学 术 论 文

可拓学在数据挖掘中的应用初探

李立希[1]，李铧汶[1]，杨春燕[2]
（1. 广东工业大学计算机学院，广州 510090；2. 广东工业大学可拓工程研究所，广州 510090）

[摘要] 可拓学在数据挖掘中的应用是多方面的，其特点是挖掘“不行变行”的规律。可拓方法丰富了数据挖掘的内容，为多值型关联规则的建立提供了新的工具。提出的可拓数据挖掘模式，有利于利用现存数据更好地为决策服务。

[关键词] 可拓学；可拓集合；可拓方法；数据挖掘
[中图分类号] TP311 [文献标识码] A [文章编号] 1009－1742（2004）07－0053－07

1 引言

数据挖掘出现于 20 世纪 80 年代后期，90 年代有了突飞猛进的发展，它是当前业界的热门技术，已经在多个应用领域产生了巨大的效益。

但是，“有关数据挖掘的理论基础研究还没有成熟。坚实的和系统的理论基础对于数据挖掘非常重要，因为它给数据挖掘技术的开发、评价和实践提供了一致的框架。”[1]现有的数据挖掘理论基础包括有数据归纳、数据压缩、模式发现、概率理论、微观经济观点、归纳数据等等。

可拓学的研究对象是客观世界中的矛盾问题，目前，已经建立了把问题进行形式化描述的模型，并利用事物的可拓性和可拓变换，建立了解决矛盾问题的可拓方法和可拓工程方法[2, 3]。另一方面，可拓学也为数据挖掘提供了新的研究方法和工具。

2 可拓数据挖掘模式

可拓学的理论基础是基元理论、可拓集合理论和可拓逻辑，统称可拓论。为了将可拓学应用于实际，研究了基于可拓论的各种方法，如发散树、分合链、相关网、蕴含系、共轭对等可拓方法；优度评价法、真伪信息判别法等评价方法；论域变换、关联规则变换和基元变换等变换方法[4]。

数据挖掘亦称数据开采，文献［5］在“SAS 的数据开采的方法论——SEMMA”一节中指出“SAS 软件研究所对数据开采所下的定义是：数据开采是按照既定的业务目标，对大量的企业数据进行探索，提出隐藏其中的规律，并进一步模型化的先进、有效的方法。”

初步研究表明，可拓学在数据挖掘中的应用是多方面的：物元的可拓性可以成为挖掘的规则，如可以利用发散分析进行分类、聚类，利用相关分析与蕴含分析提出关联规则；利用可拓集合方法挖掘“不行变行的规律”；利用事元为基础的挖掘工具；利用物元变换特别是传导变换设计数据挖掘工具。

用物元 R 或事元 I 表示业务目标，记为 g。利用相关分析找出影响目标 g 的条件 l，构成问题模型

$$P = g * l,$$

其中“$*$”表示由目标 g 和条件 l 构成问题，若问题不相容，则表示为 $g\uparrow(l)$。

当业务目标不能实现，即问题不相容时，可以选择适当的可拓变换 T_g，把问题变成相容问题：

[收稿日期] 2003－07－15；修回日期 2004－01－13
[基金项目] 国家自然科学基金资助项目（79870107；70140003；70271060）；广东省自然科学基金资助项目（010049）
[作者简介] 李立希（1945－），男，广东博罗县人，广东工业大学计算机学院副教授

$$(T_g\, g)\downarrow(l)。$$

也可以改变条件，即选择适当的可拓变换 T_l 进行操作 $T_l\, l$，从而把问题变成相容问题：

$$g\downarrow(T_l\, l)。$$

第三种是前二者的结合，使问题变成相容：

$$(T_g\, g)\downarrow(T_l\, l)^{[6]}$$

如果，T 属于传导变换，要寻找相关的主动变换 ψ_0，使 $T\Leftarrow\psi_0$。

为叙述方便，以由物元构成的问题为例进行论述，即讨论

$$P = R * r。$$

对于物元 R，可以通过可拓分析获得与 R 相关的存在物元集 $\{R_i\}$。再由 R_i 发掘与之相关的条件物元 r_i，组成集合 $\{r_i\}$。通过比较 r 与 r_i 的相关性而择优选择$r_j\in\{r_i\}$ 以及相关的主动变换 ψ'，其中有

$$T'\Leftarrow\psi',$$
$$R_j\downarrow(T'r_j)。$$

由 ψ'确定 ψ_0，实现

$$T\Leftarrow\psi_0,$$
$$R\downarrow(Tr)。$$

可拓数据挖掘与传统数据挖掘的差异之一是传统数据挖掘是知识的发现，而可拓数据挖掘可以挖掘可拓变换，即其结果可以是策略。

3 可拓集合方法在数据挖掘中的应用

3.1 物元可拓集的定义及现实意义

3.1.1 物元可拓集的定义

定义 1 给定物元集

$$W_R = \{R\} = \{R \mid R = (N, c, v),$$
$$N\in U, c\in \pounds(c), v\in V(c)\}。$$

