doi: 10.3969/j.issn.1007-7375.2020.04.010

结合区间理论熵权和TOPSIS的映前总票房区间预测

唐中君，周亚丽

（北京工业大学 经济与管理学院，北京 100124）

摘要: 为解决电影票房预测中，票房影响因素存在模糊性的问题，提出用区间数量化具有模糊性的票房影响因素，使得量化科学合理，量化结果能提高信息利用率；利用熵权法计算各影响因素权重，保证权重客观性；对票房进行分级并基于票房分级，将票房影响因素作为票房评价指标，提出用理想解法计算票房理想解贴近度区间，根据票房理想解贴近度判断票房级别，得到票房区间预测值；结合区间理论、熵权法和理想解法得到上映前总票房区间预测方法。选取2015~2017年上映的剧情类和动作类电影对该方法进行了验证。该方法在两类用于验证的电影中的平均准确率分别为79.33%和73.92%，证明本文方法具有一定的有效性和实用性。本文对于电影发行商的决策以及风险规避具有一定的意义，且对于与电影类似的短生命周期体验品的早期需求预测具有一定的参考价值。

关键词: 映前票房预测；区间数；熵权法；TOPSIS法；区间预测

中图分类号: F272.3　　**文献标志码:** A　　**文章编号:** 1007-7375(2020)04-0075-09

An Interval Prediction Method for Box Office Before Released Based on Interval Theory, Entropy Weight, and TOPSIS

TANG Zhongjun, ZHOU Yali

(School of Economics and Management, Beijing University of Technology, Beijing 100124, China)

Abstract: In order to address ambiguity of influence factors in box office prediction, interval number is used to quantify the factors, because interval number is of high utilization of information, scientific and reasonable. In order to ensure objectivity, entropy method is applied to determine weights of each factor. TOPSIS method is used to calculate box office ideal solution nearness degree interval by treating box office influence factors as box office evaluation indicators, and by dividing box office into different categories based on classification of box office. The box office level is determined based on the box office ideal solution nearness degree, and then the interval prediction value of box office is obtained. Combining interval theory, entropy weight method, and TOPSIS method, an interval prediction method for box office before released is obtained. Some action and drama movies released between 2015−2017 are used to verify the proposed method. Average accuracy of the method is 79.33% and 73.92% respectively for action and drama movies, suggesting that the method is effective and practical. The method is valuable for film distributors′ decision-making and risk aversion, and has certain reference value for early demand forecast of short life-cycle experience products similar to movies.

Key words: box office prediction before released; interval number; entropy weight method; TOPSIS; interval prediction

我国电影产业不断繁荣发展，2019年全国总票房达642.66亿元，银幕总数超越北美，成为全球最大的电影市场。虽然我国电影产业不断发展壮大，规模增长迅速，但是不同电影的投资回报存在较大的波动性。2019上半年国产电影前50部中仅有13部盈利；引进片中，《X战警：黑凤凰》等收益情况也不容乐观。因此，电影票房预测有助于制定有效决策和规避风险。

对电影票房的预测可分为上映前和上映后的预测。电影的生命周期极短，上映前的预测显得尤为

收稿日期： 2019-06-04

基金项目： 国家自然科学基金资助项目(71672004)

作者简介： 唐中君(1969-)，男，湖南省人，研究员，博士，主要研究方向为需求预测、运营与营销.

重要[1]。上映前可获得的票房影响因素有电影类型、导演票房影响力(以下简称导演影响力)、演员票房影响力(以下简称演员影响力)[1-6]、上映期票房影响力(以下简称上映期影响力)[3-6]、影片预告片播放量[6]、影片时长[7]等。导演影响力、演员影响力、上映期影响力等因素存在一定程度的模糊性。已有研究对演员影响力、导演影响力的量化方法有奖项提名数总和[5]、设置前一部电影是否成功作为虚拟变量并考虑前一部电影的票房表现[2]、搜索引擎搜索量[3]等。用上映日是否为节假日[6]或放映时间与热门档期的重合天数[5]等量化上映期影响力。这些量化方法没有考虑导演影响力、演员影响力及上映期影响力的模糊性。

对模糊性变量的量化主要有概率和非概率方法[8]。概率方法需得到模糊性变量的精确概率分布。这对于导演影响力和演员影响力等模糊性变量的量化是难以做到的[9]，然而采用非概率方法的区间数表示则是容易的。学者Young[10]提出区间数的思想，并用其解决不确定性和模糊性问题。运用区间理论时，对于模糊性变量无需知道精确值，只需给定大概范围[8]，并且在解决量化问题的同时提高信息利用率。由此可见，对于导演影响力、演员影响力等存在模糊性的变量，采用区间数量化更合理。区间理论不断发展，已广泛应用于工程分析[8, 11-12]、综合评价[13]、动态优化[14]、岩爆等级预测[15]等多个领域，但未发现将其应用于电影票房预测方面的研究。

区间数只能用于量化票房影响因素，需结合其他方法才能在上映前预测电影总票房。目前电影票房预测方法有以神经网络为代表的机器学习方法[1-4]、以线性回归为代表的回归类方法[5]、以Bass模型[16-17]为代表的扩散类方法等。这些方法与区间数结合会产生复杂的计算推演过程，但是TOPSIS法[18]则不同，其计算过程简单。该方法依据决策方案的各描述指标与理想解的距离判断决策方案的好坏[19]。方案的各指标值距正理想解越近、距负理想解越远，则该方案越优。将每部电影看作一个方案，以票房影响因素为票房评价指标，取各指标最大值和最小值为理想解，则可通过计算票房评价指标与理想解的接近程度判断电影票房。高理想解贴近度对应高票房；反之，对应低票房。不同影响因素对票房的影响程度不同，需对各因素赋权。区间数和TOPSIS法无法完成对指标的赋权，因此需要一种指标赋权方法。熵权法是一种客观赋权方法，权重结果依赖数据本身，科学合理性高[18]。电影上映前，电影口碑、在线评论、网站评分等重要信息未产生，使得上映前的票房点预测值精度难以保证。此外，电影上映前票房预测主要用于各类投资决策，区间预测值即可满足投资决策要求。不同类型的电影有不同的观众群体，具有不同的票房规律。因此，按照电影类型分别进行票房预测更具可行性。

