电影评分影响因素分析

摘要：本文研究了通过 15 项指标对电影评分进行分类的问题，通过主成分分析和因子分析降维为 5 个主成分和 5 个共性因子，构建了电影分类评估模型，实现了高维度大样本条件下的综合定量分析。采用Standard Scaler 正则化去除电影数据大量有偏分布和极端值，只保留数据的排序关系。用 5 折交叉验证方法优化参数，网格搜索进行超参数整定，得出最优参数。将判别分析、逻辑回归模型和SVM等 5 种分类器结果比较，通过混淆矩阵直观展现误判率，得SVM模型有最高的准确率，逻辑回归有最高的精确度和召回率。进一步，利用聚类分析对电影无监督分类，基于电影相似度为推荐系统提供建议。

一、背景介绍

在全球新兴电影市场特别是中国等亚洲电影市场的带动下，全球电影市场近年来保持着稳定增长的基本态势。根据统计数据显示，2017 年全球电影票房达 406 亿美元，同比 2016 年增长约 5%，创下历史新高。作为全球第二大电影市场——中国正以 30%左右的平均增长速度领跑全球市场，2017 年票房为 559.11 亿元人民币，票房同比增长 13.45%。

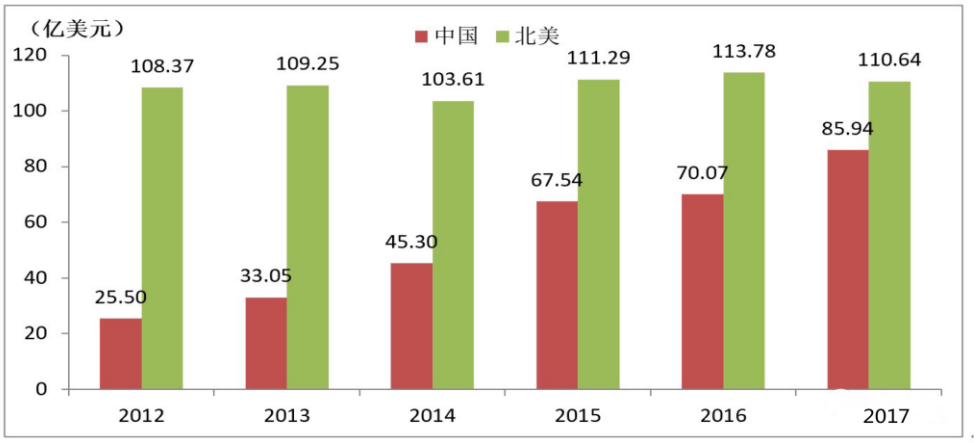

图 1-1 中国和北美 2012 至 2017 年票房

目前，用户对电影、视频网站的要求越来越多，不仅希望有丰富的电影资源，还希望能够快速找到自己感兴趣的电影资源。为了满足用户的需求，许多视频网站已经开始研究电影分类算法。目前比较流行且常用的两大类分类学习算法有：一类是单分类算法，如逻辑回归、决策树、神经网络及支持向量机等；另一类是集成算法。我们希望通过本文的研究，能够将电影根据评分高低进行分类，从而推荐评分较高的电影。

二、数据说明

本报告使用的是 Kaggle IMDB5000 电影数据集，共 4803 个样本。在数据预处理阶段，上映年份只有一个缺失值，电影时长有两个缺失值，均可通过网络查询真实值进行填充。提取原始数据中上映日期的年份作为新变量，命名为：上映年份；评分的中位数 6.6，本文将评分大于 6.6 的标记为 1，代表高分电影。反之，标记为 0。选取 15 个自变量分析电影评分影响因素，具体的变量说明如表 2-1 所示。

表 2-1 数据变量说明表

变量类型		变量名	详细说明	取值范围
因变量		评分	分类变量	0,1
自变量	外部因素	收入	连续变量	[162,7.6e+08]
		点赞数	包含三名主演、导演、编剧等人的 facebook 点赞数	[0,64w]
		评分数	连续变量	[5,1.6e+06]
		评论数	包含普通影迷和专家	[1,5060]
		上映年份	连续变量	[1929,2017]
		海报人物个数	连续变量	[0,10]
	内部因素	预算	连续变量	[218,1.2e+10]
		电影时长	连续变量	[0,276]
		纵横比	连续变量	[1.18,2.35]

三、描述性统计

在对电影评分的影响因素进行模型探究之前，首先对各变量进行描述性分析，为后续研究做铺垫。

（一）数值变量

在本案例中，评论数最大值为 13752，对应的电影是《盗梦空间》，《黑暗骑士》、《复仇者联盟》、《阿凡达》紧随其后，分别有着 12002、11776 和 11800 个评论。有 63 部电影的评

论数为 0，如《黑水船运公司》。

通过电影评论数分布直方图（图 3-1）可以看到，各部电影评论数差异较大。评论数的均值为 690，中位数为 235，中位数与均值差异较大，符合基本认知，即存在少数优质电评论数较高，拉高了整体的平均值。

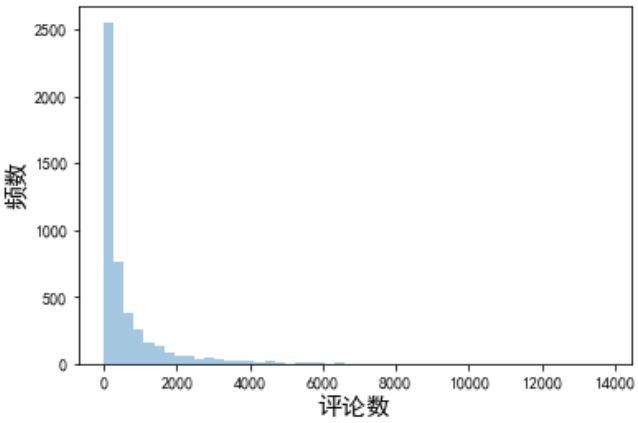

图 3-1 电影评论数分布直方图

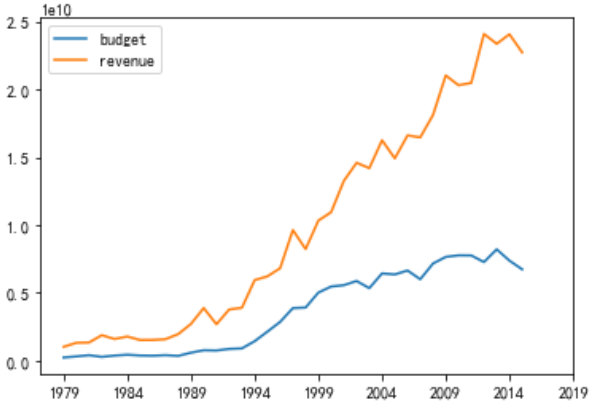

图 3-2 各年电影预算及收益折线图

与此同时，电影的预算和收益也逐年增长，收益与预算的差值越来越大，即净利润逐年增长，是电影业蓬勃发展的一个原因，越来越多人力和资金被投入到电影事业，导出越来越多受人们喜爱的电影，越来越多人去影院看电影。

