文章编号:1000 - 582x(2000)01-0063-03

⑰

基于粗糙集理论的分类规则发现

63-65,73 TP311.13

印 勇[1], 曹长修[2], 张邦礼[2]
(1. 重庆大学 通信与测控工程学院,重庆 400044; 2. 重庆大学 自动化学院)

摘 要:研究了利用粗糙集理论中核的概念,求取信息系统的最小简化策略,给出了从数据库中发现分类规则的方法。

关键词:粗糙集;分类规则;数据采掘;数据库知识发现
中图分类号:TP 181;TP 182 **文献标识码**:A

粗糙集(Rough Set)理论是波兰数学家 Z. Pawlak 在 1982 年提出的一种分析数据的数学理论[1],该理论在分类的意义下定义了模糊性和不确定性的概念,是一种处理不确定和不精确问题的新型数学工具[2]。该理论从新的视角对知识进行了定义,把知识看作是关于论域的划分,认为知识是有粒度的、它主要用于知识的简化及知识依赖性的分析。粗糙集理论的最大特点是:不需要提供问题所需处理的数据集合之外的任何先验信息,如统计中要求的先验概率和模糊集中要求的隶属度,且算法简单、易于操作。该理论提出的上下近似、核、简化等概念[3],为数据分析、决策分析等提供了新的理论和方法。笔者首先对近年兴起的粗糙集的基本理论进行了讨论,在此基础上运用粗糙集理论对从数据库中发现分类规则的方法进行了研究。

1 粗糙集的基本概念

设 U是感兴趣的对象组成的非空有限集合,称 U 为论域(Universe)、R 是 U 上的等价关系(Equivalence Relation),则序对 $A = (U, R)$ 称为一个近似空间。如果 $x, y \in U$且$(x, y) \in R$,称 x和 y在A上是不可区分的,R 被称为一个不可区分关系(Indiscernibility Relation).对于任何一个 $r \in R$,根据 r的定义域可以对U划分,称为等价类,则 R在U上导出的划分是R的所有等价类族,表示为 $U/R = \{X_1, X_2, \cdots, X_n\}$,其中,$X_i$ 为R的等价类(又称为基本集)。用$[x]_R$ 表示包含 x 的R的等价类,$x \in U$.

若 $P \subseteq R$,且 $P \neq \varnothing$,则 $\cap P$(P中全部等价关系的交集)也是一等价关系,并且称为 P上的一个不可区分关系,记为 IND(P):

$$[x]_{IND(P)} = \bigcap_{r \in P} [x]_r$$

因此,$U/IND(P)$ 表示等价关系 IND(P) 的所有等价类族,即等价关系 IND(P) 在 U上导出的划分。对于任何子集 $X \subseteq U$,定义:

X的 R- 下近似(lower approximation) 集为:

$$R_-(X) = \bigcup \{Y \in U/R;\ Y \subseteq X\}$$

X的 R- 上近似(upper approximation) 集为:

$$R^-(X) = \bigcup \{Y \in U/R;\ Y \cap X \neq \varnothing\}$$

$R_-(X)$ 表示在现有知识 R下,U中所有一定能归入 X的元素的集合;$R^-(X)$ 表示 U中可能归入X的元素的集合。

2 信息系统及简化

2.1 信息系统的表示

一个信息系统 S是一个四元组:

$$S = < U, A, V, f >$$

其中:U是对象(或事例)的有限集合,即论域;$A = C \cup D$是属性的集合。其中,C表示条件属性集,D表示决策属性集;V是属性的值域集,$V = \bigcup_{a \in A} V_a$,其中,$V_a$ 是属性 $a \in A$的值域;f是信息函数,$f: U \times A \to V$,即 $f(x, a) \in V_a$,它指定 U中每一对象的属性值。

信息系统可以方便地用数据表格形式来表示。在信息系统数据表中,列表示属性,行表示对象(如状态、过程等),并且每一行表示该对象的一条信息。因此,信

· 收稿日期:1998-11-20
作者简介:印勇(1963-),男,重庆人,重庆大学副教授、博士,主要从事信号与信息处理方面的研究。

息系统也称为信息表或决策表。信息表中的一个属性对应一个等价关系，一个信息表可以看作是定义的一族等价关系。

给定一个信息系统：$S = < U, A, V, f >$，$A = C \cup D$. 对于任一子集 $P \subseteq A$，可定义在 U 上的等价关系 IND(P) 为：$(x, y) \in U$，对于每个 $p \in P$，当且仅当 $f(x, p) = f(y, p)$，有 $(x, y) \in$ IND(P) 成立。