基于以上分析，本文将构建一种按照电影类型分别收集数据的结合区间理论、熵权法和TOPSIS法的电影上映前总票房区间预测方法。该方法选取票房的重要影响因素作为票房评价指标；采用区间数量化上映期影响力、导演影响力、演员影响力等评价指标；利用熵权法对各指标赋权；通过TOPSIS法得到票房理想解贴近度，确定不同票房级别对应的理想解贴近度区间，从而得到票房区间预测值。采用2015~2017年上映的剧情类和动作类电影验证该方法的有效性。

1 预测方法的构建

1.1 预测方法概述

结合区间理论、熵权法和TOPSIS法的电影上映前总票房区间预测方法如图1所示。该方法包括票房评价指标的选取和量化、基于训练集的票房理想解贴近度区间计算、基于测试集的方法验证和待预测电影票房预测4个阶段。其中，待预测电影票房预测阶段与基于测试集的方法验证阶段流程相同，为方便起见，在图1中，将两者画在一起。图中实线矩形为数据操作过程；平行四边形代表流程输入和输出；箭头代表流程走向。图中，η为理想解贴近度；d_{ij}^{+}、d_{ij}^{-}分别为第i部电影第j个加权规范化后的票房评价指标与正理想解、负理想解的欧几里得距离。

基于训练集的票房理想解贴近度区间计算阶段，以训练集电影数据为输入，通过熵权法确定指标权重区间、计算理想解。该阶段的输出为分级票房的理想解贴近度区间。基于测试集的方法验证(待预测电影票房预测)阶段，以测试集(待预测)电影数

据为输入，根据前述指标权重区间和理想解，计算测试集(待测试)电影的票房理想解贴近度。该阶段的输出为测试集(待预测)电影的票房预测级别，即票房区间预测值。

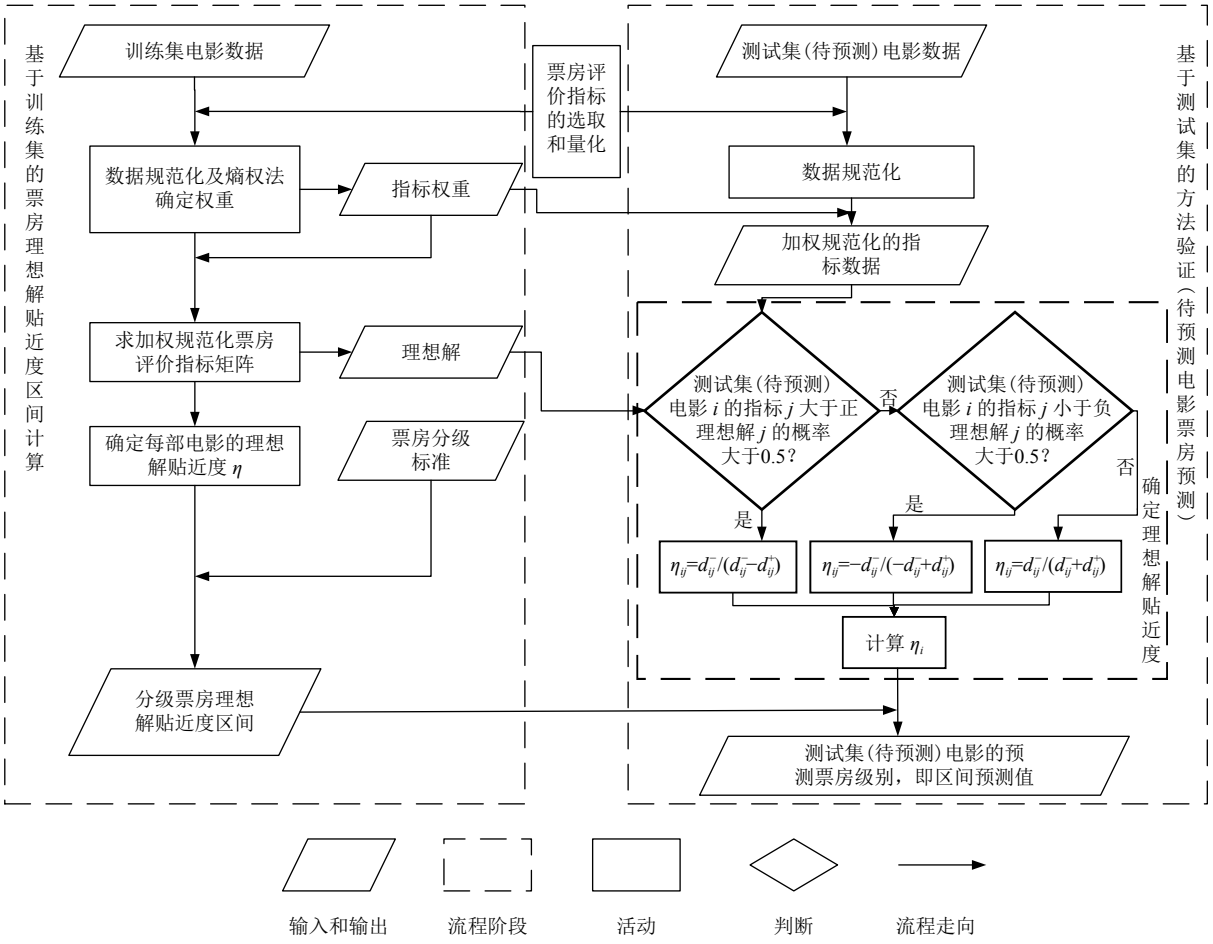

图 1　结合区间数熵权法和理想解法的电影上映前总票房预测方法

Figure 1　A prediction method for total box office before released based on interval theory, entropy weight, and TOPSIS

