线性混合模型在短视频推荐中的应用

——以抖音为例

王艳明　常　煜　李雅楠
山东工商学院　统计学院　山东烟台　264005

摘　要：进入互联网时代，社交媒体在智能终端的承载下呈现出多元化的发展态势。精准的捕捉用户喜好，可以为用户打造个人化机制。本文以抖音短视频数据为例，讨论了线性混合模型在抖音短视频推荐中的应用，同时，与二元逻辑回归模型进行了比较。结果表明，线性混合模型可以用于短视频推荐中，能够预测出用户下次再看同一类视频的可能性大小。与二元逻辑回归相比，线性混合模型得到的预测效果更好，可以突破传统的推荐算法机制和应用于其他领域的推荐机制中。同时也为线性混合模型只能适用于数值型数据的难题进行了验证。

关键词：抖音短视频；线性混合模型；二元逻辑回归；预测概率

Application of linear hybrid model in short video recommendation

-- A case study of Douyin

Yanming Wang, Yu Chang, Yanan Li
Shandong Technology and Business University, School of Statistics, Shandong, Yantai 264005

Abstract: Entering the Internet era, social media has shown a diversified development trend under the support of intelligent terminals. Precisely capturing user preferences can create personalized mechanisms for users. This paper takes Douyin (TikTok) short video data as an example and discusses the application of the linear mixed model in Douyin short video recommendations, comparing it with the binary logistic regression model. The results indicate that the linear mixed model can be used for short video recommendations and can predict the likelihood of users watching similar videos again. Compared to binary logistic regression, the linear mixed model achieves better predictive performance, breaking through traditional recommendation algorithms and being applicable to other recommendation mechanisms. Additionally, it validates the challenge of using linear mixed models solely for numerical data.

Keywords: Douyin short video; Linear mixed model; Binary logistic regression; Prediction probability

一、引言

1.抖音短视频的发展和应用

在互联网时代里，手机的使用已经成为人们日常生活中不可替代的一部分。尤其是进入到数字信息化时代，短视频的发展迅速尤其迅速。像火山短视频、西瓜视频、快手、抖音短视频等在市场取得不错的成绩。短视频的最大优点是，用户可以利用上班的休息时间和其他碎片化时间，那这些抖音短视频吸引用户的背后心理依据到底是什么？这个时候，弄清楚用户行为是非常重要的，以此为抖音短视频运营商提供精准化策略和满足用户的个性化和多样化需求。

2.研究内容和研究方法

随着，个性化的推荐越来越多的应用于各个领域中。首先是通过对原始数据的描述统计分析，将因变量以频率的方式代替概率以及缺失值的插补。最终对数据进行了模型分析以及得到了相应的概率预测的结果，并以一定的概率将该视频推荐给该用户。验证表明，线性混合模型可以适用于视频推荐，也能预测出用户观看某个视频的概率，并以多大的概率能够该用户推荐。此模型也优于二元逻辑回归的预测结果。

二、文献综述

由于网络系统发展的越来越完备，人们会对自己喜欢或不喜欢的视频进行浏览并在网上留下痕迹，以便商家通过用户浏览过的物品和电影书籍，来预测该用户的喜好。对于短视频推荐来说，最常用的算法有协同过滤、矩阵分解、聚类、机器学习。通过了解抖音的推荐算法可以知道，一个短视频发布到抖音上，主要经过4个步骤：分别是双重审核、冷启动、数据加权和叠加推荐。

但对于一个好的推荐系统，有很多的评测指标，例如，预测准确度、TOPN推荐。其中，预测准确度可通过RMSE的高低来对其推荐模型的优劣进行选择，并进行TOPN推荐是通过召回率和准确率来对其进行计算。

$$RMSE=|T|\sum\mu,i\in \mathrm{T}\left(r_{\mu i}-r_{\mu i}\right)^{2} \quad (1.1)$$

总之，哪一个评测指标会好一些，需要对其进行特定的分析。根据最优模型，选择最优的指标。

线性混合模型作为近年来应用比较广泛的一个重要的统计模型，比如在印染工业、电商销售量预测、艾滋病疗效预测等各个方面。Henderson首先推导出了一个固定效应β和随机效应μ的最佳线性无偏预测。但得到最佳的预测效果的前提是，因变量需要服从联合正态分布。高萌、张强和邓红把混合线性模型和方差分析应用到重复测量的精神科就诊数据中，对比说明两种方法的特点，指出混合线性模型的协方差分析是在选择协方差结构下可对重复测量资料进行参数估计和统计检验，而方差分析只能对重复测量资料的固定效应做出统计推断；叶辉亮将线性混合效应模型的预测进行了推广，并与最佳线性无偏预测做了比较；宋秋月、易东、伍亚舟表明线性混合效应模型能比较好的处理纵向数据[1]。

