基于决策倾向度的样本过滤与主动选择

陈 科*，唐雪飞

(1. 四川大学锦城学院计算机与软件学院 成都 611731；2. 电子科技大学信息与软件工程学院 成都 610054)

【摘要】该文提出了基于过滤函数的粗糙集样本决策倾向度动态调节与主动选择方法。首先定义样本过滤函数，从而确定样本选择或丢弃的依据；然后依次添加新增样本，根据过滤函数决定样本的去留，同时根据阈值指标调节已有样本的决策倾向度；最终建立有效的决策样本库，并在此基础上进行属性约简。本方法克服了传统变精度方法实现过程复杂、计算量大的问题，可有效地去除噪声数据，提高系统的鲁棒性。数据实验结果表明，该方法可以有效地压缩数据，提高样本分析质量。

关 键 词 属性约简； 决策倾向度； 过滤函数； 粗糙集

中图分类号 TP181 **文献标志码** A doi:10.3969/j.issn.1001-0548.2019.03.019

Active Sample Selection Method Based on Decision Making Tendency

CHEN Ke* and TANG Xue-fei

(1. School of Computer and Software, Jincheng College of Sichuan University Chengdu 611731;
2. School of Information and Software Engineering, University of Electronic Science and Technology of China Chengdu 610054)

Abstract A dynamic adjustment and active selection method for rough set decision making based on filtering function is proposed. Firstly, a sample filtering function is defined to determine the basis for sample selection or discarding; then, new samples are added in turn to determine the retention of samples according to the filtering function, and the decision-making tendency of existing samples is adjusted according to the threshold; finally, new sample library is established and attribute reduction is carried out. This method overcomes the problems of complex implementation process and large amount of calculation in traditional variable precision methods, and can effectively remove noise data and improve the robustness of the system. Experimental results show that this method can effectively compress data and improve the quality of sample analysis.

Key words attribute reduction; decision-making tendency; filter function; rough set

文献[1]在1982年提出的粗糙集理论在人工智能、数据挖掘等前沿领域都有着重要的应用价值。属性约简在粗糙集理论中占有核心地位，目前常用的方法主要是采用启发式算法的属性重要度计算[2-5]并基于此的属性约简。文献[6-9]充分考虑样本结构，引入了辨识矩阵方法，通过构造辨识函数，计算极小元素，获得决策表的约简结果，比之前的代数方法具有更高的约简精度。

无论采用何种算法，粗糙集属性约简面临两大难题：1) 属性约简对噪声非常敏感，噪声数据不仅增加了计算量，更导致约简结果的偏差；2) 约简过程是NP难问题，面对海量数据集，需要付出指数级的计算代价。为了提高抗噪声能力，文献[10-11]提出了变精度粗糙集模型，引入β-上、下近似，其中$\beta \in (0.5, 1]$是包含度阈值。变精度粗糙集能够容忍数据中的噪声，从而在一定程度上具有鲁棒性。变精度模型的最大问题是比经典粗糙集更庞大的计算量，处理大规模数据集的效率非常低下。为了提高粗糙集属性约简的效率，文献[12-15]基于辨识矩阵，提出了基于主动样本选择的粗糙集属性约简增量算法。通过设计主动样本选择机制，将加入样本归为相对于当前数据集为有用和无用样本。无用样本被永久性地过滤，从而在增量计算的过程中不予考虑；有用样本将被选择来执行增量计算，确定有信息量的属性需要加入当前约简，从而提高属性约简速度。文献[16]详细给出了增量属性约简的算法过程。

增量算法的最大问题在于大量样本在添加过程中被判定为“无用”样本被丢弃掉了，而事实上这

收稿日期：2018－05－21；修回日期：2018－11－15
基金项目：四川省重点研发项目(2017GZ0192)
作者简介：陈科(1981－)，男，博士，主要从事计算机软件体系架构、粗糙集论与应用方面的研究. E-mail: schchenke@163.com

