

超级市场零售商品的购物篮分析

王汉生¹、江明华¹、曹丽娜²、金英¹

¹北京大学光华管理学院, ²中央电视台广告部

内容提要

本文所研究的问题是如何通过有效的数量方法, 来发现各种商品在消费者购物篮中同时出现的规律。这是一个典型的商品聚类的问题。但是, 同普通聚类问题不同的是, 各种商品在同一个购物篮中的出现与否, 构成了一个典型的高维 0-1 数据 (High-Dimensional Binary Response, 即: 0-未出现, 1-出现), 而通常基于欧氏距离的聚类方法仅适用于连续数据。因此, 我们研究并介绍一个专门为高维 0-1 变量所设计的聚类方法。该方法不但可以被用来做典型的购物篮分析, 而且简单并具有良好的直观意义。我们运用此方法对某大型超市的 26 天的销售流水数据进行了探索性研究, 得到了一些有意义的结论。

关键词: 高维数据、购物篮分析、聚类分析、0-1 变量

A Basket Analysis for the Retailed Products in Supermarket

Hansheng Wang¹, Minghua Jiang¹, Lina Cao², and Ying Jin¹

¹Guanghua School of Management, PKU and ²CCTV Advertising Department

Abstract

We consider how to identify the important product co-occurrence patterns in consumer's shopping basket, via an appropriate quantitative method. This is a typical clustering problem. However, it is very different from the usual clustering analysis in that the presence or not of various products constitutes a binary data vector with an extremely high dimension, which makes the most typically used Euclidean distance based clustering method inappropriate. Therefore, we introduce in this article a clustering method, which is specifically designed for the data of similar type. The proposed method can be used for the most typical basket analysis together with its simplicity and nice interpretability. We demonstrate the usefulness of the introduced method on a major supermarket's detailed sale data over a 26 days' time interval with a number of interesting findings.

KEYWORDS: High-dimensional data; Basket analysis; Clustering analysis;
Binary variable

一、问题提出

首先，现代零售商品种类极端丰富，消费者需要处理的信息量急剧增加。消费者平均要以每秒 33 件的速度从 5 万件商品中挑选出 17 件商品。Phillips (2005) 的研究表明，当消费者面对种类繁多的商品时，并不会因为可选择的丰富多样性而得到满足。但是，消费者却能够因为超市对其商品选择的引导而感到满意。超市引导顾客的一个有效办法就是合理的商品布货。就是说，哪些商品可以摆放在一起，而哪些商品又应当分别摆放。问题是，超市进行布货的依据是什么？

其次，我们可以观察到商场和超市经常进行各种促销，其中最常见的促销方式是打折，而且，常常是全场打折。这样的打折往往不是超市最优的选择。因为，消费者在购买某些商品的时候，会同时购买另一些商品，而不管它们是否打折。在这种情况下，只要这两种商品之一处于打折状态，而另一种也极有可能受到刺激而销售量大增。如果是这样，超市只需要对一种商品打折就可以达到促销两种商品的目的，从而可以大大提高超市的效益。问题是，超市安排商品打折的依据是什么？

因此，了解消费者究竟如何在多商品类目间进行同时购买 (Simultaneous Purchase) 对于超市如何有效地引导消费者和提高效益意义重大。所以，本文的目的有二。第一、介绍一个简单而有效的数量方法，可以用来做典型的购物篮分析；第二、用此方法详细分析某中等城市的一个大型连锁超市数据，从而探索超市消费者的相关行为特征。

以下章节如下安排。下一节，详细介绍一个基于 0-1 变量的聚类方法。基于此方法的实际数据分析将在第三节中展开。最后是总结与讨论。

二、文献研究

在过去的研究中，Fader 和 Lodish (1990) 的研究表明某些消费者特征，例如，家庭渗透 (Household Penetration) 和购买频率，对零售商品定价和促销环境具有一定的解释能力。Narasimhan 等人 (1996) 的进一步研究发现，一类商品的促销弹性部分取决于该类商品的品类结构和相关消费者特征。Poel 等人 (2004) 的研究结果表明，商品之间的互补性越强，促销的交互影响越大。Raju (1992) 研究了不同类商品销量差异性，并建立了它同品类特征和营销组和变量的关系。Hoch 等人 (1995) 则研究了各类商品的商店价格弹性 (Store-Level Price Elasticities) 和所在商圈消费者人口统计特征的关系。

