第 39 卷 第 6 期
2005 年 6 月

西 安 交 通 大 学 学 报
JOURNAL OF XI′AN JIAOTONG UNIVERSITY

Vol. 39 №6
Jun. 2005

样本信息处理中一种属性约简方法的研究

夏克文，沈钧毅，李昌彪

（西安交通大学电子与信息工程学院，710049，西安）

摘要：为了剔除样本信息中存在的冗余成分和不相容性，同时提取关键信息等，根据样本信息的特点和信息具有粒度的思想，基于粗糙集的 2 个近似精度科学地定义了条件属性重要性，进而提出一种对样本信息进行属性约简的有效、简便方法. 该方法主要包括信息核的求取、可省条件属性的重要性计算和相对属性约简集的确定. 其中，为连续属性的离散化处理提供了一种基于模糊相似比原理的快速离散化算法，它能起到剔除模糊噪声的作用. 典型实例计算和在油水层识别系统中的实际应用表明，这种属性约简方法的识别准确率可达 90％以上，应用效果显著.

关键词：属性约简；样本信息；近似精度；连续属性离散化；模糊相似比

中图分类号：TP18　**文献标识码**：A　**文章编号**：0253-987X(2005)06-0558-04

Study on Attribute Reduction Method in Sample Information Processing

Xia Kewen，*Shen Junyi*，*Li Changbiao*

(School of Electronics and Information Engineering，Xi′an Jiaotong University，Xi′an 710049，China)

Abstract：In order to eliminate redundancies and incompatibilities in samples，extract the key information，and so on，the condition attribute significance was defined scientifically based on two approximation qualities of rough set according to the characteristics of sample information and the thought of information granularity. Moreover，a simple effective method of attribute reduction was presented，the main process was to determine the information core，calculate significance of dispensable condition attribute and obtain the relative attribute reduction set. For the continuous attribute processing，a speedy discretization algorithm was presented based on fuzzy similar-ratio principle which can eliminate the fuzzy noise. The typical example calculation and the actual application in oil-water formation recognition system show that discernment accuracy of the attribute reduction method can approach more than 90％，and the application effect is very notable.

Keywords：*attribute reduction*；*sample information*；*approximation quality*；*continuous attribute discretization*；*fuzzy similar-ratio*

一般来说，从数据库中选取的样本信息具有信息完备、决策属性的模式类型确定、条件属性大多为连续型等特点，但由于每个条件属性重要性不一，且常存在冗余，因而必须进行属性约简[1-5].

为了求解最佳属性约简，人们提出了不少的算法，然而 Wong 和 Ziarko 已经证明这种求解是一个 NP 难问题. 在实际应用中，其实只须求取相对属性约简，但在算法上要尽量做到快捷. 比如：Pawlak 等人[1]提出了先求核(core)，然后逐步扩展求出相对属性约简的思想；Jelonek 等人[3]提出了基于属性重要性程度的逐步扩展的算法. 这些算法均取得了较好的效果，但也存在不足，如当核不存在或其包含的

收稿日期：2004-07-21.　作者简介：夏克文(1965～)，男，高级工程师，博士后，硕士生导师；沈钧毅(联系人)，男，教授，博士生导师.　基金项目：国家自然科学基金资助项目(60173058)；中国石油天然气集团公司“九五”重点攻关资助项目(2001-6-1).

属性个数很少时，算法需要从空集（或相当于空集）开始逐个搜索，计算量较大. 另外，由于常用的 2 个属性重要性的定义不一，在约简中存在歧义[5]，虽然文献[4]从条件属性的集合和划分的近似精度出发，较合理地给出了一种衡量属性重要性程度的准则，但在应用中还存在一些问题. 为此，本文基于信息具有粒度的思想，对该准则做出修正，科学地定义了属性重要性，进而提出了一种有效的相对属性约简算法.

