基于随机森林的影片票房预测

中国电影科学技术研究所 张 鑫
天津大学 郭振宇

【摘要】提出了一种剪枝的随机森林算法，利用互联网收集到的影片上映前一个月的相关数据，建立模型，预测我国影院影片上映首周票房。该票房预测模型将预测问题转换为分类问题，把票房收入离散化为8个类别，能够预测影片在某影院的票房收入范围。以68个影院1年的票房作为数据基础进行评估验证，结果表明该模型优于一般统计模型。使用此模型可辅助制片方、发行方和营销公司进行决策，也可帮助影院经营者优化放映排期，使影院票房收入最大化。

【关键词】电影 票房预测 随机森林

引言

影片票房的预测很早以前就受到制片方、发行方、营销公司以及影院经营者的关注，但人们一直未能探索和发现出一种普适、准确的预测方法。一旦我们能够准确的预测票房，我们就可以更精准地控制投资、发行和放映，将各阶段风险降至最低，这对于整个电影工业化流程有非常现实的意义。我国电影行业正处于高速增长期，院线、影院之间的竞争越来越激烈，此时票房预测引起了院线和影院的关注，他们意识到科学预测合理排片的重要性，开始尝试使用数据分析预测来取代过去的主观臆断和头脑风暴。

研究人员曾试图开发各种统计模型预测影片票房。有人采用多层认知神经网络模型将票房分为9类进行预测，也有人使用贝叶斯信任网络和BP神经网络建立票房预测模型。最有名的当属谷歌公司提出的票房预测模型，它采用简单的线性回归模型，使用影片相关搜索查询总量提前一个月预测影片票房。

虽然各种预测模型已经能够在特定条件下得到较好表现，但市场与生俱来的不确定性仍然加大了预测的风险。百度公司曾提出一种票房预测模型，在实验阶段有较高的准确率，但在2014年预测影片《黄金时代》的票房时，预测的2亿票房与实际5千万票房相距甚远。2015年取得不错票房成绩的《港囧》，在上映前预测有20亿票房，但由于盗版等问题最终票房受到影响。

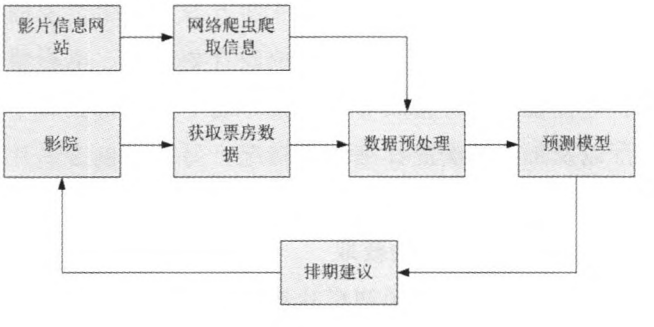

图1 票房预测系统架构

现实生活中存在的大量不可确定的因素给票房预测工作带来很大挑战。借鉴前人的研究经验和成果，本文提出了一种剪枝的随机森林算法，能提前

影片上映一个月预测影片在某影院首周的票房成绩。实际上该模型将预测问题转换为分类问题，将票房数值转换为票房的 8 个类别。使用此模型最终构建成如图 1 所示的完整预测系统。

随机森林算法在机器学习的预测和分类领域有多种场合的应用，算法本身由一定数量的决策树构成，后者是实践中应用最为广泛的机器学习算法。为了将输入向量所代表的一个对象划分为某一类型，随机森林算法将数据输入到所有决策树中。每个决策树给出一种分类结果，而最终随机森林的分类结果则由所有决策树给出的结果投票计算得出，选择获得投票最多的一种分类作为最终输出分类。本文所提出的预测算法针对随机森林的强度和相关度对其进行剪枝，使得剪枝的随机森林算法在所应用的数据上表现出优于传统随机森林算法的性能。