令 $P \subseteq C$，$Q \subseteq D$，由等价关系 IND(P) 和 IND(Q) 在 U 上导出的划分分别为：

$$U/\mathrm{IND}(P) = \{X_1, X_2, \cdots, X_n\}$$

和

$$U/\mathrm{IND}(Q) = \{Y_1, Y_2, \cdots, Y_m\}$$

则 Q 的 $P-$ 正域定义为：

$$\mathrm{POS}_{\mathrm{IND}(P)}(\mathrm{IND}(Q)) = \bigcup_{Y_i \in U/\mathrm{IND}(Q)} P_-(Y_i)$$

2.2 属性的简化和核

在信息系统 $S = < U, A, V, f >$ 中，每一个对象是利用条件属性 C 的属性值来描述的。然而，C 中的某些属性可能是冗余的，因为它们不能给 S 中的对象提供任何附加信息。

令 $B \subseteq C$ 为条件属性的非空子集，如果存在一子集 $B' \subset B$，有 $\mathrm{IND} \cap (B') = \mathrm{IND}(B)$，则 B 称为属性的依赖集，否则，称为独立集。如果 $B \subseteq C$ 是独立集，且 $\mathrm{IND}(C) = \mathrm{IND}(B)$，则 B 称为 C 的简化。通常，C 的简化不止一个，C 的所有简化族记为 RED(C).

对于任一属性 $a \in C$，如果 $\mathrm{IND}(C) \neq \mathrm{IND}(C - \{a\})$，则称 a 为 C 中不可省略的(Indispensable)；否则，a 为 C 中可省略的(Dispensable). C 中所有不可省略关系的集合称为 C 的核(Core)，记为 CORE(C)，且有：

$$\mathrm{CORE}(C) = \cap \mathrm{RED}(C)$$

根据正域的概念，将上述定义扩展，可以定义相对简化的概念。对于 $B \subseteq C$，如果存在一子集 $B' \subset B$，有 $\mathrm{POS}_{\mathrm{IND}(B')}(\mathrm{IND}(D)) = \mathrm{POS}_{\mathrm{IND}(B)}(\mathrm{IND}(D))$，则 B 称为相对于 D 的依赖集，否则，称为相对于 D 的独立集。若 $B \subseteq C$ 是相对于 D 独立集，且 $\mathrm{POS}_{\mathrm{IND}(C)}(\mathrm{IND}(D)) = \mathrm{POS}_{\mathrm{IND}(B)}(\mathrm{IND}(D))$，则称 B 为 C 相对于 D 的简化，简称为 C 的 D 简化。C 的所有 D 简化族记为 $\mathrm{RED}_D(C)$.

对于任一属性 $a \in C$，如果 $\mathrm{POS}_{\mathrm{IND}(C)}(\mathrm{IND}(D)) \neq \mathrm{POS}_{\mathrm{IND}(C-\{a\})}(\mathrm{IND}(D))$，则称 a 为 C 中 D 不可省略的；否则，a 为 C 中 D 可省略的。C 中所有 D 不可省略关系的集合称为 C 的 D 核，记为 $\mathrm{CORE}_D(C)$，且有：

$$\mathrm{CORE}_D(C) = \cap \mathrm{RED}_D(C)$$

因此，可以利用 C 的任一简化 B 来代替 C 而不会丢失信息表的任何信息，从而得到一个简化的信息表。

2.3 属性值的简化和值核

同样，对信息表中的每一个实例，也可能存在冗余的属性值，利用粗糙集方法也可以将其冗余的属性值删除。

令 $F = \{X_1, X_2, \cdots, X_n\}$ 为一集合族，$X_i \subseteq U$，$G \subseteq F$ 为 F 的非空子集族。如果存在一子集族 $G' \subset G$，有 $\cap G' = \cap G$，则 G 称为依赖集合族；否则，称为独立集合族。如果 G 是独立集合族，且 $\cap F = \cap G$，则 G 称为 F 的简化。通常，F 的简化不止一个，F 的所有简化族记为 VRED(F).

