

文章编号: 1001-0920(2012)06-0850-05

基于粗糙分类的不确定可拓群决策数据挖掘及应用

朱佳俊¹, 郑建国¹, 李金兵²

(1. 东华大学 管理学院, 上海 200051; 2. 北京大学 国家发展研究院, 北京 100871)

摘要: 针对可拓分类的不足及分类中存在的 uncertainty, 将关联函数与决策函数相结合, 研究不确定条件下可拓群决策的可拓变换与粗糙分类、属性约简与规则提取、数据分析与方案识别, 以实现决策对象在不确定条件下的比较与选择; 解决了多方案可拓分类和动态识别的系统分类问题, 提高了可拓群决策数据挖掘的准确性和可信度。

关键词: 可拓集; 可拓决策; 粗糙分类; 数据挖掘; 规则提取

中图分类号: N94

文献标识码: A

Rough classification algorithm for uncertain extension group decision-making

ZHU Jia-jun¹, ZHENG Jian-guo¹, LI Jin-bing²

(1. School of Business and Management, Donghua University, Shanghai 200051, China; 2. National School of Development, Peking University, Beijing 100871, China. Correspondent: ZHU Jia-jun, E-mail: zhujiajun@mail.dhu.edu.cn)

Abstract: For the lack of extension classification and the uncertainty in classification, this paper studies extension transformation and rough classification, attribute reduction and rules extraction, data forecast and program identification of extension group decision-making under uncertainty by combining correlation function and decision function, to improve the accuracy and the reliability of data mining by making comparison and selection of objects in changing environment, and to solve systematic classification problems of multi-project extension classification, dynamic recognition in extension group decision-making.

Key words: extension set; extension decision-making; rough classification; data mining; rules extraction

1 引言

可拓群决策^[1]是用物元可拓集描述的群决策问题, 将可拓变换与群决策优化相结合, 研究多维物元群决策的可拓关联与可拓决策空间, 以实现决策对象在可变环境中的比较与选择. 可拓群决策数据挖掘是将可拓学理论、群决策技术和数据挖掘方法相结合, 用于挖掘群决策中可拓知识的量化和形式化的工具. 基于给定的决策对象、关联准则、论域的可拓变换, 可拓群决策数据挖掘对可拓集进行划分^[2], 描述由于决策对象本身及外部条件的变换而引起的转化, 反映不确定条件下决策对象质量互变的程度、范围及边界, 通过挖掘变化知识、转化矛盾问题、提取决策规则、生成选择策略等, 满足系统化、智能化群决策的需要. 已经在数据分类^[3]与聚类^[4]、错误识别与诊断^[5]、知识表示^[6]与挖掘^[7]等方面得到了成功的应用.

基于可拓分类的群决策分类方法会出现方案分类的不确定性, 这种不确定性一方面是由可拓分类方法的局限性造成的; 另一方面, 事物、现象或决策的复杂性也会造成可拓群决策效果的不确定, 从而影响客观、有效的评价. 本文通过分析可拓群决策分类产生的不完备决策信息系统, 将属性约简和可拓转换结合起来以实现基于粗糙分类的不确定可拓群决策数据挖掘, 弥补可拓群决策分类的不足, 从而实现不确定条件下的数据动态分类、决策规则提取以及数据区间分析, 提高方案识别、判断和选择的效果.

2 不确定可拓群决策粗糙分类

2.1 可拓群决策

可拓群决策在节域和经典域的基础上, 通过可拓变化比较与观察不同条件下各方案的变化, 确定决策

收稿日期: 2010-12-01; 修回日期: 2011-05-27.

基金项目: 国家自然科学基金项目(70971020).

作者简介: 朱佳俊(1976—), 男, 博士后, 从事智能决策、知识管理的研究; 郑建国(1962—), 男, 教授, 博士生导师, 从事数据挖掘等研究.

方案集划分的依据和关键, 根据决策目标和要求对方案集归类 and 划分, 以实现方案的动态分析.

