面向非完备决策表的正向近似特征选择加速算法

钱宇华 梁吉业 王 锋

(计算智能与中文信息处理教育部重点实验室 太原 030006)

(山西大学计算机与信息技术学院 太原 030006)

摘 要 正向近似是刻画目标概念组成结构的一种有效方法.文中针对非完备决策表现有特征选择算法计算耗时过大的缺陷,提出了一种基于正向近似的通用特征选择加速算法.该算法不仅对候选属性具有保序性,而且通过在特征选择过程中减少样本数据的规模来降低计算耗时,加速特征选择过程.实验结果进一步验证了加速算法的有效性和高效性.特别指出的是,随着属性的增多和数据量的增大,加速算法的性能通常会更好,可有效应用于海量数据的特征选择.

关键词 特征选择;非完备决策表;粗糙集;正向近似

中图法分类号 TP301 **DOI 号**: 10.3724/SP.J.1016.2011.00435

A Positive-Approximation Based Accelerated Algorithm to Feature Selection from Incomplete Decision Tables

QIAN Yu-Hua LIANG Ji-Ye WANG Feng

(*Key Laboratory for Computation Intelligence & Chinese Information Processing of Ministry of Education, Taiyuan* 030006)

(*School of Computer & Information Technology, Shanxi University, Taiyuan* 030006)

Abstract Positive approximation is an effective approach to characterizing the structure of a target concept in information systems. To overcome the limitation of time-consuming of all existing feature selection algorithms in incomplete decision tables. This paper provides a general accelerated algorithm based on the positive approximation. This modified algorithm both possesses the rank preservation of attributes and reduces the time consumption through reducing the scale of data, which effectively accelerates the process of feature selection in incomplete decision tables. Experimental analyses verify the validity and efficiency of the accelerated algorithm. It is deserved to point out that the performance of these modified algorithms are getting better in time reduction with the data set becoming larger.

Keywords feature selection; incomplete decision tables; rough sets; positive approximation

1 引 言

特征选择是数据挖掘与模式识别中的基本问题之一.特征选择通过去除不相关和冗余的属性,能为特定的应用在不失去数据原有价值的基础上选择最小的属性子集.近年来,大规模数据处理问题的不断出现使数据挖掘的发展对大规模数据处理的研究提

收稿日期:2008-12-20;最终修改稿收到日期:2010-08-27.本课题得到国家自然科学基金(71031006,60903110,60773133,70971080)、国家"九七三"重大基础研究发展规划项目基金(2007CB311002)和山西省自然科学基金(2008011038,2009021017-1)资助.**钱宇华**,男,1976 年生,博士研究生,主要研究方向为数据挖掘、粒度计算.E-mail: jinchengqyh@126.com. **梁吉业**(通信作者),男,1962 年生,教授,博士生导师,主要研究领域为粗糙集理论、数据挖掘、人工智能.E-mail: ljy@sxu.edu.cn. **王 锋**,女,1984 年生,博士研究生,主要研究方向为粒度计算.

出了迫切的要求,而特征选择可提高数据的质量,加快数据挖掘的速度[1],因此,特征选择方面的研究引起了数据挖掘领域学者的高度重视.特征选择算法可从搜索方向、搜索策略、评价方法和停止标准等四个方面考察特征的选择.特征选择一般针对两类数据进行研究,即符号型数据与数值型数据.近年来,面向符号型数据的特征选择逐渐受到了人们的关注.特别是,波兰学者 Pawlak 提出的粗糙集理论在符号型数据的特征选择领域中得到了广泛应用[2-7].

应用粗糙集方法进行特征选择的过程,其实质就是对决策表进行属性约简.然而,寻找一个决策表的最小约简已被证明是一个 NP-hard 问题[8],在处理大规模数据集时计算时间代价过大.针对该问题,一些学者提出了许多有效且高效的约简算法.Hu 等[9]给出了一种较好的启发式函数,提出了基于正域的属性约简算法;Wang 等[10]用信息论观点和代数观点对知识约简进行了研究,用条件熵为启发式信息求解决策表的约简;Liu 等[11]提出了一个以区分矩阵为基础的基于属性序的完备算法;Guan 等[12]在等价关系的基础上定义了等价矩阵,通过矩阵的计算来刻画粗糙集计算;Wang 等[13]对属性约简策略进行了分析;Xu 等[14]提出了复杂度为 $\max(O(|C||U|),O(|C|^2|U/C|))$的快速约简算法.一些其他学者也做了相关的分析和讨论[15-18].以上的算法都致力于降低特征选择过程中的时间消耗,提高计算效率.

