包含度与粗糙集数据分析中的度量

梁吉业[1),2)] 徐宗本[1)] 李月香[2)]

[1)] (西安交通大学理学院信息与系统科学研究所 西安 710049)

[2)] (山西大学计算机科学系 太原 030006)

摘 要 粗糙集理论是一种新的处理模糊和不确定知识的软计算工具. 粗糙集数据分析是粗糙集理论中的主要应用技术之一, 它主要用来分析数据的性质、粗糙分类、分析属性的依赖性和属性的重要性、抽取决策规则等, 在人工智能与认知科学领域有着重要的应用. 该文通过将包含度概念引入到粗糙集理论中, 建立了包含度与粗糙集数据分析中的度量之间的关系, 证实了粗糙集数据分析中的有关度量均可归结为包含度. 这些结论有助于人们深刻理解粗糙集数据分析的本质, 可作为建立粗糙集数据分析中的度量的主要依据.

关键词 粗糙集, 包含度, 数据分析, 度量

中图法分类号: TP18

Inclusion Degree and Measures of Rough Set Data Analysis

LIANG Ji-Ye[1),2)] XU Zong-Ben[1)] LI Yue-Xiang[2)]

[1)] (*Institute for Information and System Science, Faculty of Science, Xi'an Jiaotong University, Xi'an* 710049)

[2)] (*Department of Computer Science, Shanxi University, Taiyuan* 030006)

Abstract Rough set theory is emerging as a powerful tool for reasoning about data. Rough set data analysis is one of the main application techniques arising from rough set theory. It provides a technique for gaining insights into properties of data, dependencies, and significance of individual attributes in databases, and has important applications to artificial intelligence and cognitive science, as a tool for dealing with vagueness and uncertainty of facts, and in classification. In order to analyze data effectively, many measures are defined in rough set data analysis, for example, accuracy of rough set, degree of rough belonging, accuracy of approximation of classification, measure of dependency of attributes, and accuracy of decision rule, etc. Although these measures can be applied to justifying effectiveness of data analysis, it is unclear what is the main foundation behind these measures and whether they have any common characteristics. In this paper, the relationship between inclusion degree and measures of rough set data analysis are set up by introducing a concept of inclusion degree in rough set theory. We show that: (1)Accuracy of rough set and degree of rough belonging can be reduced to inclusion degree; (2)Accuracy of approximation of classification and quality of approximation of classification can be reduced to inclusion degree; (3)Measure of dependency of attributes and measure of importance of attributes can be reduced to inclusion degree; (4)Accuracy of decision rule can be reduced to inclusion degree. These results will be very helpful for people to understand the essence of rough set data analysis, and can be regarded as the main foundation of measures which are defined for rough set data analysis.

Keywords rough set, inclusion degree, data analysis, measure

收稿日期: 2000-08-14 **本课题得到国家自然科学基金**(69805004, 69975016)**和山西省留学回国人员基金资助. 梁吉业**, 男, 1962 年生, 博士研究生, 副教授, 主要研究方向为粗糙集理论与知识发现、人工智能. **徐宗本**, 男, 1955 年生, 博士, 教授, 博士生导师, 主要研究方向为神经网络、遗传算法、非线性数值分析. **李月香**, 女, 1958 年生, 副教授, 主要研究方向为智能控制.

1 引　言

粗糙集理论[1,2]是一种新的处理模糊和不确定知识的软计算工具. 目前, 它已被成功地应用于机器学习、决策分析、过程控制、数据挖掘等领域[2-6], 并且越来越受到国际学术界的广泛关注.

粗糙集数据分析是粗糙集理论中的主要应用技术之一[7,8], 它主要用来分析数据的性质、粗糙分类、分析属性的依赖性和属性的重要性、抽取决策规则等, 在人工智能与认知科学领域有着重要的应用. 为了有效地分析和处理数据, 在粗糙集数据分析中定义了许多度量, 如粗糙集的近似精度、近似分类质量、近似分类精度、属性的依赖性度量、属性的重要性度量和规则可信度等. 那么, 这些度量主要的依据是什么? 有没有一些共同的特性?

上述问题的回答将有助于人们深刻理解粗糙集数据分析的本质和有效地利用粗糙集数据分析去解决实际问题.

