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$a \in [1,10]$.

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

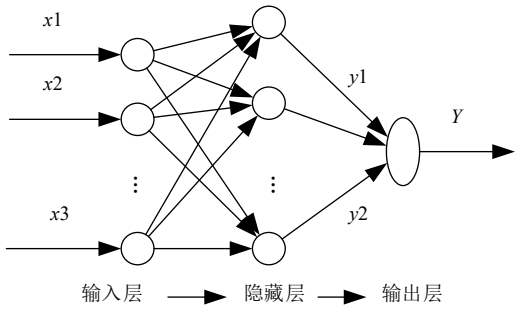

图 1 三层 BP 神经网络结构图

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.

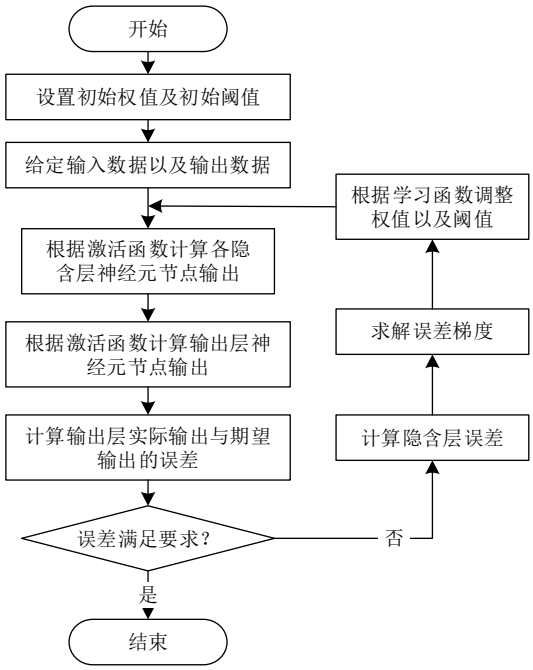

图 2 BP 神经网络模型流程图

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.

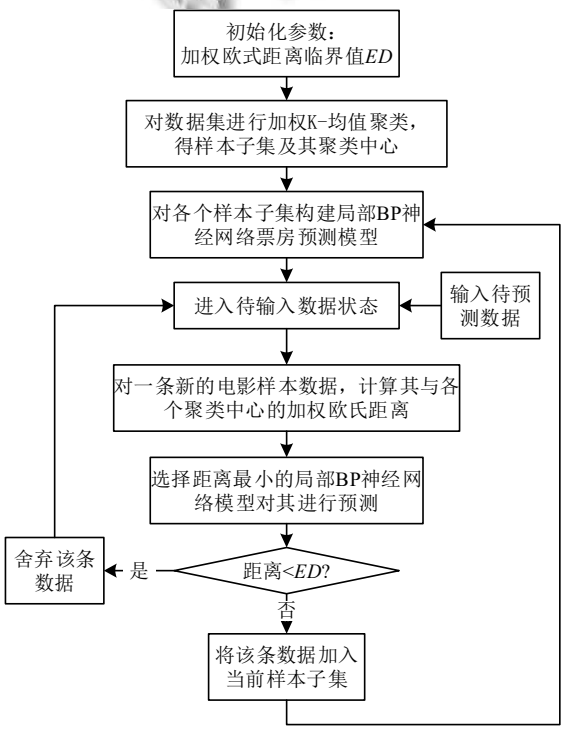

图 3 基于加权 K-means 和局部 BPNN 的票房预测流程图

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.

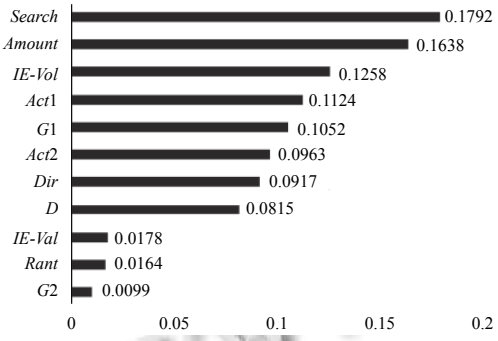

图 4 变量重要性分数结果图

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]