医疗诊断知识挖掘的区间合并与RS混合方法*)

蒋伟进　林小红

(株洲工学院计算机系　株洲412008)　(株洲市一医院内分泌科　株洲412000)

摘　要　针对医学诊断知识获取问题,提出了基于Rough Sets理论的知识获取方法,利用该理论对数据进行分析,推理出可能规则,并提出了一种概率优化规则。通过实例分析,具体说明了该方法的实现步骤,包括连续信息系统的离散化、信息系统的约简、决策规则提取、决策模型生成等。讨论了知识处理的完整过程,能够有效地解决专家系统中知识获取的瓶颈问题,为人工智能技术在医学诊断领域的应用提供了新的思路。

关键词　Rough Set,医学诊断规则,连续信息系统,离散化,规则获取

Research on Extracting Medical Diagnosis Rules Mining Based on Rough Sets Theory

JIANG Wei-Jin　LIN Xiao-Hong

(Department of Computer, Zhuzhou Institute of Technology, Zhuzhou 412008)

(Zhuzhou Institute of Physic, Xiangya of Center & South University, Zhuzhou 412008)

Abstract　Analyze how to extract medical diagnosis rules from medical cases. Based on the rough set theory, a way to acquiring knowledge is bring forward. Using this theory, the data is analyzed, possible rules are proposed, and a optimized probability formula was showed. By analyzing instances, the implement step of the way is explained, including discreting continuous information system, reducting information system, acquiring decision rules and generating decision model, and so on. In the end, the whole process of knowledge acquisition is discussed, and this option can effective solve the choke point problem of acquiring knowledge in expert system. At the same time, it also provides new brainchild to solve the artificial intelligence technology's application in the field of medicinal diagnosing.

Keywords　Rough sets, Medicine diagnose rule, Continuous information system, Discretization, Acquisition rules

1　引言

近20余年来,人工智能与专家系统在局部领域应用取得的成功激发了医学、计算机与系统工程领域的专家对医疗专家系统开发的兴趣。在开发医学专家系统过程中,知识获取是关键问题之一,对于医学这样的复杂系统,传统机器学习方法在特定领域有成功应用,基于神经网络的学习可获取病症分类知识,基于语义网络的学习可以用于医学语言的理解等。

医学诊断问题是基于知识的序贯诊断问题,诊断是基于规则的推理过程,医生通过各种途径获取知识,形成推理网络,而病例数据存储在数据库中,因此如何从病例数据库提取诊断规则成了研究主题,即知识获取。

波兰数学家Pawlak教授提出了粗集(Rough Sets,RS)理论研究不完整数据、不精确知识的表达、学习、归纳的方法。目前,RS作为一种新的数学分析工具,能够处理不确定、不精确、不完整和不一致的信息,能够进行信息简化和经验中获取知识,该理论已被广泛地应用在专家系统、决策支持系统、机器学习、智能控制、归纳推理、模式识别等领域,特别适合于信息系统的约简。但RS在医疗诊断领域中的应用,大都没有考虑:(1)决策表中存在冲突对象的约简问题;(2)决策表中的条件属性的重要性问题;(3)待诊断对象和诊断规则不完全匹配时的匹配程度问题。问题(1)使得粗糙集学习到的规则的强度同于实际水平,问题(2)和(3)采用很强的对象和诊断规则匹配方式,而实际上,可能要求对象和规则的部分匹配,按某一阈值决定诊断结论,因此,本文提出一种含有冲突对象的决策表的规则获取方法,考虑属性重要性、规则匹配程度、规则强度和诊断结论阈值进行医疗病理快速诊断。在他人工作的基础上,进一步研究从病例中获取医学诊断规则的方法。

2　Rough Sets理论

RS理论将知识与分类联系在一起,认为知识源于分类,并用等价关系形式化表示,因此,可以这样理解:知识是使用等价关系R对离散空间U的划分,记成为$U/R=\{X_1,X_2,\cdots,X_n\}$,称$X_i$为$U/R$的等价类,$|U/R|$表示分类$U/R$的等价类的个数。

