鄂旭,杨芳,侯建,等.水产养殖水质预警指标约简算方法研究[J]. 华中农业大学学报,2020,39(2):89-94.
DOI: 10.13300/j.cnki.hnlkxb.2020.02.012

水产养殖水质预警指标约简算法研究

鄂旭[1,2],杨芳[3],侯建[4],毛玫静[5],阎琦[1],励建荣[2]

1.渤海大学信息科学与技术学院,锦州 121000; 2.渤海大学食品科学研究院,锦州 121000;
3.渤海大学实验管理中心,锦州 121000; 4.渤海大学工学院,锦州 121000;
5.贵阳块数据城市建设有限公司,贵阳 550001

摘要 针对水质预警指标复杂冗余问题,提出基于关联属性重要度的水质指标属性约简算法。以属性依赖度和属性重要性为基础,根据每个条件属性对决策属性影响程度,引入属性关联重要性概念进行约简操作。通过计算属性关联重要性,挖掘水质指标之间的相互影响程度,进一步判别决策属性对条件属性的分类强度,并以鲈养殖水质为实例进行算法解析,以 UCI 数据集进行实验验证,结果表明此算法在冗余属性约简方面是有效可行的。

关键词 水产养殖安全; 水质预警; 属性约简算法
中图分类号 X 53; X 820　　**文献标识码** A　　**文章编号** 1000-2421(2020)02-0089-06

我国是水产养殖大国,水产品总量连续 20 多年位居世界第一位,占世界水产品养殖总量的 70%以上[1]。水产养殖业在改善民生增加农民收入方面发挥了重要作用。然而,“多宝鱼”等恶性事件表明我国水产品面临着越来越多来自水产养殖安全问题的挑战。因此,提高水产品质量安全是当前的一个重要课题,水产养殖安全是重要一环。水质预警是一种保障水产品质量安全的预防性措施,可以预先将许多恶性水产品安全事件扼杀在源头[2]。

然而,水产养殖中水质指标因子众多,它们之间存在着错综复杂的关系,严重影响水质预警工作展开[3-4]。将安全预警技术应用于水产品养殖过程,越来越受到专家和学者的重视[5-8],如,Da Silva 等[5]对亚硝酸盐、非离子化氨、铜、铝、锌等多种已知鱼类健康应激源进行了检测,建立了线性浓度相加(LCA)模型,并与预测毒性事件的机器学习算法相结合,提出水质预处理模型;于辉辉[7]基于机器学习方法,将多种传感器数据进行融合,利用小波降噪方法,进行数据降维等预处理工作等。从众多文献中可以看出,水质指标属性约简已成为一种水质预警行之有效的手段[8-12]。

属性约简算法本质就是在保证原有信息系统分类能力或蕴涵信息量不变的情况下,删除冗余属性,保留重要属性的一种方法[9]。属性约简通常情况下主要分为从底向上和自上向下两种方法。从底向上方法将初始集合设为空集,以一定知识作为启发式信息,依次将重要属性加入初始集合,直至满足终止条件,运算结束。自上而下方法是令初始集合为原全部属性集合,然后按照一定规则删除不重要属性,直至剩余属性集合等于或近似于原信息系统信息量[10-13]。依赖度是代数观点约简算法中具有的重要性度量,本研究以粗糙集理论为指导,前向选择为出发点,在结合属性依赖度和属性重要性的基础上,根据每个条件属性对决策属性的层次影响程度,引入属性关联重要性的概念进行约简操作。通过计算属性关联重要性,能够进一步判别决策属性对条件属性的分类强度。将这种约简思想应用到水产养殖水质指标的属性约简上,一方面能够挖掘水质指标之间的相互影响程度,另一方面可为水质指标对预警信号的准确性判断提供理论依据。

1 材料与方法

1.1 仪器与设备

美国 YSIproplus 多参数水质分析仪,可同时检测鲈水产养殖中水质的水温、氨氮含量、溶解氧含量、电导率、氧化还原电位、盐度、氯化物含量、酸碱

收稿日期: 2019-07-17
基金项目:“十三五”国家重点研发计划重点专项 (2019YFD0901605); 辽宁省社会科学规划基金项目(L19BGL016); 辽宁省自然基金重点项目(20170540005); 辽宁省教育厅基本科研项目(LQ2017002)
鄂旭,博士,教授. 研究方向:水产品冷链监测与预警. E-mail: exu21@163.com
通信作者: 励建荣,博士,教授. 研究方向: 水产品保鲜与加工. E-mail: lijr6491@163.com

