计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(2):15−23 [doi: 10.15888/j.cnki.csa.006709]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于加权 K-Means 和局部 BPNN 的票房预测模型①

米传民[1], 鲁　月[1], 林清同[2]

[1](南京航空航天大学 经济与管理学院, 南京 211106)
[2](大叶大学 资讯管理学系, 彰化 51591)
通讯作者: 鲁　月, E-mail: 18261932278@163.com

摘　要: 电影作为典型的短周期、体验型产品, 其票房收益受众多因素的共同影响, 因此对其票房进行预测较为困难. 本文主要构建了一种基于加权 K-均值以及局部 BP 神经网络 (BPNN) 的票房预测模型对目前的票房预测模型存在的不足进行改进, 从而提高票房预测的精度: (1) 构建基于随机森林的影响因素影响力测量模型, 并以此为依据对票房影响因素进行筛选, 以此来简化后续预测模型的输入; (2) 考虑到不同影响因素对票房的影响力不同的现实情况, 为了解决以往研究中对影响因素权重平均分配的问题, 本文构建了基于加权 K-均值和局部 BP 神经网络的票房预测模型, 以因素影响力为依据对样本数据进行加权的 K-均值聚类, 并基于子样本构建局部 BP 神经网络模型进行票房预测. 实验证明, 本文所构建的模型平均绝对百分比误差 (MAPE) 为 8.49%, 低于对比实验的 10.39%, 可以看出本文构建的基于加权 K-均值以及局部 BP 神经网络的票房预测模型的预测结果要优于对比模型的预测结果.

关键词: 电影票房; 预测; 加权 K-均值; BP 神经网络

引用格式: 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.计算机系统应用,2019,28(2):15–23. http://www.c-s-a.org.cn/1003-3254/6709.html

Box-Office Forecasting Model Based on Weighted K-Means Clustering and Local BPNN

MI Chuan-Min[1], LU Yue[1], LIN Ching-Torng[2]

[1](College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)
[2](Department of Information Management, Da-Yeh University, Zhanghua 51591, China)

Abstract: As a typical short cycle and experiential product, Movie's box-office is influenced by many factors, so it is hard to forecast its box-office accurately. In this study, a box-office forecasting model based on weighted K-means and local BP Neural Network (BPNN) is constructed, with aims to improve the shortcomings of the current model and improve the accuracy of box office prediction: (1) Construct the factor influence measurement model based on Random Forest (RF) and simplify the box-office influence factors according to the value of variable importance, to achieve the purpose of simplifying the input of the following forecasting model. (2) In the traditional researches, the weight of each factor was equally allocated in sample classification, which without considering the question of different factor has different influence. So a box-office forecasting model based on weighted K-means and local BPNN is constructed, using weighted K-means clustering to classify the samples based on the value of factor influence, then build several local BPNN models based on each subsample. Experiments show that the Mean Absolute Percentage Error (MAPE) of this study's model is 8.49%, which is lower than 10.39% of the contrast experiment, which proves the superiority of the box-office forecasting model built in this study.

① 基金项目: 国家社会科学基金 (17BGL055)
Foundation item: Science Fund of Social Science (17BGL055)
收稿时间: 2018-06-18; 修改时间: 2018-07-12; 采用时间: 2018-07-19; csa 在线出版时间: 2019-01-28

Key words: box-office; forecast; weighted K-means clustering; BP Neural Network (BPNN)

电影作为很典型的短周期体验型产品, 其票房收益受到很多因素的共同影响且其影响机制较为复杂, 因此对其票房进行预测是较为困难的. 据统计, 目前我国国产电影目前只有少数电影投资是盈利的, 大部分国产电影基本都难以回收成本的. 在这一背景下, 对电影票房进行预测无疑对风险控制、充分调动投资者的积极性以及扭转目前的发展局势具有巨大的现实意义. 本文主要构建一种基于加权 K-均值以及局部 BP 神经网络的票房预测模型对目前的票房预测模型存在的不足进行改进, 从而提高票房预测的精度.

目前关于票房的研究主要分为两个研究方向: 票房影响因素的研究以及票房预测模型构建方面的研究. 传统的票房影响因素研究主要是针对票房静态影响因素的研究, 这些因素在电影上映之前就已经确定且不会随着时间的变化而变化. 聂鸿迪等人[1]选取档期、电影类型以及主创阵容等因素进行研究. 罗晓芃等人[2]添加续集这一因素探究其对票房的影响. 郑坚等人[3]将演员、导演、地区、类型等量化成连续数值来提高预测准确度. 韩明忠[4]、刘涛[5]也做了类似的研究. 除此之外, 随着互联网的兴起, 在线评论、网络搜索等动态影响因素借助于网络的放大效应, 逐渐成为了票房的重要影响因素, 因此, 越来越多的研究者将这些动态因素加入到票房预测模中: 王炼等人[6]引入网络搜索量进行研究. 郝媛媛[7]、丘萍等人[8]通过对在线电影平台网络口碑数据进行分析得出网络口碑对票房收益有显著影响. Lee JH[9]等人引入熵的概念来衡量评论整体的可信度对票房的影响. 袁海霞[10]引入信息熵对网络口碑跨平台分布特征进行量化验证其与产品销量之间的关系.

