跨模型在用户意图理解任务中的对比研究

肖明魁 [1]
XIAO Mingkui

摘 要 全面评估和比较三种主流机器学习模型—随机森林（random forest，RF）、LightGBM（LGBM）与XGBoost在用户意图理解任务中的效能表现。针对当前用户意图识别精度尚存提升空间的问题，提出采用多样化的模型策略来优化预测效果。研究方法上，首先对数据集进行了细致的预处理，包括数据清洗、缺失值处理以及特征选择等步骤；接着，对每种模型实施了精细的参数调优，利用网格搜索和随机搜索策略寻找最优配置。研究内容涵盖了模型的训练、验证及测试全过程，获得模型在不同参数设置下的准确率、召回率及F_1分数等关键指标。实验结果表明，XGBoost在经过参数优化后，在用户意图理解任务上取得了最佳的整体性能，随机森林则展现出了良好的稳健性，而LGBM在训练速度上占据优势。

关键词 机器学习；随机森林；分类模型；参数优化

doi：10.3969/j.issn.1672-9528.2024.07.042

0 引言

随着人工智能技术的飞速发展，用户意图理解已成为人机交互、智能客服、推荐系统等多个领域的核心环节。准确识别用户的真实需求对于提升用户体验、优化服务流程至关重要。近年来，深度学习和机器学习模型被广泛应用于意图识别任务，其中，集成学习模型因其实现高效、泛化能力强而备受关注[1]。然而，如何在众多模型中选择最适配于特定任务的算法，以及如何有效调优以达到最优性能，仍是一个具有挑战性的课题。本研究聚焦于三种经典且强大的集成学习模型——随机森林、LightGBM（LGBM）和XGBoost在用户意图理解任务上的性能比较，为领域内应用提供实证指导和理论支撑。

1 相关研究文献综述

在用户意图识别的研究领域，学术界已经取得了重要的进展。研究者们主要集中于提升算法的性能、优化特征工程以及应用深度学习模型。特别是一些研究着重于开发基于循环神经网络（RNN）、长短期记忆网络（LSTM）和Transformer架构的识别模型，突出了捕捉用户意图中的时间依赖性的重要性[2]。同时，由于集成学习方法在解决分类任务时展现出的稳定性与效率，它们也被广泛采用于用户意图的识别工作，例如通过随机森林算法来识别多种意图，或者利用XGBoost算法增强预测的准确性。尽管如此，现有文献多集中于单个模型的优化，而对不同模型在同一任务上的性能对比和综合评估则探讨不足[3]。

2 研究问题与假设

基于以上背景，本研究旨在回答以下核心问题。

随机森林、LGBM和XGBoost在用户意图理解任务中的性能有何异同，哪种模型在综合效能上更为优越；不同模型在处理特定类型用户意图时是否存在显著的性能差异，哪些类型的意图识别更适合特定模型；如何通过模型融合或参数优化进一步提升整体识别准确率。

基于现有文献和实践观察，本研究假设：LGBM和XGBoost因优化算法的先进性，在处理大规模数据集和高维度特征时较随机森林有更高的效率和准确性；不同模型在处理具有不同特征复杂度的意图类别时会表现出不同的性能优势；通过模型集成和参数微调，可以有效提升整体系统在各类意图识别上的综合表现。

3 研究方法

研究采用了三种经典的机器学习模型进行用户意图识别任务，分别是随机森林（random forest，RF）、LightGBM（LGBM）和XGBoost。这些模型均属于集成学习方法，擅长处理分类和回归问题，尤其适用于高维度和非线性数据。

3.1 随机森林

通过集成多个决策树进行预测，每个决策树在训练时使用随机选取的特征和数据子集，最后通过多数投票或平均预测结果来决定最终输出，随机森林模型具有很好的抗过拟合能力和解释性[4]。随机森林模型的构建和优化通常分为五个