度、总溶解固体含量等多项水质指标。智能计算设备为联想台式机，操作系统采用 Windows 8.1，CPU 为 G3240@3.10GHz，内存 RAM 为 4 GB，智能软件实验开发平台为 MATLAB R2015a。

1.2 传统相关概念

本研究以粗糙集理论为指导进行属性约简，传统相关定义如下[8-9,12]：

定义 1 设信息系统 $DT=<U,A,val,f>$，其中：$U=\{u_1,u_2,\cdots,u_{|U|}\}$，称为对象空间；$A=C\cup\{d\}=\{a_1,a_2,\cdots,a_{|A|}\}\cup\{d\}$，称为属性集合，且 $C\cap\{d\}=\varnothing$，C 表示条件属性集；$\{d\}$ 表示决策属性集；对于任意 $a\in A$，存在映射 $f\rightarrow a:U\rightarrow val$，$val=\{a(u)\mid u\in U\}$，称为属性 a 的值域。

定义 2 在信息系统 $DT=<U,A,val,f>$ 中，对于每个属性子集 $X\subseteq U$，则定义不可辨识关系如下：

$$U/Ind(B)=\{X\mid \exists X(X\subseteq U\wedge \forall_{x,y\in U\times U}(\forall b\wedge b(x)=b(y)))\}$$

定义 3 对于任意一个属性子集 $X\subseteq U$ 和不可辨识关系 $U/Ind(B)$，则用

$B^-(X)=\cup\{Y_i\mid \exists Y_i(Y_i\in U/Ind(B)\wedge Y_i\cap X\neq\varnothing)\}$ 表示 X 的上近似集，$B_-(X)=\cup\{Y_i\mid \exists Y_i(Y_i\in U/Ind(B)\wedge Y_i\subseteq X)\}$ 表示 X 的下近似集。

其中，$B_-(X)=POS_B(X)$，称为 X 的 B 正域；$NEG_B(X)=U\backslash B_-(X)$ 为 X 的 B 负域。

定义 4 已知信息系统 $DT^*=<U,A,val,f>$，$A=C\cup\{d\}=\{a_1,a_2,\cdots,a_{|A|}\}\cup\{d\}$，其中 $a_k\in C$，其值域 $R_k=\{r_k{}^1,r_k{}^2,\cdots,r_k^{|A|}\}$，对象 $u_i\in U$，$\forall a_k(u_i)=r_k{}^t$ 对应的概率为 $1/|R_k|$，且 $|R_k|$ 表示值域 R_k 中包含元素的个数。假定两个对象 $u_i,u_j\in MOS(u)$，则对象 u_i 按属性相容性限制，在属性 a_k 上相似于 u_j 的概率定义为 $p_k(u_i,u_j)=1/|R_k|$。那么，按照属性相容性限制，两个对象在所有条件属性集上相似概率则被定义为：

$$\rho(u_i,u_j)=\prod_{a_k\in A}p_k(u_i,u_j)$$

定义 5 在信息系统中定义属性相关度如下：

$$K(a,b)_{u_i\in IMOS}=\frac{|POS_a(b)\cup POS_b(a)|}{|U|}$$

其中，$K(a,b)$ 称为条件属性 a 和 b 的相关度，且属性 a 和 b 是对称的，$K(a,b)\in[0,1]$，$K(a,b)$ 的值越大，则表明属性 a 和 b 的相关度越高。

定义 6 在完备信息系统 $DT=<U,A,val,f>$ 中，属性集 $A=C\cup\{d\}$，若 $\forall a\in B\subset C$，$POS_B(\{d\})=POS_C(\{d\})$ 且 $POS_{B-\{a\}}(\{d\})\neq POS_B(\{d\})POS$，则称 B 是 C 关于 $\{d\}$ 的属性约简集。

定义 7 设论域 U 上存在等价关系簇 P 和 Q，若 P 的 Q 独立子集 $S\subseteq P$，有 $POS_S(Q)=POS_P(Q)$，则称 S 为 P 的 Q 约简。

