模糊属性信息系统的规则约简

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116026)[2]

Rules Reduction for Fuzzy Attribute Information Systems

YAN De-Qin[1] CHI Zhong-Xian[2]

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

(Depertment of Computer, Dalian University of Science and Technology, Dalian 116026)[2]

Abstract A new method for rules reduction to fuzzy attribute information systems is presented in this paper. The method gives a technique to treat continue attribute problem and a practical program scheme. Besides, the paper develops fuzzy theory with the technique of probability.

Keywords Rules reduction, Fuzzy attribute, Rough sets, Fuzzy sets

1 引言

信息系统的规则约简是机器学习与数据挖掘的重要内容。在研究中，大多数算法都是以离散值为处理对象[1]，而在计算机研究与应用领域所处理的很多都是实值属性数据。实值属性信息系统没有明确的等价关系，也难以用属性值的完全相同来判定规则的一致性。因此，对实值属性信息系统的规则约简研究为研究领域所关注。模糊属性信息系统的属性值是限定在[0,1]上的实值数据，研究对模糊属性信息系统的规则约简对研究实值问题具有意义。

本文结合概率方法对模糊集进行了研究，给出了一些理论结果；并利用相关的研究结果结合粗糙集理论给出了针对模糊属性信息系统的规则约简方法；最后给出了实验结果。

2 粗糙集理论的基本概念

设 $U=\{x_1,x_2,\cdots,x_n\}$ 是一有限集，称为论域，R 是 U 上的一个等价关系，U/R 表示在 U 上导出的所有等价类；$[x]_R$ 表示包含元素 x 的 R 的等价类，$x\in U$。对任一集合 $X\subseteq U$

$R_-(X)=\{x\in U|[x]_R\subseteq X\}$

$R^-(X)=\{x\in U|[x]_R\cap X\neq\varnothing\}$($\varnothing$ 为空集)

分别称 $R_-(X)$ 与 $R^-(X)$ 为 X 的 R 下近似和 X 的 R 上近似。$BN_R(X)=R^-(X)-R_-(X)$ 称为 X 的 R 边界。

令 P 和 Q 为 U 中的等价关系，Q 的 P 正域(Positive Region)记为 $\mathrm{Pos}_P(Q)$，$\mathrm{Pos}_P(Q)=\bigcup_{(X\subseteq U/Q)}P_-(X)$ 指对于 U/P 的分类，U/Q 的正域是论域中所有通过分类 U/P 表达的知识能够确定地划入 U/Q 类的对象的集合。

一个知识库(或称为信息系统)就是一个关系系统 $K=(U,R)$，其中 U 为论域，R 是 U 上的一个等价关系族。$\cap R$(R 的所有等价关系的交)也是一个等价关系，记作 $IND(R)$。

定义2.1 令 R 为一等价关系族，且 $r\in R$，当 $IND(R)=IND(R-\{r\})$ 时，称 r 为 R 中可省略的，否则 r 为 R 中不可省略的。当对于任一 $r\in R$，若 r 为 R 中不可省略的，则 R 为独立的。

定义2.2 设 R 为一等价关系族，且 $r\in R$，如果 $P=R-\{r\}$ 是独立的，则 P 是 R 中的一个约简。

定义2.3 设 P 和 Q 都是等价关系族，如果 $\mathrm{POS}_{IND(P)}IND(Q)=\mathrm{POS}_{IND(P-R)}IND(Q)$，则称是 $R\subseteq P$ 上 Q_-可约去的；否则是 P 上 Q_-不可约去的。

规则约简即是约去等价关系中的冗余规则。

3 对非实值属性规则约简

对模糊属性信息系统的约简继承对非实值属性信息系统约简的思想。本节我们给出对非实值属性信息系统的规则约简的研究。

设信息系统 $T=(U,C,D,V,f)$ 中决策属性 D 有 M 个等价类，记为 $D_s(s=1,2,\cdots,M)$；T 的条件属性 C 中删去一列后余下的各行称为 U 的 $N-1$ 阶子向量。条件属性 C 中删去第 j 列后第 i 行 $N-1$ 阶子向量称为 $N-1$ 阶子元，记为 $\overline{U}_{ij}$(第 i 行第 j 列的属性值为 U_{ij})。由 D_s 对应的 C 中所有 $N-1$ 阶子元构成的集合记为 C_s。在 C 属性中若删去某元后使得系统不一致，则称该元为关键元(Key Elements)。由 D_s 对应的所有关键元构成的集合记为 C_{ik}。