1. 江苏经贸职业技术学院信管处 江苏南京 210000

步骤。

3.1.1 决策树的构建

决策树是随机森林的基本组成部分，其构建过程通常是基于信息增益或基尼指数进行特征选择。决策树的公式包括求信息增益公式和求基尼指数公式。

信息增益公式：

$$IG(D, A) = H(D) - \sum_{v\in\{A\}} \frac{|D_v|}{|D|} H(D_v) \quad (1)$$

式中：D 是样本集，A 是特征，Dv 是特征 A 取 v 时的数据子集。

基尼指数公式：

$$Gini(D) = 1 - \sum_{k=1}^{k} p_k^2 \quad (2)$$

式中：D 是数据集，K 是类别数量，P_k 数据集 D 中属于第 k 类的样本比例。

3.1.2 随机特征选择

在构建单个决策树时，随机森林采用了随机特征选择策略。特征选择的公式为：

$$S = \sum_{i=1}^{m} P_i \quad (3)$$

式中：S 表示特征集中所有特征的选择概率之和，P_i 表示第 i 个特征的选择概率。

3.1.3 集成学习

随机森林是一种集成学习方法，其预测过程可以用如下公式表示：

$$y = \arg\max_{c_j} \sum_{i=1}^{n} |\hat{y}_i = c_j| \quad (4)$$

式中：y 是输入数据点的预测类别，n 是决策树的个数，$\hat{y}_i$ 是第 i 棵决策树的预测结果，c_j 是类别 j 的概率。

3.1.4 优化调参

随机森林的关键超参数包括树的数量（n_estimators）、最大深度（max_depth）、最小样本分割（min_samples_split）、最小叶节点样本数（min_samples_leaf）等。通过网格搜索（GridSearchCV）或随机搜索（RandomizedSearchCV）等方法，可以找到这些超参数的最佳组合，以优化模型的性能 [5]。

3.1.5 模型性能评估

模型的性能通常通过交叉验证（如 K 折交叉验证）来评估，使用如准确率、召回率、F_1 分数以及均方误差（MSE）或平均绝对误差（MAE）等指标 [6]。

3.2 LightGBM（LGBM）

LightGBM 是一种高效的梯度提升决策树（gradient boosting decision tree，GBDT）算法框架，它在传统 GBDT 的基础上引入了几项关键技术来提升训练速度和模型效率，主要包括 gradient-based one-side sampling（GOSS）、exclusive feature bundling（EFB）以及基于直方图的优化 [7]。

3.2.1 增益计算

在 GBDT 中，增益计算通常涉及计算分裂前后损失函数的变化，对于一个特征 j 和候选分裂点 s，增益计算公式可表示为：

$$Gain = \frac{1}{2}\left[\sum_{xi\in Ileft} g_i h_i + \sum_{xi\in Iright} g_i h_i - \left(\sum_{xi} g_i h_i\right)^2 / \sum_{xi} h_i^2\right] \quad (5)$$

式中：g_i 是第 i 个样本的一阶导数（梯度），h_i 是二阶导数（Hessian），I_{left} 和 I_{right} 分别是分裂后的左右子集 [8]。

3.2.2 直方图优化

LightGBM 使用了基于直方图的算法来加速训练过程。在传统的 GBDT 中，计算最佳分裂点通常需要遍历所有特征的所有可能分割点。而 LightGBM 预先计算特征值的分布直方图，然后在直方图的边界上寻找最优分裂点，大大减少了计算量 [9]。

3.3 XGBoost

XGBoost（Extreme Gradient Boosting）是一种先进的梯度提升算法，它通过一系列优化手段和算法创新，在速度和性能上超越了传统 GBDT。通过高效地迭代构建决策树，并在每次迭代中优化目标函数，结合正则化项来控制模型复杂度，从而达到高效且准确的预测能力 [10]。其建模过程涉及以下几处关键步骤。

3.3.1 目标函数

XGBoost 的目标函数由损失函数和正则化项组成，用于衡量模型的预测误差和复杂度。损失函数衡量模型对训练数据的拟合程度，而正则化项则用于控制模型的复杂度，以防止过拟合 [11]。对于回归问题，XGBoost 的目标函数可以表示为：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (6)$$

式中：$l(y_i, \hat{y}_i)$ 是损失函数，用于衡量预测值 $\hat{y}_i$ 与真实值 y_i 之间的误差；$\Omega(f_k)$ 是正则化项，用于控制模型的复杂度；n 是样本数量；K 是决策树的数量。

3.3.2 泰勒公式展开

为了优化目标函数，XGBoost 使用泰勒公式将损失函数展开为二阶近似形式，从而将优化问题转化为二次函数优化问题。对于回归问题，泰勒展开后的损失函数可以表示为：

$$l\left(y_i, \hat{y}_i\right) \approx l\left(y_i, y_i^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right) \quad (7)$$

