

文章编号:1005—0523(2007)02—0062—04

基于变精度 Rough 集的对象甄选优化模型

邓毅雄, 黄兆华

(华东交通大学 信息工程学院, 江西 南昌 330013)

摘要:定义了对对象甄选信息系统,并根据属性值序列间的距离,给出了对象的相似关系,并利用 Rough 集理论讨论了对象的分类问题,最后给出了甄选的优化模型.

关键词:对象;信息系统;Rough 集;变精度
中图分类号:TP311 **文献标识码:**A

1 引言

在现实生活中,我们经常面对将对象进行分类的问题,分类是数据挖掘的重要任务.分类的目的是构造一个分类函数或分类模型(分类器),分类模型能把数据库中的数据项映射到某一个确定类别.分类的过程就是将区别不太大的对象分为一类,它们的关系是不分明的等价关系、相似关系或覆盖关系.在一个信息系统中,如果给定参考的属性值,对论域中的对象根据属性值的进行分类,以期以参考属性值要求尽可能匹配,这样的问题实际上就是现实生活中的对象甄选或选拔.如在国家每年一次的公务员选拔中,往往需要在经过测试得到了候选对象的基本素质表的情况下,根据不同岗位对公务员素质的要求,从候选对象中进行甄选,为每个岗位匹配一定数量的公务员.

一个信息系统是一个四元组 $S=(U, A, V, f)$, 其中 $U=\{u_1, u_2, \dots, u_n\}$ 是非空、有限集合,称为论域,论域 U 中的元素称为对象, A 是非空、有限的属性集合,设 $V=\bigcup_{a\in A} V_a$, 其中对 $a\in A, f: a\rightarrow V_a, V_a$ 称为属性 a 的值集, V 称为属性集 A 的值区域.

设 $D=(T, A)$, 其中非空、有限集 $T=\{T_1, T_2, \dots, T_k\}$ 称为类别集, $T_i(i=1, 2, \dots, k)$ 称为类别, A 仍

然是属性集.

我们将三元组 $\langle S, D, \phi \rangle$ 称为对象甄选信息系统,其中 ϕ 是映射 $S\rightarrow D$. 表示 S 与 D 之间的内在关系.此系统是按照类别的属性值要求,从论域中甄选若干个对象分别归于类别 $T_i(i=1, 2, \dots, k)$ 中,并且每个类别中至少有一个对象归属其中,这里 ϕ 表示这种归属关系,且 $k\leq n$.

我们将 D 对应的决策表称为类别要求决策表(如:表 1),将 S 对应的决策表称为对象水平决策表(如:表 2).基于表 1—2 给出的对象甄选信息系统,就是要根据类别和对象的属性值,甄选一定的对象归属于各类别中.

在一般情况下,对象甄选信息系统是不分明的,严格满足类别要求的对象可能是不存在的,所以我们总是期望对象的属性值能比较接近类别要求属性值就可以了.我们注意到, Rough 集理论正是目前处理不分明关系的较好方法.

表 1 类别要求决策表

属性 类别				
	a	b	c	d
T_1	3	4	2	4
T_2	4	3	3	2
T_3	2	2	4	4
T_4	2	3	3	4

收稿日期:2007—01—17

基金项目:江西省自然科学基金(0411035)资助

作者简介:邓毅雄(1963—),男,江西新干人,华东交通大学教授,研究方向为图论与组合优化、数据挖掘、Rough 集及其应用.

表 2 对象水平决策表

属性 对象	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>
<i>u</i> ₁	4	4	3	3
<i>u</i> ₂	4	3	4	2
<i>u</i> ₃	3	4	1	2
<i>u</i> ₄	4	3	3	3
<i>u</i> ₅	3	4	3	2
<i>u</i> ₆	3	1	4	3
<i>u</i> ₇	4	3	2	3
<i>u</i> ₈	3	4	4	2
<i>u</i> ₉	3	3	4	3
<i>u</i> ₁₀	1	3	4	2
<i>u</i> ₁₁	1	2	3	4
<i>u</i> ₁₂	4	3	2	4

2 Rough 集理论

Rough 集是 20 世纪 80 年代由 Z·Pawlak 提出的,其主要兴趣在于它恰好反映了人们用 Rough 集方法处理不分明问题的常规性,即以不完全信息和知识去处理一些不分明现象的能力,或依据观察、度量到的某些不精确的结果而进行分类数据的能力,由此,它得到了广泛的应用.

