



数 学

双论域粗糙集的近似分类精度度量

朱晓钟

(河海大学,计算机与信息学院(常州),常州 213011)

摘要 由等价关系 R 所决定的近似空间 (U, R) 上, 可用近似分类精度来表示可能的决策中正确决策的百分比。将近似分类精度概念推广到一般关系双论域粗糙集的近似空间上。通过引入独立集概念, 给出了度量公式, 最后通过实例验证了其合理性。

关键词 双论域 粗糙集 近似分类精度

中图法分类号 O144.5; **文献标志码** A

粗糙集是 20 世纪 80 年代初由波兰数学家 Z. Pawlak^[1]首先提出的处理不确定知识的数学理论, 它能有效地分析和处理不精确、不一致、不完整等各种不完备信息, 并从中发现隐含的知识, 揭示潜在的规律。粗糙集理论^[2]认为知识即为分类能力, 分类能力越强知识愈丰富。经典粗糙集理论以等价关系(自反性、对称性、传递性)为基础, 通过等价关系对论域进行划分, 而知识即表现为等价关系对论域划分的结果, 划分越细, 知识越精确, 则粒度越小, 从而又将知识与粒度紧密联系在一起。为描述知识不确定性, 粗糙集理论通过引入上、下近似运算来逼近论域中的任一概念。粗糙集理论在人工智能、机器学习、知识发现、数据挖掘和决策分析等领域得到了广泛的应用。在这些应用的推动下, 粗糙集理论得到进一步推广, 其中之一是将单一论域的粗糙集模型推广为双论域的模型^[3,4], 拓宽了研究和应用的范围。

2009 年 12 月 17 日收到

作者简介: 朱晓钟(1978—), 江苏常州人, 讲师, 硕士生, 研究方向: 粗糙集理论、数据挖掘。

近似分类精度是 Pawlak 粗糙集理论中的基本概念。指出了可能的决策中正确决策所占的百分比。本文通过引入独立集, 将近似分类精度概念推广到一般二元关系下的双论域粗糙集中。给出了度量公式, 并通过实例显示了其应用价值。

1 Pawlak 粗糙集的近似分类精度

1.1 粗糙集理论的基本概念

定义 1 四元组 $I = (U, A, V, f)$ 是一个信息系统, 其中: U 为对象的非空有限集合; A 为属性的非空有限集合; $V = \bigcup_{a \in A} V_a$, V_a 是属性 a 的值域; $f: U \times A \rightarrow V$ 是一个信息函数, 它为每个对象的每个属性赋予一个信息值, 即对任意 $a \in A, x \in U, f(x, a) \in V_a$ 。

定义 2 设 $P \subseteq A, X \subseteq U$ 。 X 关于 P 的下近似和上近似分别定义为:

$$P_* X = \{x \in U | [x]_P \subseteq X\};$$

$$P^* X = \{x \in U | [x]_P \cap X \neq \emptyset\}.$$

其中 $[x]_P$ 表示 P 划分下包含元素 $x \in U$ 的等价类。

定义 3 X 关于近似空间 A 的近似精度为:

$$\alpha_P(X) = \frac{|P_* X|}{|P^* X|}.$$

其中 $| \cdot |$ 表示集合的基数。近似精度反映了根据现有知识对 X 的了解程度。

1.2 Pawlak 粗糙集的近似分类精度度量公式^[5]

设 $I=(U, A, V, f)$ 是一个信息系统, $P \subseteq A$ 。令 $X=\{X_1, X_2, \dots, X_n\}$ 是 U 的一个划分或分类, 划分 X 独立于属性集 P 。例如, 划分 X 可能由一个专家为解决分类问题所给出。子集 $X_i(i=1, 2, \dots, n)$ 是划分 X 的一个类(或块)。 I 中的划分 X 关于 P 的下近似和上近似分别定义为 $P_* X=\{P_* X_1, P_* X_2, \dots, P_* X_n\}$ 和 $P^* X=\{P^* X_1, P^* X_2, \dots, P^* X_n\}$ 。

