Journal of Wuhan Polytechnic

基于 Probit 回归的图书馆读者图书兴趣挖掘方法

柯秀文

(商丘职业技术学院 软件学院,河南 商丘 476001)

摘 要:图书馆因缺少初始读者历史行为信息,难以为初始读者提供有针对性的图书推荐服务,为解决这一问题,提出一种基于 Probit 回归的图书馆初始读者图书兴趣挖掘方法,并根据商丘职业技术学院图书馆读者评分数据进行了模型验证实验,证实了模型的有效性,表明提出的方法能够有效地预测冷启动读者的图书兴趣问题。

关键词:初始读者;Probit 回归;图书兴趣挖掘

中图分类号:G252.0 **文献标识码**:A **文章编号**:1671-931X(2018)02-0097-04

一、引言

信息技术迅猛发展,极大的改变着人们工作、学习的方式,人们可以通过各种渠道获得需要的信息,图书馆仍然是人们获取知识、技能的重要场所,高效、个性化的图书推荐服务是现代图书馆提高服务职能新的方向和选择,然而,各级各类图书馆每天都在面对许多新的初始读者,特别是大中专院校每年都有大量的新生入校,在高校图书馆借书学习是学生学习的重要途径,因此其初始读者每年都会大量的增加,如何更好的为这些初始读者提供有针对性的个性化服务,提高图书馆服务职能,也是各级各类图书馆所要面对和解决的问题。

读者对图书的兴趣一般为一个相关的多维二值变量,变量的一个元素代表读者是否对图书感兴趣,1 表示感兴趣,0 表示不感兴趣[1]。因此,根据读者属性挖掘其图书兴趣是一种相关的多维二值变量回归问题。

为解决图书馆初始读者图书兴趣问题,本文提出一种基于 Probit[2]回归的图书馆初始读者图书兴趣挖掘方法,利用读者的属性信息做出推荐,避免了因为缺少初始读者社交信息而不能进行图书兴趣挖掘的局限。通过回归分析,获得读者属性到读者兴趣的因果关系,从而推断初始读者的图书兴趣。

二、相关定义

将读者所有可能发生的行为定义为集合 A,M_i 表示读者 i 所产生的行为数量,A_{ij} 表示读者 i 产生的第 j 个行为。K 表示所有可能的兴趣,假设读者属性可观测,将 K 作为冷启动问题辅助信息。U_i 表示读者 i 的属性,Z_i 表示读者的图书兴趣,其为不可观测的 K 维度变量。如果读者 i 对第 k 类图书感兴趣,则有 $Z_{ik}>0$,否则 $Z_{ik}=0$。如果某一种图书兴趣在读者行为上的分布为 θ_k,其可以通过话题模型(topic models)和混合成员模型(mix-membership)学习得到[3]。基于以上假设和说明,对读者图书兴趣挖掘进行建模,挖掘读者图书兴趣的目标是找到每一个读者的二值向量 Z_i,表示读者感兴趣的图书。

收稿日期:2018-03-05

基金项目:2015 年商丘市社科规划基金项目“加强高校图书馆服务职能策略研究”(项目编号:SKG-2015-211)。

作者简介:柯秀文(1981-),女,湖北黄石人,硕士,商丘职业技术学院讲师,研究方向:计算机应用技术。

三、读者兴趣挖掘模型

(一) 基于 Probit 的读者兴趣挖掘模型

为挖掘读者的图书兴趣，如图 1 所示，提出一种 2 层结构的贝叶斯生成模型。第一层根据读者的属性产生读者图书兴趣的先验分布，第 2 层根据读者的图书兴趣产生读者的行为。因为 Z_i 是一个二值向量，不能使用连续分布进行推理，因此采用多变量 Probit 回归分布[2]，如公式(1)所示。式(1)输入 X 对应读者的属性，输出 Y 对应读者的兴趣，为解决后验分布截断的情况，将输出 Y 集成到模型 2 层，这样读者兴趣 Z 的先验分布就是一个多变量的高斯分布，如公式(2)所示，该分布样本可以通过传统方法获得。第 2 层根据读者的图书兴趣产生的行为，如公式(3)所示。模型中的相关变量及说明如表 1 所示。

