数据驱动的可变精度粗糙集噪音阈值获取方法*)

赵彦钧[1,2] 王国胤[1,2] 胡 峰[1,2]
(西南交通大学信息科学与技术学院 成都 610031)[1]
(重庆邮电大学计算机科学与技术研究所 重庆 400065)[2]

摘 要 可变精度粗糙集理论是经典粗糙集理论的一种扩展理论。它通过引入噪音阈值β,增强了对噪音数据的适应性。然而噪音阈值β多是人为设定,这要求有一定先验知识。提出一种方法,完成了数据驱动的噪音阈值β的自主式获取。仿真实验结果表明,按照此方法获取的噪音阈值β能够提高可变精度粗糙集理论获取知识的性能。

关键词 可变精度粗糙集,噪音阈值,数据驱动,知识获取

Data-driven Approach to Acquisition of Variable Precision Threshold in Variable Precision Rough Set Model

ZHAO Yan-jun[1,2] WANG Guo-yin[1,2] HU Feng[1,2]
(School of Information Science and Technology, Southwest JiaoTong University, Chengdu 610031, China)[1]
(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[2]

Abstract Variable precision rough set is one of extensions of classical rough set. Through introducing variable precision threshold β, variable precision rough set enhances its ability to adapt noise data. However, the variable precision threshold β is set up by people, which needs prior domain knowledge. A data-driven approach is proposed to complete the self-learning acquisition of variable precision threshold β. The experiment results show that the variable precision threshold β acquired by this data-driven approach can improve the capability of knowledge acquisition of variable precision rough set.

Keywords Variable precision rough set, Variable precision threshold, Data-driven, Knowledge acquisition

1 引言

经典粗糙集理论[1]由波兰逻辑学家Z. Pawlak教授1982年提出。该理论利用一对上下近似集合来描述精确集合,不需要提供数据集合之外的先验知识,其在数据挖掘及数据库知识发现中的应用取得了较大进展。然而数据集中经常含有噪音,导致经典粗糙集分类效果不佳,甚至失败。为了解决这一问题,加拿大Ziarko教授提出了可变精度粗糙集理论[4],该理论通过引入噪音阈值β,将完全精确的包含关系"软化"为某种程度上的包含,增强了对噪音的适应性。

可变精度粗糙集与经典粗糙集的区别在于噪音阈值β的引入。Ziarko依据决策者的经验来选取特定的β值[4];Beynon提出在满足近似分类质量的前提下,以一个区域间隔选取β值[5];Y. Y. Yao用概率的方法研究β值设定[6]。但是这些方法都依赖于领域专家关于数据中噪音信息的先验知识。本文提出一种方法,完成了数据驱动的噪音阈值β的自主式获取。仿真实验结果表明,按照本文方法获取的噪音阈值β能够提高可变精度粗糙集理论获取知识的性能。

2 基本概念

为了便于叙述,我们先将有关经典粗糙集和可变精度粗糙集的一些基本概念做简单介绍。

2.1 有关经典粗糙集的基本概念

定义1(决策表[2]) 一个决策表 $S=\langle U,R=C\cup D,V,f\rangle$,其中$U$是对象的集合,也称为论域,$R=C\cup D$是属性集合,子集$C$和$D$分别称为条件属性集和决策属性集,$D\neq\phi$,$V$是属性值的集合,$f:U\times R\to V$是一个信息函数,它指定了$U$中每个对象$x$的属性值。

定义2(不分明关系[2]) 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$,对于每个属性子集$B\subseteq C$,我们定义一个不分明关系$IND(B)$,即

$$IND(B)=\{(x,y)\mid(x,y)\in U\times U,\forall_{b\in B}(b(x)=b(y))\}$$

显然,不分明关系是一种等价关系。

定义3(条件类和决策类[2]) 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$,$C$和$D$分别为决策表的条件属性集和决策属性集,$U/IND(C)$和$U/IND(D)$分别为论域$U$在属性集$C$和$D$上形成的划分,$U/IND(C)$中的每个等价类$X_i\in U/IND(C)$($i=1,2,\cdots,m$,$m$为等价类的个数)称为条件类;$U/IND(D)$中的每个等价类$D_j\in U/IND(D)$($j=1,2,\cdots,n$,$n$为等价类的个数)称为决策类。