以上研究,一方面为人们理解消费者的超市购物行为提供了很多有意义的理论依据和现实证据。但是另一方面,他们所研究的消费者群体据来自于欧美发达国家,其结论对于中国大陆广大消费群体的适用性值得进一步验证;而另一方面,随着中国大陆经济的迅速腾飞,人们生活水平的快速提高,以超市为代表中国大陆零售商品市场正在飞速发展。面对这样一个迅速崛起的巨大市场,一方面各大国际知名超市企业(如:沃尔玛、家乐福)迅速进入中国并快速发展,而另一方面超市企业对消费者行为的微观数据进行分析的却少之又少。在过去有限的研究中,人们仅仅把注意力集中在了相对宏观的层面上,或者纯粹的数量技术模型上。而本文试图通过翔实可靠的超级市场消费者行为数据,对大陆消费者的购物篮进行探索性分析。

购物篮指的是超级市场内供顾客购物时使用的装商品的篮子,当顾客付款时这些购物篮内的商品被营业人员通过收款机一一登记结算并记录。所谓的购物篮分析就是通过这些购物篮子所显示的信息来研究顾客的购买行为。消费者的购物篮隐含着重要且有价值的信息,等待人们去发掘。如:我们可以知道消费者的购买习惯、产品偏好、品牌忠诚度等等。而本文尝试通过合理的数量方法,研究产品的相关性。也就是说,我们关心的问题是那些产品互相之间具有很强的相关性。从而我们可以知道,当一个消费者购买其中一个产品的情况下,极有可能同时购买另外一个产品。这对于超市合理定价促销有着重要的指导意义。此类研究在国外已有成功案例,而国内市场营销界的文献资料中则很少见到。

国外对消费者同时购买行为最深入的数量研究,应该属于 Manchanda, Ansari, 以及 Gupta (1999)。他们提出了一个基于随机效用函数(Random Utility Theory)的多种类同时购买决策(Multicategory Purchase Incidence Decision)的模型。该模型通过贝叶斯多维 Probit 模型(Bayesian Multivariate Probit Model),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。类似的基于贝叶斯框架的消费者选择模型(Bayesian Consumer Choice Model),大量地存在于文献当中。有兴趣的读者,可以在最近的 Rossi, Allenby, 和 McCulloch (2005)找到很好的介绍及相关文献。

由于此类模型充分地考虑了消费者的异质性,因此有可能被用来为面向消费者的个性化的营销决策(Customized Marketing Decision)服务。但是,该模型的优点也恰恰是他的缺点。第一、对消费者异质性的精细刻画使得该模型所需要的计算量非常巨大。该计算量主要来源于贝叶斯方法所需要的 MCMC (Markov Chain Monte Carlo)算法。例如, Manchanda, Ansari, 以及 Gupta (1999)的原始文章只考虑了 4 种产品种类。这显然远远不能够满足现实的需要。以本文所研究的超市为例,该超市所涉及的产品上万种,覆盖几百个小类。如果只考虑对销

售额起决定作用的产品，那也至少几十类。这在数据上反映出来就是一个非常高维的数据，而此类方法显然无法适用。第二、此类模型通常需要能够对一给定消费者的重复购买行为予以准确跟踪。换句话说，研究者必须能够从数据辨认同一消费者在不同时间的购买纪录。这通常意味着该超市必须有完善的会员制度，以及详细准确的会员信息。这对很多超市，特别是大量的本土新兴连锁超市来说，是一个巨大的，短时期内难以克服的困难。因此，我们认为有必要研究并介绍一种简单易懂，对数据要求低，而且能够处理高维数据的分析方法，以便于探索消费者同时购买行为的规律。

