文章编号: 1001-0920(2009)02-0297-04

基于变精度粗糙信息熵的特征约简算法

丛 蓉[1,2], 王秀坤[1], 刘云飞[2], 杨南海[1]
(1. 大连理工大学 电子与信息工程学院, 辽宁 大连 116024; 2. 海军大连舰艇学院 教育技术中心, 辽宁 大连 116018)

摘 要: 为解决传统粗糙集不确定度量存在的局限,提出将变精度粗糙信息熵作为度量标准. 该度量标准不仅具有变精度粗糙集良好的抗噪声干扰性能,而且具有基于信息理论的粗糙信息熵更全面反映系统不确定性的能力. 给出了基于变精度粗糙信息熵的特征约简算法,实验结果表明该算法具有良好的运行效果.

关键词: 变精度粗糙集; 信息熵; 约简
中图分类号: TP18 **文献标识码**: A

Algorithm of feature reduction based on variable precision rough information entropy

CONG Rong[1,2], *WANG Xiu-kun*[1], *LIU Yun-fei*[2], *YANG Nan-hai*[1]
(1. School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China; 2. Education Technology Center, Dalian Naval Academy, Dalian 116018, China. Correspondent: CONG Rong, E-mail: congrong1002@sina.com)

Abstract: To solve the limitation of uncertainty measure in classical rough set, a measure criterion based on variable precision rough information entropy is proposed. The new criterion has a good tolerance to noise the same as variable precision rough set. At the same time, it has the same power to represent uncertainty as information theory based rough information entropy. An algorithm of feature reduction based on variable precision rough information entropy is presented. Experiments results show that the algorithm yields satisfying reduction results.

Key words: Variable precision rough set; Information entropy; Reduction

1 引 言

在身份融合过程中,不管在哪个级别上进行融合,均要从原始数据集合中抽取出观测对象的特征信息[1]. 这里所说的特征实际上就是对原始数据的一种抽象,特征约简是从大量的特征中选取小部分最能代表实体的特征,目的是提供一个简化的集合,使该集合能够确切、简化地表示原始信息所描述的对象. 特征约简的作用主要体现在简化了数据分析、提高了数据分析的精确性和减小了数据分析的复杂性.

特征约简具有很强的不确定性,约简结果的可靠性和准确程度将直接影响到下一步的目标识别. 这其中的不确定性包括不完整性、不一致性、随机性和模糊性中的一方面或几方面. 目前用于处理不确定性问题的方法有贝叶斯方法、概率方法、证据理论方法和模糊集方法等. 除此之外,由波兰的 Pawlak 教授提出的粗糙集理论[2],被认为是处理模糊和不确定性问题的新的数学工具. 粗糙集理论在不损失原始数据重要信息的前提下,通过约简原始数据中的信息,提取有用特征,进而生成识别规则.

Pawlak 粗糙集模型的一个局限性是它严格按照等价类进行分类,分类精确,没有某种程度上的"包含",对噪声数据十分敏感,其结果使得属性约简受噪声数据的影响很大,造成许多有价值的规则无法提取;另一个局限性是它所处理的对象已知,且从模型中得到的所有结论仅适用于这些对象集,但在实际应用中,需将一些小规模的对象集中得到的结论应用到大规模的对象集合中[3]. 为增强粗糙集模型的抗干扰能力和适用性,Ziarko 提出了变精度粗糙集(VPRS)模型[4]. 它是在 Pawlak 粗糙集模型的基础上引入参数 $\beta(0 \leq \beta < 0.5)$,即允许一定程度的错误分类率存在,有利于用粗糙集理论从认为不相关的数据中发现相关数据. VPRS 模型主要用于解决属性间无严格意义上的函数关系或存在概率上的

收稿日期: 2007-11-16; **修回日期**: 2008-01-16.
作者简介: 丛蓉(1974—),女,辽宁大连人,副教授,博士生,从事数据挖掘算法的研究; 王秀坤(1946—),女,辽宁辽阳人,教授,博士生导师,从事数据挖掘、数据库应用的研究.

不确定关系时的数据分类问题.

