BG/NBD 模型的手机阅读业务客户价值预测实证研究①

吴 昊[1,2], 王 纯[1,2], 周立娟[2]

[1](北京邮电大学 网络与交换技术国家重点实验室, 北京 100876)
[2](东信北邮信息技术有限公司, 北京 100191)

摘 要: 客户价值预测主要是对客户的消费意愿进行预测, 找出真正的忠诚用户群, 从而有针对性地进行营销. 企业通过有效地客户价值预测, 及时准确地掌握用户的消费意愿, 提供针对性的服务, 从而提升整体业务量. 以 Beta-Geometric/Negative Binomial Distribution (BG/NBD)模型为基础, 结合手机阅读用户数据进行客户价值预测的实证研究, 验证了该模型在手机阅读个性化营销方面的适用性.

关键词: BG/NBD; 客户价值预测; 个性化服务; 手机阅读; 数据挖掘

Empirical Research of Mobile Reading Customer Value Prediction with the BG/NBD Model

WU Hao[1,2], WANG Chun[1,2], ZHOU Li-Juan[1]

[1](State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)
[2](EB Information Technology Co. Ltd., Beijing 100191, China)

Abstract: Customer value prediction is mainly about forecasting the customers' consumption will. It aims at undertaking personalized marketing to those real loyal users. The Service Provide can enhance the overall business volume by promptly and accurately analyzing the customers' consumption will and providing personalized service with the method of Customer Value Prediction. This paper carries on the empirical research to validate applicability of the BG/NBD model in personalized marketing of mobile reading customers, based on the model theory and combined with mobile reading users' data.

Key words: BG/NBD; customer value prediction; personalized service; mobile reading; data mining

客户价值预测主要是对客户的付费意愿进行预测. 手机阅读领域客户与传统领域的客户有所不同, 由于消费者不需要亲自出现在书店中, 所以消费者可以排除自己身份、地位等平时在实体消费渠道可能存在的影响消费的因素, 因此消费行为呈现明显的随机性和非契约性. 随机性是指用户可能以微小的原因（如心情波动, 有无充足时间等）从放弃到建立或从建立到放弃自己的购书意愿. 非契约性是指无法跟踪到用户的状态, 无法准确预测客户发生读书行为后, 是否还会发生读书行为或下次发生读书行为的时间.

对于随机性及非契约性用户, 预测用户未来消费行为的比较有效的方式为 Beta-Geometric/Negative Binomial Distribution(BG/NBD)模型. 本文将以 BG/NBD 模型为基础, 结合手机阅读用户数据进行实证研究, 验证 BG/NBD 模型在手机阅读领域预测客户付费概率的适用性.

1 BG/NBD模型理论基础

BG/NBD 模型基于下列五个假设[1-4]:

1)在活跃时,一个客户的交易量服从交易率 λ 的

① 基金项目:国家自然科学基金(61072057,61101119,61121001,60902051);长江学者和创新团队发展计划(IRT1049);国家科技重大专项(2011ZX03002-001-01)

收稿时间:2012-08-27;收到修改稿时间:2012-09-28

泊松分布. 这相当于假设交易时间间隔服从交易率 λ 的泊松分布.

$$f(t_j \mid t_{j-1};\lambda)=\lambda e^{-\lambda(t_j-t_{j-1})},t_j>t_{j-1}\geq 0$$

2)交易率 λ 的非均匀性的概率密度函数服从 gamma 分布.

$$f(\lambda \mid r,\alpha)=\frac{\alpha^r\lambda^{r-1}e^{-\lambda\alpha}}{\Gamma(r)},\lambda>0$$

3)每次交易后,客户变得不活跃的概率为 p,因此客户退出点服从二项式分布.

$$P(\text{在第j次交易后立刻变得不活跃})=P(1-P)^{j-1},\ j=1,2,3\ldots\ldots$$

4)概率 p 的非均匀性的概率密度函数服从 beta 分布.

$$f(P \mid a,b)=\frac{P^{a-1}(1-P)^{b-1}}{B(a,b)},0\leq P\leq 1$$

其中 B (a, b) 是 beta 函数,可用 gamma 函数表示为:B (a, b) = Γ (a) Γ (b) / Γ (a + b).