些样本对决策有重要影响；另一方面，一些噪声数据却被当作有用样本参与属性约简计算，导致计算结果的偏差。为了解决以上问题，基于样本增量过程，引入过滤函数和决策倾向度指标，在提高属性约简速度的同时，提高系统的抗噪性能。

1 基本定义

令 $U=\{x_1,x_2,\cdots,x_n\}$ 为一个非空有限论域，$A=\{a_1,a_2,\cdots,a_m\}$ 是条件属性集合，$D=\{d\}$ 是决策属性集合，则 $(U,A\cup D)$ 称为一个决策表。函数 $a(x_i)$ 为 x_i 的全部属性值集合，$d(x_i)=d_i$ 是数据 x_i 对应的决策，对任意 $x\in U$，其等价类 $[x]=\{y\in U:(x,y)\in U\times U,a(x)=a(y)\}$。

$\forall x_i,x_j\in U$，$x_i\neq x_j$，如果 $a(x_i)=a(x_j)$ 且 $d(x_i)=d(x_j)$，则称样本 x_i 和 x_j 是一致的。

对于任意的 $B\subseteq A$，定义 B 的等价关系 $\mathrm{IND}(B)=\{(x,y)\in U:a(x)=a(y),\forall a\in B\}$。$\mathrm{IND}(B)$ 将论域 U 划分为等价类簇 $U/\mathrm{IND}(B)=\{[x]_B:x\in U\}$。设 $U/\mathrm{IND}(D)=\{D_1,D_2,\cdots,D_r\}$，则 D_i 的 B 下近似定义为：$\underline{B}D_i=\bigcup\{[x]_B:[x]_B\cap D_i\neq\varnothing\}$，正域 $\mathrm{POS}_U(B,D)=\bigcup_{k=1}^{r}\underline{B}D_k$。

属性子集 B 是 A 的一个约简(记作 $\mathrm{RED}(A)$)，当且仅当以下条件满足：

$$\begin{cases}\mathrm{POS}_U(B,D)=\mathrm{POS}_U(A,D)\\ \mathrm{POS}_U(B-\{a\},D)\neq\mathrm{POS}_U(A,D)\ \ \forall a\in B\end{cases}$$

所有约简结果的交集称为核属性：$\mathrm{CORE}(A)=\bigcap\mathrm{RED}(A)$。

辨识矩阵 $\boldsymbol{M}$ 是由所有 $t(x,y)$ 组成的集合矩阵，其中 $t(x,y)$ 的定义为：

$$t(x,y)=\begin{cases}\{a:a(x)\neq a(y)\} & (x,y)\in\Gamma\\ \varnothing & \text{其他}\end{cases}$$

其中：

$$\Gamma=\{(x_i,x_j)\in U\times U\}:$$
$$\begin{cases}x_i\in\mathrm{POS}_U(A,D),x_j\notin\mathrm{POS}_U(A,D)\\ x_i\notin\mathrm{POS}_U(A,D),x_j\in\mathrm{POS}_U(A,D)\\ x_i,x_j\in\mathrm{POS}_U(A,D),d(x_i)\neq d(x_j)\end{cases}$$

$t(x,y)$ 是区分 x 和 y 的辨识属性集合。显然矩阵 $\boldsymbol{M}$ 是对称的。若不存在 $t(x_i,x_j)\in\boldsymbol{M}$ 是 $t(x,y)$ 的真子集，则称 $t(x,y)$ 是 $\boldsymbol{M}$ 的一个极小元素。文献[6]证明，粗糙集约简结果是区分 Γ 中所有样本对的极小属性子集。

2 样本主动选择过程

2.1 样本增量过程

样本增量，意味着样本集并非一次性准备好，而是逐步添加的过程。样本主动选择，是指系统将根据规则自动筛选有用的样本，淘汰无用的样本，并根据有用样本计算属性约简。

在样本增量添加过程中，可能出现4种情况(设 x 是一个新增样本)：

1) 新样本 x 与已有样本库中某些不一致样本有着相同的属性值。

2) 新样本 x 与已有样本库中某些一致样本有相同的属性和决策。

3) 新样本 x 与已有样本库中一致样本有相同的属性值，但决策值不同。

4) 新样本 x 与已有样本库中任意样本都有着不同的属性值。

文献[16]在进行增量约简计算时，将情况1)和情况2)作为无用样本直接丢弃，情况3)和4)作为有用样本参与运算。然而，情况1)和2)虽然表面上对属性约简没有影响，事实上这些新样本改变了已有样本库中同属性值样本的出现频率，而被当作有用样本的情况3)，也可能是噪声数据本应当被去除。

