第 34 卷 第 6 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.6 Computer Engineering March 2008
·博士论文· 文章编号：1000—3428(2008)06—0022—03 文献标识码：A 中图分类号：TP391

决策信息系统中挖掘全部决策规则的算法

王树锋，吴耿锋，潘建国
(上海大学计算机学院，上海 200072)

摘 要：在粗糙集理论的基础上，对决策信息系统中边界区域的数据进行研究，提出一种从边界区域数据中挖掘决策规则的算法——近似序列决策规则挖掘算法。在 16 个 UCI 数据集上的测试表明，该算法在规则的准确度和平均前件长度 2 个指标上优于 ID3 算法，能简洁、高效地挖掘出决策信息系统中的全部决策规则，为挖掘未知知识提供了新的思路。针对挖掘出的全部决策规则，提出新的确定性度量和一致性度量指标，用以准确地反映决策规则的性能。

关键词：决策信息系统；粗糙集边界区域；决策规则；规则度量指标

Algorithm for Extracting All Decision Rules from Decision Information Systems

WANG Shu-feng, WU Geng-feng, PAN Jian-guo
(School of Computer, Shanghai University, Shanghai 200072)

【Abstract】 Extraction Algorithm of Approximate Sequence Decision Rules (EAASDR) extracting decision rule from border region of rough set is proposed. It can extract all knowledge from decision information systems. Comparison tests between EAASDR and ID3 in 16 UCI data sets show that the algorithm is prior to ID3 in the accuracy of rule set and the average condition number of rule sets. A new rule measure criterion of certainty and consistency is proposed in order to accurately reflect the performance of all decision rules extracted from decision table.

【Key words】 decision information system; border region of rough set; decision rule; rule measure criterion

粗糙集理论[1-3]处理的对象是一个具有对象和属性关系的数据表，称为决策信息系统。该理论建立在分类的基础上，将知识(决策规则)理解为对数据的划分，由在特定空间上的上近似集和下近似集构成。上下近似集之间的差集被认为是无法确认的数据，称为边界区域。对于海量数据，为了更加精确地描述知识，边界区域的数据一般不应忽略。基本的思路是先求出原决策信息系统的某个约简，将该约简下的数据从系统中删除，得到一个新系统。对该新决策信息系统再求得新约简，将该新约简下的数据从新系统中删除，得到一个新的决策信息系统，重复上述过程，直到系统不能有其他约简或系统中的数据全部能分类为止。本文提出一个近似序列决策规则挖掘算法(Extraction Algorithm of Approximate Sequence Decision Rules, EAASDR)，处理粗糙集边界中的数据，给出了新的决策规则度量标准，用来评价挖掘出的全部决策规则。

1 粗糙集中边界数据的处理

给定决策信息系统 $S=(U,C\cup D)$，对于任意 $X\subseteq U$，有一个满足偏序关系 $R_1\succeq R_2\succeq\cdots\succeq R_n(R_i\in R)$ 的等价关系族 $P=\{R_1,R_2,\cdots,R_n\}$。定义 $\overline{P}X=\overline{R_i}X$，称为 X 的 P 上近似集。定义 $\underline{P}X=\bigcup_{i=1}^{n}\underline{R_i}X_i$，其中，$X_1=X;X_i=X-\bigcup_{k=1}^{i-1}\underline{R_k}X_k$，称为 X 的 P 下近似集。集合 $bn_P(X)=\overline{P}X-\underline{P}X$ 称为 X 的 P 边界域。由 P 形成的近似分类结果和 X 的协调度定义为 $H(P,X)=\frac{|\underline{P}X|}{|X|}$，其中，$|\bullet|$ 表示集合的基数，显然 $H(P,X)\in[0,1]$。当 $H(P,X)=0$ 时，表示近似分类结果和待分类对象最不协调；当 $H(P,X)=1$ 时，表示近似分类结果和待分类对象完全协调，即现有的知识完全可以精确地描述待分类对象，在进行近似分类时能把待分类对象完整地分类。