对于任一子集 $X_i \subseteq F$，如果 $\cap (F - \{X_i\}) \neq \cap F$，则称 X_i 为 F 中不可省略的；否则，X_i 为 F 中可省略的。F 中所有不可省略集合的族称为 F 的核，记为 VCORE(F)，且有：

$$\mathrm{VCORE}(F) = \cap \mathrm{VRED}(F)$$

同样，将上述定义扩展，可以定义集合的相对简化的概念。

令 $F = \{X_1, X_2, \cdots, X_n\}$ 为一集合族，$X_i \subseteq U$，$G \subseteq F$ 为 F 的非空子集族。一子集 $Y \subseteq U$，使得 $\cap F \subseteq Y$. 如果存在一子集族 $G' \subset G$，有 $\cap G' \subset Y$，则 G 称为 Y 依赖集合族；否则，称为 Y 独立集合族。如果 G 为 Y 独立集合族，且 $\cap G \subseteq Y$，则 G 称为 F 的 Y 简化。F 的所有 Y 简化族记为 $\mathrm{VRED}_Y(F)$.

对于任一子集 $X_i \subseteq F$，如果 $\cap (F - \{X_i\}) \subseteq Y$，则称 X_i 为 F 中 Y 可省略的；否则，X_i 为 F 中 Y 不可省略的。F 中所有 Y 不可省略集合的族称为 F 的 Y 核，记为 $\mathrm{VCORE}_Y(F)$，且有：

$$\mathrm{VCORE}_Y(F) = \cap \mathrm{VRED}_Y(F)$$

因此，对于信息表中的任意一实例 x，可得一集合族 $F = \{[x]_{c_1}, [x]_{c_2}, \cdots, [x]_{c_n}\}$，其中，$C = \{c_1, c_2, \cdots, c_n\}$，$Y = [x]_D$，且有 $\cap F \subseteq Y$。根据上述集合族简化的概念，可得属性值简化的概念。对于实例 x，其属性值的简化为 F 的 Y 简化集合族中所对应的属性组成的集合，即，若 $C' = \{c'_1, c'_2, \cdots, c'_m\}$，$C' \subseteq C$，且 $\{[x]_{c'_1}, [x]_{c'_2}, \cdots, [x]_{c'_n}\}$ 为 F 的 Y 简化，则 C' 为实例 x 的属性值简化，它代表了每个实例的无冗余的完备信息。实例 x 的属性值的值核为 F 的 Y 核集合族中所对应的属性组成的集合，它是所有属性值简化的交集。

3 分类规则发现

3.1 算法

数据库中的关系表可被视作为一个信息系统，利用上述理论可从数据库中发现分类规则。首先删除信

息系统中的冗余属性和冗余属性值，然后，由简化的信息系统获取分类规则。对于一个信息系统来说，找出其所有的属性简化和所有的属性值简化是一个 NP 完全问题[4]。因此，可采用某种启发式方法找出最优或次优简化。在实际应用中，一般只是关心最小简化，而不需要求出所有的简化。由于核包含在所有的简化中，故从核出发求取信息表的最小简化是非常有效的方法。因此，分类规则发现的算法步骤为：

1）删除信息表中的重复实例；

2）求取条件属性相对于决策属性的属性核；

3）根据属性核删除冗余属性，求取条件属性的最小简化，并删除重复实例；

4）对于每个实例求取其属性值的值核；

5）对于每个实例删除多余的属性值，求取其最小值简化；

6）删除简化信息表中的重复实例，总结出分类规则。

3.2　算例

应用上述方法从一个病例数据库中发现肺炎和肺结核两种疾病的分类规则。在病例信息表中，每个病例有 4 种症状：a—— 发烧、b—— 咳嗽、c—— X 光阴影、d—— 听诊，即条件属性 $C=\{a, b, c, d\}$；e—— 诊断结果，即决策属性 $D=\{e\}$. 其属性值分别为：

$V_a=\{1,2,3,4\}$，其中，1—— 不发烧；2—— 低烧；3—— 中度发烧；4—— 高烧。

$V_b=\{1,2,3\}$，其中，1—— 轻微咳嗽；2—— 中度咳嗽；3—— 剧烈咳嗽。

$V_c=\{1,2,3,4\}$，其中，1—— 片状；2—— 点状；3—— 索条状；4—— 空洞。

$V_d=\{1,2,3\}$，其中，1—— 正常；2—— 干鸣音；3—— 水泡音。

$V_e=\{1,2\}$，其中，1—— 肺炎；2—— 肺结核。

病例信息表见表 1.