为了解决以上问题，本文引入新样本过滤函数，以及已有样本的决策倾向度指标，保持增量约简算法高效的同时，提高系统的鲁棒性。

2.2 决策倾向度指标及过滤函数

$\forall x_i\in U$，定义决策倾向度 $\omega(x_i)=\omega_i\in(0,1)$，所有决策倾向度值集合 $\Omega=\{\omega(x_i),x_i\in U\}$。因此，决策表可扩展为 $(U,A\cup D\cup\Omega)$。

在相同属性条件下，决策倾向度值越大，越倾向于采用该决策，反之，决策倾向度值越小，对决策影响越弱，当某个 ω 值低于阈值 β 时，将被当作噪声剔除。

定义初始倾向度数据集 $\Omega_0=\{\omega_0(a,d):\forall a\in A,d\in D\}$。即为每一组属性及对应决策值定义一个先验值 ω_0，其取值规则可根据不同样本的概率分布进行预置，比如对于最简单的均匀分布，可设置 $\forall x_i\in U,\omega_0(x_i)=1/|U|$。

随着新样本 x 的到来，$\omega(x_i)$ 的值会进行更新，新值的计算结果由样本主动选择过程及过滤函数 $f(x)$ 决定。

设已有样本库为 U，新样本 x 到来后，新样本库为 $U'=U\cup\{x\}$。

$\forall \chi \in U'$，其决策倾向度 $\omega(\chi)$ 规则如下：

$$\omega(\chi)=\begin{cases}\omega_0(a(x),d(x)) & \chi=x且\{y\in U:a(y)=a(x),d(y)=d(x)\}=\varnothing \ ①\\ \omega'(y)=\omega(x)=f(\omega(y))=\omega(y)+\Delta^+(\cdot) & \chi\in\{x\}\bigcup\{y\in U:a(y)=a(x),d(y)=d(x)\} \ ②\\ \omega'(y)=f(\omega(y))=\omega(y)-\Delta^-(\cdot) & \chi\in\{y\in U:a(y)=a(x),d(y)\neq d(x)\} \ ③\end{cases}$$

式①～③指定了新样本 x 到来时，样本主动选择过程：

1) x 是一个全新的样本，符合情况4)，即样本库 U 中不存在与 x 具有相等属性和决策值的样本，则直接将 x 添加到样本库，并取 $\omega(x)=\omega_0(a(x),d(x))$。

2) 样本库 U 中存在与 x 属性和决策相等的样本 y，符合情况2)，当前 y 的决策倾向度为 $\omega(y)$。由于 x 和 y 样本是一致的，因此 x 和 y 具有相等的决策倾向度，x 无需加入到样本库中，并且这些样本的决策倾向度将得到增强：$\omega'(y)=\omega(x)=f(\omega(y))=\omega(y)+\Delta^+(\cdot)$。其中 $\Delta^+(\cdot)\in(0,1)$ 是决策增强函数。对于均匀分布，可取 $\Delta^+(\cdot)=1/|U|$。

3) 样本库 U 中存在与 x 属性相等，但决策值不同的样本 y，符合情况3)。此时新样本 x 的决策倾向度 $\omega(x)$ 按照式①计算，而样本 y 则受 x 到来的影响，决策倾向度将降低，降低程度由函数 $\Delta^-(\cdot)$ 确定，即：$\omega'(y)=f(\omega(y))=\omega(y)-\Delta^-(\cdot)$。对于均匀分布，可取 $\Delta^-(\cdot)=\omega(y)/|U|$，即 $\omega'(y)=f(\omega(y))=\omega(y)\times(1-1/|U|)$。