1.3 新定义概念

定义 8 在完备信息系统 $DT=<U,A,val,f>$ 中，$A=C\cup\{d\}$，且 $C\cap\{d\}=\varnothing$，存在关系式 $U/Ind(C)=\{X_1,X_2,\cdots,X_n\}$，$U/Ind(\{d\})=\{Y_1,Y_2,\cdots,Y_m\}$，则定义属性依赖度为：

$$K_C(\{d\})=\frac{\sum_{i=1}^{m}|POS_C(Y_i)|}{|U|}$$

根据定义 8 可知，当 $K_C(\{d\})=0$ 时，称 C 完全独立于 $\{d\}$，也就是决策属性对于条件属性的划分无任何影响；当 $K_C(\{d\})\in(0,1]$ 时，称 C 依赖于 $\{d\}$，且当 $K_C(\{d\})=1$ 时，称 C 完全依赖于 $\{d\}$，即决策属性对于条件属性的划分是确定的。

定义 9 在完备信息系统 $DT=<U,A,val,f>$ 中，$A=C\cup\{d\}$，$C\cap\{d\}=\varnothing$。存在属性子集 $C_{sub}\subseteq C$，则 C_{sub} 关于 $\{d\}$ 的重要性定义为：

$$Sig_{\{d\}}(C-C_{sub})=K_C(\{d\})-K_{C-C_{sub}}(\{d\})$$

特别地，当 $C_{sub}=\{a\}$ 时，属性 $a\in C$ 关于 $\{d\}$ 的重要性为：

$$Sig_{\{d\}}(a)=K_C(\{d\})-K_{C-\{a\}}(\{d\})$$

通过对属性重要性大小比较，可以判别出属性 a 的重要程度。也就是说，当去掉该属性相应分类变化较大时，则说明属性 a 重要程度越高；反之，说明该属性的重要程度越低。

定义 10 完备信息系统 $DT=<U,A,val,f>$，$A=C\cup\{d\}$，$C\cap\{d\}=\varnothing$。则定义属性子集 C_{sub} 的关联重要度为：

$$H_{C_{sub}}(C)=Sig(C_{sub})*\min\{\frac{|Y_i|}{\underset{Yi\subset U/\{d\}}{|X_j|}}\}_{X_j\subset U/\{C_{sub}\}\wedge X_j\subset Y_i}$$

对于决策类 Y_i，如果在条件属性集中存在某个等价关系能唯一对应到决策类子集 Y'_i 的关系，那么在一定条件下去掉该决策类的子集会影响对应条件属性的分类能力，因此，该条件属性集对决策类 Y_i 来说是必不可少的。$H_{C_{sub}}(C)$ 就是在这个思想的基础上提出来的，$H_{C_{sub}}(C)$ 表示在论域 U 中，条件属性 C_{sub} 的分类占决策属性 Y_i 的分类最大对象数比例的倒数。在进行约简时，优先将最小 $H_{C_{sub}}(C)$ 值所对应的属性加入到约简集中，通过考虑条件属性和决策属性之间的进一步联系，可以更精确

地计算属性的重要程度。

1.4 相关定理

定理 1　假定属性集 R 是独立的，若存在属性子集 $M\subseteq R$，则 M 一定是独立的。

证明：(反证法)假设 $M\subseteq R$ 且 R 是依赖的，则存在 $S\subset R$，使得 $Ind(S)=Ind(R)$，也就是说，存在 $Ind(S\cup(R-M))=Ind(R)$，且 $S\cup(R-M)\subset R$。因此，条件 R 为依赖的假设不成立，故定理得证。

定理 2　设论域 U 上存在等价关系簇 P 和 Q，若 $RED_Q(P)$ 表示 P 的所有 Q 约简关系簇，$CORE_Q(P)$ 表示 P 的所有 Q 不可约简关系簇，则 $CORE_Q(P)=\bigcap RED_Q(P)$。

定理 3　在完备信息系统 $DT=<U,A,val,f>$ 中，$POS_C(\{d\})=\bigcup\limits_{X\in\{(U/C)\wedge(U/D)\}}X$。

定理 4　已知完备信息系统 $DT=<U,A,val,f>$，$A=C\cup\{d\}$，$C\cap\{d\}=\varnothing$。若 $C'\subseteq C$，则 $U/C=U/C'$ 的充要条件为：$K_C(\{d\})=K_{C'}(\{d\})$，$K_C(\{d\})\neq 0,1$。

