数据挖掘技术在互补品与替代品弹性测量的应用

经济学院 陶文逸

指导教师 吴力波

内容提要： 本文利用关联规则与分层聚类技术分析洋快餐食品种类价量关系，寻找目标产品的互补品与替代品,最后构建常弹性方程测算目标产品与互补品、替代品之间的交叉价格弹性。实验结果较为理想，70%以上的参数结果显著且符合预期。本文的两项技术为替代品互补品分类与构建需求模型提供了新的筛选变量方法,也为经济学研究引入数据挖掘工具提供了应用层面的案例。

关 键 词： 关联规则 分层聚类 交叉价格弹性 互补品 替代品

中图分类号： F831.4

Cross Price Elasticity of Complements and Substitutes by Implementing Data Mining Techniques

Abstract: This article implemented data-mining techniques to analyze the cross relationship of different food product. Association rule is used to detect complements and hierarchy clustering is used to find substitutes. Then constant elasticity model is used to determine the cross elasticity of target products. The result showed 70% of the estimated parameters lie within expected values.

Keywords: Elastic Constant model； Association rule；Clustering；Data Mining

JEL： D40

目　录

一、引言

（一）本文研究背景及意义

数据挖掘技术曾因硬件存储上限与计算机计算能力的制约，很难运用于实践中。但就近几年，随着硬件技术的发展突破了存储能力的限制，分布式存储数据库可以让数据分别存储在不同的物理地点；不断更新换代的高效的算法让时间复杂度和空间复杂度都大幅度降低，让计算能力大幅度提高。数据挖掘技术的运用也逐渐趋于成熟，例如决策树，分类，聚类，支持向量机，关联规则等已经被广泛地运用到工业界。其中，最广为人知的经典案例，是沃尔玛公司的利用关联规则发现“啤酒与尿布"销量关系，挖掘到商业价值。

目前这些数据挖掘工具也逐渐被用于经济学的实证研究，但运用并不广泛。笔者认为主要有两点原因：第一，现有的统计工具相比数据挖掘技术所得到的结果逻辑上更清晰，因变量自变量的线性模型更具有解释力度。工业界偏重于实用性，准确度，拟合优度，预测能力，因此聚类等算法模型也受工业界偏爱。第二，数据挖掘技术往往需要较大的数据量，然而传统的经济学研究数据来源于官方公开或购买的数据，其本身就经过统计的处理，从数据量和维度无法满足数据挖掘技术的要求。

但是笔者相信，在未来会有越来越多的像 Google Trend 百度指数等数据会被公开用于支持学术研究，大数据挖掘技术也会逐渐弥补传统统计工具的技术盲点。例如时滞，缺乏微观基础等。因此，笔者尝试性地将数据挖掘技术引入经济学实证分析中，先从简单的商品价格弹性测量入手，以数十万条快餐店销量数据为试刀石，利用数据挖掘技术中聚类与关联规则，寻找指定商品的替代品与互补品，最后再从常弹性需求线性统计模型求出的交叉替代弹性来验证结果。这样或许能为经济学研究引入数据挖掘技术提供了切入口。

（二）研究思路、创新点与研究框架

1.创新点

诚然，我们可以基于经济学常识与经验来判断替代品与互补品，在商品数量少，

属性简单的情况下，根据人工主观判断的方法，确实更为有效。然而，当商品数量达到数百种甚至上千上万种时候，通过人工筛选的方式会显得力不从心，商品属性多元化，商品之间的关系多对多，这就给人工分类造成了不小的障碍。传统的统计方法是从中进行抽样摘选，或者将同类商品指数化的方式降低维度。但在抽样或者指数化的过程中，原始的信息会或多或少有所损失。

笔者认为，数据挖掘技术相比简单人工筛选，抽样统计判断的优点如下：1.可以对全样本数据分析，信息不会因抽样而有所缺失。2.在海量多元数据处理时，可以处理更为复杂的多元关系。3.方法是扎根于每一条微观细节，可以从中发现新的规律，如啤酒尿布的案例。

