基于 Bayesian-Stacking 模型的电影票房预测

李小红[1]，韩淑淑[2]
（1. 山东工商学院 数学与信息科学学院，山东烟台，264005；
2. 山东工商学院 统计学院，山东烟台，264005）

摘要：在我国电影产业中，电影票房是整个电影产业收益的主要来源，对票房进行准确预测不仅可以优化电影投资，助力电影经营主体决策优化，还能促进整个电影产业的健康发展。本文构建了一种基于 XGBoost 的特征选取方法以及 Bayesian-Stacking 集成算法的票房预测模型。首先，我们利用 XGBoost 模型的重要性特征图筛选主要影响因素，可以提高模型特征变量的可解释性；其次，分别构建了 BP 神经网络、XGBoost、Logistic Regression、LightGBM、GBDT 以及 Stacking 模型，再利用贝叶斯优化算法实现上述模型超参数全局寻优后，对电影票房进行预测；最后，引入评价指标进行分析。结果表明：(1) 将贝叶斯优化算法与模型相结合，获得了相对于原模型更高的预测精度；(2) Bayesian-Stacking 模型的电影票房预测精度均优于其它模型。Bayesian-Stacking 模型在电影上映期间预测最终票房具有较高的参考价值，可为有关部门提供决策参考。

关键词：电影票房预测；Stacking 模型；XGBoost；贝叶斯算法
中图分类号：F 经济

Prediction of movie box office based on Bayesian-Stacking model

LI Xiao Hong[1], HAN Shu Shu[2]
(1. Shandong University of Technology and Business,School of Mathematics and Information Science,Yantai shandong,264005;
2. Shandong University of Technology and Business,School of Statistics,Yantai shandong,264005)

Abstract: In China's film industry, the box office is the main source of income for the whole film industry. Accurate prediction of the box office can not only optimize the film investment, help the decision-making optimization of film business entities, but also promote the healthy development of the whole film industry. This paper constructs a feature selection method based on XGBoost and a box office prediction model based on Bayesian-Stacking integrated algorithm. Firstly, we use the importance feature map of XGBoost model to screen the main influencing factors, which can improve the interpretability of model feature variables; Secondly, BP neural network, XGBoost, Logistic Regression, LightGBM, GBDT and Stacking models are constructed respectively, and then the box office of the film during the release period is predicted after the global optimization of the above models is realized by Bayesian optimization algorithm. Finally, the evaluation index is introduced for analysis. The results show that: (1) Bayesian optimization algorithm is combined with the model, and higher prediction accuracy is obtained compared with the original model; Bayesian-Stacking model is superior to other models in box office prediction accuracy. Bayesian-Stacking model has high reference value in predicting the final box office during the film release period, and can provide decision-making reference for relevant departments.

Key words: box office forecast; Stacking model; XGBoost ; Bayesian algorithm

基金项目：山东省自然科学基金面上项目“非线性系统区间二型模糊采样可靠控制研究（ZR2022MF278）
作者简介：李小红（1982），女，副教授、硕导，主要研究方向：偏微分方程、大数据分析. E-mail: xiaohongli@sdtbu.edu.cn

0 引言

近年来，中国的电影产业正在进入快速发展时期。电影产业作为我国文化产业的重要组成部分，促进了我国国民经济消费水平的发展。结合当前国内外文献，关于电影票房的研究方向主要有两个：票房影响因素和票房预测模型。

在电影票房影响因素方面，在早期研究中，学者更侧重于电影特征和市场因素：电影的制作方式、类型、主演、导演、上映档期、发行公司等（例如胡晓红（2018）[1]、何晓雪等人（2019）[2]、Sochay（1994）[3]、陈邦丽等人（2018）[4]和 Julian Hofmann 等人（2016）[5]等）。随着网络蓬勃发展，影评、评分和检索量等消费者因素借助于互联网应运而生且发挥的作用日益重要。申林等人（2020）实证分析了网络评价[6]、Minhoe Hur 等人（2016）分析了影评情感[7]、吴珏（2018）探讨了用户互动行为[8]、史伟（2015）[9]分析了微博评论对票房影响。学者不断完善票房影响因素体系，希望提高预测的准确度，但是随着票房市场影响指标的增多，模型指标因素的可解释性明显降低。

