基于高维 0-1 变量的 EM 算法在移动通信客户细分中的应用

胡新杰　罗荣华　江明华　王汉生[1]

摘要：现代计算机技术的迅速发展为人们提供了大量详细真实的消费者购买行为数据，而这些数据往往可以表现为高维的 0-1 变量（High dimensional binary response）。如何基于这样的数据对消费者群体合理细分是本文关心的重点。具体地说，本文提出一个基于 0-1 变量的混合模型（Mixture Model），并通过 EM 算法具体估计。本文以移动通信的实际数据为例做了详细演示，并获得了一些有意义的结论。

关键词：0-1 变量；EM 算法；市场细分

The EM algorithm based on high dimensional binary response and its application in market segmentation of China Mobile

Xinjie Hu, Ronghua Luo, Minghua Jiang and Hansheng Wang

Abstract：The advance of the information technology provides us huge amount of detailed consumer behavior data, which can typically be represented as a high-dimensional binary response. Then, how to make a good use of such a data to facilitate effective market segmentation is becomes the problem of interest. For such a purpose, we proposed a binary response mixture model with a well-developed EM algorithm. The real data from China Mobile is carefully demonstrated with a number of interesting findings.

Keywords：Binary response; EM algorithm; Market segmentation

[1]胡新杰　（北京大学光华管理学院，北京，100871，huxinjie@gsm.pku.edu.cn）
罗荣华　（北京大学光华管理学院，北京，100871，luoronghua@gsm.pku.edu.cn）
江明华　（北京大学光华管理学院，北京，100871，jmh@gsm.pku.edu.cn）
王汉生　（北京大学光华管理学院，北京，100871，hansheng@gsm.pku.edu.cn）.

一、问题提出

随着电信体制的改革和不断深化，移动通信垄断经营局面被打破，中国电信和中国网通的“小灵通”全面加入移动通信市场的竞争，现在已经形成中国移动、中国联通、中国电信和中国网通四家运营商群雄争霸的格局，移动通信已成为竞争最为激烈的通信业务。我国的移动通信市场已完全进入买方市场，消费者对电信服务的需求呈现多样化、个性化和多层次化的趋势。但移动通信运营商为了获得更多的客户资源和占有更大的市场份额，往往采取“简单”的价格竞争和“此起彼伏”的广告宣传战，其弊端显而易见。

很多理论提出要以客户需求为导向，对客户价值、消费者行为进行分析，针对不同类型的客户提供个性化的服务，设计不同的业务套餐，进行客户流失预测和管理。在营销理论中倡导以客户为核心的深度营销。可以说，对客户有效、科学的细分是营销活动非常关键的开始，只有进行有效的分组、聚类，才能很好地发现同类客户之间共同的消费行为和不同类型客户之间的差异，才能针对各种类型客户提供个性化的服务。

但现有对客户的分类和管理上还存在着大量问题，最突出的就是对客户类型简单分组。实际情况中大量的是根据客户消费额度把客户分为大客户（VIP客户）和一般客户；或者根据产品类型进行分类，如全球通客户、动感地带客户、神州行客户等等。对客户的简单分类，掩盖了同类客户之间的差异，无法获得更多的消费者行为信息，也就无法进行客户价值分析、客户流失管理，进而也就无法进行一系列的深度营销活动。而本文的目的就是要提供一套完整的数据驱动的市场细分方法，即聚类分析方法。

我们注意到，传统的聚类方法多适用于连续性变量，而移动通信用户行为中最重要的购买决策却是一个 0-1 变量（例如：0=不购买；1=购买）。因此，这些传统的聚类方法无法适用于这种特殊的数据类型。但是另一方面，王汉生等（2006）所提出的基于 0-1 变量的方法是一种粗糙的分层聚类方法（Hierarchical Clustering），由于该方法缺乏一个统一有效的模型支持，因此无法回答很多重要的问题，例如：到底应该把消费者聚成几类？为了从理论上较为彻底地解决这个问题，本文尝试发展并研究一套基于混合模型（Mixture Model）的聚类方法，并提出一个 BIC 标准用于自动选择细分市场的个数。我们以移动通信的实际数据为类，演示了该方法的有效性。

