

零售业中商品选择问题的遗传算法研究

程 岩

复旦大学管理学院 信息管理与信息系统系, 上海 200433

[摘要] 选择经营哪些商品是零售业面对的一个重大决策问题, 这个问题需要考虑到商品之间的交叉销售效果。最近的研究成果表明这一问题是 NP 难问题。采用定量关联规则技术来计算交叉销售效果, 并提出一个面向商品选择问题的遗传算法。由于交叉销售效果是一个不精确的信息, 根据遗传算法的要求, 进一步研究了面向不精确信息的适应度函数, 从而使遗传算法适用于基于交叉销售效果分析的商品选择问题。

[关键词] 遗传算法; 数据挖掘; 关联规则; 交叉销售

[中图分类号] TP311 **[文献标识码]** A **[文章编号]** 1000-7857(2007)07-0065-06

Genetic Algorithm for Item Selections in Retail Business

CHENG Yan

Department of Information Management and Information System of Management School, Fudan University, Shanghai 200433, China

Abstract: A fundamental problem in retail business is selecting items with consideration of "cross-selling effect". Recent studies show that the problem is NP-hard. In this paper, the genetic algorithm is applied to this problem, and a quantitative analysis method for "cross-selling effect" is proposed, using quantitative association rules, and a newly developed data mining technique to identify quantitative affinities in large transaction databases. Based on the features of the genetic algorithm, a method of defining imprecise fitness function is proposed, which can be used for solving item selection problems with imprecision cross-selling effect.

Key Words: genetic algorithm; data mining; quantitative association rule; cross-selling

CLC Number: TP311

Document Code: A

Article ID: 1000-7857(2007)07-0065-06

收稿日期: 2006-11-13

作者简介: 程 岩, 男, 复旦大学管理学院信息管理与信息系统系, 讲师, 研究方向为电子商务、数据挖掘;

E-mail: yancheng@fudan.edu.cn

- and coal-bed methane simultaneous extraction" technique on the basis of strata movement [J]. J China Coal Soc, 2004, 29(2): 129-132.
- [9] 刘泽功, 袁 亮, 戴广龙, 等. 开采煤层顶板环形裂隙圈内走向长钻孔法抽放瓦斯研究[J]. 中国工程科学, 2004, 6(5): 32-38.
LIU Zegong, YUAN Liang, DAI Guanglong, et al. Study on coal seam roof gas drainage from the strike of annular fracture areas by the long drill method [J]. Eng Sci, 2004, 6(5): 32-38.
- [10] 师本强, 侯忠杰. 陕北榆神府矿区保水采煤方法研究[J]. 煤炭工程, 2006(1): 63-65.
SHI Benqiang, HOU Zhongjie. Research on coal mining method with water conservation in Yushen mining area [J]. Coal Eng, 2006(1): 63-65.
- [11] 张东升, 马立强. 特厚坚硬岩层组下保水采煤技术[J]. 采矿与安全工程学报, 2006, 23(1): 62-65.
ZHANG Dongsheng, MA Liqiang. Coal mining technique with water conservation under hard-and-thick strata[J]. J Min Saf Eng, 2006, 23(1): 62-65.
- [12] 张吉雄, 缪协兴. 煤矿矸石井下处理的研究[J]. 中国矿业大学学报, 2006, 35(2): 197-200.
ZHANG Jixiong, MIAO Xiexing. Underground disposal of waste in coal mine[J]. J China Univ Min Tech, 2006, 35(2): 197-200.
- [13] 梁 杰. 煤炭地下气化技术[J]. 中国科学人, 2006(4): 82-83.
LIANG Jie. The technology of underground coal gasification[J]. Scientific Chinese, 2006(4): 82-83.

