Computer Science and Application 计算机科学与应用, 2024, 14(2), 460-467
Published Online February 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.142046

Hans汉斯

深度森林融合模型优化高校电子图书采购策略

罗　可，陈玫瑰，罗晓倩

邵阳学院图书馆，湖南　邵阳

收稿日期：2024年1月25日；录用日期：2024年2月22日；发布日期：2024年2月29日

摘　要

随着高校对电子图书采购需求的明显增加，为提升采购决策效能，文章提出了一种深度森林融合算法，即LightGBM和CatBoost融合为LHGCAT-XDF的优化模型。该模型兼具LightGBM低内存消耗、和CatBoost低时间复杂度的特点。通过实验结果显示，LHGCAT-XDF相较传统机器学习模型在综合性能上更为卓越，有效克服了传统采购模型在精准性和效率方面的限制，为高校图书馆电子图书采购提供可靠的决策支持。

关键词

深度森林，电子图书，采购模型，决策支持

Optimizing University Electronic Book Procurement Strategy with Deep Forest Fusion Model

Ke Luo, Meigui Chen, Xiaoqian Luo

Department of Library, Shaoyang University, Shaoyang Hunan

Received: Jan. 25th, 2024; accepted: Feb. 22nd, 2024; published: Feb. 29th, 2024

Abstract

With the evident increase in demand for electronic book procurement in universities, this paper proposes an optimized model, LHGCAT-XDF, by integrating the LightGBM and CatBoost algorithms to enhance the efficiency of procurement decision-making. This model has the characteristics of low memory consumption of LightGBM and low time complexity of CatBoost. Experimental results demonstrate that LHGCAT-XDF outperforms traditional machine learning models in comprehensive performance, effectively overcoming the limitations of traditional procurement models in

文章引用: 罗可, 陈玫瑰, 罗晓倩. 深度森林融合模型优化高校电子图书采购策略[J]. 计算机科学与应用, 2024, 14(2): 460-467. DOI: 10.12677/csa.2024.142046

precision and efficiency. This provides reliable decision support for electronic book procurement in university libraries.

Keywords

Deep Forest, Electronic Books, Procurement Model, Decision Support

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

在信息化社会迅猛发展和移动设备普及的背景下，纸质图书的借阅量逐年下降，与此同时，读者对电子图书的需求不断增加。这一趋势不仅推动了对电子图书种类的增长需求，也对电子图书的质量和服务提出了更高要求。

高校图书馆建设的根本目标在于满足读者对更加便捷、多样化学术资源的需求，提升服务水平，以适应信息化社会的发展趋势。这不仅包括硬件设施更新和空间优化，还需提升文献资源的数量和质量，以及全面提高图书馆服务水平的[1]。在这一过程中，图书馆必须不断转型，从传统服务模式向更智能、更符合读者需求的服务模式演变。然而，目前大多数国内高校图书馆仍采用传统的图书采购模式，主要依赖年度经费、采购者经验、师生建议和商家推荐等因素制定采购清单[2]。

虽然一些高校图书馆已经开始使用信息技术构建图书采购决策支持系统，但大多数系统仍以馆藏书目、借阅信息和读者信息为数据源，通过统计分析来指导采购决策[3]。然而，这些数据信息是动态变化的，存在一定的不确定性。如何深入了解读者的阅读需求并以有限的经费购买最符合这些需求的图书，成为图书采购过程中的一个关键问题。

为了提升高校图书馆电子图书采购的效率和质量，挖掘图书属性与读者需求之间复杂多变的潜在关系，文章选择了混合深度森林作为高校图书馆电子图书采购预测的模型。该模型不仅显着提升了预测的精度，相较于传统机器学习模型，还减少了模型预测的时间复杂度和超参数调整的难度，使其能成为图书采购预测领域中更为准确和高效的算法。

