第 36 卷 第 1 期 计 算 机 工 程 2010 年 1 月
Vol.36 No.1 Computer Engineering January 2010
·软件技术与数据库· 文章编号：1000—3428(2010)01—0049—02 文献标识码：A 中图分类号：TP311

信息表属性约简新方法

张迎春[1]，张丹枫[2]，闫德勤[1]
(1. 辽宁师范大学计算机与信息技术学院，大连 116029；2. 沈阳医学院护理学院，沈阳 110034)

摘 要：在研究区分能力大小的基础上建立一个用于指导信息表的绝对属性约简的粗糙集模型，研究区分能力和分类能力之间的关系，提出决策依赖区分精度概念，为指导决策表的相对属性约简提供了一个新的判据。给出区分精度、近似精度和决策依赖区分精度在属性约简过程中相互关系的研究结论，通过一组对比实验说明决策依赖区分精度比近似精度对分类能力的描述更细致客观。

关键词：区分精度；决策依赖区分精度；近似精度

New Methods for Attribute Reduction of Information Table

ZHANG Ying-chun[1], ZHANG Dan-feng[2], YAN De-qin[1]
(1. School of Computer and Information Technology, Liaoning Normal University, Dalian 116029;
2. School of Nursing, Shenyang Medical College, Shenyang 110034)

【Abstract】A rough set model is established to supervise the absolute attribute reduction for information table on the basis of studying the separating capacity. And a novel conception is proposed, which is called discernibility quality based on decision, on the basis of exploring the relations between the ability of discernibility and classifying, and it is an important criterion to supervise the relative attribute reduction for decision table. Several related conclusions are drawn by theoretical analyses in studying discernibility quality, approximate quality and discernibility quality based on decision. Comparison experiment shows that discernibility quality based on decision is finer than approximate quality for describing the ability of classifying.

【Key words】discernibility quality; discernibility quality based on decision; approximate quality

1 概述

波兰科学家 Pawlak 于 1982 年提出了粗糙集(rough sets)理论，其中一个重要的观点是将知识与区分事物的能力对应起来，即知识就是区分事物的能力。属性约简是有效提取知识的方法。文献[1]对知识进行量化，证明了量化的合理性，以知识量作为启发函数指导绝对属性约简，而在粗糙集理论中常用近似精度为启发函数来指导相对属性约简，其取值范围是[0, 1]上的实数，其近似精度就是相容规则数取值的归一化，即相容规则占整个信息表对象数的比例。近似精度的理论意义就是对分类能力的定性和定量描述。虽然近似精度仅仅适用于决策表，但是近似精度归一化的取值范围不仅具有明确的粗糙集理论意义，而且使得计算形式简洁统一，便于粗糙集的实际应用，从而近似精度成为粗糙集的经典标准模型。

目前知识量没有统一的取值范围，如果仅仅凭借知识量判定区分能力的增强和减弱，就如凭借相容规则个数的增减来判定分类能力的增强和减弱。因此，本文提出了知识量即区分能力大小归一化的方法，使区分能力不仅有一个定量的描述，而且有一个定性的描述，从而在对区分能力大小研究的基础上建立了一个用于指导信息表的绝对属性约简的粗糙集模型，同时在对区分能力和分类 能力两者关系深入研究的基础上提出了决策依赖区分精度概念，该概念是用于指导决策表相对属性约简的一个重要判据。

2 基本概念

设 $S=(U,Q,V,F)$ 为一信息系统，其中，$U=\{x_1,x_2,\cdots,x_n\}$ 是论域；Q 是属性集合；V 是属性取值集合；F 是 $U\times Q\to V$ 的映射；C 是条件属性集；D 是决策属性，若 D 的取值有 s 个，则由 D 导出的等价类构成 U 的一个划分：$\{Y_1,Y_2,\cdots,Y_s\}$，$Y_i=\{x\in U\mid F(x,D)=i\}$，$i=1,2,\cdots,s$。每个 x_i 和所对应的属性值称为一个信息表示的规则。当 $i\neq j$ 时，若 $F(x_i,C)=F(x_j,C)$ 但 $F(x_i,D)\neq F(x_j,D)$，则称该系统是不相容的，此时 x_i 与 x_j 所对应的规则为不相容规则。

定义 1[2] 设 $X\subseteq U$ 为论域的一个子集，$P\subseteq C$，X 关于 P 的下近似为 $P\_X=\{x\in U\mid [x]_P\subseteq X\}$，其中，$[x]_P$ 表示 U 中在等价关系 P 下的等价类元素构成的集合。