即使在过去的研究里，已经有相对成熟的推荐算法，但并未有新的方法应用于视频推荐中。首次将线性混合模型的方法应用于推荐机制的，是通过电影评分预测该用户下次观看该电影的概率，以此通过大数据机制推荐给观看过该电影的某个用户。此外该方法还可以解决模型分析所用数据中既有分类型数据和数值型数据的难题。因此，本文将线性混合模型作为短视频用户推荐分析的研究方法，以推荐系统相关理论为基础，结合线性混合模型方法的应用领域，探讨该方法是否用于视频推荐中，以得到该视频推荐给某个用户的概率，并优于传统的二元逻辑回归预测概率模型。为传统的推荐机制提供一种新的分析思路。

三、模型与方法

1.线性混合模型方法

混合模型也称为层次结构模型。在研究辅导班学生的表现时发现来自同一个班级学生的表现是不独立的，同一班级学生的表现通过不同辅导班联系起来，同一学校的不同班级则是通过学校联系起来。故从学生到班级，从班级到学校形成一个嵌套的层次。混合模型允许不同组群有不同的回归系数，对组群内部的非独立性进行矫正，同样模型也适用于重复测量资料的研究。在重复测量资料的研究中，同一个研究对象的变量随时间的多次测量值之间存在相关性，研究对象的变量随时间变化的观测值是嵌套在研究对象内的。因为线性混合模型属于方差分量模型的一种，所以接下来由方差分析原理入手，依次引出随机截距模型、随机系数模型和线性混合模型的一般形式。

对于传统的线性混合模型来说，使用的数据大多数是数值型数据，其中由固定效应和随机效应组成模型的主要成分。其中，所定义的公式为

$$Y=\mathrm{X}\beta+\mathrm{Z}\mu+\varepsilon \quad (3.1)$$

其中y表示观测值向量，β表示固定效用参数向量，μ表示随机效应参数向量，矩阵X和Z为已知的设计矩阵。其中μ服从均值为0，协方差阵为R的正态分布。同时ε服从均值为0，协方差阵为G的正态分布，即同时对协方差阵R的约束条件减少，不再限制R的主对角元素为σ_{ε}^{2}，非对角元素为0.同时要求G和R无相关关系，即$\mathrm{cov}(G,R)=0$。Y的协方差阵变为$\mathrm{var}(y)=ZGZ'$，Y的期望为$E(Y)=X\beta$，当Z为0时，$R=\sigma_{\varepsilon}^{2}I$时，线性混合模型就变成了一般的线性模型。对于传统的线型混合模型，我们会在一定的假设条件下对其进行拟合，且随机变量的条件，需要服从正态分布。否则，对于模型的建立是不合理的。因此，混合模型有如下假设[2]：

（1）观测值Y来自正态分布总体；

（2）μ服从均值向量为0，G方差–协方差阵为的正态分布；

（3）ε服从均值为0，方差–协方差阵为R的正态分布；

（4）$\mathrm{cov}(G,R)=0$即随机效应和误差之间没有相关关系；

线性混合模型最早被用于电影推荐中，将用户和项目作为两个潜在因子，对电影评分的影响。评分作为因变量$y_{ij}\sim N(s_{ij},\sigma^{2})$。可以构造的模型如下：

$$s_{ij}=f\left(x_{ij}\right)+\alpha_i+\beta_j+\mu_{ij} \quad (3.2)$$

其中，α_i和β_i为潜在因子，μ_i、v_j为随机因子。对于潜在因子模型可也被叫做隐语义模型，是通过对研究的目标量有隐含的影响变量，来预测出用户感兴趣的物品，进而将用户感兴趣的类似物品推荐，并生成推荐页面。因而，对于短视频数据的模型建立，并进行相应的分析。