粗糙集用属性之间的依赖度 γ 作为属性约简的度量标准,但 γ 存在一定的局限性:1) γ 值依赖于不可分辨函数,或是集合正域的计算,计算量较大;2) γ 值反映的是决策系统的一致性程度,但对于随机性、不完整性等问题没有考虑.

一些研究者认为信息熵能够更全面地反映属性的重要性,文献[5]从理论上比较了用代数方法和信息论方法进行属性约简的区别;文献[6]给出了基于信息熵理论的属性约简度量标准,实验表明能够更全面地反映系统的不确定性.

本文提出一种变精度粗糙信息熵作为不确定性度量标准. 该度量标准既具有变精度粗糙集抗噪声干扰的能力和适应性,还具有基于信息论的粗糙信息熵更全面反映系统不确定性的能力. 最后给出了基于变精度粗糙信息熵的特征约简算法.

2 粗糙集相关概念

为方便描述,除特别说明外,本文中所有的定义都基于如下的决策系统:

决策系统 $DS = (U, C \cup D, \{V_a\}, f_a)$, U 是论域, C 是条件属性, D 是决策属性, $B \subseteq C, U/\mathrm{ind}(C) = \{X_1, X_2, \cdots, X_m\}, U/\mathrm{ind}(D) = \{Y_1, Y_2, \cdots, Y_n\}$.

2.1 Pawlak 粗糙集模型[3]

在粗糙集理论中,知识被认为是一种将现实或抽象的对象进行分类的能力.

定义 1 决策系统中,对于 $X \subseteq U$, X 的 B 下近似集定义为

$$B_-(X) = \cup\{Y \in U/B \mid Y \subseteq X\}. \tag{1}$$

定义 2 D 对 B 的依赖度记为 γ,定义如下:

$$\gamma(B, D) = \frac{|POS_B(D)|}{|U|} = \sum_{X \in U/D} \frac{|B_-(X)|}{|U|}. \tag{2}$$

2.2 Ziarko 变精度粗糙集模型[3]

定义 3 设 X 和 Y 表示有限论域 U 的非空子集,定义二元关系

$$c(X, Y) = \begin{cases} 1 - \dfrac{|X \cap Y|}{|X|}, & |X| > 0; \\ 0, & |X| = 0. \end{cases} \tag{3}$$

其中 $|X|$ 表示集合 X 的基数,称 $c(X, Y)$ 为集合 X 关于集合 Y 的相对错误分类率. 给定 $0 \leq \beta < 0.5$,若 $c(X, Y) \leq \beta$,则称 X 以误差 β 包含于 Y,记为 $X \subseteq^{\beta} Y$.

定义 4 决策系统中,对于 $X \subseteq U$, X 的 β 下近似集定义为

$$B_{\beta-}(X) = \cup\{E \in U/B \mid c(E, X) \leq \beta\}. \tag{4}$$

X 的 β 下近似可理解为将 U 中的对象以不大于 β 的分类误差分类于 X 的集合.

定义 5 决策属性 D 对条件属性 B 的 β 近似依赖度记为 γ_{β},定义如下:

$$\gamma(B, D, \beta) = \frac{|\mathrm{POS}(B, D, \beta)|}{|U|}, \tag{5}$$

其中

$$\mathrm{POS}(B, D, \beta) = \bigcup_{X \in U/D} B_{\beta-}(X). \tag{6}$$

2.3 基于 γ 的属性约简

定义 6 条件属性 C 关于决策属性 D 的 β 约简是 C 的一个子集 $\mathrm{red}(C, D, \beta)$,且满足:

1) $\gamma(C, D, \beta) = \gamma(\mathrm{red}(C, D, \beta), D, \beta)$;

2) 从 $\mathrm{red}(C, D, \beta)$ 中去掉任何一个属性,都将使 1) 不成立.

