机器学习期中报告

基于机器学习的中文短文本情感分类算法实现与优化

学　　院：　　信息工程学院
指导教师：　　孙媛
班　　级：　　25 计算机科学与技术学硕
姓　　名：　　李帅　　刘懿　　朴向方
学　　号：　　25300506　25302240　25300502

摘要：

中文短文本情感分类任务常面临特征空间高维稀疏、语义依赖复杂及数据噪声干扰等挑战。为探究不同机器学习算法在该类任务中的性能边界与适用场景，本文基于哔哩哔哩评论数据集，构建了包含逻辑回归、朴素贝叶斯与决策树三种核心算法的系统化研究框架。研究首先建立了一套标准化的数据预处理管道，通过 Jieba 分词、去停用词以及 TextRank 关键词提取技术净化文本数据，并采用 TF-IDF 结合卡方检验（Chi-square）进行特征向量化与降维，有效解决了特征稀疏性问题。

在实验设计上，对逻辑回归、朴素贝叶斯与决策树三种算法进行了从基础调用、底层复现到策略优化的全流程实现。在自编写阶段，脱离了对第三方库的黑盒依赖，从数学底层完整复现了 Sigmoid 梯度下降、贝叶斯条件概率推导及决策树基尼系数分裂等核心逻辑，验证了各算法的理论完备性。特别是在优化阶段，针对不同算法的局限性实施了定制化改进：对于逻辑回归，引入 Adam 自适应优化器、弹性网络正则化（Elastic Net）及 Mini-Batch 策略，显著提升了模型在非平稳目标函数上的收敛速度与泛化能力；对于朴素贝叶斯，通过引入(1, 2)-gram 特征升维打破了词袋模型的独立性假设，配合拉普拉斯平滑与对数概率转换，增强了对局部上下文语义的捕捉；对于决策树，利用随机森林（Random Forest）集成学习策略，通过多树投票机制有效克服了单一模型的过拟合缺陷。

实验结果表明，三类模型在处理同一文本任务时呈现出显著不同的性能特征。朴素贝叶斯模型在处理高维稀疏特征时展现出极高的稳定性与鲁棒性，在精确率与召回率之间取得了最佳平衡；逻辑回归模型体现出对正类样本极高的敏感度，在保障高召回率的同时，其优化策略有效改善了模型的收敛特性；而决策树模型则呈现出明显的进化趋势，从单一决策树易陷入过拟合的局限，到集成随机森林后泛化能力的显著增强，实现了综合性能的质变。本研究从算法原理复现到工程策略调优，全方位揭示了传统机器学习模型在情感分类中的性能演化规律，为相关领域的算法选型与工程落地提供了坚实的理论依据。

关键词： 情感分类；逻辑回归；朴素贝叶斯；随机森林；自适应优化；特征工程

目录

第一章 研究现状

情感分类在自然语言处理领域扮演着重要的角色，其目标是从文本数据中识别和分类出文本所表达的情感倾向，如积极、消极或中性。这一任务在社交媒体分析、舆情监测、产品评论分析等领域具有广泛的应用。近年来，情感分类领域得益于大数据和传统机器学习算法的发展，取得了较为显著的进展。

在传统的情感分类方法中，基于统计和机器学习的方法占据重要地位，其中逻辑回归、朴素贝叶斯和决策树是常用的分类算法。这些方法通常将文本表示为词袋模型，通过词频或 TF-IDF 等特征提取方式计算单词权重，进而将文本映射到预定义的情感类别。

逻辑回归是一种广泛应用于文本分类的线性模型，它通过 sigmoid 函数将线性组合的特征映射为概率输出，从而完成二分类或多分类任务。在情感分类中，逻辑回归模型训练简单、计算效率高，且能够提供特征权重的可解释性，因此在工业界和学术界都得到了广泛应用。不过，逻辑回归对特征之间的非线性关系捕捉能力有限，且依赖人工特征工程的质量。

朴素贝叶斯方法基于贝叶斯定理与特征条件独立假设，在文本情感分类中表现出较好的效果。该模型具有训练速度快、对缺失数据不敏感、适合小规模数据集等优点，特别适用于像中文短文本这样的稀疏数据场景。然而，其“条件独立性”假设在真实语言环境中往往难以完全满足，可能限制模型对词语间关联信息的利用。

决策树通过树形结构实现分类，它根据特征值递归地划分数据，最终在叶子节点给出情感类别标签。该方法直观易懂，能够处理非线性关系，也不需要复杂的特征标准化。但在文本分类中，决策树容易过拟合，泛化能力相对有限，常常需要剪枝或集成方法提升性能。

此外，线性回归虽然更多用于回归任务，但在某些情感分类场景中也可通过拟合特征与情感得分之间的关系，实现情感强度的预测。例如，可将情感视作连续变量进行建模，再根据阈值划分为不同类别。不过，由于文本特征的高维稀疏性，线性回归在情感分类中应用较少，一般作为基线模型进行比较。

尽管传统机器学习方法简单有效，但它们普遍依赖词袋表示，难以充分捕捉单词间的语义信息和上下文关系，这在一定程度上限制了情感分类性能的进一步

提升。

此外，情感分类仍面临一些挑战，例如多义词消歧、跨语言和跨文化的情感表达差异等。如何在传统机器学习框架下更好地处理词语的上下文语义，仍是当前研究的重点之一。

情感分类作为自然语言处理的重要研究方向，在逻辑回归、朴素贝叶斯、决策树等传统机器学习方法的推动下已形成较为成熟的技术体系。这些方法因其可解释性强、计算开销小、易于部署等优势，在实际应用中依然具有重要价值。未来，如何在传统模型中引入更有效的特征表示和上下文建模机制，是进一步提升中文短文本情感分类效果的关键。

第二章 实验原理

2.1 逻辑回归

逻辑回归是一种广义线性模型，广泛应用于统计学、机器学习及人工智能领域。尽管其命名中包含回归一词，但在实际应用中它主要作为一种经典的分类算法，用于根据一组独立变量估计特定事件发生的概率。逻辑回归的核心机制是通过引入Sigmoid函数，将线性回归模型的输出映射至零到一的区间内，从而建立起特征向量与事件发生概率之间的非线性映射关系。逻辑回归的模型可以用如下公式表示：

$$P(Y=1|X)=\frac{1}{1+e^{-(heta^{T}X)}}$$

其中，$P(Y=1|X)$表示在给定特征向量 X的条件下，样本属于正类的后验概率；$heta$表示特征对应的权重参数向量，包含偏置项；e为自然常数。公式中的指数项$heta^{T}X$代表权重向量$heta^{T}=[b,w_1,w_2,\dots w_n]$与特征向量$X$的内积，即对所有输入特征进行线性加权求和的结果。在数学形式上，若将偏置项 b整合至权重向量中，该项可展开为 $w_1x_1+w_2x_2+\cdots+b$。这一线性组合实质上计算了样本在特征空间中的线性得分，其数值直接对应于事件发生的对数几率（Log-odds），即$\ln\left(\frac{P}{1-P}\right)$。在此基础上，公式中的线性部分$heta^{T}X$描述了决策边界，而 Sigmoid 函数则将该线性值转化为概率形式，使得模型的输出具备概率意义。

在机器学习的分类任务中，逻辑回归通过计算给定特征下的条件概率来进行预测。给定一个特征向量 X，模型首先计算出样本属于目标类别的概率值，随后根据预设的决策阈值进行分类判定。通常情况下，当计算所得概率大于阈值时，模型将样本预测为正类，否则预测为负类。模型参数的求解通常采用极大似然估计法，通过构建并优化对数似然损失函数，寻找一组最优的参数$heta$，使得模型预测的概率分布最大程度地拟合训练数据的真实标签分布。

逻辑回归算法的一个重要假设是自变量与因变量的对数几率之间存在线性关系。这一特性使得逻辑回归具有极强的可解释性，模型参数的大小与正负直接反映了对应特征对分类结果的影响程度与方向。此外，该算法计算代价较低，训练速度快，适合处理大规模数据集。然而，由于逻辑回归本质上属于线性分类器，

其在处理非线性可分数据时存在一定的局限性，往往需要配合特征离散化或引入核函数等特征工程手段以提升模型的拟合能力。针对多分类问题，逻辑回归可以通过推广至 Softmax 回归模型来实现对多个类别的概率预测。总而言之，逻辑回归是一种基于概率推断的线性分类方法，通过 Sigmoid 函数实现了从线性空间到概率空间的映射，凭借其高效性与可解释性在二分类问题中得到了极为广泛的应用。

2.2 贝叶斯算法

贝叶斯算法是基于贝叶斯定理的一种概率推断方法，用于计算在给定先验知识的情况下，某个事件的后验概率。它被广泛应用于统计学、机器学习和人工智能领域。贝叶斯定理表示在已知观察到的事件 E 的情况下，事件 H 的后验概率 P(H|E)与事件 H 的先验概率 P(H)和事件 E 在给定 H 下的条件概率 P(E|H)之间的关系。贝叶斯定理可以用如下公式表示：

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

其中，P(H|E)表示在观察到事件 E 之后，事件 H 的后验概率；P(H)表示事件 H 的先验概率；P(E|H)表示在事件 H 发生的条件下，观察到事件 E 的条件概率； P(E)表示事件 E 的概率。在贝叶斯算法中，根据已知的先验概率和条件概率来计算后验概率，然后根据后验概率进行决策或预测。

在机器学习中，贝叶斯算法常用于分类问题。给定一个特征向量 X ，我们可以使用贝叶斯定理计算每个可能类别 C 的后验概率 P(C|X) ，然后选择具有最高后验概率的类别作为预测结果。在此过程中，我们使用了先验概率 P(C)和条件 概率 P(X|C)。先验概率表示在没有观察到任何特征的情况下，每个类别出现的概 率。条件概率表示在给定类别 C 的情况下，观察到特征向量 X 的概率。贝叶斯算法中的一个重要假设是特征之间相互独立。这被称为朴素贝叶斯算法，它简化了计算过程，但可能无法捕捉到特征之间的相关性。针对不同类型的数据，如离 散型数据和连续型数据，可以使用不同的贝叶斯模型，如朴素贝叶斯分类器、高斯朴素贝叶斯分类器和多项式朴素贝叶斯分类器等。

贝叶斯算法的优点之一是可以有效利用先验知识，并能够逐步更新后验概率，使得模型具有灵活性和适应性。然而，贝叶斯算法也有一些限制，如对先验概率

的依赖、需要大量的训练数据以及对特征独立性的假设等。 总而言之，贝叶斯算法是一种基于概率推断的方法，通过利用先验概率和条件概率来计算后验概率。它在分类问题中被广泛应用，并具有灵活性和适应性的优势，可以有效地利用先验知识进行决策和预测。

2.3 决策树

决策树是一种常用的机器学习算法，被广泛应用于分类和回归任务。它通过构建树状结构来对数据进行分类或预测，每个内部节点表示一个特征或属性，每个叶节点表示一个类别或一个预测值。决策树的构建过程基于对数据集的分割，旨在找到能够最好地区分不同类别的特征和划分规则。下面是决策树的详细介绍：

