代数约简的条件信息熵表示及其高效约简算法

黄国顺[1] 曾凡智[2] 文 翰[1]
(佛山科学技术学院理学院 佛山 528000)[1] (佛山科学技术学院电子信息工程学院 佛山 528000)[2]

摘 要 给出如何保持正区域不变的语义分析,提出一种修正条件信息熵计算公式,证明保持修正条件信息熵不变与保持正区域不变相互等价。在此基础上,给出代数约简概念的修正条件信息熵表示。给出反例说明修正条件信息熵不具有单调性,导致没法给出自底向上的启发式约简算法,证明了代数协调集中不可删除属性的不可逆性质,提出一种自顶向下直接删除属性的高效约简算法。它从所有条件属性集出发,逐步删除不必要的属性,只需遍历各属性一次,即可保证得到原始决策表的一个代数约简。数值算例和实验验证了该算法的正确性和高效性。

关键词 条件信息熵,正区域,代数约简,算法
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.07.049

Conditional Information Entropy Representation of Algebraic Reduction and its Efficient Algorithm

HUANG Guo-shun[1] ZENG Fan-zhi[2] WEN Han[1]
(Science School, Foshan University, Foshan 528000, China)[1]
(Electronics and Information Engineering School, Foshan University, Foshan 528000, China)[2]

Abstract A semantic analysis on how to keep positive region unchanged was given. An improved conditional information entropy was proposed. It was proved that remaining the modified conditional information entropy unchanged and remaining positive region unchanged are equivalent. Therefore, some main concepts of algebraic reduction were described by the revised conditional information entropy. However, a counter example illustrates that its monotonicity does not hold, which means a heuristic reduction algorithm can not be constructed based on bottom-up. Any attribute in an algebraic consistent set is not irreversible if it is checked unsuppressible. An efficient algorithm based on top-down was proposed, which starts from condition attribute set, removes the unnecessary attribute step by step. It is finally guaranteed to obtain an algebraic reduction by traversing the attributes only once. Numerical example and the experimental results show that the algorithm is valid and efficient.

Keywords Conditional information entropy, Positive region, Algebraic reduction, Algorithm

1 引言

属性约简是在保持分类能力不变的情况下删除不必要的属性。Pawlazk首先提出基于正区域的属性约简[1],简称为代数约简。由于在代数表示下的直观性很差,国内外学者从其它角度对粗糙集理论进行重新定义,提出了各种不同的约简概念和方法,比较知名的有基于差别矩阵的属性约简[2]、基于条件信息熵的属性约简[3,4]、基于相对粒度的属性约简等[5-7]。但当决策表不一致时,上述约简与代数约简并不完全一致。因此对它们进行改进以保证得到原始决策表的代数约简是现有研究热点之一[8,9]。

对于条件信息熵约简,尽管目前研究成果很多[10-12],但大部分是把它作为一种独立的约简概念展开研究,很少有文献研究如何基于条件信息熵保证求到不一致决策表的代数约简。刘启和等曾提出一种基于条件信息熵求原始决策表的代数约简方法[9],它将所有不一致对象归并为一个新等价类,导出一种新的等价关系并代替原始表的决策属性关系,本质上是将不一致决策表转化为一致决策表来进行处理。类似的方法还有钱进等人提出的基于新条件信息量的属性约简算法[13]。前期我们利用差别矩阵讨论了条件熵约简与代数约简的区别与联系,提出了一种将条件信息熵约简转化为代数约简的方法[14,15]。但上述改进方法都需要中间转换过程。

本文首先分析了保持正区域不变的语义及其条件,在此基础上,提出一种修正条件信息熵计算公式,它无需任何中间转换过程,可直接利用修正条件信息熵计算原始决策表,证明了保持修正条件信息熵不变是保持正区域不变的充分必要条件,从而得到代数约简各种概念的条件信息熵表示。利用代数协调集的性质,提出一种自顶向下的启发式约简算法,它从所有条件属性集出发,每次删除不必要的属性,只需遍历各属性一次,可保证最后得到原始决策表的一个代数约简。如果采用文献[16]的方法计算等价类基数,可得其高效约简算法。

到稿日期:2013-09-13 返修日期:2013-11-25 本文受广东省自然科学基金资助项目(10452800001004185)资助。
黄国顺(1972-),男,博士,副教授,CCF高级会员,主要研究方向为粗糙集、粒度计算等,E-mail:fshgs_72@163.com;曾凡智(1965-),博士,教授,主要研究方向为数据库理论、数据挖掘等;文 翰(1977-),博士,讲师,主要研究方向为Web挖掘。

2 相关基本概念

定义 1[1] 决策表 $S=\langle U,V,f,C\cup D\rangle$是一类特殊的信息系统，其中 U 是一组对象的非空有限集合，称为论域。C 为有限的条件属性集，D 为有限的决策属性集，$C\cap D=\varnothing$，$V=\bigcup_{a\in C}V_a$，V_a 为属性 a 的值域；$f:U\times(C\cup D)\to V$ 是信息函数。对 U 上的任意属性集 $B\subseteq C\cup D$，定义不可分辨关系 $ind(B)=\{(x,y)\in U^2\mid \forall a\in B, f(x,a)=f(y,a)\}$，关系 $ind(B)$ 构成 U 的一个划分，记作 $U/ind(B)$，简记为 U/B。U/B 中的任何元素 $[x]_B=\{y\mid \forall a\in B, f(x,a)=f(y,a)\}$ 称为等价类。

定义 2[1] 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，对 $\forall P\subseteq C\cup D$，$Z\subseteq U$，记 $U/P=\{Y_1,Y_2,\cdots,Y_n\}$，称 $\underline{PZ}=\{x\in U:[x]_P\subseteq Z\}$ 为 Z 关于 P 的下近似集，称 $\overline{PZ}=\{x\in U:[x]_P\cap Z\neq\varnothing\}$ 为 Z 关于 P 的上近似集。

定义 3[1] 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，设 $P\subseteq C$，$U/D=\{D_1,D_2,\cdots,D_r\}$，$U/P=\{Y_1,Y_2,\cdots,Y_n\}$，称 $POS_P(D)=\bigcup_{D_i\in U/D}\underline{PD_i}$ 为 P 关于 D 的正区域。

记 $U_1=POS_C(D)$，$U_2=U-U_1$。

定义 4[13] 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$P\subseteq C$，若 $POS_P(D)=POS_C(D)$，则称 P 是 C 的代数协调集，若 P 是 C 的代数协调集且对于任意的 $P'\subset P$ 有 $POS_{P'}(D)\neq POS_C(D)$，则称 P 是 C 相对于 D 的一个代数约简。

所有代数约简的交称为代数约简核属性，王国胤等给出如下判定条件。

定理 1[4] 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$a\in C$ 是代数约简核属性当且仅当 $POS_{C-\{a\}}(D)\neq POS_C(D)$。

