朱衍熹, 张明西, 赵瑞, 等. 基于深度神经网络的影视评分预测[J]. 智能计算机与应用,2024,14(6):79-87. DOI:10.20169/j.issn.2095-2163.240611

# 基于深度神经网络的影视评分预测

朱衍熹, 张明西, 赵 瑞, 许星波
(上海理工大学 出版印刷与艺术设计学院, 上海 200093)

**摘 要**: 影视评分能直接反映影视作品的上映效果或收益情况,然而目前影视特征的提取方法单一,信息挖掘不充分。针对这一问题,提出一种基于混合特征表示向量的深度神经网络影视评分预测模型。根据影视作品的属性特征通过词袋模型、特征拆分、TF-IDF 文本矢量化方法生成影视混合特征表示向量,并构建基于深度神经网络的影视评分预测模型。实验结果表明:测试集 *MAE*、*MSE*、*SmoothL*1 *Loss* 指标值在模型 100 轮迭代训练后收敛,*MAE* 为 0.82,*MSE* 为 1.07,*SmoothL*1*Loss* 为 0.45, 验证了所提方法对影视作品的评分预测有很好效果,能有效评估影视作品上映后的价值。

**关键词**: 深度神经网络; 词袋模型; TF-IDF; 影视评分
**中图分类号**: TP391 **文献标志码**: A **文章编号**: 2095-2163(2024)06-0079-09

## Prediction of film ratings based on Deep Neural Network

ZHU Yanxi, ZHANG Mingxi, ZHAO Rui, XU Xingbo
(College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China)

**Abstract**: Film ratings can directly reflect the screening effect or income of film. However, the current method of extracting film features is single and the information mining is insufficient. To solve this problem, a film ratings prediction model based on Deep Neural Network and mixed feature representation vector is proposed. According to the attributes of the film, the representation vectors of film mixed features are generated by bag of words model, feature splitting and TF-IDF, and the film ratings prediction model based on Deep Neural Network is constructed by using vectors. The experimental results show that the index values of *MAE*, *MSE and SmoothL*1 *Loss* convergence after 100 iterations of training, *MAE* is 0.82, *MSE* is 1.07, *SmoothL*1 *Loss* is 0.45, which proves that the proposed method has a good effect on the ratings prediction of the film, and can effectively evaluate the value of the film after release.
**Key words**: DNN; bag of words model; TF-IDF; film rating

## 0 引 言

随着人工智能和深度学习技术的快速发展,文化产品的价值评估引起了人们的极大关注。影视作品是重要的文化产品,在文化领域受众广泛,市场价值占比高。影视评分是衡量影片质量的重要指标,构建模型预测影视评分,能够为进一步发展影视推荐系统建立基础,同时可根据影片的属性特征例如类型、概述等为用户推荐评分值较高的相关影视作品集。在影视行业,对影片的评分预测研究能拓宽对其价值评估的渠道,反映收益情况。运用深度神经网络的方法对影视评分进行预测能够为更宏观概念的文化产品的产权价值评估提供一定的思路。

深度神经网络模型已广泛应用于图片分类、语音识别、气象预测等多个领域,在分类或回归任务中取得了很好的效果。深度神经网络能够对输入特征进行全局特征的识别提取,能够抽取数据潜在的信息,挖掘属性特征与影视作品评分的关系。在影视作品的自身属性中,存在多种不同类型的数据,需要进行不同的矢量化表示,例如,短文本数据仅需要通过词汇空间进行矢量化,而要充分挖掘长文本向量信息,则需要进行词频、权重的考虑,因此对不同特征的信息提取存在一定难度。

为解决影视作品数据类型复杂的问题,提出一

**基金项目**: 国家重点研发计划项目(2021YFF0900400); 国家自然科学基金(62002225); 上海市自然科学基金(21ZR1445400)。
**作者简介**: 朱衍熹(2000-),男,硕士研究生,主要研究方向:数据挖掘; 赵 瑞(1998-),男,硕士研究生,主要研究方向:数据挖掘; 许星波(1999-),女,硕士研究生,主要研究方向:数据挖掘。
**通讯作者**: 张明西(1985-),男,博士,副教授,硕士生导师,CCF 会员(50460M),主要研究方向:数据挖掘,社会网络分析,智能媒体技术等。Email:mingxizhang10@ fudan.edu.cn
**收稿日期**: 2023-08-14

种上游任务以特征工程、词袋模型、TF-IDF 矢量化方式与下游任务深度神经网络结合的预测方法,对不同的影视作品属性数据进行不同的建模,输入到预测模型当中,充分提取序列信息,提高预测精度。