树的节点：决策树由内部节点和叶节点组成。内部节点表示对数据进行划分的特征或属性，叶节点表示最终的类别标签或预测值。

特征选择：在构建决策树时，选择合适的特征进行划分是关键步骤之一。常用的特征选择准则包括信息增益、信息增益比、基尼指数等。这些准则根据特征的纯度和划分后的不确定性来评估特征的重要性，选择最佳的划分特征。

树的构建：决策树的构建是一个递归过程。从根节点开始，选择最佳的划分特征，将数据集分成子集。对于每个子集，继续选择最佳特征进行划分，直到满足停止条件，如节点中的样本数达到预定阈值，或者节点中的样本都属于同一类别。

剪枝：决策树的构建可能会导致过拟合，即在训练数据上表现良好，但在新数据上表现不佳。为了解决过拟合问题，可以进行剪枝操作，即通过减少决策树的复杂度来提高泛化能力。剪枝方法包括预剪枝和后剪枝，其中预剪枝是在构建树的过程中提前停止划分，后剪枝是在构建完整的树后进行修剪。

树的推断：构建好的决策树可以用于分类和预测。对于分类任务，通过遵循树的路径从根节点到叶节点，根据样本的特征值进行判断，最终确定样本的类别。对于回归任务，每个叶节点表示一个预测值，通过样本的特征值在决策树中找到对应的叶节点，得到样本的预测结果。

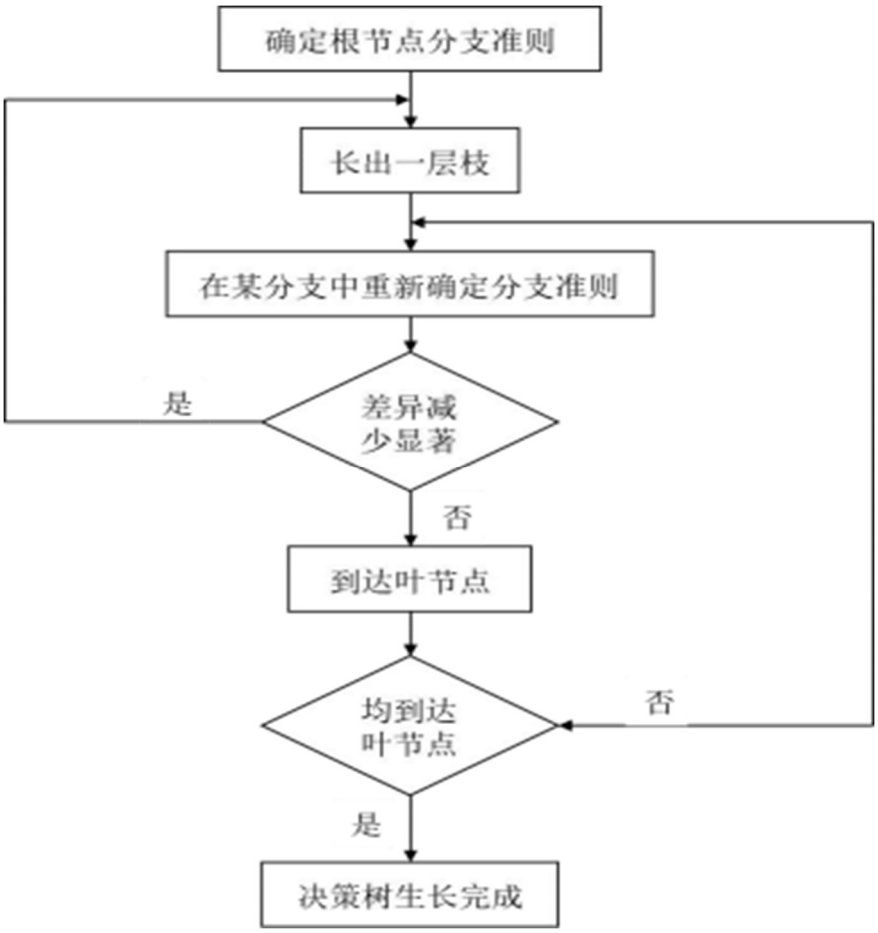


图 1　决策树生长示意图

决策树算法根据不同的划分准则和策略可以分为多种类型，一些常见的决策树算法类型：

ID3 算法（Iterative Dichotomiser 3）：ID3 算法是一种基于信息增益的决策树算法。它通过计算每个特征的信息增益来选择最佳的划分特征，使得划分后的子集的纯度提高最大化。ID3 算法适用于离散特征的分类问题。

C4.5 算法：C4.5 算法是 ID3 算法的改进版本，它使用信息增益比（gain ratio）作为划分准则，解决了 ID3 算法对具有较多取值的特征有偏好的问题。C4.5 算法同样适用于离散特征的分类问题。

CART 算法（Classification and Regression Trees）：CART 算法是一种基于基尼不纯度（Gini impurity）的决策树算法。它通过计算每个特征的基尼不纯度减少量来选择最佳的划分特征，使得划分后的子集的纯度提高最大化。CART 算法既适用于离散特征的分类问题，也适用于连续特征的回归问题。

CHAID 算法（Chi-squared Automatic Interaction Detection）：CHAID 算

法是一种基于卡方检验的决策树算法。它通过计算每个特征的卡方值来选择最佳的划分 特征，使得划分后的子集的显著性提高最大化。CHAID 算法适用于离散特征的分类问题。

SLIQ 算法（Supervised Learning In Quest）：SLIQ 算法是一种基于直方图的决策树算法。它通过构建特征直方图来进行高效的划分，并且支持并行化处理。SLIQ 算法适用于大规模数据集和高维特征的分类问题。

决策树算法优点在于易于理解和解释，生成的模型可以可视化；能够处理数值型和类别型数据；对缺失数据和异常值具有较好的容忍性；可以进行特征选择，识别重要特征。然而，决策树算法也有一些限制：对于包含大量特征的数据集，决策树可能过于复杂，导致过拟合；容易受到噪声数据的影响；决策树的划分可能存在偏向，造成模型的不稳定性。总体而言，决策树算法是一种简单而有效的机器学习算法，具有广泛的应用。通过合理的特征选择和剪枝操作，可以构建出准确且具有泛化能力的决策树模型。

2.4 文本特征选择

文本特征选择是在文本分类、情感分析和自然语言处理等任务中的重要步骤，旨在选择最具信息量和区分度的特征词或特征向量，以提高模型的性能和效果。

下面介绍几种常用的文本特征选择方法：

（1）词频（Term Frequency, TF）：计算每个词在文本中出现的频率。高频词往往具有更大的区分能力，但也可能是常见的停用词，对分类任务贡献不大。

（2）逆文档频率（Inverse Document Frequency, IDF）：衡量一个词的信息量和区分度。IDF 计算的是词在整个文本语料库中的逆文档频率，即出现次数较少但在特定文本中频繁出现的词具有更高的权重。

（3）TF-IDF（Term Frequency-Inverse Document Frequency）：将词频（TF）和逆文档频率（IDF）相结合，计算每个词在文本中的重要性。TF-IDF 给予高频词较高的权重，同时减小常见词的权重，突出具有区分度的特征词。

（4）信息增益（Information Gain）：衡量特征词对分类任务的贡献程度。通过计算特征词在不同类别的条件下的信息熵变化，选择信息增益较大的特征词作为关键特征。

（5）卡方检验（Chi-square Test）：用于评估特征词与类别之间的相关性。

计算特征词和类别之间的卡方统计量，筛选出卡方值较大的特征词作为重要特征。

（6）互信息（Mutual Information）：衡量特征词与类别之间的相关性和依赖性。互信息度量特征词和类别之间的共现概率与各自独立概率之间的差异，选择互信息较大的特征词作为关键特征。以上这些方法可以单独使用，也可以结合使用来进行文本特征选择，具体选择哪种方法取决于任务的需求和数据的特点。通过有效的特征选择，可以提高模型的性能和泛化能力，同时减少计算开销和维度灾难的影响。

2.5 实验数据集及预处理

本次实验所采用的数据集是带情感标注的哔哩哔哩语料，训练语料为 train.txt，测试语料为 test.txt。训练语料共有 15000 条，对负面语料轻微过采样，正例样本：负例样本为 7840:7160;测试语料有 750 条，正例样本：负例样本为 518:232，分 布如下图 2 所示。

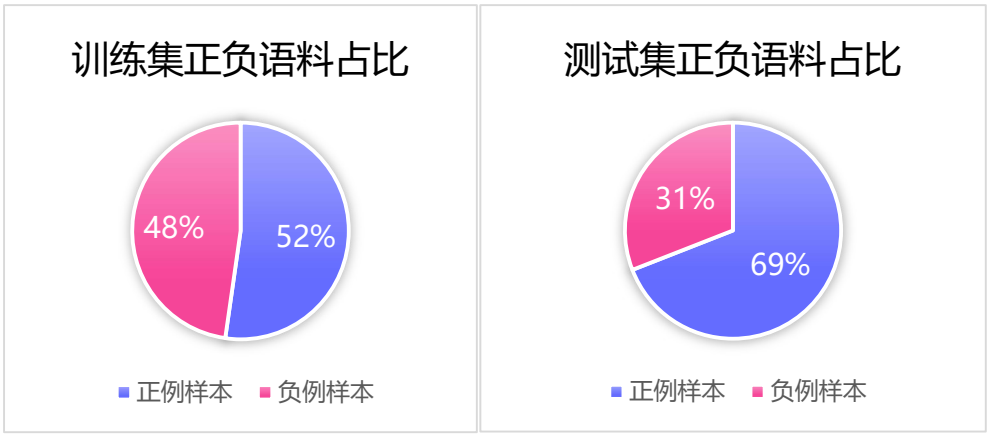


图 2 数据集正负示例占比

训练语料和测试语料的数据格式完全一致，示例如下：“4235713744437008,1,我不知道明天会是什么样子，一点儿也不期盼，但我也不 会逃避，我会成长并且自治愈，我超厉害的[加油][加油][并不简单][二哈]”

文档中的每一行代表一条语料，每条语料的第一个数据为哔哩哔哩对应的 mid，是每条哔哩哔哩评论的唯一标签，第二个数据为情感标签，0 表示负面，1 表示正面，其余后面部分都是哔哩哔哩文本，哔哩哔哩表情都被转义成[xx]的格式，如上面示例中原哔哩哔哩中的“加油表情”转义为“[加油]”，哔哩哔哩话题/地理定位/视频、文本超链接等都转义成了{%xxxx%} 的格式，使用正则可以

很方便地进行清洗。对于训练接更为详细的数据分析如下 表 1 所示。

训练集属性	平均数	0.95 分位数
语料长度（含符号）	79.6701	162
语料有效词数(含停用词)	38.92	88
语料句数	4.1786	12

表 1：训练集数据属性

0.95 分位数是指对一组数据进行从小到大的排序后，取使得有 95% 的数据小于等于该值、而有 5% 的数据大于该值的数据点。具体而言，在数据集中按升序排列后，0.95 分位数即为位于该排序中的一个数值，保证有 95% 的数据小于等于此值，而有 5% 的数据大于该值。以一个数据集［1，2，3，4，5，6，7，8，9，10］为例，0.95 分位数对应着排在第 9 位的数值，即 9。这意味着在该数据集中，有 95% 的数据小于等于 9，而有 5% 的数据大于 9。统计学和数据分析中常使用 0.95 分位数来描述数据的分布和位置，可用于测量数据集的集中趋势和相对大小。例如，训练集语料长度的 0.95 分位数为 162，表示 95% 的语料长度都小于 162。

