维普资讯 http://www.cqvip.com

基于小生境免疫算法的柔性可信度规则提取方法*)

张丽平　闫德勤
(辽宁师范大学计算机系　大连 116029)

摘　要　研究了以免疫算法为工具对信息系统进行规则提取。在免疫算法中结合了小生境技术的共享机制，利用了共享机制中对资源的共享和竞争原理，体现了规则之间竞争与互补的关系。针对普通可信度度量方法的不足，提出了柔性可信度的概念，能够获得更好的规则。实验证明，本文采用的方法不仅在处理无冲突无噪声数据时可以得到好的结果，而且在处理噪声数据、决策类不等量划分的数据中也有很好的表现。

关键词　免疫算法，小生境技术，规则提取，柔性可信度，规则度量

A Method of Flexible Confidence Rules Extraction Based on Niching Immune Algorithm

ZHANG Li-Ping　YAN De-Qin
(Department of Computer Science, Liaoning Normal University, Dalian 116029)

Abstract　In this paper, a method of extracting rules from information system used niche immunity algorithm based on sharing mechanism is proposed. The principle of sharing and competing resource in the sharing mechanism is consistent with the relation of complementing and contending among rules. Ordinary methods of confidence measure have some shortcomings. New conception of flexible confidence is introduced in this article. The experiment proved, the new method obtains the good result not only in processing the data without conflict or noise, but also in processing the noise data and the data with the imbalance of the class distribution.

Keywords　Immune algorithm, Niche method, Rule extraction, Flexibility confidence, Rule measure

1　引言

分类规则提取是数据挖掘中的重要课题。通常是在对样本数据进行学习的基础上，采用某种分类方法产生分类规则。常用的产生分类规则的方法主要有决策树方法、神经网络方法和遗传算法等。但是，决策树方法在处理大的数据集时因为数据量过大，很难完成挖掘。神经网络方法产生的规则又难于理解。标准遗传算法虽然能够解决上述两个问题，但是在进化时可能产生退化现象，染色体之间只存在竞争关系。即使利用记忆系统保存了多个最优解，这些最优解之间也不存在互补的关系[2,5]。事实上，规则空间中的规则之间既是竞争的，又是互补的。在从样本空间中提取规则时，规则之间是竞争关系，优质规则将以更大的概率被提取出来；而对已经提取出的规则，它们之间是互补关系，共同覆盖样本集甚至整个样本空间。因此，本文以免疫算法为基础，结合小生境技术中的共享机制，提出了一种新的基于柔性可信度的规则提取方法，一方面通过共享机制的嵌入体现规则之间的互补关系，进而提高规则集的预测能力；另一方面通过提出柔性可信度的概念和新的规则重要性度量函数，重新对规则的价值进行评价，进而提取出更加合理的规则。

2　规则的表示方法和重要性度量

规则的表示方法是进行规则提取的基础。规则的度量方法也直接影响到规则提取的结果。本部分首先介绍规则的表示方法，然后考察规则重要性度量的几个因素，提出柔性可信度的概念和新的重要性度量函数。

2.1　规则的表示方法

本文用 $A\Rightarrow B$ 表示一条规则，其中 A 是决策表中条件属性取值的合取，称为规则的前件；B 是决策属性值，称为规则的后件。若某条件属性对决策不重要，即无论条件属性取何值都不会影响决策值，则该位用“0”表示。

2.2　规则的重要性度量

在进行规则提取时，选取一种好的规则度量方式十分重要。现有的文献中提出了多种规则度量的方法。文[1]中使用遗传算法进行规则提取，度量规则的公式为 $(|A\&B|/|A|)^2$，作为规则个体的适应度，其中 $|\cdot|$ 表示集合的基数，平方操作的作用是为了加速收敛。在进行规则提取时，仅考虑这一方面显然是不够的。文[2]中进行规则提取的度量公式为 $\alpha^* str(r)+\beta^* cer(r)+\gamma^* co(r)$，其中 $str(r)$、$cer(r)$ 和 $co(r)$ 分别表示规则的覆盖度、可信度和完备度。α、β、γ 是 3 个附加的参数，需要根据各种不同的实际情况进行人工调整。在评价规则 $A\Rightarrow B$ 的重要性程度时，本文主要考虑以下几个因素。

