一种高效的不协调决策表约简算法*

李财莲[1],滕书华[1],孙即祥[1],康耀红[2]
(1. 国防科技大学 电子科学与工程学院,湖南 长沙 410073;
2. 海南大学 信息科学技术学院,海南 海口 570228)

摘 要:目前,不协调决策表的分布约简、最大分布约简和分配约简算法复杂度较高,不适合处理大数据集。在分析已有算法基础上,分析了基于相对可区分度的属性重要性度量的性质,解决了正域度量属性重要性的缺陷。针对不协调决策表,给出了多种简化协调决策表的定义,从而大大缩减了约简的实例数。以相对可区分度为启发函数构造了一种高效完备的不协调决策表约简算法。理论分析和实验结果表明,该约简算法解决了现有算法在复杂度和属性重要性度量上的缺陷,适合处理不协调的大数据集。

关键词:粗糙集;属性约简;不协调决策表;属性重要性
中图分类号:TP18 **文献标志码**:A **文章编号**:1001-2486(2013)01-0108-07

An efficient attribute reduction algorithm in inconsistent decision tables

LI Cailian[1], *TENG Shuhua*[1], *SUN Jixiang*[1], *KANG Yaohong*[2]
(1. College of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China;
2. Information Science Technology College, Hainan University, Haikou 570228, China)

Abstract: Existing algorithms of distribution reduct, maximum distribution reduct and assignment reduct for inconsistent decision tables are inefficient, which are not suitable for large data sets. A measurement of attribute importance based on the relative discernibility degree was presented firstly, which overcomes the shortcoming of positive domain in measuring the importance of attributes. Then, in order to simplify the decision table, some kinds of simplified consistent decision tables were defined. In the end, an efficient attribute reduction algorithm was designed based on the relative discernibility degree. Theoretical analysis and experimental results show the effectiveness and practicalbility of this algorithm on the large inconsistent data sets.

Key words: rough set; attribute reduction; inconsistent decision table; attribute importance

知识约简是粗糙集理论核心问题之一[1]。现已证明,寻找信息系统所有约简或最优约简是NP难问题。因此许多学者从不同角度对启发式约简算法做了深入研究,提出了基于正域[2-3]、属性频率[4]和条件熵[5-6]的知识约简算法。在管理决策中,大量面对的是不协调决策表,要想从复杂的不协调信息系统中获取简洁的不确定性命题,就必须对不协调决策表进行约简,因此研究不协调决策系统的约简更具现实意义[1]。

不协调决策表中由于不协调对象的存在,使得基于正域和条件熵的约简算法无法等价表示知识约简[7]。为此,张等[1]发展了Kryszkiewicz的思想[8],提出了最大分布约简的概念,并给出了基于区分矩阵的不协调决策表约简算法。基于区分矩阵方法直观、易于理解,但时间复杂度一般为$O(|C|^2|U|^2)$(其中$|C|$为属性个数,$|U|$为对象个数),不适合对大数据集处理。据此文献[9]利用基于正域的启发式算法[2]求得不协调决策表的多种约简,将约简算法的复杂度降为$O(|C|^2|U|\log|U|)$。但是,正域作为属性约简的启发式信息,并不能准确度量各个属性的重要性[10],在大多数情况下不能得到最小约简,不适合处理具有大量不协调对象和冗余属性的大数据集。

为克服现有算法的缺陷,本文首先利用文献[10]中的属性重要性度量解决了正域度量属性重要性的缺陷;然后给出了简化协调决策表的概念,利用简化协调决策表和相对可区分度设计了一种高效的不协调决策表启发式约简算法。最后通过实验验证了本文算法的有效性。

1 粗糙集基本知识

信息系统$S=(U,A)$中,$Q,P\subseteq A$,有以下

* **收稿日期**:2011-09-30
基金项目:国家自然科学基金资助项目(40901216);中国博士后科学基金项目(2012M512168)
作者简介:李财莲(1973—),女,湖南涟源人,工程师,博士,E-mail:gsstsh@sohu.com;
孙即祥(通信作者),男,教授,博士,博士生导师,E-mail:prmvnudt@sohu.com

定义:

(1)属性集 P 的不可区分关系 IND(P)定义为[10]

$$\mathrm{IND}(P)=\{(u_k,u_t)\in U\times U\mid \forall a_i\in P,\ f(u_k,a_i)=f(u_t,a_i),\ 1\leqslant k<t\leqslant|U|\}$$

如果$(u_i,u_j)\in\mathrm{IND}(P)$,则称 u_i 和 u_j 是 P 不可区分的。不可区分关系 IND(P)在 U 上导出的划分记为$\frac{U}{P}=\{[u_i]_P\mid u_i\in U\}$,其中$[u_i]_P=\{u_j\in U\mid(u_i,u_j)\in\mathrm{IND}(P)\}$为包含 u_i 的 P 等价类。

属性集 P 的可区分关系 DIS(P)定义为[10]:

$$\mathrm{DIS}(P)=\{(u_k,u_t)\in U\times U\mid \exists a_i\in P,\ f(u_k,a_i)\neq f(u_t,a_i),\ 1\leqslant k<t\leqslant|U|\}$$

若$(u_i,u_j)\in\mathrm{DIS}(P)$,则称 u_i 和 u_j 是 P 可区分的。对$\forall(u_i,u_j)\in U\times U$,$(u_i,u_j)$要么是 P 可区分,要么是 P 不可区分,因此可区分关系与不可区分关系具有互补性。