电影分为很多类型，数据集中提到的有 20 种，统计各类型电影出现次数，选取出前 10 个类型，剧情、喜剧、惊悚、动作、爱情、冒险类电影数量最多，也是目前电影院上映电影的常见类型，剧情类电影数量达到 2297 部。

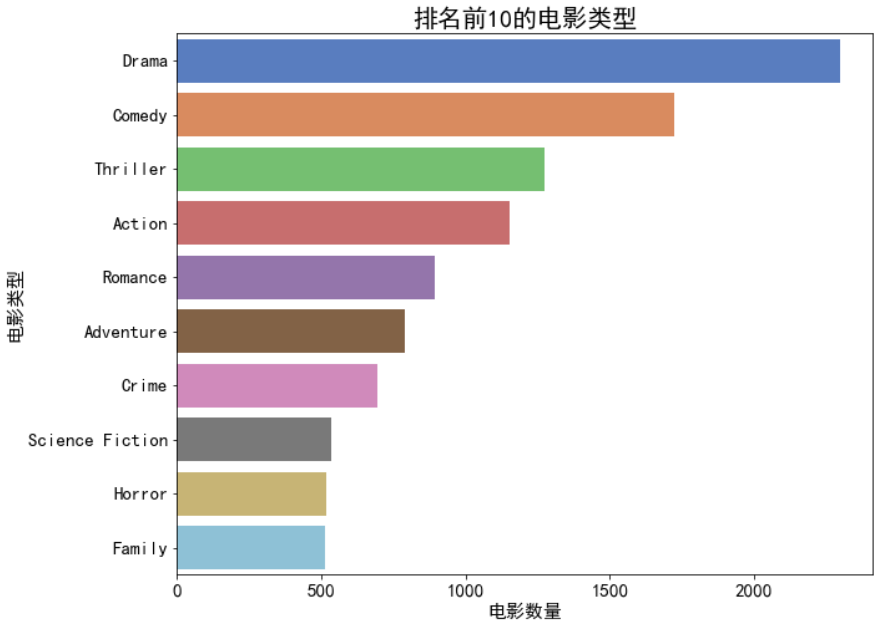

图 3-3 排名前 10 的电影类型

(二)文本变量

众多观影者会关注一部电影的导演，导演的代表作越多，其新执导的电影更容易有超高的票房。执导电影数最多的导演为 Steven Spielberg（史蒂文·斯皮尔伯格），导演了 26 部电影，代表作为《辛德勒的名单》、《拯救大兵瑞恩》等，获得过美国电影电视金球奖终身成就奖和 2018 帝国电影奖终身成就奖。

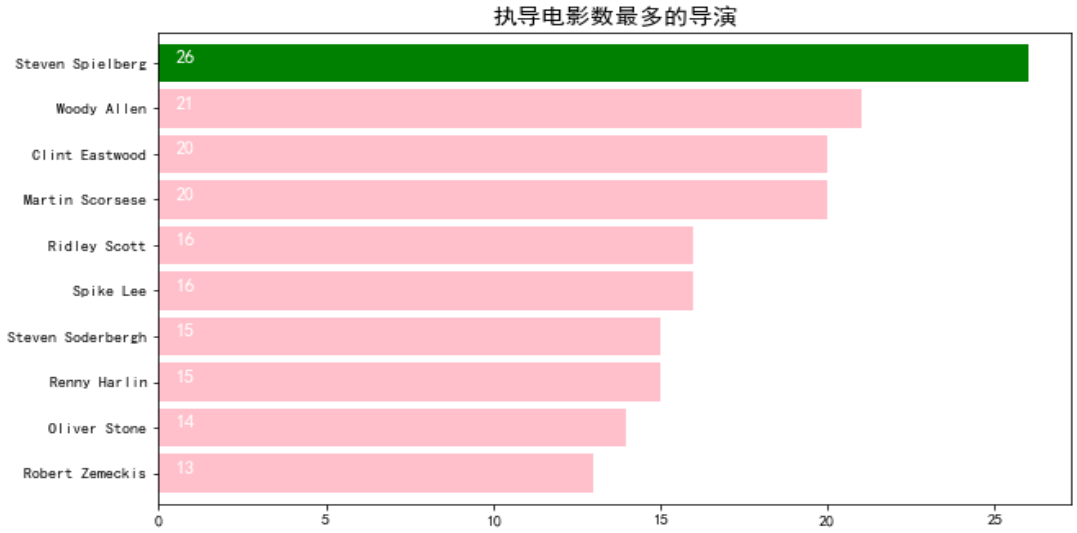

图 3-4 执导电影数排名前 10 的导演

演员同样也是一部电影的重要组成部分，Robert De Niro（罗伯特·德尼罗）是出演电影次数最多的演员，共出演 46 次，代表作品有《教父 2》、《出租车司机》等。

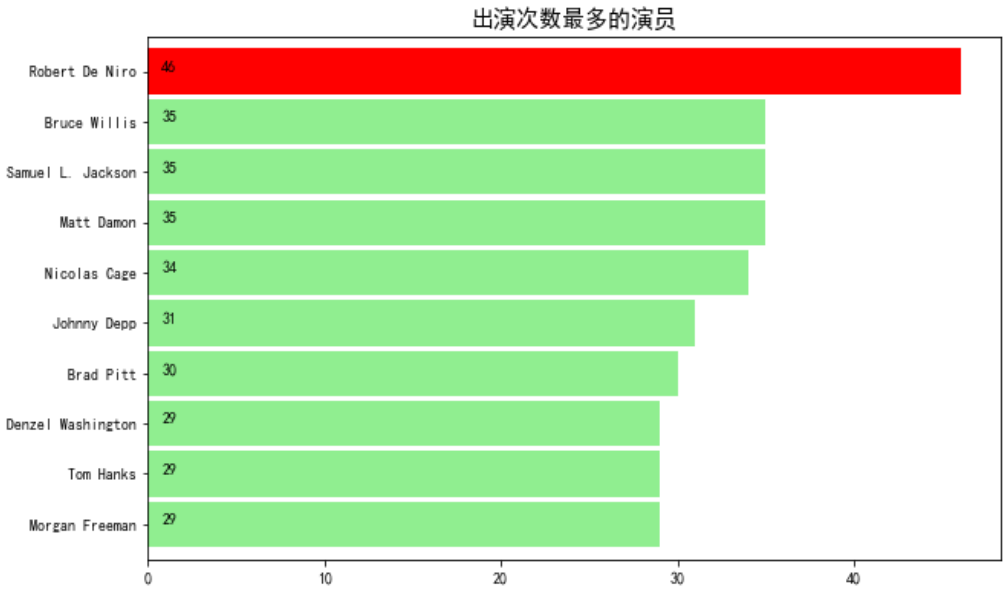

图 3-5 出演次数排名前 10 的演员

四、特征工程

（一）主成分分析

主成分分析是一种通过降维技术把多个变量化为少数几个主成分（综合变量）的统计分析方法。提取出的主成分能够反映原始变量的绝大部分信息，通常表示为原始变量的某种线性组合，两两不相关。