定义 5[4] 给定信息系统 $S=\langle U,V,f,A\rangle$，$P\subseteq A$，记 $U/P=\{Y_1,Y_2,\cdots,Y_n\}$，则 P 在论域 U 上信息熵定义为：

$$H(P)=-\sum_{Y_k\in U/P}p(Y_k)\log_2 p(Y_k)$$

式中，$p(Y_k)=|Y_k|/|U|$。

定义 6[4] 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，设 $P\subseteq C$，$U/P=\{Y_1,Y_2,\cdots,Y_n\}$，$U/D=\{D_1,D_2,\cdots,D_r\}$，那么 D 相对 P 的条件熵定义为：

$$H(D|P)=-\sum_{Y_k\in U/P}p(Y_k)\sum_{D_j\in U/D}p(D_j|Y_k)\log_2 p(D_j|Y_k) \tag{1}$$

式中，$p(Y_k)=|Y_k|/|U|$，$p(D_j|Y_k)=|Y_k\cap D_j|/|Y_k|$；$k=1,2,\cdots,n$；$j=1,2,\cdots,r$。若 $P\subseteq C$ 相对于决策属性 D 有 $H(D|P)=H(D|C)$，称 P 是 C 的信息熵协调集，若 P 是 C 的信息熵协调集且对 P 中任意属性 a 都有 $H(D|P)\neq H(D|P-\{a\})$，则称 P 是 C 的条件信息熵属性约简，简称为信息熵约简。

定理 2[3] 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$P\subseteq C$，那么 $H(D|P)=H(P\cup D)-H(P)$。

定理 3[3] 给定决策表 S，设 $Q\subseteq P\subseteq C$，$U/P=\{Y_1,Y_2,\cdots,Y_n\}$，$U/D=\{D_1,D_2,\cdots,D_r\}$，$U/Q=\{Y_1,Y_2,\cdots,Y_{i-1},Y_{i+1},\cdots,Y_{k-1},Y_{k+1},\cdots,Y_i\cup Y_k\}$，则 $H(D|Q)\geqslant H(D|P)$，等号成立的充分必要条件为对任意 $D_j\in U/D$，有 $|D_j\cap Y_i|/|Y_i|=|D_j\cap Y_k|/|Y_k|$。

定理 4[4] 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$a\in C$ 是信息熵约简下的核属性当且仅当 $H(D|C-\{a\})\neq H(D|C)$。

3 代数约简概念的条件信息熵表示

当决策表不一致时，基于条件信息熵的属性约简与代数约简并不完全一致。具体考察例 1 中的算例。

例 1 给定决策表 S_1，如表 1 所列，其中 $U=\{x_1,x_2,\cdots,x_{10}\}$，$C=\{a,b,c\}$，$D=\{d\}$。

表 1 决策表 S_1

U	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
a	1	1	1	0	0	0	0	2	2	1
b	1	1	1	0	0	1	1	1	2	1
c	1	1	1	1	1	1	1	2	2	0
d	1	1	2	1	2	1	2	3	3	2

S_1 的信息熵约简为 $\{a,c\}$，代数约简为 $\{c\}$，两者并不相同。

根据代数约简定义，代数约简即是一个保持正区域不变的极小条件属性集，其本质就是在 C 中删除一些不必要的属性，同时保持正区域不变，但在删除不必要属性集时会导致某些等价类合并。因此如果要保持正区域不变，这种合并只能属于以下两种情形：①具有相同决策属性值的等价类合并；②与正区域无关的某些等价类的合并即 U_2 中的等价类发生合并。同时注意到如下事实：当决策表一致时，$H(D/C)=0$，此时信息熵约简与代数约简是等价的；当决策表不一致时，$H(D/C)>0$，此时信息熵约简与代数约简不等价。因此为了消除这种不一致性，要求新提出的修正条件信息熵无论是一致还是不一致，决策表都能统一取 0 值。基于上述分析，提出修正条件信息熵计算公式如下。

定义 7 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$P\subseteq C$，假设 $U/P=\{Y_1,Y_2,\cdots,Y_n\}$，$U/D=\{D_1,D_2,\cdots,D_r\}$，定义修正条件信息熵如下。

$$MH(D|P)=H(D|P)-H_{U_2}(D|P) \tag{2}$$

式中，$H_{U_2}(D|P)=H_{U_2}(P\cup D)-H_{U_2}(P)$，$H_{U_2}(P)=-\sum_{Y_k\in U/P}p(Y_k\cap U_2)\log_2 p(Y_k\cap U_2)$，$H_{U_2}(P\cup D)=-\sum_{Y_k\in U/P}\sum_{D_j\in U/D}p(Y_k\cap U_2\cap D_j)\log_2 p(Y_k\cap U_2\cap D_j)$ 分别表示 P 和 $P\cup D$ 限制在 U_2 上的信息熵。

特别地，当决策表一致时，$U_2=\varnothing$，$MH(D|P)$ 退化成 $H(D|P)$，因此式(2)是对式(1)的改进和推广。

为了证明保持修正条件信息熵不变与保持正区域不变等价，需要用到如下几个引理。

引理 1 设 $f(x)$ 为区间 D 上的严凸函数，则对于 D 中任意的 $a<c<b$ 有 $\frac{f(b)-f(c)}{b-c}>\frac{f(b)-f(a)}{b-a}>\frac{f(c)-f(a)}{c-a}$。

证明：设 $p=\frac{b-c}{b-a}$，$q=\frac{c-a}{b-a}$，由 $a<c<b$，有 $p,q>0$，$p+q=1$，$c=pa+qb$，则 $\frac{f(b)-f(c)}{b-c}>\frac{f(b)-pf(a)-qf(b)}{b-pa-qb}=\frac{f(b)-f(a)}{b-a}$。类似地，$\frac{f(b)-f(a)}{b-a}>\frac{f(c)-f(a)}{c-a}$，证毕。

引理 2 (1)对任意 $a,b>0$，$(a+b)\log_2(a+b)>a\log_2 a+b\log_2 b$；

(2)对任意 $b\geqslant a\geqslant c>0$，$(a-c)\log_2(a-c)+(b+c)\log_2(b+c)>a\log_2 a+b\log_2 b$。

证明：(1)结论显然成立。(2)若 $a=c$，则退化为情形(1)，结论成立。下面只针对 $a>c$ 进行讨论。由于 $f(x)=$

$x\log_2 x$ 是$(0,+\infty)$的严凸函数，因此对任意 $0<x_1<x_2\leqslant y_1<y_2$，根据引理 1 有 $\frac{f(y_2)-f(x_2)}{y_2-x_2}>\frac{f(y_1)-f(x_1)}{y_1-x_1}$，令 $y_2=b+c, y_1=b, x_2=a, x_1=a-c$，则有 $f(b+c)-f(a)>f(b)-f(a-c)$，证毕。