(责任编辑 李慧政)

0 引言

随着科学技术的发展,商品的种类越来越丰富,市场需求的变化越来越快。面对这种严峻的市场形势,选择经营哪些商品就成为零售商的一项具有战略意义的决策行为。商品选择是一个比较困难的决策问题,这是因为,一种商品的收益不仅来自于该商品本身的直接收益,还来自于该商品对其他商品销售收益的影响,即交叉销售的效果^[1-3]。最近的研究成果表明这类商品选择问题是 NP 难问题^[4]。

针对零售业中商品选择问题的研究并不多,所提出的方法还存在着许多缺陷,目前主要有以下代表性研究成果。

最早从交叉销售效果角度来处理这一问题的研究成果是 Brijs 等^[5-6]针对自动售货机的商品选择问题提出的 PROFSET 模型,PROFSET 模型利用频繁项目集的支持度指标来识别每个交易中对一组商品的购买意向,从而分析交叉销售效果,并通过求解 0-1 规划模型来解决商品选择问题。但正如文献 [7] 所指出的那样,PROFSET 模型仅仅适用于商品种类比较小的情况。

文献[7]提出用“hub-authority”算法解决商品选择问题,“hub-authority”算法是搜索引擎中解决 Web 页排序问题的典型算法^[7]。利用这种算法的好处是可以快速处理品种数量非常大的商品收益排序问题。但是,2 个网页之间的链接调用是一对一的数量关系,与此不同,商品之间的交叉销售效果往往表现为诸如“2 件 A 商品的销售导致 5 件 B 商品的销售,也导致了 3 件 C 商品的销售”这样数量上的区别,因此,“hub-authority”算法一般只适用于单件销售的商品。

文献[4,8]提出利用定性关联规则技术来计算交叉销售效果,这样不仅从支持度的角度衡量了客户同时购买某种商品组合的意愿,还从信任度的角度加以衡量,提高了模型的精度,但是文献仍然没有从量的角度分析商品间的关联关系,只是分析了客户同时购买一组商品的意愿,因此,他们的研究结果仍然限定于商品是单件销售的零售业务。

本文认为,在客户一次可能购买多件同一种商品的零售环境中,商品间的交叉销售效果存在着数量上的关联关系。解决这类零售环境下商品选择问题的难点在于:各种商品间的交叉销售效果不可能是一个精确的数值,这种量的关系往往是含糊的不精确信息。针对上述问题,本文结合数据挖掘技术,提出了一个求解商品选择的遗传算法,其主要贡献是:① 基于定量关联规则理论,提出了一个计算交叉销售效果的方法;② 设计了基于不精确信息的求解商品选择的遗传算法。

试验结果表明,本文提出的方法具有较高的实用

性和计算精度。

1 问题的描述

设零售商经营的商品集合为 $I=\{I_1, I_2, \dots, I_n\}$, 我们需要淘汰商品集合 I 中的 m 种商品,这 m 种商品构成集合 D ,即 $D \subset I$ 。令 $V=I-D$ 表示保留下来的商品集合。令 $R(D)$ 表示淘汰 D 中的商品所损失的直接收益, $R(D, V)$ 表示淘汰集合 D 中的商品后,商品集合 V 中损失的间接收益,那么损失的总收益 R 为

$$R=R(D)+R(D, V) \quad (1)$$

计算直接收益损失是非常简单的,如式(2)所示

$$R(D)=\sum_{x \in D} S_x \cdot \text{prof}(x) \quad (2)$$

式中, $\text{prof}(x)$ 表示每销售一个数量单位的 x 带来的直接收益,商品 x 在单位时间里的平均销售量为 S_x 。

定义 1 交叉销售效果系数 $\beta(x, y)$ 是指由于交叉销售的效果,每销售一件商品 x ,将导致销售了 β 件商品 y 。

根据定义 1,商品 x 通过商品 y 的交叉销售效果所获得的间接收益按如下公式计算

$$r(x, y)=S_x \cdot \beta(x, y) \cdot \text{prof}(y) \quad (3)$$

因此,由于交叉销售效果,淘汰 D 中的商品所损失的间接收益为

$$R(D, V)=\sum_{x \in D} \left(\sum_{y \in V} r(x, y) \right) \quad (4)$$

则,式(1)又可表示为

$$\begin{aligned} R &= R(D) + R(D, V) \\ &= \sum_{x \in D} S_x \cdot \text{prof}(x) + \sum_{x \in D} \left(\sum_{y \in V} S_x \cdot \beta(x, y) \cdot \text{prof}(y) \right) \quad (5) \end{aligned}$$