设 $x=(x_1,x_2,...,x_p)'$ 为一个 p 维随机向量，并假定二阶矩存在，记 $\mu=E(x)$，$\Sigma=V(x)$。

$$\begin{cases} y_1=a_{11}x_1+a_{21}x_2+...+a_{p1}x_p=a_1'x \\ y_2=a_{12}x_1+a_{22}x_2+...+a_{p2}x_p=a_2'x \\ ... \\ y_p=a_{1p}x_1+a_{2p}x_2+...+a_{pp}x_p=a_p'x \end{cases}$$

如果用一个综合变量来代表原始的 p 个变量，为使其在 $x_1,x_2,...,x_p$ 的一切线性组合中最具代表性，应使其方差最大化，以最大限度地保留这组变量的方差和协方差结构的信息。因此 x 的第 i 主成分 $y_i=a_i'x$ 是指，在约束条件 $\|a_i\|=1$ 和 $Cov(y_k,y_i)=0,\ k=1,2,...,i-1$ 下寻找 a_i，使得 $V(y_i)=a_i'\Sigma a_i$ 达到最大，$i=1,2,...,p$。$V(y_i)=\lambda_i$，主成分 y_i 的贡献率为 $\lambda_i/\sum_{j=1}^{p}\lambda_j$，主成分 $y_1,y_2,...,y_m$ 的累计贡献率为前 m 个主成分的贡献率之和，即 $\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i$。Σ 需要通过样本来估计。

$$S=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'=(s_{ij})$$

对 15 个自变量进行主成分分析，特征值如表 4-1 所示。

表 4-1 特征值

数量	1	2	3	4	5	6	7	8
特征值	4.271	2.057	1.427	1.036	1.013	0.991	0.880	0.834
数量	9	10	11	12	13	14	15	
特征值	0.696	0.503	0.457	0.417	0.245	0.172	0.002	

选取特征值大于 1 的主成分（或因子），即在后续分析中，主成分（或因子）数量取 5。

霍特林 T^2 统计量描述了数据集（观测样本矩阵）中的每一个观测与数据集的中心之间的距

离，可用于寻找远离数据集中心的极端观测数据。将该统计量从小到大排序，发现 The Final Destination、Sholem Aleichem: Laughing in the Darkness、Anchorman: The Legend of Ron Burgundy 和 The Host 四部电影的霍特林T^2统计量最大，分别为 1327.33、1707.19、1970.07 和 2983.80，而其余电影的统计量均小于 600，所以上述四部电影可视为异常值。

表 4-2 前五个特征值、特征向量以及贡献率

特征向量	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$	$\hat{t}_4$	$\hat{t}_5$
电影时长	0.196	-0.159	-0.285	0.534	-0.034
上映年份	0.122	0.043	0.694	-0.068	-0.058
海报人物个数	0.011	0.158	0.127	0.458	0.683
导演点赞数	0.156	-0.093	-0.248	0.148	-0.139
演员 1 点赞数	0.239	0.476	-0.095	-0.011	-0.268
演员 2 点赞数	0.281	0.355	-0.039	-0.033	0.098
演员 3 点赞数	0.271	0.253	-0.017	-0.043	0.269
点赞总数	0.305	0.508	-0.090	-0.025	-0.152
电影点赞数	0.323	-0.171	0.282	-0.112	0.120
评分用户数	0.381	-0.263	-0.163	-0.062	0.069
专家评论数	0.367	-0.210	0.300	-0.113	-0.006
评论用户数	0.341	-0.293	-0.174	-0.039	0.008
预算	0.060	-0.054	0.043	0.018	-0.423
收入	0.326	-0.181	-0.119	-0.136	0.099
纵横比	0.098	-0.037	0.314	0.653	-0.354
特征值	4.271	2.057	1.427	1.036	1.013
贡献率	0.285	0.137	0.095	0.069	0.068
累计贡献率	0.285	0.422	0.517	0.586	0.654

为了研究十五个原始变量间是否存在多重共线性，观察最末一个主成分，计算结果为

$$\hat{\lambda}_{15} = 0.0018$$

$$\hat{t}_{15} = (\text{-0.0007,-0.0019,-0.0007,4.6416e-05,-0.6162,-0.1822,-0.1119,0.7579,-0.0012,}$$

$$\text{0.004,0.003,-0.00015,0.0003,-0.0134,0.0003})'$$

由于$\hat{\lambda}_{15}$非常小，所以存在这样一个多重共线性关系：

$$\text{-0.0007}x_1^* \text{-0.0019}x_2^* \text{-0.0007}x_3^* \text{+4.6416e-05}x_4^* \text{-0.6162}x_5^* \text{-0.1822}x_6^* \text{-0.1119}x_7^* \text{+0.7579}x_8^*$$

$$\text{-0.0012}x_9^* \text{+0.004}x_{10}^* \text{+0.003}x_{11}^* \text{-0.00015}x_{12}^* \text{+0.0003}x_{13}^* \text{-0.0134}x_{14}^* \text{+0.0003}x_{15}^* \approx 0$$

（二）因子分析

因子分析可看作是对主成分分析的推广和发展，也是一种重要的降维方法。因子分析是

指研究从变量群中提取共性因子的统计技术，可减少变量的数目。原始变量是因子的线性组合。因子分析的目的是，试图用几个潜在的、不可观测的随机变量（因子）来描述原始变量间的协方差或相关关系。因子的解可以有很多，表现得较为灵活（主要体现在因子旋转上），使得变量在降维后更易得到解释。

设有 p 维可观测的随机向量 $\mathrm{x}=(x_1,x_2,...,x_p)'$，其均值为 $\mu=(\mu_1,\mu_2,...,\mu_p)'$，协方差矩阵 $\Sigma=(\sigma_{ij})$。因子分析的一般模型为

$$\begin{cases} x_1=\mu_1+a_{11}f_1+a_{12}f_2+...+a_{1m}f_m+\varepsilon_1 \\ x_2=\mu_2+a_{21}f_1+a_{22}f_2+...+a_{2m}f_m+\varepsilon_2 \\ ... \\ x_p=\mu_p+a_{p1}f_1+a_{p2}f_2+...+a_{pm}f_m+\varepsilon_p \end{cases}$$