定义4(正域,边界域,负域[1]) 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$,$C$和$D$分别为决策表的条件属性集和决策属性集,设属性集合$B\subseteq C$,$U/IND(B)=\{X_1,X_2,\cdots,X_m\}$,对任意$Y\subseteq U$,其关于$B$的正域,边界域,负域分别为:

正域:$pos(Y)=\bigcup_{i=1}^{m}\{X_i\mid X_i\subseteq Y\}$

边界域:$bnr(Y)=\bigcup_{i=1}^{m}\{X_i\mid X_i\cap Y\neq 0\ 且\ \frac{|X_i\cap Y|}{|X_i|}\neq 1\}$

*)国家自然科学基金(No. 60573068, No. 60773113),新世纪优秀人才支持计划(NCET),重庆市自然科学基金(2005BA2003),重庆市教委科技项目(KJ060517)。赵彦钧 硕士生;王国胤 博士生导师,教授;胡 峰 博士生。

负域：$negr(Y)=\bigcup_{i=1}^{m}\{X_i \mid X_i \cap Y=0\}$

集合 Y 关于 B 的正域也称下近似集，集合 Y 关于 B 的边界域和正域的并集也称上近似集。

定义 5(决策表 S 的正域，边界域[2]) 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$，$C$ 和 D 分别为决策表的条件属性集和决策属性集。设属性集合 $B\subseteq C,U/IND(B)=\{X_1,X_2,\cdots,X_m\}$，且 $U/IND(D)=\{D_1,D_2,\cdots,D_n\}$，决策表 S 关于 B 的正域、边界域分别为：

正域：$pos(S)=\bigcup_{j=1}^{n}pos(D_j)$；

边界域：$bnr(S)=\bigcup_{j=1}^{n}bnr(D_j)$。

2.2 有关可变精度粗糙集的基本概念

定义 6(可变精度粗糙集中的条件概率[4]) 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$，$C$ 和 D 分别为决策表的条件属性集和决策属性集。设属性集合 $B\subseteq C,U/IND(B)=\{X_1,X_2,\cdots,X_m\}$，对任意 $Y\subseteq U$，其关于 X_i 的条件概率为：

$$P(Y|X_i)=\frac{P(Y\cap X_i)}{P(X_i)}=\frac{|Y\cap X_i|}{|X_i|}$$

定义 7(β 正域，β 边界域，β 负域[4]) 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$，$C$ 和 D 分别为决策表的条件属性集和决策属性集。设属性集合 $B\subseteq C,U/IND(B)=\{X_1,X_2,\cdots,X_m\}$，$\beta$ 是依赖于数据中噪音程度的一个取值在 $[0,0.5)$ 上的数，对任意 $Y\subseteq U$，其关于 B 的 β 正域，β 边界域，β 负域分别为

β 正域：$pos_\beta(Y)=\bigcup_{i=1}^{m}\{X_i \mid P(Y|X_i)\geqslant 1-\beta\}$

β 边界域：$bnr_\beta(Y)=\bigcup_{i=1}^{m}\{X_i \mid \beta<P(Y|X_i)<1-\beta\}$

β 负域：$negr_\beta(Y)=\bigcup_{i=1}^{m}\{X_i \mid P(Y|X_i)\leqslant\beta\}$

集合 Y 关于 B 的 β 正域也称 β 下近似集，集合 Y 关于 B 的 β 边界域和 β 正域的并集也称 β 上近似集。

定义 8(决策表 S 的 β 正域，β 边界域[2]) 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$，$C$ 和 D 分别为决策表的条件属性集和决策属性集。设属性集合 $B\subseteq C,U/IND(B)=\{X_1,X_2,\cdots,X_m\}$，且 $U/IND(D)=\{D_1,D_2,\cdots,D_n\}$，$\beta$ 是依赖于数据中噪音程度的一个取值在 $[0,0.5)$ 上的数，决策表 S 关于 B 的 β 正域、β 边界域分别为