采用近似分类方案的结果，把先验知识 X 分解为不同近似分类(处在不同粒度层次)的子类的并集 $X=X_1\cup X_2\cup\cdots\cup X_k$，第 k 次已能够最为精确地表示先验类 X，每一个子集都具有不同的粒度(约简属性)，是在相应粒度下能够精确表达知识的最大子集。

2 近似序列决策规则挖掘算法

2.1 一般描述

算法的一般描述如下：

输入 决策信息系统 $S=(U,C\cup D)$

输出 分类结果规则集

(1) 给定决策信息系统 $S=(U,C\cup D)$，计算出相对核 $CORE_D(C)$；/*通过计算各条件属性对决策属性的重要性 $\sigma_{CD}(C')$，得到相对核*/

(2)If $CORE_D(C)\neq\phi$，Then 初始属性集 $P_1=CORE_D(C)$ and 集合 $E=P_1$

Else 初始属性集 $P_1=\{c_1\}$ and $E=P_1$；/*对 $\forall c\in C$，计算 c 和 D 之间的依赖度 $\gamma_{[c]}(D)$，选择 $\gamma_{[c_1]}=\max\{r_{[c]}(D),c\in C\}$，作为初始属性集*/

(3)计算决策分类 $U/D=\{Y_1,Y_2,\cdots,Y_d\}$；

基金项目：上海市科委重点攻关基金资助项目(035115028)
作者简介：王树锋(1968－)，男，博士研究生，主研方向：人工智能，知识管理；吴耿锋，教授、博士生导师；潘建国，博士研究生
收稿日期：2007-04-23 **E-mail：**xueaimin7000@163.com

(4)等价关系族 $P=\{P_1\}$, $i=1$, $U^*=U$，动态分类结果 $B=\phi$，决策规则集 $Rule=\phi$；

(5)计算 $U^*/IND(P_i)=\{X_{i1},X_{i2},\cdots,X_{ik}\}$；

(6) $B'=\{X_k\in U^*/IND(P_i)\mid X_k\subseteq Y_j, where Y_j\in U/D, j\in\{1,2,\cdots,d\}\}$，$Rule'=\phi, \forall X_k\in B'$。输出决策规则 $Rule'=\{des_{P_i}(X_k)\to des_D(Y_j)\}$，其中，$Y_j\in U/D$ 且 $Y_j\supseteq X_k$；$Rule=Rule\cup Rule'$；$B=B\cup B'$；

(7) $B^*=\bigcup_{X\in B} x$；

If $B^*=U$ Then goto (8)

Else { $U^*=U^*-B^*$；$i=i+1$；对 $\forall c\in C-E$，计算 c 关于 E 对 D 的重要性 $\sigma_{(\{c\}\cup E)D}(\{c\})$，令 $\sigma_{(\{c_2\}\cup E)D}(\{c_2\})=\max\{\sigma_{(\{c\}\cup E)D}(\{c\}),c\in C-E\}$，则 $P_i=P_{i-1}\cup\{c_2\}$，并将 P_i 归入 P 中成为其中的一个等价关系，goto (5);}

(8) B 即为动态分类结果，$Rule$ 即为决策规则。算法结束。

2.2 算法分析

该算法通过计算决策表的相对核或条件属性对决策属性的依赖度，找到初始属性集(步骤(1)、步骤(2))，根据初始属性集，计算决策表的等价类，形成信息系统的一个约简(步骤(5))，得到相应的决策规则(步骤(6))。如果决策规则没有覆盖信息系统中的全部样本，就将规则中已经包含的样本及其属性从原信息系统中删除(步骤(7))，形成一个新的信息系统，在新信息系统下继续进行同样的规则抽取过程，直到所抽取的规则中包含了信息系统的全部样本和属性。

在粗糙集理论中，信息系统的一个约简就是寻找样本集合中无矛盾的样本，进行属性约简时，每一个属性的增减都会影响知识粒度的变化及其近似分类的结果。从信息系统中抽取知识，就是从复杂知识表示中删除例外的知识，得到“不同简洁程度”的知识，也就是不同粒度的知识[4]。该算法包含了粒度计算的思想，让知识粒度在动态变化中逐渐细化[5]。

