

粗集理论及基于集合论的学习方法*

许 琦, 李永生

(南京工业大学 机械工程学院, 江苏 南京 210009)

摘 要 运用了粗集理论的基本原理, 对不精确、不完整、不确定的数据进行上近似和下近似的计算和分类, 对数据进行简约, 寻找其隐含的关系和模式, 产生规则, 确定规则的可信度, 并通过举例加以说明。

关键词 粗集; 等价; 可信度; 精确度

中图分类号 TP18, TP392, O232 **文献标识码** A **文章编号** 1671-5322(2002)01-0018-03

自 1982 年 Pawlak 发表了经典论文 Rough Sets (简称 RS) 以来, 特别是 20 世纪 90 年代中期, 有影响的与粗集理论有关的国际性会议几乎每年召开, 大量的粗集理论的研究成果不断涌现在国际重要的学术刊物上。RS 已成为国际学术界众人瞩目, 并被认为是 21 世纪作为人工智能和认知科学中新的研究热点。粗集理论的有效性被计算机学科的基础研究人员所认可。

从一系列已有的数据中, 寻找其规律或规则, 预测问题的方向是人们不懈努力的理想目标。就科学和工程而言, 把获得的许多不精确、不完整和不确定的粗糙信息与工程科学相互交叉渗透、预测事物发展是智能科学研究发展的主要方向, 其应用范围几乎涉及到所有的控制系统设计、模式识别、语言和手写体识别, 图像处理、数字逻辑设计、机械故障诊断、信息检索、并行计算、神经网络系统的自学习和拓扑结构优化等领域。

1 粗集理论的基本原理

粗集理论最早是 Z. Pawlak 教授在研究信息系统逻辑的基础上提出的, 他称各种等价关系的等价类集合为粗集, 是某一对象集合的上近似和下近似。在粗集理论中, 基于集合论学习的基本思想, 即根据用户提出的发现任务, 由用户指定的

数据库中某一个或多个属性作为分类的标准, 形成结论属性的子集, 然后根据条件属性子集和结论属性子集, 进行一系列的上、下近似计算, 得到各个类别的分类规则^[1~3]。

1.1 等价类之间的下近似和下近似的计算

假设给定知识库 $S = (R, A)$, R 为一个非空集合, 表示数据库的所有记录, 用 e_i 表示 R 中具体的记录对象; A 是一个非空集合, 表示数据库中全部的属性, 用 a_i 表示 A 中具体属性; R 基于 A 的划分是 E , 基于 A' 的划分是 X , 对 R 的两个划分 E 和 X 之间的关系, 可用两个子集形式表示, 即上近似和下近似。

X_j 的下近似 $\underline{A}(X_j)$ 为 E 中被 X 包含的等价类中的记录并集, 是对于知识 R, A 中所有一定能归于 X_j 的元素的集合, 即:

$$\underline{A}(X_j) = \bigcup \{E_i \in E \mid E_i \subseteq X_j\} \quad (1)$$

X_j 中上近似 $\bar{A}(X_j)$ 为 E 中与 X 的交集非空集的等价类中记录的并集, 是对知识 R, A 中可能归入 X_j 的元素集合, 即:

$$\bar{A}(X_j) = \bigcup \{E_i \mid E_i \in E \text{ 且 } E_i \cap X_j \neq \emptyset\} \quad (2)$$

集合 $\bar{A}(X_j) - \underline{A}(X_j)$ 中记录, 根据 A 的属性值, 不能确定它们是否属于等价类 X_j 的, 是知识既不能归入 $\underline{A}(X_j)$, 也不能归入 $\bar{A}(X_j)$ 的元素集

* 收稿日期 2001-09-28

基金项目 江苏省自然科学基金资助项目(Bk2001095)。

作者简介 许 琦(1965-), 男, 江苏阜宁人, 盐城工学院讲师, 在职博士研究生, 主要从事压力容器的疲劳寿命和可靠性研究。

合,称为 X 的 A 边界。

1.2 粗集的非确定性的精确度 $\alpha_A(X)$ 和 $\rho_A(X)$ 粗糙度

集合的不确定性是由于边界域的存在而引起的,集合的边界域越大,其精确性越低。为了准确地表达这一点,常用精确度 $\alpha_A(X_j)$ 和粗糙度 $\alpha_A(X_j)$ 来表示:

$$\alpha_A(X_j) = \frac{|R| - |\bar{A}(X_j) - \underline{A}(X_j)|}{|R|} \quad (3)$$