(2)令 $U/P=\{P_1,P_2,\cdots,P_m\}$,$U/Q=\{Q_1,Q_2,\cdots,Q_n\}$,则 $P\leq Q$ 表示对$\forall P_i\in U/P$,$\exists Q_j\in U/Q$,使得 $P_i\subseteq Q_j$,这意味着 P 的划分比 Q 精细或 Q 的划分比 P 粗糙。

(3)决策表 $S=(U,C,D)$ 中,若 $U/C\subseteq U/D$,即 $\mathrm{IND}(C)\subseteq\mathrm{IND}(D)$,则称决策表是协调的,否则称决策表是不协调的。

(4)决策表 $S=(U,C,D)$,$u_i\in U$,$Q\subseteq C$。记 $U/D=\{D_1,D_2,\cdots,D_m\}$,$P(D_j/[u_i]_Q)=\frac{|D_j\cap[u_i]_Q|}{|[u_i]_Q|}$,$u_Q(u_i)=(P(D_1/[u_i]_Q),P(D_2/[u_i]_Q),\cdots,P(D_m/[u_i]_Q))$。显然 $u_Q(u_i)$ 是 U/D 上的概率分布函数。进一步记 $\gamma_Q(u_i)=\max\limits_{j\leqslant m}\{P(D_j/[u_i]_Q)\}$,$\delta_Q(u_i)=\{D_j:[u_i]_Q\cap D_j\neq\varnothing\}$。则有以下定义:

①若$\forall u_i\in U,u_Q(u_i)=u_C(u_i)$,则称 Q 是分布协调集. 若 Q 是分布协调集,且 Q 的任何真子集不是分布协调集,则称 Q 为分布约简。

②若$\forall u_i\in U,\gamma_Q(u_i)=\gamma_C(u_i)$,则称 Q 是最大分布协调集。若 Q 是最大分布协调集,且 Q 的任何真子集不是最大分布协调集,则称 Q 为最大分布约简。

③若$\forall u_i\in U,\delta_Q(u_i)=\delta_C(u_i)$,则称 Q 是分配协调集。若 Q 是分配协调集,且 Q 的任何真子集不是分配协调集,则称 Q 为分配约简。

2　可区分度和相对可区分度

依据知识是区分对象能力的思想,论域中所有对象,如果两两间都能被区分,那么就具有较多的知识;反之,如果论域中所有对象都属于同一个等价类,不能将任何一个对象与其他对象区分开来,这时具有的知识最少。基于这种观点,文献[10]将属性区分对象数目多少作为知识量的度量,给出可区分度、相对可区分度的概念,并对它们的性质进行了分析,下面直接给出相关定义。

定义 1　信息系统 $S=(U,A)$ 中,$P\subseteq A$。$U/P=\{P_1,P_2,\cdots,P_m\}$,知识 P 的可区分度为:

$$|\mathrm{DIS}(P)|=(|U|^2-\sum_{i=1}^{m}|P_i|^2)/2 \quad (1)$$

可区分度 $|\mathrm{DIS}(P)|$是 $\mathrm{DIS}(P)$ 中元素对的个数。知识 P 能区分的对象对越多,则可区分度越大。因此 $|\mathrm{DIS}(P)|$是对知识 P 可区分能力大小的度量。

定义 2　信息系统 $S=(U,A)$ 中,$Q,P\subseteq A$。令 $U/Q=\{Q_1,Q_2,\cdots,Q_n\}$,$U/(P\cup Q)=\{H_1,H_1,\cdots,H_k\}$,知识 P 对于知识 Q 的相对可区分度 $|\mathrm{DIS}(P/Q)|$定义为:

$$|\mathrm{DIS}(P/Q)|=(\sum_{i=1}^{n}|Q_i|^2-\sum_{j=1}^{k}|H_j|^2)/2 \quad (2)$$

相对可区分度 $|\mathrm{DIS}(P/Q)|$是论域中知识 P 能区分而知识 Q 不能区分的对象对的个数,是对知识 P 并入知识 Q 后给知识 Q 增加的区分能力的度量。因此 $|\mathrm{DIS}(P/Q)|$度量了知识 P 和知识 Q 间的差异性。

定理 1　信息系统 $S=(U,A)$ 中,$D,P,Q\subseteq A$。如果

$P\leq Q$,则 $|\mathrm{DIS}(D/Q)|\geqslant|\mathrm{DIS}(D/P)|$。

定理 1 表明在知识 D 划分粒度不变的情况下,相对可区分度 $|\mathrm{DIS}(D/Q)|$随着知识 Q 划分粒度的减小而单调递减。

性质 1　信息系统 $S=(U,A)$ 中,$P,Q\subseteq A$。$0\leqslant\left|\mathrm{DIS}(\frac{P}{Q})\right|\leqslant|\mathrm{DIS}(P)|$,当且仅当 $\mathrm{DIS}(P)\subseteq\mathrm{DIS}(Q)$ 时左等式成立,当且仅当 $\mathrm{DIS}(P)\cap\mathrm{DIS}(Q)=\varnothing$ 时右等式成立。

文献[10]在协调决策表中,利用相对可区分度构造的属性重要性测度解决了正域度量属性重要性的缺陷。与此类似,本文将此方法引入不协调决策表,进而解决不协调决策表中正域度量属性重要性的缺陷。

