Journal of Computer Applications
计算机应用,2014,34(3):742 -748
ISSN 1001-9081
CODEN JYIIDU
2014-03-10
http://www.joca.cn

文章编号:1001-9081(2014)03-0742-07
doi:10.11772/j.issn.1001-9081.2014.03.0742

基于神经网络的电影票房预测建模

郑 坚,周尚波*
(重庆大学 计算机学院,重庆 400044)
(* 通信作者电子邮箱 shbzhou@cqu.edu.cn)

摘 要: 针对电影票房预测与分类的研究中存在预测精度不高、缺乏实际应用价值等缺陷,通过对中国电影票房市场的研究,提出一种基于反馈神经网络的电影票房预测模型。首先,确定电影票房的影响因素以及输出结果格式;其次,对这些影响因子进行定量分析和归一量化处理;再次,根据确定的输入和输出变量确定各个网络层次神经元数量,建立神经网络结构,改进神经网络预测的算法和流程,建立票房预测模型;最后,用经过去噪处理的电影历史票房数据对神经网络进行训练。针对神经网络波动性的特点,对预测模型的输出结果进行改进之后,输出结果既能更可靠地反映电影在上映期间的票房收入,又能指出电影票房的波动范围。仿真结果表明,对于实验中的192部电影,基于神经网络算法的预测模型有较好的预测和分类性能(前5周票房的平均相对误差为43.2%,平均分类正确率可达93.69%),能够为电影在上映前的投资、宣传以及风险评估提供较全面、可靠的参考方案,在预测分类领域具有较好的应用价值和研究前景。

关键词: 多层反馈神经网络;电影票房预测;票房分类;影响因素量化
中图分类号: TP391.4; TP18 **文献标志码:** A

Modeling on box-office revenue prediction of movie based on neural network

ZHENG Jian, ZHOU Shangbo*
(*College of Computer Science, Chongqing University, Chongqing* 400044, *China*)

Abstract: Concerning the limitations that the accuracy of prediction is low and the classification on box-office is not significant in application, this paper proposed a new model to predict box-revenue of movie, based on the movie market in reality. The algorithm could be summarized as follows. Firstly, the factors that affected the box and format of the output were determined. Secondly, these factors should be analyzed and quantified within [0, 1]. Then, the number of neurons was also determined, aiming to build up the architecture of the neural network according to input and output. The algorithm and procedure were improved before finishing the prediction model. Finally, the model was trained with denoised historical movie data, and the output of model was optimized to dispel the randomness so that the result could reflect box more reliably. The experimental results demonstrate that the model based on back propagation neural network algorithm performs better on prediction and classification (For the first five weeks, the average relative error is 43.2% while the average accuracy rate achieves 93.69%), so that it can provide a more comprehensive and reliable suggestion for publicity and risk assessment before the movie is on, which possesses a better application value and research prospect in the prediction field.

Key words: multiple layer Back Propagation Neural Network (BPNN); movie Box-office Revenue Prediction (BRP); box-office classification; factor quantification

0 引言

作为文化生活的重要组成部分,电影不但丰富了人们的业余生活和精神世界,而且也是不同文化背景的国家和地区之间进行文化交流的重要媒介。随着电影学这门新生课题的形成,电影逐步由一种单纯的艺术形式,演变成了一种艺术形式的商品[1-2]。票房指标逐渐成为电影投资机构竞相追逐的目标,电影的投资风险评估对投资及发行机构具有重要意义。电影票房收入预测是确保电影发行投资回报、控制发行风险的重要手段,对于投资决策具有重要的实际意义[3]。作为一种具有短暂生命周期的商品,电影在上映的档期内产生票房。由于影响票房的因素量化难度大,准确预测电影在其生存周期内的盈利情况是很困难的[4]。

当前关于电影票房预测的研究较少,相关的应用更是罕见。Marshall等[5]提出使用电影历史数据预测电影上映期间的累计观众数量,使用的是简单的多元线性回归算法预测第1周的观众人数,并使用Sawhney等[6]的模型进行预测电影在上映后几周的累计观众数量。该模型可用于在电影上映的不同周期内累计观众数量的预测,其实际意义在于在电影上映的生命周期内,电影院可动态地调整放映策略,例如扩大或缩小放映厅的数量、改变放映周期等。然而,该方法存在如下缺陷:首先,Marshall等[5]在用多元线性回归算法预测第1周累积观众时考虑的电影影响因素较少(电影拷贝数、用户评价、影院数量、观众年龄),并没有考虑电影吸引观众的特殊

收稿日期:2013-09-16;修回日期:2013-11-14。 基金项目:国家自然科学基金资助项目(61103114)。
作者简介:郑坚(1988 -),男,福建三明人,硕士研究生,主要研究方向:人工神经网络、数据挖掘; 周尚波(1963 -),男,广西宁明人,教授,博士,主要研究方向:人工神经网络、混沌及其控制理论、图像处理、信息安全、物理工程计算、计算机仿真。

属性,导致第 1 周的预测误差过大。进一步地,这种误差会在使用扩散模型预测后续几周的观众数量时,不断积累,影响最后的预测精度(Marshall 得到的平均误差: 第 1 周为 117.62%,第 3 周为 126.80%,第 6 周为 105.46%) 。

Barman 等[7] 发现反馈神经网络算法在股票市场、天气预报和图像处理等领域的应用非常成功,提出使用反馈神经网络算法来预测电影的盈利与否,虽局部展示了较好的预测准确率。然而,Barman 等[7] 只把电影类型作为影响票房的单一输入,且对于输入和输出结果只是使用简单的布尔型数值表示; 其次,提出的神经网络结构过于简单(只含有 1 个隐藏层) ,这样的输入-输出映射关系忽略了譬如导演、演员等因素的影响。此外,Barman 等[7] 对电影是否盈利并没有严格的判断标准,失去了实际的应用价值。