从式(5)可以看出,交叉销售效果系数 $\beta(x, y)$ 是确定淘汰哪些商品的一个关键参数。但在现实生活中,交叉销售效果系数 $\beta(x, y)$ 不可能是一个精确的数值,我们只能以一定的置信度水平估计它的变化范围。

定义 2 不精确交叉销售效果系数 $[\beta_s(x, y), \beta_L(x, y)](c)$ 是指由于存在交叉销售效果,每销售一个数量单位的商品 x 导致销售了 S_x 数量单位的商品 y ,估计 S_x 在 $[\beta_s(x, y), \beta_L(x, y)]$ 之间的概率为 c 。规定 $(\beta_L(x, y) > \beta_s(x, y))$ 且 $0 < c < 1$ 。

根据定义 2,淘汰部分商品后损失的总收益是一个区间值 $[R_s, R_L]$,其计算公式分别为

$$R_s = \sum_{x \in D} S_x \cdot \text{prof}(x) + \sum_{x \in D} \left(\sum_{y \in V} S_x \cdot \beta_s(x, y) \cdot \text{prof}(y) \right) \quad (6)$$

$$R_L = \sum_{x \in D} S_x \cdot \text{prof}(x) + \sum_{x \in D} \left(\sum_{y \in V} S_x \cdot \beta_L(x, y) \cdot \text{prof}(y) \right) \quad (7)$$

假设零售商放弃经营 m 种商品,将营销资源让给更有市场潜力的新商品。本文要研究的问题是,如何将

商品集合 I 最优划分为 2 个集合 D 和 V , 使得淘汰 D 中的 m 个商品后损失的收益最小。

2 交叉销售效果分析

为了计算交叉销售效果系数 $\beta(x, y)$, 需要从历史数据中分析那些被客户共同购买的关联模式。关联规则是数据挖掘的一个主要内容, 它反映了客户在购买了一种商品时还会购买另外一种商品的必然规律。

本文采用文献[7]所提出的简单二元关联规则来进行交叉销售效果分析。简单二元关联规则可表示为: $x \Rightarrow y(s)(c)$, 规则中 $x \in I, y \in I$ 。规则的含义是: 如果在某个交易中出现商品 x , 那么该交易也会包含商品 y 。关联规则是否成立需要用支持度(s)和信任度(c)两个指标进行判断。支持度 s 是指所有包含 x 和 y 的交易记录数占总交易数的比例, 信任度 c 是指同时包含 x 和 y 的交易记录数占包含 x 的交易记录数的比例。

上述关联规则仅仅定性地描述一个交易是否包含某些商品, 而对交易量没有定量描述, 这类关联规则被称为定性关联规则。事实上, 不同的商品间不仅存在定性的关联关系, 也存在量的关联关系。对此, 文献[9-10]提出了定量关联规则的概念和挖掘算法。为了论述方便, 做如下定义。

定义 3 令 $Q(x)$ 表示在一次交易中商品 x 的交易数量, 定量关联规则表示为 $x[V_s^x, V_L^x] \Rightarrow y[V_s^y, V_L^y](s)(c)$, 规则中 $x \in I, y \in I$ 。规则的含义是, 如果某交易包含商品 x , 且有 $V_s^x < Q(x) \leq V_L^x$, 那么, 该交易中也包含商品 y , 且有 $V_s^y < Q(y) \leq V_L^y$ 。支持度(s)是指同时购买了商品 x 和 y , 且购买的数量分别满足 $V_s^x < Q(x) \leq V_L^x$ 和 $V_s^y < Q(y) \leq V_L^y$ 的交易数量占总交易数的比例。信任度(c)是指购买了商品 x 且购买数量满足 $V_s^x < Q(x) \leq V_L^x$ 的交易中同时还购买了商品 y , 且购买数量满足 $V_s^y < Q(y) \leq V_L^y$ 的交易所占的比例。