在图 1 的票房预测系统中，票房数据来自 68 个国内影院，影片信息从不同的电影网站爬取。与其他票房预测模型相比，本文模型具有以下显著特点：首先，模型加入了百度指数和预告片搜索量，这两个特征能够代表影片的受关注程度；其次，模型加入了最低票价，这一特征可以反映片方和发行方对影片的某种预期。

1 影片票房的预测方法

1.1 影片票房历史数据

本文所使用的影片票房历史数据为影院分厅分场详细售票数据，来自国内某城市的 68 个影院，时间跨度为 2013 年 6 月至 2014 年 6 月。每个影院的数据包括售票系统日报表、放映计划报表、电影票预定报表和支付报表等。数据体现了影院每天每个厅以及每个场次的详细售票情况，另外还包含影片的最低票价等有价值的信息。

1.2 影片信息的获取

为了达到准确预测影片票房的目的，单纯依靠影片票房历史数据是远不够的，需要更多的影片信息作为特征才能科学地建立起预测模型。影响影片票房的因素众多，参考其他票房预测方法，对各种因素进行评价和筛选，将影片的导演、演员、类型、制片厂、档期等作为影片自身重要特征加入模型。为了提升模型的准确度，参考谷歌公司的预测模型，加入百度指数、预告片搜索量作为额外的重要特征。另外，意识到准确可靠的数据是成功建立模型的关键，同种类的数据会从不同网站获取，选择其中质量最好的数据来建模。数据获取由图 2 所示分布式网络爬虫系统负责，针对不同网站由不同的网络爬虫插件处理，实现实时抓取互联网数据，存储数据到 MySQL 数据库中。

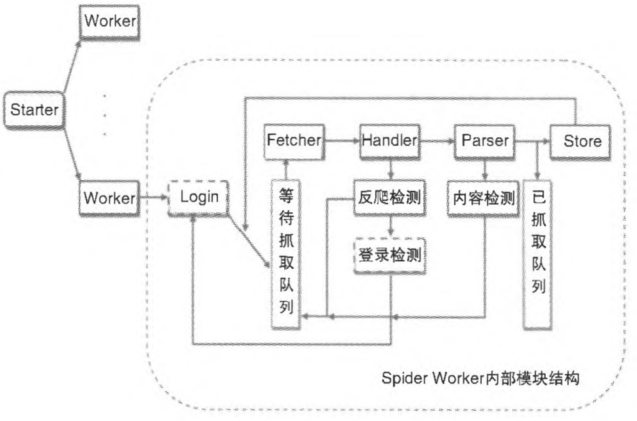

图 2 分布式网络爬虫系统

(1) 百度指数

百度指数是百度公司对公众开放的互联网大数据平台，百度公司使用特有的算法为用户所要查询的关键词计算指数。我们发现使用影片作为关键词在百度指数上进行查询，指数在影片上映前一个月开始有较为集中的增长，在影片上映后随着影片临近下映，指数随热度的逐渐下降而降低。影片在百度指数上体现的数值与影片票房具有正相关关系。影片的百度指数在临近影片上映的前几周逐渐增大，我们使用影片上映前 5 周的百度指数作为预测模型的特征。因为在同一时间段内可能有多个影片即将上映或已经上映，且它们之间能够产生相互作用，所以把那些影片的百度指数也作为重要的特征。

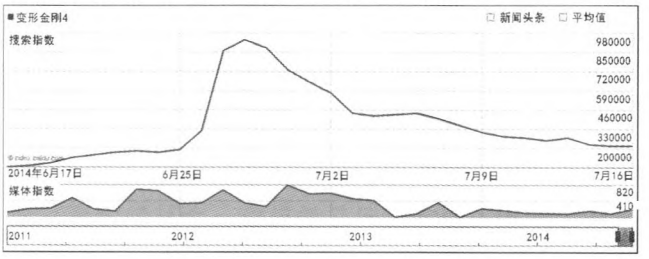

图 3 影片的百度指数趋势

（2）同档期影片

同档期影片即在同一时间可供影院选择排片的新片，它们是有着“竞争”关系的影片。影片除了自身质量外，选择在什么时间上映是极为重要的商业运作技巧，这在行业内广为人知。《小时代》当时取得了令人刮目的票房成绩，其中一个重要原因是影片选择了当年 6 月底上映，在此时间点没有特别值得观众关注的影片，使得该片获得了非常高的排片占比。而且，这个时间点正值学生暑假，影片的目标观众有足够的时间去影院观影。