设 c_0 为 R 的评价特征，量值为 $c_0(R) = x$，以 $V(c_0)$ 为论域，X_0 为正域，$X_0\subseteq V(c_0)$，作静态可拓集合 $\widetilde{X}_0$，关联函数为 $k(x)$。对给定变换 $T = (T_W, T_k, T_R)$，称

$$\widetilde{M}(R)(T) = \{(R, Y, Y') \ R\in T_{W_R}W_R,$$
$$Y = K(R) = k(x)\in(-\infty, +\infty),$$
$$Y' = T_K\, K(T_R\, R) = T_k\, k(x')\in(-\infty, +\infty),$$
$$x' = c_0(T_R\, R)\}$$

为 W_R 上的一个物元可拓集。这里规定：当 $R\in T_{W_R}W_R - W_R$ 时，$y = K(R) < 0$。称$y = K(R)$为关联函数，$Y' = T_K K(T_R\, R)$为可拓函数，其中 T_W 表示对论域 W 的变换，T_K 为对关联函数 K 的变换，T_R 为对元素 R 的变换[7]。

数据挖掘是一种提供商业优势的活动，主要结合可拓营销与客户资源管理 (CRM) 进行研究。

3.1.2 可拓域和稳定域的现实意义

称

$$M_+(R;T) = \{(R, y, y') \mid R\in T_W W,$$
$$y = K(R)\leqslant 0, y' = T_K\, K(T_R\, R)\geqslant 0\}$$

为正可拓域。它指由于变换 T 的实施，使原来“不行”的变“行”。因此挖掘正可拓域即挖掘“不行变行”的规律。

可拓域在数据挖掘中有其重要的现实意义。从可拓营销的角度，假如 R 为顾客，就是挖掘各种使非顾客变为顾客的经营之道。称

$$M_-(R;T) = \{(R, y, y') \mid R\in T_W W,$$
$$y = K(R)\geqslant 0, y' = T_K\, K(T_R\, R)\leqslant 0\}$$

为负可拓域。它指由于变换 T 的实施，使原市场的顾客变为非顾客的集合。随着行业竞争愈来愈激烈和获得一个新客户的开支愈来愈大，保持原有的客户工作愈来愈有价值。通过数据挖掘技术可以发现负可拓域，从而预测哪些客户具有高风险转移的可能性，可以预测客户的流失率。称

$$M_+(R;T) = \{(R, y, y') \mid R\in T_W W,$$
$$y = K(R)\geqslant 0, y' = T_K\, K(T_R\, R)\geqslant 0\}$$

为正稳定域。正稳定域属于客户的保持问题。保持客户的忠诚度将对提高客户的盈利能力有重要作用。忠诚度高的顾客是指那些多年来一直是公司客户的消费者，顾客和企业的这种关系可以给企业带来较大的利润。

若设 $y = K(R)$ 为顾客的盈利函数，并设阀值 $d > 0$，则当 $y = K(R)\geqslant d$ 时为“黄金级”客户，而 $0 < y < d$ 成为“青铜级”的客户。通过数据挖掘，要了解：有多少“黄金级”的客户会在下一年度变成“青铜级”客户？或有多少“青铜级”客户会在接下来的 12 个月内成为“黄金级”客户？销售和市场投资的最基本的标准就是保留住那些有价值的客户和促使那些价值不大的顾客转变成有价值的顾客。

3.2 关联函数的建立方法

可拓集合是用关联函数来刻画的，要把解决不相容问题的过程量化，首先必须建立可拓集合的关联函数。针对要解决的问题，可以运用实验或经营数据构造关联函数。其中，数据挖掘是十分重要的手段。通过数据挖掘建立关联函数，属于获取模