3 信息熵相关概念

3.1 信息熵和条件熵[7]:

定义 7 决策系统中, B 的信息熵定义为

$$H(B) = -\sum_{i}^{m} \frac{|X_i|}{|U|} \log_2 \frac{|X_i|}{|U|}. \tag{7}$$

定义 8 条件属性 B 相对于决策属性 D 的条件熵 $H(D \mid B)$ 为

$$H(D \mid B) = -\sum_{i}^{m} \sum_{j}^{n} \frac{|X_i \cap Y_j|}{|U|} \log_2 \frac{|X_i \cap Y_j|}{|X_i|}. \tag{8}$$

从信息熵的物理意义可知,信息熵 $H(B)$ 反映了 B 提供的平均信息量,或表示知识 B 的平均不确定性,或表示知识 B 的随机性. 从知识的粗糙性含义可知,知识 B 的粗糙性越小,则其提供的平均信息量越大,它的平均不确定性及其随机性也越小,那么它的熵越小. 条件熵 $H(D \mid B)$ 表示在已知知识 B 的情况下,对知识 D 仍然存在的平均不确定度.

3.2 基于熵的属性约简

粗糙集理论是从代数角度来进行属性约简,熵是从信息角度进行属性约简.

属性约简就是在保证与原决策系统相同的相关系数的前提下,选择尽可能简单的条件属性集. 文献[6]描述了用于粗糙集属性约简的最小描述长度原则理论(MDLP),该理论认为选择决策规则中的条件属性时,最理想的预测属性是使如下两者之和为最小的条件属性 B:

1) 对条件属性 B 进行编码的复杂性,用熵 $H(B)$ 表示;

2) 在已知条件属性 B 的情况下,由于不确定性规则而使决策 D 存在的统计不确定性,用某些合适的条件熵 $H(D \mid B)$ 表示.

定义 9 属性约简描述为使 $H(B \rightarrow D)$ 为最小的条件属性集合 B,粗糙信息熵 $H(B \rightarrow D)$ 定义如

下:

$$H(B \rightarrow D) = H(B) + H(D \mid B). \quad (9)$$

4　基于变精度粗糙信息熵的约简

4.1　依赖度与粗糙信息熵的比较

依赖度 γ 是对执行准确的对象分类能力的评价,近似依赖度 γ_β 是对执行具有分类误差 β 的对象分类能力的评价,表示只有 $\gamma_\beta \times 100\%$ 的对象能通过属性 B 以不大于 β 的分类误差划分到的等价类 U/D 中. γ 越大,分类集合的不一致性就越小.

然而,不确定性问题不仅仅只包括不一致性,还包括随机性、不完整性等等. γ 度量的是不一致性的程度,因此单用 γ 来度量不确定性是不准确的[5,6].

粗糙信息熵 $H(B \rightarrow D)$ 比 γ 和条件熵 $H(D \mid B)$ 更适用于属性约简. 因为它既包括了不确定性的度量 $H(D \mid B)$,又包括了条件属性的编码复杂度 $H(B)$,能够更全面地反映出条件属性 B 的预测分类能力,并能从多个方面来度量不确定性.

4.2　变精度粗糙信息熵

粗糙集认为,论域 U 中的对象分为确定性的和非确定性的,确定性对象就是能够根据条件属性的值通过决策规则推导出唯一的决策属性. VPRS 扩大了 U 中确定性对象的范围,认为那些可能是受噪声影响的对象是弱确定性的. 根据第 2 节的定义,传统粗糙集和变精度粗糙集中确定性对象的集合分别为 Z 和 Z_β,即

$$Z = \mathrm{POS}(B, D) = X_1 \cup \cdots \cup X_b,$$

$$Z_\beta = \mathrm{POS}(B, D, \beta) = X_1 \cup \cdots \cup X_b \cup \cdots \cup X_c.$$

将式(7),(8)代入(9),可得

$$H(B \rightarrow D) = -\sum_{i=1}^{m} \frac{|X_i|}{|U|} \log_2 \frac{|X_i|}{|U|} - \sum_{i=b+1}^{m} \sum_{j=1}^{n} \frac{|X_i \cap Y_j|}{|U|} \log_2 \frac{|X_i \cap Y_j|}{|X_i|}. \quad (10)$$

由定义 8 可知,$H(D \mid B)$ 没有区分确定性规则和不确定性规则,而是根据条件属性分区和决策属性分区的交集来计算熵值. 认为 Z 中的对象都是确定性的,位于 $U-Z$ 中的边界区域内对象,根据其属于某一决策类的可能性大小进行猜测. 而实际上边界区域内的对象的决策值都是不可知的,因而文献[6]对 $U-Z$ 内对象不进行猜测,而把边界区域内的每个对象均当作单独的一个类,边界区域内每个对象的不确定性都设为 $1/|U| \log_2 |U|$.

