含序信息的粗集方法研究*)

孙成敏　刘大有　孙舒杨

(吉林大学计算机科学与技术学院 吉林大学符号计算与知识工程教育部重点实验室　长春 130012)

摘　要　经典粗集理论给出了不可识别、上近似、下近似、简式和核等概念，其核心思想是运用条件属性集导致的知识粒子来近似决策属性集导致的知识粒子，进而推导出规则。这些知识粒子的实质是根据存在于属性值间的等价关系得到的，而事实上可能存在某些属性，其属性值内部存在序关系，与其它某属性间存在语义关系，这样的属性称为标准。本文所研究的粗集方法，考虑标准所携带的这些信息，推导出含有序信息的规则，并探讨使推导的规则更加完全和一致。本文给出了含序粗集方法(CORS)的定义、数据分析以及规则生成方法，并提出了一种更加合理的质量近似公式以及生成规则的四条原则。

关键词　标准，优先序，有序决策表，支配关系，含序粗集方法

Containing Order Rough Set Methodology

SUN Cheng-Min　LIU Da-You　SUN Shu-Yang

(College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012)

Abstract　In classical rough set theory which gives definitions of indiscernibility relation, upper approximation, lower approximation, reduct and core, the main idea is approximating knowledge granules which come from decision attributes set employing knowledge granules from condition attributes set, hence generating rules. These knowledge granules are obtained according to equivalence relation in essence, it is possible there exist attributes which contain preference order relation among their values and correlate semantically with other attributes. Such attributes are called criteria. Rough set methodology involved in this paper takes into account these information which criteria carry, deduces rules containing order information, and discusses to keep rules set more complete and consistent. In this paper definition of containing order rough set (CORS) methodology and other concerned notations are introduced, method of data analysis and rules generation are formalized. Moreover a more rational approximation quality measure and four principles of generating rules are provided.

Keywords　Criteria, Preference order, Ordered decision table, Dominance relation, CORS

1　简介

初始的粗集方法(RS)[1]使用属性集及其对应的属性值来描述论域中的对象，用于近似知识的粒子是根据不可识别关系(等价关系)建立的。然而该方法不能揭示与标准(含有优先顺序的属性，criteria)有关的不一致问题。这些原始粗集方法未必能检测到不一致，可能丢失重要的信息。而且，原始的粗集方法不能推导出含有序信息的规则，即得不到更加有意义和泛化的规则。本文使用常规属性(不含有优先顺序的属性)和标准共同描述对象，我们将这样的含有顺序信息的粗集方法称为含序粗集方法(Containing Order Rough Set CORS)，也称为MCDA方法[4](Multicriteria Decision Analysis)。在CORS中，给定一个对象的集合，描述该集合对象的条件属性中至少有一个是标准，并且所有对象被划分进含有优先顺序的决策类(即决策属性是一个标准)，而且条件属性中的标准和含序的决策标准语义相关，这种新的粗集方法CORS能检测到基于支配原则的不一致并且通过支配关系实现决策类的近似。

所谓支配关系，指的是相同的常规属性和标准描述对象 x 和对象 y，如果说对象 x 支配对象 y，则必须满足两个条件：x 和 y 在任一常规属性上的取值是相等的；x 在任一标准上的取值都优于 y 在该标准上的取值。建立于支配关系上的支配原则如下：考察两个对象 x 和 y，二者包含相同的常规属性和标准，且满足条件部分的每个标准与决策部分中的至少一个标准语义相关；如果对象 x 的条件描述支配对象 y 的条件描述，则 x 的决策描述也应该支配对象 y 的决策描述，否则 x 与 y 不满足支配原则。

在本文中，将属性从性质上分为常规属性和标准，其中常规属性建立于不可识别关系之上，而标准建立于支配关系之上。S. Greco 等人将常规属性分为定性的(qualitative)属性和定量的(quantitative)属性，分别研究不可识别关系和相似关系[2]。Gunther Gediga 等人形式化地推导了这种含序粗集

*)国家自然科学基金重大项目(60496321)：非规范知识处理的基本理论和核心技术；国家自然科学基金项目(60373098，60173006)；国家863高技术研究发展计划项目(2003AA118020)；吉林省科技发展计划重大项目(20020303)；吉林省科技发展计划项目(20030523)。**孙成敏**　博士研究生，讲师，从事机器学习、数据挖掘、粗集的研究；**刘大有**　教授，博士生导师，从事知识工程与专家系统、多Agent系统、不确定性推理、数据挖掘、算法与数据结构、空间推理与GIS应用的研究。

方法[6]。

一般地讲，RS中的不一致出现于条件属性部分不可识别的两个(或多个)对象，而分别属于不同的决策类。而本文主要研究不满足支配原则的不一致问题，这种不一致是RS所无法觉察的，需要使用CORS进行数据分析和规则推导。

我们举例来说明。表1[2]所示是一个决策表，其中Return on investment(return)和Location(location)是条件属性，Bankruptcy risk(risk)是决策属性。

表1 公司破产风险评估举例

Firms	Return on	Location	Bankruptcy risk
Firm A	5	City X	High risk
Firm B	8	City X	Medium risk
Firm C	3	City Y	Medium risk
Firm D	3	City X	Medium risk

使用RS，我们会得出如下的规则集合，显然不存在不一致：

if(return =3)then risk= Medium risk;

if(return=8) and(location=city X) then risk=Medium risk;

if(return = 5) and(location = city X) then risk = High risk;