实际上，知识就是“规则＋例外”，知识粒度依赖于例外的数量，例外越少，知识越精确，知识粒度的动态细化意味着知识的简洁程度(精度)在变化，因此，该算法具有变精度粗糙集模型[6-7]的思想，突破了Pawlak模型分类是完全确定的限制，使分类具有某种程度的“包含”和“属于”特性。

Pawlak 模型处理的对象是已知的，从模型中得到的结论仅适用于这些对象，基于变精度粗糙集模型的思想，可以将小规模对象集中得到的结论应用到大规模数据集中去，因此，该算法具有一定的泛化能力，它不像使用变精度粗糙集模型那样，需要做大量的计算工作。

2.3 示例验证

用关于肺炎诊断的例子来表示算法的执行，抽取该决策信息系统中的全部决策规则。肺炎诊断表如表 1 所示。

表 1 肺炎诊断表

病例号	发烧	咳嗽	X 光成像	血沉	听诊	诊断
1	高	剧烈	片状	正常	水泡声	肺炎
2	中	剧烈	片状	正常	水泡声	肺炎
3	低	轻微	点状	正常	干鸣声	肺炎
4	高	中度	片状	正常	水泡声	肺炎
5	中	轻微	片状	正常	水泡声	肺炎
6	无	轻微	索条状	正常	正常	肺结核
7	高	剧烈	空洞	快	干鸣声	肺结核
8	低	轻微	索条状	正常	正常	肺结核
9	中	轻微	点状	快	干鸣声	肺结核
10	低	中度	片状	快	正常	肺结核
11	低	轻微	点状	正常	干鸣声	肺炎
12	高	剧烈	空洞	快	干鸣声	肺结核

条件属性集 $C=\{a,b,c,d,e\}$，其中，a 代表发烧；b 代表咳嗽；c 代表 X 光成像；d 代表血沉；e 代表听诊；决策属性集 D=$\{f\}$，f 代表诊断。$U/D=\{\{1,2,3,4,5,11\},\{6,7,8,9,10,12\}\}$，$U/C=\{\{1\},\{2\},\{3,11\},\{4\},\{5\},\{6\},\{7,12\},\{8\},\{9\},\{10\}\}$。

算法执行过程如下：

(1)求相对核 $CORE_D(C)$，其计算公式为 $\sigma_{CD}(C')=\gamma_C(D)-\gamma_{C-C'}(D)$，其中，$C'\subseteq C$。经计算，$\sigma_{CD}(\{a\})=0$，$\sigma_{CD}(\{b\})=0$，$\sigma_{CD}(\{c\})=0$，$\sigma_{CD}(\{d\})=0$，$\sigma_{CD}(\{e\})=0$，因此，$CORE_D(C)=\phi$。

(2)计算各属性和 D 间的依赖度。其计算公式为 $\gamma_{\{C'\}}(D)=|pos_{C'}(D)|/|U|$，其中，$C'\subseteq C$。经计算，$\gamma_{\{a\}}(D)=1/12$，$\gamma_{\{b\}}(D)=0$，$\gamma_{\{c\}}(D)=4/12$，$\gamma_{\{d\}}(D)=4/12$，$\gamma_{\{e\}}(D)=7/12$。$P_1=\{e\}$，$E=P_1$。

(3) $U/D=\{\{1,2,3,4,5,11\},\{6,7,8,9,10,12\}\}$。

(4) $P=\{P_1\}$，$i=1$，$U^*=U$，$B=\phi$，$Rule=\phi$。

(5) $U/P_1=\{\{1,2,4,5\},\{3,7,9,11\},\{6,8,10\}\}$。

(6)可得 $B=\{\{1,2,4,5,\},\{6,8,10\}\}$，决策规则

$Rule=\{des_{\{e\}}(\{1,2,4,5\})\to des_D(\{1,2,3,4,5,11\})$

$des_{\{e\}}(\{6,8,10\})\to des_D(\{6,7,8,9,10,12\})\}$

如果只用粗糙集的上、下近似，只能抽取出这 2 条规则，采用本算法，还可在边界区域中继续抽取规则，直到抽取出全部规则。

(7) $\bigcup_{x\in B} x=\{1,2,4,5,6,8,10\}\neq U$，执行 else 语句，计算剩下的属性 a,b,c,d 关于 e 对 D 的重要度。经计算，

