文章编号:1000-582X(2007)12-0115-04

# 基于粗集理论的一种相容决策表的学习算法

李文生
(重庆大学 经济与工商管理学院,重庆 400030)

摘　要:基于粗集理论,针对相容系统规则,提出一种新的相容系统决策表归纳学习算法,并通过实际例子说明该算法的有效性和可信度,与以往的相容决策表的归纳学习算法相比较,这种算法比较简单,而且能够全面地获取规则且没有冗余。给出了具有较高可信度的规则挖掘过程。

关键词:粗集理论;相容; 决策规则;归纳学习
中图分类号:TP18　　文献标志码:A

机器学习近年来受到广泛关注,还没有实用化,知识的获取主要依赖于知识工程师和专家的沟通,而归纳学习又是这一领域中最好理解的问题。许多专家系统需人工对不同领域专家知识进行归纳编码建立,因而需要计算机人员和不同领域专家的紧密合作,这一过程不仅费时,且效率低下,并且不同的应用领域需要进行相同的枯燥乏味的工作。因此,迫切需要建立一种归纳学习方法,能够从一个精心选取的专家决策样本集中自动归纳和提炼决策规则。

上世纪 80 年代初,波兰数学家 Pawlak 教授等提出用粗集理论(Rough Set Theory)研究不完整、不精确知识的表达、学习、归纳等方法,该理论目前主要用于知识的约简和知识依赖性的分析,在医疗诊断、模式识别、专家系统、机器学习、图像处理等领域获得了广泛的应用。笔者基于粗集理论,针对相容决策表,提出相应的归纳学习算法,并通过实例说明算法的有效性。

## 1　粗糙集理论基础

粗集理论的主要思想是用不可分辨关系划分知识,用上下近似逼近描述概念,通过知识约简导出决策或分类规则。

### 1.1　知识表达系统和决策系统

基于粗集理论的观点,一个知识表达系统可表示为 $S = <U, A, V, f>$,其中,$U$ 是对象的有限集合,$U = \{\chi_1, \chi_2, \cdots, \chi\}$;$A$ 是属性的有限集合;$V = \bigcup_{a \in A} V\ a$ 是属性的值域集,其中 $Va$ 是属性 $a \in A$ 的值域;$f$ 是信息函数,$f: U \times A \rightarrow V$, $f(x_i, a) \in Va$. 知识表达系统可以方便地用数据表表示,表的行代表对象 $e \in U$,列代表属性 $a \in A$,表中的数值代表着对象 $e$ 对应属性 $a$ 的属性值 $a(e)$,$(a, a(e))$ 表示对象的属性 - 值对,每行表示该对象的一条信息.

### 1.2　不可分辨关系和上下近似

对于 $B \subseteq A$,则 $B$ 在 $U$ 上的不可分辨关系定义为:$\mathrm{ind}(B) = \{(\chi_1, \chi_2) \mid f(\chi_1, b) = f(\chi_2, b)$, for any $b$ in $B\}$,不可分辨关系也是等价关系. $\mathrm{ind}(B)$ 把 $U$ 划分为 $k$ 个等价类 $\chi_1, \chi_2, \cdots, \chi$,记 $U\ /\ \mathrm{ind}(B) = \{\chi_1, \chi_2, \cdots, \chi\}$,用 $\mathrm{des}\{Xi\}$ 表示 $U$ 上的一个等价类的描述.

对任意一个对象集合 $X \subseteq U$,$X$ 的 $B$ 下近似定义为:$B\_(X) = \cup \{Y \in U\ /\ \mathrm{ind}(B): Y \subseteq X)\}$,$X$ 的 $B$ 上近似定义为:$B(X) = \cup \{Y \in U\ /\ \mathrm{ind}(B): Y \cap X \neq \varphi\}$. 下近似 $B\_(X)$ 表示所有一定能归入 $X$ 的等价类元素的集合,而上近似 $B(X)$ 表示所有可能归入 $X$ 的等价类元素的集合.