但是，如果我们使用CORS来分析，显然location是常规属性，return是标准。决策属性risk的两个值High risk和Medium risk将对象分成含有优先顺序的两个决策类，Medium risk "优先于" High risk。实际上，表1的风险评估结果基本上是合情合理的。例如Firm A和Firm B坐落于同一座城市(City X)，Firm B比Firm A有较高的投资回报率，因此破产风险相对较小(Medium risk优先于High risk)。而Firm A和Firm D这两个对象却出现了不一致，两个公司同样坐落于一座城市，Firm A比Firm D有较高的投资回报率，而破产风险却较大。也就是说，条件属性部分，Firm A支配Firm D；而决策属性部分，Firm D支配Firm A。显然不满足支配原则。

从上面的例子可以看出，经典的粗集理论无法推导含有标准的决策表。本文基于标准研究了包含优先顺序信息的粗集方法，问题描述如下：给定一个对象的集合，它们由标准和属性共同描述，这些对象被分在一些决策类中，并且这些决策类是含有预先定义好的顺序的。本文第2部分基于文[2～4]介绍与RS和CORS相关的定义和符号；第3部分和第4部分讲述如何应用CORS进行数据分析和决策规则推导；第5部分介绍了一个例子，用来解释和说明前面的概念和原理；最后是对本文的总结和对未来工作的展望。

2 定义和符号

定义 2.1 称四元组 $T=(U, A, V, f)$是一个信息表，其中U是对象x的非空有限集合(称为论域)；A是属性的非空有限集合；$V=\bigcup_{a\in A} V_a$，其中V_a是属性$a\in A$的取值范围；$f=\{f_a : a\in A\}$是U到V上的映射，其中$f_a : U\to V_a$，若x是对象，$f_a(x)$表示对象x关于属性a的取值，简记为x_a或者$a(x)$。

定义 2.2 $T=(U, A, V, f)$是一个信息表，如果$A=C\cup D$，且$C\neq\Phi, D\neq\Phi, C\cap D=\Phi$，则称$T$是一个决策表，$C$是条件属性集合，$D$是决策属性集合。

在经典粗集理论(RS)的研究中，一般认为属性$a\in A$的值域V_a中元素的关系R_a是恒等关系，即$R_a=\{(v_{a1}, v_{a1}) : v_{a1}\in V_a\}$，此时称$a$为常规属性。在CORS中，存在某些属性$a$，$V_a$的元素间既存在恒等关系，也存在序关系(包含部分序关系、全序关系)，记为$\leqslant_a$，并且a与其它某个(些)属性语义相关，则称这样的属性为标准(criteria)。所谓语义相关，指的是两个标准：在其它属性值保持不变的前提下，如果一个标准其值变化趋向好，不会导致另一个标准的值趋向差。在本文中，如果不做特殊说明，我们将常规属性简称为属性，而含有序关系的属性简称为标准。

让我们举例说明序关系和语义相关的含义。在某个决策表中，如果学生的"数学成绩"、"物理成绩"等具体学科成绩是条件属性，而"总成绩"是决策属性，不妨假设各属性的值域为{好、中、坏}。那么显然，该值域上存在序关系(坏≤中≤好)，即所有这些属性都是标准。这一点是合情合理的：在其它成绩不变的情形下，如果学生"数学成绩"得到提高，那么绝对不会导致总成绩的更差，这样就说明"数学成绩"和"总成绩"是语义相关的。事实上，说两个属性语义相关，也就表明它们符合支配原则。

定义 2.3 决策表$T=(U, A, V, f)$中，如果某个属性$a\in C$，其域值间存在序关系，则称该属性为标准。如果C和D中分别至少含有一个标准，并且C中的所有标准都和D中的标准语义相关，则称该决策表为有序决策表。

在有序决策表中，存在于属性a的值域V_a中的关系R_a既有恒等关系，也有部分序关系(包含全序关系)，但我们认为在知识推导过程中发挥主导作用的是，恒等关系和序关系二者必居其一，且只居其一。另外有序决策表的条件属性集合$C=C_n\cup C_o$，且$C_n\cap C_o=\Phi$，其中C_n和C_o分别表示常规属性和标准的集合。

定义 2.4 $T=(U, A, V, f)$是一个决策表，$a\in A$为一个常规属性，如果对象x和y关于属性a的值x_a和y_a，存在$x_a=_a y_a$，则称x和y是a-不可识别的。$B_n\subseteq A$，且B_n中的每个元素都是常规属性，则称$IND(B_n)$为U上的一个不可识别关系，其中$IND(B_n)=\{(x, y)\in U^2, \forall a\in B_n, x_a=_a y_a\}$。若有$(x, y)\in IND(B_n)$，则称$x$和$y$是$B_n$不可识别的，记为$x\, I_{B_n} y$。

定义 2.5 $T=(U, A, V, f)$是一个决策表，$a\in A$为一标准，如果对象x和y关于属性a的值x_a和y_a，存在$x_a\leqslant_a y_a$，则称y是a-支配x的。$B_o\subseteq A$为一标准集合，则称$DOM(B_o)$为U上的一个支配关系，其中$DOM(B_o)=\{\langle x, y\rangle\in U^2, \forall a\in B_o, x_a\leqslant_a y_a\}$，若有$\langle x, y\rangle\in DOM(B_o)$，则称$y$是$B_o$支配$x$的，简记为$y\, D_{B_o}\, x$。进一步，如果$B_o\subseteq A$是标准的集合，$B_n\subseteq A$是常规属性的集合，不妨设$B=B_o\cup B_n$，则称$DOM(B)=\{\langle x, y\rangle\in U2, \langle x, y\rangle\in DOM(B_o)\ \wedge (x, y)\in IND(B_n)\}$为$U$上的一个完全支配关系。且若有$\langle x, y\rangle\in DOM(B)$，则称$y$是$B$完全支配$x$的，记为$y\, D_B\, x$。当$B_o$或者$B_n$为空集时，是完全支配关系两类特殊的情形。