式中：$\hat{y}_i^{(t-1)}$ 是前 t-1 棵决策树的预测值；g_i 和 h_i 分别是损失函数关于 $\hat{y}_i^{(t-1)}$ 的一阶和二阶导数。

3.3.3 决策树的构建

XGBoost 通过构建决策树来优化目标函数。在每一步中，XGBoost 会根据损失函数的负梯度来选择最佳的分裂特征和分裂点，以目标函数最小化。具体来说，XGBoost 会计算每个特征的增益（Gain），选择增益最大的特征进行分裂 [12]。

3.3.4 叶子节点输出

对于每个叶子节点，其输出权重可以通过解决以下优化表达式获得：

$$W_j^* = \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{8}$$

式中：I_j 是分配到第 j 个叶子节点的样本集合。

4 实验设计和分析

本研究采用的数据集来源于 Kaggle 数据网站上的酒店评论数据集，（网址：https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews），其中包含了丰富的用户反馈信息，旨在通过这些评论内容识别用户的评价意图。数据集中每条记录包含两部分关键信息：用户评论文本（Review）和对应的评分（Rating），评分范围通常为 1 至 5 星，代表了用户对酒店服务、设施等方面的满意度。本实验总共随机抽取 3000 个样本作为研究对象，每个评分级别样本数量为 600，分别通过三种建模算法，最后完成结果验证和模型评估。

4.1 数据预处理和词向量化

首先对数据集进行了必要的预处理，包括数据清洗、去除缺失值、随机抽取平均分布的样本等，确保数据质量。并且基于 OpenAI 的 Embedding 技术提升数据特征表达能力，将每个样本数据中的 Review 文本转化为 1536 个词向量后保存于 Embedding 列中。生成结果如图 1 和表 1 所示。

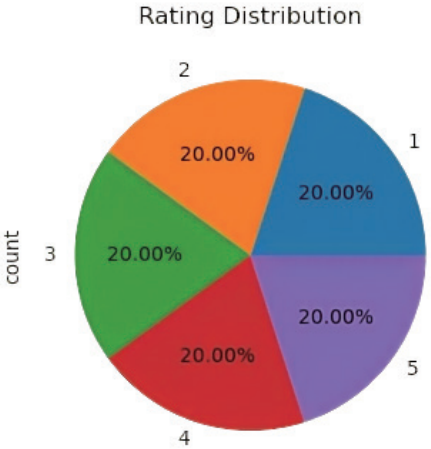

图 1 样本评级分布

表 1 数据预处理结果

	Review	Rating	embedding
2196	hanks staffs stayed not dissapointed, promoti…	5	[0.01656026393175125, -0.0020536514930427074, …
903	conveniently located, feel bit mean giving 2 r…	2	[-0.05223320052027702, -0.008319692686200142, …
2317	suprise needed place stay days roof house did…	5	[0.025705602020025253, 0.002284435322508216, …

4.2 模型训练和参数调优

研究选用了随机森林（random forest）、LightGBM（LGBM）和 XGBoost 作为主要的机器学习模型。对于每种模型进行实例化，并设定了初始的超参数，如学习率、树的最大深度等。采用网格搜索（GridSearchCV）和随机搜索（RandomizedSearchCV）策略来优化模型的超参数。例如，为三种模型定义了参数网格，并通过交叉验证寻找最佳参数组合。最后生成分类问题的性能评估指标，包括精确度（precision）、召回率（recall）、F_1 分数（F_1-score）以及每个类别的支持度（support），并且绘制精确率 - 召回率曲线，通过分析不同类别的曲线，评估每个类别的分类性能。

4.3 实验结果对比分析

本次研究通过模型评估指标和精确率 - 召回率曲线，深入分析了测试集中，三种主流的机器学习模型在用户意图识别任务上的效能表现，以下是对这些模型性能的深入解析，以及它们在特定分类任务上的优势与不足。

