

优势-模糊目标 VPRSM 及其应用

黄 兵¹ 周献中² 史迎春³

(南京审计学院信息科学学院 南京 210029)¹ (南京大学工程管理学院 南京 210093)²

(武汉通信指挥学院仿真中心 武汉 430010)³

摘 要 在群决策理论中,如何获取合理的决策规则是一个重要的研究内容。针对条件属性具有优势关系及决策值为模糊数的多决策信息系统,构建了基于优势关系的模糊目标信息系统变精度粗糙集模型,给出了该模型的几种知识约简定义;通过构造适当的启发式函数,得到了该模型的优势下分布约简算法。最后将该模型应用于计算机审计风险评估,得到了较为合理的评估规则。

关键词 变精度粗糙集模型,优势关系,群决策,计算机审计风险

中图法分类号 N94-0,C934 文献标识码 A

Dominance Relation-based VPRSM and its Application in Fuzzy Objective Information Systems

HUANG Bing¹ ZHOU Xian-zhong² SHI Ying-chun³

(School of Information Science, Nanjing Audit University, Nanjing 210029, China)¹

(School of Engineering & Management, Nanjing University, Nanjing 210093, China)²

(Simulation Center, Wuhan Institute of Communication Command, Wuhan 430010, China)³

Abstract In group decision-making theory, how to acquire reasonable decision rules is an important issue. This paper constructed a dominance relation-based variable precision rough set model (VPRSM) in fuzzy objective information systems, and proposed some knowledge reduction definitions such as dominance-based lower/upper distribution reduction. Using a heuristic function, we devised a lower distribution reduction algorithm. Finally, this model was applied to computer audit risk appraisal and some reasonable appraisal rules were acquired.

Keywords VPRSM, Dominance relation, Group decision-making, Computer audit risk

在多属性群决策问题中,属性值有优劣关系的属性称为准则,而属性值无优劣关系的称为属性。群决策分类问题主要考虑属性值无优劣关系的决策,这时决策者通过对属性值的主观偏好将不同决策方案或对象分配到预先设定的类型中,从而形成自己的偏好。然后将每个决策者的偏好用某种方法进行集结,形成群体偏好^[1]。

粗糙集理论是一门处理不精确、不确定信息的数学理论,经过二十多年的发展,在理论上已取得了长足进步,并成功应用于智能信息处理、知识获取、决策分析和数据挖掘等领域。近年来,将粗糙集理论应用于群决策方法的研究逐步成为一个热点。如刘业政等^[2]利用粗糙集理论建立了一种判断矩阵的构造方法;王珏^[3]等提出了一种群决策语言信息处理的粗糙集方法;文献^[4]利用粗糙集进行了区间多维刻划。

在群决策中,多个专家在进行决策时往往不一致。而变精度粗糙集模型可以较好地克服噪声数据的影响,因此,变精度粗糙集理论在群决策中得到了广泛应用。如文献^[5-8]利用变精度粗糙集模型分别讨论了 IT 项目外购风险评估、信用风险评估、软件项目投标风险规避和群决策分类方法等。

在实际问题的群决策过程中,各对象属性描述的属性值

往往具有一定的优劣关系,如方案的成本费用、实施难度、收益指标等;同时,单个专家在进行决策时往往很难准确地将对象评判为某一类,常常给出的判断结果是一个模糊的、不精确的结论。本文针对属性值具有优劣关系及决策值为模糊值的群决策问题,提出一种基于优势关系的模糊目标变精度粗糙集模型,并探讨模型的一些基本性质,提出模型上几种知识约简的定义,给出适当的启发式函数,进而设计出一种知识约简算法,最后将模型应用于计算机审计风险评估的规则获取。

1 基本概念

定义 1 称 $S=(U, A=C \cup D, V, f)$ 是一个信息系统,其中 U 是非空有限对象集合, $U=\{x_1, x_2, \dots, x_n\}$; C 是非空有限条件属性集合, D 是非空有限决策属性集合, 且 $C \cap D = \emptyset$ 。 f 是一个 $U \times A$ 到属性值集合 V 上的一个映射, 表示每个对象在每个属性上对应一个值, 称为信息函数。若 $\forall x_i \in U, d_j \in D, f(x_i, d_j) \in [0, 1]$, 则称信息系统是模糊目标信息系统。