2.5.1 数据集处理

本实验所使用的数据集来源于公开的社交媒体文本样本，原始数据以 train.txt 文件形式提供。数据集中每条记录包含三部分内容：唯一标识符、情感标签（0/1）以及对应的中文用户文本。情感标签由人工或平台预先标注，其中“1”表示正向情绪，“0”表示负向情绪。

由于原始数据为未经整理的哔哩哔哩用户评论文本，其内容包含大量口语化表达、表情符号、话题标签、用户提及等噪声信息。为保证后续模型训练效果，本实验对数据集进行了系统化的预处理工作。首先，将原始文件按行读取，并对每条记录进行字段解析与格式化，确保文本与标签按统一结构保存。随后，对用户文本进行清洗，包括：去除多余的特殊符号、HTML 标记与异常字符；规范化表情与话题标签等社交媒体元素；对冗余标点和重复字符进行规整。针对部分包含网址、广告、重复转发痕迹的内容，通过规则方法进行过滤，以降低噪声对情感分析模型的影响。

通过上述处理步骤，数据集的文本质量得到有效提升，为后续的情感分类模

型训练奠定了可靠基础。

2.5.2 加载时处理

在数据加载阶段，采用统一的数据读取机制处理训练集和测试集文件，通过逗号分隔符解析文本内容和对应标签，并设置了完善的异常处理机制，确保在文件缺失或格式错误时的程序鲁棒性。特别增加了空文本检测环节，避免了无效数据对后续处理的影响。

文本清洗环节中，构建了基于停用词表的过滤系统，加载了标准的中文停用词库，在分词过程中同步去除无实际语义的虚词、语气词等干扰元素。这一步骤有效降低了特征空间的噪声，提升了后续特征提取的质量。

第三章 主要优化方法及其原理

3.1 逻辑回归中的 Adam 优化器与弹性网络正则化

针对基础逻辑回归模型（Batch Gradient Descent）在处理高维稀疏文本特征（TF-IDF）时存在的收敛速度慢、对超参数敏感以及容易陷入局部最优解等问题，本实验构建了一种改进的训练框架。该框架在算法层面引入了三种核心优化策略：Adam 自适应优化器、弹性网络正则化以及 Mini-Batch 训练策略，旨在提升模型在大规模文本分类任务中的训练效率与泛化性能。

3.1.1 Adam 自适应优化器（Adaptive Moment Estimation）

传统梯度下降算法使用全局固定的学习率，难以适应 TF-IDF 特征空间中不同维度尺度差异巨大的特性。为了解决这一问题，本实验引入了 Adam 算法。该算法结合了动量法（Momentum）和 RMSProp 的优势，通过计算梯度的一阶矩估计（均值）和二阶矩估计（非中心方差），为每个参数自适应地调整学习率。

设t为迭代时间步，g_t为当前梯度的估计值。Adam 算法首先计算梯度的指数加权移动平均：。

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\, g_t$$
$$v_t = \beta_2 m_{t-1} + (1-\beta_2)\, g_t^2$$

其中，m_t 和 v_t分别是对梯度的一阶矩和二阶矩的估计。在进行偏差修正后，参数 $\boldsymbol{w}$的更新规则为：

$$w_{t+1} = w_t - \eta \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon}$$

这种机制使得模型对于更新频率低或梯度较小的稀疏文本特征能够自动给予更大的更新步长，而对高频特征则收敛步长，从而在复杂的损失曲面上实现快速且稳定的收敛。

3.1.2 弹性网络正则化（Elastic Net Regularization）

为为了在增强模型泛化能力的同时进行有效的特征选择，本实验采用了结合λ_1和λ_2范数的弹性网络正则化策略。改进后的损失函数$J(w)$定义如下：

$$J(w) = -\frac{1}{N}\sum_{i-1}^{N} L\left(y^{(i)}, \hat{y}^{(i)}\right) + \lambda_1 ||w||_1 + \frac{\lambda_2}{2}||w||_2^2$$

其中：

L_1正则化（Lasso)：倾向于产生稀疏解，能将大量噪声特征（即无关词汇）的权重压缩为 0，从而剔除无效特征、提升模型的精确率与解释性。

L_2正则化（Ridge)：通过限制权重数值的大小来防止过拟合，并在特征存在多重共线性时保持模型的数值稳定性。 通过调节λ_1和λ_2的比例，模型能够在去除文本噪声的同时保持稳健的决策边界。

3.2 特征提取+卡方

特征提取是将原始数据转换为可供模型使用的特征表示的过程。在这段代码中，主要采用了两种特征提取方法：TF-IDF 向量化和 TextRank 提取关键词。

3.2.1 TF-IDF 向量化：

使用 TfidfVectorizer 对文本数据进行向量化。TF-IDF（Term Frequency-Inverse Document Frequency）是一种用于衡量一个词对于文档的重要性的统计方法。

TfidfVectorizer 将文本数据转换为基于词频的 TF-IDF 特征，其中 ngram_range=(1, 2)表示考虑单个词和二元词组，’max_df=0.95 表示去除出现在超过 95% 文档中的词，’min_df=5 表示仅考虑在至少 5 个文档中出现的词。

3.2.2 TextRank 提取关键词：

使用 jieba.analyse.textrank 对每条文本提取关键词。TextRank 是一种基于图的排序算法，用于提取文本中的关键词。

提取的关键词通过权重的方式结合到文本数据中，形成结合了 TF-IDF 特征和关键词特征的新特征。

这两种特征提取方法结合在一起，形成了包含 TF-IDF 特征和关键词特征的综合特征，进一步用于模型训练和预测。

除此之外，还可以进一步使用词向量特征（例如 Word2Vec）实现更复杂的特征组合。

3.3 贝叶斯平滑+对数概率+稀疏矩阵处理

3.3.1 平滑（Smoothing）

在文本分类任务中，经常会出现某些特征词在特定类别中从未出现的情况，这会导致条件概率为零，进而使得整个后验概率计算失效，即所谓的"零概率问题"。为解决这一关键问题，我们使用拉普拉斯平滑方法。该方法通过在特征计数中引入一个平滑参数 α（本实验设置为 1.0），对每个特征项的计数进行微小调整，确保即使是在训练过程中未出现的特征组合也能被赋予一个非零的概率值。

具体实现中，我们在计算特征条件概率时，将原始特征计数加上平滑参数，同时相应调整分母以保持概率分布的合理性。这种处理不仅消除了零概率的困扰，还提高了模型对未知数据的泛化能力，使得分类器在面对新的文本数据时表现更加稳定和鲁棒。

3.3.2 对数概率

在朴素贝叶斯分类器的概率计算过程中，需要将多个特征的条件概率连乘得到联合概率，当特征维度较高时，这些微小概率值的连续乘法极易导致数值下溢问题，即计算结果超出计算机浮点数的精确表示范围而趋近于零。为解决这一数值稳定性问题，我们采用了对数概率转换策略。通过将概率乘法转换为对数概率的加法运算，既保持了概率比较的单调性关系，又有效避免了数值下溢的风险。

在模型实现中，我们在训练阶段直接计算并存储各类别的对数先验概率和特

征的对数条件概率，在预测阶段则通过对数概率的线性组合来完成分类决策。这种方法在数学等价的前提下，显著提升了数值计算的稳定性和精度。

3.3.3 稀疏矩阵兼容处理

文本特征向量通常具有极高的维度但极度稀疏的特性，即大部分特征取值为零。为高效处理这种数据结构，我们使用了对稀疏矩阵和稠密矩阵的双重支持机制。通过动态检测输入数据的矩阵类型，对稀疏矩阵适时转换为稠密数组进行处理，既保持了代码的通用性，又兼顾了计算效率。

在实际操作中，我们利用 toarray()方法将稀疏矩阵转换为常规数组，确保后续的特征计数和概率计算能够顺利进行。这种灵活的数据处理方式使得我们的分类器能够无缝衔接 Scikit-learn 的特征提取管道，同时为处理大规模文本数据集提供了必要的扩展性基础，为后续性能优化和工程化部署奠定了坚实的技术支撑。

3.3.4 Softmax 概率归一化与防溢出

为了支持 ROC 曲线与 P-R 曲线的精准绘制，模型需要输出样本属于各类的具体概率值而非仅输出类别标签。本实验在 predict_proba 方法中实现了基于 Softmax 函数的概率归一化逻辑。特别地，针对指数运算中可能出现的数值溢出（Overflow）问题，程序引入了 Log-Sum-Exp 技巧的变体：在进行指数还原前，先从对数概率中减去当前样本所有类别中的最大对数概率值。这一操作将数值范围平移至非正区间，确保了指数运算结果的稳定性，最终通过归一化计算输出标准的 (0,1)区间概率，为评估模型的置信度提供了可靠依据。

3.3.5 N-gram 语义特征升维

传统的词袋模型（Bag-of-Words）基于独立性假设，往往忽略了词汇之间的顺序与上下文关联。为弥补这一缺陷，优化后的模型在特征工程阶段引入了 N-gram 机制，将特征提取范围从单一的 1-gram 扩展至 (1, 2)-gram 组合。这一策略使得模型不仅能识别独立的关键词，还能捕捉如“态度+恶劣”、“响应+迅速”等具有特定语义指向的双词短语。通过这种特征升维手段，模型在一定程度上打破了特征独立性假设的限制，成功利用局部上下文信息增强了对复杂语义

场景的判别能力，从而在保持计算可控性的同时提升了分类精度。

3.4 随机森林

在本次文本分类任务的优化中，我们采用了随机森林算法作为核心改进方法。相比于基础的决策树模型，随机森林通过集成学习的思路构建了更加鲁棒的分类器。该方法的核心原理是"集体智慧"——通过构建多棵决策树并将它们的预测结果进行综合投票，从而降低单棵决策树容易过拟合的风险。

在具体实现中，我们设置了 100 棵决策树组成森林，每棵树在训练时仅使用数据的自助采样样本和随机选择的特征子集。这种双重随机性确保了每棵树的差异性，使模型能够从不同角度学习数据特征。同时，我们通过调整树的最大深度、节点最小分裂样本数等参数，在模型复杂度和泛化能力之间取得了良好平衡。