4) 对于情况1)，事实上是式②和式③同时存在的状态，由于样本库中已经存在同属性值的样本，新样本 x 不会加入到样本库中，但式②会增强样本库中那些与 x 同决策值的 ω，同时削弱那些与 x 不同决策的样本 ω 值。

5) 过滤：设定阈值 β，$\forall \chi \in U'$，样本 χ 将从样本库中删除当且仅当 $\omega(\chi)<\beta$，即：新样本库 $U_{\text{update}}=U'-\{\chi:\omega(\chi)<\beta,\chi\in U'\}$。根据经验，$\beta$ 的取值一般在 $0.1/|U|$ ～ $0.3/|U|$ 之间。

式①～式③涵盖了样本增量的所有情况，与文献[16]不同的是，任何新样本都不会是“无用样本”被直接丢弃，由于决策倾向度和过滤函数的引入，随着新样本的到来，一致样本的决策得到巩固和加强，不一致样本决策得到削弱，在过滤同质样本(同属性同决策样本)的同时，“极少数决策”(阈值以下样本)也将视作噪声被过滤掉，达到压缩样本、去除噪声的双重目的。

主动样本选择算法描述如下：

输入：原始决策表 $(U,A\bigcup D\bigcup\Omega)$，阈值 β
输出：　新决策表 U'
bool　is_new=true，$U'=\varnothing$　　//初始化
for　x　in　U //从原始决策表 U 依次取出样本 x
for　y　in　U'
　if　$a(y)=a(x)$　and　$d(y)=d(x)$
　　set is_new=false　//样本库中存在一致样本
　　set $\omega(y)=f(\omega(y))=\omega(y)+\Delta^+(\cdot)$　　// 更 新 $\omega(y)$
　else if　$a(y)=a(x)$ and $d(y)\neq d(x)$
　　set $\omega(x)=\omega_0(a(x),d(x))$
　　set $\omega(y)=f(\omega(y))=\omega(y)-\Delta^-(\cdot)$
　end if
　if　$\omega(y)<\beta$
　　$U'=U'-\{y\}$　　//从 U' 中删除样本 y
　end if
end for　y
　if　is_new=true　and　$\omega(x)\geqslant\beta$
　　set $\omega(x)=\omega_0(a(x),d(x))$
　　$U'=U'\bigcup\{x\}$　　//将新样本 x 添加到 U' 中
end if
end for　x
return　U'

示例：根据样本增量选择算法，新样本是从原样本库 U 中依次添加到新库 U' 中的，假设经过若干次样本添加后，样本库 U' 当前状态如表1所示。

表1　样本库 U' 当前状态

U	a_1	a_2	a_3	d	ω
x_1	1	0	1	1	0.5
x_2	1	0	1	0	0.3
x_3	1	0	0	1	0.6

继续添加新样本。设 $x=(1,0,1,1)$ 是后续新样本，显然已有样本库中 x_1 与 x 是一致样本，而 x_2 是不一致样本。设 $\Delta^+=0.1$，$\Delta^-=0.05$，$\beta=0.28$，则根据过滤规则，各样本决策倾向度变化为：

$$\omega(\chi)=\begin{cases}\omega(x_1)=\omega(x)=0.5+0.1=0.6\\ \omega(x_2)=0.3-0.05=0.25\\ \omega(x_3)=0.6(保持不变)\end{cases}$$

由于 $\omega(x_2)<\beta$，样本 x_2 将被视为噪声过滤掉，

更新后新样本库如表2所示。

表2 更新后样本库

U	a_1	a_2	a_3	d	ω
x_1	1	0	1	1	0.6
x_3	1	0	0	1	0.6

3 数据实验

为了验证过滤函数和决策倾向度规则的有效性，选取来自于UCI数据库[17]的3种不同数量级的样本：Soybean、Kr-vs-Kp、Letter。

3.1 数据准备

为了检验算法的去噪能力，为每个数据集加入5%的随机噪声。

噪声数据 η 满足条件 $\{x:x\in U,a(x)=a(\eta)\}\neq\varnothing$ 但 $\{x:x\in U,a(x)=a(\eta),d(x)=d(\eta)\}=\varnothing$，即数据集中不存在与噪声属性相等且决策相等的一致样本。