引理 3 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，则 $MH(D|C)=0$。

证明：假设 $U/C=\{X_1,X_2,\cdots,X_m\}$，$U/D=\{D_1,D_2,\cdots,D_r\}$。$X_i\subseteq U_1$ 或 $X_i\subseteq U_2$ 必取其一。当 $X_i\subseteq U_1$ 时，存在某个 $D_{j_0}\in U/D$ 使得 $X_i\subseteq D_{j_0}$，从而 $\sum_{D_j\in U/D}p(X_i\cap D_j)\log_2 p(X_i\cap D_j)=p(X_i)\log_2 p(X_i)$，$\sum_{X_i\in U/C\wedge X_i\subseteq U_1}\sum_{D_j\in U/D}p(X_i\cap D_j)\log_2 p(X_i\cap D_j)=\sum_{X_i\in U/C\wedge X_i\subseteq U_1}p(X_i)\log_2 p(X_i)$。所以，

$$\begin{aligned}H(D|C)&=\sum_{X_i\in U/C}p(X_i)\log_2 p(X_i)-\sum_{X_i\in U/C}\sum_{D_j\in U/D}p(X_i\cap D_j)\log_2 p(X_i\cap D_j)\\&=\sum_{X_i\in U/C\wedge X_i\subseteq U_2}p(X_i)\log_2 p(X_i)-\sum_{X_i\in U/C\wedge X_i\subseteq U_2}\sum_{D_j\in U/D}p(X_i\cap D_j)\log_2 p(X_i\cap D_j)\end{aligned}$$

另一方面，当 $X_i\subseteq U_1$，$|X_i\cap U_2|=0$，因此

$$\begin{aligned}H_{U_2}(D|C)&=\sum_{X_i\in U/C}p(X_i\cap U_2)\log_2 p(X_i\cap U_2)-\sum_{X_i\in U/C}\sum_{D_j\in U/D}p(X_i\cap U_2\cap D_j)\log_2 p(X_i\cap U_2\cap D_j)\\&=\sum_{X_i\in U/C\wedge X_i\subseteq U_2}p(X_i\cap U_2)\log_2 p(X_i\cap U_2)-\sum_{X_i\in U/C\wedge X_i\subseteq U_2}\sum_{D_j\in U/D}p(X_i\cap U_2\cap D_j)\log_2 p(X_i\cap U_2\cap D_j)\\&=\sum_{X_i\in U/C\wedge X_i\subseteq U_2}p(X_i)\log_2 p(X_i)-\sum_{X_i\in U/C\wedge X_i\subseteq U_2}\sum_{D_j\in U/D}p(X_i\cap D_j)\log_2 p(X_i\cap D_j)\end{aligned}$$

所以 $MH(D|C)=H(D|C)-H_{U_2}(D|C)=0$。

类似地，当 $X_i\subseteq U_2$ 时同样有结论成立，证毕。

引理 3 表明，不管决策表是否一致，都有 $MH(D|C)=0$，这是我们希望得到的结果，也说明这种改进是有效和合适的。

引理 4 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$P\subseteq C$，那么 $MH(D|P)\geqslant 0$ 并且等号成立的充分必要条件是 U/C 中参与合并的等价类都来自正区域 U_1 且具有相同决策属性值，或者都来自 U_2。

证明：设 $U/P=\{Y_1,Y_2,\cdots,Y_n\}$，$U/D=\{D_1,D_2,\cdots,D_r\}$，记 $MH(D|P)=\sum_{Y_k\in U/P}G(Y_k)$，其中

$$G(Y_k)=p(Y_k)\log_2 p(Y_k)-\sum_{D_j\in U/D}p(Y_k\cap D_j)\log_2 p(Y_k\cap D_j)-p(Y_k\cap U_2)\log_2 p(Y_k\cap U_2)+\sum_{D_j\in U/D}p(Y_k\cap U_2\cap D_j)\log_2 p(Y_k\cap U_2\cap D_j)$$

因 $P\subseteq C$，所以对于任意 $Y_k\in U/P$，它都可表示成 U/C 中若干个等价类的并。不失一般性，假设 $Y_k=[x_s]_C\cup[x_t]_C$，$[x_s]_C\neq[x_t]_C$ 且它们非空，此时又需分 4 种情形分别讨论：(1)$[x_s]_C\subseteq U_1$，$[x_t]_C\subseteq U_1$ 且具有相同的决策属性值；(2)$[x_s]_C\subseteq U_1$，$[x_t]_C\subseteq U_1$ 但具有不同的决策属性值；(3)一个在 U_1 中，另一个在 U_2 中；(4)两个都在 U_2 中。

对于情形(1)，必存在某个 $D_{j_0}\in U/D$ 使得 $[x_s]_C\subseteq D_{j_0}$ 且 $[x_t]_C\subseteq D_{j_0}$，从而 $Y_k\subseteq D_{j_0}$，$\sum_{D_j\in U/D}p(Y_k\cap D_j)\log_2 p(Y_k\cap D_j)=p(Y_k)\log_2 p(Y_k)$，$Y_k\cap U_2=\emptyset$，$Y_k\cap U_2\cap D_j=\emptyset$，所以 $G(Y_k)=0$。

对于情形(2)，存在 $D_{j_1},D_{j_2}\in U/D$ 且 $D_{j_1}\neq D_{j_2}$，使得 $[x_s]_C\subseteq D_{j_1}$，$[x_t]_C\subseteq D_{j_2}$，因为 $D_{j_1}\cap D_{j_2}=\emptyset$，所以 $Y_k\cap D_{j_1}=[x_s]_C$，$Y_k\cap D_{j_2}=[x_t]_C$，$\sum_{D_j\in U/D}p(Y_k\cap D_j)\log_2 p(Y_k\cap D_j)=p([x_s]_C)\log_2 p([x_s]_C)+p([x_t]_C)\log_2 p([x_t]_C)$；又因 $[x_s]_C\subseteq U_1$，$[x_t]_C\subseteq U_1$，从而 $Y_k\cap U_2=\emptyset$，$Y_k\cap U_2\cap D_j=\emptyset$，所以 $G(Y_k)=p([x_s]_C\cup[x_t]_C)\log_2 p([x_s]_C\cup[x_t]_C)-p([x_s]_C)\log_2 p([x_s]_C)-p([x_t]_C)\log_2 p([x_t]_C)$，根据引理 2 的结论(1)有 $G(Y_k)>0$。

对于情形(3)，不妨设 $[x_s]_C\subseteq U_1$，$[x_t]_C\subseteq U_2$，由 $[x_s]_C\subseteq U_1$ 知存在某个 $D_{j_1}\in U/D$，使得 $[x_s]_C\subseteq D_{j_1}$。由 $[x_t]_C\subseteq U_2$ 知存在若干个(至少两个)U/D 的不同等价类，使得 $[x_t]_C$ 与它们相交不空且 $[x_t]_C$ 被它们的并所包含。不妨设就是两个这样的等价类，即 $D_{j_2},D_{j_3}\in U/D$，$D_{j_2}\neq D_{j_3}$，$[x_t]_C\cap D_{j_2}\neq\emptyset$，$[x_t]_C\cap D_{j_3}\neq\emptyset$，$[x_t]_C\subseteq D_{j_2}\cup D_{j_3}$。此时又分为两种情况分别讨论：