对任意 $B \subseteq C$, 记 $R_B = \{(x_i, x_j) : f(x_i, b) \leq f(x_j, b), \forall b \in B\}$, 这里, $f(x_i, b) \leq f(x_j, b)$ 表示 x_j 在属性 b 上的取值不劣于 x_i 。显然 R_B 是 U 上的相似关系(满足自反性和传递性), 称

到稿日期:2009-04-08 返修日期:2009-07-09 本文受江苏省自然科学基金(BK2009395),江苏省高校自然科学基金基础研究面上项目(07KJD120087),南京审计学院科研重点项目(NSK2008/A04)资助。

黄 兵(1972-),男,博士后,副教授,研究方向为粗糙决策理论与应用,E-mail:hbhuangbing@126.com。

为由条件属性子集 $B \subseteq C$ 确定的不可区分关系, 记为 $U/R_B = \{[x]_B: x \in U\}$ 。其中 $[x]_B = \{y \in U: (x, y) \in R_B\}$ 表示 x 关于条件属性子集 B 的优势类。

定义 2 $X \subseteq U$ 是一个精确集, 由条件属性子集 $B \subseteq C$ 确定的 X 的下近似和上近似定义为

$$\begin{aligned} \underline{R}_B(X) &= \{x \in U: [x]_B \subseteq X\} = \bigcup \{[x]_B: [x]_B \subseteq X\} \\ \overline{R}_B(X) &= \{x \in U: [x]_B \cap X \neq \emptyset\} = \bigcup \{[x]_B: [x]_B \cap X \neq \emptyset\} \end{aligned}$$

类似文献[9], 袁修久^[10]给出了如下基于优势关系的模糊目标信息系统的粗糙集模型。

定义 3^[10] 优势模糊目标信息系统 $S = (U, C \cup D, V, f)$, $d \in D$ 关于 $B \subseteq C$ 的下近似 $\underline{B}(d)$ 和上近似 $\overline{B}(d)$ 定义为

$$\begin{aligned} \underline{B}(d)(x) &= \min\{f(y, d): y \in [x]_B\} \\ \overline{B}(d)(x) &= \max\{f(y, d): y \in [x]_B\} \end{aligned}$$

2 优势-模糊目标变精度粗糙集模型

群决策是多个决策主体(专家)针对同一对象集根据对象的已知属性特征进行的分类。设已知属性值具有优势关系, 专家给出的分类结果是一系列模糊值, 这就构成了一个基于优势关系的模糊多决策信息系统。由于决策的不一致性, 允许同一对象的所有决策值存在一定的误差。下面针对这种信息系统建立优势-模糊变精度粗糙集模型。

定义 4 优势模糊目标信息系统 $S = (U, C \cup D, V, f)$, $B \subseteq C, D = \{d\}$ 。对 $\beta \in (0.5, 1], \lambda (0 < \lambda \leq 1)$, 设 $d_i(x) (1 \leq i \leq m)$ 表示第 i 个决策主体对对象 $x \in U$ 给出的决策值(模糊值), 则 D 关于 B 的 β 下近似 $\underline{B}_\lambda^d(D)$ 和上近似 $\overline{B}_\lambda^d(D)$ 定义为

$$\begin{aligned} \underline{B}_\lambda^d(D)(x) &= \begin{cases} \min_{x \in [x]_B, d_i(x) \geq \lambda (1 \leq i \leq m)} d_i(x), & \frac{|\{d_i(x): d_i(x) \geq \lambda, x \in [x]_B\}|}{m |[x]_B|} \geq \beta \\ 0, & \text{otherwise} \end{cases} \\ \overline{B}_\lambda^d(D)(x) &= \begin{cases} \min_{x \in [x]_B, d_i(x) \geq \lambda (1 \leq i \leq m)} d_i(x), & \frac{|\{d_i(x): d_i(x) \geq \lambda, x \in [x]_B\}|}{m |[x]_B|} \geq 1 - \beta \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

称定义 4 的上下近似为优势-模糊目标信息系统变精度粗糙集模型。

定理 1 优势模糊目标信息系统 $S = (U, C \cup D, V, f)$, $E \subseteq B \subseteq C$, 则有

$$(1) \underline{B}_\lambda^d(D) \subseteq \underline{B}_\lambda^d(D) \quad (2) \overline{B}_\lambda^d(D) \subseteq \overline{B}_\lambda^d(D)$$

