Modeling and Simulation 建模与仿真, 2024, 13(1), 358-364
Published Online January 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.131034

Hans汉斯

基于预训练－微调策略的电影票房预测

赵　瑞，张明西*，杨　薪，钟昌梅，王博闻，符云杰

上海理工大学出版印刷与艺术设计学院，上海

收稿日期：2023年11月28日；录用日期：2023年12月15日；发布日期：2024年1月16日

摘　要

有监督学习模型对数据量有着较高的依赖，然而现有电影票房数据集较少，导致预测精度降低。针对上述问题，提出一种基于预训练－微调策略的电影票房预测模型。利用电影评分和电影票房之间的相关性，在电影评分数据集上采用预训练的方式，使模型提前获取有关电影的先验知识，同时利用电影间的属性差异信息进行数据增强，最后在电影票房数据集上进行微调，实现对电影票房的预测。实验结果表明，所提方法R^2指标提升了7%，MSE下降了69%。

关键词

电影票房预测，预训练，微调，集成学习模型

Movie Box Office Prediction Based on Pre-Train and Fine-Tuning

Rui Zhao, Mingxi Zhang*, Xin Yang, Changmei Zhong, Bowen Wang, Yunjie Fu

College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai

Received: Nov. 28th, 2023; accepted: Dec. 15th, 2023; published: Jan. 16th, 2024

Abstract

Supervised learning models have a high dependence on the amount of data, however, the existing movie box office dataset is small, which leads to lower prediction accuracy. To address the above problems, a movie box office prediction model based on a pre-training and fine-tuning strategy is proposed. Using the correlation between movie ratings and movie box office, pre-training is used on the movie ratings dataset to make the model acquire a priori knowledge about movies in advance. At the same time, data enhancement is carried out by using the information of attribute differences between movies. Finally fine-tuning is applied on the movie box office dataset to real-

*通讯作者。

文章引用: 赵瑞, 张明西, 杨薪, 钟昌梅, 王博闻, 符云杰. 基于预训练-微调策略的电影票房预测[J]. 建模与仿真, 2024, 13(1): 358-364. DOI: 10.12677/mos.2024.131034

ize the prediction of movie box office. Experimental results show that the proposed method improves the R^2 index by 7% and decreases the MSE by 69%.

Keywords

Movie Box Office Prediction, Pre-Train, Fine-Tuning, Ensemble Learning Model

1. 引言

预测电影票房已经成为一种新型需求，投资人通过衡量电影票房收入和电影制作预算来判断投资回报，同时电影票房预测可以为投资决策、广告策略等提供参考信息，对电影行业发展具有重要意义。

在电影票房预测研究中，研究人员使用计量经济学和机器学习方法对电影票房预测进行广泛研究，基于电影属性分析电影票房的影响因素，再利用影响因素进行电影票房预测[1] [2] [3]。Dai 等人[4]使用灰色关联度计算分析电影票房的影响因素，再使用神经网络对票房进行预测。Wang 等人[5]利用动态异构网络学习演员、导演和公司之间潜在表征，使用深度学习模型从预告片中提取电影质量的高级表示，基于学习到的特征来训练预测模型，进行票房预测。

随着社交媒体的兴起，研究人员开始从数据驱动角度考虑电影票房预测问题[6] [7] [8] [9]。从社交媒体上获取消费者对不同电影的评价，消费者对不同明星的喜爱程度等各种信息，通过对这些信息进行情感分析等处理，再对电影票房进行预测。Asur 等人[10]提出了一种正负极性模型，该模型可以对网络评论进行情感分析，基于这些分析进行电影票房预测。Kim 等人[2]基于社交网络服务数据，提出了非累计票房的预测模型。Shen 等人[11]基于社交网络评论分析电影演员的社会网络特征，结合电影元数据特征和演员社交网络度量等特征，提出 FC-GRU-CNN 电影票房预测模型。QIU 等人[12]利用微博上的影评，计算网络指数和影评来进行票房预测。然而，现有的有监督模型不能很好地适应电影票房数据量较低的预测场景，导致模型预测精度较低。

随着深度学习模型规模的扩大，Bert [13]、GPT [14] [15] [16]等预训练模型取得成功，预训练-微调策略已经被应用在多模态模型[17] [18]、计算机视觉[19] [20] [21]、自然语言处理[22] [23]等领域。在自然语言处理领域，基于在大量文本数据集的预训练后，可以将模型应用在各种自然语言处理的子任务中，并有着较好的性能，例如机器翻译[24] [25]，实体类型推断[26] [27]，序列标记[28]等。预训练可以帮助模型在目标任务中，仅需要少量的目标任务数据，即可完成任务。