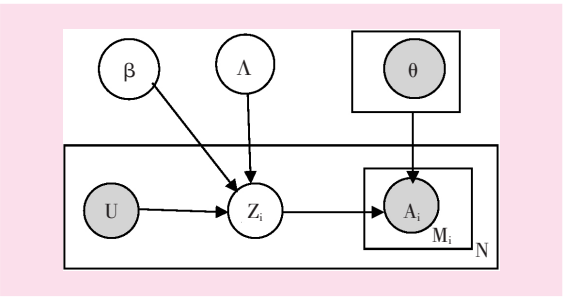

图 1 基于 Probit 回归的读者图书兴趣挖掘模型

表 1 Probit 图书兴趣挖掘模型相关变量及说明

变量名字	变量意义
K	读者用户存在的图书兴趣的数量
M_i	读者 i 的行为数量
N	读者数量
U_i	读者 i 的属性数量
β	读者属性与图书兴趣的回归系数
Λ^{-1}	图书兴趣的协方差矩阵
Z_i	读者 i 的兴趣指示向量
θ_k	图书兴趣 k 在行为上的分布
A_{ij}	读者 i 产生的第 j 个行为

$$\begin{cases} P(Y_i=y_i \mid \beta,\Sigma)=\int_{B_{iJ}}\cdots\int_{B_{iJ}} N_J(t \mid X_i\beta,\Sigma)dZ_i \\ B_{ij}\begin{cases}(0,\infty), \text{if } y_{ij}=1\\(-\infty,0], \text{if } y_{ij}=0\end{cases}\end{cases} \quad (1)$$

$$P(Z_i \mid \beta,\Lambda,U_i)=N(U_i\beta,\Lambda^{-1}) \quad (2)$$

$$P(Aij \mid Z_i,\theta)\frac{\sum_k^K \theta_{k,Aij}1(Z_{ik}>0)}{\sum_m^M\sum_k^K \theta_{k,m}1(Z_{ik}>0)} \quad (3)$$

为 2 个参数分别设定先验分布，以得到一个全概率的贝叶斯模型。假设 β 的先验分布是一个多变量的高斯分布，如公式(4)所示。β 的超参数分别是 μ_0 和 Λ_0。

$$\pi(\beta \mid \mu_0,\Lambda_0^{-1}) \quad (4)$$

假设 Λ 的先验分布是一个 Wishart 分布，如公式(5)所示。Λ_0 的超参数分别是 V_0 和 n_0。

$$\pi(\Lambda \mid V_0,n_0)=W(V_0,n_0) \quad (5)$$

基于以上 2 个先验分布可以得到公式(6)所示的联合概率分布：

$$p(\beta,\Lambda,Z,A \mid \theta,U,\mu_0,\Sigma_0,\Psi_0,v_0)=\pi(\beta \mid \mu_0,\Lambda_0)\pi(\Lambda \mid V_0,n_0)p(Z \mid \beta,\Lambda,U)p(A \mid Z,\theta) \quad (6)$$

后验分布根据贝叶斯理论可以得到公式(7)，因无法直接采样，使用 Gibbs 采样[4]方式进行处理。

$$p(\beta,\Lambda,Z \mid A,\theta,U,\mu_0,\Lambda_0,V_0,n_0)=\frac{p(\beta,\Lambda,Z,A \mid \theta,U,\mu_0,\Lambda_0,V_0,n_0)}{p(A \mid \theta,U,\mu_0,\Lambda_0,V_0,n_0)}\propto p(\beta,\Lambda,Z,A \mid \theta,\mu_0,\Lambda_0,V_0,n_0) \quad (7)$$