Sharda 等[8] 结合影响票房的多个电影属性,以多层神经网络算法为基础,提出一种电影票房分类模型,并使用分类正确率作为评估模型分类性能的主要指标,取得了较好的分类效果。然而,该方法使用二进制的离散数来量化电影票房的各个影响因素,显然是一种模糊的处理方式,没有根据实际情况对这些变量进行不同的量化处理,因此不能完整地体现影响因素中不同变量的差异性,例如: 导演、演员的影响程度不能只是简单地用 0 和 1 表示。而各个影响因素的量化处理,能通过神经网络的权值连接,影响神经网络训练收敛的程度,进而影响神经网络的分类性能。此外,该预测模型在输出层对票房的分类同样显得模糊,使得每一个票房等级分类的跨度过大(例如第二类的票房范围在 [100 000,1 000 000]) 。这样的分类对于电影投资者和电影院控制电影制作、放映成本而言,参考价值不大。

针对当前票房预测研究领域存在的问题,本文结合反馈神经网络(Back Propagation Neural Network, BPNN) 与中国内地电影市场的实际情况[9] ,提出一种基于多层反馈神经网络的票房预测(Box-office Revenue Prediction, BRP) 模型,该模型从多个维度考虑电影票房的影响因素,能够较准确地预测票房的具体数值。考虑到电影票房的随机波动性,为了提高预测精度,模型对算法和预测流程进行改进,给出了票房波动的范围区间,既能保证最终的预测结果具备较好的预测精度,又能对电影的风险控制提供有价值的投资参考,具有实际的应用前景。

1　反馈神经网络

BPNN 是一种根据误差逆扩散算法训练的多层前馈网络。上下层之间实现全连接,每层内神经元之间无连接。图 1为 BP 网络结构示意图,其中 f 为隐含层的激活函数,可以是线性的,也可以是非线性的,主要由输入、输出映射关系确定。

通过调整 BP 神经网络的规模(输入节点数、输出层节点数、隐含层层数及隐层节点数) 及网络中连接权值,就可以实现非线性分类问题,并且能以任意精度逼近任何非线性函数。

BPNN 能学习和存储大量的输入-输出模式映射关系,而无需事前揭示描述这种映射关系的数学函数或是方程。因此,BPNN 成功地应用于图像压缩[10] 、天气预报[11] 、破产预测[12] 、卫星图像分类[13] 、不规则形状分类[14] 、邮件分类[15] 。

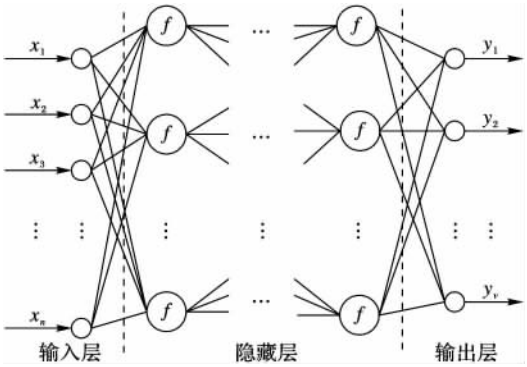

图 1　BP 网络结构

2　BRP 模型

BRP 模型的建立经历如下阶段: 首先,结合中国内地电影市场的实际情况,确定电影票房的影响因素; 其次,对这些影响因素进行量化处理,形成神经网络能够接受的数据格式; 此外,还需要对电影的历史数据进行去噪处理,构建 BPNN 的训练集,由此便可以初步确定神经网络的预测模型以及输入-输出的映射结构。最后,在 BPNN 预测结构的基础上,通过对预测算法以及流程的改进,确定最终的 BRP 模型。BRP 模型的具体建模过程如图 2 所示。

图 2　BRP 建模过程

2.1　影响因子的确定

为了简化计算并提高系统的性能,需合理选择变量。Chua 等[16] 提出使用摇摆的值定义输入变量,但该方法缺乏理论依据,而且算法的具体实现并不直观。本文的变量选择参考 Sharda 等[8]245-247 的研究,同时,在对中国内地电影票房历史数据统计分析的基础上,结合中国内地电影市场的实际情况[9] ,选取导演、第一主演、第二主演、第一类型、第二类型、发行地区、上映档期作为输入因素,并赋予各因素不同的权重,式(1) 表示电影的输入向量。考虑到连续的变量值可以提高神经网络的敏感性,本文将所有的输入变量都处理成 [0,1] 区间的连续数值,2.3 节将具体阐述处理过程。

$$\boldsymbol{Input} = \{DirectorWeight, ActorWeight_i, GenreWeight_i, NationWeight, DateWeight\} \quad (1)$$

其中: $DirectorWeight$ 表示导演的影响力权值,$ActorWeight_i$($i=1,2$) 表示主演的影响力权值,$GenreWeight_i$($i=1,2$) 表示电影类型的影响力权值,$NationWeight$ 表示上映地区的权值,$DateWeight$ 表示上映档期的权值。

2.2　数据预处理

电影在首映的第 1 周票房数据对于该电影的预测具有重要意义,它往往能体现观众对这部电影的关注程度,Marshall

等[5-1803]的研究中把第 1 周的观众作为扩散模型的潜在观众。然而,由于每部电影在首映第 1 周的实际上映天数不尽相同(如《锦衣卫》于周三上映,第 1 周实际上映 5 天;《大兵小将》于周日上映,第 1 周实际上映 1 天),因此,在对影响因子量化处理之前,本文使用式(2)对所用电影的第 1 周票房数据做出修正,消除这种由于数据不规整性而带来的实验误差:

$$b_{j1} = \frac{7}{days_j} \cdot b'_{j1} \tag{2}$$

其中:b'_{j1} 表示第 j 部电影首映的第 1 周实际票房,$days_j$ 表示第 j 部电影在首映第 1 周内实际上映的天数,b_{j1} 表示第 j 部电影修正后的首映第 1 周票房数据。

