具有非一致性数据预处理的粗糙集特征选择算法*)

袁 赣 张 巍 蔡庆生

(中国科学技术大学计算机科学与技术系 合肥230027)

摘 要 大多数特征选择算法面临着对非一致性数据缺乏有效的处理的问题。本文提出了一种处理非一致性数据的方法,采用阈值将非一致性数据做归类处理,当某一类非一致性数据的某个取值比例超过了该阈值,则该类数据都取该值,并只保留一条记录。在此基础上,本文提出了一种改进的基于粗糙集理论的特征选择算法。

关键词 特征选择,粗糙集,信息论,非一致性数据

A Rough Set Feature Selection Approach Dealing with Inconsistent Data

YUAN Gan ZHANG Wei CAI Qing-Sheng

(Department of Computer Science and Technology, University of Science and Technology of China, HeFei 230027)

Abstract Most of feature selection algorithms can not deal with inconsistent data. This article constructs an approach to combine inconsistent data to consistent data. When the ratio of one of the values of a kind of inconsistent data is more than a threhold, the kind of inconsistent data takes this value and only one data is kept. Based on this, the article gives a rough set feature selection approach which can deal with inconsistent data.

Keywords Feature selection, Rough set, Information theory, Inconsistent data

1 前言

数据库的建立是为了更有效地管理信息资源。然而,所存贮的数据往往含有大量冗余或者不完整的属性,严重降低了数据挖掘算法的时间效率和算法质量。如何删除冗余,是一个极其具有挑战性的工作。这就是特征选择所需要完成的工作。

对于数据中存在的非一致性数据,以往的特征选择算法要么将其删除,要么采用了极其简略的处理办法。本文采用阈值,将非一致性数据作出合理的归类处理,当某一类非一致性数据的某个取值比例超过了该阈值,则该类数据都取该值,并只保留一条记录。其思想超过了将非一致性数据删除或者简单的归为一个新类的想法[6]。

近来,粗糙集理论在特征选择算法中得到了广泛的应用。粗糙集理论的特点是不需要预先给定某些特征的数量描述,而是直接从给定问题的描述集合出发,通过不可分辨关系和不可分辨类确定给定问题的近似域,从而找出该问题的内在规律[5]。

在非一致性数据处理的基础上,本文提出了一种可以有效处理非一致性数据的粗糙集特征选择算法。对于该算法得到的多个简约,利用信息论基础[2],找出属性之间关联性最小的简约作为最终得到的结果[1]。

本文第2节介绍了粗糙集的基本概念与理论,第3节介绍了非一致性数据预处理的方法;第4节介绍相关的信息论基础,第5节介绍具有非一致性数据预处理的粗糙集特征选择算法,第6节举例阐述该算法,第7节进行算法的实验与比较,最后给出本论文的结论。

2 粗糙集

粗糙集理论的出发点是,根据目前已有的对给定问题的知识的论域进行划分,然后对划分后的每一个组成部分确定其对某一概念的支持程度:肯定支持,肯定不支持,可能支持,分别用三个近似集合表示为正域、负域、边界。

定义1 称 $S=(U,A,C,D)$ 为决策系统,其中 U 为非空有限集,称为论语;A 为属性集合;A 由条件属性 C 和决策属性 D 组成,$C\subset A$, $D\subset A$, 且 $C\cup D=A$, $C\cap D=\Phi$。

定义2 对于决策系统,令 $a\in A$, $P\subseteq A$, 二元关系 $IND(P)$ 称为不可分辨关系,如下定义:

$IND(P)=\{(x,y)\in U\times U:\forall a\in P, a(x)=a(y)\}$,用 $U/IND(P)$ 代表二元关系 $IND(P)$ 的所有等价类,简写为 U/P。

定义3(下近似) 令 $R\subseteq C$, $X\subseteq U$, X 的 R 下近似集,是通过知识 R 能肯定划归到集合 X 中的 U 的所有元素的集合。可形式地定义为:

$\underline{R}X=\{Y\in U/R: Y\subseteq X\}$

定义4(正域) D 的 C 正域是指通过属性集 C 能肯定划归到 U/D 的等价类的 U 中元素集合。形式地定义为:

$$POS_C(D)=\bigcup_{X\in U/D}\underline{C}X$$

定义5(必要属性和不必要属性) 令 $c\in C$, c 是不必要的,如果 $POS(C-\{c\})(D)=POS_C(D)$; 否则 c 是必要的。

定义6(简约,REDUCT) 属性集 $R\subseteq C$ 称为 C 的简约,如果 $T=(U,A,R,D)$ 是独立的,并且 $POS_R(D)=POS_C(D)$。