其中 $f_1,f_2,...f_m$ 为公共因子，$\varepsilon_1,\varepsilon_2,...\varepsilon_p$ 为特殊因子，都是不可观测的随机变量。用矩阵、向量表示为

$$x=\mu+Af+\varepsilon$$

其中 $f=(f_1,f_2,...f_m)'$ 为公共因子向量，$\varepsilon=(\varepsilon_1,\varepsilon_2,...\varepsilon_p)'$ 为特殊因子向量，$A=(a_{ij}):p\times m$ 称为因子载荷矩阵。假定

$$\begin{cases} E(f)=0 \\ E(\varepsilon)=0 \\ V(f)=I \\ V(\varepsilon)=D=diag(\sigma_1^2,\sigma_2^2,...,\sigma_p^2) \\ Cov(f,\varepsilon)=E(f\varepsilon')=0 \end{cases}$$

即公共因子彼此不相关且具有单位方差，特殊因子也彼此不相关且和公共因子也不相关。

设 $x_1,x_2,...,x_n$ 是一组 p 维样本，μ 和 Σ 的估计值为

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad S=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'$$

需要估计因子载荷矩阵 $A=(a_{ij}):p\times m$ 和特殊方差矩阵 $D=diag(\sigma_1^2,\sigma_2^2,...,\sigma_p^2)$，常用的三种参数估计方法为主成分法、主因子法和极大似然法。

因子模型的参数估计完成后，需要对模型中的公共因子进行合理的解释，因子的解释带有一定的主观性，可通过旋转因子的方法来减少主观性且使之容易解释。最常见的因子旋转

方法为正交旋转中的最大方差旋转法。

$$\mathrm{d}_{ij}=\frac{a_{ij}^{*}}{h_i},\ \bar{d}_j=\frac{1}{p}\times\sum_{i=1}^{p}\mathrm{d}_{ij}^{\ 2}$$

上式消除了公共因子对各原始变量的方差贡献不同的影响，取平方可以消除 d_{ij} 符号不同的影响。其中 $A^{*}=(a_{ij}^{\ *})$，则 A^{*} 的第 j 列元素平方的相对方差可定义为

$$V_j=\frac{1}{p}\sum_{i=1}^{p}(\mathrm{d}_{ij}^{\ 2}-\bar{d}_j)^2$$

最大方差旋转法选择正交矩阵 T，使得矩阵 A^{*} 所有的 m 个列元素平方的相对方差之和 $V=V_1+V_2+...V_m$ 达到最大。

表 4-3 m=5 时旋转后的主成分解

变量	因子载荷				
	f_1	f_2	f_3	f_4	f_5
电影时长	0.423	0.054	-0.420	0.509	0.108
上映年份	0.066	0.098	0.846	0.182	-0.015
海报人物个数	-0.068	0.097	0.057	0.180	0.845
导演点赞数	0.328	0.108	-0.288	0.188	-0.117
演员 1 点赞数	-0.001	0.870	-0.020	0.098	-0.168
演员 2 点赞数	0.216	0.7346	0.063	-0.012	0137
演员 3 点赞数	0.304	0.584	0.086	-0.070	0.272
点赞总数	0.103	0.972	0.012	0.066	-0.063
电影点赞数	0.662	0.118	0.432	0.069	0.044
评分用户数	0.882	0.142	-0.081	0.048	-0.042
专家评论数	0.744	0.132	0.460	0.140	-0.080
评论用户数	0.829	0.067	-0.113	0.077	-0.096
预算	0.065	0.037	0.048	0.212	-0.391
收入	0.733	0.170	-0.021	-0.048	-0.021

纵横比	0.011	0.024	0.208	0.839	-0.06
累计贡献率	0.230	0.410	0.507	0.584	0.654

表 4-4 m=5 时旋转后的主因子解

变量	因子载荷				
	f_1	f_2	f_3	f_4	f_5
电影时长	0.430	0.042	-0.005	-0.010	0.048
上映年份	-0.120	0.047	0.602	0.101	0.017
海报人物个数	-0.078	0.052	0.034	0.130	0.031
导演点赞数	0.307	0.057	0.009	0.009	0.042
演员 1 点赞数	0.101	0.986	0.009	0.079	0.041
演员 2 点赞数	0.203	0.304	0.130	0.354	0.821
演员 3 点赞数	0.275	0.146	0.079	0.866	0.167
点赞总数	0.179	0.893	0.111	0.298	0.262
电影点赞数	0.422	0.028	0.598	0.119	0.024
评分用户数	0.860	0.075	0.213	0.010	0.021
专家评论数	0.520	0.052	0.697	0.055	0.028
评论用户数	0.809	0.035	0.167	-0.048	0.011
预算	0.086	0.001	0.070	0.010	0.010
收入	0.637	0.060	0.178	0.111	0.051
纵横比	0.056	0.034	0.241	-0.004	0.025
累计贡献率	0.181	0.308	0.403	0.472	0.524

表 4-5 m=5 时旋转后的极大似然解

变量	因子载荷				
	f_1	f_2	f_3	f_4	f_5
电影时长	0.386	0.038	0.047	0.004	0.039
上映年份	-0.121	0.048	0.563	0.074	0.022

海报人物个数	-0.057	0.052	-0.001	0.118	0.032
导演点赞数	0.301	0.054	0.036	0.009	0.039
演员 1 点赞数	0.105	0.985	0.091	0.094	0.042
演员 2 点赞数	0.199	0.292	0.149	0.372	0.845
演员 3 点赞数	0.275	0.127	0.146	0.864	0.142
点赞总数	0.181	0.887	0.132	0.317	0.250
电影点赞数	0.396	0.026	0.618	0.067	0.036
评分用户数	0.900	0.071	0.209	-0.031	0.032
专家评论数	0.470	0.045	0.831	-0.006	0.033
评论用户数	0.811	0.026	0.209	-0.060	0.014
预算	0.064	0.001	0.087	0.011	0.007
收入	0.642	0.047	0.196	0.143	0.036
纵横比	0.048	0.034	0.196	-0.000	0.018
累计贡献率	0.179	0.304	0.413	0.483	0.538

求出残差矩阵 $\hat{R}-(\hat{A}\hat{A}'+\hat{D})$ 并加和，得到主成分解的误差平方和为 1.268909，主因子解的误差平方和为 0.1454512，极大似然估计解的误差平方和为 0.1534746，因为 0.1454512<0.1534746<1.268909, 极大似然解和主因子解都比主成分解拟合的好，主因子解和极大似然解类似，主因子解拟合的略好一点。