型。

现实生活和工程技术中的大量问题，其量值都是可以用实数来表示的，对于非数量化的值都可以通过数量化用实数来描述。因此，以实数为论域是基本的情形。常用的有简单关联函数、初等关联函数。

例如，商品的价格是变化的，如何确定合理的价位是决策的重要内容。无论是商家或者是顾客，对价格而言一般都可以概括为满意价与可接受价。设满意价的范围为 $X_0 = \langle a, b\rangle$，可接受价的范围为 $X = \langle c, d\rangle$ 其中 $X_0 \subseteq X$，对应的关联函数为

$$K(x) = \rho(x, X_0)/D(x, X_0, X)^{[4]},$$

其中

$$\rho(x, X_0) = |x - (a+b)/2| - (b-a)/2,$$

$$D(x, X_0, X) = \begin{cases} \rho(x, X) - \rho(x, X_0) & \text{其他}, \\ -1 & x \in X_0。 \end{cases}$$

主要参数 a，b，c，d 可以用统计方法从营销数据中获得。

也可以利用蕴含分析通过数据挖掘建立关联函数。若 $A@$，必有 $B@$，则称 A 蕴含 B，记作 $A \Rightarrow B$ 或 $B \Leftarrow A$，A 与 B 之间的关系称为蕴含关系，符号@表示存在。利用蕴含关系于数据挖掘，一种用法是对蕴含关系进行逆变换，即从结果反向找原因。

例如，当世界经济由传统工业走向知识经济时，企业目标也从成本最低或利润最大转向客户最满意和服务新颖。因此，消费者的抱怨信息十分重要。获取顾客抱怨信息的途径有多种多样，其中的一种方案是通过营销数据挖掘。如顾客对商品某些特征量值抱怨，则有如下蕴含关系：

（抱怨商品特征量值）⇒（不买商品）⇒（商品卖不出去）⇒（销售比例低）。

因此，可以从销售数据中挖掘销售比例，并利用抱怨信息与销售比例的相关联性而获取顾客对某些特征量值的抱怨信息。

以价格为例，设 c 为价格特征，$V(c)$ 为价格的量值域，了解顾客对某一范围的价格抱怨，可建立函数如下：

记 $V(c)$ 的幂集为 $\rho(V(c))$，函数 $f: \rho(V(c)) \to [0,1]$ 为顾客对某一价格范围的抱怨函数。对于 $A \subseteq V(c)$ 定义 $f(A) = f(M, N)$，其中 M 为价格 A 范围的商品的总数量，而 N 是在价格 A 范围已售出的商品总数。一般可设

$$f(M, N) = 1 - N/M,$$

其中 N/M 称为销售比例。

在抱怨函数基础上可以建立关联函数：根据实际经验或需要设定一抱怨阈值 $\alpha \in [0,1]$，定义关联函数

$$K(A) = \alpha - f(A)。$$

故有

$$K(A) = \alpha - f(M, N) = \alpha - (1 - N/M) = N/M - (1 - \alpha)。$$

令 $d = 1 - \alpha$，则

$$K(A) = N/M - d,$$

其中 d 称为销售比例目标值，$K(A) \in (-1,1)$。

3.3　挖掘可拓市场

市场的定义有很多种。在营销学中，把市场定义为“有能力购买且愿意购买某产品的顾客的全体”。在可拓营销中，用形式化的方法给出了可拓市场[8, 9]的概念。

定义 2　设论域 U 为人群集合，对 $u \in U$，令

$$R = \begin{pmatrix} u, & c_1, & v_1 \\ & c_2, & v_2 \end{pmatrix} = \begin{pmatrix} u, & \text{购买意愿}, & v_1 \\ & \text{购买能力}, & v_2 \end{pmatrix}$$

在 c_1 和 c_2 的值域 $V(c_1)$ 和 $V(c_2)$ 上，分别建立关联函数 $K_1(v_1)$ 和 $K_2(v_2)$，$v_1 \in V(c_1)$，$v_2 \in V(c_2)$，在物元集

$$W = \left\{ R \middle| R = \begin{pmatrix} u, & c_1, & v_1 \\ & c_2, & v_2 \end{pmatrix}, u \in U \right\}$$

上，对给定的变换 $T = (T_W, T_K, T_R)$，建立物元可拓集合

$$\widetilde{M}(R; T) = \{(R, y, y') \mid R \in T_W W,$$
$$y = K(R) = K_1(v_1) \wedge K_2(v_2),$$
$$y' = T_K K(T_R R)\},$$