1.5 实验方法描述

输入：完备决策预警信息表 $DT=<U,A,val,f>$，$U=\{u_1,u_2,\ldots,u_n\}$，$A=C\cup\{d\}$。

输出：C 相对 $\{d\}$ 的最小约简 RED。

步骤 1：初始化。令 $RED=CORE_{\{d\}}(C)=\varnothing$。

步骤 2：计算 C 相对于 $\{d\}$ 的核集 $CORE_{\{d\}}(C)$。

(1)依次计算出 $U/Ind(C)$、$U/Ind(\{d\})$、$POS_C(Y_j)$，从而求出条件属性对决策属性的依赖度 $K_C(\{d\})$；

(2)计算每个条件属性 $a_i\in C$(其中 i 为条件属性的个数)在 C 中对决策属性 $\{d\}$ 的重要性 $Sig_{\{d\}}(C-a_i)$。若 $Sig_{\{d\}}(C-a_i)\neq 0$，则 $CORE_{\{d\}}(C)=CORE_{\{d\}}(C)\cup\{a_i\}$；

步骤 3：更新 $RED=CORE_{\{d\}}(C)=CORE_{\{d\}}(C)\cup\{a_i\}$，计算 $U/Ind(RED)$、$K_{RED\{d\}(C)}(\{d\})$，并判断是否 $K_{RED\{d\}(C)}(\{d\})=K_C(\{d\})$；

若 $K_{RED\{d\}(C)}(\{d\})=K_C(\{d\})$，则终止计算，即此时 $CORE_{\{d\}}(C)$ 为所求的最小相对约简集。否则，继续步骤 4；

步骤 4：分别计算剩余各个属性对决策属性 $\{d\}$ 的重要性 $Sig_{\{d\}RED}(C-a_i)$。

若 $Sig_{\{d\}RED}(C-a_i)$ 值不相等，则选择按序号将其对应属性加入到 RED 尾部；否则计算属性子集 a_i 对决策属性的关联重要性 $H_{a_i}^C(\{d\})$ 并排序；优先选择 $H_{a_i}^C(\{d\})$ 小的值添加到 RED 中，返回步骤 3；

步骤 5：判断是否满足条件 $K_{RED\{d\}(C)}(\{d\})=K_C(\{d\})$，若满足条件，则转到步骤 6，否则，返回步骤 4；

步骤 6：结束操作，输出 C 相对 $\{d\}$ 的最小约简 RED。

1.6 实验用例

借鉴相关文献研究[8]，以鲈养殖为研究对象，给定一个养殖池中水质指标因素预警信息表为 $DT=<U,A,val,f>$。其中，决策属性 $\{d\}$ 表示养殖是否安全，N 和 P 分别表示安全和不安全；条件属性中，a_1 表示养殖池中的溶解氧，mg/L；a_2 表示 pH 值；a_3 表示水温，℃；a_4 表示非离子氨氮量，mg/L。

为了约简的简洁快速，首先将给定的初始水质指标因素预警信息表进行离散化，离散化后的预警信息表如表 1 所示。

表 1　离散化后的预警信息表

Table 1　Discretized decision table

U	a_1	a_2	a_3	a_4	$\{d\}$
u_1	0	2	1	1	0
u_2	0	2	1	0	0
u_3	1	2	1	1	1
u_4	2	1	1	1	1
u_5	2	0	0	1	1
u_6	2	0	0	0	0
u_7	1	0	0	0	1
u_8	0	1	1	1	0
u_9	0	0	0	1	1
u_{10}	2	1	0	1	1
u_{11}	0	1	0	0	1
u_{12}	1	1	1	0	1
u_{13}	1	2	0	1	1
u_{14}	2	1	1	0	0
u_{15}	0	2	1	0	0
u_{16}	0	2	1	0	1

初始化。令 $RED=CORE_{\{d\}}(C)=\varnothing$。

计算 $U/Ind(C)$、$U/Ind(\{d\})$，从而求出条件属性对决策属性的依赖度 $K_C(\{d\})$，由：

$$U/Ind(C)=\{\{u_1\},\{u_2,u_{15},u_{16}\},\{u_3\},\{u_4\},\{u_5\},\{u_6\},\{u_7\},\{u_8\},\{u_9\},\{u_{10}\},\{u_{11}\},\{u_{12}\},\{u_{13}\},\{u_{14}\}\},U/Ind(\{d\})=\{\{u_1,u_2,u_6,u_8,u_{14},u_{15}\},\{u_3,u_4,u_5,u_7,u_9,u_{10},u_{11},u_{12},u_{13},u_{16}\}\},$$

$$Y_1=\{u_1,u_2,u_6,u_8,u_{14},u_{15}\},Y_2=\{u_3,u_4,u_5,u_7,u_9,u_{10},u_{11},u_{12},u_{13},u_{16}\};$$