2.研究思路

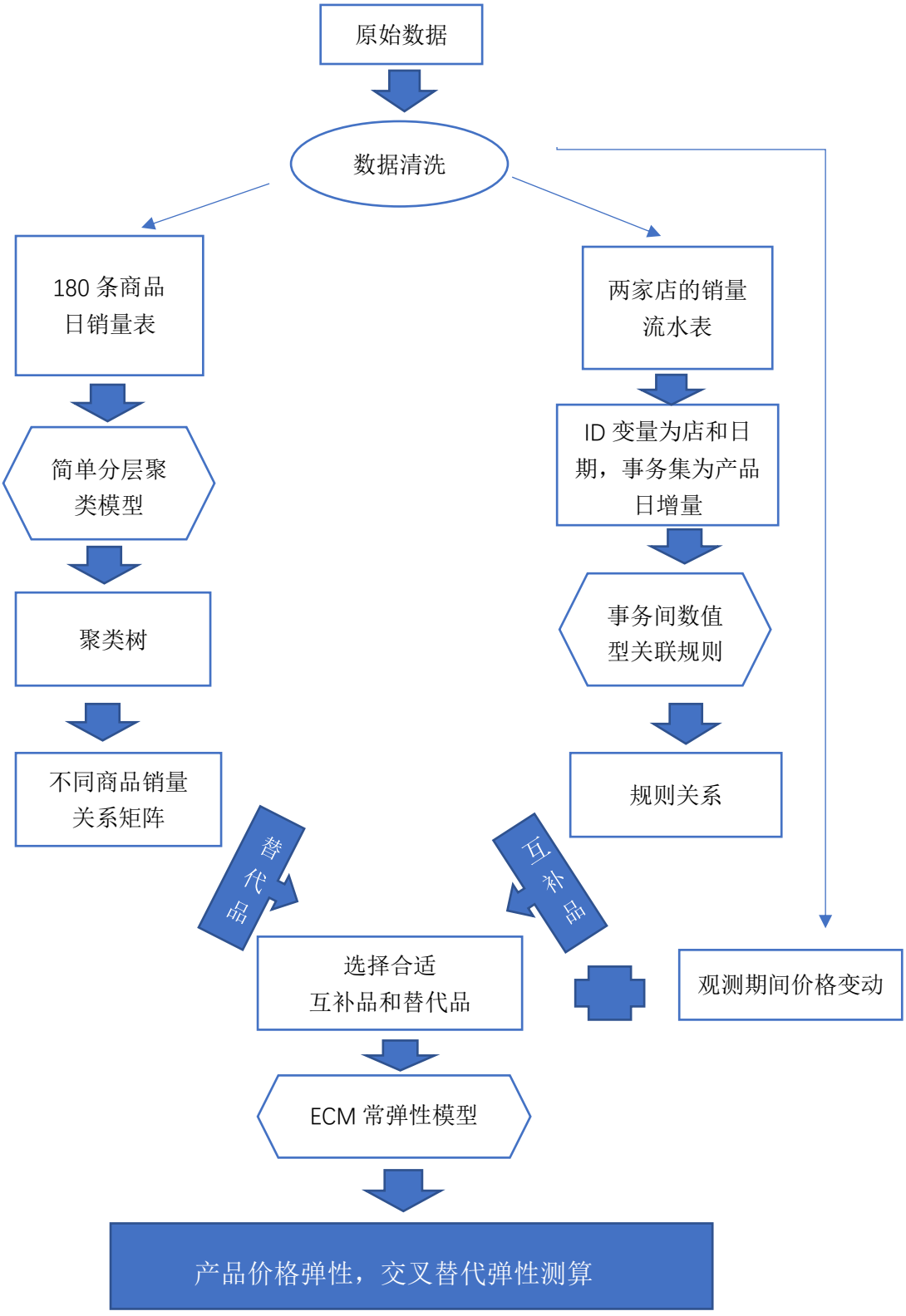
本文运用了两项数据挖掘技术，聚类与关联规则。聚类，是将具有相同属性的商品聚成一类，餐饮类的商品大多基于商品的功能，例如主食，小食，配餐等，然而并不是同分类下的产品就互为替代，判断往往是基于商品多重属性的综合考量。以餐饮业为例，纳入消费者考量范围的就有：口味，颜色，含水量，香味，菜系，碳水化物蛋白质比例，食物热量等，不同属性的商品搭配城篮子会形成对原有商品的替代与互补关系。因此，理想情况下，按照属性聚类需要事先设计问卷对消费者进行调研。聚类的另外一种方法是利用连续性变量的相关系数作为变量的相似性度量，本文中将采用销量相关系数的相反数作为矩阵的输入，其经济学依据是，消费者在面临预算约束的情况下，在替代品中只会选择有限个数的商品，因此具有替代关系的产品的销量具有此消彼长的关系。

关联规则是基于个体的购物清单进行商品篮子统计。如果大量的顾客在购买产品 A 的同时又购买了产品 B，那么商品 A 与商品 B 可能是互补品。关联规则相比与线性相关模型有天然的优势，具体有两点：首先，它是扎根落实于每一个个体的微观交易行为，能够从个体的交易，即每一笔的购物行为的累计成为一个具有统计意义的商品规律。其次，当数据量特别大时，而数据密度却比较稀疏时，而相关性分析依赖于数据的完整性，而关联规则对缺失值不敏感，所以可以很好的处理大型离散型数据集。

笔者准备以商品中的主餐的 10 款产品入手，来看与该 10 款产品相关的规则。有效地对规则进行剪枝，对于支持度可信度与提升度等权重选最高的前 10 项规则，最后将日销量数据做一阶差分，根据正负号转化成仅包含 0，1 元素的矩阵。1 代表销量增加，0 代表销量减少或持平。所得到规则结果即是，当改天产品 A 销量上涨时，产品 B 的销量上涨概率较大。因此，产品 A 与产品 B 很可能为互补品。

最后，互为替代品与互补品之间的价量关系用交叉价格弹性来验证，聚类求出的替代品交叉价格弹性期望值为正数，关联规则求出的互补品交叉价格弹性期望值为负数。

3. 研究框架


原始数据
数据清洗
180 条商品
日销量表
简单分层聚
类模型
聚类树
不同商品销量
关系矩阵
两家店的销量
流水表
ID 变量为店和日
期，事务集为产品
日增量
事务间数值
型关联规则
规则关系
替代品
互补品
选择合适
互补品和替代品
观测期间价格变动
ECM 常弹性模型
产品价格弹性，交叉替代弹性测算

二、文献综述

在分析商品价量关系方面，学者们的研究对象大多是农产品、工业原材料等。构建需求价格弹性模型也分为长短期，短期模型会根据 GDP 分组看国家间差异，长期还会引入宏观经济变量等。Asatryan，Armen A.（2004）利用三阶段选择调整模型分析了英国境内的猪肉销量。Laura Cornelsen, Rosemary Green(2016) 等人在全球范围内分析了食品销量与食品之间的交叉替代弹性，同时将人群分为高收入，中等收入，低收入组，分组分析食物价格弹性的大小比价，并给出政策建议。John Baffes and Allen Dennis(2013) 将宏观经济变量，汇率，利率与通胀率引入价格决定模型中，分析了影响小麦，大豆橄榄油等五种农产品价格的长期因素。Carlos M.Guerero-Lopez(2017)测算了 2012-2013 年智利软饮料之间的交叉价格弹性。

也有部分学者利用数据挖掘技术对产品销量预测做出了尝试，试图改进经济学的线性模型。A.J. Feelders(2002) 讨论了决策树与线性回归的单调性的差异，并以单调树重新计算了影响房价的模型。Sebastien Thomassey, Antonio Fiordaliso(2006) 利用决策树和聚类方法预测了服装品牌的销量。Dennis Maa, Marco Spruit(2014) 利用数据挖掘技术分析了短周期的商品需求，最终结果是利用单调二叉树的数据挖掘技术并不能提高短期的需求模型预测能力，并且认为模型变量方差越大，需求模型方法越简单。