票房预测模型构建方面的研究主要涉及预测方法、样本处理、模型构建过程等方面. 票房预测中应用较多的预测方法主要有线性回归以及机器学习等方法: 李特曼、斯格特·苏凯的模型都是经典的线性回归模型[1]. 部分学者研究了线性回归以及机器学习方法哪种方法更适用于票房预测: 聂鸿迪[1]、Du J[11]、Hur M 等人[12]主要运用线性回归与 SVM、ANN、CART、SVR 等方法对票房进行预测, 得出机器学习优于线性回归的结论, 表明机器学习方法更适用于电影这种短周期体验型产品的预测. Kim T[13]等人将三种机器学习方法得到的结果进行平均, 结果优于单一的机器学习方法. 韩忠明等人[4]对特征与电影票房建立 GBRT 模型, 对票房进行预测. 因此, 目前进行票房预测的首选方法主要是机器学习方法: 魏明强[14]利用神经网络方法分析了网络评价在不同时段对票房走势的影响. 刘涛[5]分别采用 SVM 以及 ANN 对票房进行分类预测, 结果证明 ANN 的预测效果优于 SVM. 因此目前大部分学者对票房进行预测时都会选择神经网络相关方法, 其中最为常用的是 BP 神经网络: 郑坚[3]、 Zhang L[15]分别构建了基于多层 BP 神经网络的票房预测模型对票房进行预测. 除此之外还有部分学者对预测模型构建过程的其他方面进行改进: Hur M[12]考虑到电影上映的不同时段影响票房的因素侧重点会有所变化, 分别构建了六个票房预测模型来提高预测的准确度. 李金芝[16]在构建票房预测模型中应用灵敏度分析确定各参数对模型输出结果的影响力大小, 对输入变量进行筛选.

通过对电影票房预测相关研究的总结可以得出, 在票房影响因素方面虽然目前很多已经将网络口碑相关信息加入到了预测模型中, 但大部分研究仅仅考虑了单一平台, 并没有深入考虑到网络口碑的跨平台分布特征, 并且针对单一平台的网络口碑影响力研究并不能很全面的反映网络口碑对票房的影响; 在预测模型构建方面, 目前大多数学者都选择基于神经网络的预测方法, 另外还有一些学者对票房预测模型的构建过程进行优化, 但大部分研究者都用整体样本对模型进行训练. 在此情况下, 很难有一个预测模型能对如此复杂现实票房进行很好的拟合. 因此有的研究者在对模型进行训练之前对样本数据进行分类, 但目前的主要是应用简单的 K-均值聚类, 在聚类过程中不同的影响因素被赋予同等的权重, 而实际情况中, 不同的影响因素影响力是不同的, 因此简单的 K-均值聚类虽然在一定程度上提高了训练集的质量, 但是由于没有考虑到不同因素的影响力问题, 会在一定程度上影响最终预测结果.

基于上述的问题, 本文构建了一种基于加权 K-均

值聚类和局部 BP 神经网络的票房预测模型: ① 构建基于随机森林的影响因素影响力测量模型, 并以此为依据对票房影响因素进行筛选, 以此来简化后续预测模型的输入; ② 考虑到不同影响因素对票房的影响力不同的现实情况, 为了解决以往研究中对影响因素权重平均分配的问题, 构建了基于加权 K-均值和局部 BP 神经网络的票房预测模型, 以因素影响力为依据对样本数据进行加权的 K-均值聚类, 并基于子样本构建局部 BP 神经网络模型进行票房预测.

1 理论方法

1.1 随机森林

随机森林 (Random Forest, RF)[17]是一种由多个独立的决策树组合而成的集成分类器. 其决策原理可以描述为[18]: 若干个专家聚集在一起对某个特定的任务进行分析并根据自身“经验”给出自己认定的正确结果, 最后随机森林通过专家投票的方法, 采用“少数服从多数”的原则得出最后分类结果. 其生成过程主要可以分为以下几个步骤:

Step 1. 通过 Bootstrap 方法从整体的训练集数据中随机抽取, 生成k个子样本集, 以及k个袋外数据;

Step 2. 根据随机抽取生成的k个子样本集, 依据构建决策树的原理及方法选择合适的节点分裂算法来构建k棵相互独立的决策树;

Step 3. 将 Step 2 中生成的k棵决策树进行集成, 构建随机森林集成分类器;

Step 4. 将测试集输入到随机森林分类器中, 利用 Step 3 构建的随机森林分类器对其进行分类.

1.2 加权 K-均值聚类

K-均值算法是一种很有代表性的基于距离的聚类方法, 它将距离作为评价样本之间相似性的依据, 即越近的两个对象其类似度越大. 假设有n个样本且每个样本包含m个属性, 形成了一个包含n个m维数据点的样本数据集, 则聚类过程主要可以概括为以下几个步骤:

Step 1. 选取k个样本点作为初始聚类中心 (质心);

Step 2. 计算每个样本与各质心的距离, 并将其指派到距离最近的质心, 完成一次迭代;

Step 3. 对每个分组内的质心进行更新;

Step 4. 判断是否满足算法终止条件 (质心不变/距离平方和最小): 若满足则聚类完成; 否则, 重复 Step 2~Step 3 直到满足终止条件.