β 正域：$pos_\beta(S)=\bigcup_{j=1}^{n}pos_\beta(D_j)$

β 边界域：$bnr_\beta(S)=\bigcup_{j=1}^{n}bnr_\beta(D_j)$

3 数据驱动的可变精度粗糙集噪音阈值获取算法

可变精度粗糙集中，对于一个多决策类的决策表，可能存在这样的条件类：由于它与每一个决策类的交集或者为空或者包含的对象非常少，从而属于每一个决策类的 β 负域，我们将这样的条件类组成的集合称为决策表 S 的 β 绝对负域，其定义如下：

定义 9(决策表 S 的 β 绝对负域) 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$，$C$ 和 D 分别为决策表的条件属性集和决策属性集，设属性集合 $B\subseteq C,U/IND(B)=\{X_1,X_2,\cdots,X_m\}$，且 $U/IND(D)=\{D_1,D_2,\cdots,D_n\}$，$\beta$ 是依赖于数据中噪音程度的一个取值在 $[0,0.5)$ 上的数，决策表 S 的 β 绝对负域为

$$nnegr_\beta(S)=\bigcup_{i=1}^{m}\{X_i \mid \forall_{D_j\in U/IND(D)}P(D_j|X_i)\leqslant\beta\}$$

由定义 8 和定义 9 可知，在可变精度粗糙集中，论域 U 可以划分为 3 个部分：

$U=pos_\beta(S)+bnr_\beta(S)+nnegr_\beta(S)$

当 $\beta=0$ 时，$pos_\beta(S)=pos(S)$，$bnr_\beta(S)=bnr(S)$，$nnegr_\beta(S)=\phi$，$U(S)=pos_\beta(S)+bnr_\beta(S)+nnegr_\beta(S)=pos(S)+bnr(S)$，此时 $pos_\beta(S)$ 和 $nnegr_\beta(S)$ 取到最小值，$bnr_\beta(S)$ 取到最大值，可变精度粗糙集等同于经典粗糙集。

当 β 在 $[0,0.5)$ 上逐渐增大时，$bnr_\beta(S)$ 中对象将被划入到 $pos_\beta(S)$ 或 $nnegr_\beta(S)$ 中。即当 β 逐渐增大时，$bnr_\beta(S)$ 将变小，而 $pos_\beta(S)$ 和 $nnegr_\beta(S)$ 将变大。

根据上述分析，我们可以得到定理 1。

定理 1 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$，$C$ 和 D 分别为决策表的条件属性集和决策属性集，设属性集合 $B\subseteq C$，$U/IND(B)=\{X_1,X_2,\cdots,X_m\}$，且 $U/IND(D)=\{D_1,D_2,\cdots,D_n\}$，若 $0\leqslant\beta_1\leqslant\beta_2<0.5$，则 $pos_{\beta_1}(S)\leqslant pos_{\beta_2}(S)$，$bnr_{\beta_1}(S)\geqslant bnr_{\beta_2}(S)$，$nnegr_{\beta_1}(S)\leqslant nnegr_{\beta_2}(S)$。

由定义 7，我们可以得到定理 2。

定理 2 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$，$C$ 和 D 分别为决策表的条件属性集和决策属性集，设属性集合 $B\subseteq C$，$U/IND(B)=\{X_1,X_2,\cdots,X_m\}$，且 $U/IND(D)=\{D_1,D_2,\cdots,D_n\}$，对任意 $D_j(j=1,2,\cdots,n)$，令 $\delta(D_j)=1-\min\{P(D_j|X_i)|P(D_j|X_i)>0.5\}(i=1,2,\cdots,m)$，当 $\delta(D_j)\leqslant\beta<0.5$ 时，$pos_\beta(D_j)=\bigcup_{i=1}^{m}\{X_i \mid P(D_j|X_i)>0.5\}=pos_{\delta(D_j)}(D_j)$，即 $pos_{\delta(D_j)}(D_j)$ 为最大值。

证明：

$\because pos_\beta(D_j)=\bigcup_{i=1}^{m}\{X_i \mid P(D_j|X_i)\geqslant 1-\beta\}$，$pos_{\delta(D_j)}(D_j)=\bigcup_{i=1}^{m}\{X_i|P(D_j|X_i)\geqslant 1-\delta(D_j)\}$

