文章编号：2095-2163(2019)04-0132-05 中图分类号：TP311 文献标志码：A

基于决策树算法的电影票房预测研究

李振兴[1]，韩丽娜[1,2]，史 楠[1]
（1 西安石油大学 计算机学院，西安 710065；2 陕西学前师范学院，西安 710100）

摘 要： 决策树是一种具有树形结构的机器学习算法，能够在短时间内处理数据，并能直观地显示数据特性。具有速度快、直观、精度高等特点。本文在大数据分析的基础上，以 2018 年国内上映的 30 部国产电影的信息数据作为训练模型，选取 C4.5 算法作为工具，构建出基于决策树算法的票房预测模型。经过测试，该模型的准确率为 78%，并从中分析出影响票房的关键因素是演员。

关键词： 决策树算法；电影票房；预测

Research on the prediction of film box office based on decision tree algorithm

LI Zhenxing[1], HAN Lina[1,2], SHI Nan[2]
(1 School of Computer Science, Xi'an Shiyou University, Xi'an 710065, China;
2 Shaanxi Xueqian Normal University, Xi'an 710100, China)

【Abstract】 Decision tree is a machine learning algorithm with a tree structure that processes data in a short period of time and visually displays data features. It has the characteristics of fast speed, intuitiveness and high precision. Based on the analysis of big data, this paper takes the information data of 30 domestic films released in China in 2018 as the training model, and selects C4.5 algorithm as the tool to construct the box office prediction model based on decision tree algorithm. After testing, the accuracy of the model is 78%, and the key factor affecting the box office is analyzed, which is the actor.

【Key words】 decision tree algorithm; film box office; prediction

0 引 言

电影产业是一项高投资、高收益、高风险的行业，当今社会已进入了大数据时代，可以将数据挖掘技术应用到电影票房的预测研究中，为投资者智能规避电影投资风险，并帮助影院运营商优化放映计划，实现收益的最大化[1-2]。本文提出了一种基于决策树算法的票房预测模型，该模型将预测问题转换为分类问题，将电影类型、演员流量程度、导演知名度作为自变量，电影票房类别作为因变量。与以往的主观假设和头脑风暴相比，这是一种更可靠、更科学的方法[3]。

1 决策树算法及相关概念

1.1 决策树

决策树算法采用的是自顶向下的贪婪算法，在每个节点上选择出最优属性进行分类。算法包括 ID3、C4.5、CHAID、CART、SLIQ、SPRINT 等。其中 C4.5 算法在 2006 年 12 月举行的国际数据挖掘会议（ICDM）上，排在十大数据挖掘算法之列[4]。

1.2 C4.5 算法

C4.5 算法是一种基于信息熵的机器学习算法，主要采用信息增益率作为条件属性的判断标准，信息增益率越高，数据分类能力越强。因此，分别计算每一个条件属性的信息增益率，选取信息增益率最高的属性作为下一个分裂节点，以此递归即可构建 C4.5 决策树[5-7]。相关公式如下：

（1）信息熵 $E(S)$：信息量的数学期望，是对不确定性的度量。

$$E(S) = -\sum_{k=1}^{K} \frac{|S_k|}{|S|} \log_2 \frac{|S_k|}{|S|}, \quad (1)$$

其中，S 表示训练样本集；K 表示划分的类别数；$\frac{|S_k|}{|S|}$ 表示第 k 个类别在训练样本集中出现的概率。

（2）条件熵 $E(S,A)$：表示属性 A 对训练样本集 S 划分的期望信息。

$$E(S,A) = -\sum_{i=1}^{n} \frac{|S_i|}{S} \sum_{k=1}^{K} \frac{|S_{ik}|}{|S_i|} \log_2 \frac{|S_{ik}|}{|S_i|}, \quad (2)$$

其中，n 表示属性 A 将训练样本集 S 划分的子集

作者简介：李振兴（1994-），男，硕士研究生，主要研究方向：图像处理、数据挖掘；韩丽娜（1976-），女，博士，教授，硕士生导师，CCF 会员，主要研究方向：数据挖掘、图像处理；史 楠（1995-），男，硕士研究生，主要研究方向：图像处理、数据挖掘。

收稿日期：2019-04-19

数量，$\frac{|S_i|}{S}$ 表示第 i 个属性的权重。

属性 A 划分样本集 S 的信息增益表示为：

$$Gain(A) = E(S) - E(S,A), \quad (3)$$

属性 A 划分样本集 S 的信息增益率表示为：

$$GainR(A) = Gain(A)/E(A). \quad (4)$$

1.3　决策树修剪

由于决策树是由训练数据集生成的，许多分支反映的是噪声或孤立点，这可能会增加决策树分类的错误率，因此有必要对决策树进行修剪[10]。修剪决策树一般分为：预剪枝法和后剪枝法。预剪枝法是在树生长的过程中设置一定的标准来阻止树木继续生长。后剪枝法是待决策树完全生成后再进行剪枝。后剪枝方法比预剪枝方法需要更多的计算量，但通常可以产生更可靠的树[11-12]。