表 1　病例信息表

U	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
a	4	3	1	3	4	2	4	3	3	4	3	2	1	3	4	4	3	1	2	4
b	3	3	1	1	3	1	2	1	2	3	1	2	2	2	2	3	1	3	1	3
c	1	1	3	2	4	3	1	1	1	2	2	1	3	2	2	2	2	2	4	3
d	3	3	1	1	2	1	3	3	3	1	3	3	1	1	3	3	2	1	1	2
e	1	1	2	1	2	2	1	1	1	1	2	2	2	1	1	1	2	2	2	2

该信息表中无重复实例，可直接计算条件属性 C 相对于决策属性 D 的核，通过计算可知，属性集 $\{a, c, d\}$ 是 C 的 D 核，而属性 b 是冗余属性，即条件属性 C 仅有一个简化属性集 $\{a, c, d\}$. 去掉冗余属性 b 并删除重复实例后，对每一个实例属性值简化删除冗余属性值后的最小值简化信息表见表 2.

表 2　最小值简化表

U	1	2	3	4	5	6	7	8	9	10	11	12
a	4	4	3	1	*	3	*	*	2	4	4	3
c	1	*	1	*	3	*	4	*	*	2	*	2
d	*	3	*	*	*	1	*	2	*	*	1	3
e	1	1	1	2	2	1	2	2	2	1	1	2

表 2 即是最小简化表。根据最小简化表总结分类规则为：

（高烧）且具有症状（X 光所见为片状）或（X 光所见为点状）或（听诊正常）或（听诊为水泡音）之一，则为肺炎；

（中度发烧）且具有症状（X 光所见为片状）或（听诊正常）之一，则为肺炎；

具有症状（不发烧）或（低烧）或（X 光所见为索条状）或（X 光所见为空洞）或（听诊为干鸣音）之一，则为肺结核；

具有症状（中度发烧）且（X 光所见为点状）且（听诊为干鸣音），则为肺结核。

参 考 文 献

[1] PAWLAK Z. Rough sets[J]. Inter J of Computer and Information Sciences, 1982, 11(2): 341～356.

[2] PAWLAK Z. Vagueness and uncertainty: A Rough Set Prospective[J]. Inter J of Computer Intelligence, 1995, 11(2): 227～232.

[3] PAWLAK Z. Rough Classification[J]. Inter J of Man－Machine Studies, 1984, 20: 469～483.

[4] ZIARKO W. The Discovery, Analysis and Representation of Data Dependencies in Databases[A]. Piatetsky－Shapiro G, Frawley W J eds. Knowledge Discovery in Databases[C], AAA/MIT Press, 1990.213～228.

（下转 73 页）

A Video Coding Algorithm Based on Wavelet and Lower Complexity Vector Quantization

WANG Guang-xue[1], ZHANG Wen-ge[2], CAO Chang-xiu[1]

(1. College of Automation, Chongqing University, Chongqing 400044, China; 2. Guizhou Education Institute, Guiyang 550003, China)

ABSTRACT: A new video coding algorithm based on wavelet and lower complexity vector quantization is presented. The complexity of vector quantization is reduced by 76 percent by the replacement of calculating of the whole distortion with distortion of a part of components, using the interband correlation, the complexity of motion compensation is diminished considerably by the prediction of motion vector of the other subimages with the motion vector in the lowest frequency subimage. Tests are taken on Miss sequence images resulting a compression ratio 168∶1 and a PSNR 37.8, which show that the algorithm results a better performanec with a lower complexity.

KEYWORDS: video coding; vector quantization / wavelet

(责任编辑　吕赛英)

(上接 65 页)

Classification Rule Discovery Based on Rough Set Theory

YIN Yong, CAO Chang-xiu, ZHANG Bang-li

(College of Communication and Information Engineering, Chongqing University, Chongqing 400044, China)

ABSTRACT: A strategy for finding smallest reduction of information system is studied by using concept of core in rough set theory. A method of discovering classification rule in databases is proposed.

KEYWORDS: rough set; classification rule; data mining; knowledge discovery in databases

(责任编辑　吕赛英)