注意，支配关系是存在方向的，即若$\langle x, y\rangle\in DOM(B_o)$，则$y$是主动的，而$x$是受动的。

定义 2.6 如果$R\subseteq C$是属性集合，$P\subseteq R$是标准集合，则$Q=R-P$是常规属性集合，且$x\in U$是受动的。那么可以定义一个关于属性集合R的支配x的对象$y\in U$的集合，称为R-支配x集合，$D_R^{\geqslant}(x)=\{y\in U : yD_{Px}\wedge xI_{Qy}\}$，同样可以定义一个关于属性集合$R$的被$x$支配的对象$y\in U$的集合，

称为 R-被 x 支配集合，$D_R(x)=\{y\in U: xD_{P_y}\wedge xI_Qy\}$。

定义 2.5 和定义 2.6 都是关于条件属性集合的，下面我们讨论与决策属性相关的定义。本文中假设决策属性集 D 将 U 中的对象划分为有限数目的决策类，记为 $CL=\{Cl_t, t\in T\}$，$T=\{1,2,\cdots,n\}$，且满足 $x\in U$ 属于且仅属于一个决策类 $Cl_t\in CL$，并进一步假设这些决策类是有全序关系的，不妨设 $r<s\in T$，Cl_r 中的对象都是"低于"或"差于" Cl_s 中的对象的。事实上，当前存在这样两种关系：包含于条件属性集的标准导致的完全支配关系和决策类间的全序关系，这两种关系的性质和联系，以及如何利用这两种关系进行粗集数据分析，是本文研究的主要内容。

首先，U 中的对象根据决策属性的不可识别关系，形成若干决策类。由于决策属性中标准的存在，致使这些决策类构成序关系。事实上，根据决策属性集的构成，决策类间可能形成全序关系，可能形成部分序关系，但本文只研究形成全序关系的决策类。因此不妨设决策属性集合 $D=\{d\}$，且 d 是一个全序关系。

定义 2.7 设 $D=\{d\}$，则 d 将 U 划分为有限数目的类 $CL=\{Cl_t, t\in T\}$，$T=\{1,2,\cdots,n\}$。每个 $Cl_t\in CL$ 的向上和向下合并分别归结为：

$$Cl_t^{\geqslant}=\bigcup_{s\geqslant t}Cl_s,\quad Cl_t^{\leqslant}=\bigcup_{s\leqslant t}Cl_s,\quad t=1,2,\cdots,n$$

推论：对于 $t=2,3,\cdots,n$，有 $Cl_{t-1}^{\leqslant}=U-Cl_t^{\geqslant}$ 且 $Cl_t^{\geqslant}=U-Cl_{t-1}^{\leqslant}$。

该推论表明，不属于 Cl_t 或者更好类的所有对象，一定属于 Cl_{t-1} 或者更坏的类。

事实上，如果标准是全序关系，那么也存在向上、向下合并。例如前面例子中的"数学成绩"取值分别为"坏、中、好"，且满足"坏≤中≤好"，那么"坏"的向上合并是{坏、中、好}，"坏"的向下合并是{坏}；同理"中"的向上合并是{中、好}，"中"的向下合并是{中、坏}；"好"的向上合并是{好}，向下合并是{坏、中、好}。为了与决策属性的向上、向下区别开来，我们一般称标准的向上、向下合并为有序上合并和有序下合并。

3 应用 CORS 数据分析

经典 RS 利用条件属性对于决策类的"上近似"和"下近似"来近似知识，这主要依赖于不可识别关系。而在 CORS 中，我们将支配关系融入不可识别关系，使二者共同作用，进行知识的近似。根据前面的定义，我们可以看出，用来进行粗集分析的知识粒子是 $D_P^+(x)$、$D_P^-(x)$、$Cl_t^{\geqslant}$ 和 $Cl_t^{\leqslant}$，其中 $P\subseteq C$，$t\in\{1,2,\cdots,n\}$。在本文中，被近似的知识是决策类及其决策类的向上或者向下合并，即 Cl_t、$Cl_t^{\geqslant}$ 和 $Cl_t^{\leqslant}$，而用来近似知识的粒子是 $D_Q^+(x)$ 和 $D_Q^-(x)$。所推导的分类模式是用 $D_Q^+(x)$ 和 $D_Q^-(x)$ 表示 Cl_t、$Cl_t^{\geqslant}$ 和 $Cl_t^{\leqslant}$ 的函数，而函数的集合就是规则集。

对于 $P\subseteq C$，确定属于 $Cl_t^{\geqslant}$ 和 $Cl_t^{\leqslant}$ 的对象分别构成了 $Cl_t^{\geqslant}$ 和 $Cl_t^{\leqslant}$ 的下近似 $\underline{P}(Cl_t^{\geqslant})$ 和 $\underline{P}(Cl_t^{\leqslant})$，可能属于 $Cl_t^{\geqslant}$ 和 $Cl_t^{\leqslant}$ 的对象分别构成了 $Cl_t^{\geqslant}$ 和 $Cl_t^{\leqslant}$ 的上近似 $\overline{P}(Cl_t^{\geqslant})$ 和 $\overline{P}(Cl_t^{\leqslant})$。那么，如何来确认这些确定的和可能的元素呢？CORS 理论认为，当 x 的 P-支配 x 集合中的每个元素都属于 $Cl_t^{\geqslant}$ 时，则 x 确定属于 $Cl_t^{\geqslant}$；$Cl_t^{\geqslant}$ 中元素 x 的 P-支配 x 集合的并集，形成可能属于 $Cl_t^{\geqslant}$ 的所有元素，即

$$\underline{P}(Cl_t^{\geqslant})=\{x\in U: D_P^+(x)\subseteq Cl_t^{\geqslant}\}$$

$$\overline{P}(Cl_t^{\geqslant})=\bigcup_{x\in Cl_t^{\geqslant}}D_P^+(x)\quad t=1,2,\cdots,n$$

