一种粗糙集知识挖掘算法的实现*

肖 萍[1] 张 彤[2] 姜迪刚[1] 邓 晶[1]
(1. 中山大学 北校区生物医学工程系, 广东 广州 510080;
2. 中山大学 北校区生理学教研室, 广东 广州 510080)

摘 要: 给出了一个基于粗糙集的知识规则的挖掘算法系统。系统具有通用性, 可以对各个领域内容的数据进行挖掘。挖掘系统理论上不限定条件字段及决策字段的个数, 可以对各种离散型的数据进行处理并形成规则库。

关键词: 数据挖掘; 粗糙集; 规则; 可信度
中图分类号: TP311.13 文献标识码: A 文章编号: 0529-6579 (2005) S1-0168-03

随着计算机技术和信息技术的快速发展, 特别是伴随着存贮容量技术与数据库及数据仓库技术飞速发展, 各类数据越积越多, 其中隐含的信息、知识也无可估量。面对这样的海量数据, 要想获取隐含的知识, 需要借助强有力的工具、合适的理论方法和技术。本文以粗糙集理论为依据, 尝试建立一个基于粗糙集的数据挖掘算法系统, 用于从大量数据中得到知识及规则。

1 粗糙集理论

粗糙集理论是一种处理模糊和不确定知识的数学工具, 由波兰学者 Z. Pawlak 在 1982 年首次提出[1]。在数据挖掘领域, 粗糙集最初主要用于分类, 由于它不需要依赖先验知识对不确定性的定量描述, 而只依赖数据内部的知识, 用数据之间的近似来表示知识的不确定性。适合于发现数据隐含的、潜在的知识, 而且粗糙集发现的知识是显式的描述, 可被人理解, 因此近年来, 受到国际上广泛的关注[2]。

粗糙集将研究对象信息系统以如下形式表示: $S=(U, A, V, F)$, 其中:

U: 为论域, 是一个非空有限对象 (元组) 的集合, 即系统中的所有记录对象的集合; $U=\{x1, x2, \cdots, xn\}$, 其中 xi 为对象。

A: 为对象的属性 (Attributes) 的集合; 分为两个不相交的子集, 即条件属性 C 和决策属性 D 即 $A=C\cup D$, 应用中常有 $C\cap D=\emptyset$, $C\neq\emptyset$。

V: 是属性值的集合 $V=\bigcup Va$, $Va\in A$, Va 表示属性 a 的值域。

F: 为映射关系, $f: U\times A\rightarrow V$。它为每个对象的每个属性赋予一个属性值, 即 $\forall a\in A$, $x_i\in U$, $f(xi, a)\in Va$。

1.1 等价集 (不可分辨关系) 的定义

设任一子集 $B\subseteq A$, $R(B)$ 称为不可分辨关系, 如果:

$R(B)=\{(x_i, x_j)\in U2 \mid \forall a\in B, f(x_i, a)=f(x_j, a)\}$

B 将 U 划分成若干个等价类, 此分类称为 U 基于属性集 A 的划分, 表示为

$U=\{U_1, U_2, \cdots U_n\}$, 一个 U 集里的任意两个对象 x, y, 根据 B 是不可区分的, 称 $B\_$ 为不可分辨的 ($B\_$ induscernible)。即对 $\forall a\in B$, $f(x, a)=f(y, a)\}$, 我们说 x, y 在属性集合 B 上不可分辨。$B(x)$ 表示对象 x 所属的 B 属性。

1.2 等价集的上、下近似

对于以上有限个对象的一个集合 X, 如果满足:

$B\_(X)=\{x\in U: B(x)\subseteq X\}$

$B\_(X)=\{x\in U: B(x)\cap X\emptyset\neq\}$

集合 $B\_(X)$ 称为 X 的 $B-$ 下近似, $B-(X)$ 称为 X 的 $B-$ 上近似。

2 粗糙集算法系统的实现

本文给出了一个基于粗糙集的知识规则的挖掘算法, 建立了一个计算机知识规则挖掘系统。系统主要分为: 系统管理、数据库管理、算法功能三个大模块。系统的核心部分是“算法功能”模块, 该模块是一个通用的程序模块, 可以对离散型的

* 收稿日期: 2005-02-02
作者简介: 肖萍 (1966 年生), 女, 硕士, 讲师; E-mail: jsjxp@gzsums.edu.cn

Access97 至 Access 2002 版本的数据表进行知识的提取并形成规则库。

2.1 系统主要功能模块设计

本系统主要划分为三大功能模块

系统管理模块：本系统是一个通用的算法系统，允许对 Access 97－Access2002 格式的任一张二维表进行操作，允许打开不同的文件，还可以支持由用户新建自己的数据表。以及对当前操作的数据表进行保存等操作。

