基于混合特征的电影评分预测系统

黄东晋,耿晓云,李 娜,丁友东
(上海大学,上海 200072)

摘 要:电影评分是衡量一部电影优劣的重要标准,对于投资商和观影者极具参考价值,因此电影评分的预测成为电影领域的研究热点。然而目前的评分预测系统由于特征信息不足,特征工程处理方法过于简单,机器学习算法较为单一,所以预测误差偏大。针对这一问题,结合自然语言处理技术提出一种基于混合特征的预测模型,并应用到电影评分预测系统中。数据集来源是某常用电影网站,同时为了获取更好的训练数据,需要对电影特征信息进行复杂的特征工程处理。利用训练完成的 Bert 模型矢量化电影数据集中的文本信息得到文本矢量特征,并采用支持向量机(SVM)算法初步训练预测评分。将该评分作为一维新特征和电影特征信息一起通过随机森林(random forest)算法训练预测最终评分。实验结果表明,该预测模型是可行的,预测值与真实值的误差较小,准确性显著提升。

关键词:电影评分预测;机器学习;自然语言处理;文本矢量特征;Bert
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2020)12-0136-06
doi:10.3969/j.issn.1673-629X.2020.12.024

Film Rating Prediction System Based on Mixed Features

HUANG Dong-jin, GENG Xiao-yun, LI Na, DING You-dong
(Shanghai University, Shanghai 200072, China)

Abstract: Film rating is an important criterion for measuring the pros and cons of a film, which is of great reference value for investors and moviegoers. Therefore, the prediction of film rating has become a research hotspot in the film field. However, the current film rating prediction system has insufficient feature information, the feature engineering processing method is too simple, and the machine learning algorithm is relatively simple, so the prediction error is too large. Aiming at this problem, a prediction model based on mixed features is proposed in combination with natural language processing technology and applied to the film rating prediction system. The source of the dataset is a commonly used film website. At the same time, in order to obtain better training data, complex feature engineering processing of film feature information is required. The trained Bert is used to vectorize the text information in the film dataset to obtain the text vector features, and the support vector machine (SVM) algorithm is used to initially train and predict the text rating. The rating is used as a one-dimensional new feature along with film feature information to train and predict the final rating through the random forest algorithm. The experiment shows that the prediction model is feasible, the error between the predicted value and the real value is small, and the accuracy is significantly improved.

Key words: film rating prediction; machine learning; natural language processing; text vector features; Bert

0 引 言

现如今,观影已经逐渐成为人们日常生活中不可或缺的一种娱乐消遣方式,作为观众,希望每一次观影体验都是物超所值的,而电影评分很大程度上决定人们是否选择这部电影;对于投资商来说,准确地预测出评分可以有效减少利益损失。时至今日,国内外在电影方面的预测系统多数集中在票房的预测,电影评分预测系统很少且大多忽略电影文本信息对于评分的影响,往往只采用了一些常用的特征信息或者电影评论信息,并且特征工程处理方式不够完善,机器学习算法较单一,最终导致误差普遍较高。

2012 年,Andrei 等人[1]基于社交媒体数据,确定电影的定性和定量活动指标,通过提取两组表面和文本特征进行电影评分的预测,但评价电影的用户的人口统计数据可能与分享评论的人不同且数据处理过于复杂。2014 年,Rajitha 等人[2]提出一种基于视频中观

收稿日期:2020-02-18 **修回日期**:2020-06-19
基金项目:国家自然科学基金项目(61402278);上海市自然科学基金项目(19ZR1419100)
作者简介:黄东晋(1982-),男,博士,副教授,硕导,研究方向为虚拟现实、计算机图形学、人工智能等;耿晓云(1995-),女,硕士研究生,研究方向为机器学习、自然语言处理。

