计算机科学2003Vol. 30№. 4

基于不一致数据库的缺省加权规则挖掘算法*)

刘开第　庞彦军　王义闹
(河北建筑科技学院　河北邯郸056038)

The Default Weight Regular Mining Algorithm Based on Inconsistent Data Base

LIU Kai-Di　PANG Yan-Jun　WANG Yi-Nao
(He bei Institute of Architectural Science and Technology, Handan 056038)

Abstract　In paper [2], Mollestad and Skowron propose an algorithm excavating the default regular from the inconsistent data. But, the algorithm can't fitter noises effectively and its operation is large, needs too much time and its efficiency is lower because its "up and down" searching strategy begins to search unavoidably from the upper layer which includes the most attributes. So, the paper first proposes the method of determining the attribute weight. ON the basis of the method, the paper defines the concept of the weight regular support degree and the concept of the weight regular trust degree and givs the MDWRBR algorithm, which can filter noises effectively and determine the searching direction and the stop condition and can end the operation before the regular mining conducts to the upper layer. So the algorithm reduces the operation and saves time and has some practical value.

Keywords　Rough set, Inconsistent database, Default regular mining, Attribute weight, Weight regular support

1　前言

由于获取和存储数据量的急剧膨胀，人们面对着从海量数据中发现、提取隐藏在这些数据中的有用模式。然而，实用的智能数据分析技术目前还很不成熟，这使得数据产生、存储与数据分析之间极不平衡，甚至人们在海量数据面前显得无能为力。所以，寻求快速的、有效的、智能的数据分析方法是非常重要的，这是数据挖掘领域中一项意义深远的工作。由于知识获取和知识表示等种种原因，给定的数据库常常存在着不一致性，即一些具有相同条件属性值的数据而具有不同的分类，这使得数据挖掘系统不能根据条件属性值将对象进行分类。Mollestad 在文[2]中提出了从不一致性数据中提取命题缺省规则的方法。该方法的好处在于能从不一致性数据中提取规则，特点是自上而下在网络中搜索各个节点，在每个节点上生成规则，直至所搜索的节点遍历完毕。不足之处是：①不能有效地去除噪声；②提取的规则数量通常是巨大的；③由于搜索深度是$|C|$($|C|$表示条件属性集的基数)，在规则提取和实际分类时需要大量的时间消耗，因为越是顶部节点，在计算约简时所花费的时间越多，并且产生的规则越特殊，适用范围也就越小。针对这种情况，本文依据 Rough 集理论，引入加权支持度概念，提出缺省规则挖掘算法 MDWRBR，该算法能有效过滤噪声，并且提高规则挖掘效率。

2　Rough 集理论

1982年，波兰学者 Z. Pawlak 提出 Rough 集理论，这是一种处理不完整、不确定、不一致数据的数学方法，现已被广泛应用到数据挖掘领域。为了本文需要，下面简要回顾 Rough 集理论中的有关概念。

定义1　设 $K=\langle U,A,V\rangle$ 是一个信息系统，其中 U 称为论域，由有限个研究对象构成，记作 $U=\{x_1,x_2,\cdots,x_n\}$；A 为属性集，$A=\{a_1,a_2,\cdots,a_m\}$；$V=\bigcup V_a$，V_a 称为 a 的值域，满足：任给 $x\in U$，$a\in A$，$a(x)\in V_a$。当 $A=C\cup D$，其中 C 为条件属性集，D 为决策属性集，并且满足 $C\cap D=\phi$，则称此时的信息系统为决策系统，记作 $T=\langle U,C\cup D\rangle$，为简单起见，信息系统 K 与决策系统 T 中的 V 常省略。

定义2　设 $K=\langle U,A\rangle$ 为信息系统，$S\subseteq A$，称 $IND(S)$ 为等价关系，其定义为

$$IND(S)=\{(x,y)\in U\times U\mid a(x)=a(y),\forall\ a\in S\}$$

等价关系 $IND(S)$ 将论域 U 划分为若干个等价类，记作 $U/IND(S)=\{E_1,E_2,\cdots,E_r\}$，其中每个等价类 E_i 中的对象 x 在 S 意义下是不可辩识的，或说是无法区分的。