$$\sigma_{(\{a\}\cup\{e\})D}(\{a\})=\gamma_{\{a,e\}}(D)-\gamma_{\{e\}}(D)=5/12$$

$$\sigma_{(\{b\}\cup\{e\})D}(\{a\})=\gamma_{\{b,e\}}(D)-\gamma_{\{e\}}(D)=2/12$$

$$\sigma_{(\{c\}\cup\{e\})D}(\{a\})=\gamma_{\{c,e\}}(D)-\gamma_{\{e\}}(D)=2/12$$

$$\sigma_{(\{d\}\cup\{e\})D}(\{a\})=\gamma_{\{d,e\}}(D)-\gamma_{\{e\}}(D)=5/12$$

由于 a,d 关于 e 对 D 的重要度相等，因此任意选择其中一个属性 a 进行计算。$P_2=\{a,e\}$，$P=\{P_1,P_2\}$，转到步骤(5)。重复执行可得

$B=\{\{1,2,4,5\},\{3,11\},\{6,8,10\},\{7,12\},\{9\}\}$

$Rule=\{1:des_{\{e\}}(\{1,2,4,5\})\to des_D(\{1,2,3,4,5,11\}),$

$2:des_{\{e\}}(\{6,8,10\})\to des_D(\{6,7,8,9,10,12\}),$

$3:des_{\{a,e\}}(\{3,11\})\to des_D(\{1,2,3,4,5,11\}),$

$4:des_{\{a,e\}}(\{7,12\})\to des_D(\{6,7,8,9,10,12\}),$

$5:des_{\{a,e\}}(\{9\})\to des_D(\{6,7,8,9,10,12\})\}$

由于此时 $\bigcup_{x\in B} x=U$，因此算法结束。最终抽取出的全部规则集 $Rule$ 如表 2 所示。

表 2 抽取出的全部决策规则

规则号	发烧	听诊	诊断
1	低	水泡声	肺炎
2	低	干鸣声	肺炎
3	低	正常	肺结核
4	高	干鸣声	肺结核
5	中	干鸣声	肺结核

按照上述步骤，抽取得到了全部决策规则。对表 1 中的数据，采用属性约简和值约简结合的方法可以抽取出相同的规则，表明该算法运行的结果是正确可信的。该算法将属性约简和值约简方法统一起来，是挖掘全部规则的新思路。

2.4 模拟实验

模拟实验在 UCI 的 16 个数据集上进行，用 EAASDR 和 ID3 算法作对比实验。对每一个数据集，75%的数据用于训练算法产生规则，其余 25%用作测试。规则集的准确度和复

杂度 2 个指标作为标准评价算法。准确度用测试集实验得到的规则的准确度来描述，复杂度包括规则的个数和每个规则的平均前件(条件)个数这 2 个指标。

表 3 列出了 2 个算法在准确度和平均数 2 个指标上的数据。结果表明，EAASDR 算法在 11 个数据集上的性能优于 ID3。对于抽取的规则集的数量，ASDREA 算法要比 ID3 多，这是因为用 EAASDR 抽取出了数据集中的全部规则，而 ID3 算法则进行了一定的修剪。表 4 列出了 EAASDR 在训练集、测试集上规则的平均前件个数和 ID3 平均前件个数。虽然在某些数据集上，EAASDR 算法抽取的规则前件比 ID3 要多，但其平均值优于 ID3。