定义7(核,CORE) C 的所有必要属性所构成的集合称

*)本文得到国家科学基金(No. 701710525和 No. 60075015)的资助。袁 赣 硕士研究生,研究方向为数据挖掘和知识发现。张 巍 博士研究生,研究方向为数据挖掘和知识发现、COM 组件和 XML 通用数据交换技术、MIS。蔡庆生 教授,博士生导师,主要研究方向为人工智能,机器学习,知识发现。

为核，记为 $CORE(C)$。有如下性质

$CORE(C)=\bigcap RED(C)$

3 非一致性数据的预处理

在数据的收集过程中，会导致不一致数据的出现。在决策系统中，表现为论域中元素具有相同的条件属性值，可是具有不同的决策属性值，也即属于不同的类。如表1所示。

定义8（非一致性数据的预处理规则） 给定一个决策系统 $S=(U,A,C,D)$，当有非一致性数据出现的时候，按照如下规则构建一个新的决策系统 $S'=(U',A,C,D)$：给定一个阈值，当具有相同条件属性值的元素具有不同的决策属性值时，如某一个决策属性值的取值比例超过了该阈值时，则该类元素的决策属性即取该决策属性值，并在新的决策表中只保留一条数据。

如表1所示，设定阈值为60%，x_1,x_8,x_9三元素为非一致性数据，决策属性 E 的取值分别为1，1，0，取值为1的比例超过了阈值60%。故而删去 x_8,x_9，留下 x_1，其决策属性取值为1，得到新的决策系统如表2。

表1

	a	b	c	d	E
x1	1	0	2	1	1
x2	1	0	2	0	1
x3	1	2	0	0	2
x4	1	2	2	1	0
x5	2	1	0	0	2
x6	2	1	1	0	2
x7	2	1	2	1	1
x8	1	0	2	1	1
x9	1	0	2	1	0

表2

	a	b	c	d	E
x1	1	0	2	1	1
x2	1	0	2	0	1
x3	1	2	0	0	2
x4	1	2	2	1	0
x5	2	1	0	0	2
x6	2	1	1	0	2
x7	2	1	2	1	1

这种合并规则的主要思想是，考虑到具有相同条件属性的元素取多值的情况，当某一取值超过一定的比例，就认为该类元素属于该类。在一个极端的情况，具有相同条件属性的元素共有100条，其中90条属于类1，5条属于类2，则理所当然的认识该类元素属于类1。有一些处理这种非一致性数据的过程，简单的将它归为不属于任一类别的 * 类，显然是不合乎情理的[6]。

4 信息论基础

定义9 属性集 P 的信息熵 $H(P)$ 定义为：

$H(P)=-\sum_{i=1}^{n}p(x_i)\log(p(x_i))$，其中，$p(x_i)=card(x_i)/card(U)$，$X=U/P=\{x_1,x_2,\cdots,x_n\}$。

定义10 知识 $Q(U/Q)=\{Y_1,Y_2,\cdots,Y_m\}$ 对于知识 $P(U/P)=\{X_1,X_2,\cdots,X_n\}$ 的条件信息熵定义为：

$H(Q|P)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))$，其中 $P(Y_j/X_i)$ 为条件概率。$P(Y_j/X_i)=card(X_i\bigcap Y_j)/card(U)$。

当条件信息属性越小，说明两者之间的关联性越小。

5 具有非一致性数据预处理的粗糙集特征选择算法

在粗糙集理论中，从区别矩阵出发来计算 CORE 和 REDUCT 是一种比较简洁且效率较高的算法。所谓区别矩阵，就是将论域中元素按照行和列排列，矩阵中元素就是将属于不同类别的行列元素区别开来的属性集合。当只需要一个属性就可以将属于不同类的元素区别开来时，该元素就为 CORE。矩阵中不包括 CORE 中任意属性的元素采用布尔联结，展开后，得到的不同的表达式，再加上 CORE 中属性，即为 REDUCT。事实上，REDUCT 是一个属性集，它至少包含区别矩阵中每个非空元素的一个属性[3]。