然而,上述特征选择算法都以完备意义下的决策表为研究对象.现实生活中,信息的非完备(对象的属性值缺损)现象是广泛存在的.比如,在医学诊断智能决策系统中,可能存在这样一组病人,他们不能执行所有要求的检查,因此不能够获得病人(对象)的某些检查结果(属性值).此时,如何从非完备信息系统中获取有用知识就显得更为重要,经典粗糙集理论的等价关系不再适合.于是,完备信息系统被推广到了非完备信息系统[19-20].针对非完备意义下的信息系统或决策表的特征选择,近年来一些作者也做了初步探索[21-22].Liang 等[23]给出了非完备信息系统中粗糙熵的定义,并提出了基于粗糙熵的知识约简算法;Huang 等[24]通过引入信息量来刻画属性的重要度,提出基于信息量的启发式约简算法;Meng 等[25]提出了一种针对非完备决策表属性约简的快速算法.综上分析,对于大规模数据集,现有的基于非完备决策表的约简算法都不同程度地存在耗时较大的问题.因此,非完备决策表中特征选择的加速机制就成为亟待解决的一个关键问题.

Qian 和 Liang 等[26-27]提出的正向近似是一种刻画目标概念的有效方法.在粒度变化过程中,有逐渐细化和逐渐粗糙两种情况.前者主要处理对研究对象刻画和描述过于粗糙、仍需作进一步更精细的刻画的情况;而后者则相反,是处理目前所进行的刻画和描述过于精细、丢失了一些对象的抱团性质、需要使之粗糙一些的情形.用一组具有偏序关系的等价关系族来刻画目标概念称为动态粒度下的粗糙集近似.如果采用的是逐渐细化的思想,称为正向近似;如果采用的是逐渐粗糙的思想,称为逆向近似.Qian 和 Liang 等[28]进一步研究了非完备意义下的正向近似,讨论了非完备意义下如何通过正向近似的方法来刻画粗糙集的粒度结构.动态粒度下的正向近似思想,为粒度计算和粗糙集理论提供了新的研究角度,并且在规则提取和属性约简中也得到了应用.运用正向近似,本文设计了一种通用算法加速器,提出了一种面向非完备决策表的通用特征选择加速算法.该算法利用正向近似思想,通过在启发式迭代中不断减少样本规模来降低计算量,有效地加速了特征子集的求解过程,更适合处理大规模数据集的特征选择.

本文在第 2 节中介绍非完备信息系统及决策表的基本概念;第 3 节中分析非完备意义下正向近似的相关概念及性质;第 4 节选取两种有代表性的启发式信息的度量[9,29],证明正向近似下启发式信息度量的保序性,提出一种面向非完备决策表的正向近似特征选择加速算法;第 5 节,选取 UCI 数据库中的 4 组常用的非完备数据集进行实验分析,结果进一步验证了该加速算法的高效性.值得指出的是,随着数据量的增大,该加速算法的优势更加明显.本文提出的加速算法有效地加快了特征子集的求解过程,降低了大规模数据集特征选择的计算耗时,为海量信息的数据挖掘提供了新方法.

2 基本概念

设 $S=(U,A)$为一个信息系统,其中 U 表示对象的非空有限集合,称为论域;A 表示属性的非空有限集合;对任意 $a\in A$ 有 $a:U\to V_a$,其中 V_a 称为属性 a 的值域;如果至少有一个属性 $a\in A$ 使得 V_a 含有空值,则称 S 为一个非完备信息系统,并用 $*$ 表示空值.

设 $P\subseteq A$,相容关系定义如下[5]:

$$SIM(P)=\{(u,v)\in U\times U\mid \forall a\in P, f(u,a)=f(v,a) \text{ 或 } f(u,a)=* \text{ 或 } f(v,a)=*\}.$$

显然，$SIM(P)=\bigcap_{a\in P} SIM(\{a\})$，令 $S_P(u)$ 表示对象集 $\{v\in U\mid (u,v)\in SIM(P)\}$. $S_P(u)$ 是与 u 可能不可区分的对象的最大集合(相对 P 而言).

设 $U/SIM(P)$ 表示分类或信息粒度，即一簇集合 $\{S_P(u)\mid u\in U\}$. $U/SIM(P)$ 中的元素称为相容类或信息颗粒. $U/SIM(P)$ 中的相容类一般不构成 U 的划分，它们构成 U 的覆盖，即对于每一个 $u\in U$ 有 $S_P(u)\neq\varnothing$，且 $\bigcup_{x\in U} S_P(u)=U$.

设 $X\subseteq U$ 且 $P\subseteq A$，X 的下近似 $\underline{P}X$ 和上近似 $\overline{P}X$ 定义如下[5]：

$$\begin{cases}\underline{P}X=\{u\in U\mid S_P(u)\subseteq X\},\\ \overline{P}X=\{u\in U\mid S_P(u)\cap X\neq\varnothing\}.\end{cases}$$

与完备信息系统一样，$\underline{P}X$ 是肯定属于 X 的对象的集合，而 $\overline{P}X$ 是可能属于 X 的对象的集合，其中 X 的正域为 $POS_P(X)=\underline{P}X$.

对非完备信息系统 $S=(U,A)$，如果 $A=C\cup D$，C 称为条件属性集合，D 表示决策属性集合，且 $C\cap D=\varnothing$，则称非完备信息系统 S 为非完备决策表.