3　高效的不协调决策表约简算法

基于正域的约简算法中,Xu[3] 通过构造简化决策表,降低了约简算法时间复杂度。但此快速

约简算法并不能处理多种不协调决策表的约简。下面借鉴简化决策表思想,定义了三种简化协调决策表,进而给出原决策表及其简化协调决策表约简之间的关系,在此基础上提出了不协调决策表的高效约简算法。

3.1 简化协调决策表

在不协调决策表 $S=(U,C,D,V,f)$ 中,令 $U/C=\{[u_1']_C,[u_2']_C,\cdots,[u_m']_C\}$, $U'=\{u_1',u_2',\cdots,u_m'\}$,下面给出 S 对应的三种简化协调决策表的定义。

定义3 称 $S_u'=(U',C,D_u',V,f)$ 为 S 对应的简化分布协调决策表。其中 D_u' 为分布决策属性,满足 $\mathrm{IND}(D_u')=\{(u_i',u_j')\in U'\times U'\mid u_C(u_i')=u_C(u_j')\}$。令 $U'/D_u'=\{[u_1'']_{D_u'},[u_2'']_{D_u'},\cdots,[u_l'']_{D_u'}\}$,对 $\forall u_i'\in U'$,若 $u_i'\in[u''_k]_{D_u'}$,$1\leqslant k\leqslant l$,则 D_u' 在对象 u_i' 上的值为 $f(u_i',D_u')=k$。

定义4 称 $S_\gamma'=(U',C,D_\gamma',V,f)$ 为 S 对应的简化最大分布决协调策表。其中 D_γ' 为最大分布决策属性,满足 $\mathrm{IND}(D_\gamma')=\{(u_i',u_j')\in U'\times U'\mid\gamma_C(u_i')=\gamma_C(u_j')\}$。令 $U'/D_\gamma'=\{[u_1'']_{D_\gamma'},[u_2'']_{D_\gamma'},\cdots,[u_l'']_{D_\gamma'}\}$,对 $\forall u_i'\in U'$,若 $u_i'\in[u_k'']_{D_\gamma'}$,$1\leqslant k\leqslant l$,则 D_γ' 在对象 u_i' 上的值为 $f(u_i',D_\gamma')=k$。

定义5 称 $S_\delta'=(U',C,D_\delta',V,f)$ 为 S 对应的简化分配协调决策表。其中 D_δ' 为分配决策属性,满足 $\mathrm{IND}(D_\delta')=\{(u_i',u_j')\in U'\times U'\mid\delta_C(u_i')=\delta_C(u_j')\}$。令 $U'/D_\delta'=\{[u_1'']_{D_\delta'},[u_2'']_{D_\delta'},\cdots,[u_l'']_{D_\delta'}\}$,对 $\forall u_i'\in U'$,若 $u_i'\in[u_k'']_{D_\delta'}$,$1\leqslant k\leqslant l$,则 D_δ' 在对象 u_i' 上的值为 $f(u_i',D_\delta')=k$。

三种简化协调决策表分别简写为 $S_u'=(U',C,D_u')$, $S_\gamma'=(U',C,D_\gamma')$ 和 $S_\delta'=(U',C,D_\delta')$。显然在简化协调决策表中的条件属性和原决策表中的条件属性相同。

推论1 不协调决策表 $S=(U,C,D)$ 对应的简化分布协调决策表、简化最大分布协调决策表和简化分配协调决策表分别为 $S_u'=(U',C,D_u')$, $S_\gamma'=(U',C,D_\gamma')$ 和 $S_\delta'=(U',C,D_\delta')$。则

(1) $\mathrm{DIS}(D_u')=\{(u_i',u_j')\in U'\times U'\mid u_C(u_i')\neq u_C(u_j')\}$;

(2) $\mathrm{DIS}(D_\gamma')=\{(u_i',u_j')\in U'\times U'\mid\gamma_C(u_i')\neq\gamma_C(u_j')\}$;

(3) $\mathrm{DIS}(D_\delta')=\{(u_i',u_j')\in U'\times U'\mid\delta_C(u_i')\neq\delta_C(u_j')\}$。

3.2 不协调决策表的约简

定理2 不协调决策表 $S=(U,C,D)$ 中,$Q\subseteq C$。$U/Q=\{Q_1,Q_2,\cdots,Q_n\}$。

(1) Q 为 S 的分布协调集,当且仅当 $\forall u_k,u_l\in Q_i$, $2\leqslant|Q_i/C|$,$1\leqslant i\leqslant n$,满足 $u_C(u_l)=u_C(u_k)$;

(2) Q 为 S 的最大分布协调集,当且仅当 $\forall u_k,u_l\in Q_i$,$2\leqslant|Q_i/C|$,$1\leqslant i\leqslant n$,满足 $\gamma_C(u_l)=\gamma_C(u_k)$;

(3) Q 为 S 的分配协调集,当且仅当 $\forall u_k,u_l\in Q_i$, $2\leqslant|Q_i/C|$,$1\leqslant i\leqslant n$,满足 $\delta_C(u_l)=\delta_C(u_k)$。

证明 充分性:由分布协调集的定义,如果 Q 为 S 的分布协调集,则对 $\forall u_k,u_l\in U$,$u_Q(u_k)=u_C(u_k)$, $u_Q(u_l)=u_C(u_l)$。而对 $\forall u_k,u_l\in Q_i$,$2\leqslant|Q_i/C|$,$1\leqslant i\leqslant n$,显然有 $u_Q(u_k)=u_Q(u_l)$。因此 $\forall u_k,u_l\in Q_i$,$2\leqslant|Q_i/C|$,$1\leqslant i\leqslant n$, 满足 $u_C(u_l)=u_C(u_k)$。

