类器可能存在局限性。

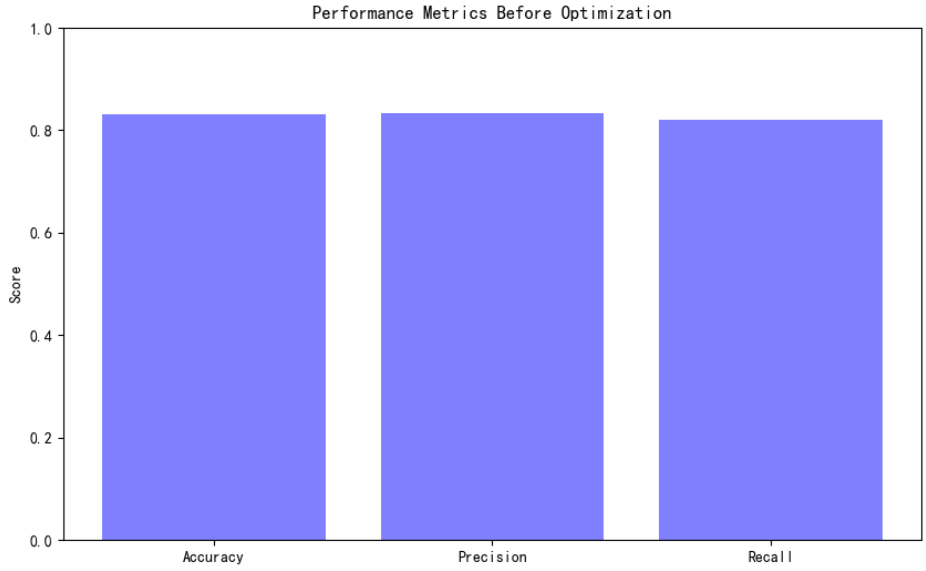

图 3 优化前评价指标

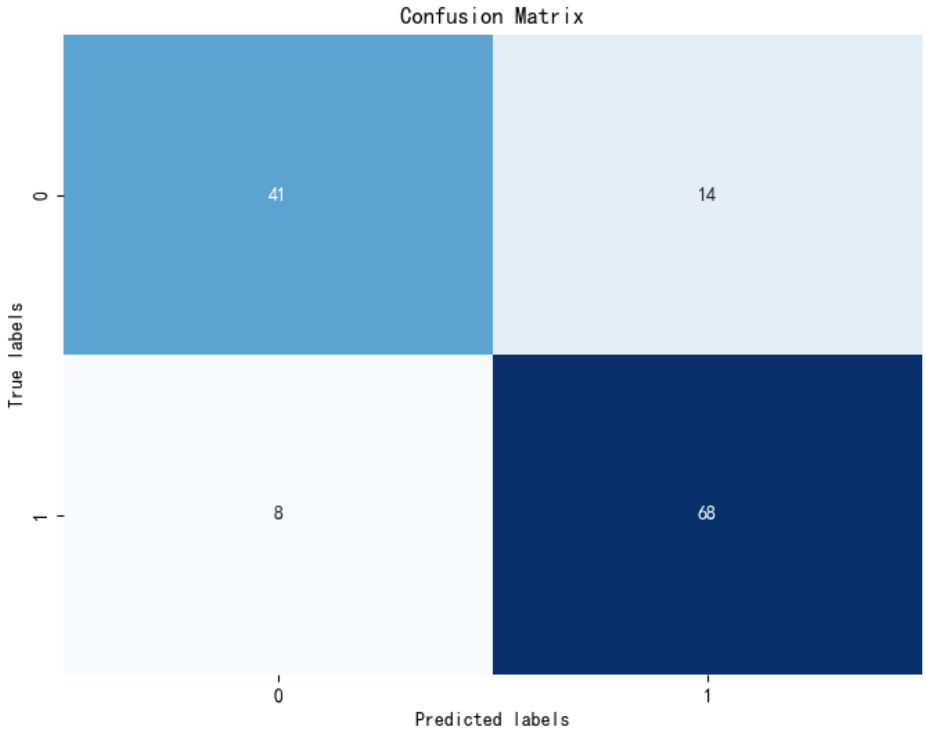

图 4 优化前混淆矩阵

算法组合性能提升：通过将多个贝叶斯分类器与随机森林结合起来，可以看到各评价指标准比如准确度与精确度均均有提高，整体模型分类性能得到了提升。组合算法综合了贝叶斯分类器的特征提取能力和随机森林的泛化能力，从而提高了分类预测的准确性和稳健性。

