2015 年第四届全国大学生

统计建模大赛

论 文 题 目 ： **大数据背景下消费者网络购物分析[1]**

参 赛 学 校 ： **中南财经政法大学**

参 赛 者 ： **程曦 杨律迅 张冰洁**

指 导 教 师 ： **蒋锋**

完 成 时 间 ： **2015 年 6 月**

中南财经政法大学

[1] 注:该论文获得由中国统计教育学会举办的“ 2015 年（第四届）全国大学生统计建模大赛 ”大数据统计建模类本科生组三等奖。

摘要

如今，“网购”已成为人们生活的一部分。在购物的同时，人们的购物记录被保留在网站数据库内。网店老板经常关心的问题是顾客的购物习惯。他们想知道：“什么商品组或集合顾客会在一次购物时同时购买”。本论文通过“蘑菇街”购物网站一段时期所有顾客购买物品的清单和相应商品的利润，分析顾客购物习惯，制定促销方案，从而增加销量。 问题的数据来源为“蘑菇街”购物网站。本论文旨在解决三个问题中，首先，根据给出的购买数据，建立一种数学模型来定量表达顾客在网购时采购的多种商品之间的关联密切程度，其次，寻找一种快速有效的方法分析出哪些商品是最频繁被同时购买的，最后，给出使得经营者受益最大的促销方案。

在问题一中，参照 Manhattan 距离公式，定义参数 λ 反映不同商品间的密切程度，但考虑到原公式受“有顾客未同时选购这组商品”的影响过于巨大，参照 Jaccard 距离对公式进一步优化，得出求解密切程度的最优算法模型，并举例计算出几组相关商品的密切程度。

在问题二中，我们从商品数量为 2 开始逐个递增，依次计算不同商品数量下所有组合的购买频率，每次计算都对结果进行筛选并作为下一次计算的参考数据。最后，我们定义优度系数这一概念，按优度系数对所有组合进行降序排序，并列举出最优的前 20 位组合方案。

在问题三中，我们首先计算出每组商品组合的利润和，再用利润和与优度系数共同表示利润指标，以比较组合的预期获利大小。按照利润指标降序对所有筛选出的组合排序后，找出利润指标最大的若干商品组合，并由此设计相应的促销方案。以利润指数最高的商品组合：368 号、529 号商品为例，该组商品的利润指数高达 189487.6，在所有商品组合中排首位，说明促销可能带来的利润最大，商家可以通过将其放在首页醒目处、加大宣传、组合减价销售等方法进行促销。

关键词网购商品 Manhattan 距离 密切程度 预期获利 促销方案

一、问题描述

随着网络的发展，网购已经成为越来越多人的消费习惯。为了了解顾客的购物习惯与消费心理，网店老板通过调查“什么商品组或集合顾客会在一次购物时同时购买”，可以把这些“同类商品”相互关联在网页内，以便于顾客在一次网购行为中购买更多的商品，从而引导顾客消费增加销量。现在已知某购物网站一段时期所有顾客购买物品的清单和相应商品的利润，需要研究得出合理的顾客购物习惯分析报告，并提供一个促销计划的初步方案。

具体地说，我们需要完成以下任务：

1、通过分析一段时期内 4625 个顾客对 999 种商品的购买记录数据，建立一种数学模型，定量的表达网购商品中多种商品间的关联关系的密切程度；

2、根据在问题 1 中建立的模型，寻找一种快速有效的方法分析出哪些商品是最频繁被同时购买的；

3、已知这 999 种商品对应的利润，试根据前面建立的模型，给出一种初步的促销方案，使网店经营者的效益最大。

对于问题一，我们首先想到引入一个能够定量反映某些商品组合间密切程度的参数，通过比较各个参数的大小便可直观反映出不同商品组合的密切程度。首先我们将某一顾客购买某一商品的情况用不同参数表示，用 1-0 二值量表示这一购买行为是否发生。接着我们参考聚类分析中对元素距离的定义方法，用参数对不同组合的密切程度进行定义并比较，还对参数模型进行了优化，得出最合理的模型来表示参数大小。

问题二的目标是寻找尽可能多的商品被同时购买的信息。这一目标是使商品数量和同时购买频率这两个量都尽可能的大，而注意到这两个量的关系是此消彼长的，所以我们用一个与二者同时相关的系数来刻画二者大小，用此系数最大化来寻找目标。对于某一固定商品数量，直接枚举寻找最大值的计算次数太多，运

