一种基于决策表的分类规则挖掘新算法

谢娟英　冯德民
（陕西师范大学计算机科学学院　西安710062）

A New Algorithm of Mining Classification Rules Based on Decision Table

XIE Juan-Ying　FENG De-Min
(Computer Science College, Shanxi Normal University, Xi'an, 710062)

Abstract The mining of classification rules is an important field in Data Mining. Decision table of rough sets theory is an efficient tool for mining classification rules. The elementary concepts corresponding to decision table of Rough Sets Theory are introduced in this paper. A new algorithm for mining classification rules based on Decision Table is presented, along with a discernable function in reduction of attribute values, and a new principle for accuracy of rules. An example of its application to the car's classification problem is included, and the accuracy of rules discovered is analyzed. The potential fields for its application in data mining are also discussed.

Keywords Decision table, Rough sets, Data mining, Classification rules

1 引言

分类是通过分析训练集数据，产生关于分类的精确描述。这种类别描述常由分类规则组成，可以用于对未来的数据进行分类预测，有着广泛的应用前景[1]。

Rough Sets 是对不完整数据进行分析、学习的主要方法[2]。该方法只依赖于数据内部的知识，用数据之间的近似来表示知识的不确定性。其中的决策表理论能在条件满足的情况下，根据这些条件描述相应的决策[3]，对相应数据做出类别判断。故可用来分析训练数据集，产生关于分类的精确描述，生成相应的分类规则。但是，现有的文[1，4～6]，在应用粗集的决策表来挖掘分类规则时，所挖掘规则表示为 $des(X_i)\overset{c}{\rightarrow}des(Y_j)$，其中 $X_i\in U/C$，$Y_j\in U/D$，C 是条件属性集，D 是决策属性集，$c=card(X_i\cap Y_j)/card(X_i)$ 是规则的确定性。但这样的规则是没有泛化的。本文将可辨识函数引入到属性值的约简中，以得到泛化的分类规则，并利用决策属性的基本集关于条件属性的上下近似来判断所挖掘分类规则的确定性。并实例说明如何利用该算法在数据库中发现分类规则。

2 相关概念

2.1 信息系统[7]

一个信息系统（近似空间）是一个有序四元组：$IS=(U, A, V, f)$。其中，U 是全域（对象构成的集合，$U=\{x_1, x_2, \cdots, x_m\}$）；$A$ 是属性（特征，变量）集；$V=\bigcup_{a\in A}V_a$ 是属性值的集合，V_a 是属性 a 的值域；$f: U\times A\rightarrow V$ 是一个信息函数，对每一个 $a\in A$，和 $x\in U$ 定义了一个信息函数 $f(x,a)\in V_a$，即信息函数 f 指定 U 中每一个对象 x 的属性值。

2.2 不可辨识关系(Indiscernibility relation)[4]

不可辨识关系是用来表达这样一个事实，由于缺乏一定的知识而不能将已知信息系统中的某些对象区别开[2]。

设 $\forall B\subseteq A$，不可辨识关系 $Ind(B)=\{(x_i,x_j)\mid(x_i,x_j)\in U^2, \forall b\in B(b(x_i)=b(x_j))\}$，表示对象 x_i 和 x_j 关于属性集 A 的子集 B 是不可辨识的。

显然 $Ind(B)$ 是一个等价关系[8]，关于 $Ind(B)$ 的等价类称为 B 的基本集。对 $\forall x_i\in U$，关于 $Ind(B)$ 的等价类记为 $[x_i]_{Ind(B)}$。

2.3 下近似和上近似（Lower and Upper approximations)[5]

设 $X\subset U$，$B\subseteq A$，则 X 关于 B 的下近似 $\underline{BX}$，是所有真包含于 X 的 B 基本集的并。即：$\underline{BX}=\{x_i\in U\mid[x_i]_{Ind(B)}\subseteq X\}$；$X$ 关于 B 的上近似 BX，所有与 X 的交不为空的 B 基本集的并。即：$BX=\{x_i\in U\mid[x_i]_{Ind(B)}\cap X\neq\Phi\}$。

如果 $\underline{BX}=BX$，则集合 X 是可定义集；否则，集合 X 在 U 中是不可定义集，即 Rough 集。

2.4 X 的 B 正域[5]

X 的 B 正域 $POS_B(X)=\underline{BX}$，是所有根据知识 B 能确定地划入集合 X 的 U 中对象的集合。

2.5 近似的精度(Accuracy of approximation)[5]