知识表达系统KBS(knowledge based system)可表示为$S=\langle U,C,D,V,f\rangle$,$U$是对象的集合,$A=C\cup D$是属性集合,子集$C$和$D$分别称为条件属性集和决策属性集,$V=\bigcup_{a\in A}V_a$是属性值的集合,$V_a$表示属性$a\in A$的属性值范围,$f:U\times A\to V$是一个信息函数,它指定$U$中每一个对象$x$的属性值,知识表达系统可以方便地用表格表达,称为决策表。

知识系统S,$\forall X\subseteq U$,当X为某些R基本范畴的并时,称X是R可定义{Rdefinable},否则X为不可定义(Rundefinable)。R可定义集是论域的子集,它可在S中被精确定义,称作R精确集(Rexact sets);而R不可定义集不能在S中被定义,不可定义集也称作R的粗集(Rrough sets)。RS的概念

*)基金项目:国家自然科学基金资助课题(60373062);湖南省杰出中青年专家科技项目基金资助(02JJYB012);教育部重点科研项目基金资助(02A056);湖南省卫生厅科技基金资助项目(2001-Y89)。蒋伟进　副教授,高级工程师,硕士,主要从事人工智能及软件理论研究。林小红　教授,主任医师,主要研究方向:代谢内分泌与生物信息处理。

就是由此而来，可以定义两个精确的集合来描述：

$R_-(X)=\bigcup\{\in U/R;Y\subseteq X\}$

$R^-(X)=\bigcup\{\in U/R;Y\cap X\neq\phi\}$

分别称它们为 X 的 R 下近似(R lower approximation)和 R 上近似(R upper approximation)。

集合 $BN_R=R^-(X)-R_-(X)$ 定义为 X 的边界；$POS_R(X)=R_-(X)$ 定义为 X 的正域；$NEG_R(X)=U-R_-(X)$ 定义为 X 的负域。

RS 的不精确性是由于边界域的存在引起的，引入精度的概念定义为

$$a_R(X)=\frac{|R_-(X)|}{|R^-(X)|}$$

其中，$|\cdot|$ 表示集合的基数，$a_R(X)$ 反映了在知识 R 下对集合 X 的了解程度，显然有 $0\leqslant a_R(X)\leqslant 1$。

知识 R 粗糙度(roughness)定义为 $\rho_R=1-a_R(X)$

信息系统 S 的决策属性 D 和条件属性 C 的依赖记为 $C\Rightarrow_k D$，其中 k 为下面将定义的依赖度，依赖度定义为

$$k=\gamma_C(D)=\frac{|POS_C(D)|}{|U|}$$

依赖度 $\gamma_C(D)$ 表示条件属性 C 下能够确切划入决策类 U/D 的元组的比率，即表达了决策属性对条件属性的依赖程度，称决策属性 D 是 C 的 k 度可导。

3 诊断规则的获取

3.1 连续属性的离散化

RS 理论只处理离散知识系统，而医学诊断知识系统中很多属性取连续值，连续取值属性需要离散化，关键是怎样离散化可减少信息损失。

医学诊断的经验数值，医学中“金标准”，是基于大量数据的统计结果，有鉴于此，这里将 Kerber 提出的基于区间合并 CHI-MERGE 与 RS 方法结合，提出一种属性离散化以分析研究胃溃疡的 HSV(highly selective vagotomy)信息系统分析离散化问题为例，该系统采集了胃溃疡的116病例数据，HSV 信息系统主要数据如表1。表1中各属性的含义见表2。

表1 HSV 信息系统

编号	1	2	3	4	5	6	7	8	9	10	11	类别
1	1	46	12	0	5.6	79	50	4.4	19	119	22.6	1
2	0	27	3	1	12.5	58	15	7.3	26	120	120	31.2
…	…	…	…	…	…	…	…	…	…	…	…	…
116	0	21	4	0	14.7	182	31	26.8	27.5	379	104.2	4