根据误差平方和的结果，主要根据主因子解和极大似然解来解释共性因子。从表 4-4 和表 4-5 中可以看到，评分用户数、评论用户数和收入在因子 f_1 上都具有大的正载荷，电影时长、导演点赞数、电影点赞数和专家评论数具有中等的正载荷，其他变量只有小的载荷，由于评论用户和评论用户大多数是贡献了票房的，因此 f_1 可称为盈利因子。在因子 f_2 上，演员 1 点赞数、点赞总数有大的正载荷，演员 2 点赞数和演员 3 点赞数有中等偏小的正载荷，其他变量载荷均小于 0.1，因此 f_2 可称为主角光环因子。在因子 f_3 中，电影点赞数、专家评论数和上映年份有较大的正载荷，其余变量载荷均较小，近年来，越来越多人喜欢电影 ，且有自己的见解，因此 f_3 可称为专业因子。在因子 f_4 中，演员 3 点赞数有大的正载荷，除此之外，演员 2 点赞数和点赞总数有中等正载荷，其余变量载荷均非常小，在现实生活中，无论是电影还是电视剧，能够给观众留下深刻印象，突然爆红的往往不是主角，而是饰演配

角的娱乐圈新人，因此f_4可称为新潮因子。在因子f_5中，演员 2 点赞数有大的正载荷，点赞总数和演员 3 点赞数有中等偏小的正载荷，其余变量载荷均特别小，因此f_5可称为配角因子，综上所述，演员点赞数是非常重要的一个变量，对评分会有较大的影响，这也与我们的日常认知相同。

五、模型建立

（一）判别分析

线性判别分析和二次线性判别分析是两个经典的分类器，分别是线性和二次决策表面。

这两个分类器没有需要调整的超参数，计算简便。线性判别分析只能学习线性边界，而二次判别分析可以学习二次边界。因此，二次判别分析风具有灵活性。

本文挖掘数据已有分布$P(X|y=k)$（k 为类别），作为先验信息，利用最大后验概率法进行 Bayes 判别。

$$P(y=k|X)=\frac{P(X|y=k)P(y=k)}{P(X)}=\frac{P(X|y=k)P(y=k)}{\sum_l P(X|y=l)P(y=l)}$$

最大后验概率法是采用如下的判别规则：

$$x\in\pi_l, 若P(\pi_l|x)=\max_{1\le i\le k}P(\pi_i|x)$$

对于线性判别分析和二次线性判别分析,$P(X|y)$为多元正态分布：

$$P(X|y=k)=\frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}}\exp(-\frac{1}{2}(X-\mu_k)^t\Sigma_k^{-1}(X-\mu_k))$$

其中，d 为特征个数。

线性判别分析中，假设协方差矩阵相同,$\Sigma_k=\Sigma$,则线性决策边界可以如下表示：

$$\log\left(\frac{P(y=k|X)}{P(y=l|X)}\right)=\log\left(\frac{P(X|y=k)P(y=k)}{P(X|y=l)P(y=l)}\right)=0\Leftrightarrow$$

$$(\mu_k-\mu_t)^t\Sigma^{-1}X=\frac{1}{2}(\mu_k^t\Sigma^{-1}\mu_k-\mu_l^t\Sigma^{-1}\mu_l)-\log(\frac{P(y=k)}{P(y=l)})\ .$$

二次判别分析中，并未假设协方差矩阵相同。我们基于以上算法，建立判别模型来预测一部电影的评分是否为 1（电影的评分为 1 表明评分值超过 6.6）。根据特征工程主成分分析选取的 5 个变量来判断电影的评分高低。将电影的历史数据作为判别分析的训练集，根据用户反馈信息和电影特征属性建立一个分类模型，估计用户评分值。

在将清洗干净的数据代入分类器进行训练前，本文首先对数据进行正则化处理。采用

Standard Scaler 正则化去除电影数据大量有偏分布和极端值，只保留数据的排序关系。正则化的过程是把数据集中的每个样本所有数值缩放到(-1,1)之间，将每个样本缩放到单位范数。

准确率是预测正确的个数除以总样本个数，表明了预测值和真实值的差异，反映了分类器对整个样本的判定能力。在样本不均衡的时候，即使精度可以达到很高，但是模型没有用处，准确率具有欺骗性。因此，本文又引入精确度和召回率来综合评估模型预测效果。精确度反映了被分类器判定的正例中真正的正例样本的比重。召回率反映了被正确判定的正例占总的正例的比重。

用混淆矩阵来表示分类器的评估效果。

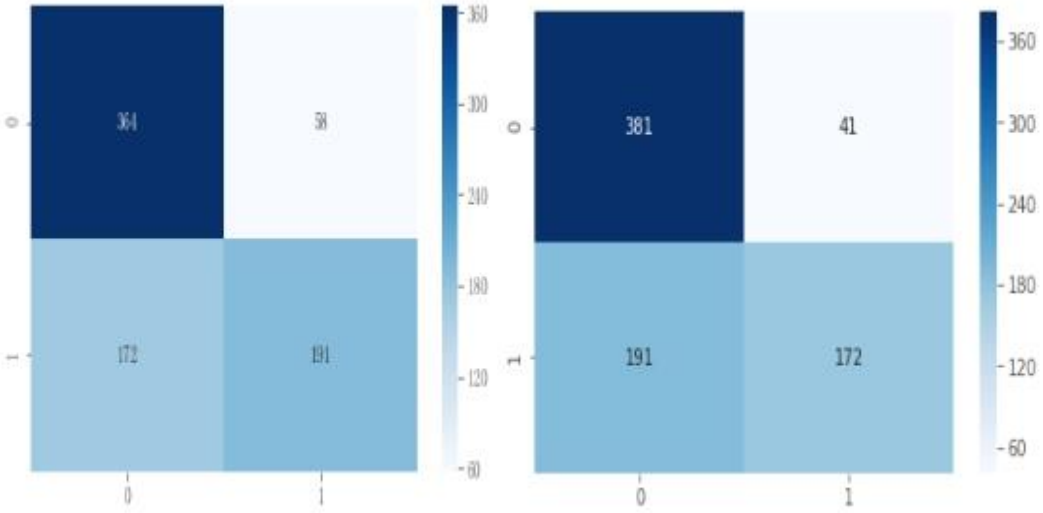

图 5-1 混淆矩阵

混淆矩阵展示了分类器的误判率。右上角是误杀值，左上角和右下角为成功预测的样本个数。左图为线性判别分析的结果，误判率为$\hat{P}(1|0)=\frac{n(1|0)}{n_0}=14.07\%$,$\hat{P}(0|1)=\frac{n(0|1)}{n_1}=47.38\%$.右图为二次判别分析的结果，误判率为$\hat{P}(1|0)=\frac{n(1|0)}{n_0}=9.71\%$,$\hat{P}(0|1)=\frac{n(0|1)}{n_1}=52.61\%$。