### 1.3　简　约

在决策表中,设 $U\ /\ \mathrm{ind}(C) = X = \{\chi_1, \chi_2, \cdots, \chi\}$, $U\ /\ \mathrm{ind}(D) = Y = \{y_1, y_2, \cdots, y\}$,则 $Y$ 的 $C$ 正域定义为:$POS^c(D) = \cup \{C^-(Xi) \mid Xi \in Y\}$,正域包含着基于条件属性所得的等价类能够归入基于决策属性所得的

收稿日期:2007-07-12
作者简介:李文生(1979-),男,重庆大学硕士,主要从事非线性系统理论,管理科学与工程等领域的研究,(E-mail)cqulws@ yahoo. com。

等价类的所有对象集合。$Y$ 的 $C$ 负域定义为:$NEG^c(D)=U^-\cup\{C^-(Xi)\mid Xi\cap D\neq\varphi\}$,它包含着基于条件属性所得的等价类不能够归入基于决策属性所得的等价类的所有对象集合. $D$ 对于 $C$ 的依赖度定义为:$K(C,D)=\mathrm{card}(POS^c(D)+NEG^c(D))/\mathrm{card}(U)$,$K(C,D)$反映了把对象 $U$ 划分为概念 $D$ 和 $-D$ 的分类能力。

对于属性 $p\in C$,如果 $K(C^-\{p\},D)=K(C,D)$,则 $p$ 在决策中相对于 $D$ 是冗余的,否则是不可缺少的. 而属性 $p$ 在 $C$ 和 $D$ 中的重要度定义为:$\mathrm{SGF}(p,C,D)=K(C,D)-K(C^-\{p\},D)$. 如果 $C$ 中的任意一个属性关于 $D$ 是不可缺少的,则 $C$ 关于 $D$ 是正交的;对于 $BC$,如果 $K(B,D)=K(C,D)$,而且 $B$ 是正交的,则 $B$ 是 $C$ 的 $D$ 简约,简约可以理解为在丢失信息的前提下,可以最简单地表示决策系统的决策属性对条件属性集的依赖和关联,记为:$\mathrm{REDU}(C,D)$. 一般情况下,$C$ 的 $D$ 简约有多个。

## 2 一种相容系统规则提取

决策规则的提取是粗集理论的具体应用。采用常用的相容决策表归纳学习方法分析实例(决策表1)说明笔者的方法,提出其中的问题。

对于决策表1,先给出一种常用的归纳学习算法分析[3],其过程如下:

输入:相容知识表达系统 $S=\{U,C,D,V,f\}$,$U=\{\chi_1,\chi_2,\cdots,\chi\}$,$C=\{c_1,c_2,\cdots,c\}$ 导师分类 $y=\{y_1,y_2,\cdots,y\}$;

输出:对于导师分类的每一类决策规则。

1)$j=1$;

2)$U'=U,C'=C,B=\varphi,y=yj$;

3)对于所有的条件属性 $c\in C',B'=B\cup\{c\}$,

求使 $\max\alpha^{\beta}{}'(y)$ 的属性 $c_{\max}$;令 $B'=B\cup\{c_{\max}\}$,$B=B'$;

4)若 $B\_y=\varphi$,转5);否则 $B\_y=\{\chi_1,\chi_2,\cdots,\chi\}$,输出确定性规则:$\mathrm{des}(\chi k)\rightarrow\mathrm{des}(y)$,$k=1,2,\cdots,r$;

5)$U'=U'-((U'-By)\cup B\_y)$,$y=y-B-y$. 若 $U'=\varphi$,转6);否则令 $C'=C'-B$,若 $C'\neq\varphi$,转3);

6)$j=j+1$,若 $j\leqslant n$,转2);否则结束。

为了便于叙述,现举例一个知识表达系统,如表1所示表中,$U=\{e_1,e_2,\cdots,e\}$,$C=\{c_1,c_2,\cdots,c\}$,其中 $C_1$=发烧,$C_2$=咳嗽,$C_3$=头痛,$C_4$=体温,$C_5$=听诊。则 $C_1$ 中包含4个元素:高、中度、低和无,分别记为 $c_{11}$、$c_{12}$、$c_{13}$和 $c_{14}$;$C_2$ 包含3个元素:剧烈、中度、轻微,分别记为 $c_{21}$、$c_{22}$和 $c_{23}$;$C_3$ 中包含4个元素:强烈、中度、轻微和无,分别记为 $c_{31}$、$c_{32}$、$c_{33}$和 $c_{34}$;$C_4$ 中包含2个元素:高和底,分别记为 $c_{41}$、$c_{42}$;$C_5$ 中包含3个元素:干鸣音、水泡音和正常,分别记为 $c_{51}$、$c_{52}$和 $c_{53}$;决策分类中包含2个元素:肺炎和流感,分别记为 $d_1$ 和 $d_2$。