在上述 K-均值聚类算法中, 样本的每个属性被赋予了同等权重$1/m$, 若对不同属性赋予不同的权重, 即加权 K-均值聚类. 简单来讲, 加权 K-均值聚类在计算样本点到质心的距离时, 用各个属性对应的权重替代原来的等权权重$1/m$, 加权 K-均值聚类算法中第i个样本点到质心的距离计算公式为公式 (1)[19]:

$$ED_{i*}(w)=\sqrt{w_1^2(x_i(1)-x*(1))^2+\cdots+w_m^2(x_i(m)-x*(m))^2} \tag{1}$$

1.3 BP 神经网络

在多种神经网络模型中, 多层前向神经网络由于其成熟的算法, 较强的非线性映射能力、泛化能力以及容错能力成为了应用最为广泛一类神经网络模型, 其中最为典型的算法为误差反向传播算法——BP (Back-Propagation) 算法, BP 算法对应的模型即为 BP 神经网络模型, BP 神经网络是一种典型的信号单向传播的多层前向神经网络. BP 神经网络的训练过程主要包括两个部分: 信号的正向传播、误差的反向传播. 在正向传播过程中, 信号由输入层经过隐含层到输出层生成输出结果与期望输出进行对比, 若结果不理想则启用误差的反向传播过程, 误差信息将由输出端开始逐层进行反向传播从而对网络中的权值进行调节, 从而使得信号正向传播过程中得到的输出结果更接近理想输出.

2 基于随机森林的重要票房影响因素筛选

2.1 影响因素量化

2.1.1 电影类型

结合较为权威的电影类型分类以及我国国产电影类型的发展现状在本文的电影类型中主要包含剧情、爱情、喜剧、动作、惊悚、奇幻、悬疑其中类型. 在对类型变量进行量化时, 主要借助于各个类型的历史票房数据对其影响力进行衡量, 其求解公式如下:

$$G_i=\frac{\sum_{j=1}^{N_{gi}}Box_j}{N_{gi}} \tag{2}$$

其中, G_i表示第i个电影类型的影响力, N_{gi}代表的是在所收集的样本中属于第i个类型的电影数量, Box_j表示第j个属于第i个类型的电影的票房. 本文主要考虑电影的第一类型和第二类型.

2.1.2 演员

考虑到名品演员影响力的持久性以及人气偶像演员的瞬时性, 本文在对演员影响力进行量化时主要从两个方面进行, 一方面从演员的历史参演电影的平均票房入手衡量其持久影响力, 另一方面借助于百度搜索这一平台提取电影上映时相关演员的平均搜索量——网络搜索量 (network search volume) 来衡量其瞬时影响力. 其求解公式为:

$$Act_i = \alpha * \overline{Box_i} + \beta * \overline{NSV_i} \tag{3}$$

其中, Act_i表示第i个演员的影响力, α和β表示历史票房以及网络搜索量的重要性系数, $\overline{Box_i}$表示该演员近期内作为主演参演电影的平均票房, $\overline{NSV_i}$表示在电影上映时该演员的平均网络搜索量. 一般情况下一部电影会有很多个演员参演, 在此我们只考虑第一主演和第二主演.

2.1.3 导演

在对导演影响力进行量化时不仅要考虑到其作为导演身份的影响力还要考虑到其本身具有的其他身份的影响力, 本文主要通过该导演作为导演参与的电影票房以及作为演员参与的电影票房、其他身份的影响力主要通过网络搜索量来衡量, 因此导演影响力的求解公式为:

$$Dir_i = \alpha * \overline{Box_i} + \beta * \overline{NSV_i} \tag{4}$$

其中, Dir_i表示第i个导演的影响力, α和β表示历史票房以及网络搜索量的重要性系数, $\overline{Box_i}$表示其作为导演以及其作为主演参演电影的平均票房, $\overline{NSV_i}$表示在电影上映时该导演的平均网络搜索量.

2.1.4 档期

本文在对前人对电影档期研究做了充分总结的基础之上, 最终将电影档期分为以下几种: 贺岁档 (前一年的 11 月底至下一年的二月底)、五一档 (每一年的 4 月底到 5.3)、暑期档 (每一年的 6 月初到 8.31)、十一档 (每一年的 9 月底到 10.7). 本文在对档期变量进行量化时, 借助于往年各个档期的票房数据对档期影响力进行衡量, 其求解公式如下:

$$D_i = \frac{\sum_{j=1}^{N_i} Box_j}{N_i} \tag{5}$$

其中, D_i表示第i个档期的影响力, N_i代表的是第i个档期所包含的天数, Box_j表示在第i个档期内的第j天所有电影所产生的总票房.

2.1.5 网络搜索量

一部电影在上映期间对应的网络搜索量从一个侧面反映了潜在观影者对其的关注度, 虽然不同的潜在观影者会在搜索之后做出不同的观影决策, 但是从另一个层面来讲, 越多的人关注就表明可能有更多的潜在观影者会选择去观看这部电影, 因此本文将网络搜索量作为一个潜在观影者对电影的关注度的衡量指标, 由于百度是目前国内用户基础最大的搜索引擎, 其搜索数据具有较强的代表性, 因此本文变量网络搜索量$Search_i$具体量化数据来自百度搜索指数.