从影片票房历史数据能够看出，对于一部普通的影片，只要选择了准确的上映时间点，就可以获得不错的票房成绩。在 2014 年“春节档”，《西游记大闹天宫》、《爸爸去哪儿》、《澳门风云》三部国产影片都取得了非常不错的票房成绩，但如果将这些影片安排到其他时间上映，则肯定不会享受“春节档”带来的这般福利。

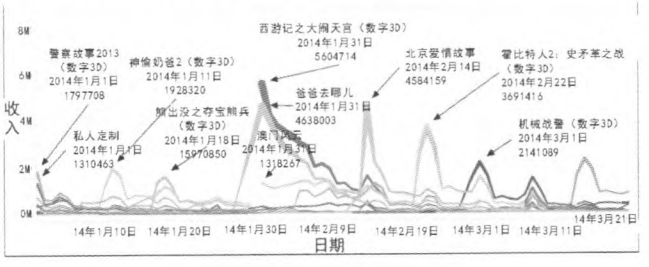

图 4　2014 年春节档前后影片票房曲线

（3）导演和演员

研究表明影响影片票房最重要的因素是导演和演员，甚至超过了影片故事内容本身。导演是影片创作团队的领导者和组织者，决定了影片的质量和影片艺术风格。演员则具有特殊的票房号召力，观众会被演员的个人魅力所吸引，甚至成为他的忠实粉丝。研究人员曾使用《好莱坞报道》提供的演员、导演影响力指标作为特征预测票房。在本文的模型中，假定导演和演员的影响力可以表现为他们曾经参与影片的票房成绩。以著名演员成龙为例，将他的影片《大兵小将》、《十二生肖》和《警察故事 2013》作为特征，来预测他的新近影片《天将雄师》。

（4）影片类型

影片有科幻片、喜剧片、爱情片等类型，不同观众在不同的时期可能会有不同的喜好，因此影片类型对于影片票房也很重要。影片类型决定了影片内容的表现方式、观众基础和影响力。不同观众有不同的文化背景和社会背景，也有不同的消费习惯和精神诉求。这种现象对影院的影响主要在于影院周围的社会人群，因此会出现局部地区观众对某些类影片有特殊的关注。

（5）预告片搜索量

谷歌公司对票房预测的研究表明，在票房和预告片相关搜索量之间存在正线性相关性。因为优酷是国内最大的视频网站，其数据相较其他视频网站更有参考价值，所以本文通过网络爬虫从优酷指数网站获取搜索和播放量。如同百度指数，搜索量和播放量数值越大，则影片受关注程度越高，预期票房数值越高。

（6）制片厂

国内较受欢迎的电影制片厂或制片公司有中影集团、华谊兄弟、光线传媒等。大的制片厂所代表的往往是较高的影片质量。他们也参与决定最低票价，最低票价也是本文模型的一个有用特征。

（7）场次数量

无论从直觉上还是统计上都表明，影片在影院被安排的场次数量越多，越能有更高的票房。因此，本文将场次数量以连续变量的形式加入到预测模型中。

表 1　独立变量及数据来源

变量名	数据来源
百度指数	index. baidu. com
同档期影片	影院票房数据
导演和演员	www. cbooo. cn
影片类型	www. gewara. com
预告片搜索量	index. youku. com
制片厂	www. cbooo. cn
场次数量	影院票房数据