众观影时的表情和肢体动作的电影评分预测系统,但由于该实验对于观众的要求较高,自我报告实现比较困难。2017 年,Mustafa 等人[3]提出一种基于混合属性使用和集成学习的电影用户评分预测系统,能够较好地预测评分。同年刘明昌基于豆瓣电影数据构建了混合评分预测系统,有效提高了预测准确性。2018 年,黄幸颖等人[4]提出了一种克服了协同过滤算法中稀疏性影响的基于自编码网络的电影评分预测系统,但同时它带来的非凸函数的优化问题使得实验结果并不稳定。

针对这些问题,该文结合自然语言处理技术,提出一种基于混合特征的电影评分预测系统,巧妙综合了文本特征[5]和常用的电影特征的优势,实验结果表明,混合特征能够显著降低预测误差,使得该系统能够较为准确地预测电影评分。

1　系统框架

该文提出的基于混合特征的电影评分预测系统主要由文本矢量化、文本评分预测模型以及基于混合特征的评分预测模型这三部分组成。

具体流程如图 1 所示。

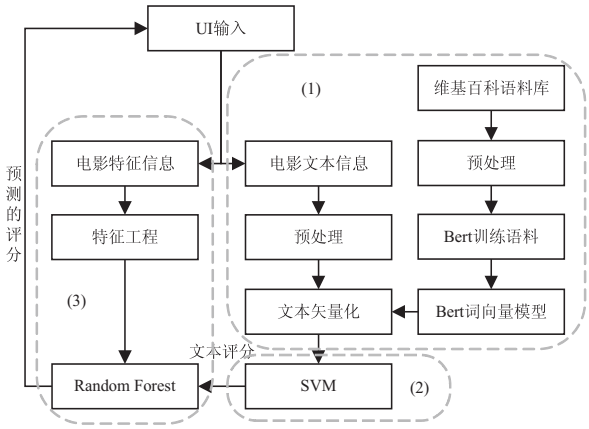

图 1　系统框架

(1)UI 输入导演、三位主演、上映时间、电影类型等电影特征信息和电影简介文本信息传入后端,文本信息经过预处理后由维基百科语料库预训练好的 Bert 模型进行矢量化,从而得到文本矢量特征;

(2)利用 SVM 算法对该矢量特征进行训练建模,预测文本评分;

(3)将文本评分作为一维新特征加入到预处理好的电影特征信息中,通过 Random Forest 算法预测最终的电影评分并返回到 UI 界面。

2　系统关键技术

2.1　文本矢量化

在自然语言处理中首先要考虑的就是词在计算机中的表示方法,对于文本信息,词嵌入要做的就是把单词嵌入到低维空间中用向量来表示,因此近义词的词向量距离理应较近。由于任意一个词都可以用它的相邻词来表示,所以一般情况下,可以通过统计学方法或者基于不同结构的神经网络的语言模型来生成词向量。

One-hot 编码使用参数 0 和 1 表示,词向量的维度等于词的总数量且仅一位有效,通常用来处理离散型数值特征,但是当特征较多时,矩阵会过于稀疏。而词袋模型则认为大量独立无序词汇的集合形成文本,其在文本里出现的次数作为这个词的向量,但仍然存在矩阵稀疏性,另外该方法丢失了上下文信息。2013 年谷歌推出一种引起业界轰动的分为 Skip-gram 和 CBOW 两种模型的词向量工具-Word2vec[6],其中前者是由中心词预测上下文,后者是由上下文预测中心词,通过层次 softmax 和负采样技术大幅度改进了词向量模型的性能。由于该模型引入了上下文,使得词向量带有语义信息,所以近义词的词向量具有相似性,但是该模型只考虑到了局部信息,而忽视了全局统计信息。为了解决这个问题,斯坦福 NLP 实验组于 2014 年推出全局词向量表达工具 Glove[7],该工具的本质是将全局矩阵分解和局部文本框捕捉两大技术进行结合,提高了很多 NLP 基础任务的准确率。然而它们都忽略了一个问题,即一词多义性,2018 年华盛顿大学提出的使用双向长短期记忆模型(bi-directional long short-term memory,Bi-LSTM)[8]的基于语言模型的词向量(embedding from language models,ELMO)工具[9]就是针对这一问题进行了优化,但 LSTM 序列模型有两大缺陷,一是无法双向考虑上下文信息,二是并行计算能力差。所以 2018 年 Google 推出了采用 Transformer 编码器的 Bert[10]词向量模型,通过与自注意力机制相结合,真正实现了双向编码。