二、文献回顾

国内的相关研究非常缺乏，最近的主要相关研究为王汉生等（2006）。他们以超市零售商品数据为例，提出了一种基于高维 0-1 变量的聚类方法。作者关注于 m 种不同商品的

相关性。用 $x_i=(x_{i1},\cdots,x_{im})$ 描述顾客的某次购买行为，其中 m 是维度，即所研究的商品数量；$x_{ij}=1$ 表示在该消费者 i 的购物篮中发现了第 j 种商品，$x_{ij}=0$ 表示在该消费者的购物篮中未发现第 j 种商品。用：

$$s_{jk}=\sum_{i=1}^{n}I\left\{x_{ij}=x_{ik}\right\}\left\{\sum_{i=1}^{n}I\left\{x_{ij}+x_{ik}>0\right\}\right\}^{-1}$$

来度量商品 j 和商品 k 之间的距离，代表了至少购买商品 j 和 k 中的某一种的消费者中，有多少消费者同时购买了两种商品；如果 s_{jk} 很大，说明消费者一旦决定购买商品 j 和 k 中的任何一种，那么另一种就也有很大可能性被同时购买，这两种相关性很大的商品被聚为一类。在分层聚类的过程中，距离近的、相似性最强的商品被首先聚为各个“小类”；然后将相似的小类聚为“大类”。该方法有其简单有效的优点，但是也有明显的缺点。主要局限有二：第一、该方法仅仅对商品聚类，而没有对顾客进行同步聚类。因此，丧失了细分顾客市场的机会。第二、即使对商品聚类，由于该方法仅仅是一个统计算法，缺乏一个合理有效的统计模型支持，因此无法客观确定“类”的个数。

如果我们将消费者同时购买或者选择多个产品或者服务的行为称为同时购买（Simultaneous Purchasing），那么国外最相关的研究莫过于 Manchanda, Ansari, 以及 Gupta（1999）。具体地说，基于随机效用理论（Random Utility Theory），他们发展了一个多种类同时购买决策模型（Multicategory Purchase Incidence Decision Model）。该模型通过分层贝叶斯多维 Probit 模型（Hierarchical Bayesian Multivariate Probit Model）得到了很好地实现。该模型的优点在于他精细地刻画了各个消费者的同时购买行为特征，并同时考虑到了消费者异质性（Heterogeneity）的影响。但是该方法的优点同时也是其缺点，对于实际工作者来说，模型算法太过复杂，难以实现；另一方面，模型所产生的基于每个消费者个体的分析结果，更难以在技术手段还有限的国内企业中实现。

同样的问题在计算机领域也备受关注。如果我们将消费者同时购买或者选择多个产品或者服务的行为称为“同时发生（Co-occurrence）”，那么最重要的理论研究由 Hofmann 和 Puzicha（1998）完成。本文所采用的模型框架主要来源于此。但是，不一样的是本文在此基础上提出了一个合理的 BIC 选择标准，用于自动决定细分市场的个数。

三、混合模型结构

为了讨论方便，我们用 $x_i=\left(x_{i1},\cdots x_{im}\right)^T$ 来代表来自于第 $i\ \left(i=1,\cdots,n\right)$ 个个体的观

测到的高维0-1变量，即 $x_{ik} \in \{0,1\}$； $k(k=1,\ldots,m)$ 即个体 i 选择的第 k 个商品。其中 n 是样本量，即所研究的客户数量；m 是维度，也即所研究的商品数量。此外，我们还假设对任意样本 i，有一个隐含的观测不到的分组变量 $z_i = \left(z_{i1},\cdots z_{ig}\right)^T$。其中，如果该个体（如移动客户）来自于第 $j\left(j=1,\cdots,g\right)$ 个细分市场，则 $z_{ij}=1$；否则 $z_{ij}=0$。因此，$\left\{z_{ij}:1\le j\le g\right\}$ 中永远有且仅有一个为1。并且，我们用 ϕ_{jk} 来表示来自于第 j 个市场细分的个体选择第 k 个商品的概率。