必要性:如果对 $\forall u_k,u_l\in Q_i$,$2\leqslant|Q_i/C|$,$1\leqslant i\leqslant n$,满足 $u_C(u_l)=u_C(u_k)$,要证 Q 为 S 的分布协调集,只需证 $\forall u_k\in Q_i$,$2\leqslant|Q_i/C|$,$u_Q(u_k)=u_C(u_k)$(如果 $|Q_i/C|=1$,即 $[u_k]_C=[u_k]_Q$,显然有 $u_Q(u_k)=u_C(u_k)$)。下面证明之。

$\forall u_k\in Q_i$,不妨设 $|Q_i/C|=2$,即令 $Q_i=[u_l]_C\cup[u_k]_C$。对 $\forall D_j\subseteq U/D$,令 $|[u_l]_C|=f$, $|[u_k]_C|=y$, $|[u_l]_C\cap D_j|=e$, $|[u_k]_C\cap D_j|=x$, 则 $P(D_j/[u_l]_C)=\dfrac{e}{f}$, $P(D_j/[u_k]_C)=\dfrac{x}{y}$。因为 $u_C(u_l)=u_C(u_k)$,则 $\dfrac{e}{f}=\dfrac{x}{y}$,从而 $xf=ey$。因 $[u_l]_C\cap[u_k]_C=\varnothing$,所以 $[u_k]_Q$ 在 D_j 上的隶属度为

$$P(D_j/[u_k]_Q)=\frac{|Q_i\cap D_j|}{|Q_i|}=\frac{|([u_l]_C\cup[u_k]_C)\cap D_j|}{|[u_l]_C\cup[u_k]_C|}=\frac{x+e}{y+f}$$

因为 $\dfrac{x+e}{y+f}-\dfrac{x}{y}=\dfrac{ey-xf}{y^2+yf}=0$,即 $\dfrac{x+e}{y+f}=\dfrac{x}{y}$,所以 $P(D_j/[u_k]_Q)=P(D_j/[u_k]_C)$。由 D_j 的任意性得 $u_Q(u_k)=u_C(u_k)$, 即对 $\forall u_k\in Q_i$, $|Q_i/C|=2$ 时有 $u_Q(u_k)=u_C(u_k)$。$2<|Q_i/C|$ 时的证明与此类似。

(2)和(3)的证明与(1)类似。

定理2给出了一种新的判定属性子集是分布协调集、最大分布协调集和分配协调集的方法。由此可得以下推论:

引理1 不协调决策表 $S=(U,C,D)$ 中,$Q\subseteq C$。令 $U/C=\{C_1,C_2,\cdots,C_m\}=\{[u_1']_C,[u_2']_C,\cdots,[u_m']_C\}$, $U'=\{u_1',u_2',\cdots,u_m'\}$, $U/Q=\{Q_1,Q_2,\cdots,Q_n\}$。则

(1) $|U'/Q|=n$;

(2) $\forall Q_i\in U/Q$,存在唯一的 $Q_i'\in U'/Q$ 使得 $Q_i=\{[u_k']_C\mid u_k'\in Q_i'\}$。

证明 (1)由 $Q\subseteq C$ 可知 $\forall C_j\in U/C$, $\exists Q_i\in U/Q$,使得 $C_j\subseteq Q_i$,即存在 $\{1,2,\cdots,m\}$ 的一个划

分 $E=\{E_1,\cdots,E_n\}$ 使 $Q_i=\bigcup_{j\in E_i}C_j=\bigcup_{j\in E_i}[u'_j]_C$,$1\leqslant i\leqslant n$。显然存在唯一的 $Q'_i\in U'/Q$ 使得 $Q'_i=\bigcup_{j\in E_i}u'_j$,即 Q'_i和 E_i 一一对应。由 $|E|=n$ 可得 $|U'/Q|=n$。

(2)由(1)的证明可知 $Q'_i\in U'/Q$ 和 E_i 是一一对应的,且 $Q_i=\bigcup_{j\in E_i}C_j$,$Q'_i=\bigcup_{j\in E_i}u'_j$,$C_j=[u'_j]_C$,因而存在唯一的 $Q'_i\in U'/Q$ 使得 $Q_i=\{[u'_k]_C\mid u'_k\in Q'_i\}$。

推论 2　$S'_u=(U',C,D'_u)$,$S'_\gamma=(U',C,D'_\gamma)$和 $S'_\delta=(U',C,D'_\delta)$分别为不协调决策表 $S=(U,C,D)$对应的简化分布协调决策表、简化最大分布协调决策表和简化分配协调决策表,$Q\subseteq C$。令 $U/Q=\{Q_1,Q_2,\cdots,Q_n\}$,$U'/Q=\{Q'_1,Q'_2,\cdots,Q'_n\}$,则:

(1) Q 为 S 的分布协调集$\Leftrightarrow|\mathrm{DIS}(D'_u/Q)|=0$;

(2) Q 为 S 的最大分布协调集$\Leftrightarrow$ $|\mathrm{DIS}(D'_\gamma/Q)|=0$;

