文章编号:1673-9469(2008)01-0096-04

一种缺省规则挖掘算法

刘志民,范 杰,杨 珠,庞彦军
(河北工程大学 理学院,河北 邯郸 056038)

摘要:针对不一致数据库,定义属性权重及缺省规则加权支持度概念,在此基础上给出一种缺省规则挖掘算法。

关键词:Rough集;不一致数据库;属性权重;缺省规则;加权支持度
中图分类号:TP18 **文献标识码**:A

An mining algorithm of default regulars

LIU Zhi-min, FAN Jie, YANG Zhu, PANG Yan-jun
(College of Science, Hebei University of Engineering, Handan 056038, China)

Abstract: The attribute weight and weighted support of default regular are defined by using the conditional entropy and a mining algorithm of default regulars are given for inconsistent database.

Key words: Rough set; inconsistent database; attribute weight; default regular; weighted support

由于知识获取和知识表示等方面的原因,大部分数据库存在着不一致性,即具有相同条件属性值的样本具有不同类别。一致性数据库产生确定性规则,不一致性数据库将产生非确定性规则。在不一致数据库中,数据挖掘系统不能根据条件属性值将样本分类,Mollestad提出一种从不一致数据库中提取命题缺省规则的方法[1],但该算法不能有效地过滤噪声,提取的规则在数量上往往是巨大的。并且文献[1]由上而下的搜索策略不可避免地要从包含属性最多的顶层节点开始搜索,要作大量的运算和消耗大量时间。文献[2]在置信度和支持度下给出一种缺省规则挖掘算法,但没有考虑各属性的不同重要性,定义的规则支持度概念不够完善,支持度阈值设置有较大的随意性。本文借助Rough集[3]方法定义属性权重,在此基础上定义规则加权支持度概念,采取由下而上的搜索策略,给出搜索停止条件,建立能有效过滤噪声、有较高挖掘效率的缺省规则挖掘算法。

1 条件属性权重及其确定方法

1.1 规则置信度

定义1 设 $T=<U,C\cup D>$ 为决策系统,$S\subseteq C$,

$U/IND(S)=\{E_1,E_2,\cdots,E_k\}$,

$U/IND(D)=\{Z_1,Z_2,\cdots,Z_l\}$。

规则 $Des(E_i,S)\rightarrow Des(Z_j,D)$ 的置信度为

$$\mu_c(E_i,S)=|E_i\cap Z_j|/|E_i| \quad (1)$$

若条件等价类 E_i 全部被映射到同一个决策类 Z_j 中,此时能产生确定性规则,规则置信度 $\mu_c(E_i,S)$ 为1。但是,由于属性、属性值的遗漏或噪音的影响,条件等价类 E_i 中的所有对象不能完全映射到同一个决策类 Z_j 中,而是映射到两个以上的决策类。这时,所产生规则的置信度小于1,即产生缺省规则。

1.2 规则支持度

为了减弱或消除噪音的影响,只考虑规则的

收稿日期:2007-10-31
基金项目:国家自然科学基金资助项目(60474019),河北省自然科学基金资助项目(F2005000482) 特约专稿
作者简介:刘志民(1979-),男,河北临漳人,助教,从事不确定性信息数学处理研究。

置信度 $\mu_c(E_i,S)$是不够的。如,一个特定元组在数据库中比例极小,而产生的规则置信度为 1。如果只考虑规则置信度,那么,这条规则将作为有效规则提交给用户。但是,这条规则完全可能是因为噪音影响所至。为了排除噪音干扰,需引入规则"支持度"概念。

定义 2 设 $E_i\in U/IND(S)$, $Z_j\in U/IND(D)$, $E_i\cap Z_j\neq\phi$,规则 $Des(E_i,S)\rightarrow Des(Z_j,D)$的支持度定义为

$$\mu_s(E_i,Z_j)=|E_i\cap Z_j|/|U| \tag{2}$$

有了规则支持度概念,当设定支持度阈值 μ_s 后,对支持度 μ_s 小于的规则可认为是噪音所致而舍弃相应规则。

1.3 条件属性权重的确定方法

规则支持度应与生成等价类 E_i 的属性集 S 中包含的属性的重要性有关,要合理的定义规则支持度,首先要确定 C 中各属性的权重。

给定决策系统 $T=<U,CUD>$, $S\subseteq C$, $E_i\in U/IND(S)$, $Z_j\in U/IND(D)$。则由知识 S 在 U 的子集组成的 σ 代数上可定义概率分布

$$[S,p]=\left\{\frac{E_1}{p(E_1)},\frac{E_2}{p(E_2)},\cdots,\frac{E_r}{p(E_r)}\right\}$$