此外，连续属性离散化处理是属性约简中必不可少且最为关键的环节，如果区间划分得不合理，将直接影响应用效果. 目前，人们提出了很多离散化算法，如等宽区间法、等频区间法、归并法、划分法和基于信息熵的二元分割法等[6]，但这些算法有的过于简单，有的则太复杂，计算量较大，另外还未考虑模糊噪声的影响. 为此，本文根据样本信息特点，基于模糊相似比快速算法的原理，提出了一种快速离散化算法.

1　属性约简法的原理

1.1　基本概念[2-4]

（1）**信息系统**. 记信息系统 $L=(U,Q,V_q,F_q)$，$q\in Q$. 其中：U 是论域；Q 是属性集合，分为条件属性集 C 和决策属性集 D，$Q=C\cup D$，$C\cap D=\varnothing$；V_q 是属性取值的集合；F_q 是 $U\times Q\to V_q$ 的映射.

（2）**不可辨识关系**. 对于 $x,y\in U$，$P\subseteq Q$，如果满足 $\forall q\in P:F_q(x)=F_q(y)$，则称对象 x、y 对于属性集合 P 是不可辨识的，否则称 x、y 是可辨识的. 由 P 决定的不可辨识关系记为 $\mathrm{ind}(P)$.

（3）**集合的上近似、下近似**. 设 $P\subseteq Q$，$Y\subset U$，$x\in U$，$[x]_P=\{y\in U\mid x\ \mathrm{ind}(P)\ y\}$，定义集合 Y 的下近似为 $\underline{PY}=\{x\in U\mid [x]_P\subseteq Y\}$，上近似为 $\overline{PY}=\{x\in U\mid [x]_P\cap Y\neq\varnothing\}$.

（4）**近似精度**. 设 $P\subseteq Q$，L 为由决策属性 D 所决定的 U 的划分 $\{Y_1,Y_2,\cdots,Y_k\}$，$Y\in U$，则集合 Y 关于属性集 P 的近似精度

$$\alpha_P(Y)=\mathrm{card}(\underline{PY})/\mathrm{card}(\overline{PY})$$

式中：$\mathrm{card}(Y)$ 表示集合 Y 所含元素的个数.

划分 L 关于属性集 P 的近似精度

$$\gamma_P(L)=\sum_{i=1}^{k}\mathrm{card}(\underline{PY_i})/\mathrm{card}(U)$$

根据算法要求，定义 L 关于属性集 P 的粗糙近似精度

$$\alpha_P(L)=\sum_{i=1}^{k}\mathrm{card}(\underline{PY_i})/\sum_{i=1}^{k}\mathrm{card}(\overline{PY_i})$$

（5）**一般约简**. 设 $p\in P$，若 $\mathrm{ind}(P)=\mathrm{ind}(P-\{p\})$，称 p 是 P 中可省的（或可约简的）属性，否则称 p 是不可省的（或不可约简的）属性.

如果对 $p\in P$ 都是不可省的，称集合 P 是独立的，否则称集合 P 是相关的. 若 $Q\subseteq P$ 是独立的，且 $\mathrm{ind}(Q)=\mathrm{ind}(P)$，则称 Q 是 P 的一个简化. P 中所有不可省属性的集合称为 P 的核，记为 $\mathrm{core}(P)$.

（6）**相对属性约简**. 设 $P\subseteq Q$，$R\subseteq Q$ 且 $R\subseteq P$，若 $\gamma_P(L)=\gamma_R(L)$ 且 R 是 P 中满足该等式的最小集合，则称 R 为 P 的一个约简，记为 $\mathrm{red}(P)$. 由此看出，约简前后的划分 L，其近似精度是不变的.

1.2　基于属性重要性的属性约简算法

1.2.1　**属性重要性定义**　属性重要性主要是基于属性依赖度和信息熵这 2 种定义，文献[5]则通过 2 个反例说明了这 2 种定义是不完备的，并综合这 2 种属性重要性做了加权平均处理，虽然可取得较好的效果[5]，但其权值的确定全凭经验且不易操作.