简单的线性回归的结果在销量的预测能力上似乎占优，因此也有学者不做销量预测，只分析消费者选择偏好。黄玉佳（2015）分析了运动品牌服装的消费者偏好。同本文一样采用了关联规则和聚类方法，先将多种产品按销量进行聚类，作为结果输入 BASS 模型，预测消费者下一期的偏好，同时基于客户购物单，利用关联规则分析了消费者在颜色，店铺，品种上的选择。夏旻旻（2015）利用关联规则完成了茶叶消费者的偏好测度，具体方法是以问卷调查所得的每个潜在消费者的属性打分表，接着以打分表的单条数据为关联规则的一项事务，最后构建出消费者偏好模型。

本文所做的工作有别于上述文献，旨在利用数据挖掘工具寻找到具有相关关系的变量，再利用传统的统计线性回归方法来验证变量符合替代品与互补品假设。

三、相关技术原理

（一） 交叉替代弹性

替代品是经济学中的概念，指一种物品价格的上升会引起另一种物品需求的增加，那么这两种物品相互被称之为替代品。一般而言，替代品之间在功能上能部分或者全部替代另一种产品。公共交通与私家轿车互为替代品。

互补品同样在经济学中是重要的研究对象，他指的是共同满足一种欲望的两种商品，它们之间是相互补充的。比方说汽车与汽油之间互为互补品，他们都满足人自驾交通的需求。

一般商品的价量之间存在一定的关系，销量变动百分比与价格变动百分比的比值为商品的弹性。互补品与替代品的价量之间也同样存在一定的关系，商品 X 的价格变动会对互补品 Y 的销量存在影响，我们把互为互补品或替代品之间的弹性称之为交叉价格弹性。大多数情况下，互补品的交叉价格弹性小于零，替代品的交叉价格弹性大于零。公式如下：

$$E_{xy} = \frac{dQ_{dx}}{dP_y} * \frac{P_y}{Q_{dx}} \tag{1}$$

（二）关联规则原理

关联规则在超市应用中比较经典，称之为购物篮分析。基本原理是对单个客户购买的流水单进行商品频率统计，如果在海量流水单中我们商品 X 出现后与商品品 Y 也同时出现，那么就形成一个关联规则 $R: X \Rightarrow Y$，商品 X 在全样本流水单中出现的概率即为规则支持度 ：

$$\text{Support}(X \Rightarrow Y) = P(X) \tag{2}$$

商品 X 与商品 Y 共同出现的次数占商品 X 出现的次数比值为规则的可信度：

$$\text{Confidence}(X \Rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \tag{3}$$

除此之外，关联规则还有一个重要指标是提升度：

$$\text{Lift}(\text{X} \Rightarrow \text{Y}) = \frac{\text{Confidence}(\text{X} \Rightarrow \text{Y})}{\text{Support}(\text{Y})} = \frac{P(X \cup Y)}{P(X) * P(Y)} \quad (4)$$

上述公式的分母是商品 X 与商品 Y 出现的概率乘积，分子是商品 X，Y 共同出现的概率，如果比值大于 1 则为规则意味着商品 X 出现，提高了 Y 出现的概率。比值小于 1 则意味着 X 出现后，降低了 Y 出现的概率。

从上述内容可知，关联规则的原理并不复杂，在利用关联规则寻找目标商品的互补品时，只需要关注支持度高，可信度高，提升度大于 1 的规则即可。原理虽然简单，但是计算机的任务却非常繁重，对计算的开销非常大[①]。并且找一个复杂度较低的算法并不是一件简单的事情，因此算法是数据分析技术的核心。本论文采用 apriori 算法，算法的具体内容超出了论文讨论的范畴，不做赘述。

除了计算开销大的缺点外，它还存在一些其他的问题。第一，规则太多，比较杂乱，很难进行有效的摘取，大量的规则中有价值的信息比较少，很多都是频繁琐碎没有价值的规则。第二，作为筛选标准的支持度，可信度，提升度等参数是人为设定，缺乏绝对客观的标准，比较考验数据分析者的实践水平或行业经验。第三，模型本身的缺陷，事务集内的变量为简单的对称二元变量，即 0，1 类型，这往往忽略了数据的一些重要信息，例如销量与价格等，因此关联规则在面对连续型变量时处理力不从心。

（三）分层聚类原理

聚类的本质是将具有相似性特征的实体或者属性聚在一起，具有更多相同属性的产品相似度就越高，产品之间的替代性也就越强。因而聚类有助于在海量商品中挖掘替代品。聚类中的核心指标是相似性度量指标，往往统计的是离散变量的差平方和与离散变量的频数统计。聚类结果往往取决于聚类方式的选择，与相似性度量的选择，因此也是一个主观性较强的数据挖掘方法。因此我们不仅要分析分类结果，还要从聚类方式上分析类与类之间的相对性远近关系。

[①] 当数据集非常大时，测试的规则变多，哪怕简单的 5 个交易单，7 个产品需要测试的规则就有 602 条，若测试每条都要扫描一遍数据库，计算开销非常大。

一般而言，聚类的相似性度量是基于商品的属性值，以餐饮业为例，产品的属性有，蛋白质含量，蔬菜或肉类，口感，香味，卡路里含量等，然后采用欧式距离法，也就是不同项目在不同属性上的取值的差的平方和作为产品间的距离远近值，最后将这些产品之间的相似性矩阵作为聚类模型的输入。当然也可以采用相关系数作为变量输入。

本文将采用层次聚类方法，即开始每一个观测值自称一类，然后这些类每次两两合并，直到所有的类被据称一类为止。类与类的合并方式有三种，单链接，完全链接与平均链接。

单链接取类中距离最近的作为代表值，完全链接取最远的距离为代表值，平均链接取类中所有的平均距离。

本文中采用相关系数矩阵法来进行聚类，相关系数矩阵由每日销量组成，根据替代品的经济学定义可知，一件商品的价格上涨，导致另一样商品的销量上涨，因而两种产品有很好的替代属性。根据一般经济学原理，消费者因为受预算约束在 A 和 B 产品中进行选择，最终选择了 A 产品，可以认为是 A 与 B 之间存在替代效应。因而可以得出结论，两者的销量呈现出负相关，即 A 产品销量的增加导致了 B 产品销量的减少。所以替代产品的衡量标准即为相关系数的相反数。