定义 1 设 $W_i = W_1 \times W_2 \times \dots \times W_n$ 为决策论域, $a_i = \{a_1, a_2, \dots, a_n\}$ 为可拓群决策方案集, a_i 的决策者集表示为 $c_j = \{c_1, c_2, \dots, c_m\}$, $c_j(a_i) = (u_{i1}, u_{i2}, \dots, u_{im})$ 为决策者 c_j 对方案 a_i 的评价向量值, 则可拓群决策的方案物元为 $a_i = (N, c_j, u_{ij})$. 给定变换 $T = (T_W, T_K, T_a)$, 称

$$\begin{aligned} \hat{A}(a) &= \{(a, y, y') | a \in T_W W, \\ & y = K(a) \in (-\infty, +\infty), \\ & y' = T_K K(T_a a) \in (-\infty, +\infty)\} \end{aligned} \quad (1)$$

为实域 W 上的一个可拓群决策集. 当 $a \in T_{W_a} W_a - W_a$ 时, $y = K(a) < 0$ 为关联函数, $y' = T_K K(T_a a)$ 为可拓函数. 其中: T_W 表示对论域 W 的变换, T_K 表示对关联函数 $K(a_i)$ 的变换, T_a 表示对决策方案 a_i 的变换; $i = 1, 2, \dots, n, j = 1, 2, \dots, m$.

定义 1 中, $K(a)$ 为方案 a_i 关于决策者 c_j 的综合关联度^[8], γ_i 为方案 a_i 的权重系数, $K_i(a_i)$ 为方案 a_i 关于决策者 c_j 的方案关联度, χ_j 为决策者 c_j 的权重系数. 即

$$K(a) = \sum_{i=1}^n (\gamma_i \cdot K_i(a_i)); \quad (2)$$

$$K_i(a_i) = \frac{1}{\max_{1 \leq j \leq m} |k_j(u_{ij})|} \sum_{j=1}^m \chi_j \cdot k_j(u_{ij}); \quad (3)$$

$$k_j(u_{ij}) = \begin{cases} -\rho(u_{ij}, v_{lj})/|v_{lj}|, & u_{ij} \in v_{lj}; \\ \frac{\rho(u_{ij}, v_{lj})}{\rho(u_{ij}, v_{pj}) - \rho(u_{ij}, v_{lj})}, & u_{ij} \notin v_{lj}. \end{cases} \quad (4)$$

$k_j(u_{ij})$ 为评价量值与区间的关联度, $v_{pj} = [q_{pj}, b_{pj}]$ 为节域, $v_{lj} = [q_{lj}, b_{lj}]$ 为经典域, $v_{lj} \in v_{pj} (l = 1, 2, \dots, g)$, $\rho(u_{ij}, v_{lj})$ 为点 u_{ij} 与经典域 v_{lj} 的距离, $\rho(u_{ij}, v_{pj})$ 为点 u_{ij} 与节域 v_{pj} 的距离.

2.2 不确定可拓群决策的粗糙分类

将可拓群决策与粗糙集的分类方法相结合, 通过可拓变化实现方案的动态分类, 并运用属性约简实现方案的再分类, 可以提高可拓分类的效果.

定义 2 设可拓群决策 $a_i = (N, c_j, u_{ij})$, 经可拓变化后形成 $\tau = (A, C, F, d)$. 如果任意 $L(c_j(a_i))$ 且 $d(a_i)$ 是唯一确定的, 则称为完备的可拓群决策信息系统; 否则, 称为不完备的信息系统.

设 $\tau = (A, C, F, d)$ 是不完备可拓群决策信息系统, A 为决策方案集, C 为决策者属性集, $a_i \in A$, $BS \in C$, d 为决策结果属性. 记

$$A/R_{\{d\}} = \{D_1, D_2, \dots, D_t\}, \quad (5)$$

$$m_{BS}(a_i) = \max\{D(D_j/S_{BS}(a_i)) | j \leq t\}. \quad (6)$$

其中: $D(E/F) = |E \cap F|/|F|$ 为 $P(A)$ 上的包含度; 而

$$S_{BS}(a_i) = \{a_{i+1} \in A | (a, a_{i+1}) \in \text{SIM}(BS)\} \quad (7)$$

为 a 的相似类; $h_l \in BS$, $\text{SIM}(BS) = \{(a_i, a_{i+1}) \in A \times A | h_l(a_i) \cap h_l(a_{i+1}) \neq \emptyset\}$ 为 BS 的相似关系;