我们知道，粗糙集理论的一个重要思想是，信息是具有粒度的，根据某个等价关系可以把论域划分为正域、负域和边界域，因此可以从粗糙集中的 2 个近似分类精度出发来定义属性重要性.

设条件属性集合 $C=\{c_1,c_2,\cdots,c_m\}$ 为有限集，决策属性集合为 D，由 D 决定的划分 L 为 $\{Y_1,Y_2,\cdots,Y_k\}$. 对每个条件属性计算 $k+2$ 个参数，即 $\alpha_{c_i}(Y_j)$，$j=1,2,\cdots,k$，以及 $\alpha_{c_i}(L)$ 和 $\gamma_{c_i}(L)$. 令 α_{c_i} 和 σ_{c_i} 分别为这 $k+2$ 个参数的均值和方差.

定义 1　属性 c_i 的重要性 $S_{c_i}=\alpha_{c_i}$.

定义 2　当属性重要性相同时，使 σ_{c_i} 最小的属性 c_i 的重要性为优.

由定义 1 和定义 2 可知：当 $k+2$ 个参数均为非 0 时，表明该属性对划分的各子集均有影响；当属性重要性相同时，$k+2$ 个参数的方差能体现出各自的差别.

1.2.2　**相对属性约简算法**　该算法步骤如下.

（1）利用一般约简方法，求得条件属性 C 的核 $\mathrm{core}(C)$.

（2）令 $R=\mathrm{core}(C)$，计算 $\gamma_R(L)$，若 $\gamma_R(L)<\gamma_C(L)$，令 $A=C-R$，下转（3），否则下转（6）.

（3）对每个属性 $a\in A$，计算属性重要性 S_A.

（4）选择 A 中使 S_A 最大的属性 a（出现多个属性重要性相同时，选择重要性为优的属性，如果属性重要性均为优，就任选一个属性），$R=R\cup\{a\}$，同时

在集合 A 中去掉属性 a.

(5)如果 $\gamma_R(L)<\gamma_C(L)$,上转(4),否则下转(6).

(6)输出 R,即 R 为条件属性 C 的一个相对约简.

显然,这种算法的计算复杂度为 $O(m^2)$,其中 m 为决策表中条件属性的个数.

2 连续属性离散化算法

2.1 模糊相似比快速算法描述

模糊相似比快速算法[7]主要是将排序问题转化为海明距离的计算,并由小到大排序,然后进行综合排序. 在应用中最为关键的 2 个环节描述如下.

(1)以海明距离确定相似关系. 设 X 为由某属性元素构成的样本集$\{x_1,x_2,\cdots,x_n\}$,x_0 为选定的固定样本元素. 计算海明距离 $D_i=|x_i-x_0|$,$i=1,2,\cdots,n$,并将其按由小到大排序. 序号越小的元素表示相似程度越高,反之相似程度低.

(2)固定样本 x_0 的选取. x_0 的选取直接影响分类结果,根据聚点性质,应将求出的聚点作为固定样本. 在实际处理中,选取相似程度高的元素集,并取均值为聚点(固定样本). 显然,这种均值处理能起到剔除模糊噪声的作用.

2.2 连续属性离散化优化算法

在样本信息中决策属性的类型数是确定的,因此可以按类型数分别对各种条件属性的值进行模糊聚类,进而将连续区间离散化,描述如下.

(1)决策属性的泛化. 令决策属性 $D=\{d\}$,$d=\{d_i=i,i=0,1,\cdots,k-1\}$.

(2)针对信息表的第 1 个条件属性 c_1,令 $X=\{c_{11},c_{21},\cdots,c_{n1}\}$为样本集.