设 X 为任一连续条件属性，d 为其定义域，在信息系统中样本数为 n，将 d 划分成 k 个区间 $I_j(j=1,2,\cdots,k;k\geqslant 2)$，采用平均分配的办法实现初步区间划分，划分的粒度达到每个区间取值相对集中的要求，得到 X 的初始划分点为 x_p，$1\leqslant p\leqslant k-1$ 且 $k<n$，不妨设 $x_1\leqslant x_2\leqslant\cdots\leqslant x_{k-1}$，设决策属性 X 存在 m 个分类 $X_i(I=1,2,\cdots,m)$，则有

n_{ij}——区间 I_j 中属于分类 X_i 的样本数目；

n_j——区间 I_j 中的样本数目，且有 $n_j=\sum_{i=1}^{m}n_{ij}$；

n_i——分类 X_i 中的样本数目，且有 $n_i=\sum_{j=1}^{k}n_{ij}$；

样本数 $n=\sum_{i=1}^{m}n_i=\sum_{j=1}^{k}n_j=\sum_{i=1}^{m}\sum_{j=1}^{k}n_{ij}$

因此，得到一个 m 行 k 列的矩阵 T

$$T=\begin{pmatrix}n_{11}&n_{12}&\cdots&n_{1k}\\n_{21}&n_{22}&\cdots&n_{2k}\\\vdots&\vdots&\vdots&\vdots\\n_{m1}&n_{m2}&\cdots&n_{mk}\end{pmatrix}$$

T 的第 j 列为 $T_j=(n_{1j}\,n_{2j}\cdots n_{mj})^T$，则有 $T=(T_1\,T_2\cdots T_k)$。

Kerber 判断准则为

$$\mathscr{R}^2(T_q,T_{(q+1)})<X^2(\alpha,m-1)$$

其中 $\mathscr{R}^2(T_q,T_{q+1})=\sum_{i=1}^{m}\sum_{j=q}^{q+1}\dfrac{(n_{ij}-n_j\sum_{k=q}^{q+1}n_{ik})^2}{n_j\sum_{k=q}^{q+1}n_{ik}}$

满足准则的相邻间隔 I_q 和 I_{q+1} 合并，直到所有的间隔点都不满足上述准则。得到新的间隔点序列为：$x_1\leqslant x_2\leqslant\cdots\leqslant x_l$；$l\leqslant k-1$。

基于 Kerber 准则给出求取区间分割点的算法步骤为：

(1)均匀离散化取值区间，并达到粒度要求。

(2)建立矩阵 T。

(3)用 Kerber 准则考察所有相邻区间，若没有相邻区间满足准则，则转(6)，否则继续。

(4)合并(3)中得出的满足准则的相邻区间。

(5)转(3)。

(6)输出得到的区间分割点值的集合。

经处理后 HSV 信息系统中连续取值的分割点的取值见表3。

上述分割点的求取没有考虑其它属性，考虑到其余属性的分类，分割点可能存在冗余，需进一步合并分割区间，使离散后在条件属性下的等价类最小化。下面进一步讨论全局一致分割点的求取。

对于分割点 x_i，若去除了分割点 x_i 不改变样本在信息系统中相容性，则该分割点为冗余分割点，所谓不改变其相容性，即不改变条件属性和决策属性的决定关系。

这里给出基于信息熵的全局一致分割点的求取方法，R 的一个等价类为 $R=\{X_1,X_2,\cdots,X_k\}$，则

$$H(R)=-\sum_{i=1}^{k}\frac{|X_i|}{U}\log_2\frac{|X_i|}{|U|}$$

分割点的改变，关系所蕴涵的信息熵将改变。

给出全局一致分割点的求取算法如下：

(1)根据冗余分割点的定义，判断所有的分割点，将冗余分割点添加到集合 RP 中；

(2)若 $RP=\phi$，则转(5)，否则继续；

(3)从 RP 中，选择信息熵最小的分割点 x_l；

(4)令 $RP=\phi$，转(1)；

(5)用新的分割点重新离散化样本数据，得到新的信息系统。

全局一致分割点算法求得的 HSV 信息系统的分割点见表4，该分割点与医学统计经验数据基本一致。

HSV 信息系统离散化处理后的信息系统见表5。

表2 HSV信息系统中属性的含义

编号	属性含义	属性单位	取值					说明
			0	1	2	3	4	
1	性别		男	女				
2	年龄	年	连续取值					
3	持续时间	年	连续取值					
4	溃疡并发症		无	出血	大量出血	穿孔	幽门狭窄	
5	HCl浓度	mmol/100ml	连续取值					基本分泌物
6	胃液分泌	ml/h	连续取值					
7	剩余胃液	ml	连续取值					
8	排除胃酸	mmol/h	连续取值					
9	HCl浓度	mmol/100ml	连续取值					组胺刺激下的分泌物
10	胃液分泌	ml/h	连续取值					
11	最大胃酸排除量	mmol/h	连续取值					
类别	分类			很好	好	尚可	不好	