2　应用决策树技术预测电影票房

2.1　数据准备

本次研究中数据信息来源于“中国电影票房年度总排行榜”网，从中抽取 48 部电影，将其中的 30 部作为训练样本数据，剩余的 18 部作为测试样本数据。数据源主要包括电影类型、电影导演、电影主演。原始数据见表 1。

表 1　30 个训练样本原始数据

Tab. 1　Raw data for 30 training samples

电影名	类型	导演	主演	票房
红海行动	剧情	林超贤	张译 / 黄景瑜	36.494 4 亿
我不是药神	喜剧	文牧野	徐峥/王传君	30.969 1 亿
西虹市首富	喜剧	闫非	沈腾	25.462 5 亿
捉妖记 2	奇(科)幻	许诚毅	梁朝伟 / 白百何	22.366 5 亿
后来的我们	爱情	刘若英	井柏然 / 周冬雨	13.639 4 亿
一出好戏	喜剧	黄渤	黄渤 / 舒淇	13.513 1 亿
无双	剧情	庄文强	周润发	12.72 亿
西游记女儿国	奇(科)幻	郑保瑞	郭富城 / 冯绍峰	7.271 1 亿
影	剧情	张艺谋	邓超 / 孙俪	6.281 4 亿
四大天王	奇(科)幻	徐克	赵又廷 / 冯绍峰	6.064 2 亿
李茶的姑妈	喜剧	吴昱翰	黄才伦 / 艾伦	6.038 5 亿
邪不压正	剧情	姜文	彭于晏	5.837 5 亿
爱情公寓	喜剧	韦正	陈赫 / 袁弘	5.547 2 亿
动物世界	剧情	韩延	李易峰	5.089 8 亿
快把我哥带走	奇(科)幻	郑芬芬	张子枫 / 彭昱畅	3.740 6 亿
幕后玩家	剧情	任鹏远	徐峥 / 王丽坤	3.584 4 亿
悲伤逆流成河	爱情	落落	赵英博 / 任敏	3.546 5 亿
找到你	剧情	吕乐	姚晨 / 马伊琍	2.84 亿
南极之恋	爱情	吴有音	赵又廷 / 杨子姗	2.339 5 亿
猛虫过江	喜剧	小沈阳	小沈阳 / 潘斌龙	2.035 8 亿
祖宗十九代	喜剧	郭德纲	岳云鹏 / 吴京	1.684 1 亿
江湖儿女	爱情	贾樟柯	赵涛 / 廖凡	6 938.1 万
卧底巨星	喜剧	谷德昭	陈奕迅 / 李荣浩	3 891 万
我是你妈	喜剧	张骁	闫妮	3 682 万
泡芙小姐	爱情	张歆艺	张歆艺	1 933.3 万
我说的都是真的	喜剧	刘仪伟	小沈阳 / 陈意涵	1 781.2 万
英雄本色 2018	剧情	丁晟	王凯 / 马天宇	6 300 万
遇见你真好	剧情	顾长卫	白客	5 094 万
时空偷渡少女	奇(科)幻	吴琴	王安琪	198.5 万
盲道	剧情	李杨	李杨 / 杜函梦	52.5 万

2.2　数据预处理

(1)电影类型。每位观众在不同阶段可能会有不同的喜好，因此电影类型对于电影票房很重要。变量值有：剧情、喜剧、奇(科)幻、爱情。

(2)电影导演。导演是影片制作的领导者和组织者，决定着影片的质量和影片艺术风格。通过对这些导演的获奖情况和近三年来执导电影所获票房的均数进行分析。将国内顶级的大导演划分为高层次，知名导演划分为中等层次，非知名导演划分为低层次。

(3)电影主演。演员具有一定程度的票房号召力，观众会因为喜欢的演员而选择电影。通过对

“2018 年中国内地演员排行榜”的数据分析,将排名前 50 的演员划分为高流量演员,排名 51-300 的演员划分为中等流量演员,排名 300 以后的演员划分为低流量演员。

(4)电影票房。作为数据的因变量,参考国外学者 Ramesh[13] 的票房划分方法,将票房收益高于 6 亿的电影划分为高票房,将票房收益介于 1 亿至 6 亿之间的电影划分为中等票房,将票房收益低于 1 亿的电影划分为低等票房。