为进一步比较判别分析分类器的效果，本文另外选取了 3 种算法：逻辑回归、支持向量机和朴素贝叶斯。

逻辑回归(LR)算法具有快捷、稳健、可解释性强的特点，是工业界最常用的模型之一。LR模型对变量关系的线性限制，使得其难以达到最优。在建模的过程中，增加L2 惩罚函数，减少过拟合。本文将LR模型作为基准，对数据预处理和模型表现作出快速评估。

为评估电影评分值，需要求解出 16 个参数$\theta_0,\theta_1,\ldots,\theta_{15}$,并根据所得参数值得到一个概率值。由概率值转化成 0,1 分类，这就是建立分类器的过程。在建立分类器之前，需设定阈

值，根据阈值判断违约结果。若设定阈值为 0.5，电影评分的概率值大于 0.5 被判定为高分电影，反之，被判定为低分电影。

为减少过拟合，本文采用 5 折交叉验证法估计模型参数。先进行洗牌操作，将样本数据随机的切分成 5 份互不相交的大小相同的子集；每次取 1 份为测试集，其他 4 份为训练集。对此训练集进行训练，得到一个模型。依次取不同的 1 份为测试集，重复上述过程 5 次，可得 5 个模型。最终选出 5 次评测中平均测试误差最小的模型，将测试集数据代入该模型，得出预测值。 利用网格搜索法在训练集上进行超参数整定，调整或更改默认参数值以获得更好的模型。

SVM是一个强大的分类算法，模型的精准度与参数集有很大关系。当C 很大时，虽然可以导致分错的点更少，但是过拟合情况会较严重。我们使用网格搜索法在训练集上寻找最好的参数，得出惩罚因子C为 0.5，核函数为线性核函数最佳。同理，我们得出朴素贝叶斯分类算法的最优参数。

针对imdb电影数据，分别采用上述 5 种算法进行分类，给出不同模型下的结果对比，见表 5-1。

表 5-1 不同模型结果分析

分类器	准确率(%)	精确度(%)	召回率(%)
逻辑回归	76.07	72.36	58.68
线性判别分析	76.71	70.70	52.62
二次判别分析	80.75	70.45	47.38
支持向量(SVM)	82.57	71.85	49.59
朴素贝叶斯	72.04	66.38	41.76

由表 5.1 结果可得，二次判别分析的准确率较高，支持向量机有最高的准确率，逻辑回归有较高精确度和召回率。权衡准确率、精确度和召回率三项评估指标，逻辑回归的表现效果最好。

（二）聚类分析

（1）Ward 方法

Ward 方法属于聚集系统聚类，思想为：初始 n 个样本各自作为一类，将距离最近的两类合并成新类，计算新类与其他类的距离；重复进行两个最近的类的合并，每次减少一个类，直至所有的样本合并为一类。

类中各个样本到类重心的平方欧式距离之和称为类内离差平方和。设类G_K和G_L合并成新类G_M，则离差平方和分别是：

$$W_K = \sum_{i \in G_K} \left(x_i - \bar{x}_K\right)^T \left(x_i - \bar{x}_K\right)$$

$$W_L = \sum_{i \in G_L} \left(x_i - \bar{x}_L\right)^T \left(x_i - \bar{x}_L\right)$$

$$W_M = \sum_{i \in G_M} \left(x_i - \bar{x}_M\right)^T \left(x_i - \bar{x}_M\right)$$

对于固定的类内样品数，它们反映了各自类内样本的分散程度。如果两个类相距较近，则合并后增加的离差平方和较小；否则，应较大。于是我们定义G_K，G_L之间的平方距离为：

$$D_{KL}^2 = W_M - W_L - W_K$$

（2）k-means 方法

k-means 聚类的目的是：把 n 个点（可以是样本的一次观察或一个实例）划分到 k 个聚类中，使得每个点都属于离他最近的均值（此即聚类中心）对应的聚类，以之作为聚类的标准。这个问题将归结为一个把数据空间划分为 Voronoi cells 的问题。

这个问题在计算上是 NP 困难的，不过存在高效的启发式算法。一般情况下，都使用效率比较高的启发式算法，它们能够快速收敛于一个局部最优解。这些算法通常类似于通过迭代优化方法处理高斯混合分布的最大期望算法（EM 算法）。而且，它们都使用聚类中心来为数据建模；然而 k-平均聚类倾向于在可比较的空间范围内寻找聚类，期望-最大化技术却允许聚类有不同的形状。

已知观测集$\left(x_1, x_2, \cdots, x_n\right)$，其中每个观测都是一个 d-维实向量，k-means 聚类要把这 n 个观测划分到 k 个集合中(k≤n),使得组内平方和（WCSS within-cluster sum of squares）最小。换句话说，它的目标是找到使得下式满足的聚类S_i。

$$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \left\| x - u_i \right\|^2$$

其中u_i是S_i中所有点的均值。

给定初始的 k 个均值点 $m_1^{(1)},\cdots,m_k^{(1)}$,算法的按照下面两个步骤交替进行[7]：

分配：将每个观测分配到聚类中，使得组内平方和（WCSS）达到小。 因为这一平方和就是平方后的欧氏距离，所以很直观地。把观测分配到离它近得均值点即可[8]。（数学上，这意味依照由这些均值点生成的 Voronoi 图来划分上述观测）。

$$S_i^{(t)}=\left\{x_p:\left\|x_p-m_i^{(t)}\right\|^2\le\left\|x_p-m_j^{(t)}\right\|^2,\forall j,1\le j\le k\right\}$$

其中每个 x_p 都只被分配到一个确定的聚类 S_t 中，尽管在理论上它可能被分配到 2 个或者更多的聚类。

更新：对于上一步得到的每一个聚类，以聚类中观测值的图心，作为新的均值点。

$$m_i^{(t+1)}=\frac{1}{\left|S_i^{(t)}\right|}\sum_{x_j\in S_i^{(t)}}x_j$$

因为算术平均是最小平方估计，所以这一步同样减小了目标函数组内平方和（WCSS）的值。

算法将在对于观测的分配不再变化时收敛。由于交替进行的两个步骤都会减小目标函数 WCSS 的值，并且分配方案只有有限种，所以算法一定会收敛于某一（局部）优解。注意：但使用这一算法无法保证得到全局优解。