在电影票房预测模型研究中，早期研究者以 Litman（1989）等人为代表，主要通过多元线性回归来进行预测[10]。但目前线性回归模型主要用于验证新变量的引入，例如何晓雪和姜绳（2018）[2]。随着技术的发展，越来越多学者把机器学习应用在票房预测上，如申林等人（2020）[6]、Minhoe Hur 等人（2016）[7]、Ting Liu 等人（2014）[11]和米传民等人（2019）[12]等。杨威（2015)研究了机器学习和线性回归两种模型，通过对比实验发现机器学习模型的精度要优于线性回归[13]。目前，越来越多的学者选择深度学习预测电影票房，尤其以集成算法最为流行。张涛和陈潇潇（2023）采取 Stacking 集成算法，结果显示集成学习预测模型拟合优度优于单一模型[14]，但是没有考虑到影响因素的筛选且优化的 Stacking 模型精度没有显著提升。

针对上面问题，首先，我们利用 XGBoost 模型的重要性特征图筛选主要影响因素，提高模型特征变量的可解释性；其次，利用贝叶斯算法优化 Stacking 组合模型，实现模型超参数全局寻优，有效提升模型的预测性能。

1 指标体系的构建

本研究从电影特征因素、市场因素、消费者因素和微博因素 4 个维度共选取 18 个指标构建电影票房影响因素体系，各指标数据均来自于艺恩网、中国票房网、豆瓣网、时光网、微博官网以及百度指数等网站。由于我国电影市场是从 2010 年以后迅速发展，2010 年之前的电影数据质量层次不齐，收集数据难度大，研究分析也存在一定难度；而 2010 之后的电影，不管在数量方面还是票房方面都有了极大的提升，所以本文研究样本为 2012 年 1 月 1 日至 2022 年 12 月 31 日期间在中国上映的电影，通过对数据筛选、预处理以及分析后，最终选取了 1305 个样本，数据量有保证，横跨年份较长，具有较强的代表性。

表 1 变量的统计分析

一级指标	二级指标	变量	变量说明
电影票房	电影票房	Box office	电影在上映期间的收入
电影特征	电影类别	FM	电影所属的类型，按照公式（1）进行量化

	电影市场	Duration	电影的播放时间
	电影档期	D	档期是一部电影自上映到下映的时间间隔，按照公式（2）处理
	电影制式	Schedule	电影的制作方式一般为 2D、3D、中国巨幕和 IMAX。
	阵容因素	Actor,Director,Editor	本文把导演因素、编剧因素、演员（演员表前 4 的演员）因素统称为阵容因素
	想看人数	Num_pepole	在电影下映前想看某部电影的人数
	银幕数	Screen	银幕是电影放映最为基础性的设施
市场因素	发行公司	Company	承担电影早期宣传和发行的公司
	首日票房	First_day_box	电影上映第一日所取得的票房收入
	百度指数检索量	Baodi_index	消费者对某部电影的百度指数网络检索量
消费者因素	首日影评情感倾向	Film review	对电影上映首日影评进行分词，并利用中文情感分析库进行情感分析，挖掘评论文本的情感倾向
微博因素	微博话题讨论量	Weobo topic	消费者对某部电影的话题讨论量

在电影特征方面，电影产品的感知易用性主要体现在观影体验与情感价值方面，综合了电影的审美性、娱乐性等功能给消费者带来直观的体验与感受，电影本身的制作水平与观看方式直接决定了观众的感知易用性，在指标选取时，将电影类型、电影时长、上映档期、电影制式、影片类型以及阵容等因素加入评估指标中。在市场因素方面，电影作品的感知有用性主要体现在电影市场和消费者因素两方面，市场因素主要受电影的宣传力度、银幕数和潜在消费者的影响，在选取特征变量时，将想看人数、银幕数、发行公司以及首日票房加入指标中。在消费者因素方面，电影票房是整个电影产业收益的主要来源，消费者的态度直接影响着电影票房的收入，而且消费者对电影的网络检索数值越高，则说明该电影的热度越高，也就代表着其潜在消费者多，进而影响着电影的票房。目前网络搜索指数众多，国外学者使用的网络搜索数据大多来自 Google 搜索引擎，而对于中文内容的搜索而言，百度指数在一众搜索指数中脱颖而出，更具代表性。所以选择把消费者影评的情感分析和百度搜索检索量这两个因素加入消费者影响体系中。在微博因素方面，近几年，微博在我国凭借其独特信息的简短性和发布的实时性等优点在一众社交平台中崭露头角。随着其地位的提升，越来越多的学者在研究票房因素指标中愈发重视这一因素，本文也考虑到这一现实因素，把微博中关于影片的话题讨论量也加入到票房影响因素体系中。

