基于互信息的模糊粗糙集属性约简

徐菲菲[①] 苗夺谦[①] 魏 莱[①] 冯琴荣[①] 毕玉升[②]
[①](同济大学计算机科学与技术系 上海 201804)
[②](同济大学经管学院 上海 200092)

摘 要：模糊粗糙集知识约简是模糊粗糙集理论的核心内容之一。该文从粗糙集知识熵出发，结合模糊集隶属度函数，将其应用于模糊环境下，推广了互信息的度量概念，使其能评价模糊决策表中属性的重要性。并给出了一种模糊决策表的启发式属性约简算法，通过实例验证了它的可行性，为模糊决策表的属性约简提供了一种有效的方法。

关键词：模糊粗糙集；属性约简；模糊决策表；互信息
中图分类号：TP181 **文献标识码**：A **文章编号**：1009-5896(2008)06-1372-04

Mutual Information-Based Algorithm for Fuzzy-Rough Attribute Reduction

Xu Fei-fei[①] Miao Duo-qian[①] Wei Lai[①] Feng Qin-rong[①] Bi Yu-sheng[②]
[①](*Department of Computer Science and Technology*, *Tongji University*, *Shanghai* 201804, *China*)
[②](*School of Economics and Management*, *Tongji University*, *Shanghai* 200092, *China*)

Abstract: Fuzzy-rough attribute reduction is one of the important topics in the research on fuzzy-rough set theory. In this paper, the information entropy is generalized in rough set so that it could be used to value the importance of attribute under fuzzy circumstance. A new heuristic algorithm based on mutual information for fuzzy-rough attribute reduction is introduced and illustrated with a simple example.

Key words: Fuzzy-rough set; Attribute reduction; Fuzzy decision table; Mutual information

1 引言

波兰科学家 Pawlak 于 1982 年提出的粗糙集理论[1]是一种处理不相容、不精确或不完全数据的强有力工具，已成功地应用于机器学习、模式识别、过程控制等各种领域。但是，经典的粗糙集理论处理的是符号值，是清晰的数据。而在实际生活中遇到的更多的都是模糊概念和模糊知识，所以经典粗糙集理论具有较大的局限性。为此，学者们对粗糙集理论进行了推广，提出了粗糙模糊集理论[2]和模糊粗糙集理论[3]，进一步拓宽了粗糙集理论的应用范围。

知识约简是粗糙集理论的核心内容之一，很多学者从不同的角度对其进行了研究[4, 5]。文献[4]中提出的基于互信息的决策表属性约简算法，是信息观下经典粗糙集知识约简理论的一种有效算法。此外，我们发现关于模糊粗糙集属性约简方法的研究[6, 7]相对于模糊粗糙集模型上、下近似算子的研究[2, 3, 8−10]是比较少的。为此本文在文献[4]的基础上，提出了模糊粗糙集的基于互信息的知识相对约简算法，可以看到其是经典粗糙集下基于互信息的决策表属性约简算法的推广，并且通过实例证明其有效性。

2006-11-27 收到，2007-05-21 改回
国家自然科学基金(60475019)和教育部博士点专项基金(20060247039)资助课题

2 基于互信息的粗糙集属性约简

为了下文的阐述，在这里对信息观下粗糙集理论及基于互信息的粗糙集属性约简做必要的介绍。

定义 1[11] 设 U 为一个论域，P,Q 为 U 上的两个等价关系(即知识)。P,Q 在 U 上导出的划分分别为 X,$Y: X=\{X_1, X_2, \cdots, X_n\}$，$Y=\{Y_1, Y_2, \cdots, Y_m\}$，则 P,Q 在 U 的子集组成的 σ^- 代数上定义的概率分布为

$$[X;p]=\begin{bmatrix} X_1 & X_2 & \cdots & X_n \\ p(X_1) & p(X_2) & \cdots & p(X_n) \end{bmatrix}$$

$$[Y;p]=\begin{bmatrix} Y_1 & Y_2 & \cdots & Y_m \\ p(Y_1) & p(Y_2) & \cdots & p(Y_m) \end{bmatrix}$$

