基于 BGE-M3 模型的中文图书自动分类研究

何戈锐 [1]
HE Gerui

摘 要 图书在版编目数据中的图书分类号无法适用于图书馆所特有的馆藏分类原则，若执行图书馆特有的分类规则则需要大量的人工参与。为节省人力资源并提高图书加工速度，文章将 BGE-M3 模型作为文本嵌入层，结合文本卷积神经网络进行中文图书自动分类。基于图书馆真实馆藏数据集进行实证实验，结果表明：所提出的方法相较于主流文本嵌入模型（Word2Vec、BERT）在综合分类准确率上提升 2%~10%，达到 83.76%；在进一步的模型对比中分类精确度（Precision）、召回率（Recall）和 F_1 分数上也明显优于 BERT 模型，所提方法显著提升了图书分类的精度，有效降低了图书分类中人工干预需求。

关键词 图书分类；人工智能；文本嵌入；TextCNN

doi：10.3969/j.issn.1672-9528.2025.12.047

0 引言

图书分类作为图书馆知识组织系统的核心要素，其准确度直接关系到图书馆资源的可用性与图书馆的服务效能。在智慧图书馆建设与阅读推广开展背景下，精确且符合特定馆藏分类要求的图书分类不仅影响到图书的可发现性和读者的检索效率，更是智慧服务和学科服务实施精准知识服务的关键基础。而目前的图书分类在实践层面依旧面临挑战：尽管图书在版编目（cataloguing in publication, CIP）数据含有分类号，但 CIP 数据由出版社主导完成，不仅有分类错误发生 [1]，更是无法适配图书馆特色资源建设的需求。因此，在实际应用馆藏分类策略时难免需要人工的介入，造成编目效率下降的问题。而通过人工智能技术对馆藏分类原则学习后进行自动化的图书分类，就可以大幅提升编目速度，缩短图书加工周期。

从技术实现角度分析，图书分类本质上属于文本分类任务，主流的分类流程是在抽取图书的文本信息后使用文本嵌入（text embedding）技术将离散的文本转化为连续的向量数据，之后将图书的向量表示和类别标签输入到分类模型中进行学习，最终得到一个分类器实现图书的分类任务。该任务的核心要素在于文本嵌入技术与分类模型，尤其是文本嵌入部分能否从文本中提取足够的文本特征直接决定了后续分类模型的性能好坏。

在中文图书分类领域，早期的研究主要依赖于 Word2Vec 等静态词向量模型，例如谢日敏等 [2] 使用 Word2Vec 作为文本嵌入模型进行中文图书分类，马思丹等 [3] 使用 Word2Vec 作为文本嵌入层，使用 K 近邻分类算法进行文本分类。随着深度学习技术的发展，基于注意力机制的预训练模型（bi-directional encoder representations from transformers, BERT）逐渐成为文本嵌入的主流选择，欧阳涛 [4]、刘磊 [5] 和赵旸等 [6] 使用预训练模型 BERT 作为文本嵌入模型进行中文图书分类。然而 BERT 模型在中文语义理解方面存在局限性，针对这一问题北京智源人工智能研究院（Beijing academy of artificial intelligence, BAAI）于 2024 年发布了预训练模型（BAAI general embedding multi-linguality multi-granularity multi-functionality, BGE-M3），BGE-M3 模型提出了一种新颖的自知识蒸馏方法，并通过 2.5 TB 多语言语料训练，展现出对中文文本更强的理解能力 [7]。基于 BGE-M3 模型的强大表现，本研究构建了融合 BGE-M3 文本嵌入层与文本卷积神经网络特征提取层的混合模型架构，采用 Softmax 分类器实现中文图书分类。通过图书馆真实图书数据集的实证测试，验证 BGE-M3 在中文文本表征中的有效性，进而解决传统分类方法在准确率与效率方面的双重困境，最终实现图书加工周期的优化与图书馆服务能力的提升。

1 文本嵌入方法概述

在中文图书分类任务中，文本嵌入技术作为连接原始文本和分类模型的核心环节，承担着将离散字符序列转换为连续向量空间的关键作用。这一过程不仅要捕捉文本中诸如词汇共现关系等表层特征，更需要挖掘语义层面的抽象表示。