除了电影票房之外，电影评分也是评判一部电影质量的重要指标，且与电影票房具有相关性。本文提出了基于预训练–微调策略的电影票房预测模型，在电影评分数据集上采用预训练方式，使模型提前接触到更多的电影数据，学习到更多电影的相关特征信息。然后在少量电影票房数据集上做训练，实现电影票房的精准预测。

2. P-EL 预测模型

本文构建基于预训练–微调策略的电影票房预测模型 P-EL (Pre-Training and Fine-Tuning Ensemble Learning)，通过预训练策略，在一定程度上解决了因电影票房数据量不足导致的预测精度下降的问题。

预训练为模型提供了更加有效的初始化参数，使模型预先学习到电影数据与电影评分之间的变化规律，在电影票房数据集上进行微调后，针对电影票房预测具有较高的预测精度。

2.1. 问题定义

给定电影评分数据集 $A=\{(X_t,u_t)\}_{t=1}^{N}$，表示电影评分数据集中包含 N 部电影的数据，以及电影票房数据集 $B=\{(X_t,y_t)\}_{t=1}^{T}$，表示电影票房数据集中包含 T 部电影的数据。其中 $X_t=(x_1,x_2,x_3,x_4,x_5)$ 为第 t 部电影的电影属性，x_1 表示导演，x_2 表示演员，x_3 表示预算，x_4 表示电影类型，x_5 表示上映时间，u_t 表示电影评分，y_t 表示电影票房。

2.2. 预训练－微调策略

预训练－微调策略表示为：

$$\begin{aligned} Model_{pretrain} &= pretrain(A) \\ Model_{final} &= fine\_tuning(Model_{pretrain},B) \end{aligned} \tag{1}$$

其中，$pretrain$ 作为预训练函数；$fine\_tuning$ 作为微调函数。在电影评分数据集上经过预训练得到 $Model_{pretrain}$，$Model_{pretrain}$ 在目标电影票房数据集上进行微调，得到最终的预测模型 $Model_{final}$，最后利用 $Model_{final}$ 对目标电影票房进行预测。

因为电影评分和电影票房有着高度的相关性，电影评分数据量相比电影票房数据量更多，所以选择电影评分数据作为预训练的数据。经过预训练后，保存训练完的预训练模型。利用在电影评分数据中提前学习到的电影数据和电影评分之间的变化关系，为预测电影票房提供了一个更合理的初始化参数，将训练好的模型和电影票房数据输入到集成学习模型中，并在电影票房数据上进行微调，最终得到电影票房预测模型 P-EL。基于预训练－微调策略的 P-EL 框架图如图 1 所示。

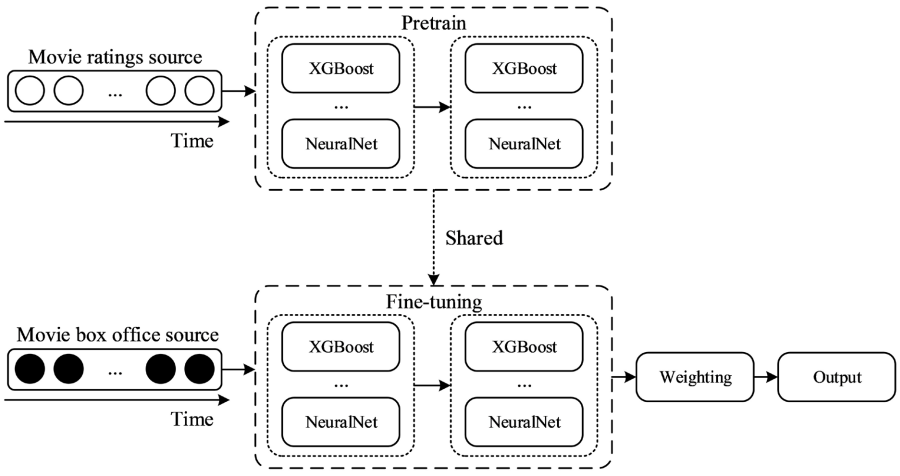

Figure 1. Framework of P-EL
图 1. P-EL 框架图

2.3. 预测模型

P-EL 采用集成学习模型来预测电影票房，集成学习模型可以融合多个模型的优势，避免单个模型的局限性，提高模型预测的整体性能。P-EL 模型中数据处理部分采用时序属性增强的数据增强方法，时序