其中 $p(X_i)=|X_i|/|U|, i=1,2,\cdots,n$ ； $p(Y_j)=|Y_j|/|U|$, $j=1,2,\cdots,m$；符号 $|E|$ 表示集合 E 的基数。则知识 P 的熵 $H(P)$ 定义为

$$H(P)=-\sum_{i=1}^{n} p(X_i)\log p(X_i) \tag{1}$$

知识 Q 相对于知识 P 的条件熵 $H(Q\mid P)$ 定义为

$$H(Q\mid P)=-\sum_{i=1}^{n} p(X_i)\sum_{j=1}^{m} p(Y_j\mid X_i)\log p(Y_j\mid X_i) \tag{2}$$

通过上述定义，文献[11]中将知识与信息熵联系起来并且证明了粗糙集理论的信息表示与原来的代数表示完全等

价。而且说明信息表示比代数表示更加直观，在信息表示下，能够导出高效的知识约简算法。

但是，为了能够进行有效的知识约简，必须要建立一个衡量属性重要性的标准。在粗糙集理论的信息观点下，文献[4]中提出在决策表中添加某个属性所引起的互信息的变化大小可以作为该属性重要性的度量。

设 $T=(U,C\bigcup D,V,f)$ 为一个决策表，且 $R\subseteq C$ 。那么，在 R 中添加一个属性 $a\in C$ 之后互信息的增量为：

$$I(R\bigcup\{a\};D)-I(R;D)=H(D\mid R)-H(D\mid R\bigcup\{a\}) \quad (3)$$

这里，$I(x;y)$ 表示 x 与 y 的互信息；$H(y\mid x)$ 表示已知 x 时，y 的条件熵。

定义 2 [4]　设 $T=(U,C\bigcup D,V,f)$ 是一个决策表，且 $R\subseteq C$ 。则对于任意属性 $a\in C-R$ 的重要性 $\mathrm{SGF}(a,R,D)$ 定义为

$$\begin{aligned}\mathrm{SGF}(a,R,D)&=I(R\bigcup\{a\};D)-I(R;D)\\&=H(D\mid R)-H(D\mid R\bigcup\{a\})\end{aligned} \quad (4)$$

若 $R=\varnothing$，则 $\mathrm{SGF}(a,R,D)$ 变为 $\mathrm{SGF}(a,D)=H(D)-H(D\mid a)=I(a;D)$ 即为属性 a 与决策 D 的互信息。$\mathrm{SGF}(a,R,D)$ 的值越大，说明在已知 R 的条件下，属性 a 对于决策 D 就越重要。

有了上述理论准备，文献[4]完整地提出基于互信息的知识相对约简(MIBARK)算法。它是以 bottom-up 的方式求相对约简的。它以决策表的相对核为起点，依据上述定义的属性重要性，逐次选择最重要的属性添加到相对核中，直到终止条件满足。

算法 1 MIBARK(Mutual Information-Based Algorithm for Reduction of Knowledge)

步骤 1　计算决策表 T 中条件属性 C 与决策属性 D 的互信息 $I(C;D)$；

步骤 2　计算 C 相对于 C 的核 $C_0=\mathrm{CORE}_D(C)$；一般来说，$I(C_0;D)<I(C;D)$；有时，相对核 $C_0=\varnothing$，此时 $I(C_0;D)=0$；

步骤 3　令 $B=C_0$，对条件属性集 $C-B$ 重复：

(1)对每个属性 $p\in C-B$，计算条件互信息 $I(p;D\mid B)$；

(2)选择使条件互信息 $I(p;D\mid B)$ 最大的属性，记作 p (若同时有多个属性达到最大值，则从中选取一个与 B 的属性值组合数最少的属性作为 p)；并且 $B\Leftarrow B\bigcup\{p\}$；

(3)若 $I(B;D)=I(C;D)$，则终止；否则，转(1)；

步骤 4　最后得到的 B 就是 C 相对于 D 的一个相对约简。

值得提出的是，MIBARK 算法的复杂度为 $O(M^2)$，而且它并不是一个完备的属性约简算法，但是在大多数情况下确实能够有效地得到决策表的最小约简。

3　基于互信息的模糊粗糙集属性约简

模糊粗糙集理论是对粗糙集理论的推广，它将粗糙集中讨论的对象集合拓展为模糊集，并且将等价关系 R，转换为模糊等价关系 $\Re$，扩大了粗糙集理论的应用范围，有着广泛的理论和应用价值。因此，模糊粗糙集的知识约简也就显得必要，但是这方面的研究却并不多。文献[7]给出一种基于属性依赖度的知识约简的方法，但该方法是以代数表示为基础的，运算的直观性较差。为此，本文通过模糊粗糙集的信息熵表示提出了模糊粗糙集知识约简的算法。