根据定义 3, 定量关联规则反映了客户购买了一定数量的某商品时, 还会购买一定数量的另外一种商品的概率。因此, 利用定量关联规则, 可以计算出两种商品的交叉销售效果。

定义 4 设定量关联规则 $R: x[V_s^x, V_L^x] \Rightarrow y[V_s^y, V_L^y](s)(c)$, 规则 R 对两种商品交叉销售效果的最大贡献和最小贡献可以分别按式(8)和式(9)计算。

$$\gamma_R^L(x, y) = \frac{V_L^y}{V_L^x} \cdot c \quad (8)$$

$$\gamma_R^S(x, y) = \frac{V_S^y}{V_L^x} \cdot c \quad (9)$$

式(8)的含义是在乐观的情况下, 每卖出 1 件商品 x , 导致 $\gamma_R^L(x, y)$ 件商品 y 的销售, 而式(9)的含义是在悲观的情况下, 每卖出 1 件商品 x , 导致 $\gamma_R^S(x, y)$ 件商品 y 的销售。例如, 有规则 $a[3, 5] \Rightarrow b[20, 30][5\%][80\%]$, 规则的含义是如果客户购买了 3~5 件商品 a , 那么他有 80% 的可能还会购买 20~30 件商品 b , 因此, 在最乐观的情况下, 客户购买 3 件 a , 就会有 80% 的概率购买 30 件 b , 那么, 该规则对交叉销售效果的最大贡献为 $\gamma_R^L(a, b) = \frac{30}{3} \times 80\% = 8$ 。同理, 根据式(9), 该规则的交叉销售最小贡献为 3.2。

定义 5 设定量关联规则 $R: x[V_s^x, V_L^x] \Rightarrow y[V_s^y, V_L^y](s)(c)$, 如果某交易包含商品 x , 且 x 的购买量在 $[V_s^x, V_L^x]$ 的区间范围内, 则称该交易满足规则 R 的前提或规则 R 的前提覆盖该交易。

两件商品间可能存在多个数量上的关联关系, 为了计算出交叉销售效果系数, 需要综合两件商品间所有定量关联规则的贡献。不同的规则覆盖的交易数量是不同的, 因此, 不同的规则对交叉销售效果的贡献也是不同的。设规则前提为商品 x , 结论为商品 y 的定量关联规则有 k 条, 分别为 $R_1, R_2, \dots, R_k$, 令 S_x 为商品 x 的销售量, $S_x(R_i)$ 表示满足规则 R_i 前提的所有交易中商品 x 的交易量, 那么, 规则 R_i 对交叉销售效果的贡献权重可以按式(10)计算。

$$\omega_i = \frac{S_x(R_i)}{S_x} \quad (10)$$

根据每条规则的最大贡献和最小贡献, 可以分别按式(11)和式(12)计算出商品 x 对商品 y 的最大交叉销售效果系数 $\beta_L(x, y)$ 和最小交叉销售效果系数 $\beta_S(x, y)$ 。

$$\beta_L(x, y) = \sum_{i=1}^k \omega_i \cdot \gamma_{R_i}^L \quad (11)$$

$$\beta_S(x, y) = \sum_{i=1}^k \omega_i \cdot \gamma_{R_i}^S \quad (12)$$

3 商品选择的遗传算法

在商品集合 D 中的商品之间也存在交叉销售效果, 由这些商品之间的交叉销售所带来的间接收益是不应该记入收益损失的, 我们需要考虑的是集合 D 中的商品对集合 V 中的商品的交叉销售效果。显然, 本文研究的问题属于对集合进行最优划分的问题, 我们采用遗传算法求解。所谓用遗传算法求解一个优化问题就是对

