

F-模糊集属性信息系统的知识获取^{*}

闫德勤¹ 迟忠先²

(辽宁师范大学计算机系 大连116029)¹ (大连理工大学计算机系 大连116024)²

摘要 模糊集属性信息系统的知识获取是一个新的研究领域,与其相关的理论和方法具有重要意义。本文提出了一种新的概念F-模糊集,给出了一种在F-模糊集属性信息系统中进行知识获取的方法。

关键词 F-模糊集, 模糊集, 知识获取

Knowledge Acquisition with F-Fuzzy Sets Attributes

YAN De-Qin¹ CHI Zhong-Xian²

(Department of Computer Science, Liaoning Normal University, Dalian 116029)¹

(Department of Computer Science, Dalian University of Technology, Dalian 116024)²

Abstract Knowledge acquisition for information systems with fuzzy sets attributes is a newly developed research area, for which the related theory and method are of important significance. In this paper, a new concept F-fuzzy sets is proposed, and a method for knowledge acquisition with fuzzy sets attributes of a information system is developed.

Keywords F-Fuzzy sets, fuzzy sets, knowledge acquisition

1 引言

对于属性数据具有经典(crisp)形式的信息系统应用粗糙集已有丰富的理论与方法^[4,5]。对于连续属性数据的处理也有一些有效的方法^[6]。当信息系统的属性为模糊集时,就要关联到模糊集与粗糙集结合的问题,这方面工作的理论和方法目前还很少,有待进一步的发展^[2,3]。de Korvin 在文[2]中给出了一种对属性具有类似模糊集形式的信息系统知识获取的方法,其属性的数据具有模糊集表示形式,但隶属度是由模糊方式导出的经典集合元素形式。这种数据形式在模糊集理论应用的很多领域都存在,因此对有关这类数据信息系统知识处理的研究具有实际意义。为了更好地处理此类数据本文提出了一种F-模糊集的概念,作为对模糊集的一种补充,给出了相关的运算。同时,给出了一种对属性为F-模糊集形式的信息系统知识获取的方法。

2 F-模糊集

文[2]中处理的是一个医学诊断的信息表(文[2]的表1),在表中的条件属性数据(c_1, c_2)和决策属性数据(d)有如下表示:

$$c_1 = \text{not likely}/L + \text{likely}/S$$

$$c_2 = \text{not likely}/H + \text{very likely}/P$$

$$d = \text{not likely}/DA + \text{somewhat likely}/DB$$

其中, L, S, H, P 表示一些特征, DA, DB 表示病症的类型。not likely, likely等表示又由多项真实因素与可信因素对确定,如:

$$\text{not likely} = 0.9/0.2 + 0.8/0.3 + 0.2/0.4$$

$$\text{likely} = 0.9/0.8 + 0.8/0.9 + 0.2/0.6$$

式中符号“/”两侧的数据表示真实因素与可信因素的隶属度值(在 $[0, 1]$ 范围取值),加号“+”模糊集表示的加号具有同样

意义。

这种类型的数据具有模糊集合的表示结构,但相应的隶属度表示为经典集合形式而不是在 $[0, 1]$ 范围取实数值的模糊集合隶属度形式。

对实际应用而言,这样的表示形式有两方面的不足。一个不足是数据丢失很多信息,如“not likely”等元素的确定是在主观性认定下的结论,而不是原始数据本身。另一个不足是导致在对信息系统处理时部分地采用对经典集合数据处理的方法,不能准确体现数据间的关系。

文[2]中要处理的数据实际是模糊集中隶属度为模糊集的情况。该文之所以对数据采取那样的表示形式是由于在现有的模糊集理论与方法中没有谈及关于此类数据的工具和方法。针对这样的问题,下面本文提出一种新的概念:F-模糊集,作为对模糊集的补充。