在文本嵌入方法发展过程中，早期研究主要使用基于词频统计的词袋模型（bag of words, BoW）。BoW 通过构建词典后使用独热编码（one-hot embedding）技术将词编码为词

1. 湘南学院图书馆 湖南郴州 423000

典索引的独热向量，该方法的优点在于简单快捷并易于理解。但 BoW 的核心缺陷在于忽视了语法结构和语序信息等更高层次的抽象信息，仅仅依赖于词频统计数据难以捕捉并表示词与词之间的语义关联，更无法应对句 - 句和段 - 段等更高抽象层次的信息。同时 BoW 的表层特征提取方式在处理同义词、多义词和复杂语义任务时表现出明显的局限性，导致模型在语义理解层面存在明显不足。

为应对 BoW 模型只能提取表层特征的困境，Mikolov 等[8]于 2013 年提出了 Word2Vec 模型，使用神经网络模型将词汇映射到连续的向量空间，从而捕获词之间的深层信息，提升了语义表示能力。Word2Vec 模型有两种架构：连续词袋模型（continuous bag-of-words, CBoW）和跳字模型（SkipGram），前者通过上下文预测目标词实现高效训练，后者则通过目标词预测上下文以捕捉更细致的语义关系。尽管 Word2Vec 在提取语义信息方面取得了进展，但是依旧无法应对多义词，并且静态词向量的特点导致了该模型无法适应动态语义变化，在复杂任务中的表现受到限制。

2018 年，Google 基于自注意力机制的 Transformer 架构提出了双向编码表示器 BERT，BERT 的核心创新在于使用了双向 Transformer 编码器实现对上下文的深度理解，能够同时考虑上下文的前向信息和后向信息[9]。使用掩码语言模型和下一句预测的预训练手段增强了双向语义关系捕获能力和句间逻辑的建模能力，这两个预训练手段使得 BERT 能够学到更加丰富更加深层的语义信息，在后续的文本分类等任务上表现出良好的性能。

尽管 BERT 在自然语言处理领域取得了突破性进展，但该模型及后续衍生模型（如 RoBERTa、ALBERT 等）在中文语料学习方面存在明显不足。这种语言特定性限制导致其在中文文本处理任务中表现出性能局限。2024 年 BAAI 发布了 BGE-M3 模型，该模型基于 XLM-RoBERTa 进行优化，使用了 2.5 TB 的多语言语料进行预训练，支持从短句到 8 192 个 Token 的长文档处理[5]。BGE-M3 模型在训练中使用创新性的自知识蒸馏的方法提升模型质量，将不同的检索任务中的输出得分进行融合，生成一个综合的“伪目标”，然后使用这个伪目标指导模型自身的训练，实现了一种平衡多种任务的自监督训练方法。

2 中文图书分类架构

在文本分类任务中文本卷积神经网络（text convolutional neural networks, TextCNN）是一种经典的架构，TextCNN 通过多尺度卷积核能够有效捕捉文本的局部语义特征，同时相对于循环神经网络和 Transformer 等模型 TextCNN 模型的参数量更少、计算效率更高，更加适合处理 Word2Vec、BERT、BGE-M3 等大规模文本嵌入模型输出的高维特征向量。Softmax 作为多类别分类器，能够通过交叉熵损失函数直接优化类别分布，确保模型在大规模中文图书分类任务中实现高精度预测，是中文图书分类任务中理想的基线模型。因此，本研究所采用的模型为 TextCNN+Softmax，整体架构总共分为七层：输入层、嵌入层、卷积层、池化层、融合层、全连接层和分类层。

输入层主要功能为将原始的文本信息转化为满足文本嵌入模型输入要求，例如 Word2Vec 和 BERT 模型的输入数据为词序列，输入层则需要将原始文本信息进行分词和去除停用词等处理。

嵌入层负责将输入层输入的文本信息转化为向量形式的嵌入表示。在本研究中分别使用了 4 种不同的嵌入模型用于性能对比，分别是：Word2Vec-CBoW、Word2Vec-Skip-Gram、BERT、BGE-M3。

卷积层是 TextCNN 的核心部分之一，主要用于提取文本的特征信息，本研究通过使用多个不同尺寸的卷积核对输入信息进行卷积操作，提取了多个不同抽象层次的语义信息和文本特征。