在对指标进行量化时，需要对文本指标进行转化。受篇幅影响本文只列举前两位指标的具体计算。电影类型量化：考虑到各种电影题材的质量层次不齐，进而导致票房的高低，如果只采取简单的取均值来代表这一类型的影响力，无法保证类型影响力的合理性，对此采取将各题材电影平均票房与各题材电影占比相乘的方法对电影题材因素进行量化，量化公式为

$$G_i = \frac{\sum_{j=1}^{n_i} box_j}{n_i} \times p_i \tag{1}$$

电影档期处理过程：过对前人文献总结的基础上，把电影档期划分为五类，即春节档（1/20-2/20）、暑假档期（7/1-9/1）、国庆档期（9/30-10/10）、跨年档（2/25-1/5）和普通档。前人在研究档期时大多引入哑变量，该方法固然易懂便于计算，但没有充分考虑到档期之间的差异性，因此可以引用电影票房对档期变量进行量化，计算方法如下：

$$D_i = \frac{\sum_{j=1}^{n_i} box_j}{n_i} \tag{2}$$

其中，G_i 表示第 i 个电影类型的影响力，D_i 表示的是第 i 个档期的影响力，box_j 表示第 j 个电影的票房，p_i 为该题材电影的占比，n_i 表示第 i 个电影的档期数量。

2 预测模型的构建

集成学习是使用多种学习器进行学习，并使用某种规则将各个学习结果进行集成的一种机器学习方法。Stacking 集成学习算法作为集成学习的一种，主要思想是训练多个机器学习模型，将每个模型的训练输出作为新的训练集，再使用另一个机器学习模型进行最终训练，能够综合比较多个学习器的组合效果，减少标准差和偏差，从而提高模型预测精度。Stacking 通常都是组合多种不同的基学习器形成的学习模型，该策略可以构建两层模型，如图 1 所示：

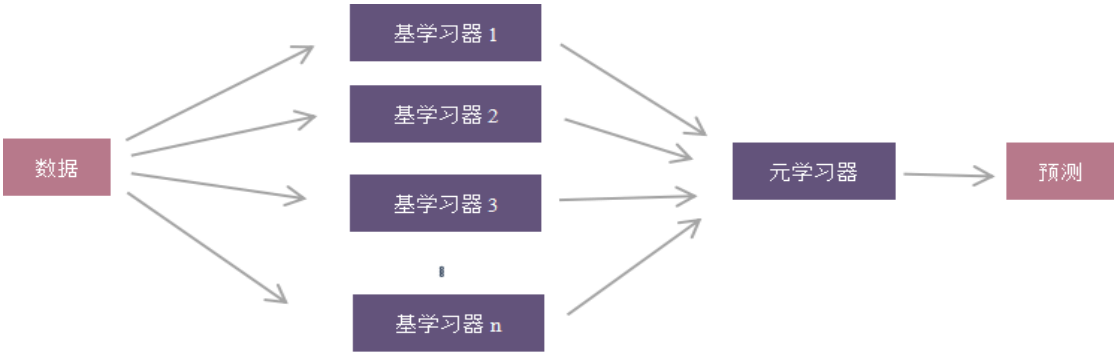

图 1 Stacking 模型流程图

本研究所构建的 XGBoost 特征选取方法和 Bayesian-Stacking 集成票房预测模型，具体步骤介绍如下:

步骤 1：贡献度。对样本(x_i, y_i),其中$i = 1,2...m$，计算负梯度r_{ti}，从而求出(x_i, r_{ti})，然后拟合分类与回归树得到第 t 颗回归树，其对应叶子节点区域为R_{ti},，其中$i = 1,2...m$，其中为 J 回归树 t 的叶子节点个数。在叶子节点的分裂过程中，选取最大的特征及其切分点作为最优特征和最优切分点进行分裂。对叶子区域$j = 1,2...J$，计算最佳拟合值c_{ti}。进而更新强学习器，求出个各个自变量对因变量的贡献度。其作为衡量变量重要性的指标，值越高说明该变量对因变量越重要，从而进行指标筛选。