$$\text{rp}_B(a_i) = \{D_{ji} | D(D_{ji}/S_B(a_i)) = m_B(a_i)\} \quad (8)$$

为决策函数. 若对于任意 $a_i \in A$, 有 $\text{rp}_{BS}(a_i) = \text{rp}_C(a_i)$ 成立, 则称 BS 是 τ 的最大分布协调集. 若 BS 是最大分布协调集, 且 BS 的任何真子集都不是 τ 的最大分布协调集, 则称 BS 是 τ 的最大分布约简集^[9].

设 S_δ 是不完备可拓群决策信息系统 τ 的所有选择中的一个, BS_δ 是 S_δ 的最大分布约简集, $BS_h = \bigcap BS_\delta \neq \emptyset$ 是 τ 全部选择的最大分布约简集的最小集合 ($\delta = 1, 2, \dots; h = 1, 2, \dots$). 当 $CS = \bigcap BS_h \neq \emptyset$ 时, 称 CS 为划分核心决策者集; $JS = \bigcup BS_h - CS$ 为划分相对必要决策者集; $QS = C - \bigcup BS_h$ 为划分不必要决策者集.

根据决策者对方案的决策评价价值 $L(c_j(a_i))$ 与方案综合评价价值 $d(a_i)$ 离差绝对值之和 $f(c_j)$ 最小来确定分类的根属性, $c_j \in QS$, 有

$$f(c_j) = \min_{j=1}^m \left(\sum_{i=1}^n |L(c_j(a_i)) - d(a_i)| \right). \quad (9)$$

当 $BS_h > 1$ 时, 通过 BS_h 与方案综合评价价值 $d(a_i)$ 离差绝对值之和 $f(B_\psi)$ 最小来确定分类的子属性 B_ψ .

$$f(B_\psi) = \min(f(c_j) + f(c_{j+1})). \quad (10)$$

其中: $B_\psi \in BS_h, c_j \cup c_{j+1} = B_\psi$. 用根属性 c_j 创建分类的始节点, 并以该决策评价属性值, 对根属性的每个值创建分枝, 并将全部选择的最大分布约简集的最小集合 B_ψ 作为子属性, 再引出分枝, 并据此划分样本.

3 基于不确定可拓粗糙分类的规则提取

3.1 可拓群决策的决策规则及决策函数

不确定可拓群决策信息系统 τ 中, $L(c_j(a_i))$ 表示决策者对方案的评价类别, $d_g(a_i)$ 表示确定方案的评价类别, $rL = \bigwedge (c_j, L(c_j(a_i)))$ 构成决策规则的前件, $rD = \bigvee (d_g, d_g(a_i))$ 构成决策规则的后件. 方案对象 $a \in A$ 支持规则 $rL \rightarrow rD$ 的充分必要条件是 τ 中的对象 a 同时拥有条件部分 rL 和结论部分 rD .

定义 3 设 $\tau = (A, C, F, d)$ 是协调的不确定可拓群决策信息系统, $B \subseteq C$, 则 $R_d^>$ 条件下的辨识集是 $D([a_i]_C, [a_{i+1}]_C)$, 当且仅当 $R_B^> \subseteq R_d^>$, $B \cap D([a_i]_C, [a_{i+1}]_C) \neq \emptyset$ ($D([a_i]_C, [a_{i+1}]_C) \neq \emptyset, i = 1, 2, \dots, n, i + 1 \in n$).