证明: 由定义 4 直接可得。

定理 1 的性质 2 说明, 当条件属性减少时, 下近似单调减小, 上近似单调增加。即随着条件属性的减少, 不确定性将单调上升。这与经典粗糙集模型是一致的。

定义 5 优势模糊目标信息系统 $S = (U, C \cup D, V, f)$, $B \subseteq C, \beta \in (0.5, 1], \lambda (0 < \lambda \leq 1)$, 记

$$LB_\lambda^d = \frac{\sum_{x \in U} |\underline{B}_\lambda^d(D)(x)|}{|U|} \quad UB_\lambda^d = \frac{\sum_{x \in U} |\overline{B}_\lambda^d(D)(x)|}{|U|}$$

定义 6 优势模糊目标信息系统 $S = (U, C \cup D, V, f)$, 对 $\beta \in (0.5, 1], \lambda (0 < \lambda \leq 1), B \subseteq C$ 。

(1) $\forall x \in U$, 若 $\underline{C}_\lambda^d(D)(x) > 0$, 均有 $\underline{B}_\lambda^d(D)(x) = \underline{C}_\lambda^d(D)(x)$, 则称 B 为优势 β - λ 下分布协调集。若 B 是优势 β - λ 下分布协调集, 而它的任何真子集都不是优势 β - λ 下分布协调

集, 则称 B 是优势 β - λ 下分布约简。

(2) $\forall x \in U$, 若 $\overline{C}_\lambda^d(D)(x) > 0$, 均有 $\overline{B}_\lambda^d(D)(x) = \overline{C}_\lambda^d(D)(x)$, 则称 B 为优势 β - λ 上分布协调集。若 B 是优势 β - λ 上分布协调集, 而它的任何真子集都不是优势 β - λ 上分布协调集, 则称 B 是优势 β - λ 上分布约简。

(3) 若 $\underline{B}_\lambda^d(D) = \underline{C}_\lambda^d(D)$ 且 $\overline{B}_\lambda^d(D) = \overline{C}_\lambda^d(D)$, 则称 B 为优势 β - λ 分布协调集。若 B 是优势 β - λ 分布协调集, 而它的任何真子集都不是优势 β - λ 分布协调集, 则称 B 是优势 β - λ 分布约简。

(4) 若 $LB_\lambda^d = LC_\lambda^d$, 则称 B 是 β - λ 下近似协调集。若 B 是 β - λ 下近似协调集, 而它的任何真子集都不是 β - λ 下近似协调集, 则称 B 是 β - λ 下近似约简。

(5) 若 $UB_\lambda^d = UC_\lambda^d$, 则称 B 是 β - λ 上近似协调集。若 B 是 β - λ 上近似协调集, 而它的任何真子集都不是 β 上近似协调集, 则称 B 是 β - λ 上近似约简。

优势 β - λ 下分布约简是保持每个对象在决策上的最小隶属程度不变的最小条件属性集合, 其确保原信息系统所有优势类决策值至少为 λ 的可信程度不低于 β 的规则; 优势上分布约简是保持每个对象在决策上的最大隶属程度不变的最小条件属性集合, 其确保原信息系统所有优势类决策值至少为 λ 的可信程度不低于 $1 - \beta$ 的规则; 而 β - λ 下(上)近似约简只保持所有对象在决策上总隶属度之和不变, 因此由它得到的规则与原信息系统可能发生冲突。

3 优势-模糊目标变精度粗糙集模型约简算法

下面讨论 β - λ 下分布约简的启发式算法。

定义 7 优势模糊目标信息系统 $S = (U, C \cup D, V, f)$, $\beta \in (0.5, 1], \lambda (0 < \lambda \leq 1), x, y \in U, \underline{C}_\lambda^d(D)(x) > 0, \exists c \in C$, 使得 $y \notin [x]_c$ 而 $y \in [x]_{C \setminus \{c\}}$, 若 $\exists i (1 \leq i \leq m)$, 有 $\lambda \leq d_i(y) < \underline{C}_\lambda^d(D)(x)$, 则 c 是 β - λ 下分布约简的核属性。