由于每部电影实际放映的周数也不尽相同,因此本文约定将每部电影上映前 5 周的累计总票房作为该部电影上映期间的累计票房收入,2.3 节的量化处理也将使用这一标准。另一方面,考虑到动画类型的电影与其他电影在观众群体上的差异性,为了能研究中国内地电影市场的一般化规律,本文将剔除动画类型的电影。

2.3 影响因子归一量化

本节将具体阐述影响电影票房的因素并给出相应的定义,同时给出电影票房数据各属性的量化处理过程,为 BP 神经网络训练集的构建做准备。

2.3.1 导演

定义 1 导演 i 的票房影响力指标 Dir_i:

$$Dir_i = \left(\sum_{j=1}^{m} \sum_{k=1}^{5} b_{jk} \right) / m; \quad m = \min\{5, m\} \tag{3}$$

其中:i 表示导演序号;j 表示导演 i 参与拍摄的第 j 部电影;k 表示上映的周次;m 表示导演 i 参与拍摄的所有电影中,上映时间距离现在时间最近的 m 部电影;b_{jk} 表示最近第 j 部电影在上映的第 k 周内产生的票房。

进一步地,可以得到衡量导演 i 的影响力对其执导电影的票房影响力权值 $DirectorWeight_i$ 如下:

$$DirectorWeight_i = \left(\lg \frac{Dir_i}{Dir_{\min}} \right) / \left(\lg \frac{Dir_{\max}}{Dir_{\min}} \right) \tag{4}$$

其中:

$$Dir_{\max} = \max\{Dir_1, Dir_2, \cdots\} \tag{5}$$

$$Dir_{\min} = \min\{Dir_1, Dir_2, \cdots\} \tag{6}$$

i 表示导演序号,Dir_i 表示第 i 个导演的影响力。

2.3.2 演员

定义 2 演员 i 的票房影响力指标 A_i:

$$A_i = \left[\sum_{j=1}^{m} u_{ij} \left(\sum_{k=1}^{5} b_{jk} \right) \right] / m; \quad m = \min\{5, m\} \tag{7}$$

其中:i 表示演员序号;j 表示主演 i 参与拍摄的第 j 部电影;k 表示上映的周次;m 表示主演 i 参与拍摄的所有电影中,上映时间距离现在时间最近的 m 部电影;b_{jk} 表示最近第 j 部电影在上映的第 k 周内产生的票房;u_{ij} 为演员 i 参与的最近第 j 部电影的参演系数,定义如下:

$$u_{ij} = \begin{cases} 1 - (n-1)/10, & n \in [1, 5] \\ 0.5, & n \in (5, +\infty) \end{cases} \tag{8}$$

其中 n 为正整数,表示演员 i 在第 j 部电影中主演的名次顺序。

进一步地,可以得到衡量主演 i 的影响力对参与的电影的票房影响力权值 $ActorWeight_i$:

$$ActorWeight_i = \left(\lg \frac{A_i}{A_{\min}} \right) / \left(\lg \frac{A_{\max}}{A_{\min}} \right) \tag{9}$$

其中:

$$A_{\max} = \max\{A_1, A_2, \cdots\} \tag{10}$$

$$A_{\min} = \min\{A_1, A_2, \cdots\} \tag{11}$$

其中:i 表示演员序号,A_i 表示第 i 个演员的影响力。

2.3.3 电影类型

互联网电影资料库(Internet Movie Database,IMDB)是目前最具权威的电影网站,它可以为电影内容的分类提供参考。结合 IMDB 的电影类型种类,本文将电影按内容分成如下 12 种类型:爱情片、灾难片、悬疑片(冒险、犯罪)、恐怖片(惊悚)、战争片、纪录片(传记、历史)、家庭片、戏剧(音乐、戏曲)、科幻片(魔幻、奇幻)、喜剧片、动作片(武侠、古装)、剧情片(故事)。

下面给出各类型的票房影响力指标的定义。

定义 3 电影类型 i 的票房影响力指标 G_i:

$$G_i = \sum_{j=1}^{m} \sum_{k=1}^{5} b_{jk} \tag{12}$$

其中:i 表示类型序号($i = 1, 2, \cdots, 12$,分别对应上述 12 种电影类型);k 表示上映的周次;m 表示类型 i 的电影总量;j 表示属于类型 i 的第 j 部电影;b_{jk} 表示内容类型为 i 的第 j 部电影在上映的第 k 周内产生的票房。

进一步地,可以得到衡量类型 i 对归属于该类型电影的票房影响力权值 $GenreWeight_i$:

$$GenreWeight_i = \left(\lg \frac{G_i}{G_{\min}} \right) / \left(\lg \frac{G_{\max}}{G_{\min}} \right) \tag{13}$$

其中:

$$G_{\max} = \max\{G_1, G_2, \cdots\} \tag{14}$$

$$G_{\min} = \min\{G_1, G_2, \cdots\} \tag{15}$$

2.3.4 国家地区

根据电影制片公司所在的国家或地区,本文将电影分为欧美、日韩、港台、中国内地和其他地区 5 类。

定义 4 国家或地区 i 的票房影响力指标 N_i:

$$N_i = \sum_{j=1}^{m} \sum_{k=1}^{5} b_{jk} \tag{16}$$

其中:i 表示发行地区序号($1 \leqslant i \leqslant 5$),$i$ 的取值与地区的对应关系如表 1 所示;k 表示上映的周次;m 表示发行地区属于地区 i 的电影总数;j 表示发行地区属于地区 i 的第 j 部电影;b_{jk} 表示发行地区为 i 的第 j 部电影在上映的第 k 周内产生的票房。