根据 VPRS 和文献[6],本文定义一种新的条件熵

$$H_\beta(D \mid B) = \sum_{i}^{c} \sum_{j}^{n} \frac{|X_i \cap Y_j|}{|X_i|} \log_2 \frac{|X_i|}{|X_i \cap Y_j|} + \frac{|U - Z_\beta|}{|U|} \log_2 |U| = (1 - \gamma_\beta) \log_2 |U|, \quad (11)$$

相应的变精度粗糙信息熵如下:

$$H_\beta(B \rightarrow D) = \sum_{i}^{m} \frac{|X_i|}{|U|} \log_2 \frac{|U|}{|X_i|} + (1 - \gamma_\beta) \log_2 |U|. \quad (12)$$

$H_\beta(D \mid B)$ 由两部分组成,一部分是确定性规则和受到噪声影响的弱确定性规则;另一部分是不确定性规则. 因此 $H_\beta(B \rightarrow D)$ 能区分确定性规则和不确定性规则,并且通过适当选取 β 值,能够正确区分不确定性规则与受到噪声影响的规则,是一种较为理想的基于 MDLP 原理的熵度量方法.

为了比较不同的条件属性子集,定义归一化变精度粗糙信息熵为

$$S_\beta(B \rightarrow D) \overset{\mathrm{def}}{=} \begin{cases} 1, & H(D) = \log_2 |U|; \\ 1 - \dfrac{H_\beta(B \rightarrow D) - H(D)}{\log_2 |U| - H(D)}, & \text{others.} \end{cases} \quad (13)$$

$S_\beta(B \rightarrow D)$ 值越接近 1,说明条件属性 B 的选择越优.

4.3　算法设计

本文提出了基于变精度粗糙信息熵的约简算法,其利用文献[8]提出的算法求核集,再将其他属性按照信息熵排列,并采用了前向型启发式发现算法.

输入:决策系统 $DS = (U, C \cup D, \{V_a\}, f_a)$;错误分类率 $\beta (0 \leq \beta < 0.5)$;阈值 ε.

输出:条件属性的约简 $\mathrm{red}_D(C)$.

步骤:

1) 根据决策系统区分矩阵得到条件属性的核 $\mathrm{core}_D(C)$;

2) $Z = S_\beta(C \rightarrow D)$,$\mathrm{red}_D(C) = \mathrm{core}_D(C)$;

3) 计算 $Z' = S_\beta(\mathrm{core}_D(C) \rightarrow D)$,如果 $Z/Z' \geq \varepsilon$,算法结束;

4) 令 $W = C - \mathrm{core}_D(C) = \{w_1, \cdots, w_k\}$,按照信息熵排序 $H(w_1) \leq \cdots \leq H(w_k)$,$i = 1$;

5) 令 $\mathrm{red}_D(C) = \mathrm{red}_D(C) \cup \{w_i\}$,计算 $Z' = S_\beta(\mathrm{red}_D(C) \rightarrow D)$,如果 $Z/Z' \geq \varepsilon$,算法结束;

6) $i = i + 1$;如果 $i \leq |W|$,转到 5),否则算法结束.

4.4　算法复杂度分析

计算 $\mathrm{core}_D(C)$ 的时间复杂度是 $O((|C| + |D|)^2 |U| \cdot \log |U|)$,计算 $H(B)$ 的时间复杂度

是 $O(|B||U|)$. 在第 5 步的每次循环中,计算 $S_\beta(B \to D)$ 的时间复杂度是 $O((|B|+|D|)|U| \cdot \log|U|)$,最坏情况下需要执行 $|C|$ 次. 一般决策系统中 $|D|=1$,则算法总的时间复杂度是 $O(|C|^2|U| \cdot \log|U|)$,算法空间复杂度是 $O(|C||U|)$.