池化层是 TextCNN 的另一个核心部分，通过最大池化提取关键特征，对卷积层的输出特征进行降采样，将其转为固定长度的特征向量，能够有效降低模型的过拟合风险。

融合层用于实现多层次特征的融合，将不同卷积核提取到的特征进行拼接融合，将不同层次的语义特征进行融合表示。

全连接层将融合后的特征进行整合重组，通过非线性激活函数使模型能学习到更加复杂抽象的特征组合，以增强模型的特征表示能力。

Sofmax 层将特征转化为类别的概率分布，用于模型在训练过程中的损失计算和模型最终的结果输出。

3 中文图书分类实证研究

3.1 实验数据来源以及数据处理

本研究的数据来源于湘南学院图书馆的馆藏信息，馆藏信息的 MARC 数据为格式化数据，便于数据的获取和清洗。在使用图书管理系统导出馆藏数据后，过滤掉图书摘要信息为空的条目，同时对于上下册、丛书等多本图书摘要信息相同的数据进行了去重操作。最终获得有效数据共计 157 315 条，数据的内容为图书的《中国图书馆分类法》（下称中图法）的一级分类号以及图书的摘要信息。

3.2 实验工具及参数

本次实验使用 PyTroch 作为深度学习的框架，Word-2Vec 模型使用 gensim 包进行训练，分词工具为 jieba 包，

停用词表使用哈工大停用词表，预训练模型 BERT 使用工具包 transformers 进行加载，加载代码为 AutoModel.from_pretrained("bert-base-chinese") 和 AutoTokenizer.from_pretrained("bert-base-chinese")，加载的模型和分词器为 google 使用中文进行预训练的 BERT 模型。预训练模型 BGE-M3 使用工具包 FlagEmbedding 进行加载，加载代码为 BGEM3FlagModel('BAAI/bge-m3', use_fp16=True)。

本研究的实验涉及 4 个文本嵌入模型，两个预训练模型的数据维度分别是 BERT 的 768 和 BGE-M3 的 1 024，而 Word2Vec 的两种架构需要自己进行训练，考虑到一般情况下更高维度的数据能够表示更多的信息，本研究在训练 Word2Vec 模型时将词向量的输出维度设置为 1 024。在使用词向量表示句向量时使用平均法，即每个句子的向量是该句子中所有词的词向量和的平均值。

在图书分类中，数据的批处理大小设置为 128，损失函数使用交叉熵损失函数，优化算法使用 Adam 算法，学习率设置为 0.000 1，遗忘率设置为 0.5。因训练数据有所波动，实验中并未给模型设置早停策略，而是首先训练 50 轮确保训练过拟合，每一轮训练都输出训练数据的精度和损失，同时使用验证集检查验证精度和验证损失，通过可视化精度和损失的方式确定模型最佳的训练轮次。

3.3 模型训练

在模型训练中，分别使用了四种不同的文本嵌入方法：Word2Vec-CBoW、Word2Vec-SkipGram、BERT、BGE-M3，获得文本的嵌入表示后使用 TextCNN+Softmax 进行训练和分类，按照 0.8:0.2 的比例划分训练集和验证集。本次训练的目的是初步判断嵌入模型的性能并通过可 视化的方式确定模型最佳训练轮次，因此模型训练 50 轮，并将验证数据集的精度和损失使用可视化显示为图 1，模型训练结果如表 1 所示。