2. 深度森林理论模型

深度森林(Deep Forest, DF)是由周志华和冯杰于 2017 年提出的一种基于判定树的集成方法[4]，属于决策树集成方法的一种。相较于深度神经网络(Deep Neural Networks, DNN)，深度森林具有更强的竞争力，主要体现在需要调整较少的超参数，从而降低了超参数调整的时间成本，适应各种大小的数据集以及展现出良好的泛化性等优势，使得深度森林在多个领域得到广泛应用，证明了其在分类和预测任务中的鲁棒性[5] [6]。深度森林主要由两个部分构成，即多粒度扫描(Multi-Grained Scanning)和级联森林(Cascade Forest)。

2.1. 多粒度扫描

多粒度扫描是对输入的特征进行分析，以挖掘特征之间的顺序关系为目的。其具体流程如下：

1) 输入数据：初始时，输入具有 p 维特征的数据。

2) 滑动窗口特征提取：使用多种长度为 k 的滑动窗口在输入特征向量上进行扫描。通过设置步长为 n，计算得到 s 个 k 维特征片段。这样的操作以便更好的提取多尺度、局部的特征信息。

3) 特征片段输入模型：将每个特征片段分别输入随机森林(RF)和完全随机树森林(CRTF)模型。

4) 类概率向量拼接：将所有森林(包括 RF 和 CRTF)输出的类概率向量进行拼接。

5) 生成转换特征向量：将拼接后的类概率向量作为转换特征向量，该向量将作为级联森林的输入。

在整个流程中，可以提取多尺度特征片段和随机森林的信息，来生成更丰富的表示，以供级联森林模型进一步学习和建模。

2.2. 级联森林

在多粒度级联森林的结构中，每个级别(Level)包含多个集成学习分类器，其中决策树森林用作示例，但也可以替换为其他集成学习器，如 XGBoost、LightGBM 或 CatBoost。这层次结构的目的是通过层级组织来构建一个更强大、具有更好泛化性能的集成。

在多粒度级联森林的结构中，每个级别(Level)包含多个集成学习分类器。其中，决策树森林被用作示例，但也可以替换为其他集成学习器，如 XGBoost、LightGBM 或 CatBoost。这一层次结构的目的是通过层级组织来构建一个更强大、具有更好泛化性能的集成。深度森林采用了深度学习的层次结构，每一层的输入由前一层的输入和输出连接而成，从而使模型能够更灵活地学习和组合特征。为了避免过度拟合，每个森林的训练都采用 k-折交叉验证。具体而言，每个训练样本在森林中被使用 $k-1$ 次，产生 $k-1$ 个类别列表，然后将其平均作为下一个级联结构的输入。在级联结构扩展到新的级别时，通过验证集来评估之前所有级联结构的性能。如果评估结果未显着改变或提升，训练过程便停止。因此，级联结构的级别数量由训练过程的动态调整确定。与神经网络不同，深度森林能够通过训练过程自适应地调整模型的复杂度，并在适当的时机停止增加级别，为在控制训练过程中的损失或限制计算资源方面提供更大的灵活性。

2.3. 基于 Lightgbm 和 CatBoost 算法的优化深度森林算法

在小规模数据集上，深度森林算法仍存在性能提升的空间。尽管引入多粒度扫描有助于挖掘特征关系，但也增加了模型的复杂性。文章采用了基于决策树的 LightGBM 和 CatBoost 算法，通过简化多粒度扫描结构和优化随机森林数量，进一步提升了深度森林模型的性能。

在深度森林模型中，大量的决策树会显着增加训练过程的时间和空间成本。为了解决这一问题，文章采用了 LightGBM 进行单边梯度采样[7]，减少了只有小梯度的数据实例的数量。在计算信息增益时，它只使用具有高梯度的数据，大大节省了时间和空间的开销。采用 CatBoost 算法[8]通过优化的梯度提升方法、结合对称树模型和特征量化度量，无需进行独热编码，从而简化了模型训练并减少了数据预处理的复杂性。这为电子图书采购预测模型提供了高效、精准的解决方案，特别是在处理文本类型的电子图书属性时更加灵活和可靠。重新构建后的级联森林结构如图 1 所示。