5) 交易率 λ 和流失率 p 是相互独立的分布.

从以上假设中, 可以推导出一个结论: 设 T 为时间长度, x 为某一用户在 T 时间内的消费次数, t_x 为某一用户最后一次消费的时间. 则购买历史为（X=x,t_x,T）的消费者将来在长度为 t 的时间内的期望消费次数为:

$$E(Y(t)\mid X=x,t_x,T,r,\alpha,a,b)=\frac{\dfrac{a+b+x-1}{a-1}}{1+\delta_{x>0}\dfrac{a}{b+x-1}\left(\dfrac{a+T}{a+t_x}\right)^{r+x}}\left[1-\left(\frac{a+T}{a+T+t}\right)^{r+x}\cdot F_1\left(r+x,b+x,a+b+x-1;\frac{\alpha}{\alpha+T+t}\right)\right] \quad (1)$$

以上式子包含（r, α ,a,b）四个参数, 这些参数可以通过最大似然方法进行估计.

$$L(r,\alpha,a,b \mid X=x,t_x,T)=A_1\cdot A_2\cdot(A_3+\delta_{x>0}A_4)$$

其中,

$$A_1=\frac{\Gamma(r+x)\alpha^r}{\Gamma(r)} \qquad A_2=\frac{\Gamma(a+b)\Gamma(b+x)}{\Gamma(b)\Gamma(a+b+x)}$$

$$A_3=\left(\frac{1}{(\alpha+T)}\right)^{r+x} \qquad A_4=\left(\frac{a}{b+x-1}\right)\left(\frac{1}{\alpha+t}\right)^{r+x}$$

假设有一个含有 N 个客户的样本,客户 i 在(0, T_i)内有 X_i= x_i 次交易,最后一次交易发生在 t_{xi} 时,样本的似然函数对数为[4-6]:

$$LL(r,\alpha,a,b)=\sum_{i=1}^{N}\ln[L(r,\alpha,a,b \mid X_i=x_i,t_x,T_i)] \quad (2)$$

因此在我们知道客户历史消费行为的情况下, 先通过式(2)求出(r, α ,a,b)这四个参数, 然后依据把这些参数代入式(1). 对于每个历史消费行为是（x_i, t_x, T_i）, 就可以预测该客户在未来一段时间 t 内发生交易的次数.

2 数据准备

由于 BG/NBD 模型是从客户历史消费行为导出模型进行预测用户未来消费行为, 因此需要选取两部分数据. 一部分作为预测样本, 反映手机阅读客户历史付费行为, 用于导出 BG/NBD 模型; 另一部分作为验证样本, 反映手机阅读客户未来消费行为, 用来验证生成的 BG/NBD 模型. 预测样本的内容为前一段时期内（取 2011 年 11 月份为例）一批用户的按章订购天数. 验证样本为后一段时间内（取 2011 年 12 月份为例）该批用户整段时间内按章订购天数. 表 1 为样本数据示例:

表 1 样本数据示例

手机号	按章订购天数（预测样本）				付费章节数（验证样本）
	11/1	11/2	……	11/30	12 月
用户 1	1	0	……	1	21
用户 2	0	1	……	0	4

3 模型建立

模型建立整体步骤如图 1 所示:

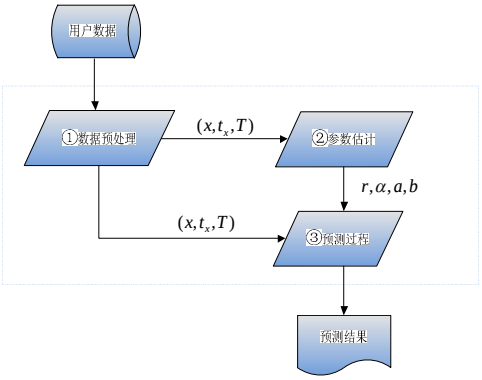

图 1 模型建立步骤

1) 数据预处理. 数据预处理的任务是接受预测样本中每个用户的原始数据, 输出相应的三元组（x, t_x, T）. 其中 x 为交易次数 , T 为每个客户第一次购买到观察期截止时间长度, tx 为客户最后一次购买发生的时间. 对于付费潜力预测来说, T 与 t_x 以天计算, 消费次数则按照按章订购天数计算（指标可替换）. 例如设观察期为 11 月份的 30 天, 一个用户在 11 月份的消费行为如表 2 所示.