4.3.1 随机森林

如表 2 所示，随机森林在五类评分任务中展现了一定的性能平衡。其在第一类别的精确率和召回率分别达到了 0.66 和 0.81，F_1 分数为 0.72，表明模型在该类别上的识别能力较强，但仍有提升空间。然而，在第二至第四类别的表现相对较低，特别是第二类，其精确率和召回率分别为 0.48 和 0.44，说明模型在区分这些类别时存在困难。总体而言，随机森林的准确率为 0.57，宏观平均 F_1 分数也为 0.56，表明模型在各个类别上的表现相当，没有显著偏向某一类别。随机森林的优势在于其鲁棒性好、不易过拟合，且能提供特征重要性信息，适合于特征选择和理解模型决策过程。

表 2 随机森林性能评估指标

	precision	recall	F_1-score	support
1	0.66	0.81	0.72	180
2	0.48	0.44	0.46	180
3	0.54	0.42	0.47	180
4	0.46	0.45	0.46	180
5	0.66	0.74	0.70	180
accuracy			0.57	900
Macro avg	0.56	0.57	0.56	900
Weighted avg	0.56	0.57	0.56	900

如图 2 中所示，RF 模型的 Rating 1 曲线显示了最高精确度和召回率，表明该等级下 RF 模型能够很好地平衡精确度和召回率。Rating 2 曲线略低于 Rating 1，但整体性能仍然较好。Rating 3 曲线增长速度较慢，表明在该等级下 RF 模型的性能受到限制。Rating 4 曲线随着召回率的增加，精确度急剧下降，可能无法处理某些类别的数据。Rating 5 曲线是最低的分类结果，精确度和召回率都较低。

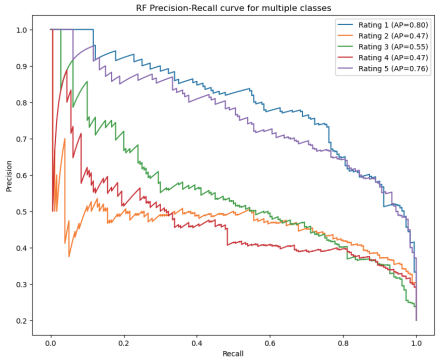

图 2 随机森林 PR 曲线

4.3.2 LightGBM（LGBM）

如表 3 所示，LGBM 在某些类别上的表现有所提升，尤其是在第一类别的精确率提高到了 0.68，召回率也达到了 0.76，F_1 分数为 0.71，相较于随机森林在该类别上的表现有所进步。然而，在第二类和第三类的精确率和召回率上，LGBM 的表现与随机森林相近，表明在这两个类别上的分类效果没有显著改善。LGBM 的准确率同样为 0.57，宏观平均 F_1 分数与随机森林相同，说明在整体效能上两者表现接近。LGBM 的优点在于训练速度快，能高效处理大规模数据集，并且通过优化的算法结构和特征选择机制，能够在某些情况下实现更好的性能。

表 3 LGBM 性能评估指标

	precision	recall	F_1-score	support
1	0.68	0.76	0.71	180
2	0.50	0.45	0.47	180
3	0.50	0.48	0.49	180
4	0.48	0.48	0.48	180
5	0.67	0.69	0.68	180
accuracy			0.57	900
Macro avg	0.57	0.57	0.57	900
Weighted avg	0.57	0.57	0.57	900

如图 3 所示，LGBM 模型的 Class 1 曲线展示了最高的精确度和召回率，与 RF 模型类似，表明该等级下 LGBM 模型能够很好地平衡精确度和召回率。Class 2 曲线略低于 Class 1，但整体性能仍然较好。Class 3 曲线增长速度较慢，表明在该等级下 LGBM 模型的性能受到限制。Class 4 曲线随着召回率的增加，精确度急剧下降，可能无法处理某些类别的数据。Class 5 曲线是最低的分类结果，精确度和召回率都较低。