## 1 相关工作

影视作品的评分或票房预测研究得到广泛关注[1],现有预测方法主要分为 2 类。一类是传统机器学习的预测方法,何琦等学者[2]依托大数据展开实证研究,运用机器学习与模型融合方法构造有更高拟合性与精度的票房预测模型。李香君等学者[3]研究了支持向量机(SVM)回归预测对电影评分预测,取得较好的效果。李旺泽[4]提出 Lasso-XGBoost 组合预测模型,相比传统机器学习提高了准确率。李振兴[5]建立了决策树模型、朴素贝叶斯模型和随机森林模型,对电影票房进行了分类预测。张红丽等学者[6]使用逐步回归方法筛选出变量构建评分预测模型。任丹[7]提出多元线性回归的票房预测模型,并基于 SSH 框架,采用 MVC 开发模式实现了可对电影票房预测的系统。为了提高电影评分的预测精度, Zhang[8]建立改进鲸鱼算法对 IMDB 电影评分进行预测。You 等学者[9]提出了指数形式的多元非线性回归电影评分预测模型,并揭示了评分与相关变量之间的关系。Zahabiya 等学者[10]利用随机森林和 XGBoost 算法对电影评分预测。Sathiya 等学者[11]基于特征工程的方法建立预测评分模型。Soojin 等学者基于推荐系统技术的预测系统,使用协同过滤和模糊系统来解决协同过滤问题,预测用户对电影的评分。另一类是基于神经网络的预测方法,魏明强等学者[12]运用神经网络模型根据网络评价得分探讨在不同阶段网络评分变化对于票房走势的影响。Mohammed 等学者[13]利用反向传播和 Delta 规则技术训练的特征神经网络建立电影评分模型。Su 等学者[14]设计了一系列电影评分指标,提出了一种基于神经网络算法的电影评分预测模型。

科研人员运用了传统的机器学习模型和神经网络等方法对于评分或票房进行预测,能够达到一定的预测效果。然而,现有的方法对特征的提取手段单一,对属性信息挖掘不充分,精度提升有限,为解决该问题,本文提出一种基于混合特征表示向量的深度神经网络影视评分预测模型。

## 2 模型流程的框架

影视评分预测模型的构建包括 2 个阶段,分别是:图 1 中,混合特征表示向量构建阶段和模型应用阶段,如图 1 所示。图 1 中,混合特征表示向量构建阶段为建模上游任务,主要目标是构建影视属性信息的表示向量,步骤包括影视数据预处理、空值异常值处理、文本属性清洗、分词等;特征表示,对 json 格式属性字段拆分,按类别构建双值表示向量;词袋模型,对文本信息较少的属性忽略语法及词元出现的顺序构建短文本表示向量;TF-IDF 模型,对文本内容较多的属性考虑词元对于样本的重要性程度,结合词频构建长文本表示向量;构建表示向量,合并双值表示向量、短文本表示向量、长文本表示向量组成影视混合特征表示向量并进行归一化处理。

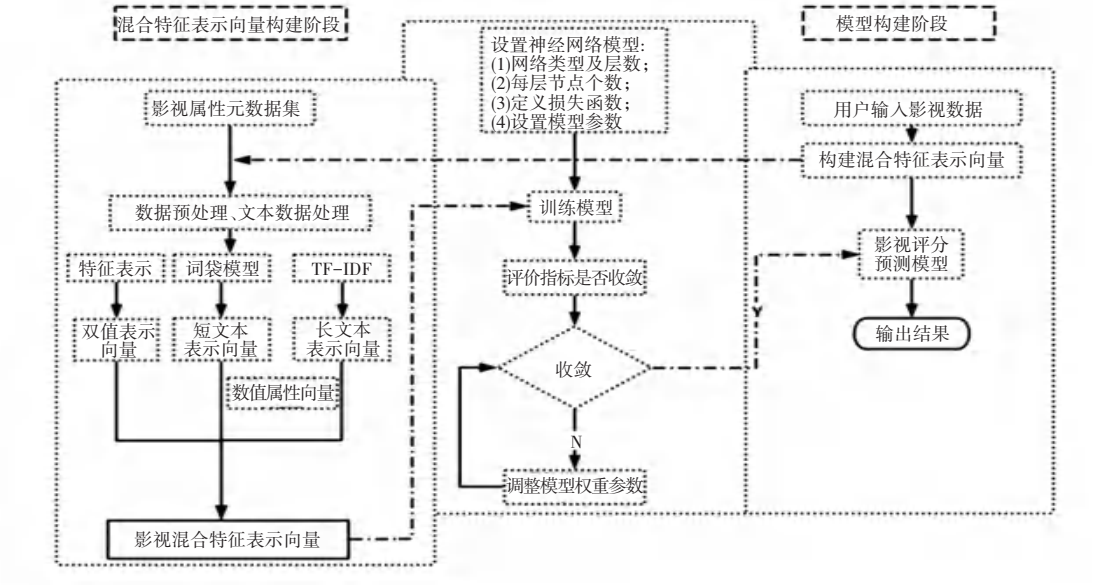


**图 1 模型框架**

**Fig. 1 The framework of the system**

模型应用阶段，构建基于深度神经网络的影视作品评分模型，根据用户输入的影视属性预测作品最终的评分情况。具体的步骤为：输入影视混合特征表示向量对神经网络模型进行训练，调整模型权重参数并保存，此后可得到对新影视作品属性数据的评分结果输出。

构建基于混合表示特征的深度神经网络影视作品评分预测模型利用影视自身属性信息进行评分预测，能使影视领域从业者根据预测结果从多角度对影视作品上映效果进行分析，规避风险。

## 3　影视混合特征表示向量

### 3.1　双值表示向量

影视双值表示向量构建采用特征拆分的方式。特征拆分是特征工程中的一种方式，数据的特征表示建模会影响模型最终的预测效果，流程如图 2 所示。为了能够充分挖掘数据信息，将部分原始数据进行整合转换，对影视属性中 json 格式的字段影视类型、影视语言、出版公司、出版国家通过特征拆分的方法进行处理并转化成多个维度的布尔值，进而构建双值表示向量，并根据下游任务的测试和评估效果反馈，调整双值表示向量结构，使得影视作品元数据中的特征更具体化、表示效果更好。