经过数据预处理后,量化表示数据表中的描述性文字,得到了 30 个处理后的数据训练样本,见表 2。

表 2 30 个训练样本

Tab. 2 30 training samples

编号	类型	导演知名程度	演员流量程度	票房高低
1	喜剧	中	高	高
2	喜剧	中	中	高
3	喜剧	中	中	中
4	喜剧	低	中	中
5	喜剧	低	中	低
6	喜剧	低	中	低
7	喜剧	中	中	低
8	喜剧	中	高	高
9	喜剧	低	高	高
10	喜剧	低	低	中
11	奇(科)幻	中	中	中
12	奇(科)幻	中	高	高
13	奇(科)幻	中	高	高
14	奇(科)幻	高	高	高
15	奇(科)幻	低	低	低
16	剧情	高	中	高
17	剧情	中	中	高
18	剧情	中	中	中
19	剧情	中	中	低
20	剧情	高	高	高
21	剧情	高	高	中
22	剧情	中	高	中
23	剧情	中	高	中
24	剧情	低	高	低
25	剧情	低	低	低
26	爱情	中	中	中
27	爱情	低	中	低
28	爱情	低	高	高
29	爱情	高	低	低
30	爱情	低	低	中

2.3 C4.5 构造决策树

(1)计算信息熵。在表 2 的 30 个训练样本中,属于高票房的有 11 个样本,中等票房的有 10 个样本,低票房的有 9 个样本。设训练样本集为 S, 根据公式(1)得到样本分类的期望信息为:

$$entropy(S) = -\frac{11}{30}\log_2\frac{11}{30} - \frac{10}{30}\log_2\frac{10}{30} - \frac{9}{30}\log_2\frac{9}{30} = 1.58.$$

(2)计算条件熵。属性导演知名度 director 将票房划分为 3 部分,即:高知名度导演($S_{h\_director}$)、中等知名度导演($S_{m\_director}$)、低知名度导演($S_{l\_director}$)。取值分别是 5,14,11。而在高知名度导演中,高票房样本 3 个,中等票房样本 1 个,低票房样本 1 个。根据公式(1)可得到 $S_{h\_director}$ 的熵为:

$$entropy(S_{h\_director}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{1}{5}\log_2\frac{1}{5} - \frac{1}{5}\log_2\frac{1}{5} = 1.37.$$

同理,可以计算出 $S_{m\_director}$ 和 $S_{l\_director}$ 的熵分别为 1.45 和 1.44,根据公式(2),使用属性 director 划分样本集 S 的期望信息为:

$$entropy(S,director) = \frac{5}{30}entropy(S_{h\_director}) + \frac{14}{30}S_{m\_director} + \frac{11}{30}S_{l\_director} = 1.43.$$

(3)计算信息增益和信息增益率。根据公式(3)、公式(4) 可得到属性 *director* 的信息增益和信息增益率为:

$$gain(S,director) = 1.58 - 1.43 = 0.15;$$

$$gain\_ratio(S,director) = \frac{0.15}{1.43} = 0.10.$$

同理可得,属性演员流量程度(actor)和属性电影类型(sort)的信息增益和信息增益率分别为:

$$gain(S,actor) = 1.58 - 1.31 = 0.27;$$

$$gain\_ratio(S,actor) = \frac{0.27}{1.31} = 0.21;$$

$$gain(S,sort) = 1.58 - 1.53 = 0.05;$$

$$gain\_ratio(S,sort) = \frac{0.05}{1.53} = 0.03.$$

(4)建立决策树。因为属性 actor 的信息增益率最大,所以选择属性 actor 作为根结点。按照 actor 的取值,对 30 个样本进行分支得到 3 个子集,如图 1 所示。并对每个子集按照以上方法创建分支,最后得到 C4.5 决策树,如图 1 所示。最后采用后修剪方式,修剪后的 C4.5 决策树如图 2 所示。

2.4 模型评估

为了验证模型的可靠性,根据图 3 的决策树对 18 个测试样本数据进行了测试,其中 14 条数据与模型结果一致,准确率达到 78%。结果表明,该模型具有较好的预测效果,可为电影票房预测提供一定的参考

价值。通过对电影票房的预测和分析，影响票房预测的最重要的因素是演员。演员流量程度越高，其主演的电影票房水平就越高。因此，选择受欢迎程度高和具有票房号召力的演员才是票房收益的关键[14]。