为了达到以上目的，我们将同时购买行为规范成一个典型的聚类问题（Clustering Problem）。具体地说，具有某种“相似性”的商品容易被同时购买。基于传统数据的聚类分析，已被广泛研究并发展完善。有兴趣的读者可以通过 Johnson 和 Wichern(2003)以及 Hastie, Tibshirani, 和 Friedman (2001) 获得一个完整的介绍。但是，超市购物篮数据有其独特特征。具体地说，与普通数据不同的是，各种商品在同一个购物篮中出现与否，构成了一个典型的高维 0-1 数据（High-dimensional binary response，即：0-未出现，1-出现），而普通的基于欧氏距离的聚类方法（Johnson 和 Wichern, 2003）仅适用于连续数据。因此，有必要发展并介绍一个专门为高维 0-1 变量所设计的聚类方法。该方法应该不但可以被用来做典型的购物篮分析，而且简单易行并具有良好的直观意义。

三、基于 0-1 变量的聚类方法

简而言之，我们的目的就是要对购物篮中商品的相关性予以分析，并根据其相关性的的大小予以聚类。假设我们关注于 p 种不同产品的相关性。对于第 i 个消费者，我们可以用向量 $x_i = (x_{i1}, \dots, x_{ip})$ 来描述他的某次购买行为。其中， $x_{ij} = 1$ ，如果在该消费者的购物篮中发现了第 j 种商品，否则 $x_{ij} = 0$ 。假设，我们有总共 n 个消费者，那么我们可以定义向量 $v_j = (x_{1j}, \dots, x_{nj})$ 。该向量刻画了第 j 种商品被 n 个消费者购买的情况。如果， v_j 由大量的 1 构成，那么我们就知道该商品被消费者购买的频率很高。另一方面，如果 v_j 由大量的 0 构成，那么我们就知道该商品被购买的频率很低。

另外，如果有两个共同的商品 j 和 k ，我们还可以通过比较向量 v_j 和 v_k 的相似性来获得对他们相关性的度量。具体地说，如果我们发现 v_j 和 v_k 的各个分量非常相似，这说明商品 j 和 k 很容易被同时购买，或者被同时不购买。因此，我们可以简单地认为这两种产品的

相关性很强。因此，我们第一种度量商品相关性数量指标定义如下：

$$r_{jk} = \frac{1}{n} \sum_{i=1}^n I\{x_{ij} = x_{ik}\},$$

其中示性函数 $I\{x_{ij} = x_{ik}\} = 1$ ，如果确实有 $x_{ij} = x_{ik}$ ；否则 $I\{x_{ij} = x_{ik}\} = 0$ 。简单地说 r_{jk} 就是对商品 j 和 k 有相同购买行为（同时购买，或者同时不买）的消费者在总共 n 个消费者中所占的比例。因此，如果 r_{jk} 值很大，这说明商品 j 和 k 的相关性很强，因此应该被聚为一类，否则说明相关性很弱。

但是，遗憾的是这个度量在实际数据应用中并不理想。主要原因在于对于超市数据，对于任意两个产品我们都能否发现他们有很大的 r_{jk} 值。这并不说明这两个产品的相关性都很强，而是由于产品种类繁多，大多数消费者都会同时不购买这两种产品，因此造成

$$\frac{1}{n} \sum_{i=1}^n I\{x_{ij} = x_{ik} = 0\}$$

的值很大。因此，我们转而考虑如下相关性度量：

$$s_{jk} = \frac{\sum_{i=1}^n I\{x_{ij} = x_{ik}\}}{\sum_{i=1}^n I\{x_{ij} + x_{ik} > 0\}}。$$

请注意，由于 x_{ij} 和 x_{ik} 为取值只可能为 0 或者 1 的 0-1 变量，因此条件 $x_{ij} + x_{ik} > 0$ 隐含着 x_{ij} 和 x_{ik} 中至少有一个取值为 1。也就是说，商品 j 和 k 中至少有一种被第 i 个消费者购买。因此，