设 U 是论域, $X \subseteq U$, R 是 U 上的一个等价关系, (U, R) 是一个近似空间,在 (U, R) 上,如果 X 是一些 R -基本类的并集,则称 X 是 R -可定义的;否则称 X 是 R -不可定义的. R -可定义集称为 R -一致集或 R -恰当集; R -不可定义集称为 R -不一致集或 R -Rough 集(简称不一致集或 Rough 集).

如果存在一个等价关系 $R \in IND(U)$,其中 $IND(U)$ 是 U 上给定的所有等价关系的交集,使得 $X \subseteq U$ 是 R -一致的,则集合 X 被称为 U 中一致集;如果 $X \subseteq U$ 对任意 $R \in IND(U)$ 都是 R -Rough 的,则称 X 为 U 上不一致集或 Rough 集.

X 关于 R 的下近似 $R^*(X)$ 和上近似 $R^+(X)$ 分别定义:

$$R^*(X) = \bigcup \{Y \mid Y \in U/R, Y \subseteq X\}$$
$$R^+(X) = \bigcup \{Y \mid Y \in U/R, Y \cap X \neq \emptyset\}$$

总一个集合 X 不能利用有效的等价关系被恰当地分类时,则可通过集合 $R^*(X)$ 和 $R^+(X)$ 来近似这个集合.

设 R 是一个等价关系族, $r \in R$,如果 $IND(R) = IND(R - \{r\})$,则称 r 在 R 中是可被约去的知识;如果 $R - \{r\}$ 是独立的,则 P 是 R 中一个约简.属性约简的方法有数据分析法和分明矩阵法,其中

分明矩阵法是 1991 年由 A·Skowron 提出的.

3 属性值序列的距离

设属性子集 $B = \{a_1, a_2, \dots, a_m\} \subseteq A$,对任意一个对象(或类别) $w \in U$ (或 $w \in T$)(这里用 w 表示对象本身也表示属性值序列)其关于 B 的属性值按一定顺序构成一个序列 $w = (v_1^{(w)}, v_2^{(w)}, \dots, v_m^{(w)})$,其中 $v_j^{(w)}$ 为 w 对应属性 a_j 的属性值.

考虑不同对象的属性值序列之间的所谓距离,目的是用数量关系来衡量对象之间的差异程度.当属性及其属性值具有定性的特征,我们只要考虑对应属性值之间的异同性即可;当属性及其属性值具有定量的特征,我们通常考虑对应属性值之间的数量距离(差的绝对值).属性的特征不同,衡量标准也可能有所不同,下面分别给出两种属性值序列的距离:

定义 1 设两个属性值序列分别为 $x = (v_1^{(x)}, v_2^{(x)}, \dots, v_m^{(x)})$, $y = (v_1^{(y)}, v_2^{(y)}, \dots, v_m^{(y)})$,令

$$v_j^{(x)} + v_j^{(y)} = \begin{cases} 0, & v_j^{(x)} = v_j^{(y)} \\ 1, & v_j^{(x)} \neq v_j^{(y)}, j = 1, 2, \dots, m, \end{cases}$$

称
$$d_B(x, y) = \sum_{j=1}^m (v_j^{(x)} + v_j^{(y)})$$
 为 x 与 y 的汉明距离,相似程度 $s_B(x, y) = m - d_B(x, y)$.

定义 2 设两个属性值序列分别为 $x = (v_1^{(x)}, v_2^{(x)}, \dots, v_m^{(x)})$, $y = (v_1^{(y)}, v_2^{(y)}, \dots, v_m^{(y)})$,称

$$d_B(x, y) = \left[\sum_{j=1}^m (v_j^{(x)} + v_j^{(y)})^p \right]^{1/p}$$

为 x 与 y 的 p -距离,当 $p = 2$ 时,就是我们常见的欧氏距离.

这两种距离均表示两个属性值序列之间在属性值方面的总体偏差,用数量指标反映了对象间的差异,刻划了在一定的属性情况下,属性值序列之间在性质上的偏差,也就是刻划了对象之间的相似程度.如果两个属性值序列的距离小,说明这两个属性值更接近,以更大的可能归属于同一类.但它们在具体应用中根据问题的对象及其属性的特征的不同,有选择地使用.如果问题的属性反应的是某种性质的比较,那么使用汉明距离较合适;如果问题的属性反应的是某种数量的大小,那么使用欧氏距离更准确.