又 $\because\delta(D_j)\leqslant\beta<0.5$

$\therefore$ 由定义 7 知，$pos_\beta(D_j)\geqslant pos_{\delta(D_j)}(D_j)$

$\because\delta(D_j)=1-\min\{P(D_j|X_i)|P(D_j|X_i)>0.5\}$

$\therefore pos_{\delta(D_j)}(D_j)=\bigcup_{i=1}^{m}\{X_i|P(D_j|X_i)\geqslant\min\{P(D_j|X_i)|P(D_j|X_i)>0.5\}\}$

此时任意关于决策类 D_j 的条件概率大于 0.5 的条件类 X_i 都将归入决策类 D_j 的 β 正域

即 $pos_{\delta(D_j)}(D_j)=\bigcup_{i=1}^{m}\{X_i \mid P(D_j|X_i)>0.5\}$

$\therefore$ 由定义 7 知，此时 $pos_{\delta(D_j)}(D_j)$ 为最大值

又 $\because pos_\beta(D_j)\geqslant pos_{\delta(D_j)}(D_j)$

$\therefore pos_\beta(D_j)=pos_{\delta(D_j)}(D_j)=\bigcup_{i=1}^{m}\{X_i|P(D_j|X_i)>0.5\}$。

由定理 1、定理 2 和定义 8，我们可以得到定理 3。

定理 3 给定决策表 $S=\langle U,R=C\cup D,V,f\rangle$，$C$ 和 D 分别为决策表的条件属性集和决策属性集，设属性集合 $B\subseteq C$，$U/IND(B)=\{X_1,X_2,\cdots,X_m\}$，且 $U/IND(D)=\{D_1,D_2,\cdots,D_n\}$，若 $\delta(D_j)=1-\min\{P(D_j|X_i)|P(D_j|X_i)>0.5\}(i=1,2,\cdots,m;j=1,2,\cdots,n)$，$\delta(S)=\max\{\delta(D_1),\delta(D_2),\cdots,\delta(D_n)\}$，则当 $\delta(S)\leqslant\beta<0.5$ 时，$pos_\beta(S)=\bigcup_{i=1}^{m}\{X_i \mid \exists_{D_j}P(D_j|X_i)>0.5\}=pos_{\delta(S)}(S)$，即 $pos_{\delta(S)}(S)$ 为最大值。

证明：

对于任意决策类 $D_j(j=1,2,\cdots,n)$，

$\because\delta(D_j)=1-\min\{P(D_j|X_i)|P(D_j|X_i)>0.5\}(i=1,2,\cdots,m)$，

$\therefore$ 由定理 2 知，当 $\delta(D_j)\leqslant\beta<0.5$ 时，$pos_\beta(D_j)=\bigcup_{i=1}^{m}\{X_i|$

$P(D_j|X_i)>0.5\}=pos_{\delta(D_j)}(D_j)$为最大值。

又∵$\delta(S)=\max\{\delta(D_1),\delta(D_2),\cdots,\delta(D_n)\}$，

∴由定理1知，当$\delta(S)\leqslant\beta<0.5$时，决策表的所有决策类的$\beta$正域都为最大值。

又∵$pos_\beta(S)=\bigcup_{j=1}^{n}pos_\beta(D_j)$，

∴当$\delta(S)\leqslant\beta<0.5$时，$pos_\beta(S)=\bigcup_{i=1}^{m}\{X_i|\exists_{D_j}P(D_j|X_i)>0.5\}=pos_{\delta(S)}(S)$为最大值。

可变精度粗糙集噪音阈值β的选取原则：(1)首先使决策表S的$pos_\beta(S)$达到最大，即有尽可能多的对象确定分类；(2)在满足(1)的情况下，使$nnegr_\beta(S)$尽可能小，即有尽可能少的对象无法分类。

根据定理3，选取$\beta=\delta(S)$，可以满足以上可变精度粗糙集噪音阈值β的选取原则。

证明：

∵由定理3知，当$\beta\in[\delta(S),0.5)$时，$pos_\beta(S)=\bigcup_{i=1}^{m}\{X_i|\exists_{D_j}P(D_j|X_i)>0.5\}=pos_{\delta(S)}(S)$为最大值。

又∵由定理1知，当β在$[\delta(S),0.5)$上逐渐增大时，$nnegr_\beta(S)$可能会随之变大，即$bnr_\beta(S)$中对象有可能被划入到$nnegr_\beta(S)$中。