算法 具有非一致性数据预处理的粗糙集特征选择算法

设 U 是论域，$THREHOLD$ 是非一致性数据归并处理门限值。

STEP1 非一致性数据归并处理

```
for( row1=0;row1<最大论域元素数目 && 该元素未被处理;row1++)
{ for(row2=row1+1;row2<最大论域元素数目 && 该元素未被处理;row2++)
    if(row1元素与row2元素条件属性值都相同)
    { 标记row2元素为已处理
      if(row2的决策属性值已出现过)
        该值出现的的次数加1
      else
        该值出现的次数设为1
    }
  计算与row1元素具有相同条件值的所有元素的决策属性的取值比例，比例最大的值超过设定的门限值 THREHOLD 时，该值设为 row1元素的决策属性值，并删除其余的该类元素。
}
```

STEP2 计算区别矩阵。令 $T=(U,A,C,D)$ 是一个决策表，其中 $U=\{x_1,x_2,x_3,\cdots x_n\}$

T 的区别矩阵 $M(T)$，定义 $n\times n$ 矩阵：

$m_{ij}=\{a\in C,a(x_i)!=a(x_j)$ 且 $(d\in D,d(x_i)!=d(x_j))\}$，$i,j=1,2,\cdots\cdots,n$

m_{ij} 是将 x_i 与 x_j 分类到不同类别的属性的集合。

STEP3 计算 *CORE*

CORE 是区别矩阵中单个元素的集合：

$CORE=\{a\in C:m_{ij}=\{a\}$，对于某个 $i,j\}$

实际上，*CORE* 中元素就是靠单个属性就能区别 U 中元素的属性集合。

STEP4 计算 *REDUCT*

矩阵中不包括 *CORE* 中任意属性的元素采用布尔联结，展开后，得到的不同的表达式，再加上 *CORE* 中属性，即为 *REDUCT*。

令任意 $P_i=(a_{i1},a_{i2},\cdots,a_{ij})$ 是 $M(T)$ 中元素，且 $P_i\bigcap CORE=$ 空集，$Q=\wedge(a_{i1}\vee a_{i2}\vee\cdots\vee a_{ij})$，将 Q 展开化简，得到析取式 $Q_1,Q_2,\cdots,Q_n$。

得到 *REDUCT*：$RED_i=Q_i\bigcup CORE$ $(i=1,\cdots,n)$。

STEP5 选取 *REDUCT*

1)取出属性个数最小的 *REDUCT*，如只有一个，则该 *REDUCT* 即为最终所求的 *REDUCT*

2)属性个数最小的 *REDUCT* 有多个时，计算每个 *REDUCT* 的任意两个属性之间的条件信息熵(参见定义10)

3)对于任意的 *REDUCT*，计算平均条件信息熵

4)平均信息熵最小的 *REDUCT* 即为最终所求的 *REDUCT*

6 举例

如表1，令预处理门限值为60%，经过非一致性数据欲处理后，得到表2。

表2的区别矩阵为

表3

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1							
x_2							
x_3	b,c,d	b,c					
x_4	b	b,d	c,d				
x_5	a,b,c,d	a,b,c		a,b,c,d			
x_6	a,b,c,d			a,b,c,d			
x_7		a,b	c,d,a,b	cd	c,d		

从表3可知，$CORE=\{b\}$，$Q=c\vee d$

得到 $RED_1=\{b,c\}$，$RED_2=\{b,d\}$

$U/b=\{\{x_1,x_2\},\{x_3,x_4\},\{x_5,x_6,x_7\}\}$

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

$U/d=\{\{x_1,x_4,x_7\},\{x_2,x_3,x_5,x_6\}\}$

$H(c|b)=0.302$，$H(d|b)=0.221$

取信息熵最小的 $H(d|b)$，可知，最终所的的 *REDUCT* $=\{b.d\}$。

7 实验

KAN LI，YU-SU LIN 提出的算法[1]，不能对包含有非一致性元素的数据进行处理，在进行运算前，必须确保论域中数据为一致性的。本算法与之相比，能对非一致性数据进行预处理，然后进行特征选择。而且本算法与普通的算法比，也较为简洁易懂。

以下选择 UCI 标准数据集进行测试，并与文[1]中算法结果进行比较，文[1]中算法明确的指出不可以处理非一致性数据，故而文[1]算法只取一致性数据的试验结果。其中非一致性数据预处理门限值为60%。