其中 $|R|$ 和 $|\bar{A}(X_j) - \underline{A}(X_j)|$ 分别集合 R 、 $(\bar{A}(X_j) - \underline{A}(X_j))$ 中的记录总数,精确度 $\alpha_A(X_j)$ 用来反映 X_j 的知识的完整程度。对于每一个 A ,且 $X_j \subseteq R$ 且 $0 \leq \alpha_A(X_j) \leq 1$ 。当 $\alpha_A(X_j) = 1$ 时,表示 R 中的全部记录数据,根据 A 的属性就可以确定是属于 X_j , X 的边界域是空集;当 $0 < \alpha_A(X_j) < 1$ 时,表示 R 中的记录数据,根据 A 的属性不能确定全部属于 X_j ;当 $\alpha_A(X_j) = 0$ 时,表示 R 中的全部记录都不能根据 A 的属性,确定属于 X_j 。

也可以用 A 的粗糙度来定义集合 X 的不确定的程度:

$$\rho_A(X_j) = 1 - \alpha_A(X_j) \quad (4)$$

与概率论或模糊集合不同,不精确的数值不是事先假定的,而是通过表达知识不精确性的概念近似 计算得到的,这是不精确性的数值表示有限知识的结果。

1.3 粗集的等价类^[4]

对于确定记录 e 的属性 a 取值, $\forall a \in A$ (A 中包含一个或多个属性), $e_i \in R, e_i \in R$, 有一个 a 和 e 的函数 $f(e, a)$, 当 $f(e_i, a) = f(e_j, a)$ 成立, 那么 e_i 和 e_j 基于属性集 a 等价。

在数据库记录集合 R 中,基于属性集合 A 所有等价类的记录集合,称之为基于属性 A 的等价类。在数据库记录集合 R 中,基于属性集 A 的所有等价类,称之为集合 R 的属性集 A 的划分。

粗集的等价概念和传统的集合论的基本概念有本质的区别,在传统集合论中,当两个集合有完全相同的元素时,它们是等价的;而粗集等价理论中,只是在某一个属性之下,集合的取值相等,它是集合间的拓扑结构,不是构成集合的元素间的比较。

1.4 分类规则的形成^[5~6]

应用粗集理论,对数据库进行学习,从中寻找隐含的模式和关系,对数据进行约简,评价数据的重要性,从数据中产生分类规则。通过分析 R 中

两个划分 E 和 X 之间的关系,把 E 视为分类条件, X 视为分类结论,可得到下面的分类规则:

A. 当时 $E \cap X_j \neq \phi$ 时,则有 $r_{ij} : Des(E_i) \rightarrow Des(X_j)$ 。 $Des(E_i)$ 和 $Des(X_j)$ 分别是等价类 E_i 和等价类 X_j 中的记录的特性描述:

(1) 当时 $E_i \cap X_j = E_i$ (E_i 全部被 X_j 包含),建立的规则 r_{ij} 是确定的,规则的可信度 $cf = 1$;

(2) 当时 $E_i \cap X_j \neq E_i$ (E_i 部分被 X_j 包含),建立的规则 r_{ij} 是不确定的,规则的可信度 $cf = \frac{|E_i \cap X_j|}{|E_i|}$;

B. 当时, E_i 和 X_j 不能建立规则。

2 举例

表 1 是描述某医院感冒检测临床病例的数据库,每个个体就是一个病人。属性是各种症状,如头痛(10 表示头不痛,11 表示头痛),发烧(20 表示不发烧,21 表示发低烧,22 表示发高烧),流鼻涕(30 表示不流鼻涕,31 表示流鼻涕),感冒(40 表示普通感冒,41 表示流感),试用粗集理论寻找其规则。

表 1 感冒检测临床病例的数据库

Table 1 Database of Cold Clinical case

	头痛 a_1	发烧 a_2	流鼻涕 a_3	感冒 a_4
e_1	μ_{10}	μ_{20}	μ_{31}	μ_{40}
e_2	μ_{10}	μ_{21}	μ_{30}	μ_{41}
e_3	μ_{10}	μ_{21}	μ_{31}	μ_{40}
e_4	μ_{11}	μ_{20}	μ_{31}	μ_{40}
e_5	μ_{11}	μ_{21}	μ_{31}	μ_{40}
e_6	μ_{11}	μ_{22}	μ_{30}	μ_{41}
e_7	μ_{11}	μ_{22}	μ_{31}	μ_{41}
e_8	μ_{11}	μ_{20}	μ_{30}	μ_{40}
e_9	μ_{11}	μ_{21}	μ_{31}	μ_{41}

在表 1 中, R 基于 $X = \{a_4\}$ 进行划分得 $X = \{X_1, X_2\}$, $X_1 = \{e_1, e_3, e_4, e_5, e_8\}$, $X_2 = \{e_2, e_6, e_7, e_9\}$ 。

R 基于 $A = \{a_1, a_2\}$, 进行划分得, $E = \{E_1, E_2, E_3, E_4, E_5\}$, $E_1 = \{e_1\}$, $E_2 = \{e_2, e_3\}$, $E_3 = \{e_4, e_8\}$, $E_4 = \{e_5, e_9\}$, $E_5 = \{e_6, e_7\}$ 。

