文章编号: 1001-0920(2001)05-0577-04

一种基于粗糙集的信息系统决策规则提取方法

夏雨佳, 李少远, 席裕庚
(上海交通大学 自动化研究所, 上海 200030)

摘 要: 以粗糙集理论为基础, 引入相似性的概念, 并提出其衡量方法。改进了粗糙集理论中不可辨关系的确定条件, 给出了基于新的相似关系的上下近似空间定义, 并举例说明了基于粗糙集的相似性规则提取方法。

关键词: 粗糙集; 决策规则; 不可辨关系; 近似空间
中图分类号: TP 14 **文献标识码**: A

A Method of Inducing Decision Rules Based on Rough Set Theory

XIA Yu-jia, LI Shao-yuan, XI Yu-geng
(Institute of Automation, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: Based on rough set theory, the lower and upper approximations are redefined according to a new more general definition using the similarity relation. A general method for defining similarity relations on objects described by a set of attributes is given. An example shows that the decision rules induced from similarity-based lower approximations of decision classes are more robust than those induced from indiscernibility-based lower approximations.

Key words: rough set; decision rules; indiscernibility relation; approximation

1 引 言

粗糙集理论[1~3]的一个重要特点是其具有很强的定性分析能力, 即不需要预先给定某些特征或属性的数量描述(如统计学中的概率分布、模糊集理论中的隶属度或隶属函数等), 而直接从给定问题的描述集合出发, 通过不可分辨关系和不可分辨类确定给定问题的近似域, 从而找出该问题的内在规律。

本文以粗糙集理论为基础, 将知识与分类联系在一起, 运用相似性分类的概念, 改进粗糙集理论中不可辨关系的确定条件。从而提高对噪音数据的容错能力, 更合理地利用各属性之间的权重关系, 增强信息系统中知识发现的鲁棒性。

2 粗糙集理论

粗糙集理论的出发点是根据目前已有的描述对问题的论域进行划分, 然后针对划分后的每一个组成部分确定其对某一概念的支持程度, 即肯定支持此概念、肯定不支持此概念和可能支持此概念。在

收稿日期: 2000-05-26; 修回日期: 2000-09-29
基金项目: 国家自然科学基金项目(69934020)
作者简介: 夏雨佳(1974—), 男, 上海人, 博士生, 从事智能控制、人工智能及数据挖掘等研究; 席裕庚(1946—), 男, 上海人, 从事复杂系统控制理论及智能机器人等研究。

粗糙集理论中,以上 3 种情况分别用 3 个近似集合表示为正域、负域和边界。为描述方便,可采用知识表示系统和决策系统对问题进行描述。这样,粗糙集的方法和模型便可建立在一种直观的二维表的基础上。

定义 **1**(粗糙集的上近似和下近似) 对于知识表示系统 $S = (U, A)$,设 $B \subseteq A$, $x \subseteq U$,称

$$\begin{cases} \underline{B}X = \{x \in U \mid [x]_{\mathrm{IND}(B)} \subseteq X\} \\ \overline{B}X = \{x \in U \mid [x]_{\mathrm{IND}(B)} \cap X \neq \varnothing\} \end{cases} \tag{1}$$

分别为 x 的 B-下近似和 B-上近似。

定义 **2**(正域、负域和边界)

$$\begin{cases} \mathrm{POS}_B(X) = \underline{B}X \\ \mathrm{NEG}_B(X) = U - \overline{B}X \\ \mathrm{BN}_B(X) = \overline{B}X - \underline{B}X \end{cases} \tag{2}$$

分别称为 X 在 B 下的正域、负域和边界。

当对象属性很多并用数值精确表示时,按照粗糙集的定义,两个对象的属性值必须完全相同,它们之间才具有等价关系,即不可辨关系。这就极大地限制了规则提取的广泛性和准确性,容易受到数据微小差别的影响。在人们认识客观世界、分析解决问题的过程中,存在着大量的相似问题[3]。粗糙集理论实际上与人的认知特性很相似。本文根据相似性的概念给出在粗糙集中近似关系的定义如下:

定义 **3** R 是论域 U 上的不可辨关系(等价关系),$\tilde{R}$ 是 R 扩展的近似关系,当且仅当

$$\begin{cases} \forall x \in U, \quad R(x) \subseteq \tilde{R}(x) \\ \forall x \in U, \quad \forall y \in R, \quad R(y) \subseteq \tilde{R}(x) \end{cases} \tag{3}$$