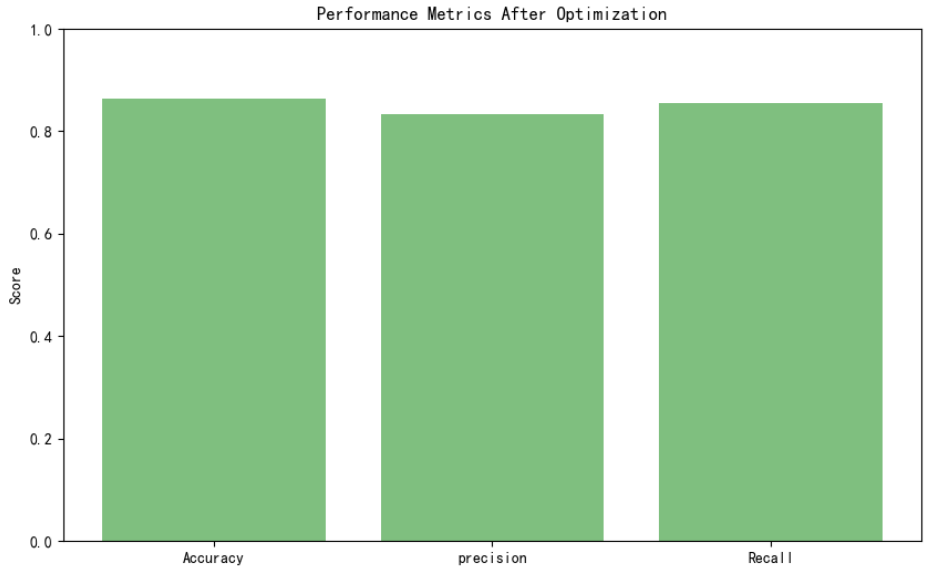

图 5 优化后评价指标

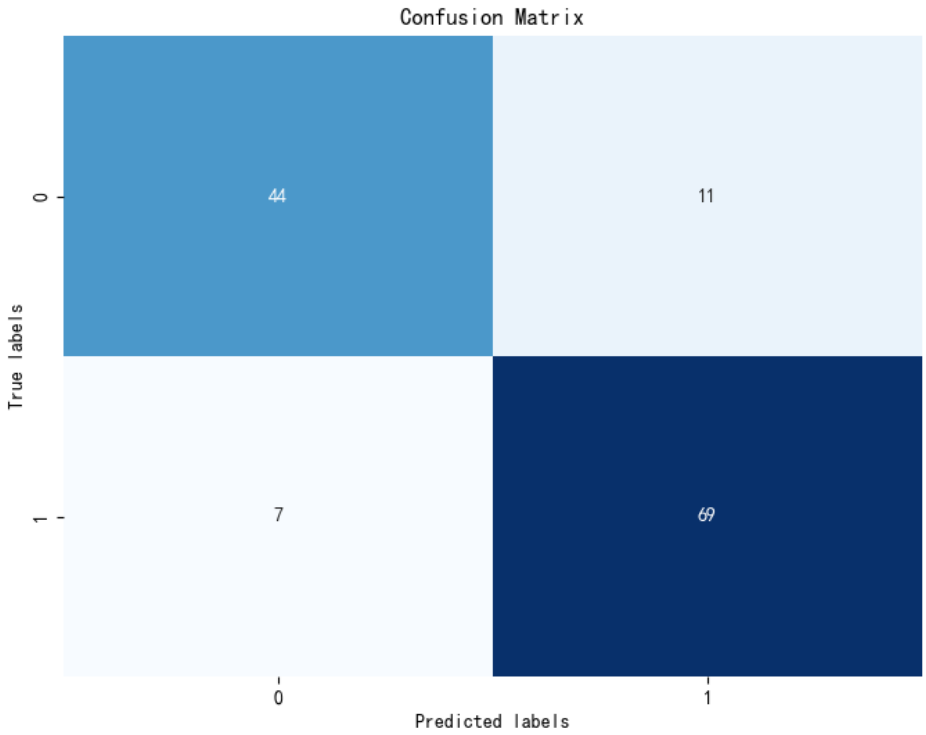

图 6 优化前混淆矩阵

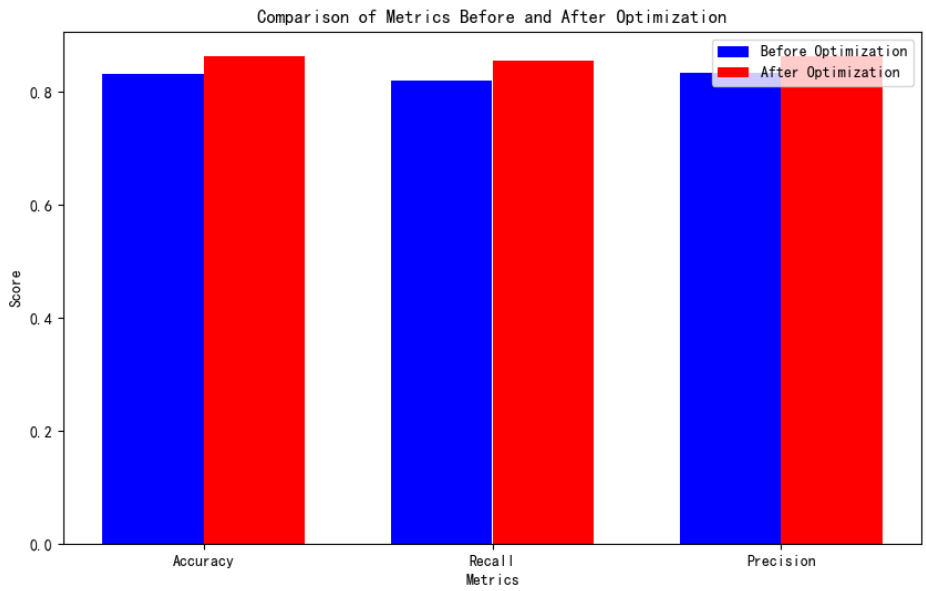

图 7 优化前后指标对比

```
[690 rows x 16 columns]
优化前的评估指标:
准确率:  0.8320610687022901
召回率:  0.8200956937799043
精确度 0.833001493280239
优化后的评估指标:
准确率:  0.8625954198473282
召回率:  0.8539473684210527
精确度 0.8626225490196079

Process finished with exit code 0
```

图 8 优化前后指标对比（代码截图）

4. 结论和讨论

4.1 主要结论

本研究证明了将多个贝叶斯分类器与随机森林结合起来可以有效提高客户信用评分的分类预测性能。

算法组合能够充分利用不同分类器的优势，进一步优化了预测结果，提高了整体性能。

4.2 未来工作

后续工作可以进一步优化算法组合的参数，探索更多的特征工程方法，以进一步提高客户信用评分预测的准确性和稳健性。

本报告提供了关于客户信用评分数据集分类预测与算法组合的研究方法和

结果，为金融领域的信用风险管理提供了有益的参考和启示。

4.3 团队贡献

本项目中，组长蒋国威负责设计项目，收集 数据，编写代码，制作 ppt；单嘉乐负责编写代码，设计项目；李俊成负责改进项目，编写 代码，优化算法，制作 ppt；吕承骏负责提供研究方向与背景，总结结论，代码审查；陈泽豪负责编写代码，收集数据。

组长: 22 计算机 G1 蒋国威 5112221011512
组员: 22 计算机 G1 单嘉乐 5112221011503
22 计算机 G1 李俊成 5112221011504
22 计算机 G1 陈泽豪 5112221011510
22 计算机 G1 吕承骏 5112221011501

5. 参考文献

• UCI Machine Learning Repository. (n.d.). Credit Screening Data. Retrieved from https://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/crx.data.