令 $S=(U,C\cup D)$ 是一个非完备决策表，$P\subseteq C$，$U/D=\{D_1,D_2,\cdots,D_r\}$，则 P 相对于 D 的正域定义为 $POS_P(D)=\bigcup_{k=1}^{r}\underline{P}D_k$. 由于 $P\subseteq C$，所以条件属性子集 P 的分类为 $U/SIM(P)=\{S_P(u_1),S_P(u_2),\cdots,S_P(u_{|U|})\}$，如果记目标决策为 $U/SIM(D)=\{S_D(u_1),S_D(u_2),\cdots,S_D(u_{|U|})\}$，则正域的定义也可以等价地描述如下 $POS_P(D)=\{u\mid S_P(u)\subseteq S_D(u), u\in U\}$.

3　正向近似相关概念

正向近似是一种新的集合近似方法，通过粒度序来刻画粗糙集的粒度结构，为粒度计算和粗糙集理论提供了新的研究角度，也逐渐得到了应用[26-27]. 由于本文关注非完备决策表的特征选择，本节对非完备意义下正向近似的定义及其相关性质作简要介绍[28].

令 $S=(U,A)$ 是一个非完备信息系统，在 2^A 上定义偏序关系 $\preceq$：设 $P\preceq Q$(或 $Q\succeq P$)表示 Q 比 P 粗糙(或 P 比 Q 精细)，当且仅当满足对任意 $i\in\{1,2,\cdots,|U|\}$，都有 $S_P(u_i)\subseteq S_Q(u_i)$，如果 $P\preceq Q$ 且 $P\neq Q$ 表示 Q 严格比 P 粗(或 P 严格比 Q 细)记作 $P\prec Q$(或 $Q\succ P$).

定义 1[28]. $S=(U,C\cup D)$ 是一个非完备决策表，$X\subseteq U$，$B=\{P_1,P_2,\cdots,P_n\}$ 是一组属性集且 $P_1\geq P_2\geq\cdots\geq P_n$($P_i\in 2^A$). X 的上近似 $\overline{B}X$ 和下近似 $\underline{B}X$ 分别定义如下：

$$\overline{B}X=\overline{SIM(P_n)}(X),\quad \underline{B}X=\bigcup_{i=1}^{n}\underline{SIM(P_i)}(X_i),$$

其中 $X_1=X$，$X_i=X-\bigcup_{k=1}^{i-1}\underline{SIM(P_k)}(X_k)$，$i=2,\cdots,n$.

我们记 X 的边界域为 $BN_P=\overline{B}(X)-\underline{B}(X)$，记 X 的正域为 $POS_P(X)=\underline{B}(X)$，且 $NEG_P(X)=U-\overline{B}(X)$. 显然，$\overline{B}(X)=POS_P(X)\cup BN_P(X)$.

为了说明在非完备决策表中，正向近似的本质是关注目标概念 X 结构的变化，我们使用以上的相容类重新定义了 X 的正向近似. X 正向近似的上近似 $\overline{B}(X)$ 和下近似 $\underline{B}(X)$ 的结构可表示如下：

$$[\overline{B}(X)]=\{S_{P_n}(u)\mid S_{P_n}(u)\cap X\neq\varnothing, u\in U\},$$
$$[\underline{B}(X)]=\{S_{P_i}(u)\mid S_{P_i}(u)\subseteq X_i, i\leq n, u\in U\},$$

其中 $X_1=X$，$X_i=X-\bigcup_{k=1}^{i-1}\underline{SIM(P_k)}(X_k)$，$i=2,\cdots,n$，且 $[\cdot]$ 表示一个粗糙近似的结构.

定义 2[28]. 设 $S=(U,C\cup D)$ 是一个非完备决策表，$X\subseteq U$，$B=\{P_1,P_2,\cdots,P_n\}$ 是一组属性集且 $P_1\geq P_2\geq\cdots\geq P_n$($P_i\in 2^C$)，且 $U/D=\{D_1,D_2,\cdots,D_r\}$ 是 U 上的一个决策，则 D 相对于 P 的上近似和下近似定义如下：

$$\overline{B}D=\{\overline{B}D_1,\overline{B}D_2,\cdots,\overline{B}D_r\},$$
$$\underline{B}D=\{\underline{B}D_1,\underline{B}D_2,\cdots,\underline{B}D_r\}.$$

$\underline{B}D$ 也称为 D 相对于粒度序 P 的正域，即 $POS_P(D)=\bigcup_{k=1}^{r}\underline{B}D_k$.

4　基于正向近似的特征选择算法

特征选择的目的在于去除多余特征，特征选择是寻找满足一定准则的最优特征子集的过程. 其主要思想是在保持分类能力不变的前提下，导出问题的决策或分类规则. 用于特征选择的方法主要有穷尽搜索、随机搜索和启发式搜索三种，目前最常用的特征选择方法为启发式搜索. 本节将正向近似的方法应用到启发式的特征选择中，提出面向非完备决策表的一种正向近似特征选择加速算法. 本文中选取非完备意义下以正域和 Liang 的条件熵[29] 作为

启发式信息的两种特征选择算法来研究.

4.1 属性重要度的度量

基于上述介绍,非完备决策表中基于正域的属性重要度的度量定义如下.

定义 3. 设 $S=(U,C\cup D)$是一个非完备决策表,$P\subseteq C$. 任意属性 $a\in P$ 的属性重要度定义为

$$SIG_1^{\text{in}}(a,P,D,U)=\gamma_P(D)-\gamma_{P-\{a\}}(D),$$

其中 $\gamma_P(D)=|POS_P(D)|/|U|$.