这一算法经常被描述为“把观测按照距离分配到近的聚类”。标准算法的目标函数是组内平方和（WCSS），而且按照“ 小平方和”来分配观测，确实是等价于按照小欧氏距离来分配观测的。如果使用不同的距离函数来代替（平方）欧氏距离，可能使得算法无法收敛。然而，使用不同的距离函数，也能得到 k-均值聚类的其他变体，如球体 k-均值算法和 k-中心点算法。

将 imdb 电影数据作为数据集，首先对数据进行规范化处理，使得每个的特征的范数统一为 1，对规范化后的数据分别进行系统聚类以及动态聚类：

系统聚类：采用聚集型层次聚类方法，选取 ward 方法作为簇间距离（离差平方和法）。

动态聚类：采用 k-means 方法，以 k-means++方法选取初始点（使得初始聚类中心相聚较远），距离为欧式距离。

首先令聚类簇数 k 为 2，并对数据降维至二维平面显示聚类结果如下：

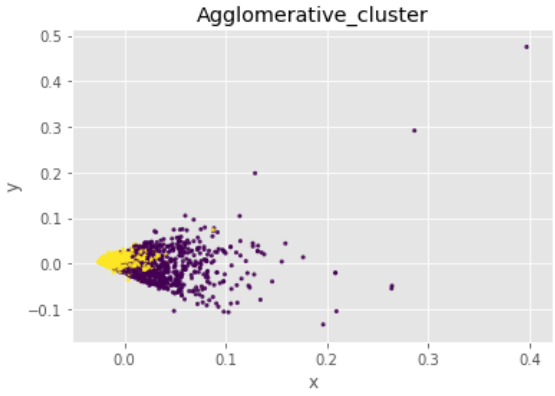

图 5-2 ward 方法聚类

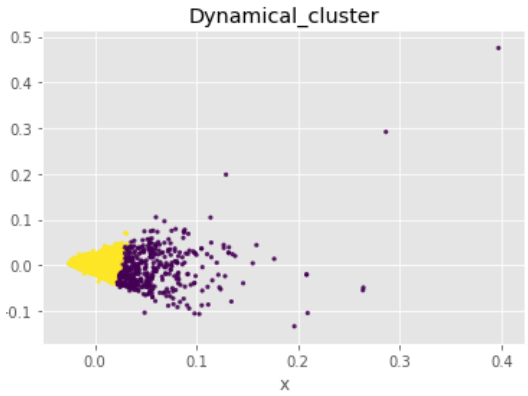

图 5-3 k-means 聚类

可见，k-means 聚类分界线明显，效果较好，而层次聚类类别间界限较模糊。

分别对两个类别的数据进行预览，做出箱线图：

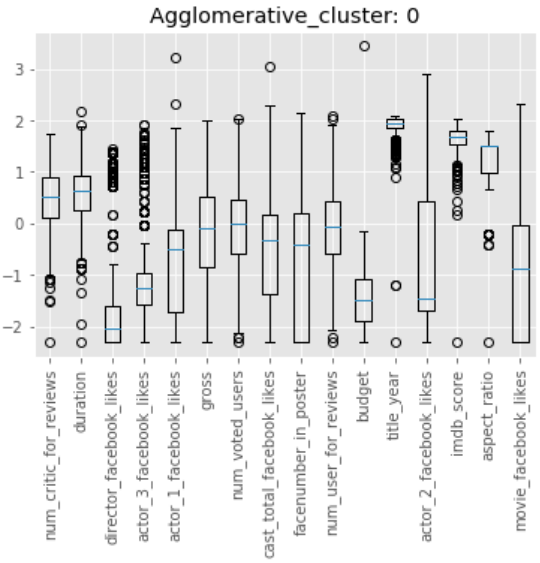

图 5-4 ward 方法类别 0 各特征箱线图

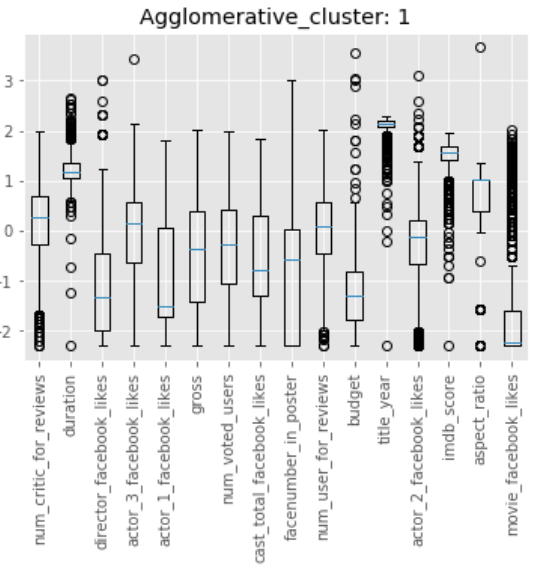

图 5-5 ward 方法类别 1 各特征箱线图

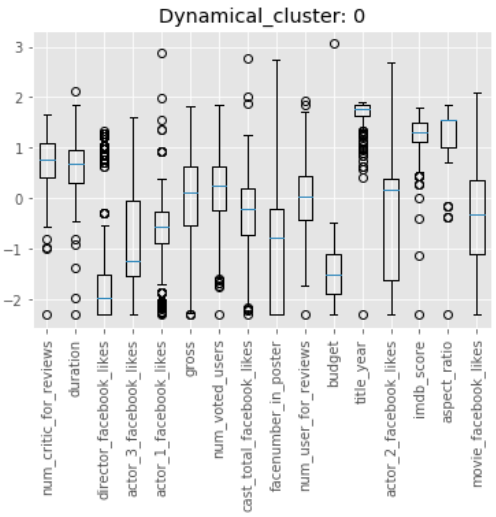

图 5-6 k-means 方法类别 0 各特征箱线图

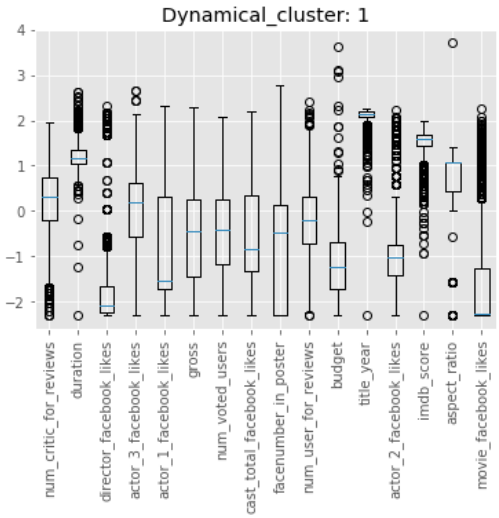

图 5-7 k-means 方法类别 1 各特征箱线图

可以看出：类型 0 的电影 facebook 喜爱数，imdb 评分，收入，投票数等明显较高，而类型 1 的导演以及演员的 facebook 喜爱数却相反较类型 0 高，此时可以解释聚类结果为类型 0 为实力内涵型电影，而类别 1 为演员人气型电影。