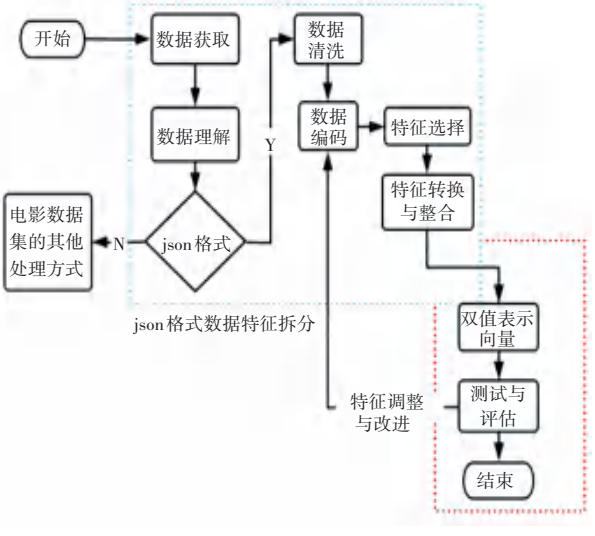


**图 2　特征拆分流程**

**Fig. 2　Splitting process of features**

下面以影视类型字段处理为例，主要步骤如下：

**步骤 1**　统计数据集中每种类型的影片数量，得到影视类型数量降序序列 $[genre_1, genre_2, \cdots, genre_c]$；

**步骤 2**　设定选择影视类型拆分数量 $n(n < c)$ 并剔除剩余类型（从高到低），得到新序列 $[genre_1, genre_2, \cdots, genre_n]$；

**步骤 3**　每个数据样本新增 $n$ 个字段，将影视作品对应类型字段设置为 1，非对应的类型字段设置为 0；每一样本新增字段 $[genre\_feature_1, \cdots, genre\_feature_n]_{1\times n}$，其中 $genre\_feature_{1\sim n}$ 为离散值 0 或 1。

影视语言、影视出版公司、影视出版国家均按照同样步骤处理。通过在原始影视作品数据的基础上将 json 格式字段进行扩展，可使得影片特征进一步细化，较全面地凸显出各数据样本的属性特点，表达能力更强。

### 3.2　短文本表示向量

影视短文本表示向量利用词袋模型进行构建。词袋模型是自然语言处理中对文本建模常用的表示方法，不考虑文本中词与词之间的上下文关系，忽略语法及词元出现的顺序，将文本语句仅看作是若干个词汇的集合，并只考虑所有词的权重[15]，使用一组无序的单词来表示文本，其核心思想是将文本语句转换成机器可识别的向量。影视数据字段中的影视名字、发行状态、发行日期、原始语言等文本型数据单样本内容较少，考虑词之间的联系或者顺序的意义不大，针对此类属性使用词袋模型转换成短文本表示向量，主要步骤如下：

**步骤 1**　短文本分词、去停用词，建立包含所有词的特征空间 $[word_1, word_2, \cdots, word_L]$，其中 $L$ 为词空间的长度大小；

**步骤 2**　利用特征词空间建立词袋模型；

**步骤 3**　每个样本对应的文本型数据转换为向量 $[S_1, \cdots, S_L]$，向量维度为词空间的长度大小 $L$。

短文本表示向量仅考虑词汇的权重情况，能充分提取上下文关系较弱文本的字段特征，向量转化关系如图 3 所示。

### 3.3　长文本表示向量

影视属性数据中存在长文本字段，例如影视作品标语、影视作品概述等，将每一个长文本字段作为一个文档，利用 TF-IDF 构建长文本表示向量。TF-IDF 是一种统计方法，用以评估一个字词对于一个文件集或一个语料库中的一份文件的重要程度。样本字词的重要性与该字词在文件中出现的次数成正比，但同时会随着该字词在语料库中出现的频率成反比下降。主要的思想如下：某个单词在一篇文章中出现的频率 $TF$ 高，并且在其他文章中很少出现，则认为此词或短语具有很好的类别区分能力。

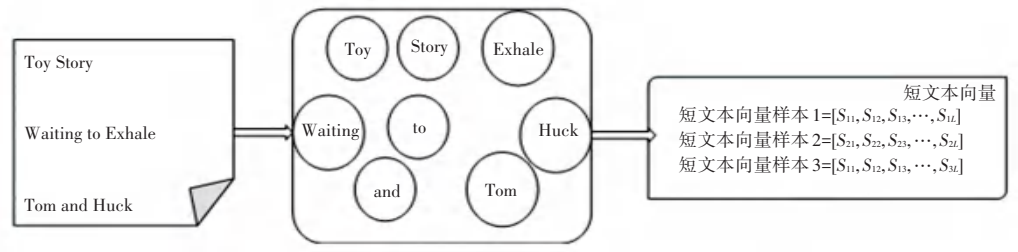


**图 3 短文本向量**

**Fig. 3 Short text representation vector**

*TF* 的计算公式如下[15]：

$$TF(t,d)=\frac{\text{出现在文档 } d \text{ 中的次数}}{\text{文档 } d \text{ 中所有词总数}} \tag{1}$$