算效率很低，所以我们从 2 件商品开始，依次增加商品数量进行计算，每次计算以上一次计算的结果为基础，只取其中频率较大的结果参与新的组合并计算频率。这样，我们在找到每个商品数量下的较多频率的结果后，再通过系数最大化找出该问题需要寻找的目标。

在考虑如何增加购物网站的效益时，需要综合考虑密切度高的商品组合种类数与购买的次数。在得到总利润之后，我们用每组商品的利润和与该组商品销售频率的乘积表示销售这种商品组合的利润指标。通过比较不同商品组合的利润指标的相对大小，并进行排序，得到利润指标最大的若干组合。根据这一数据合理安排促销方式，使购物网站的效益达到最大。

就具体促销方案而言，我们考虑把销量大的、不同类别的商品组合中的利润最大商品放到网站醒目位置，引起消费者关注；在商品网页的下方，显示它的一组相关商品，便于消费者找到需要的整套商品组合。这些广告宣传的方法，在一定程度上也可以增加销量。

二、指标选择

表 1 指标选择表

变量	说明
x_{ij}	顾客 i 对商品 j 的购买情况
y_j	表示商品 j 被顾客购买的情况
$\lambda_{a_1a_2\ldots a_j}$	模型优化前，a_1，a_2，…，a_j 这几种商品间的密切程度
$S_{a_1a_2\ldots a_j}$	模型优化后，a_1，a_2，…，a_j 这几种商品间的密切程度
n	商品件数
$T(n,s)$	n 件商品（编号集合为 s）的同时购买频率

$A_i\ (2 \le i \le 999)$	i 件商品同时购买的所有组合的集合
$A_i'(2 \le \mathrm{i} \le 999)$	A_i 中所有频率不小于 N 的组合
N	判断组合是否有效的频率界限
$A_n(i)$	集合 A_n 中的第 i 个元素
$A_2(\mathrm{i})_s$	集合 A_n 中的第 i 个元素的商品编号组合属性，为 n 个编号的集合
$A_2(\mathrm{i})_s(j)$	集合 A_n 中的第 i 个元素的商品编号组合中的第 j 个商品的编号
Ω	自定义的优度系数
B	所有有效方案的集合
$B(i)_n$	第 i 方案同时购买频率
$B(i)_{\mathrm{t}}$	第 i 方案的商品数量
θ	自定义的利润指数
$B(i)_\theta$	集合 B 中第 i 个组合的利润指数值

三、数据描述

附件 1 中的表格数据显示了“蘑菇街”购物网站主要购物数据，一段时期内 4625 个顾客对 999 种商品的购买记录，表格中每一行代表一个顾客的购买记录，数字代表了其购买商品的网站内部编号。

附件 2 给出了这 999 中商品的对应的利润。

四、模型建立

（一）模型假设

（1）假设一段时间内购买的商品组合为几乎同时购买，存在关联性；

（2）假设每位顾客购买某种商品时只购买一件而不是多件；

（3）假设顾客对每种商品的需求量都是一定的，购买频率只受该商品与其他商品关联性密切程度的影响；

（4）假设模型二中舍去的小于 60 的组合数据，在组合中的产品数增加时，其购买频率并不会增长或不变，即舍去数据不会对所求结果产生影响。

（5）假设模型三中考虑利润时，消费者在购买网站的某种推荐商品时，有很大可能性会购买与该商品形成商品组合中的全部产品。

（6）假设市场需求弹性是大于一的，当价格降低时，需求量（即销量）会增加，薄利多销。

（二）问题一

本文问题一中需要我们对每位顾客购买每种商品的购买记录进行分析，从而找出多种商品间关联关系的密切程度。我们联想到聚类算法中的距离系数，即密切程度越大的商品其距离越近。在选取距离系数类型时，我们有如下思路。

1.Manhattan 距离系数算法

Manhattan 距离是两点间的折线距离，对于 0-1 的二值型变量，恰为两者有差异的属性项数和。借此思路，我们考虑以下算法。

为了用数学符号表示顾客购买商品的情况，假设有 m 名顾客（m=4625），n 种商品。我们用二值型变量 x_{ij} 表示第 i 个顾客购买商品 j 的情况：若顾客 i 购买了商品 j，则 $x_{ij}=1$；若顾客 i 并未购买商品 j，则 $x_{ij}=0$。同时，从商品的角