表1 决策表1

| 样本 $U$ | 条件属性 | | | | | 决策属性 |
|---|---|---|---|---|---|---|
| | $(C_1)$ | $(C_2)$ | $(C_3)$ | $(C_4)$ | $(C_5)$ | $(D)$ |
| $e_1$ | $c_{14}$ | $c_{23}$ | $c_{34}$ | $c_{41}$ | $c_{52}$ | $d_1$ |
| $e_2$ | $c_{13}$ | $c_{23}$ | $c_{34}$ | $c_{41}$ | $c_{52}$ | $d_1$ |
| $e_3$ | $c_{12}$ | $c_{21}$ | $c_{32}$ | $c_{41}$ | $c_{51}$ | $d_1$ |
| $e_4$ | $c_{14}$ | $c_{22}$ | $c_{34}$ | $c_{41}$ | $c_{52}$ | $d_1$ |
| $e_5$ | $c_{13}$ | $c_{21}$ | $c_{34}$ | $c_{41}$ | $c_{52}$ | $d_1$ |
| $e_6$ | $c_{11}$ | $c_{21}$ | $c_{33}$ | $c_{41}$ | $c_{53}$ | $d_2$ |
| $e_7$ | $c_{14}$ | $c_{23}$ | $c_{31}$ | $c_{42}$ | $c_{51}$ | $d_2$ |
| $e_8$ | $c_{12}$ | $c_{21}$ | $c_{33}$ | $c_{41}$ | $c_{54}$ | $d_2$ |
| $e_9$ | $c_{11}$ | $c_{21}$ | $c_{32}$ | $c_{42}$ | $c_{51}$ | $d_2$ |
| $e_{10}$ | $c_{12}$ | $c_{22}$ | $c_{34}$ | $c_{42}$ | $c_{53}$ | $d_2$ |

上述归纳学习算法强调了一些突出规则(从步骤3)可以看出),而难以得到所有的规则,且规则中存在冗余。算法得到的规则如下

1)$C_5=c_{52}\rightarrow D=d_1$;

2)$(C_1=c_{13}\vee c_{41}=C_4)\wedge C_5=c_{51}\rightarrow D=d_1$;

3)$C_5=c_{53}\rightarrow D=d_2$;

4)$C_4=c_{41}\wedge C5=c_{51}$(冗余)$\rightarrow D=d_2$;

5)$C_1=c_{11}\wedge C5=c_{51}\rightarrow D=d_2$;

6)$C_1=c_{14}\wedge C5=c_{51}$(冗余)$\rightarrow D=d_2$。

为了能从决策表尽可能多地得到简单实用的规则,提高决策水平,提出一种归纳学习算法,大致过程如下

输入:相容知识表达系统 $S=\{U,C,D,V,f\}$,$U=\{\chi_1,\chi_2,\cdots,\chi\}$,$C=\{c_1,c_2,\cdots,c\}$ 导师分类 $y=\{y_1,y_2,\cdots,y\}$;

输出:对于导师分类的每一类决策规则。

1)$C$ 的一阶幂集 $T1(C)=\{c_1,c_2,\cdots,c\}$

For $i=1$ to $n$

计算 $U/\mathrm{IN}\,D\{c_i\}=\{c_{1i},c_{i2},\cdots,c\}$;

For $j=1$ to $s$

For $k=1$ to $m$

if $(c_{ij}\subseteq y_k)$

then

{输出规则:$\mathrm{des}(c_{ij})\rightarrow\mathrm{des}(y_k)$;

//$\mathrm{des}(c_{ij})$、$\mathrm{des}(y_k)$分别为 $c_{ij}$和 $y_k$ 的描述

$c_{i\_}x\leftarrow x(c_{ij})$;

//将拥有属性 $c_{ij}$的对象存进数组 $c_{i\_}x$。

}

2) $C$ 的二阶幂集 $T2(C)=\{(c_r,c_p)\in 2U, r=1,2,\cdots,n-1; P=2,3,\cdots,n; r<P\}$

For $r=1$ to $n-1$

{For $p=2$ to $n$

if ($c_{i\_}x\neq\varphi <\parallel c_{p-x}\neq\varphi$)

then

{$x\in c_{i-x}x \parallel c_{p-x}$入栈 $X$;

erase ($x$); / / 删除对象 $x$

}

计算 $U$ / IN $D(c,c_p)=\{c_{rp1},c_{rp2},\cdots,c_{rpq}\}$;