1.2　票房评价指标的选取及量化

一部电影的成功与否，受多方面因素影响。能否从众多因素中选取最关键的因素，关系着电影票房预测有效性。本文选择如下9个因素，建立票房评价指标体系。

1.2.1　百度指数

百度作为全球最大的中文搜索引擎，其指数是分析网民行为的重要数据。电影上映前，电影发行方通常会组织大量的营销宣传活动。百度指数能体现电影营销宣传活动的力度。基于此，选取百度指数为票房评价指标之一。为了保证数据一致性，以电影名称为搜索关键词收集电影百度指数。百度指数区间

$$\mathrm{DBD}_i = [\mathrm{DBD}_i^{\mathrm{L}}, \mathrm{DBD}_i^{\mathrm{U}}]。\tag{1}$$

其中，DBD_i代表第i部电影百度指数区间。由于遗忘效应的存在，本文只收集电影上映前7 d的百度指数。其中，百度指数以d为基本单位；$\mathrm{DBD}_i^{\mathrm{L}}$、$\mathrm{DBD}_i^{\mathrm{U}}$分别代表第$i$部电影上映前7 d百度指数的最小值和最大值。

1.2.2　微博话题关注度与微博电影视频播放量

电影上映前，发行商通常在国内主流社交媒体“新浪微博”上宣传。宣传方式包括创建电影话题，发布电影宣传片、预告片、花絮片等方式。区别于百度指数，社交媒体数据反映潜在观众对电影的关注程度，是电影票房的重要影响因素[20-21]。因此，选择微博话题关注度、微博电影视频播放量为电影票房评价指标。以上2个指标量化后的值为点数据，为便于计算，将其转化为区间数

$$DWG_i = [DWG_i^L, DWG_i^U], \tag{2}$$

$$DSP_i = [DSP_i^L, DSP_i^U]。 \tag{3}$$

其中，DWG_i代表第i部电影的微博话题关注度区间；DWG_i^L、DWG_i^U为第i部电影的微博话题关注度的最小值和最大值(人)；DSP_i代表第i部电影的微博电影视频播放量区间；DSP_i^L、DSP_i^U分别为第i部电影的微博电影视频播放量的最小值和最大值(万次)。

1.2.3 上映期影响力

已有研究证实，票房表现与上映期密切相关[5]。上映期影响力指上映期对票房的影响程度。本文运用区间数量化上映期影响力为

$$DDQ_i = [DDQ_i^L, DDQ_i^U]。 \tag{4}$$

其中，DDQ_i代表第i部电影的上映期影响力区间；DDQ_i^L、DDQ_i^U分别代表第i部电影上映期影响力区间的左端点和右端点。本文在文献[4]有关上映期影响力量化的基础上，将上映期影响力转化为区间数。具体日期的影响力区间如表1所示。

表 1 上映期影响力区间

Table 1 The influence range of release date

月份	节日	节日影响力(DDQ)	其余日期影响力(DDQ)
1月	贺岁档(正月初一至正月十五)	[8.75, 10]	[6.25, 7.5]
		[8.75, 10]	
2月	情人节(02.14)	[7.5, 8.75]	[6.25, 7.5]
3月	(03.01-03.15)	[1.25, 2.5]	[0, 1.25]
4月	劳动节(04.23-04.30)	[7.5, 8.75]	[2.5, 3.75]
5月	劳动节(05.01)	[7.5, 8.75]	[0, 1.25]
6月			[2.5, 3.75]
7月			[5, 6.25]
8月			[3.75, 5]
9月	国庆节(09.23-09.30)	[8.75, 10]	[3.75, 5]
10月	国庆节(10.01)	[8.75, 10]	[0, 1.25]
11月			[2.5, 3.75]
12月	圣诞节(12.20-12.31)	[8.75, 10]	[7.5, 8.75]

1.2.4 想看人数

电影上映前的营销宣传会增强观众的观影意向。体现潜在观众观影意向的指标有时光网、豆瓣电影网统计的想看人数。同种类型的电影在相同的网站统计想看人数数据。类似于微博话题关注度，想看人数为点数据，将其转化为区间数为

$$DXK_i = [DXK_i^L, DXK_i^U]。 \tag{5}$$

其中，DXK_i代表第i部电影的想看人数区间；DXK_i^L、DXK_i^U分别为第i部电影的想看人数的最小值和最大值。

1.2.5 导演影响力

电影导演作为一部电影的执导者，对电影票房的成功起着至关重要的作用。文献[3]研究表明，新电影最大的魅力有包括电影导演在内的超级明星。导演影响力指导演对票房的影响程度为

$$DDY_i = [DDY_i^L, DDY_i^U]。 \tag{6}$$

其中，DDY_i代表第i部电影的导演影响力区间；DDY_i^L、DDY_i^U分别代表第i部电影导演影响力区间的左端点和右端点。当该导演在第i部电影之前执导的全部电影数目$dr \neq 0$时，

$$\begin{cases} DDY_i^U = \max\limits_{dp=1}^{dq}(d - Boxoffice_{idp}), \\ DDY_i^L = \min\limits_{dp=1}^{dq}(d - Boxoffice_{idp})。 \end{cases} \tag{7}$$