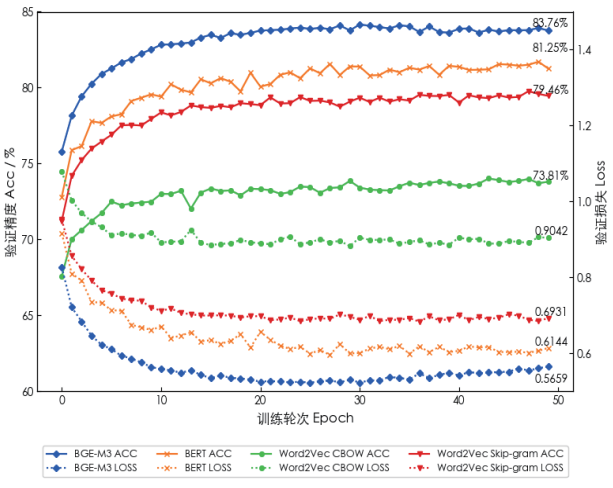

图 1 文本嵌入模型训练中验证精度及损失

表 1 文本嵌入模型训练相关数据

模型名称	训练精度/%	训练损失	验证精度/%	验证损失
Word2Vec-CBow	74.52	0.833 3	73.81	0.904 2
Word2Vec-Skipgram	81.43	0.587 7	79.36	0.693 1
BERT	82.60	0.542 2	81.25	0.614 4
BGE-M3	**92.34**	**0.233 9**	**83.76**	**0.565 9**

表 1 数据显示，以 BGE-M3 作为文本嵌入模型在训练和验证中都表现出最高的精度，训练精度 92.34%，验证精度 83.76%。这表明 BGE-M3 模型在文本嵌入步骤中捕获了更多的信息，使得分类模型在训练数据上有良好的拟合能力，在验证数据上也保持了较高的泛化能力。相比之下，BERT 和 Word2Vec 等模型的表现就较差。

同时，在损失上，BGE-M3 模型也以 0.233 9 的训练损失和 0.565 9 的验证损失显著低于其他模型，表明该 BGE-M3 嵌入表示的文本在训练过程中收敛效果较好。且图 1 结果显示验证损失和验证精度两条曲线的波动性都显著低于其他 3 个文本嵌入模型，表明 BGE-M3 模型的嵌入步骤更加适配中文图书分类这一任务，能更有效地捕捉文本的语义特征。

四种不同的嵌入模型的验证损失都大于训练损失，并且图 1 显示在训练轮次 epoch 达到 20 的时候就出现了验证精度和验证损失几乎已经收敛的趋势，并且在训练到 30 轮附近时验证损失已经最低。综合考虑下比较理想的训练轮次应该在 20 轮，能在性能和时间上取得较好的平衡。

3.4 模型分类性能对比

表 1 的结果显示在训练过程中出现了验证损失高于训练损失，表明模型存在过拟合可能，但验证精度没有下降，该矛盾情况一般出现在数据类别不平衡数据集中。图 2 显示了本次实验中 22 种图书分类的数据。数据集确实存在着较为严重的类别不平衡问题，有 9 类图书的数量都不超过 1 000 条，9 类图书的数量总和占比 2.97%，而数量排名前 4 类的图书（R、T、I、F），数量总和占比则达到了 53.89%。

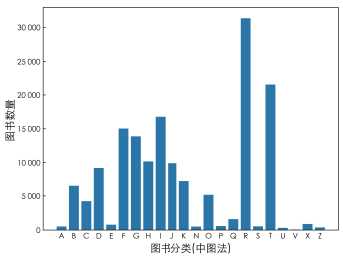

图 2 中文图书分类测试数据

训练结果显示 Word2Vec 模型的两种架构在验证精度上相近且明显低于 BERT 和 BGE-M3，因此，本次实验最终使用 BERT 和 BGE-M3 两种嵌入模型做更详细的分类性能比较，模型的训练 Epoch 设置为 20，数据集按照 0.7:0.15:0.15 的比

例划分训练集、验证集和测试集。在模型训练结束后使用测试集对模型进行测试，生成混淆矩阵并计算 22 个中图法分类号的精确度（precision）、召回率（recall）和 F_1 分数，模型的 F_1 分数结果如图 3 所示。

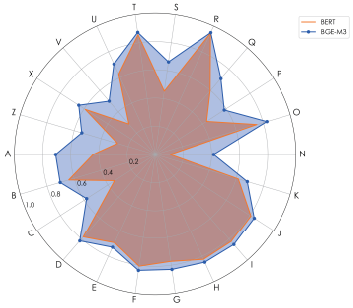

图 3 BERT 和 BGE-M3 模型的 F_1 分数对比

对于类别不平衡的数据，F_1 分数因综合了精确度和召回率能更公平地体现模型性能。图 3 显示了 BERT 和 BGE-M3 两个模型作为嵌入层的 F_1 分数对比，图上清晰地显示 BGE-M3 仅在 R 类图书上与 BERT 模型性能相当，在其他 21 个图书分类上都体现出了对 BERT 模型的领先。