定义 4. 设 $S=(U,C\cup D)$是一个非完备决策表,$P\subseteq C$. 任意属性 $a\in C-P$ 的属性重要度定义为

$$SIG_1^{\text{out}}(a,P,D,U)=\gamma_{P\cup\{a\}}(D)-\gamma_P(D),$$

其中 $\gamma_P(D)=|POS_P(D)|/|U|$.

在非完备信息系统中,Liang 等[29]定义了一种新的信息熵,并应用于消去冗余特征,其描述如下.

定义 5[29]. 设 $S=(U,A)$是一个非完备信息系统,$U/SIM(A)=\{S_A(u_1),S_A(u_2),\cdots,S_A(u_{|U|})\}$,则 A 的信息熵定义为

$$E(A)=\sum_{i=1}^{|U|}\frac{1}{|U|}\left(1-\frac{|S_A(u_i)|}{|U|}\right),$$

且有 $0\leq E(A)\leq 1-1/|U|$.

基于这个定义,容易给出其在非完备决策表中的条件信息熵.

定义 6. 设 $S=(U,C\cup D)$是一个非完备决策表,$P\subseteq C$,则条件信息熵定义为

$$E(D|P)=\frac{1}{|U|^2}\sum_{i=1}^{|U|}(|S_P(u_i)|-|S_P(u_i)\cap S_D(u_i)|).$$

根据定义 6,基于 Liang 条件熵的相对属性重要度可定义如下.

定义 7. 设 $S=(U,C\cup D)$是一个非完备决策表,$P\subseteq C$. 任意属性 $a\in P$ 的属性重要度定义为

$$SIG_2^{\text{in}}(a,P,D,U)=E(D|P-\{a\})-E(D|P).$$

定义 8. 设 $S=(U,C\cup D)$是一个非完备决策表,$P\subseteq C$. 任意属性 $a\in C-P$ 的属性重要度定义为

$$SIG_2^{\text{out}}(a,P,D,U)=E(D|P)-E(D|P\cup\{a\}).$$

对于给定的非完备决策表,所有约简的交集称为核(可能是空集),核中的每一个属性都是必要的. 为求解决策表的核属性,给出如下的定理.

定理 1. 设 $S=(U,C\cup D)$是一个非完备决策表,$a\in C$,如果 $SIG_j^{\text{in}}(a,C,D)>0(j=1,2)$,则属性 a 是非完备决策表 S 的一个核属性.

基于上述介绍的两种属性重要度的度量,根据定理 1 可求解得到决策表的核属性,将度量信息作为启发式搜索的迭代准则,便可构造启发式的特征选择算法,进而求解决策表的相对约简.

4.2 基于正向近似的特征选择算法

以上介绍了属性重要度的度量,并证明了正向近似下候选属性的属性重要度具有保序性. 在此基础上,本小节给出非完备意义下基于正向近似的启发式特征选择方法,算法主要思想是:对非完备决策表 $S=(U,C\cup D)$,C 为条件属性集,D 为决策属性. 首先计算非完备决策表的核属性集 P,以核属性集为起点,根据正向近似方法逐次从论域中去掉协调部分的对象,记剩下的论域为 U',同时选择使 $EF^{U'}(D|P\cup\{a_0\})$最小的条件属性 $a_0(a_0\in C-P)$添加到 P 中,直到满足终止条件 $EF^{U'}(D|P)=EF^U(D|C)$,则 P 是决策表的约简;其中 $EF^U(D|P)$表示对属性重要度的度量,如在基于正域的特征选择算法中$EF^U(D|P)$表示为$\gamma_P(D)$,而在基于 Liang 的条件熵的特征选择算法中表示为 $E^U(D|P)$.

算法 1. 面向非完备决策表的正向近似特征选择加速算法(IN-FSPA).

输入:非完备决策表 $S=(U,C\cup D)$

输出:特征选择结果 red

过程:

1. 令 $red=\varnothing$;
2. for($j=1$; $j\leq|C|$; $j++$)
   {计算 $Sig^{\text{in}}(a_j,C,D,U)$;
   如果 $Sig^{\text{in}}(a_j,C,D,U)>0$,则 $red=red\cup\{a_j\}$;
   }
3. 令 $i=1$, $P=red$, $U_i=U$;
4. while ($EF^{U_i}(P,D)\neq EF^U(C,D)$) do
   {$i=i+1$;
   $U_i=U_{i-1}-POS_P^{U_{i-1}}(D)$;
   依次计算并选取
   $Sig^{\text{out}}(a_0,red,D,U_i)=\max\{Sig^{\text{out}}(a_j,red,D,U_i),a_j\in C-red\}$;
   $red=red\cup\{a_0\}$;
   $P=red$;
   }
5. 输出特征选择结果 red.

由于求解正域和条件熵要多次计算非完备决策表的相容类,所以本文使用文献[30]中的相容类快速算法来计算相容类.