文中的电影文本信息首先要经过预处理,包括过滤特殊字符、去停用词等工作,然后使用预训练好的 Bert 词向量模型对文本进行向量化。

2.1.1　Bert 词向量模型

真正实现双向编码的 Bert 模型能够有效联系上下文,显著提高泛化能力。Bert 模型将 Transformer 编码器和注意力机制结合起来,比 RNN 的效率更高,同时对于长文本的效果更好[10]。

其中 Transformer 模型是由 1 个编码器组和 1 个解码器组构成,而它们又分别由 6 个编码器和 6 个解码器组成。如图 2 所示,每个编码器包括一个前馈神经网络和一个帮助编码器在编码单词的过程中理解输入序列中的其他单词的自注意力机制,而每个解码器在编码器的基础上增加了一层用来帮助当前节点获取

当前需要关注的重点内容的编码-解码注意力层。

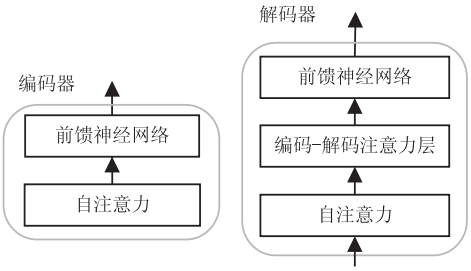

图 2 编码-解码器

Transformer 模型的具体流程是:首先将通过 Embedding 处理后的输入数据送到编码器中,然后相继由自注意力机制和 Feed Forward 进行处理,输出到下一个编码器,最后将解码器的输出通过一个全连接层和一个 softmax 进行映射,选取最大概率的词。

2.1.2 文本矢量化

Bert 的使用分为预训练和微调,其中预训练包括 MLM 和 NSP[10],前者是指随机选取 15% 的词,其中 80% 的概率采用 mask 标记,10% 的概率采用随机词替代,剩下 10% 的概率不做替换,然后利用上下文来预测这些词。后者是指判断输入 Bert 的两个文本的连续性,相当于二分类任务。

Bert 词向量模型以字为最小单位,不需要对文本进行分词,输入由三个嵌入特征构成。Bert 模型在预处理好的维基百科语料上进行预训练,获得 Bert 词向量模型,然后对预处理过的电影文本信息进行矢量化,输出 768 维文本向量。

2.2 回归算法

当前的回归预测模型主要是机器学习算法和深度学习算法的应用,常用的有线性回归、决策树回归、SVM、Random Forest、xgboost、LSTM、CNN 等。

2011 年 Dong Nguyen 等人[11]提出基于线性回归算法的文本作者年龄预测模型,实验表明话语模式与个人识别之间的相关性高达 0.74,平均绝对误差在 4.1 至 6.8 之间。2015 年蔡慧苹等人[12]提出基于 Word embedding 和 CNN 的情感分类模型,其准确率比传统 ML 高了约 5 个百分点。2017 年 Ashok 等人[13]提出一种利用支持向量回归算法在线模式开发基于机器视觉的铁矿石等级预测模型,实验表明测试样品的观察值和预测值之间的相关系数为 0.824 4,该模型对于铁矿石等级的预测性能较好。同年胡西祥针对微博评论构建基于 DL 的情感分类模型,其准确率为 84.5%。随后 Torlay 等人[14]提出基于 xgboost 算法的癫痫患者分析及分类,AUC 指标为 96%。同时王斌构建了基于 LSTM 的交通流量预测系统,实验表明平均精度为 95%。对比实验表明,SVM 和 Random Forest 在该电影数据集上的表现最优,所以下面就这两种算法做一些简要的介绍。