式中，$d - Boxoffice_{idp}$表示第i部电影导演在该电影之前执导的第dp部电影的票房(万元)，$dq = \min(dr, 3)$。当$dr = 0$时，

$$DDY_i^L = DDY_i^U = 0。 \tag{8}$$

1.2.6 演员影响力

电影的呈现靠演员实现。演员专业水平、角色塑造能力影响着电影的质量，进而影响观众的观影感受。演员影响力指演员对票房的影响程度。Allbert[22]的研究证明，当前电影的票房受演员前一部电影表现的影响。因此，本文以第一主演和第二主演在当前电影之前参演的电影的票房为基础，量化演员的票房影响力区间，见式(9)~(12)。

$$DZY_i(k) = [DZY_i^L(k), DZY_i^U(k)]。 \tag{9}$$

其中，$DZY_i(k)$代表第i部电影第$k(k = 1,2)$主演影响力区间。当该演员在第i部电影之前参演的全部电影数目$sr(k) \neq 0$时，

$$\begin{cases} DZY_i^U(k) = \max\limits_{sp(k)=1}^{sq(k)}\left(s - Boxoffice_{isp(k)}(k) \times \left(\dfrac{10 - t_{sp(k)}}{10}\right)\right), \\ DZY_i^L(k) = \min\limits_{sp(k)=1}^{sq(k)}\left(s - Boxoffice_{isp(k)}(k) \times \left(\dfrac{10 - t_{sp(k)}}{10}\right)\right)。 \end{cases} \tag{10}$$

式中，$s - Boxoffice_{isp(k)}(k)$表示第$i$部电影第$k$主演在参演第$i$部电影之前参演的第$sp(k)$部电影的票房

(万元)。$t_{sp(k)}$表示演员在之前参演的第sp(k)部电影的角色排名，本文只取演员角色排名在10以内的电影。sq(k) = min(sr(k),3)，sr表示该主演在第i部电影之前主演的全部电影数目，当sr(k)=0时，

$$\mathrm{DZY}_i^{\mathrm{U}}(k)=\mathrm{DZY}_i^{\mathrm{L}}(k)=0。\tag{11}$$

1.2.7　电影时长

在正常的电影时长范围和同等花费的条件下，观众倾向于观看时长更长的影片。电影时长对电影票房有正向影响[7]。类似于微博话题关注度，电影时长是点数据

$$\mathrm{DDS}_i=[\mathrm{DDS}_i^{\mathrm{L}},\mathrm{DDS}_i^{\mathrm{U}}]。\tag{12}$$

其中，DDS_i代表第i部电影时长区间；$\mathrm{DDS}_i^{\mathrm{L}}$、$\mathrm{DDS}_i^{\mathrm{U}}$分别为第$i$部电影时长的最小值和最大值(min)。

1.3　基于训练集的票房理想解贴近度区间计算

基于训练集的票房理想解贴近度区间计算阶段由数据规范化及熵权法确定权重、求加权规范化票房评价指标矩阵、确定每部电影的理想解贴近度和确定分级票房的理想解贴近度区间4部分组成。

1.3.1　数据规范化及熵权法确定权重

1) 数据规范化。

根据票房评价指标选取与量化阶段选取的n个票房评价指标及量化方法，收集m部同类型电影原始数据，构建如式(13)所示的原始的区间数票房评价指标矩阵$\widetilde{\boldsymbol{X}}_{m\times n}$。其中，$\widetilde{X}[i,j]=x_{ij}=[x_{ij}^{\mathrm{L}},x_{ij}^{\mathrm{U}}]$代表第$i$部电影第$j$个票房评价指标区间。

$$\widetilde{\boldsymbol{X}}=\begin{pmatrix}[x_{11}^{\mathrm{L}},x_{11}^{\mathrm{U}}] & \cdots & [x_{1n}^{\mathrm{L}},x_{1n}^{\mathrm{U}}]\\ \vdots & & \vdots\\ [x_{m1}^{\mathrm{L}},x_{m1}^{\mathrm{U}}] & \cdots & [x_{mn}^{\mathrm{L}},x_{mn}^{\mathrm{U}}]\end{pmatrix}。\tag{13}$$

票房评价指标选取及量化阶段选取的指标均为效益型指标，即指标值越大对票房越有益。针对效益型指标的规范化方式[23](见式(14))，$\tilde{y}_{ij}=[y_{ij}^{\mathrm{L}},y_{ij}^{\mathrm{U}}]$代表规范化后的第$i$部电影第$j$个票房评价指标区间。对$\widetilde{\boldsymbol{X}}$规范化，得到式(15)所示的规范化的区间数票房评价指标矩阵$\widetilde{\boldsymbol{Y}}_{m\times n}$，$\widetilde{\boldsymbol{Y}}[i,j]=\tilde{y}_{ij}=[y_{ij}^{\mathrm{L}},y_{ij}^{\mathrm{U}}]$。

$$\begin{cases}y_{ij}^{\mathrm{L}}=\dfrac{y_{ij}^{\mathrm{L}}}{\sqrt{\sum\limits_{i=1}^{m}(x_{ij}^{\mathrm{U}})^2}},\\ y_{ij}^{\mathrm{U}}=\dfrac{y_{ij}^{\mathrm{U}}}{\sqrt{\sum\limits_{i=1}^{m}(x_{ij}^{\mathrm{L}})^2}}。\end{cases}\tag{14}$$

$$\widetilde{\boldsymbol{Y}}=\begin{pmatrix}[y_{11}^{\mathrm{L}},y_{11}^{\mathrm{U}}] & \cdots & [y_{1n}^{\mathrm{L}},y_{1n}^{\mathrm{U}}]\\ \vdots & & \vdots\\ [y_{m1}^{\mathrm{L}},y_{m1}^{\mathrm{U}}] & \cdots & [y_{mn}^{\mathrm{L}},y_{mn}^{\mathrm{U}}]\end{pmatrix}。\tag{15}$$

2) 熵权法确定权重。

根据规范化之后的区间数票房评价指标矩阵$\widetilde{\boldsymbol{Y}}_{m\times n}$，分3步利用熵权法计算各指标权重区间$\widetilde{\omega}$[11, 18]，得到各指标的权重区间。

首先，根据规范化后的区间数票房评价指标矩阵$\widetilde{\boldsymbol{Y}}_{m\times n}$得到列标准化的票房评价指标矩阵$\widetilde{\boldsymbol{F}}_{m\times n}$，$\widetilde{F}[i,j]=\widetilde{f}_{ij}=[f_{ij}^{\mathrm{L}},f_{ij}^{\mathrm{U}}]$，$\widetilde{f}_{ij}$表示列标准化后的第$i$部电影的第$j$个票房评价指标区间。

$$\begin{cases}f_{ij}^{\mathrm{L}}=y_{ij}^{\mathrm{L}}/\sum\limits_{i=1}^{m}y_{ij}^{\mathrm{L}},\\ f_{ij}^{\mathrm{U}}=y_{ij}^{\mathrm{U}}/\sum\limits_{i=1}^{m}y_{ij}^{\mathrm{U}}。\end{cases}\tag{16}$$