算法 IN-FSPA 的时间复杂度分析:文献[30]中计算相容类的时间复杂度是 $\max(O(|U||C|),O(\sum_{j=1}^{|C|}\sum_{k=1}^{j-1}|*_k||v_k|))$,其中$|*|$和$|v|$分别表示第 j 个属性下取值为缺省值的对象个数和属性取值个数. 所以步 2 总的时间复杂度是 $O(|U||C|^2+$

$|C|\sum_{j=1}^{|C|}\sum_{k=1}^{j-1}|*_k||v_k|)$；步 4 往核属性集中添加属性并求得约简结果的时间复杂度是 $O(\sum_{i=1}^{|C|}|U_i||C|^2+|C|\sum_{j=1}^{|C|}\sum_{k=1}^{j-1}|*_k^{U_i}||v_k^{U_i}|)$；算法中其余步骤的时间复杂度都为常数，所以算法 IN-FSPA 总的时间复杂度是 $O(|U||C|^2+|C|\sum_{j=1}^{|C|}\sum_{k=1}^{j-1}|*_k||v_k|+\sum_{i=1}^{|C|}|U_i||C|^2+|C|\sum_{j=1}^{|C|}\sum_{k=1}^{j-1}|*_k^{U_i}||v_k^{U_i}|)$. 由此可分析得，在计算过程中随着论域 U_i 的减小，加速算法有效地加速了特征子集的求解过程，减少了时间消耗.

5　实验及分析

为验证算法 IN-FSPA 的高效性和有效性，本文选取了 UCI 数据集中的 4 组非完备数据集进行测试，比较加速后的启发式特征选择算法与加速前算法的计算耗时. 为表示方便，本节中将非完备意义下的基于正域的启发式特征选择算法记为 IN-PR，基于 Liang 熵的算法记为 IN-LCE. 实验中使用的数据集见表 1.

表 1　数据集

	数据集	样本数	条件属性	类别
1	adult	48842	14	2
2	shuttle	58000	9	7
3	ticdata2000	5822	85	2
4	mushroom	8124	22	2

为方便比较加速前后算法的计算时间，首先分别将表 1 中的每组数据平均分为 10 份，记为 $x_i(i=1,2,\cdots,10)$，实验中使用的 10 份数据记为 $X_i(i=1,2,\cdots,10)$，其中 $X_1=x_1$，$X_2=x_1+x_2$，$\cdots$，$X_{10}=x_1+x_2+\cdots+x_{10}$，最后一份数据即是数据集本身. 为了更好地说明算法效果，我们在硬件配置为 CPU Pentium 3.40GHz、内存 1GB 的计算机上，用 C# 语言编程实现算法，开发平台为 Microsoft Visual Studio 2005.

(1) IN-PR 和基于正向近似加速后的 IN-PR (IN-FSPA-PR)

首先比较算法 IN-PR 和 IN-FSPA-PR 分别在 4 组数据集上的时间消耗，实验结果见图 1～4，其中 X 轴表示上述 10 份样本数目由少到多的数据集 X_i $(i=1,2,\cdots,10)$，Y 轴表示算法在不同数据集上的计算时间(单位：min).

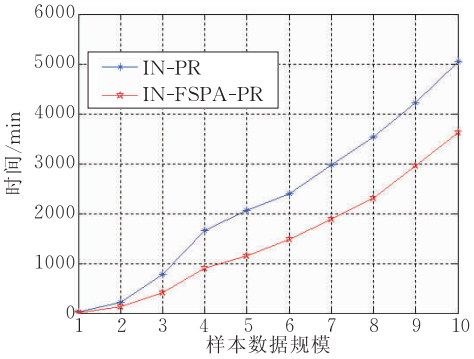

图 1　IN-PR 和 IN-FSPA-PR 不同数据集下的计算时间(数据集 adult)

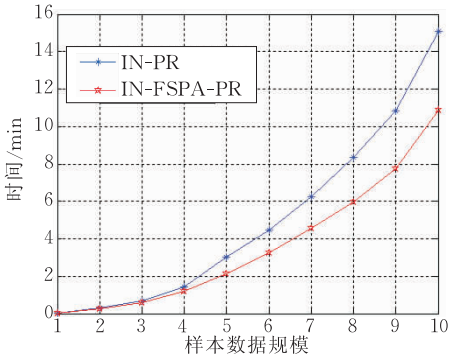

图 2　IN-PR 和 IN-FSPA-PR 不同数据集下的计算时间(数据集 shuttle)

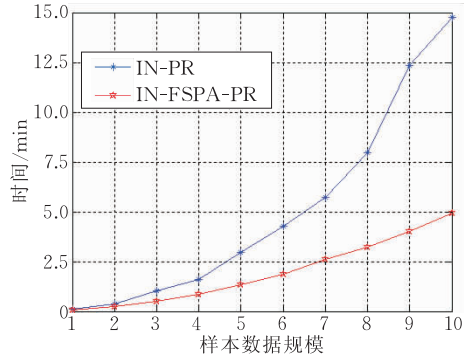

图 3　IN-PR 和 IN-FSPA-PR 不同数据集下的计算时间(数据集 ticdata2000)