得出 $K_C(\{d\})=\dfrac{\sum\limits_{i=1}^{m}|POS_C(Y_i)|}{|U|}=\dfrac{13}{16}$；

计算各个属性重要性：

$$U/Ind(a_1)=\{\{u_1,u_2,u_8,u_9,u_{11},u_{15},u_{16}\},\{u_3,u_7,u_{12},u_{13}\},\{u_4,u_5,u_6,u_{10},u_{14}\}\};$$

$$X_1=\{u_1,u_2,u_8,u_9,u_{11},u_{15},u_{16}\},$$

$$X_2=\{u_3,u_7,u_{12},u_{13}\},$$

$$X_3=\{u_4,u_5,u_6,u_{10},u_{14}\};$$

$$U/Ind(a_2)=\{\{u_1,u_2,u_3,u_{13},u_{15},u_{16}\},\{u_4,u_8,u_{10},u_{11},u_{12},u_{14}\},\{u_5,u_6,u_7,u_9\}\};$$

$$X_1=\{u_1,u_2,u_3,u_{13},u_{15},u_{16}\},$$

$$X_2=\{u_4,u_8,u_{10},u_{11},u_{12},u_{14}\},$$

$$X_3=\{u_4,u_5,u_6,u_{10},u_{14}\};$$

$$U/Ind(a_3)=\{\{u_1,u_2,u_3,u_4,u_8,u_{12},u_{14},u_{15},u_{16}\},\{u_5,u_6,u_7,u_9,u_{10},u_{11},u_{13}\}\};$$

$$X_1=\{u_1,u_2,u_3,u_4,u_8,u_{12},u_{14},u_{15},u_{16}\},$$

$$X_2=\{u_5,u_6,u_7,u_9,u_{10},u_{11},u_{13}\};$$

$$U/Ind(a_4)=\{\{u_1,u_3,u_4,u_5,u_8,u_9,u_{10},u_{13}\},\{u_2,u_6,u_7,u_{11},u_{12},u_{14},u_{15},u_{16}\}\};$$

$$X_1=\{u_1,u_3,u_4,u_5,u_8,u_9,u_{10},u_{13}\},$$

$$X_2=\{u_2,u_6,u_7,u_{11},u_{12},u_{14},u_{15},u_{16}\};$$

$$Sig_{\{d\}}(C-a_1)=K_C(\{d\})-K_{C-a1}(\{d\})=\frac{13}{16}-\frac{5}{16}=\frac{1}{2};$$

$$Sig_{\{d\}}(C-a_2)=K_C(\{d\})-K_{C-a2}(\{d\})=\frac{13}{16}-\frac{13}{16}=0;$$

$$Sig_{\{d\}}(C-a_3)=K_C(\{d\})-K_{C-a3}(\{d\})=\frac{13}{16}-\frac{13}{16}=0;$$

$$Sig_{\{d\}}(C-a_4)=K_C(\{d\})-K_{C-a4}(\{d\})=\frac{13}{16}-\frac{8}{16}=\frac{5}{16};$$

所以 $CORE_{\{d\}}(C)=\{a_1,a_4\}$；

更新 $RED=CORE_{\{d\}}(C)=\{a_1,a_4\}$后，有：

$U/Ind(RED)=\{u_1,u_8,u_9\},\{u_2,u_{11},u_{15},u_{16}\},\{u_3,u_{13}\},\{u_7,u_{12}\},\{u_4,u_5,u_{10}\},\{u_6,u_{14}\}\}$；

从而求得 $K_{RED\{d\}(C)}(\{d\})=\frac{9}{16}$，又因为 $K_C(\{d\})=\frac{13}{16}$，得 $K_{RED\{d\}(C)}(\{d\})\neq K_C(\{d\})$，则分别计算属性 a_2 和 a_3 在 RED 中对决策属性$\{d\}$的重要性 $Sig_{\{d\}RED}(C-a_i)$：