∴当$\beta=\delta(S)$时，$pos_\beta(S)$为最大值，并且在满足此条件下，$nnegr_\beta(S)$为最小值。

我们用一个例子来说明。表1为一个决策表。

表1　决策表$S(a_1,a_2,a_3,a_4$为条件属性，d为决策属性)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
a_1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1
a_2	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0
a_3	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
a_4	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
d	1	1	1	2	1	0	0	3	0	1	1	3	3	2	0	1	0

当$\beta=\delta(S)=0.25$时，$pos_{\delta(S)}(S)=\bigcup_{i=1}^{4}\{X_i|\exists_{D_j}P(D_j|X_i)>0.5\}=\{1,2,3,4,5,6,7,8,9\}$为最大值，$bnr_{\delta(S)}(S)=\{10,11,12,13,14,15,16,17\}$，$nnegr_{\delta(S)}(S)=\phi$。

当$\beta=0.4$时，$pos_{0.4}(S)=\bigcup_{i=1}^{4}\{X_i|\exists_{D_j}P(D_j|X_i)>0.5\}=\{1,2,3,4,5,6,7,8,9\}$仍为最大值，然而$bnr_{0.4}(S)=\{15,16,17\}$，$nnegr_{0.4}(S)=\{10,11,12,13,14\}$。

显然，当$\beta>\delta(S)$时，$pos_\beta(S)$保持最大值不变而$nnegr_\beta(S)$变大。

算法1　数据驱动的可变精度粗糙集噪音阈值获取

输入：决策表$S=\langle U,R=C\cup D,V,f\rangle$

输出：噪音阈值β

设$U/IND(C)=\{X_1,X_2,\cdots,X_m\}$，且$U/IND(D)=\{D_1,D_2,\cdots,D_n\}$

Step1：计算$P(D_j|X_i)(i=1,2,\cdots,m;j=1,2,\cdots,n)$

$$P(D_j|X_i)=\frac{|D_j\cap X_i|}{|X_i|};$$

Step2：计算$\delta(D_j)(j=1,2,\cdots,n)$

$\delta(D_j)=1-\min\{P(D_j|X_i)|P(D_j|X_i)>0.5\}$；

Step3：计算$\delta(S)$

$\delta(S)=\max\{\delta(D_1),\delta(D_2),\cdots,\delta(D_n)\}$；

Step4：返回$\delta(S)$。

4　数据驱动的可变精度粗糙集约简算法

与经典粗糙集不同，可变精度粗糙集中的决策表属性约简有可能产生决策表的β正域变大、β边界域和β绝对负域变小的情况。

定义10　给定决策表$S=\langle U,R=C\cup D,V,f\rangle$，$C$和$D$分别为决策表的条件属性集和决策属性集，$a_i$为任意条件属性$(i=1,2,\cdots,|C|)$，$\beta$是依赖于数据中噪音程度的一个取值在$[0,0.5)$上的数，定义$a_i$的属性重要性为$m_i$：

$$m_i=\frac{|nnegr_\beta^{C-a_i}(S)|}{|U|}-\frac{|pos_\beta^{C-a_i}(S)|}{|U|}$$

m_i越大，表示a_i越重要。

算法2　可变精度粗糙集属性重要性排序

输入：决策表$S=\langle U,R=C\cup D,V,f\rangle$，噪音阈值$\beta$

输出：按属性重要性升序排序后的条件属性序列C'