可以验证, 定义 7 给出的核属性定义与经典粗糙集理论中核属性的含义是一致的。

β - λ 下分布约简算法

输入: 一个优势模糊目标系统 $S = (U, C \cup D, V, f), \beta \in (0.5, 1], \lambda (0 < \lambda \leq 1)$

输出: $S = (U, C \cup D, V, f)$ 的一个优势 β - λ 下分布约简 R

第一步 计算所有的 $[x]_c, x \in U$ 及 $\underline{C}_\lambda^d(D)(x)$;

第二步 对每个 $\underline{C}_\lambda^d(D)(x) > 0$, 根据核属性的定义求出所有核属性, 并将它们加入到 R ;

第三步 对所有的 $\underline{C}_\lambda^d(D)(x) > 0$, 判断 $\underline{C}_\lambda^d(D)(x) = \underline{R}_\lambda^d(D)(x)$ 是否成立, 若成立, 转第五步;

第四步 对每个 $c \in C \setminus R$, 计算 $R \cup \{c\}_\lambda^d(D)(x)$, 记 $Q(c) = \{R \cup \{c\}_\lambda^d(D)(x) > 0; \underline{C}_\lambda^d(D)(x) > 0\}$, 取 $c_0 \in \{c_k: |Q(c_k)| = \max_{c \in C \setminus R} |Q(c)|\}$, 做 $R = R \cup \{c_0\}$, 转第三步;

第五步 对 R 中的每个非核属性 c , 对所有 $\underline{C}_\lambda^d(D)(x) > 0$ 检查 $\underline{C}_\lambda^d(D)(x) = \underline{R \setminus \{c\}}_\lambda^d(D)(x)$ 是否成立, 若成立, 则 $R = R \setminus \{c\}$;

第六步 输出 R , 即为一个优势 β - λ 下分布约简。

算法说明: 在第四步中, $Q(c)$ 是多重集合, 即元素可以重复出现。若 $Q(c) = \emptyset$, 则选择使所有 $\underline{C}_\lambda^d(D)(x) > 0$ 的对象 x 的优势类减小最快的 c ; 若 $\{c_k: |Q(c_k)| = \max_{c \in C \setminus R} |Q(c)|\}$ 中有多个元素, 则选取与所有 $\underline{C}_\lambda^d(D)(x) > 0$ 值的总误差最小者。

下面分析算法的时间复杂度。

设 $|U| = n, |C| = l$, 每个对象有 m 个决策值, 则第一

步、第二步、第四步和第五步的时间复杂度均为 n^2lm ; 第三步为 n , 故算法的总时间复杂度为 n^2lm 。

4 计算机审计风险评估算例

审计风险评估是审计风险导向的核心,也是审计风险决策的重要依据。随着信息技术的高度发展和普及运用,多层次的数据存取界面、审计线索和内容的变化使得以核查纸质账目的传统审计转变为融合会计、审计、计算机、通讯和网络技术的计算机审计。这种转变给审计风险赋予了新的含义并增添了新的内容。计算机审计是指为了信息系统的安全、可靠与有效,由独立于审计对象的计算机审计师,以第三方的客观立场对以计算机为核心的信息系统进行综合检查与评价,向计算机审计对象的最高领导提出问题与建议的一连串活动。计算机审计风险是指由于审计人员未能正确合理地运用计算机审计技术对被审计单位计算机会计信息系统运行的有效性、数据处理的合法性和正确性、最终报表的真实性和公允性进行审计,进而对被审计单位含有重要错报或漏报的财务报表发表不恰当时审计意见的风险^[11]。

与传统审计风险因素不同,计算机审计风险成因可分为 5 大要素:系统环境风险、系统控制风险、财务数据风险、审计软件风险和人员操作风险^[12]。

表 1 给出一个计算机审计风险评估模糊多目标决策信息系统,其中 $U=\{x_1, x_2, \dots, x_{10}\}$ 表示共有 10 个审计对象, $C=\{c_1, c_2, \dots, c_5\}$ 表示对象现有 5 个属性,分别表示系统环境风险、系统控制风险、财务数据风险、审计软件风险和人员操作风险,其中取值为“1”表示风险程度为轻度、“2”表示适中、“3”表示较高、“4”表示极高。 $D=\{d_1, d_2, d_3, d_4\}$ 表示有 4 个专家对这 10 个对象进行评估,对应取值为专家认为该审计案件总风险(终极风险)高的程度。为从该表中获取有用知识辅助审计风险决策,利用优势-模糊目标变精度粗糙集模型来获取规则。