其中，$t$ 表示词，$d$ 表示文档。

*IDF* 是逆文档频率，*IDF* 表示一个词在文档集中出现的次数，即在文档集中的多少个文档中出现。*IDF* 值越小，即在很少的文档中出现，那么这个词就有越强的文档区分能力。*IDF* 的计算公式如下：

$$IDF(t,D)=\log\frac{|D|}{|\{d\in D: t\in d\}|} \tag{2}$$

其中，$|D|$ 表示文档集中的文档总数；$|\{d\in D: t\in d\}|$ 表示文档集中出现词 $t$ 的文档数。

TF-IDF 算法是 TF 算法和 IDF 算法的综合使用。其计算公式如下：

$$TFIDF(t,d,D)=TF(t,d)\times IDF(t,D) \tag{3}$$

TF-IDF 在影视作品长文本字段的转化步骤如下：

**步骤 1** 长文本分词、去停用词，建立特征的词空间[$word_1$,$word_2$,…,$word_L$]，其中，$L$ 为词空间的长度；

**步骤 2** 计算每一个文档中词汇的 *TF*、*DF*、*IDF*、*TF* - *IDF*；

**步骤 3** 将每个样本对应的长文本文本型数据根据 *TF* - *IDF* 计算结果转换为向量[$L_1$,$L_2$,…,$L_L$]，维度为词空间的长度 $L$。

影视作品长文本数据的向量由文档中每个词汇的 *TF* - *IDF* 值构成，能够考虑词汇的重要性情况，长文本向量转化过程如图 4 所示。

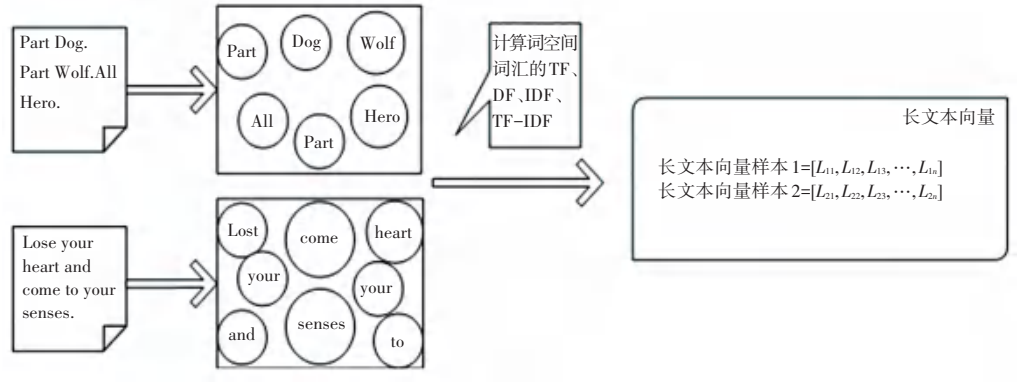


**图 4 长文本向量**

**Fig. 4 Long text vector transformation**

### 3.4 影视混合特征表示向量

影视混合特征表示向量由双值表示向量、短文本表示向量、长文本表示向量及数值数据组合生成。表示向量有着不同的量纲和单位，会影响数据矢量化后的分析结果，将整个训练样本向量矩阵进行标准化归一化处理，以消除数据在数值指标的差异性和指标之间的量纲影响，将所有值映射到 0~1 范围之间。对此给出阐释分述如下。

(1)标准化(Standardization)。数学公式为：

$$\frac{x-\mu}{\sigma} \tag{4}$$

其中，$x$ 表示原始数据；$\mu$ 表示数据均值，$\sigma$ 表示标准差。

(2)归一化(Normalization)。数学公式为：

$$\frac{x-\min(x)}{\max(x)-\min(x)} \tag{5}$$

## 4　影视评分预测模型

影视评分预测模型采用有监督学习的训练方式，基于深度神经网络对训练集中已标记好的评分数据进行迭代学习，有着较强的表达能力和拟合能力。深度神经网络包括输入层、隐藏层和输出层，其中输入层接收影视作品混合特征表示向量，数据经隐藏层计算流向输出层，输出层输出影视作品评分值，其层次结构体系是一个有向无环图。

### 4.1　深度神经网络

模型初始化相关参数，输入上游任务构建的影视混合特征表示向量，迭代训练，计算梯度并反向传播更新权重，当满足收敛条件或者达到所需效果后停止训练并输出结果。每层神经网络的节点数按照从少到多的原则设计，输入的特征维度先降后升，网络风格窄而深。数据流从每一层网络（除输出层）中输出后均经过 *ReLU* 激活函数计算，增加网络模型的非线性效果。为减少过拟合情况，在模型中加入 2 层 Dropout 层，比例权重参数设定为 10%，即去掉其中 2 层神经网络中 10%的参数。神经网络大致结构如图 5 所示。