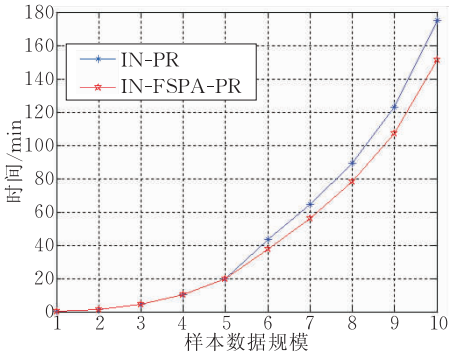

图 4　IN-PR 和 IN-FSPA-PR 不同数据集下的计算时间(数据集 mushroom)

图 1～4 分别表示 4 组数据集上算法 IN-PR 和加速后的算法 IN-FSPA-PR 的计算时间. 如图中实验结果所示,通常随着样本数据规模的增大,时间消耗的增多,两种算法的计算时间的差值会增大. 所以实验结果表明加速算法不仅有效地加速了特征选择的求解过程,而且通常对规模较大的数据集,新算法的性能会更好,高效性也更明显. 实验中图 1～4 的实验结果都比较有效,很好地体现了加速算法的高效性. 其中图 4 中前几份数据在两种算法上的计算时间比较接近,这主要与计算过程中约简结果和核属性的数目有关. 所以,加速算法 IN-FSPA-PR 与原算法 IN-PR 相比较,有效地提高了算法效率,降低了计算量并减少了时间消耗;而且通常对大规模数据集,新算法的高效性更明显.

(2) IN-LCE 和基于 Liang 熵加速后的 IN-LCE (IN-FSPA-LCE)

算法 IN-LCE 和 IN-FSPA-LCE 在 4 组数据集上的时间消耗见图 5～8,其中 X 轴表示上述 10 份样本数目由少到多的数据集 $X_i(i=1,2,\cdots,10)$, Y 轴表示算法在不同数据集上的计算时间(单位: min).

图 5～8 表示原算法 IN-LCE 和加速算法 IN-FSPA-LCE 在 4 组数据集上的计算时间. 图中的

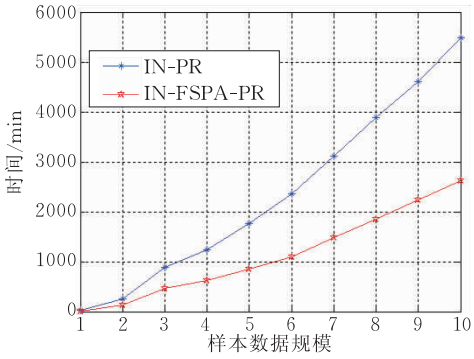

图 5 IN-LCE 和 IN-FSPA-LCE 不同数据集下的计算时间(数据集 adult)

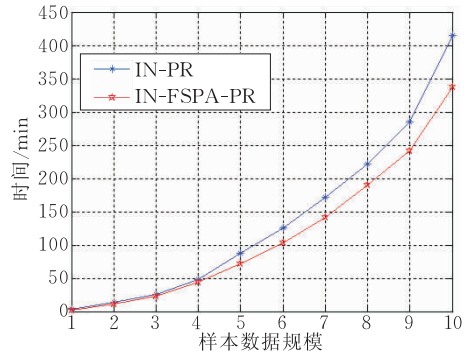

图 6 IN-LCE 和 IN-FSPA-LCE 不同数据集下的计算时间(数据集 shuttle)

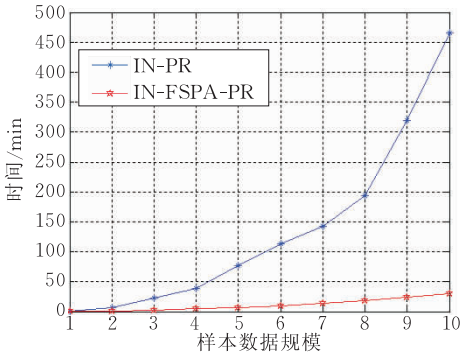

图 7 IN-LCE 和 IN-FSPA-LCE 不同数据集下的计算时间(数据集 ticdata2000)

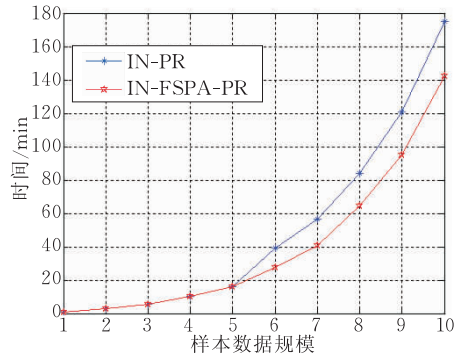

图 8 IN-LCE 和 IN-FSPA-LCE 不同数据集下的计算时间(数据集 mushroom)

实验结果表明,通常随着样本数据规模的增大和时间消耗的增多,两种算法时间消耗的差值会逐渐增大. 所以实验结果进一步有效地验证了加速算法的高效性,加速算法不仅加速了特征选择的求解过程,而且通常随着数据规模的增大,算法的高效性会更明显. 所以加速算法可以更有效地处理大规模数据集的特征选择. 实验中的实验结果都比较有效,很好地体现了加速算法的高效性. 图 8 中前几份数据集上两种算法计算时间的差值不太明显,原因主要受计算过程中约简结果和核属性的数目的影响,但随着数据规模的增大,加速算法便有效地降低了时间消耗. 所以加速算法 IN-FSPA-LCE 相比较原算法,有效地提高了计算效率,且针对大规模数据集性能会更好.