三、数据准备工作与描述性统计

（一）数据源说明与数据清洗

原始数据源自于某餐饮业数据库，数据规模为 304879 条[②]，涵盖从 2013 年 5 月 20 日至 2015 年 5 月 17 日 728 天的两家店的销量数据，表 1 为原始数据结构：

表 1　原始数据信息

标签	说明	例子	补充说明
Day_id	日期的 ID	ID4888	每一天对应的 ID 不同
Store_code	商店编码	SHA227	全样本两家店
Item_id	食品编码	ID6040	总共 180 款产品
Sell_id	卖出方式	ID67373	每种产品的卖出方式
Unit_sold	销售量	34 个	有复数，产品退回或者操作错误
Alc_price	原价	9 元	剔除 0 元商品
Sell_price	实际卖价	9 元	套餐价，折扣价以及外卖等方式会对
Amount	总金额	306 元	卖出的总金额 Amount=sell_price*unit_sold
Combo_flag	套餐标识	0	如果是套餐，则为 2，不是套餐则为 0
Category	总分类	16	6 大总分类，ID9 或 ID37
Sub_category	细分子类	78	属于 category=16 下，snack 下一共 4 个子分类。
Calendar_date	时间	05/20/13	13 年 5 月 20 日，与 day_id 相互对应
Item_name	商品标签	“某某炸鸡块”	

资料来源：笔者根据数据库内容自行整理，本文中全部数据均来自于该数据库

该张关系表的主码是日期编号,商店编号,产品编号与售卖方式编号，表明可以用

[②] 因个人计算机处理能力有限，因此仅抽取了数千家店中的两家店的数据，即使如此规模也达到 30 万，并且在数据清洗与模型测试的时候已经倍显吃力。

日期，店，物品，卖出方式来唯一确定关系表里的一条记录。以上图例子来说，即在 2013 年 5 月 13 日这天，ID6040 这款产品以 ID67373 这种方式卖出了 34 个。

因为存在一些缺失值，异常值以及标签格式错误，数据挖掘工作之前要进行数据清洗，数据清洗工作异常繁重，往往占据数据挖掘工作的工程量的 60%以上的时间。处理的事务包括，剔除重复性元组，缺失值分析，异常值删除，以及数据传递过程中出现的乱码问题。

（二）产品分类、销售时间、销售金额占比

该快餐连锁店一共推出 6 大类产品，分别是主餐，饮料，甜品，配餐，小吃以及非食物类。前前后后两年间一共有 180 款产品。主餐是汉堡与肉卷居多，甜品以蛋挞为主，小食有薯条炸鸡块等，配餐有色拉与蔬菜汤，非食物类是一些玩偶手办纪念品。

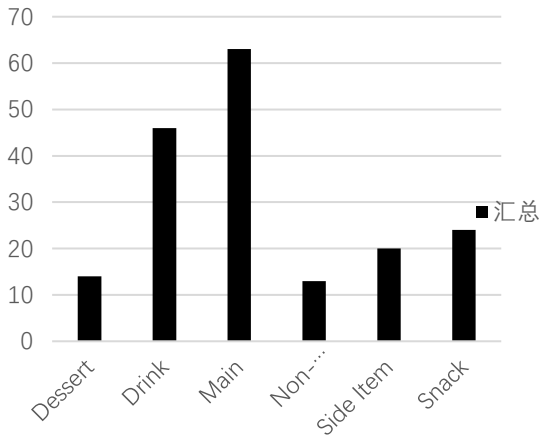


图 1 产品数量统计条形图

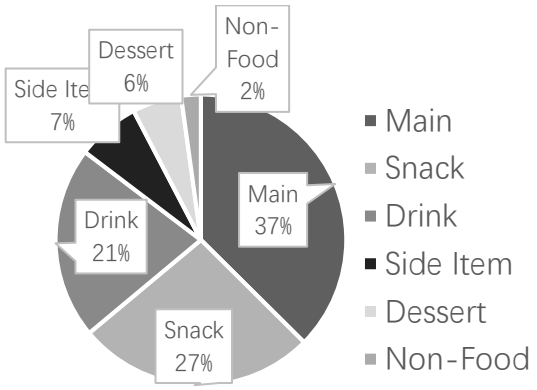


图 2 产品销售金额占比图

从各类产品销售金额占比来看，小吃与主食类占了大头，占比 64%，配餐，甜品，非实物类占比较小，只占了 15%。

从产品的销售时间上来看，并不是所有的产品都是在 2 年多的观测期间一直销售的，商家经常会对部分产品线进行调整，销量逐渐下滑的会停止售卖，同时也会时不时地推出一些新产品。非食物类的题材类纪念品存续期一般较短，大部分主食推出后销售时间都比较长，可能是产品线的退出成本较其他类产品高，并且会对原有产品销量产生一定冲击，所以保持相对比较稳定。考虑到数据的完整性，本文以 10 款经典主食类产品为切入点，以销量相关系数看该 10 款产品之间的替代效应，随后按照替代性强弱进行聚类，再以关联规则来看与产品相关的互补品。

图 3：部分产品销售时间图

四、分层聚类结果下的同类替代品

就如前文所讨论，聚类是为了看产品的相似程度，把相似程度越近的产品聚成一类。根据替代品的含义，替代品是能满足消费者的需求的具有相近功能的产品，因此理想情况下，通过将商品不同属性进行相似度计算得到的产品距离可以很好地满足聚类的要求。

然而原始数据并没有包含商品具体属性的内容，所以商品的具体形态，口味我们也不得而知。时间充裕与条件允许的情况下，应进行实地考察，或者找一批消费者进

行访谈，这样可以得到更加满意的结果。

本文采取了一种差强人意的替代方法。根据经济学的原理，替代品较强的产品具有相似的功能，能够满足消费者的某一特定需求。因此在预算条件约束的情况下，两款替代品的销量会出现此消彼长的情况。给予这一点，同类产品下，销量出现此消彼长的关系，我们认为该对产品可能互有较强的替代性。因此该方法仅要求产品的日销量与大类别标签。