其次，计算左右端点信息熵$\widetilde{\boldsymbol{H}}_j=[H_{ij}^{\mathrm{L}},H_{ij}^{\mathrm{U}}]$。$\widetilde{\boldsymbol{H}}_j$表示第$j(j=1,2,\cdots,n)$个指标的信息熵。

$$\begin{cases}H_j^{\mathrm{L}}=\dfrac{\sum\limits_{i=1}^{m}f_{ij}^{\mathrm{L}}\ln f_{ij}^{\mathrm{L}}}{\ln m},\\ H_j^{\mathrm{U}}=\dfrac{\sum\limits_{i=1}^{m}f_{ij}^{\mathrm{U}}\ln f_{ij}^{\mathrm{U}}}{\ln m}。\end{cases}\tag{17}$$

最后，根据信息熵得到指标权重区间$\widetilde{\omega}_j=[\omega_j^{\mathrm{L}},\omega_j^{\mathrm{U}}]$，$j=1,2,\cdots,n$。

$$\begin{cases}\omega_j^1=\left(1-H_j^{\mathrm{L}}\right)/\left(n-\sum\limits_{j=1}^{n}H_j^{\mathrm{L}}\right),\\ \omega_j^2=\left(1-H_j^{\mathrm{U}}\right)/\left(n-\sum\limits_{j=1}^{n}H_j^{\mathrm{U}}\right)。\end{cases}\tag{18}$$

$$\begin{cases}\omega_j^{\mathrm{L}}=\min(\omega_j^1,\omega_j^2),\\ \omega_j^{\mathrm{U}}=\max(\omega_j^1,\omega_j^2)。\end{cases}\tag{19}$$

1.3.2　求加权规范化票房评价指标矩阵

根据数据规范化及熵权法确定权重阶段得到的规范化区间数票房评价指标矩阵及各指标权重，建立加权规范化区间数票房评价指标矩阵$(\boldsymbol{c}_{ij})_{m\times n}$为

$$\boldsymbol{c}_{ij}=[c_{ij}^{\mathrm{L}},c_{ij}^{\mathrm{U}}]=[y_{ij}^{\mathrm{L}},y_{ij}^{\mathrm{U}}]\times[\omega_j^{\mathrm{L}},\omega_j^{\mathrm{U}}]。\tag{20}$$

其中，$C[i,j]=c_{ij}=[c_{ij}^{\mathrm{L}},c_{ij}^{\mathrm{U}}]$表示第$i$部电影加权规范后的第$j$个指标。

进而根据求得的加权规范化区间数票房评价指标矩阵，得到如式(21)和(22)所示的正负理想解。s_j^+代表第j个电影票房评价指标的正理想解；s_j^-代表第j个电影票房评价指标的负理想解。

$$s_j^+ = [s_j^{+\mathrm{L}}, s_j^{+\mathrm{U}}] = [\max_{i=1}^{m}(c_{ij}^{\mathrm{L}}), \max_{i=1}^{m}(c_{ij}^{\mathrm{U}})], \tag{21}$$

$$s_j^- = [s_j^{-\mathrm{L}}, s_j^{-\mathrm{U}}] = [\min_{i=1}^{m}(c_{ij}^{\mathrm{L}}), \min_{i=1}^{m}(c_{ij}^{\mathrm{U}})]。 \tag{22}$$

1.3.3 确定每部电影的理想解贴近度

确定票房理想解贴近度之前首先要确定票房评价指标与正负理想解之间的距离。欧几里得距离是常用的一种距离定义。对于任意的2个区间数$a = [a^{\mathrm{L}}, a^{\mathrm{U}}]$，$b = [b^{\mathrm{L}}, b^{\mathrm{U}}]$，$a$和$b$之间的欧几里得距离[24]为

$$L(a,b) = \sqrt{\frac{1}{2}\left[(a^{\mathrm{L}} - b^{\mathrm{L}})^2 + (a^{\mathrm{U}} - b^{\mathrm{U}})^2\right]}。 \tag{23}$$

根据求加权规范化票房评价指标矩阵阶段得到的理想解与加权规范化区间数票房评价指标矩阵，由式(23)可以得到每部电影票房评价指标与理想解的欧几里得距离为

$$\begin{cases} d_{ij}^+ = L(c_{ij}, s_j^+), \\ d_{ij}^- = L(c_{ij}, s_j^-)。 \end{cases} \tag{24}$$

其中，d_{ij}^+、d_{ij}^-代表第i部电影的第j个加权规范化后的票房评价指标与正理想解、负理想解的欧几里得距离。

根据每部电影各指标与理想解的距离，确定各指标理想解贴近度η_{ij}与电影票房理想解贴近度η_i $(i = 1,2,\cdots,m)$[25]为

$$\eta_{ij} = d_{ij}^- / (d_{ij}^- + d_{ij}^+), \tag{25}$$

$$\eta_i = \left(\sum_{j=1}^{n} \eta_{ij}\right) / n。 \tag{26}$$

1.3.4 确定分级电影的理想解贴近度区间

2017年票房过亿的92部电影中，票房2亿以上的电影占据70%以上的份额。票房超过5亿无疑是好营销和好作品的结合。基于此，将电影票房分为4个等级：Ⅰ(5 000万以下)、Ⅱ(5 000万至2亿)、Ⅲ(2亿至5亿)、Ⅳ(5亿以上)。根据确定每部电影的理想解贴近度阶段得到的每部电影的理想解贴近度及票房分级标准，确定每个等级的电影理想解贴近度区间。