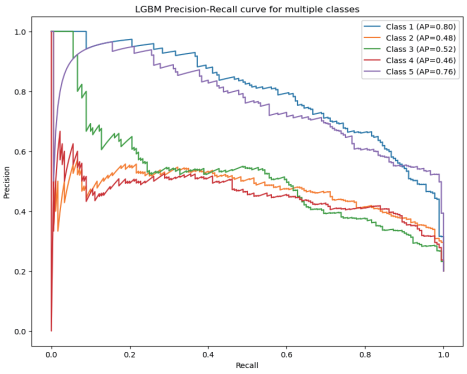

图 3 LGBM PR 曲线

4.3.3 XGBoost

如表 4 所示，XGBoost 在多个类别的表现上略微优于前两者，尤其是在第一类别，其精确率达到了 0.67，召回率更是提升至 0.77，分数为 0.72，显示了更强的分类能力。尽管在第二至第四类别的表现上，XGBoost 的提升并不显著，但在第五类别上，XGBoost 的精确率和召回率分别为 0.67 和 0.73，与随机森林和 LGBM 相比，略有优势。XGBoost 的总准确率为 0.59，宏观平均 F_1 分数为 0.58，略高于随机森林和 LGBM，显示出其在整体分类性能上的优越性。XGBoost 的强项在于其高级的正则化策略，可以有效控制模型复杂度，减少过拟合，同时通过梯度提升框架，能够逐步优化模型，适应复杂的非线性关系。

表 4 XGBoost 性能评估指标

	precision	recall	F_1-score	support
1	0.67	0.77	0.72	180
2	0.52	0.46	0.49	180
3	0.53	0.49	0.51	180
4	0.53	0.48	0.51	180
5	0.67	0.73	0.70	180
accuracy			0.59	900
Macro avg	0.58	0.59	0.58	900
Weighted avg	0.58	0.59	0.58	900

如图 4 所示，XGBoost 模型的 Rating 1 曲线显示了最高精确度和召回率，表明该等级下 XGBoost 模型能够很好地平衡精确度和召回率。

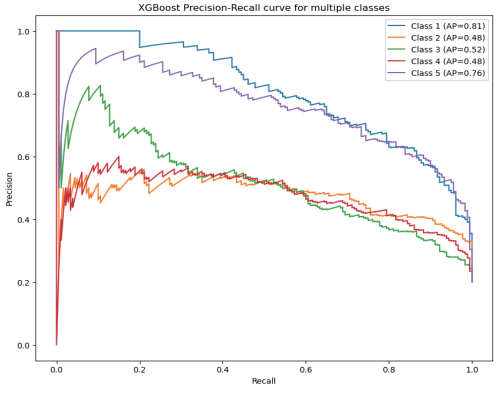

图 4 XGBoost PR 曲线

Rating 2 曲线略低于 Rating 1，但整体性能仍然较好。Rating 3 曲线增长速度较慢，表明在该等级下 XGBoost 模型的性能受到限制。Rating 4 曲线随着召回率的增加，精确度急剧下降，可能无法处理某些类别的数据。Rating 5 曲线是最低的分类结果，精确度和召回率都较低。

4.3.4 综合比较

如表 5 所示，在深入分析随机森林（random forest，RF）、light gradient boosting machine（LGBM）和 XGBoost 三种机器学习模型在分类任务上的表现后，可以从几个关键维度进行比较：精确率 - 召回率平衡、模型复杂度与过拟合风险、资源消耗与速度，以及它们在不同场景下的适用性 [13]。

表 5 模型综合比较

模型	精确率与召回率平衡	模型复杂度与过拟合	资源消耗与速度	适用场景
随机森林（RF）	在某些类别上表现良好（如 Rating 1），但整体平衡性较弱，对类别区分度低的数据处理不足。	较强的抗过拟合能力，但模型复杂度较高，尤其是在处理高维数据时。	训练过程内存消耗大，适合中等规模数据集。	需要解释性强的场景，如金融风控、医疗诊断。
LightGBM（LGBM）	表现均衡，特别是在关键类别上（如 Class 1 和 Class 5），但在某些类别上的表现仍有波动。	通过优化减少资源消耗，降低过拟合风险，但需仔细调参。	高效且资源消耗小，特别适合大数据集。	大数据处理，资源敏感型应用，如推荐系统。
XGBoost	在多数类别上展现了最佳的精确率与召回率平衡，尤其在处理类别区分度高的数据时。	强大的正则化机制有效控制过拟合，但模型复杂度相对较高，需谨慎调参以避免过拟合。	训练时间可能较长，但并行计算能力强，适合中到大型数据集。	追求高性能，对精确率和召回率均有高要求的场景，如竞赛、复杂模式识别。

5 结语

通过以上实验对比分析可以看出，三种模型各有千秋。随机森林在解释性和鲁棒性方面表现出色，但对某些类别区分度较低；LGBM 以其高效的训练速度和处理大规模数据的能力著称，但在某些分类任务上与随机森林相近；XGBoost 则在精确度和召回率上取得了较好的平衡，尤其是在某些关键类别上表现突出，整体准确率最高，展现出其在复杂分类任务中的优势。