(二) 后验分布采样

利用 Gibbs 采样获得后验分布的近似，这里涉及 3 个随机变量 β、Λ、Z，其中 β、Λ 是参数，Z 是需要推理的随机参数。采样时，每次固定其他随机变量，选择一个随机变量产生样本，得到样本依赖于其他随机变量的条件分布，这样对每一个变量进行采样，直到分布达到混合状态，这样得到的样本接近于真实分布得到的采样[5]。3 个条件分布的采样方法如下：

1.在给定 β、Z 的情况下，计算 Λ 的后验分布。根据贝叶斯理论，得到 Λ 的后验分布公式(8)。

$$p(\Lambda \mid \beta,Z,A,\theta,U,\mu_0,\Sigma_0,V_0,n_0)\propto\pi(\Lambda \mid V_0,n_0)p(Z \mid \beta,\Lambda,U) \quad (8)$$

因为 $\pi(\Lambda \mid V_0,n_0)$ 是分布 $p(Z \mid \beta,\Lambda,U)$ 的先验，给定 β 后，Λ 的后验分布如式(9)所示。

$$\Lambda \mid (Z,\beta)\sim W\left(\left(V_0^{-1}+\sum_{i=1}^N (Z_i-U_i\beta)(Z_i-U_i\beta)'\right)^{-1},N+n_0\right) \quad (9)$$

2.在给定 Λ、Z 的情况下，获得 β 的条件分布。根据式(6)和贝叶斯理论得到式(10)中的后验分布。

$$\begin{aligned} &p(\beta \mid \Lambda,Z,A,\theta,U,\mu_0,\Lambda_0,V_0,n_0)\\ &=\frac{p(\beta,\Lambda,Z \mid A,\theta,U,\mu_0,\Lambda_0,V_0,n_0)}{p(\Lambda,Z \mid A,\theta,U,\mu_0,\Lambda_0,V_0,n_0)}\\ &\propto p(\beta,\Lambda,Z \mid A,\theta,U,\mu_0,\Lambda_0,V_0,n_0)\\ &\propto p(\beta,\Lambda,Z,A \mid \theta,U,\mu_0,\Lambda_0,V_0,n_0)\\ &\propto\pi(\beta \mid \mu_0,\Lambda_0)p(Z \mid \beta,\Lambda,U)\end{aligned} \quad (10)$$

根据式(4)，$\pi(\beta \mid \mu_0,\Lambda_0)$ 是一个高斯分布，而且是似然 $p(Z \mid \beta,\Lambda,U)$ 的共轭先验。因此，β 的条件后验也是一个高斯分布，如式(11)所示。

$$\beta \mid (Z,A)\sim N(\beta \mid \hat{\beta},\hat{\Lambda}) \quad (11)$$

其中 $\hat{\Lambda}=\Lambda_0+\sum_{i=1}^N U_{i'}\Lambda U_i$，$\hat{\beta}=\hat{\Lambda}^{-1}\left(\Lambda_0\mu_0+\sum_{i=1}^N U_{i'}\Lambda Z_i\right)$。可以直接通过高斯采样获得该分布样本。

3.分析在给定 β、Λ 的情况下，计算 Z 的后验分布，如式(12)所示。

$$p(Z_i \mid \beta,\Lambda,A,\theta,U,\mu_0,\Lambda_0,V_0,n_0)$$
$$\propto p(Z_i \mid \beta,\Lambda,U_i)p(A_i \mid Z_i,\theta) \quad (12)$$
$$=N(U_i\beta,\Lambda^{-1})\prod_{j=1}^{Mi}\frac{\sum_{k}^{K}\theta_{k,Aij}1(Z_{ik}>0)}{\sum_{m}^{M}\sum_{k}^{K}\theta_{k,m}1(\hat{Z_{ik}}>0)}$$