相似地，我们可以定义 $Cl_t^{\leqslant}$ 的上、下近似 $\underline{P}(Cl_t^{\leqslant})$ 和 $\overline{P}(Cl_t^{\leqslant})$：

$$\underline{P}(Cl_t^{\leqslant})=\{x\in U: D_P^-(x)\subseteq Cl_t^{\leqslant}\}$$

$$\overline{P}(Cl_t^{\leqslant})=\bigcup_{x\in Cl_t^{\leqslant}}D_P^-(x)\quad t=1,2,\cdots,n$$

模糊属于 $Cl_t^{\geqslant}$ 和 $Cl_t^{\leqslant}$ 的对象分别构成 $Cl_t^{\geqslant}$ 和 $Cl_t^{\leqslant}$ 的边界，记作 $Bn_P(Cl_t^{\geqslant})$ 和 $Bn_P(Cl_t^{\leqslant})$，用其上下近似表示如下：

$$Bn_P(Cl_t^{\geqslant})=\overline{P}(Cl_t^{\geqslant})-\underline{P}(Cl_t^{\geqslant})$$

$$Bn_P(Cl_t^{\leqslant})=\overline{P}(Cl_t^{\leqslant})-\underline{P}(Cl_t^{\leqslant})$$

粗集理论的另一个重要概念就是分类的近似质量度量 $\lambda_P(Cl)$，该度量的实质是能够被确定规则所捕捉的相对对象数目，其值为确定分类的对象数目和论域所有对象数目的比率。在 CORS 相关的文章[2,5,6]中，$\lambda_P(Cl)$（如式(1)）虽然形式上和经典粗集理论很相像，但都没有准确反映其实质。因此在文[7]中，我们提出一个新的度量公式（式(2)），该公式比较合理地反映了近似质量度量的实质。

$$\lambda_P(Cl)=\frac{|(U-(\bigcup_{t\in T}Bn_P(Cl_t^{\leqslant})))|}{|U|}\tag{1}$$

$$\lambda_P(Cl)=\frac{|\bigcup_{t=1}^{n-1}(\underline{P}(Cl_t^{\leqslant}))\cup\bigcup_{t=2}^{n}(\underline{P}(Cl_t^{\geqslant}))|}{|U|}\tag{2}$$

由于公式(1)的分子是不属于任何边界的对象数目，事实上它夸大了可疑对象的数目。而在推理过程中，因为有些对象可能属于某个边界，但同时也可能属于某个下近似。而一旦某个对象属于某个下近似，那么它就是我们的确定规则集所能涵盖的。但是我们能看到公式(1)在分母中摒弃了这些对象，这导致正确分类对象数目的减少。因此与公式(1)相比较，我们选择公式(2)作为质量近似的度量标准，实际上它也是更趋向于合理。相应的说明可以参见第 5 部分中的例子。

4 基于支配关系的规则推导

从知识发现的角度出发，基于支配原则的关于向上和向下合并的粗近似，能够从数据表中推导出对对象更加泛化的描述。向上或向下合并的 P-下近似代表来自于 $P\subseteq C$ 中标准和常规属性提供的确定知识。对于一个给定的向上合并 $Cl_t^{\geqslant}$（或向下合并 $Cl_t^{\leqslant}$），我们认为属于 $\underline{P}(Cl_t^{\geqslant})$（或 $\underline{P}(Cl_t^{\leqslant})$）的对象是正向而确定的，它们或者属于 Cl_t 类，或者属于好于（或差于）Cl_t 的类，而其它的对象是反向否定的。基于此，我们根据向上和向下合并的下近似，推导出形如"if …，then …"形式的规则。

这里要阐明一个问题，经典的粗集理论 RS 能够发现基于不可识别关系的不一致，而包含优先顺序的 CORS 能帮助我们发现基于支配原则的不一致。简单地说，基于支配关系的不一致存在于标准和决策属性间：当两个对象常规属性集是不可识别的，并存在至少一个标准和决策属性不满足支配原则，而其它标准（若存在）与决策属性都满足支配原则，此时就出现了不一致。一旦出现这种逻辑上的不一致，我们就要试图消除它。最简单的方法是删除不一致的对象，维持数据的一致性。至于采取何种方法更加有效，本文暂不作更多探讨。

在 CORS 中，假设数据存在序关系，我们主要考虑如下两种确定的规则：

确定的 $D_{\leqslant}$-rules：对于那些确定属于 $Cl_t^{\leqslant}$ 的对象，即属于 $\underline{R}(Cl_s^{\leqslant})$ 的对象，提供了如下的描述：

if for $\forall q\in Qx_q=_q r_q\wedge$ for $\forall p\in Px_p\leqslant_p r_p$ then $x\in Cl_t^{\leqslant}$；

确定的 $D_{\geqslant}$-rules：对于那些确定属于 $Cl_t^{\geqslant}$ 的对象，即属于 $\underline{R}(Cl_t^{\geqslant})$ 的对象，提供了如下的描述：

if for $\forall q \in Q x_q =_q r_q \wedge$ for $\forall p \in P x_p \geqslant_p r_p$ then $x \in Cl_t^{\geqslant}$

其中 $R=P \cup Q$，Q 是常规属性集合，P 是标准的集合。

当然，还可以考虑如下的不确定规则：模糊的 $D_{\leqslant}$-rules、模糊的 $D_{\geqslant}$-rules、模糊的 $D_{\geqslant\leqslant}$-rules 等，本文只考虑生成确定规则。