For $i=1$ to $q$

For $k=1$ to $m$

if ($c_{rpi}\subseteq y_k$)

then

{输出规则:des($c_{rpi}$)→des($y_k$);

$c_{i\_}x\leftarrow x(c_{rpi})$;

//将拥有属性组合 $c_{rpi}$的对象存进数组 $c_{rp}-x$;

}

$X$ 出栈;

}

3)对三阶以上的幂集采用相似于 2)的步骤,直到由 $Ti(L), i\geqslant 3$ 得到的规则集合为空。运用上述算法,得到决策表 1 的规则如下:

1) $C_1=c_{13}\rightarrow D=d_1$;

2) $C_1=c_{11}\rightarrow D=d_2$;

3) $C_3=c_{33}\rightarrow D=d_2$;

4) $C_3=c_{31}\rightarrow D=d_2$;

5) $C_4=c_{42}\rightarrow D=d_2$;

6) $C_5=c_{52}\rightarrow D=d_1$;

7) $C_5=c_{41}\rightarrow D=d_2$;

8) $C_1=c_{14}\wedge C_2=c_{13}\rightarrow D=d_1$;

9) $C_1=c_{12}\wedge C_2=c_{13}\rightarrow D=d_2$;

10) $C_1=c_{14}\wedge C_3=c_{34}\rightarrow D=d_2$;

11) $C_1=c_{12}\wedge C_3=c_{32}\rightarrow D=d_1$;

12) $C_1=c_{12}\wedge C_3=c_{34}\rightarrow D=d_2$;

13) $C_1=c_{14}\wedge C_4=c_{41}\rightarrow D=d_1$;

14) $C_1=c_{12}\wedge C_5=c_{51}\rightarrow D=d_1$;

15) $C_1=c_{14}\wedge C_5=c_{51}\rightarrow D=d_2$;

16) $C_2=c_{23}\wedge C_3=c_{34}\rightarrow D=d_1$;

17) $C_2=c_{21}\wedge C_3=c_{34}\rightarrow D=d_1$;

18) $C_2=c_{23}\wedge C_4=c_{41}\rightarrow D=d_1$;

19) $C_2=c_{22}\wedge C_4=c_{41}\rightarrow D=d_1$;

20) $C_2=c_{23}\wedge C_5=c_{51}\rightarrow D=d_2$;

21) $C_3=c_{34}\wedge C_4=c_{41}\rightarrow D=d_1$;

22) $C_3=c_{32}\wedge C_4=c_{41}\rightarrow D=d_1$;

23) $C_4=c_{41}\wedge C_5=c_{51}\rightarrow D=d_1$。

可以看出,笔者提出的算法不仅可以得到决策表所有的规则,而且得到的规则简单,规则没有冗余现象。不考虑规则的冗余,常用的归纳学习方法得到的规则集仅是本文算法得到结果的一个子集。

算法的复杂性:一般讲来,由三阶以上的幂集得到的规则很少(决策表 1 三阶幂集得到的规则集为空)。算法的时间复杂性大致为:$O(m(C_n^1+C_n^2))=O(n2.m)$,$n$ 为属性个数,$m$ 为例子个数,即算法可以以多项式时间复杂性获得决策表的所有规则,易编程实现。

下面举例说明算法的简单性。

[例 1]现在比较文献[4]和文章提出的算法取简单的例子决策表 2。

文献[4]的大致思想是:决策表的简化就是简化决策表的条件属性。化简后的决策表在保持信息系统决策能力的前提下,条件属性最少,这样就可以得到最小决策规则集。决策表化简分两步:1 决策表条件属性的简化;2)消去决策规则的冗余条件属性值。在这个过程中,要进行诸多的迭代和计算。具体的算法过程请参考文献[4]。