1.4 基于测试集的方法验证

1) 数据规范化。将测试集电影票房评价指标数据规范化处理，得到规范化的票房评价指标。

2) 确定加权规范化票房评价指标。按照式(20)对第1)步求得的规范化的测试集电影票房评价指标进行加权，得到加权规范化的测试集票房评价指标数据。其中，指标权重源自基于训练集的票房理想解贴近度区间计算阶段。

3) 确定测试集电影票房理想解贴近度。首先判定第2)步中得到的加权规范化的票房评价指标与基于训练集的票房理想解贴近度区间计算阶段得到的理想解的大小。其中，对于任意的2个非负区间数$a = [a^{\mathrm{L}}, a^{\mathrm{U}}]$，$b = [b^{\mathrm{L}}, b^{\mathrm{U}}]$，则称$P(a \geqslant b)$为$a \geqslant b$的可能度[26]

$$P(a \geqslant b) = \max\left\{1 - \max\left(\frac{b^{\mathrm{U}} - a^{\mathrm{L}}}{(a^{\mathrm{U}} - a^{\mathrm{L}}) + (b^{\mathrm{U}} - b^{\mathrm{L}})}, 0\right), 0\right\}。 \tag{27}$$

如果测试集中电影i的第j个加权规范化的票房评价指标大于正理想解s_j^+的可能度大于0.5，则记该指标与理想解的贴近度$\eta_{ij} = d_{ij}^- / (d_{ij}^- - d_{ij}^+)$；反之，如果小于负理想解$s_j^-$的可能度大于0.5，则该指标与理想解的贴近度记$\eta_{ij} = -d_{ij}^- / (-d_{ij}^- + d_{ij}^+)$；如果不是以上2种情况，则按照式(24)和(25)计算η_{ij}。最后按照式(26)计算待测试电影的理想解贴近度η_i。

4) 测试集电影票房级别的预测。判定测试集电影i票房理想解贴近度属于哪一票房理想解贴近度区间。该票房理想解贴近度区间对应的票房级别即为测试集电影i的票房预测级别，从而得到票房预测区间值。

由于待预测电影的票房预测流程与基于测试集的方法验证流程相同，不再对待测试电影的票房预测流程加以赘述。

2 预测方法的验证

2.1 数据来源

收集了2015~2017年上映的168部电影相关数据，最终选择数量多且票房波动性居前2类的68部动作类电影和65部剧情类电影验证本文提出的方法。首先，随机抽取8部动作类和8部剧情类电影作为测试集，60部动作类和57部剧情类电影为训练集，分别进行预测方法的验证；其次，为进一步验证方法的可靠性，利用样本数据进行K-折交叉验证。

根据本文选取的票房评价指标的特点及各网站数据显示情况，本文各指标数据来源见表2。

表 2　各指标数据来源

Table 2　Data source of each indicator

指标	数据来源
百度指数(DBD)	百度指数(http://index.baidu.com)
上映日期票房影响力(DDQ)、导演票房影响力(DYZ)、主演票房影响力(DZY(k))、电影时长(DDS)	中国票房网(http://www.cbooo.cn) 猫眼电影(http://pf.maoyan.com) 豆瓣电影(https://movie.douban.com/)
微博话题关注度(DWG)、微博电影视频播放量(DSP)	微博网页端(http://www.weibo.com)
想看人数(DXK)	时光网(http://www.mtime.com) 豆瓣电影(https://movie.douban.com/)

2.2　基于训练集的剧情类和动作类电影票房理想解贴近度区间计算

用57部剧情类和60部动作类电影分别作为训练集，按照图1所示的方法计算票房理想解贴近度区间。

按照式(17)~(19)求得剧情类和动作类电影各指标左右端点的熵以及指标权重，结果见表3。

根据指标权重，按照式(19)对规范化的区间数票房评价指标矩阵进行加权。按照式(21)和(22)求得各指标理想解，结果见表4。

表 3　指标熵及权重

Table 3　The index entropy and weight

评价指标	剧情类电影				动作类电影			
	H^{L}	H^{U}	ω^{L}	ω^{U}	H^{L}	H^{U}	ω^{L}	ω^{U}
DBD	0.860 0	0.885 9	0.086 5	0.090 3	0.871 6	0.869 1	0.090 9	0.132 3
DWG	0.801 2	0.801 2	0.122 7	0.157 3	0.767 2	0.767 2	0.164 8	0.235 2
DSP	0.692 4	0.692 4	0.189 9	0.243 4	0.782 9	0.782 9	0.153 7	0.219 4
DDQ	0.943 2	0.970 5	0.023 3	0.035 0	0.942 7	0.969 7	0.030 6	0.040 6
DXK	0.905 7	0.905 7	0.058 2	0.074 6	0.901 4	0.901 4	0.069 8	0.099 6
DDY	0.731 3	0.791 4	0.165 0	0.165 9	0.752 9	0.840 9	0.160 7	0.175 0
DZY(1)	0.677 6	0.864 9	0.106 9	0.199 1	0.794 1	0.905 0	0.096 0	0.145 8
DZY(2)	0.770 7	0.825 8	0.137 9	0.141 6	0.776 8	0.976 0	0.024 3	0.158 1
DDS	0.998 3	0.998 7	0.001 0	0.001 4	0.998 3	0.998 3	0.001 2	0.001 7

