第21卷 第7期 工 程 数 学 学 报 Vol. 21 No. 7
2004年12月 CHINESE JOURNAL OF ENGINEERING MATHEMATICS Dec. 2004

文章编号:1005-3085(2004)07-0064-07

数据挖掘技术在小型商业网点布局问题的应用

艾乐强, 孙 健, 刘艳平

指导教师: 徐 丹

(东北农业大学理学院, 哈尔滨 150030)

编者按: 本文是2004年全国大学生数学建模比赛中A题获奖论文中的一篇有特色的论文。该文使用数据挖掘的方法去寻找数据中反映的有用规则, 是所有参赛论文中较少使用的好方法。由于数据挖掘方法是近年兴起的数据分析的好方法, 希望通过这篇论文的发表, 推动数据挖掘方法使用的普及。

摘 要: 本文采用数据挖掘技术中Apriori算法, 进行关联规则挖掘, 逐层进行频度分析, 编制MATLAB程序得出了各种因素之间的相关性和各因素影响规律性的权重系数, 并结合概率分析方法对已给的数据进行精确处理, 得出了全面而简明的观众出行、用餐和购物等方面的规律。对问题2采用明确的等式简化原则和等概率原则将最短路径问题转化为简单合理的概率模型, 结合已经得到的规律对数据进行信息提取、分析, 得出了与实际情况吻合的人流量分布和购物欲望分布结果。根据得到的两个分布构造出商业区获利的多目标函数, 并对其进行约束优化得出了符合实际情况的MS网点分布方案。

关键词: MS 网点设计; 人流量分布; 数据挖掘; 多目标规划

分类号: AMS(2000) 90C29 **中图分类号**: 221 **文献标识码**: A

1 问题重述 (略)

2 模型假设

1) 假设每次出行的出发线路和返回线路不变;

2) 假设每个商业区的面积相等;

3) 假设观众一日内平均出行两次。

3 符号说明

j: A、B、C 三个比赛主场馆区 (分别取值 1 、 2 、 3);

$P(j)$: 进入比赛主场馆j 区的人座位在各个看台的概率;

$Q(j)$: 比赛主场馆j 区的最大观众容量;

P_j: 从相应入场口进出 (或途经) j 区的观众经过j 区内各个商区的概率;

P_A: 进 (出) 比赛主场馆j 区时, 以该区作为目标场馆区 (或作为初始场馆区) 的观众占所有观众的概率;

P_B: 进 (出) 比赛主场馆j 区时, 以该区作为抵达目的地所途经的场馆的观众占所有观众的概率;

P_j': 不同乘车方式的观众从相应入场口进出(或途经)j 区时经过j区内各个商区的概率;

P_j'': 不同餐饮方式的观众从相应入场口进出(或途经)j区时经过j 区内各个商区的概率;

α_i: MS点的规模;

β_j: 商区的购买力;

其他符号在题中出处将给予说明。

4 问题分析

问题 1： 根据已有的问卷调查数据，运用数据挖掘技术找出观众在出行、用餐和购物等方面所反映的规律，找出数据之间的相关性，判断某个因素是否具有反映规律的特性。

问题 2： 测算20个商业区的人流量在于观众一天两次出行中有多少人经过该商业区。解决题中谈到观众出行选择路径的问题，采用等价代换和估算的方法。

问题 3： 设计MS网点达到三个基本的标准，建立与人流量和购物欲望密切相关的购买力函数。在此基础上进行了函数优化得出了MS网点的设计方案。

5 模型建立与求解

问题 1：

由问题 1 的分析可知需要对数据处理，采用数据挖掘技术中Apriori算法，指定需求的关联规则，初始候选项集如表 1：

$$I = \{i_1, i_2, i_3 \cdots i_{21}\}$$

表1 初始候选集表

I1	I2	I3	I4	I5	I6	I7	I8	I9
年龄 1	年龄 2	年龄 3	年龄 4	女	男	公交南北	公交东西	出租
I10	I11	I12	I13	I14	I15	I16	I17	I18
私车	地铁东	地铁西	中餐	西餐	餐饮	消费等级 1	消费等级 2	消费等级 3
I19	I20	I21						
消费等级 4	消费等级 5	消费等级 6						

而后确定最小支持度阈值(min_sup)，和最小置信度(min_conf)；然后从初始项集中选定任意 2 -项集，$I = \{I_{k1}, I_{k2}\}$; $k1, k2 \in (1, 2 \cdots m)$且$k1 \neq k2$ 求出这些 2 -项集的出现频率，即支持计数；最终得到满足强规则的频繁项集，对这些有特性的频繁项集中的元素可以根据其对应的柱状图进行分析，从而得到相对应的人们就餐，购物，和出行方式的规律。