2.1.6 网络口碑数量与效价

考虑到现实情况中一般潜在观影者不会在单一平台搜集信息之后就马上作出观影决策, 而是通过多个平台搜索之后经过对比衡量之后最后才作出观影决策, 所以本文在对网络口碑数量以及效价进行量化时采用多平台评论数量求平均值的方法, 并且考虑到不同平台之间的用户基数以及评分机制的不同, 本文在对口碑数量以及口碑效价进行平均之前, 首先对其进行归一化, 最终得到网络口碑数量变量值$Amount_i$以及网络口碑效价的量化结果$Rant_i$.

2.1.7 网络口碑离散度

网络口碑离散度指的是网络口碑在不同平台之间的传播程度, 即: 网络口碑的跨平台分布特征. 为了更为全面的对网络口碑的跨平台分布特征进行量化, 本文从口碑数量和口碑效价两个方面进行探究: 引入信息熵 (information entropy) 这一概念, 构造数量信息熵 (IE_Vol_i) 以及效价信息熵 (IE_Val_i) 对口碑离散度进行量化. 信息熵是信息论中用于测算所有可能发生情况的平均不确定性的指标, 信息熵越大, 说明整体系统越混乱, 即各个事件发生的概率分布越平均. 本文在对网络口碑离散度进行量化时主要思路是将信息熵求解公式中的事件发生的概率替换为网络口碑各个特征值, 并通过公式 (6) 和公式 (7) 进行求解:

$$IE\_Vol_i = -\sum_{Total\_Vol_i > 0} \frac{Vol_i^j}{Total\_Vol_i} \log\left(\frac{Vol_i^j}{Total\_Vol_i}\right), \tag{6}$$

$$IE\_Val_i = -\sum_{Total\_Val_i > 0} \frac{Val_i^j}{Total\_Val_i} \log\left(\frac{Val_i^j}{Total\_Val_i}\right), \tag{7}$$

其中, j代表第j个电影网络口碑平台, $Total_Vol_i$代表第i部电影在各个平台的评论数的总和, $Total_Val_i$代表第i部电影在各个平台的总评分的总和. Vol_i^j代表第i部电影在第j个电影网络口碑平台的网络口碑数量特征值, Val_i^j代表第i部电影在第j个电影网络口碑平台的网络口碑效价特征值.

2.2 基于随机森林的因素影响力判定和指标筛选

2.2.1 基于重要性分数的因素影响力

利用随机森林算法对变量重要性进行判定时主要采用变量重要性分数 (variable importance score), 其主要作用是对各个条件属性对于决策属性的影响程度进行衡量. 本文主要采用基于置换的变量重要性分数. 将整体训练样本集的集合设为 D, 并且将用向量$X_j, j=\{1,2,\cdots,11\}$表示影响电影票房的因素, 对整体训练样本采用 Bootstrap 抽样生成 K 个子训练样本集, 则第k个样本子集则表示为D_k, 则变量重要性分数则表示为向量$VIS=\{VIS_1, VIS_2,\cdots, VIS_j,\cdots, VIS_{11}\}$, 则通过变量重要性分数对票房影响因素进行衡量可以总结为以下几个步骤:

Step 1. 首先将k值取 1;

Step 2. 并在其对应的子训练集D_k的基础上构建决策树T_k, 同时将对应的袋外数据用D_k^{oob}表示;

Step 3. 应用 Step 2 中生成的决策树Tk对对应的袋外数据D_k^{oob}进行分类, 并计算其分类准确率R_k^{oob};

Step 4. 对于变量$X_j, j=\{1,2,\cdots,11\}$, 对其变量值进行变换直至其原始袋外数据D_k^{oob}样本自变量与因变量之间的关系被打断, 并将针对该变量扰动之后的袋外数据用D_{kj}^{oob}表示;

Step 5. 应用 Step 2 中生成的决策树T_k对扰动后的袋外数据D_{kj}^{oob}进行分类, 并计算其分类准确率R_{kj}^{oob};

Step 6. 分别另$k=1,2,\cdots,K$, 对其重复进行 Step 2~ Step 5 的操作, 得出各个子训练集对应下的扰动前后的分类正确率;

Step 7. 通过公式计算特征X_j的变量重要性分数, 其求解公式为式 (8):

$$VIS_j=\frac{1}{K}\sum_{k=1}^{k}(R_k^{oob}-R_{kj}^{oob}) \tag{8}$$

Step 8. 对$j=\{1,2,\cdots,11\}$重复上述过程, 得出所有变量重要性分数, 输出重要性分数向量$VIS=\{VIS_1, VIS_2,\cdots, VIS_j,\cdots, VIS_{11}\}$.

通过对样本数据集进行上述操作得到票房影响因素的重要性分数, 可以看出, 当对一个变量的对应值进行变换前后分类准确率减少量越大, 表明这一变量重要程度越强, 反之则表明该变量不是很重要, 因此对其变量值进行扰动不会对最终分类结果造成影响.

2.2.2 票房影响因素筛选

通过构造随机森林并通过随机森林的重要性分数对影响电影票房的各个影响因素的重要性进行衡量, 并以各个变量的重要性分数作为其对票房重要性的依据, 从而对各个影响因素的重要性进行比较, 进行指标筛选, 从中选出重要性较高的票房影响因素用于后续票房预测任务. 但是由于随机森林的特性, 当依据样本数据对票房影响因素的重要性分数进行求解时, 同样的数据在多次试验中得出的各个因素的重要性分数是不同的, 但是观察多次试验的结果可以看出, 每个影响因素的重要性分数的值都在一定的范围内波动, 因此本文在对因素重要性进行衡量时, 采取多次试验求平均值的方法.