表 1 计算机审计风险评估模糊决策表

U	c_1	c_2	c_3	c_4	c_5	d_1	d_2	d_3	d_4
x_1	1	2	1	1	1	0.2	0.4	0.2	0.4
x_2	1	2	2	1	1	0.4	0.2	0.4	0.2
x_3	1	1	2	2	3	0.2	0.2	0.2	0.4
x_4	1	1	1	2	3	0.2	0.2	0.4	0.2
x_5	2	2	1	1	2	0.4	0.4	0.6	0.4
x_6	2	3	1	1	1	0.4	0.6	0.6	0.4
x_7	2	2	1	3	3	0.6	0.6	0.6	0.6
x_8	3	4	3	4	3	0.6	0.8	0.8	0.8
x_9	3	4	3	3	4	0.8	0.8	0.8	0.6
x_{10}	4	3	3	3	4	0.8	0.8	0.8	0.8

假定我们关心的是总风险为“高”的模糊决策规则,取 $\beta=0.8, \lambda=0.6$ 。下面利用本文提出的算法求该模糊决策信息系统的优势 β - λ 下分布约简,并进而求得约简规则。

第一步 计算所有的 $\underline{C}_{0.8}^{0.6}(D)(x)$ 。

$$\underline{C}_{0.8}^{0.6}(D)(x_1) = \underline{C}_{0.8}^{0.6}(D)(x_2) = \underline{C}_{0.8}^{0.6}(D)(x_3) = \underline{C}_{0.8}^{0.6}(D)(x_4) = 0$$

$$\underline{C}_{0.8}^{0.6}(D)(x_5) = \underline{C}_{0.8}^{0.6}(D)(x_6) = \underline{C}_{0.8}^{0.6}(D)(x_7) = \underline{C}_{0.8}^{0.6}(D)(x_8) = 0.6$$

$$\underline{C}_{0.8}^{0.6}(D)(x_9) = 0.6 \quad \underline{C}_{0.8}^{0.6}(D)(x_{10}) = 0.8$$

第二步 计算核属性。

考虑 x_9 和 x_{10} , $\underline{C}_{0.8}^{0.6}(D)(x_{10}) = 0.8 > d_4(x_9) = 0.6, x_9 \notin$

$[x_{10}]_C$, 而 $x_9 \in [x_{10}]_{C \setminus \{c_1\}}$, 故 c_1 是核属性。由于其它条件属性都不是核属性,故 $R = \{c_1\}$ 。

第三步 由于 $\underline{R}_{0.8}^{0.6}(D)(x_5) = 0 < \underline{C}_{0.8}^{0.6}(D)(x_5) = 0.6$, 故 $R = \{c_1\}$ 还不是约简。

第四步 对 $R = \{c_1\}$, 分别计算 $Q(c_i) (2 \leq i \leq 5)$, 有 $Q(c_2) = \{0.6, 0.6, 0.6, 0.6, 0.8\}$, $Q(c_3) = \{0.6, 0.6, 0.6, 0.8\}$, $Q(c_4) = \{0.6, 0.6, 0.6, 0.6, 0.8\}$, $Q(c_5) = \{0.6, 0.6, 0.6, 0.6, 0.8\}$

于是 $R = \{c_1, c_5\}$ 。又 $\underline{R}_{0.8}^{0.6}(D)(x_6) = 0 < \underline{C}_{0.8}^{0.6}(D)(x_6)$, 故 $R = \{c_1, c_5\}$ 不是约简。对 $R = \{c_1, c_5\}$, 分别计算, 有

$Q(c_2) = \{0.6, 0.6, 0.6, 0.6, 0.6, 0.8\}$, $Q(c_3) = \{0.6, 0.6, 0.6, 0.6, 0.6, 0.8\}$, $Q(c_4) = \{0.6, 0.6, 0.6, 0.6, 0.6, 0.8\}$, 故 $R = \{c_1, c_2, c_5\}$ 。经检验, $R = \{c_1, c_2, c_5\}$ 即为一个约简。

进一步采用粗糙集理论中的值约简算法, 可得到如下阶段决策规则。

IF “ $c_1 \geq 4$ ” THEN 风险度为“高”的程度至少是 0.8, 置信度为 100%, 支持度为 10%;