在决策过程中选择最合适的模型时，需要考虑包括任务的具体需求、数据的特定特征以及可用的计算资源在内的多种因素。如果任务对模型的可解释性有较高要求，并且处理的数据量相对适中，随机森林算法可能成为一个合适的选择 [14]。对于那些更注重训练效率和能够处理庞大数据集的应用程序，LGBM 算法可能更为适宜。对于寻求最高性能，特别是在解决非线性问题和需要精细化调整的场景，XGBoost 算法，凭借其先进的优化特性和卓越的性能，往往是首选。在实际应用中，还可以考虑模型融合策略，结合各模型的优势，进一步提升分类性能。

参考文献：

[1] 任元凯 , 谢振平 . 大语言模型领域意图的精准性增强方法 [J/OL]. 计算机应用研究 :1-8[2024-04-10].https://doi.org/10.19734/j.issn.1001-3695.2024.02.0022.

[2] 安锐 , 陈海龙 , 艾思雨 , 等 . 基于 BERT-LSTM 模型的航天文本分类研究 [J/OL]. 哈尔滨理工大学学报 :1-10[2024-04-10].http://kns.cnki.net/kcms/detail/23.1404.N.20240405.0028.002.html.

[3] 龙志 , 陈湘州 . 基于图注意力 LSTM 深度学习的季度 GDP 预测应用 [J]. 湖南工程学院学报 (社会科学版), 2024, 34(1):54-64+118.

[4] 叶丽珠 , 郑冬花 , 刘月红 , 等 . 基于鲸群优化随机森林算法的非平衡数据分类 [J]. 南京邮电大学学报 (自然科学版), 2022, 42(6):99-105.

[5] 王诚 , 唐振坤 . 基于随机森林算法的负载预警研究及并行化 [J]. 计算机技术与发展 ,2022,32(11):204-207+220.

[6] 王磊 , 刘雨 , 刘志中 , 等 . 处理不平衡数据的聚类欠采样加权随机森林算法 [J]. 计算机应用研究 ,2021,38(5):1398-1402.

[7] 王震铎 , 马时来 . 基于 LightGBM 和 Stacking 融合算法的高校学生成绩预测研究 [C]// 中国计算机用户协会网络应用分会 . 中国计算机用户协会网络应用分会 2023 年第二十七届网络新技术与应用年会论文集 . 北京 : 北方工业大学 ,2023:5.

[8] 王宏平 , 马雪静 , 彭玉蛟 , 等 . 基于 KMeans 和 LightGBM 模型的大学生公益人群画像分析 [J]. 电脑知识与技术 , 2023, 19(19):39-42.

[9] 赵小强 , 乔慧 . 基于不完整数据的 IHB-LightGBM 心脏病预测模型 [J]. 中国医学物理学杂志 ,2023,40(4):512-520.

[10] 温廷新 , 白云鹤 . 融合 RF-GA-XGBoost 和 SHAP 的虚假新闻群体互动质量可解释模型 [J/OL]. 数据分析与知识发现 :1-18[2024-04-05].http://kns.cnki.net/kcms/detail/10.1478.G2. 20240117.1108.018.html.

[11] 王琦 , 熊莎丽娜 , 詹柔 , 等 . 非平衡数据集下基于 XGBoost 模型的财务舞弊识别研究 [J]. 计算机时代 , 2023 (12): 59-63.

[12] 刘巧红 , 马雨生 , 蔡雨晨 . 基于 XGBoost 算法的糖尿病分类预测模型及应用 [J]. 现代仪器与医疗 ,2023,29(4):1-6+11.

[13] 周化 , 张沁蕙 , 袁志 . 基于文本挖掘与语义识别的用户消费行为分析 [J]. 企业技术开发 ,2016,35(19):7-10.

[14] 黄微 , 张耀之 , 李瑞 . 网络舆情信息语义识别关键技术分析 [J]. 图书情报工作 ,2015,59(21):33-37.

【作者简介】

肖明魁（1979—），男，江苏南京人，本科，工程师，研究方向：机器学习、人工智能、数据挖掘。

（收稿日期：2024-05-08）