综上分析,加速算法 IN-FSPA 与原算法相比较,作以下几点分析:

(1) 由于基于正向近似的加速算法对候选属性具有保序性,所以加速后的算法不改变原算法的特征选择结果;

(2) 新算法在得到相同特征选择结果的同时有效地加速了特征选择过程,降低了时间消耗;

(3) 随着样本数据规模和属性的增多,加速算

法与原算法时间消耗的差值会增大，高效性更明显，但是，对于核属性集即是特征选择结果的数据集不适合于本算法.

6 结　论

针对现有面向非完备决策表的特征选择算法计算耗时过大的缺陷，本文构建了一个通用特征选择算法加速器.通过分析非完备正向近似的机理，证明了基于正域的和基于 Liang 的条件熵的两个启发式度量的保序性.基于这个保序性质，提出了一种基于正向相似的通用特征选择加速算法.该算法通过不断地缩小论域的规模，有效地降低了特征选择求解中的计算耗时，实现了对特征选择算法的加速.进一步，我们选取了 UCI 数据库中 4 组常用的非完备数据集进行实验分析，结果有效地验证了该加速算法的有效性和高效性.特别指出的是，通常随着数据规模的增大，加速算法 IN-FSPA 的高效性会更明显，可有效应用于海量数据的特征选择中.

参　考　文　献

[1] Liu Huan, Motoda H. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, 1998

[2] Pawlak Z. Rough Sets Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, 1991

[3] Zhang Wen-Xiu, Leung Yee, Wu Wei-Zhi. Information System and Knowledge Discovery. Beijing: Science Press, 2003: 7-12(in Chinese)
(张文修，梁怡，吴伟志. 信息系统与知识发现. 北京：科学出版社，2003：7-12)

[4] Zhang Wen-Xiu, Wu Wei-Zhi, Liang Ji-Ye, Li De-Yu. Rough Sets Theory and Method. Beijing: Science Press, 2005: 3-39, 206-211(in Chinese)
(张文修，吴伟志，梁吉业，李德玉. 粗糙集理论与方法. 北京：科学出版社，2005：3-39，206-211)

[5] Liang Ji-Ye, Li De-Yu. Uncertainty and Knowledge Acquisition in Information Systems. Beijing: Science Press, 2005: 37-46(in Chinese)
(梁吉业，李德玉. 信息系统中的不确定性与知识获取. 北京：科学出版社，2005：37-46)

[6] Liu Qing. Rough Sets and Rough Reasoning. Beijing: Science Press, 2001(in Chinese)
(刘清. Rough 集及 Rough 推理. 北京：科学出版社，2001)

[7] Liang Ji-Ye, Qian Yu-Hua. Information granular and entropty theory in information systems. Science in China Series F: Information Sciences, 2008, 51(10): 1427-1444

[8] Wong S K M, Ziarko W. On optimal decision rules in decision tables. Bulletin of Polish Academy of Sciences, 1985, 33(11-12): 693-696

[9] Hu Xiao-Hua, Cercone N. Learning in relational databases: A rough set approach. International Journal of Computational Intelligence, 1995, 11(2): 323-338

[10] Wang Guo-Yin, Yu Hong, Yang Da-Chun. Decision table reduction based on conditional information entropy. Chinese Journal of Computers, 2002, 25(7): 759-766(in Chinese)
(王国胤，于洪，杨大春. 基于条件信息熵的决策表约简. 计算机学报，2002，25(7)：759-766)

[11] Liu Shao-Hui, Sheng Qiu-Jian, Wu Bin, Shi Zhong-Zhi, Hu Fei. Research on efficient algorithms for Rough set methods. Chinese Journal of Computers, 2003, 26(5): 524-529(in Chinese)
(刘少辉，盛秋戬，吴斌，史忠植，胡斐. Rough 集高效算法的研究. 计算机学报，2003，26(5)：524-529)

[12] Guan Ji-Wen, Bell David A, Guan Z. Matrix computation for information systems. Information Sciences, 2001, 131: 129-156

[13] Wang Jue, Miao Duo-Qian. Analysis on attribute reduction strategies of rough set. Journal of Computer Science and Technology, 1998, 13(2): 189-192

[14] Xu Zhang-Yan, Liu Zuo-Peng, Yang Bing-Ru, Song Wei. A quick attribute reduction algorithm with complexity of $\max(O(|C||U|), O(|C|^2|U/C|))$. Chinese Journal of Computers, 2006, 29(3): 391-399(in Chinese)
(徐章艳，刘作鹏，杨炳儒，宋威. 一个复杂度为 $\max(O(|C||U|), O(|C|^2|U/C|))$ 的快速属性约简算法. 计算机学报，2006，29(3)：391-399)