该优化问题存在的许多解, 计算每个解对应的适应函数, 优化的过程就是要寻找这样的解, 使得与之对应的适应函数最大或最小。

3.1 商品选择问题的图论描述

为了论述方便, 我们对商品选择问题用图论术语描述如下: 在图 $G(I, A)$ 中, I 表示顶点集合, $I = \{I_1, I_2, \dots, I_n\}$, 每个顶点 I_i 代表一种商品, 对 G 中的每个顶点相应地有一个数 $\text{prof}(I_i)$ 表示商品 I_i 的直接收益。对 G 中的某一条有向边 $(I_i \rightarrow I_j)$, 相应地有一个数值区间 $[\beta_s(I_i, I_j), \beta_L(I_i, I_j)]$ 表示商品 I_i 对商品 I_j 存在交叉销售效果。如果商品 I_i 对商品 I_j 不存在交叉销售效果, 那么与有向边 $(I_i \rightarrow I_j)$ 对应的数值为 0。

顶点之间是否有边联结由邻接矩阵 A 决定。邻接矩阵 $A = [a_{ij}]$ 是一个 $N \times N$ 阶方阵, 如果某定量关联规则的前提为商品 I_i , 结论为 I_j , 那么 $a_{ij} = 1$, 否则 $a_{ij} = 0$ 。如果 $i = j$, $a_{ij} = 0$ 。如果 $a_{ij} = 0$, 则可认为顶点 I_i 和 I_j 之间没有边联结。

商品选择问题是将图 $G(V, A)$ 划分为两个子图, $G_D(D, A_D)$ 和 $G_V(V, A_V)$, 图 G_D 中顶点集合 D 包含 m 个淘汰商品, 且 $D \cap V = \Phi, D \cup V = I$ 。

3.2 染色体编码

本文按文献[11-12]的原理对染色体进行编码。对于一个给定的图模型, 将图 G 中各个顶点按照顶点号自然排序, 然后按此顺序将每个待选顶点作为染色体的一个基因, 当基因值为 1 时, 表示相应的顶点被选入子图 G_D 中, 如果基因值为 0 时, 表示相应的顶点没有被选入子图 G_D 中, 而是被划分到子图 G_R 中。染色体的长度等于图 G 中顶点的数目。例如, 某个染色体编码为 001100..., 该染色体的含义是 3 号和 4 号商品被选入了子图 G_D 中, 而 1, 2, 5, 6, ... 号商品被划分到子图 G_R 中。

3.3 基于不精确信息的适应度函数

根据式(6),(7), 可以计算出淘汰子图 G_D 中所有商品所损失的收益, 其结果是一个区间值 $[R_s, R_L]$ 。我们将收益损失定义为适应度函数, 由于适应度函数是一个区间值, 为了比较两个染色体的优劣, 做如下规定。

令符号“ $>$ ”表示关系“优于”, 符号“ $<$ ”表示关系“劣于”, 符号“ $\geq$ ”表示关系“弱优于”, 符号“ $\leq$ ”表示关系“弱劣于”。如果有两条染色体 X 和 Y , 且 $X > Y$ 或 $X \geq Y$, 则认为染色体 X 所代表的划分损失的收益更小, 所以染色体 X 优于 Y 。我们按下面的 4 条准则对任意两个染色体 X, Y 进行比较。

准则 1 如果 $R_s(X) > R_L(Y)$, 那么 $X < Y$ 。

准则 2 如果 $R_L(X) < R_s(Y)$, 那么 $X > Y$ 。

但满足准则 1 和 2 的情况很少出现。例如, 染色体 X 的适应度函数值为 $[4000, 6000]$, 而染色体 Y 的适应

度函数值为 $[3000, 8000]$, 这时无法利用上面的准则进行比较。在这种情况下, 则根据最小后悔的原则进行比较^[13-14]。令指标 $Z(X, Y)$ 表示选择染色体 X 而不选择染色体 Y 的最大后悔程度, $Z(X, Y)$ 可按式进行计算

$$Z(X, Y) = R_L(X) - R_s(Y) \quad (13)$$

同理, $Z(Y, X)$ 可按式进行计算

$$Z(Y, X) = R_L(Y) - R_s(X) \quad (14)$$

利用指标 $Z(X, Y)$ 和 $Z(Y, X)$, 我们可按如下准则进行染色体的比较, 并对所有染色体进行排序。

准则 3 如果 $Z(X, Y) < Z(Y, X)$, 那么 $X > Y$ 。

准则 4 如果 $Z(X, Y) > Z(Y, X)$, 那么 $X < Y$ 。

3.4 初始解群的形成

首先我们计算出每个商品的收益 (包括直接收益和间接收益), 由于间接收益是个区间值, 因此, 商品 x 的收益仍然是个区间值 $[R_s(x), R_L(x)]$, 其计算公式为