步骤 2：基学习器。在选择将所有训练集均纳入学习器的迭代训练中，并通过交叉验证的方法对训练集加以控制，规避或减小过拟合风险。在多数研究中，一般选择 5 或者 10 折交叉验证。本文采用十折交叉验证的方法，将训练集划分为 10 等份，9 份样本作为训练集，剩余的一份作为验证集，训练集训练模型，验证集代入训练好的模型中，得到了一个预测结果，因为有 10 个训练模型，所以最终会得到 10 种预测结果，并将 10 种测试集的输出结果求取均值得到新的测试集。对其他模型进行同样的训练过程，总共训练 T 个模型，得到新

训练集 $x_{1,}x_2\ldots x_T$，将 T 个新训练集作为元学习器的训练集 X，同理将新的测试集 $Y_{1,}Y_2\ldots Y_T$ 作为元学习器模型的测试集 Y。基学习器的目的是要建立起原始数据集和标签之间的关系，因此一般选择复杂度高、学习能力强的学习算法，所以本文选择 LightGBM、XGBoost、BP 神经网络和 GBDT。

步骤 3：元学习器。第二层训练框架可以理解为通过模型在反向训练标签。由于输入的训练集特征数量等于基学习器个数，导致训练集特征变量数量过少，如果这时元学习器依然选用复杂的强模型，可能由于过拟合而让模型融合效果适得其反。因此通常选用简单学习模型，本文选择逻辑回归（Logistic Regression 简称 LR），通过在训练集 X 和测试集 Y 上训练得到元模型的输出。

步骤 4：优化。贝叶斯优化算法是一种样本有效的全局优化技术，可以根据对未知目标函数获取的信息，找到下一个评估位置，从而最快地达到最优解。为了寻找适用于 Stacking 模型的超参数组合，节省调优时间，采用贝叶斯优化算法自动选择最优超参数组合。

3 实证分析

3.1 影响因素的筛选

利用 XGBoost 模型筛选主要影响因素，主要是通过计算出各个自变量在 XGBoost 模型中对电影票房预测的贡献度，其值越高说明该变量对因变量越重要，把贡献度值作为衡量变量重要性的指标，根据该指标数值的高低进行筛选变量。在进行票房影响因素的选择时只选取影响力较大的因素，剔除贡献数值相对较小的变量，从而在输入层对预测模型进行简化[12]。但由于 XGBoost 随机性的特征，当对因变量的重要性分数进行求解时，各个自变量的重要性分数在多次试验中得出的结果是有差异的，但通过观察多次实验结果发现各个指标的贡献度数值在一定范围内上下波动。因此，本文为使数据更具有稳定性采取多次试验求均值的方法对指标因素重要性特征进行选取，具体票房影响因素筛选思路如下所示：

Step1：通过对 XGBoost 进行多次构建以求解指标体系中各自变量的多组贡献度值；

Step2：将求出的多组指标因素的重要性分数计算其平均值，并按照从大到小的顺序排列得到的最终重要性分值；

Step3：根据筛选原则和现实情况选取重要性数值较大的若干个影响因素，从而完成基于 XGBoost 的影响力指标的筛选，得出最终的票房影响因素集合，并在此基础上进行后续票房的预测。

按照上述步骤将特征重要性按照得分大小排序，结果如图 2 所示。在电影票房影响因素中，从总体来看，首日票房重要性分值最高，则说明其重要性程度远高于其他变量；电影评论因素、演员 1、导演和演员 2 的影响力分值排名靠前，由此可见，消费者参与的评论以及影片主创团队的创作核心能力在电影票房影响因素体系发挥着举足轻重的作用。经观察发现在图 1 中出现三个明显的断层，由上到下依此定为断层 1、断层 2 和断层 3，其中在电影制式和演员 3 影响力之间出现的第 3 个断层，其累计重要性数值占据总值的 90%左右，能代表着因变量影响指标体系中绝大部分信息，所以以此为界限进行变量的筛选。因此筛选后的票房影响因素共包含首日票房、百度指数、导演、主演 1、想看人数、主演 2、短评情感倾向、微博话题讨论量、屏幕数、上映档期、编辑和电影制式等因素。