5 实验研究

5.1 实例分析

下面给出一个例子对算法作出解释. 表 1 是一个决策系统的数据,其中 $\{C_1, C_2, C_3, C_4\}$ 是条件属性,$\{d\}$ 是决策属性. 首先,用决策系统区分矩阵,得到条件属性的核 $\{C_2\}$;计算非核条件属性的信息熵得到 $H(C_1) = H(C_4) = 2.3922 > H(C_3) = 1.6422$;$\beta = 0.1$ 时,由归一化变精度粗糙信息熵 $S_\beta((C_1, C_2) \to d) = S_\beta((C_1, C_2, C_3, C_4) \to d)$,根据算法可得属性约简集合 $\{C_1, C_2\}$. 也就是对于决策系统表 1,只需条件属性 $\{C_1, C_2\}$,便可预测决策属性的值.

表 1 实例数据

	C_1	C_2	C_3	C_4	d
1	1	1	1	0	1
2	1	0	1	2	0
3	0	0	1	1	0
4	2	2	1	2	1
5	2	1	1	2	0
6	2	2	1	0	1
7	1	2	0	1	1
8	0	0	0	0	0

5.2 算法性能测试

实验数据采用 5 个 UCI[9] 分类问题数据集(diabetes,heartdisease,iris,monk1,soybean)来对算法的有效性进行测试,5 个集合的属性如表 2 所示. 其中 heartdisease 和 soybean 集合因含有噪声,造成数据缺失、信息不完整,部分属性值为空.

表 2 数据集属性

数据集	记录数	条件属性个数	决策属性个数	类别数目	属性中是否有空值
diabetes	768	8	1	2	无
heartdisease	270	13	1	2	有
iris	150	4	1	3	无
monk1	432	6	1	2	无
soybean	307	35	1	19	有

实验软件采用华沙大学开发的粗糙集系统 RSES(Rough Set Exploration System)和 Delphi 平台开发的约简算法. 实验过程将每个数据集分成训练集(70 %)和测试集(30 %). 算法分别采用基于依赖度 γ(算法 1)、基于粗糙信息熵 $H(B \to D)$(算法 2)和本文提出的基于变精度粗糙信息熵(算法 3,$\beta = 0.1$,$\varepsilon = 0.95$)作为不确定性度量标准,先对训练集进行属性约简,得到决策规则,再对测试集进行分类. 每组数据采用每种算法运行 10 次,求平均属性个数、规则个数和平均识别精度,结果如表 3 所示. 实验结果表明,相比较其他度量标准,能够提供较少的约简属性数目和规则数目. 在对 diabetes,heartdisease,iris 三个数据集的测试中,本文提出的算法识别精度最高,在对 monk1 和 soybean 两个数据集的测试中,识别精度略低于算法 2. 实验比较结果表明,本文提出的算法对于含有噪声类型的数据也具有较好的效果.

表 3 约简结果

数据集	算法 1		算法 2		算法 3	
	平均属性/规则个数	平均识别精度	平均属性/规则个数	平均识别精度	平均属性/规则个数	平均识别精度
diabetes	3.4/198	0.714	2.9/371	0.681	2.3/110	0.723
heartdisease	4.2/34.5	0.79	3/130	0.716	2.8/20	0.802
iris	2/29	0.867	2/29	0.867	1.6/19.6	0.933
monk1	3/21	0.952	3/21	0.972	3.5/11.5	0.964
soybean	4.1/61.8	0.943	3.4/42	0.962	2.3/14.3	0.956

6 结 论

特征约简是身份融合中非常关键的步骤,但由于其存在的大量不确定性影响了约简的效果. 粗糙集作为解决不确定问题的方法,采用依赖度作为不确定性度量,在目标识别领域取得了一定的应用,但存在抗噪声干扰能力差、计算量大、不能全面反映不确定性等局限性. VPRS 能够提高粗糙集对噪声数据的处理,发现那些受噪声影响的弱确定性规则. 基于粗糙信息熵的不确定性度量能够反映不确定性的多个方面. 本文提出一种新的基于 VPRS 和信息熵的变精度粗糙信息熵作为决策系统中的不确定性度量,并给出了优化的算法. 实验结果表明,相比较其他度量标准,在不损失识别精度的前提下,能够提供较少的约简属性数目和规则数目.

参考文献(References)

[1] 杨万海. 多传感器数据融合及其应用[M]. 西安: 西安电子科技大学出版社, 2004.
(Yang W H. Multi-sensor data fusion and its applications[M]. Xi'an: Publishing House of University of Electronic Science and Technology, 2004.)