(3)对应决策属性,将 X 划分为 k 个区间集合 X_i,$i=1,2,\cdots,k$,如果 k 个区间集合中不存在包含关系,下转(4)处理,如果区间集合中存在物理意义上的包含关系,可以作集合减法处理. 例如 $X_1\subset X_2$,可以按 $X_1=X_1$,$X_2=X_2-X_1$ 细分区间,若$(X_1\cup X_2)\subset X_3$,则 $X_1=X_1$,$X_2=X_2$,$X_3=X_3-(X_1\cup X_2)$. 显然,在出现包含关系的区间中,存在对应一些决策属性的多个区间,即 X 被划分为 k_1 个区间集合X_i,$i=1,2,\cdots,k_1$,$k_1>k$.

(4)求被划分的各区间集合的聚点,即计算各区间元素的平均值,而相邻 2 个聚点的平均值点为区间分割点 t_i,$i=1,2,\cdots,k-1(k_1-1)$(利用相似比原理不难证明). 这样,可得到如下形式的划分,即$(-\infty,t_1),(t_1,t_2),\cdots,(t_{k-2},t_{k-1}),(t_{k-1},\infty)$或$(-\infty,t_1),(t_1,t_2),\cdots,(t_{k_1-2},t_{k_1-1}),(t_{k_1-1},\infty)$. 在样本点数较少时,若分割点不在两相邻区间之间,则可令两相邻区间之间的中点为分割点.

(5)按决策属性泛化值 $d=\{d_i=i,\ i=0,1,\cdots,k-1\}$,将区间相应元素泛化为 $0\sim k-1$ 的自然数.

(6)对于信息表中其他条件属性 $c_2\sim c_m$,按照上述(2)~(5)处理.

3 应用举例

3.1 约简算法应用与对比

表 1 为,在油水层模式识别系统中某地区以地层为样品点的信息,其中条件属性 c_1、c_2、c_3 和 c_4 分别表示自然伽马位比、孔隙度、侵入系数和含油饱和度,$C=\{c_1,c_2,c_3,c_4\}$.

表 1 样本信息表

样本号	c_1	c_2	c_3	c_4	D
1	0.200	0.33	4.30	83.4	油层
2	0.138	0.21	0.75	72.8	油层
3	0.346	0.27	1.82	62.1	油层
4	0.359	0.21	1.37	61.9	油层
5	0.480	0.18	1.20	60.0	油水层
6	0.520	0.27	3.00	58.1	油水层
7	0.420	0.14	0.80	54.7	油水层
8	0.620	0.24	6.20	54.3	水层
9	0.500	0.27	1.46	52.0	水层
10	0.560	0.20	3.00	22.0	水层

令决策属性 $D=\{d\}$,$d=\{d_i=i,i=0,1,2\}$,其中 0、1、2 分别代表油层、油水层、水层.

首先对连续属性 c_1 进行离散化,步骤如下.

(1)按对应于元素 $d_i(i=0,1,2)$的大小划分的区间集合为[0.138,0.359]、[0.42,0.52]和[0.50,0.62].

(2)计算每个区间内的均值点(聚点),得到 $p_1=0.261$、$p_2=0.473$ 和 $p_3=0.56$.

(3)计算分割点,得到离散化区间$(-\infty,0.367\ 0],(0.367\ 0,0.516\ 5],(0.516\ 5,\infty)$. 同样,对于属性 c_2、c_3、c_4,得到离散化区间分别为

$(-\infty,0.217],(0.217,0.246],(0.246,\infty)(-\infty,1.118\ 75],(1.118\ 75,2.793\ 50],(2.793\ 50,\infty)$

$(-\infty,54.50],(54.50,60.95],(60.95,\infty)$

这样，就得到如表 2 所示的离散化后的样本信息.