度来考虑, 我们用二值变量集合 y_j 表示商品 j 被顾客购买的情况, $y_j=(x_{1j}, \cdots, x_{nj})$ 显然$1 \le i \le m, 1 \le j \le n$。

假设此时有两种的商品 p 和 q 需要我们比较其关联关系的密切程度, 即比较 y_p 与 y_q 中各个分量的相似性。若相似性很高, 说明这两种商品 p 和 q 容易被同时购买, 即这两种商品的密切程度很高。为了直观地表现这种密切程度, 我们引入 λ 这一变量表现这种密切程度。则 λ 的表达式为

$$\lambda_{pq}=\frac{1}{m}\sum_{i=1}^{m}K\left\{x_{ip}=x_{iq}\right\}$$

其中, 当 $x_{ip}=x_{iq}=1$ 时, $K\left\{x_{ip}=x_{iq}\right\}=1$, 否则 $K\left\{x_{ip}=x_{iq}\right\}=0$。

也就是说, λ_{pq} 表示所有消费者中, 对 p,q 这两种商品购买行为相同（同时购买 p,q 或者同时不购买 p, q）的顾客, 在所有顾客当中所占的比例。因此, 如果 λ_{pq} 的值很大, 则说明商品 p, q 的关联性密切程度越大。

当有三种商品 a,b,c 时

$$\lambda_{abc}=\frac{1}{m}\sum_{i=1}^{m}K\left\{x_{ia}=x_{ib}=x_{ic}\right\}$$

其中, 当 $x_{ia}=x_{ib}=x_{ic}=1$ 时, $K\left\{x_{ia}=x_{ib}=x_{ic}\right\}=1$, 否则 $K\left\{x_{ia}=x_{ib}=x_{ic}\right\}=0$

以此类推, 考虑商品 a_1, a_2, …, $a_j(1 \le j \le n)$ 这几种商品的密切程度时,

$$\lambda_{a_1a_2\ldots a_j}=\frac{1}{m}\sum_{i=1}^{m}K\left\{x_{ia_1}=x_{ia_2}=\ldots=x_{ia_j}\right\},(1 \le j \le n)$$

然而, 由于商品数量繁多, 每位顾客每次只能购买其中的极小一部分商品, 这使得存在大量的 $x_{ij}=0$, 那么几乎所有的 $\lambda_{a_1a_2\ldots a_j} \to 0,(1 \le j \le n)$。事实上, 每位顾客不购买大量的商品并不能说明这些商品具有相关性, 所以, 我们决定对此算

法进行优化，以剔除大量不购买数据的影响。

2. Jaccard 距离系数算法

Jaccard 距离系数[3]是针对二值型变量的。它考虑的是几个个体同为 1 的属性个数占至少有一个个体属性为 1 的属性个数的比例。借此思路，我们的密切程度刻画如下。

我们对上述模型进行一定优化，改用相关性系数 S 表示多种商品间的密切程度，当有 a_1，a_2，…，$a_j(1\le j\le n)$ j 种商品时，其表达式如下：

$$S_{a_1a_2\ldots a_j}=\frac{\sum_{i=1}^{m}K\left\{x_{ia_1}=x_{ia_2}=\ldots=x_{ia_j}\right\}}{\sum_{i=1}^{m}K\left\{x_{ia_1}+x_{ia_2}+\ldots+x_{ia_j}>0\right\}},(1\le j\le n)$$

该表达式中，由于 a_1，a_2，…，$a_j(1\le j\le n)$ 只能取 0 或 1，因此 $x_{ia_1}+x_{ia_2}+\ldots+x_{ia_j}>0$ 表示这 j 种商品中至少有一种商品被顾客 i 购买，$\sum_{i=1}^{m}K\left\{x_{ia_1}+x_{ia_2}+\ldots+x_{ia_j}>0\right\}$ 即表示在所有的顾客中，有多少人至少购买了这 j 种商品中的一种。$S_{a_1a_2\ldots a_j}$ 这一参数衡量了在购买这 j 种商品中至少一种商品的顾客里，又有多少人同时购买了这 j 种产品。如果 $S_{a_1a_2\ldots a_j}$ 较大，说明顾客一旦购买这 j 种商品中的某一种，就很可能购买其他的 j-1 种商品，这就说明这 j 种商品的密切程度很高。