表 1 i 的取值与电影发行地区的对应关系

i	发行地区	i	发行地区
1	欧美	4	中国内地
2	日韩	5	其他地区
3	港台		

进一步地,可以得到衡量发行地区 i 的影响力对发行地区属于该类别的电影的票房影响力权值 $NationWeight_i$:

$$NationWeight_i = N_i / \sum_{j=1}^{5} N_j \tag{17}$$

其中:i 表示发行地区序号,N_i 表示发行地区 i 的影响力,N_j 表示发行地区 j 的影响力权值。

2.3.5 上映档期

根据电影在中国内地上映前后 3 天所处的节假日,本文将电影上映档期分为五一档(4 月 27 日至 5 月 10 日)、暑期

档(7 月 1 日至 9 月 1 日)、国庆档(9 月 27 日至 10 月 10 日)、贺岁档(正月初一至正月十五)和其他档期共 5 类。

定义 5　档期 i 的票房影响力指标 D_i:

$$D_i = \sum_{j=1}^{m}\sum_{k=1}^{5} b_{jk} \tag{18}$$

其中: i 表示档期序号($1 \leqslant i \leqslant 5$),$i$ 的取值与地区的对应关系如表 2 所示; k 表示上映的周次; m 表示上映日期在档期 i 的电影总数; j 表示上映日期在档期 i 的第 j 部电影; b_{jk} 表示上映日期在档期 i 的第 j 部电影在上映的第 k 周内产生的票房数据。

表 2　i 的取值与上映档期的对应关系

i	档期	i	档期
1	五一档	4	贺岁档
2	暑期档	5	其他档期
3	国庆档		

进一步地,可以得到衡量档期 i 的影响力对在该档期内上映的电影票房的影响力权值 $DateWeight_i$:

$$DateWeight_i = \left(\lg\frac{D_i}{D_{\min}}\right) \Big/ \left(\lg\frac{D_{\max}}{D_{\min}}\right) \tag{19}$$

其中:

$$D_{\max} = \max\{D_1, D_2, \cdots\} \tag{20}$$

$$D_{\min} = \min\{D_1, D_2, \cdots\} \tag{21}$$

其中: i 表示上述档期序号,D_i 表示档期 i 的影响力。

2.4　确定模型结构

2.4.1　输入层

根据 2.1 节对影响变量的分析,决定 BRP 模型的输入层神经元共有 7 个,即导演、第一主演、第二主演、第一类型、第二类型、发行地区、上映档期,用于处理式(1)的权值向量输入。

2.4.2　输出层

本文中设计 BRP 模型的输出值为电影票房,因此输出层只含一个神经元,其激活函数使用式(22)的 Sigmoid 函数:

$$f(x) = \frac{1}{1+e^{-x}}; \quad x \in \mathbf{R} \tag{22}$$

其中 x 表示上一层的所有神经元对本神经元的净输入。

由于 Sigmoid 函数的值域在[0,1]区间,因此需要使用用式(23)对输入到 BPNN 中用于训练的电影票房做如下的归一处理:

$$box_{jk} = \left(\sum_{i=1}^{k} b_{ji}\right) \Big/ \max\left\{\sum_{i=1}^{k} b_{1i}, \sum_{i=1}^{k} b_{2i}, \cdots\right\} \tag{23}$$

其中: j 表示电影序号; k 表示上映的周次; b_{ji} 表示第 j 部电影在上映的第 i 周内产生的票房; box_{jk} 表示可用于 BPNN 训练的第 j 部电影上映截止到第 k 周结束时的累计票房,$box_{jk} \in [0,1]$。

2.4.3　隐藏层

1988 年 Cybenko[17] 指出,当各节点均采用 S 型函数时,一个隐含层就足以实现任意判决分类问题,两个隐含层则足以表示输入图形的任意输出函数。Lippman[18-19] 利用它对多层网络功能的几何解释,指出第二隐含层的节点数应为 $M \times 2$; 这里 M 为输出层的节点数。在高维输入时,第一隐含层与第二隐含层的最佳节点数的比例为 3∶1[20-22]。

基于上述理论,本文采用两个隐藏层,当输出层节点数量为 1 时,第二隐藏层的节点数为 $M \times 2 = 2$,第一隐藏层的节点数为 $3 \times (M \times 2) = 6$,且所有隐藏层节点都使用式(22)的 Sigmoid 函数。表 3 表示 BRP 模型中各输入变量以及输出阈值,图 3 给出了本文提出的 BRP 模型结构。

表 3　BRP 模型的输入以及输出值域

变量	变量含义	值域	变量	变量含义	值域
输入变量 1	导演	[0,1]	输入变量 5	第二类型	[0,1]
输入变量 2	第一主演	[0,1]	输入变量 6	发行地区	[0,1]
输入变量 3	第二主演	[0,1]	输入变量 7	上映档期	[0,1]
输入变量 4	第一类型	[0,1]	输出变量	票房(box_{jk})	[0,1]