1.3 剪枝的随进森林算法

随机森林算法作为一种集成算法已经被广泛采用，通常能带来出色的预测结果。本文通过强度和

相关性两个参数剪枝随机森林算法。

给定一个随机森林分类器集合 $h_1(x)$，$h_2(x)$，…，$h_k(x)$，和一个模型训练集。定义随机森林的边差函数为：

$$mr(X, Y) = \frac{1}{N}\sum_{n=1}^{N}\{I(h(X)=Y) - \max_{j\neq Y} I(h(X)=j)\} \quad (1)$$

这里 $I(\cdot)$ 为指示函数，边差是投票给正确类别的估计概率与除了正确类别以外的最优可能类别的估计概率之间的差别度量。边差越大则预测置信度越高。随机森林的强度函数定义为：

$$strength = E_{X,Y}\, mr(X, Y) = \frac{1}{N}\sum_{i=1}^{N} mr(x_i, y_i) \quad (2)$$

定义随机森林的相关函数为：

$$\bar{\rho} = \frac{(var(mr))}{sd(h)^2} \quad (3)$$

其中 sd（h）为随机森林的标准差。

为了改善精度，需要在增加随机森林强度的同时最小化相关性。算法研究人员已经研究了很多种随机森林的剪枝算法，如基于树相似度的随机森林剪枝和边差距离最小化的随机森林剪枝，这些方法都主要聚焦在如何改进每棵树的预测精度。

本文通过调整强度和相关性两个参数达到对随机森林剪枝的目的，两个参数都需要计算随机森林的边差。边差对于集成分类器算法很重要，如 Adaboost 算法、Mdboost [13] 算法以及 Arc－Gv 算法等，通过改变边差来改进他们的泛化能力。在本算法中我们假设随机森林为，通过其观察强度和相关性在移除某基树时的减少量，来评价基树的贡献度和重要性，这等同于评估子系综 {H/h}。当多余的树被剪以后，H 缩减为其子集 H′= {H/h}。在这个过程中，我们首先计算 H 的强度和相关性，接着计算每棵树 H_k 的强度和相关性，最后去掉 H 中具有较小强度和较大相关性的不重要的树。在训练集 S 中每棵树的强度和相关性用下面的评估矩阵计算：

$$Strength(h_i, H_k, S) = strength(x, y, H) - strength(x, y, \frac{H}{h_i}) \quad (4)$$

$$Correlation(h_i, H_k, S) = \bar{\rho}(x, y, H) - \bar{\rho}(x, y, \frac{H}{h_i}) \quad (5)$$

2 实验结果

2.1 算法的分类性能

表 2　68 个影院预测结果

		实际类型								平均
预测类型		1	2	3	4	5	6	7	8	
	1	1517	349	58	6	2	0	0	0	
	2	287	1024	448	76	12	4	1	0	
	3	28	378	885	417	80	11	0	0	
	4	7	54	344	942	398	62	7	0	
	5	0	3	40	368	919	375	57	5	
	6	0	2	4	42	359	975	349	38	
	7	0	0	0	7	32	347	1058	303	
	8	1	1	1	3	6	46	344	1469	
	Bingo	82.45	56.54	49.72	50.62	50.83	53.57	58.26	80.94	60.37
	1－Away	98.04	96.69	94.21	92.80	92.70	93.24	96.42	97.63	95.22

如前所述，本文剪枝的随机森林算法最终将影片按票房归为 8 个类型中的一类。我们采用 8 折交叉验证方法来评估算法的性能。用平均命中率（APHR）度量算法的预测性能。实验中，我们使用两种不同的命中率，分别为：绝对准确率（Bingo）－分类到正确类型的概率；相对准确率（1－Away）