从以上比较可看出，本算法扩展了计算数据集的范围，对于一致性数据取得和文[1]算法大致相同的效果，而本算法还可以对包含非一致性数据的数据集进行有效的处理。显然，本论文所提出的算法要优越。

结论 本论文提出了一种具有非一致性数据预处理的粗糙集特征选择算法。利用非一致性数据处理得到一致性论域后，用粗糙集方法得到 CORE 和多个可能的 REDUCT。然后用信息论知识计算各 REDUCT 中各元素的相关性，平均信息熵最小说明各元素的相关性小，取该 REDUCT 作为最后的属性集。通过实验，证明本算法具有合理的非一致性数据处理方法，选取 REDUCT 具有较好的效果。

数据集	实例个数	非一致性实例个数	属性个数	REDUCT 个数	
				本算法	文[1]算法
Ballon(1)	16		4	2	2
Ballon(2)	16		4	2	3
Ballon(3)	16		4	2	2
Ballon(4)	16		4	4	4
tic-tac-toe	958		9	8	8
Chess(kr vs kp)	3196		36	29	32
Balance scale	625	190	4	4	
Mushroom	8124	2684	22	12	
Postoperative Patient	90	6	8	8	
Restricted (primary-tumor)	339	27	17	16	
restricted (breast-cancer)	286	6	9	8	

在计算 REDUCT 时，虽然方法简单，但是当实例特别多或者属性个数比较多时，从区别矩阵计算 REDUCT 仍然有计算量大的问题，采用新的方法，比如启发式等别的方法是今后研究的重点[4]。

参考文献

1 LI K, LIN Y S. Rough set based attribute reduction approach in data mining. In: Proc. of the First Intl. Conf. on Machine Learning and Cybernetics, Beijing, China, Nov. 2002

2 Miao Duoqian, Wang Jue. Information-based algorithm for reduction of knowledge. In: 1997 IEEE Intl. Conf. on Intelligent Proceeding Systems, Beijing, China, Oct. 1997

3 Questier F, Arnaut-Rollier I, Walczak B, Massart D L. Application of rough set theory to feature selection for unsupervised clustering. Chemometrics and Intelligent Laboratory Syatems, 2002, 63: 155～167

4 Zhong N, Dong J, Ohusuga S. Using Rough Sets with Heuristics for Feature Selection, Journal of Intelligent Information Systems, 2001, 16: 199～214

5 Pawlak Z, Rough Sets. International Journal of Computer and Information Sciences, 1982, 11: 341～356

6 Zheng Zheng, Wang Guoying, Yu Wu. Object's Combination Based Simple Computation of Attribute Core. In: Proc. of the 2002 IEEE Intl. Symposium on Intelligent Control, Vancouver, Canana, Oct. 2002

7 张祥德，张巍，刘玉蓉. 数据挖掘分类问题的贪婪粗糙集约简算法. 东北大学学报(自然科学版)，2001，22(5)

(上接第195页)

3 吴浩扬，常炳国，朱长纯，刘君华. 基于模拟退火机制的多种群并行遗传算法. 软件学报，2000，11(3)：416～420

4 Herrera F, Lozano M. Heterogeneous Distrubuted Genetic Algorithms Based on the Crossover Operator. Genetic Algorithms In Engineering Systems: Innovations And Applications, 1997. GALESIA 97. Second Intl. Conf. On (Conf. Publ. No. 446), Sept. 1997. 203～208

5 Fung C C, Chow S Y, Wong K P. Solving the Economic Dispatch Problem with an Integrated Parallel Genetic Algorithm, Power System Technology, 2000. In: Proc. PowerCon 2000. Intl. Conf. on, Volume: 3, Dec. 2000. 1257～1262

6 杨华军，苏德富. 基于 Windows 95/NT 的 PVM 并行计算平台. 计算机工程，1999，25(2)：24～25

7 王凌著. 智能优化算法及其应用. 清华大学出版社，2001

8 http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95/

9 Grefenstette J, et al. Genetic Algorithms for the Traveling Salesman Problem. In: Proc. of 1st Int. Conf. on Genetic Algorithms and Their Applications, Lawrence Erlbaum Associates, 1985. 154～159

10 Xiong Shengwu, Li Chengjun. A Distributed Genetic Algorithm to TSP. Intelligent Control and Automation, In: 2002. Proc. of the 4th World Congress on, Volume: 3, June 2002. 1827～1830