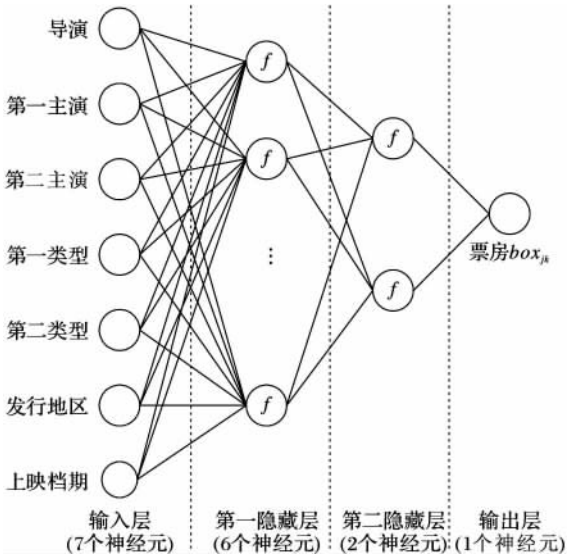

图 3　BRP 模型结构

2.5　预测算法改进

随着 BPNN 隐藏层节点数量的增加,BRP 模型的计算容纳能力也随着增加,但同时对预测的结果带来波动性,这是因为输出层之前任意两个节点之间的权值收敛方向发生变化,都会对结果造成影响。因此,需要对 BRP 模型的预测算法进行改进,本文通过对同一组数据的多次预测,找出数据中心点,进而确定结果波动的范围。改进后预测算法的伪代码表示如下。

```
Input:
  TrainSet[n][8]: 用于训练的 n 部电影,其中,TrainSet[i][j] = {DirectWeight, ActorWeight1, ActorWeight2, GenreWeight1, GenreWeight2, NationWeight, DateWeight, boxik};
  PredictSet[m][7]: 需要预测第 k 周累计票房的 m 部电影,其中 PredictSet[i][j] = {DirectWeight, ActorWeight1, ActorWeight2, GenreWeight1, GenreWeight2, NationWeight, DateWeight};
  E0: 允许的全局最小误差;
  t: 连续预测的次数;
  p: 预测结果的可信百分比。
Output:
  avg[m]: m 部电影第 k 周预测得出的累计票房值;
  range[m]: m 部电影第 k 周预测得出的累计票房波动值。
Begin                                                    //算法开始
  maxbox = max(TrainSet[][8]);                           //获得票房最大值
  for i = 1 to m
    for j = 1 to t
      net = init(η, W0);                                 //初始化神经网络
      net = train(E0, TrainSet);                         //训练神经网络
      out[j][0] = predict(net, PredictSet[i][]);
```

```
    out[i][j] = out[i][j] × maxbox;          //预测票房
  end
end
q = ⌈t × (1 - p)⌉;
for i = 1 to m
  avg[i] = average(out[i]);
  //删除 out[i] 中距离 avg[i] 最远的 q 个数:
  deleteFurthest(out[i],q,avg[i]);
  avg[i] = average(out[i]);
  max[i] = max(out[i]);
  min[i] = min(out[i]);
  range[i] = max(max[i] - avg[i],avg[i] - min[i]);
end
End                                          //算法结束
```

在本文的仿真实验中,取经验值参数 $E_0 = 10^{-3}$,$t = 10$,$p = 70\%$。

根据上述算法,第 i 部电影最终预测结果可用式(24)表示:

$$avg[i] \pm range[i] = \text{BRPmodel}(PredictSet[i]) \quad (24)$$

其中:$PredictSet[i]$ 表示第 i 部预测电影的7个影响因素权值向量(参考式(1)),$avg[i]$ 表示第 i 部电影的累计票房中心预测值,$range[i]$ 表示第 i 部电影的累计票房波动范围预测值,第 i 部电影的预测结果可表示为 $[avg[i] - range[i], avg[i] + range[i]]$。

3 仿真实验

本文用于仿真实验的电影数据来源于艺恩咨询(www.entgroup.cn)。艺恩咨询是中国专业的咨询机构,提供包括电影票房在内的权威商业数据,本文收集了2008至2010年之间192部电影在中国内地上映的票房以及电影数据,作为本实验的数据集。由于数据集中上映超过5周的电影数据较少,因此本文的预测只针对电影上映前5周的票房。表4是将票房按上映周次分类统计的结果。

表4 上映前5周的电影数量分布

上映周次	电影数量	上映周次	电影数量
第1周	192	第4周	146
第2周	190	第5周	100
第3周	190		

在实验工具选择上,考虑到Matlab在科学计算领域有着较高精度,因此本文使用Matlab实现BRP模型与Marshall的扩散模型,并对比它们的性能。

3.1 评价方法

由于BRP模型的输出结果是由票房值和票房波动范围两部分组成(见式(24)),因此需要对这两部分的预测性能分别进行评估。

Marshall等[5]1802的扩散模型如式(25)表示:

$$N(t) = \frac{N}{s-h}\left[(s-h) + he^{-st} - se^{-ht}\right] \quad (25)$$

其中:N 表示电影的潜在观众数量,即第1周观众人数,为了便于比较,本文将其作为电影第1周的票房数据;t 表示电影上映周次,$t \geqslant 2$,$N(t) > 0$ 表示在时间 t 内,实际去看电影的观众数,本文将其视为时间 t 内的累积票房;s 表示观众决定看电影的相关系数;h 表示观众实际去看电影的相关系数。本节的模型评估中将随机抽取一定数量的训练集,使用非线性最小二乘法确定上述两个系数。

对于票房值的评价,本节采用式(26)定义的平均相对误差作为票房预测的评价指标,并使用相同的数据集分别对本文提出的BRP模型输出的票房值($avg[n]$)和扩散模型[5]1802输出的票房($N(t)$)进行对比。

$$\bar{E} = \left(\sum_{j=1}^{n} \frac{|box_{jk} - box'_{jk}|}{box_{jk}}\right)/n \times 100\% \quad (26)$$