根据 Hofmann 和 Puzicha（1998）的模型设定，我们假设人们所看到的产品相关性，完全由消费者所处不同的市场细分所决定。统计上这就是假设，一旦 z_i 给定，x_i 的全部分量就变得相互独立。具体地说，根据 Hofmann 和 Puzicha 的理论，$P\left(x_i \mid z_{ij}=1\right)=\prod_{k=1}^{m}\phi_{jk}^{x_{ik}}$，而 $P\left(z_{ij}=1\right)=\pi_j$。由此可见，$\pi_j$ 定义了消费者群体中第 j 个细分市场占总体消费者的比例。为了获得一个合理的 EM 算法，我们暂时假设 z_i 是观测到的，那么可以得到联合似然函数如下：

$$\prod_{i=1}^{n} p\left(x_i\right)=\prod_{i=1}^{n}\prod_{j=1}^{g}\left[\prod_{k=1}^{m}\phi_{jk}^{x_{ik}}\left(1-\phi_{jk}\right)^{\left(1-x_{ik}\right)}\right]^{z_{ij}} \qquad (1)$$

如果我们假设 ϕ 服从均匀分布，那么对一固定的 j，ϕ_{jk} 的后验分布为：

$$\begin{aligned}\left\{\phi_{jk}:1\le k\le m\right\} &\propto \prod_{i=1}^{n}\left[\prod_{k=1}^{m}\phi_{jk}^{x_{ik}}\left(1-\phi_{jk}\right)^{\left(1-x_{ik}\right)}\right]^{z_{ij}} \\ &=\prod_{i=1}^{n}\left[\prod_{k=1}^{m}\phi_{jk}^{z_{ij}x_{ik}}\left(1-\phi_{jk}\right)^{z_{ij}\left(1-x_{ik}\right)}\right] \qquad (2) \\ &=\prod_{k=1}^{m}\phi_{jk}^{\sum_{i=1}^{n}z_{ij}x_{ik}}\left(1-\phi_{jk}\right)^{\sum_{i=1}^{n}z_{ij}\left(1-x_{ik}\right)}\end{aligned}$$

可见：$\left\{\phi_{jk}\right\}$ 的后验分布是一个参数为 $\left(\sum_{i=1}^{n}z_{ij}x_{ik}+1,\ \sum_{i=1}^{n}z_{ij}\left(1-x_{ik}\right)+1\right)$ 的 β 分布。因此，在已知 z_{ij} 的情况下，我们可以得到 ϕ_{jk} 的估计量为：

$$\hat{\phi}_{jk} = \left(\sum_{i=1}^{n} z_{ij} x_{ik} + 1 \right) \left(\sum_{i=1}^{n} z_{ij} + 2 \right)^{-1} \qquad (3)$$

而这成为我们下面 EM 算法的一个重要理论基础。

四、EM 估计算法

由于 z_i 是观测不到的隐含变量，获得极大化的联合似然函数并不容易。所以，我们采用了广为接受的EM算法（Dempster、 Laird 和 Rubin，(1977)）。模型中：$\phi_{jk}^{(l)}$ 代表EM算法第 l 步中对 ϕ_{jk} 的估计，$\pi_j^{(l)}$ 代表代表EM算法第 l 步中对 π_j 的估计，$z_{ij}^{(l)}$ 代表EM算法第 l 步中对 z_{ij} 的估计。

第一步：随机分组。初次分组的方法：计算每种产品的总购买次数，对购买次数排序，找出购买次数最高的产品，然后找出所有购买该产品的客户，把这部分客户归为第一组；从数据库中删除第一组客户，然后计算每种产品的总购买次数，找出总购买次数排在第一位的产品，所有购买该产品的客户归为第二组；遵循前两步的原则，直到找出第 $g-1$ 组，剩余的客户归为第 g 组。