属性增强可以帮助模型更好地理解电影间的差异，从而提高预测精度。

预测模型的输入为电影的基本属性。我们根据之前的研究和行业专家的建议，选择了 5 个电影基本属性：导演、演员、预算、电影类型和上映时间。不同电影间的属性差异信息导致了不同电影票房的差距，而越短上映时间间隔内的差异信息，对电影票房的影响越大，参考价值也就越高。这里我们选择时序属性增强的方法，对数据进行数据增强。将电影按照时间进行排序，逐一计算相邻电影间的属性差异信息 D_t：

$$D_t = X_t - X_{t-1} \tag{2}$$

将属性差异信息作为预测电影票房时的补充信息。为了加速训练过程，我们对模型的输入进行归一化：

$$f\left(s_i\right) = \frac{s_i - s_i^{mean}}{s_i^{std}} \tag{3}$$

其中，s_i 表示 i 类型的电影属性值，s_i^{mean} 和 s_i^{std} 分别表示 i 类型电影属性的平均值和标准差。

P-EL 模型采用基于多层 stacking 的集成学习模型，P-EL 选取 CatBoost，RandomForest，LightGBM，XGBoost 和 NeuralNet 作为基础模型。P-EL 模型中每一层的堆叠器均为 5 个基础模型，多层 stacking 会将 5 个基础模型输出与原始输入合并，作为下一层堆叠器的输入，类似于残差连接：

$$Z_2 = concat\left(m_1(Z), \cdots, m_5(Z), Z\right) \tag{4}$$

其中，Z_2 表示第二层堆叠器的输入，m_i 表示第 i 个基础模型，Z 表示 P-EL 的输入。最后一层使用集成选择，以加权和的方式聚合模型的预测，并作为 P-EL 模型的输出。

3. 实验

3.1. 实验设置

3.1.1. 实验环境

实验使用的 CPU 是 Intel(R) Xeon(R) Silver 4310，显卡为 NVIDIA GeForce RTX3090，内存为 256 GB。数据预处理和电影票房预测部分由 Python 编写，模型部分使用 PyTorch 框架实现。

Table 1. Statistics of datasets
表 1. 数据集统计

含评分的电影数据	含票房的电影数据			数据类别
	训练集	验证集	测试集	
8792 条	3000 条	970 条	1000 条	演职人员、情节关键词、预算、票房、发行时间、语言、制作公司、评分

3.1.2. 数据集

数据来自 Kaggle 上的电影数据集，包含 2017 年 9 月前发布的电影数据。如表 1 所示，其中，含有电影评分数据的电影 8792 部，含有电影票房数据的电影 4970 部。

3.1.3. 评价指标

模型的预测性能评价指标选用决定系数(R^2)、均方误差(MSE)以及平均绝对误差(MAE)。

R^2 可以反应预测值和真实值的拟合程度。R^2 在 0 到 1 之间，越接近 1，说明拟合程度越好：

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{5}$$

其中，y_i 表示第 i 个预测值，$\hat{y}_i$ 表示第 i 个真实值，$\bar{y}$ 表示真实值的平均值，m 表示测试集中数据总数。

MSE 可以衡量预测值和真实值之间的偏差，*MSE* 越小，说明模型预测效果越好：

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{6}$$

MAE 可以准确反应预测值和真实值的误差大小，*MSE* 越小说明模型预测的误差越小：

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|(y_i - \hat{y}_i)\right| \tag{7}$$

3.1.4. 对比模型

为了验证 P-EL 模型的性能，现与以下基准模型进行对比：

1) 多层感知机(Multi-Layer Perceptron, MLP)：一种最基本的前馈神经网络，MLP 的网络架构为：输入层、隐藏层和输出层。在此之前的研究中，MLP 是用于票房预测最好的模型。