(a)若 D_{j_1} 与 D_{j_2}，D_{j_3} 都不相同，亦即 $D_{j_1}\neq D_{j_2}\neq D_{j_3}$，则

$$G(Y_k)=p([x_s]_C\cup[x_t]_C)\log_2 p([x_s]_C\cup[x_t]_C)-p([x_s]_C)\log_2 p([x_s]_C)-p([x_t]_C)\log_2 p([x_t]_C)$$

根据引理 2 的结论(1)知 $G(Y_k)>0$。

(b)若 D_{j_1} 与 D_{j_2}，D_{j_3} 其中一个相同，不妨设为 $D_{j_1}=D_{j_2}\neq D_{j_3}$，那么

$$G(Y_k)=p([x_s]_C\cup[x_t]_C)\log_2 p([x_s]_C\cup[x_t]_C)-p(([x_s]_C\cup[x_t]_C)\cap D_{j_2})\log_2 p(([x_s]_C\cup[x_t]_C)\cap D_{j_2})-p([x_t]_C)\log_2 p([x_t]_C)+p([x_t]_C\cap D_{j_2})\log_2 p([x_t]_C\cap D_{j_2})$$

令 $b=p(([x_s]_C\cup[x_t]_C)\cap D_{j_2})$，$a=p([x_t]_C)$，$c=p([x_t]_C\cap D_{j_3})$，则 $b+c=p([x_s]_C\cup[x_t]_C)$，$a-c=p([x_t]_C\cap D_{j_2})$，根据引理 2 的结论(2)知 $G(Y_k)>0$。

对于情形(4)，由于 $[x_s]_C\subseteq U_2$，$[x_t]_C\subseteq U_2$，从而 $Y_k\subseteq U_2$，$Y_k\cap U_2=Y_k$，$Y_k\cap U_2\cap D_j=Y_k\cap D_j$，因此 $G(Y_k)=0$。

综上所述，对任意 $Y_k\in U/P$，都有 $G(Y_k)\geqslant 0$，从而 $MH(D|P)=\sum_{k=1}^{n}G(Y_k)\geqslant 0$。上述证明过程同时表明，只有情形(1)和(4)才保持修正条件信息熵不变，对于其它两种情形则会引起修正条件信息熵变大，因此 $MH(D|P)=0$ 当且仅当 U/C 中参与合并的等价类同时在正区域 U_1 中且具有相同的决策属性值，或都在 U_2 中，证毕。

引理 4 表明，保持修正条件信息熵不变与保持正区域不变在语义上是一致的，从而有如下定理 5 成立。

定理 5 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$P\subseteq C$，那么 $POS_P(D)=POS_C(D)$ 当且仅当 $MH(D|P)=MH(D|C)$。

证明：先证必要性。假设 $U/P=\{Y_1,Y_2,\cdots,Y_n\}$，$U/D=\{D_1,D_2,\cdots,D_r\}$，记 $MH(D|P)=\sum_{Y_k\in U/P}G(Y_k)$，其中

$$G(Y_k)=p(Y_k)\log_2 p(Y_k)-\sum_{D_j\in U/D}p(Y_k\cap D_j)\log_2 p(Y_k\cap D_j)-p(Y_k\cap U_2)\log_2 p(Y_k\cap U_2)+\sum_{D_j\in U/D}p(Y_k\cap U_2\cap D_j)\log_2 p(Y_k\cap U_2\cap D_j)$$

若 $POS_P(D)=POS_C(D)$，那么 $U-POS_P(D)=U-POS_C(D)=U_2$，对任意的 $Y_k\in U/P$，分两种情况讨论，(1)$Y_k\subseteq POS_P(D)$；(2)$Y_k\subseteq U-POS_P(D)$。

(1)当 $Y_k\subseteq POS_P(D)$ 时，必存在某个 $D_{j_0}\in U/D$ 使得 $Y_k\subseteq D_{j_0}$，从而 $\sum_{D_j\in U/D}p(Y_k\cap D_j)\log_2 p(Y_k\cap D_j)=p(Y_k)\log_2 p$

(Y_k)，又因 $POS_P(D)\subseteq U_1$，所以 $|Y_k\cap U_2|=0$，易得 $G(Y_k)=0$。

(2)若 $Y_k\subseteq U-POS_P(D)$，由 $POS_P(D)=POS_C(D)$ 知 $U-POS_P(D)=U_2$，那么 $|Y_k\cap U_2|=|Y_k|$，$|Y_k\cap U_2\cap D_j|=|Y_k\cap D_j|$，因此 $G(Y_k)=0$。

综上所述，对任意的 $Y_k\in U/P$，只要 $POS_P(D)=POS_C(D)$，必有 $G(Y_k)=0$，从而 $MH(D|P)=MH(D|C)$。

反之，如果 $MH(D|P)=MH(D|C)$，则一定有 $POS_P(D)=POS_C(D)$。

由于 $P\subseteq C$，U/P 中的等价类都可以由 U/C 的等价类合并而成，根据引理 4，有 $MH(D|P)\geqslant 0$。假设 $POS_P(D)=POS_C(D)$ 不成立，由 $P\subseteq C$ 知，必有 $POS_P(D)\subset POS_C(D)$，这意味着至少存在一个 $x_0\in POS_C(D)$，但 $x_0\notin POS_P(D)$，从而在 $[x_0]_P$ 中至少存在两个相异元素 $x_s,x_t\in[x_0]_P$，使得 $f(x_s,D)\neq f(x_t,D)$(否则 $[x_0]_P\subseteq POS_P(D)$，与 $x_0\notin POS_P(D)$ 矛盾)，又因 $[x_0]_P$ 是若干个 U/C 的等价类的并，且一定包含 U_1 中的等价类 $[x_0]_C$，所以 $[x_0]_P$ 不可能属于引理 4 中等号成立的两种情形，从而 $MH(D|P)>0$，但 $MH(D|C)=0$，与题设 $MH(D|P)=MH(D|C)$ 矛盾。从而知当 $MH(D|P)=MH(D|C)$ 时一定有 $POS_P(D)=POS_C(D)$。

定理 5 表明，保持修正条件信息熵不变与保持正区域不变等价。从而可以给出代数约简各种概念的修正条件信息熵描述。

定义 8 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$P\subseteq C$，对任意的 $B\subset P$，若 $MH(D|P-B)=MH(D|P)$，则称 B 为 P 中的不必要属性集，特别地，如果 $B=\{a\}$，则称 a 是 P 中的不必要属性，否则称 a 为 P 中必要的属性。