其中 $T_K K(T_R R)$ 为关于变换 T 的可拓函数，称

$$\dot{M}(R; T) = \{R \mid R \in T_W W,$$
$$y = K(R) \leqslant 0, y' = T_K K(T_R R) \geqslant 0\}$$

为原市场 $M(R) = \{R \mid R \in W, K(R) \geqslant 0\}$ 关于变换 T 的可拓市场。

在大多数商业领域中，在业务发展的主要指标内，包括新客户的获取能力，要从企业数据中挖掘变换 $T = (T_W, T_K, T_R)$，挖掘 T_W 就是挖掘在数据中的扩大旧市场，用新市场置换旧市场、市场细分等经验。挖掘元素变换 T_R，就是挖掘各种成功地提高顾客购买意愿与购买能力的措施，包括各种行销措施、优惠措施与付款方式。

4 可拓关联规则

4.1 可拓关联规则的定义

可拓关联规则是多值型关联规则的开拓。

关联规则是数据挖掘中一种重要的模式，其应用很广泛，包括货篮计划、商品广告邮寄分析、仓储规划、银行欺诈甄别、网络故障分析等。关联规则可以分为2种：布尔型关联规则和多值型规则。目前对布尔型关联规则的挖掘算法研究得比较多，而有关多值型关联规则的文献较少。

多值型关联规则可表示为

$$\wedge P_i(X) \Rightarrow \wedge Q_j(X),$$

其中 $P_i(X), Q_j(X)$ 就是（语言变量）+（语言值）的形式[10]。

支持度（support）和可信度（confidence）是描述关联规则的2个重要概念。若 support $\geqslant$ minsupport（最小支持度）并且 confidence $\geqslant$ minconfidence（最小可信度），称关联规则为强规则，否则称关联规则为弱规则。传统数据挖掘仅限于研究强关联规则，而笔者对二者都进行研究，并统称二者为可拓关联规则。并用挖掘出来的可拓关联规则构造基础规则库，它是应用的基础。这也是可拓数据挖掘的特色之一，而且有很大的实用价值：有的应用问题需要对低支持度问题进行研究，这是一个新的研究方向；而可信度低的规则可以看作是问题的一种反映。这些可以扩大传统数据挖掘的应用范围。

在形式上，把多值型关联规则推广如下：

定义3 设 $G(g) = \{g \mid g = (N, c, v)$ 或 (d, h, u) 或 $(s, a, z)\}$ 称之为基元集合，其中 $R = (N, c, v)$ 称为物元，$I = (d, h, u)$ 称为事元，$Q = (s, a, z)$ 称为关系元。

定义4 $D = \{S \mid S \subseteq G(g)\}$，$S$ 称为事务，是一组基元，D 是一组事务集，$|D|$ 表示 D 中事务的数量。

定义5 设 $A \subseteq G(g)$ 是一个基元集，事务 S 支持 A 当且仅当 $A \subseteq S$，可拓关联规则是形如 $A \Rightarrow (l)B$ 的条件蕴含式，其中 $A \subseteq G(g)$，$B \subseteq G(g)$，并且 $A \cap B = \varnothing$。规定可拓关联规则的条件 $l = (\text{support}, \text{confidence}, k)$ 为一个三维向量，表示以支持度（support）、可信度（confidence）的大小来判断关联规则 $A \Rightarrow B$ 的真实程度，或对 A 与 B 的相关程度进行定量化分析。令 $|A|$ 表示在 D 中支持 A 的事务数，$|A \cup B|$ 表示在 D 中同时支持 A 与 B 的事务数。$A \Rightarrow B$ 的支持度为 support $= |A \cup B| / |D|$，可信度为 confidence $= |A \cup B| / |A|$。用关联函数 k 来表示“兴趣度”，以便对规则进行评价。

4.2 可拓关联规则的表示形式

可拓关联规则的一般形式为

$$\wedge g_i \Rightarrow (l) \wedge g_j, \text{其中 } g_i, g_j \in G(g) \quad (1)$$