诚然，本方法具有不可忽视的缺陷。第一，回避了聚类算法的本质。聚类算法的本质是根据项目的属性来判断项目的相似程度。然而该方法采用的是销量为唯一属性值。第二，日销量数据形成的相关系数矩阵，并不能严格地证明产品的替代性，我们只能说明产品之间可能有替代关系。如果同类产品之间销量相关系数为正，我们并不能解释两者之间不存在替代关系。这种正相关关系很可能是由第三个解释变量所造成的，比方说收入效应带来的消费者的购买力增强，使得两类产品随着时间日销量同时增加。

如果我们做一些假设，两年期间不存在收入效应等其他变量的影响，并且销量变化能够很好地反应消费者偏好变化，是产品之间替代性较好的代理变量的话，那么，我们可以通过日销量的相关系数矩阵作为层次聚类模型的数据输入。

表 2 十大热销主食类产品销量相关系数矩阵（成对）

ID 代码	5976	5992	5968	6107	5969	11399	12324	5306	5977	6106
5976	1.000									
5992	0.264	1.000								
5968	-0.304	-0.107	1.000							
6107	-0.441	-0.225	0.173	1.000						
5969	-0.079	0.342	-0.107	0.115	1.000					
11399	-0.272	-0.195	0.453	0.541	-0.187	1.000				
12324	0.023	-0.109	-0.015	-0.169	-0.150	0.057	1.000			
5306	-0.351	-0.190	0.434	0.260	-0.198	0.475	0.212	1.000		
5977	-0.296	-0.132	0.427	0.452	-0.107	0.563	0.083	0.610	1.000	
6106	-0.048	-0.117	0.248	0.415	-0.177	0.374	0.129	0.369	0.275	1.000

在上表 2 中，我们发现 ID 5976 款产品与其他产品具有很好的替代效应，但是该

产品与其他产品相关系数波动范围较大，主要原因可能是该产品的销售时间比较集中，样本量比其他产品较少，销售天数仅有百天，而与之相比的覆盖了整个观测期 728 天的其他同类型产品相比，时长较短，笔者推测可能是在观测期间商家推出的新产品，新产品的引入对原有系列产品有或多或少的价量冲击。笔者再次注意到另一款产品 ID5992 也具有较好的替代品性质，其他产品销量有线性负相关。该款产品的销量最高，推测是该品牌的主打产品之一，回头客在首次光顾该店后可能会选择其他同类产品，因而 ID5592 与其他产品也同样具有较强的替代性。剩余的 8 款产品销量都分别与其他 3 到 4 款产品互有线性负相关，但并不是全部所有产品销量之间都是负相关系数，这说明了，某一大类产品并不是所有产品都相互具有替代性，产品还可以根据其属性来划分为更小的类别。

基于表 2 的销量相关系数矩阵，笔者针对其进行分层聚类。来看不同产品之间的相似性远近。根据传统的聚类方法，项目之间的距离矩阵中的数值是标准化后的$[0,+\infty)$之间的连续变量。相似系数越高，数值越接近于 1，距离越远数值越大。因此笔者将相关距离矩阵做简单处理，$\text{distance} = \rho + 1, \text{distance} \epsilon (0,2)$ 变换为距离矩阵，作为分层聚类的数据输入。随后按照单链接，完全链接，平均链接的方式绘制分层聚类树。

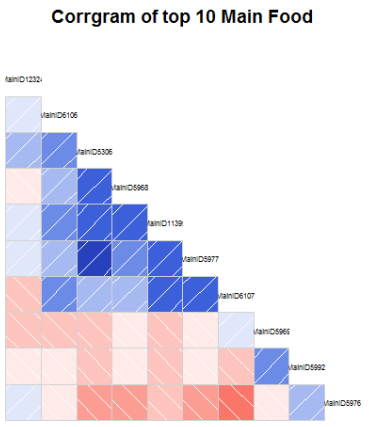


图 4 相关关系马赛克图

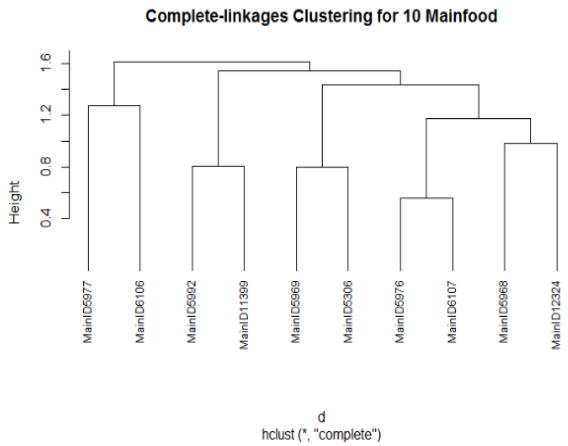


图 5 完全链接分层聚类树

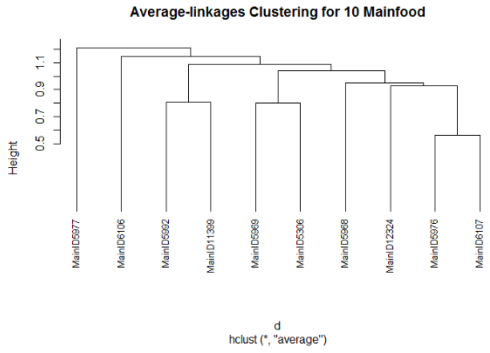


图 6 平均链接分层聚类树

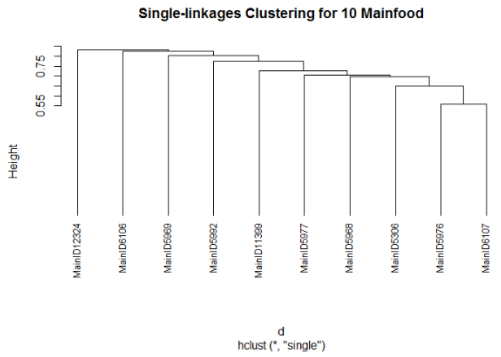


图 7 单链接分层聚类树

层次聚类的操作方式是，即开始每一个观测值自称一类，然后这些类按照距离最近的每次两两合并，直到所有的类被据称一类为止。因此数据结构为简单二叉树。按照不同的聚类方式，我们可以得到不同的聚类结果。不管什么聚类方式，第一次合并的都是相同的，差别在于如何选择新聚成的一类与其他剩余类别之间的距离，单链接选择一个最短距离，完全链接选择最长距离，平均链接选择平均距离带入新的距离矩阵。