$R_d^> = \{([a_i]_C, [a_{i+1}]_C) | d([a_i]_C) \geq d([a_{i+1}]_C)\}$ 是

可拓群决策属性 d 上的优势关系^[10], 辨识矩阵为 $D = (D([a_i]_C, [a_{i+1}]_C) | [a_i]_C, [a_{i+1}]_C \in A/R_C)$, 则辨识属性集为

$$D = \begin{cases} \{a_i \in A | (a_i, a_{i+1}) \notin R_{\{a_i\}}^{\geq}\}, (a_i, a_{i+1}) \notin D; \\ \emptyset, (a_i, a_{i+1}) \in D. \end{cases} \quad (11)$$

用 λ 表示 A/R_d^{\geq} 的个数, $\lambda = 1, 2, \dots, \partial, i+1 \in n, i+1 \leq n$, 则 R_d^{\geq} 条件下的决策函数 DEF λ 为^[11]

$$\text{DEF}\lambda = \bigwedge_{\lambda=1}^{\partial} \bigvee_{c=1}^j D([a_i]_C, [a_{i+1}]_C), \quad (12)$$

\bigwedge 和 \bigvee 分别表示合取和析取运算. 于是得到 λ 的决策规则为

$$r_a : rL \rightarrow rD = (c_1, \geq) \wedge (c_2, \geq) \wedge \dots \rightarrow (d_g(a_i), \geq). \quad (13)$$

3.2 基于粗糙分类的可拓群决策的规则提取

定义 4 设 $r_a : rL \rightarrow rD$ 是不确定可拓群决策信息系统 $\tau = (A, C, F, d)$ 方案对象 a 的决策规则, 于是规则的确度、覆盖度和支持度分别为

$$\text{cer}_{\tau}(rL \rightarrow rD) = \frac{|[a]_{rL} \cap [a]_{rD}|}{|[a]_{rL}|}, \quad (14)$$

$$\text{cov}_{\tau}(rL \rightarrow rD) = \frac{|[a]_{rL} \cap [a]_{rD}|}{|[a]_{rD}|}, \quad (15)$$

$$\text{sup}_{\tau}(rL \rightarrow rD) = \frac{|[a]_{rL} \cap [a]_{rD}|}{|A|}. \quad (16)$$

定义 5 设 $r_a = (c_1, \geq) \wedge (c_2, \geq) \wedge (c_3, \geq) \rightarrow (d_g(a_i), \geq)$ 为不确定可拓群决策信息系统 τ 的一条决策规则, $c \in C$, 如果存在决策规则

$$r'_a = r_a = (c_1, \geq) \wedge (c_2, \geq) \rightarrow (d_g(a_i)), \quad (17)$$

则决策规则 $r_a : rL \rightarrow rD$ 所匹配的方案集能同时满足规则 $r'_a : rL \rightarrow rD$ 的前提和结论, 记为 $r_a \subset r'_a$.

4 基于可拓群决策粗糙分类的方案识别

4.1 不确定可拓群决策规则的数据区间分析

可拓群决策将关联函数与决策函数结合起来, 运用可拓转换得到决策规则, 使得决策规则具有一定的数据区间, 这种数据区间为不确定条件下识别方案提供了数据依据.

定义 6 设 $\tau = (A, C, F, d)$ 是协调的不确定可拓群决策信息系统, $c_j \in C, d_g(a_i) \in V$. 当 c_j 为 r_a 的增序排列 $(c_j, \geq, d_g(a_i))$ 或降序排列 $(c_j, \leq, d_g(a_i))$ 的一个属性时, 在节域 v_{pj} 和经典域 v_{lj} 条件下, 通过关联度 $K_v(u_{ij})$ 得到决策规则 r_a 下属性 c_j 的数据区间

$$\tilde{u}_{c_j} = \begin{cases} (b_{lj}, b_{pj}], f_{c_j}(a) > v_l; \\ [q_{pj}, q_{lj}), f_{c_j}(a) < v_l; \\ [q_{lj}, b_{lj}], f_{c_j}(a) = v_l. \end{cases} \quad (18)$$