我们知道, 包含度理论[9]是一种描述不确定性关系的有效度量方法, 在人工智能、专家系统和模糊集理论等领域有着重要的应用[9,10]. 那么, 包含度与粗糙集数据分析中的度量之间有没有关系? 它们的关系如何呢?

本文主要讨论包含度与粗糙集数据分析中的度量之间的关系, 并且证实了粗糙集数据分析中的有关度量均可归结为包含度.

2 粗糙集理论的基本概念

2.1 信息系统

四元组 $S=(U,A,V,f)$ 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. A 中的属性又可分为两个不相交的子集, 即条件属性集 C 和决策属性集 D, $A=C\cup D$.

$V=\bigcup_{a\in A}V_a$, V_a 是属性 a 的值域.

$f: U\times A\to V$ 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 $a\in A$, $x\in U$, $f(x,a)\in V_a$.

2.2 不可区分关系

令 $P\subseteq A$, 定义由属性集 P 决定的不可区分关系 $\tilde{P}$ 为

$$\tilde{P}=\{(x,y)\in U\times U \mid \forall a\in P, f(x,a)=f(y,a)\}.$$

如果 $(x,y)\in\tilde{P}$, 则称 x 和 y 是 P 不可区分的. 容易证明对于任意 $P\subseteq A$, 不可区分关系 $\tilde{P}$ 是 U 上的等价关系. 符号 $U/\tilde{P}$ (简记为 U/P) 表示不可区分关系 $\tilde{P}$ 在 U 上导出的划分, 即由 $\tilde{P}$ 决定的等价类的集合. $\tilde{P}$ 的等价类称为 S 中的 P-基本集. 符号 $[x]_P$ 表示包含元素 $x\in U$ 的 $\tilde{P}$ 的等价类.

2.3 近似集

设 $P\subseteq A$, $X\subseteq U$. X 关于 P 的下近似定义为

$$\underline{P}X=\cup\{Y \mid Y\in U/P, Y\subseteq X\};$$

X 关于 P 的上近似定义为

$$\overline{P}X=\cup\{Y \mid Y\in U/P, X\cap Y\neq\varnothing\};$$

X 关于 P 的边界定义为

$$BN_P(X)=\overline{P}X-\underline{P}X.$$

3 包含度

定义 1[9]. 设 $(L,\prec)$ 是偏序集. 若对于任意 $a,b\in L$ 有数 $D(b/a)$ 对应, 且满足:

(1) $0\le D(b/a)\le 1$.

(2) 当 $a\prec b$ 时, $D(b/a)=1$.

(3) 当 $a\prec b\prec c$ 时, $D(a/c)\le D(a/b)$.

(4) 当 $a\prec b$ 时, 对 $\forall c\in L$ 有 $D(a/c)\le D(b/c)$,

则称 D 为 L 上的包含度.

在定义 1 中, (1) 是对包含度的规范化, 包含度在 $[0,1]$ 中取值. (2) 表示包含度与经典包含的谐调性, 经典包含关系是包含度为 1 的特殊情况. (3) 与 (4) 是包含度的单调性. 在使用包含度时, 有时可仅利用 (3) 或 (4) 之一即可.

例 1　设 X 是有限集合, $F=\{Y \mid Y\subseteq X\}$, $\subseteq$ 为 F 上的偏序关系. 对任意 $A,B\in F$, 记

$$D_0(B/A)=\begin{cases}\dfrac{|B\cap A|}{|A|}, & A\neq\varnothing\\ 1, & A=\varnothing\end{cases}$$

则 $D_0(B/A)$ 为 A 关于 B 的包含度, 即 B 包含 A 的程度. 这里 $|A|$ 表示集合 A 的基数.