定义3　设 $K=\langle U,A\rangle$ 为信息系统，$S\subseteq A$，令

$$M(i,j)=\{a\in S\mid a(x_i)\neq a(x_j),x_i,x_j\in U,i\neq j,i,j=1,2,\cdots,n\}$$

称 $M(i,j)$ 为差别元素，显见差别元素是由属性构成的集合。由差别元素 $M(i,j)$ 构成的 $n\times n$ 阶矩阵记作 $M(S)=(M(i,j))_{n\times n}$，称 $M(S)$ 为 S 的差别矩阵。由于当 $i=j$ 时，$M(i,j)=\phi$，且 $M(i,j)=M(j,i)$，故差别矩阵 $M(S)$ 可用其下三角部分表示。

定义4　设 $K=\langle U,A\rangle$ 为信息系统，$S\subseteq A$，$a\in S$，若 $IND(S)=IND(S-\{a\})$，则称 a 是 S 中可约去的知识。否则，称 a 是 S 中不可约去的知识。如果任意 $a\in S$ 是 S 中不可约去的，则称等价关系族 S 是独立的。否则，称 S 是相关的。

定义5　设 $K=\langle U,A\rangle$ 为信息系统，$S\subseteq A$，若 $IND(S)=IND(A)$，并且 S 是独立的，则称 S 是 A 的一个约简，A 的所有约简集合记作 $RED(A)$，用 $RED(E,A)$ 表示 A 在对象集 E 上约简的集合。

定理1　设 $K=\langle U,A\rangle$ 为信息系统，$S\subseteq A$，$P\subseteq A$，E_{S_i} 是 S

*)本文是国家自然科学基金课题(60075013)和河北省自然科学基金课题(601312)。刘开第　教授，硕士生导师，主要研究方向为不确定信息处理的理论与方法。

上的类；E_{P_j}是 P 上的类；如果 E_{S_i}是 E_{P_j}的子集，则 E_{S_i}是 $S\cup P$ 的类；若 $E_{S_i}\cap E_{P_j}\neq\phi$，则 $E_{S_i}\cap E_{P_j}$为属性集 $S\cup P$ 上的类。

3 属性权重

文[2]中的缺省规则挖掘算法是在规则置信度意义下给出的。规则置信度定义如下。

定义6 $T=\langle U,C\cup D\rangle$为决策系统，$S\subseteq C$，$E_i\in U/IND(S)$，$Z_j\in U/IND(D)$，且 $E_i\cap Z_j\neq\phi$，一条规则为：

$Des(E_i,S)\rightarrow Des(Z_j,D)$

则规则置信度定义为：

$$\mu_r(E_i,Z_j)=|E_i\cap Z_j|/|E_i| \tag{1}$$

称(1)式定义的置信度为标准置信度。

由于文[2]中仅仅考虑了规则的置信度，使该算法不能有效地过滤噪声。例如，当一个特定元组在数据库的百分比极小，而由此产生的规则的置信度为1时，则这条规则将作为有效规则提交给用户。由于产生这条规则的元组的百分比极小，完全可能是由于噪音所致，因此，为了排除噪音对规则的影响，仅仅考虑规则置信度是不够的，还需引入能够过滤噪声影响的规则"支持度"概念。而规则支持度与属性的权重有关，并且数据库中各条件属性的权重是不同的。为此，首先给出属性权重的概念，并确定属性权重的赋值方法。

3.1 Rough 集中知识与信息的关系

Rough 集中对知识作了严格定义，把知识看作是对论域的划分，从而能够对知识进行严密的分析和处理。事实上，Rough 集中的知识和信息有着密切的关系，可以用信息去表示知识。

设 $T=\langle U,C\cup D\rangle$为决策系统，$S\subseteq C$ 是 U 上的等价关系(即知识)。U 上的任一等价关系都可以看作是定义在 U 的子集组成的 σ—代数上的一个随机变量。如由 S 导出的等价类集合$\{E_1,E_2,\cdots,E_r\}$，可以看作是一个随机变量 X，其概率分布为