2.2.1 Random Forest

Random Forest[15]是一种由多个弱分类器对数据进行训练并预测的集成算法,一个样本数据有多个分类输出结果,而最终的类别由投票机制确定。

该算法流程如图 3 所示。对于每个分类器,首先采用有放回机制在所有的数据样本中随机选取部分样本,然后从这些样本的特征中再随机选取部分特征,并挑选出最好的特征。同时每棵决策树都无剪枝的尽可能的生长直至输出一个分类结果,通过多数为胜的投票机制确定最终输出类别。

Random Forest 是一种实用性很强的算法,在目前所有的算法中具有较好的准确率,而且在大数据集和高维特征上都有很好的表现。

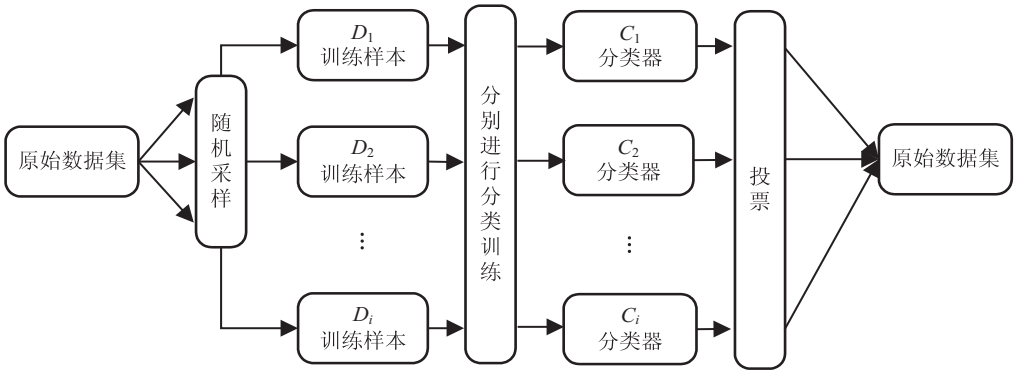

图 3 Random Forest 结构

2.2.2 SVM

SVM[16]是一种适用于处理分类和回归问题的基于统计学的有监督的二分类器,主要问题是如何在特征空间中使数据点与超平面的距离最大化。该算法在训练过程中首先利用拉格朗日乘子法[17]与对偶学习法来处理最优化问题,然后由序列最小优化(sequential minimal optimization,SMO)来求解。

其中线性可分 SVM 适用于严格线性可分的数据集,假设超平面为:

$$y = w^{\mathrm{T}}x + b \tag{1}$$

则数据点 (x_i, y_i) 与超平面的距离为:

$$\hat{r}_i = |w^{\mathrm{T}}x_i + b| = y_i(w^{\mathrm{T}}x_i + b) \tag{2}$$

其几何距离为:

$$r = \min_i r_i = \min_i \frac{\hat{r}_i}{\|w\|} \tag{3}$$

距离最大化为 $\max_{w,b} r, \text{s.t. } r_i \geqslant r, \text{for } i = 1,2,\cdots,n$,令 $\hat{r} = 1$,最大化 $\frac{1}{\|w\|}$ 即最小化 $\frac{1}{2}\|w\|^2$,由拉格朗日乘子法求解最优解 (w^*, b^*)。已知拉格朗日函数为:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(w^{\mathrm{T}}x_i + b) - 1]$$
$$\text{s.t. } \alpha_i \geqslant 0, \text{for } i = 1,2,\cdots,n \tag{4}$$

令 $\theta = \max_{\alpha,\alpha\geqslant 0} L(w,b,\alpha)$,由 $\theta = \frac{1}{2}\|w\|^2$,即:

$$\min_{w,b}\theta = \min_{w,b}\max_{\alpha,\alpha\geqslant 0} L(w,b,\alpha) =$$
$$\min_{w,b}\max_{\alpha,\alpha\geqslant 0}\{\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(w^{\mathrm{T}}x_i + b) - 1]\}$$
$$= p^* \tag{5}$$