定理3.1 如下关系成立：

$C_{ik}=\{U_{ij}|\overline{U}_{ij}\notin D_-(C_s)\}$

其中，$D_-(C_s)$ 为 C_s 的 D 下近似，D 为决策等价关系。

证明：对于 $\overline{U}_{ij}\in D_-(C_s)$，则相应的 $N-1$ 阶子元属于同一个决策等价类 C_s，因此若删去 U_{ij} 元后不会使系统不一致。而当 $\overline{U}_{ij}\notin D_-(C_s)$ 时，C_s 中相应的 $N-1$ 阶子元必属于另一个决策等价类，若删去 U_{ij} 元后会使系统不一致。证毕。

记由 C_s 构成的等价关系为 P，则由定理3.1得：

定理3.2 下式成立：

$\bigcup_{s=1}^{M}BN_P(C_s)=U-\mathrm{Pos}_P(D)$

其中，$BN_P(C_s)$ 为 $P^-(C_s)-P_-(C_s)$ 所对应的 U 中的元素。

证明：由定理3.1及正域的定义可证。

4 关于模糊集的概率方法

设在论域 $U=\{x_1,x_2,\cdots,x_n\}$ 上有模糊集合 $A_k(k=1,2,\cdots,m)$，相应的隶属度为 $\mu_{Ak}(x_i)(i=1,2,\cdots,n;k=1,2,\cdots,m)$。我们把 $\mu_{Ak}(x_i)$ 形式地看作关联 x_i 与 A_k 的概率。

定义4.1 设论域中元素 x 对模糊集合 A 的隶属度为 $\mu_A(x)$，形式地定义

$P_A(x)=\mu_A(x)$

为元素在 x 确定条件下模糊集合 A 实现的概率。

事实上对于任一 x_i 有：$P_{A_j}(x_i)\geqslant 0$，

闫德勤 博士，副教授，主要研究领域为模式识别。迟忠先 教授，博士生导师，主要研究领域为知识发现、数据仓库、数据挖掘等。

$$\sum_{j=1}^{m} P_{A_j}(x_i)=\sum_{j=1}^{m}\mu_{A_j}(x_i)=1$$

所有的隶属度总和为 $\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{A_j}(x_i)=n$。

令每个 A_k 对应的隶属度总和为：

$$Q_k=\sum_{i=1}^{n}\mu_{A_k}(x_i)。$$

根据概率的意义我们得到：

$$P(A_k)=\frac{1}{n}\sum_{i=1}^{n}\mu_{A_k}(x_i)=\frac{Q_k}{n} \tag{1}$$

$$P(x_i/A_k)=\frac{\mu_{A_k}(x_i)}{Q_k} \tag{2}$$

$$P(A_k/x_i)=\mu_{A_k}(x_i) \tag{3}$$

$(i=1,2,\cdots,n;k=1,2,\cdots,m)$

其中，P 表示概率，$P(A/B)$表示在事件 B 发生情况下事件 A 发生的概率。我们称上式(1)、(2)、(3)为模糊集的概率表示。

定理4.1 在模糊集的概率表示下有下式成立：

$$P(A_k/x_i)=\frac{P(x_i/A_k)P(A_k)}{\sum_{j=1}^{m}P(x_i/A_j)P(A_j)} \tag{4}$$

$(i=1,2,\cdots,n;k=1,2,\cdots,m)$。

证明：分别把(1)、(2)式代入(4)式右端得：

$$\frac{\frac{\mu_{A_k}(x_i)}{Q_j}\times\frac{Q_k}{n}}{\sum_{j=1}^{m}\frac{\mu_{A_j}(x_i)}{Q_j}\times\frac{Q_j}{n}}=\mu_{A_k}(x_i)=P(A_k/x_i)$$

即得到(4)式。从另一方面看，由于 m 个 A_k 集合在概率意义下具有(1)式的表达形式并且有 $\sum_{k=1}^{m}P(A_j)=1$，所以可以看作是对概率样本空间的一个划分。因此，(4)式符合贝叶斯(Bayes)公式。证毕。

由(4)式得：

$$P(A_k/x_i)=\frac{P(x_i/A_k)P(A_k)}{P(x_i)}$$

$$P(x_i)=\frac{P(x_i/A_k)P(A_k)}{P(A_k/x_i)}=\frac{\frac{\mu_{A_k}(x_i)}{Q_k}\times\frac{Q_k}{n}}{\mu_{A_k}(x_i)}=\frac{1}{n}$$