其中 $p(E_i)=|E_i|/|U|$, $i=1,2,\cdots,r$。

知识 S 的熵[4] $H(S)$定义为

$$H(S)=-\sum_{i=1}^{r}p(E_i)\cdot\log p(E_i) \tag{3}$$

知识 D 相对于知识 S 的条件熵定义为 $H(D/S)$

$$H(D/S)=-\sum_{i=1}^{r}p(E_i)\cdot\sum_{j=1}^{m}p(Z_j/E_i)\cdot\log p(Z_j/E_i) \tag{4}$$

其中 $p(Z_j/E_i)=|E_i\cap Z_j|/|E_i|$。条件熵 $H(D/S)$是知识 D 相对于知识 S 具有的信息量。

显见,若 S、P 是 U 上的两等价关系族,并且 $IND(S)=IND(P)$,则 $H(S)=H(P)$,这说明两个代数表示下等价的知识库具有相同的信息量。当 $S\subseteq P$, $H(S)=H(P)$时,可证明 $IND(S)=IND(P)$[5]。说明,当两个知识库存在包含关系时,由知识的信息量相等可得出它们在代数表示下是等价的。

任意属性 $a\in C$,其重要性并非由 a 自身确定,而是体现在从属性集 S 中将 a 去除(当 $a\in S$)或将 a 加入 S 中(当 $a\in C-S$)后知识 D 关于知识 S 的信息量的改变量。从这个意义上讲,属性 a 的重要性是相对的。事实上,若 $S\subseteq P\subseteq C$, $a\in S$,对 P 来说 a 可能是可去除的,而对 S 来讲 a 可能是必要的。所以,属性 a 的重要性是相对于某种固定的知识 $S\subseteq C$ 而言的,离开知识 S 孤立地讨论属性 $a\in C$ 的重要性是没有实际意义的。

属性 $a\in C$ 关于知识 $S\subseteq C$ 的重要性可用将 a 从 S 中去除或将 a 加进 S 后条件熵的增量来度量,故给出属性关于知识 S 的权重如下定义。

定义 3 给定决策表 $T=<U,CUD>$, $S\subseteq C$, $E_i\in U/IND(S)$, $Z_j\in U/IND(D)$, $a\in C$, a 关于 S 的权重定义为

$$\omega^S(a)=|\triangle H_a|/\sum_{a\in C}|\triangle H_a| \tag{5}$$

其中

$$\triangle H_a=\begin{cases}H(D/S)-H(D/(S-\{a\})),\text{当 } a\in S\\ H(D/(S+\{a\}))-H(D/S),\text{当 } a\in C-S\end{cases} \tag{6}$$

当 $S=\phi$ 时,没有属性对 U 划分,即 $U/IND(S)=U$,所以

$$H(D/S)=-\sum_{i=1}^{r}p(E_i)\sum_{j=1}^{k}p(Z_j/E_i)\log p(Z_j/E_i)$$
$$=-\sum_{j=1}^{k}p(Z_j/U)\log p(Z_j/U)$$
$$=-\sum_{j=1}^{k}p(Z_j)\log p(Z_j)=H(D)$$

所以,当 $S=\phi$ 时, a 关于 S 的权重为

$$\omega^{\phi}(a)=|H(D/\{a\})-H(D)|/\sum_{i=1}^{m}|H(D/\{a_i\})-H(D)| \tag{7}$$

2 规则加权支持度

用 Rough 集理论进行分类规则挖掘时往往产生大量的分类规则,其中包括了由于噪音影响产生的分类规则。为了消除噪音影响生成的分类规则,除了要考虑规则置信度外,还要考虑规则支持度,而支持度与产生分类的 S 中的属性有关,即与属性权重有关。

定义 4 设决策系统为 $T=<U,C\cup D>$, $S\subseteq C$, $E_i\in U/IND(S)$, $Z_j\in U/IND(D)$, $E_i\cap Z_j\neq\phi$,规则 $Des(E_i,S)\rightarrow Des(Z_j,D)$的加权支持度定义为

$$\mu_s(E_i,S)=\frac{|E_i\cap Z_i|}{|U|\cdot|S|}\cdot\sum_{a_j\in S}W^S(a_j) \tag{8}$$