由Apriori算法可知各关联因素的支持度为:Support$(A) = P(A)$;

由于考虑到公交，地铁的方向不能影响其他因素之间的相关性，在选定 1 -项集时将地铁东、西；公交南北、东西的支持度分别求和，作为地铁，公交的总支持度如表 2：

表2 候选 1 -项集各关联因素支持度表

相关因素	男	女	公交	地铁	出租	私车	中餐
Support(%)	52.37	47.62	33.96	38.05	18.95	9.04	22.48
相关因素	西餐	餐饮	年龄 1	年龄 2	年龄 3	年龄 4	消费等级 1
Support(%)	52.50	25.64	11.11	57.93	20.19	10.74	17.45
相关因素	消费等级 2	消费等级 3	消费等级 4	消费等级 5	消费等级 6		
Support(%)	23.19	40.94	2.52	1.23	0.70		

我们设定min_support = 30%,只要support > min_support的相关因素就可以进入到 1 -项集中，得到 1 -项集如表 3

表3 1 -项集表

男	女	公交	地铁	年龄 2	消费等级 3	西餐
A1	A2	A3	A4	A5	A6	A7

由表3得到候选2-项集，根据算法中每两个相关因素的支持度$\text{support}(A,B)=P(AB)$;

$$\text{confidence}(A\Longrightarrow B)=P(A\mid B)=\frac{\text{support}(A\cup B)}{\text{support}(A)}$$

对各相关因素的支持度和置信度进行计算，得到支持度，置信度值如表4：

表4 候选2-项集表

二项集	{A1,A2}	{A1,A3}	{A1,A4}	{A1,A5}	{A1,A6}	{A1,A7}	{A2,A3}
Confidence(%)	-	42.55	39.21	50.96	43.48	53.78	25.86
Support(%)	-	22.2	20.46	29.54	22.69	28.06	12.37
二项集	{A2,A4}	{A2,A5}	{A2,A6}	{A2,A7}	{A3,A4}	{A3,A5}	{A3,A6}
Confidence(%)	35.01	49.04	47.55	51.08	-	31.99	32.64
Support(%)	16.74	28.43	22.74	24.43	-	18.54	14.83
二项集	{A3,A7}	{A4,A5}	{A4,A6}	{A4,A7}	{A5,A6}	{A5,A7}	{A6,A7}
Confidence(%)	33.42	39.23	39.12	39.13	57.86	62.00	49.27
Support(%)	17.54	22.74	17.77	20.54	33.54	35.94	25.86

考虑到消费等级4、5、6的指标比较接近，1、2等级的也比较接近，因此将消费等级重新定义为初等消费（1,2），中等消费（3）和高等消费（4,5,6）；置信度为各等级置信度之和，选定性别作为计算置信度标准，$\min\_\text{support}=30\%$，$\min\_\text{confidence}=20\%$，当满足$\text{support}(A,B)>\min_\text{support}$并且$\text{confidenc}>\min_\text{confident}$就可认为此2-项集满足强关联规则，由此得到强关联规则联表，如表5，其他置信度的值由此原理可以得到。

表5 强关联规则置信度

关联因素	公交(K1)	地铁(K2)	西餐(K3)	初等消费(K4)	中等消费(K5)	高等消费(K6)
男(%)	41.54	40.56	52.24	53.38	46.72	4.7
女(%)	25.7	35.3	52.92	34.58	45.99	19

由图1可以看到：

1. 女士与乘公交和地铁的关联性相对要低于男士；
2. 男、女都与吃西餐有强关联性；
3. 女士与高等消费的关联性要明显高于男士；
4. 男士与初等消费的关联性高于女士，主要取决于男士与女士的消费观念不同；
5. 大部分人与中等消费存在强关联性；

对于其它的2-项集：

1. 年龄与乘车方式、用餐类型无关；
2. 年龄1、4时消费等级主要为1、2,年龄2、3时消费等级主要为2、3；
3. 乘车方式与用餐类型无关；
4. 消费等级高的喜欢吃餐饮；
5. 乘私车、出租的消费等级高。

问题2

由前面分析得出，题中没有路径长度的量度，而且比例尺也不确定，因此不可能进行直接的数值上路径的比较。本模型解决该问题采取"相对最短路径"，结合"等概率原则"对路径进行分析得出了人们出行，就餐的路径方案。