令 $M=CORE_{\{d\}}(C)=\{a_1,a_4\}$，对 $C-M=\{a_2,a_3\}$进行：

$$Sig^{RED}_{\{d\}}(C-a_2)=K_C{}^{RED}(\{d\})-K_{C-a_2}{}^{RED}(\{d\})=\frac{13}{16}-\frac{9}{16}=\frac{1}{4};$$

$$Sig^{RED}_{\{d\}}(C-a_3)=K_C{}^{RED}(\{d\})-K_{C-a_3}{}^{RED}(\{d\})=\frac{13}{16}-\frac{9}{16}=\frac{1}{4};$$

可知属性 a_2 和 a_3 的重要性相等，则计算两者的关联重要性 $H^C_{a_i}(\{d\})$。

$$H_{a2}(C)=Sig^{RED}_{\{d\}}(C-a_2)\times\min\{\frac{|Y_i|}{Y_i\subset U/\{d\}}/\frac{|X_j|}{X_j\subset U/\{C'\}\wedge X_j\subset Y_i}\}=\frac{1}{4}\times\frac{1}{3}=\frac{1}{12};$$

$$H_{a3}(C)=Sig^{RED}_{\{d\}}(C-a_3)\times\min\{\frac{|Y_i|}{Y_i\subset U/\{d\}}/\frac{|X_j|}{X_j\subset U/\{C'\}\wedge X_j\subset Y_i}\}=\frac{1}{4}\times\frac{1}{4}=\frac{1}{16};$$

因此选择属性 a_3 插入到 RED 中，更新$RED=\{a_1,a_4\}\bigcup\{a_3\}=\{a_1,a_3,a_4\}$。重新计算 $U/Ind(RED)$和 $K_{RED\{d\}(C)}(\{d\})$。

$$U/Ind(RED)=\{\{u_1,u_8\},\{u_2,u_{15},u_{16}\},\{u_3\},\{u_4\},\{u_5,u_{10}\},\{u_6\},\{u_7\},\{u_9\},\{u_{11}\},\{u_{12}\},\{u_{13}\},\{u_{14}\}\};$$

又由于 $K_{RED\{d\}(C)}(\{d\})=K_C(\{d\})=\frac{13}{16}$，最终属性的约简结果为 $RED=\{a_1,a_3,a_4\}$，输出最终 C 相对$\{d\}$的最小属性约简 RED，如表2所示。

表2 相对最小属性约简信息

Table 2 Relatively minimal attribute reduction information

U	a_1	a_3	a_4	$\{d\}$
u_1	4.48	13.3	1.89	N
u_2	4.48	13.3	1.80	N
u_3	4.62	13.3	1.89	P
u_4	4.70	13.3	1.89	P
u_5	4.70	12.5	1.89	P
u_6	4.70	12.5	1.8	N
u_7	4.62	12.5	1.8	P
u_8	4.48	13.3	1.89	N
u_9	4.48	12.5	1.89	P
u_{10}	4.70	12.5	1.89	P
u_{11}	4.48	12.5	1.8	P
u_{12}	4.62	13.3	1.8	P
u_{13}	4.62	12.5	1.89	P
u_{14}	4.70	13.3	1.8	N
u_{15}	4.48	13.3	1.8	N
u_{16}	4.48	13.3	1.8	P

2 结果与分析

对于算法的时间复杂度问题，实验步骤2只需整体搜索一遍对象集 U，就可以计算出条件属性集、决策属性集对应的整体划分及条件属性相对于决策属性的依赖度，算法复杂度为 $O(|C||U|)+O(|C|^2|U|\log|U|)$；实验步骤3主要是对核属性集不断进行扩充，并以此进行目标对象等价划分，计算复杂度为 $O(|C||U|^2)$；实验步骤4是计算剩余条件属性相对于决策属性的重要性程度，计算复杂度最大为 $O(|C||U|^2)$；实验步骤5是迭代计算约简属性集与决策属性集依赖度，并判断是否与原信息表中的依赖度相等，计算复杂度为 $O(|C|^2|U|\log|U|)$。

所以整个算法的时间复杂度为$O(|C||U|)+O(|C|^2|U|\log|U|)+O(|C||U|^2)$。

上述实例已证明本研究约简算法在水产养殖水质指标预警信息预处理是可行的。为了进一步验证算法的有效性，选用文献[9]中的 A 算法——基于属性重要性的约简算法、B 算法——基于信息熵的属性约简算法和本文中基于关联属性重要度的约简算法，对业界常用的国际标准测试数据集 UCI 中的 Iris 数据集、Car Evaluation 数据集、Tic-Tac-Toe 数据集和 Soybean 数据集进行验证与分析。对于信息系统中的连续属性离散化问题，采用 Rosetta 中的 Entropy/MDL 算法进行处理。表 3 为相关 UCI 数据集说明，表 4 为 3 种算法最终约简实验对比结果。