常用的可拓关联规则有

$$\wedge g_i \Rightarrow (l)g, \text{其中 } g_i, g \in G(g) \quad (2)$$

如果 g_i，g_j，g 可以取物元、事元或关系元，则称为混合型可拓关联规则。特殊情况有2种，一种是物元型可拓关联规则

$$\wedge R_i \Rightarrow (l) \wedge R_j, \text{或} \wedge R_i \Rightarrow (l)R;$$

另一种是事元型可拓关联规则

$$\wedge I_i \Rightarrow (l) \wedge I_j, \text{或} \wedge I_i \Rightarrow (l)I。$$

4.3 组合型关联规则

组合型关联规是指在规则中既有基元项又有可拓变换项。常用的形式有

$$A_1 \wedge A_2 \wedge \cdots \wedge A_n \Rightarrow (l)B \quad (3)$$

其中 $A_i (i = 1, 2, \cdots, n), B \in G(g) \cup \mathrm{T}$。

组合型规则的挖掘适合于研究复杂系统的关联规律。例如，如果顾客为教工，工资在2000元以上，已买电脑，而且与其专业相关的新软件价格七折，则他们会买该软件。用组合型关联规则表示为

$$R_1 \wedge I \wedge \mathrm{T} \Rightarrow I_0,$$

其中

$$R_1 = \begin{bmatrix} x, & \text{职业}, & \text{教师} \\ & \text{工资额}, & 2000\text{元}/\text{月} \end{bmatrix},$$

$$I = \begin{bmatrix} \text{购买}, & \text{支配对象}, & \text{电脑 } A \\ & \text{施动对象}, & x \end{bmatrix},$$

$$\mathrm{T}(\text{软件 } B, \text{价格}, v) = (\text{软件 } B, \text{价格}, 0.7v),$$

$$I_0 = \begin{bmatrix} \text{购买}, & \text{支配对象}, & \text{软件 } B \\ & \text{施动对象}, & x \end{bmatrix}。$$

4.4 挖掘可拓关联规则

传统挖掘关联规则算法的代表是 Apriori 算法[1]，它的基础是频繁集概念，主要包含连接步与剪枝步，一般不能直接使用 SQL 挖掘。可拓关联规则的挖掘是利用计算机技术对可拓相关性与蕴含性定量分析的工具，其基础是等价关系及其交运算后的等价分类，因而可以直接利用 SQL 中的 GROUP BY 语句来完成。

5　可拓数据挖掘方法应用案例分析

数据挖掘作为决策支持新技术在近 10 年得到迅速发展。利用可拓方法挖掘“不行变行的规律”，将进一步开拓数据挖掘在决策支持系统的应用。

以商品房营销决策支持为例。商品房是一种较复杂的商品，一些特征的量值之间存在相关性，并共同影响顾客的态度，进而影响销售。如何挖掘营销数据中的有用信息，为决策提供参考方案？在可拓数据挖掘方法框架下，采用编程语言 Delphi 6.0，数据库系统为 MS SQL Server 2000，利用标准 SQL 查询，从商品房营销数据中挖掘关联规则。

5.1　建立可拓模型

把某一段时间内销售额大于或等于 50 % 作为目标，建立可拓模型如下：

$$P = I * r = \begin{bmatrix} 销售, & 支配对象, & \{房子\ A\} \\ & 比例, & \geqslant 50\% \end{bmatrix} * \begin{bmatrix} \{房子\ A\}, & 建设套数, & M \\ & 销售套数, & N \end{bmatrix}。$$

根据实践经验与专家意见，房子的销售状况与单价、景观、户型、装修等特征有关。因此，每套房子用多维物元表示如下：

$$R_{ijk} = \begin{bmatrix} 房子\ A_{ijk}, & 单价\ c_1, & v_{ijk1} \\ & 景观\ c_2, & v_{ijk2} \\ & 户型\ c_3, & v_{ijk3} \\ & 装修\ c_4, & v_{ijk4} \\ & 销售状况\ c_5, & v_{ijk5} \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{bmatrix},$$

$V_1(c_1) = \{3000,3500,4000,4500,5000\}$，

$V_2(c_2) = \{优,良,中,差\}$，

$V_3(c_3) = \{地下花园大,地下花园小,$
$复式大,复式小,普通大,普通小\}$，

$V_4(c_4) = \{高档,中档,低档\}$，

$V_5(c_5) = \{已售,待售\} = \{1,0\}$。

5.2　相关分析

利用关联规则挖掘潜在知识，对于给定的与任务相关的数据集，数据挖掘过程可能会发现许多规则，但并不都是能为用户服务的。为了使挖掘更具针对性，用户可提供各种约束。

在案例中，“为什么房子卖不出去？什么样的房子卖不出去？”与“什么样的房子容易售出？如何使房子卖出去？”，为下一步的销售及楼盘的二期建设提供决策依据。

从上面的模型可以看出，要提高销售比例，即要改变销售状况；而销售状况与单价、景观、户型、装修有关。故策略的关键在于寻求变换 ψ_0，使单价、景观、户型、装修实现最优组合，以提高顾客的购买意愿与购买能力。以上分析可以表述为模式