则其对偶问题为:

$$\max_{\alpha,\alpha\geqslant 0}\min_{w,b} L(w,b,\alpha) =$$
$$\max_{\alpha,\alpha\geqslant 0}\min_{w,b}\{\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[y_i(w^{\mathrm{T}}x_i + b) - 1]\}$$
$$= Q^* \tag{6}$$

且 $Q^* \leqslant P^*$,根据 Slater 条件可知存在 x_i 使得 $Q^* = P^*$,对 w 和 b 求偏导后通过 SMO 算法来求解。

线性 SVM 适用线性不可分数据集,增加了松弛变量,其超平面和决策函数与前面的相同。而非线性 SVM 引入了核函数,通过非线性映射使该数据线性可分,然后按照线性 SVM 的方法求解。

3　实验与结果分析

本次实验的硬件配置是基于 macOS High Sierra 系统,CPU 型号为 3.5GHz 6-Core Intel Xeon E5,内存为 16G;软件配置:编程工具为 Pycharm2018.3.2,基于 Python3.7 编程语言和 Tensorflow1.13.1 框架,此外还使用了 gensim 库、jieba 分词库、pandas 库和 scikit_learn 库等。

如图 4 所示,系统 UI 包括输入和输出两部分,其中输入是指用户输入电影名、电影特征信息以及电影简介文本信息,输出是指前端输入的信息传输到后端,后端进行处理后,将预测出的评分返回到 UI 界面。

该文设计了两组实验,分别是文本评分预测模型实验和基于混合特征的评分预测模型实验,实验性能指标采用均方根误差 RMSE。

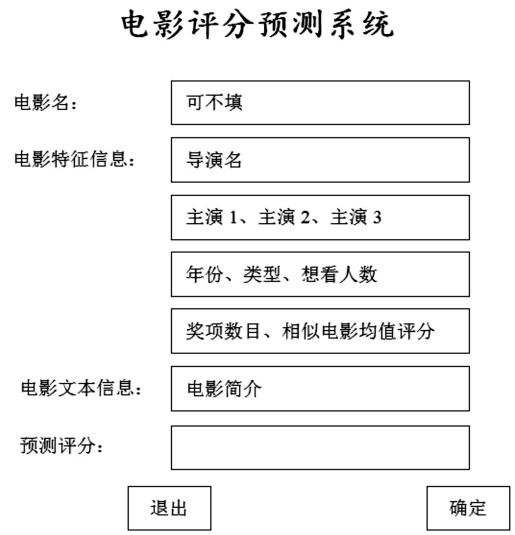

图 4　系统 UI 界面

3.1　文本评分预测模型实验

第一组实验是文本评分预测模型实验,使用的数据是豆瓣爬取的 12 491 条电影的文本信息,即电影简介,其中以 0.8 和 0.2 的比例切分训练测试样本。

该实验首先使用预训练好的 Bert 和 Word2vec 模型分别矢量化输入数据得到文本矢量特征,然后基于 SVM 算法建模,调参后使用最佳参数训练测试。同时为了验证该模型性能的优劣性,使用了另外十种算法进行对比实验,实验结果如图 5 所示,横坐标为十一种算法,纵坐标为 RMSE 值。