表 3　UCI 数据集说明

Table 3　Explanation of UCI data sets

序号 ID	数据集 Data set	实例数 Instance amount	条件属性数 Conditional amount	决策属性数 Decisional amount	是否完备 Complete or not
1	Iris	150	5	3	是 Yes
2	Car Evaluation	1 728	6	4	是 Yes
3	Tic-Tac-Toe	958	10	8	是 Yes
4	Soybean	47	36	2	是 Yes

表 4　3 种算法的约简实验结果比较

Table 4　Comparison of experiment results for three reduction algorithms

数据集 Data set	A 算法 Algorithm A		B 算法 Algorithm B		本文算法 Algorithm C	
	约简后的属性数 Reduced number	运行时间/s Run time	约简后的属性数 Reduced number	运行时间/s Run time	约简后的属性数 Reduced number	运行时间/s Run time
Iris	4	0.082	4	0.095	4	0.074
Car Evaluation	6	0.443	6	0.467	6	0.361
Tic-Tac-Toe	8	0.407	8	0.452	8	0.282
Soybean	2	0.037	2	0.040	2	0.033

通过对表 4 中数据分析可以看到，3 种算法约简后的属性个数大体上是一致的，说明这 3 种算法的约简质量差不多，且针对 UCI 数据集都得到了最小约简集 RED。但在时间耗能上，本研究算法明显要优于前 2 种算法。这是因为本研究算法相对于 A 算法而言，通过计算属性依赖度并引入关联重要度的同时，不需要重复计算删除某个属性后所引起的正域划分程度，这样有效减少了约简的时间复杂度；相对于 B 算法而言，当考虑信息熵时会涉及到概率分布，在计算 $H(Q|P)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i))$时就需要消耗一定时间，在每次更新约简集时都要重新计算该值，这样也会使约简时间延长。

最后，通过对 Car Evaluation 数据集的约简结果可知，3 种算法在处理实例数较多的数据时，约简质量是相同的，但本研究算法在时间性能上更具优势。运行时间对比如图 1 所示。

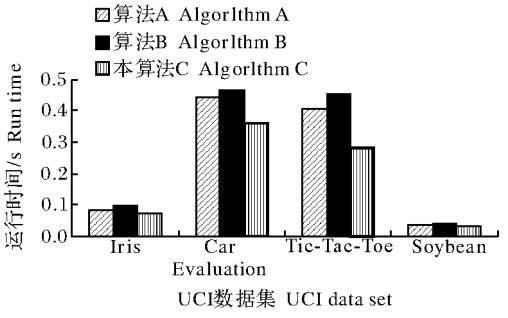

图 1　属性约简结果对比图

Fig.1　Attribute reduction results comparison chart

3　讨　论

笔者针对水质预警指标的属性存在冗余性问题进行了研究，在结合属性依赖度和属性重要性的基础上，根据每个条件属性对决策属性的层次的影响程度，引入了属性关联重要性的概念进行约简操作，提出了基于关联属性重要度的水质指标属性约简算法。将这种属性约简思想应用到水产养殖水质指标的属性约简上，一方面能够挖掘水质指标之间的相互影响程度，另一方面为水质指标对预警信号的准确性判断提供理论参考依据。并用实例和 UCI 国际公共测试数据集对本算法进行了实验分析，验证了本算法的可行性和有效性。

目前，本算法只适用于离散型属性，未来连续型属性约简方法将是进一步研究的方向。

参考文献 References

[1] 农业部渔业局.中国渔业统计年鉴 2017[M].北京：中国农业出版社，2017. Bureau of fisheries, ministry of agriculture. China fishery statistics yearbook 2017[M]. Beijing: China Agricul-

ture Press,2017(in Chinese).

[2] 唐晓纯. 国家食品安全风险监测评估与预警体系建设及其问题思考[J]. 食品科学,2013,34(15):342-348. TANG X C. Construction of national food safety risk monitoring and evaluation and early warning system and its problems [J]. Food science,2013,34(15):342-348(in Chinese with English abstract).