其中:j 表示电影序号;k 表示上映的周次;box_{jk} 表示第 j 部电影在上映截至第 k 周结束时的实际累计票房;box'_{jk} 表示第 j 部电影在上映截至第 k 周结束时的累计票房预测值(对于BRP模型,$box'_{jk} = avg[j]$;对于扩散模型,$box'_{jk} = N(k)$);n 表示电影的数量;$\bar{E}$ 表示 n 部测试电影的平均相对误差。

对于票房波动范围的评价,本文采用式(28)定义的预测正确率作为评价指标[8],用于评估BRP模型在预测票房波动范围时的分类性能,并与Sharda等[8]247提出的BP票房分类模型进行对比。

$$bingo_j = \begin{cases} 1, box_{jk} \in [avg[j] - range[j], avg[j] + range[j]] \\ 0, box_{jk} \notin [avg[j] - range[j], avg[j] + range[j]] \end{cases} \quad (27)$$

$$BingoRate = \left(\sum_{j=1}^{n} bingo_j\right)/n \quad (28)$$

其中:j 表示电影序号;k 表示上映的周次;box_{jk} 表示第 j 部电影在上映截至第 k 周结束时的实际累计票房;$avg[j]$ 表示电影在上映截至第 k 周结束时的累计票房预测值;$range[j]$ 表示电影在上映截至第 k 周结束时的累计票房波动范围预测值;n 表示电影的数量;$BingoRate$ 表示正确率;$bingo_j$ 表示预测第 j 部电影的分类情况(1表示分类正确,0表示分类错误)。

3.2 误差对比

针对第1周票房的预测,Marshall等[5]1803在研究智利的电影市场时把广告花费、电影拷贝数、用户评价、戏院数量、观众年龄分布作为变量并使用多元线性回归预测电影第1周的票房(潜在观众),本节在进行对比时,除了给出使用BRP模型的预测误差,还将上文提及的票房影响力因素作为输入变量,采用多元线性回归方法进行预测并给出相应的预测误差,实验中BRP模型和Regress1所使用的训练集与测试集比例为182:10。误差比较如表5所示。

表5 3种方法在第1周票房平均相对误差对比 %

方法	平均相对误差
BRP模型	45.42
Regress1	103.91
Regress2[5]1805	117.62

表5中:BRP模型即为本文方法;Regress1表示把导演权值、第一主演权值、第二主演权值、第一类型权值、第二类型权值、发行地区权值、档期权值作为输入变量,使用多元线性回归进行预测;Regress2表示在Marshall等[5]1805的研究中得到的预测误差。该对比结果表明,中国内地电影市场与智利电影市场的票房影响因素不同,其预测结果也存在差异,本文所列举的因素对中国内地电影票房影响更大,更适合中国内地电影票房的预测研究。

在进行第2周至第5周的票房预测时,本文将收集的可用样本随机地分成两部分:一部分为训练集;另一部分为测试集,测试集与训练集没有重复数据。根据相关研究结果和作者已有经验,结合数据样本数量,经过多次随机实验,发现在

训练集与测试集比例约9∶1时，BRP 模型和 Marshall 等[1802]的扩散模型有较好的预测精度。

2 种模型对比实验的条件为: 在选取的训练集基础上，使用非线性最小二乘法确定式(25) 中 s 和 t 两个系数。同时，使用相同的训练集训练 BRP 模型中的 BP 神经网络，直至收敛。由此，便可建立可直接用于预测票房的两种模型。

图 4 ~ 7 是分别使用 BRP 模型和 Marshall 等[1802]的扩散模型预测电影第 2 周到第 5 周的累计票房值的误差曲线比较; 表 6 为分别使用 BRP 模型和扩散模型进行预测的平均相对误差对比。