可以看出，Z 的后验分布是一个非标准分布，无法产生对应样本。因此采用 Metropolis-Hastings[6]算法获得相应样本。假设建议分布为 $q(Z_1^* \mid Z_1,\sum,\beta)=N(Z_1^*;U_i\beta,\Lambda^{-1})$，那么 Z_i 的样本可以通过如下 2 个采样步骤得到：

(1)根据分布 $N(U_i\beta,\Lambda^{-1})$产生样本 Z_i^*。

(2)以概率 $\alpha(Z_i^*,Z_i)=\min\left\{1,\frac{p(W_i \mid \theta,Z_i^*)}{p(W_i \mid \theta,Z_i)}\right\}$接受 Z_i^*，否则接受 Z_i。

四、实验

为了证实评估模型的有效性，研究人员以商丘职业技术学院图书馆读者评分数据进行分析，以证多变量 probit 模型能够很好的获得读者的兴趣所在。以教师、学生为主体的商丘职业技术学院能够有效获得图书馆读者性别、年级、专业等信息。因此，本研究以商丘职业技术学院图书馆读者作为研究用户，将读者的读书类别选择作为读者的行为，以此来挖掘他们的兴趣。将 2015 年以来三个年级的大专学生的读书行为作为行为记录，共收集到 11873 个学生读者，101487 本图书和 350591 的读书行为。读者的性别、年级、专业都被当作属性，以此推断他们的读书兴趣，本文将商丘职业技术学院图书馆图书进行归纳分类，共分为哲学；政治、法律；军事；经济、管理；文学；艺术；历史、地理；工业技术；机械、仪表工业；无线电电子学、电信技术；电工技术；自动化技术、计算技术等 20 个类别，将各专业学生分为文、理、工、农四个学科大类，图 2 给出了 20 个种类图书分布。

图 2 给出的是读者属性和图书类别之间的关系，图 2(a)读者图书兴趣与性别之间的关系，男同学阅读的文学书比女同学一般要少；图 2(b)是不同专业大类学生读者对各类图书的兴趣分布，学生阅读和自己专业相近的图书显著高于其它类别图书。从图中可以看出，根据读者性别或者所在专业大类的不同可以得到不同的图书类别分布，进而说明读者属性和图书兴趣存在一定的关系。下面根据读者属性，验证本文提出的模型对读者图书兴趣预测的有效性。

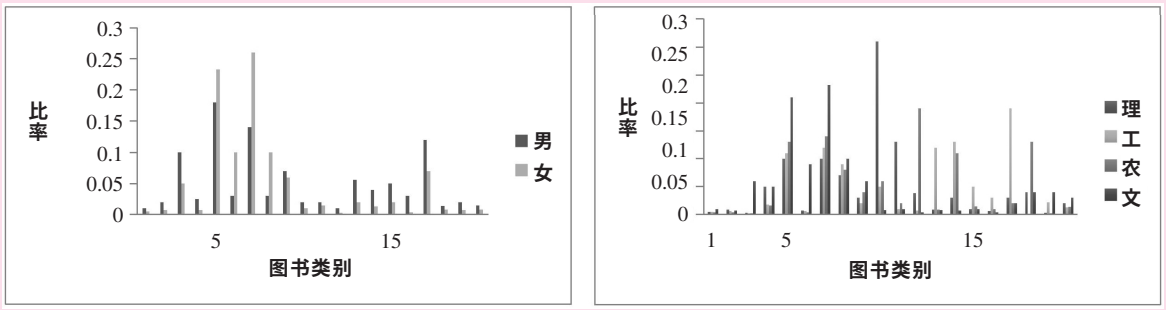

图 2 图书类别分布情况

对于读者真正的兴趣由于实际条件受限无法有效获得，这里采用读者经常借阅的图书类别作为读者的兴趣，以此来验证本文模型的有效性。将已借阅过的图书的类别作为真值，同时也将此方法作为比较的基准方法，因此基准方法没有利用读者的相关属性信息，所以读者的训练行为较少时，很难获得较好的性能。为了验证先验信息对预测冷启动读者图书兴趣的重要性，提出方法与基准方法在不同比例训练集上的性能差异。采用余弦相似度度量测试值与真实值之间的差异，实验结果如图 3 所示。