定义 2[2] 设 $P\subseteq C$ 对于划分 $\{Y_1,Y_2,\cdots,Y_k\}$ 的 P 的近似精度为

$$\gamma_P=\sum_{i=1}^{k}card(P\_Y_i)/card(U)$$

定义 3[2] 设 $P\subseteq C$，若 $\gamma_P=\gamma_C$，且不存在 $R\subset P$ 使得 $\gamma_R=\gamma_P$，则称 P 为 C 的一个相对属性约简。

定义 4[1] 设 P，Q 是 U 上的等价关系簇，若 P，Q 在 U 上导出的所有等价类相同，则 P 与 Q 具有相同的区分能力。

定义 5[1] 设 P，Q 是 U 上的等价关系簇，若 Q 与 P 具有相同的区分能力，且 Q 中没有多余的关系，则称 Q 是 P 的一

基金项目：国家自然科学基金资助项目(60372071)；中国科学院自动化研究所复杂系统与智能科学重点实验室开放课题基金资助项目(20070101)

作者简介：张迎春(1980－)，女，硕士研究生，主研方向：模式识别，粗糙集理论；张丹枫，本科生；闫德勤，教授

收稿日期：2009-06-06 **E-mail：**zhangyingchun1871@163.com

个绝对属性约简。

3 新的粗糙集模型

3.1 知识量

知识量性质及定理可以参看文献[1]，以下仅对用到的部分内容进行简单说明。

定理 1[1] 若某属性集合将论域 U 分成 m 个等价类，每个等价类分别有元素 $n_1,n_2,\cdots,n_m$ 个，则该属性集合具有的知识量 $W(n_1,n_2,\cdots,n_m)=W(1,1)\times\sum_{1\le i<j\le m}n_i\times n_j$，$W(1,1)$ 可以看作知识量的基本单位，文献[1]将其定义为常数 1。

3.2 区分精度

本文利用线性代数知识对文献[1]中定理 1[1] 提出的知识量表达式进行了深入研究，提出了另外一个知识量表达式，即本文的定理 2，然后提出了区分精度这个新概念，即本文的定义 6。

为证明定理 2，首先证明以下引理：

引理 $\sum_{1\le i<j\le m}n_i\times n_j=\frac{1}{2}(N^2-\sum_{1\le i\le m}n_i^2)$

其中，$N=\sum_{1\le i\le m}n_i$。

证明：因为 $N^2=(n_1+n_2+\cdots+n_m)\times(n_1+n_2+\cdots+n_m)=$

$n_1n_1+n_1n_2+\cdots+n_1n_m+n_2n_1+n_2n_2+\cdots+n_2n_m+\cdots=$

$n_mn_1+n_mn_2+\cdots+n_mn_m=2\sum_{1\le i<j\le m}n_i\times n_j+\sum_{1\le i\le m}n_i^2$

所以 $\sum_{1\le i<j\le m}n_i\times n_j=\frac{1}{2}(N^2-\sum_{1\le i\le m}n_i^2)$

定理 2 若论域 U 含有 N 个对象，某属性集将论域分成 m 个等价类，每个等价类含有对象个数分别为 $n_1,n_2,\cdots,n_m$，则该属性集具有的知识量为

$$W(n_1,n_2,\cdots,n_m)=W(1,1)\times[\frac{1}{2}N(N-1)-\sum_{1\le i\le m}\frac{1}{2}n_i(n_i-1)]$$

证明：因为 $W(1,1)\times\sum_{1\le i<j\le m}n_i\times n_j=W(1,1)\times\frac{1}{2}(N^2-\sum_{1\le i\le m}n_i^2)=$

$$W(1,1)\times\frac{1}{2}(N^2-N-\sum_{1\le i\le m}n_i^2+\sum_{1\le i\le m}n_i)=$$

$$W(1,1)\times[\frac{1}{2}N(N-1)-\sum_{1\le i\le m}\frac{1}{2}n_i(n_i-1)]$$

所以 $W(n_1,n_2,\cdots,n_m)=W(1,1)\times[\frac{1}{2}N(N-1)-\sum_{1\le i\le m}\frac{1}{2}n_i(n_i-1)]$