3 基于加权 K-均值和局部 BP 神经网络的票房预测

3.1 基于加权 K-均值的训练数据分类模型

简化后的指标体系中各个票房影响因素的个数为n, j代表第j个影响因素, 则w_j则表示第j个影响因素的权重, 最佳聚类数用k表示, 另外i表示第i个电影样本数据. 则加权 K-均值聚类算法中第i个样本点到质心的加权欧式距离ED_i*计算公式为式 (9):

$$ED_{i*}=\sqrt{w_1^2(x_i(1)-x*(1))^2+\cdots+w_n^2(x_i(n)-x*(n))^2} \tag{9}$$

基于加权 K-均值的样本分类可以分为以下步骤:

Step 1. 随机选取k个样本点作为初始聚类中心 (质心);

Step 2. 依据式 (9) 计算其余每个样本与各个质心的加权欧式距离, 并将其指派到距离最近的质心, 完成一次迭代;

Step 3. 对每个分组内的质心进行更新;

Step 4. 判断是否满足算法终止条件: 满足的话, 聚类完成; 否则, 重复 Step 2~Step 3 直到满足终止条件, 完成聚类.

通过对样本数据进行加权 K-均值聚类, 对不同影响因素赋予不同的权重, 弥补了一般 K-均值聚类中各

因素权重平均分配忽略不同影响因素影响力之间差异的问题, 因此, 在考虑到不同影响因素对电影票房影响力的差异的基础上对样本数据进行分类可以使得最终的分类结果更为科学.

3.2 基于 BP 神经网络的票房预测模型

3.2.1 BP 神经网络结构设计

BP 神经网络的结构设计主要包含网络层数确定、输入层和输出层设计以及隐含层设计三个方面: 根据 Kosmogorov 定理, 在合理的条件下, 一个三层 BP 神经网络可以拟合出任意复杂的连续函数. 因此本文所构建的 BP 神经网络为三层神经网络 (如图 1); 输入层以及输出层所包含的节点数主要由数据本身特征所决定, 输入层的节点数为自变量的数目, 输出层的节点数为目标因变量的数目. 因此本文所构建的 BP 神经网络预测模型中, 输入层节点数目为简化后对应的影响电影票房的因素的个数. 输出层节点只有一个, 代表票房变量; 隐含层设计的主要是确定隐含层所包含神经元的数目, 其确定公式为公式 (10), 其中nh代表隐含层神经元的数目, ni表示输入层神经元的数目, no表示输出层神经元的数目, a为认为设定的可变常数并且$a \in [1,10]$.

$$nh = \sqrt{n_i + n_o} + a \tag{10}$$

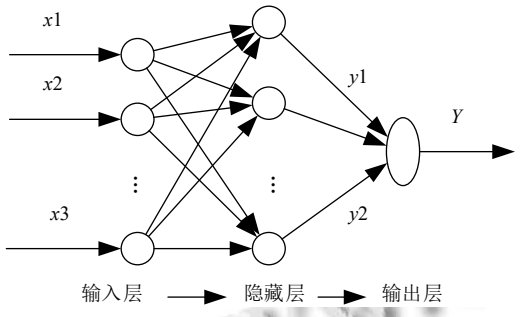

图 1 三层 BP 神经网络结构图

3.2.2 BP 神经网络参数选取

BP 神经网络的参数选取主要包含初始权值及阈值选取、学习速率的选取、激活函数以及学习函数的选择三个方面: 在对初始权值以及阈值进行确定时, 本文选择采用随机生成初始权值及阈值的方法; 学习速率η的值通过 BP 神经网络在训练过程中权值的修正量来影响神经网络的学习过程. 通过对相关理论以及文献的学习以及总结, 常用的学习速率的取值范围在 0.01 到 0.8 之间. 常用的激活函数有单/双极性 Sigmoid 函数、正弦函数等. 本文在进行 BP 神经网络建模时选择单极性 Sigmoid 函数, 其数学表达式如公式 (11):

$$f(x) = \frac{1}{1+e^{-x}}, (x \in (0,1]) \tag{11}$$

目前常用的学习函数有: 动量 BP 算法、拟牛顿法及 L-M 算法等等. 同时 L-M 算法由于其具有较高的学习速率以及较快的收敛速度最为常用, 因此本文在进行 BP 神经网络建模时也选择 L-M 算法作为学习函数.

3.2.3 BP 神经网络模型构建

通过前文的 BP 神经网络结构设计以及 BP 神经网络的主要参数选取, 确定了本文 BP 神经网络模型的基本结构, 在对本文 BP 神经网路进行建模以及训练时主要流程以及思路如图 2 所示.