系数

$$d_p(X) = \frac{\sum_{i=1}^n |P_* X_i|}{\sum_{i=1}^n |P^* X_i|}.$$

称为分类 X 关于 P 的近似精度, 简称为近似分类精度。

2 双论域粗糙集的近似分类精度

2.1 双论域粗糙集基本概念

定义4^[6] 设 U, W 为两个非空有限集合, $R \subseteq U \times W$ 是一个从 U 到 W 的二元关系, 称三元组 (U, W, R) 为广义近似空间。 R 可以看成是从 U 到 2^W 的一个集值映射 $r: U \rightarrow 2^W$, 对任意的 $x \in U$, 记: $r(x)=\{y \in W \mid (x, y) \in R\}$, $r(x)$ 表示 W 中所有与元素 x 相关的元素的集合。对于任意的 $Y \subseteq W$, Y 关于近似空间 (U, W, R) 的下近似和上近似分别定义为:

$$R_*(Y) = \{x \mid r(x) \subseteq Y\},$$

$$R^*(Y) = \{x \mid r(x) \cap Y \neq \emptyset\}.$$

当 $U=W$ 时, 可将 $r(x)$ 看成 x 的邻域, 这时得到的模型就退化为一般关系下的单论域粗糙集模型。

2.2 独立集基本概念^[7]

定义5 设 U 为非空有限论域, $R \subseteq U \times W$ 为二元关系, 如果 $x \in U$ 且 $r(x) = \emptyset$, 则称 x 为 R 关系下的一个独立元素, 由 R 关系下所有独立元素组成的集合称为独立集, 用符号 S 表示。

$$S = \{x \mid x \in U, r(x) = \emptyset\}$$

关于独立集 S 的性质有:

(1) $R_*(\emptyset) = S, R^*(\emptyset) = \emptyset, R_*(W) = U, R^*(W) = S^c$ 。 $(S^c$ 表示 S 在 U 中的补集)

(2) $S \subseteq R_*(Y), R^*(Y) \subseteq S^c$ 。

(3) $R_*(Y)-S \subseteq R^*(Y)$ 。

(4) 如果 $S \neq \emptyset$, 则 $R_*(Y) \neq R^*(Y)$ 。

其中 $Y \in 2^W$, 证明从略, 参见文献[7]。

2.3 双论域粗糙集的近似精度

定义3给出了Pawlak粗糙集的近似精度度量公式, 但此公式不适合一般关系双论域粗糙集。由定义4得一般关系双论域粗糙集的下近似和上近似分别为:

$$R_*(Y) = \{x \in U \mid r(x) \subseteq Y\} = \{x \in U \mid r(x) = \emptyset\} \cup \{x \in U \mid \emptyset \neq r(x) \subseteq Y\};$$

$$R^*(Y) = \{x \in U \mid r(x) \cap Y \neq \emptyset\} \subseteq \{x \in U \mid r(x) \neq \emptyset\}.$$

显然, $R_*(Y) \not\subseteq R^*(Y)$, 若按 $\alpha_R(Y) = \frac{|R_* Y|}{|R^* Y|}$ 计算, 则 $\alpha_R(Y) \notin [0, 1]$, 与近似精度 $\alpha_R(Y) \in [0, 1]$ 的定义相违背。因此, 一般关系双论域粗糙集中, 根据现有知识对 $Y(Y \in 2^W)$ 的了解程度, 即近似精度应定义为:

$$\alpha_R(Y) = \frac{\{x \in U \mid \emptyset \neq r(x) \subseteq Y\}}{\{x \in U \mid r(x) \cap Y \neq \emptyset\}} = \frac{|R_* Y - S|}{|R^* Y|}.$$

2.4 双论域粗糙集的近似分类精度度量公式

设 $I=(U, W, A, V, f)$ 是一个信息系统, 令 $Y=\{Y_1, Y_2, \dots, Y_n\}$ 是 W 的一个划分或分类, 划分 Y 与属性无关。例如, 划分 Y 可能由一个专家为解决分类问题所给出。子集 $Y_i(i=1, 2, \dots, n)$ 是划分 Y 的一个类(或块)。 I 中的划分 Y 关于 R 的下近似和上近似分别定义为 $R_* Y=\{R_* Y_1, R_* Y_2, \dots, R_* Y_n\}$ 和 $R^* Y=\{R^* Y_1, R^* Y_2, \dots, R^* Y_n\}$ 。由双论域粗糙集的上、下近似集概念及独立集概念, 通过修正Pawlak粗糙集的度量公式, 得到双论域粗糙集的近似分类精度的度量公式。

系数

$$d'_R(Y) = \frac{\sum_{i=1}^n |R_* Y_i - S|}{\sum_{i=1}^n |R^* Y_i|}.$$

称为分类 Y 关于关系 R 的近似分类精度, S 为独立集。

例 1 在社区医疗管理系统中,若某居民迁出社区,通常不能删除该居民留存在社区系统中的信息,该居民的既往医疗信息有被查询的可能,需要继续保留。一般的做法是将该居民与系统中疾病集之间的映射关系终止,所以独立集 S 常常非空。