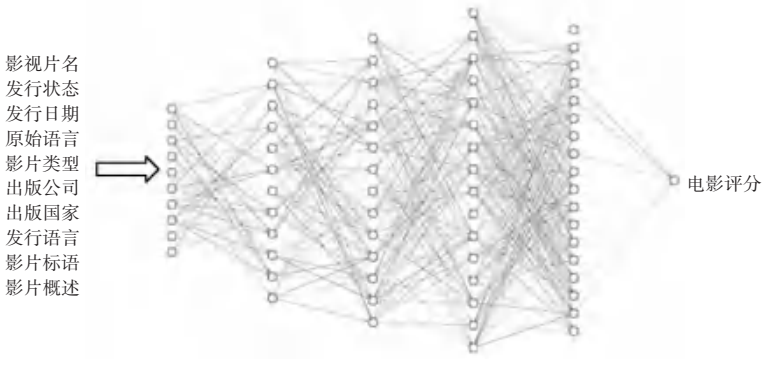


**图 5　神经网络结构**

**Fig. 5　Neural network structure**

深度神经网络能够很好地提取输入属性的全局特征，下面将阐述网络的前向传播过程。基于深度神经网络的影视评分预测模型共 6 个隐藏层，权重矩阵定义如下：

$$\begin{bmatrix} W_{11}^k & W_{21}^k & \cdots & W_{i1}^k \\ W_{12}^k & W_{22}^k & \cdots & W_{i2}^k \\ W_{13}^k & W_{23}^k & \cdots & W_{i3}^k \\ \vdots & \vdots & \ddots & \vdots \\ W_{1j}^k & W_{2j}^k & \cdots & W_{ij}^k \end{bmatrix} \tag{6}$$

权重矩阵为 $i \times j$ 形状的矩阵，其中 $W_{ij}^k$ 上标 $k$ 表示第 $k$ 层神经网络，下标 $i$ 表示当前神经元节点，下标 $j$ 是当前神经元节点与下一层所连接的神经元节点，$ij$ 表示神经元 $i$ 与神经元 $j$ 连接。模型的前向传播计算公式如下所示：

$$\begin{cases} a_1^k = \sigma(W_{11}^k a_1^{k-1} + W_{21}^k a_2^{k-1} + \ldots + W_{i1}^k a_i^{k-1} + b_1^k) \\ a_2^k = \sigma(W_{12}^k a_1^{k-1} + W_{22}^k a_2^{k-1} + \ldots + W_{i2}^k a_i^{k-1} + b_2^k) \\ a_3^k = \sigma(W_{13}^k a_1^{k-1} + W_{23}^k a_2^{k-1} + \ldots + W_{i3}^k a_i^{k-1} + b_3^k) \\ \vdots \\ a_j^k = \sigma(W_{1j}^k a_1^{k-1} + W_{2j}^k a_2^{k-1} + \ldots + W_{ij}^k a_i^{k-1} + b_j^k) \end{cases} \tag{7}$$

其中，$a_j^k$ 定义隐藏层输出节点；$k$ 表示第 $k$ 层神经网络；$j$ 表示该层神经元编号；$\sigma$ 表示 *ReLU* 激活函数；$b$ 表示偏置项。

影视混合特征表示向量将经过输入层、隐藏层后在输出层返回结果，输出层仅有一个神经元，即输出影视作品的评分预测值。

### 4.2　评分预测模型

基于深度神经网络的影视作品评分预测模型的结构如图 6 所示。由图 6 可知，模型由输入层、隐藏层、输出层组成，其中隐藏层由 6 层线性网络构成和 2 层 Dropout 组成。影视混合特征表示向量依次输入到各层网络，并且均经过 *ReLU* 激活函数，采用 Adam 优化器进行模型的优化，迭代训练模型得到最优的影视作品评分预测模型并输出结果。

#### 4.2.1　损失函数和 Adam 优化器

实验采取均方损失函数 *MSE Loss*，影视作品评分预测模型在训练迭代时通过计算评分预测值和真实值的损失函数进行反向传播，并对模型中的权重参数进行更新，数学定义公式如下：

$$MSE\ Loss(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{8}$$

其中，$y_i$ 表示真实值；$\hat{y}_i$ 表示预测值；$N$ 表示样本数量。

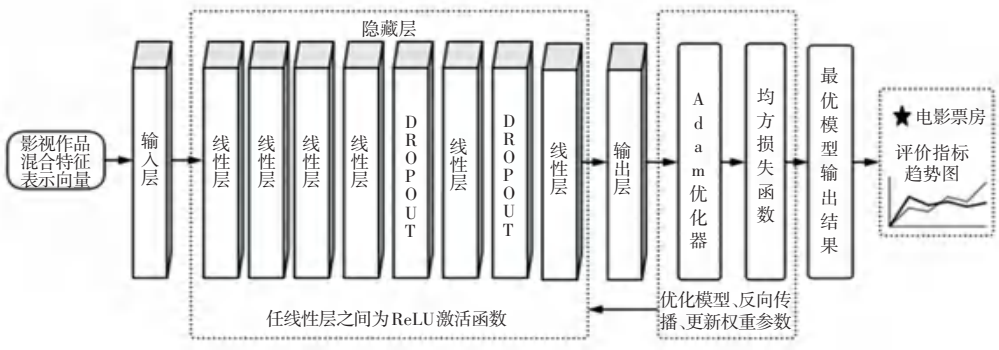


**图 6 影视作品评分预测模型**

**Fig. 6 Prediction model of film and television scores**

影视评分预测模型选用 Adaptive Moment Estimation 优化器,一种对随机梯度下降法的扩展,是带动量的梯度下降算法和 RMSProp 算法的结合,计算效率高,内存需求少,其更新的步长能通过设置初始学习率被限制在大致的范围内。Adam 的定义公式如下:

$$\begin{cases} m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ \widehat{m}_t = \dfrac{m_t}{1-\beta_1^t} \\ \widehat{v}_t = \dfrac{v_t}{1-\beta_2^t} \\ w_t = w_{t-1} + \dfrac{a}{\sqrt{\widehat{v}_t} + \varepsilon} \widehat{m}_t \\ g = \dfrac{1}{m} \nabla_{w_t} \sum_i L(f(x^{(i)}; w_t), y^{(i)}) \end{cases} \tag{9}$$