其中 $\tilde{R}(x) = \{y \in U : y \tilde{R} x\}$ 是 x 的相似类。

定义 **4** X 的 $\tilde{R}$ 下近似和 $\tilde{R}$ 上近似分别定义为

$$\begin{cases} \tilde{R}_*(X) = \{x \in X : \tilde{R}(x) \subseteq X\} \\ \tilde{R}^*(X) = \bigcup_{x \in X} \tilde{R}(x) \end{cases} \tag{4}$$

$\tilde{R}$ 的边界定义为

$$\mathrm{RN}_{\tilde{R}}(X) = \tilde{R}^*(X) - \tilde{R}_*(X) \tag{5}$$

3 基于 ϵ 不可辨关系的近似程度计算方法

对于信息系统中的两个对象 x 和 y,每一个对象都有 n 个属性。计算近似程度的方法通常是用距离来衡量,这种方法的主要缺点是不能判断是正近似还是负近似。另一方面,距离测量已隐含了对称性。我们通过引入文献[4, 5]中的概念来计算 x 和 y 之间各属性的符合程度和分歧程度,经过 ϵ 不可辨关系计算,最后得到综合的近似程度。

x 近似于 y 的部分和谐度用 $C_j(x, y)$ 表示,代表在多大程度上 x_j 近似于 y_j, $j = 1, 2, \cdots, n$。可以看出,$C_j(x, y)$ 代表了属性 a_j 支持近似性的程度。$C_j(x, y)$ 的取值范围为集合{0, 1}的值或[0, 1]之间的值:1) 当属性 a_j 与 x_j 近似于 y_j 完全不符时,$C_j(x, y) = 0$;2) 当属性 a_j 与 x_j 近似于 y_j 完全相符时,$C_j(x, y) = 1$。

x 近似于 y 的部分分歧度用 $D_j(x, y)$ 表示,指在多大程度上 x_j 近似于 y_j。可以看出,$D_j(x, y)$ 反映了属性 a_j 对近似性的负面削弱作用。$D_j(x, y)$ 的取值范围为集合{0, 1}的值或[0, 1]之间的值:1) 当属性 a_j 与 x_j 近似于 y_j 完全不存在分歧时,$D_j(x, y) = 0$;2) 当属性 a_j 与 x_j 近似于 y_j 完全分歧时,$D_j(x, y) = 1$。

当得到部分和谐程度和分歧程度后,便可通过聚合分析来计算整体的和谐程度和分歧程度。整体和谐度为

$$C(x, y) = g(C_1(x, y), \cdots, C_n(x, y)) \tag{6}$$

其中函数 g 是聚合算子,可从如下两种算子簇中选择:

1) 平均算子中加入权重

$$C(x, y) = \sum_{j=1}^{n} w_j C_j(x, y) \tag{7}$$

2) 与运算,如“min”或“乘积”运算

$$C(x, y) = \min_{j=1, 2, \cdots, n} \{C_j(x, y)\} \tag{8}$$

或

$$C(x, y) = \prod_{j=1}^{n} C_j(x, y) \tag{9}$$

整体分歧度为

$$D(x, y) = h(D_1(x, y), \cdots, D_n(x, y)) \tag{10}$$

其中函数 h 应有利于选择最重要的部分分歧度,可采用如下算子

$$D(x, y) = \max_{j=1, 2, \cdots, n} \{D_j(x, y)\} \tag{11}$$

$$D(x, y) = 1 - \prod_{j=1}^{n} (1 - D_j(x, y)) \tag{12}$$

$$D(x, y) = 1 - \prod_{j: D_j(x, y) > C(x, y)} \left(\frac{1 - D_j(x, y)}{1 - C(x, y)} \right) \tag{13}$$

近似度是通过整体和谐度与分歧度综合得到的,即

$$s(x, y) = f(C(x, y), D(x, y)) \tag{14}$$

f 满足条件

$$\begin{cases} f(v,0) = v \\ f(0,v) = 0, \quad \forall v \in [0,1] \\ f(v,1) = 0 \end{cases} \tag{15}$$

f 应有“与”算子的特性，以同时满足和谐度和分歧度。比如

$$s(x,y) = \min\{C(x,y), 1 - D(x,y)\} \tag{16}$$

或

$$s(x,y) = C(x,y)(1 - D(x,y)) \tag{17}$$

在信息系统中，众多属性值难免存在噪音数据，数据的采集过程也会带来偏差。通过 ϵ 不可辨关系可将一些属性之间的微小偏差滤除，从而提高系统决策的鲁棒性。ϵ 不可辨关系是建立在和谐程度计算基础上的，即

$$C_j(x,y) = \begin{cases} 1, & |x_j - y_j| \le \epsilon_j(y_j) \\ 0, & \text{其 它} \end{cases} \tag{18}$$