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

其中 $p(E_i)=|E_i|/|U|$，$i=1,2,\cdots r$，$|*|$表示集合 $*$ 的基数。

有了知识概率分布的定义，根据信息论可定义知识的熵和条件熵的概念。

定义7 设 $T=\langle U,C\cup D\rangle$为决策系统，$S\subseteq C$，$U/IND(S)=\{E_1,E_2,\cdots,E_r\}$，$U/IND(D)=\{Z_1,Z_2,\cdots,Z_k\}$，则知识 S 的熵 $H(S)$定义为

$$H(S)=-\sum_{i=1}^{r}p(E_i)\log\ p(E_i) \tag{2}$$

知识 D 相对于知识 S 的条件熵 $H(D/S)$定义为

$$H(D/S)=-\sum_{i=1}^{r}p(E_i)\sum_{j=1}^{k}p(z_j/E_i)\log\ p(Z_j/E_i) \tag{3}$$

其中 $p(Z_j/E_i)=|Z_j\cap E_i|/|E_i|$

显见，若 $S\subseteq C$，$P\subseteq C$，并且 $IND(S)=IND(P)$，则 $H(S)=H(P)$。这说明两个代数表示下的等价的数据库具有相同的信息量。

3.2 属性权重的确定方法

设 $T=\langle U,C\cup D\rangle$为决策系统，$a\in C$，则 a 的重要性大小并不在于 a 的自身，而是体现在把 a 从 $S\subseteq C$ 中去除(若 $a\in S$)或把 a 加入 S(此时 $a\in C-S$)后信息量的改变上。所以，属性 a 的权重是相对的，是相对于某个条件属性集 S 而言的。事实上，若 $S\subseteq P\subseteq C$，属性 a 对 P 来讲可能是不必要的，而 a 对 S 来说可能是必要的。所以，同一种属性 a 相对于不同的属性集合而言通常具有不同的权重；从这个意义上讲，离开属性集单独考虑属性的权重没有实际意义，并且条件属性集 C 中的各种属性不可能具有相同的权重，也不可能具有不变的权重。知识的条件熵表达了知识的信息量，那么条件熵的增量给出了属性权重的定量描述。

定义8 设 $T=\langle U,C\cup D\rangle$为决策系统，$S\subseteq C$，$C=\{a_1,a_2,\cdots,a_m\}$，则任意 $a_i\in C$，a_i 关于 S 的权重 $W_{a_i}^S$可定义为：

$$W_{a_i}^S=|\Delta H(S/a_i)|/\sum_{j=1}^{n}|\Delta H(S/a_j)| \tag{4}$$

其中 $\Delta H(S/a_i)=$

$$\begin{cases}H(D/S)-H(D/S-\{a_i\}), & \text{当 } a_i\in S\\ H(D/S\cup\{a_i\})-H(D/S), & \text{当 } a_i\in C-S\end{cases}^{(i=1,2,\cdots,m)} \tag{5}$$

显然上述定义的权重 $W_i(i=1,2,\cdots,m)$满足 $W_i\geqslant 0$，$\sum_{i=1}^{m}W_i=1$。

当 $S=\phi$ 时，说明没有属性对论域 U 划分，即 $U/INS(S)=U$，所以此时条件熵

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

即当 $S=\phi$ 时，有 $H(D/S)=H(D)$，故有下述定理。

定理2 设 $T=\langle U,C\cup D\rangle$为决策系统，$C=\{a_1,a_2,\cdots,a_m\}$，若 $S=\phi$，$a\in C$，则 a 关于 S 的权重 W_a^s 为：

$$W_a^s=|H(D/\{a\})-H(D)|/\sum_{i=1}^{m}|H(D/\{a\})-H(D)|$$

至此，对任意属性集 $S\subseteq C$，任意 $a\in C$，均可唯一地求出 a 关于 S 的权重 W_a^S。

4 缺省规则挖掘算法

设 $T=\langle U,C\cup D\rangle$为决策系统，$S\subseteq C$，$E_i\in U/IND(S)$，$Z_j\in U/IND(D)$，$E_i\cap Z_j\neq\phi$，并且 $E_i\not\subset Z_j$，这时产生缺省规则，即产生的规则不具备百分之百的置信度。缺省规则虽然不是百分之百的正确，但在大多数情况下并不影响其使用，比如，作为鸟的"鸵鸟不会飞"，并不影响"鸟都会飞"这条缺省规则的使用。由于人们知识的局限，通常不能满足特定的确定规则的条件，而只能得到缺省规则；由于缺省规则通常具有简洁、方便使用的优点，使得在实际中缺省规则有时比确定性规则更有效。由于仅仅考虑规则置信度不能有效过滤噪声，为此引入规则加权支持度的概念。