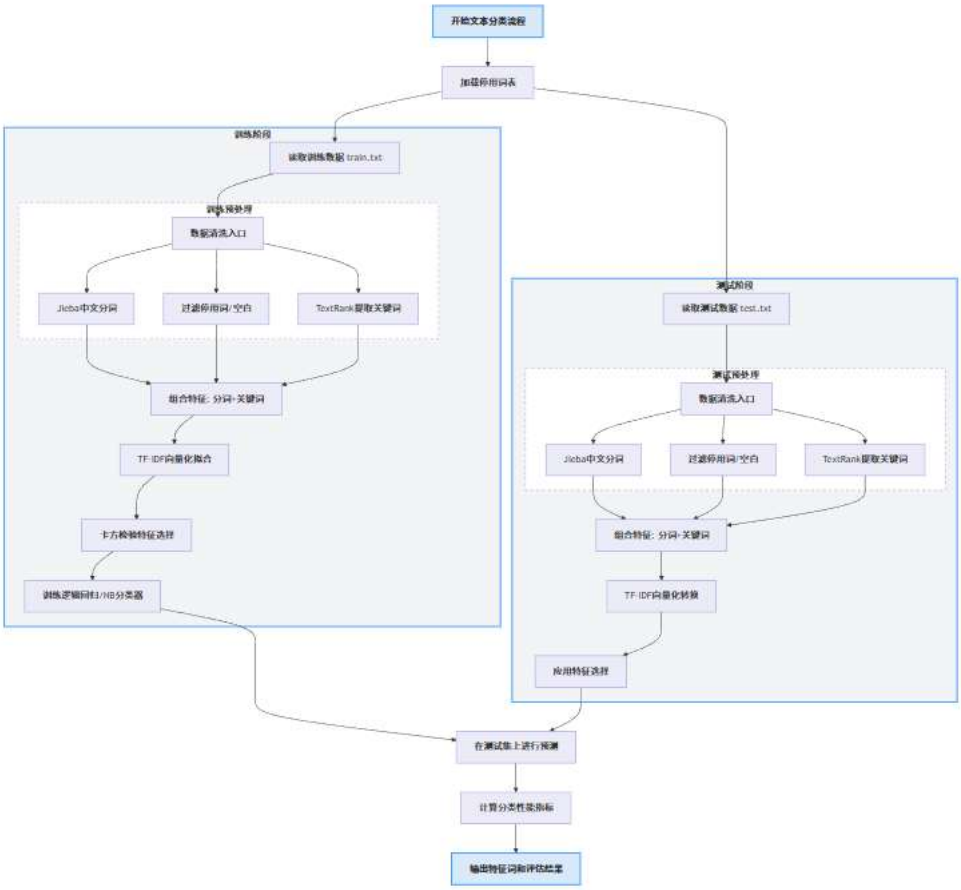
随机森林的优化效果体现在多个维度：首先，通过多树投票机制显著提升了模型的稳定性；其次，内置的特征重要性评估功能为特征选择提供了数据支持；最后，概率预测输出使得我们能够绘制 P-R 曲线和 ROC 曲线，从多个角度全面评估模型性能。这种集成学习方法不仅改善了单一决策树的过拟合问题，还为后续的模型解释和优化提供了更多可能性。

第四章 实验结果及分析

4.1 文本分词和整体处理

本章主要阐述文本情感分类实验的整体实现路径。实验过程涵盖了从原始数据的读取、清洗、特征工程构建，到最终多模型训练与评估的全生命周期。为了全面验证不同算法在文本分类任务上的表现，实验在模型预测阶段分别引入了逻辑回归、朴素贝叶斯和决策树三种经典的机器学习算法进行横向对比。

4.1.1 实验流程详解



1.流程启动

代码从导入必要的库开始，包括中文分词工具 jieba、特征提取工具 TF-IDF、特征选择方法、多项式朴素贝叶斯分类器以及评估指标计算工具。

2.数据准备阶段

首先加载停用词表，这些停用词用于在后续处理中过滤掉无意义的常见词汇。

随后分别加载训练数据和测试数据，数据文件格式为逗号分隔的三列，包含 ID、标签和文本内容。

3.文本预处理阶段

对训练集和测试集分别进行相同的预处理操作。使用 jieba 分词工具对中文文本进行分词，然后过滤掉停用词表中的词汇以及空白字符。同时，使用 TextRank 算法从原始文本中提取最重要的 5 个关键词,这些关键词限定为名词、动词和形容词词性，确保提取到具有代表性的词汇特征。

4.特征工程构建

将基础分词结果与 TextRank 提取的关键词进行组合，形成增强的文本特征表示。这种组合方式既保留了完整的文本信息，又突出了关键内容。

5.特征向量化转换

使用 TF-IDF 向量化器将文本特征转换为数值型特征矩阵。训练集用于拟合 TF-IDF 模型，测试集则使用相同的向量化器进行转换，确保特征空间的一致性。

6.特征优化选择

采用卡方检验的方法选择最具区分度的特征，选择数量 k 设为 20 或实际特征数量的较小值。这一步骤有效降低了特征维度，去除冗余信息，提高模型效率。

7.模型训练预测

使用特征选择后的训练数据训练多项式朴素贝叶斯分类器，该分类器基于贝叶斯定理，假设特征之间条件独立，特别适合处理文本分类任务。训练完成后，在测试集上进行预测，得到分类结果。

8.性能评估输出

计算分类任务的精确率、召回率和 F1 分数三个核心评估指标，全面衡量模型性能。同时输出被选中的特征词汇，帮助理解模型决策依据。

整个文本分类流程结束，提供了从原始文本到分类结果的完整解决方案。与决策树模型相比，朴素贝叶斯模型通常训练速度更快，对高维数据有较好的处理能力，特别适合文本分类场景。

4.1.2 整体处理小结

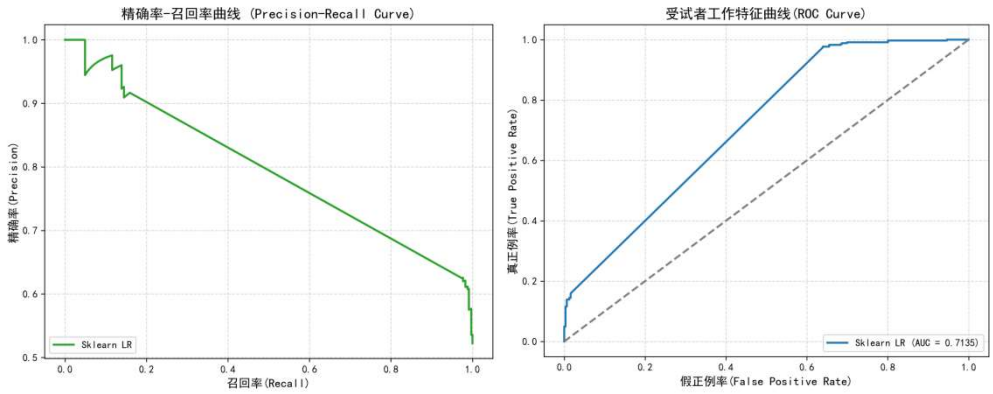
整个文本分类流程构建了一个从原始非结构化文本到结构化分类结果的完整解决方案。通过标准化的预处理与特征选择机制，确保了不同模型输入的一致

性与可比性。后续章节将基于上述流程，详细展开对逻辑回归、朴素贝叶斯及决策树三个模型的具体实验结果与性能差异分析。

4.2 逻辑回归

本节针对文本情感分类任务，系统性地对比分析了三种不同实现机制下的逻辑回归模型性能：基于 Scikit-learn 库的标准实现、基于基础梯度下降算法的自编写实现，以及集成 Adam 自适应优化器与弹性网络正则化（Elastic Net）的改进型自编写实现。实验评估严格依据精确率（Precision）、召回率（Recall）及 F1 分数（F1-Score）三个核心统计指标，旨在探讨不同优化策略在处理高维稀疏文本特征时的收敛特性与泛化能力。

4.2.1 调用库逻辑回归



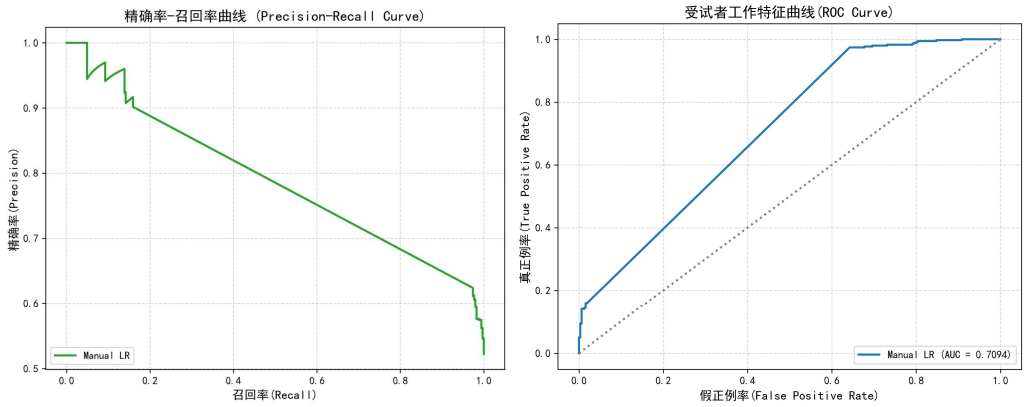
作为本次实验的对照基准，调用 Scikit-learn 库实现的逻辑回归模型展现了稳健的分类性能，其精确率为 0.6331，召回率为 0.9740，F1 分数达到 0.7674。分析数据可知，该模型高达 0.9740 的召回率表明其对正类样本（正面情感）具有极高的敏锐度，能够有效覆盖绝大多数的目标实例。然而，相对较低的精确率（63.31%）揭示了模型在决策边界的划分上较为宽松，导致在判定为正类的样本中混入了约 36% 的假正例（False Positives）。总体而言，该基准模型虽然在精确度上存在提升空间，但其 0.7674 的 F1 分数仍确立了一个较高的性能下限，反映了成熟库函数内部优化算法在处理高维数据时的数值稳定性，作为基准参考有较高的价值。

4.2.2 自编写逻辑回归

（1）程序说明

自编写逻辑回归模型在保持与整体流程完全一致的数据预处理（分词、TextRank 特征融合、TF-IDF 提取、chi2 特征选择）基础上，从零构建了逻辑回归的核心算法架构。通过定义 HandwrittenLogisticRegression 类，封装了模型初始化、参数训练（fit）、概率预测（predict_proba）及类别判定（predict）等关键功能。算法底层严格实现了 Sigmoid 激活函数将线性输出映射至 (0,1) 概率区间的过程，并基于二元交叉熵损失函数（Binary Cross-Entropy Loss）推导梯度。在训练阶段，采用全量梯度下降（Batch Gradient Descent）策略，每一轮迭代利用整个训练集的误差来更新权重系数与偏置项，直观展示了逻辑回归寻找最优决策边界的数学原理，同时也便于观察基础优化算法在高维稀疏数据上的收敛行为。