（三）问题二

1.寻找 n 商品购买频率算法

假设指定 n 件商品，其编号分别为 $s_1,s_2,s_3,\ldots,s_n$。同时沿用问题一中的 $x_{i,j}$ 来表示顾客 i 对商品 j 的购买情况，即

$$x_{i,j}=\begin{cases}1,\text{顾客i购买了商品j}\\0,\text{顾客i未购买商品j}\end{cases}$$

也沿用是否同买的判断函数：

$$\mathrm{K}(\mathrm{x}_1=\mathrm{x}_2=...=\mathrm{x}_n)=\begin{cases}1,\mathrm{x}_1=\mathrm{x}_2=...=\mathrm{x}_n=1\\0,\mathrm{x}_1\mathrm{x}_2...\mathrm{x}_n=0\end{cases}$$

则同时购买编号为$s_1,s_2,s_3,...,s_n$的 n 件商品个数为

$$T(n,s)=\sum_{i=1}^{4625}\mathrm{K}(\mathrm{x}_{i,s_1}=\mathrm{x}_{i,s_2}=...=\mathrm{x}_{i,s_n})$$

T(n,s)即是 n 件商品同时购买的频率。其中，n 为商品个数，s 为$s_1,s_2,s_3,...,s_n$的集合，即$s=\{s_1,s_2,s_3,...,s_n\}$。

2.计算 2 件商品购买组合频率

对于选取 2 种商品的情况，共有$C_{999}^2=\frac{999\times998}{2\times1}=498501$种组合，可以通过枚举法全部遍历。因此，设立结果集合A_2包含 498501 元素，每个元素是一个包含频数和商品组合的双元素集合。用$A_2(\mathrm{i})_n$表示第 i 个结果中的频数值，$A_2(\mathrm{i})_s$表示第 i 个结果中的商品编号组合（这里 n,s 不是数，而是标示符）。即

$$\begin{cases}A_2(i)_n=T(2,A_2(i)_s)\\A_2(i)_s=\{s_1,s_2\}\end{cases},1\le i\le 498501,1\le s_1<s_2\le 999.$$

这样，对满足条件的s_1,s_2枚举遍历后，集合A_2便记录了所有同时购买两件商品的所有组合以及组合出现的同时购买频数。

3.计算n $(n\ge3)$件商品购买组合频率

理论上, 此时仍可沿用 2 件商品时的算法, 枚举出所有组合并求频数, 但是, 当$3\le n\le996$时，循环次数$C_{999}^n\ge C_{999}^3\approx1.66\times10^8$远大于一般软件在较短时间内可

完成的循环计算次数。所以，枚举遍历法不可行。

事实上，在枚举遍历法中，有很多组合是无需进行计算的。例如，对于某 n 个商品的同时购买频率已经为 0 或者很小，那么包括这 n 个商品的所有 n+1,n+2,…,999 个商品的组合的同时购买频率一定为 0 或者很小。而由于问题目标是追求最大化，这些频率很小的结果是完全不必要记录下来参与统计和研究的。

介于此，我们对枚举遍历法进行优化，采用递推的选择枚举法，即对于 $n\,(n \geq 3)$ 件商品同时购买的情况，以 n-1 件商品同时购买的计算结果分析为基准，只考虑包含频率较多的 n-1 件商品组合的所有 n 件商品组合，以此递推出每个 n 值的结果，有效控制每次的结果数目。

假设 n-1 件商品的计算结果集合 A_{n-1} 中的有效组合（即频率不小于一定值 N 的组合，N 值在模型求解中根据计算结果分析确定）共有 j 组，其集合记为 A_{n-1}'，即

$$A_{n-1}' = \{A_{n-1}(i) \mid A_{n-1}(i)_n \geq N\}$$

其中对于每种方案，都以角标 n 的元素表示该方案的频数，角标 s 的元素表示商品编号组合，即

$$A_{n-1}'(i) = \{A_{n-1}'(i)_n, A_{n-1}'(i)_s\}, 1 \leq \mathrm{i} \leq \mathrm{j}$$

这样，集合 A_{n-1}' 中都是频数较高的 n-1 件商品的方案，而对于 n 件商品的方案，要使频率有可能达到较大，只需考虑包含这些组合中的 n-1 个商品的所有方案，即将 A_{n-1}' 中的每组方案都增加一件第 n 件商品即可。这件商品编号按照比第 n-1 件商品编号大的数枚举（因为 n 件商品的编号是由小到大排列的）。