其中 ϵ 是关于参考对象属性 a_j 的函数。

实际应用中，我们采用

$$\epsilon_j(y_j) = \alpha_j y_j + \beta_j$$

特别地，当 $\alpha_j = 0$ 时，代表一固定的域值；当 $\beta_j = 0$ 时，代表域值随参考对象属性值变化。

根据整体和谐度 $C(x,y) = \min\limits_{j=1,2,\cdots,n}\{C_j(x,y)\}$，只要有一个属性 a_k 其和谐度 $C_k(x,y) = 0$，则 $C(x,y) = 0$，同时 $s(x,y) = 0$。这里不需考虑 $D(x,y)$ 的影响，近似度可简化为 $s(x,y) = C(x,y)$。

4　决策规则提取方法

在信息表 $T = (U,A)$ 中，$A = C \cup D$，C 为一系列的条件属性，D 为决策属性。集合 D 引入分区将 U 分成不同的决策类 U_i，$i = 1,2,\cdots,m$。决策表中的每一条对象都是规则的实例。将每一类实例都提取出来，便形成了判别规则。其形式如下：

$$\text{if } c_1, c_2, \cdots, c_k \text{ then } d_i$$

c_j 是判别条件。在上述基于 ϵ 不可辨关系近似度规则判别的算法中，如果近似度与整体和谐度相同，则对 $k \le n$ 个条件属性，有 $r = c_1 \wedge c_2 \wedge \cdots \wedge c_k$，$r$ 为规则条件。选择 c_j 涉及的属性 a_j 具有如下形式

$$x_i \in [y_j - \epsilon_j(y_j), y_j + \epsilon_j(y_j)]$$

本文用一个数据表来说明基于相似性原则的粗糙集的应用。表 1 描述了 12 家企业 1 年的经营情况。企业用 3 个条件属性来描述，即 a_1：固定资产；a_2：上一年销售额；a_3：经营性质。a_1 和 a_2 为数量值，a_3 属于{0, 1, 2}。决策属性 $d = 1$ 表示企业盈利，$d = 2$ 表示企业亏损。

用上述基于 ϵ 不可辨关系近似度规则判别的算法，设 $\alpha_1 = 0.2, \beta_1 = 0; \alpha_2 = 0, \beta_2 = 10; \alpha_3 = \beta_3 = 0$。近似类为{1}，{2}，{3}，{4, 9}，{5}，{6}，{7}，{8}，{9}，{10}，{11}，{12, 4}，其中主对象在前。上近似空间算法如下：

表 1　12 家企业模拟数据分析表

企业编号	属性 a_1	属性 a_2	属性 a_3	决定 d
1	43	78	0	1
2	54	75	0	2
3	124	50	1	1
4	102	65	1	2
5	98	80	2	2
6	88	102	2	2
7	130	57	0	1
8	128	92	1	2
9	82	59	1	2
10	134	103	2	2
11	58	55	0	1
12	126	71	1	2

参数

P：属性集合

X：对象集合

结果

关于属性集合 P 和 X 的上近似空间

算法

```
{UA = X;
while (I goes through all elements of X)
while (J goes through all elements
        of - X)
if(objects I an J are indiscernible
   with respect to P)
UA = UA + J;
return UA;
}
```

由属性 d 分隔的上下近似空间分别为

$$\tilde{R}_*(U_1) = \{1, 3, 7, 9, 11\}$$

$$\tilde{R}^*(U_1) = \{1, 3, 7, 9, 11\}$$

$$\tilde{R}_*(U_2) = \{2, 5, 6, 8, 10, 12\}$$

$$\tilde{R}^*(U_2) = \{2, 4, 5, 6, 8, 9, 10, 12\}$$

现在通过改变系数来放松相似条件，使 $a_1 = 0.25$ 而不是 0.2，其它参数保持不变。这时近似类的范围扩大为{1}，{2, 1}，{3}，{4, 9, 12}，{5}，{6}，{7}，{8}，{9, 4}，{10}，{11}，{12, 4}。U_1 与 U_2 的近似空间更新如下