2.2.1　完备度

定义 1　规则的完备度：

$x_1=|A\&B|/|B|$

在决策类分布不均衡的情况下，较小的类因为包含的样本个数少，在提取规则时经常被忽略。完备度因素使得较小的类仍能够提取出规则，x_1 越大，说明该规则对 B 类决策越重要。

2.2.2　覆盖度

*)国家自然科学基金(60372071)资助；辽宁省教育厅高等学校科学研究基金(2004C031)资助；辽宁师范大学校基金资助。**张丽平**　硕士研究生；**闫德勤**　博士，教授，主要研究领域为模式识别、数据挖掘等。

维普资讯 http://www.cqvip.com

定义 2 规则的覆盖度：

$x_2=|A|/M$

其中，M 为样本的总数。覆盖度体现了规则在样本集合中进行归纳的能力。x_2 越大，说明该规则的归纳能力越强，越重要。

2.2.3 规则的大小

定义 3 规则的大小 $N=|A|'(|A|'\geqslant|A|)$，指在整个样本空间(不仅仅是样本集)中，满足前件 A 的个体的数量。这一因素体现着规则的预测能力。在提取规则时，希望得到 N 尽可能大的规则。N 越大，说明该规则的预测能力越强。在本文采用的方法中，该值可以根据规则的表示方法通过简单计算得出，并不需要遍历整个样本空间。

2.2.4 可信度

可信度指规则的准确程度。

定义 4 规则的原始可信度 $x_3=|A\&B|/|A|$。x_3 越大，说明该规则准确度越高，越可信。在进行规则提取时，更期望得到可信度高的规则。但是，过高的可信度要求，一方面将使得小规则被提取出的概率增加，提取出规则的个数增加；另一方面，对包含少量噪声的数据，不能体现出容噪能力。因此，在对规则的可信度要求不是特别高的情况下，可以适当降低规则的可信度，以提取出具有更高概括能力的规则。

一般地，在处理可信度时有两种方法：一是直接使用原始可信度作为规则度量的一个因素[1,2]，这样将使可信度过低的规则仍然会参与到规则的竞争中，虽然它的竞争力会减弱，但是不十分明显。这显然不符合规则提取的初衷。二是通过设置阈值 $Tr(0\leqslant Tr\leqslant 1)$，限制规则的可信度，可信度低于 Tr 的规则完全不具备竞争能力，即取阶梯函数

$$imp=\begin{cases}1 & x_3\geqslant Tr\\ 0 & x_3<Tr\end{cases}$$

但阶梯函数对边界处的处理过于严格。若存在规则 $r1$ 和 $r2$，规则的大小 $N_{r1}\gg N_{r2}$，覆盖度和完备度不相上下。设它们的可信度为 $x_3(r1)=Tr-k$，$x_3(r2)=Tr+k$，这里 k 为一个很小的值。根据阶梯函数将得到 $imp_{r1}=0$，$imp_{r2}=1$ 的结论。这样，规则 $r1$ 将完全丧失竞争力。实际上，从提取规则的角度讲，规则 $r1$ 并不比规则 $r2$ 差。考察整个规则空间，在任何阈值 Tr 附近，规则之间原始可信度 x_3 都十分接近，采用阶梯函数"一刀切"的办法显然不够合理。

因此，对可信度的度量很重要。一方面要体现出可信度高的规则更重要，另一方面还要柔性地处理可信度，对那些可信度略低的规则既适当降低其重要性，又要允许它们保留一定的竞争能力，再结合规则度量的其它因素进行衡量。

为此引入新的柔性可信度函数。

定义 5 规则的柔性可信度。

$imp=(tanh((x_3-Tr)\times a)+1)/2$

其中，a 称为陡度因子，a 越大，imp 值在 Tr 附近变化越明显，见图 1。

$$f_k=\begin{cases}1 & rule(i,k)=rule(j,k)\\ 0 & rule(i,k)\neq rule(j,k)\ \ and\ \ rule(i,k)\times rule(j,k)\neq 0\\ -1/a_k & rule(i,k)=0\ \ and\ \ rule(j,k)\neq 0\\ 1/a_k & rule(i,k)\neq 0\ \ and\ \ rule(j,k)=0\end{cases}$$