第二步：*Maximization Step.* 给定 $\phi_{jk}^{(l)}, \pi_j^{(l)}, z_{ij}^{(l)}$，根据（3）得：

$$\hat{\phi}_{jk}^{(l)} = \left(\sum_{i-1}^{n} z_{ij}^{(l)} x_{ik} + 1 \right) \left(\sum_{i-1}^{n} z_{ij}^{(l)} + 2 \right)^{-1}, \quad \pi_j^{(l)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{l}$$

第三步：*Expectation-Step.*给定 $\phi_{jk}^{(l)}, \pi_j^{(l)}$，根据 Bayesian 公式，我们可以得到：

$$z_{ij}^{(l+1)} = \frac{P(z_{ij} = 1,\ x_i,\ \phi_{jk}^{(l)}) \times \pi_j^{(l)}}{\sum_{j'=1}^{g} P(z_{ij'} = 1,\ x_i,\ \phi_{j'k}) \times \pi_{j'}^{(l)}} = \frac{\prod_{k=1}^{m} \phi_{jk}^{x_{ik}} \times \sum_{i=1}^{n} z_{ij}^{(l)}}{\sum_{j'=1}^{g} (\prod_{k=1}^{m} \phi_{j'k}^{x_{ik}} \times \sum_{i=1}^{n} z_{ij'}^{(l)})}$$

第四步：迭代。重新进行以上步骤，直至参数 ϕ 和 π 收敛，得到最终的后验概率。

第五步：根据以下公式，计算相应的 BIC 取值：

$$\text{BIC(g)=-2}\times\sum\log\left\{\prod_{i=1}^{n}p\left(x_i\right)\right\}+\log(n)\times g\times m$$

其中 $p(x_i)$ 由（1）式定义。在实际工作中，我们可以从 $g=1$ 出发，尝试不同的 g 取值，并从中选择最优的分组方案（即：产生最小 BIC 值的分组方案）。

五、 实际数据分析

1、 数据介绍

我们的数据来自于中国移动通信下属某公司，共包含 3048 个随机抽取的样本。其中每一个样本代表了该移动通信公司的某个用户在该月的移动通信消费情况。根据该公司所提供的服务类型，我们将户用所产生的消费情况归纳为以下 9 种不同的类型：（1）本地通话费、（2）国内长途费、（3）国内漫游费、（4）IP 国内长途费、（5）短信费、（6）梦网费用、（7）基本月租费、（8）主叫号码显示费用、以及（9）彩铃使用费。对于各个产品，如果该消费者在当月确有消费，那么定义相应的变量为 1，否则为 0。这样，我们就将常见的关于客户消费额的数据转化为一个高维 0-1 变量的数据。对于变换后的高维 0-1 数据，我们通过简单计算和作图，可以看到所研究的客户选择各种产品的平均概率。

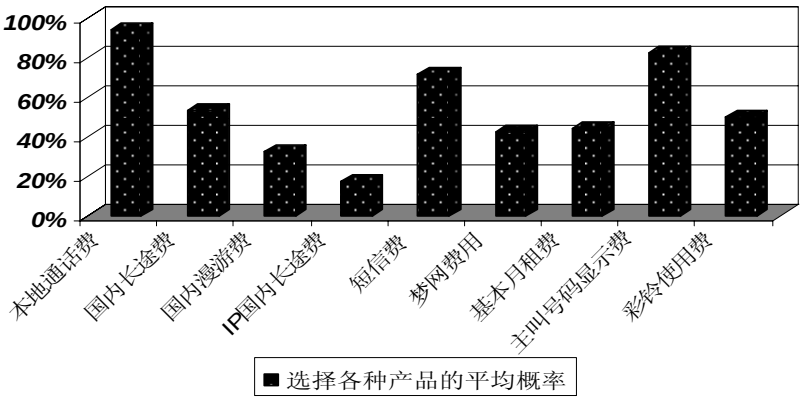

图 1　选择产品的平均概率

从图 1 可以看到，客户对于本地通话功能的选择概率最大，其次是主叫号码显示功能。在实际生活中，大部分的客户都会开通主叫号码显示功能，也会产生本地通话费。然后就是短信，处于第三位。还可以看到，梦网和彩铃的选择率已经超过国内漫游费和 IP 国内长途费，略低于国内长途费；可以说明很多人可能对通过 IP 拨打长途的业务不是很熟悉，更多地是直接拨打长途；漫游产品的选择一般是商务人士到外地出差所产生，也有部分出外旅游的人会产生一些漫游费用；梦网和彩铃的选择说明已有更多的人开始使用这两项业务，对于新业务的推广及第三代移动通信的发展有很好的启示作用。