3. 构建基于优化深度森林算法的电子图书预测模型

以 S 院图书馆所提供的近五年馆藏访问记录为样本，构建精准的采购预测模型。在获取原始数据后，进行了相关数据预处理。同时，通过价值指标筛选出与影响电子图书采访决策相关的字段。采用 BM25 [9] 进行文本数据的特征工程，处理后的样本再以 LHGCAT-XDF 方法建立预测模型。整体流程如图 2 所示。

3.1. 特征分析

不同高校图书馆的采访人员在进行电子图书采访时，各自的决策因素存在差异，而且对这些因素的

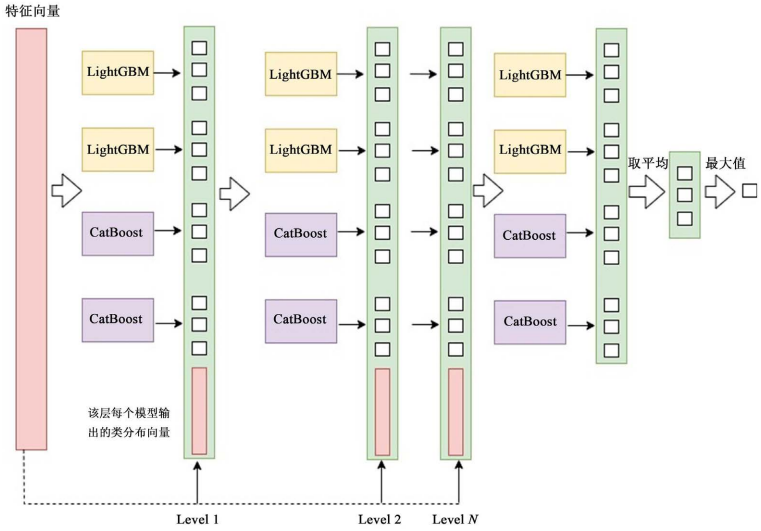

Figure 1. Schematic diagram of the LHGCAT-XDF algorithm
图 1. LHGCAT-XDF 算法的示意图

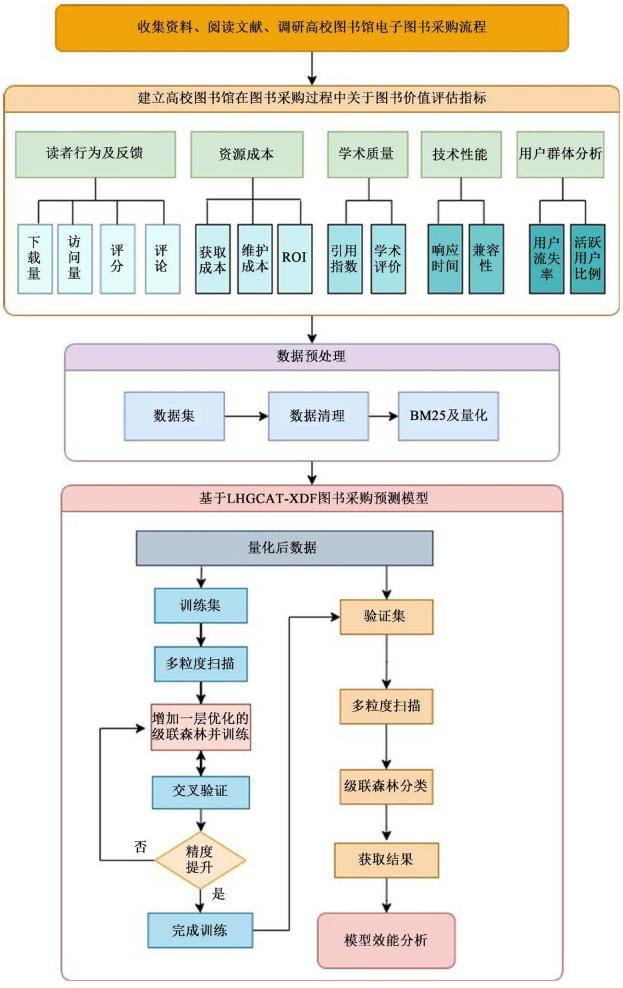

Figure 2. Flowchart of the electronic book procurement forecasting model
图 2. 电子图书采购预测模型流程图