表3 分割点取值表

属性编号	属性分割点						说明
2	21	35	71				
3	0	0.5	3.2	32.0			
5	1.0	1.9	4.5	26.1			基本分泌物
6	5.0	71.2	153.5	387.7	525.0		
7	2.0	13.8	48.9	104.3	202.9	254.0	
8	0.48	1.97	3.82	16.70	39.10		
9	1.6	9.6	18.7	42.3			组胺刺激下的分泌物
10	21.0	99.9	152.0	263.2	627.0		
11	2.1	16.2	25.4	42.0	123.7	151.4	

表4 分割点表

属性编号	属性分割点					
		0	1	2	3	
2	21	35	71			
3	0	0.5	3.2	32.0		
5	1.0	1.9	4.5	26.1		
6	5.0	71.2	153.5	525.0		
7	2.0	48.9	104.3	254.0		
8	0.48	1.97	16.70	39.10		
9	1.6	9.6	18.7	42.3		
10	21.0	99.9	263.2	627.0		
11	2.1	16.2	25.4	42.0	151.4	

表5 离散化处理后的信息系统

编号	1	2	3	4	5	6	7	8	9	10	11	类别
1	1	1	1	0	2	1	1	1	2	1	1	1
2	0	0	0	1	2	1	0	1	2	1	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…
116	0	0	1	0	2	2	0	2	2	2	3	4

3.2 信息约简

经上述方法离散化后的信息系统需进一步约简，约简的目的是去除相同元组、去除冗余的属性。

信息系统$S=(U,A,V,f)$中，$\forall x_1,x_2\in U$，$A=C\cup D$，若$\forall a\in A$，都有$f(x_1,a)=f(x_2,a)$，则x_1,x_2是重复的元组。

求取信息系统$S=(U,A,V,f)C$的D简化，可以通过构造信息系统的差别矩阵$M(C,D)$，求得差别函数$f_D(C)$，应用布尔代数中的分配律和吸收律进行化简得到$f_D(C)$的析取范

式。

差别矩阵 $M(C,D)$的元素为

$$\delta(x,y)=\begin{cases}\{a\in C:f(x,a)\neq f(y,a)\}\\ [x]_c\neq[y]_c \text{ and } [x]_D\neq[y]_D\\ \phi[x]_C=[y]_C \text{ or} [x]_D=[y]_D\end{cases}$$

基于上述数学描述，给出信息系统约简的算法：

(1)建立信息系统 S 的差别矩阵；

(2)化简差别矩阵 $M(C,D)$，化简的方法是，对矩阵中任何一个 m，如果存在另外的元素 m'，使得 $m\subseteq m'$，则以 ϕ 替代矩阵元素 m；