根据式(1)得 X 的下近似:

$$\underline{A}(X_1) = \{e_1, e_4, e_8\}, \underline{A}(X_2) = \{e_6, e_7\}$$

根据式(2)得 X 的上近似:

$$\bar{A}(X_1) = \{e_1, e_2, e_3, e_4, e_5, e_8, e_9\},$$

$$\bar{A}(X_2) = \{e_2, e_3, e_5, e_6, e_7, e_9\},$$

$$\bar{A}(X) - \underline{A}(X) = \{\{e_6, e_7\}, \{e_1, e_4, e_8\}\}$$

根据式(3)得确定或精确度:

$$\alpha_A(X_1) = 0.778, \alpha_A(X_2) = 0.667$$

推理规则的形成:

(1) 当 E 为前提, X_1 为结论, 可得到如下规则:

$$r_{11} : Des(E_1) \rightarrow Des(X_1); \text{即}(a_1 = u_{10}) \wedge (a_2 = u_{20}) \rightarrow (a_4 = u_{40}) (cf = 1)$$

$$r_{21} : Des(E_2) \rightarrow Des(X_1); \text{即}(a_1 = u_{10}) \wedge (a_2 = u_{21}) \rightarrow (a_4 = u_{40}) (cf = 1)$$

$$r_{31} : Des(E_3) \rightarrow Des(X_1); \text{即}(a_1 = u_{11}) \wedge (a_2 = u_{20}) \rightarrow (a_4 = u_{40}) (cf = 1)$$

$$r_{41} : Des(E_4) \rightarrow Des(X_1); \text{即}(a_1 = u_{11}) \wedge (a_2 = u_{21}) \rightarrow (a_4 = u_{40}) (cf = 1)$$

r_{51} 不存在。

根据 E 判断 R 中, 记录属于 X_1 类的精确度为 0.778。

(2) 当 E 为前提, X_2 为结论, 可得到如下规则:

$$r_{22} : Des(E_2) \rightarrow Des(X_2); \text{即}(a_1 = u_{10}) \wedge$$

$$(a_2 = u_{21}) \rightarrow (a_4 = u_{41}) (cf = 1)$$

$$r_{42} : Des(E_4) \rightarrow Des(X_2); \text{即}(a_1 = u_{11}) \wedge$$

$$(a_2 = u_{21}) \rightarrow (a_4 = u_{41}) (cf = 1)$$

$$r_{52} : Des(E_5) \rightarrow Des(X_2); \text{即}(a_1 = u_{11}) \wedge$$

$$(a_2 = u_{22}) \rightarrow (a_4 = u_{41}) (cf = 1)$$

r_{12}, r_{32} 不存在。

根据 E 判断 R 中, 记录属于 X_2 类的精确度为 0.667。

由于对感冒的症状没有全部列出, 这里只是说明应用粗集理论推导规则的过程, 得到的规则不一定完全符合实际情况。

3 结论

(1) 应用粗集理论, 能够处理不依赖于数据集之外任何先验信息的数据库, 从现有的数据库中寻找关系和模式, 对数据进行约简, 产生规则, 并能提供规则的可信度;

(2) 在粗集理论中, 粗糙的不确定性来自于对象之间的不可分辨性, 在不可分辨关系下, 集合只能被近似定义, 用可信度函数来刻画粗糙的不确定性, 它完全由数据决定, 比较客观。

参考文献:

- [1] 陈文伟. 智能决策技术[M]. 北京: 电子工业出版社, 1998.
- [2] 许志兴. 粗集理论的若干技术及其应用研究[D]. 南京: 南京航空航天大学, 2001.
- [3] 李文, 孙辉, 陈善本. 一种建立模糊模型的粗糙集方法[J]. 控制理论与应用, 2001, 18(1): 69-75.
- [4] Slowinski K. Rough Set Approach to Analysis of Data form Peritoneal Lavage in Acute Pancreatitis[J]. Med. Inform, 1988, 13(3): 143-159.
- [5] 李永敏, 朱善后. 基于粗糙集理论的数据挖掘模型[J]. 清华大学学报, 1999, 39(1): 110-113.
- [6] Pawlak Z. Rough Sets in Theoretical Aspects of Reasoning about Date[M]. The Netherlannds: Kluwer Academic Publishers, 1991.

Rough Set Theory and Learning Method Based on Set Theory

XU Qi, LI Yong-sheng

(College of Mechanic Engineering of Nanjing Polytechnical University, Jiangsu Nanjing 210009, China)

Abstract Based on the principle of rough set theory, calculation of upper and lower approximation and categorization were performed on inaccurate, insufficient and indefinite data. Data reduction was carried out and implied relation and mode between data were figured out. Thus relative rule were set up and its credibility was established which can be illustrated by examples.

Keywords rough set theory; equivalence; credibility; accuracy