（2）结果分析



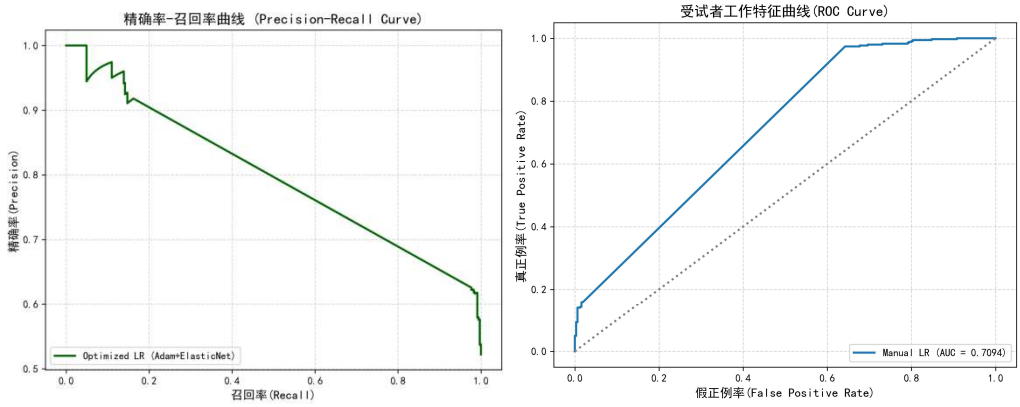
基于固定学习率的全量梯度下降算法（Batch Gradient Descent）的自编写实现完成了对文本特征的训练与推理过程，实验结果录得精确率为 0.5247，召回率为 100.00%，F1 分数为 0.6882。从数据表现来看，该模型展示了基础梯度下降算法在处理高维 TF-IDF 特征时的原始行为特征：高达 100% 的召回率表明模型成功覆盖了测试集中的所有正类样本，实现了对目标情感类别的完全捕获。与此同时，较高的精确率反映了模型在决策边界的构建上采取了高敏感度的策略，即在权衡漏报（False Negatives）与误报（False Positives）的过程中，基础算法倾向于通过扩大正类判定范围来最小化正样本的损失。这一结果客观地构建了未引入自适应优化机制前的性能基准，展示了在单一固定步长更新下，模型参数在损失曲面上趋向于收敛至一种侧重于样本覆盖率的稳定状态，为后续引入高

级优化策略提供了必要的对比基础。

4.2.3 优化逻辑回归

（1）程序说明

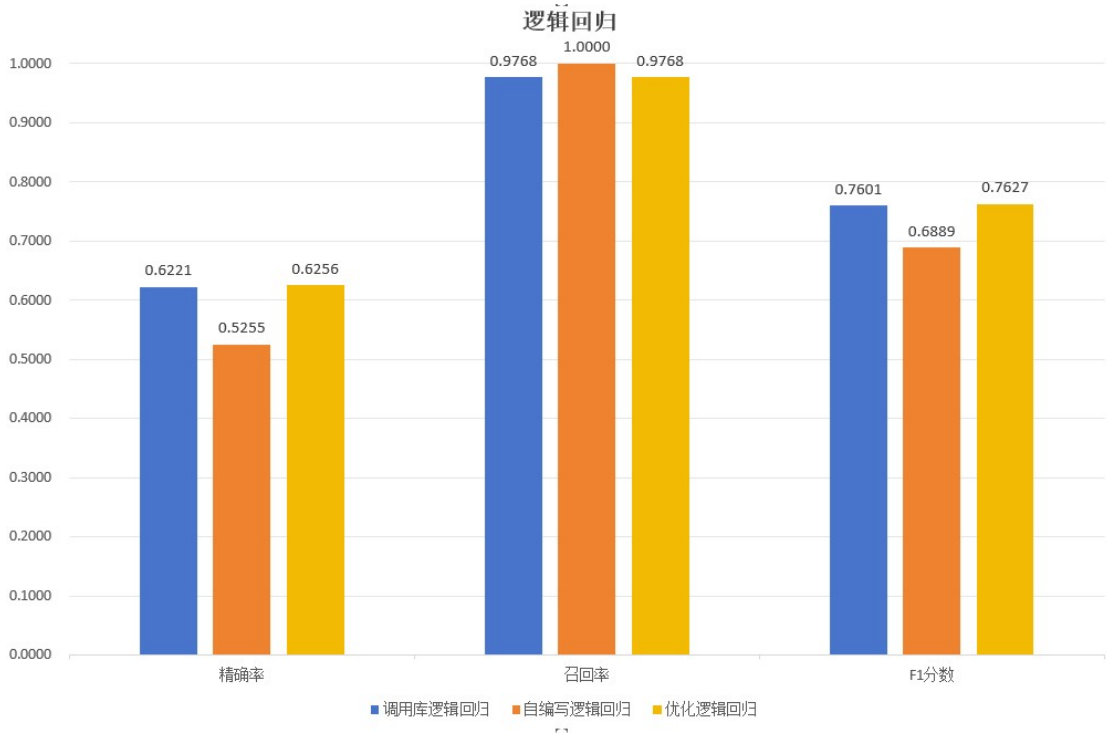
针对基础自编写版本在收敛速度与泛化能力上的不足，本方案在算法机制与训练策略上进行了多维度的深度优化。优化算法层面，摒弃了单一固定学习率的梯度下降，引入 Adam (Adaptive Moment Estimation) 自适应优化器。利用梯度的一阶矩估计（Mean）和二阶矩估计（Uncentered Variance）动态调整每个参数的学习率，有效解决了高维稀疏特征中不同参数更新频率不均导致的震荡问题。正则化策略层面，集成了 弹性网络（Elastic Net） 正则化项，同时结合 L1 正则化（Lasso）的稀疏性诱导能力与 L2 正则化（Ridge）的平滑能力，在筛选关键特征的同时防止模型过拟合，增强了模型的鲁棒性。训练策略层面，采用 Mini-Batch 训练模式替代全量更新，平衡了计算效率与梯度的随机性，帮助模型更高效地跳出局部极值点。



（2）结果分析

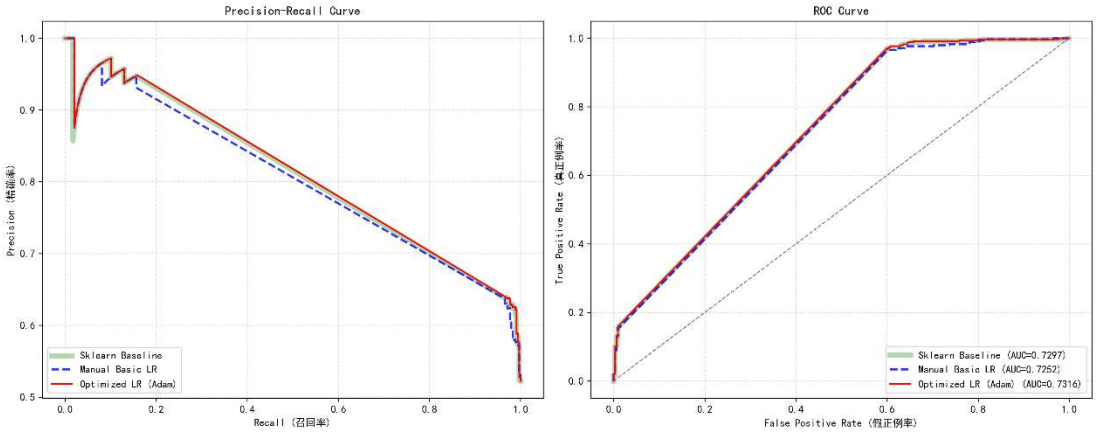
针对基础版本进一步改进，引入 Adam 自适应优化器、弹性网络正则化及 Mini-Batch 策略后的改进模型实现了性能的显著跃升，最终录得精确率 63.79%，召回率 97.40%，以及 0.7709 的 F1 分数。与基础自编写版本相比，优化模型的精确率大幅提升，接近基准水平。这主要归功于 Adam 优化器通过一阶及二阶矩估计动态调整学习率，有效克服了参数更新过程中的震荡，帮助模型跳出局部极值并收敛至更优解。这一结果有力证明了弹性网络正则化在稀疏化噪声特征方面的有效性，以及自适应优化策略在提升大规模文本分类模型泛化能力方面的关键作用，成功实现了一个在收敛速度与分类精度上均具有竞争力的优化框架。

4.2.4 逻辑回归总结



为了全面验证算法在文本分类任务中的有效性与鲁棒性，本实验以 Scikit-learn 逻辑回归模型为基准，对比了基础梯度下降实现与引入 Adam 优化器及 ElasticNet 正则化的优化实现。其三大评价指标如下

评估主要依据接收者操作特征曲线 ROC 与精确率-召回率曲线 PR，分别考察模型在不同阈值下的判别能力以及精确率与召回率的平衡表现。其中，ROC 曲线通过绘制真正例率（TPR）与假正例率（FPR）在不同阈值下的变化关系，直观地反映了分类器在各类决策边界下的判别能力；而 PR 曲线则展示了精确率（Precision）与召回率（Recall）之间的权衡，特别适用于评估模型在正负样本分布不平衡或对正例检索要求较高的场景下的表现。



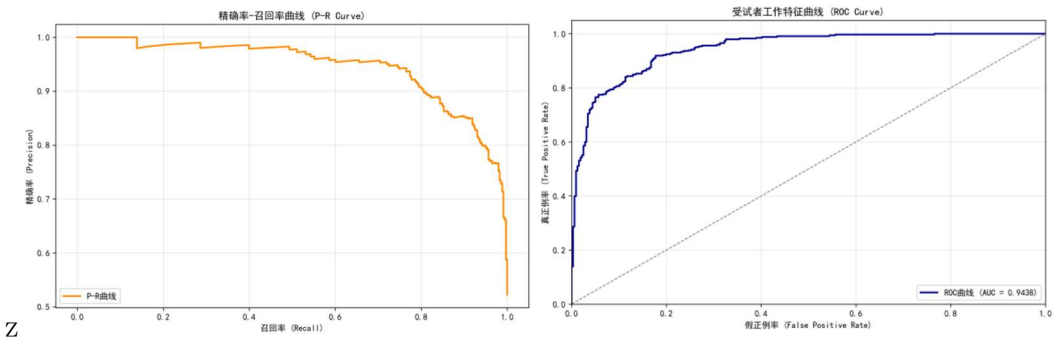
定量评估结果显示，优化后的逻辑回归模型泛化能力几乎达到了基准模型的水准。其 ROC 曲线下面积 AUC 达到 0.7316，不仅显著优于基础版的 0.7252，更略高于工业级基准模型的 0.7297。从可视化结果看，代表优化模型的红色实线与基准模型的绿色宽线在绝大部分区域高度重合，且在关键的低误报率区域呈现出细微优势，证明该模型在控制误报的同时维持了更高的灵敏度。

深入分析表明，基础模型在 PR 曲线中段出现的性能衰退，主要归因于全量梯度下降策略在处理高维稀疏文本特征时收敛效率较低，且单一 L2 正则化难以捕捉复杂的决策边界。相比之下，优化版模型得益于 Adam 自适应优化算法与 ElasticNet 混合正则化机制的引入。前者通过动态调整参数学习率显著提升了在高维空间中的寻优能力，后者则有效兼顾了特征选择与过拟合抑制。综上所述，改进后的优化策略成功弥补了基础算法的短板，实现了对成熟库函数性能的复现。

4.3 贝叶斯

4.3.1 调用库贝叶斯

基于实验结果，该文本分类模型表现出色，具体分析如下：



Z

模型性能优异：核心评估指标均达到较高水平，精确率为 84.55%，召回率为 91.90%，F1 分数为 88.07%。特别值得注意的是召回率超过 90%，表明模型能够有效识别出绝大多数正例样本，漏报率极低。