(3) Q 为 S 的分配协调集$\Leftrightarrow|\mathrm{DIS}(D'_\delta/Q)|=0$。

证明　(1)充分性:对 $\forall(u'_k,u'_l)\in\mathrm{IND}(Q)$,则 $u'_k,u'_l\in Q'_i$,$Q'_i\in U'/Q$。由引理 1 可知$[u'_k]_C\cup[u'_l]_C\in Q_i$,$1\leqslant i\leqslant n$。因 Q 为 S 的分布协调集,由定理 2 得 $u_C(u'_l)=u_C(u'_k)$。根据分布协调决策属性的定义可知$(u'_k,u'_l)\in\mathrm{IND}(D'_u)$,由$(u'_k,u'_l)$的任意性得 $\mathrm{IND}(Q)\subseteq\mathrm{IND}(D'_u)$,即 $\mathrm{DIS}(D'_u)\subseteq\mathrm{DIS}(Q)$。由性质 1 可知$|\mathrm{DIS}(D'_u/Q)|=0$。

必要性:$|\mathrm{DIS}(D'_u/Q)|=0$,由性质 1 可知 $\mathrm{DIS}(D'_u)\subseteq\mathrm{DIS}(Q)$,即 $\mathrm{IND}(Q)\subseteq\mathrm{IND}(D'_u)$。对 $\forall Q'_i\in U'/Q$,不妨设 $Q'_i=\{u'_k,u'_l\}$,则 $u'_k,u'_l\in[u'_k]_{D_u}$。根据分布协调决策属性的定义可知 $u_C(u'_k)=u_C(u'_l)$。由引理 1 可知 $Q_i=[u'_k]_C\cup[u'_l]_C$。显然对于 $\forall u_g\in[u'_k]_C$ 或 $\forall u_g\in[u'_l]_C$,有 $u_C(u_g)=u_C(u'_k)$。即对 $\forall u_k,u_l\in Q_i$,$|Q_i/C|=2$,$1\leqslant i\leqslant n$ 时,有 $u_C(u_l)=u_C(u_k)$。同理可证当 $2<|Q'_i|$,即 $2<|Q_i/C|$时也满足:$\forall u_k,u_l\in Q_i$,$1\leqslant i\leqslant n$ 时,有 $u_C(u_l)=u_C(u_k)$。由定理 2 可知 Q 为 S 的分布协调集。

(2)和(3)的证明过程与(1)类似。

推论 3　不协调决策表 $S=(U,C,D)$对应的简化分布协调决策表、简化最大分布协调决策表和简化分配协调决策表分别为 $S'_u=(U',C,D'_u)$,$S'_\gamma=(U',C,D'_\gamma)$和 $S'_\delta=(U',C,D'_\delta)$.

(1)如果 $Q\subseteq C$ 为 S 的分布协调集,则 $|\mathrm{DIS}(D'_u)|\leqslant|\mathrm{DIS}(Q)|$;

(2)如果 $Q\subseteq C$ 为 S 的最大分布协调集,则 $|\mathrm{DIS}(D'_\gamma)|\leqslant|\mathrm{DIS}(Q)|$;

(3)如果 $Q\subseteq C$ 为 S 的分配协调集,则 $|\mathrm{DIS}(D'_\delta)|\leqslant|\mathrm{DIS}(Q)|$。

证明　(1)因 Q 为 S 的分布协调集,则由推论 2 得 $|\mathrm{DIS}(D'_u/Q)|=0$。又由性质 1 可得 $\mathrm{DIS}(D'_u)\subseteq\mathrm{DIS}(Q)$。根据可区分度的意义可得 $|\mathrm{DIS}(D'_u)|\leqslant|\mathrm{DIS}(Q)|$。

(2)和(3)的证明过程与(1)类似。

由推论 2 和推论 3 可知不协调决策表的分布协调集至少能区分 $\mathrm{DIS}(D'_u)$中的对象对,最大分布协调集至少能区分 $\mathrm{DIS}(D'_\gamma)$中的对象对,而分配协调集至少能区分 $\mathrm{DIS}(D'_\sigma)$中的对象对。因此若 $0<|\mathrm{DIS}(D'_u/Q)|$ $(0<|\mathrm{DIS}(D'_\gamma/Q)|$或 $0<|\mathrm{DIS}(D'_\sigma/Q)|)$,即当知识 Q 不能完全区分 $\mathrm{DIS}(D'_u)$($\mathrm{DIS}(D'_\gamma)$或 $\mathrm{DIS}(D'_\sigma)$)中的对象对时,则 Q 不是分布协调集(最大分布或分配协调集)。推论 2 给出了判断属性子集是否为协调集的另一种方法,由此可进一步得到不协调决策表约简定理如下:

定理 3　(基于相对可区分度的约简)不协调决策表 $S=(U,C,D)$对应的简化分布协调决策表、简化最大分布协调决策表和简化分配协调决策表分别为 $S'_u=(U',C,D'_u)$,$S'_\gamma=(U',C,D'_\gamma)$和 $S'_\delta=(U',C,D'_\delta)$,$Q\subseteq C$。则:

(1)Q 为 S 的分布约简当且仅当对 $\forall q\in Q$ 满足 $|\mathrm{DIS}(D'_u/\{Q-q\})|>0$ 且 $|\mathrm{DIS}(D'_u/Q)|=0$;