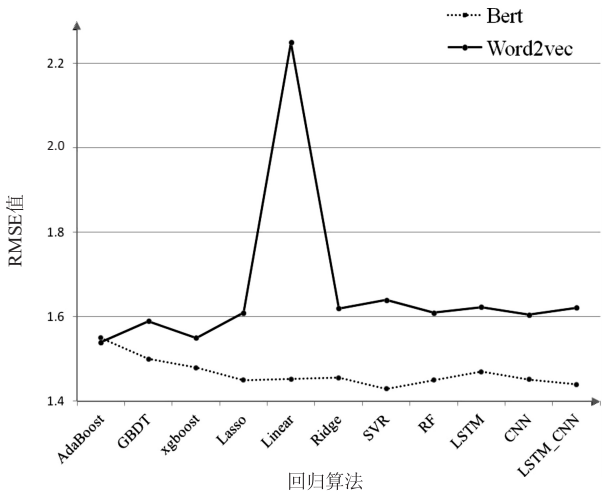

图 5　Bert 和 Word2vec 效果对比

实验结果表明,在该数据集上 Bert 词向量效果要优于 Word2vec,且 SVM 算法的表现最好,RMSE 为 1.43。为了更加直观地观察模型性能,在测试集中随机选取 50 条数据,评分预测值与实际值的比较结果如图 6 所示,容易看出基本走势大致相同,但整体误差较大,说明仅仅依靠文本特征无法很好地预测评分。

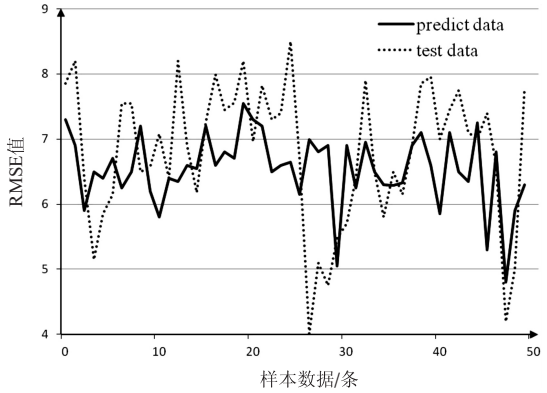

图 6 基于 SVM 算法的评分预测值与真值的结果对比

3.2 基于混合特征的评分预测模型实验

第二组实验是基于混合特征的评分预测模型实验,基于上一组实验中的 2 500 条测试样本,依据是否含有文本特征这一条件设计一组对比实验。样本中的主要信息包括电影 ID、名称、上映年份、类型、导演、演员和豆瓣评分等,同样以0.8和0.2的比例切分训练测试样本。

该实验首先根据是否包含电影文本信息将数据集分为两组,在将数据标准化后,分别基于 Random Forest 算法构建模型,通过网格搜索调参获得最佳模型参数,然后随机划分数据集进行 100 次训练预测实验,计算 RMSE 的平均值,同时为了验证模型性能的优劣性,分别使用了另外十种算法进行对比实验,两组实验结果对比如表 1 所示。

实验结果表明,电影特征信息和文本信息的特征混合能够显著提升模型性能,且在该数据集上表现最优的算法为 Random Forest,其 RMSE 为 0.564 3,在测试集中随机选取 50 条数据,评分预测值与实际值的比较结果如图 7 所示。最后为了评估该系统的用户体验度,邀请了 30 位同学来体验,为该系统打分,采用 10 分制,问卷统计结果如表 2 所示,可以看出该系统操作性、流畅性以及实用性很好,但是 UI 设计和耗时性有待改进。

表 1 对比结果

算法	基于电影特征信息的评分预测模型	排名(1)	基于混合特征的评分预测模型	排名(2)
Adaboost	0.607 7	6	0.605 5	9
GradientBoosting	0.620 4	9	0.576 7	3
KNN	0.688 3	11	0.682 4	11
Lasso	0.578 0	2	0.577 0	6
MLP	0.615 4	8	0.626 1	10
RandomForest	0.574 3	1	0.564 3	1
Ridge	0.578 1	3	0.576 8	4
SVR	0.622 6	10	0.575 0	2
XGBoost	0.610 2	7	0.580 9	7
Linear	0.578 2	4	0.576 9	5
LSTM	0.579 8	5	0.582 4	8