定义1 在论域 U 上给定映射 $A:U \rightarrow F$, F 为论域 V 上的模糊集合族,称 A 为F-模糊集。

例如对于论域 $U = \{x_1, x_2, \dots, x_n\}$,和论域 $V = \{y_1, y_2, \dots, y_m\}$,一个F-模糊集 A 可表示为 $A = A_1/x_1 + A_2/x_2 + \dots + A_n/x_n$ 。其中, $A_i = A_i(y_1)/y_1 + A_i(y_2)/y_2 + \dots + A_i(y_m)/y_m$ ($i = 1, 2, \dots, n$)。即每个 A_i 为论域 V 上的模糊集。下面即是一个F-模糊集合的具体例子:

$$A = (0.2/y_1 + 0.9/y_2 + 0.3/y_3)/x_1 + (0.8/y_1 + 0.6/y_2 + 0.5/y_3)/x_2 + (0.5/y_1 + 0.6/y_2 + 0.7/y_3)/x_3 + (0.3/y_1 + 0.2/y_2 + 0.8/y_3)/x_4$$

F-模糊集 A 也可形式地写为:

$$A = \sum_{i=1}^n \sum_{j=1}^m (A_i(y_j)/y_j)/x_i$$

记F-模糊集合的全体为 $F(U, V)$,对于 $F(U, V)$ 中的F-模糊集合的写法在本文以下内容中采用与相应模糊集合一致的方式。如F-模糊集 A ,对应的诸模糊集为 A_i, B 所对应的模

^{*} 辽宁师范大学校基金资助。闫德勤 博士,教授,主要研究领域为人工智能、图像处理等。迟忠先 教授,博士生导师,主要研究领域为知识发现、数据库、数据挖掘等。

模糊集为等 B_i 。其表示式为：

$$A = A_1/x_1 + A_2/x_2 + \dots + A_n/x_n$$

$$B = B_1/x_1 + B_2/x_2 + \dots + B_n/x_n$$

$$C = C_1/x_1 + C_2/x_2 + \dots + C_n/x_n$$

.....

以下对 F-模糊集定义几种基本的运算。

定义2 设 $A, B \in F(U, V)$ ，若对每一个 i 都有 $A_i \supseteq B_i$ ，则称 A 包含 B ，记作 $A \supseteq B$ 。

定义3 设 $A, B \in F(U, V)$ ，若对每一个 i 都有 $A_i = B_i$ ，则 A 称与 B 相等，记作 $A = B$ 。

定义4 设 $A, B, C \in F(U, V)$ ，若 $\forall y \in V$ 对每一个 i 都有 $C_i(y) = A_i(y) \vee B_i(y)$ 即 $C_i(y) = \max(A_i(y), B_i(y))$ ，则称 C 为 A 与 B 的并，记作 $C = A \cup B$ 。

定义5 设 $A, B, C \in F(U, V)$ ，若 $\forall y \in V$ 对每一个 i 都有 $C_i(y) = A_i(y) \wedge B_i(y)$ 即 $C_i(y) = \min(A_i(y), B_i(y))$ ，则称 C 为 A 与 B 的交，记作 $C = A \cap B$ 。

定义6 设 $A, B \in F(U, V)$ ，若对每一个 i 都有 $B_i = A_i^c$ ，则称 B 为 A 的补集，记作 $B = A^c$ 。

定义7 设 $A, B \in F(U, V)$ ，称 $A - B = A \cap B^c$ 为 A 与 B 的差集。

可见，关于 F-模糊集的运算都可转化为相应的模糊集的运算。因此，在这种定义下有关模糊集的其他一些运算都适用于 F-模糊集，以下给出关于 F-模糊集的几种运算性质，利用模糊集的运算性质可以推出，这里证明从略。

幂等律： $A \cup A = A, A \cap A = A$

交换律： $A \cup B = B \cup A, A \cap B = B \cap A$

结合律： $(A \cup B) \cup C = A \cup (B \cup C)$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

分配律： $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

吸收律： $(A \cup B) \cap A = A$

$$(A \cap B) \cup A = A$$

对偶律： $(A \cup B)^c = A^c \cap B^c$

$$(A \cap B)^c = A^c \cup B^c$$

研究集合的距离度量与相似度量对于理论发展和实际应用具有意义。这里以模糊集间的距离度量与相似度量为基础给出 F-模糊集间的距离度量与相似度量。

设 $A, B \in F(U, V)$ ，其距离可表示为：

$$D(A, B) = \frac{1}{n} \sum_{i=1}^n D(A_i, B_i)$$

其相似度可表示为：

$$S(A, B) = \frac{1}{n} \sum_{i=1}^n S(A_i, B_i)$$

其中， $D(A, B), S(A, B)$ 分别表示相应模糊集的距离和相似度。

3 F-模糊集属性信息系统的知识获取

模糊集间确定的包含和可能的包含反映着知识的关系。计算包含度的以下公式被应用于文[2]，在其它文中也有应用：

$$I(A \subset B) = \min_x \max\{1 - A(x), B(x)\} \quad (1)$$

$$J(A \# B) = \max_x \min\{A(x), B(x)\} \quad (2)$$

其中， A, B 为模糊集合， $A(x), B(x)$ 分别为 A, B 的隶属度函数。 $I(A \subset B)$ 表示 B 对 A 的确定的包含程度。 $J(A \# B)$ 表示 B 对 A 可能的包含程度。