$\tilde{R}^{'}_{*}(U_1) = \{1, 3, 7, 11\}$

$\tilde{R}^{*}(U_1) = \{1, 3, 4, 7, 9, 11\}$

$\tilde{R}^{'}_{*}(U_2) = \{5, 6, 8, 10, 12\}$

$\tilde{R}^{*}(U_2) = \{1, 2, 4, 5, 6, 8, 9, 10, 12\}$

根据粗糙集理论[6,7] 可推导出如下规则, 以精确地描述决策类的性质。比如可得出 $d = 2$, 即运营亏损公司的特点为: 从事产业 2; 产值 82 ～ 133; 固定资产 43.2 ～ 64.8, 且产值为 65 ～ 85; 固定资产 78.4 ～ 117.6, 且产值为 70 ～ 90; 固定资产 100.8 ～ 151.2, 且产值为 61 ～ 80; 产值 61 ～ 81, 且从事产业 1。上述特征可作为预测一家企业盈利可能性的依据。

5 结 论

本文所讨论的基于近似空间的粗糙集规则提取方法, 能提高粗糙集中不可辨关系的应用范畴和适应范围。通过引入相似程度的概念和计算方法, 可较好地替换传统粗糙集和推理学习的严格 "与" 计算。基于相似性下近似空间得出的决策规则比基于不可辨关系下近似空间得到的决策规则鲁棒性更强, 前者对属性值的微小变化不会过于敏感。相似关系可以更好地处理数值属性, 对于具有离散值的属性可通过正规变换使之数值化后, 再进行相似性处理。近似标准的放宽扩大了研究对象的范围, 使所得到的决策规则更具普遍意义。

参考文献:

[1] Pawlak Z, Grzymala Busse J, Slowinsk R *et al*. Rough sets[J]. Communications of the ACM, 1995, 38(11): 88-95.

[2] Pawlak Z. Rough sets, theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[3] 李元. 利用相似理论来指导建模[J]. 计算机仿真, 1999, 16(1): 49-53.

[4] D Dubois, H Prade. Criteria aggregation and ranking of alternatives in the framework of fuzzy set theory[J]. Fuzzy Sets and Decision Analysis, 1984, 20: 209-240.

[5] R S Lowinski. A generalization of the indiscernibility relation for rough set analysis of quantitative information[J]. Rivista Di Matematica Perle Scienze Economichee Sociali, 1992, 15(1): 65-78.

[6] J W Grzymala Busse. LERS —— A system for learning from examples based on rough sets[A]. Intelligent Decision Support[C]. Dordrecht: Kluwer Academic Publishers, 1992. 3-18.

[7] J Stefanowski, Vanderpooten D. A general two-stage approach to inducing rules from examples[A]. Rough Sets, Fuzzy Sets and Knowledge Discovery[C]. Berlin: Springer-Verlag, 1994. 317-325.

招 聘 启 事

《信息与控制》是我国自动控制及信息科学的中文核心期刊, 其影响因子名列同类学术期刊的前列。为建立一个与知识创新工程相适应的学术期刊编辑部体系, 现面向国内外公开招聘《信息与控制》专职常务副主编一名。

招聘条件: 1. 年龄在 45 周岁以下(女性 40 周岁以下); 2. 原则上要求具有博士学位, 至少有 3 年以上科研或期刊工作经历, 并具有课题研究或独立期刊编辑能力, 国内应聘者应具有副高级以上专业技术职称; 3. 要求具有自动控制、机械电子工程、信息处理及相近专业学历; 4. 掌握学科发展动态, 并对学科发展有前瞻性、创新性构想, 在专业研究或期刊学领域发表过较高水平的论文, 或出版过专著; 5. 具有团结、协作精神及相应的组织管理和领导能力, 热爱学术期刊编辑工作。

待遇: 1. 受聘者享受国家规定的工资、保险、福利待遇; 2. 在 3 年聘用期内, 每年可获得 10 万元(共 30 万元)专项经费资助, 用于科学研究、岗位津贴等; 3. 聘用期内享受博士后人员的住房待遇。

聘后管理: 1. 聘用期为 3 年, 在聘期间, 被聘人与中科院沈阳自动化研究所签订聘用合同, 明确双方的责、权、利; 聘任期满, 根据情况可以续聘, 并另签合同; 2. 实行严格的聘期目标管理制, 进行年度目标考核, 并将考核结果报中科院人教局, 合同期满的评估工作由院人教局组织进行。

应聘程序: 应聘者可从网上(www.ms.sia.ac./bjb) 下载 "中国科学院文献和期刊领域引进优秀人材自荐表", 按栏目填好后从 E-mail 发到 xk@sia.ac.cn 张主任收, 如不能下载此表格者, 可来函向编辑部索取(沈阳市 南塔街 114 号, 110015)。

联系人: 张主任

联系电话: (024) 23970056