为了下面的讨论，先在经典粗糙集理论下对式(1)，式(2)改写一下。

假设论域 $U=\{x_1,x_2,\cdots,x_n\}$，P,Q 为 U 上的两个等价关系(即知识)。P,Q 在 U 上导出的划分分别为 $X,Y:X=\{X_1,X_2,\cdots,X_n\}$，$Y=\{Y_1,Y_2,\cdots,Y_m\}$。其中 $\forall X_i\in X,Y_j\in Y$ 都是可定义集合(crisp set)。引入模糊集中的隶属度函数，于是对于 $\forall X_i\in X$ 及 $x_k\in U$ 有

$$\mu_{X_i}(x_k)=\begin{cases}1, & x_k\in X_i\\ 0, & x_k\notin X_i\end{cases},\quad \mu_{Y_j}(x_k)=\begin{cases}1, & x_k\in Y_j\\ 0, & x_k\notin Y_j\end{cases}。$$

因此 $p(X_i)=\dfrac{\mid X_i\mid}{\mid U\mid}$ 亦即可以表示为 $p(X_i)=\dfrac{\sum_{k=1}^{\mid U\mid}\mu_{X_i}(x_k)}{\left|U\right|}$，$i=1,2,\cdots,n$；同理 $p(Y_j)=\dfrac{\sum_{k=1}^{\mid U\mid}\mu_{Y_j}(x_k)}{\left|U\right|}$，$j=1,2,\cdots,m$ 于是式(1)变成如下形式：

$$\begin{aligned}H(P)&=-\sum_{i=1}^{n}p(X_i)\log p(X_i)\\&=-\sum_{i=1}^{n}\frac{\sum_{k=1}^{\mid U\mid}\mu_{X_i}(x_k)}{\mid U\mid}\log\frac{\sum_{k=1}^{\mid U\mid}\mu_{X_i}(x_k)}{\mid U\mid}\end{aligned} \quad (5)$$

而式(2)可以表述为

$$\begin{aligned}H(Q\mid P)&=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j\mid X_i)\log p(Y_j\mid X_i)\\&=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}\frac{p(Y_j\cap X_i)}{p(X_i)}\log\frac{p(Y_j\cap X_i)}{p(X_i)}\\&=-\sum_{i=1}^{n}\frac{\sum_{k=1}^{\mid U\mid}\mu_{X_i}(x_k)}{\mid U\mid}\sum_{i=1}^{m}\frac{\sum_{k=1}^{\mid U\mid}\mu_{X_i\cap Y_j}(x_k)}{\sum_{k=1}^{\mid U\mid}\mu_{X_i}(x_k)}\log\frac{\sum_{k=1}^{\mid U\mid}\mu_{X_i\cap Y_j}(x_k)}{\sum_{k=1}^{\mid U\mid}\mu_{X_i}(x_k)}\end{aligned} \quad (6)$$

通过这样的形式转换，就可以将其应用到模糊粗糙集中。先定义一个模糊决策表。

定义 3　设论域 $U=\{x_1,x_2,\cdots,x_N\}$，模糊属性集 $\widetilde{A}$ 是由一族模糊属性 $\left\{\widetilde{A}^1,\widetilde{A}^2,\cdots,\widetilde{A}^M,\widetilde{A}^{M+1}\right\}$ 组成，其中 $D=\left\{\widetilde{A}^{M+1}\right\}$ 是模糊决策属性，其他为模糊条件属性 $C=\left\{\widetilde{A}^1,\widetilde{A}^2,\cdots,\widetilde{A}^M\right\}$。每一个模糊属性可以将论域划分成 p_j 个模糊等价类，即 $F(\widetilde{A}^j)=\left\{\widetilde{F}_1^j,\widetilde{F}_2^j,\cdots,\widetilde{F}_{p_j}^j\right\}(j=1,2,\cdots,M+1)$，其中 $\widetilde{F}_i^j(1\le$