$$R_1 \wedge R_2 \wedge R_3 \wedge R_4 \Rightarrow (l) R_5 \qquad (4)$$

但这些仅仅是定性分析，还要在已有的营销数据基础上，利用计算机技术进行定量分析，即求 l = (support, confidence, k)。对相关性进行定量分析的重要工具之一是数据挖掘，案例的挖掘类型属于式 (2) 中的特例，即物元型可拓关联规则

$$\wedge R_i \Rightarrow (l) R。$$

案例中某类型的房子的可信度 confidence = 销售套数/建筑套数 (即销售比例)。

5.3　系统的结构

系统由规则生成与可拓变换及评价组成。

5.3.1　*规则生成*　系统首先生成基础数据库，然后根据需要可生成问题数据库、强关联数据库等。

1) *基础数据库*　基础数据库存放的是支持度 support>0 而且可信度 confidence≥0 的所有元组，它是产生其他规则的基础。在应用时，也可以根据用途筛选，使数据库存量减少，如表 1 所示。

表 1　基础数据库 (部分)

Table 1　Basic database (part)

序号	条件 1 (单价/元)	条件 2 (景观)	条件 3 (户型)	条件 4 (装修)	结果 (销售)	支持度	可信度	关联函数 k
1	3000				已售	0.125	0.567	0.067
22	3500	良			已售	0.113	0.704	0.204
112	3500	良	普通大		已售	0.075	0.556	0.056
245	4000	良	普通小	高档	已售	0.075	0.667	0.167
251	4000	中	普通小	高档	已售	0.025	0.333	−0.167

2) *问题数据库*　问题库来源于基础数据库，其条件为关联函数 $k \leqslant 0$，其中 k = confidence −

d。通过问题库可以发现问题以及与问题相联系的因素，这是问题不相容的定量形式。首先，利用关联函数提出问题，一般可设定关联函数 $k \leqslant 0$ 为有问题，并对问题数据库按关联函数 k 进行升序排序，则问题的大小就一目了然。在案例中，把某一段时间内销售比例大于或等于 50 % 作为目标，故设

$$k = \text{confidence} - 0.5,$$

$k = -0.5$ 表示最大问题。由基础数据库中关联函数 $k \leqslant 0$ 的元组组成问题数据库，部分问题库见表2。其中，用统计性量值支持度、可信度和以及在此基础上生成的关联函数表示问题不相容的量值，而其他则是与问题相关联的因素，对于相关联的因素又可进一步分为条件属性与结果属性。

表2 问题库（部分）

Table 2 Question database (part)

序号	条件1（单价/元）	条件2（景观）	条件3（户型）	条件4（装修）	结果（销售）	支持度	可信度	关联函数 k
251	4000	中	普通小	高档	已售	0.025	0.333	-0.167
224	3000	良	地下花园大	中档	已售	0.038	0.333	-0.167
230	3500	差	普通小	中档	已售	0.058	0.357	-0.143
242	4000	差	普通大	高档	已售	0.021	0.400	-0.100
229	3500	差	普通大	中档	已售	0.025	0.500	0.000
237	3500	中	普通大	中档	已售	0.025	0.500	0.000

3) *强关联数据库* 从强关联数据库中发现的规则即是传统的关联规则，它符合 support ≥ minsupport，而且 confidence≥minconfidence，强关联数据库见表3。

表3 强关联数据库（support≥0.06，confidence≥0.6）

Table 3 Association rules

序号	条件1（单价/元）	条件2（景观）	条件3（户型）	条件4（装修）	结果（销售）	支持度	可信度	关联函数 k
56	4500			中档	已售	0.075	0.611	0.111
75		优	普通小		已售	0.075	0.667	0.167
128	4000	优	普通小		已售	0.075	0.667	0.167
160	4500	优		中档	已售	0.075	0.722	0.222
245	4000	良	普通小	高档	已售	0.075	0.667	0.167
247	4000	优	普通大	中档	已售	0.075	0.833	0.333
248	4000	优	普通小	中档	已售	0.075	0.667	0.167