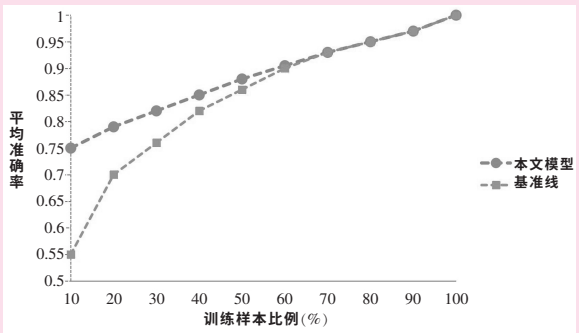

图 3 算法的平均准确率

实验表明，当训练样本的比例低于 40%时，基于 Probit 回归的读者图书兴趣挖掘方法和基准方法相比具有显著的优越性。随着训练样本比例增加，两种方法的准确性都同时增加，最后收敛道真实的读者图书兴趣分布。虽然真值与基准方法都采用同样的策略推断读者的图书兴趣，但在训练样本缺少的时候，基准方法与真实分布的差距较大。基于 Probit 回归的读者图书兴趣挖掘方法，能够在加入先验信息之后获得更真实的读者图书兴趣所在。随着训练样本的增加，提出方法也能获得真实的读者图书兴趣。

基于 Probit 回归的读者图书兴趣挖掘方法能够在读者行为比较少的情况下，较好的解决冷启动读者问题，较为准确的推断图书馆初始读者的图书兴趣。

五、结束语

本文提出了一种基于 Probit 回归的图书馆读者图书兴趣挖掘方法，同时，给出了基于 Gibbs 采样和 Metropolis-Hastings 采样推理方法。在商丘职业技术学院图书馆读者评分数据上进行实验，证实了模型的有效性。该模型能够在读者行为欠缺的情况下，能够根据先验信息推断出较为准确的读者图书兴趣，解决了图书馆初始读者图书兴趣的问题，为图书馆初始读者图书推荐奠定了基础。在获得充分的读者行为之后，模型能够获得更加准确的读者图书兴趣。

参考文献：

[1] 景民昌，于迎辉.基于借阅时间评分的协同图书推荐模型与应用[J].图书情报工作，2012，(2)：117-120.
[2] Chib S，Greenberg E.Analysis of multivariate Probit models [J].Biometrika，1998，85(2)：347-361.
[3] Ahmed A，Low Yucheng，Aly M，et a.Scalable distributed inference of dynamic user interests for behavioral targeting [C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.2011：114-122.
[4] Casella G，George E I.Explaining the Gibbs sampler[J]. The American Statistician，1992，46(3)：167-174.
[5] 张建平，张立波，程浩忠，等.基于改进拉丁超立方抽样的概率潮流计算[J].华东电力，2013，(10)：40-46.
[6] Hastings W K.Monte Carlo sampling methods using Markov chains and their applications [J].Biometrika，1970，57(1)：97-109.

[责任编辑 刘 骋]

Library Reader Book Interest Mining Method Based on Probit Regression

KE Xiu-wen
(College of Software, Shangqiu Polytechnic, Shangqiu476001, China)

Abstract: Because of the lack of initial readers' historical behavior information, it is difficult to provide targeted reader recommendation services for the initial readers. To solve this problem, a library based on Probit regression is proposed for library initial readers' book interest mining methods. A model validation experiment was conducted on the readership rating data of the library of the vocational and technical college, which confirmed the validity of the model and showed that the proposed method can effectively predict the book interest of cold-start readers.

Key words: initial reader; Probit regression; book interest mining