设 $X\subset U$，$B\subseteq A$，则 X 在空间 B 的精度 $\mu_B(X)=card(\underline{BX})/card(BX)$。通常，$0\leqslant\mu_B(X)\leqslant 1$。但是，若 X 在 U 是可定义的，则 $\mu_B(X)=1$；若 X 在 U 中不可定义，则 $0\leqslant\mu_B(X)<1$。

2.6 属性独立性(Independence of attributes)[6]

如果 $Ind(A)=Ind(A-\{a_i\})$，则属性 a_i 相对于属性集 A 是冗余的。否则属性 a_i 在空间 A 中是独立的。

如果 $\forall a_i\in A, Ind(A-\{a_i\})\neq Ind(A)$，则 A 是独立的。

一个信息系统，去掉冗余属性可得到与原信息系统具有相同特性且简化了的信息系统，因此，属性集之间是依赖的[9]。这是我们在分类规则挖掘中进行属性约简的依据。

2.7 核与属性约简(Core and reduct of attributes)[6]

设 $B\subseteq A$，如果 B 是独立的，且 $Ind(B)=Ind(A)$，则 B 是 A 的一个约简。显然，A 可能存在多种约简。A 的所有必要关系构成的集合是 A 的核 $core(A)$。显然，$core(A)=\bigcap red(A)$。其中 $red(A)$ 表示 A 的所有约简。

2.8 可辨识矩阵与可辨识函数[6]

设 $IS=(U, A, V, f)$ 是一个信息系统，$card(U/Ind(A))=n$，则 IS 的可辨识矩阵是 $n\times n$ 矩阵，其任一元素 $a(x,y)=$

$\{a\in A|f(x,a)\neq f(y,a)\}$。显然，可辨识矩阵是对称阵，因此，仅考虑其下三角部分。

可辨识函数 $f(A)$ 是一个布尔函数，它是可辨识矩阵中所有非空元素的布尔合取，对于 $\forall\ \alpha(x,y)=\{a_1,a_2,\cdots,a_k\}\neq\Phi$，指定一个布尔函数 $a_1\vee a_2\vee\cdots\vee a_k$ 或 $a_1+a_2+\cdots+a_k$，并用 $\Sigma\alpha(x,y)$ 表示。则 $f(A)=\prod_{(x,y)\in(U/Ind(A))\times(U/Ind(A))}\Sigma\alpha(x,y)$。如果，属性集为空，则指定其为布尔常量1。

计算 $f(A)$ 的简化形式，可利用吸收率[8]。通常，一个 IS 的可辨识函数的最后简化形式需要表示为一个析取标准式，代表了一个 IS 的属性集的多个约简，其中的每一个合取式是原 IS 的一个约简 R。即：$Ind(R)=Ind(A)$。

2.9 核与属性值约简(Core and reducts of attribute values)[9]

构造基于约简了的 IS 的可辨识矩阵，以及构造与 IS 的基本集个数相同的可辨识函数。每一个可辨识函数可将相应基本集与其余基本集区别开。则根据 $f_i(A),i=1,2,\cdots,n$ 可将 IS 进一步在属性值上化简。这是本文挖掘分类规则的特点。

2.10 分类[9]

设 $F=\{X_1,X_2,\cdots,X_n\},X_i\subset U$，且 $X_i\cap X_j=\Phi,\cup X_i=U,i=1,\cdots,n$，则 F 称为 U 的分类(划分)，其中，X_i 是类。

F 关于 $B(B\subseteq A)$ 的下近似和上近似分别定义为：

$\underline{B}(F)=\{\underline{B}(X_1),\underline{B}(X_2),\cdots,\underline{B}(X_n)\}$

$B(F)=\{B(X_1),B(X_2),\cdots,B(X_n)\}$

分类 F 的质量定义为：

$\eta_BF=\cup card\ \underline{B}(X_i)/cardU$

分类 F 在 B 中的精度可通过下式计算：

$\beta_BF=\cup card\ \underline{B}(X_i)/\cup cardB(X_i)$

2.11 决策表

决策表是一个信息系统[2,3]，只是在该信息系统中属性被分为两类：条件属性和决策属性。即：$A=C\cup D,C\cap D=\Phi$。其中 C 是条件属性集，D 是决策属性集。$card(D)\geqslant 1$。基于决策表的分类规则挖掘，就是根据决策表找出决策属性与条件属性间的依赖关系。用条件属性来表达决策属性。

2.12 D-冗余属性(D-superfluous attributes)[8]

设 $a_i\in B(B\subseteq C)$，若 $POS_B(D)=POS_{B-\{a_i\}}(D)$，则属性 a_i 是 D-冗余的(也称为是 B 中 D-冗余的)；否则，属性 a_i 是 D-必要的(也称为是 B 中 D-必要的)。