则对于该用户, 数据预处理的输出为:

$x=1+1=2$（11 月 5 日订购及 11 月 20 日订购）

$t_x=20$（11 月 20 日为观察期内用户消费的最后一天）

$T=30-5+1=26$（11 月 5 日到 11 月 30 日有 26 天）

2) 参数估计. 参数估计将数据预处理步骤输出的所有三元组（x, t_x, T）输入到式 2 中:

$$LL(r,\alpha,a,b)=\sum_{i=1}^{N}\ln[L(r,\alpha,a,b\mid X_i=x_i,t_x,T_i)]$$

然后对该似然函数分别对 r, α ,a,b 求偏导并另参数分别为 0, 得到似然方程组. 通过该似然方程组便可得到 4 个参数的值.

3) 预测过程. 预测过程就是对于预测样本中的每个用户, 将其在数据预处理中得到的 x, t_x, T 以及参数估计中得到的 r, α ,a,b 代入到式(1)中:

$$E(Y(t)\mid X=x,t_x,T,r,\alpha,a,b)=\frac{\frac{a+b+x-1}{a-1}\left[1-\left(\frac{a+T}{a+T+t}\right)^{r+x}\cdot F_1\left(r+x,b+x,a+b+x-1;\frac{\alpha}{\alpha+T+t}\right)\right]}{1+\delta_{x>0}\frac{a}{b+x-1}\left(\frac{a+T}{a+t_x}\right)^{r+x}}$$

即可得到该用户在未来某段时间内期望消费次数.

4 模型适用性验证

对于交易次数 x 的选取, 需要根据实际业务进行确定. 手机阅读业务主要涉及的交易有: 包月、按本订购、按章订购, 由于包月和按本订购行为发生频次相对较低（包月和按本均属于一次性消费）, 所以本部分以按章订购为基础. 下面对 x -按章订购次数、 x -按章订购天数, 分别进行模型适用性验证.

4.1 首次订购用户数据验证

选取数据说明:

1) 样本用户: 考察期之前无订购行为、2011.11.1 至 2011.11.10 期间首次按章订购的用户.

2) 观察期: 首次订购时间至 2011.12.31.

3) 预测期: 2012.1.1 至 2012.1.31.

结果说明:

抽取 6000 个用户（3000 个 1 月实际有交易行为的用户以及 3000 个 1 月实际无交易行为的用户）, 按照上述组合 2 种情况分别进行验证, 发现 x -按章订购次数的情况, 模型预测效果不甚理想, 此种情况不适合用该模型, 具体结果不再展示, 只对 x -按章订购天数结果进行展示.

准确性定义:

对于预测结果的准确性, 需要根据业务需求目标而对应不同的标准. 如下所示两种情况.

第一种情况

目标: 预测用户是否来消费

标准:

a. 实际消费天数=0 and 预测消费天数<0.5;

b. 实际消费天数>=2 and 预测消费天数>=1;

c. 实际消费天数=1 and 预测消费天数>=0.5;

第二种情况

目标: 预测用户消费天数

标准:

a. 实际消费天数=0 and 预测消费天数<0.5;

b. 实际消费天数>=20 and (预测消费天数-实际消费天数)<=10;

c. 实际消费天数>=10 and 实际消费天数<20 and (预测消费天数-实际消费天数)<=8;

d. 实际消费天数>=5 and 实际消费天数<10 and 预测消费天数>=1 and (预测消费天数-实际消费天数)<=7;

e. 实际消费天数>=3 and 实际消费天数<5 and 预测消费天数>=1 and (预测消费天数-实际消费天数)<=6

f. 实际消费天数=2 and 实际消费天数>0 and 预测消费天数>=0.5 and (预测消费天数-实际消费天数)<=5

g. 实际消费天数=1 and 预测消费天数<=0.5 and (预测消费天数-实际消费天数)<=4

根据上述准确性定义, 结果如表 2 所示.