$i \le p_j$) 为一模糊集。我们称由这样的论域与模糊属性集构成的信息系统 $S=(U,\widetilde{A})$ 为模糊决策表。

在上述模糊决策表下，定义模糊粗糙集的知识信息熵和知识条件熵。

定义 4 设模糊决策表 $S=(U,\widetilde{A})$，P,Q 为模糊属性构成的模糊等价关系(也即知识)，$U/\mathrm{IND}(P)=\{\widetilde{X}_1,\widetilde{X}_2,\cdots,\widetilde{X}_n\}$，$U/\mathrm{IND}(Q)=\{\widetilde{Y}_1,\widetilde{Y}_2,\cdots,\widetilde{Y}_m\}$，这里 $\forall\widetilde{X}_i\in U/\mathrm{IND}(P)$, $\widetilde{Y}_j\in U/\mathrm{IND}(Q)$ 都是论域 U 上的模糊集，则定义知识 P 的熵为

$$H(P)=-\sum_{i=1}^{n}p(\widetilde{X}_i)\log p(\widetilde{X}_i)=-\sum_{i=1}^{n}\frac{\sum_{k=1}^{|U|}\mu_{\widetilde{X}_i}(x_k)}{|U|}\log\frac{\sum_{k=1}^{|U|}\mu_{\widetilde{X}_i}(x_k)}{|U|}\tag{7}$$

知识 Q 相对于知识 P 的条件熵 $H(Q\mid P)$ 定义为

$$H(Q\mid P)=-\sum_{i=1}^{n}p(\widetilde{X}_i)\sum_{j=1}^{m}p(\widetilde{Y}_j\mid\widetilde{X}_i)\log p(\widetilde{Y}_j\mid\widetilde{X}_i)=-\sum_{i=1}^{n}\frac{\sum_{k=1}^{|U|}\mu_{\widetilde{X}_i}(x_k)}{|U|}\sum_{i=1}^{m}\frac{\sum_{k=1}^{|U|}\mu_{\widetilde{X}_i\cap\widetilde{Y}_j}(x_k)}{\sum_{k=1}^{|U|}\mu_{\widetilde{X}_i}(x_k)}\cdot\log\frac{\sum_{k=1}^{|U|}\mu_{\widetilde{X}_i\cap\widetilde{Y}_j}(x_k)}{\sum_{k=1}^{|U|}\mu_{\widetilde{X}_i}(x_k)}\tag{8}$$

其中 $U/\mathrm{IND}(P)=\otimes U/\mathrm{IND}\{\tilde{A}^j\},\tilde{A}^j\in P$，$U/\mathrm{IND}(Q)=\otimes U/\mathrm{IND}\{\tilde{A}^j\},\tilde{A}^j\in Q$。定义 $\widetilde{T}_1\otimes\widetilde{T}_2=\{\widetilde{X}\cap\widetilde{Y}:\forall\widetilde{X}\in\widetilde{T}_1,\forall\widetilde{Y}\in\widetilde{T}_2,\widetilde{X}\cap\widetilde{Y}\}\neq\varnothing$。此外，$\mu(\cdot)$ 为模糊集的隶属度函数，且 $\mu_{\widetilde{T}_1\cap\widetilde{T}_2\cap\cdots\cap\widetilde{T}_n}(x)=\min\{\mu_{\widetilde{T}_1}(x),\mu_{\widetilde{T}_2}(x),\cdots,\mu_{\widetilde{T}_n}(x)\}$，$\widetilde{T}_i$ 是 U 上的模糊集。

特别地，当模糊等价关系退化为经典等价关系时，$H(P)$ 即退化为经典粗糙集理论下知识 P 的信息熵，$H(Q\mid P)$ 也就退化为知识 Q 相对于知识 P 的条件熵 $H(Q\mid P)$。

有了这些定义后，我们将互信息的概念引入到模糊粗糙集中，用其来度量模糊决策表中模糊属性的相对重要性。

设模糊决策表 $S=(U,\widetilde{A})$，$\Re$ 是模糊条件属性集合。那么，在 $\Re$ 中添加一个模糊属性 $\tilde{A}^j$ 之后互信息的增量为

$$I(\text{侣}\cup\{\tilde{A}^j\};D)-I(\ ;D)=H(D\mid\text{侣})-H(D\mid\ \cup\{\tilde{A}^j\})\tag{9}$$