当 $\{a_{\varphi}\}$ 为决策规则 r_a 的增序排列或降序排列

的方案集时, $\varphi \in n$, 根据可拓群决策的关联对应性^[12], $\{a_{\varphi}\}$ 的最小关联度 $0 \leq K(\min(u_{ij})) \leq 0.5$, 其中 $\min(u_{ij})$ 的对应点为 $\xi = q_{lj} + b_{lj} - \min(u_{ij})$. 于是在 $f_{c_j}(a) = d_g(a_i)$ 的条件下, $u = \min(u_{ij})$ 或 $u = \max(u_{ij})$, 经典域中点 $v_{AV}(l_j) = (q_{lj} + b_{lj})/2$, 则方案集 $\{a_{\varphi}\}$ 关于属性 c_j 的核心数据区间和可拓数据区间分别为

$$\text{core}(\tilde{u}_{c_j}) = \begin{cases} [u, \xi], u < v_{AV}(l_j); \\ [\xi, u], u \geq v_{AV}(l_j). \end{cases} \quad (19)$$

$$\text{exte}(\tilde{u}_{c_j}) = \begin{cases} [q_{lj}, u) \cup (\xi, b_{lj}], u < v_{AV}(l_j); \\ [q_{lj}, \xi) \cup (u, b_{lj}], u \geq v_{AV}(l_j). \end{cases} \quad (20)$$

4.2 不确定可拓群决策规则的方案识别

不确定可拓群决策的方案识别是在决策规则和可拓数据区间的基础上对不确定条件下的新方案进行识别和分类, 对决策规则的更新以及为辅助决策提供相应的策略.

定义 7 对于不确定条件下可拓群决策信息系统 $\tau = (A, C, F, d)$ 在 R_d^{\geq} 条件下的决策规则 r_a 及符合其规则的方案集 $\{a_{\varphi}\}, \varphi \in n$, 属性集 c_j 的核心数据区间集为 $\text{core}(\tilde{u}_{c_j})$, 可拓数据区间集为 $\text{exte}(\tilde{u}_{c_j})$. 当一新方案的物元为 $a'' = (N, c_j, u''_{ij})$ 时, 如果其决策者的评价集 $u''_{ij} \subseteq \text{core}(\tilde{u}_{c_j})$, 则方案 a'' 完全符合 R_d^{\geq} 条件下的决策规则 $r_a, a'' \subset \{a_{\varphi}\}$; 如果 $u''_{ij} \subseteq \text{exte}(\tilde{u}_{c_j})$, 则方案 a'' 基本符合决策规则 $r_a, a'' \in \{a_{\varphi}\}$.

定义 7 表明, 由于物元的经典域 v_{lj} 是一个区间, $\text{core}(\tilde{u}_{c_j}) \cup \text{exte}(\tilde{u}_{c_j}) = v_{lj}$. 由初始数据生成的决策规则 r_a 关于属性集 c_j 的核心数据区间集 $\text{core}(\tilde{u}_{c_j})$, 以及根据 $\text{exte}(\tilde{u}_{c_j})$ 和 v_{lj} 生成的可拓数据区间, 都符合规则 r_a 的数据区间要求. 因此, 可拓群决策规则的基础是一个数据区间, 通过结合核心划分集 CS, 为方案识别和缺失数据估计提供了适当的条件和范围, 从而提高了可拓规则的有效性.

5 实例分析

2010 年末, 汕头某大型玩具制造公司为了迎接元旦的销售旺季, 拟开展产品组合促销活动. 公司通过建立可拓群决策模型, 对产品组合销售方案进行了系统分析与决策, 实现方案分类并确定最佳方案.

公司根据各产品系列主题的生产规模和销售业绩, 抽取各系列产品组成 12 个组合销售方案, 建立方案集 a_i . 同时建立专家集 $c = \{c_1, c_2, c_3, c_4, c_5\}$, 并设专家初始权重 $\chi_j = (0.2, 0.2, 0.2, 0.2, 0.2)$. 设定评价节域 $v_{pj} = [6, 10]$, 经典域 $v_{lj} = \{l_1, l_2, l_3, l_4\}$. 其中: $l_1 = [6, 7.5]$ 为合格, $l_2 = [7.5, 8.5]$ 为一般, $l_3 = [8.5, 9.5]$ 为良好, $l_4 = [9.5, 10]$ 为优.