对于线性混合模型来说，是由固定因子和随机因子所构成。用短视频数据所建立的模型中，观看该视频的用户数为随机因子，因为观看同一视频的用户数是不确定的，同一个视频有可能只会有一个人会观看，也有可能被多个人观看完。固定因子为视频ID，视频的ID是一定的，就像人的性别只有男和女。因变量为该用户浏览完成该视频的概率。那么短视频线性混合模型可以表示为：

$$Y_i=\alpha+\beta_i x+e_i \quad (3.3)$$

y_i代表第i个视频的浏览并完成的概率

β_i代表观看该视频的用户数的影响系数

e_i表示概率的随机变化

α代表固定因子

在实际的工作研究中，统计学家为了方便研究，对于协方差矩阵的结构也做了界定，总结了几种常用的协方差结构，以便建模时提高模型的性能，主要有以下几种结构：

（1）简单结构，协方差矩阵中只含有一个参数，也称独立结构；

（2）含有两个参数的协方差矩阵主要有两种结构，分别是复合对称结构结构（CS）和一阶自回归结构（AR）；

（3）含有t个参数的协方差矩阵主要有两种形式，分别是循环相关结构以及带状柱对角结构；

在进行实证分析时，选取固定效用和随机效用后，根据AIC和BIC准则选取拟合效果最好的协方差矩阵结构，通常选取使AIC和BIC值同时达到最小的为最优协方差结构。

2.固定效应的估计

$$\begin{aligned}&Y=\mathrm{X}\beta+\mathrm{Z}\mu+\varepsilon\\&\mu\sim N(0,G)\\&\varepsilon\sim N(0,R)\\&\mathrm{var}(y)=ZGZ'\end{aligned} \quad (3.4)$$

μ已知线性混合模型满足以上基本条件，假定G、R已知，运用最小二乘估计方法估计β，可得正则方程：

$$X'\Sigma^{-1}X\beta^{*}=X'\Sigma^{-1}y \quad (3.5)$$

可得$\beta^{*}=\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}y$，由此可得任何的可估函$c'\beta$的最佳线性无偏估计为：

$$c'\beta^{*}=c'\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}y \quad (3.6)$$

但事实上G、R的形式我们是不知道的，用估计值$\hat{G}$、$\hat{R}$代替，即用$\hat{\Sigma}=Z\hat{G}Z'+R$可代替$\Sigma$，则可得：

$$c'\tilde{\beta}=c'\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}y \quad (3.7)$$

3.方差分量的估计

极大似然方法最早应用在统计学中，后来被应用在线性混合模型中，极大似然法在处理线性混合模型的方差分量估计中取得较好的效果。考虑一般的线性混合模型：

$$y=X\beta+U_1\xi_1+U_2\xi_2+\cdots U_k\xi_k \quad (3.8)$$

假设$\xi_i\sim N(0,\sigma_i^2 I_{ti})$，$i$等于$1,2\cdots k$。所有的$\xi_i$都相互独立。令$V_i=U_iU_i'$，则方差$\sigma^2=(\sigma_1^2,\sigma_2^2,\cdots\sigma_k^2)$变为：

$$\mathrm{cov}(y)=\sum_{i=1}^{k}\sigma_i U_i U_i'=\sum_{i=1}^{k}\sigma_i V_i\triangleq\Sigma(\sigma^2) \quad (3.9)$$

假设$\Sigma(\sigma^2)>0$，则$y\sim N(0,\Sigma(\sigma^2))$构造似然函数为：

$$L(\beta,\sigma^2|y)=(2\pi)^{-n/2}\left|\Sigma\sigma\right|^{-1/2}\exp\left\{(-\frac{1}{2}(y-x\beta)'\Sigma\sigma^{2-1}y-x\beta)\right\} \quad (3.10)$$

对以上等式两边同时取对数可得：

$$\begin{aligned}&l(\beta,\sigma^2|y)=-\ln\left|\Sigma(\sigma^2)\right|-(y-x\beta)'\Sigma(\sigma^2)^{-1}y-x\beta)\\&=\ln\left|\Sigma(\sigma^2)\right|-tr\Sigma(\sigma^2)^{-1}(y-x\beta)'(y-x\beta)\end{aligned} \quad (3.11)$$

然后对其求偏导，并令其导数为零。由此可推出可估函数$c'\beta$的似然估计为：$c'\beta=c'\left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}y$，可知其为$\sigma^2$的最大似然估计。

四、实证分析

1.数据来源与介绍

视频数据集的采集来源于一次数学建模的竞赛数据，竞赛数据是关于抖音APP的数据，数据集中包括100多万个数据，13个分类变量，其中，有用户ID、视频ID、用户所在城市、作者ID、作品城市、观看该视频的来源、是否浏览完成作品、是否对该作品点赞、音乐ID、设备ID、时间、时长。对于这些数据而言，首先对其进