4.1 规则加权支持度

定义9 设 $E_i\in U/IND(S)$，$S\subseteq C$，$Z_j\in U/IND(D)$，且 $E_i\cap Z_j\neq\phi$，一条规则为：

$Des(E_i,S)\rightarrow Des(Z_j,D)$

则规则加权支持度定义为：

$$\mu_i=|E_i\cap Z_j|\cdot(\sum_{a_i\in S}W_{a_i}^S)/|U|\cdot|S| \tag{6}$$

其中，$|E_i\cap Z_j|/|U|$是在不考虑属性权重在通常意义下的规则支持度，它表示的是 E_i 中落在 Z_j 中的对象在 U 中占的比例，若该比例数甚小，不超过预先设置的规则支持度阈值，则可认为是噪声影响所至；$W_{a_i}^S$是 S 中属性 a_i 相对于 S 的权重，

$(\sum_{a_i\in S}W_{a_i}^S)/|S|$表示 S 中属性的平均权重。称⑤式定义的支持度为加权规则支持度。

设 E_i 和 E_i' 分别是两个不同的条件属性集 S 和 S' 的等价类，如果有

$$|E_i\cap Z_j|/|E_i|=|E_i'\cap Z_j|/|E_i'|$$

由文[2]关于规则置信度的定义，S 与 S' 没有区别。事实上 S 与 S' 中含有不同数量的条件属性，为了表明这种差别，在缺省规则置信度定义中可以引入属性权重的概念。

定义10 若一条规则为 $Des(E_i,S)\rightarrow Des(Z_j,D)$，$E_i\cap Z_j\neq\phi$，则称

$$\mu_c=\frac{|E_i\cap Z_j|}{|E_i|\cdot|S|}\cdot\sum_{a\in S}W_a^S \tag{7}$$

为规则加权置信度。显然，规则加权置信度比 Mollestad 定义的标准置信度更精细，可以区分在标准置信度相同情况下规则的优劣。

4.2 搜索策略

Mollested 和 Skowron 在文[2]给出的规则挖掘算法，是"从上而下、从细到粗"的搜索过程，由于越是靠近顶部节点包含的属性越多，顶部节点包含最多的属性。所以，从上向下搜索不可避免地要计算最耗时的多属性的顶部节点。我们采用由下而上、由粗到细的搜索策略，先从属性少的底层节点开始，由下向上进行搜索。比如，最底层可在空条件属性集上挖掘，得到的规则是决策属性值的概率分布，然后在第一层、第二层…直到第$|C|$层上挖掘。

任何一种搜索算法，必须明确地回答两个问题：①搜索进行到何时停止；②搜索的方向和原则是什么。为此给出下面定义和定理。

定义11 设 $T=\langle U,C\cup D\rangle$ 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 $E_{S_i}\in U/IND(S)$，$Z_j\in U/IND(D)$，使下式成立

$$|E_{S_i}\cap Z_j|\geqslant\frac{\mu_S^{(\min)}\cdot|U|\cdot|S|}{\sum_{a_i\in S}W_{a_i}^S} \tag{8}$$

则称 S 为期望属性集。

定理3 设 $T=\langle U,C\cup D\rangle$ 为决策系统，$S\subseteq C$，预先设定最低加权支持度阈值 $\mu_S^{(\min)}$，若对所有 $E_{S_i}\in U/IND(S)$ 和所有的 $Z_j\in U/IND(D)$ 都有：

$$|E_{S_i}\cap Z_j|<\frac{\mu_S^{(\min)}\cdot|U|\cdot|S|}{\sum_{a_i\in S}W_{a_i}^S+\max_{a\in(C-S)}W_a^S} \tag{9}$$