$$R_s(x) = S_x \cdot \text{prof}(x) + \sum_{y \in V} S_x \cdot \beta_s(x, y) \cdot \text{prof}(y) \quad (15)$$

$$R_L(x) = S_x \cdot \text{prof}(x) + \sum_{y \in V} S_x \cdot \beta_L(x, y) \cdot \text{prof}(y) \quad (16)$$

利用前面介绍的区间值比较准则, 根据每个商品的总收益损失对商品进行由小到大的排序, 并相应地设置不同的选择概率, 收益损失小的商品对应的选择概率大, 收益损失大的商品对应的选择概率小。对一组个体进行排序的算法是一种被广泛使用的算法, 近年来, 有很多基于时间复杂度的改进算法被提了出来, 将时间复杂度由 $O(n^2)$ 改进为 $O(n)$, 这一问题的解决有大量典型算法^[15-16], 本文不再论述。

根据相应的选择概率, 我们随机地进行淘汰商品的选择, 如果选择出 m 个商品, 一个染色体就形成了, 然后再重新开始选择, 以便形成另一个染色体, 这样反复, 直到形成 k 个不同的染色体。

3.5 交叉和变异操作

3.5.1 双亲染色体的挑选

根据前代染色体适应度函数值的大小确定作为双亲染色体的选择概率, 利用前面介绍的区间值排序方法, 对前代染色体进行排序, 并设置相应的概率, 收益损失小的个体被选中的概率较大。根据每个染色体的选择概率, 随机选择两个染色体作为双亲染色体。反复进行双亲选择, 这样就形成了若干双亲对。

3.5.2 交叉继承操作

比较两个双亲染色体的基因编码, 将两个染色体中相同的基因片段直接遗传给子代染色体。这时子代染色体上还有一些基因的值为空, 说明这些基因片段还没有从双亲染色体上继承基因值。为了保证在继承过程中, 子代染色体中刚好有 m 个值为 1 的基因, 我们按如下顺序实现从双亲染色体中进行交叉继承。

1) 计算出子代染色体中已经继承了基因值的基因

中值为 1 的基因数量, 假设为 r 个, 那么, 子代染色体上值为空的基因中应该有 $(m-r)$ 个基因的值置为 1, 其余值为空的基因置为 0。

2) 将子代染色体中值为空的基因的位置号取出, 根据这些位置号, 依次浏览双亲染色体中相同位置的基因值, 如果双亲染色体中有任一染色体在该位置的基因值为 1, 那么, 将该位置号放入集合 H 中。

3) 按相同的概率在集合 H 中选择 $(m-r)$ 个位置号, 在子代染色体上, 将这 $(m-r)$ 个位置号上的基因值置为 1, 其他值为空的基因值置为 0。

3.5.3 变异操作

通过交叉继承操作, 从前代染色体中遗传出了一些子代染色体, 以一个较小的概率 (例如 0.5%) 从子代染色体中随机选择一个染色体进行变异操作, 为了保证变异后的染色体中有 m 个值为 1 的基因, 对该染色体做如下变异操作。

1) 以等概率从值为 0 的基因中挑选出一个基因, 将其值变为 1。

2) 以等概率从值为 1 的基因中挑选出一个基因, 将其值变为 0。

4 模拟实验结果

4.1 实验过程

笔者观测了上海某连锁便利店的交易数据, 这些数据包含了 100 种商品连续 15 天的交易样本, 并分析了各种商品以及商品组合被购买的概率分布。基于对样本数据的分析结果, 笔者设计了一个可以生成交易记录的模拟程序, 利用该程序, 可以模拟出一定时间里发生的交易数据。