对于 F-模糊集而言，相当于把模糊集的隶属度函数用另

一个模糊集代替，所以集间确定的包含度和可能的包含度对应于(1)式和(2)式应是一种模糊集表示。为此，对于 F-模糊集这里给出如下的确定的包含度和可能的包含度算式：

$$FI(A \subset B) = \sum_{j=1}^m \frac{\min \max\{1 - A_i(y_j), B_i(y_j)\}}{y_j} \quad (3)$$

$$FJ(A \# B) = \sum_{j=1}^m \frac{\max \min\{A_i(y_j), B_i(y_j)\}}{y_j} \quad (4)$$

其中， A, B 为 F-模糊集合，(3)、(4)式的等号右端为模糊集表示。

如果把 F-模糊集中的模糊集写为向量形式，则对于一个 F-模糊集合 A 有：

$$\begin{aligned} A &= \sum_{i=1}^n \sum_{j=1}^m (A_i(y_j)/y_j)/x_i \\ &= \sum_{i=1}^n (A_i(y_1), A_i(y_2), \dots, A_i(y_m))/x_i \\ &= (A_1(y_1), A_1(y_2), \dots, A_1(y_m))/x_1 + \\ &\quad (A_2(y_1), A_2(y_2), \dots, A_2(y_m))/x_2 + \\ &\quad \dots \dots \dots \\ &\quad (A_n(y_1), A_n(y_2), \dots, A_n(y_m))/x_n \end{aligned} \quad (5)$$

若把上式(即(5)式)进一步省略去 x_i ，则 F-模糊集合 A 可写为 $n \times m$ 阶矩阵式：

$$\tilde{A} = \begin{bmatrix} A_1(y_1) & A_1(y_2) & \dots & A_1(y_m) \\ A_2(y_1) & A_2(y_2) & \dots & A_2(y_m) \\ \dots & \dots & \dots & \dots \\ A_n(y_1) & A_n(y_2) & \dots & A_n(y_m) \end{bmatrix}$$

有了矩阵表示，F-模糊集合 A 也可写为矩阵与向量的积的形式：

$$A = X \tilde{A} Y^T$$

其中， $X = (x_1, x_2, \dots, x_n), Y = (y_1, y_2, \dots, y_m)$ 。 Y^T 为 Y 的转置向量。

例1 设 F-模糊集合为：

$$\begin{aligned} A &= (0.8/y_1 + 0.5/y_2 + 0.3/y_3)/x_1 + \\ &\quad (0.7/y_1 + 0.6/y_2 + 0.5/y_3)/x_2 + \\ &\quad (0.9/y_1 + 0.6/y_2 + 0.7/y_3)/x_3 + \\ &\quad (0.1/y_1 + 0.2/y_2 + 0.8/y_3)/x_4 + \\ &\quad (0.2/y_1 + 0.9/y_2 + 0.7/y_3)/x_5 \end{aligned}$$

则其向量形式为：

$$\begin{aligned} A &= (0.8, 0.5, 0.3)/x_1 + (0.7, 0.6, 0.5)x_2 \\ &\quad + (0.9, 0.6, 0.7)/x_3 + (0.1, 0.2, 0.8)/x_4 + (0.2, \\ &\quad 0.9, 0.7)/x_5 \end{aligned}$$

对应的矩阵为：

$$\tilde{A} = \begin{bmatrix} 0.8 & 0.5 & 0.3 \\ 0.7 & 0.6 & 0.5 \\ 0.9 & 0.6 & 0.7 \\ 0.1 & 0.2 & 0.8 \\ 0.2 & 0.9 & 0.7 \end{bmatrix}$$