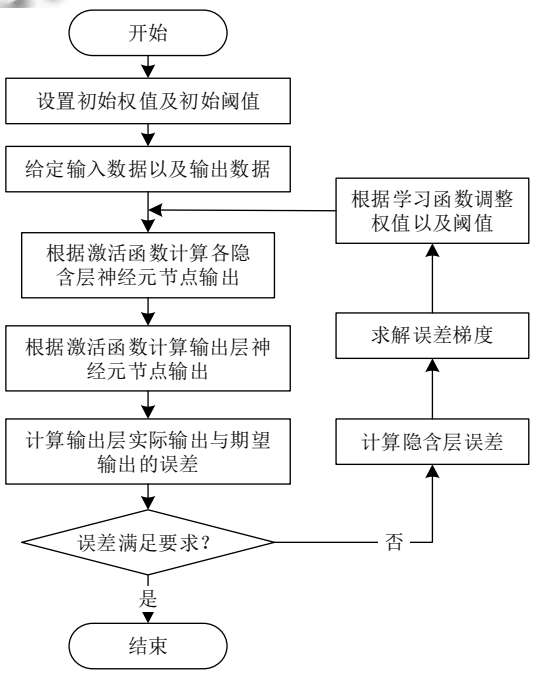

图 2 BP 神经网络模型流程图

3.3 基于局部 BP 神经网络的票房预测模型构建

基于加权 K-均值聚类的局部 BP 神经网络票房预测模型的主要思路为: 通过加权 K-均值聚类将原始样本数据分为若干个样本子集, 并基于各个样本子集构建对应的局部 BP 神经网络票房预测模型, 并且对新的电影数据进行票房预测时, 通过判断其与各个样本子集的聚类中心的加权欧式距离来决定调用哪一个局部 BP 神经网络对其进行预测, 并在这一过程中加入判断条件, 来决定是否要将新数据加入样本子集中; 另外随着新数据的加入, 整体样本的分类效果可能在某一时刻不再是最佳分类, 所以在过程中加入了整体数据分

类效果的判定, 决定是否需要对整体样本数据重新进行分类. 具体可以分为以下几个步骤 (如图 3 所示).

Step 1. 初始化参数: 加权欧氏距离临界值ED;

Step 2. 对数据集内的所有数据进行加权 K-均值聚类, 得到若干个样本子集以及各样本子集的聚类中心;

Step 3. 对这若干个样本子集构建对应的局部 BP 神经网络票房预测模型, 使得样本子集、样本子集聚类中心、局部 BP 神经网络预测模型一一对应;

Step 4. 输入待预测数据, 计算其与各个样本子集聚类中心的加权欧氏距离, 并选择距离最小的对应局部 BP 神经网络模型对其进行预测, 得到预测结果;

Step 5. 判断该条数据与最近聚类中心的加权欧氏距离是否小于设定的加权欧氏距离临界值ED, 若在临界值内则将该条数据加入该样本子集, 转 Step 3, 否则舍弃该条数据, 转 Step 4.

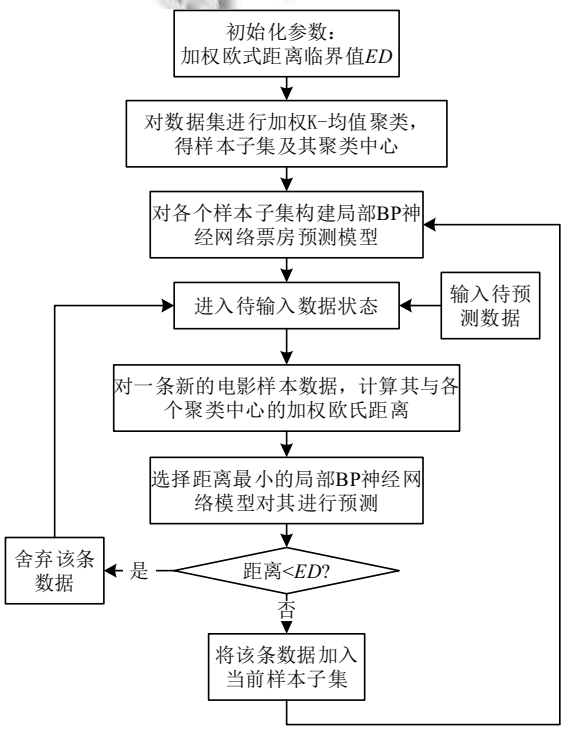

图 3 基于加权 K-means 和局部 BPNN 的票房预测流程图

4 实验验证

4.1 数据来源与量化

4.1.1 数据来源

本文样本主要包含 2016–2017 年间的电影数据, 主要来源于艺恩咨询、百度指数、豆瓣网、时光网以及猫眼电影等平台. 其中票房、类型、演员、导演、档期等数据来源于艺恩咨询. 网络搜索量相关数据来自于百度指数. 网络口碑相关信息从豆瓣网、时光网、猫眼电影收集得到. 本文收集到的原始数据共包含 415 部国产电影, 在此基础之上, 剔除数据不全、票房过低以及特殊题材的电影后用于实证分析的电影数据共有 327 部.

4.1.2 样本数据量化

在对样本数据进行量化时, 考虑到不同的变量量化之后具有不同的量级, 不同量级的数值可能会对接下来的影响因素重要性判断造成影响, 本文通过归一化数据来去除数据的不同量级对因素重要性判别的影响, 进一步归一化之后的数据描述性统计如表 1 所示.