上面的描述只是从形式上给出了基于支配关系的 CORS 规则。至于从具体的数据表中如何生成简洁而完备的规则集，需要我们提供合理的算法，本文对如何生成确定的完备规则集进行了讨论。

由于本文中假设决策属性集 D 将 U 中的对象划分为有限数目的决策类，即 $CL=\{Cl_t, t \in T\}$，$T=\{1,2,\cdots,n\}$，显然确定的 $D_{\geqslant}$-rules 不应该包含 $x \in Cl_n^{\leqslant}$ 的规则描述；同理，确定的 $D_{\geqslant}$-rules 不应该包含 $x \in Cl_t^{\geqslant}$ 的规则描述，因为所有的对象都是属于 $Cl_n^{\leqslant}$，且属于 $Cl_t^{\geqslant}$ 的，所以这样的规则是平凡的，没有任何指导意义。因此我们只需要推导出后件为 $x \in Cl_t^{\leqslant}$，$t=1,2,\cdots,n-1$ 和 $x \in Cl_t^{\geqslant}$，$t=2,3,\cdots,n$ 的规则即可。

同样地，CORS 中规则的推导主要包含简式约简和值约简。如何进行值约简，从而发现数据中基于支配关系的不一致，并挖掘出基于支配关系的知识，是我们最关心的问题。假设 $R \subseteq C$ 是决策表的一个简式，且 $R=P \cup Q$，Q 是常规属性集合，P 是标准的集合，那么如何产生规则，得到完备而简约的规则集呢？下面我们给出根据等价类和有序合并来近似决策类及向上、向下合并，进而推导规则的四条原则：

原则 4.1 常规属性集 Q 导致的不可识别关系中若有某个等价类包含于某决策类（包括向上或向下合并），则生成相应规则。

原则 4.2 条件属性集中某标准集 P 导致的有序合并，如果包含于某决策类（包括向上或向下合并），生成相应的规则。

原则 4.3 条件属性集中某个（些）常规属性导致的等价类与标准导致的有序合并的交集，包含于某决策类（包括向上向下合并），则生成相应规则。

原则 4.4 如果某个等价类、有序合并或者它们的交集，基于上述三条原则形成规则，则其不再参加结论部分同方向的规则推导。

关于对于上述原则的说明：原则 4.1 表明，若某些常规属性集 Q 及取值所确定的某个等价类包含于某个决策类，则可以推导出形如“if $\forall q \in Q x_q =_q r_q$ then $x \in Cl_t^{\leqslant}$ (or $Cl_t^{\geqslant}$)”的规则；而原则 4.2 表明，若标准的等价类或者其有序合并包含于某决策类，则可以推导出形如 if for $\forall p \in P x_p \leqslant_p (\text{or} \geqslant_p) r_p$ then $x \in Cl_t^{\leqslant}$ (or $Cl_t^{\geqslant}$) 的规则。显然，这两类规则是上面介绍的确定规则的特例。容易看到原则 4.3 根据常规属性集 Q 及标准集 P 推导规则，推导出的规则形如 if for $\forall p \in P x_p \leqslant_p (\text{or} \geqslant_p) r_p \wedge$ for $\forall q \in Q x_q = r_q$ then $x \in Cl_t^{\leqslant}$ (or $Cl_{\geqslant t}$)。原则 4.4 说明，如果某知识粒子或者知识粒子的交集包含于某决策知识粒子，它不需要再参加相同方向的规则推导。至于相同方向的规则，我们指的是那些其后件同为 $Cl_t^{\leqslant}$（或 $Cl_t^{\geqslant}$）的规则。原则 4.4 保证我们无需做冗余的工作。

上面，我们给出了规则推导的原则。至于依据标准集 P 和常规属性集 Q 如何获得等价类及有序合并，并由此应用上述原则推导规则，我们将在以后的研究中给出相应的算法。

5 一个例子

下面的例子来源于文[2]，经过修改后来说明前面所涉及到的定义和符号，以及显示原始 RS 与当前扩展模型的主要区别。在表 2 中，记录了使用如下三个条件属性 $C=\{a, b, c\}$ 和一个决策属性 d 描述的 9 个大型超市：a，销售员工的能力；b，地理位置；c，经营类型；d，赢利情况（赢利、持平和亏损）。其中将 d 看作决策属性，a 看作标准，b 和 c 看作常规属性。

表 2 超市信息表

Warehouse	*a*	*b*	*c*	*d*
w1	*Medium*	*A*	*X*	*Loss*
w2	*Good*	*A*	*X*	*Profit*
w3	*Bad*	*A*	*X*	*Loss*
w4	*Good*	*B*	*X*	*Balance*
w5	*Good*	*A*	*Y*	*Profit*
w6	*Medium*	*B*	*Y*	*Balance*
w7	*Medium*	*B*	*X*	*Profit*
w8	*Bad*	*B*	*Y*	*Loss*
w9	*Medium*	*A*	*Y*	*Balance*