定义 5 设模糊决策表 $S=(U,\widetilde{A})$，$\Re$ 是模糊条件属性集合。则对于任意属性 $\tilde{A}^j\in C-\Re$ 的重要性 $\mathrm{SGF}(\tilde{A}^j,\Re,D)$ 定义为

$$\mathrm{SGF}(\tilde{A}^j,\text{侣},D)=I(\ \cup\{\tilde{A}^j\};D)-I(\ ;D)=H(D\mid\text{侣})-H(D\mid\ \cup\{\tilde{A}^j\})\tag{10}$$

若 $\Re=\varnothing$，则 $\mathrm{SGF}(\tilde{A}^j,\Re,D)$ 即为 $\mathrm{SGF}(\tilde{A}^j,D)=H(D)-H(D\mid\tilde{A}^j)=I(\tilde{A}^j;D)$ 即为模糊属性 $\tilde{A}^j$ 与模糊决策属性 D 的互信息。$\mathrm{SGF}(\tilde{A}^j,\Re,D)$ 的值越大，说明在已知 $\Re$ 的条件下，模糊属性 $\tilde{A}^j$ 对于模糊决策属性 D 就越重要。

有了以上的一些基本概念和原理后，我们正式提出基于互信息的模糊粗糙集知识相对约简(MIBAFRAR)算法。它同样是以 bottom-up 的方式求相对约简的，但以空集为起点，依据上述定义的属性重要性，逐次选择最重要的属性添加到集合中，直到终止条件满足。

算法 2 MIBAFRAR(Mutual Information-Based Algorithm for Fuzzy-Rough Attribute Reduction)：

步骤 1 计算模糊决策表中条件属性 C 与决策属性 D 的互信息 $I(C;D)$；

步骤 2 令 $\Re=\varnothing$，对条件属性集 $C-\Re$ 重复：

(1)对每个属性 $\tilde{A}^j\in C-\Re$，计算条件互信息 $I(\tilde{A}^j;D\mid\Re)$；

(2)选择使条件互信息 $I(\tilde{A}^j;D\mid\Re)$ 最大的属性，记作 $\tilde{A}^j$ (若同时有多个属性达到最大值，则从中选取一个相似类个数最少的属性作为 $\tilde{A}^j$)；并且 $\text{侣}\Leftarrow\ \cup\{\tilde{A}^j\}$；

(3)若 $I(C;D)=I(\Re;D)$，则终止；否则，转(1)；

步骤 3 最后得到的 $\Re$ 就是条件属性 C 相对于 D 的一个相对约简。

4 实例分析

为了考察 MIBAFRAR 算法的有效性，利用本文的 MIBAFRAR 算法对表 1 所示的模糊决策表进行约简。表 1 中前 3 个属性是条件属性，最后一个“Class”为决策属性。第一个条件属性“Temperature”有 3 个模糊等价类。

(1) 对该表，计算得 $I(C;D)=0.1480$。

(2) 令 $\Re=\varnothing$，对 $\forall\tilde{A}^j\in C-\Re$，计算条件互信息 $I(\tilde{A}^j;D\mid\Re)$：

$I(\tilde{A}^1;D)=0.0458$，$I(\tilde{A}^2;D)=0.1205$，$I(\tilde{A}^3;D)=-0.0024$。

可以看出，使条件互信息最大的属性为“Humidity”，所以更新后的 $\Re=\{\tilde{A}^2\}$，并且，新的 $I(\text{侣};D)=I(\{\tilde{A}^2\};\ )=0.1205$。

下一步，用同样的方法可求得，使 $I(\tilde{A}^j;D\mid\Re)$ 最大的属性为“Windy”，所以，更新后的 $\Re=\{\tilde{A}^2,\tilde{A}^3\}$；并且，新的 $I(\Re;D)=0.1472$。

此时，$|I(C;D)-I(\Re;D)|\le 10^{-3}$，我们可视为在允许的误差范围内，故程序终止。因此，属性集{Humidity，Windy}就是该模糊决策表的一个相对约简，约简后的模糊信息表为表 2。