$\sum_{i=1}^n I\{x_{ij} + x_{ik} > 0\}$ 计算了 n 个消费者中，有多少人至少购买了商品 j 和 k 中的一种。那么，

指标 s_{jk} 就度量了在购买了商品 j 和 k 中至少一种的消费者中，有多少消费者同时购买了两种产品。由此可见，如果 s_{jk} 很大，这说明消费者一旦决定购买商品 j 和 k 中任何一种，那么另外一种就也有很大可能性被同时购买；进而我们知道，这两种商品的相关性很大，应该被聚为一类，否则相关性很小。一旦有了相似性的度量，我们就可以通过变换 $d_{jk} = 1 - s_{jk}$ 来获得关于差异性的距离度量。

值得注意的是，以上的距离定义仅仅适用于两种具体的商品。在分层聚类（Hierarchical

Clustering)的过程中,距离最近的,相似性最强的商品被首先聚为个各“小类”。然后,我们需要在此基础上,再将相似的“小类”聚为“大类”。因此,我们需要定义“类”与“类”之间的距离。假设我们有两个“小类”,记为: $A = \{i_1, i_2, \dots, i_p\}$ 与 $B = \{j_1, j_2, \dots, j_q\}$ 。即:第一个“小类”中总共包含了 p 个不同的产品,而第二个“小类”中包含了另外 q 种产品。研究中常用的定义“小类”A和B之间距离的方法有以下三种:

单连接法 (Single Linkage):

$$d_{AB} = \min \{d_{j_a j_b} : 1 \leq a \leq p, 1 \leq b \leq q\}$$

完全连接法 (Complete Linkage):

$$d_{AB} = \max \{d_{j_a j_b} : 1 \leq a \leq p, 1 \leq b \leq q\}$$

平均连接法 (Average Linkage):

$$d_{AB} = \frac{1}{pq} \sum_{a=1}^p \sum_{b=1}^q d_{j_a j_b}。$$

简单地说,单连接法要求比较弱。只要两个“小类”中有两个产品相似,那么单连接法就会认为这两个“小类”非常相似。而完全连接法则不然。根据定义,完全连接法要求两个“小类”中的所有产品都相似,这两个小类才可以被称为相似。而平均连接法则居于单连接和完全连接中间。对于各种连接方法的详细讨论,可以参阅 Johnson 和 Wichern (2003)。对于我们的超市数据,以上三种连接方法都有所尝试,最后发现完全连接方法最为适合。一旦距离被合理地定义了出来,多数现有的软件都可以被用来作相应的聚类分析,如: SAS、SPSS、Splus、R 等等。本文所有计算和图表主要使用的是 SAS 和 R。

四、实际数据分析

本文中所采用的数据来自于我国北方某中型城市,处于垄断地位的某大型连锁超市 26 天的销售流水数据,共 65,535 条记录,主要包括交易代码,交易时间,商品代码,商品名称,销售数量,销售金额等等。交易定义为顾客的每次购买行为,也即是一个购物篮,一次交易涉及一个或多个商品,一条销售记录指一次交易中某种商品的销售数量和金额。其中有效销售记录为 65,348 条,其余 187 条销售记录所对应的商品并未包含在商品基本信息数据库中。而这 65,348 条有效销售记录来自于 10,216 笔交易,即包含了 10,216 个购物篮的商品信息。这些购物篮中共包括了总共 4,833 种商品,其平均购买金额为 40.48 元。

在发生销售的 4,833 种商品中,有 2,989 种商品的购买频率不超过 5 次。一般来讲,顾客购买频率较高的商品对于超市具有更加显著的经济意义,这些商品的价格及销售量的变动