图 4 显示了 BERT 和 BGE-M3 两个模型作为嵌入层的分类精确度对比，精确度表示模型预测为正例中实际正例的比例，例如 BGE-M3 模型在医学类（R）图书上的 Precision 是 94.96%，表示模型预测出来有一万本 R 类图书，其中真正属于 R 类有 9 496 本，不属于 R 类的图书有 504 本。

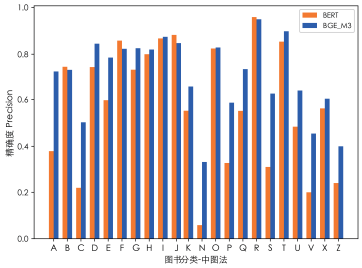

图 4 BERT 和 BGE 模型的分类精度 Precision 对比

图 5 显示了 BERT 和 BGE-M3 两个模型作为嵌入层的分类召回率对比，召回率表示模型能够将正例预测出来的比例，例如 Z 类图书的 Recall 是 85%，表示模型能够将 100 本 Z 类图书中的 85 本预测正确。

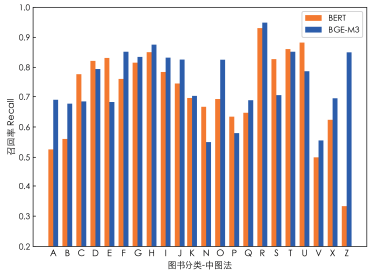

图 5 BERT 和 BGE 模型的分类召回率 Recall 对比

综合实验结果表明，BGE-M3 模型在文本嵌入表示上的性能最好，在 F_1 分数上的结果显示 BGE-M3 模型全面领先 BERT 模型。

4 结语

本文将 BGE-M3 模型作为中文图书分类任务中的文本嵌入模型，大大提高了中文图书分类的精确度，有望加速图书加工步骤，节省采编过程中的人力和时间成本。但本研究也存在一些缺陷，首先只使用了 TextCNN 模型，没有使用更多种类的模型进行验证；其次是研究的数据集存在类别不平衡的问题，导致分类模型的损失出现明显的过拟合特征；最后是分类的数据只使用了图书摘要，没有使用书名，主题词等信息。下一步工作考虑引入更多的分类模型，同时引入更多的图书信息，以进一步提升中文图书分类的准确性。

参考文献：

[1] 杨海燕 . 图书在版编目数据中图书分类错误探析 [J]. 科技与出版 , 2012(8):56-58.

[2] 谢日敏 , 陈杰 , 游贵荣 , 等 . 基于 Word2Vec 的中文图书分类研究 [J]. 云南民族大学学报 (自然科学版), 2018, 27(4): 335-339.

[3] 马思丹 , 刘东苏 . 基于加权 Word2vec 的文本分类方法研究 [J]. 情报科学 ,2019,37(11):38-42.

[4] 欧阳涛 . 基于预训练模型的中文图书自动分类研究 [D]. 昆明 : 云南师范大学 ,2023.

[5] 刘磊 . 基于中文图书自动分类的图书管理系统的研究与实现 [D]. 南昌 : 江西师范大学 ,2021.

[6] 赵旸 , 张智雄 , 刘欢 . 基于层次分类法的中文医学文献分类研究 [J]. 图书馆学研究 ,2021,(21):49-55.

[7]CHEN J L, XIAO S T, ZHANG P T, et al. M3-Embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation[C]//Findings of the Association for Computational Linguistics: ACL 2024. Brussels: ACL, 2024:2318–2335.

[8]MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL].(2013-09-07)[2025-06-29].https://doi.org/10.48550/arXiv.1301.3781.

[9]DEVLIN J, CHANG M W, LEE K, et al. BERT:pre-training of deep bidirectional transformers for language understanding[C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Brussels:ACL,2019: 4171-4186.

【作者简介】

何戈锐（1995—），男，湖南郴州人，硕士，助理馆员，研究方向：智慧图书馆、人工智能，email:libhgr@xnu.edu.cn。

（收稿日期：2025-07-02 修回日期：2025-12-12）