F-模糊集的矩阵表示相当于把模糊集向量表示的一维扩展到二维。在一定意义上，F-模糊集即是二维模糊集合。

另一方面，在给出一定的运算法则下，利用矩阵运算可进一步研究与 F-模糊集相关的理论和应用。

表1给出的是一个医疗分析信息表， L, S 和 H, P 分别表示病例的特征， A 和 B 分别表示病症。其中的模糊集表示依照一定的衡量方式所得到的关于病例 $x_i (i=1, 2, \dots, 5)$ 对于有关特征和病症的一种模糊确定。这种模糊确定比结论性的指标更充分利用客观数据。

(BTP),基本处所短语(BNS),基本名词短语(BNP)和基本动词短语(BVP)。一种组块类型有两种边界标记,加上不属于任何组块类型的边界标记O,一共定义了15种汉语组块类型边界标记。

本文仅使用了词、词性信息。在将来的研究过程中,我们将结合语义、搭配、共现的知识,以取得更好的汉语组块识别效果。

参考文献

- 1 张昱琪,周强.汉语基本短语的自动识别.中文信息学报,2002,16(6):1~8.
- 2 朱德熙.语法讲义.商务印书馆,1982
- 3 Kudo T, Matsumoto Y. Chunking with Support Vector Machines. ACL,2001
- 4 周强,孙茂松,黄昌宁.汉语句子的组块分析体系.计算机学报,1999,22(11):1158~1165
- 5 赵军,黄昌宁.基于转换的汉语基本名词短语识别模型.中文信息学报,1998,13(2):1~7
- 6 Kudo T, Matsumoto Y. Use of Support Vector Learning for

- Chunk Identification, CoNLL,2000
- 7 Nakagawa T, Kudoh T, Matsumoto Y. Unknown word guessing and part-of-speech tagging using support vector machines. In: Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, 2001. 325~331
- 8 Yamada H, Kudoh T, Matsumoto Y. Japanese named entity extraction using support vector machines (in Japanese). In IPSJ SIG Notes NL-142~17, 2001
- 9 Joachims T. Learning to Classify Text using Support Vector Machines. Dissertation, Kluwer, 2002
- 10 Abney S. Parsing by chunks. In Principle-Based Parsing Kluwer Academic Publishers, 1991
- 11 Erik F, Sang T K, Buchholz S. Introduction to the CONLL-2000 Shared Task: Chunking[a], 2000. In: Proc. of CONLL-2000 and LLL-2000. 127~132
- 12 Vapnik V N, Chervonenkis A Y. Theory of Pattern Recognition. (in Russian) Nauka, Moscow, 1974
- 13 Introduction to Support Vector Machines. Dustin Boswell. Aug. 2002
- 14 Platt J C. Probabilities for SV machines. In: A. J. Smola, P. L. Bartlett, B. Scholkopf, D. Schuurmans, eds. Advances in Large Margin Classifiers, MIT Press, 2000. 61~71

(上接第133页)

表1 医疗分析信息表

Patients	Condition1	Condition2	Decision
x_1	$(0.2, 0.3, 0.2)/L + (0.8, 0.8, 0.2)/S$	$(0.4, 0.3, 0.2)/H + (0.9, 0.9, 0.7)/P$	$(0.3, 0.2, 0.2)/A + (0.9, 0.7, 0.6)/B$
x_2	$(0.3, 0.3, 0.2)/L + (0.9, 0.7, 0.7)/S$	$(0.4, 0.3, 0.1)/H + (0.8, 0.8, 0.2)/P$	$(0.8, 0.8, 0.2)/A + (0.9, 0.7, 0.6)/B$
x_3	$(0.8, 0.8, 0.3)/L + (0.3, 0.2, 0.2)/S$	$(0.9, 0.7, 0.5)/H + (0.9, 0.7, 0.6)/P$	$(0.9, 0.7, 0.4)/A + (0.8, 0.8, 0.3)/B$
x_4	$(0.9, 0.6, 0.7)/L + (0.9, 0.8, 0.5)/S$	$(0.1, 0.3, 0.2)/H + (0.9, 0.6, 0.8)/P$	$(0.9, 0.7, 0.6)/A + (0.3, 0.2, 0.2)/B$
x_5	$(0.2, 0.3, 0.3)/L + (0.9, 0.6, 0.8)/S$	$(0.1, 0.3, 0.2)/H + (0.9, 0.7, 0.7)/P$	$(0.1, 0.2, 0.2)/A + (0.4, 0.1, 0.2)/B$