就算法来说，即是对于 A_{n-1}' 中的每一项 $A_{n-1}'(\mathrm{i})$，都按照

$$\begin{cases} A_n(k)_s = \{A_{n-1}'(i)_s(1), A_{n-1}'(i)_s(2), ..., A_{n-1}'(t)_s(n-1), \mathrm{w}\}, \quad A_{n-1}'(i)_s(n-1) \le \mathrm{w} \le 999 \\ A_n(k)_n = T(n, A_n(k)_s) \end{cases}$$

(其中$A_{n-1}'(i)_s(\mathrm{m})$表示商品组合$A_{n-1}'(i)_s$中第m件商品的编号，k是自然形成的顺序编号）的方式枚举遍历所有的 w，形成 n 个商品组合的所有可能有效的方案的总集 A_n。

接下来，重复进行上述 $A_{n-1} \to A_{n-1}' \to A_n$ 的递推计算过程，就能找到所有有参考意义的 $A_n(\mathrm{n} \ge 3)$，以及每种商品数量n下满足频率不小于N的方案集合 A_n'。

4.最频繁购买商品组合的算法

由于问题目标是使商品数量 n 和出现频率 t 同时最大化，而两者是此消彼长的关系，所以我们定义优度系数来刻画两者的综合大小，并以此为依据来对方案排序。由于商品数量是逐个递增，而频率是可能大幅变化的，所以这个系数应当受两者变化比例影响程度基本相当，而非两者绝对数量产生等效影响。因此，我们定义优度系数为

$$\Omega = \mathrm{n} \bullet \mathrm{t}$$

在计算过程中，当商品数量 n 增加到一个较大的数时，同时购买频率会非常小，失去参考价值，即当 $n \ge M$ 时，$A_n' = \varnothing$。所以，目标方案只存在在 $A_2, A_3, ..., A_M$ 中。我们将这些方案全部抽取出来，得到集合 $B = A_2 + A_3 + ... + A_M$，假设其共有 L 个元素。对于 B 中的每个元素 B(i)，现已有属性 $B(i)_n$ 表示方案同时购买频率，$B(i)_s$ 表示方案所有商品编号集合 ，增加属性 $B(i)_{\mathrm{t}}$ 表示该方案的商品数量，$B(i)_\Omega$ 表示该方案的优度系数。其中 $B(i)_{\mathrm{t}}$ 取自该方案来自的集合 A_{n} 中的 n 值，而每种方案的优度系数按照前段给出的公式，为

$$B(i)_\Omega = B(i)_n \bullet B(i)_t$$

使 i 枚举遍历 B 中的每种方案，即可求得所有方案的优度系数。

接下来，我们通过对每种方案的优度系数大小比较，对所有方案进行降序排

列。排在前面的即是最符合问题目标的结果。

（四）问题三

1.“ 利润指数 ” 刻画预期利润

当组合中的全部商品被同时购买时，获得的是全部商品的总利润和。在安排促销时，我们希望促销方案能够使总利润大的商品组合卖出较多数量。因此，用来刻画整体预期利润的“ 利润指数 ” 应当与组合总利润和组合同时购买频率这两个量相关，受两个量大小同时影响。介于此，我们定义“ 利润指数 ” 为

$$\theta = n \bullet \sum_{i=1}^{n} p_i \quad (\text{n 是商品个数，} \ p_i \text{是第 i 件商品的利润})$$

2.集合 B 每种组合的利润指数

沿用问题二中包括所有较优组合的集合 B。为每一种组合 $B(i)$ 定义一个新的属性 $B(i)_\theta$ 用以记录利润指数，同时保留问题二中已定义的同时购买频率属性 $B(i)_n$、商品编号集合属性 $B(i)_s$、商品数量属性 $B(i)_t$，则 $B(i)_\theta$ 的计算方法为

$$B(i)_\theta = B(i)_n \bullet \sum_{j=1}^{B(i)_t} p(B(i)_s(j))$$

其中 $B(i)_s(j)$ 是 $B(i)_s$ 属性中的第 j 个商品编号，$p(B(i)_s(j))$ 则是该商品编号对应的商品利润。将 i 从 1 开始枚举遍历集合 B 中的所有元素个数，即可求出所有组合的利润指数属性值。