行预处理，适应所建立的模型[3]。在此次的分析中，我们只选择用户ID、视频ID、是否浏览完成作品作为此次数据的分析。对于一个用户来说，会观看过很多的视频，以及观看过的视频是否浏览完。进而对于线性混合模型来说，可以通过该视频的浏览完成的作为因变量。

2.短视频数据预处理

通过参考竞赛的结果示例，以及如何利用电影评分数据来建立线性混合模型并得到了相应的预测概率，以便将该电影或者该视频推荐用户。因此，首先对抖音视频数据进行了预处理。其次，从十三个分类变量中选择了视频ID、用户ID以及该用户是否浏览完成该视频，进而对所有的数据用excel进行了排序，最后选择了对这3个变量分别进行了统计，统计出看同一个视频的用户数，还有该视频有多少用户浏览并看完的概率，最终以频率代替概率[4]。因数据中也有一部分人并未浏览该视频，这部分缺失的数据就利用均值插补法进行了填补。这就得到了符合建立线性混合模型的数据集。再而通过对数据进行相应的录入，汇合成能够进行分析的数据。由于数据集太庞大，只选择了适合软件容载的数据。这为模型的建立提供了便利，才能在接下来的分析中继续进行下去。

尤其是，对于短视频这种特殊类型数据而言，取决于不同的模型需要，要应具体的方法对数据进行相应的预处理，以便适合模型的建立。

对原始数据做描述统计可知数据的基本特征，在经过对数变换之后，在其峰值在正负3之间，无明显的异常点。因变量浏览完成该作品的概率满足正态分布的要求，符合线性混合模型的基本要求，所以该数据能适用于接下来的模型分析。

3.视频数据存在的问题

从原始数据的分析中，可以看出，所有的变量是分类变量。如果直接用于数据分析以及线型混合模型的建立，结果肯定会出现问题，以及数据有稀疏性问题，需要通过一定的方法进行处理，在前面，提到过相应的方法。如果短视频数据用于成熟的机器学习算法，可以利用特征提取方法，比如，特征提取的线性方法有主成分分析、独立成分分析、线性判别分析来进行降维，进而选择相应显著的变量。所以对此次线性模型的建立，就要对这一万多个数据进行预处理。否则，无法进行模型的建立。

4.固定效应结果分析

表1　各类信息准则表

受限对数似然	AIC	AICC	CAIC	BIC
35.457	39.457	39.582	46.647	44.647

这一结果也说明了，通过计算我们可以看出AIC准则、BIC准则以及极大似然对数与CAIC准则选择的子集相同。说明这几种准则对于模型选择的作用都是相同的。以及视频ID对于用户来说其视频浏览水平同视频的各项指标都有着密切的关系，从而可以通过评估其各项指标及能力来为其用户推荐视频。

表2　III类固定效应估计

源	分子自由度	分母自由度	F	显著性
截距	1	99	164.8	0
itemid	1	99	11.14	0.001

表3　固定效应估算

参数	估算	标准误差	自由度	t	显著性	95%置信区间	
						下限	上限
截距	0.66554	0.051837	99	12.839	0.000	1	0.768395
itemid	−0.00057	0.000169	99	−3.337	0.001	0	−0.00023

在固定效应的表2和表3中，得到视频ID的显著性、T值以及相应的置信区间，还有相应的估计值。且从中表2得到的分析结果是，并能够知道，其中，视频ID的显著性，在0.01的显著性水平下，不能拒绝原假设，固定效应变量对该视频浏览完成的概率具有统计学意义。并得出该模型在假设检验下是合适的，进而得到固定效应的线性方程是：

$$Y = 0.66554 - 0.00057x$$

并从中可以看出，固定效应的参数结果中，其中截距的显著性为0.000，说明拒绝原假设，说明线性混合模型中的截距可以为零，就拒绝所有的视频的完成概率不为零的假设，用户数作为随机因子，显著性为0.000，说明用户数对看完视频的概率有显著的影响，说明不同的视频类型对用户浏览完该视频的概率就的确有差别。因而，此次建立的线性混合模型时可以用于抖音短视频的推荐中。

表4　协方差参数估算值

参数	估算	标准误差	瓦尔德Z	显著性	95%置信区间	
					下限	上限
残差	0.034461	0.009796	3.518	0	0	0.060159
users[主体=itemid]方差	0.034461b	0.	0.	0.		.