我们也注意到,在一个信息系统中,可能这两种属性类型均存在,有的属性是定性分析的,有的是定量分析的,那么也可以是它们的综合.

为讨论方便,在下面的对象甄选信息系统中,我们只讨论属性值表示性质的情形,即用汉明距离刻画.如果 $x \in U, y \in T$,那么 $d_B(x, y)$ 反映了 x 与 y 之间的接近程度.当此距离值较小时,说明对象 x 更应该归属于类别 y 中.下面我们再给出 x, y 的平均距离 $\bar{d}_B(x, y)$ 与平均相似程度 $\bar{s}_B(x, y)$:

$$\begin{aligned}\bar{d}_B(x, y) &= \frac{d_B(x, y)}{m}, \\ \bar{s}_B(x, y) &= \frac{s_B(x, y)}{m} = 1 - \frac{d_B(x, y)}{m}.\end{aligned}$$

显然 $0 \leq \bar{d}_B(x, y) \leq 1, 0 \leq \bar{s}_B(x, y) \leq 1$.

4 基于变精度的 Rough 集模型

定义 3 对于对象甄选信息系统 $S, D, \phi, x \in U, B \subseteq A, \alpha$ 是一实数,定义 x 的 $B-\alpha$ (邻域:
 $S_B^\alpha(X) = \{y \mid \bar{d}_B(x, y) \leq \alpha, y \in U, 0 \leq \alpha \leq 1\}$.
它表示在 B 下与对象 x 的平均距离小于等于 α 的所有对象的集合.

设 $T_j \in T$,类似定义 T_j 的 $B-\alpha$ 邻域:
 $S_B^\alpha(T_i) = \{T_i \mid \bar{d}_B(T_j, T_i) \leq \alpha, T_i \in T, 0 \leq \alpha \leq 1\}$.
事实上,由上面的定义,有
 $S_B^\alpha(X) = \{y \mid \bar{s}_B(x, y) \geq 1 - \alpha, y \in U, 0 \leq \alpha \leq 1\}$.
 $S_B^\alpha(T_i) = \{T_i \mid \bar{s}_B(T_j, T_i) \geq 1 - \alpha, T_i \in T, 0 \leq \alpha \leq 1\}$.
当 $B=A$ 时, $S_A^\alpha(x)$ 和 $S_A^\alpha(T_j)$ 简记为 $S^\alpha(x)$ 和 $S^\alpha(T_j)$.

定义 4 设 $T_j \in T, 0 \leq \beta \leq 1, T_j$ 在 U 中的甄选下、上近似分别定义为:

$$\begin{aligned}R_B^\beta(T_j) &= \bigcup \{x \mid \forall y \in S_B^\alpha(x), \bar{d}_B(y, T_j) \leq \beta\}, \\ \bar{R}_B^\beta(T_j) &= \bigcup \{x \mid \exists y \in S_B^\alpha(x), \bar{d}_B(y, T_j) \leq \beta\}.\end{aligned}$$

将 R_A^β 和 $\bar{R}_A^\beta$ 分别简记为 $R^\beta(T_j)$ 和 $\bar{R}^\beta(T_j)$.
对应于 T_j 的 Rough 集为 $(R_B^\beta(T_j), \bar{R}_B^\beta(T_j))$, 用此上、下近似可以确定 T_j 的甄选范围.

性质 1 $R_B^\beta(T_j) \subseteq \bar{R}_B^\beta(T_j)$.
性质 2 设 $0 \leq \beta_1 \leq \beta_2 \leq 1, T_j \in T$, 有 $R_B^{\beta_1}(T_j) \supseteq R_B^{\beta_2}(T_j), \bar{R}_B^{\beta_1}(T_j) \supseteq \bar{R}_B^{\beta_2}(T_j)$.