表 2 首次订购用户验证结果

目标	实际用户数	预测准确用户数	准确率	合计
是否消费	3000（x>0）	2030	67%	77%
	3000（x=0）	2587	86%	
消费天数	3000（x>0）	1523	51%	69%
	3000（x=0）	2587	86%	

4.2 有订购历史行为用户数据验证

选取数据说明:

1) 样本用户: 考察期之前无订购行为、2011.11.1 至 2011.11.10 期间首次按章订购的用户.

2) 观察期: 首次订购时间至 2011.12.31.

3) 预测期: 2012.1.1 至 2012.1.31.

结果说明:

抽取 6000 个用户（3000 个 1 月实际有交易行为的用户以及 3000 个 1 月实际无交易行为的用户），按照上述组合 2 种情况分别进行验证, 发现-按章订购次数的情况, 模型预测效果不甚理想, 此种情况不适合用该模型, 具体结果不再展示.

准确性定义:

准确性定义同首次订购用户, 这里不再赘述.

数据验证结果如表 3 所示.

表 3 有订购历史行为用户验证结果

目标	实际用户数	预测准确用户数	准确率	合计
是否消费	3000（x>0）	2676	89%	77%
	3000（x=0）	1947	65%	
消费天数	3000（x>0）	1952	65%	65%
	3000（x=0）	1947	65%	

3 结果分析

根据验证结果显示, 依据不同的业务目标, 该模型能较好的抓住用户的付费意愿和付费次数. 通过客户价值预测, 可帮助业务关注者定位忠诚用户, 对该类客户进行进一步的细分和特征分析, 并能根据用户的其他业务特征制定有针对性的营销策略.

通过用户是否有消费意愿的预测, 对消费意愿减弱或意愿消失倾向的用户, 可提前制定措施有针对性的营销挽回, 同时结合付费次数预测, 在合理限值内进行营销以避免形成过打扰.

综上所述, 本文通过手机阅读领域用户数据验证了适当业务目标的定义下, 使用 BG/NBD 模型能够有效而且准确的进行客户价值预测, 对客户消费意愿进行预测, 定位忠诚用户群, 从而有针对性地进行营销能够起到很好的作用.

参考文献

1 张春莲. BG/NBD 模型对客户购买行为的预测分析.时代经贸,2008,6(97):51–52.

2 王永贵,董大海.客户关系管理的研究现状、不足和未来展望.中国流通经济,2004,(6):52–56.

3 周洁如.客户关系管理中的价值创造研究.上海管理科学,2003,(4):55–56.

4 Reichheld F. Learning from Customer Defections. Harvard Business Review,1999,(2):56–69.

5 陈明亮.客户重复购买意向决定因素的实证研究.科研管理,2003,24(1):110–115.

6 万里,廖建新,王纯.基于社会网络信息流模型的协同过滤算法.吉林大学学报,2011,41(1):270–275.

(上接第 139 页)

参考文献

1 王娟,郭永冲,王强.基于 BHO 的网络隐蔽通道研究.计算机工程,2009(5):159–161.

2 桑庆兵,吴小俊.基于 BHO 的网站过滤系统研究与实现.计算机工程与应用,2009,45(31):18–19.

3 http://baike.baidu.com/view/362533.htm.

4 东方人华,吕伟臣编.Delphi7.0 入门与提高.北京:清华大学出版社,2006,369–380.

5 ../Delphi7/Source/Internet/SHDocVw.pas[CP/OL]:478–580.

6 ../Delphi7/Source/rtl/win/ActiveX.pas[CP/OL]:3068–4813.

7 ../Delphi7/Source/rtl/sys/System.pas[CP/OL]:254–275.

8 魏志强,王忠华.程序设计 Delphi 5.0.北京:中国铁道出版社,2000.238–252.