实验分为以下步骤。

1) 规定一个固定的时间长度 T , 利用模拟程序生成时间长度 T 内发生的交易记录。

2) 采用 Srikant 在 1996 年提出的算法发现模拟数据中的定量关联规则, 并将发现的规则存入知识库中。

3) 根据定量关联规则计算交叉销售效果系数。

4) 采用遗传算法选择 m 个收益能力低的商品, 其选择的结果为 D_1 , 将选择的商品淘汰后, 剩下的商品组成集合 $V_1=I-D_1$, 再利用商品直接收益排序法 (仅仅根据商品的直接收益排序来选择商品) 选择 m 个收益能力低的商品, 其选择的结果为 D_2 , 剩下的商品组成集合 $V_2=I-D_2$ 。

5) 分别用商品集合 V_1 和 V_2 中的商品再生成两组模拟交易数据, 其模拟的时间长度仍然为 T , 然后, 计算出这两组交易数据的收益。

6) 比较遗传算法和商品直接收益排序法对商品收

益的影响。

4.2 实验结果分析

在定量关联规则的挖掘过程中, 需要确定最优的支持度与信任度阈值, 如果阈值过高, 尽管规则的精度很高, 但发现的规则数量很少, 许多对决策有帮助的规则不能被发现出来; 如果阈值过低, 尽管发现的规则很多, 但规则的精度太低, 降低决策的质量。为此, 笔者用不同的参数组合进行了 12 次挖掘运算。利用遗传算法, 分别根据 12 种数据挖掘结果进行商品选择, 在这些实验中, 规定淘汰 8 种效益低下的商品, 选择结果对收益的影响如表 1 所示。表 1 中, 收益栏是指删除某些商品后, 零售商从剩下的商品中获得的收益, 因此, 收益值越大越好。

表 1 不同参数组合对商品选择效果的影响
Table 1 Influence of different parameter groups on the effectiveness of item selection

参数值		规则数	淘汰部分商品后零售商的收益	
支持度 阈值/%	信任度 阈值/%		遗传算法	直接收益 排序/万元
1	80	205	8.4	10.4
1	85	180	9.6	
1	90	156	10.1	
3	80	185	12.5	
3	85	152	15.1	
3	90	96	13.4	
5	80	117	13.7	
5	85	96	12.2	
5	90	78	11.1	
7	80	82	10.9	
7	85	44	10.7	
7	90	12	10.5	

从表 1 可以看出以下几点。

1) 用支持度阈值为 3%、信任度阈值为 85% 的参数组合时, 商品选择的效果最优。

2) 当参数值偏低时, 商品的收益急剧降低, 原因是许多规则的精度很低, 这样会夸大交叉销售的效果, 结果导致一些直接收益大的商品被错误地淘汰了。

3) 当参数值偏大时, 商品的收益趋于稳定, 这是因为阈值很高时, 许多规则不能成立, 选择商品时交叉销售效果几乎没有被考虑进来, 这样, 基于遗传算法的商品选择结果接近于用直接收益排序法的选择结果。

根据对表 1 的分析, 笔者确定最优支持度阈值为 3%, 信任度阈值为 85%, 并用这组参数组合对应的关联规则来计算交叉销售效果系数。令 m 为需要选择出的淘汰商品的数量, 表 2 是 m 取不同数值时两种商品选择方法对收益影响的对比。从表 2 可以发现遗传算法

表 2 两种商品选择方法的收益比较
Table 2 Comparison of revenue between two different approaches for item selection

淘汰商品的数量	淘汰部分商品后的收益	
	遗传算法	直接收益排序/万元
1	24	24
2	21	20.8
3	19.6	19.1
4	18.5	16.2
5	17	14.2
6	16	12.7
7	15.3	11.2
8	15.1	10.4
9	14.9	9.2
10	13.1	7.1

优于直接收益排序法,但是,当 m 很小时,遗传算法和直接收益排序法的结果没有太大区别,当 m 变大时,利用遗传算法进行商品选择可以获得更大的收益。这是因为,直接收益排序法没有考虑交叉销售效果的影响,如果 m 很小,交叉销售效果的影响也很小;反之,如果 m 很大,交叉销售效果的影响也很大。