通过数据采集,得到专家对方案集评价的可拓群决策的复合物元 a_i ,即

$$a_i = (N, c_j, u_{ij}) = \begin{bmatrix} 7.2 & 8.6 & 7.9 & 7.6 & 7.3 \\ 7 & 8.6 & 7.8 & 7.9 & 7.4 \\ 9.6 & 9.3 & 9.8 & 9.6 & 9.7 \\ 9 & 8.7 & 8.9 & 8.3 & 8.4 \\ 6.9 & 8.8 & 8.1 & 7.7 & 6.9 \\ 8 & 8.6 & 7.6 & 7.6 & 6.5 \\ 8 & 8.2 & 7.7 & 8.8 & 7.8 \\ 8.5 & 8.4 & 9 & 8.4 & 8.2 \\ 9.6 & 9.3 & 9.7 & 9.4 & 9.6 \\ 8.3 & 8.1 & 8.2 & 8.9 & 8 \\ 8.8 & 8.9 & 8.8 & 8.4 & 8.4 \\ 9.8 & 9 & 9.6 & 9.5 & 9.7 \end{bmatrix}.$$

进而通过可拓转换和分类,得到可拓群决策评价价值,见表 1.

表 1 可拓群决策评价价值

a	$L(c_1)$	$L(c_2)$	$L(c_3)$	$L(c_4)$	$L(c_5)$	d
1	1	3	2	2	1	1
2	1	3	2	2	1	1
3	4	3	4	4	4	4
4	3	3	3	2	2	3
5	1	3	2	2	1	1
6	2	3	2	2	1	1
7	2	2	2	3	2	2
8	2,3	2	3	2	2	2,3
9	4	3	4	3	4	4
10	2	2	2	3	2	2
11	3	3	3	2	2	3
12	4	3	4	3,4	4	4

根据定义 2, $a_i = (N, c_j, v_{lj})$ 是一个不完备信息系统,因此所有选择的最大属性约简集的最小集合是 $\{c_1, c_2\}$ 和 $\{c_1, c_5\}$,核心决策者集为 $C = \{c_1\}$. 运用式

(9)和(10)得到分类的根属性 $\{c_3\}$ 和最大分布约简集的最小属性集合 $\{c_1, c_5\}$,并得到如下决策规则:

$r_1 : (c_5, \leq, 1) \rightarrow (d, =, 1)$, 确定度 $\text{cer}_\tau(r_1) = 1$, 覆盖度 $\text{cov}_\tau(r_1) = 0.75$, 支持度 $\text{sup}_\tau(r_1) = 0.33$.

规则 r_1 关于 c_5 的核心数据区间 $\text{core}(\tilde{u}_{c_j}) = [6.5, 7.4]$, 可拓数据区间 $\text{exte}(\tilde{u}_{c_j}) = [6, 6.5] \cup [7.4, 7.5]$.

$r_2 : (c_5, \geq, 4) \rightarrow (d, =, 4)$, 确定度 $\text{cer}_\tau(r_2) = 1$, 覆盖度 $\text{cov}_\tau(r_2) = 1$, 支持度 $\text{sup}_\tau(r_2) = 0.25$.

规则 r_2 关于 c_5 的核心数据区间 $\text{core}(\tilde{u}_{c_j}) = [9.6, 9.7]$, 可拓数据区间 $\text{exte}(\tilde{u}_{c_j}) = [9.5, 9.6] \cup [9.7, 10]$.

$r_3 : (c_1, \geq, 4) \vee (c_2, \geq, 3) \rightarrow (d, \geq, 4)$, 决策规则的确定度 $\text{cer}_\tau(r_3) = 1$, 覆盖度 $\text{cov}_\tau(r_3) = 1$, 支持度 $\text{sup}_\tau(r_3) = 0.25$.