表 1

	Temperature			Humidity		Windy		Class	
	Hot	Mild	Cool	High	Normal	Flase	Ture	Positive	Negative
1	0.9	0.1	0	0.8	0.2	0.7	0.4	0.4	0.7
2	0.8	0.2	0.1	0.9	0.2	0.1	0.8	0.3	0.7
3	0.9	0.1	0.1	0.9	0.1	0.9	0.1	0.8	0.3
4	0.1	0.9	0	0.6	0.5	0.8	0.3	0.6	0.5
5	0	0.1	0.9	0	0.1	0.8	0.2	0.9	0.2
6	0	0.2	0.9	0.1	0.9	0.1	0.9	0.3	0.8

表 2　约简后的模糊信息表

	Humidity		Windy		Class	
	High	Normal	Flase	Ture	Positive	Negative
1	0.8	0.2	0.7	0.4	0.4	0.7
2	0.9	0.2	0.1	0.8	0.3	0.7
3	0.9	0.1	0.9	0.1	0.8	0.3
4	0.6	0.5	0.8	0.3	0.6	0.5
5	0	0.1	0.8	0.2	0.9	0.2
6	0.1	0.9	0.1	0.9	0.3	0.8

5　结束语

粗糙集是数据分析的一种强有力的理论工具，然而它不能直接应用到模糊环境中。本文首先改写了粗糙集中知识熵与条件熵公式，由此推广到模糊数据下，从信息的角度描述了模糊知识。然后，在此基础上，提出了一种基于互信息的模糊知识相对约简的启发式算法。最后，通过实例分析表明，在多数情况下该算法能够得到模糊决策表的最小约简。

参 考 文 献

[1] Pawlak Z. Rough sets. *International Journal of Information and Computer Science*, 1982,11(5): 341–356.

[2] Banerjee M and Pal Sankar K. Roughness of a fuzzy set. *Information and Computer Science*, 1996, 93(3): 235–245.

[3] Dubois D and Prade H. Rough fuzzy sets and fuzzy rough sets. *Information and Computer Science*, 1990, 17(2): 191–209.

[4] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法. 计算机研究与发展, 1999, 36(6): 681-684.

[5] 叶东毅, 陈昭炯. 一个新的差别矩阵及其求核方法. 电子学报, 2002, 30(7): 1086–1088.

[6] Wang Xi Zhao, Ha Yan, and Chen De Gang. On the reduction of fuzzy rough sets. In: Proceeding of the Third International Conference on Machine Learning and Cybernetics[C], Guangzhou, 2005,18–21: 3175–3178.

[7] Jensen R and Shen Q. Fuzzy–rough sets for descriptive dimensionality reduction. Proc. 11th Internat. Conf. on Fuzzy Systems, Hawaii, 2002: 29–34.

[8] Tsang C C, Chen De Gang, Lee W T, and Yeung S. On the upper approximation of covering generalized rough sets. In: Proceeding of the Third International Conference on Machine Learning and Cybernetics [C], Shanghai, 2004, 26–29: 4200–4203.

[9] Yeung S, Chen De Gang, Tsang C C, and Lee W T T. On the generalization of fuzzy rough sets. *IEEE Trans. on Fuzzy System*, 2005, 13(3): 343–361.

[10] Wu Weizhi, Mi Jusheng, and Zhang Wenxiu. Generalized fuzzy rough sets. *Information Science*, 2003, 151(5): 263–282.

[11] 苗夺谦, 王珏. 粗集理论中概念与运算的信息表示. 软件学报, 1999, 2: 113–116.,

徐菲菲：女，1983 年生，博士生，研究方向为模式识别与智能系统、粗糙集理论、粒度计算.

苗夺谦：男，1961 年生，教授，博士生导师，研究方向为人工智能、模式识别、知识发现、粗糙集理论等.

魏　莱：男，1980 年生，博士生，研究方向为模式识别与智能系统、流形学习、模糊粗糙集、粒度计算.

冯琴荣：女，1972 年生，博士生，研究方向为计算机代数、人工智能、模式识别、粗糙集理论.

毕玉升：男，1982 年生，博士生，研究方向为金融数学.