(2)Q 为 S 的最大分布约简当且仅当对 $\forall q\in Q$ 满足 $|\mathrm{DIS}(D'_\gamma/\{Q-q\})|>0$ 且 $|\mathrm{DIS}(D'_\gamma/Q)|=0$;

(3)Q 为 S 的分配约简当且仅当对 $\forall q\in Q$ 满足 $|\mathrm{DIS}(D'_\delta/\{Q-q\})|>0$ 且 $|\mathrm{DIS}(D'_\delta/Q)|=0$。

定理 3 的证明由推论 2,推论 3 和不协调决策表约简定义可得。定理 3 给出了一种计算不协调决策表分布约简、最大分布约简和分配约简的新方法。下面根据相对可区分度给出一种新的属性重要性测度。

定义 6　不协调决策表 $S=(U,C,D)$对应的简化协调决策表为 $S'_t=(U',C,D'_t)$,其中 $t\in\{u,\gamma,\delta\}$。令 $Q\subseteq C$,$a_i\in\{C-Q\}$。则属性 a_i 的重要性 $SGF(a_i,Q,D)$定义为:

$$SGF(a_i,Q,D'_t)=\frac{|\mathrm{DIS}(D'_t/Q)|-|\mathrm{DIS}(D'_t/(Q\cup\{a_i\}))|}{|\mathrm{DIS}(D'_t)|}\tag{3}$$

若 $Q=\varnothing$,则 $|\mathrm{DIS}(D'_t/Q)|=|\mathrm{DIS}(D'_t)|$。$SGF(a_i,Q,D'_t)$越大,说明在已知 Q 的条件下 a_i 对 D_t 越重要。约简过程中,对 $\forall b\in\{C-Q-\{a_i\}\}$,若 $SGF(b,Q,D'_t)\leqslant SGF(a_i,Q,D'_t)$,则将属性 a_i 加入 Q 中作为约简元素。这样就可保证重要的属性首先被加入到约简中,最终得到含属性较少且都比较重要的属性约简集。

3.3 一种高效的不协调决策表约简算法

不协调决策表约简首先要求得简化协调决策表 $S'_t=(U',C,D'_t)$, $t\in\{u,\gamma,\delta\}$,然后依次选取使 $SGF(a_i,Q,D'_t)$ 最大的属性 a_i 放到所求的属性集合 Q 中,直到满足约简的条件为止。下面给出约简算法过程。

算法 RDDBARIDT (relative discernibility degree-based algorithm for reduction of inconsistent decision table):

输入:不协调决策表 $S=(U,C,D)$。

输出:该决策表的一个相对约简 Q_t, $t\in\{u,\gamma,\delta\}$。

step 1 采用基数排序法[3]求得 $U/C=\{C_1,C_2,\cdots,C_m\}$, $U/D=\{D_1,D_2,\cdots,D_n\}$;

step 2 对 $C_i(1\leqslant i\leqslant m)$ 中的对象按照决策属性值的大小排序,从头到尾扫描排序后的 C_i,记录 C_i 中对象在 $D_k(1\leqslant k\leqslant n)$ 上的取值个数 g_{ik},进而求得 C_i 在 D_k 上的隶属度为 $g_{ik}/|C_i|$;

step 3 任选 C_i 中的某个元素(不妨选择第一个元素)作为 u'_i,令 $U'=\{u'_1,u'_2,\cdots,u'_m\}$,显然 $[u'_i]_C$ 在 D_k 上的隶属度为 $g_{ik}/|C_i|$;

step 4 对 $\forall u'_i\in U'$,根据 $g_{ik}/|C_i|$ 求 $t_C(u'_i)$;

step 5 令 D'_t 为新的决策属性,D'_t 满足 $IND(D'_t)=\{(u'_i,u'_j)\in U'\times U'\mid t_C(u'_i)=t_C(u'_j)\}$,令 $U'/D'_t=\{D'_1,D'_2,\cdots,D'_L\}$, $\forall u'_i\in D'_j$, $1\leqslant j\leqslant L$, $f(u'_i,D'_t)=j$,则求得简化协调决策表为 $S'_t=(U',C,D'_t)$;

step 6 令 $Q=\varnothing$, $T=C$;

step 7 对 $\forall a_i\in T$,计算 $SGF(a_i,Q,D'_t)$;

若 $SGF(a_k,Q,D'_t)=\max\limits_i SGF(a_i,Q,D'_t)$(若同时有多个属性达到最大值,则从中选取可区分度最小的属性作为 a_k),令 $Q=Q\cup a_k$, $T=T-\{a_k\}$;

step 8 若 $|\mathrm{DIS}(D'_t/Q)|=0$,则转 step 9,否则转 step 7;

step 9 若 Q 中元素数目大于1,则转 step10,否则 Q 即为所求约简;

step 10 从后向前依次对 Q 中每个属性 q 进行判断:若 $|\mathrm{DIS}(D'_t/\{Q-\{q\}\})|=0$,令 $Q=Q-\{q\}$;

step 11 最终求得的 Q 即为所求约简。

算法 RDDBARIDT 通过前向添加搜索策略保证了约简结果的紧凑性,step 6 保证了知识约简是完备的。

3.4 算法复杂度分析

第1~5步计算简化协调决策表的时间复杂度为 $O(|U||C|)$;由性质3可知计算单个属性相对可区分度时间复杂度为 $O(|U'|)$,因此算法第7~9步最差的时间复杂度为 $O(|C|^2|U'|)$;同理算法第10、11步最差时间复杂度为 $O(|Q|^2|U'|)$。由于 $|U'|=|U/C|$,因此本文约简算法的最差时间复杂度为 $\max(O(|U/C||U/D|),O(|C|^2|U/C|))$,通常情况下 $|U/D|$ 相对较小,因此基于相对可区分度的约简算法最坏时间复杂度为 $O(|C|^2|U/C|)$。