2.13 属性的相对核与相对约简(Relative core and relative reducts of attributes)[8]

设 $B\subseteq C$，若 $POS_B(D)=POS_C(D)$，且 $\forall\ a_i\in B$，$POS_{(B-\{a_i\})}(D)\neq POS_B(D)$，则 B 是 C 的 D-约简(相对约简)。

条件属性 C 中所有决策属性 D 所必要的属性构成的集合称为 C 的 D-核(相对核)，是 C 的 D-可辨识矩阵中所有单个元素组成的集合。

2.14 C 的 D-辨识矩阵与 C 的 D-核[9]

C 的 D-辨识矩阵是一个 $n\times n$ 的对称矩阵，$n=card(U)$，其任一元素 $\alpha(x,y)=\{a\in A|f(x,a)\neq f(y,a),x\in[x]_{Ind(D)},y\in[y]_{Ind(D)},[x]_{Ind(D)}\neq[y]_{Ind(D)}\}$。$C$ 的 D-核是 C 的 D-可辨识矩阵中所有单个元素的集合。

3 基于决策表的分类规则挖掘

3.1 基于决策表的分类规则挖掘算法

① 对原始数据信息离散化，并编码；

② 计算 U/D，得 D-空间的基本集；

③ 计算 $\underline{C}(U/D)$，$C(U/D)$，并利用上下近似计算规则的确定性；

④ 利用 C 的 D-可辨识矩阵计算关于条件属性集 C 的 D-核与 D-约简，选择最佳 D-约简对决策表进行约简；

⑤ 对④所得约简决策表进行对象合并；

⑥ 对⑤所得约简决策表利用 C' 的 D-可辨识矩阵计算关于的条件属性 C' 的值的 D-核与 D-约简，得到分类规则。

⑦ 合并相同的规则，得到所挖掘的最终规则。

⑧ 规则还原，并表示。

3.2 决策规则

由3.1决策表最终化简为描述了如下形式的一组决策(分类)规则的决策表(规则表)：

$a_{k_i}\Rightarrow d_j$

其中 a_{k_i} 表示属性 a_k 的值为 i，符号⇒表示隐含。在决策规则 $\theta\Rightarrow\Phi$ 中，公式 θ 和 Φ 分别称为条件和决策。

3.3 属性的类型

现实中，有两类不同类型的属性。量的属性：表示对象可测量的特性。它们的值按定义有序。例如：温度，pH 值，浓度，等等。质的属性：根据相关术语表达。它们可分为两类：(1)有序的质的属性。这类属性的值将按其重要性排序。值的顺序可表示为按其编码值递增或递减。例如：极性＝(低，中，高)可编码为低＝1，中＝2，高＝3。(2)无序的质的属性(名字属性)。这类属性的值不可能被排序；即，它们是不能根据任一重要性排序的。

粗糙集理论应用于表示质的属性是最直接的。对于名字属性，粗糙集理论与其他分类器相比有明显的优势。对于典型的分类器来说为使用这类属性必须对其进行特别的编码。而基于粗糙集理论的分类器却不需要。

然而，对于连续的条件属性，在使用粗糙集方法前，必须经过离散化。区域的个数和区域的范围必须是优化的。区域的个数决定着逻辑规则的个数。

通常有两种可能的离散化方法。一是仅考虑对象在属性空间的相似性来优化编码，二是在编码阶段将信息系统的确定性特性最大化。尽管方法二听起来可行，但有它的局限性[10]。一个弱粗糙集(即，有多个区域)将导致弱规则(即，每一规则的支持对象少)，这可能导致与专家经验的矛盾。而当粗糙参数设置得高时，将会产生通用的规则，这些规则的支持例很多，从而发现很强的规则。

通常，条件属性的离散化，需要一定的领域知识。在对实际问题进行基于决策表的分类归则挖掘时，首先应根据实际问题的特点，进行属性的选取，属性离散化，属性的编码，将其表示为决策表。然后再运用相应分类规则挖掘算法进行分类规则的挖掘。

4 实例分析

我们以一个汽车数据库为例，来挖掘汽车的油耗与其相应属性的关系，从而得到相应的分类规则。我们以10辆不同类型的汽车作为训练对象，所选汽车属性有：制造型号(make-model)、发动机排气量(displace)、压缩比(compress)、功率(power)、传动类型(trans)、重量(weight)和油耗(mileage)。其中，前6个属性构成条件属性 C，依次表示为 $\{a_1,a_2,a_3,a_4,a_5,a_6\}$，最后一个属性是决策属性 $D=\{d\}$。这些属性值离散