经过计算得到，该决策表的简式只有一个，即 $Rec_{CL}=Core_{CL}=C=\{a, b, c\}$。根据常规属性 b 得到的等价类是 $\{\{1,2,3,5,9\},\{4,6,7,8\}\}$，根据常规属性 c 得到的等价类是 $\{\{1,2,3,4,7\},\{5,6,8,9\}\}$，属性集 $\{b, c\}$ 得到的等价类是 $\{\{1,2,3\},\{4,7\},\{5, 9\},\{6,8\}\}$。根据标准 a 的支配关系得到的有序类是 $\{\{3,8\} \leqslant \{1,6,7,9\} \leqslant \{2,4,5\}\}$，其有序下合并为 $\{\{3,8\} \leqslant \{1,3,6,7,8,9\} \leqslant \{1,2,3,4,5,6,7,8,9\}\}$，有序上合并为 $\{\{2,4,5\} \geqslant \{1,2,4,5,6,7,9\} \geqslant \{1,2,3,4,5,6,7,8,9\}\}$。很容易看到，对象 4 和对象 7 由于在属性集 $\{b, c\}$ 上不可识别，且标准 $\{a\}$ 和决策属性 $\{d\}$ 间不满足支配关系，因此对象 4 和对象 7 是不一致的。为简单起见，我们通过消除对象 4 和对象 7 来处理不一致问题。含序决策属性将所有对象划分为 $Cl_1=\{1,3,8\}$，$Cl_2=\{4,6,9\}$，$Cl_3=\{2,5,7\}$，其中 Cl_3 好于 Cl_2，Cl_2 好于 Cl_1，所以 $Cl_1^{\leqslant}=\{1,3,8\}$，$Cl_2^{\leqslant}=\{1,3,4,6,8,9\}$，$Cl_3^{\geqslant}=\{2,5,7\}$，$Cl_2^{\geqslant}=\{2,4,5,6,7,9\}$。

首先推导“$D_{\leqslant}$-rules”。根据原则 4.1 得到规则(1)：

(1)if $x_b=B \wedge x_c=Y$ then $x_d \leqslant$Balance (涵盖 6,8)

根据原则 4.2 得到规则(2)：

(2)if $x_a \leqslant Bad$ then $x_d \leqslant$Loss (涵盖 3,8)

进一步推导会得到如下规则：

(3)if $x_a \leqslant Medium \wedge x_b=A$ then $x_d \leqslant$Balance (涵盖 1,3,9)

(4)if $x_a \leqslant Medium \wedge x_c=Y$ then $x_d \leqslant$Balance (涵盖 6,8,9)

我们继续推导 $D_{\geqslant}$-rules”。根据原则 4.1 得到规则(5)：

(5)if $x_b=A \wedge x_c=Y$ then $x_d \geqslant$Balance(涵盖 5,9)

根据原则 4.2 得到规则(6)：

(6)if $x_a \geqslant Good$ then $x_d \geqslant$Profit (涵盖 2,5)

同样地，进一步推导会得到如下规则：

(7)if $x_a \geqslant Medium \wedge x_b=B$ then $x_d \geqslant$Balance (涵盖 6)

(8)if $x_a \geqslant Medium \wedge x_c=Y$ then $x_d \geqslant$Balance (涵盖 5,6,9)

根据上面的例子，我们看到原则 4.4 保证我们避免了许

多冗余的工作。例如规则(2),知识粒子{3,8},即$a \leqslant bad$,满足原则4.2,因此我们不将推导其前件条件强于$a \leqslant bad$、后件形如$Cl^{\leqslant}$的规则,但是我们可以继续推导后件形如$Cl^{\geqslant}$的规则。

通过上面的规则推导,我们很容易看到规则集合覆盖了除对象4和7以外的所有对象,并且在某种程度上很简化。当然,也可能存在某些对象被几条规则覆盖。例如,对象9被规则(3)和规则(5)覆盖,但它们的含义是不同的。而且标准a与决策属性d语义相关,规则也准确地反映了这一点。至于如何进一步简化这些规则,从而基于要求提炼出更加精炼和完备的规则集,本文不做深入研究。

除此而外,根据计算近似质量的公式,让我们来计算上述例子的$\lambda_P(Cl)$,此处$P=\{a\}$。我们来比较前面章节中提到的公式哪个更合理。

表3 Cl_t的下近似和上近似

Cl_t	$\underline{P}(Cl)_t$	$\overline{P}(Cl)_t$	$Bnp(Cl_t)$
$Cl_1^{\leqslant}=\{1,3,8\}$	{3,8}	{1,3,6,8,9}	{1,6,9}
$Cl_2^{\leqslant}=\{1,3,6,8,9\}$	{1,3,6,8,9}	{1,3,6,8,9}	Φ
$Cl_2^{\geqslant}=\{2,5,6,9\}$	{2,5}	{1,2,5,6,9}	{1,6,9}
$Cl_3^{\geqslant}=\{2,5\}$	{2,5}	{2,5}	Φ

根据公式(1),我们有

$$\lambda_P(Cl)=\frac{|(U-(\bigcup_{t\in T}Bn_P(Cl_t^{\leqslant})))|}{|U|}=\frac{|\{1,2,3,5,6,8,9\}-\{1,6,9\}|}{9}=\frac{4}{9}$$

根据公式(2),我们有

$$\lambda_P(Cl)=\frac{|\bigcup_{t=1}^{n-1}(\underline{P}(Cl_t^{\leqslant}))\cup\bigcup_{t=2}^{n}(\underline{P}(Cl_t^{\geqslant}))|}{|U|}=\frac{|\{1,3,6,8,9\}\cup\{2,5\}|}{9}=\frac{7}{9}$$