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 $C-S$ 中任选一种属性加入 S 后都不可能是期望属性集。任取 $a\in C-S$，$E_{(S\cup\{a\})_t}\in U/IND(S\cup\{a\})$ 则对任意 $E_{(S\cup\{a\})_t}$ 和任意 $Z_j\in U/IND(D)$，必存在 $E_{S_i}\in U/IND(S)$，使 $E_{S_i}\supseteq E_{(S\cup\{a\})_t}$，并且

$$|E_{(S\cup\{a\})_t}\cap Z_j|\leqslant|E_{S_i}\cap Z_j|<\mu_i^{(\min)}\cdot|U|\cdot|S|/((\sum_{a_j\in S}W_{a_j}^S)+\max_{a\in(C-S)}W_a^S)<\mu_i^{(\min)}\cdot|U|\cdot(|S|+1)/\sum_{a_j\in(S+\{a\})}W_{a_j}^S$$

所以，$S+\{a\}$ 不是期望属性集。

由定理知，当(9)式成立时，搜索即可停止。下面回答搜索方向的选择。

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，并入 S 构成包含集 $P\supseteq S$，则 P 给出搜索的方向。至此，可给出如下挖掘算法。

4.3 MDWRBR 算法

通过上述分析，给出基于 Rough 集的缺省加权规则挖掘算法，简记作 MDWRBR 算法。

Input：决策系统 $T=\langle U,C\cup D\rangle$，加权支持度阈值 $\mu_S(0\leqslant\mu_S\leqslant 1)$，加权置信度阈值 $\mu_C(0\leqslant\mu_C\leqslant 1)$。

$C=\{a_1,a_2,\cdots,a_m\}$，a_i 关于属性集 S 的权值记作 $W_i^S(i=1,2,\cdots,m)$；

Output：在给定决策系统 T 上的确定性规则和缺省加权规则集合。

Begin

(1)C_{N_i} 表示第 i 层节点 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 $R=\phi$；

(2)设底层节点 N_0 所对应的属性集为 C_{N_0}，在节点 N_0 上生成的加权规则 $R_{N_0}=CreateRole(U,D,C_{N_0},\mu_s,\mu_C)$；$R=R+R_{N_0}$；

(3)循环：i 从1到$|C|$，执行：

①循环：j 从1到 $C_{|C|}$，执行：

对第 i 层上的节点 N_{ij} 上生成规则

$R_{N_{ij}}=CreateRole(U,D,C_{N_{ij}},\mu_S,\mu_C)$；

$R=R+R_{N_{ij}}$；

②生成第 i－1层上产生的缺省加权规则的例外(blocks)；

(4)End。

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 $W_j^{C_{cur}}$，$j=1,2,\cdots,m$；

循环对每一个 $X\in U/IND(D)$执行

循环对每一个 $E_i\in U/IND(C_{cur})$执行

如果 $|E_i\cap X|(\sum_{a_j\in C_{cur}}W_j^{C_{cur}})/|E_i|\cdot|C_{cur}|\geqslant\mu_c|E_i\cap X|(\sum_{a_j\in C_{cur}}W_j^{C_{cur}})/|U|\cdot|C_{cur}|\geqslant\mu_S$

则循环对每一个 $a\in RED(E_i,C_{cur})$执行：

R=R∪{Des(E$_i$,a)→Des(Z$_j$,D)}；

Retarn(R)；

End

结束语 本文根据 Rough 集理论，由知识与信息关系定义了属性权重。在此基础上定义了缺省规则加权支持度、缺省规则加权置信度概念，在加权支持度和加权置信度下给出 MDWRBR 算法，该算法能有效过滤噪声，并且"由下而上"的搜索策略、明确的搜索方向和停止搜索条件，往往在到达顶层节点之前搜索就停止，降低计算复杂度减少时耗，提高缺省规则挖掘效率，算法合理，有一定实用性。

参 考 文 献

1 Pawlak Z. Rough sets. International Journal of Information and Computer, 1982,11(5):341～356

2 Mollested T, Skowron A. A Rough set framework for data mining of propositional default rules. In: The 9th In'l Sympon Methodologies for Intelligent System. ISM IS'96, Poland, 1996

3 苗本谦，王珏. 粗糙集理论中概念与运算的信息表示. 软件学报，1999,10(2)

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