表 1 归一化数据描述性统计分析表

变量	变量描述	最小值	最大值	均值
Box	电影票房	0.0002	1	0.0332
*G*1	第一类型	0.0106	1	0.3439
*G*2	第二类型	0.0316	1	0.4041
*Act*1	第一主演	0.0044	1	0.1388
*Act*2	第二主演	0.004	1	0.1324
Dir	导演	0.0025	1	0.0953
D	档期	0.0761	1	0.2674
Search	网络搜索量	0.002	1	0.0644
Amount	网络口碑数量	0.0006	1	0.0434
Rant	网络口碑效价	0.3677	1	0.7165
IE-Vol	口碑数量离散度	0.093	1	0.5581
IE-Val	口碑效价离散度	0.7083	1	0.875

4.2 基于随机森林的重要票房影响因素筛选

根据前文介绍的基于随机森林的票房影响因素变量重要性分数的求解过程对各个变量的重要性进行求解, 由于随机森林的算法特性导致在利用随机森林算法进行变量重要性分数求解时其结果会具有一定的波动性, 因此本文在进行实验时采用多次建模求平均值的方法对变量重要性进行判定, 最终求解结果如图 4 所示.

通过对结果的观察可以看出在所有的影响因素中, 网络搜索量的对应的重要性分数最高, 说明在影响票房的所有因素中, 这一因素发挥的作用最大, 其次是网络口碑数量、口碑数量离散度等影响因素, 另外通过对图 4 中变量重要性分数分布结果图的观察可以看出, 有部分影响因素的重要性分数很小几乎接近于零, 表明这些因素在对票房的影响方面发挥的作用很小, 相对于其他的重要性分数较大的因素其作用几乎可以忽

略不计, 这些因素包括: 口碑效价离散度、口碑效价以及第二类型, 因此为了简化后续的票房预测模型输入, 本文在进行票房影响因素的选择时只选取影响力较大的因素, 去掉一些作用很小的影响因素, 从而在输入层对预测模型进行简化. 因此, 筛选后的票房影响因素共包含网络搜索量、口碑数量、口碑数量离散度、第一主演、第一类型、第二主演、导演和档期等因素.

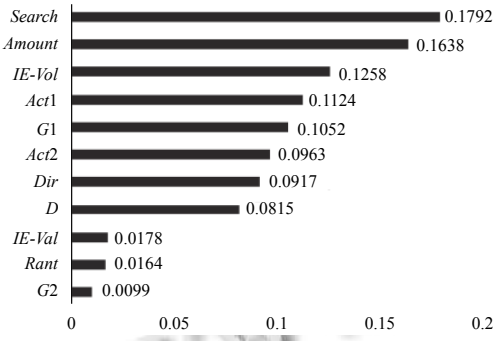

图 4 变量重要性分数结果图

4.3 基于加权 K-均值和局部 BP 神经网络的票房预测

通过对筛选后的影响因素的变量重要性分数进行归一化处理得到各个影响因素的对应权重, 影响因素及其对应权重结果如表 2 所示.

表 2 影响因素及其权重结果表

变量	重要性分数	权重
*G*1	0.1052	0.1101
*Act*1	0.1124	0.1176
*Act*2	0.0963	0.1007
Dir	0.0917	0.0959
D	0.0815	0.0853
Search	0.1792	0.1875
Amount	0.1638	0.1714
IE-Vol	0.1258	0.1316

在对最优聚类数进行确定时本文所采用的方法为: 通过对每个聚类数对应的 F 值 (组间离差平方和的平均值除以组内离差平方和的平均值) 进行比较, 当聚类数发生变化而跟其相对应 F 值不变化或者变化很小的话, 对应的聚类数即为最佳聚类数. 通过计算得出电影样本数据分类的最佳聚类数为 3, 通过加权 K-均值聚类将电影样本数据分为 3 类, 分别以三类子样本为依据构建局部 BP 神经网络模型, 本文采用 Python 编程来实现 BP 神经网络预测的功能, 其中部分参数设置如表 3 所示.

表 3 BP 神经网络参数设置

参数	数值
输入层神经元个数	8
输出层神经元个数	1
隐含层神经元个数	12
迭代次数	1000
学习率	0.05
训练精度	0.01
训练集占比	80%
测试集占比	20%

为了对本文构建模型的效果进一步进行验证, 本文同时设置了对比实验, 在对比实验中首先采用简单 K-均值聚类对样本数据进行聚类, 并在此基础上构建 BP 神经网络进行票房预测, 同样采用 Python 编程实现, 从而对本文的改进效果进行验证.

4.4 结果对比及分析

平均绝对百分比误差 (Mean Absolute Percentage Error, MAPE) 是对预测模型进行评估时常用的一种指标, 其值可以通过公式 (12) 求得, 其中Vpi表示第i个样本的票房预测值 (Predictive Value), Vai表示第i个样本的实际票房值 (Actual Value), n表示用于预测实验的样本数.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Vpi - Vai|}{Vai} * 100\% \tag{12}$$

在采用两种模型进行预测时, 由于受 BP 神经网络模型自身特征影响, 其预测结果会在一个特定范围内产生一定的波动, 因此本文在对两个模型的预测效果进行衡量时, 采用多次预测求平均值的方式, 实验结果如表 4 所示, 最后得出基于本文构建的模型进行的票房预测的平均绝对百分比误差 (MAPE) 控制在 8.49%, 对比模型平均绝对百分比误差 (MAPE) 控制在 10.39%. 可以看出本文构建的基于加权 K-均值以及局部 BP 神经网络的票房预测模型的预测结果要优于对比模型的预测结果, 从而证明了本文所构建的票房预测效果.