4 实验结果及分析

为验证本文约简算法的有效性,我们利用UCI机器学习数据库(http://www.ics.uci.edu/~mlearn/ MLRe-pository.html)中的数据 Votes 和 Dermatology 构造了10个不协调数据,如表1所示,其中"Votes×5"表示 Votes 中每个对象重复5次,构造的新数据的决策属性值分为五类,由随机函数随机赋值。其他数据与此同。在 Windows XP, CPU2.4GHz, RAM512MB, Matlab 上进行了编程实验,数据中的缺失值被作为一种新的属性值来处理。

文献[12]已证明基于信息熵的属性约简[5]与分布约简等价。因此本文选用文献[5]中 CEBARKNC 算法(简称为算法1)、文献[9]中基于正域的约简算法(简称为算法2)和本文中 RDDBARIDT 算法(简称为算法3)进行比较。三种约简算法均采用基数排序法求划分。由于算法1对不协调数据集仅能求得分布约简,因此表1中仅包含算法1求得的分布约简。表1中,$|U|$ 表示数据集的实例数,$|C|$ 表示原始条件属性数;带有下标 u,γ 和 δ 的 $|Core|$ 和 $|Q|$ 分别表示分布约简、最大分布约简和分配约简下的核属性个数和约简后的条件属性个数;表中加黑数字表示求得的约简是最优约简。图1给出了三种约简算法在不协调数据集(表1)上的执行时间。

从表1和图1可以看出:

(1)对于表1中的不协调数据集,由于 Data1~Data5 的核属性和约简结果几乎相同,因而三种算法基本都能求得最优约简(Data3 的分配约简的核属性和约简结果差别较大,因而算法2没有求得最优分配约简)。但数据集 Data6~Data10 的核属性和约简结果差别很大,从而使得算法1均没有求得最优分布约简,算法2在 Data7 上的分配约简、在 Data8 上的最大分布约简都不是最优的,算法3在 Data9 上没有求得最优分布约简。总体上,算法3的约简结果是最

优的,在多数数据集上能找到最优约简;算法 2 的约简结果其次。

(2)图 1 中,算法 3 的执行时间比算法 2 快了一个数量级以上,比算法 1 快了近一个数量级。这是由于算法 3 通过简化协调决策表去除了大量不协调对象,缩减了约简的实例数;而算法 1 对于冗余属性较少的数据集 Data1 ~ Data3,和算法 2 的执行时间相差不大,随着不协调对象和相对冗余属性的增多(如数据 Data6 ~ Data10),算法 1 执行时间相对于算法 2 的优势更加明显,这是由于算法 1 和算法 2 采取的搜索策略不同导致的。

以上实验表明,本文算法是可靠高效的,尤其适用于有大量冗余属性和不协调数据集的约简。

表 1　不协调数据集约简结果

Tab. 1　The reduct of inconsistent data sets

	Data	$\lvert U\rvert$	$\lvert C\rvert$	$\lvert Core\rvert_u$	$\lvert Core\rvert_\gamma$	$\lvert Core\rvert_\delta$	Algorithm 2 $\lvert Q_u\rvert$	Algorithm 2 $\lvert Q_u\rvert$	Algorithm 2 $\lvert Q_\gamma\rvert$	Algorithm 2 $\lvert Q_\delta\rvert$	Algorithm 3 $\lvert Q_u\rvert$	Algorithm 3 $\lvert Q_\gamma\rvert$	Algorithm 3 $\lvert Q_\delta\rvert$
Data1:	Votes ×5	2175	16	15	14	15	15	15	14	15	15	14	15
Data2:	Votes ×10	4350	16	15	15	14	15	15	15	14	15	15	14
Data3:	Votes ×20	8700	16	15	15	5	15	15	15	11	15	15	10
Data4:	Votes ×40	17 400	16	15	14	15	15	15	14	15	15	14	15
Data5:	Votes ×60	26 100	16	15	9	15	15	15	10	15	15	10	15
Data6:	Dermatology ×5	1830	34	3	3	3	17	12	12	12	12	12	12
Data7:	Dermatology ×10	3660	34	3	3	2	15	12	12	13	12	12	12
Data8:	Dermatology ×20	7320	34	3	3	1	17	12	12	9	12	11	9
Data9:	Dermatology ×40	14 640	34	1	0	1	12	7	6	7	8	6	7
Data10:	Dermatology ×60	21 960	34	1	0	1	11	8	6	6	8	6	6
Average							14.7	12.6	11.6	11.7	12.7	11.5	11.5