4 包含度与粗糙集数据分析中的度量之间的关系

4.1 粗糙集的近似精度与粗糙隶属度可以归结为包含度

设 $S=(U,A,V,f)$ 是一个信息系统. $P\subseteq A$, $X\subseteq U$, 且 $X\neq\varnothing$. 粗糙集 X 关于 P 的近似精度定义为[1]

$$\alpha_P(X) = \frac{|\underline{P}X|}{|\overline{P}X|},$$

易知

$$\alpha_P(X) = \frac{|\underline{P}X \cap \overline{P}X|}{|\overline{P}X|} = D_0(\underline{P}X / \overline{P}X).$$

元素 $x \in U$ 关于集合 X 的粗糙隶属度定义为[3]

$$\mu_X^P(x) = \frac{|X \cap [x]_P|}{|[x]_P|},$$

易知

$$\mu_X^P(x) = D_0(X / [x]_P).$$

因此, $\alpha_P(X)$ 与 $\mu_X^P(x)$ 均可归结为包含度

4.2 近似分类精度与近似分类质量可以归结为包含度

设 $S = (U, A, V, f)$ 是一个信息系统, $P \subseteq A$. 令 $Y = \{Y_1, Y_2, \cdots, Y_n\}$ 是 U 的一个划分或分类, 划分 Y 独立于属性集 P. 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 ($i = 1, 2, \cdots, n$) 是划分 Y 的一个类(或块). S 中的划分 Y 关于 P 的下近似和上近似分别定义为 $\underline{P}Y = \{\underline{P}Y_1, \underline{P}Y_2, \cdots, \underline{P}Y_n\}$ 和 $\overline{P}Y = \{\overline{P}Y_1, \overline{P}Y_2, \cdots, \overline{P}Y_n\}$.

系数

$$d_P(Y) = \frac{\sum_{i=1}^{n} |\underline{P}Y_i|}{\sum_{i=1}^{n} |\overline{P}Y_i|}$$

称为分类 Y 关于 P 的近似精度[1], 简称为近似分类精度

系数

$$\gamma_P(Y) = \frac{\sum_{i=1}^{n} |\underline{P}Y_i|}{|U|}$$

称为分类 Y 关于 P 的近似质量[1], 简称为近似分类质量

下面设 $Y = \{Y_1, Y_2, \cdots, Y_n\}$ 是 U 上的一个划分, $F = \{\{F_1, F_2, \cdots, F_n\} \mid F_i \subseteq Y_i, i = 1, 2, \cdots, n\}$, $X = \{X_1, X_2, \cdots, X_n\} \in F$, $Z = \{Z_1, Z_2, \cdots, Z_n\} \in F$.

在 F 上定义偏序关系 $\prec$ 如下:

$X \prec Z$ 当且仅当 $X_i \subseteq Z_i$, $i = 1, 2, \cdots, n$.

对于任意 $X, Z \in F$, 定义

$$D_1(X / Z) = \frac{|\left(\bigcup_{i=1}^{n} X_i\right) \cap \left(\bigcup_{i=1}^{n} Z_i\right)|}{|\bigcup_{i=1}^{n} Z_i|}.$$

易证 D_1 为 F 上的包含度

由于 $d_P(Y) = D_1(\underline{P}Y / \overline{P}Y)$, $\gamma_P(Y) = D_1(\underline{P}Y / Y)$, 因此, $d_P(Y)$ 与 $\gamma_P(Y)$ 可以归结为包含度

4.3 属性依赖性度量与属性重要性度量可以归结为包含度

设 $S = (U, A, V, f)$ 是一个信息系统, $A = C \cup D$, 其中 C 为条件属性集, D 为决策属性集. 令 $P \subseteq C$, $Q \subseteq D$. P 与 Q 之间的依赖性度量定义为[1]

$$\gamma(P, Q) = \frac{|POS_P(Q)|}{|U|},$$

其中 $POS_P(Q) = \bigcup_{Y \in U/Q} \underline{P}Y$.

令 F 表示 U 上所有划分的集合, $X = \{X_1, X_2, \cdots, X_n\} \in F$, $Z = \{Z_1, Z_2, \cdots, Z_m\} \in F$. 在 F 上定义偏序关系 $\prec$ 如下:

$X \prec Z$ 当且仅当对于任意 $X_i \in X$, 存在 $Z_j \in Z$, 使得 $X_i \subseteq Z_j$.