[15] Zhang Wen-Xiu, Mi Ju-Sheng, Wu Wei-Zhi. Knowledge reductions in inconsistent information systems. Chinese Journal of Computers, 2003, 26(1): 12-18(in Chinese)
(张文修，米据生，吴伟志. 不协调目标信息系统的知识约简. 计算机学报，2003，26(1)：12-18)

[16] Yang Ming. An incremental updating algorithm for attribute reduction based on improved discernibility matrix. Chinese Journal of Computers, 2007, 30(5): 815-822 (in Chinese)
(杨明. 一种基于改进差别矩阵的属性约简增量式更新算法. 计算机学报，2007，30(5)：815-822)

[17] Ye Dong-Yi. An improvement to Jelonek's attribute reduction algorithm. Acta Electronic Sinica, 2000, 28(12): 81-82 (in Chinese)
(叶东毅. Jelonek 属性约简算法的一个改进. 电子学报，2000，28(12)：81-82)

[18] Han Su-Qing, Wang Jue. Second attribute algorithm based on tree expression. Journal of Computer Science and Technology, 2006, 21(3): 383-392

[19] Kryszkiewicz M. Rough set approach to incomplete information systems. Information Sciences, 1998, 112: 39-49

[20] Slowinskir R, Vsnderprooten D. A generalized definition of rough approximations based on similarity. IEEE Transactions on Data and Knowledge Engineering, 2000, 12(2): 331-336

[21] Leung Yee, Wu Wei-Zhi, Zhang Wen-Xiu. Knowledge acquisition in incomplete information systems: A rough set approach. European Journal of Operational Research, 2006 (68): 164-183

[22] Sun Hui-Qin, Zhang Xiong, Finding minimal reducts from incomplete information systems//Proceedings of the International Conference on Machine Learning and Cybernetics. 2003: 350-354

[23] Liang Ji-Ye, Xu Zong-Ben. The algorithm on knowledge reduction in incomplete information systems. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 2002, 10(1): 95-103

[24] Huang Bing, Zhou Xian-Zhong, Zhang Rong-Rong. Attribute reduction based on information quantity under incomplete information systems. Systems Engineering-Theory and Practice, 2005, 4(4): 55-60(in Chinese)
(黄兵, 周献中, 张蓉蓉. 基于信息量的不完备信息系统属性约简. 系统工程理论与实践, 2005, 4(4): 55-60)

[25] Meng Zu-Qiang, Shi Zhong-Zhi. A fast approach to attribute reduction in incomplete decision systems with tolerance relation-based rough sets. Information Sciences, 2009, 179: 2774-2793

[26] Liang Ji-Ye, Qian Yu-Hua, Chu Cheng-Yuan, Li De-Yu, Wang Jun-Hong. Rough set approximation based on dynamic granulation//Lecture Notes in Artificial Intelligence 3641, 2005: 701-708

[27] Qian Yu-Hua, Liang Ji-Ye, Dang Chuang-Yin. Converse approximation and rule extraction from decision tables in rough set theory. Computers and Mathematics with Applications, 2008, 55: 1754-1765

[28] Qian Yu-Hua, Liang Ji-Ye. Positive approximation and rule extracting in incomplete information systems. International Journal of Computer Science and Knowledge Engineering, 2008, 2(1): 51-63

[29] Liang Ji-Ye, Shi Zhong-Zhi, Li De-Yu, Wireman M J. The information entropy, rough entropy and knowledge granulation in incomplete information systems. International Journal of General Systems, 2006, 34(1): 641-654

[30] Wang Feng, Liang Ji-Ye, Qian Yu-Hua. Quick computation for tolerant classes from incomplete information systems. Computer Engineering and Applications, 2009, 45(27): 133-136(in Chinese)
(王锋, 梁吉业, 钱宇华. 非完备信息系统的相容类快速计算. 计算机工程与应用, 2009, 45(27): 133-136)

QIAN Yu-Hua, born in 1976, Ph. D. candidate. His research interests include data mining and granular computing.

LIANG Ji-Ye, born in 1962, professor, Ph. D. supervisor. His research interests include rough set theory, data mining and artificial intelligence.

WANG Feng, born in 1984, Ph. D. candidate. Her research interests focus on granular computing.

Background

Feature selection, also called attribute reduction, is a challenging problem in such areas as pattern recognition, machine learning and data mining. It has been proven that finding the minimal reduct of a decision table is a NP hard problem, to overcome the limitation of time-consuming, many kinds of attribute reduction have been developed in rough set theory. But most of them are study the decision tables under complete decision tables, how to reduce the time consumption of feature selection algorithms in incomplete decision tables still is a problem need to be resolved. This paper provides a general accelerated algorithm based on the positive approximation. This modified algorithm both possesses the rank preservation of attributes and reduces the time consumption through reducing the scale of data, which effectively accelerates the process of feature selection in incomplete decision tables.

This work was partially supported by the National Natural Science Foundation of China (Nos. 71031006, 60903110, 60773133, 70971080), the National Basic Research Program of China (973 Program) of China (No. 2007CB311002), the Natural Science Foundation of Shanxi Province of China (Nos. 2008011038, 2009021017-1).