根据定理 1 和定理 5 可给出代数核的修正条件信息熵表示。

定义 9 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，对任意的 $a\in C$，若 $MH(D|C-\{a\})>0$，则称 a 为代数约简核属性，简称为代数核。

根据定理 5 和定义 4 可以给出代数约简的条件信息熵描述。

定义 10 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$P\subseteq C$，如果 $MH(D|P)=MH(D|C)$，则称 P 是 C 的代数协调集，如果 P 是 C 的代数协调集，且对任意的 $P'\subset P$，$MH(D|P')>MH(D|C)$，则称 P 是 C 的代数约简。如果 P 是 C 的代数协调集，$P'\subset P$，有 $MH(D|P')=MH(D|C)$，则称 P' 是 P 的真子代数协调集。

文献[3]提出的条件信息熵与本文提出的修正条件信息熵有如下关系。

性质 1 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$P\subseteq C$，如果 $H(D|P)=H(D|C)$，则一定有 $MH(D|P)=MH(D|C)$。

证明：假设 $U/D=\{D_1,D_2,\cdots,D_r\}$，由于 $P\subseteq C$，那么对任意的 $Y_k\in U/P$ 都可表示成 U/C 中若干等价类的并，不失一般性，假设 $Y_k=X_i\cup X_l$，$X_i,X_l\in U/C$。若 $H(D|P)=H(D|C)$，那么根据定理 3，对任意的 $D_j\in U/D$，有 $|D_j\cap X_i|/|X_i|=|D_j\cap X_l|/|X_l|$。由于 $\sum_{j=1}^{r}|D_j\cap X_i|/|X_i|=1$，因此对 $|D_j\cap X_i|/|X_i|$ 的取值有两种情况：(1)存在某个 $D_{j_0}\in U/D$，使得 $|D_{j_0}\cap X_i|/|X_i|=1$，从而 $X_i\subseteq D_{j_0}$；(2)存在某些 $D_j\in U/D$(至少两个)，使得 $0<|D_j\cap X_i|/|X_i|<1$，从而 $X_i\subseteq U_2$。

对于情形(1)，根据 $|D_j\cap X_i|/|X_i|=|D_j\cap X_l|/|X_l|$ 知道同时有 $X_l\subseteq D_{j_0}$，这意味着参与合并的两等价类 X_i,X_l 都属于正区域 $POS_C(D)$ 且具有相同的决策属性值，根据引理 4，$MH(D|P)=MH(D|C)$。

对于情形(2)，根据 $|D_j\cap X_i|/|X_i|=|D_j\cap X_l|/|X_l|$ 知 $X_l\subseteq U_2$，根据引理 4 同样有 $MH(D|P)=MH(D|C)$。

所以不管哪种情况，只要 $H(D|P)=H(D|C)$，就一定有 $MH(D|P)=MH(D|C)$，证毕。

反之结论不成立，如例 1 的 S_1，$MH(D|\{c\})=MH(D|C)$，但 $H(D|\{c\})\neq H(D|C)$，具体计算过程参见例 2。

推论 1 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$P\subseteq C$，如果 P 是 C 的信息熵协调集，则 P 是 C 的代数协调集。

证明：根据性质 1 和定义 4 和定义 10 即知结论成立。

推论 2 给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$a\in C$ 是代数核，则它一定是信息熵核属性。

证明：令 $P=C-\{a\}$，根据性质 1 的逆否命题即知结论成立。

例 2 利用例 1 的决策表 S_1 阐述怎样用修正条件信息熵求其代数约简和代数核。

由于 $U/C=\{X_1,X_2,X_3,X_4,X_5,X_6\}$，$U/D=\{D_1,D_2,D_3\}$，其中 $X_1=\{x_1,x_2,x_3\}$，$X_2=\{x_4,x_5\}$，$X_3=\{x_6,x_7\}$，$X_4=\{x_8\}$，$X_5=\{x_9\}$，$X_6=\{x_{10}\}$，$D_1=\{x_1,x_2,x_4,x_6\}$，$D_2=\{x_3,x_5,x_7,x_{10}\}$，$D_3=\{x_8,x_9\}$，可计算 $POS_C(D)=\{x_8,x_9,x_{10}\}$，$U_2=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$。

根据代数约简定义，可判定 $P=\{c\}$ 是 C 的一个代数约简。

另一方面，根据 U/C 和 $U/(C\cup D)=\{\{x_1,x_2\},\{x_3\},\{x_4\},\{x_5\},\{x_6\},\{x_7\},\{x_8\},\{x_9\},\{x_{10}\}\}$，可计算 $H(D|C)=(3\log_2 3+2\log_2 2)/10$，$H_{U_2}(D|C)=(3\log_2 3+2\log_2 2)/10$，所以 $MH(D|C)=0$。对于属性集 $P=\{c\}$，根据 $U/P=\{\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\},\{x_8,x_9\},\{x_{10}\}\}$ 和 $U/(P\cup D)=\{\{x_1,x_2,x_4,x_6\},\{x_3,x_5,x_7\},\{x_8,x_9\},\{x_{10}\}\}$，可计算条件信息熵 $H(D|P)=(7\log_2 7-4\log_2 4-3\log_2 3)/10$，$H_{U_2}(D|P)=(7\log_2 7-4\log_2 4-3\log_2 3)/10$，所以 $MH(D|P)=MH(D|C)$，根据定理 5 知 $POS_P(D)=POS_C(D)$，因 $P=\{c\}$ 只有一个属性，可判定 $P=\{c\}$ 是 C 的一个代数约简。

再深入分析会发现，由于 $P\subseteq C$，因此 U/P 中的等价类都可以由 U/C 的等价类合并而成。根据引理 4 的结论，当参与合并的等价类具有相同的决策属性值，或它们能同时被包含在 U_2 中，合并过程不会导致修正条件信息熵变大。具体到本例中，$U/P=\{\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\},\{x_8,x_9\},\{x_{10}\}\}$，其中 $\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$ 是由 U/C 的 3 个等价类 X_1、X_2 和 X_3 合并而成，它们都是 U_2 的子集，根据引理 4，这种合并不会引起修正条件信息熵变大；类似地，$\{x_8,x_9\}$ 是由 U/C 的一致对象集 X_4 和 X_5 合并得到，由于参与合并的 X_4 和 X_5 具有相同的决策属性值，因此它们的合并也不会引起修正条件信息熵变化；从而有 $MH(D|P)=MH(D|C)$，继而保证 $POS_P(D)=POS_C(D)$ 成立。但 $H(D|P)\neq H(D|C)$，这是由于当 X_1、X_2 和 X_3 合并而成 $\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$ 时，$|D_1\cap X_1|/|X_1|\neq|D_1\cap X_2|/|X_2|$，根据定理 3 知 $P=\{c\}$ 不是 C 的一个条件信息熵约简。