表 2　离散化后的样本信息表

样本号	c_1	c_2	c_3	c_4	D
1	0	2	2	2	0
2	0	0	0	2	0
3	0	2	1	2	0
4	0	0	1	2	0
5	1	0	1	1	1
6	2	2	2	1	1
7	1	0	0	1	1
8	2	1	2	0	2
9	1	2	1	0	2
10	2	0	2	0	2

对条件属性集 C 的不可分辨关系 $\mathrm{ind}(C)$，有等价类 $U/\mathrm{ind}(C)=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\},\{9\},\{10\}\}$，因为 $U/\mathrm{ind}(C-\{c_i\})=U/\mathrm{ind}(R)$，$i=1,4$，$U/\mathrm{ind}(C-\{c_2\})=\{\{1\},\{2\},\{3,4\},\{5\},\{6\},\{7\},\{8,10\},\{9\}\}$，$U/\mathrm{ind}(C-\{c_3\})=\{\{1,3\},\{2\},\{4\},\{5,7\},\{6\},\{8\},\{9\},\{10\}\}$，则得 $\mathrm{core}(C)=\bigcap \mathrm{red}(C)=\bigcap\{\mathrm{ind}(C-\{c_1\}),\mathrm{ind}(C-\{c_4\})\}=\{c_2,c_3\}$.

按照一般约简算法也能得到 $\{c_1,c_2,c_3\}$ 和 $\{c_2,c_3,c_4\}$ 这 2 个约简，但并不知道哪一个约简效果最好，因此还须通过属性重要性来分辨.

下面计算可省略属性 c_1、c_4 的重要性. 为了与文献[4]的算法进行对比，将所有属性重要性的计算列于表 3，其中划分

$$L=\{Y_1,Y_2,Y_3\}=\{\{1,2,3,4\},\{5,6,7\},\{8,9,10\}\}$$

表 3　属性重要性表

参数	c_1	c_2	c_3	c_4
$\alpha_{c_i}(Y_1)$	1.000 0	0.000 0	0	1
$\alpha_{c_i}(Y_2)$	0.000 0	0.000 0	0	1
$\alpha_{c_i}(Y_3)$	0.000 0	0.100 0	0	1
$\alpha_{c_i}(L)$	0.250 0	0.035 7	0	1
$\gamma_{c_i}(L)$	0.400 0	0.100 0	0	1
S_{c_i}	0.330 0	0.047 1	0	1
σ_{c_i}	0.135 6	0.002 0	0	0

而划分 L 关于属性集 C 和核 $\mathrm{core}(C)$ 的近似精度分别为 $\gamma_C(L)=1.0$ 和 $\gamma_{\mathrm{core}(C)}(L)=0.2$.

利用本文算法，选择重要性最大的属性 c_4 与核共同组成集合 $R=\{c_2,c_3,c_4\}$，计算得 $\gamma_R(L)=1.0$，即有 $\gamma_R(L)=\gamma_C(L)$，因此 R 为相对约简属性表. 另外，从模式识别角度看，显然属性 c_4 比属性 c_1 优越.

如果利用文献[4]中的算法，最后得到的约简表是 $\{c_4,c_1,c_2,c_3\}$. 对比上述结果，显然文献[4]算法没有得到相对约简表，并且也没有起到真正的属性约简的作用.

3.2　在模式识别中的应用与对比

在油水层模式识别中，先对样本信息进行属性约简，再利用神经网络方法进行学习训练和识别. 现将 c_2、c_3、c_4 这 3 个属性的 10 个层位样点作为神经网络训练样本集，以 c_2、c_3、c_4 作为输入，以油层、油水层、水层作为输出，来建立 3×3 的非线性连接权的神经网络结构[8]. 对样本进行学习训练，再用训练好的网络对某地区的另外 10 个未知油层进行识别，识别结果如表 4 所示.

表 4　不同方法对未知油层的识别结果与试油结论对比

油层编号	试油结论	识别结果					
		M1	M2	M3	M4	M5	M6
1	油	√	√	√	√	√	√
2	水	√	√	√	√	√	√
3	油水	油	√	√	油	√	√
4	水	油水	油水	油水	油水	油水	√
5	油	油水	√	油水	√	√	√
6	油水	√	√	√	油	√	√
7	油水	油	√	√	√	√	√
8	油	油水	√	油水	√	√	√
9	水	油水	油水	油水	油水	油水	油水
10	油	√	√	√	√	√	√

注：√表示归类正确；M1 表示贝叶斯线性判别法；M2 表示贝叶斯二次判别法；M3 表示费歇线性判别法；M4 表示降维法；M5 表示神经网络法；M6 表示本文方法.