规则 r_3 关于决策者 $c_1 \cup c_2$ 的核心数据区间 $\text{core}(\tilde{u}_{c_j}) = ([9.6, 9.8])_{c_1} \cup ([9, 9.3])_{c_2}$, 可拓数据区间 $\text{exte}(\tilde{u}_{c_j}) = ([9.5, 9.6] \cup [9.8, 10])_{c_1} \cup ([8.5, 9] \cup [9.3, 9.5])_{c_2}$.

a_{new} 是由 a_{13}, a_{14} 和 a_{15} 组成的新产品组合方案物元,即

$$a_{\text{new}} = \begin{bmatrix} 9.6 & 8.8 & 9.5 & 9.6 & 9.6 \\ 9.1 & 8.6 & 9 & 9.1 & 9.5 \\ 7 & 7.3 & 6.8 & 7.1 & 6.6 \end{bmatrix}.$$

利用生成的主要规则可以判断: $a_{13} \in d_4$, 该产品组合方案的市场接受度较高; $a_{15} \in d_1$, 该方案市场接受度较低,销售前景不好;而方案 a_{14} 可能属于 d_3 ,即 $a_{14} \in d_3$, 该方案的市场接受度尚可.

将 C4.5^[13], Naive Bayes^[14] 和 BP^[15] 3 种方法与可拓群决策数据挖掘方法进行比较, 所得结果如表 2 所示. 从方案划分、规则生成以及识别结果来看, 本文

表 2 4 种方法分类效果比较

方法	方案分类	规则生成	识别结果
可拓群决策数据挖掘	$\{a_1, a_2, a_5, a_6\} \in d_1, \{a_7, a_8, a_{10}\} \in d_2$ $\{a_4, a_8, a_{11}\} \in d_3, \{a_3, a_9, a_{12}\} \in d_4$	$r_1 : (c_5, \leq, 1) \rightarrow (d, =, 1); r_2 : (c_5, \geq, 4) \rightarrow (d, =, 4)$ $r_3 : (c_1, \geq, 4) \vee (c_2, \geq, 3) \rightarrow (d, \geq, 4)$	$a_{13} \in d_4$ $a_{15} \in d_1$ $a_{14} \in d_3$
BP	$L(c_1(a_8)) = 3$ 时: $\{a_1, a_2, a_5\} \in d_1, \{a_6, a_7, a_{10}\} \in d_2$ $\{a_4, a_8, a_{11}\} \in d_3, \{a_3, a_9, a_{12}\} \in d_4$	—	$a_{13} \in d_4$ $a_{14} \in d_4$ $a_{15} \in d_2$
	$L(c_1(a_8)) = 2$ 时: $\{a_1, a_2, a_5, a_6\} \in d_1, \{a_7, a_8, a_{10}\} \in d_2$ $\{a_4, a_{11}\} \in d_3, \{a_3, a_9, a_{12}\} \in d_4$	—	$a_{13} \in d_4$ $a_{14} \in d_3$ $a_{15} \in d_1$
C4.5	$\{a_1, a_2, a_5, a_6\} \in d_1,$ $\{a_7, a_8, a_{10}\} \in d_2,$ $\{a_4, a_8, a_{11}\} \in d_3,$ $\{a_3, a_9, a_{12}\} \in d_4$	$r_1 : (c_3, =, 2) \vee (c_5, =, 1) \rightarrow (d, =, 1); r_2 : (c_3, =, 2) \vee (c_5, =, 2) \rightarrow (d, =, 2);$ $r_3 : (c_3, =, 3) \rightarrow (d, =, 3); r_4 : (c_3, =, 4) \rightarrow (d, =, 4);$ $r_5 : (c_5, =, 1) \rightarrow (d, =, 1); r_6 : (c_5, =, 4) \rightarrow (d, =, 4);$ $r_7 : (c_2, =, 2) \rightarrow (d, =, 2);$ $r_8 : (c_2, =, 3) \vee (c_5, =, 2) \rightarrow (d, =, 3);$ $r_9 : (c_2, =, 3) \vee (c_5, =, 1) \rightarrow (d, =, 1);$ $r_{10} : (c_2, =, 3) \vee (c_5, =, 4) \rightarrow (d, =, 4)$	$a_{13} \in d_4$ $a_{15} \in d_1$ $a_{14} \in d_3$
	Naive Bayes	—	—

方法的优势主要体现在 3 个方面: 1) 该方法结合了可拓变换和动态分类的长处, 适用于小样本、数据不确定及描述性决策; 2) 该方法的规则提取更快、更简洁, 对于不确定问题的判断更贴切; 3) 该方法不仅可以动态地反映类内的差别, 还可以生成规则的可拓数据区间, 用以观察方案数值的可变范围及其对分类和判断的影响, 对于方案的识别更加灵活.