这里，a_k 为属性 k 取值的个数。如果两个规则在某属性位上取值相同，则认为它们在该属性位上完全相似；若在某属性位上取值不同，且都不为 0(0 表示可取任意值)，则认为它们在该属性位上完全不相似。规则 i 在该属性位上取值为 0，而规则 j 在该属性位上取值不为 0，那么它们有 $1/a_k$ 的概率相似。又因为可取任意值的是规则 i，它在该属性位上是包含规则 j 的，故其相似度要被降低，取负值。反之则取正值。两个规则的相似度就是它们所有属性位上相似度的平均值。

通过调节 a 的值，得到以阈值 Tr 为中心、随 a 增大而减小的邻域区间，该区间称为柔性度量区间。对落在该区间的规则，产生一个(0,1)区间上的数，为其新的柔性可信度。对原始可信度很高($\gg Tr$)的规则，新的柔性可信度仍然很高，接近于 1。对原始可信度很低($\ll Tr$)的规则，新的柔性可信度也会很低，接近于 0。对原始可信度在阈值 Tr 附近的规则，其新的柔性可信度将分布在(0,1)区间上。在大致满足阈值要求的前提下，把可信度阈值看作一个强制性不强的因素，成为本文规则重要性度量公式的一部分。

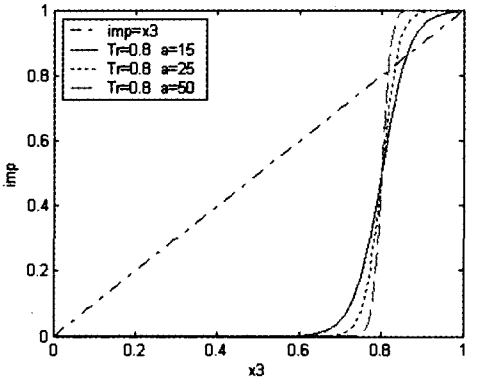


图 1 不同梯度因子下的柔性可信度函数

综合以上因素，本文提出的规则重要性度量的公式为

$$f(x)=(x_1+x_2)\times N\times imp \tag{1}$$

3 小生境共享机制的实现方法

小生境技术源于进化论中的小生境概念[8]，指含有有限资源的生态环境。共享机制指特定环境下，物种共享有限资源，物种之间既要通过协调达到共同进化的目的，又要相互竞争有限的资源，最后维持一种合理的平衡状态。规则之间也是这样的竞争和合作关系。把整个样本空间看作有限的资源，提取出的规则集看作共享该资源的物种。物种要最大化利用资源，即最大化覆盖样本空间。对样本空间中的任意位置，规则之间要进行竞争和进化，使得最合理的规则被保留。本文通过使用相似度函数调整抗体的亲和力来体现共享机制。

相似度指规则 $A\Rightarrow B$ 与其他规则的相似程度，在本文讨论的方法中，特指规则与免疫记忆系统中抗体的相似程度。进行规则提取时，相似度大的规则，尤其是被包含的规则要受到抑制。

定义 6 规则 i 的相似度函数为

$\mathrm{sim}_i=\max(s(i,j))$

其中，$s(i,j)$ 表示规则 i 与免疫记忆系统中规则 j 的相似程度，设 n 为条件属性的个数，则 $s(i,j)=(\sum_{k=1}^{n}f_k)/n$。用 $rule(i,k)$ 表示规则 i 在属性 k 上的取值，f_k 为规则 i 与规则 j 在属性 k 上的相似程度。

维普资讯 http://www.cqvip.com

而一个规则与免疫记忆系统的相似度就是该规则与免疫记忆系统中规则相似度的最大值。

例如：设决策表中有 1 个决策属性、3 个条件属性，且每个条件属性可取值只有 1 和 2 两种，用 0 表示可取任意值。若免疫记忆系统中有规则 $isr1$：010→1，$isr2$：111→1，$isr3$：201→1，讨论规则 ra：110→1 和 rb：220→1 的相似度 sim_{ra} 和 sim_{rb}。