一分类到正确类型或其临近类型的概率。两种平均命中率用下式进行计算：

$$APHR_{Bingo}=\frac{1}{n}\sum_{i=0}^{c}p_i \tag{6}$$

$$APHR_{1-Away}=\frac{1}{n}\sum_{i=0}^{c}(p_i+p_{i-1}+p_{i+1}) \tag{7}$$

其中，c 为所有类型的总数，n 为样本众数，p_i 为被分为类的样本总数。

表 2 是对 68 个国内影院进行 8 折交叉验证所得到的实验结果，以混淆矩阵的形式给出。

每行的总数表示预测为该类别的样本数量，每列的总数表示总样本中实际某类型的样本数量。

2.2 与其他算法的比较

除了本文所采用的随机森林，应用中还存在多种流行的分类算法。常用的统计分类算法有：多元逻辑回归、CART、SVM 以及神经网络等。本文对随机森林剪枝，得到了 60%的绝对准确率和和 95%的相对准确率，与传统随机森林（RF）、决策树（DT）、支持向量机（SVM）和多层感知机（MLP）等算法相比具有明显的优势，如下表 3 和图 5 所示：

表 3 预测结果比较

	PRF	RF	DT	SVM	MLP
Bingo（%）	60.09	54.27	52.9	32.2	19.7
1－Away（%）	95.26	90.96	92.1	59.6	48.7

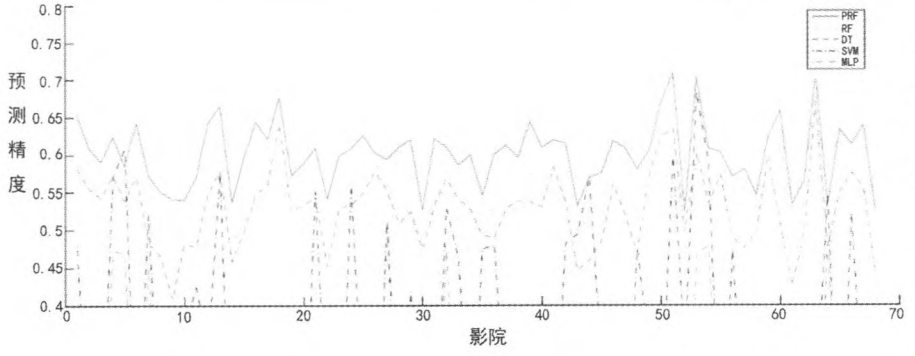

图 5 预测精度曲线

3 总结

本文提出了一种基于剪枝的随机森林算法，使用来自国内 68 家影院的票房数据和互联网影片信息数据，运用机器学习的方法得出票房预测模型，实验结果表现较好。实践过程中数据源的获取以及数据预处理是关键，没有准确的数据就没有准确的预测。背景知识对于建立模型相当重要，数据背后的故事需要丰富的电影行业背景知识帮助解读。随机森林算法通常能够得到好的效果，但也带来过拟合的风险。综上所述，无论是数据还是算法在票房预测模型的建立过程中都需要认真谨慎对待。通过研究实践，可以肯定的是，采用算法预测票房能够得到较好的准确率，并可作为一种较可靠的方法取代主观臆断和头脑风暴，帮助人们做出科学的决策。

❖

参考文献

[1] Ainslie, A., Dr'eze, X., Zufryden, F.: Modeling movie life cycles and market share. Marketing Science 24 (3), 508 - 517 (2005)

[2] Breiman, L.: Prediction games and arcing algorithms. Neural computation 11 (7), 1493 - 1517 (1999)

[3] Breiman, L.: Random forests. Machine learning 45 (1), 5 - 32 (2001)

[4] Eliashberg, J., Elberse, A., Leenders, M. A.: The motion picture industry: Critical issues in practice, current research, and new research directions. Marketing Science 25 (6), 638 - 661 (2006)

[5] Hennig－Thurau, T., Houston, M. B., Walsh, G.: Determinants of motion picture box office and profitability: an interrelationship approach. Review of Managerial Science 1 (1), 65 - 92 (2007)

[6] Lee, K. J., Chang, W.: Bayesian belief network for box－office performance: A case study on korean movies. Expert Systems with Applications 36 (1), 280 - 291 (2009)

[7] Liaw, A., Wiener, M.: Classification and regression by randomforest. R news 2 (3), 18 - 22 (2002)

[8] 张鑫，陈健宁，陈伟基．数字签名在数字电影放映监管中的应用［J］．现代电影技术，2015，1：23～27.