6 结 论

针对可拓分类的不足与不确定性, 本文将物元分析、可拓变换和粗糙集结合起来, 综合了可拓变化与软划分的优点, 分析了可拓群决策分类、属性约简、规则提取、数据预测及方案识别的过程. 从静态和动态的角度观察决策者对方案分类的影响度以及决策者评价受可拓变化的影响, 确定主要影响因素和可拓变量, 实现了在群体交互和个体偏好集结基础上不确定群决策的数据挖掘, 提高了多方案动态分类、可拓识别的有效性和系统性. 结合产品组合销售的目的和组合方案, 运用可拓数据挖掘方法, 分析不确定条件下产品组合的状态及方案的优劣, 并对新方案进行判断和识别, 突出了可拓数据挖掘方法在不确定条件下数据分类、筛选、规则生成、方案识别等方面的系统性与合理性.

将基于粗糙分类可拓群决策数据挖掘方法与其他 3 种方法进行比较, 所得结果表明, 本文方法的方案分类受不确定数据影响较小, 生成的规则较简单、有效, 识别结果较为准确和全面.

参考文献(References)

- [1] Cai W. Extension theory and its application[J]. Chinese Science Bulletin, 1999, 44(17): 1538-1548.
- [2] 李兴森, 石勇, 李爱华. 基于可拓集的企业数据挖掘应用方案初探[J]. 哈尔滨工业大学学报, 2006, 38(7): 1124-1128.
(Li X S, Shi Y, Li A H. Study on enterprise data mining solution based on extension set[J]. J of Harbin Institute of Technology, 2006, 38(7): 1124-1128.)
- [3] 李琴, 王国会. 可拓分类方法及其在流动单元分类中的应用[J]. 地球物理学进展, 2007, 22(6): 1975-1979.
(Li Q, Wang G H. Research of the extension classification method and classification of flow units[J]. Progress in Geophysics, 2007, 22(6): 1975-1979.)
- [4] 邓卫红, 周科平. 矿岩可崩性的可拓聚类预测研究[J]. 中国安全科学学报, 2008, 18(1): 34-39.
(Deng W H, Zhou K P. Study on extension cluster prediction for orerock cavability[J]. China Safety Science, 2008, 18(1): 34-39.)
- [5] Lay Y L, Tsaich, Yang H J. The application of extension neural network on computer-assisted lip reading recognition for hearing impaired[J]. Expert Systems with Applications, 2008, 34(2): 1465-1473.
- [6] 钟诗胜, 张艳. 可拓知识表示及知识库系统的开发[J]. 计算机集成制造系统, 2008, 14(11): 2184-2190.
(Zhong S S, Zhang Y. Extension-based knowledge representation and development of knowledge base system[J]. Computer Integrated Manufacturing Systems, 2008, 14(11): 2184-2190.)
- [7] Wang M H. Extension neural network and its application[J]. Neural Networks, 2003, 16(5/6): 779-784.
- [8] Cai W. Extension management engineering and applications[J]. Int J of Operations and Quantitative Management, 1999, 5(1): 59-72.
- [9] Pawlak Z. Rought set[J]. Int J of Computer and Information Sciences, 1982, 11(5): 341-356.
- [10] Jelonek J. Rough set: Reduction of attributes and their domains for neural networks[J]. Computational Intelligence, 1995, 11(2): 339-347.
- [11] Krysckiewicz M. Rough set approach to incomplete information systems[J]. Information Sciences, 1998, 112(1): 39-49.
- [12] Zhu J J, Zheng J G, Qin C Y. Analysis and application of extension association and correspondence under uncertainty[J]. Advances in Natural Science, 2010, 3(3): 18-31.
- [13] Quinlan J R. C4.5: Programs for machine learning[M]. San Mateo: Morgan Kaufmann, 1993.
- [14] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifier[J]. Machine Learning, 1997, 29(1): 131-163.
- [15] Haykin S. Neural networks: A comprehensive foundation[M]. Prentice Hall, 2001.