表 4　理想解

Table 4　The ideal solution

评价指标	剧情类电影		动作类电影	
	$[s_j^{-\mathrm{L}}, s_j^{-\mathrm{U}}]$	$[s_j^{+\mathrm{L}}, s_j^{+\mathrm{U}}]$	$[s_j^{-\mathrm{L}}, s_j^{-\mathrm{U}}]$	$[s_j^{+\mathrm{L}}, s_j^{+\mathrm{U}}]$
DBD	[0.000 1, 0.000 4]	[0.014 3, 0.097 5]	[0.000 1, 0.001 2]	[0.012 1, 0.185 1]
DWG	[0,0.000 1]	[0.067 2, 0.086 1]	[0,0]	[0.132 3, 0.188 7]
DSP	[0,0]	[0.123 1, 0.158 1]	[0,0]	[0.072 0, 0.102 7]
DDQ	[0,0.001 2]	[0.004 6, 0.009 6]	[0,0.001 3]	[0.005 6, 0.010 3]
DXK	[0.000 1, 0.000 1]	[0.028 6, 0.036 7]	[0.000 1, 0.000 1]	[0.025 8, 0.036 9]
DDY	[0,0]	[0.048 9, 0.166 4]	[0,0]	[0.039 3, 0.135 7]
DZY(1)	[0,0]	[0.033 1, 0.303 4]	[0,0]	[0.011 1, 0.226 9]
DZY(2)	[0,0]	[0.016 1, 0.453 6]	[0,0]	[0.003 7, 0.267 5]
DDS	[0.000 1, 0.000 1]	[0.000 2, 0.000 2]	[0.000 125, 0.000 178]	[0.000 2, 0.000 3]

根据理想解，用式(26)计算训练集各电影的票房理想解贴近度，进一步得到各电影票房等级的票房理想解贴近度η的取值区间，结果见表5。

2.3　基于测试集的剧情类和动作类电影票房预测方法的验证

用随机选择的8部动作类和8部剧情类电影分别验证提出的电影上映前总票房区间预测方法。验证结果见表6。

由表6可知，只有编号7、8剧情类电影和编号1、2、6动作类电影预测错误。所有预测结果没有出现预测级别和实际级别差超过一个级别的样本，预测结果可用来为影院排片以及发行商决策提供指导。

表 5 票房理想解贴近度区间

Table 5 The ideal solution nearness degree interval of box office

票房级别	剧情类η的取值区间	动作类η的取值区间
Ⅰ(小于5 000万)	[0, 0.135 5)	[0, 0.102 7)
Ⅱ(5 000万-2亿)	[0.135 5, 0.274 3)	[0.102 7, 0.221)
Ⅲ(2亿-5亿)	[0.274 3, 0.4)	[0.221, 0.321 4)
Ⅳ(5亿以上)	[0.4, 1]	[0.321 4, 1]

2.4 动作类和剧情类电影的5-折交叉验证

为了进一步验证预测方法的有效性，基于动作类和剧情类样本数据进行K-折交叉验证。K为进行交叉验证的次数，K取5，将两种类型电影样本通过随机抽样分别分成5个样本子集，轮流将其中4份做训练集，1份做验证集，用票房预测正确的电影数占测试集电影总数的比例计算预测准确率，将5次交叉验证准确率的均值作为K-折交叉验证结果。本文提出的预测方法在动作类和剧情类电影的K-折交叉验证结果见表7。平均准确率分别为79.33%和73.92%，说明本文提出的预测方法具有一定有效性。

表 6 测试集验证结果

Table 6 The validation results of test set

电影类型	编号	电影名称	票房	票房理想解贴近度	实际级别	预测级别
剧情类	1	睡在我上铺的兄弟	12 782.1	0.230 7	Ⅱ	Ⅱ
	2	梦想合伙人	8 100.9	0.234 3	Ⅱ	Ⅱ
	3	魔宫魅影	8 843.9	0.261 4	Ⅱ	Ⅱ
	4	致青春·原来你还在这里	33 685.6	0.294 0	Ⅲ	Ⅲ
	5	七月与安生	13 589.8	0.192 9	Ⅱ	Ⅱ
	6	李雷和韩梅梅	4 072.2	0.072 6	Ⅰ	Ⅰ
	7	上海王	1 365.3	0.149 4	Ⅰ	Ⅱ
	8	中国推销员	1 014.9	0.164 9	Ⅰ	Ⅱ
动作类	1	王牌特工2：黄金圈	47 385.3	0.352 1	Ⅲ	Ⅳ
	2	使徒行者	60 630.4	0.301 9	Ⅳ	Ⅲ
	3	变形金刚5：最后的骑士	15 124.3	0.628 8	Ⅳ	Ⅳ
	4	决战食神	12 191.1	0.207 3	Ⅱ	Ⅱ
	5	超级快递	6 164	0.151 4	Ⅱ	Ⅱ
	6	真相禁区	4 203.8	0.140 5	Ⅰ	Ⅱ
	7	缉枪	1 645.3	0.089 4	Ⅰ	Ⅰ
	8	窦娥奇冤	447	0.040 7	Ⅰ	Ⅰ

表 7 交叉验证结果

Table 7 The cross-validation results %

类型	验证次数					平均准确率
	1	2	3	4	5	
动作类	92.31	76.92	69.23	76.92	81.25	79.33
剧情类	85	61.54	92.31	61.54	69.23	73.92

3 结论

因不同类型电影有不同需求规律，本文提出按电影类型分类的结合区间理论、熵权法和TOPSIS法的电影上映前总票房区间预测方法。为解决模糊性票房影响因素的量化问题，提出用区间数量化票房评价指标；考虑数据本身信息的效用值，采用熵权法对各指标赋权；根据TOPSIS法求解每部电影的票房理想解贴近度，根据每个级别电影的票房理想解贴近度区间判断待预测电影票房所属的级别，从而得到区间预测值。用2015~2017年上映的68部动作类和65部剧情类电影验证了提出的预测方法的有效性。该方法对于与电影类似的短生命周期体验品的早期需求预测具有一定的参考价值。本方法的预测结果存在一定误差，主要原因是没有考虑口碑和电影制作成本等指标。以后的研究有必要考虑增加指标。