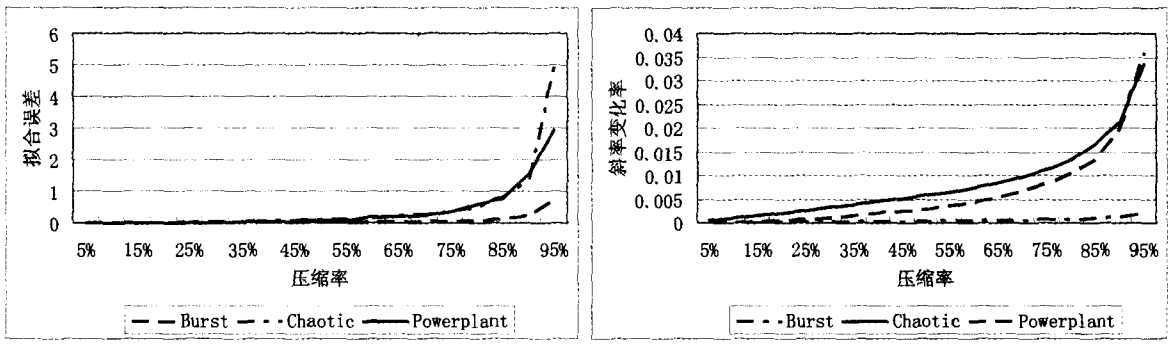
通过表3,我们能看到公式(1)夸大了可疑对象的数目。显然,$Bn_{\{a\}}(Cl_1^{\leqslant})=\{1,6,9\}$,但是$\{1,6,9\}\subseteq\underline{P}(Cl_2^{\leqslant})$,因此这些对象被确定规则所捕捉,而公式(2)保证对象集{1,6,9}不再是可疑对象。

结论 本文基于经典粗集理论,研究探讨了含序信息粗集方法。本文定义了标准、向上合并、向下合并、有序决策表等概念,并通过标准和属性共同刻画决策类的有序合并。本文提出了根据得到的知识粒子形成规则的四条原则,最后通过例子阐述上述概念和思想。

本文只是初步研究了含序粗集方法,该方法的许多方面和细节还在进一步研究和讨论中,未来需要做的工作还有许多,如:

(下转第163页)

(上接第142页)



(a) 在不同压缩率下的拟合误差变化情况　　(b) 在不同压缩率下的斜率变化率变化情况

图5 在不同压缩率下的拟合误差和斜率变化率的变化情况

时间序列的SEEP表示简单直观,具有很强的数据压缩能力和一定的除噪能力,能突出时间序列的模式变化特征。

在来自不同领域的时间序列数据集上的实验表明:对于斜率变化范围比较集中时间序列,SEEP表示方法有着非常好的效果,与以往的分段线性表示方法相比,SEEP表示方法与原始时间序列之间的拟合误差更小,而且要小很多,并且当压缩率小于75%的时候拟合误差的增长极其缓慢;对于斜率变化范围比较大的时间序列,SEEP表示方法与原始时间序列之间的拟合误差,和以往的分段线性表示方法相比,也相差不大。而且SEEP表示方法计算简单,易于实现,并且由于算法的时间复杂度仅为$O(n)$,因此即使在数据量比较大的时间序列中仍然具有很高的执行效率。

致谢 感谢Keogh等人提供的来自于不同领域的实验数据集,使本文可以顺利完成。

参考文献

1 Keogh E. Fast similarity search in the presence of longitudinal scaling in time series databases [C]. In: Proceedings of the IEEE 9th International Conference on Tools with Artificial Intelligence, Washington: IEEE Computer Society, 1997. 578～584

2 Keogh E, Folias T. The UCR Time Series Data Mining Archive [EB/OL]. http://www.cs.ucr.edu/～eamonn/TSDMA/index.html. Irvine, CA: University of California, Department of Information and Computer Science, 2002

3 Prat K B, Fink E. Search for patterns in compressed time series [J]. International Journal of Image and Graphics, 2002, 2(1): 89～106

4 Keogh E J, Chakrabarti K, Pazzani M J. Sharad Mehrotra. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases [J]. Knowl. Inf. Syst, 2001, 3(3): 263～286

5 Yi B K, Faloustsos C. Fast Time Sequence Indexing for Arbitrary Lp Norms [C]. In: Proceedings of the 26th International Conference on Very Large Data Bases, San Francisco: Morgan Kaufmann Publishers Inc, 2000. 385～394

6 Xiao Hui, Feng Xiao-Fei, Hu Yun-Fu. A new segmented time warping distance for data mining in time series database [C]. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics, Shanghai, China, 2004. 1277～1281

7 肖辉. 时间序列的相似性查询与异常检测: [博士论文]. 上海: 复旦大学, 2005

出了一种改进的变精度粗糙集模型(简称 $VPRS_{l,u}$ 模型)。

$VPRS_{l,u}$ 模型中提供了两个人为的可变参数:l 和 u,满足条件 $0\leqslant l<P(X)<u\leqslant 1$。$P(X)$ 是以论域 U 为样本空间,集合 X 为样本点的概率表示。相关定义如下:

设等价关系 R 在论域 U 上的等价类为 $\tilde{R}=\{E_1,E_2,\cdots,E_k\}$。对于 $X\subset U$:

u-正域为

$$POS_u(X)=\bigcup\{E\in\tilde{R}:P(X|E)\geqslant u\}$$

l-负域为

$$NEG_l(X)=\bigcup\{E\in\tilde{R}:P(X|E)\leqslant l\}$$

(l,u)-边界为

$$BND_{l,u}(X)=\bigcup\{E\in\tilde{R}:l<P(X|E)<u\}$$

其中,$P(X|E)$ 为条件概率。

下面我们把 $VPRS_\beta$ 和 $VPRS_{l,u}$ 两个模型做一下比较。

由于 $C(X,Y)=1-card(X\cap Y)/card(X)$ 可以得到

$C(E,X)=1-card(E\cap X)/card(E)=1-P(X|E)$,

因此,在 $VPRS_\beta$ 模型下,β-正域、负域和边界可写为:

$$\underline{R}_\beta(X)=\bigcup\{E\in\tilde{R}:P(X|E)\geqslant 1-\beta\}$$

$$NEGR_\beta(X)=\bigcup\{E\in\tilde{R}:P(X|E)\leqslant\beta\}$$

$$BNR_\beta(X)=\bigcup\{E\in\tilde{R}:\beta<P(X|E)<1-\beta\}$$

β-上近似定义可写为:

$$\overline{R}_\beta(X)=\bigcup\{E\in\tilde{R}:P(X|E)>\beta\}$$

在 $VPRS_{l,u}$ 模型下,可以得到对于 $X\subset U$,上近似的表示:

$$POS_l(X)=\bigcup\{E\in\tilde{R}:P(X|E)>l\}$$

由此可见 $VPRS_{l,u}$ 模型是把 $VPRS_\beta$ 模型中的参数 β 和 $1-\beta$ 分别用两个参数 l 和 u 代替的一种改进。改进后的模型使得粗糙集边界更有弹性。