化后见表1。

表1 决策表

U	a_1	a_2	a_3	a_4	a_5	a_6	d
x_1	USA	Medium	High	High	Auto	Medium	Medium
x_2	USA	Small	High	Medium	Auto	Medium	Medium
x_3	USA	Medium	Medium	Medium	Manual	Medium	Medium
x_4	USA	Medium	Medium	High	Manual	Medium	Medium
x_5	USA	Medium	Medium	High	Auto	Medium	Medium
x_6	USA	Small	High	Medium	Manual	Medium	High
x_7	USA	Medium	High	High	Manual	Light	High
x_8	JAPAN	Small	High	Low	Manual	Light	High
x_9	JAPAN	Medium	Medium	Medium	Manual	Medium	High
x_{10}	JAPAN	Small	High	High	Manual	Medium	High

对表1中的信息数据编码见表2。

表2 属性编码表

Attributes	Codes		
	1	2	3
a_1	USA	JAPAN	
a_2	Small	Medium	
a_3		Medium	High
a_4	Low	Medium	High
a_5	Manual	Auto	
a_6	Light	Medium	
d		Medium	High

由此可得编码后的信息系统。见表3。

表3 编码后的决策表

U	a_1	a_2	a_3	a_4	a_5	a_6	d
x_1	1	2	3	3	2	2	2
x_2	1	1	3	2	2	2	2
x_3	1	2	2	2	1	2	2
x_4	1	2	2	3	1	2	2
x_5	1	2	2	3	2	2	2
x_6	1	1	3	2	1	2	3
x_7	1	2	3	3	1	1	3
x_8	2	1	3	1	1	1	3
x_9	2	2	2	2	1	2	3
x_{10}	2	1	3	3	1	2	3

从表3可得决策属性集 D 的基本集关于条件属性集 C 的上、下近似，以及每一 D 基本集的精度。见表4。

表4 决策属性 D 的基本集关于条件属性集 C 的上、下近似

Class number	Number of objects	Lower approximation	Upper approximation	Accuracy
1	5	5	5	1.0
2	5	5	5	1.0

下一步是构造关于条件属性的 C-核与 D-约简。有2个 D-约简，它们是：

Set # 1=$\{a_1,a_3,a_5\}$

Set # 2=$\{a_1,a_2,a_5,a_6\}$

由于条件属性 C 的所有 D-约简的交集是 D-核，因此，该系统的相对 D-核是$\{a_1\}$，这说明属性$\{a_1\}$是分类的最重要的属性，因此，在不降低分类质量的情况下，该属性不能从属性集中剔除。约简的势为3或4，因此，在任何情况下将会有3或2个属性是冗余的，剔除这些冗余属性不会影响分类的结果。最好(最小)的约简是 Set # 1。这样，原信息系统将由6个属性简化为3个属性，得到约简了的决策表。然后合并相同得对象。在此基础上，计算属性值的约简，得到规则集，合并相同规则，得到所挖掘的最终规则。基于约简 Set # 1的规则集见表5。

表5 规则集

Rules	a_1	a_3	a_5	d
1	*	*	2	2
2	1	2	*	2
3	*	3	1	3
4	2	*	*	3

由于约简 Set # 1给出的分类精度等于1，因此，表5中的规则都是确定的。挖掘的逻辑规则如下：

(1)在如下任一条件下，汽车的油耗等级为中等。

(a) 传动类型为自动；

(b) 美国制造，且压缩比中等。

(2) 在如下任一条件下，汽车的油耗等级为高。

(a) 压缩比高，且传动类型是手动；

(b) 日本制造。

使用上述所挖掘的分类规则对汽车数据库中的任一汽车数据进行检验，分类结果与汽车的实际油耗等级一致，分类正确率100%。

表6给出了剔除约简 Set # 1的任一属性后，每一类的精度的变化。从中可以看出约简 Set # 1中的属性对分类精度的影响。

表6 约简 Set # 1的任一属性对每一类精度的影响

Class number	Removed attribute			
	none	a_1	a_3	a_5
1	1.0000	0.5000	0.4286	0.4286
2	1.0000	0.5714	0.4286	0.4286

讨论 本文提出的分类规则挖掘算法，将可辨识矩阵引入到属性值的约简过程中，从而得到泛化了的分类规则；同时，对规则确定性的度量采用决策属性基本集相对于条件属性的上下近似来计算，从而得到泛化规则的确定性。该算法，在 $card(D)>1$时，通过决策属性的合并与展开，也可进行关

(下转第77页)