(3)根据简化的差别矩阵 $M'(C,D)$，构造差别函数 $f_D(C)$；

(4)化简差别函数 $f_D(C)$为析取范式；

(5)选择 $f_D(C)$中的任一析取项，析取项所包含的属性组成的集合即为信息系统 S 的约简属性集。

医疗快速诊断系统约简 HSV 信息系统的简化属性组中属性的编号为{3,4,6,9,10}。

3.3 诊断规则获取

在离散决策表中，往往含有冲突对象，这时一般的做法是剔除冲突对象，仅对非冲突的对象进行规则获取，这样虽然规则强度为1，但并不适合所有的对象，考虑这种情况，提出一种含有冲突对象的离散决策表的规则获取方法，具体步骤如下：(1)设有初始决策表 DTori，合并 DTori 中的重复对象，形成决策表 DTcom，设 DTcom 中对象的数目为 m。记录 DTcom 中每一个对象代表 DTori 中的对象的个数，$C_i^{ori\text{-}obj}$，$1\leqslant i\leqslant m$；(2)消除 DTcom 中的冲突对象构成决策表 DTlast，设 DTlast 中对象的个数为 n，对于 DTcom 中的冲突对象 Obj，$1\leqslant i\leqslant m$。若 $C_i^{ori\text{-}obj}$，$1\leqslant s\leqslant m$，为最大者，则选择 Obj，为 DTlast 中的一个对象，若 $C_j^{ori\text{-}obj}=C_i^{ori\text{-}obj}$，$1\leqslant i,j\leqslant m$，$i\neq j$，则任选其一作为 DTlast 中的一个对象，同时 DTlast 中每一对象 Obj_k，$1\leqslant k\leqslant n$，代表 DTori 中对象中的个数 $C_k^{ori\text{-}obj}$，$1\leqslant k\leqslant n$，和一个比值 $Ratio_k$，$1\leqslant k\leqslant n$，该比值为所选代表对象在 DTlast 进行条件属性简化；(4)去除 DTlast 的冗余属性值；(5)对 DTlast 中的每一行产生决策规则；(6)规则合并，并设置规则强度，设对象 Obj_k，$1\leqslant k\leqslant p\leqslant n$，产生相同的规则 r_k，$1\leqslant k\leqslant p\leqslant n$；对象 Obj_k，$1\leqslant k\leqslant p\leqslant n$，的记录为 $C_k^{ori\text{-}obj}$，$Ratio_k$，$1\leqslant k\leqslant p\leqslant n$；记 $STH=Round\left(\left(\sum_{k=1}^{p}Ck^{ori\text{-}obj}\times Ratio_k\right)\right)/\sum_{k=i}^{p}Ck^{ori\text{-}obj}$，其中 $Round(\cdot)$表示取整，则规则 $r_k(1\leqslant k\leqslant p\leqslant n)$合并后的强度为 STH。

这样，在获取的规则中，规则强度为1.0的规则为确定性规则，而其他的规则为非确定性规则。

通过上文约简，可以获得约简信息系统 $S=(U,A,V,F)$，从约简信息系统提取决策规则。

对信息系统 $S=(U,A,V,f)$中，$A=C\cup D$，C 为条件属性，D 为决策属性，任一元组确定了一条 C 基本公式和 D 基本公式，$des_C(x)$表示规则的条件部分，$des_D(x)$表示规则的结论部分。

$$des_C(x)=\bigwedge_{c\in C}(c,f(x,c))$$

$$des_D(x)=\bigwedge_{d\in D}(d,f(x,d))$$

规则可以表示为：$r_x^{C,D}:des_C(x)\rightarrow des_D(x)$

用$[des_C(x)]\in U/ind(C)$表示包含 x 的用条件属性集 C 描述的一个范畴，同样，$[des_D(x)]\in U/ind(D)$表示包含 x 的用决策属性集 D 描述的一个范畴，若$[des_C(x)]$包含于$[des_D(x)]$，则 $r_x^{C,D}$ 为确定性决策规则；否则为可能规则，也可以用上、下近似来表示，若 $x\in C_-([des_D(x)])$，则该规则为确定性规则，记为：

$$r_x^{C,D}:des_C(x)\xrightarrow{D}des_D(x)$$

若 $x\in C^-([des_D(x)])$，则该规则为可能性规则，记为：

$$r_x^{C,D}:des_C(x)\xrightarrow{P}des_D(x)$$

为了衡量规则的可信度和下文规则匹配的度量，对每条规则都引入下列指标，这些指标的计算是基于原始的信息系统，而不是上文讨论的约简后的信息系统：

定义规则的精度：

$$a_r=\frac{|[des_C(x)]\cap[des_D(x)]|}{|[des_C(x)]|}$$

定义规则的可能性：

$$k_r=\frac{|[des_C(x)]\cup[des_D(x)]|}{|[des_D(x)]|}$$

定义规则的强度 $s_r=|des_C(x)\rightarrow des_D(x)|$。也就是说属性取值与规则 $des_C(x)\rightarrow des_D(x)$的属性取值相同的元组数。

定义了规则和描述规则的各类指标，下文在规则匹配和应用中将讨论这些指标在系统中的应用。规则 $r_x^{C,D}$ 统一表示为

$$r_x^{C,D}:des_C(x)\rightarrow des_D(x)\quad a_r,k_r,s_r$$

以上给出规则描述，但是在规则求取中，需要约简，采用文[7]中提出的构造决策矩阵 M、化简决策矩阵、构造决策函数 B_i，求取信息系统的约简规则。无论确定性规则还是可能性规则，提出约简算法：