[2] Pawlak Z. Rough sets theory and its applications to data analysis[J]. Cybernetics and Systems, 1998, 29(7): 661-688.

(下转第 304 页)

器参数取为 $Q = 1, R = 20, P = 40.5$,预测时域 $T = 0.5$ s,采样周期 $\delta = 0.1$ s. 系统初始条件为 $x_1(0) = 1, x_2(0) = 0.5, x_3(0) = 0.8$,假设系统存在模型失配,$\dot{x}_1$ 变为 $\dot{x}_1 = \sin x_1 - x_1 + 3x_2 + x_3 + u_1$. 采用上述控制器参数,状态 x_3 响应曲线如图 1 所示.

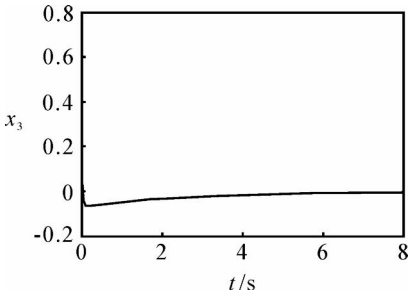

图 1 状态 x_3 的响应曲线

由仿真结果可看出,系统数学模型虽然存在失配,但控制器仍能保持系统内部动态的稳定.

5 结 论

本文针对仿射非线性 MIMO 非最小相位系统,提出了一种输入输出反馈线性化和模型预测控制结合的控制方案. 首先通过反馈线性化将系统分解为外部动态和内部动态,对外部动态用极点配置保证其稳定性,针对零动态不稳定的非最小相位系统,用模型预测控制镇定系统内部动态,并将模型预测控制信号作为系统的参考输出,实现闭环控制系统的指数稳定. 能够很好地解决存在模型失配时非最小相位系统的镇定问题. 仿真算例说明了控制器的设计过程和性能.

参考文献(References)

[1] Guardabassi G O, Savaresi S M. Approximate linearization via feedback — An overview [J]. Automatica, 2001, 31(1): 1-15.

[2] Panjapornpon C, Soroush M, Seider W D. Model-based control of unstable, non-minimum-phase, nonlinear processes[C]. Proc of the 42nd IEEE Conf on Decision and Control. Hawaii: IEEE Press, 2003: 6151-6156.

[3] Guemghar K, Srinivasan B, Mullhaupt Ph, et al. Analysis of cascade structure with predictive control and feedback linearization[J]. IEE Proc on Control Theory and Applications, 2005, 152(3): 317-324.

[4] Michael P Niemieca, Costas Kravaris. Nonlinear model-state feedback control for nonminimum-phase processes [J]. Automatica, 2003, 39(7): 1295-1302.

[5] Khalil H K. Nonlinear system [M]. New York: Prentice Hall, 2002.

[6] Isidori A. Nonlinear control systems [M]. London: Springer-Verlag, 1995.

[7] Camacho E F, Bordons C. Model predictive control [M]. Berlin: Springer, 1999.

(上接第 300 页)

[3] 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.
(Zhang W X, Wu W Z, Liang J Y. The theory and method of rough set[M]. Beijing: Publishing House of Science, 2001.)

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

[5] Wang G Y. Rough reduction in algebra view and information view[J]. Int J of Intelligent System, 2003, 18(6): 679-688.

[6] Duntsch I, Gediga G. Uncertainty measures of rough set prediction[J]. Artificial Intelligence, 1998, 106(1): 109-137.

[7] 苗夺谦, 王迁. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.
(Miao D Q, Wang Q. An information representation of the concepts and operations in rough set theory[J]. J of Software, 1999, 10(2): 113-116.)

[8] 唐建国, 谭明术. 粗糙集理论中的求核与约简[J]. 控制与决策, 2003, 18(4): 449-452.
(Tang J G, Tan M S. On finding core and reduction in rough set theory[J]. Control and Decision, 2003, 18(4): 449-452.)

[9] Newman D J, Hettich S, Blake C L. UCI Repository of machine learning databases [DB/OL]. (1998-03-20). http://www.ics.uci.edu/~mlearn/MLRepository.html. 1998.