信息表的所有可能区分对数 $\frac{1}{2}N(N-1)$ 是一定的，将所有可能区分对数乘以 $W(1,1)$ 看作信息表的所有可能知识量，即信息表所有可能知识量是一定的。无论信息表是动态还是静态的，无论信息表的属性集合具有的实际知识量是多少，如果实际知识量与信息表的所有可能知识量相当，那么所得值的取值范围是区间[0, 1]上的实数。这样，对信息表某属性集所具有的知识量有了一个程度的衡量标准，也就为区分能力提供了一个定性分析描述。因此，提出如下区分精度概念。区分精度是信息表某属性集具有的实际知识量占整个信息表所有可能知识量的比。用 λ 表示区分精度，由于这个比的分子和分母都含有 $W(1,1)$，可以约去，因此有

定义 6 $\lambda=\dfrac{\frac{1}{2}N(N-1)-\sum_{1\le i\le m}\frac{1}{2}n_i(n_i-1)}{\frac{1}{2}N(N-1)}$，$\lambda\in[0,1]$

定义7 设 $P\subseteq C$，若 $\lambda_P=\lambda_C$，且不存在 $R\subset P$，使得 $\lambda_R=\lambda_P$，则称 P 为 C 的一个绝对属性约简。

3.3 决策依赖区分精度

对于决策表这个特殊的信息表而言，无论初始决策表是否含重复样本点，在属性约简过程中很有可能出现新的重复样本点。重复样本点对是不可区分的，从而改变了条件属性集合的区分能力。而重复样本点间的规则却可能是相容的，即如果该重复样本点不与其他样本点产生矛盾就不改变条件属性集合的分类能力，从而本文提出了决策依赖区分精度概念。决策依赖区分精度也是对分类能力的一个定性定量的描述。

定义 8 决策依赖区分精度

$\zeta=\dfrac{\frac{1}{2}N(N-1)-\sum_{1\le i\le m}\frac{1}{2}n_i(n_i-1)+\sum_{1\le i\le m}d_i}{\frac{1}{2}N(N-1)}$，$\zeta\in[0,1]$，其中，$d_i$ 表示第 i 个等价类含有相容重复样本点的对数。

定义9 设 $P\subseteq C$，若 $\zeta_P=\zeta_C$，且不存在 $R\subset P$，使得 $\zeta_R=\zeta_P$，则称 P 为 C 的一个相对属性约简。

定理 3 在信息表中，随着属性集合的单调递减，属性集合的区分精度、近似精度、决策依赖区分精度保持不变或单调递减；当属性减少，出现等价类的划分与属性减少前不一样时，属性集合的区分精度严格单调递减(证明略)。

4 区分精度、决策依赖区分精度和近似精度的比较

以下几种性质的比较结论是对含决策属性的信息表而言的：

(1)若决策表为相容决策表且不含重复样本点，则 $\gamma_p=\zeta_p=\lambda_p=1$。

(2)当决策表为不相容决策表，若 $\zeta_p=0$，则 $\gamma_p=\lambda_p=0$，反之未必成立。

(3)如果 $\gamma_p=\zeta_p=\lambda_p<1$，那么 $\gamma_p,\zeta_p,\lambda_p$ 对应的全部属性约简集不相等。

(4)若 $\gamma_p=\zeta_p=\lambda_p=1$，$\gamma_p$ 和 ζ_p 对应的属性约简集一定都相等，在一般情况下，λ_p 对应的属性约简与 γ_p 和 ζ_p 对应的属性约简集都不相等，但也有特殊情况，UCI 数据集 Heart-s 就是一个特例，γ_p 和 ζ_p 及 λ_p 对应的全部属性约简相等。

(5)以区分精度为启发函数令 λ_p =1 求得的属性约简集对应的近似精度和决策依赖精度一定为 1，但该属性约简未必是近似精度和决策依赖精度的属性约简。

(6) ζ_p 的每个取值都有唯一 γ_p 的值与之对应。

(7)若以区分精度为启发函数的属性约简存在，那么分别以近似精度和决策依赖区分精度为启发函数的属性约简一定存在，反之未必成立，例如 UCI 数据集 Iris。

5 区分精度、决策依赖区分精度和近似精度的实验

采用文献[3]方法生成所有可能属性约简组合，再根据文献[4]的约简组合搜索策略结合相应启发函数生成所有属性约简。因为在理论上早已证明求全部属性约简是一个 NP 问题，所以特别选用一个汽车数据集[5]来说明问题，该数据集含有 21 个对象、9 个条件属性和 1 个决策属性，以近似精度和决策依赖区分精度为启发函数求得的全部属性约简结果是一样的：000111001, 100110001, 010101101, 100101101, 110100101, 110101001, 111100011，共有 7 个属性约简、2 个最小约简，这与文献[5]讨论的该数据集的所有相对属性约简

（下转第 72 页）