Step1：计算a_i的属性重要性$m_i(i=1,2,\cdots,|C|)$；

$$m_i=\frac{|nnegr_\beta^{C-a_i}(S)|}{|U|}-\frac{|pos_\beta^{C-a_i}(S)|}{|U|};$$

Step2：将a_i按m_i升序排序，并输出排序后的条件属性序列C'。

在算法1和算法2的基础上，我们可以得到算法3。

算法3　数据驱动的可变精度粗糙集属性约简

输入：决策表$S=\langle U,R=C\cup D,V,f\rangle$

输出：决策表S的属性约简

设$U/IND(C)=\{X_1,X_2,\cdots,X_m\}$，且$U/IND(D)=\{D_1,D_2,\cdots,D_n\}$

Step1：判断S的属性值是否为连续属性值，若是则进行离散化处理，否则跳到Step 2；

Step2：调用算法1，获取噪音阈值$\delta(S)$；

Step3：计算$pos_{\delta(S)}^{C'}(D_j)(j=1,2,\cdots,n)$；

Step4：计算$nnegr_{\delta(S)}^{C'}(S)$；

Step5：调用算法2，获取按属性重要性升序排序后的条件属性序列C'；

Step6：依次删除C'中条件属性$a_i(i=1,2,\cdots,|C|)$，观察是否同时满足以下两个条件：

(1)$pos_\beta^{C'}(D_1)\subseteq pos_\beta^{C'-a_i}(D_1)$，$pos_\beta^{C'}(D_2)\subseteq pos_\beta^{C'-a_i}(D_2)$，$\cdots$，$pos_\beta^{C'}(D_n)\subseteq pos_\beta^{C'-a_i}(D_n)$

(2)$nnegr_\beta^{C'}(S)\supseteq nnegr_\beta^{C'-a_i}(S)$

若同时满足(1)和(2)说明a_i为冗余属性，可以删除，若至少有一个不满足，说明a_i为必要属性，在C'中保留a_i；

Step7：返回C'中剩余属性。

5　仿真实验

本文选取UCI数据集中的三组数据glass，ecoli和zoo做仿真实验。对于每组数据，为了增大数据中的噪音，我们随机删除若干属性，然后随机取50%记录做训练集，用整组数据做测试集。

仿真实验分两步：

第一步，求出基于本文算法获取的噪音阈值β下的测试正确率和拒识率；

第二步，在$[0,0.5)$上，以0.025为步长获取噪音阈值β，并求出相关测试正确率和拒识率。

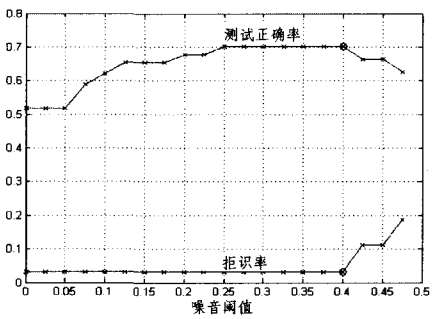
实验结果如下：

(1) glass数据实验结果(训练集107条记录，测试集214条记录)见表2和图1。

表 2　glass 数据实验结果(r 为测试正确率,n 为拒识率,β 为噪音阈值,加星号数值是按照本文算法获取的噪音阈值)

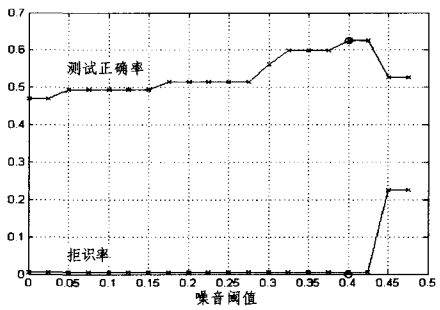
β	0.000	0.025	0.050	0.075	0.100	0.125	0.150	0.175	0.200	0.225
r	51.87%	51.87%	51.87%	58.88%	62.15%	65.42%	65.42%	65.42%	67.76%	67.76%
n	3.27%	3.27%	3.27%	3.27%	3.27%	3.27%	3.27%	3.27%	3.27%	3.27%

β	0.250	0.275	0.300	0.325	0.350	0.375	0.400	0.425	0.450	0.475	0.400*
r	70.09%	70.09%	70.09%	70.09%	70.09%	70.09%	70.09%	66.36%	66.36%	62.62%	70.09%
n	3.27%	3.27%	3.27%	3.27%	3.27%	3.27%	3.27%	11.21%	11.21%	18.69%	3.27%



叉号标注以 0.025 为步长获取的噪音阈值 β,圆圈标注按照本文方法获取的噪音阈值 β。

图 1　glass 数据实验结果



叉号标注以 0.025 为步长获取的噪音阈值 β,圆圈标注按照本文算法获取的噪音阈值 β。

图 2　ecoli 数据实验结果

表 3　ecoli 数据实验结果(r 为测试正确率,n 为拒识率,β 为噪音阈值,加星号数值是按照本文算法获取的噪音阈值)