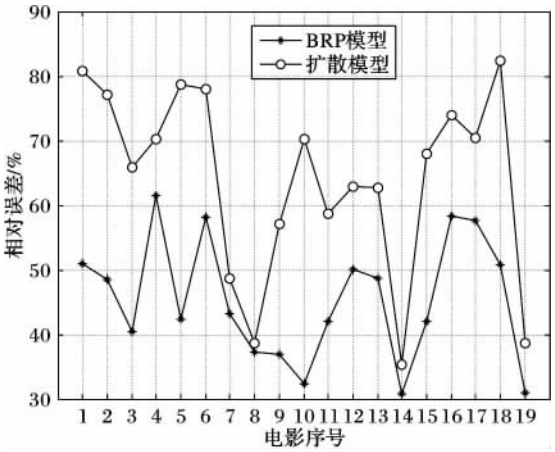

图 4　2 种模型在上映第 2 周票房预测的相对误差对比

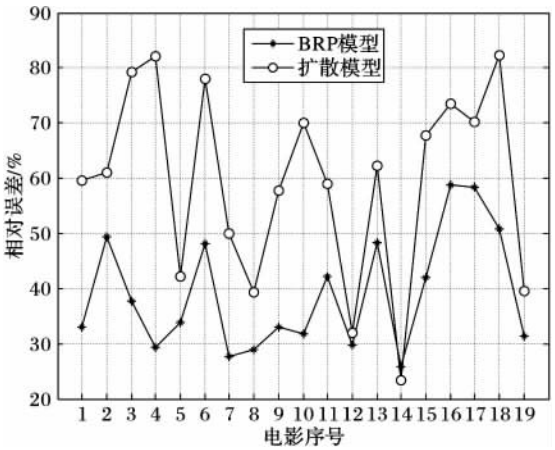

图 5　2 种模型在上映第 3 周票房预测的相对误差对比

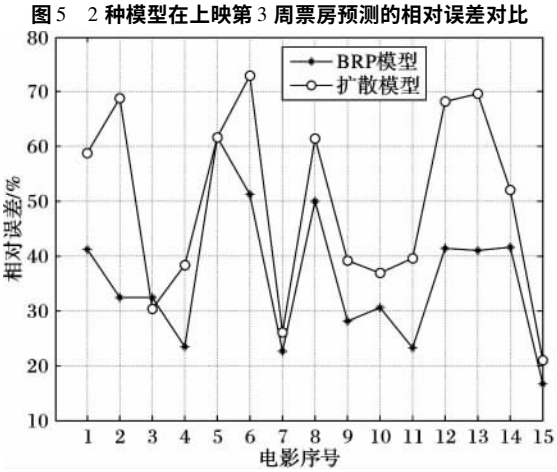

图 6　2 种模型在上映第 4 周票房预测的相对误差对比

从上述对预测的平均误差对比实验中可以看出，本文所设计的 BRP 模型预测性能明显优于 Marshall 等[1802]的扩散模型，有效解决了 Marshall 等[1802]在研究中预测精度过低的问题。

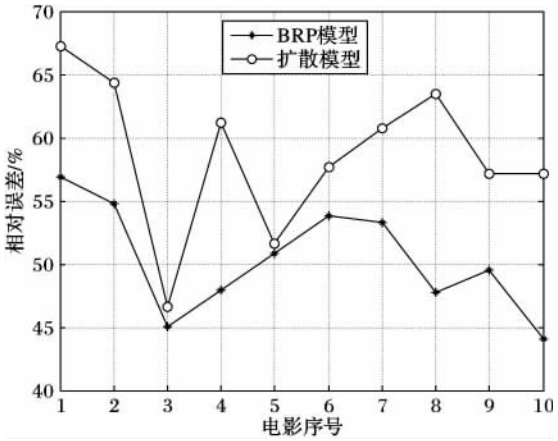

图 7　2 种模型在上映第 5 周票房预测的相对误差对比

表 6　2 种模型在不同上映周次的平均相对误差对比　%

上映周次	平均相对误差		上映周次	平均相对误差	
	BRP 模型	扩散模型[1802]		BRP 模型	扩散模型[1802]
第 2 周	45.42	64.14	第 4 周	35.84	49.67
第 3 周	38.94	59.46	第 5 周	50.38	58.71

3.3　分类结果对比

本节利用式(27)、(28) 的评估指标，分别使用 BRP 模型和 Sharda 等[247]在研究中提出的 BP 分类方法，对相同的数据集进行测试，每次测试随机抽取的训练集与测试集不重复。其中，BP 分类的正确率计算式如(29) 所示:

$$\text{BP 分类正确率} = \frac{\text{分类正确的样本数}}{\text{样本总数}} \times 100\% \tag{29}$$

图 8 ~ 10 表示当使用的训练集比例分别为 60%、70% 和 80% 时的分类正确率比较。

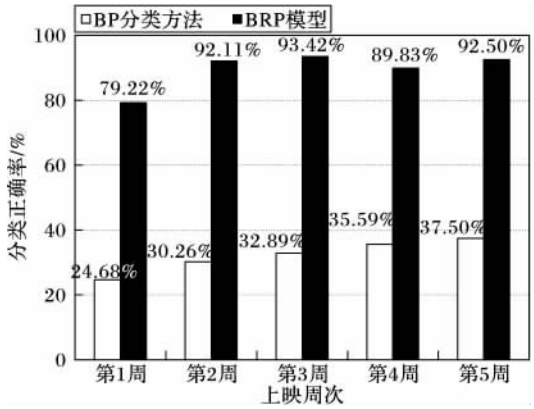

图 8　2 种模型在训练集为 60% 时的分类正确率对比

上述实验结果表明，随着训练集所占比例的提高，2 种模型的分类正确率都有显著的提高。然而，与 Sharda 等[247]的方法相比，本文提出的 BRP 模型的电影分类性能更好。

4　结语

本文从实际的中国内地电影票房市场出发，提出一种基于反馈神经网络的电影票房预测模型，将多层反馈神经网络算法应用于电影票房预测领域，把导演、第一主演、第二主演、第一类型、第二类型、发行地区、上映档期作为影响票房的因素，并对其进行不同的归一量化处理。此外，还对 BP 神经网络的输出、隐藏层、神经元结构做了调整，改进了算法和预测流程，并建立电影票房预测的 BRP 模型。通过对比实验，

BRP模型比Marshall等[5]1802的传播模型在预测票房数值的平均相对误差更低。同时,BRP模型预测的票房波动范围比Sharda等[8]247的BP分类方法在预测票房分类时的正确率更高。综上所述,本文提出的BRP模型既解决了电影票房预测精度过低的问题,又能较准确地给出票房预测的波动范围,能够为电影的投资和放映提供有价值的参考,具有实际的意义。