13显示了一个手写体字的笔画提取的对比实验。

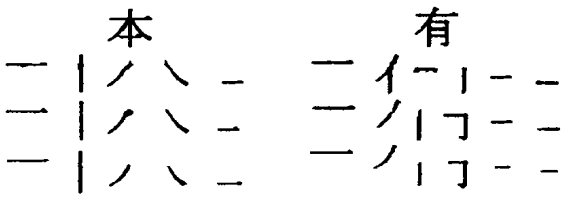

图12 对比实验

用[3]中模型提取的笔画　　用本模型提取的笔画

图13 对比实验

从图12和13可以看出，用本模型进行笔画提取的实验结果不比用文[3]中的模型的结果差，且在某些笔画的提取效果上更好。而本模型和文[3]中模型的提取效果都比文[5]中模型的要好。同时本模型比文[3]中模型的时间耗费大大降低，因为本模型只考虑分叉点区域的点，而文[3]中的模型要考虑字符的所有点，如表1所示。

结论 本文在研究已存在的几种笔画提取模型的基础上，提出了一种新型的基于笔画段分割和组合的汉字笔画提取模型。该模型基于文[3]中宽笔画二值字符图像区域分解的原理，并结合了笔画段分割与组合的新技术。实验证明，与已存在的几种模型相比，该模型在保证笔画提取准确性的前提下，能大大降低时间耗费。从表1可以看出，本模型与文[3]中的模型的时间耗费比都低于20%。同时，该模型应用于汉字印刷体和手写体字符笔画提取都能够达到很好的效果。

表1 时间耗费对比(时间单位为单个像素的PBOD曲线计算时间)

模型	步骤 / 字符	本	有	膀	力
文[3]中的模型	计算所有像素的PBOD曲线	1344	1337	2177	723
	计算所有像素的BBOD曲线	1344	1337	2177	723
	总计	2688	2674	4354	1446
本模型	计算分叉点区域像素的BBOD曲线	204	184	183	85
	笔画段4-连接选择与组合	48	80	96	12
	笔画修正	204	184	183	85
	总计	456	448	464	182
本模型与文[3]中的模型的时间耗费比		16.96%	16.75%	10.66%	12.59%

参考文献

1 Chang H-H, Yan Hong. Analyse of Stroke Structures of Handwritten Chinese Characters. IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS, 1999, 29(1)

2 Lin Feng, Tang Xiaoou. Off-line Handwritten Chinese Character Stroke Extraction

3 Cao Ruini, Tan C L. A Model of Stroke Extraction from Chinese Character Images

4 Zhang T Y, Suen C Y. A Fast Parallel Algorithm For Thinning Digital Patterns. Communications of the ACM, 1984, 27(3)

5 Chen Y S, Hsu W H. An interpretive model of line continuation in human visual perception. pattern recognition, 1989, 22: 619～639

(上接第63页)

联规则的挖掘。因为分类规则与关联规则的不同之一在于规则的结果中(then…)属性的个数。在关联规则中，结果属性可多于一个[11]。该算法还可用于医学临床数据的分析、银行数据的分析、人口数据分析、股市的预测、生物数据的分析、化学实验数据分析、最新的SARS病例数据分析等等。难点在于，对于数量类型的属性需要选择合适子范围来对其进行离散化，这常常需要一些领域知识。

参考文献

1 邢乃宁，孙志挥. 一种基于粗集理论的分类规则挖掘的实现方法. 计算机应用，2001，21(12)：29～31

2 Pawlak Z. Rough set approach to knowledge-based decision support. European Journal of Operational Research, 1997(99): 48～57

3 Pawlak Z. Rough sets and intelligent data analysis. Information sciences, 2002(147): 1～12

4 王国胤. Rough集理论与知识获取. 西安交通大学出版社，2001

5 曾黄麟. 粗糙集理论及其应用—关于数据推理的新方法. 重庆大学出版社，1996

6 张文修，吴伟志，梁吉业等编著. 粗糙集理论与方法. 科学出版社，2003

7 Pawlak Z. Rough classification. Int. J. Human-Computer Studies, 1999(51): 369～383

8 [美]C. L. Liu 著，刘振宏译. 离散数学基础. 人民邮电出版社，1982

9 Walczak B, Massart D L. Rough sets theory, Chemometrics and Intelligent Laboratory System, 1999 (47): 1～16

10 Ziarko W, Golan R, Edwards D. An Application of Datalogic/R Knowledge Discovery Tool to Identify Strong predictive Rules in Stoc Market Data. AAAI-93 Workshop on Knowledge Discovery in Databases, Washington: DC, 1993

11 Freitas A A. A Survey of Evolutionary Algorithm for Data Mining And Knowledge Discovery. http://www.ppgia.pucpr.br～alex