5.3.2 *可拓变换及其评价*

对于从问题库中选择出来的问题，为了解决问题就要进行各种变换，而对变换进行评价的常用方法是优度评价法，通过优度评价，可择优选择变换。也可以利用范例库对可拓变换评分，基于范例库的推理是20世纪80年代末90年代初期新崛起一项重要技术，其核心在于用过去的实例和经验来解决问题[11]，把数据挖掘技术应用于范例库的构造，将有利于提高知识获取的自动化程度。详细的评价方法将另文论述。

表4表示问题与变换，其中的问题来源于问题库，然后利用可拓变换对与问题关联的条件进行变换，变换结果的支持度、可信度与关联函数则来源于强关联数据库。

表4 问题与变换

Table 4 Question and transformation

类型	条件1（单价/元）	条件2（景观）	条件3（户型）	条件4（装修）	结果（销售）	支持度	可信度	关联函数 k
问题	4000	中	普通小	高档	已售	0.025	0.333	-0.167
变换1	4000	良	普通小	高档	已售	0.075	0.667	0.167
变换2	4000	优	普通小	中档	已售	0.075	0.667	0.167
变换3	4000	优	普通大	中档	已售	0.075	0.833	0.333

对选择出来的满足关联函数 $k \geqslant 0$ 的有 3 个变换，可以生成如下解决矛盾问题的策略：

在销售期间发现问题，想改变户型与装修不容易，因此优先考虑变换 1，即景观由中变为良。对二期建设也可改变户型、装修与景观，故同时考虑变换 2 与变换 3。

6　结语

可拓学在数据挖掘中的应用是多方面的，笔者研究了可拓集合方法、可拓变换方法与可拓分析方法等在数据挖掘中的应用，并提出可拓数据挖掘的一种模式，为数据挖掘提供了新的方法。传统数据挖掘得出的结果是知识，而可拓数据挖掘的结果是策略，预料在营销决策支持系统中的应用更具有优势，并为可拓策略生成系统的研究提供有力的工具。

参考文献

[1] Jiawei Han. 数据挖掘概念与技术 [M]. 范　明，孟小峰，等译. 北京：机械工业出版社，2001

[2] 涂序彦. 可拓学——研究“矛盾转化，开拓创新”的新学科 [J]. 中国工程科学，2002，2(2)：97

[3] 蔡　文，杨春燕，何　斌. 可拓学与人工智能 [A]. 中国人工智能进展 [M]. 北京：北京邮电大学出版社，2001. 1064～1068

[4] 蔡　文，杨春燕，林伟初. 可拓工程方法 [M]. 北京：科学出版社，2000

[5] 高洪深. 决策支持系统(DSS) [M]. 北京：清华大学出版社，2000

[6] 李立希，李　嘉. 可拓知识库系统及其应用 [J]. 中国工程科学，2001，3(3)：61～64

[7] 蔡　文，杨春燕，何　斌. 可拓逻辑初步 [M]. 北京：科学出版社，2003

[8] 杨春燕，张拥军. 可拓市场的类型与实现方式研究 [J]. 工业工程，2002，5(3)：46～49

[9] 杨春燕，张拥军. 可拓策划 [M]. 北京：科学出版社，2002

[10] 杨炳儒，孙海洪，熊范纶. 利用标准 SQL 查询挖掘多值型关联规则及其评价 [J]. 计算机研究与发展，2002，39(3)：307～312

[11] 陈文伟，施平安，何　义，陈　鹏. 数据仓库的可拓决策分析工具 [J]. 中国人工智能进展. 北京：北京邮电大学出版社，2001. 1085～1089

Study on the Application of Extenics in Data Mining

Li Lixi[1], Li Huawen[1], Yang Chunyan[2]

(1. *Institute of Computer, Guangdong University of Technology, Guangzhou* 510090, *China*; 2. *Research Insitute of Extension Engineering, Guangdong University of Technology, Guangzhou* 510090, *China*)

[**Abstract**] The application of extenics in data mining is multi-aspect, and its characteristic is to mine the rule of “unable to able”. The extension method enriches the content of data mining, and provides new tools for building multivalue correlative criteria. The extension data mining model, which is given in this article, is favorable to using existing data for making decision.

[**Key words**] extenics; extension set; extension method; data mining