IF “ $c_2 \geq 3$ ” THEN 风险度为“高”的程度至少是 0.6, 置信度为 87.5%, 支持度为 35%;

IF “ $c_5 \geq 4$ ” THEN 风险度为“高”的程度至少是 0.6, 置信度为 100%, 支持度为 20%。

这些规则对于确定审计的重点环节, 采取何种审计策略和措施有效控制审计风险, 辅助信息系统审计师作出正确的审计风险决策有着重要的指导意义。

结束语 针对具有优势关系的模糊目标信息系统的规则获取, 建立了优势-模糊目标变精度粗糙集模型, 并定义了此模型的几种知识约简; 通过定义优势下分布约简的一种启发式函数, 设计出相应的约简算法。最后将优势-模糊目标变精度粗糙集模型应用于计算机审计风险评估, 得到了较为可信的评估规则。研究结果表明, 优势-模糊目标变精度粗糙集模型用于审计风险评估是合理的, 有助于审计主体作出合理有效的审计风险决策。

参考文献

- [1] Zopounidis C, Doumpos M. Multicriteria classification and sorting methods a literature review [J]. European Journal of Operational Research, 2002(138): 229-246
- [2] 刘业政, 杨善林, 刘心报. 基于 rough set 理论的判断矩阵构造方法[J]. 系统工程学报, 2002, 17(2): 182-187
- [3] 王珏, 刘三阳, 张杰. 群决策中基于语言信息处理的一种粗糙集方法[J]. 系统工程学报, 2006, 21(1): 18-23
- [4] Huang Jih-Jeng, Ong Chorng-Shyong, Tzeng Gwo-Hshiung. Interval multidimensional scaling for group decision using rough set concept [J]. Expert Systems with Applications, 2006, 31(3): 525-530
- [5] Cong Guodong, Zhang Jinlong, Yu Beihai. A variable precision fuzzy rough group decision-making model for IT outsourcing risk evaluation in nuclear power industry [C]//Lecture notes in computer science. 2007, 4489: 960-963
- [6] Xie Gang, Zhang Jinlong, Lai K K, et al. Variable precision rough set for group decision-making: An application [J]. Approximation Reasoning, 2008, 49(2): 331-343

(下转第 241 页)

同类型的事件要素聚集在同一个类簇中。其中,从特征选择到输出特征权值矩阵属于特征加权算法(FWA)。整个 FWEAI 算法的框架如图 1 所示。

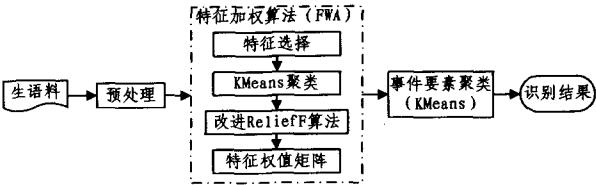


图 1 FWEAI 算法框架

3 实验结果及分析

为了评价 FWEAI 算法的性能,我们使用 java 语言实现了该算法。实验中分词和句法分析工具采用了哈工大 LTP 语言技术平台提供的模块。由于 ACE 语料不对外公开评测,本文实验语料收集了来自于中国政府网以及新浪、搜狐、新华网等关于突发事件的报道,共 8 类事件,每类事件各 100 篇,共 800 篇文档。不同事件类中的相同事件要素(比如时间要素和地点要素)视为一类要素,共包含 16 类事件要素,如表 2 所列。

表 2 实验采用的突发事件语料

事件名称	事件要素个数	语料篇数
台风	3	100
地震	3	100
交通事故	5	100
爆炸	3	100
食物中毒	5	100
禽流感	4	100
骚乱	3	100
空袭	4	100

图 2 是 FWA 算法迭代 20 次后得到的各个特征所对应的权值,其中特征 1,2 和 3 分别表示中心词 w_i 词本身、词性及句法特征;特征 4,5 和 6 分别表示 w_{i-1} 词本身、词性及句法特征;特征 7,8 和 9 分别表示 w_{i+1} 词本身、词性及句法特征;特征 10,11,12 和 13 分别表示触发词 t 本身、触发词 t 所属事件类别、 t 的词性及句法特征;特征 14 表示中心词 w_i 与触发词 t 的位置关系。可以看出,中心词 w_i 本身所具备的特征以及 w_i 与 t 的位置关系对聚类贡献比较大。