相对最短路径：在没有数值比较任两条路径长短的条件下，仍然要给出路径之间的对比。总体上采用目测法，给出两条路径的长度相对对比。

等概率原则：假设要比较的两条路径不能用相对最短路径原则比较，那么本模型认为人们选择每条路径的概率相等，各为50%；如有更多的路径（m 条）同时可以到达目的地，选择每条路径的概率相等为$\frac{1}{m}$。

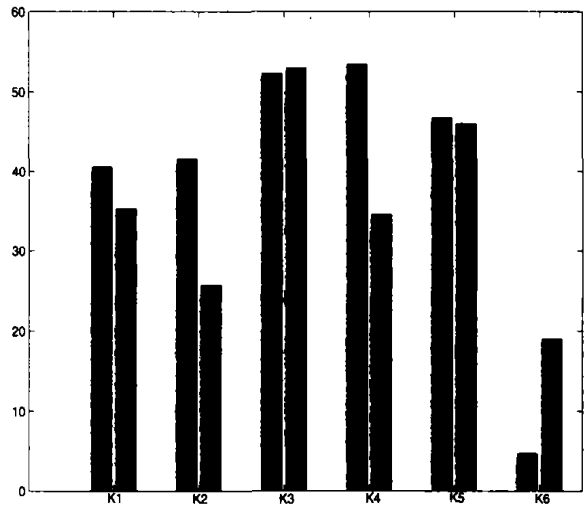

图 1: 强关联规则置信度值（左男，右女）

观众进出j区均有两种方式,一种为从南侧口进出，另一种为从北侧口进出，相应有两种决策值。

$$X_{ij}=\begin{cases}1\cdots\text{从南侧入口进入}j\text{区}\\0\cdots\text{从北侧入口进入}j\text{区}\end{cases}$$

经分析得，进j 区各看台观众的$P(j)$相等

$$P(j)=\frac{1}{Q(j)}$$

则$P(1)=\frac{1}{10},P(2)=\frac{1}{6},P(3)=\frac{1}{4}$。

将观众分两类考虑:A类为以到达该场馆为目的地的观众;B类为到达另一目标场馆途经此场馆的观众。可用表达式$P(j)=P_A+P_B$反映出观众到达各个商区的内在关联性。

经过以上分析，考察每个主场馆各个商区的人流分布时，将到达每个主场馆的观众分为A、B两类，之后，结合上面得出的规律，运用计算机编程将观众到达各个商区的概率进行统计，即对影响比赛主场馆j区人流量分布的观众进行分析，求出选择不同入场决策（j=1或0）时，从相应入场口进入该主场馆各个商区的概率，且由于对B类观众，j=1与j=0时对结果无影响，故可将两种情况合并在一起考虑，具体结果可由向量表示（见附表）。

人流量的分析

考虑对各个商区的人流量大小的影响因素时，分两步进行。

第一步：第一次出行，即进出比赛主场馆区。

由假设可知观众进出比赛主场馆区采用相同的路径，所以不妨只考虑观众进入场馆的路径，以各个主场馆区为考虑人流量的核心。只需分别考虑到达各个主场馆区的观众采用什么乘车方式。现根据上面对"采用最短路径"这一因素的分析得出的"相对最短路径" 及"等概率原则"，可以给出到达各个比赛主场馆区的观众采用不同乘车方式的路径。具体路径仅举一例见表 6，其余同理。在此次出行中，考虑影响各个比赛主场馆区人流量因素时得出函数$f_1=\sum_j P_j'(P_A+P_B)$,运用此关系式可得出观众采用不同乘车方式各个比赛主场馆区人流量的量化结果。