根据图 7 来看，单链接后的结果层次分明，二叉树深度为 10，具有较好的聚合性质，以其中一个变量为中心点，层层包裹，越在外层就距离就越远，相似度就差，该产品与其他产品之间的替代性就越弱。我们发现 ID5976 处于中心的最内层，这并不出乎意料，因为其他观测值都与其有较近的距离。所以如果观测值中出现中心点时，单链接的聚类结果是以该值为中心，一层一层同心圆似的聚集在其周围。单链接有助于我们发现中心点。不助于我们发现除中心点外其他观测值之间的关系。

完全链接后的结果类别清晰，图 5 显示二叉树的深度为 4，同层次类与类之间距离较远，具有较好的区分性质。大都两两互聚成一类，因此可以很好地看清与各个变量距离最近变量的关系，但是由于选取的是最远距离，所以外层变量与内层变量之间的联系不够紧密。根据此我们可以看出每个产品替代性最强的产品是哪个。

图 7 的结果是平均链接，选取的是平均距离，因此分类的情况介于单链接与完全链接之间。每层聚类后的取平均距离的权数相同综合单链接与完全链接的优点。笔者基于平均链接的结果，综合考量单链接与完全链接，以处于相邻内外层和同层次选择每款产品的替代品。遴选出的替代品在表 3 中：

表 3 聚类结果下指定产品的替代品

指定商品	1 号替代品	2 号替代品	3 号替代品
Main_ID 5977	Main_ID 5992	Main_ID 5969	
Main_ID 6106	Main_ID 5977	Main_ID 5992	Main_ID 5969
Main_ID 5992	Main_ID 11399	Main_ID 6107	Main_ID 5968
Main_ID 11399	Main_ID 5992	Main_ID 5969	
Main_ID 5969	Main_ID 5977	Main_ID 6106	Main_ID 5306
Main_ID 5306	Main_ID 5992	Main_ID 5969	
Main_ID 6107	Main_ID 5992	Main_ID 5969	
Main_ID 5968	Main_ID 5969	Main_ID 5992	
Main_ID 12324	Main_ID 5992	Main_ID 5969	

五、关联规则结果下的异类互补品

如前文所述，关联规则是基于购物篮数据结构，从单项商品的购买记录来统计某些商品同时出现频率的数据挖掘工具。该方法对数据精细化要求程度较高，这也是数据挖掘技术与传统统计的区别。传统统计所处理的大多是汇总后的聚集数据，已经初步对数据进行了汇总，例如日销量，因而在记录数据的时候，简单的求和让数据集失去了一些有用的重要信息。同时在面对日销量这种连续性变量时候，要使用关联规则会比较棘手。处理方式大多是转化成离散性变量，但是划分区域人为干扰因素较强，并且若是区域划分较多，会导致规则极剧膨胀。但是当数据仅仅只有每日汇总值时，处理连续型变量就是一个无法绕过的问题。

经过仔细权衡，笔者建议所采取的权宜之计是，对每个产品日销量进行一阶差分，以正值为 1，负值为 0，然后进行关联规则测试。规则的$X \Rightarrow Y$的含义就是，当 X 销量上涨时，该天 Y 销量上涨可能性比较大。这种方法与简单的测算相关系数差别在于，可以根据支持度，置信度与提升度来看，X 与 Y 以及$X \cup Y$在全部事件中出现的频率。

这样处理数据存在的问题是显而易见的，我们只能够通过结果推断销量有可能上涨，但是上涨的幅度是未知的，同时与最原始的购物篮分析相比是差强人意的。原始购物篮所能得到的规则是客户在购买 X 产品的时候很可能会购买 Y 产品，这种强规则是基于大量微观的个体消费行为的，而统计汇总后的日销量遗失了这部分有价值的信

息。

通过 Apriori 算法所求出的结果包含大量的规则，一般情况下，数据挖掘分析师将规则划分为三类：1.可执行的规则 2.无关紧要的规则 3.无法解释的规则。可执行的规则提供非常清晰有用的洞察，例如啤酒与尿布之间的规则。无关紧要的规则是平凡且显而易见的规则，例如奶粉与婴儿尿布。还有一些是无法解释的规则，可能是因为巧合或是需要专业人事进行进一步分析的规则[③]。

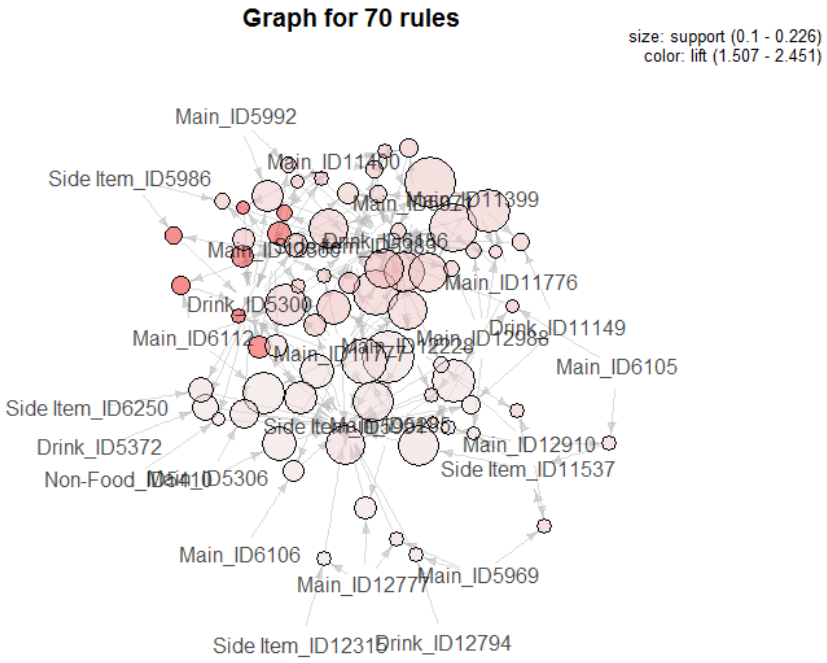


图 8 提升度前 70 的规则图

笔者按照提升度进行排序取前 70 条关联规则，并将其绘制在同一张图上，可以看到观测量之间的复杂关系之多，要在这些规则中进行有效的筛选是比较考验技术能力的。笔者能力有限，采取了一种简便方法。提取关联规则是左手边 X 中凡是包含 10 款主餐产品中任意一种的，右手边是除去主餐以外的其他产品，并按照提升度从高到低排序，取前 15 条，显示在表 4 中。