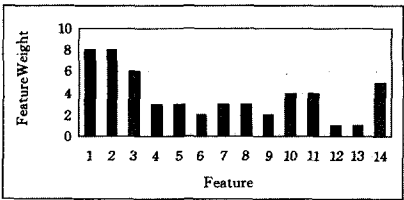


图 2 FWA 算法得到的各个特征所对应的权重

我们将 FWA 算法计算得到的权值矩阵用于 FWEAI 算法的后续事件要素聚类,并且将 FWEAI 算法与没有经过特

征加权的 KMeans 聚类算法识别事件要素进行比较。分别抽取语料的 20%,40%,60%,80%和 100%进行试验,实验结果采用准确率 P (Precision)评价。实验结果如表 3 所列。

表 3 FWEAI 算法与 KMeans 算法识别事件要素结果比较

语料比例 (%)	Precision (%)	
	KMeans Only	FWEAI
20	50.32	67.05
40	51.16	66.90
60	52.39	67.29
80	50.55	67.65
100	52.69	67.55
平均	51.42	67.28

从表 3 可以看出,仅采用 KMeans 算法不对特征进行加权,在不同比例语料上的平均准确率为 51.42%,而 FWEAI 算法在不同比例语料上的平均准确率提高到 67.28%。因此,采用特征加权算法根据不同特征对聚类的不同贡献分配相应的权重,可以明显提高事件要素识别的准确率。

结束语 本文分析了事件抽取和事件要素识别的研究现状。针对目前研究中的不足,提出了一种基于特征加权的事件要素识别算法(FWEAI)。算法采用 FWA 算法计算权值矩阵,根据不同特征对聚类的不同贡献,分配相应的权值,然后采用 KMeans 算法对事件要素进行聚类。在生语料上的实验表明,FWEAI 算法可以提高事件要素识别的准确率。由于 FWEAI 算法采用的是无监督的聚类学习,聚类的结果需要人工加以判断。用半监督的聚类算法进一步提高聚类准确率,并且通过标记对象自动推导聚类的结果,是下一步的研究方向。

参考文献

- [1] Doddington G, Mitchell A, Przybocki M, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation[C]//Proc. LREC. 2004;837-840
- [2] Ahn D. The stages of event extraction[C]//Proceedings of the COLING-ACL 2006 Workshop on Annotating and Reasoning About Time and Events. 2006;1-8
- [3] 赵妍妍,秦兵,车万翔,等. 中文事件抽取技术研究[J]. 中文信息学报,2008,22(1):3-8
- [4] Tan H, Zhao T, Zheng J. Identification of Chinese Event and Their Argument Roles[C]//Computer and Information Technology Workshops, 2008 CIT Workshops 2008 IEEE 8th International Conference on. 2008;14-19
- [5] Wetttschereck D, Aha D. Weighting features [M]. Springer, 1995; 347-358
- [6] 李洁,高新波,焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报,2006(1)
- [7] Kononenko I. Estimating attributes; analysis and extensions of RELIEF[M]. New York; Springer-Verlag, 1994; 171-182
- [8] Robnik-Sikonja M, Kononenko I. Theoretical and Empirical Analysis of Relief and RRelief[J]. Machine Learning, 2003, 53 (1/2):23-69

(上接第 229 页)

- [7] Xie Gang, Zhang Jinlong, Lai K K. A group decision-making model of risk evasion in software project bidding based on VPRS [C]//Lecture notes in computer science. 2005,3642:738-742
- [8] 毕文杰,陈晓红. 一种基于可变精度粗糙集的群体分类决策方法[J]. 系统工程,2007,25(8):94-97
- [9] Zhang Mei, Wu Wei-zhi. Knowledge reductions in information systems with fuzzy decisions [J]. 工程数学学报,2003,20(2):

53-58

- [10] 袁修久,何华灿. 优势关系下模糊目标信息系统约简的辨识矩阵[J]. 空军工程大学学报:自然科学版,2006,7(2):81-84
- [11] 林琳. 计算机审计风险分析及防范对策[J]. 中国管理信息化:综合版,2005(7):44-46
- [12] 蒋锡元. 会计电算化条件下的审计风险与防范措施[J]. 中国管理信息化,2006(5)