具体绘制两个类别电影 Facebook 喜爱数的直方图进行对比：类别 0 的喜爱数总体偏高。

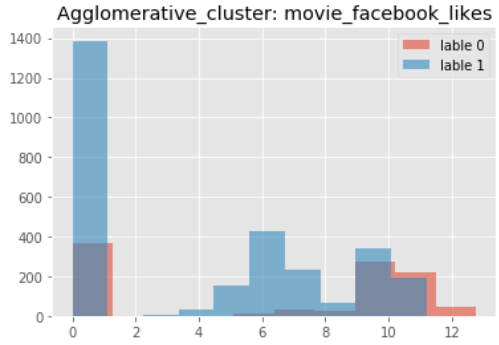

图 5-8 ward 电影 facebook 喜爱数直方图

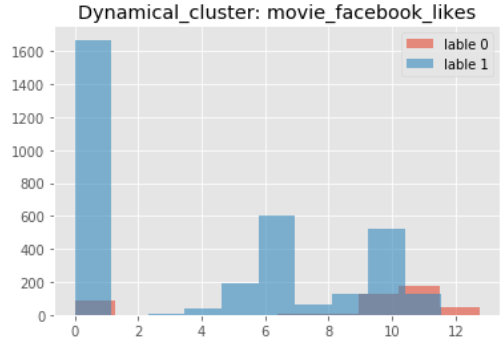

图 5-9 k-means 电影 facebook 喜爱数直方图

以及两个类别的 imdb 电影评分直方图：类别 0 评分同样整体偏高。同样也可以发现，高质量的电影数目较低质量电影数目少。

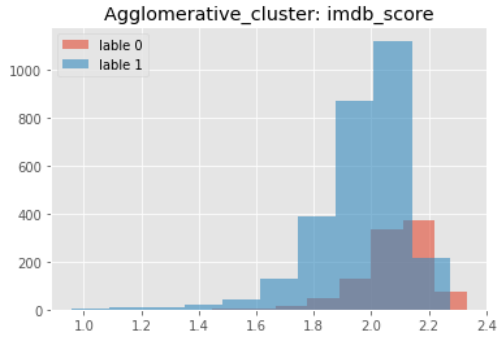

图 5-10 ward：imdb 评分直方图

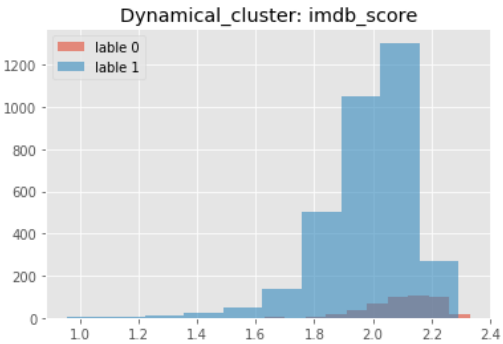

图 5-11 k-means：imdb 评分直方图

六、结论与建议

本案例对 1929 年-2017 年之间上映的部分电影进行统计分析，得到如下结论：

通过判别分析得出影响电影评分的主要因素有：（1）外部因素：评论数、点赞数均对电影的评分有正向影响；（2）内部因素：电影的预算、电影时长适中更倾向于有较高的评分。电影所属类别对评分值的影响有所区别。动画类、虚幻类、冒险类和科幻类评分最高，而纪录片、外国和电视电影类评分最低。

通过聚类分析可以基于相似度为用户推荐电影。为喜爱实力内涵型电影的用户推荐电影 facebook 点赞数，imdb 评分，收入，投票数等较高的电影；为喜欢演员人气型电影的用户推荐导演、演员 facebook 点赞数较高的电影。

参考文献

[1] Steinhaus, H. Sur la division des corps matériels en parties. Bull. Acad. Polon. Sci. 1957, 4 (12): 801–804. MR 0090073. Zbl 0079.16403.

[2] MacQueen, J. B. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press: 281– 297. 1967 [2009-04-07]. MR 0214227. Zbl 0214.46201 .

[3] Lloyd, S. P. Least square quantization in PCM. Bell Telephone Laboratories Paper. 1957. Published in journal much later: Lloyd., S. P. Least squares quantization in PCM (PDF). IEEE Transactions on Information Theory. 1982, 28 (2): 129–137 [2009-04-15]. doi:10.1109/TIT.1982.1056489 .

[4] E.W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. 1965, 21: 768–769.

[5] J.A. Hartigan. Clustering algorithms. John Wiley & Sons, Inc. 1975.

[6] Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A kMeans Clustering Algorithm. Journal of the Royal Statistical Society, Series C. 1979, 28 (1): 100–108. JSTOR 2346830 .

[7] MacKay, David. Chapter 20. An Example Inference Task: Clustering (PDF). Information Theory, Inference and Learning Algorithms. Cambridge University Press. 2003: 284–292. ISBN 0-521-64298-1. MR 2012999.

[8] Since the square root is a monotone function, this also is the minimum Euclidean distance assignment.

[9] Aloise, D.; Deshpande, A.; Hansen, P.; Popat, P. NPhardness of Euclidean sum-of-squares clustering. Machine Learning. 2009, 75: 245–249. doi:10.1007/s10994-009-5103-0 .

[10] Dasgupta, S. and Freund, Y. Random Projection Trees for Vector Quantization. Information Theory, IEEE Transactions on. July 2009, 55: 3229–3242. arXiv:0805.1390. doi:10.1109/TIT.2009.2021326 .

[11] Mahajan, M.; Nimbhorkar, P.; Varadarajan, K. The Planar k-Means Problem is NP-Hard. Lecture Notes in Computer Science. 2009, 5431: 274–285. doi:10.1007/978-3-642-00202-1_24 .

[12] Inaba, M.; Katoh, N.; Imai, H. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. Proceedings of 10th ACM Symposium on Computational

Geometry: 332–339. 1994. doi:10.1145/177424.178042

[13] Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine. 1901, 2(6): 559-572

[14] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. Journal of Educational Psychology, 1993, 24 : 417-441.

[15] “The Elements of Statistical Learning”, Hastie T., Tibshirani R., Friedman J., Section 4.3, p.106-119, 2008.

[16] Ledoit O, Wolf M. Honey, I Shrunk the Sample Covariance Matrix. The Journal of Portfolio Management 30(4), 110-119, 2004.

[17] 基于数据挖掘的我国 P2P 网络借贷违约预测模型研究[D]. 大连理工大学, 2016.

[18] 刘汉文, 陆佳佳. 2017 年中国电影产业发展分析报告[J]. 当代电影, 2018(3).