表 4 两模型预测效果对比表 (%)

实验次数	平均绝对百分比误差 (*MAPE*)	
	本文模型	对比模型
1	9.61	11.82
2	7.18	9.21
3	9.86	9.36
4	8.27	10.26
5	7.52	11.28
AVG_MAPE	8.49	10.39

5 总结与展望

电影作为很典型的短周期体验型产品, 其票房收益受到很多因素的共同影响且其影响机制较为复杂, 因此对其票房进行预测是较为困难的. 本文在对电影票房预测研究进行了较为全面的总结与分析的基础上, 对电影票房预测建模过程进行了一定的优化与改进, 构建了基于加权 K-均值聚类以及局部 BP 神经网络的票房预测模型, 本文的研究可以总结为以下几个方面:

(1) 构建基于随机森林的影响因素影响力测量模型, 并以此为依据对票房影响因素进行筛选, 以此来简化后续预测模型的输入; (2) 考虑到不同影响因素对票房的影响力不同的现实情况, 为了解决以往研究中对影响因素权重平均分配的问题, 本文构建了基于加权 K-均值和局部 BP 神经网络的票房预测模型, 以因素影响力为依据对样本数据进行加权的 K-均值聚类, 并基于子样本构建局部 BP 神经网络模型进行票房预测. 同时通过实际电影数据实验可以看出, 本文构建的基于加权 K-均值聚类以及局部 BP 神经网络的票房预测模型可以减小票房预测误差, 提高预测的准确度.

本文应用随机森林进行影响力测算以及采用加权 K-均值聚类对数据进行聚类, 并采用 BP 神经网络模型进行票房预测. 在后续的研究中, 需要进一步对 BP 神经网络模型的构建过程进行优化, 并对其中一些参数的选择以及设置方法进行改进, 进一步提高整体票房预测模型的精确度.

参考文献

1 聂鸿迪. 中国电影票房的影响因素及其实证研究[硕士学位论文]. 北京: 北京交通大学, 2015.

2 罗晓芃, 齐佳音, 田春华. 电影首映日后票房预测模型研究. 统计与信息论坛, 2016, 31(11): 94–102. [doi: 10.3969/j.issn.1007-3116.2016.11.016]

3 郑坚, 周尚波. 基于神经网络的电影票房预测建模. 计算机应用, 2014, 34(3): 742–748.

4 韩忠明, 原碧鸿, 陈炎, 等. 一个有效的基于 GBRT 的早期电影票房预测模型. 计算机应用研究, 2018, 35(2): 410–416. [doi: 10.3969/j.issn.1001-3695.2018.02.020]

5 刘涛. 面向社交媒体的电影票房预测技术的研究与应用[硕士学位论文]. 石家庄: 河北科技大学, 2016.

6 王炼, 贾建民. 基于网络搜索的票房预测模型——来自中国电影市场的证据. 系统工程理论与实践, 2014, 34(12): 3079–3090. [doi: 10.12011/1000-6788(2014)12-3079]

7 郝媛媛, 邹鹏, 李一军, 等. 基于电影面板数据的在线评论情感倾向对销售收入影响的实证研究. 管理评论, 2009, 21(10): 95–103.

8 丘萍, 张鹏. 第三方网络口碑对短生命周期产品销量的影响研究. 河海大学学报 (哲学社会科学版), 2017, 19(2): 39–46.

9 Lee JH, Jung SH, Park JH. The role of entropy of review text sentiments on online WOM and movie box office sales. Electronic Commerce Research and Applications, 2017, 22: 42–52. [doi: 10.1016/j.elerap.2017.03.001]

10 袁海霞. 网络口碑的跨平台分布与在线销售——基于 BP 人工神经网络的信息熵与网络意见领袖敏感性分析. 经济管理, 2015, 37(10): 86–95. [doi: 10.3969/j.issn.1007-5097.2015.10.013]

11 Du JF, Xu H, Huang XQ. Box office prediction based on microblog. Expert Systems with Applications, 2014, 41(4): 1680–1689. [doi: 10.1016/j.eswa.2013.08.065]

12 Hur M, Kang P, Cho S. Box-office forecasting based on sentiments of movie reviews and independent subspace method. Information Sciences, 2016, 372: 608–624. [doi: 10.1016/j.ins.2016.08.027]

13 Kim T, Hong J, Kang P. Box office forecasting using machine learning algorithms based on SNS data. International Journal of Forecasting, 2015, 31(2): 364–390. [doi: 10.1016/j.ijforecast.2014.05.006]

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

15 Zhang L, Luo JH, Yang SY. Forecasting box office revenue of movies with BP neural network. Expert Systems with Applications, 2009, 36(3): 6580–6587. [doi: 10.1016/j.eswa.2008.07.064]

16 李金芝. 基于泛函网络的票房预测研究与应用[硕士学位论文]. 重庆: 重庆大学, 2015.

17 姚登举. 面向医学数据的随机森林特征选择及分类方法研究[博士学位论文]. 哈尔滨: 哈尔滨工程大学, 2016.

18 曹正凤. 随机森林算法优化研究[博士学位论文]. 北京: 首都经济贸易大学, 2014.

19 陈小雪, 尉永清, 任敏, 等. 基于萤火虫优化的加权 K-means 算法. 计算机应用研究, 2018, 35(2): 466–470. [doi: 10.3969/j.issn.1001-3695.2018.02.031]