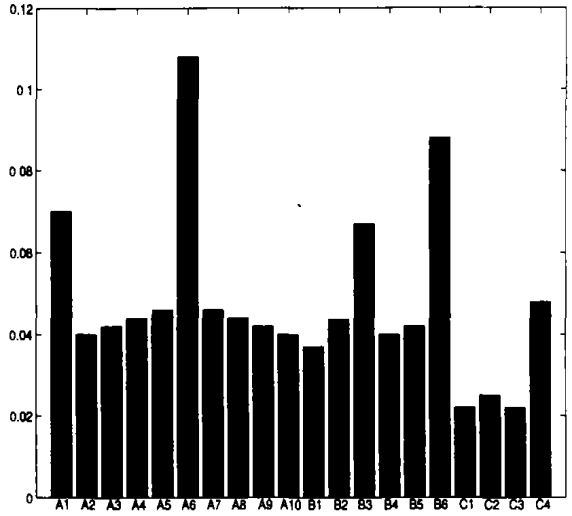

图2: 人流量分布

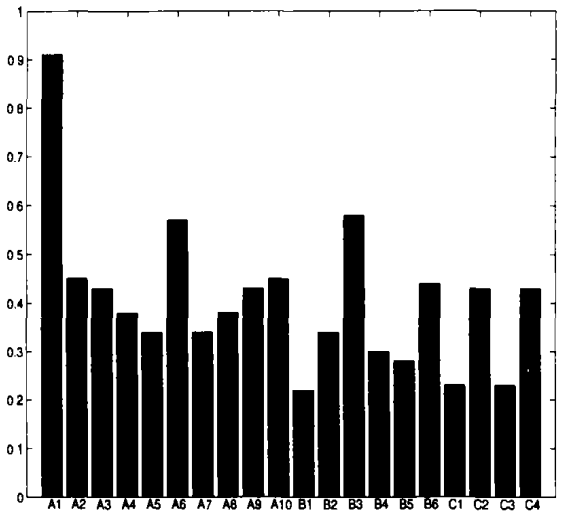

图3: 购物欲望

表6 到达比赛主场馆A的观众采用不同乘车方式的路径

	未途径其它场馆而直接到达目标场馆A							途径其它场馆而直接到达目标场馆A	
乘车方式	I8	I8	I10	I12	I9	I7	I11	I7	I11
入场决策	1	1	1	0	1	0	0	0	0

第二步：第二次出行，即餐饮。

假设可知观众为就餐进出比赛主场馆区采用相同的路径，所以不妨只考虑观众出场馆的路径，此时仍可以各个主场馆区为考虑人流量的核心。只需分别考虑到达各个主场馆区的观众采用何种餐饮方式时所走的路径。仍利用上文中给出的"相对最短路径"及"等概率原则"确定各个比赛主场馆区的观众采用不同用餐方式所走的路径，具体路径仅举一例如表 7，其余同理。在此次出行中，考虑影响各个比赛主场馆区人流量因素时得出函数$f_2 = \sum_j P_j''(P_A + P_B)$,运用此关系式同样可得出观众采用不同餐饮方式各个比赛主场馆区人流量的量化结果。

表7 影响场馆A区人流量的观众就餐不同的路径

	以场馆A为就餐出发点			到目标餐馆就餐时途经场馆A
就餐方式	I13	I14	I15	不存在
出厂决策	1	0	0	

经过以上两步分析可知，需综合考虑观众乘车方式、餐饮方式对各个商区人流量的影响, 给出相应的权值ω_1和ω_2，建立了如下函数关系刻画它们之间的关联：

$$f(t) = \omega_1 f_1 + \omega_2 f_2$$

f_1为观众乘车方式不同对各个商区人流量的影响函数，f_2观众餐饮方式不同对各个商区人流量的影响函数。在对量化结果所包含的信息进行挖掘之后，且得出ω_1与ω_2之比近似为 1 。最后得出了20个商区人流量的规律如图 2 所示：

购物欲望的分析

由于商店选址不仅取决于商圈内的人流量，还取决于购物欲望，所以需要对购物欲望进行分析。根据求解问题 1 所统计的数据结果，反复拟和得到购物欲望指数变化规律，表达形式如

下：

$$g(t) = e^{\sum\sum \theta_i t_j}$$

θ_i为第i年龄对消费等级影响的权重($i = 1 \cdots 4$)，t_j 为消费等级为j的比率($j = 1 \cdots 6$)。据此得出了购物欲望分布图 3，其可以明显、准确地反映诸多因素对购物欲望影响。

结合人流量、购物欲望对商店选址的影响，引入"购买力"的概念。购买力：某一商区观众的实际消费额。它与购物欲望成正比，与人流量成正比。现给出购买力与人流量、购买欲望之间的关系函数：