参考文献：

[1] GHIASSI M, LIO D, MOON B. Pre-production forecasting of movie revenues with a dynamic artificial neural network[J]. Ex-

pert Systems with Applications, 2015, 42(6): 3176-3193.
[2] LEE K J, CHANG W. Bayesian belief network for box-office performance: a case study on Korean movies[J]. Expert Systems with Applications, 2009, 36(1): 280-291.
[3] ZHANG L, LUO J, YANG S. Forecasting box office revenue of movies with BP neural network[J]. Expert Systems with Applications, 2009, 36(3): 6580-6587.
[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[J]. 计算机应用研究, 2018(2): 410-416.
HAN Zhongming, YUAN Bihong, CHEN Yan, et al. Effective box-office revenue prediction model based on GBRT[J]. Application Research of Computers, 2018(2): 410-416.
[5] 王铮, 许敏. 电影票房的影响因素分析——基于Logit模型的研究[J]. 经济问题探索, 2013(11): 96-102.
[6] 孙春华, 刘业政. 电影预告片在线投放对票房的影响——基于文本情感分析方法[J]. 中国管理科学, 2017, 25(10): 151-161.
SUN Chunhua, LIU Yezheng. The effects of online pre-launch movie trailers on the box office revenue—based on text sentiment analysis method[J]. Chinese Journal of Management Science, 2017, 25(10): 151-161.
[7] 夏丹. 我国3D电影票房影响因素的实证分析[J]. 现代传播-中国传媒大学学报, 2012, 34(9): 139-140.
[8] ALBA S, EUGENIA R. A novel interval finite element method based on the improved interval analysis[J]. Computer Methods in Applied Mechanics & Engineering, 2016, 311: 671-697.
[9] GRECHUKA B, ZABARANKIN M. Direct data-based decision making under uncertainty[J]. European Journal of Operational Research, 2018, 267: 200-211.
[10] YOUNG R C. The algebra of many-valued quantities[J]. Mathematische Annalen, 1931, 104(1): 260-290.
[11] GAO W, WU D, GAO K, et al. Structural reliability analysis with imprecise random and interval fields[J]. Applied Mathematical Modelling, 2017, 55: 49-67.
[12] YANG C, TANGRAMVONG S, GAO W, et al. Interval elastoplastic analysis of structures[J]. Computers & Structures, 2015, 151: 1-10.
[13] TANGARAMVONG S, TIN-LOI F, YANG C, et al. Interval analysis of nonlinear frames with uncertain connection properties[J]. International Journal of Non-Linear Mechanics, 2016, 86: 83-95.
[14] PÉREZ-GALVÁN C, BOGLE I David L. Global optimisation for dynamic systems using interval analysis[J]. Computers & Chemical Engineering, 2017, 107(5): 343-356.
[15] 张晓君, 郑怀昌. 基于区间理论的高应力巷(隧)道围岩岩爆预测[J]. 采矿与安全工程学报, 2011, 28(3): 401-406.
ZHANG Xiaojun, ZHENG Huaichang. Rockburst prediction of surrounding rocks in high-stress roadways (tunnel) based on interval theory[J]. Journal of Mining & Safety Engineering, 2011, 28(3): 401-406.
[16] SAWHNEY M, ELIASHBERG J. A parsimonious model for forecasting gross box-office revenues of motion pictures[J]. Informs, 1996, 15(2): 113-131.
[17] 李波, 陆凤彬, 赵秀娟, 等. 我国电影生命周期模型及实证分析[J]. 系统工程理论与实践, 2010, 30(10): 1790-1797.
LI Bo, LU Fengbin, ZHAO Xiujuan, et al. Chinese movies' life cycle model and empirical analysis[J]. Systems Engineering-Theory & Practice, 2010, 30(10): 1790-1797.
[18] 肖冬香. 基于区间数熵权TOPSIS法的数字图书馆馆藏评价[J]. 情报理论与实践, 2016(12): 99-102.
XIAO Dongxiang. Evaluation of digital library collection based on interval entropy weight TOPSIS method[J]. Information Studies: Theory & Application, 2016(12): 99-102.
[19] 许静, 何桢, 袁荣, 等. 基于主成分分析与TOPSIS模型相结合的函数型产品质量特性的优化方法研究[J]. 工业工程与管理, 2016, 21(3): 59-67.
XU Jing, HE Zhen, YUAN Rong, et al. Optimization of function product quality characteristic problems based on PAC and TOPSIS model[J]. Industrial Engineering and Management, 2016, 21(3): 59-67.
[20] CHONG O, ROUMANI Y, NWANKPA J K, et al. Beyond likes and tweets: consumer engagement behavior and movie box office in social media[J]. Information & Management, 2016, 54(1): 25-37.
[21] BAEK H, OH S, YANG H D, et al. Electronic word-of-mouth, box office revenue and social media[J]. Electronic Commerce Research & Applications, 2017, 22: 13-23.
[22] ALBERT S. Movie stars and the distribution of financially successful films in the motion picture industry[J]. Journal of Cultural Economics, 1999, 23(4): 319-323.
[23] 赵萌, 任嵘嵘, 李刚. 基于模糊熵-熵权法的混合多属性决策方法[J]. 运筹与管理, 2013(6): 78-83.
ZHAO Meng, REN Rongrong, LI Gang. Method based on fuzzy entropy-entropy weight for hybird multi-attribute decision making[J]. Operations Research and Management Science, 2013(6): 78-83.
[24] DANG Y, LIU S, MI C. Multi-attribute grey incidence decision-model for interval number[J]. Kybernetes, 2006, 35(7/8): 1265-1272.
[25] 刘书庆, 李春花. 基于改进型TOPSIS的协同创新参与主体组合模式优选模型研究[J]. 工业工程与管理, 2016, 21(5): 1-8.
LIU Shuqing, LI Chunhua. The optimal selection of collaborative innovational participants combination model based on improved TOPSIS[J]. Industrial Engineering and Management, 2016, 21(5): 1-8.
[26] 岳中亮, 贾玉英. 基于区间数的高校图书馆网站的综合测评[J]. 情报杂志, 2007, 26(12): 142-144.
YUE Zhongliang, JIA Yuying. The integration evaluation of university library website based on interval number[J]. Journal of Intelligence, 2007, 26(12): 142-144.

(责任编辑: 刘敏仪)