其中,$m_t$ 和 $v_t$ 分别表示对梯度的一阶矩估计和二阶矩估计;$\widehat{m}_t$ 和 $\widehat{v}_t$ 表示修正后的值;$g$ 和 $a$ 分别表示梯度和学习率;$\beta_1$ 和 $\beta_2$ 表示衰减速率;$\varepsilon$ 为 $10^{-8}$,$(f(x^{(i)}; w_t), y^{(i)})$ 表示样本。

4.2.2 Dropout 层和 *ReLU* 函数

Dropout 层是指在深度神经网络训练过程中,按一定的概率随机地丢弃神经网络中训练的神经元,从而达到一种对神经网络模型正则化的作用。深度网络拟合的参数较多,因此在某些层之间加入 Dropout 层,去除权重比例为 10%。

*ReLU* 激活函数是常用的神经网络激活函数,可增加模型的非线性的因素,增加模型的表达能力,加强模型的训练效果,其图像如图 7 所示。函数实质是将所有的小于 0 的负值均变成 0,而大于等于 0 的正值保持不变,定义公式如下:

$$ReLU(X) = \begin{cases} X, & X \geqslant 0 \\ 0, & X < 0 \end{cases} = \max(X, 0) \tag{10}$$

其中,$X$ 表示输出矩阵。

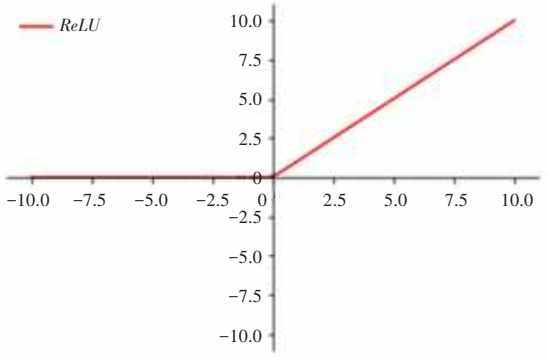


**图 7 *ReLU* 函数**

**Fig. 7 *ReLU* function**

## 5 实验与结果分析

### 5.1 实验环境及参数设置

本文所有实验均基于 Windows 系统实现,使用 Python3.7 和 PyTorch1.12 框架,硬件为 Intel i5 的 CPU。实验中参数设置如下:初始化学习率,0.005;训练批量(*batch_size*),64;dropout 层数,2($p$ = 0.2);优化算法,Adam;损失函数,*MSE Loss*;训练批量(*eopch*),150。训练集不乱序 *shuffle* = False,测试集乱序 *shuffle* = True。

### 5.2 实验数据

影视作品数据集来自 kaggle 网站公开数据集,由 2017 年 7 月及之前发布的影视作品组成,囊括了多种类型影片,包含大量已上映的热门影视作品和小众影片资源。该数据集能较好反映 20 世纪以来主流的影视作品方向,覆盖领域广,数据泛化性强。同时,影视作品涉及的年份跨度广,时间轴长,研究价值较为突出,电影发行年份统计如图 8 所示。

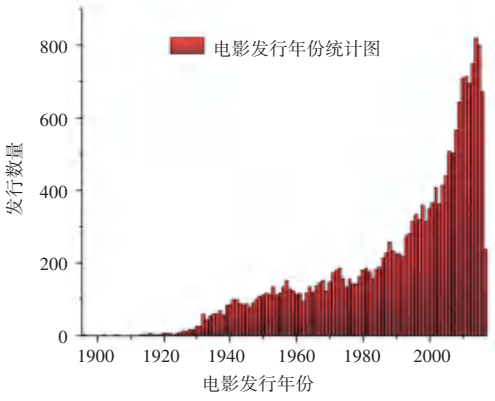


**图 8　影视作品发行年份统计**

**Fig. 8　Statistics of the year of publication**

原始影视作品数据集共 45 466 条数据，预处理包括缺失值/异常值处理，去除干扰项等。过滤评分的异常值，得到19 162条影视作品样本数据，评分值分布情况前后对比如图 9 所示。

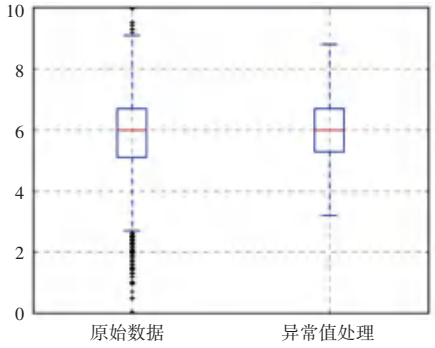


**图 9　影视作品评分分布**

**Fig. 9　Score distribution**

筛选后数据集中的文本型数据转换成机器能够识别的向量模式，对于不同类型的文本型数据，分别采用词袋模型、TF-IDF、特征拆分的方法进行向量化，归并数值型数据构成影视混合特征表示向量。