将显著影响到超市的效益。在后面的分析中，我们将选择顾客购买频率最高的前 50 种商品进行分析，这是因为该 50 种产品的销售额占到该超市销售额很大的比例。为了方便起见，我们仅仅将处于销售频率前十位的商品进行描述（参见表 1）：这 10 种商品的销售金额约为 10 万元，占总销售额的 26.55%，具有很好的代表性，而且对超市意义重大。为直观起见，因此，利用整理好的数据，采用 0-1 变量距离进行聚类分析，分别采用单连接法、完全连接法和平均连接法，经过比较发现完全连接法的结果最为显著，所得到的聚类结果的树状结构图（参见图 1）。从中可以看出，这 50 种商品最终被划分为 8 类。最为相似的两种商品是“主食厨房”和“生鲜蔬菜”两种商品。其次，是“仙岛增鲜酱油 1 级 380ml”和“新碘盐 500g”，主食与蔬菜同时购买，比较符合家庭的购买习惯，酱油与盐则同时属于烹饪调味料。值得注意的是我们发现“260 克三鹿纯牛奶”与“伊利百利包纯牛奶”也表现出了很强的相关性，它们之间的距离约为 0.88（参见图 1），这隐含着相关性强度为 $s_{jk} = 1 - 0.88 = 12\%$ 。这说明，一旦一个消费者决定购买这两种不同品牌牛奶中的任何一种，有 12% 的可能性会同时购买这两种牛奶。

表 1：销售金额最高的十类商品

排序	商品类别	销售金额（元）	占总销售额的百分比
1	鲜猪肉	22121.9	5.35
2	色拉油	14849.3	3.59
3	散蛋类	10769.2	2.60
4	盒装白酒	9820.6	2.37
5	果汁	9179.4	2.22
6	洗发类	9152.2	2.21%
7	散粮	8769.1	2.12
8	纯牛奶	8734.1	2.11
9	蔬菜类	8233.5	1.99
10	白面	8153.3	1.97
合计		109782.6	26.55

五、总结与讨论

我们通过对超市消费者的购物篮进行分析，发现了一些很有意义的现象。有的是我们可以预见的，例如，仙岛生鲜酱油与新碘盐的高度相关性；而有些则是出乎我们意料的，例如，三鹿纯牛奶与伊利纯牛奶。我们推测原因可能是消费者对两种牛奶的品质难以区分的情况下采取尝试性购买，或者可能是消费者处于品牌转换阶段，开始尝试性购买；也可能是为了满足不同家庭成员对不同品牌的牛奶有不同的偏好；还可能是其他我们不知道的原因

等等。我们无法对这些行为背后的原因进行推断，深入研究这些背后的推动因素是一个很有意义的未来方向。而这恰恰就是本方法的特点所在。我们所提出的方法的能够在没有很多先验知识的情况下，发现有意义的相关现象，并为未来的研究指明方向。不论怎样，发现这些相关性对于指导未来的超市经营管理具有一定的借鉴意义。例如，可以考虑对高度相关产品的某一种大力促销，而对另外一种保持原价，甚至提高价格。消费者会因为其中打折促销产品的吸引而光顾超市，但是他们除了会购买打折产品之外，还会购买与之高度相关的产品。而超市由于部分促销让利所造成的损失，可以由此而获得相应的甚至更多的补偿。当然，本研究不可避免地具有一定的局限性，具体地说：

第一、由于本文的重点在于发现具有强相关关系的产品，而不是对所发现的同时购买行为的隐含推动因素进行深入研究和分析。因此，我们却无法对消费者同时购买这些相关性很强的商品的深层次原因进行具体分析。造成该局限性的另外一个原因在于数据的局限性，即绝大多数消费者拒绝透露详细的个人信息。我们深信，对于消费者个体信息的获得会极大地提高本研究的价值。

第二、本文所提出的数量方法也是有一定的局限性。例如，我们所发现的“260 克三鹿纯牛奶”与“伊利百利包纯牛奶”的相关性为 12%。但是对于实际工作而言，也许该相关性强度是有限的，因此能够起到的指导作用也是有限的。究其原因，我们认为这主要是因为本方法没有对消费者予以同时聚类。也就是说，一个更好的方法应该能够做到对消费者以及产品同时聚类。那么基于细分的消费者群体，我们预期能够发现的相关性产品会更多、其相关性会更强、因此对实际工作的指导意义更大。这是我们努力的另外一个重要方向。

综上所述，本文提出了一个简单适用的购物篮聚类方法。我们的初步研究表明该方法是有效的，并有希望得到进一步的发展。我们深信对该问题的进一步研究是很有意义的！

参考文献

- [1]. Fader, P. and Lodish, L. M. (1990) “A cross-category analysis of category structure and promotional activity for grocery products”, *Journal of Marketing*, **54**, 52-56.
- [2]. Hastie, T. and Tibshirani, R. and Friedman, J. (2001) “The Elements of Statistical Learning”, New York: Springer.
- [3]. Hoch, S. J., Kim, B. D., Montgomery, A. L. and Rossi, P. E. (1995). “Determinants of store level price elasticities”, *Journal of Marketing Research*, **32**, 17-29.
- [4]. Johnson, R. A. and Wichern, D. W. (2003). “Applied Multivariate Statistical Analysis” 5th ed.