[③] 例如在图 8 中，我们可以看到形成的规则很可能是杂乱无章的，图中的箭头代表一条规则，圆圈大小代表支持度高低，颜色深度越深代表替升度越大。

表 4 提取的前 15 条关联规则

Rules	Support	Confidence	Lift
{Main_ID11399,Main_ID12988} => {Drink_ID6186}	0.103164	0.824176	1.697382
{Main_ID5977,Side Item_ID5984} => {Snack_ID6034}	0.176066	0.805031	1.706291
{Main_ID5977,Snack_ID6040} => {Snack_ID6034}	0.210454	0.831522	1.762438
{Main_ID5977,Side Item_ID5319} => {Snack_ID6034}	0.18707	0.824242	1.747009
{Main_ID12360,Main_ID5992} => {Drink_ID6186}	0.112792	0.803922	1.655668
{Main_ID11399,Main_ID12360} => {Drink_ID6186}	0.114168	0.813725	1.67586
{Main_ID12910,Main_ID5969} => {Side Item_ID11537}	0.105915	0.836957	1.709178
{Drink_ID11149,Main_ID11399} => {Drink_ID6186}	0.116919	0.825243	1.699579
{Drink_ID12794,Main_ID5969} => {Side Item_ID5991}	0.108666	0.814433	1.529956
{Main_ID12777,Main_ID5969} => {Side Item_ID5991}	0.10729	0.821053	1.542391
{Drink_ID5300,Main_ID5992} => {Drink_ID6186}	0.160935	0.806897	1.661795
{Drink_ID5300,Main_ID5306} => {Side Item_ID5991}	0.154058	0.817518	1.535751
{Main_ID11777,Main_ID6106} => {Side Item_ID5991}	0.126547	0.814159	1.529441
{Main_ID11777,Main_ID5306} => {Side Item_ID5991}	0.169188	0.803922	1.510209
{Main_ID11399,Main_ID11777} => {Drink_ID6186}	0.185695	0.828221	1.705713
{Main_ID11399,Side Item_ID5383} => {Drink_ID6186}	0.222834	0.81407	1.67657

以第一条规则为例，{Main_ID11399,Main_ID12988} => {Drink_ID6186} 当 Main_ID11399,Main_ID12988 两款产品销量同时上涨时，Drink_ID6186 销量上涨概率较大，支持度 0.131 是指前两款产品上涨情况出现次数占总天数的 13.1%，置信度 0.824 是指两款主食类产品销量上涨后，82.4%的情况下产品饮料类 ID6186 销量也同时上涨。提升度是指前两款产品出现上涨的情况下，比不出现上涨情况下，提高了 0.69 倍 ID6186 产品销量的上涨概率。因而我们可以推测 Drink_ID6186 是上述两款产品的互补品。

基于表 4 的结果，我们将主食类产品的互补品统计如下表：

表 5 主食类产品的互补品

	1 号互补品	2 号互补品	3 号互补品
Main_ID 5977	Snack_ID6034	Side Item_ID5984	Snack_ID6040
Main_ID 6106	Side Item_ID5991		
Main_ID 5992	Drink_ID6186		
Main_ID 11399	Drink_ID6186	Drink_ID11149	Side Item_ID5383
Main_ID 5969	Side Item_ID11537	Side Item_ID5991	
Main_ID 5306	Side Item_ID5991		

六、交叉价格弹性测量

本文的最后一部分是为了验证上述结果所找到的互补品与替代品的关系是准确的，因此有必要测算产品之间的交叉替代弹性。根据经济学的基本原理，互补品的交叉替代弹性大于小于零，替代品的交叉替代弹性大于零。这比较好理解，因为替代品价格的上涨，互补品价格的下降，会使得消费者在选择的时候更倾向于选择价格便宜的本产品。相反同理。

那么余下的问题就是看每款产品的价格在观测期间的变动情况。然而，较为麻烦的是，该寡头垄断厂商的定价策略为了追求利润的最大化，采取的是价格歧视策略。即定价并不是单一的，店内订餐与网上订餐价格不同，处于机场火车站附近的店价格高于其他店面，该商家还根据就餐时间不同推出不同的促销活动，例如 8 点前早餐优惠等，当然还有更为常见的是大量折价券发放。这使得该产品即使同一款产品在同一天以不同售卖方式售出的价格是不同的，因此增加的问题分析的难度。

笔者的处理方式是采用价格的加权平均数，以售出的数量为权重：

$$P_i = \sum_{j=1}^{n} p_{ij} w_j = \sum_{j=1}^{n} P_{ij} \frac{Q_{ij}}{\sum_{j=1}^{n} Q_{ij}} \qquad (5)$$

P_i 第 i 款产品的加权平均价格

p_{ij} 第 i 款产品以第 j 种方式售卖出去的价格

Q_{ij} 第 i 款产品以第 j 种方式售卖出去的销量

最后笔者采用常弹性方程(Elastic Constant Model)来测算交叉价格弹性，该模型假定观测期间价格弹性不发生变化，为固定的常数：

$$Q_{it} = AP_t^{\beta_1} P_{ct}^{\beta_2} P_{st}^{\beta_3} e_{it} \qquad (6)$$

Q_{it} 第 i 款产品的 t 时刻的销量

P_t 第 i 款产品以 t 时刻的加权平均价格

P_{ct} 该款产品的互补品的 t 时刻的加权平均价格

P_{st} 该款产品的替代品的 t 时刻的加权平均价格

对上式左右两边取对数：

$$\ln(Q_{it}) = \beta_0 + \beta_1 \ln P_{it} + \beta_2 \mathrm{In} P_{ct} + \beta_3 \ln P_{st} + U_{it} \qquad (7)$$