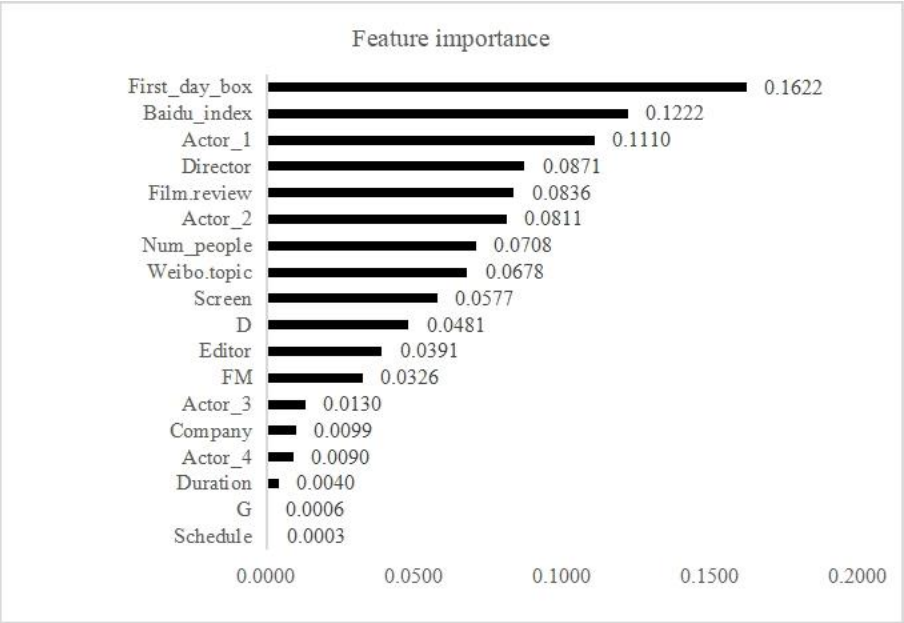

图 2 电影票房影响因素变量重要性分数图

3.2 模型评估

3.2.1 模型的评价指标

评估函数进行模型评估工作也是进行机器学习预测研究中的关键一环。为了科学评估对比各模型的预测效果，本文参考了米传民[12]和杨威[13]的指标选取方法，本文采用三种评价指标来对比模型预测效果，即平均绝对误差（Mean Absolute Error 简称 MAE）、平均绝对百分比误差 (Mean Absolute Percentage Error 简称 MAPE) 和可决系数（R-Squared 简称 R^2)。

表 2 模型评价标准的对比

模型	MAE	MAPE	R^2（%）
XGBoost	0.549	0.745	93.24
GBDT	0.571	0.798	92.35
LightGBM	0.564	0.726	90.12
BP	0.820	0.853	88.70
LR	0.786	0.924	84.37
Stacking	0.493	0.609	97.46

表 2 是各模型评价指标的对比，可知 Stacking 模型的预测结果在 MAE、MAPE 和 R^2 三项评估指标中均优于其他单一模型，其余模型按照拟合优度 R^2 由大到小排序，依次为 XGBoost、GBDT、LightGBM、BP 神经网络、LR，为进一步提升模型预测效果，利用贝叶斯优化算法进行优化参数。选择预测效果较好的四个模型进行优化，其中 B-GBDT、B-XGBoost、B-LightGBM、B-Stacking 表示贝叶斯优化后的 GBDT、XGBoost、LightGBM、Stacking 模型,其预测指标如下所示：

表 3 基于贝叶斯优化模型评价标准的对比

模型	MAE	MAPE	R^2（%）
B-XGBoost	0.467	0.657	95.74
B-GBDT	0.482	0.793	92.05
B-LightGBM	0.559	0.625	97.36
B-Stacking	0.431	0.519	98.12

表 3 是基于贝叶斯优化模型评价标准的对比，由上表可知，各模型的预测效果均得到了一定提升，将贝叶斯优化算法与模型相结合，获得了相对于原模型更高的预测精度。从总体上看，Bayesian-Stcaking 模型的预测结果在 MAE、MAPE 和 R^2 三项评估指标中均优于其他模型。

3.2.2 案例电影的对比

为验证 Bayesian-Stcaking 实际预测效果，随机选取了 10 部受消费者呼吁较高的电影（记为案例电影），本文选择相对误差的评价指标，计算公式见（3）所示。其中，E 表示相对误差，P 为预测值，X 为真实值。

$$E = \frac{P - X}{X} \tag{3}$$

表 4 案例电影预测票房的对比

序号	电影名	真实值	预测值	相对误差（%）
1	唐人街探案 3	452234.30	450403.30	1.34
2	复仇者联盟 4：终局之战	423881.70	437073.70	1.86
3	红海行动	364726.20	368453.20	0.27
4	唐人街探案 2	339768.80	340368.80	0.52
5	我和我的祖国	317118.90	320784.90	2.23
6	独行月球	310291.10	316061.10	1.46
7	唐人街探案 3	452234.30	453795.30	1.23
8	人生大事	171231.40	170205.40	4.31
9	送你一朵小红花	143246.00	146767.00	2.57
10	湄公河行动	118417.40	132574.40	5.25