β	0.000	0.025	0.050	0.075	0.100	0.125	0.150	0.175	0.200	0.225
r	47.02%	47.02%	49.40%	49.40%	49.40%	49.40%	49.40%	51.49%	51.49%	51.49%
n	0.60%	0.60%	0.60%	0.60%	0.60%	0.60%	0.60%	0.60%	0.60%	0.60%

β	0.250	0.275	0.300	0.325	0.350	0.375	0.400	0.425	0.450	0.475	0.400*
r	51.49%	51.49%	56.25%	59.82%	59.82%	59.82%	62.50%	62.50%	52.68%	52.68%	62.50%
n	0.60%	0.60%	0.60%	0.60%	0.60%	0.60%	0.60%	0.60%	22.62%	22.62%	0.60%

显然,按照本文算法,噪音阈值 β 取 0.400 时,测试正确率达到最大值 70.09%,拒识率达到最小值 3.27%。

(2)ecoli 数据实验结果(训练集 168 条记录,测试集 336 条记录)见表 3 和图 2。

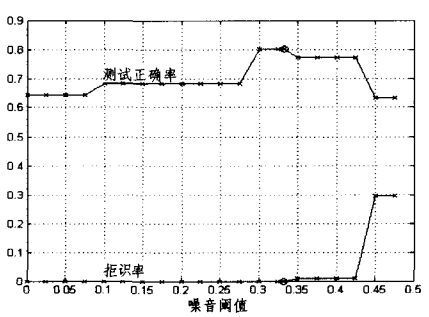
显然,按照本文算法,噪音阈值 β 取 0.400 时,测试正确率达到最大值 62.50%,拒识率达到最小值 0.60%。

(3) zoo 数据实验结果(训练集 50 条记录,测试集 101 条记录)见表 4 和图 3。

表 4　zoo 数据实验结果(r 为测试正确率,n 为拒识率,β 为噪音阈值,加星号数值是按照本文算法获取的噪音阈值)

β	0.000	0.025	0.050	0.075	0.100	0.125	0.150	0.175	0.200	0.225
r	64.36%	64.36%	64.36%	64.36%	68.32%	68.32%	68.32%	68.32%	68.32%	68.32%
n	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

β	0.250	0.275	0.300	0.325	0.350	0.375	0.400	0.425	0.450	0.475	0.333*
r	68.32%	68.32%	80.20%	80.20%	77.23%	77.23%	77.23%	77.23%	63.37%	63.37%	80.20%
n	0.00%	0.00%	0.00%	0.00%	1.00%	1.00%	1.00%	1.00%	29.70%	29.70%	0.00%



叉号标注以 0.025 为步长获取的噪音阈值 β,圆圈标注按照本文算法获取的噪音阈值 β。

图 3　zoo 数据实验结果

显然,按照本文算法,噪音阈值 β 取 0.333 时,测试正确率达到最大值 80.20%,拒识率达到最小值 0.00%。

结束语　可变精度粗糙集通过引入噪音阈值 β,将经典粗糙集中完全精确的包含关系"软化"为某种程度上的包含,具有一定的容错能力,增强了对噪音的适应性。然而噪音阈值 β 多是人为设定,这要求有一定先验知识。本文提出一种方法,完成了数据驱动的噪音阈值 β 的自主式获取。仿真实验表明,按照本文方法选取的噪音阈值 β 能够提高可变精度粗糙集理论获取知识的性能。

参 考 文 献

[1] Pawlak Z. Rough Set [J]. International Journal of Computer and Information Sciences, 1982, 11: 341-356

[2] 王国胤. Rough 集理论与知识获取[M]. 西安交通大学出版社, 2001

[3] 王国胤,何晓. 一种不确定性条件下的自主式知识学习模型. 软件学报,2003,14(6):1096-1102

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

[5] Beyon M. Reducts within the variable precision rough sets model: a further investigation [J]. International Journal of Operational Research,2001, 134:592-605

[6] Yao Y Y. Probabilistic Approaches to Rough Set [J]. Expert Systems, 2003,20(5):287-297

[7] An A, Shan N, Chan C, et al. Discovering rules for water demand prediction: an enhanced rough set approach [J]. Engineering Application and Artificial Intelligence,1996, 9(6):645-653