2) XGBoost：一个端到端的梯度提升树系统，针对分类和回归任务有着良好的预测效果。

3) LightGBM：一种基于决策树梯度提升的机器学习方法，其训练速度快，效率高，内存占用低。

4) RandForset：以决策树为基本单元，集成构建随机森林的算法，泛化能力强。

5) CatBoost：一种基于 GBDT 改进的决策树梯度提升的机器学习算法。该方法对预测偏移的处理可以减少模型的过拟合现象，从而提升模型预测效果。

6) NeuralNet [29]：为了适配集成学习模型专门设计的一种神经网络。

7) 集成学习(Ensemble Learning, EL)：无预训练策略的集成学习模型。

3.2. 结果对比

Table 2. Overall performance comparison
表 2. 整体性能比较

模型	R^2	*MSE*	*MAE*
P-EL	**0.77697**	**0.19314**	**0.34584**
EL	0.70742	0.48936	0.38278
MLP	0.53506	1.12587	0.60878
XGBoost	0.63750	0.52337	0.42781
RandomForest	0.62103	0.54037	0.42141
LightGBM	0.63312	0.53642	0.42262
CatBoost	0.66157	0.50885	0.40882
NeuralNet	0.60756	0.61967	0.46815

模型的整体性能如表 2 所示，在所有模型中，基于预训练－微调策略的电影票房预测模型的性能明显优于其他基准模型。P-EL 模型较无预训练－微调策略的 EL，在 R^2 指标上最高提升了 6.96%，在 *MSE* 和 *MAE* 指标上分别降低了 60.53%和 9.65%。EL 内部模型均为采用预训练－微调策略，与这些模型相比，P-EL

在 R^2 上最高提升了 16.94%，最低提升了 11.54%，在 *MSE* 上最高降低了 68.83%，最低降低了 62.04%，在 *MAE* 上，最高降低了 26.13%，最低降低了 15.41%。因为电影评分和电影票房之间的相关性，所以利用含电影评分的电影数据集进行预训练，使模型学习到了关于电影票房的先验知识，提升了模型的预测性能。

4. 结论

本文提出了一种基于预训练－微调策略的电影票房预测模型 P-EL，利用电影评分数据集进行预训练，在电影票房数据集中进行微调，并在 Kaggle 电影数据集上进行模型评估。实验结果表明，P-EL 性能优于其他基准模型，能在一定程度上解决因电影票房数据量少而导致的模型预测精度不足的问题。在未来的研究工作中，将引入电影之外的影视作品数据集，进一步提升模型的预测精度。

基金项目

国家重点研发计划项目(2021YFF0900400)；国家自然科学基金项目(62002225)；上海市自然科学基金项目(21ZR1445400)。

参考文献

[1] Edwards, D.A., Buckmire, R. and Ortega-Gingrich, J. (2014) A Mathematical Model of Cinematic Box-Office Dynamics with Geographic Effects. *IMA Journal of Management Mathematics*, **25**, 233-257. https://doi.org/10.1093/imaman/dpt006

[2] Kim, T., Hong, J. and Kang, P. (2015) Box Office Forecasting Using Machine Learning Algorithms Based on SNS Data. *International Journal of Forecasting*, **31**, 364-390. https://doi.org/10.1016/j.ijforecast.2014.05.006

[3] Du, J., Xu, H. and Huang, X. (2014) Box Office Prediction Based on Microblog. *Expert Systems with Applications*, **41**, 1680-1689. https://doi.org/10.1016/j.eswa.2013.08.065

[4] Dai, D. and Chen, J. (2021) Research on Mathematical Model of Box Office Forecast through BP Neural Network and Big Data Technology. *Journal of Physics*: *Conference Series*, **1952**, Article ID: 042118. https://doi.org/10.1088/1742-6596/1952/4/042118

[5] Wang, Z., Zhang, J., Ji, S., *et al*. (2020) Predicting and Ranking Box Office Revenue of Movies Based on Big Data. *Information Fusion*, **60**, 25-40. https://doi.org/10.1016/j.inffus.2020.02.002

[6] Arias, M., Arratia, A. and Xuriguera, R. (2014) Forecasting with Twitter Data. *ACM Transactions on Intelligent Systems and Technology* (*TIST*), **5**, 1-24. https://doi.org/10.1016/j.inffus.2020.02.002

[7] Liu, T., Ding, X., Chen, Y., Chen, H.C. and Guo, M.S. (2016) Predicting Movie Box-Office Revenues by Exploiting Large-Scale Social Media Content. *Multimedia Tools and Applications*, **75**, 1509-1528. https://doi.org/10.1007/s11042-014-2270-1

[8] Ghiassi, M., Lio, D. and Moon, B. (2015) Pre-Production Forecasting of Movie Revenues with a Dynamic Artificial Neural Network. *Expert Systems with Applications*, **42**, 3176-3193. https://doi.org/10.1016/j.eswa.2014.11.022