数据集信息如表3所示(样本数量加号后面是人为增加的噪声数量)。

表3 实验样本

数据集	样本数量	属性数量	决策类数
Soybean	307+15	35	19
Kr-vs-Kp	3 196+160	36	2
Letter	20 000+1 000	16	26

3.2 应用实例

为了说明算法的运行过程，以Soybean为例，说明数据集增量式主动选择过程。关于Soybean样本的描述可参见表3。算法步骤如下：

1) 初始化决策表 $(U,A\cup D\cup\Omega)$，其中 $|U|=322$，$|A|=35$，$|D|=19$。新决策表 $U'=\varnothing$。设初值 $\omega_0=1/|U|=3.1\times10^{-3}$，$\beta=0.1/|U|=3.1\times10^{-4}$，$\Delta=1/|U|=3.1\times10^{-3}$。

2) 依次取出样本 $x\in U$，对已经加入新样本库的所有 $y\in U'$，按照以下规则进行计算：

① 若 x 是全新样本，即 $\forall y\in U',a(y)\neq a(x)$，则直接将 x 添加到新样本库 U'。

② 若 $a(y)=a(x),d(y)=d(x)$，即 y和x 是一致样本，则更新 $\omega'(y)=\omega(y)+\Delta$，并丢弃样本 x。

③ 如 $a(y)=a(x),d(y)\neq d(x)$，即 y和x 不一致，则将样本 x 加入到 U' 中，并更新 $\omega'(y)=\omega(y)(1-1/|U|)$，若 $\omega'(y)<\beta$，则从 U' 中删除样本 y。

3) 经过增量添加和过滤，得到新样本库 U'，根据文献[18]的方法构造辨识矩阵 $\boldsymbol{M}$，并按照文献[19]实现的样本对选择算法，计算数据集的核属性及属性约简，计算依据为：$\mathrm{CORE}(U')=\{a:t(x,y)=a\}$，即约简核是辨识矩阵中所有单元素集合的并集。文献[19]给出了该计算依据的详细证明过程。

本实例的运行结果如表4所示。

表4 运行结果

初始样本数量	过滤后样本数量	约简后属性数量	运行时长/ms
322	296	32	754

3.3 样本过滤测试

本实验采用样本增量过滤计算方法，表5列出了各数据样本的实验结果。

表5 实验结果

数据集	初始样本数量	$\beta\times100$	过滤后样本数量	压缩率/%
Soybean	322	0.031	296	92
Kr-vs-Kp	3356	0.029	3 121	93
Letter	21 000	0.004 7	18 284	87

实验结果表明，人为加入的噪声数据都得到了有效地过滤，并且同时也清除了了原样本中的噪声和重复数据。

图1显示了3个数据集的数据量和运行时间(单位ms)的对比关系，从图上可以看出，过滤时间不会随着样本数据的增加而急剧增长。

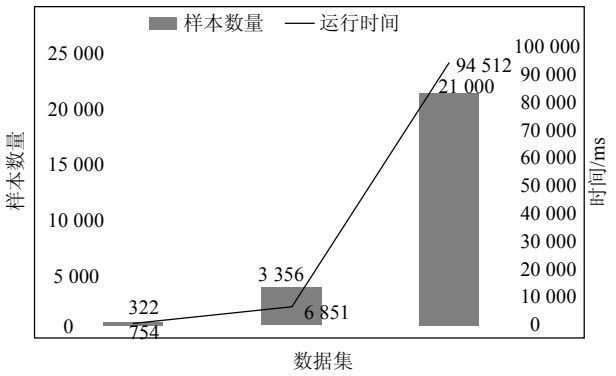

图1 数据量与时间关系图

4 结 束 语

通过引入决策倾向度指标，可以很方便地量化在不同属性值条件下，每种决策的可靠程度，通过设计过滤函数，为每一份新增的样本计算决策倾向度，并根据阈值指标决定是否过滤该样本，同时刷新已有样本库的决策倾向度，随着样本的增加，可以很好地加固稳定决策，同时去除噪声数据，实验证明，该方法可以有效地压缩数据，提高样本分析质量。

参 考 文 献

[1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341-356.