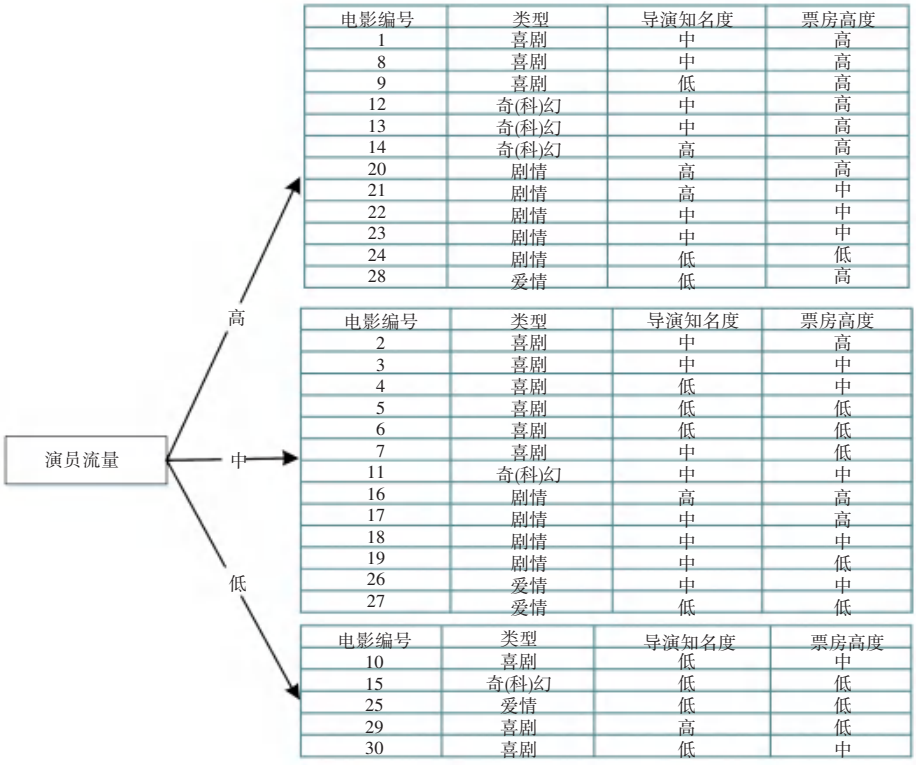

图 1　属性 actor 为根建立决策树分支

Fig. 1　Attribute actor creates decision tree branches for root

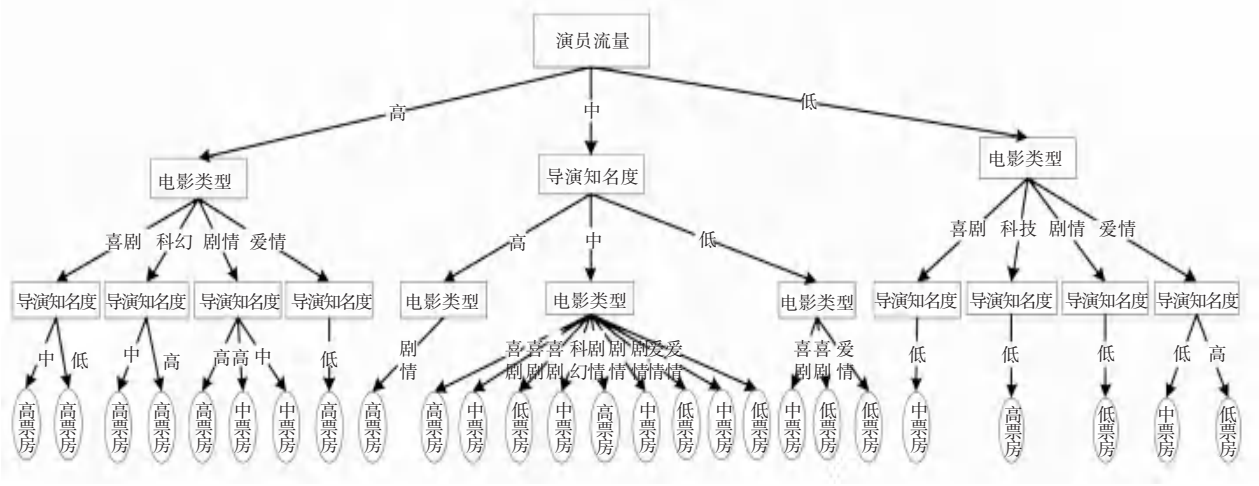

图 2　电影票房预测模型建立的决策树

Fig. 2　Decision tree established for a movie box office forecasting model

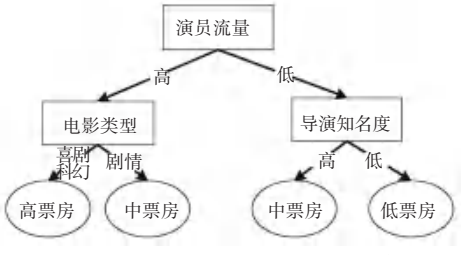

图 3　电影票房预测模型修剪后的决策树

Fig. 3　Shredded decision tree for movie box office prediction model

3　结束语

文章将决策树算法 C4.5 应用于电影票房的预测研究，通过对电影票房信息数据进行分析处理，建立完整的预测模型。实验结果说明，基于决策树算法的电影票房预测模型简单、快速，为电影票房的预测提供一定的科学依据[15]。不足之处在于模型中

（下转第 139 页）