$s(ra,isr1)=(1/2+1+1)/3=5/6$

$s(ra,isr2)=(1+1-1/2)/3=3/6$

$s(ra,isr3)=(0+1/2-1/2)/3=0/6$

$\mathrm{sim}_{ra}=\max(5/6,3/6,0/6)=5/6$

$s(rb,isr1)=(1/2+0+1)/3=3/6$

$s(rb,isr2)=(0+0-1/2)/3=-1/6$

$s(rb,isr3)=(1+1/2-1/2)/3=2/6$

$\mathrm{sim}_{rb}=\max(3/6,-1/6,2/6),3/6$

sim_r 越大，说明规则 r 与免疫记忆系统越相似，这样的规则将越不重要。由例可知，相似度 $\mathrm{sim}_{ra}>\mathrm{sim}_{rb}$，规则 rb 比规则 ra 对规则的提取更为重要。事实上，规则 ra 被免疫记忆系统中的规则 $isr1$ 包含，这样的规则没必要再被提取出来。

本文使用相似度对抗体的亲和力进行修改，使得与免疫系统相似的规则被适当抑制，最终达到增大记忆系统覆盖度的目的。

4 基于免疫算法的规则提取方法

免疫算法[6]是受生物免疫系统的运行原则和机制启发而产生的一种智能搜索方法。在对样本集合进行学习时，把每个样本看作一个抗原，由免疫系统产生相应的抗体群进行进化，从中选择亲和力 aff 最高的赋予记忆细胞，形成免疫记忆系统。由于记忆细胞个数有限，在记忆细胞用尽后，采用置换策略，用亲和力最高的抗体取代免疫记忆系统中最差个体。免疫算法具有学习、识别、记忆和自我平衡的能力，能够提取出较好的分类规则。但是，免疫算法只负责提取较好的规则。事实上，这些规则之间还存在合作关系，共同覆盖原始样本空间，而基本免疫算法不能体现这一特点。因此，在基本免疫算法的基础上，结合了小生境技术中的共享机制进行规则提取。共享机制的使用主要体现在抗体亲和力的修改上，即使用公式(2)调整抗体的亲和力，使得与免疫记忆系统相似的抗体亲和力下降，与免疫记忆系统差异较大的抗体亲和力上升。亲和力调整后，抗体在免疫过程中的竞争能力被改变，进而使免疫结果体现出抗体之间的合作关系。

$$aff=aff^*(1-sim) \tag{2}$$

算法流程如下：

```
初始化可信度阈值 Tr 和陡度因子 a；
while(未识别全部样本或未达到学习次数)
{
  for 样本集中每个样本：
  {
    if(样本已被识别)continue;
    else
    {
      初始化原始抗体种群；
      接种疫苗；
       使用公式(1)计算抗体亲和力 aff；
       使用公式(2)进行小生境处理，调整抗体亲和力；
       while(未达到免疫进化代数)
       {
         在繁殖池中进行繁殖和变异；
         选取最优且未在免疫记忆系统中出现过的变异体 x1 和种
         群中最差抗体 x2；
         If(aff(x1)>aff(x2))用 x1 置换 x2；
       }
       进化结束，寻找尚未在记忆池中出现过的抗体赋予记忆细
       胞；
    }
  }
}
```

5 实验结果及分析

分别选取了无冲突无噪声、有噪声、决策类分布不均衡的数据进行了相关实验。

5.1 MONK1 数据

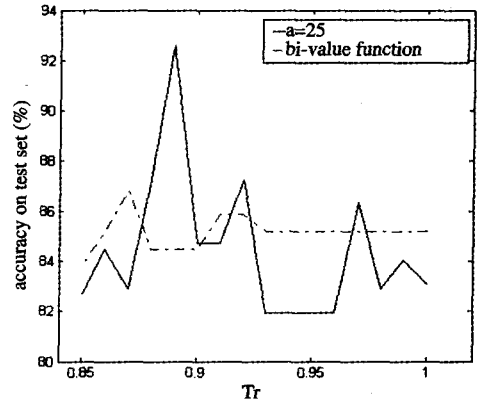
MONK1 数据是无冲突无噪声的数据。对其中的训练集 MONK1. train，样本数据个数为 $M=124$，条件属性个数为 $C=6$，决策属性个数为 $D=1$。共有两个决策类，两个决策类是等量划分的。设种群大小 $P=20$，记忆池大小 $N=M$(即假设记忆池足够大)，抗体繁殖池大小为 $Q=100$，免疫进化最大代数 $GG=50$。