表 3　规则平均准确度和规则平均数的比较

Data set	Average accuracy/(%)		Average no. of rules	
	EAASDR	ID3	EAASDR	ID3
Arrhythmia	75.7±1.3	76.8±0.9	33.7±3.4	16.0±0.5
Breast-cancer	96.5±0.5	93.9±0.5	66.0±2.5	63.2±1.5
Breast-W	96.5±0.3	93.8±0.4	59.2±3.6	39.2±2.6
Dermatology	95.7±0.5	88.5±0.5	30.9±5.8	16.6±6.4
Echocardiogram	68.5±3.9	68.6±2.8	19.9±4.7	12.7±2.5
Heart-C	80.6±1.4	78.9±0.9	23.7±2.6	16.5±1.2
Heart-H	79.6±0.9	77.3±1.4	21.4±4.2	13.0±4.2
Hepatitis	79.2±2.8	81.8±1.7	17.0±2.9	9.0±2.8
Horse-colic	82.1±1.5	79.0±1.3	52.5±2.5	35.3±2.1
Hypothyroid	99.1±0.1	99.0±0.1	334.5±1.2	188.3±0.9
Liver-disorders	70.1±0.9	65.5±1.5	28.6±1.4	15.6±1.2
lymphography	80.4±2.0	81.6±2.3	18.5±1.6	16.8±1.9
Pima-Indians	75.4±1.1	74.2±0.9	78.6±1.3	35.9±0.9
Primary-tumors	40.4±2.3	41.9±1.6	29.8±0.9	19.8±1.0
Sick	98.1±0.1	98.1±0.2	348.6±4.6	119.6±3.9
Thyroid-benchmark	98.4±0.1	99.2±0.1	589.7±5.9	265.5±4.9
average	82.3±1.2	81.1±1.1	109.5±3.1	55.2±2.4

表 4　规则前件平均长度的比较

Data set	Average no. of condition accuracy		
	EAASDR (training)	EAASDR (test)	ID3
Arrhythmia	141.7	150.4	115.3
Breast-cancer	51.1	63.8	33.7
Breast-W	27.4	39.3	71.5
Dermatology	71.0	85.3	67.5
Echocardiogram	13.5	18.5	62.2
Heart-C	74.3	88.0	65.2
Heart-H	44.5	55.8	81.3
Hepatitis	18.1	27.6	60.5
Horse-colic	59.0	73.6	89.0
Hypothyroid	42.4	51.1	165.6
Liver-disorders	101.7	127.6	92.7
lymphography	41.1	48.5	45.9
Pima-Indians	92.5	98.2	134.3
Primary-tumors	156.9	187.1	369.7
Sick	61.0	77.9	115.3
Thyroid-benchmark	85.4	88.2	46.5
average	67.6	80.1	101.2

3　度量全部决策规则性能的新指标

利用 EAASDR 算法可以抽取出决策信息系统中的全部决策规则，对这些决策规则如何度量和评价是要研究的另外一个问题。本文将传统的决策规则确信度和支撑度进行扩充，形成了度量全部决策规则的 2 个新指标：总体确信度度量 C 和总体支撑度度量 A。

3.1　修正决策规则评价指标的依据[1,8]

设 $S=(U,A)$ 是一个决策信息系统，$A=C\cup D$，$C\cap D=\phi$，其中，C 为条件属性集；D 为决策属性集。令 X_i 和 Y_j 分别代表 U/C 与 U/D 中的各个等价类，$des(X_i)$ 和 $des(Y_j)$ 分别表示等价类 X_i 和等价类 Y_j 的描述。决策规则定义为 $r_{ij}:des(X_i)\rightarrow des(Y_j), Y_j\cap X_i\neq\phi$，规则的确信度定义为 $\mu(X_i,Y_j)=|X_i\cap Y_j|/|X_i|$，$0<\mu(X_i,Y_j)\le 1$。规则的支撑度定义为 $s(X_i,Y_j)=|X_i\cap Y_j|/|U|$，$0<s(X_i,Y_j)\le 1$。