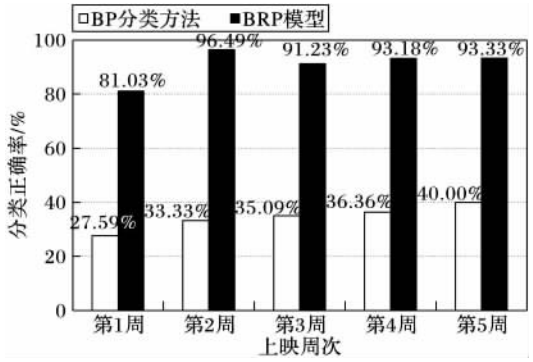

图9 2种模型在训练集为70%时的分类正确率对比

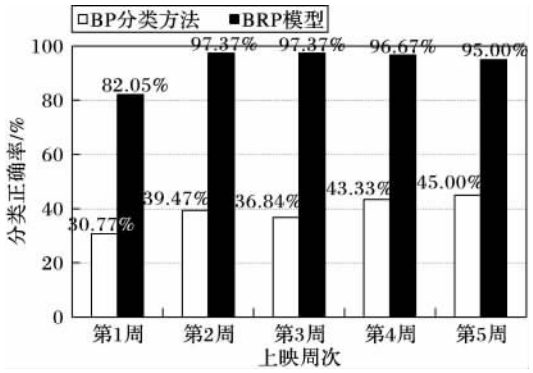

图10 2种模型在训练集为80%时的分类正确率对比

然而,作为一次对电影票房预测问题的探索和大胆尝试,电影票房模型对预测领域的研究讨论是十分有价值的。将反馈神经网络应用于票房预测,本身就是一次大胆的创新,下一步的研究方向如下:首先,及时地更新电影库中的电影信息。对于距今年代较远的电影、导演、演员,可以在模型中适当增加或者减弱它的票房影响力;其次,对于没有历史票房记录的新导演或新演员,尝试从其他角度衡量他们对票房的影响力(如在媒体中出现的频率);再次,尝试研究某些导演和演员之间的"强强合作"效应,使用组合实验探讨不同电影属性组合后的综合效应。

参考文献:

[1] GINSBURGH V A, THROSBY D. Handbook on the economics of art and culture: volume 1 [M]. Amsterdam: North-Holland, 2006: 615 - 659.

[2] CAVES R E. Creative industries: contracts between art and commerce [M]. Cambridge: Harvard University Press, 2000: 2 - 17.

[3] JEHOSHUA E, ELBERSE A, LEENDERS M A A M. The motion picture industry: critical issues in practice, current research, and new research directions [J]. Marketing Science, 2006, 25(6): 638 - 661.

[4] CHANG B H, KI E J. Devising a practical model for predicting theatrical movie success: focusing on the experience good property [J]. Journal of Media Economics, 2005, 18(4): 247 - 269.

[5] MARSHALL P, DOCKENDORFF M, IBÁÑEZ S. A forecasting system for movie attendance [J]. Journal of Business Research, 2013, 66(13): 1800 - 1806.

[6] SAWHNEY M S, ELIASHBERG J. A parsimonious model for forecasting gross box-office revenues of motion picture [J]. Marketing Science, 1996, 15(2): 113 - 131.

[7] BARMAN D, CHOWDHURY N, SINGHA R K. To predict possible profit/loss of a movie to be launched using MLP with back-propagation learning [C]// Proceedings of the 2012 International Conference on Communications, Devices and Intelligent Systems. Piscataway, NJ: IEEE Press, 2012: 322 - 325.

[8] SHARDA R, DELEN D. Predicting box-office success of motion pictures with neural networks [J]. Expert Systems with Applications, 2006, 30(2): 243 - 254.

[9] WU X. Six factors which effect on office-box [J]. China Movie Market, 2006(4): 14 - 15. (吴宣文. 影响电影票房的六大因素[J]. 中国电影市场, 2006(4): 14 - 15.)

[10] ANNA DURAI S, ANNA SARO E. Image compression with back-propagation neural network using cumulative distribution function [J]. International Journal of Applied Science, Engineering and Technology, 2007, 3(4): 185 - 189.

[11] BABOO S S, SHEREEF I K. An efficient weather forecasting system using artificial neural network [J]. International Journal of Environmental Science and Development, 2010, 1(4): 321 - 326.

[12] ZHANG G Q, HU M Y, EDDY PATUWO B, *et al*. Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis [J]. European Journal of Operational Research, 1999, 116(1): 16 - 32.

[13] SAPKAL A T, BOKHARE C, TARAPORE N Z. Satellite image classification using the back propagation algorithm of artificial neural network [EB/OL]. [2013-07-25]. http://pdf.aminer.org/000/347/478/a_neural_network_classifier_for_occluded_images.pdf.

[14] LIN S-W, CHOU S-Y, CHEN S-C. Irregular shapes classification by back-propagation neural networks [J]. International Journal of Advanced Manufacturing Technology, 2007, 34(11/12): 1164 - 1172.

[15] AYODELE T, ZHOU S, KHUSAINOV R. Email classification using back propagation technique [J]. International Journal of Intelligent Computing Research, 2010, 1(1/2): 3 - 9.

[16] CHUA D K H, KOG Y C, LOH P K, *et al*. Model for construction budget performance-neural network approach [J]. Journal of Construction Engineering and Management, 1997, 123(3): 214 - 222.

[17] CYBENKO G. Approximation by superpositions of a sigmoidal function [J]. Mathematics of Control, Signals and Systems, 1989, 2(4): 303 - 314.

[18] LIPPMANN R P. An introduction to computing with neural nets [J]. ASSP Magazine, 1987, 4(2): 4 - 22.

[19] LIPPMANN R P. Pattern classification using neural networks [J]. Communications Magazine, 1989, 27(11): 47 - 50.

[20] MIRCHANDANI G, CAO W. On hidden nodes for neural nets [J]. IEEE Transactions on Circuits and Systems, 1989, 36(5): 661 - 664.

[21] GORMAN R P, SEJNOWSKI T J. Analysis of hidden units in a layered network trained to classify sonar targets [J]. Neural Networks, 1988, 1(1): 75 - 89.

[22] KUNG S Y, HU Y H. A Frobenius Approximation Reduction Method (FARM) for determining optimal number of hidden units [C]// IJCNN-91: Proceedings of the 1991 Seattle International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 1991: 163 - 168.