Pearson Education, Hong Kong. P. R. China.

- [5]. Manchanda, P., Ansari, A. and Gupta S. (1999). “The ‘shopping basket’: a model for multicategory purchase incidence decisions”, *Marketing Science*, **18**, 95-114.
- [6]. Narasimhan, C., Neslin, S. A. and Sen, S. (1996) “ Promotional elasticities and category characteristics”, *Journal of Marketing*, **60**, 17-30.
- [7]. Phillips, H. (2005) “Theory and reality of choice in retail markets”, *Consumer Policy Review*; May **15**, 99-103.
- [8]. Poel, D. V., Schamphelaere, J. D. and Wets, G. (2004) “Direct and indirect effects of retail promotions on sales and profits in the do-it-yourself market”, *Expert Systems with Applications*, **27**, 53-62.
- [9]. Raju, J. S. (1992) “The effect of price promotions on variability in product category sales”, *Marketing Science*, **11**, 207-220.
- [10]. Rossi, Allenby, and McCulloch (2005). “Bayesian Statistics and Marketing”, Wiley, NY.

图 1: 聚类结果树状图

表2：聚类分析中顾客购买频率最高的50种商品

商品名称	购买频率(次)	销售金额(元)	商品名称	购买频率(次)	销售金额(元)
生鲜蔬菜	5126	7924.30	超级六丁目五连包香辣牛肉95gX5	205	564.60
主食厨房	1563	2851.81	盼盼蛋黄派 500克	201	1011.90
鸡蛋	1073	9701.24	洽洽怪味豆 52g	190	187.60
桃李现场制作面包蛋糕	659	2162.69	秦大鸡肉丸 300克	189	447.00
新腆盐 500克	568	728.78	散胖哥汤圆	181	858.81
仙岛增鲜酱油一级 380ml	520	770.40	小康家庭100红烧牛肉95克	180	553.10
散大米	486	6832.42	南峰园鸡蛋面 300克	179	395.70
自制糖蜜果	453	1296.29	秦大香辣鸡柳 300克	178	579.30
水果(燕大)	450	4996.96	散花生米	176	766.95
带皮前槽	425	6676.18	纯滴大豆色拉油51	174	6246.80
豆制品	392	1137.87	铁山楂(山楂精卷)	170	611.80
糕点	380	2111.21	开口酥	168	505.11
散白糖	344	1131.93	京京小孜然烤肠 400克	167	836.40
豆浆	334	725.66	华龙小康100五连包香辣面95克x5	167	503.00
素拌菜	333	840.96	双汇香脆肠 225克	162	531.40
带皮后丘	319	6284.41	汇源10%真鲜橙 500ml	161	1291.50
永真香油麻酱现场制作	281	2226.50	华龙小康家庭100红烧排骨面五连包95克X5	158	494.40
统一鲜橙多鲜橙汁饮料 330ml	272	777.00	仙岛果糖醋 380ml	158	150.40
涮羊肉调料 150g	256	315.90	庞业酱菜	157	392.94
白薯粉	235	585.38	麻辣拌	154	1185.11
大宝SOD蜜 100ml	223	1107.50	雪豹高级鞋油 335克(黑)	142	137.70
仙岛二级增鲜酱油 380ml	219	267.00	三鹿纯牛奶 250g	142	627.60
伊利百利包纯牛奶 227ml	213	1365.60	盼盼真实惠家庭号薯片 152克	141	262.80
可口可乐	211	1600.80	亲亲大薯片 120克	141	193.50
双汇鸡肠 30g*10	209	629.30	福临门色拉油 51	141	5573.00