置信度与支持度是本质不同的概念,随着属性集扩充,划分变细,置信度递增。当细分到一定程度某些规则的置信度可取到最大值1。而加权支持度通常随着 S 的扩充呈下降趋势。

当 $E_i\cap Z_j\neq\phi$,且 $E_i\not\subset Z_j$ 时,所产生的缺省规则不具有百分之百的置信度,但在大部分情况下可以正确使用。比如作为鸟的鸵鸟不会飞,但这并不影响"鸟都会飞"这条缺省规则在大部分情况下的使用。领域专家使用的经验规则并不完全正确,有使规则不成立的反例,但在实际应用中却非常有效。因为缺省规则通常比较简洁,使用方便。

3 缺省加权规则挖掘算法

3.1 搜索策略

本文采用"由下而上"的搜索策略,即从条件属性少的底层节点开始,向上到条件属性多的上层节点逐一进行搜索。最底层对空条件属性集进行挖掘,得到的规则是决策属性值的概率分布,然后是第一层,第二层…直到第$|C|$层逐层进行挖掘。这种搜索策略符合人们由粗到细的认知过程。

3.2 搜索停止条

定义5 设 $T=<U,C\cup D>$ 为决策系统,$S\subseteq C$,对预先设定的最低加权支持度阈值 $\mu_S^{\min}$,如果存在 $E_s\in U/IND(S)$,$Z_j\in U/IND(D)$,满足

$$|E_{S_i}\cap Z_j|\geq\mu_S^{\min}\cdot|U|\cdot|S|/\sum_{a_j\in S}W^S(a_j) \tag{9}$$

则称 S 是期望属性集。

定理 设 $T=<U,C\cup D>$ 是决策系统,$S\subseteq C$,$\mu_S^{\min}$ 是最低加权支持度阈值。若对任意的 $E_{s_i}\in U/IND(S)$,和任意 $Z_j\in U/IND(D)$都有

$$|E_{S_i}\cap Z_j|<\mu_S^{\min}\cdot|U|\cdot|S|/\sum_{a_j\in S}W^S(a_j)+\max_{a\in C-S}W^S(a) \tag{10}$$

则在缺省加权规则挖掘中 S 属性集构成的节点不可能有上层节点。

证明:只需证明,在 $C-S$ 中任取一种属性 a 加入 S 后都不可能是期望属性集。

任取 $a\in C-S$,$E_{(s+\{a\})_i}\in U/IND(S+\{a\})$,则对任意的 $E_{(S+\{a\})_i}$,由定理一知,必存在 $E_{S_i}\in U/IND(S)$,使 $E_{(s+\{a\})_i}\subseteq E_{S_i}$,并且对任意 $Z_j\in U/IND(D)$ 有

$$\begin{aligned}&|E_{(S+\{a\})_i}\cap Z_j|\leq|E_{S_i}\cap Z_j|\\&<\mu_S^{\min}\cdot|U|\cdot|S|/\sum_{a_j\in S}W^S(a_j)+\max_{a\in C-S}W^S(a)\\&<\mu_S^{\min}\cdot|U|\cdot(|S|+1)/\sum_{a_j\in(S+\{a\})}W^S(a_j)\end{aligned}$$

所以,$S+\{a\}$不是期望属性集。

定理说明,若搜索进行到第 k 层,而第 k 层的任一节点上属性集 S 的类 $E_{S_i}\in U/IND(S)$都满足(10)式,则在缺省规则挖掘中 S 属性集构成的节点不可能有上层节点,意味着搜索可以停止。说明搜索深度为 k,而通常情况下 $1\leq k<|C|$。

3.3 搜索方向

设 N_i 表示第 i 层的节点,$C_{N_i}=S\subseteq C$ 是节点 N_i 上的属性集,则从 $C-S$ 中选择对 S 具有最大权重的前 t 种属性依次并入 S 构成 S 的包含集 P,称 P 是备选属性集,则备选集指示了搜索方向。

3.4 算法步骤

通过上述分析,给出缺省加权规则挖掘算法,简称为 *MDWRBR* 算法。

输入:决策系统 $T=<U,C\cup D>$,加权支持度阈值 $\mu_s(0\leq\mu_s\leq1)$,置信度阈值 $\mu_c(0\leq\mu_c\leq1)$。$C=\{a_1,a_2,\cdots,a_m\}$,a_i 关于属性集 S 的相对权值为 $W^S(a_j)(i=1,2,\cdots,m)$。