重视程度也不同。通过对大多数高校图书馆电子图书采访实际工作的综合研究，结合学者的观点，分析总结出目前影响高校图书馆采访决策的因素。在此基础上，基于信息增益计算出特征变量，并进行量化，以确定后续电子图书预测模型构建所需的特征变量。

3.1.1. 高校图书馆电子图书采访决策影响因素分析

读者在图书馆门户等系统平台上进行电子图书下载、浏览、评论等操作时，会生成大量的数据。这些数据不仅是图书馆的日志记录，更是支持电子图书采购决策的重要依据。读者的信息可以分为基本信息和行为信息两大类。基本信息包括个人资料和借阅记录，而行为信息涵盖了通过信息检索和数据库访问所产生的历史信息。在行为信息中，包括读者对电子图书的评价和搜索次数等直观数据，也包含浏览过的网页和对同类图书的潜在兴趣等需要深入挖掘的隐性信息。这些信息共同构成一个全面的读者画像，为图书馆提供了深入了解读者需求和偏好的机会。这种深入挖掘的数据不仅有助于满足读者的潜在阅读需求，还在电子图书采购模型中发挥着至关重要的作用。在省内各高校进行调研中，基于对电子图书采购工作的全面了解，总结出影响电子图书采访决策的关键因素如表 1。

Table 1. Overview of influencing factors in electronic book procurement
表 1. 电子图书采购影响因素一览表

影响因素	衡量指标	数据来源	获取方式
用户行为	下载量	电子图书平台后台统计	由电子图书平台提供的下载次数统计
	浏览量	电子图书平台后台统计	通过电子图书平台记录的用户浏览次数
	评分	电子图书平台	通过平台获取评分等级
	评论	用户评论反馈	通过收集用户评论反馈来获取
资源成本	获取成本	购买/订阅电子图书的实际成本	从财务记录或购买合同中获取
	维护成本	电子图书库运营的维护成本	通过财务记录或运营开支明细获取
	ROI	投资回报率	计算电子图书采访的效益和成本比例来获取
学术质量	引用指数	学术数据库或引用工具的数据	通过学术数据库、引用工具等获取
	学术评价	学术评审结果	从相关学术出版物、期刊或平台获取
技术性能	响应时间	电子图书平台性能监测	通过性能监测工具获取
	兼容性	平台测试报告、用户反馈	通过进行平台测试，并收集用户反馈
用户群体分析	用户流失率	用户行为数据分析工具	通过用户行为数据分析工具计算
	活跃用户比例	电子图书平台后台统计	通过统计活跃用户数量和总用户数的比例

通过使用信息增益算法从影响因素中确定特征变量，并对这些特征变量进行量化，可以更准确地分析出影响电子图书采购决策的关键因素，从而有助于制定更为科学和有效的电子图书采购策略。

3.1.2. 数据预处理

根据研究需要，对已获取的数据(包括电子图书数据、读者评论、用户操作日志、图书馆内历年财务数据以及教务系统中读者基本数据情况)进行数据预处理。从中筛选出与电子图书采购相关的信息，构成数据集，并从基本特征、数量特征、时序特征和组合特征中提取具体特征。在处理数据时，发现存在较多缺失值，因此采用网络爬虫等技术手段进行缺失字段值的补全。通过数据分析工具，将电子图书使用数据与电子图书采访数据进行比较，以查找过去五年中哪些电子图书曾被访问过。经过均衡处理后，得