方差-协方差成分的参数估计值由表4给出，本次分析中所有的方差成分估计值均为正，说明各个用户之间确实存在斜率和截距的变异。且得到残差为0.034461。

最终得到，线性混合模型不仅解决了数据同时包含分类变量和数值型变量时，无法进行数据分析的难题。而且打破了传统的推荐方法，为未来在推荐系统中遇到新的问题，提供一种新的思路和分析模式。

5.线性混合模型与二元逻辑回归模型对比分析

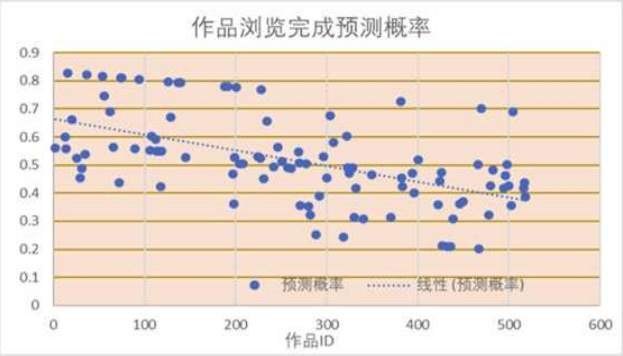

图1　线性混合模型预测概率图

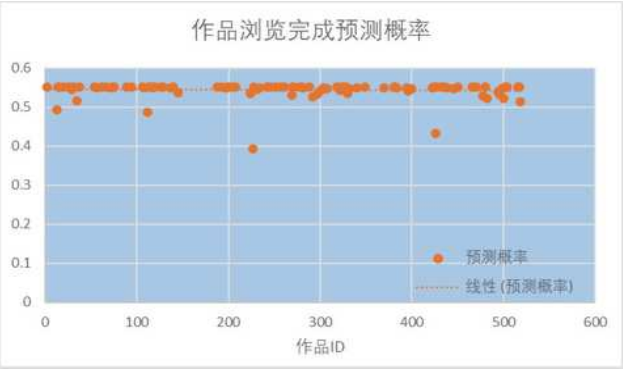

图2　二元逻辑回归预测概率图

通过对两种模型所得到的预测概率进行对比分析来看，线性混合模型比二元逻辑回归所得出的预测效果更好。从数据类型来看，二元逻辑回归不需要对原始数据进行相应的处理，但预测出的效果却并不理想。相反，线性混合模型时对原始数据进行一定的处理后，得到的预测效果比二元逻辑回归更好[5]。这说明在模型拟合效果对数据精度的也有一定的要求。最重要的是，线性混合模型可以解决变量中存在分类和数值两种类型数据的分析中，也为以后的推荐系统，提供了新的推荐机制和用于其他类型的用户推荐中，并给大数据分析提供一种新的思路与解决方式。

五、结论

抖音短视频的数据特征建立线性混合模型，并将其与二元逻辑回归模型所得出的预测效果进行了模型对比。传统的推荐方法有协同过滤、矩阵分解、逻辑回归等算法，最早将线性混合模型用于电影推荐机制中，用来对电影数据进行分析，并将电影评分作为因变量，得到该电影的预测概率，以多大概率将其推荐给看过该类型的电影的用户。模型建立时，数据类型也同样影响模型预测效果的优劣，但恰好线性混合模型就解决了自变量中即含有分类数据和数值型数据的难题，也为模型最终预测效果提供了良好的模型基础。经过一系列分析，线性混合模型可以用于推荐系统中。且与逻辑回归模型相比，得到预测效果更好。在未来用户推荐分析中，将不再局限于传统的推荐算法，为接下来的用户推荐机制提供新的思路和分析模型。

参考文献：

[1]宋秋月，易东，伍亚舟.基于纵向数据线性混合效应模型的老年人抑郁影响因素研究[J].第三军医大学学报，2019，41（04）：384-387.

[2]周兰凤，麻双克，付正，et al.基于复杂属性商品的混合协同过滤推荐模型[J].华东师范大学学报（自然科学版），2017，2017（5）：154-161.

[3]蔡润芹.中国最具影响力的综合搜索引擎比较研究[J].电脑知识与技术，2018，v.14（17）：217-219.

[4]李鹏，于晓洋，孙渤禹.基于用户群组行为分析的视频推荐方法研究[J].电子与信息学报，2014，36（6）：1485-1491.

[5]杨博，郭宇莎，于雪婷，et al.基于混合线性模型的广西某市县级医院住院费用影响因素研究[J].右江民族医学院学报，2018，v.40；No.188（01）：71-74.