与试油结论对比可知，贝叶斯二次判别法和神经网络法的识别准确率可达 80%，而本文中的基于样本信息属性约简的神经网络方法，其识别准确率能高达 90%.

（下转第 602 页）

(a)一辆车 (b)两辆车 (c)多辆车 (d)大车

(e)图 6a 的处理结果 (f)图 6b 的处理结果 (g)图 6c 的处理结果 (h)图 6d 的处理结果

图 6 原始图像及用本文消除阴影方法处理后的图像

参考文献:

[1] Cucchiara R, Grana C, Piccardi M, et al. Detecting moving objects, ghosts, and shadows in video streams [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(10):1 337-1 342.

[2] Cucchiara R, Grana C, Piccardi M, et al. Improving shadow suppression in moving object detection with HSV color information [A]. IEEE Transportation Systems Conference Proceedings, Oakland, USA, 2001.

[3] Fung G S, Yung N H, Grantham K H, et al. Effective moving cast shadow detection for monocular color traffic image sequences [J]. Optical Engineering, 2002, 41(6):1 425-1 440.

[4] Yoneyama A, Yeh C H, Kuo C J. Moving cast shadow elimination for robust vehicle extraction based on 2D joint vehicle [A]. The IEEE Conference on Advanced Video and Signal Based Surveillance, Miami, USA, 2003.

[5] 刘直芳,游志胜,徐 欣,等. 基于阴影轮廓差分投影方法的快速定位车体算法 [J]. 四川大学学报,2003,40(4):662-666.

(编辑 刘 杨)

(上接第 561 页)

4 结束语

属性约简是粗糙集理论研究中的一个核心内容,在信息处理中能起到很好的作用. 对样本信息进行属性约简有利于信息处理系统的优化和快捷、有效的应用. 本文根据信息具有粒度性的特点,用粗糙集近似精度来定义属性重要性,进而提出了信息核求取与属性重要性计算相结合的属性约简算法. 典例表明,该算法是切实可行的,优于一般属性约简算法和文献[4]的算法. 实际油水层识别表明,采用属性约简后的样本信息,输入神经网络经训练学习,能够很好地进行模式识别,且效果显著. 本文针对于连续属性的离散化处理提出的基于模糊相似比的快速离散化算法简便、实用,且能有效剔除模糊噪声.

参考文献:

[1] Pawlak Z, Wong S K M, Ziarko W. Rough sets: probabilistic versus deterministic approach [J]. International Journal of Man-Machine Studies, 1988, 29 (1): 81-95.

[2] Pawlak Z, Slowinski R. Rough set approach to multi-attribute decision analysis [J]. European Journal of Operational Research, 1994, 72(3):443-459.

[3] Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domains for neural networks [J]. International Journal of Computational Intelligence, 1995, 11(2):339-347.

[4] 叶东毅,黄翠微,赵 斌. 基于逼近精度的一个粗糙集属性约简算法 [J]. 福州大学学报,2000,28(1):7-10.

[5] 石 峰,娄臻亮,张永清. 一种改进的粗糙集属性约简启发式算法 [J]. 上海交通大学学报,2002,36(4):478-481.

[6] 韩秋明,赵轶群. Rough Set 中基于聚类的连续属性离散化方法 [J]. 计算机工程,2003,29(4):81-87.

[7] 夏克文. 模糊相似比方法的改进 [J]. 煤田地质与勘探, 1994,22(5):59-60.

[8] 夏克文,宋建平. 应用带有非线性连接权的神经网络识别水泥胶结质量 [J]. 西安交通大学学报,2003,37(2):192-195.

(编辑 苗 凌)