表 2 问卷统计结果

UI 设计	操作性	系统流畅性	耗时性	实用性
6.266 7	7.666 7	7.933 3	6.2	7.8

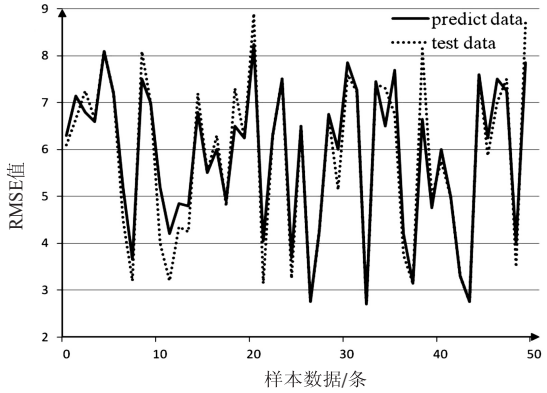

图 7 基于 Random Forest 算法的评分预测值与实际值的结果对比

4 结束语

利用自然语言处理技术与机器学习算法的优势,提出了基于混合特征的电影评分预测系统,通过对比实验可看出 Bert 具有更好的词向量效果且混合特征可显著提升模型性能,另外实验结果表明机器学习算法中 SVM 和 Random Forest 算法在该电影数据集上的表现最好,系统预测准确率较高。当然该系统还存在不足之处,比如实时性不高且 UI 设计不够完美,后期可以针对 Bert 模型进行改进,加快文本的向量化过程,另外在 UI 设计方面,可以在得到评分预测结果的同时背景中加入对应电影海报的展示,使得系统界面

更加智能美观。

参考文献:

[1] OGHINA A, BREUSS M, TSAGKIAS M, et al. Predicting IMDB movie ratings using social media[C]//The European conference on information retrieval. Barcelona, Spain: Springer, 2012: 503-507.

[2] NAVARATHNA R, LUCEY P, CARR P, et al. Predicting movie ratings from audience behaviors[C]//IEEE winter conference on applications of computer vision. Steamboat Springs, CO, USA: IEEE, 2014: 1058-1065.

[3] EREN A O, SERT M. Movie rating prediction using ensemble learning and mixed type attributes[C]//Signal processing and communications applications. Antalya, Turkey: IEEE, 2017: 1-4.

[4] 黄幸颖,梁　路,滕少华. 电影评分的自编码网络预测研究[J]. 小型微型计算机系统,2018,39(9):2035-2038.

[5] 黄东晋,纪　浩,耿晓云,等. 基于文本矢量特征的电影评分预测模型[J]. 现代电影技术,2019(3):44-50.

[6] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//Proceedings of the international conference on learning representations. Scottsdale, Arizona, USA: ICLR, 2013: 1-12.

[7] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the conference on empirical methods in natural language processing. Stroudsburg, PA, USA: ACL, 2014: 1532-1543.

[8] 李　洋,董红斌. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析[J]. 计算机应用,2018,38(11):3075-3080.

[9] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies. Stroudsburg, PA, USA: ACL, 2018: 2227-2237.

[10] 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. Minneapolis, USA: NAACL, 2019: 4171-4186.

[11] NGUYEN D, SMITH N A. Author age prediction from text using linear regression[C]//Proceedings of the association for computational linguistics. Stroudsburg, PA, USA: ACL, 2011: 115-123.

[12] 蔡慧苹,王丽丹,段书凯. 基于 word embedding 和 CNN 的情感分类模型[J]. 计算机应用研究,2016,33(10):2902-2905.

[13] PATEL A K, CHATTERJEE S. Development of online machine vision system using support vector regression (SVR) algorithm for grade prediction of iron ores[C]//Fifteenth IAPR international conference on machine vision applications. Tokyo, Japan: IEEE, 2017: 149-152.

[14] TORLAY L, PERRONE-BERTOLOTTI M, THOMAS E. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy[J]. Brain Informatics, 2017, 4(3): 159-169.

[15] HO T K. Random decision forests[C]//Proceedings of 3rd international conference on document analysis and recognition. Montreal, Quebec, Canada: IEEE, 1995: 278-282.

[16] SMOLA A J, SCHÖLKOPF B. A tutorial on support vector regression[J]. Statistics and Computing, 2004, 14(3): 199-222.

[17] FRIEDMAN J, HASTIE T, TIBSHIRANI R. The elements of statistical learning[M]. New York: Springer, 2001: 417-438.