2、用混和模型和 EM 算法分析实际数据

（1）首先要确定细分市场的个数 g 。为可靠起见，我们共尝试了 10 种不同的分组方案，即 $g=1,\cdots,10$ 。其相应 BIC 取值，如图 2 所示。从中可以清晰看到，最优细分方案应为 $g=5$ 。

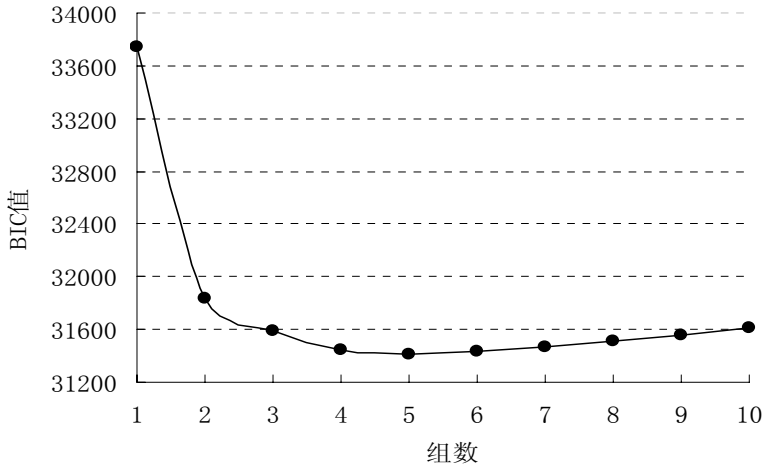

图 2　通过 BIC 确定组数

（2）然后在固定 $g=5$ 的情况下，计算每个市场细分所占比例（即 π_j ），如下图所示：

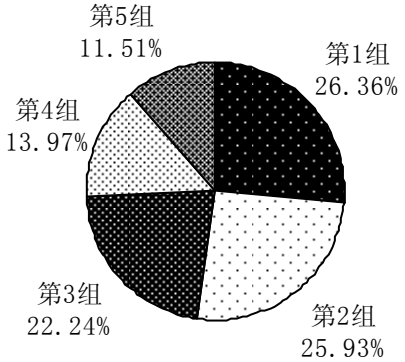

图 3　市场细分中客户数的比例

从图 3 可以看到，各组人数差别不是很大，其中第 1、2、3 组客户数较多，第 4、5 组客户数相对较少。

（3）然后通过结果中的后验概率值 ϕ_{jk} ，可以得到每个小组选择不同产品的概率，如表 1 所示，并对其结果具体分析如下：

[1] 第 1 组（时尚追求型客户群）：该组客户对短信、彩铃的选择概率最高，对梦网费用、本地通话费的选择概率也比较高，而对国内漫游费的选择概率接近于 0。在实际生活中，年轻人更为倾向于短信、彩铃的功能使用。我们认为该组客户是时尚追求型客户群。

[2] 第 2 组（高端商务型客户群）：国内长途费、本地通话费的选择概率最高，国内漫游费的选择概率也远大于 1、3、5 组。国内漫游费主要是由于客户漫游到外地进行通信所产生费用，一般是由客户出差时所产生。由此，我们可以认定第 2、4 组为商务型客户。而对其他各项产品的选择概率也几乎都高于其他各组中的选择概率，尤其是基于商业需求而产生的梦网费用的选择概率远高于其他各组。所以我们认为第 2 组是高端商务型客户群。

[3] 第 3 组（费用节省型客户群）：国内长途费、国内漫游费、IP 国内长途费的选择概率接近 0，本地通话费、基本月租费、主叫号码显示费的选择概率最低。本地通话功能是客户进行通信所需要的最基本的产品，选择概率越小说明客户价值越低。本组的客户对各个产品的选择概率几乎都是最低的。我们认为该组客户是费用节省型客户群。