出一个特征矩阵作为深度森林的输入值。具体步骤如下：

第一步，对获取的原始数据进行数据清理，包括处理缺失值和错误数据，采用直接删除方式处理。同时，进行标点符号和停止词的预处理。

第二步，去除完全重复的行数据，并对所有数值型数据进行描述统计分析，以初步了解数据的特征。

第三步，使用 BM25 计算单词与文档之间的相关性，并使用所得相关性建构模型。

3.2. 模型预测

基于 LHGCAT-XDF 的电子图书采购预测模型的工作原理如下：首先，对预处理清洗后的数据按照 8:2 的比例划分为训练集和验证集。随后，将训练集输入模型中的多粒度扫描进行特征选择。所选的特征数据随后被送入优化的级联森林，经过两个 LightGBM 和两个 CatBoost 模型进行训练，并进行 10 折交叉验证。在此过程中，每个 LightGBM 和 CatBoost 模型生成不同比例和种类的样本。最后，对两个 LightGBM 和两个 CatBoost 的输出结果取平均值，将概率最大的类作为样本的预测结果。如果模型精度显着提升，将生成一个新的级联森林层。该层的输入由上一层和当前层的结果拼接而成，作为下一层的训练输入。循环进行训练直至模型精度无显着提升而结束。随后，将验证集输入模型，获得结果后计算模型的各项指标。

3.2.1. 评价指标

在评估模型性能时，通常采用准确率(Accuracy)、精确度(Precision)、召回率(Recall)、特异性(Specificity)和 F1 分数等指标(详见表 2)，以全面评估模型的性能[10]。

Table 2. Model evaluation criteria
表 2. 模型评估标准

评估标准	含义	公式
准确率(Accuracy)	表示整个样本预测正确的样本数量与总体数量的比值	$\frac{TP+TN}{TP+TN+FP+FN}$
精准度(Precision)	表示分类正例样本占所有被分类为正例样本的比例	$\frac{TP}{TP+FP}$
召回率(Recall)	表示在所有正样本中，预测为正样本的概率	$\frac{TP}{TP+FN}$
特异性(Specificity)	表示被正确预测为负类别的样本占所有实际负类别样本的比例。	$\frac{TN}{TN+FP}$
F1 值	表示为精准度与召回率的加权平均值	$\frac{2\times Precision\times Reccall}{Precision+Reccall}$

其中，TP (True Positive)表示真正例，TN (True Negative)表示真负例，FP (False Positive)表示假正例，FN (False Negative)表示假负例。

3.2.2. 实验设计与结果分析

深度森林模型的构建是关键步骤，其中森林的建立尤为重要。为了提高模型的准确度，必须通过反复调整森林的各个参数。文章引入了 LightGBM 和 XGBoost 到级联森林结构中。LightGBM 提供了丰富的参数设置，通过交叉验证方法以优化参数，其中，Learning_rate 是模型的学习率。学习率越大，模型沿损失梯度下降的速度越快，反之亦然。Num_leaves 表示每棵树上的叶片数量，而 Max_depth 则是用于设定判定树回归模型树的最大深度。Feature_fraction 用于进行特征的子抽样，以提高训练速度并防止过拟合。参数设置详见表 3。

Table 3. LightGBM classification model parameter settings
表 3. LightGBM 分类模型参数设定

参数	数值	参数	数值
Learning_rate	0.005	Feature_fraction	0.8
N_estimator	927	Num_leaves	10
Max_depth	−1	Max_bin	245
Bagging_fraction	0.6	Bagging_freq	0

通过对这些参数进行调整，重新构建了级联森林。为了在模型运行时间和准确率之间取得平衡，进行了多次实验。最终，选择了参数 N_estimators = 927，使得模型的准确率达到了 79%。

3.2.3. 模型对比

为了突显深度森林模型预测的优越性，对样本数据使用了传统的学习模型(LightGBM、随机森林、KNN、CNN)进行了预测，并将各个模型的评价指标进行了对比(详见表 4)。具体数值上，深度森林的准确率达到了 79.0%，明显超过其他模型。在精确度方面，深度森林的 83.72%也远高于其他模型。召回率、特异性和 F1 分数同样展现了深度森林相对优秀的趋势。尽管传统的机器学习模型运行时间较短，但是它们的各个评价指标都不如深度森林的结果优秀。