[2] 张文修, 吴伟志, 梁吉业. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001: 19-23.
ZHANG Wen-xiu, WU Wei-zhi, LIANG Ji-ye. Rough set theory and method[M]. Beijing: Science Press, 2001: 19-23.

[3] 胡清华, 于达仁. 应用粗糙计算[M]. 北京: 科学出版社, 2012: 56-60.
HU Qing-hua, YU Da-ren, Application of rough computing [M]. Beijing: Science Press, 2012: 56-60.

[4] 刘清. Rough集及Rough推理[M]. 北京: 科学出版社, 2001: 12-18.
LIU Qing, Rough sets and Rough reasoning[M]. Beijing: Science Press, 2001: 12-18.

[5] 廖启明, 龙鹏飞. 基于属性重要性的粗糙集属性约简方法[J]. 计算机工程与应用, 2013, 49(15): 130-132.
LIAO Qi-ming, LONG Peng-fei. Rough set reduction method of attribute based on importance of attribute[J]. Computer Engineering and Applications, 2013, 49(15): 130-132.

[6] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems[M]. Theory & Decision Library, 1992, 11: 331-362.

[7] 杨明. 一种基于改进差别矩阵的属性约简增量式更新算法[J]. 计算机学报, 2007, 30(5): 815-822.
YANG Ming. An incremental updating algorithm for attribute reduction based on improved discernibility matrix[J]. Chinese Journal of Computers, 2007, 30(5): 815-822.

[8] GUAN Li-he. An incremental updating algorithm of attribute reduction set in decision tables[C]//2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery. Tianjin, China: IEEE, 2009: 421-425.

[9] FENG S R, ZHANG D Z. Increment algorithm for attribute reduction based on improvement of discernibility matrix[J]. Journal of Shenzhen University Science and Engineering, 2012, 29(5): 405-411.

[10] ZIARKO W. Variable precision rough set model[J]. Journal of Computer and System Sciences, 1993, 46(1): 39-59.

[11] ZIARKO W. Analysis of uncertain information in the framework of variable precision rough sets[J]. Foundations of Computing & Decision Sciences, 1993, 18(3): 381-396.

[12] YANG Y, CHEN D, WANG H. Active sample selection based incremental algorithm for attribute reduction with rough sets[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(4): 825-838.

[13] YANG Y, CHEN D, WANG H, et al. Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving[J]. Fuzzy Sets and Systems, 2017, 312: 66-86.

[14] DONG Z, SUN M, YANG Y. Fast algorithms of attribute reduction for covering decision systems with minimal elements in discernibility matrix[J]. International Journal of Machine Learning and Cybernetics, 2016, 7(2): 297-310.

[15] CHEN D, YANG Y, DONG Z. An incremental algorithm for attribute reduction with variable precision rough sets[J]. Applied Soft Computing, 2016, 45: 129-149.

[16] 杨燕燕. 基于粗糙集的增量属性约简机理与算法研究[D]. 北京: 华北电力大学, 2017.
YANG Yan-yan. Rough set based mechanisms and algorithms for incremental attribute reduction[D]. Beijing: North China Electric Power University, 2017.

[17] BACHE K, LICHMAN M. UCI machine learning repository[EB/OL]. [2017-08-13]: http://archive.ics.uci.edu/ml/datasets.html.

[18] YAO Y, ZHAO Y. Discernibility matrix simplification for constructing attribute reducts[J]. Information Sciences, 2009, 179(7): 867-882.

[19] CHEN D G, ZHAO S Y, ZHANG L. Sample pair selection for attribute reduction with rough set[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(11): 2080-2093.

编 辑 叶 芳