3.制定促销方案

我们将所求出的集合 B 的各个利润指数进行排序，希望其能直观反映出不同商品组合能够为商家带来的利润。接着再根据所排列出的利润指数的大小，将较高的利润指数所对应的商品组合对商家进行促销建议。

例如，将利润指数最大的组合 X 中利润最大的商品 y 放在网站的醒目位置，宣传其功效、优点及产品的相关特性，或者以“新品”、“特价”等字眼以吸引顾客眼球。商品选购页面中，再增加捆绑销售，减价组合销售以及关联产品推荐等方式，使得顾客在购买这一产品时，同时考虑购买与该产品在同一组合中的其他产品。以此达到尽可能多地销售利润指标最高的商品组合的目的。

五、模型求解和检验

（一）问题一

我们将密切程度算法输入 Matlab 软件（附录 1），得到可以输入 n 件编号分别为 $a_1,a_2,...,a_n$ 的商品自动求出该组商品密切程度 $S_{a_1a_2...a_n}$ 的程序。该程序格式为

$$jac(n,[a_1,a_2,...,a_n],liebiao);$$

其中 n 是输入的商品数目，$a_1,a_2,...,a_n$ 是全部商品编号，liebiao 是源数据生成的二值矩阵常量，直接输入这个常量名即可。

这里列举一些组合的计算结果：

表 2　问题一计算结果举例

商品组合（所有编号）	输入命令	所得密切程度值
368,914	jac(2,[368,914],liebiao)	0.1366
413,826,956	jac(3,[413,826,956],liebiao)	0.0754
74,177,298,579,745,881	jac(6,[74,177,298,579,745,881],liebiao)	0.0578
74,177,275,298,375,489,745,809	jac(9,[74,177,275,298,375,489,745,809,956],liebiao)	0.0041

,956		
74,177,275,298,489,579,648,710,745,809,881,934,956	jac(13,[74,177,275,298,489,579,648,710,745,809,881,934,956],liebiao)	0.0036

（二）问题二

将问题 2 的全部算法在 Matlab 软件中进行编程（源码见附录），得到执行结果。任意 2 件商品同时购买组合的频率的主要结果如下。

该项计算得出了 2 件商品同时购买的全部组合的频率，在对频率进行降序排序后，取前 20 种列举如下：

表 3　两件商品同时购买组合频率前 20 种列举

序号	同时购买频率	商品 1	商品 2	序号	同时购买频率	商品 1	商品 2
1	329	368	529	11	251	529	692
2	307	368	829	12	246	368	720
3	289	368	489	13	246	529	829
4	286	368	682	14	237	438	529
5	280	217	368	15	236	217	529
6	266	368	419	16	235	692	829
7	258	368	937	17	234	419	829
8	256	368	510	18	227	205	368
9	255	368	692	19	222	368	722
10	255	368	914	20	221	368	438

在计算过程中，根据计算量的合理控制原则和结果中频数的大小数据，我们设立有效方案的频率界限为 N=60，并依此作为每次 $A_n \to A_n'$ 的筛选依据。任意 n $(n \geq 3)$ 件商品同时购买组合的频率的主要结果如下。

计算发现当 n=14 时，已没有频率超过 60 的方案，即 $M=14, A_{14}'=\varnothing$，因此不需要继续计算 n>14 的情况。

于是，计算得出了 n $(2 \leq n \leq 14)$ 件商品同时购买的全部组合的频率，在对频率进行降序排序后，取每种数量的频率最大方案列举如下：

表 4　每种商品数量下频率最大的组合列举

n 值	最大频率	商品 1	商品 2	商品 3	商品 4	商品 5	商品 6	商品 7	商品 8	商品 9	商品 10	商品 11	商品 12	商品 13	商品 14
2	329	368	529	-	-	-	-	-	-	-	-	-	-	-	-
3	122	368	489	682	-	-	-	-	-	-	-	-	-	-	-
4	105	413	572	797	956	-	-	-	-	-	-	-	-	-	-
5	102	413	424	572	797	956	-	-	-	-	-	-	-	-	-
6	99	413	424	538	797	826	956	-	-	-	-	-	-	-	-
7	97	41	42	53	57	79	82	95	-	-	-	-	-	-	-