### 5.3　评估方法

本研究采用在回归任务中常用的 3 个模型评价指标：*MSE*、*MAE*、*SmoothL*1 *Loss*。影视作品数据集进行训练集和测试集划分后，深度神经网络模型利用训练集进行模型训练，迭代更新权重参数，并统计评价指标值，在每轮模型迭代中，统计测试数据集的平均值。各指标的研究剖析具体如下。

（1）均方误差（Mean Square Error，*MSE*）。反映估计量与被估计量之间差异程度的一种度量：

$$MSE(y_i,\hat{y}_i)=\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i-y_i)^2 \tag{11}$$

其中，$y_i$ 表示真实值；$\hat{y}_i$ 表示预测值；$N$ 表示样本数量，下同。

（2）平均绝对误差（Mean Absolute Error，*MAE*）。预测值和真实值之间绝对误差的平均数：

$$MAE(y_i,\hat{y}_i)=\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i-y_i| \tag{12}$$

（3）*SmoothL*1 *Loss*。当误差在(-1,1)上是平方误差，其他情况是 L1 损失，分段使用均方误差和平均绝对误差，用于回归模型：

$$SmoothL1\ Loss(y_i,\widehat{y}_i)=\begin{cases}\frac{1}{2}(\widehat{y}_i-y_i)^2, & \text{if } |\widehat{y}_i-y_i|<1\\ |\widehat{y}_i-y_i|-\frac{1}{2}, & \text{otherwise}\end{cases} \tag{13}$$

### 5.4　实验结果

当指标值逐渐收敛并达到平稳时，对 *MSE*、*MAE*、*SmoothL*1*Loss* 评价指标进行统计，保存深度神经网络的模型结构和参数并输出测试集中的影视作品评分预测值。部分影视作品评分真实值和预测值可视化如图 10 所示。

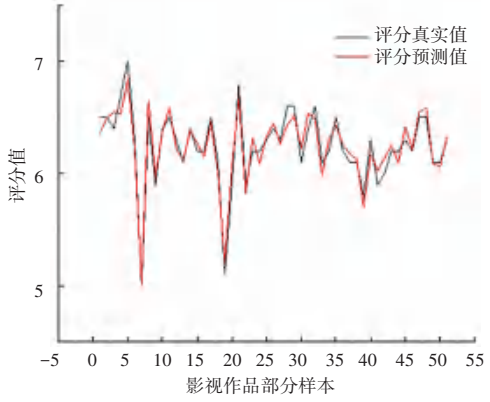


**图 10　影视作品数据预测值和真实值**

**Fig. 10　Predicted value and true value of publication**

测试集中的影视作品评分真实值和预测值在总的趋势上吻合，走势大致相同，基于深度神经网络的影视作品评分预测模型有很好的预测效果，在影视作品价值评估的应用领域具有一定的研究意义。

### 5.5　模型评价指标效果

基于深度神经网络的影视作品评分模型利用训练集进行训练，*MAE* 在前 90 轮迭代训练中出现一定程度的波动，90 轮迭代后开始收敛，经过 100 轮迭代后数值保持平稳，训练集 *MAE* 为 0.83。

在对影视作品评分预测模型的迭代更新中，统计每轮测试集评分预测值和真实值的 *MSE*，*MAE*，*SmoothL*1 *Loss*，3 个评价指标值于同一次实验中进行统计，趋势相同，随训练轮数 *epoch* 变化如图 11 所示。

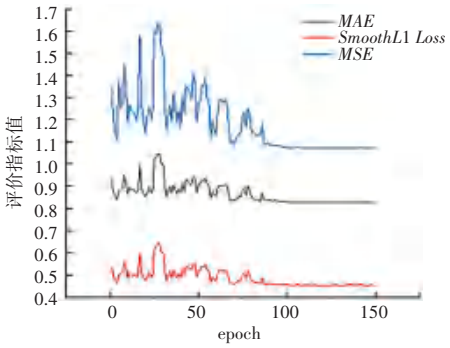


**图 11 *SmoothL1 Loss*、*MAE*、*MSE* 趋势图**

**Fig. 11 Trend chart of *SmoothL1 Loss*, *MAE*, *MSE***

在影视作品数据集上共进行 150 轮的训练，对 *MAE*、*MSE*、*SmoothL1 Loss* 的观察分析，模型在 90 轮迭代后开始收敛，100 轮迭代后 3 个指标值波动情况逐渐减少，趋向平稳。*MAE*、*MSE*、*SmoothL1 Loss* 数值收敛后统计结果见表 1。

模型评价指标值反映影视作品不同属性经过特征提取、词袋模型、TF-IDF 三种文本矢量化方法构建影视混合特征表示向量，输入到网络训练得出的评分预测值与真实值差距较小，预测效果良好，能充分评估未上映影视作品潜在的市场价值，可为决策人员提供必要的参考，将作品的最终价值最大化。

**表 1 评价指标**

**Table 1 Evaluation index**

| 指标 | 指标值 |
|---|---|
| *MSE* | 1.075 1 |
| *SmoothL1 Loss* | 0.453 2 |
| *MAE* | 0.827 1 |

### 5.6 案例研究

在数据集中随机选取 3 条影视作品数据来对基于混合特征表示向量的深度神经网络影视评分预测模型进行实例演示，评分预测结果见表 2。通过对模型的迭代训练得到最优模型后，将“影视片名”、“发行状态”、“发行日期”、“原始语言”、“影视类型”、“出版公司”、“出版国家”、“发行语言”、“影片标语”、“影片概述”输入到最优模型当中，预测模型将会输出影视作品的评分预测结果。