若令 $Q=\{a,c\}$，$U/Q=\{\{x_1,x_2,x_3\},\{x_4,x_5,x_6,x_7\},$

$\{x_8,x_9\},\{x_{10}\}\}$，它也是由 U/C 中某些等价类合并而得到，其中 $\{x_4,x_5,x_6,x_7\}$ 由等价类 X_2,X_3 合并而成，$\{x_8,x_9\}$ 由 X_4 和 X_5 合并得到。由于对任意的 $D_j\in U/D$，有 $|D_j\cap X_2|/|X_2|=|D_j\cap X_3|/|X_3|$ 和 $|D_j\cap X_4|/|X_4|=|D_j\cap X_5|/|X_5|$ 同时成立，因此根据定理 3 知 $H(D|Q)=H(D|C)$，同时可验证它是 C 的一个信息熵约简。由于 $X_2,X_3\subseteq U_2,X_4,X_5\subseteq U_1$ 且具有相同的决策属性值，它们的合并不会导致修正条件信息熵变化，根据定理 5 知 $Q=\{a,c\}$ 也是 C 的一个代数协调集，但不是代数约简，因为还有冗余属性 a。

核属性方面，考察属性集 $C-\{c\}$，$U/C-\{c\}=\{\{x_1,x_2,x_3,x_{10}\},\{x_4,x_5\},\{x_6,x_7\},\{x_8\},\{x_9\}\}$，由于它引起不一致对象集与一致对象集合并（$\{x_1,x_2,x_3\}$ 和 $\{x_{10}\}$ 合并成 $\{x_1,x_2,x_3,x_{10}\}$），因此根据引理 4 有 $MH(D|C-\{c\})>0$，根据定义 8 知属性 c 是一个代数核。

4 代数约简的高效约简算法

很遗憾的是，修正条件信息熵不具有单调性，即对于 $Q\subseteq P\subseteq C$，不一定有 $MH(D|Q)\geqslant MH(D|P)$ 成立，具体算例如下。

例 3　将例 $1S_1$ 中的对象 x_4 和 x_5 在属性 a 的取值分别改成“1”，其它不变，得到给定决策表 S_2，如表 2 所列。

表 2　决策表 S_2

U	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
a	1	1	1	1	1	0	0	2	2	1
b	1	1	1	0	0	1	1	1	2	1
c	1	1	1	1	1	1	1	2	2	0
d	1	1	2	1	2	1	2	3	3	2

经简单计算得到 $MH(D|\{a,b\})=(4\log_2 4-3\log_2 3-2\log_2 2)/10$，$MH(D|\{a\})=(4\log_2 4-5\log_2 5+3\log_2 3)/10$，显然有 $MH(D|\{a,b\})>MH(D|\{a\})$，但 $\{a\}\subset\{a,b\}$。因此没法利用修正条件信息熵设计度量属性重要性的度量公式及相应的自底向上的启发式约简算法，但能否设计出自顶向下的约简算法呢？下文的性质 2 和推论 3 确保了代数协调集中的不可删除属性具有不可逆性质，即在较大的代数协调集中已经被证明是不可删除属性在其真子代数协调集中仍是不可删除的。

性质 2　给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$Q\subseteq P\subseteq C$ 且 P,Q 都是 C 的代数协调集，属性 $a\in P\cap Q$，若 a 在 P 中是必要的，那么 a 在 Q 中也是必要的。

证明：根据题设有 $MH(D|P)=MH(D|Q)=0$，$MH(D|P-\{a\})\neq MH(D|P)$，要证 $MH(D|Q-\{a\})\neq MH(D|Q)$。

（反证法）由于 $Q-\{a\}\subseteq C$，因此任意 $Q_i\in U/(Q-\{a\})$ 都可表示成 U/C 中若干等价类的并。假设性质 2 的结论不成立，从而有 $MH(D|Q-\{a\})=MH(D|Q)=0$。根据引理 4，U/C 中等价类合并成 Q_i 只能有两种情况发生，即引理 4 中情形(1)和(4)。

由 $P-\{a\}\subseteq C$ 知，任意的 $Y_k\in U/(P-\{a\})$ 也可表示成 U/C 中若干等价类的并，但 $MH(D|P-\{a\})\neq 0$，因此至少存在一个 $U/(P-\{a\})$ 中的等价类 $Y_{k_0}\in U/(P-\{a\})$，使得它是由引理 4 中情形(2)和(3)的 U/C 等价类合并而成。但同时注意到，$Q-\{a\}\subseteq P-\{a\}$，$U/(P-\{a\})\preceq U/(Q-\{a\})$，因此对于上述 Y_{k_0}，存在 $Q_{i_0}\in U/(Q-\{a\})$，使得 $Y_{k_0}\subseteq Q_{i_0}$，从而 Q_{i_0} 只能由引理 4 中情形(2)和(3)的 U/C 等价类合并而成，与前述结论矛盾。

推论 3　给定决策表 $S=\langle U,V,f,C\cup D\rangle$，$Q\subseteq P\subseteq C$ 是 C 的代数协调集，对任意的 $a\in Q$，若 a 在 Q 中是不必要的属性，那么 a 在 P 中也是不必要的。

证明：根据性质 2，其逆否命题也成立，证毕。

特别地，如果 P 是 C 的代数协调集，且 $a\in P$ 是 P 中不必要的属性，那么 a 在 C 中也是不必要的。

根据性质 2 和推论 3，在检验某属性是否是代数协调集中可删除的，不必反复检验，只需检验一次即可。

由于所有条件属性集本身已是一代数协调集，以它为出发点，每次删除不必要的属性仍然保持修正条件信息熵不变，根据定理 5，仍然是一代数协调集。如果某属性 a 被检验是必要的，那么保留它对修正条件信息熵没任何影响，一直这样下去，直至遍历完所有条件属性，最后得到的仍是一个代数协调集且没有真子集是代数协调集，从而得到一个代数约简。

由于直接删除属性算法依赖于原始决策表属性集的排列顺序，顺序不同，可能会得到不同的约简结果（具体见例 4 计算结果），考虑到 $MH(D|\{a\})$ 刻画的是决策属性与条件属性的矛盾程度，$MH(D|\{a\})$ 越大则说明 a 对决策的参考程度越低；$MH(D|\{a\})$ 越小则越重要，特别地如果 $MH(D|\{a\})=0$，则表示属性 a 是最重要的且其它条件属性集都可约去。因此为了排除自然顺序的不良影响，得到较好的约简结果，在检验属性前先以决策属性 D 相对于条件属性 a 的修正条件信息熵 $MH(D|\{a\})$ 的大小作为排序标准，$MH(D|\{a\})$ 越大，则越排在前面优先检验，于是得到基于修正条件信息熵的直接约简算法。