表 2　决策表 2(检测与诊断实测数据)

| $U$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $y$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 0 | 1 | 1 | 1 |
| 3 | 1 | 2 | 1 | 2 | 2 | 1 |
| 4 | 1 | 2 | 1 | 0 | 2 | 1 |
| 5 | 1 | 0 | 1 | 0 | 0 | 2 |
| 6 | 1 | 1 | 2 | 0 | 1 | 2 |
| 7 | 2 | 1 | 1 | 1 | 1 | 2 |
| 8 | 2 | 1 | 2 | 0 | 1 | 2 |

两种方法都可以得到下列规则

1)如果(故障指示灯,不亮)且(过压指示,闪烁),则(电源电压不正常);

2)如果(过压指示,闪烁)且(稳压输出,正常),则(电源电压不正常);

3)如果(故障指示灯,闪烁)且(过压指示,常亮),则(电源电压不正常);

4)如果(故障指示灯,闪烁)且(过压指示,不亮),则(电路元件故障);

5)如果(过压指示,闪烁)且(稳压输出,高),则(电路元件故障);

6)如果(故障指示灯,常亮)且(过压指示,闪烁),则(电路元件故障)。

得到优化规则见表 3

| $U$ | $X_1$ | $X_2$ | $X_3$ | $y$ |
|---|---|---|---|---|
| 1 | 0 | 1 | × | 1 |
| 2 | × | 1 | 0 | 1 |
| 3 | 1 | 2 | × | 1 |
| 4 | 1 | 0 | × | 2 |
| 5 | × | 1 | 2 | 2 |
| 6 | 2 | 1 | × | 2 |

比较两种算法,笔者的方法省去了决策表约简繁琐的步骤,过程比较简单,对复杂的例子也是如此。

讨论:决策表的学习质量,与表中实例的质量有很大关系。通常,决策表中会含有噪声,这样无论归纳学习算法有多完善,得到的知识(规则)就不一定可信,要由专家评判。笔者提出的算法,对噪声比较敏感,故对样本的质量要求较高。在算法执行前,最好对决策增加一个校验过程。

## 3 结 论

粗集理论作为一种新的分析和处理不精确和不确定性知识的数学工具,由于其不需要预先给定某些特征或属性的数量,可从现有的数据出发给出知识的简化和相对简化,从决策表抽取规则是获取知识、辅助决策的一种重要途径。运用粗糙集理论,对相容决策表进行分析,提出了相应的发现算法。笔者采用的算法简单,能够归纳出比较简洁、没有冗余的规则。而且最重要的是,发现的规则比较全面,可以得到较高可信度的规则。随着数据的动态变化,原有规则可能失效或部分失效。若用原有算法重新导出新规则十分麻烦,而基于动态数据的更新知识算法是对原有规则进行修正,从而得出基于新数据的规则。知识的动态获取是知识发现的难点,目前基于动态数据的更新知识算法还不多见,需进一步研究。

参考文献:

[1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences,1982 ,11 :341-356.

[2] PAWLAK Z. Rough Sets[J]. Communication of The ACM, 1995,38(11) :89-91

[3] 曾黄麟. 粗集理论及其应用[M]. 重庆:重庆大学出版社, 1998.

[4] 李岩. 基于粗集理论的规则知识获取[J]. 兵工自动化网络信息技术,2003,22(3):24-26.

[5] QUINLAN J R. Induction of decision trees[J]. Machine Learning,1986,1(1):81-106.

[6] HONG J R. AE1: An extension matrix approximate method for the general covering problem [J]. International Journal of Computer and Information Science,1985,14(6):421-437.

[7] BRODLEY C E,UTGOFF PE. Multivariate decision trees[J]. Machine Learning,1995,19:45-77.

[8] PAWLAK Z. Rough Sets-Theoretical Aspects of Reasoning About Data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

# A Kind of Compatibility Decision-making List Learning Means Based on Rough Sets

*LI Wen-sheng*

(College of Economics and Business Administration,Chongqing University,Chongqing 400030,P. R. China)

**Abstract**: The paper introduces basic concept of Rough Set Theory. The compatibility decision-making list inductive learning is an important field in where the rough set theory is used,Base on to consistent decision-making list ,lodging a new kind of deduce means . An example illustrates the efficiency of those methods. Contrast to old means,the means is more easy,it can gain completely rules and no any redundancy. Moreover, is introduced the study of incompatibility decision-making list,concluding the rule digging process is high credibility.

**Key words**: rough sets; compatibility; decision-making rule; inductive learning

(编辑 张小强)