3 关于 Bayesian 粗糙集模型

虽然 $VPRS_{l,u}$ 模型具有很大程度的灵活性,但其中的参数是人为的,为了克服这一缺陷 Slezk 和 Ziarko 最近提出了一种新的模型:Bayesian 粗糙集模型[4]。

Bayesian 粗糙集模型基于贝叶斯(Bayesian)推理的思想定义出关于集合 X 的正域、负域和边界:

$$POS^*(X)=\bigcup\{E\in\tilde{R}:P(X|E)>P(X)\}$$

$$NEG^*(X)=\bigcup\{E\in\tilde{R}:P(X|E)<P(X)\}$$

$$BND^*(X)=\bigcup\{E\in\tilde{R}:P(X|E)=P(X)\}$$

但是在 Bayesian 粗糙集模型的定义中我们看到一些不足:

(1)当 $P(X)$ 很小时,满足 $P(X|E)>P(X)$ 的所有等价类 E 都成为了正域的集合,这就意味着正域中有一些等价类含有很少的 X 中的元素。这不符合定义正域的思想。

(2)当 $P(X)$ 很大时,使用负域的定义会损失大量的信息。这也不符合改进传统粗糙集模型以获得更多信息的思想。

(3)对于边界的定义,Slezk 和 Ziarko[4] 认为边界区域完全与 X 无关,并且从概率事件的角度认为 X 与其它等价类 (E) 事件是相互独立的。事实上这样的认为是不合理的。这是因为边界区域中存在着与 X 有关的信息,而信息量的比例与等价类有关。在该模型中,边界的定义存在的另一个问题是满足 $P(X|E)=P(X)$ 条件的等价类在实际计算中很少,因此在很多情况下,边界区域是空集。这不符合改进传统粗糙集模型更柔性化的思想。

针对以上 3 点不足我们做如下改进尝试:

定义 1 设等价关系 R 在论域 U 上的等价类为 $\tilde{R}=\{E_1,E_2,\cdots,E_k\}$。$X$ 为论域 U 上的集合,令 $K(X)=\max\{P(X),1-P(X)\}$,关于集合 X 的正域、负域和边界定义为:

$$POS(X)=\bigcup\{E\in\tilde{R}:P(X|E)>K(X)\}$$

$$NEG(X)=\bigcup\{E\in\tilde{R}:P(X|E)<1-K(X)\}$$

$$BND(X)=\bigcup\{E\in\tilde{R}:1-K(X)<P(X|E)<K(X)\}$$

结论 本文讨论了 Bayesian 粗糙集模型以及相关变精度模型。指出了 Bayesian 粗糙集模型存在的不足,提出了改进方法。对 Bayesian 粗糙集模型的进一步研究和应用具有重要性。

参 考 文 献

1 Pawlak Z. Rough set. International Journal of Computer and Information Science, 1982, 11(5): 341~356
2 Ziarko W. Variable precision rough set model. Journal of Computer and System Science, 1993, 46(1): 39~59
3 Katzberg JD, Ziarko W. Variable precision extension of rough sets. Fundamenta Informaticae, 1996, 27: 155~168
4 Slezk D, Ziarko W. The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning, 2005, 40: 81~89
5 曾黄麟. 粗集理论及其应用(修订版). 重庆:重庆大学出版社, 1998

(上接第 161 页)

当条件属性中包含的标准多于一个的时候,多个标准会使条件属性部分的知识粒子更加复杂,如何更有效地推得决策规则;

根据条件属性和决策属性部分的知识粒子推导规则的算法并不容易给出,本文根据提出的原则形成规则,是否存在能得到完备规则集且时间效率较高的算法;

含序粗集方法中,如何有效地获得决策表的简式以及核;

当存在基于支配原则不一致时,如何消除之及进行化简;

如何从决策表中推导得出标准的有序信息,而不是由决策者作为先验知识给出等。

参 考 文 献

1 Pawlak Z. Rough sets: theoretical aspects of reasoning about data. In: System Theory, Knowledge Engineering and Problem Solving, Kluwer Academic Publishers, Dordrecht, 1991, 9
2 Greco S, Matarazzo B, Slowinski R. Rough sets methodology for sorting problems in presence of multiple attributes and criteria. European Journal of Operational Research, 2002, 138: 247~259
3 Greco S, Matarazzo B, Slowinski R. Rough approximation of a preference relation by dominance relations. European Journal of Operational Research, 1999, 117: 63~83
4 Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis. European Journal of Operational Research, 2001, 129: 1~47
5 Slowinski R, Greco S, Matarazzo B. Rough Set Analysis of Preference-Ordered Data. Rough Sets and Current Trends in Computing, 2002. 44~59
6 Gediga G, Dumtsch I. Approximation quality for sorting rules. Computational Statistics & Data Analysis, 2002, 40
7 Sun Chengmin, Liu Dayou, Sun Shuyang. Containing Order Rough Set Methodology. In: Proceedings of 2005 International Conference on Machine Leaning and Cybernetics, vol 3 of 9, Guangzhou, China, 2005. 1722~1727