β_1 产品的价格弹性

β_2 互补品的交叉替代弹性

β_3 替代品的交叉替代弹性

根据表 6 的结果，我们发现产品的价格弹性都符合预期，有两款商品的价格弹性与预期相反。ID 6106 与 ID 5968 的价格弹性都是正号显著，这说明这两款产品的价格与销量存在正相关关系。该结果有些出乎意料，笔者推测，这种关系的形成可能是由暂时无法观测的其他变量导致的，例如有可能是竞争对手在该产品线上退出，或消费者偏好随时间转移等，商家乘势提价以扩大利润率。

另外，36 个交叉替代弹性中，有 7 个不显著，5 个异常值。笔者将不显著的变量从回归模型中剔除，以避免影响其他变量的显著程度。3 个互补品的交叉替代弹性为正值显著，2 个替代品的交叉替代弹性为负值显著，进一步发现异常值所出现的模型平均 VIF 在 1.9 至 2.3 之间，一般认为 VIF 大于 2，模型就存在多重共线性问题，并且增减解释变量后发现模型部分参数估计值发生交大变化。因而笔者推测是多重共线性导致了异常值的出现。对此，暂无解决办法。

表 6 产品弹性测算表

	变量表	Main_ID 5977	Main_ID 6106	Main_ID 5992	Main_ID 11399	Main_ID 5969	Main_ID 5306	Main_ID 6107	Main_ID 5968	Main_ID 12324
价格弹性	P	-1.9780***	2.0949***	-1.9252***	-0.84712***	-1.63625***	-1.7276***	-1.71***	2.09867***	-0.51541*
交叉价格弹性	Main_ID 5977		-1.9194***		-1.13540**					
	Main_ID 6106					不显著				
	Main_ID 5992	不显著	0.2998**				2.6541***	0.16450**	不显著	不显著
	Main_ID 11399	1.1043***		1.0934*						
	Main_ID 5969		3.27***		不显著		0.6036***	2.68217***	5.59905***	22.277***
	Main_ID 5306					2.51584***				
	Main_ID 6107			1.5052***						不显著
	Main_ID 5968			0.6218**	0.3183***					-0.34884*
	Main_ID 12324					0.52032*			-0.62529***	
	Snack_ID6034									
	Side Item_ID5991	-0.90099***	-1.1764***			-2.60776***		不显著		
	Drink_ID6186			-3.3179***	0.8223***					
	Side Item_ID11537					10.617***				
	Side Item_ID5984	-0.6040*								
	Snack_ID6040	0.18408*								
	Side Item_ID5383				-2.07570*					
	Drink_ID11149						不显著			
	调整 R 方	0.1837	0.4657	0.2363	0.1368	0.3774	0.2482	0.2502	0.4644	0.1721
	Prob>F	2.56e-14	2.2e-16	2.2e-16	2.895e-10	2.2e-16	2.2e-16	2.2e-16	2.2e-16	5.144e-10
	自由度：	683	609	346	515	571	674	664	568	248

注：（1）*、**、***分别表示在 1%、0.1%、0.001%水平上显著。

七、结果与不足之处

笔者在不存在严重的多重共线性问题模型中，进行变量的删减，发现所加的其他产品价格与销量大多都不显著，这也从侧面说明了数据挖掘技术所获取的互补品与替代品的有效性。

关联规则技术在商品数量巨大的时候，优势就比较明显。传统方法需要把全部变量放入模型中并逐个按照显著性删除，比较拟合优度的方式机械且费时费力。然而，数据挖掘技术也并不是没有缺点。聚类算法对操作者的要求较高，属性与模型的选择都依赖于分析者主观判断，聚类的结果也比较难以解释，因此该项技术优势并不太好操作。

本文采用的数据挖掘与实际的工业大数据挖掘实践还存在着一些距离，有以下几点可以改进的地方。首先是数据量虽然有 30 万，但与全样本数百万数据比起来只占冰山一角，数据量的膨胀带来的是计算机性能的开销，以个人台式笔记本的内存及 CPU 并不能足以应付海量数据处理。其次数据本身也存在一些缺陷，一些重要的信息没有包涵进入模型中来，聚类缺少产品的详细属性，例如产品包装，口味等，关联规则缺少交易单类数据类型，产品销量的模型中缺少，广告投入，竞争对手的产品价格等其他变量。最后，常弹性方程的假设弹性在观测期间内不变，该假设较为严格，如果用弹性随时间而变的自回归分布滞后模型，可能会有更好的结果。

参考文献

[1] 黄玉佳：数据挖掘技术在消费者偏好中的应用[D]. 北京林业大学, 2015。

[2] 夏旻旻：基于置信规则库的消费者偏好测度研究[D]. 合肥工业大学, 2015。

[3] Asatryan A A. Data mining of market information to assess at-home pork demand[J]. Texas A & M University, 2004.

[4] Baffes J, Dennis A. Long-Term Drivers of Food Prices[J]. Social Science Electronic Publishing, 2013.

[5] Cornelsen L, Mazzocchi M, Green R, et al. Estimating the Relationship between Food Prices and Food Consumption—Methods Matter[J]. Applied Economic Perspectives and Policy, 2016, 38(3):ppw010.

[6] Feelders A J. Prior Knowledge in Economic Applications of Data Mining[C]// European Conference on Principles of Data Mining and Knowledge Discovery. Springer-Verlag, 1999:395-400.

[7] Guerrero-López C M, Unar-Munguía M, Colchero M A. Price elasticity of the demand for soft drinks, other sugar-sweetened beverages and energy dense food in Chile[J]. Bmc Public Health, 2017, 17(1):180.

[8] Thomassey S, Fiordaliso A. A hybrid sales forecasting system based on clustering and decision trees[J]. Decision Support Systems, 2006, 42(1):408-421.

后　记

随着毕业论文的完成，大学本科的学习生活就要结束了。我非常感谢我的导师吴力波，每个学期一次的导师见面会，给我带来的积极影响很大。后来参加的学术课题小组让我首次接触到大数据概念及其领域，我萌生了想继续朝着该领域深造的想法。我还要感谢家人，感谢父母对我的养育之恩，感谢他们对我的无条件的支持。