由表 4 可以看出，所预测的十部电影的相对误差值均在 10%以下，这表明，整体预测效果较好。其中，相对误差最高的是电影《湄公河行动》为 5.25%，最低的是电影《红海行

动》仅为 0.27%，说明所建立的模型拟合效果良好，可以较为准确地预测出电影票房。

4 结论

电影的票房收入作为国内电影市场最重要的指标，是每一部电影最终要追求的目标，关系到了制片方，出品方的决策和电影市场的投资方向，同时也关系到电影生产和营销的各个环节，包括电影题材的选择，剧本的设定，导演、演员、编剧团队的组建，电影宣发策略和营销方式等各个方面的因素。本文在对电影票房体系总结梳理的基础上从以下三方面展开分析本研究。

1.考虑到目前票房预测研究领域很少有学者将微博话题讨论量以及电影首日影评情感分析因素作为电影票房影响因素，所以本研究在加入电影首日影评情感因素的基础上，同时连结电影特征与市场因素构建了一种更为全面电影票房影响因素体系，并在总结文献的基础上对票房各影响指标采用了较为合理的量化方法。

2.为达到简化后续预测模型的输入和提高模型的精度目的，通过构建一种基于 XGBoost 算法的影响力测量模型来进行变量的筛选。

3.通过对以往文献的梳理发现，深度学习在票房预测模型相比具有很多优势，但精度还不够准确，故本文基于 Bayesian-Stacking 集成模型来构建电影票房预测模型，通过对比发现该模型对电影票房的预测性能得到了提升。

因此，研究我国电影票房收入的影响因素并对其进行预测研究，可以为国产电影投资商与制作发行商的投资决策提供建议，使其在投资过程中做出正确的价值判断，实现收益的稳步提升，有利于电影产业与其它产业实现良性循环，共同促进经济发展。

[参考文献] (References)

[1] box office forecast; Stacking model; XGBoost ; Bayesian algorithm
[2] [2]何晓雪,毕圆梦,姜绳.基于网络数据预测电影票房的多元线性回归方程构建[J].新媒体研究,2018,4(05):41-48.
[3] [3]Sochay S. Predicting the Performance of Motion Pictures[J]. Journal of Media Economics,1994,7(4):1-20.
[4] [4]陈邦丽,徐美萍.基于 LARS-SVR 的电影总票房预测模型研究[J].陕西师范大学学报(自然科学版),2018,46(01):10-15.
[5] Hofmann J,Clement M,V ö ckner F. Empirical Generalizations on the Impact of Stars on the Economic Success of Movies[J]. International Journal of Research in Marketing, 2016.
[6] [6]申林,王靖舒.从豆瓣电影看网络评价对电影票房的影响-以 2019 年院线电影为例[J].中国电影市场,2020(08):13-17.
[7] [7]Minhoe Hur,Pilsung Kang,Sungzoon Cho.Box-office forecasting based on sentiments of movie reviews and Independent subspace method[J].Information Sciences,2016,372.
[8] [8]吴珏,潘徐.基于用户内容消费数据的电影票房预测模型探索[J].全球传媒学刊,2018,5(03):96-107.
[9] [9]史伟,王洪伟,何绍义.基于微博情感分析的电影票房预测研究[J].华中师范大学学报(自然科学版),2015,49(01):66-72.
[10] [10]Barry R. Litman Linda S. Kohl. Predicting financial success of motion pictures:The 80s experience[J]. Journal of Media Economics,1989,2(2).
[11] Ting Liu, Xiao Ding, Yiheng Chen. Predicting movie Box-office revenues by exploiting large-scale social media content. Multimedia Tools And Applications, 2014, 12.
[12] [12]米传民,鲁月,林清同.基于加权 K-Means 和局部 BPNN 的票房预测模型[J].计算机系统应用,2019,28(02):15-23.DOI:10.15888/j.cnki.csa.006709.
[13] 杨威.基于微博数据的电影票房预测模型研究[D].安徽大学,2015.
[14] [14]张涛,陈潇潇.基于集成学习的电影票房预测[J].电子制作,2023,31(14):67-70./j.cnki.cn11-3571/tn.2023.14.018.