$$\delta(t) = f(t) \times g(t)$$

即：

$$\delta(t) = (\omega_1 f_1 + \omega_2 f_2) \cdot e^{\sum\sum \theta_i t_j}$$

据此函数关系可准确刻画影响商店选址的因素，具体规律见购买力分布图 4 。以上分析为后文合理给出MS网点的设计方案提供了有利的支持。

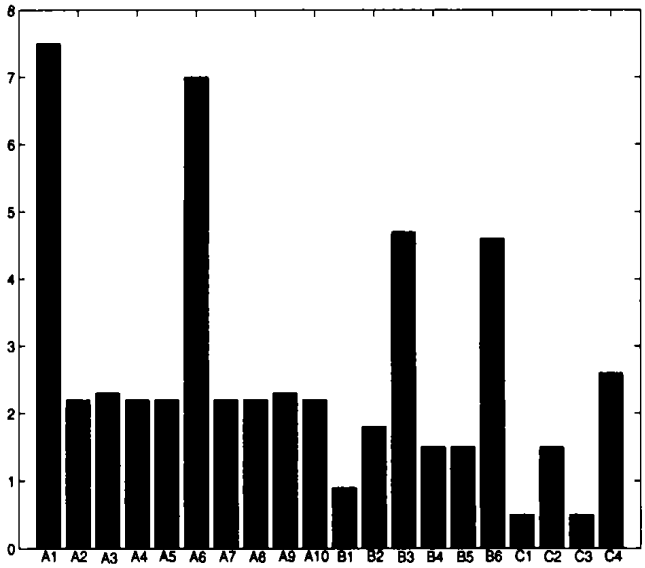

图4: 购买力分布

问题 3

设某一商区内两个MS点的距离为d，S为商圈覆盖范围（即商店覆盖面积）由下式可知它与α_i^2成正比（即规模大的MS，商圈覆盖范围大），与β_j^2成反比，为保障赢利购买力低需要更大的商圈范围来削弱MS之间的竞争。n_1 为大型MS网点的个数，n_2为小型MS网点的个数。$L(x)$为利润函数。

$$d_{ij} = \frac{\alpha_i}{\beta_j}$$

$$s_{ij} = (\frac{\alpha_i}{\beta_j})^2$$

购买力、规模越大则利润越大，而MS个数越少越赢利，通过以上分析，建立如下规划模型：

$$L(x) = \frac{\alpha_1 \beta_j}{n_1} + \frac{\alpha_2 \beta_j}{n_2}, \qquad max\,(n_1 + n_2 + kL(x))$$

即：

$$\max\left(n_1 + n_2 + k\left(\frac{\alpha_1\beta_j}{n_1} + \frac{\alpha_2\beta_j}{n_2}\right)\right)$$

$$\text{s.t.}\begin{cases} n_1(\frac{\alpha_1}{\beta_j})^2 + n_2(\frac{\alpha_2}{\beta_j})^2 \le \beta_j \\ n_1 \ge 0 \\ n_2 \ge 0 \end{cases}$$

设大小规模比为3:1。我们运用Lingo软件求解模型 3 得出结果见表 8：

问题 4 （**略**）

6 模型的评价推广（略）

表8 MS网点分布

商区	大MS网点的个数	小MS网点的个数
A1	5	13
A2	1	5
A3	1	5
A4	1	5
A5	1	5
A6	5	13
A7	1	5
A8	1	5
A9	1	5
A10	1	5
B1	0	4
B2	1	5
B3	3	7
B4	0	4
B5	0	4
B6	3	7
C1	0	4
C2	0	4
C3	0	4
C4	2	6

参考文献:

[1] Jiawei Han Micheline Kamber著. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001
[2] 朱明等著. 数据挖掘[M]. 合肥：中国科学技术出版社，2002
[3] 景体华著. 北京蓝皮书2004年：中国首都发展报告[M]. 北京：社会科学文献出版社，2004
[4] 袁亚湘著. 最优化理论与方法[M]. 北京：科学出版社，2003
[5] 洪文著. LINGO 4.0 for Windows最优化软件及其应用[M]. 北京：北京大学出版社，2001

The Application of Data Mining in the Distribution of Miniature Commerce Net Sites

AI Le-qiang, SUN Jian, LIU Yan-ping
Advisor: XU Dan
(College of Science, Northeast Agricultural University, Harbin 150030)

Abstract: In this model, we use the Apriori algorithm in data mining to write the MATLAB code. By the analysis of inter-relationships as well as the frequency, we get the relation between factors and the weight coefficient of them. In virtue of probability on the available data, we can obtain the all-round rules for people's activities on dining and shopping, etc. As to problem 2, we convert the minimization problem to the simple probability model by the rules of equation simplification and equi-probability. From the information got by data analyzing, we could get the flux distribution of consumers as well as their appetite. Finally by the aid of the two distributions, we can construct the multi-functions for the profit of the commercial drive and by constrained optimization to obtain the MS net distribution corresponding to reality.

Keywords: designing of MS net sites; data mining; multiple programming