[4] 第 4 组（普通商务型客户群）：国内长途费选择概率较高，略低于第 2 组，而远高于其他组；其次为国内漫游费，是 5 个小组中选择概率最高的。但第 4 组的本地通话费和梦网费用的选择概率远低于第 2 组，其他各产品的选择概率也均低于第 2 组。所以，我们认为第 4 组是普通商务型客户群。

[5] 第 5 组（普通消费型客户群）：本地通话费的选择概率最高，但国内漫游费选择概率接近 0，而对其他各项产品的选择概率在各组中都处于中间水平。我们认为该组客户是普通消费型客户群。

表 1：不同市场细分的选择概率对比表

组号	本地通话费	国内长途费	国内漫游费	IP 国内长途费	短信费用	梦网费用	基本月租费	主叫号码显示	彩铃使用费
1	0.817	0.259	0.023	0.123	0.901	0.540	0.398	0.900	0.628
2	0.942	0.986	0.739	0.360	0.883	0.613	0.805	0.984	0.520
3	0.295	0.025	0.007	0.014	0.521	0.270	0.142	0.608	0.481
4	0.476	0.922	0.879	0.205	0.654	0.248	0.238	0.694	0.451
5	0.758	0.621	0.027	0.128	0.405	0.250	0.557	0.861	0.224

五、 结论和研究建议

通过以上分析，我们最终把客户分为 5 个市场细分，分别为：时尚追求型客户群、高端商务型客户群、费用节省型客户群、普通商务型客户群、普通消费型客户群。分类的同时，我们可以得到每个顾客被分到哪个组，以及每组的客户选择不同产品的概率，对于移

动运营商提供个性化服务、确定客户价值、进行套餐设计和实行深度营销等都有很强的现实意义。当然，本研究也有一定的局限性：

1、所研究的移动产品不是很多。现实中的移动产品种类很多，如果能对客户所消费的所有产品进行分析，可能分类结果会更加准确。

2、样本量相对不大。由于数据来源问题，样本量的选取相对不是很多，如果有更多的样本供研究，研究的推广价值应该更大。

3、如果能对连续很多期的数据进行研究，应该可以对客户未来的消费情况进行预测，也是客户流失管理所关心的内容之一。

4、如果能结合客户的背景资料综合分析，对现实的指导意义会更大。

综上所述，本文所使用的 EM 算法在移动客户聚类分组中的应用是相对有效的，对于各类行业中关于客户数据的聚类分组同样具有很强的指导意义。同时希望在今后的研究中能够通过对更加完备的数据的分析，来完善该研究方法。

参考文献

[1] 王汉生，江明华，曹丽娜，金英．超级市场零售商品的购物篮分析[J]．营销科学学报，2006（2）：71-77．

[2] ANDREAS M, THOMAS R．An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data[J]．Journal of Retailing and Consumer Services, 2002（10）：123–133．

[3] CHEN Y H, GEROGE E I．A Bayesian model for collaborative filtering[J]．Online Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics,1999．

[4] CONSONNI G, VERONESE P．A Bayesian method for combining results from several binomial experiments[J]．Journal of the American Statistical Association,1995（90-431）：935-944．

[5] DEMPSTER A P, LAIRD N M, RUBIN D B．Maximum likelihood from incomplete data via the EM algorithm[J]．Journal of the Royal Statistical Society Series,1977,1–39．

[6] DOMINGOS P, PAZZANI M．On the optimality of the simple Bayesian classifier under zero-one loss[J]．Machine Learning, 1997（29）：103-130．

[7] HOFMANN T, PUZICHA J．Statistical models for cooccurrence data[J]．Artifical Intelligence Laboratory Memo No.1625,1998．

[8] KONSTAN J, MILER B, MALTZ D, HERLOCKER J, GORDON L, RIEDL J．GroupLens: applying collaborative filtering to usenet news[J]．Communications of the Association for Computing Machinery, 1997（40-3）：77-87．