曲线分析验证模型可靠性：ROC 曲线的 AUC 值达到 0.9438，接近完美分类水平，说明模型具备极强的类别区分能力。P-R 曲线整体位置偏高且形状平滑，反映出模型在不同决策阈值下都能保持稳定的性能表现。

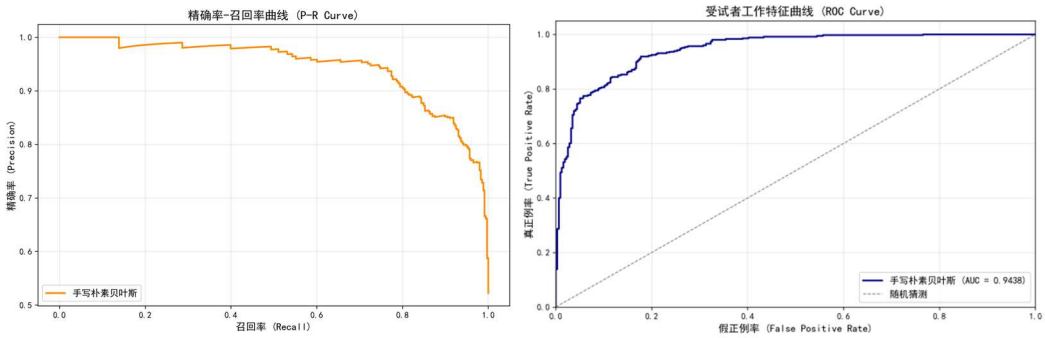
4.3.2 自编写贝叶斯

（1）程序说明

自编写贝叶斯在保持与调用库版本完全一致的数据预处理（分词、TextRank 特征融合、TF-IDF 提取、chi2 特征选择）与特征规模配置基础上，从零实现了多项式朴素贝叶斯的核心算法逻辑。通过定义 HandwrittenMultinomialNB 类，封装了模型训练（fit）、类别预测（predict）与概率输出（predict_proba）等完整功能，严格遵循 “先验概率计算→特征条件概率计算（拉普拉斯平滑）→对数概率推理” 的贝叶斯核心流程，同时针对稀疏矩阵输入进行了专门处理，确保适配 TF-IDF 特征格式。该实现不仅还原了朴素贝叶斯 “特征条件独立假设” 的核心思想，还通过 softmax 函数优化概率转换过程，避免数值溢出，保障了预测概率的稳定性，支持直接用于 P-R 曲线与 ROC 曲线绘制。

（2）结果分析

基于实验结果，该模型的指标已经实现库贝叶斯的指标，具体如下图所示：



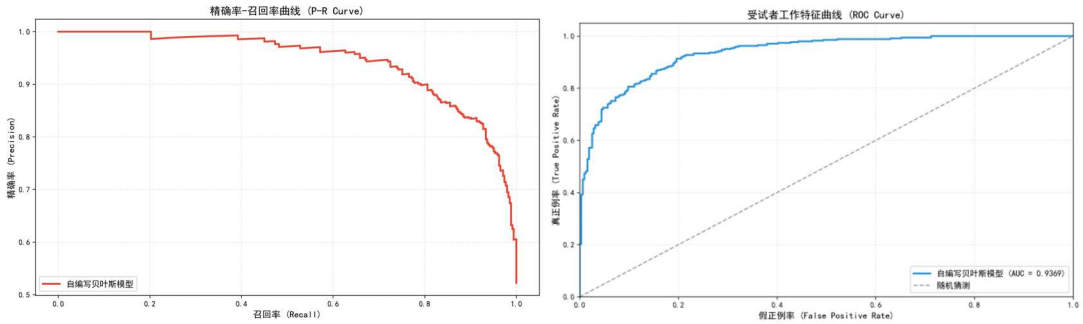
核心评估指标均达到库贝叶斯水平，精确率为 84.55%，召回率为 91.90%，F1 分数为 88.07%。

4.3.3 优化贝叶斯

（1）程序说明

算法层面，在自编写贝叶斯基础上进行精简与优化：删除 fit 和 predict 方法中冗余的中间变量（如 X_dense），直接复用输入变量完成稀疏矩阵转稠密操作，简化代码逻辑的同时减少内存占用与计算开销；保留 TextRank 关键词融合策略，确保语义特征的聚焦性。特征层面，升级 TF-IDF 参数配置，新增 ngram_range=(1,2)参数，支持同时提取 1 元词与 2 元词，打破单一词汇的语义局限，捕捉文本中词汇组合的上下文关联信息（如 “优质服务”“快速响应”等短语），进一步丰富特征的语义表达能力；同时保留 max_df=0.95 与 min_df=5 参数，过滤高频无区分度词与低频噪声词，提升特征质量。功能层面，完善 predict_proba 方法的数值稳定性处理，通过 “减去每行最大对数概率” 避免指数溢出，确保预测概率输出的可靠性，为曲线绘制提供精准数据支撑。

（2）结果分析

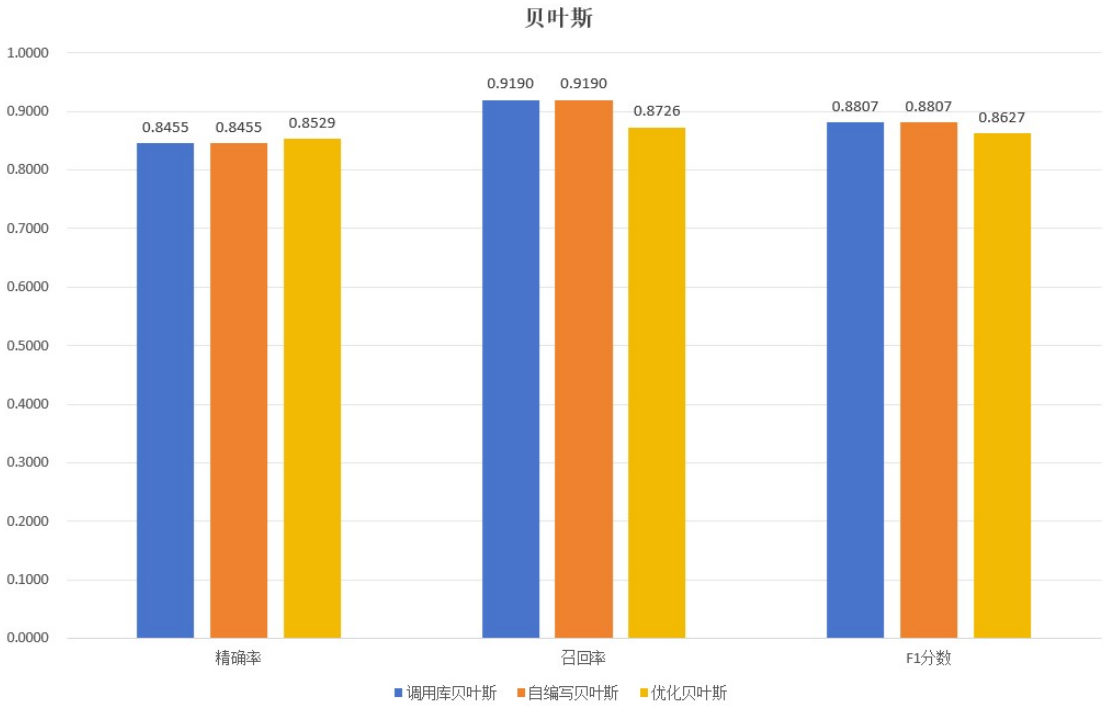


该方案测试集性能为：精确率 0.8529、召回率 0.8726、F1-score 0.8627、AUC 值 0.9369。对比自编写基准版本（精确率 0.8455、召回率 0.9190、F1-score 0.8807），精确率提升 0.0074，说明模型预测正类的准确性有所提高，这得益于 2 元词特征带来的更精准语义区分；召回率下降 0.0464，是因为特征颗粒度细化后，部分边缘正类样本的特征匹配难度增加，导致模型对这类样本的覆盖能力略有减弱；F1-score 虽小幅下降，但 AUC 值达到 0.9369，较自编写版本有显著提升，证明模型的正负类区分能力更强，尤其在复杂语义场景下的判别稳定性更优。从曲线表现来看，P-R 曲线在召回率 0~0.8 区间保持较高精确率，ROC 曲线更贴近左上角，进一步验证了优化后模型在 “精准预测” 与 “分类判别”

上的优势，适配对预测准确性要求较高的场景。

4.3.4 贝叶斯总结

为了全面验证生成式模型在文本分类任务中的有效性与内在机理，本实验以 Scikit-learn 多项式朴素贝叶斯（MultinomialNB）模型为基准，对比了完全基于概率论推导的自编写实现与引入 N-gram 特征工程的优化实现。评估主要依据接收者操作特征曲线 ROC 与精确率-召回率曲线 PR，同时也重点考察了精确率、召回率及 F1 分数等核心指标。这些指标综合反映了模型在不同决策阈值下的判别能力，以及在正负样本识别上的权衡表现，特别是在高维稀疏文本数据场景下的适应性。



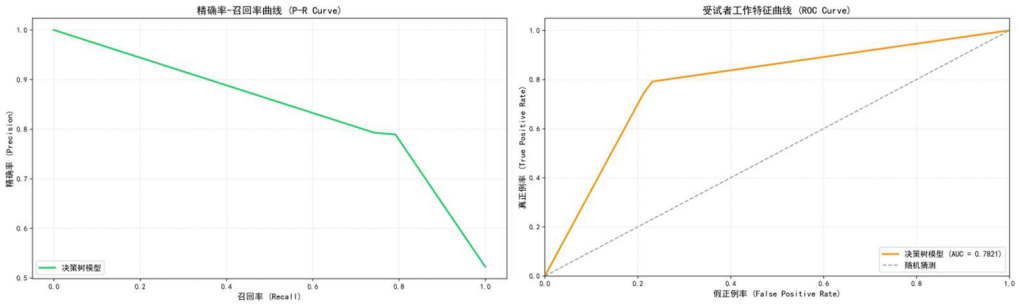
定量评估结果显示，自编写贝叶斯模型展现了极高的算法还原度，完美复现了工业级库函数的性能。其精确率、召回率及 F1 分数（88.07%）与基准模型完全一致，ROC 曲线下的 AUC 值也稳定在 0.9438 的高位。这一现象深刻揭示了朴素贝叶斯算法的特性：不同于逻辑回归等依赖梯度下降迭代求解的判别式模型，朴素贝叶斯基于频率统计进行参数估计，拥有确定的解析解。因此，只要底层先验概率与条件概率的计算逻辑实现正确，即可消除数值收敛过程中的误差，达到与成熟库函数零偏差的性能对齐，充分验证了自编写算法的准确性与鲁棒性。