		3	4	8	2	7	6	6							
8	76	74	177	275	489	579	710	745	881	-	-	-	-	-	-
9	75	74	177	275	489	579	710	745	881	934	-	-	-	-	-
10	74	74	177	275	298	489	579	710	745	881	934	-	-	-	-
11	73	74	177	275	298	375	489	579	710	745	881	934	-	-	-
12	71	74	177	275	298	375	489	579	710	745	809	881	934	-	-
13	69	74	177	275	298	375	489	579	648	710	745	809	881	934	-
14	12	74	177	275	298	375	489	579	648	710	745	809	881	914	934

查找商品数量尽可能大的最频繁同时购买的商品组合的计算结果如下。按照优度算法计算出集合 B，并按优度系数降序进行排序，得到前 20 位组合方案如下：

表 5　优度系数最高的 20 种组合列举

优度系数	商品数量	频率	商品1	商品2	商品3	商品4	商品5	商品6	商品7	商品8	商品9	商品10	商品11	商品12	商品13

897	13	69	74	177	275	298	375	489	579	648	710	745	809	881	934
852	12	71	74	177	275	298	375	489	579	710	745	809	881	934	-
852	12	71	74	177	275	298	375	489	579	648	710	745	881	934	-
840	12	70	74	177	275	298	375	489	579	648	710	745	809	881	-
840	12	70	74	177	275	298	375	489	579	648	710	745	809	934	-
840	12	70	74	177	275	298	375	489	579	648	710	809	881	934	-
840	12	70	74	177	275	298	375	489	579	648	745	809	881	934	-
840	12	70	74	177	275	298	489	579	648	710	745	809	881	934	-
840	12	70	74	177	275	298	375	489	648	710	745	809	881	934	-
828	12	69	74	177	275	298	375	579	648	710	745	809	881	934	-
828	12	69	74	275	298	375	489	579	648	710	745	809	881	934	-
828	12	69	177	275	298	375	489	579	648	710	745	809	881	934	-

828	12	69	74	177	275	375	489	579	648	710	745	809	881	934	-
828	12	69	74	177	298	375	489	579	648	710	745	809	881	934	-
803	11	73	74	177	275	298	375	489	579	710	745	881	934	-	-
792	11	72	74	177	275	298	375	489	579	710	745	809	881	-	-
792	11	72	74	177	275	298	375	489	579	710	745	809	934	-	-
792	11	72	74	177	275	298	489	579	710	745	809	881	934	-	-
792	11	72	74	177	275	298	375	489	579	710	809	881	934	-	-
792	11	72	74	177	275	298	375	489	579	745	809	881	934	-	-

这一结果也与我们对结果数据的分析相符。当$n \leq 13$时，随着 n 的增加，最大频率的减少是较为缓慢且超过 60 的，这些都是有效组合，所以当 n=13 时达到目标；而 n=14 时最大频率仅为 12，相对 n=13 产生了大幅减少，结果不具有参考性。因此恰在 n=13 时达到目标最大化。

可见，最频繁被同时购买的且商品数量最多的方案是第一排所示的，商品数量为 13，频率达到 69，商品编号为 74，177，275，298，375，489，579，648，710，745，809，881，934 的组合。

（三）问题三

利用模型算法对集合 B 中的每种组合计算利润指数值，按此值降序排序，列举前 20 个结果如下：

表 5 利润指数最高的 20 种组合列举

利润指数	商品数量	同时购买频率	商品1	商品2	商品3	商品4	商品5	商品6	商品7	商品8	商品9	商品10	商品11	商品12
189487.6	2	329	368	529	-	-	-	-	-	-	-	-	-	-
156389.4	2	266	368	419	-	-	-	-	-	-	-	-	-	-
147808.4	11	72	74	177	275	298	375	489	579	648	710	881	934	-
147433.4	10	72	74	177	275	298	375	579	648	710	881	934	-	-
147249.5	2	307	368	829	-	-	-	-	-	-	-	-	-	-
146125.4	12	71	74	177	275	298	375	489	579	648	710	745	881	934
146055.3	10	72	177	275	298	375	489	579	648	710	881	934	-	-

146055.3	10	72	74	177	275	375	489	579	648	710	881	934	-	-
145755.5	11	71	74	177	275	298	375	579	648	710	745	881	934	-
145680.3	9	72	177	275	298	375	579	648	710	881	934	-	-	-
145680.3	9	72	74	177	275	375	579	648	710	881	934	-	-	-
144444.9	10	73	74	177	275	298	489	579	648	710	881	934	-	-
144396.6	11	71	177	275	298	375	489	579	648	710	745	881	934	-
144396.6	11	71	74	177	275	375	489	579	648	710	745	881	934	-
144302.2	9	72	177	275	375	489	579	648	710	881	934	-	-	-
144067.3	12	70	74	177	275	298	375	489	579	648	710	809	881	934
144064.6	9	73	74	177	275	298	579	648	710	881	934	-	-	-
144026.8	10	71	177	275	298	375	579	648	710	745	881	934	-	-
144026.8	10	71	74	177	275	375	579	648	710	745	881	934	-	-