Table 4. Performance evaluation table for various models
表 4. 各类模型效能评估表

模型	LightGBM	随机森林	KNN	CNN	LHGCAT-XDF
准确率(Accuracy)	71.0%	71.5%	69.5%	72.5%	79.0%
精准度(Precision)	77.63%	77.22%	76.71%	77.78%	83.72%
召回率(Recall)	0.59	0.61	0.56	0.63	0.72
特异性(Specificity)	0.83	0.82	0.83	0.82	0.86
F1 值(F1-Score)	67.05%	68.16%	64.74%	69.61%	77.42%

4. 结论

电子图书采购的精准预测对于高校图书馆建设至关重要，然而当前的预测模型存在单一性且准确度不高的问题。为此，提出了一种基于 LightGBM 和 CatBoost 的深度森林算法，即 LIGHT-XDF。在级联森林中引入了 LightGBM 和 CatBoost，通过 CatBoost 提高了模型的预测精准度，同时通过 LightGBM 降低了模型的复杂度。LIGHT-XDF 算法使用读者的行为数据和电子图书馆藏数据进行采购预测。实验结果显示，相较于其他模型，LIGHT-XDF 在综合性能上表现最优。未来工作将通过在不同图书馆馆藏数据集上进一步进行性能测试，来验证 LIGHT-XDF 算法的鲁棒性和泛化能力，并尝试应用多种新技术，提高整体电子图书采购预测的准确度。

基金项目

湖南省哲学社会科学基金项目“人工智能技术在高校图书精准采购中的应用研究”(编号：21YBA179)；湖南省高校图工委科研课题(项目编号：2023L033)。

参考文献

[1] Yang, H.J. (2011) Study on Book Purchase Strategy Based on Web Data Mining Technology. 2011 *International Conference on Control, Automation and Systems Engineering*, Singapore, 30-31 July 2011, 311-313. https://doi.org/10.1109/ICCASE.2011.5997591

[2] Wang, R.H., Tang, Y. and Li, L. (2012) Application of BP Neural Network to Prediction of Library Circulation. 2012 *IEEE* 11*th International Conference on Cognitive Informatics and Cognitive Computing*, Kyoto, 22-24 August 2012, 31-39. https://doi.org/10.1109/ICCI-CC.2012.6311183

[3] 阎世竞, 庞丽川, 张鹏, 等. 基于统计分析的高校图书采购决策[J]. 图书馆工作与研究, 2011(7): 61-63.

[4] Zhou, Z.H. and Feng, J. (2019) Deep Forest. *National Science Review*, **6**, 74-86. https://doi.org/10.1093/nsr/nwy108

[5] AlJame, M., Imtiaz, A., Ahmad, I., *et al*. (2021) Deep Forest Model for Diagnosing COVID-19 from Routine Blood Tests. *Scientific Reports*, **11**, Article No. 16682. https://doi.org/10.1038/s41598-021-95957-w

[6] Dong, L., Qi, J.F., Yin, B.S., *et al*. (2022) Reconstruction of Subsurface Salinity Structure in the South China Sea Using Satellite Observations: A LightGBM-Based Deep Forest Method. *Remote Sensing*, **14**, Article 3494. https://doi.org/10.3390/rs14143494

[7] Ke, G., Meng, Q., Finley, T., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, December 2017, 3146-3154.

[8] 顾崇寅, 徐潇源, 王梦圆, 等. 基于 CatBoost 算法的光伏阵列故障诊断方法[J]. 电力系统自动化, 2023, 47(2): 105-114.

[9] Robertson, S. and Zaragoza. H. (2009) The Probabilistic Relevance Framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, **3**, 333-389. https://doi.org/10.1561/1500000019

[10] 胡陈陈, 吕卫东, 郑江怀, 等. 基于深度森林的在线课程购买行为预测研究[J]. 应用数学进展, 2022, 11(7): 4306-4312.