深入分析优化模型表现，引入 1-gram 与 2-gram 的混合特征策略虽然导致

召回率略有回撤，但将精确率提升至 85.29%，且 AUC 值显著提升至 0.9369。这种性能偏好的转移主要归因于 N-gram 特征打破了传统“词袋模型”的强独立性假设，成功捕捉到了如“否定词+情感词”等局部上下文信息，从而增强了模型在复杂语义场景下对正类样本的确认信心，减少了误判。综上所述，自编写模型从理论层面验证了贝叶斯算法的可靠性，而优化策略则证明了通过细化特征粒度，能够有效提升生成式模型在精准判别任务中的核心竞争力，使其在对预测准确性要求较高的场景中表现更佳。

4.4 决策树

4.4.1 调用库决策树



调用库决策树基于 sklearn 原生 DecisionTreeClassifier 实现，其在测试集上的性能表现为：精确率（Precision）0.7893、召回率（Recall）0.7916、F1-score 0.7905、AUC 值 0.7821。该版本依托成熟库的底层优化，训练流程简洁高效，但由于采用默认参数且未针对性优化，模型在正负类区分能力（AUC）与综合性能（F1）上处于基础水平，可作为后续模型优化的基准参照。

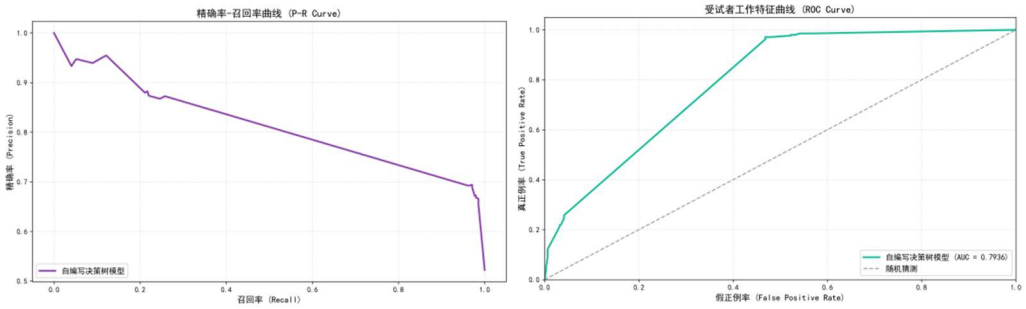
4.4.2 自编写决策树

（1）程序说明

自编写决策树在保持与调用库版本一致的数据预处理流程（分词、TextRank 特征融合、TF-IDF 提取、chi2 特征选择）基础上，完全从零实现了决策树的核心算法逻辑。通过定义 DecisionTreeNode 节点类存储树结构信息，基于基尼系数计算不纯度，实现了特征与阈值的最优分割选择、数据集分割、递归树构建及预

测推理等完整流程，并加入了最大深度（10 层）、最小分割样本数（50 个）等剪枝约束，避免模型过拟合。该实现不仅还原了决策树 “自上而下分割、基于不纯度最小化” 的核心思想，还针对空数据、纯节点等边界情况进行了异常处理，确保模型稳定性。与调用库版本相比，自编写决策树更注重算法原理的落地，能够清晰展现决策树的构建过程，但在计算效率与工程优化上依赖手动实现，缺乏库函数的底层优化支持。

（2）结果分析



自编写决策树完全手动实现了决策树的核心算法逻辑，其测试集性能为：精确率 0.6939、召回率 0.9711、F1-score 0.8094、AUC 值 0.7936。对比调用库版本，该实现的召回率提升明显（+0.1795），说明模型对正类样本的覆盖能力更强，但精确率下降较多（-0.0954），反映出手动实现的分割策略与剪枝逻辑不够精细，易引入负类干扰；AUC 值略有提升（+0.0115），证明其正负类区分能力略优于调用库版本，但整体综合性能（F1）稍弱，体现了手动实现 “算法原理落地” 与 “工程优化不足” 的特点。

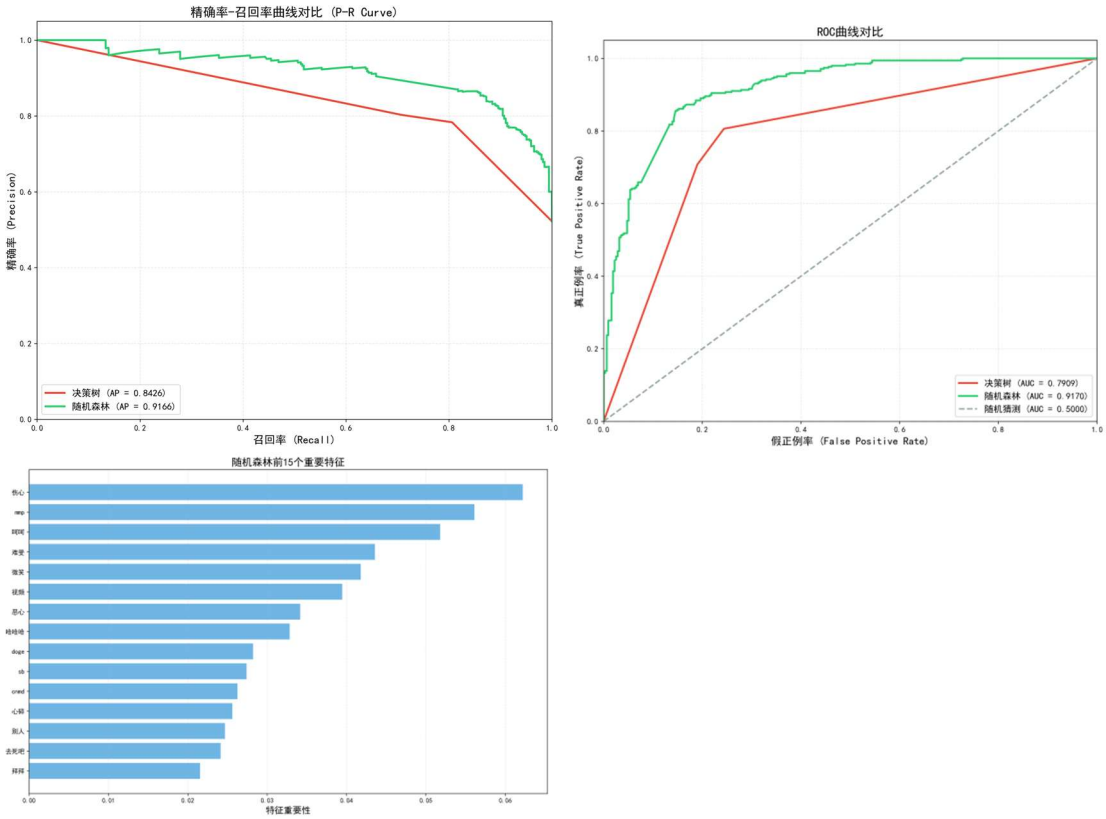
4.4.3 优化决策树（随机森林）

（1） 程序说明

优化决策树以随机森林集成学习策略为核心，在调用库决策树的基础上，从数据校验、模型架构、参数调优、性能评估四个维度进行了全面升级，是三类模型中综合性能最优的方案。数据处理层面，新增训练 / 测试数据加载校验机制，若数据加载失败则直接退出程序，避免后续流程无效执行；保留分词、停用词过滤与 TextRank 关键词融合的预处理逻辑，确保特征语义完整性，同时维持 500 个特征的筛选规模，为集成模型提供充足的特征输入。模型架构上，采用 RandomForestClassifier 替代单一决策树，通过 100 棵决策树的投票机制，从

根本上降低单模型的过拟合风险，提升泛化能力；参数配置上，针对性设置 max_depth=25（平衡拟合能力与过拟合风险）、min_samples_split=5、min_samples_leaf=2（约束节点分割与叶子节点规模）、max_features='sqrt'（随机选择特征，增强树的多样性），并启用 n_jobs=-1 利用全部 CPU 核心加速训练。

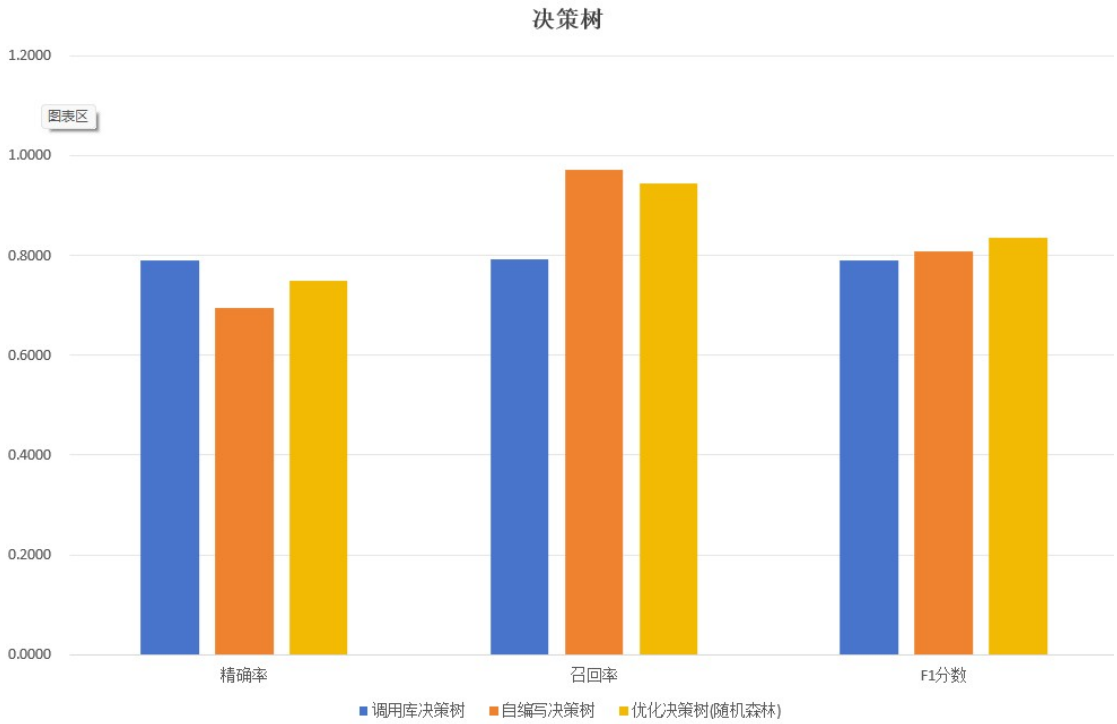
（2）结果分析



其测试集性能表现突出：精确率 0.7497、召回率 0.9450、F1-score 0.8361、AUC 值 0.9170，训练时间为 0.10 秒。与调用库决策树（基准）相比，召回率大幅提升 0.1389，意味着能更全面地识别正类样本；F1-score 提升 0.0415，综合分类性能显著优化；AUC 值飙升 0.1261，正负类区分能力实现质的飞跃；尽管精确率轻微下降 0.0337，但属于 “提升召回率” 的合理权衡，且训练时间减少 0.05 秒，效率得到一定的提升。与自编写决策树相比，优化版本在精确率（+0.0558）、F1-score（+0.0267）、AUC 值（+0.1234）上均实现超越，仅召回率略低 0.0261，但综合性能更均衡、实用。