取可信度函数为

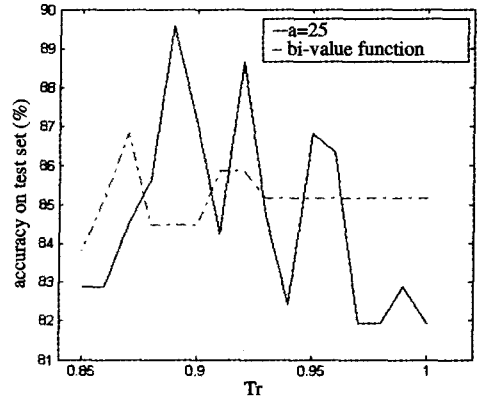
$imp(x)=(tanh((x-0.85)\times 25)+1)/2$

时，可以得到标准的结果，说明该方法适用于无冲突无噪声的数据。

5.2 MONK3 数据



(a)



(b)

图 2

MONK1 数据是有 5% 噪声的数据。对其中的训练集 MONK3. train，样本数据个数为 $M=122$，条件属性个数为 $C=6$，决策属性个数为 $D=1$。设种群大小 $P=20$，记忆池大小 $N=M$(即假设记忆池足够大)，抗体繁殖池大小为 $Q=100$，免疫进化最大代数 $GG=20$。

使用取不同可信度阈值 Tr 和陡度因子 a 的可信度函数与使用阶梯函数进行对比实验，效果如图 2 所示。图中曲线的最高点是用训练集提取出的规则集合在测试集上的最大正

维普资讯 http://www.cqvip.com

确识别率。可见，选用了新的柔性可信度函数后，对提取出的规则集而言，最大识别率要优于使用阶梯函数的效果。

5.3 lenses 数据集

lenses 数据集是完备、无冲突、无噪声的，其中有 24 个样本，每个样本包含 4 个条件属性和 1 个决策属性。决策属性有 3 个决策类，且这 3 个决策类是不等量划分的，分别占样本总数的 16.67%，20.83%和 62.5%。在可信度阈值 $Tr=0.95$ 和陡度因子 $a=25$ 的条件下，得到 9 条规则，使得样本的识别率为 100%，与标准结果相同。这说明，本文的规则度量函数可以很好地处理决策类不等量划分的情况。

结论 本文使用免疫算法作为分类规则提取的工具，结合小生境技术，提出了柔性可信度的概念和新的规则重要性度量方法，将其作为免疫算法中抗体亲和力的度量公式进行了仿真实验。实验结果表明，无论是处理无噪声的数据，还是有噪声或决策类不均衡的数据，本文的方法都可以取得较好的效果。

参 考 文 献

1 Khoo L P, Zhai L Y. A prototype genetic algorithm-enhanced rough set-based rule induction system. Computers in Industry, 2001, 46(1): 95～106
2 He M, et al. Genetic algorithm based multi-knowledge extraction method for rough set. Mini-Micro System, 2005, 26(4): 651～654
3 Freitas A A. On rule interestingness measure. Knowledge-Based System, 1999, 12(4): 309～315
4 黄席樾, 等. 现代智能算法理论及应用. 北京: 科学出版社, 2005
5 Ghosh A, Nath B. Multi-objective rule mining using genetic algorithms. Information Sciences, 2004, 163(1-3): 123～133
6 Wu Z J, et al. A Chromosome-based Evaluation Model for Computer Defense Immune Systems. In: Proceedings of 2003 IEEE Congress on Evolutionary Computation, 2003. 1363～1369
7 Zheng J G, Wang L, Jiao L C. An immune algorithm for generalized rule induction. In: Proceedings of CIE International Conference of Radar Proceedings, 2001. 1023～1026
8 Goldberg D E, Richardson J J. Genetic algorithms with sharing for multimodal function optimization. Genetic algorithms and their application. In: Proceedings of the Second International Conference of Genetic Algorithms, 1987. 41～49

（上接第 184 页）

由表 3 所示集值决策信息系统在关系 $R_A^{\cap}$ 和 $R_A^{\cup}$ 下的广义决策辨识矩阵，可以得到基于 $R_A^{\cap}$ 的广义决策辨识矩阵中的非空元素是基于 $R_A^{\cup}$ 的广义决策辨识矩阵中的相应元素的子集。但是由关系 $R_A^{\cap}$ 得到的最优广义决策规则与由关系 $R_A^{\cup}$ 得到的最优广义决策规则并无关联。