算法 1　基于修正条件信息熵的直接约简算法

输入：决策表 S

输出：该决策表一个代数约简 R

初始化：R=C；

步骤 1　计算 $POS_C(D)$ 和 U_2；

步骤 2　分别计算 D 相对 C 中各属性 x 的修正条件信息熵 MH(D/{x})，并按 MH(D/{x})降序重新排列 C 中各属性顺序；

步骤 3　对排序后的各属性 x 重复执行以下操作，直至所有属性被遍历一次：计算 MH(D/R−{x})，如果 MH(D/R−{x})=0，则属性 x 可约，令 R=R−{x}，否则 R 保持不变；

如果采用文献[16]的基排序计算正区域，整个算法的时间复杂度为 $O(|C|^2|U|)$。

例 4[9]　在表 3 的决策表 S_3 中，$U=\{x_1,x_2,\cdots,x_{10}\}$，$C=\{a,b,c,e,f\}$。

表 3　决策表 S_3

U	a	b	c	e	f	D
x_1	0	0	0	0	1	0
x_2	0	1	1	1	0	1
x_3	1	1	0	1	1	1
x_4	0	1	1	1	0	0
x_5	0	0	1	0	1	0
x_6	1	1	0	1	0	1
x_7	0	1	1	1	1	1
x_8	1	1	1	0	1	1
x_9	1	1	0	1	1	0
x_{10}	0	1	1	1	1	0

如果按从左到右的顺序 $a\to b\to c\to e\to f$ 检验各属性，根据算法 1 得到约简结果 $R=\{b,c,e,f\}$。如果将属性列 b 与属性 a 列对调，再按从左至右利用算法 1 检验各属性，最后得到

约简结果 $R=\{a,e,f\}$。可见不同的原始排列顺序确实会对最后的约简结果产生一定的影响，尽管上述两个约简集都是 C 的代数约简集。如果先对各属性计算其修正条件信息熵 $MH(D/\{x\})$并按降序排序，得到的计算结果如表 4 所列。按降序排列得到新的检验顺序为：$c\rightarrow f\rightarrow e\rightarrow a\rightarrow b$，利用算法 1 计算得到约简结果 $R=\{a,e,f\}$，它是 S_3 的最小代数约简。

表 4　各属性的修正条件信息熵 $MH(D/\{x\})$

	a	b	c	e	f
MH(D/{x})	$\frac{1}{10}\log\frac{6^6}{3^3\times4^4}$	$\frac{1}{10}\log\frac{8^8}{5^5\times6^6}$	$\frac{1}{10}\log\frac{6^6}{3^3}$	$\frac{1}{10}\log\frac{7^7}{2^8\times4^4}$	$\frac{1}{10}\log\frac{7^7\times3^3}{4^8}$

5　仿真实验

为了对不一致决策表展开实验，我们利用 UC[17] 的原始数据集删除一些条件属性生成实验所需的不一致数据集，其中 Mushroom1 和 Mushroom2 由 Mushroom 原始数据分别删除 5 个和 8 个条件属性及对应的取值所得，其它数据集类似，它们的参数特征具体如表 5 所列。

表 5　仿真数据集

决策表	总实例数	一致实例数	不一致实例数	属性数	核属性数
例 3 的 S_3	10	4	6	5	1
Mushroom1	8124	8024	100	17	4
Mushroom2	8124	8024	100	14	5
Voting-recorder1	435	431	4	13	7
Zoo1	101	90	11	16	6
Chess end game1	3196	2840	356	29	25
Tic-tac-toe1	958	904	54	7	7

对表 5 提供的不一致决策表分别用文献[9,16]的算法（分别简记为算法 A 和算法 B，其中算法 A 以核属性集为起点）和本文提出的算法 1（记作算法 C）进行实验比较。实验环境为 Intel Xeon 2.0G，1G RAM，Windows Server 2003 专业版 SP2，C# 编程，运行在 dotnetfx3.0 上，具体计算结果如表 6 所列，其中执行时间标出"<0.01"的表示实际运行结果为 0s。

就约简结果而言，算法 A 所得的结果最好，在所做实验中，它都能得到最小约简集。其次是算法 C，除了在 Voting-recorder1 数据集上它不能得到最好约简结果外，在其它仿真数据集上都能得到最小约简集。算法 B 所得约简结果最差，在仿真数据集 S_3、Mushroom1、Mushroom2 上约简结果都差于其它两种算法，这主要是由于算法 B 以近似分类质量为启发式信息，而近似分类质量的计算会产生"非此即彼"的问题，导致并不能很好地度量出属性的重要性。另外该算法不完备，如果增加二次约简过程，则能进一步得到较小的约简集（表 6 算法 B 列括号中的数据即为增加二次约简后的计算结果）。

表 6　3 个约简算法的比较

决策表	算法 A		算法 B		算法 C	
	约简后属性数	执行时间/(s)	约简后属性数	执行时间/(s)	约简后属性数	执行时间/(s)
S_3	3	<0.01	4	<0.01	3	<0.01
Mushroom1	6	1.25(0.52)	7(6)	0.17	6	0.19
Mushroom2	6	0.94(0.38)	7(6)	0.14	6	0.16
Voting-recorder1	9	0.03(0.02)	11(10)	0.02	10	0.02
Zoo1	8	0.01(<0.01)	9(8)	<0.01	8	<0.01
Chess-end-game1	26	0.78(0.56)	27(26)	0.36	26	0.27
Tic-tac-toe1	7	0.03(0.02)	7	0.02	7	0.02

就时间效率而言，算法 B 除了在 Ches-end-game1 外，在其余数据集上计算效率最高，这主要是因为算法 B 以简化决策表加快计算速度；本文提出的算法 C 也非常快，在数据集 Ches-end-game1 上它用时最少，优于其它两种算法，在其它数据集上比算法 B 慢些，但几乎相当，主要是因为两者都采用基排序思想，同时算法 C 计算过程省略了自底向上收集属性的过程，只需判断各属性是否可约即可。算法 A 的耗时最多，主要是因为它以快速排序来计算正区域以及以核属性为起点，算法 A 执行时间列括号中的数据为求核属性所用时间。

结束语　由于现有条件信息熵约简与代数约简并不等价，因此没法通过条件信息熵刻画代数约简及其相关概念。为了给出代数约简的条件信息熵描述，本文提出一种修正条件信息熵计算公式；当决策表一致时，它退化为现有条件信息熵公式，当决策表不一致时，它能考虑到不一致对象对属性约简的影响，因此它是现有条件信息熵公式的推广和改进。业已证明，保持修正条件信息熵不变与保持正区域不变相互等价，在此基础上给出了代数约简的知识粒度表示。令人意外的是，本文提出的修正条件信息熵不具有单调性，从而无法利用它设计出计算属性相对重要性的度量公式及相应的自底向上启发式约简算法。对代数协调集的性质展开讨论，证明了在较大代数协调集中不可约属性在其真子代数协调集中仍是不可约的，同时注意到所有条件属性集本身已经是一个代数协调集的事实，提出一种自顶向下的约简算法，它从所有条件属性集出发，依次检验各属性是否可约，只需遍历一次即可保证得到一个代数约简。由于条件信息熵与知识粒度紧密相关[18]，下一步我们将本文方法与知识粒度联系起来研究，对属性约简的知识粒度本质展开讨论，建立基于知识粒度求属性约简的一般理论和方法。