5 结论

随着市场需求的动态变化,商品选择成为市场营销领域的一项重大决策。商品之间的交叉销售效果是影响商品选择结果的一个重要因素。在现实生活中,交叉销售效果不可能是一个精确的数值,而是一个具有一定置信度的变化区间,本文提出的以定量关联规则技术来计算交叉销售效果的方法正反映了这一客观实际。鉴于商品选择问题是一个 NP 难问题,本文根据遗传算法的原理对问题进行求解,提出的基于不精确区间值数据的适应度函数可以使遗传算法很好地应用于商品选择问题。实验结果表明,本文提出的商品选择方法具有较高的精确度。

参考文献 (References)

[1] FELDMAN J M. Package deal [J]. Air Transport World, 2003, 40(3): 40-42.
[2] NASH D, STERNA-KARWA A. An application of DEA to measure branch cross selling efficiency [J]. Computers and Operations Research, 1996, 23(4): 385-392.
[3] KAMAKURA W A, WEDEL M, ROSA F. Cross-selling through database marketing: a mixed data factor analyzer doe data augmentation and prediction [J]. International Journal of Research in Marketing, 2003, 20(1): 45-65.
[4] RAYMOND C W W, ADA W C. Maximal-profit Item Selection with Cross-Selling Considerations[C]. Proceedings of the third IEEE international conference on data

mining (ICDM'03), Los Alamitos, CA, USA, 2003.
[5] BRIJS T, SWINNEN G, VANHOOF K, *et al.* Using association rules for product assortment decisions: A case study [C]. Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, San Diego: ACM Press, 1999: 254-260.
[6] BRIJS T, GOETHALS B, SWINNEN G, *et al.* A data mining framework for optimal product selection in retail supermarket data: the generalized profset model [C]. Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, Boston: ACM Press, 2000: 300-304.
[7] WANG K, YEN M, SU T. Item selection by "hub-authority" profit ranking[C]. Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton: ACM Press, 2002: 652-657.
[8] RAYMOND W C W, ADA F W C, KE W. Data mining for inventory item selection with cross-selling considerations [J]. Data Mining and Knowledge Discovery, 2005, 11(1): 81-112.
[9] SRIKANT R, AGRAWAL R. Mining quantitative association rules in large relational tables [C]. Proceedings of ACM SIGMOD international conference on data management. Tucson: ACM Press, 1996: 1-12.
[10] 程岩, 卢涛, 黄梯云. 在数据库中挖掘定量关联规则的方法研究[J]. 管理科学学报. 2001, 4(4): 41-48.
CHENG Yan, LU Tao, HUANG Tiyun. Research on the method of mining quantitative association rules in database[J]. Journal of Management Sciences in China, 2001, 4(4): 41-48.
[11] GREFFENSTETTE J J. Optimization of control parameters for genetic algorithm [J]. IEEE Transaction on Systems Man and Cybernetics, 1986, 16(1): 122-128.
[12] REEVES C R, GENETIC A. Algorithm for Flow Shop Sequencing [J]. Computers and Operations Research, 1995, 22(1): 5-13.
[13] 程岩, 解帆. 一种基于不精确信息的科研项目评估决策支持系统[J]. 情报学报, 2003, 22(6): 730-734.
CHENG Yan, XIE Fan. DSS for R&D project assessment based on imprecision information [J]. Journal of the China Society for Scientific and Technical Information, 2003, 22(6): 730-734.
[14] JAE K K, SAND H C. A utility range-based interactive group support system for multi-attribute decision making [J]. Computer & Operations Research, 2001, 28(3): 486-503.
[15] 方同祝, 胡正国, 田铮, 等. 一种节省空间的排序算法[J]. 小型微型计算机系统, 2005, 26(7): 1200-1201.
FANG Tongzhu, HU Zhengguo, TIAN Zheng, *et al.* Space-minimized sorting algorithm [J]. Mini-micro Systems, 2005, 26(7): 1200-1201.
[16] HE W M, CUI J Z. A new mixed sorting method by grading[J]. Mini-micro Systems, 2003, 24(10): 1804-1807.

(责任编辑 王芷)