输出:给定决策系统 T 上的确定性规则和缺省加权规则集合。

步骤1 C_{N_i}表示节点 N_i 上的属性集,R_{N_i}表示节点 N_i 上的加权规则集;算法产生的加权规则集 $R=\phi$;

步骤2 设底层节点 N_0 所对应的属性集为 C_{N_0},在 N_0 节点上生成加权规则

$R_{N_0}=Createrule(U,D,C_{N_0},\mu_s,\mu_c)$;$R=R+R_{N_0}$;

步骤3 循环:i 从1到$|C|$,执行:

①循环:J 从1到 $C_{|C|}^i$,执行:

对第 I 层上的节点 N_{ij}生成缺省加权规则

$R_{N_{ij}}=CreateRule(U,D,C_{N_{ij}},\mu_s,\mu_c)$;$R=R+R_{N_{ij}}$;

②生成第 $i-1$ 层上产生的缺省加权规则的

例外(*blocks*);

步骤 4 结束。

算法说明:

函数 $CR(U,D,C_{cur},\mu_s,\mu_c)$

begin

$R=\phi$

计算所有类 $U/IND(C_{cur})$,计算 a_j 关于 C_{cur} 的权重 $W^{C_{cur}}(a_j)(j=1,2,\cdots,m)$。

循环对每一个 $Z\in U/IND(D)$执行:

循环对每一个 $E_i\in U/IND(C_{cur})$执行:

如果 $|E_i\cap Z|/|E_i|\geq\mu_c$,并且 $|E_i\cap Z|\geq\mu_s\cdot|U|\cdot|C_{cur}|/\sum_{a_j\in C_{cur}}W^{C_{cur}}(a_j)$,则循环对每一个 $a\in RED(E_i,C_{cur})$执行:

$R=R\cup\{Des(E_i,a)\rightarrow Des(Z,D)\}$;

Return(R);

end。

4 结束语

本文对文献[1]和文献[2]算法进行了改进,指出规则支持度与属性权重有关,由知识和信息的关系给出基于条件熵的属性权重赋值法。在属性权重基础上定义缺省规则加权支持度概念,在置信度和加权支持度下给出挖掘算法。该算法能有效去除噪声,降低挖掘深度,提高缺省规则挖掘效率,对支持度阈值设置较文献[2]有很大改进,具有合理性和一定实用性。

参考文献:

[1] MOLLESTAD T, SKOWRON A. A Rough set framework for data mining of propositional default rules[A]. The 9th International Symptom Methodologies for Intelligent System[C]. ISMIS'96, Poland, June 1996, 1-4.

[2] 尹旭日,陈世福. 一种基于 Rough 集的缺省规则挖掘算法[J]. 计算机研究与发展, 2000, 37(12): 1441-1445.

[3] PAWLAK Z. Rough sets[J]. International Journal of Information and Computer, 1982, 11(5): 341-356.

[4] 刘清. Rough 集与 Rough 推理[M]. 北京: 科学出版社, 2001.

[5] 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.

[6] 张文修, 仇国芳. 基于粗糙集的不确定决策[M]. 北京: 清华大学出版社. 2005.

(责任编辑　闫纯有)

一种缺省规则挖掘算法

作者： 刘志民， 范杰， 杨珠， 庞彦军， LIU Zhi-min， FAN Jie， YANG Zhu， PANG Yan-jun
作者单位： 河北工程大学,理学院,河北,邯郸,056038
刊名： 河北工程大学学报（自然科学版） ISTIC
英文刊名： JOURNAL OF HEBEI UNIVERSITY OF ENGINEERING(NATURAL SCIENCE EDITION)
年，卷(期)： 2008,25(1)

参考文献(6条)

1. MOLLESTAD T;SKOWRON A A Rough set framework for data mining of propositional default rules 1996

2. 尹旭日;陈世福 一种基于Rough集的缺省规则挖掘算法[期刊论文]-计算机研究与发展 2000(12)

3. PAWLAK Z Rough sets 1982(05)

4. 刘清 Rough集与Rough推理 2001

5. 苗夺谦;王珏 粗糙集理论中概念与运算的信息表示[期刊论文]-软件学报 1999(02)

6. 张文修;仇国芳 基于粗糙集的不确定决策 2005