结论 通常我们所处理的信息表中属性的值是单一的并且是确定的，但这只是一种理想状态。在实际情况中，我们所得到的数据往往既不单一也不确定，可能缺失部分数据，可能是数据集的一个子集。本文用集值信息系统来处理数据不完全的信息系统。集值信息系统是处理不完备信息系统的一种方法。本文在集值信息系统上定义了两种关系：相容关系和优势关系。同时，给出了集值信息系统在这两种关系下的上下近似的定义及其性质。在关系 $R_A^{\cup}$ 下关于属性集 B 的下近似和上近似优于在关系 $R_A^{\cap}$ 下关于属性集 B 的下近似和上近似。讨论了集值决策信息系统基于两种不同关系的广义决策约简，给出了广义决策约简的判定定理及辨识矩阵，进而从集值决策信息系统中提取了最优广义决策规则。

参 考 文 献

1 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Boston, 1991
2 Stefanowski J, Tsoukias A. On the Extension of Rough sets under Incomplete Information. In: Zhong N, et al. eds. RSFDGrC 1999, LNAI 1711, 1999. 73～82
3 Kryszkiewicz M. Rule in incomplete information systems. Information Science, 1999, 113: 271～292
4 Grzymala-Busse J W. Incomplete Data and Generalization of Indiscernibility Relation, Definability, and Approximations. In: Slezak, et al. eds. RSFDGrC 2005, LNAI 3641, 2005. 244～253
5 Slowinski R, Vanderpooten D. A Generalized Definition of Rough Approximations Based on Similarity. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(2): 331～336
6 张文修, 梁怡, 吴伟志. 信息系统与知识发现. 北京: 科学出版社, 2003
7 王国胤. Rough 集理论与知识获取. 西安: 西安交通大学出版社, 2001

（上接第 187 页）

小结 本文探讨了粗糙集理论中最小决策规则的求取方法，提出决策依赖度的定义，尝试从最短的条件属性组合中提取尽可能多的决策规则。只有现有长度的决策规则无法完全覆盖所有样本时，才会考虑增加决策规则的长度。由于条件属性的所有组合是条件属性数的幂集，本文提出了 3 种减少计算复杂性的方案：1）引入跳跃系数 λ，先计算较长的条件属性组合的决策依赖度 β。如果 β 为 0，则表明该属性组合的所有子集均不必再计算；如果 β 不为 0，再回溯到其较短的子集进行计算。如此跳跃式计算，在属性较多的情况下可以大大减少计算复杂性。而回溯功能则可以避免决策树学习方法经常会遇到的收敛到局部最优答案的情况。2）在计算中只对具有相同决策值的样本进行等价类划分，这样处理的好处是避免了对含有不同决策值的等价类的无用划分。3）设计 Remain 集合，只针对 Remain 中的样本进行等价类的划分。随着 Remain 中样本数的减少，计算量会大幅下降。实验表明，本文所提出的基于决策依赖度的跳跃式决策规则求取方法可以直接应用于不完备信息系统，因此具有良好的实用价值。

参 考 文 献

1 曾华军, 张银奎译. 机器学习. 北京: 机械工业出版社, 2003
2 文硕频, 乔胜勇, 陈彩云, 等. 基于决策树的不完全决策表的数据及规则提取. 计算机应用, 2003, 23(11): 17～22
3 孙长嵩, 董西国, 张健沛. 一个基于粗糙集和决策树的最简分类规则集生成算法. 哈尔滨工程大学学报, 2002, 23(5): 87～91
4 Yang Jing, Wang Hao, Hu Xue-Gang, et al. A new classification algorithm based on rough set and entropy. International Conference on Machine Learning and Cybernetics, 2003, 1: 364～367
5 Quinlan J R. C4. 5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann
6 Witten I H, MacDonald B A. Using concept learning for knowledge acquisition. International Journal of Man-Machine Studies, 1988, 27: 349～370
7 Mitchell T M, Utgoff P E, Banerji R. Learning by experimentation: acquiring and modifying problem-solving heuristics. Machine Learning, 1: 163～190
8 王东锴, 梁梁. 不完全决策表中的决策规则发现. 计算机科学, 2001, 28(5. 专刊): 83～85
9 文硕频, 乔胜勇, 陈彩云, 等. 基于决策树的不完全决策表的数据及规则提取. 计算机应用, 2003, 23(11): 17～22