在决策信息系统 $S=(U,A)$ 中，全体决策规则是由 $\{Z_{ij}\mid Z_{ij}:des(X_i)\rightarrow des(Y_j), X_i\cap Y_j\neq\phi\}$ 所构成的。如果 $\psi=\{Z_1,Z_2,\cdots,Z_s\}$ 是一个规则簇，那么 ψ 的确信度定义为 $\mu(\psi)=\frac{1}{S}\sum_{k=1}^{s}\mu(Z_k)$，它是规则簇的整体确信度度量。决策信息系统的规则生成的方法很多，对同一问题可能得到不同的结果。如果 Z 是决策表 $S=(U,C\cup D)$ 中的一条决策规则，经过将条件属性集 $P\subseteq C$ 的取值细化为规则簇 $\psi=\{Z_1,Z_2,\cdots,Z_s\}$，这个变化使得 $\mu(\psi)\ge\mu(Z)$；经过将决策属性集 $Q\subseteq D$ 的取值细化为规则簇 $\psi=\{Z_1,Z_2,\cdots,Z_s\}$，这个变化使得 $\mu(\psi)\le\mu(Z)$。经典的度量规则方法都是针对某一具体规则，没有给出决策信息系统整体决策效果的度量。虽然可以将定义 $\mu(\psi)=\frac{1}{S}\sum_{k=1}^{s}\mu(Z_k)$ 推广为整个决策信息系统决策性能评价参数，但是这个度量仅仅反映了整体决策规则的简单平均度量，忽略了规则本身的权重、支撑度以及规则之间的一致性问题，因此，有必要给出度量决策信息系统整体决策性能评价的新评价指标。

3.2　新指标

设 $S=(U,A)$ 是一个决策信息系统，$A=C\cup D$，$C\cap D=\phi$，$U/IND(C)=\{X_1,X_2,\cdots,X_m\}$，全体规则由 $\{Z_{ij}\mid Z_{ij}:des(X_i)\rightarrow des(Y_j), X_i\cap Y_j\neq\phi\}$ 构成，决策表 $S=(U,A)$ 的总体确信度度量 C 定义为

$$C(S)=\frac{1}{|U|}\sum_{i=1}^{m}\frac{|X_i|}{N_i}\sum_{j=1}^{Ni}\frac{|Z_{ij}|}{|U|}\mu(Z_{ij})$$

其中，N_i 为等价类 X_i 对应的决策规则个数；$\frac{|Z_{ij}|}{|U|}$ 表示规则 Z_{ij} 在论域 U 中的支撑度，其中，$|Z_{ij}|=|X_i\cap Y_j|$；$\mu(Z_{ij})$ 表示 Z_{ij} 的规则确信度。

决策信息系统 $S=(U,A)$ 的总体支撑度度量 A 定义为

$$A(S)=\sum_{i=1}^{m}\frac{|X_i|}{|U|}[1-\sum_{j=1}^{Ni}\mu(Z_{ij})(1-\mu(Z_{ij}))]$$

其中，N_i 为等价类 X_i 对应的决策规则个数；$\mu(Z_{ij})(1-\mu(Z_{ij}))$ 表示规则 Z_{ij} 的不一致性；$\mu(Z_{ij})$ 表示 Z_{ij} 的规则确信度。

3.3　新指标的性质

总体确信度度量 C 不仅反映了决策信息系统的整体决策的支撑效果，也反映了决策信息系统整体决策的确信度情况。利用该指标得出结论：在条件分类不变的情况下，决策分类越细，决策信息系统的整体决策效果越差；在决策分类不变的情况下，条件分类越细，决策信息系统的整体决策效果越差。

总体支撑度度量 A 给出了决策表整体决策的一致性度量。在决策表中，如果决策粒度不变，条件粒度越细，评价参数 A 越小，决策表整体决策的一致性效果越差。

4　结束语

用粗糙集处理海量信息时，粗糙边界中的数据规模可能很大，本文提出的近似序列决策规则挖掘算法利用粗糙集的上下近似，逐步从粗糙边界中获取规则，直到挖掘出海量信息中全部决策规则或挖掘出满足用户需求的规则为止。该算法使规则在动态变化中逐步细化，包含了粒度计算的思路，同时也体现了变精度粗糙集的思想，具有一定的泛化能力。该算法稍加改造，就可用于本体的自动构建中。

(下转第 27 页)