这是一个F-模糊集属性信息系统表,在这个表中 Condition 1和 Condition2为条件属性, Decision 为决策属性。

本文以考察有关特征和病症的关系来说明关于F-模糊集属性信息系统的知识获取方法。

由表1可以得到:

$$L = (0.2, 0.3, 0.2)/x_1 + (0.3, 0.3, 0.2)/x_2 + (0.9, 0.6, 0.7)/x_3 + (0.9, 0.6, 0.7)/x_4 + (0.2, 0.3, 0.3)/x_5$$

$$S = (0.8, 0.8, 0.2)/x_1 + (0.9, 0.7, 0.7)/x_2 + (0.3, 0.2, 0.2)/x_3 + (0.9, 0.8, 0.5)/x_4 + (0.9, 0.6, 0.8)/x_5$$

$$H = (0.4, 0.3, 0.2)/x_1 + (0.4, 0.3, 0.1)/x_2 + (0.9, 0.7, 0.5)/x_3 + (0.1, 0.3, 0.2)/x_4 + (0.1, 0.3, 0.2)/x_5$$

$$P = (0.9, 0.9, 0.7)/x_1 + (0.8, 0.8, 0.2)/x_2 + (0.9, 0.7, 0.6)/x_3 + (0.9, 0.6, 0.8)/x_4 + (0.9, 0.7, 0.7)/x_5$$

$$A = (0.3, 0.2, 0.2)/x_1 + (0.8, 0.8, 0.2)/x_2 + (0.9, 0.7, 0.4)/x_3 + (0.9, 0.7, 0.6)/x_4 + (0.1, 0.2, 0.2)/x_5$$

$$B = (0.9, 0.7, 0.6)/x_1 + (0.9, 0.7, 0.6)/x_2 + (0.8, 0.8, 0.3)/x_3 + (0.3, 0.2, 0.2)/x_4 + (0.4, 0.1, 0.2)/x_5$$

这是一种病例对诸特征和病症的支持度的F-模糊集合表示。若要考察某些特征对病症的必然或可能的关联程度用式(3)和(4)即可得到。下面仅就特征L和H对于病症A必然和可能的关联程度进行计算。

$$L \cap H = (0.2, 0.3, 0.2)/x_1 + (0.3, 0.3, 0.1)/x_2 + (0.9, 0.6, 0.5)/x_3 + (0.1, 0.3, 0.2)/x_4 +$$

$$(0.1, 0.3, 0.2)/x_5$$

表示同时具有特征L和H的F-模糊集。利用式(3)和(4)可以得到:

$$FI((L \cap H) \subset A) = (0.8, 0.7, 0.5)$$

$$FJ((L \cap H) \# A) = (0.9, 0.8, 0.6)$$

结论 本文提出了F-模糊集的概念,给出了利用F-模糊集在信息系统中进行知识获取的方法。模糊集合是对经典集合的补充,解决了由经典集合难以表示和解决的问题。F-模糊集是对模糊集的补充,以解决在实际应用中直接利用模糊集难以解决的问题。文中对F-模糊集导出的矩阵表示对于理论研究及应用具有一定意义。利用矩阵表示和运算对F-模糊集属性信息系统规则约简和属性约简是本项目进一步研究的一个方面。

参考文献

- 1 Zadeh L A. Fuzzy sets. Inform. and Control, 1965, 8: 338~353
- 2 de Korvin A, McKeegan C. Knowledge acquisition using rough sets when membership values are fuzzy sets. J. Intelligent and Fuzzy Systems, 1998, 6: 237~244
- 3 Hong T P, et al. Learning a coverage set of maximally general fuzzy rules by rough sets. Experts Systems with Applications, 2000, 19: 97~103
- 4 曾黄麟. 粗集理论及其应用. 重庆: 重庆大学出版社, 1996
- 5 张文修. 粗糙集理论与方法. 北京: 科学出版社, 2001
- 6 闫德勤, 迟忠先. 一种实值属性信息系统的粗集约简方法. 小型微型计算机系统, 2003, 24(3): 517~519
- 7 张文修. 模糊数学引论. 西安: 西安交通大学出版社, 1991