由此可见 x_i 的概率相当于 n 个元素中取一个的情况。因此可推出对任两元素 x_i,x_j 有：

$$P(x_i,j_j)=\frac{1}{C_n^2}=\frac{2}{n(n-1)}$$

其中，x_i,x_j 表示积事件。

定义4.2 对论域上的元素 x_i 定义：均值、方差、相关系数分别如下：

$$E(x_i)=\sum_{j=1}^{m}\mu_{A_j}(x_i)P(x_i)$$

$$D(x_i)=E\{[x_i-E(x_i)]^2\}$$

$$s(x_i,j_j)=\frac{E\{[x_i-E(x_i)][x_j-E(x_j)]\}}{\sqrt{D(x_i)D(x_j)}} \tag{5}$$

由定义4.2可以推出关于均值与方差相应的一系列性质，这里略去。下一节我们将利用相关系数确定模糊信息系统中规则间的等价性，从而利用粗集理论进行规则约简。

5 模糊属性规则的约简

设模糊属性信息系统有 n 个元素($U=\{x_1,x_2,\cdots,x_n\}$)。每个元素对应 m 个条件属性 $A_k(k=1,2,\cdots,m)$和一个决策属性 D，形成一个规则。x_i 对应的第 i 条规则第 j 个属性值为：

$$U_{ij}=\mu_{A_j}(x_i)$$

决策属性有 M 个等价类 $D_k(k=1,2,\cdots,M)$，每个等价类含有 S_k 个元素。

对模糊属性信息系统的规则约简首先是根据规则间的相关性找出相关度最大的等价类，然后依决策属性的等价类划分区域，在各区域中删去关于相关性最大的上近似集合对应的规则。

对模糊属性信息系统的规则约简方法如下：

(1)获取最大与最小相关度标记

For $L=1$ to M-1

$$k=\sum_{i=1}^{L}S_{i-1}+1 \quad (S_0=0)$$

For $i=k+1$ to $k+S_L$

利用(5)式分别计算第 i 规则与在 $L+1$到 M 子块中的第 j 个规则的相关系数。同时用 max 与 min 标记第 i 规则与第 j 个规则所在块的相关系数最大与最小值所对应的规则。

Next i

Next L

(2)根据等价类约简规则

根据标记把信息系统分为三个等价类，分别记为：

$R^-(U)=\{x_i|$标记全部为 max$\}$，

$R_-(U)=\{x_i|$标记全部 min$\}$，

$B(U)=\{x_i|$ 标记既含 max 又含 min$\}$。

删除 $R^-(U)$所对应的所有规则。对 $B(U)$中的约简设定一阈值，删除 max 标记比例超过阈值以上的规则。

6 实验结果

使用本文所给出的方法对 IRIS 数据进行了规则约简实验。IRIS 数据是著名的植物分类数据，共150个，分三类。其数据属性是实数型的，我们利用归一化的方法使其变为模糊表示，使其代表实验的对象为模糊数据，同时也说明该方法也适用于对实值属性信息系统的规则约简。IRIS 数据约简后的结果由表1给出。其中 U 表示记录，A 表示属性，D 表示所属类别。从实验结果看本文所给出的方法是有效可行的。

表1 约简结果

U	A_1	A_2	A_3	A_4	D
1	0.5	0.32	0.16	0.02	1
3	0.49	0.38	0.1	0.03	1
6	0.5	0.3	0.16	0.04	1
13	0.47	0.37	0.14	0.02	1
39	0.53	0.27	0.15	0.05	1
51	0.43	0.2	0.25	0.12	2
54	0.42	0.17	0.3	0.11	2
56	0.4	0.2	0.31	0.09	2
58	0.42	0.21	0.27	0.1	2
101	0.34	0.18	0.33	0.15	3
103	0.39	0.16	0.32	0.13	3
118	0.37	0.18	0.33	0.11	3

参考文献

1 Chen M S, Han J, Yu P S. Data Mining: An overview from a database perspective. IEEE Transaction on Knowledge and Data Engineering, 1996, 8(6):866～883

2 Ziarko W. Introduction to the special issue on rough sets and knowledge discovery. International Journal of Computational Intelligence, 1995, 11(2):223～226

3 曾黄麟. 粗集理论及其应用. 重庆：重庆大学出版社，1996

4 郭桂蓉，庄钊文. 信息处理中的模糊技术. 长沙：国防科技大学出版社，1996