设 $F_j = S_B^\alpha(T_j) (j=1, 2, \dots, k)$, 下面进一步将定义 3 推广:

定义 5 对 $F_j (j=1, 2, \dots, k), 0 \leq \beta \leq 1, F_i$ 在 U 中的甄选下、上近似分别定义为:

$$\begin{aligned}R_B^\beta(F_i) &= \bigcup \{E_i \mid \forall y \in E_i, \forall T_j \in F_i, \bar{d}(y, T_j) \leq \beta\}, \\ \bar{R}_B^\beta(F_i) &= \bigcup \{E_i \mid \exists y \in E_i, \forall T_j \in F_i, \bar{d}(y, T_j) \leq \beta\}.\end{aligned}$$

性质 3 $R_B^\beta(F_i \cap F_j) = R_B^\beta(F_i) \cap R_B^\beta(F_j), \bar{R}_B^\beta(F_i \cap F_j) = \bar{R}_B^\beta(F_i) \cap \bar{R}_B^\beta(F_j)$

5 优化模型

设 T_j 同时表示第 j 类别所甄选的对象集合,所要甄选的对象总数为 $N (N \leq n), N_j$ 为第 j 类别的甄选对象数,即 $N_j = |T_j| (j=1, 2, \dots, k)$, 且 $N_1 + N_2 + \dots + N_k = N$.

对象甄选的目的是使相似程度大的对象归属于相应的类,并使得总的相似程度达到最大,即距离最小,所以其优化模型为:

$$\begin{aligned}V^- &= \min \sum_{j=1}^k \sum_{x \in R^\alpha} \bar{d}(x, T_j), \\ \text{或 } V^- &= \min \sum_{j=1}^k \sum_{x \in \bar{R}^\alpha} \bar{d}(x, T_j).\end{aligned}$$

在对象甄选过程中,首先给定精度 α , 确定 Rough 集,将对象进行分类,并在此分类的基础上进行甄选,使得总体相似程度最优.如果不能得到满意的结果,可对精度 α 进行调整.

参考文献:

[1] 刘清. Rough 集及 Rough 推理[M]. 北京: 科学出版社, 2001.
[2] 邵峰晶, 于忠清. 数据挖掘: 原理与算法[M]. 北京: 中国水利水电出版社, 2003.
[3] H. Tanaka, Dual Mathematical Models Based on Rough Approximations in Data Analysis, Rough Sets, Fuzzy Sets[J]. Data mining and Granular Computing, 2003, (5): 52~59.
[4] Tatsuo Nishino, etl Mitsuo Nagamachi, and Hideo Tanaka. Variable Precision Bayesian Rough Set Model and Its Application to Human Evaluation Data[J]. 10th International Conference, SFDGrC 2005, 294~303.
[5] Ziarko, W. Variable precision rough set model[J]. Journal of Computer and System Sciences, 1993, (46): 39~59.
[6] Stepaniuk, J. Knowledge Discovery by Application of Rough Set Models[J]. L. Polkowski, S. Tsumoto and T. Y. Lin, Rough Set Methods and Applications, Physica-Verlag (2000): 137~233.

The Optimized Model of Selecting Object Information Systems Based on Variable Precision Rough Sets

DENG Yi-xiong, HUANG Zhao-hua

(School of Information Engineering, East China Jiaotong Univ. Nanchang 330013, China)

Abstract: This paper defines the selecting object information system and the similitude relation of objects according to the distance of attribute value sequence, discusses the classify of objects using variable precision rough sets theory and finally gives the optimized model of selecting object.

Key words: object; information system; rough set; variable precision

(上接第 36 页)

Study on the Disposal Measures of Karst Subgrade in Express Highway

YUAN Hong-qing¹, WANG Zai-Xi², WANG Hai-sheng¹, YAN Dong¹

(1. Hubei Provincial Road & Bridge Co. Ltd., Hubei 430056 Wuhan; 2. Hurongxi Express Highway headquarters of Hubei Province, Enshi 445000 Hubei, China)

Abstract: Based on the developping disciplinarian of karst, it is pointed out that distortion and water damage are two basic types of Subgrade damage, and asymmetric settlement and stability lose of subgrade are the material reflections of karst subgrade damage. According to the karstic characters of X¹ where clint, funnel, billabong, dissolved cavern and karstic spring existe at the same time, considering the scale effects of karstic configuration, abiding by the preventting and disposing principles by dredging and leading, spanning, reinforcing and jamming, the disposal material measures for karst subgrade of X¹ section are put forward.

Key words: subgrade engineering; karstic configuration; scale effect; asymmetric settlement; stability; disposal measures