(1)建立信息系统 S 的决策矩阵 M；

(2)化简决策矩阵 M。化简的方法是，对矩阵 M 每一行的元素，任何一个元素 m，如果存在另外的 m'，使得 $m\subseteq m'$，则以 ϕ 替代矩阵元素 m；

(3)根据简化的决策矩阵 M'，构造决策函数 B_i；

(4)化简决策函数 B_i 为析取范式；

(5)B_i 中的任一析取项，所包含的属性的对应与信息系统 S 中的取值即为信息系统 S 一条规则。

根据该算法求取 HSV 信息系统的决策规则以表格的形式表示(见表6)。

由确定性规则、可能性规则以及描述规则的参数的定义，显然有 $a_r=1$对应着信息系统 S 的确定性规则，而 $a_r<1$对应着 S 的可能性规则。

3.4 诊断规则的表达

对医疗诊断而言，连续条件属性为诊断征兆，离散属性类型为诊断结论，因此仅对条件属性值离散化。由决策表中的第 s 列 a 的聚类的上下界 $CLas(i)$、$CUas(i)$，$1\leqslant i\leqslant p$，和式(2)，得诊断规则，第 r 条规则的形式如下：

R:if $CL(i)\leqslant Fe(x,a1)\leqslant CU(i)$

and $CL(j)\leqslant Fc(x,a2)\leqslant CU(J)\cdots$

and $CL(k)\leqslant Fc(x,aq)\leqslant CU(k)\cdots$

then $f(x,k,)=vd$

其中，$F(x,aq)$ 表示属性 a_{qi}的连续值，$q=1,2,\cdots,1\leqslant q\leqslant q\leqslant n,i,j,\cdots,k$ 分别表示属性列 $a_1,a_2,\cdots,a_k$ 的第 $I,j,\cdots,k$ 个聚类，其中，$2\leqslant i\leqslant \max i,2\leqslant j\leqslant \max j,\cdots,\max k$ 分别为属性列 $a_1,a_2,\cdots,a_k$ 的最大聚类数目，ud 表示决策表中诊断结论第 t 个离散属性值，其中 $t=1,2\cdots,m_D$，这里，m_D 为诊断结论的数目。

应用决策矩阵提取出来的决策规则组成决策模型，其中规则是存在冗余的，且随着信息系统 $S=(U,A,V,f)$的规模的增加，规则数成几何级数增长，需进一步简化。

表6 经简化决策属性后的决策结果

简化的决策属性编号					决策结果				简化的决策属性编号					决策结果			
3	4	6	9	10	类别	a_r	k_r	s_r	3	4	6	9	10	类别	a_r	k_r	s_r
0	3				1	1	0.01	1	2	3		2		2	1	0.05	1
1	0		2		1	1	0.11	9	1	2		0		2	0.43	0.16	3
2	2		2		1	1	0.02	2	2	2		1		2	0.5	0.05	1
2	1		1		1	1	0.06	5	0	1	1	0		2	0.05	0.05	1
2	1		2		1	0.75	0.07	6	2	0	2	1		2	0.33	0.11	2
1	1		2		1	1	0.05	4	0	0				3	1	0.25	2
2	3		0		1	1	0.02	2	1	3	0	1		3	0.75	0.38	3
2	0	1	2		1	0.83	0.06	5	1	3	2	0		3	1	0.13	1
1	1	0	1		1	1	0.01	1	1	1	2	0	2	3	1	0.63	5
1	1	2	1		1	1	0.01	1	1	0	2	0	2	3	0.67	0.25	2
1	0	1	1		1	1	0.06	5	2	4		2		4	1	0.14	2
2	1	0	0		1	1	0.01	1	1	2		1		4	1	0.21	3
2	4	2	0		1	1	0.01	1	1	3	2	1		4	1	0.07	1
1	1	2	0	1	1	0.31	0.05	4	2	1	1	0		4	0.33	0.14	2
2	0	1	1	2	1	1	0.02	2	2	0	0	0	0	4	1	0.14	2
1	0	0	0	0	1	1	0.02	2	2	0	2	2	2	4	0.6	0.21	3
2	0	0	0	1	1	1	0.05	4	1	0	2	0	1	4	1	0.07	1
2	0	0	1	1	1	1	0.02	2	2	0	0	1	0	4	0.63	0.36	5