[3] EHSAN O,HAMID Z A,ALI D M. A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River[J]. Geoscience frontiers,2017,8(3):517-527.

[4] 李道亮. 农业物联网导论[M]. 北京:科学出版社,2012. LI D L. Introduction to agricultural Internet of things [M]. Beijing: Science Press,2012(in Chinese).

[5] DA SILVA L F B A, YANG Z C,PIRES N M M,et al. Monitoring aquaculture water quality: design of an early warning sensor with aliivibrio fischeri and predictive models[J]. Sensors (Basel,Switzerland),2018,18(9):2848-2864.

[6] PIRES N,MIGUEL M,DONG T,et al. A fluorimetric nitrite biosensor with polythienothiophene-fullerene thin film detectors for on-site water monitoring[J]. Analyst,2019,144(14): 4342-4350.

[7] 于辉辉. 基于机器学习的池塘养殖水质关键因子预测方法研究[D].北京:中国农业大学,2018. YU H H. Prediction research of water quality in aquaculture based on machine learning method[D]. Beijing: China Agricultural University,2018 (in Chinese with English abstract).

[8] 毛玫静,鄂旭,谭艳,等. 基于属性相关度的缺失数据填补算法研究[J]. 计算机工程与应用,2016,52(6):74-79. MAO M J,E X,TAN Y,et al. Algorithm study on missing data imputation based on attribute relevancy[J]. Computer engineering and applications,2016,52(6):74-79(in Chinese with English abstract).

[9] 王国胤.Rough集理论与知识获取[M].西安:西安交通大学出版社,2001. WANG G Y. Rough theory and knowledge acquisition [M]. Xi'an: Xi'an Jiaotong University Press,2001(in Chinese).

[10] 陈昊,杨俊安,庄镇泉.变精度粗糙集的属性核和最小属性约简算法[J].计算机学报,2012,35(5):1011-1017. CHEN H, YANG J A,ZHUANG Z Q. The core of attributes and minimal attributes reduction in variable precision rough set[J]. Chinese journal of computers,2012,35(5):1011-1017(in Chinese with English abstract).

[11] 葛浩,李龙澍,杨传健.基于差别集的启发式属性约简算法[J]. 小型微型计算机系统,2013,34(2):380-385. GE H,LI L S, YANG C J. Heuristics attribute reduction algorithm based on discernibility set[J]. Journal of Chinese computer systems, 2013,34(2):380-385(in Chinese with English abstract).

[12] 鄂旭,谭艳,励建荣,等.属性约简算法在海产品安全评估中的应用[J]. 计算机工程与应用,2017,53(2):98-102,150. E X, TAN Y,LI J R,et al. Application of attribute reduction algorithm in seafood safety assessment[J]. Computer engineering and applications, 2017, 53(2): 98-102, 150(in Chinese with English abstract).

[13] NGUYEN N T,SARTRA W. A new approach for reduction of attributes based on stripped quotient sets[J]. Pattern recognition,2019,97(1):1-13.

Reduction algorithm of aquaculture water quality early warning indexes

E Xu[1,2],YANG Fang[3],HOU Jian[4],MAO Meijing[5],YAN Qi[1], LI Jianrong[2]

1.*College of Information Science and Technology,Bohai University,Jinzhou* 121000,*China*;
2.*Research Institute of Food Science,Bohai University,Jinzhou* 121000,*China*;
3. *Experiment Center of Bohai University,Jinzhou* 121000,*China*;
4.*College of Engineering,Bohai University,Jinzhou* 121000,*China*;
5.*Guiyang Block Data City Construction Co.,LTD,Guiyang* 550001,*China*

Abstract To solve the problem of complex redundancy of water quality early warning indexes,an attribute reduction algorithm based on correlation attribute importance was proposed. On the basis of attribute dependence and attribute importance,the concept of attribute relevance importance was introduced to reduce the decision attribute according to the influence of each conditional attribute on the decision attribute. By calculating the attribute importance and mining the interaction degree among water quality indicators,classification of decision attribute of condition attribute intensity was further distinguished,then the water quality of bass culture was taken as an example to analyze the algorithm,and the UCI data set was used for experimental verification. The results indicate that the algorithm is effective and feasible on redundancy attribute reduction.

Keywords aquaculture safety; water quality early warning; attribute reduction algorithm

(责任编辑:边书京)