参 考 文 献

[1] Pawlak Z. Rough set [J]. Commucation of the ACM, 1995, 38(11): 89-95

[2] Hu X H, Cerene N. Learning in relational databases: a rough set approach [J]. International Journal of Computation Intelligence, 1995, 11(2): 323-338

[3] 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

[4] Wang G Y, Zhao J, An J J, et al. A comparative study of algebra viewpoint and information viewpoint in attribute reduction [J]. Fundamenta Informaticae, 2005, 68(3): 289-301

（下转第 274 页）

加了 61.6%，而本文的基于期望的方法多样性则增加到559.8，增加了 182.7%。同样在其它实验设置下，基于推荐期望的方法表现了同样的优势，尤其是当准确度稍微再下降时，多样性增加更为显著。

结束语 推荐多样性作为评价推荐质量的一个重要方面，最近引起了人们的关注。传统推荐系统通常向用户推荐Top-N个具有最高预测评分的物品，因此提供了较好的预测准确度，但是推荐多样性方面的性能却较差。本文提出的基于候选物品推荐期望的方法较文献[4]中提出的重排名方法在准确度-多样性的权衡上具有更好的性能，并且保持了较低的时间复杂度，保证了推荐的效率。另外本文提出的方法具有出色的灵活性，能够与不同的评分预测算法结合使用，可参数化选择准确度-多样性的平衡点。本文通过结合多种热门评分预测算法，在实际的 MovieLens 数据集上对提出的基于推荐期望的推荐产生方法进行测试，证明了该方法的有效性。

参 考 文 献

[1] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating Collaborative Filtering Recommender Systems[J]. ACM Transactions on Information Systems, 2004, 22(1): 5-53

[2] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques For Recommender Systems[J]. IEEE Computer, 2009, 42(8): 30-37

[3] Guy Shani, Gunawardana A. Evaluating Recommendation Systems[M]// Recommender Systems Handbook. Springer, 2011: 257-297

[4] Adomavicius G, Kwon Y. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(5): 896-911

[5] Bradley K, Smyth B. Improving Recommendation Diversity [C]// Proceeding of 12th Irish Conference on Artificial Intelligence and Cognitive Science. Ireland, 2001

[6] Brynjolfsson E, Hu Y J, Simester D. Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales[J]. Management Science, 2011, 57(8): 1373-1386

[7] Fleder D, Hosanagar K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity[J]. Management Science, 2009, 55(5): 697-712

[8] Zhang M, Hurley N. Avoiding monotony: improving the diversity of recommendation lists[C]// Proceedings of the 2nd ACM Conference on Recommender Systems. Lausanne, 2008

[9] Goldstein D G, Goldstein D C. Profiting from the Long Tail[J]. Harvard Business Review, 2006, 84(6): 24-28

[10] Kim H K, Kim J K, Ryu Y U. A Local Scoring Model for Recommendation[C]// Proceedings of the 20th Workshop on Information Technologies and Systems. St. Louis, Missouri, USA, 2010

[11] Park Y J, Tuzhilin A. The Long Tail of Recommender Systems and How to Leverage It[C]// Proceedings of the 2nd ACM Conference on Recommender Systems. Lausanne, 2008

[12] Adomavicius G, Kwon Y. Maximizing aggregate recommendation diversity: a graph-theoretic approach[C]// Proceedings of Workshop on Novelty and Diversity in Recommender Systems. Chicago, Illinois, USA, 2011

[13] Cremonesi P, Koren Y, Turrin R. Performance of Recommender Algorithms on Top-N Recommendation Tasks[C]// Proceedings of the fourth ACM conference on Recommender systems. Barcelona, 2010

[14] Smyth B, McClave P. Similarity vs. Diversity[C]// Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development. 2001

[15] Shannon C E. A Mathematical Theory of Communication [J]. Reprinted with corrections from The Bell System Technical Journal, 1948, 27: 379-423, 623-656

[16] Brynjolfsson E, Hu Yu, Smith M D. Long Tails vs. Superstars: The Effect of Information Technology on Product Variety and Sales Concentration Patterns [J]. Information Systems Research, 2010, 21(4): 736-747

[17] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749

[18] Sarwar B, Karypis G, Konstan J, et al. Item-Based Collaborative Filtering Recommendation Algorithm [C] // Proceedings of WWW10. Hong Kong, 2001

(上接第 241 页)

[5] Liang J Y, Wang F, Dang C Y, et al. An efficient rough feature selection algorithm with a multi-granulation view [J]. International Journal of Approximate Reasoning, 2012, 53(6): 912-926

[6] 陈玉明，吴克寿，谢荣生. 基于相对知识粒度的决策表约简[J]. 山东大学学报：工学版，2012，42(6)：8-12

[7] Qian Y H, Liang J Y, Pedrycz W, et al. Positive approximation: an accelerator for attribute reduction in rough set [J]. Artificial Intelligence, 2010, 174(9/10): 597-618

[8] 葛浩，李龙澍，杨传健. 新的可分辨矩阵及其约简方法[J]. 控制与决策，2010，25(12)：1891-1895

[9] 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

[10] 马胜蓝，叶东毅. 信息熵最小约简问题的若干随机优化算法[J]. 模式识别与人工智能，2012，25(1)：96-104

[11] 钱文彬，杨炳儒，徐章艳，等. 基于信息熵的核属性增量式高效更新算法[J]. 模式识别与人工智能，2013，26(1)：42-49

[12] 蒋云良，杨章显，刘勇. 不协调信息系统快速属性分布约简方法[J]. 自动化学报，2012，38(3)：382-388

[13] 钱进，叶飞跃，孟祥萍，等. 一种基于新的条件信息量的属性约简算法[J]. 系统工程与电子技术，2007，29(12)：2154-2157

[14] 黄国顺，刘云生. 不一致决策表信息熵约简与代数约简的核计算与转化[J]. 小型微型计算机系统，2008，29(2)：308-312

[15] 黄国顺，刘云生. 不一致决策表各种属性约简的不一致性分析与转化[J]. 小型微型计算机系统，2008，29(4)703-708

[16] 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

[17] Blake C L, Merz C J. UCI Repository of Machine Learning Databases, University of Califonia at Irvine[OL]. http://www.ics.uci.edu/~mlearn/MLRepositioy.html, 1998

[18] Liang J Y, Qian Y H. Information granularity and entropy theory in information system [J]. Science in China, (F), 2008, 51(10): 1427-1444