143927.1	8	72	177	275	375	579	648	710	881	934	-	-	-	-

六、模型结果分析

综上分析，利润指数最高的组合并非都是商品数量较多的，说明对某些商品数量很少的组合，虽然每单位组合的商品的总利润可能没有商品数量多的组合多，但由于其同时购买频率较大，也能产生非常可观的预期利润。因此，我们要结合商品数量较多的和较少的组合综合考虑我们的促销方案，由此得出下面的促销方案。

以利润指数最高的商品组合：368 号、529 号商品为例，该组商品的利润指数为 189487.6，在所有商品组合中排首位，在为商家设计促销方案时，首先选取这一组商品组合中利润最高的商品即 368 号商品（利润为 290.91 元）放置在购物网站的首页等醒目位置，并对其优点、功能进行宣传，或标上“ 新品 ”、“ 促销 ”等醒目字眼吸引消费者眼球，引导他们进行消费。在 368 号商品的购物页面中，再将其与 529 号商品捆绑销售，或组合减价销售，或作为推荐商品供消费者选购，通过这种方式实现商品组合能够尽可能多地被更多消费者同时选购，以实现最大利益。

而对于商品数量较多的组合（如第二组），则可将其中每件商品的购物页面中加上“ 推荐类似商品 ”栏目，其中展示同组合其他商品的信息，吸引消费者组合购买。由于同组商品相关程度很高，消费者在没看到同组商品时可能想不到购买，而有此栏目后很可能会选择同时购买。

七、结论和建议

通过本文综合分析，我们得出以下三个结论：

(1) 在问题一中灵活运用聚类中的距离概念引入参数λ，并建立公式反映不同商品间的密切程度，通过计算同时购买商品组合中所有商品的顾客与购买商品组合中任意一种商品的顾客的比值，对公式进一步优化，得出求解密切程度的最优算法模型,使结果更加准确。但该模型只能定量检测制定商品间的密切度，无法自动对全体商品间的密切度进行全面自动的计算。

(2) 在问题二中，我们结合追求频率和商品组数都尽可能高的问题要求，剔除一些对结果影响不大的数据，从两组商品的若干组合中挑选购买频率大于 60 的组合进行进一步运算，提高了运算效率。并在找出购买频率与商品种数均较大的商品组合时，定义了优度系数Ω来刻画商品数量和同时购买的频率的综合大小，从而能够科学地表示各组合与问题二目标的契合程度。但是由于在问题二中剔除小于 60 的数据，可能会导致得到的结果并不完备。同时，枚举法的运算次数太多，导致运算效率不高。

(3) 在问题三中，我们定义了利润指标这一概念使不同种类数的商品组能够在统一标准下进行比较，从而能够科学对对销售方式进行建议。同时，我们只列举出利润指数最高的前二十组商品组合也使得对促销方案的建议更有针对性。但在问题三中，为了强调多安排销售哪些组别的商品组合能够获得最大利润这一目标，我们忽略了成本、消费者心理、商品的需求弹性等因素，导致我们无法提出更加细致具体的促销方式。

八、参考文献

[1] Imad A. Moosaa, Bernard Bollen. LINGO-1regulatesoligodendrocyte differentiation by inhibiting ErbB2 translocation and activation in lipid rafts [J]. International Review of Financial Analysis, 2002.

[2] 桑杨阳, 朱万红, 但兵兵. 非线性规划建模与 LINGO 软件的编程应用[J].电脑知识与技术,2012(04).

[3] 卓金武, MATLAB 在数学建模中的应用, 北京航空航天大学出版社, 2011.4.

[4] 姜启源, 谢金星, 数学模型（第三版）, 高等教育出版社, 2003.8.

[5] 姜启源, 谢金星, 数学建模案例选集, 高等教育出版社, 2006.7.

[6] 夏骄雄, 数据资源的聚类预处理, 上海科学普及出版社, 2011.11.