设 $U = \{a, b, c, d, e, f\}$ 为该社区居民的集合, $W = \{A, B, C, D, E, F, G, H, I\}$ 为该社区常见疾病的集合, $R = \{(c, A), (c, E), (d, A), (d, E), (e, B), (e, C), (e, D), (e, I), (f, H)\}$ 为居民与疾病之间的关系。设 $Y = \{Y_1, Y_2\}$, 其中 $Y_1 = \{A, B, E\}$, $Y_2 = \{C, D, F, G, H, I\}$ 。现实意义为, Y_1 为近阶段多发的疾病集, Y_2 为近阶段相对少发的疾病集。

则, $r(a) = r(b) = \emptyset$, $r(c) = r(d) = \{A, E\}$,

$$r(e) = \{B, C, D, I\}, r(f) = \{H\};$$

$$S = \{a, b\};$$

$$R_*(Y_1) = \{a, b, c, d\}, R^*(Y_1) = \{c, d, e\};$$

$$R_*(Y_2) = \{a, b, f\}, R^*(Y_2) = \{e, f\};$$

$$d'_R(Y) = \frac{\sum_{i=1}^n |R_* Y_i - S|}{\sum_{i=1}^n |R^* Y_i|} = \frac{(4-2)+(3-2)}{3+2} = \frac{3}{5}.$$

以上例中 $Y_1 = \{A, B, E\}$, $R_*(Y_1) = \{a, b, c, d\}$ 为例分析, $r(a) = \emptyset$, 而 \emptyset 并非是 Y_1 中的元素 A, B, E 中的任何一个,说明 a 不患有 A, B, E 这三种疾病。但在下近似集的计算中,因为 $\emptyset \subseteq Y_1$,使得 $a \in R_*(Y_1)$ 。同理,使得 $b \in R_*(Y_1)$ 。在社区医疗管理系统中, $r(a) = r(b) = \emptyset$, 只是表明 a, b 两位居民已经搬离了本社区。将孤立集中元素放入下近似是不正确的决策。在上近似集的计算中因为 $\emptyset \cap Y_1 = \emptyset$, 所以孤立

集 S 中的元素不可能进入上近似集,即孤立集中的元素不是可能的决策。为了体现“可能的决策中正确决策的百分比”,应将集合计算上属于而逻辑上不该属于的元素,也就是孤立集中的元素排除。

例 1 的现实意义表明,当前社区中的 c, d 两位居民患病的可能性很大,应重点关注;而 f 居民患病的可能性较小,可降低关注。

3 总结

近似分类精度是粗糙集中的重要概念,是考查分类效果的重要指标。本文通过引入独立集概念,给出了双论域粗糙集中对近似分类精度进行度量的公式,并对公式的合理性进行了分析。这将有助于对双论域粗糙集的各种其它性质展开深入研究。

参 考 文 献

- 1 Pawlak Z. Rough sets. International Journal of Computer and Information Science, 1982;11:341—356
- 2 张文修,吴伟志,梁吉业,等.粗糙集理论与方法.北京:科学出版社, 2001
- 3 Pei D W, Xu Z B. Rough set models on two universes. International Journal of General Systems, 2004;33:569—581
- 4 余 杨.双论域的粗糙集模型.科学技术与工程, 2005;5(10): 661—662
- 5 Pawlak Z. Rough set: theoretical aspects of reasoning about data. Dordrecht: Kluwer Academic Publishers, 1991
- 6 Yao Y Y, Wong S K M, Lin T Y. A review of rough set models. Rough Sets and Data Mining: Analysis for Imprecise Data. Boston: Kluwer Academic Publishers, 1997:47—75
- 7 Liu G, Zhu W. The algebraic structures of generalized rough set theory. Information Sciences, 2008;178(21): 4105—4113

Approximation Classified Precision Based on Two Universal Rough Sets

ZHU Xiao-zhong

(College of Computer and Information (Changzhou), Hohai University, Changzhou 213011, P. R. China)

[Abstract] On Pawlak rough set model, the concept of approximation classified precision expressed the percentage which the correct decision occupied in all possible decision-making. The mentioned concept is promoted to an approximation space which considering an arbitrary binary relation on two universal sets. By introducing the concept of solitary set, proposed the measurement formula, and finally an example is used to show its rationality.

[Key words] two universal sets rough set approximation classified precision