[9] Zhou, Y., Zhang, L. and Yi, Z. (2019) Predicting Movie Box-Office Revenues Using Deep Neural Networks. *Neural Computing and Applications*, **31**, 1855-1865. https://doi.org/10.1007/s00521-017-3162-x

[10] Asur, S. and Huberman, B.A. (2010) Predicting the Future with Social Media. 2010 *IEEE*/*WIC*/*ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, 31 August-3 September 2010, 492-499. https://doi.org/10.1109/WI-IAT.2010.63

[11] Shen, D. (2020) Movie Box Office Prediction via Joint Actor Representations and Social Media Sentiment. arXiv: 2006.13417.

[12] Qiu, X. and Tang, T.Y. (2018) Microblog Mood Predicts the Box Office Performance. *Proceedings of the* 2018 *Artificial Intelligence and Cloud Computing Conference*, Tokyo, 21-23 December 2018, 129-133. https://doi.org/10.1145/3299819.3299839

[13] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805.

[14] Radford, A., Narasimhan, K., Salimans, T., *et al*. (2018) Improving Language Understanding by Generative Pre-Training. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

[15] Radford, A., Wu, J., Child, R., *et al*. (2019) Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, **1**, 9.

[16] Brown, T., Mann, B., Ryder, N., *et al*. (2020) Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, **33**, 1877-1901.

[17] Jemni, S.K., Ammar, S., Souibgui, M.A., *et al*. (2023) ST-KeyS: Self-Supervised Transformer for Keyword Spotting in Historical Handwritten Documents. arXiv: 2303.03127.

[18] Wang, G., Yu, F., Li, J., *et al*. (2023) Exploiting the Textual Potential from Vision-Language Pre-Training for Text-Based Person Search. arXiv: 2303.04497.

[19] Guo, Y., Wang, P., Zhou, X., *et al*. (2022) An Improved Imaging Algorithm for HRWS Space-Borne SAR Data Processing Based on CVPRI. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **16**, 126-140. https://doi.org/10.1109/JSTARS.2022.3224194

[20] Guo, Y., Xiao, X., Wang, X., *et al*. (2023) A Two-Stage Real Image Deraining Method for GT-RAIN Challenge CVPR 2023 Workshop UG $^{\textbf {2}} $ + Track 3. arXiv: 2305.07979.

[21] Wang, L., Guo, H. and Liu, B. (2023) A Boosted Model Ensembling Approach to Ball Action Spotting in Videos: The Runner-Up Solution to CVPR'23 SoccerNet Challenge. arXiv: 2306.05772.

[22] Zhou, Y., Ringeval, F. and Portet, F. (2023) A Survey of Evaluation Methods of Generated Medical Textual Reports. *Proceedings of the* 5*th Clinical Natural Language Processing Workshop*, Toronto, July 2023, 447-459. https://doi.org/10.18653/v1/2023.clinicalnlp-1.48

[23] Abdelhalim, N., Abdelhalim, I. and Batista-Navarro, R.T. (2023) Training Models on Oversampled Data and a Novel Multi-class Annotation Scheme for Dementia Detection. *Proceedings of the* 5*th Clinical Natural Language Processing Workshop*, Toronto, July 2023, 118-124. https://doi.org/10.18653/v1/2023.clinicalnlp-1.15

[24] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 30. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[25] Liu, X., Duh, K., Liu, L., *et al*. (2020) Very Deep Transformers for Neural Machine Translation. arXiv: 2008.07772.

[26] Tang, W., Xu, B., Zhao, Y., *et al*. (2022) UniRel: Unified Representation and Interaction for Joint Relational Triple Extraction. *Proceedings of the* 2022 *Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, December 2022, 7087-7099. https://doi.org/10.18653/v1/2022.emnlp-main.477

[27] Ji, S., Pan, S., Cambria, E., *et al*. (2021) A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, **33**, 494-514. https://doi.org/10.1109/TNNLS.2021.3070843

[28] Cui, L. and Zhang, Y. (2019) Hierarchically-Refined Label Attention Network for Sequence Labeling. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, November 2019, 4115-4128. https://doi.org/10.18653/v1/D19-1422

[29] Erickson, N., Mueller, J., Shirkov, A., *et al*. (2020) Autogluon-Tabular: Robust and Accurate Automl for Structured Data. arXiv: 2003.06505.