用户输入影视作品“Toy Story”，“Jumanji”，“Grumpier Old Men”的多项属性，模型将会分别输出 7.912，6.853，6.489 为最终的预测评分值。

**表 2 案例研究**

**Table 2 Case study**

| 作品属性 | 影视片名 | | |
|---|---|---|---|
| | Toy Story | Jumanji | Grumpier Old Men |
| 发行状态 | Released | Released | Released |
| 发行日期 | 1995/10/30 | 1995/12/15 | 1995/12/22 |
| 原始语言 | EN | EN | EN |
| 影视类型 | Animation, Comedy, Family | Adventure, Fantasy, Family | Romance, Comedy |
| 出版公司 | Pixar Animation Studios | TriStar Pictures, Teitler Film, Interscope Communications | Warner Bros, Lancaster |
| 出版国家 | United States of America | United States of America | United States of America |
| 发行语言 | English | English | English |
| 影片标语 | Led by Woody, Andy´s toys live happily in his room until... | When siblings Judy and Peter discover an enchanted... | A family wedding reignites the ancient... |
| 影片概述 | Now, empty that safe! | Roll the dice and unleash the excitement! | Still Yelling. Still Fighting. Still Ready for Love. |
| 预测评分 | 7.912 | 6.853 | 6.489 |

## 6 结束语

影视作品是文化领域的重要组成部分，利用其自身的属性信息通过特征拆分、词袋模型、TF-IDF 矢量化方法构建影视混合特征表示向量，基于深度神经网络建立影视作品评分预测模型，能达到很好的预测效果。实验表明，模型评价指标收敛情况很好，在 100 次迭代训练后收敛，测试集的 3 个指标值为：*MSE*，1.075 1、*SmoothL1 Loss*，0.453 2、*MAE*，0.827 1。通过挖掘影视作品属性信息建立影视作

品评分模型可以有效预测作品上映后的评分，满足影视行业对作品收益情况初步评估要求。

在对属性信息提取过程中，还存在表示向量维度过高或样本矩阵局部稀疏、字段不完整、语义偏差等问题。另外，影视作品导演、影视作品演员阵容、主要制作团队、影视作品关键词等信息未加入到考虑当中。因此，在未来对影视作品评分预测模型的优化探索中，可进一步考虑更多属性信息，并可通过降维等方法提取表示向量的主要特征，以达到更好的预测效果。

## 参考文献


[1] JEON T，CHO J，LEE S，et al. A movie rating prediction system of user propensity analysis based on collaborative filtering and fuzzy system[C]//2009 IEEE International Conference on Fuzzy Systems. Jeju，Republic of Korea：IEEE，2009：507-511.

[2] 何琦，袁芳英. 数字经济时代电影消费影响因素及票房预测研究—基于机器学习与模型融合视角[J]. 价格理论与实践，2021(9)：163-167，204.

[3] 李香君，肖小玲. 基于机器学习的电影评分预测研究[J]. 电脑知识与技术，2021，17(27)：109-111.

[4] 李旺泽. 基于监督学习的国产电影票房影响因素研究[D]. 武汉：湖北工业大学，2020.

[5] 李振兴. 机器学习在电影票房预测中的应用研究[D]. 西安：西安石油大学，2020.

[6] 张红丽，刘济郢，杨斯楠，等. 基于网络用户评论的评分预测模型研究[J]. 数据分析与知识发现，2017，1(8)：48-58.

[7] 任丹. 基于多元线性回归模型的电影票房预测系统设计与实现[D]. 广州：中山大学，2015.

[8] ZHANG Cong. Research on IMDB film score prediction based on improved whale algorithm[J]. Procedia Computer Science，2022，208：361-366.

[9] YOU Xuemei，LIU Yongdong，ZHANG Mingming，et al. Movie score predication model based on multiple nonlinear regression[J]. Tehnicki Vjesnik-technical Gazette ，2021，28(3)：914-921.

[10] ZAHABIYA M，SULTHANA R A，SHETTY D S. Movie rating prediction using ensemble learning algorithms[J]. International Journal of Advanced Computer Science and Applications(IJACSA)，2020，11(8)：383-388.

[11] SATHIYA D S，PARTHASARATHY G. Feature engineering based approach for prediction of movie ratings[J]. International Journal of Information Engineering and Electronic Business(IJIEEB)，2019，11(6)：24-31.

[12] 魏明强，黄媛. 网络评价对电影票房走势的影响[J]. 中国传媒大学学报(自然科学版)，2017，24(3)：68-71.

[13] MOHAMMED H，HAMADA M. Performance comparison of featured neural network trained with backpropagation and delta rule techniques for movie rating prediction in multi - criteria recommender systems[J]. Informatica (Slovenia)，2016，40(4)：409-414.

[14] SU Yumin，ZHANG Yuan，YAN JinYao. Neural network based movie rating prediction[C]// 2018 International Conference on Big Data and Computing(ICBDC'18). Shenzhen，China：ACM，2018：33-37.

[15] 阎亚亚. 词袋模型和 TF-IDF 在文本分类中的比较研究[J]. 电脑知识与技术，2021，17(28)：138-140.