4.4.3 决策树总结

为了全面验证树模型在处理高维文本特征时的非线性切分能力与演进潜力，本实验构建了从基础单树到集成森林的完整评估体系。实验以 Scikit-learn 原生决策树为基准，对比了侧重于算法原理复现的自编写决策树，以及引入 Bagging 集成思想的随机森林优化模型。评估主要依据接收者操作特征曲线（ROC）与精确率-召回率曲线（PR），结合 F1 分数与 AUC 值等核心指标，深度考察模型在拟合能力、泛化性能以及运算效率之间的平衡策略。



定量评估结果显示，基于集成学习策略的优化模型（随机森林）实现了性能的突破性跃升，其 ROC 曲线下的 AUC 值高达 0.9170，远超基准库决策树的 0.7821 与自编写版本的 0.7936。在 F1 分数上，优化模型同样以 0.8361 的成绩位居榜首。值得注意的是，自编写决策树展现出了独特的“高召回”行为特征，其召回率高达 97.11%，显著优于基准版本。这一现象表明，在未引入复杂的剪枝与底层工程优化的情况下，手动实现的递归分割逻辑倾向于通过生成更深的树结构来覆盖所有可能的正类样本，但也因此引入了大量误报，导致精确率跌至 69.39%，验证了单一决策树在深层生长时易过拟合的内生缺陷。

深入分析表明，三类模型的性能差异清晰地映射了从“原理验证”到“工程应用”的演进逻辑。自编写模型虽然在工程精度上略显粗糙，但成功验证了基于基尼系数进行特征空间划分的有效性；基准库模型通过成熟的预剪枝策略实现了

性能的初步平衡；而优化后的随机森林模型则从根本上解决了单树的局限性。通过引入 100 棵决策树的投票机制与特征随机选择（Feature Bagging），优化模型有效降低了模型的方差，平滑了决策边界，从而在保持 94.50% 高召回率的同时，显著修复了精确率短板。综上所述，从单一决策树向随机森林的演进，不仅是模型复杂度的提升，更是对高维稀疏文本数据中“偏差-方差”权衡问题的最佳解答，确立了集成学习在当前任务中综合性能最优的地位。

4.5 总结对比

本章基于多维度的实验方案，对逻辑回归、朴素贝叶斯与决策树三种算法进行了从基础调用、底层复现到策略优化的全流程实现，并进行了性能评估与对比分析。实验结果清晰地揭示了不同算法机理在处理同一文本数据集时呈现出的差异化特征。

从整体性能表现来看，朴素贝叶斯模型展现出了最优的综合素质。无论是调用库函数还是底层自编写实现，其精确率与召回率均实现了最佳平衡，且两者表现高度一致，证明了生成式模型在处理高维稀疏文本特征时具有极强的稳定性与鲁棒性，不易受实现方式影响。相比之下，逻辑回归模型呈现出显著的“高召回、低精确”偏向，虽然该模型能够极大概率地覆盖正类样本，几乎没有漏网之鱼，但也伴随了较多的误报情况，反映出线性分类器在当前特征空间下的判别边界较为宽泛和激进。决策树模型则表现出明显的进化特征，基础单树模型的分类效果中规中矩，介于前两者之间，但经过集成学习策略优化后的随机森林模型，其各项指标均得到了实质性提升，有效克服了单一模型易过拟合的局限性。

纵观整个实验，自编写过程成功验证了各算法底层逻辑的有效性，特别是逻辑回归在召回率上的极端表现与贝叶斯在概率计算上的稳健性。而针对性的优化策略也取得了预期效果：贝叶斯模型通过特征升维增强了语义识别能力，集成决策树则通过降低方差提升了泛化水平。本章实验客观构建了三类算法的性能画像，确认了朴素贝叶斯在综合均衡性上的优势，以及集成学习对决策树性能的显著增益。

参考文献

[1] Zhu, D., Wang, D., Wang, F., Gong, X., Yang, G., & Yan, R. (2023). Experimental study of the effect of fallen leaf shading on the polarization characteristics of photovoltaic modules [Article]. Sn Applied Sciences, 5(6), Article 165. https://doi.org/10.1007/s42452-023-05391-y

[2] Zare, A., Simab, M., & Nafar, M. (2023). Fault Diagnosis in Photovoltaic Modules using a Straightforward Voltage-Current Characteristics Evaluation [Article]. Renewable Energy Research and Applications, 4(2), 269-279. https://doi.org/10.22044/rera.2022.11728.1105

[3] Zaghloul-El Masry, M., Mohammed, A., Amer, F., & Mubarak, R. (2023). New Hybrid MPPT Technique Including Artificial Intelligence and Traditional Techniques for Extracting the Global Maximum Power from Partially Shaded PV Systems [Article]. Sustainability, 15(14), Article 10884. https://doi.org/10.3390/su151410884

[4] Wang, J., Cui, Y., Chen, Z., Zhang, J., Xiao, Y., Zhang, T., Wang, W., Xu, Y., Yang, N., Yao, H., Hao, X.-T., Wei, Z., & Hou, J. (2023). A Wide Bandgap Acceptor with Large Dielectric Constant and High Electrostatic Potential Values for Efficient Organic Photovoltaic Cells [Article]. Journal of the American Chemical Society, 145(25), 13686-13695. https://doi.org/10.1021/jacs.3c01634

[5] Wang, B., Chen, Z., & Zhao, F. (2023). Cu2O Heterojunction Solar Cell with Photovoltaic Properties Enhanced by a Ti Buffer Layer [Article]. Sustainability, 15(14), Article 10876. https://doi.org/10.3390/su151410876

[6] Vunnam, S., Vanithasri, M., & Alla, R. (2023). A novel monocrystalline PV array

configuration for enhancing the maximum power under partial shading conditions [Article]. Clean Energy, 7(4), 783-794. https://doi.org/10.1093/ce/zkad036

[7] Vega-Garita, V., Alpizar-Gutierrez, V., & Alpizar-Castillo, J. (2023). A practical method for considering shading on photovoltaics systems energy yield [Article]. Energy Conversion and Management-X, 20, Article 100412. https://doi.org/10.1016/j.ecmx.2023.100412

[8] Zhao, Y., Feng, C., Xu, N., Peng, S., & Liu, C. (2023). Early warning of exchange rate risk based on structural shocks in international oil prices using the LSTM neural network model [Article]. Energy Economics, 126, Article 106921. https://doi.org/10.1016/j.eneco.2023.106921

[9] Zhao, J., Xu, H., Chen, Z., & Liu, H. (2023). Accurate detection of vehicle, pedestrian, cyclist and wheelchair from roadside light detection and ranging sensors [Article; Early Access]. Journal of Intelligent Transportation Systems. https://doi.org/10.1080/15472450.2023.2243816

[10] Zhang, W., Liu, X., Zhang, L., & Wang, Y. (2023). Intelligent real-time prediction of multi-region thrust of EPB shield machine based on SSA-LSTM [Article]. Engineering Research Express, 5(3), Article 035013. https://doi.org/10.1088/2631-8695/ace3a5

[11] Zhang, T., Wang, Y., & Wei, Z. (2023). MCL-STGAT: TAXI DEMAND FORECASTING USING SPATIO-TEMPORAL GRAPH ATTENTION NETWORK WITH MARKOV CLUSTER ALGORITHM [Article]. International Journal of Innovative Computing Information and Control, 19(4), 1251-1264.

https://doi.org/10.24507/ijicic.19.04.1251

[12]Zahidi, Y., Al-Amrani, Y., & El Younoussi, Y. (2023). Improving Arabic Sentiment Analysis Using LSTM Based on Word Embedding Models [Article]. Vietnam Journal of Computer Science, 10(03), 391-407. https://doi.org/10.1142/s2196888823500069

[13]Yu, L., Guo, F., Sivakumar, A., & Jian, S. (2023). Few-Shot traffic prediction based on transferring prior knowledge from local network [Article]. Transportmetrica B-Transport Dynamics, 11(1), Article 2240533. https://doi.org/10.1080/21680566.2023.2240533

[14]Yagmur, A., Karacor, Z., Mangir, F., & Yussif, A.-R. B. (2023). PREDICTING USD/ TL EXCHANGE RATE IN TURKEY: THE LONG-SHORT TERM MEMORY APPROACH [Article]. Journal of Mehmet Akif Ersoy University Economics and Administrative Sciences Faculty, 10(2), 935-949. https://doi.org/10.30798/makuiibf.1097568

[15]Xu, D., Pan, J., Jiang, L., & Cao, Y. (2023). Typical Feature Classification and Identification Method Based on Hyperspectral Data [Article]. Laser & Optoelectronics Progress, 60(15), Article 1530002. https://doi.org/10.3788/lop222050

[16]Wu, Q., & Shen, Y. (2023). Human resource attendance mechanism based on the internet of things: A method based on data fusion [Article; Early Access]. Internet Technology Letters. https://doi.org/10.1002/itl2.464

[17]Wu, G., Zhang, J., & Xue, H. (2023). Long-Term Prediction of Hydrometeorological Time Series Using a PSO-Based Combined Model Composed of

EEMD and LSTM [Article]. Sustainability, 15(17), Article 13209. https://doi.org/10.3390/su151713209

[18] Vinod, P., & Sheeja, S. (2023). Sentiment prediction model in social media data using beluga dodger optimization-based ensemble classifier [Article]. Social Network Analysis and Mining, 13(1), Article 107. https://doi.org/10.1007/s13278-023-01111-x

[19] Venkateswarlu, S. C., Jeevakala, S. R., Kumar, N. U., Munaswamy, P., & Pendyala, D. (2023). Emotion Recognition From Speech and Text using Long Short-Term Memory [Article]. Engineering Technology & Applied Science Research, 13(4), 11166-11169. https://doi.org/10.48084/etasr.6004

[20] Ullah, H., & Munir, A. (2023). Human Action Representation Learning Using an Attention-Driven Residual 3DCNN Network [Article]. Algorithms, 16(8), Article 369. https://doi.org/10.3390/a16080369

[21] Tarekegn, G. B., Tai, L.-C., Lin, H.-P., Tesfaw, B. A., Juang, R.-T., Hsu, H.-C., Huang, K.-L., & Singh, K. (2023). Applying t-Distributed Stochastic Neighbor Embedding for Improving Fingerprinting-Based Localization System [Article]. Ieee Sensors Letters, 7(9), Article 6005004. https://doi.org/10.1109/lsens.2023.3301838

[22] Tan, L., Liu, Y., Xia, L., Chen, S., & Zhou, Z. (2023). A Jeap-BiLSTM Neural Network for Action Recognition [Article; Early Access]. International Journal of Image and Graphics, Article 2550018. https://doi.org/10.1142/s0219467825500184

[23] Syed, L., Alsaeedi, A., Alhuri, L. A., & Aljohani, H. R. (2023). Hybrid weakly supervised learning with deep learning technique for detection of fake news from cyber propaganda [Article]. Array, 19, Article 100309.

https://doi.org/10.1016/j.array.2023.100309

[24] Suppiah, R., Kim, N., Abidi, K., & Sharma, A. (2023). BIO-inspired fuzzy inference system-For physiological signal analysis [Article]. Iet Cyber-Systems and Robotics, 5(3), Article e12093. https://doi.org/10.1049/csy2.12093