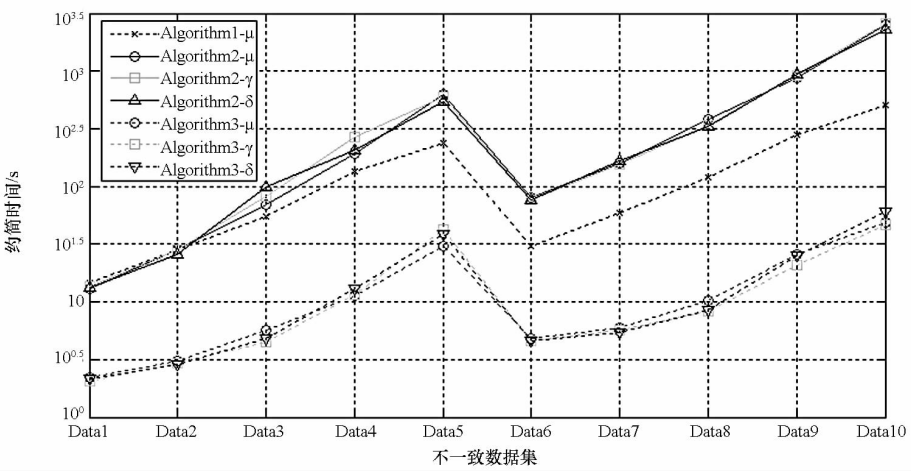

图 1　三种约简算法对不协调数据集的执行时间

Fig. 1　The run time of three reduction algorithms for inconsistent data sets

5　结论

现有约简算法中,或需数据集保持协调,或计算效率较低,很大程度上限制了约简算法的应用。由于误差或干扰等原因,实际问题中的数据往往是不协调的,因而需要约简算法在能够很好地支持不协调数据的同时具有良好的计算效率。基于此,针对不协调决策表,本文首先利用相对可区分度构造属性重要性,避免了基于正域的重要测度选择属性随机性大的缺陷;然后给出了简化协调决策表的概念,大大缩减了约简实例数;最后通过把不协调决策表的约简转化为求简化协调决策表的约简,提出了一种基于相对可区分度的高效完备不协调决策表约简算法。通过对 UCI 数据的实验结果表明,相比现有的不协调决策表约简算法,新算法不仅有较好的约简质量,而且具有较高的约简效率,适合处理具有大量冗余属性和不协调的海量数据集。

参考文献(References)

[1]　张文修, 米据生, 吴伟志. 不协调目标信息系统的知识约简[J]. 计算机学报, 2003,26(1): 12 - 18.
ZHANG Wenxiu, MI Junsheng, WU Weizhi. Knowledge

reductions in inconsistent information systems [J]. Chinese Journal of Computers, 2003,26(1): 12 – 18. (in Chinese)

[2] 刘少辉, 盛秋戬, 吴斌, 等. Rough 集高效算法的研究[J]. 计算机学报, 2003,26(5): 524 – 529.
LIU Shaohui, SHENG Quijian, WU Bin, et al. Research on efficient algorithms for rough set methods[J]. Chinese Journal of Computers, 2003, 26(5): 524 – 529. (in Chinese)

[3] 徐章艳, 刘作鹏, 杨炳儒, 等. 一个复杂度为 max(O(|C||U|), O(|C|^2|U/C|))的快速属性约简算法[J]. 计算机学报, 2006,29(3): 391 – 399.
XU Zhangyan, LIU Zhuopeng, YANG Bingru, et al. A quick attribute reduction algorithm with complexity max(O(|C||U|), O(|C|^2|U/C|))[J]. Chinese Journal of Computers, 2006, 29(3): 391 – 399. (in Chinese)

[4] Wang J. Reduction algorithms based on discernibility matrix: the ordered attributes method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489 – 504.

[5] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002,25(7): 759 – 766.
WANG Guoyin, YU Hong, YANG Dachun. Decision table reduction based on conditional information entropy [J]. Chinese Journal of Computers, 2002, 25(7): 759 – 766. (in Chinese)

[6] 滕书华, 周石琳, 孙即祥, 等. 基于条件熵的不完备信息系统属性约简算法[J]. 国防科技大学学报, 2010,32(1): 90 – 94.
TENG Shuhua, ZHOU Shilin, SUN Jixiang, et al. Attribute reduction algorithm based on conditional entropy under incomplete information system [J]. Journal of National University of Defense Technology, 2010,32(1): 90 – 94. (in Chinese)

[7] 刘启和, 李凡, 闵帆, 等. 一种基于新的条件信息熵的高效知识约简算法[J]. 控制与决策, 2005,20(8): 878 – 882.
LIU Qihe, LI Fan, MIN Fan, et al. An efficient knowledge reduction algorithm based on new conditional information entropy[J]. Control and Decision, 2005,20(8): 878 – 882. (in Chinese)

[8] Marzena K. Comparative study of alternative types of knowledge reduction in inconsistent systems[J]. International Journal of intelligent Systems, 2001,16(1): 105 – 120.

[9] 李凡, 刘启和, 叶茂, 等. 不一致决策表的知识约简方法研究[J]. 控制与决策, 2006,21(8): 857 – 862.
LI Fan, LIU Qihe, YE Mao, et al. Approaches to knowledge reductions in inconsistent decision tables [J]. Control and Decision, 2006,21(8): 857 – 862. (in Chinese)

[10] Teng S H, Wu J W, Sun J X, et al. An efficient attribute reduction algorithm [C]//The 2nd IEEE International Conference on Advanced Computer Control. Shenyang, 2010: 471 – 475.

[11] Zhao Y, Yao Y Y, Luo F. Data analysis based on discernibility and indiscernibility[J]. Information Sciences, 2007,177(22): 4959 – 4976.

[12] Xu Z Y, Yang B R, Wei S. Comparative study of different attribute reduction based on decision table [J]. Chinese Journal of Electronics, 2006,15(4 A): 953 – 956.