针对决策矩阵 M，求取信息系统的最小决策模型即转化求取 M 的最小覆盖问题是NP难题。将决策属性 C 在不同的决策类下多个约简规则中，选择精度因子、可能性因子和强度因子都较高的规则组成决策规则，这样筛去了噪声规则，同时也会去除一定的有用信息，生成的决策模型代替最小决策模型(minimal decision model，MDM)，取得较好的应用效果。

在医学诊断中，精度因子、可能因子和强度因子取值小的规则并不一定表示是噪声，因为这样的规则往往代表了一种"顿悟"或"灵感"，在医学里，不可以否认"顿悟"或"灵感"的存在。有鉴于此，不应该删除那些连接强度低的规则，而是同样作为一种可能留待医生进一步处理，由这样的决策规则组形成的决策模型实际上是一决策推理网络。

推理网络表示了规则间的关系，然而描述规则的量化值精度因子、可能因子及强度因子该怎样合成呢，规则的合成可以有最小、最大、乘积、有界和、代数和、Yager 算子等等，可以根据现实问题选用。这里对于精度因子、可能因子选用乘积算子，而强度因子选用最小算子，直观上是符合一般的思维过程的。

设有两条规则

$r_x^{C,D}1: des_C(x) \rightarrow des_D(x)\ \alpha_1, k_1, s_1$

$r_x^{C,D}2: des_C(x) \rightarrow des_D{}'(x)\ \alpha_2, k_2, s_2$

两条规则满足 $D=C'$，组合而成的新规则为

$r_x^{C,D}1: des_C(x) \rightarrow des_D(x)\ \alpha, k, s$

其中 $\alpha=\alpha_1\alpha_2, k=k_1k_2, s=\min(s_1, s_2)$

根据规则推理网络，医生可以将医学诊断的前因后果联系起来，合成推理的结论为医生提供决策依据。

结论 提出了一种基于区间合并和粗糙集理论的辅助医疗诊断方法，该方法包括粗糙集规则学习和诊断规则匹配两个过程。其中过程考虑了样本中的重复对象和冲突对象，使获得的诊断规则能够覆盖所有的学习样本，并得到规则强度；在诊断规则匹配时，根据规则中条件属性的属性重要性、条件属性匹配的程度、规则强度以及诊断结论阈值得到诊断结论，从而使得到的结论更客观，最后，通过实验验证了该方法的有效性。

参考文献

1 Pawlak Z. Rough sets[J]. Intl. Journal of Information and Computer Science, 1982,11(5): 341～356

2 Missaoui R. Extracting expert and approximate rules formdatabases [J]. Corum. of ACM,1995,38(11): 201～211

3 Chen C C. Rough set boundaries as a tool for learning rulesfrom examples[J]. In methodologies for Intelligent Systems. North-Holland,New York,1989. 281～288

4 Ras Z W. Learning concept in rough environment: an optimization procedure[J]. In methodologies for Intelligent Systems, North-Holland,New York,1987. 256～263

5 Grzynala Busse J W. Learning from examples on rough multi-sets [J]. In Methodologies for Intelligent Systems, North-Holland, New York,1987. 268～272

6 Shan N. An incremental Learning algorithm for constructing decision rules[R]. Technical CS 93 02,Department of Computer Science,University of Regina,Canada. 1993

7 Slowinski R,et al. 'RoughDAS' and 'RoughCLASS' software implications of the rough sets approach[A]. Ziarko W. Rough Sets, Fuzzy Sets and Knowledge Discovery [C]. Sastatchewan: 1999. 221～229

8 Yao Y Y. Constructive and algebraic methods of the theory of rough sets [J]. Journal of Information Sciences,1998,109(1): 21～47

9 Jiang Weijin. Technologies of Virtual Enterprise and Dynamic Modeling Based on MA & BP. Information and Control,2002,31(4): 329～335

10 Jiang Weijin. Research and Implementation of Distributed MSP Algorithm Based on GA & MAS. Computer Science,2002,29(9): 443～447