对于任意 $X, Z \in F$, 定义

$$D_2(Z / X) = \frac{|\bigcup_{Z_j \in Z}\left(\bigcup_{X_i \subseteq Z_j} X_i\right)|}{|U|},$$

下证 $D_2(Z / X)$ 为包含度

(1) $0 \le D_2(Z / X) \le 1$ 显然成立

(2) 设 $X = \{X_1, X_2, \cdots, X_n\} \in F$, $Z = \{Z_1, Z_2, \cdots, Z_m\} \in F$, 且 $X \prec Z$, 则有 $m \le n$ 且存在 $\{1, 2, \cdots, n\}$ 的一个划分 $E = \{E_1, E_2, \cdots, E_m\}$ 满足

$$Z_j = \bigcup_{i \in E_j} X_i, \quad j = 1, 2, \cdots, m,$$

所以

$$\bigcup_{Z_j \in Z}\left(\bigcup_{X_i \subseteq Z_j} X_i\right) = \bigcup_{Z_j \in Z} Z_j = U,$$

因此

$$D_2(Z / X) = \frac{|U|}{|U|} = 1.$$

(3) 令 $X = \{X_1, X_2, \cdots, X_n\} \in F$, $Z = \{Z_1, Z_2, \cdots, Z_m\} \in F$, $Y = \{Y_1, Y_2, \cdots, Y_l\} \in F$, 且 $X \prec Z \prec Y$, 则 $l \le m$ 且存在 $\{1, 2, \cdots, m\}$ 的一个划分 $E = \{E_1, E_2, \cdots, E_l\}$ 使得

$$Y_j = \bigcup_{i \in E_j} Z_i, \quad j = 1, 2, \cdots, l.$$

下面证明

$$\bigcup_{X_j \in X}\left(\bigcup_{Y_i \subseteq X_j} Y_i\right) \subseteq \bigcup_{X_j \in X}\left(\bigcup_{Z_i \subseteq X_j} Z_i\right) \tag{1}$$

事实上, 对于 $\forall X_j \in X$, $\forall Y_{i_0} \in Y$ 且 $Y_{i_0} \subseteq X_j$, 由 $Z \prec Y$ 知, $Y_{i_0} = \bigcup_{i \in E_{i_0}} Z_i$. 又对于 $\forall i_1 \in E_{i_0}$, 有 $Z_{i_1} \subseteq Y_{i_0}$, 从而 $Z_{i_1} \subseteq X_j$, 所以 $Z_{i_1} \subseteq \bigcup_{X_j \in X}\left(\bigcup_{Z_i \subseteq X_j} Z_i\right)$, 即 $Y_{i_0} \subseteq \bigcup_{X_j \in X}\left(\bigcup_{Z_i \subseteq X_j} Z_i\right)$. 因此式(1)成立

由式(1)可知

$$D_2(X / Y) \le D_2(X / Z).$$

(4) 令 $X, Z \in F$, $X \prec Z$. 对于 $\forall Y \in F$, 一定有

$$\bigcup_{X_j\in X}\left(\bigcup_{Y_i\subseteq X_j}Y_i\right)\subseteq\bigcup_{Z_j\in Z}\left(\bigcup_{Y_i\subseteq Z_j}Y_i\right)\qquad(2)$$

成立

事实上, 对于$\forall X_j\in X$, $\forall Y_{i_0}\in Y$ 且 $Y_{i_0}\subseteq X_j$, 由 $X\prec Z$ 知, 存在 $Z_{j_0}\in Z$ 使得 $X_j\subseteq Z_{j_0}$, 从而 $Y_{i_0}\subseteq Z_{j_0}$, 即 $Y_{i_0}\subseteq\bigcup_{Z_j\in Z}\left(\bigcup_{Y_i\subseteq Z_j}Y_i\right)$. 因此式(2)成立

由式(2)可知

$$D_2(X/Y)\leq D_2(Z/Y),$$

因此, D_2 为 F 上的包含度

由于 $\gamma(P,Q)=D_2((U/Q)/(U/P))$, 从而 $\gamma(P,Q)$ 可以归结为包含度, 亦即划分 U/Q 包含划分 U/P 的程度

在粗糙集数据分析里, 条件属性子集 $C\subseteq C$ 关于决策属性集 D 的重要性度量定义为[1]

$$\gamma(C,D)-\gamma(C-C',D).$$

特别当 $C'=\{c\}$ 时, $\gamma(C,D)-\gamma(C-\{c\},D)$ 即为属性 $c\in C$ 关于 D 的重要性度量

由于 $\gamma(C,D)-\gamma(C-C',D)=D_2((U/D)/(U/C))-D_2((U/D)/(U/(C-C')))$, 所以, $C\subseteq C$ 关于 D 的重要性度量也可以归结为包含度计算