[9] Martinez－Muoz, G., Hern'andez－Lobato, D., Su'arez, A.: An analysis of ensemble pruning techniques based on ordered aggregation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31 (2), 245 - 259 (2009)

（下转第 35 页）

《环境保护法》、《环境影响评价法》、《自然保护区条例》、《风景名胜区条例》、《文物保护法》、《文物保护法实施条例》等法律法规，严格履行报批手续，并且在拍摄活动结束后对拍摄地环境进行恢复，由有关政府部门组织验收，对于验收不合格者，应责令其在一定期限内治理恢复，治理不达标者则要求剧组采取补救措施甚至罚款。在限制拍摄区，剧组应根据其拍摄活动对环境的影响程度，对于可能造成重大环境影响的，应当编制环境影响报告书，对拍摄活动产生的污染和对环境的影响进行全面、详细的评价；对于可能造成轻度环境影响的，应当编制环境影响报告表，对拍摄活动产生的污染和对环境的影响进行分析或者专项评价；对于环境影响很小，不需要进行环境影响评价的，应当填报环境影响登记表。经批准的环境影响评价文件作为准予影视拍摄许可、备案和批准的依据[5]。

三是加强政府监督管理、树立绿色影视先锋模范。在影视剧组履行相关审批后，政府要实施统一监督管理，特别是在环境敏感区拍摄，不仅主管环境问题的国家机关要进行监督管理，当地政府的有关职能部门也应加强监管，各地环保、土地、矿产、林业、农业、水利行政主管部门等都有管理、监管环境的职责和义务。政府应加大宣传，倡导绿色影视，树立绿色影视先锋模范，比如：由成龙主演，威尔·史密斯监制的好莱坞电影《功夫梦》在武当山风景区拍摄时自己带了移动厕所，所有垃圾都回收处理；华纳兄弟影业公司更是立志成为电影业中的“环保领导者”，草坪使用生物除草法避免化学除草剂、所有车辆的机油经过再提炼后重复使用、拍摄场地搭建场景使用可重复金属支架等，政府主管部门应当将他们树立为“绿色影视”先锋，发挥模范作用。❖

参考文献

[1] 李克．2011. 新世纪东北“绿色影视”的环保意识源流．长春工程学院学报（社会科学版）．12（4）：78－80.

[2] 中国电影科学技术研究所，中国电影电视技术学会环保专业委员会．2000. 全国电影环保工作大事记．

[3] 李北陵．2006. 透过美国绿色电影产业看差距．环境．7（7）：13.

[4] 环境保护部环境工程评估中心．2014. 环境影响评价相关法律法规．中国环境科学出版社．

[5] 李婧．论敏感区影视拍摄的法律规制．2012. 硕士学位论文．山西财经大学．

（上接第 15 页）

[10] Panaligan, R., Chen, A.: Quantifying movie magic with google search. Google WhitepaperIndustry Perspectives + User Insights (2013)

[11] R¨atsch, G., Onoda, T., M¨uller, K.R.: Soft margins for adaboost. Machine learning 42 (3), 287 - 320 (2001)

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

[13] Shen, C., Li, H.: Boosting through optimization of margin distributions. Neural Networks, IEEE Transactions on 21 (4), 659 - 666 (2010)

[14] Yang, F., Lu, W.h., Luo, L.k., Li, T.: Margin optimization based pruning for random forest. Neurocomputing 94, 54 - 63 (2012)

[15] Zhang, H., Wang, M.: Search for the smallest random forest. Statistics and its Interface 2 (3), 381 (2009)

[16] Zhang, L., Luo, J., Yang, S.: Forecasting box o _ ce revenue of movies with bp neural network. Expert Systems with Applications 36 (3), 6580 - 6587 (2009)