4.4　规则可信度可以归结为包含度

设 $S=(U,A,V,f)$ 是一个信息系统, $A=C\cup D$, 其中 C 为条件属性集, D 为决策属性集. 令 X_i 和 Y_j 分别代表 U/C 和 U/D 中的各个等价类, $Des(X_i)$ 表示对等价类 X_i 的描述, 即等价类 X_i 对于各条件属性值的特定取值; $Des(Y_j)$ 表示对等价类 Y_j 的描述, 即等价类 Y_j 对于各决策属性值的特定取值

决策规则定义如下[6]:

$$r_{ij}: Des(X_i)\rightarrow Des(Y_j),$$

规则可信度

$$\mu(X_i,Y_j)=\frac{|Y_j\cap X_i|}{|X_i|},$$

其中 $0\leq\mu(X_i,Y_j)\leq 1$.

当 $\mu(X_i,Y_j)=1$ 时, r_{ij} 是确定的; 当 $0<\mu(X_i,Y_j)<1$ 时, r_{ij} 是不确定的; 当 $\mu(X_i,Y_j)=0$ 时, $Des(X_i)$ 与 $Des(Y_j)$ 不能建立规则

由于 $\mu(X_i,Y_j)=D_0(Y_j/X_i)$, 所以, 规则可信度 $\mu(X_i,Y_j)$ 可以归结为包含度

5　结束语

粗糙集数据分析是粗糙集理论中的主要应用技术之一, 本文通过将包含度概念引入到粗糙集理论中, 建立了包含度与粗糙集数据分析中的度量之间的关系, 证实了粗糙集数据分析中的有关度量均可归结为包含度. 这些结论有助于人们深刻理解粗糙集数据分析的本质, 也可作为建立粗糙集数据分析中的度量的主要依据

参　考　文　献

1 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic Publishers, 1991

2 Pawlak Z *et al*. Rough sets. Communications of the ACM, 1995, 38(11): 89- 95

3 Pawlak Z. Rough set theory and its application to data analysis. Cybernetics and Systems, 1998, 29(9): 661- 668

4 Wang Jue, Wang Ren, Miao Duo-Qian *et al*. Data enriching based on rough set theory. Chinese Journal of Computers, 1998, 21(5): 393- 400(in Chinese)
(王　珏, 王　任, 苗夺谦等. 基于 Rough Set 理论的"数据浓缩". 计算机学报, 1998, 21(5): 393- 400)

5 Liang Ji-Ye, Xu Zong-Ben, Miao Duo-Qian. Reduction of knowledge in incomplete information systems. In: Shi Zhong-Zhi *et al* eds. Proceedings of Conference on Intelligent Information Processing. Beijing: Publishing House of Electronics Industry, 2000. 528- 532

6 Zhang Wen-Xiu, Wu Wei-Zhi, Liang Ji-Ye, Li De-Yu. Theory and Method of Rough Set. Beijing: Science Press, 2001(in Chinese)
(张文修, 吴伟志, 梁吉业, 李德玉. 粗糙集理论与方法. 北京: 科学出版社, 2001)

7 Gediga G, Düntsch I. Uncertainty measures of rough set prediction. Artificial Intelligence, 1998, 106: 109- 137

8 Guan J W, Bell D A. Rough computational methods for information systems. Artificial Intelligence, 1998, 105: 77- 103

9 Zhang Wen-Xiu, Leung Yee. The Uncertainty Reasoning Principles. Xi'an: Xi'an Jiaotong University Press, 1996(in Chinese)
(张文修, 梁　怡. 不确定性推理原理. 西安: 西安交通大学出版社, 1996)

10 Leung Yee, Zhang Wen-Xiu. The degree of the consistency on fuzzy rules and the method to delete rules. Chinese Journal of Computers, 1997, 20(10): 947- 952(in Chinese)
(梁　怡, 张文修. 模糊规则的谐调度和矛盾规则的排除方法. 计算机学报, 1997, 20(10): 947- 952)