数据库管理模块：该模块又分为：记录的浏览、修改、删除、添加等功能。允许用户对自己的数据表作必要的修改。同时，还可以对不必要的整张数据表进行删除。

算法模块：算法模块是整个系统的核心部分，当用户打开了一张符合要求的决策表后，便可以由用户选取任意的字段作为条件属性或决策属性。若用户的选取不合法，系统将会给出判断，如所选取条件字段和决策字段不能为同一个字段，或者条件字段、决策字段其中之一不为空。当条件字段和决策字段均按要求选择后，程序首先给出条件字段及决策字段的等价集，用户认可，便可生成规则，同时给出规则的可信度及覆盖度。

2.2 程序实现中的几个关键问题

2.2.1 数据库的连接　要实现对数据库的操作，必须建立数据绑定，本系统主要使用了 ADO (Microsoft ActiveX Data Objects) 数据绑定方法。应用程序能够通过 OLEDB 提供者访问和操纵数据库。

以下是用 ADO 建立数据库连接的关键代码：

strcon ＝″ Provider ＝ Microsoft. Jet. OLEDB. 4. 0; Data Source＝″&主程序．MdbName &″; Mode＝Share Deny None; Persist Security Info＝False″

con. Open strcon

SQL＝″SELECT ＊ FROM″ &主程序. formName

rs. Open SQL, con, adOpenStatic, adLockOptimistic

2.2.2 等价集的划分算法　算法的基本思路：用一个容器 EContainer (k, EContainerNo) 来存放属于同一个等价类的记录号，有多少个等价类就设计多少个容器；用指针标记 flag (i) 来记录已入容器的记录；用数据表指针定位 (rs. AbsolutePosition＝i) 及指针向前移动 MoveNext 方法来控制指针的方向。算法用了两层循环，实现所有记录的遍历及等价集归类。算法流程图如图 1 所示。

2.2.3 规则获取算法　条件字段及决策字段的等价类均划分完成后，接着就可以生成规则集。当条件等价集 (E_i) 与决策等价集 (Y_i) 的交集不为空时，

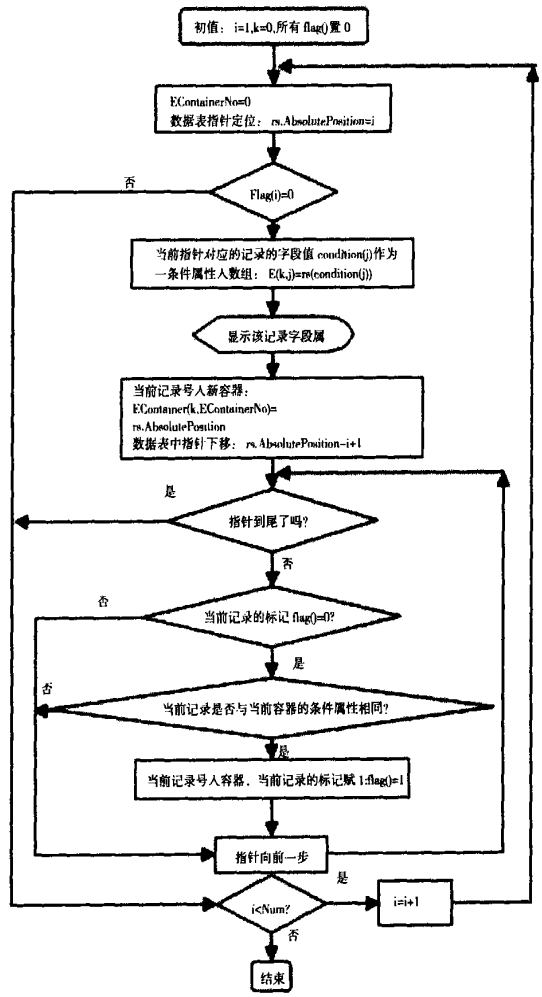

图 1　等价集算法流程图

Fig. 1　Flowchart of equipollence sets arithmetic

即 $(E_i \cap Y_i \neq \varnothing)$，则有规则：

R_{ij}：Des $(E_i) \rightarrow$ Des (Y_j)

Des (E_i) 和 Des (Y_j) 分别是等价集 E_i 和等价集 Y_i 中的特征描述。

(1) 当 $E_i \cap Y_i = E_i$ 时，$(E_i$ 完全被 Y_i 包含) 建立的规则 r_{ij} 是确定的，规则的可信度为 $cf = 1.0$

(2) 当 $E_i \cap Y_i \neq E_i$ 时，$(E_i$ 部分被 Y_i 包含) 建立的规则 r_{ij} 是不确定的，规则的可信度为：$cf = \frac{|E_i \cap Y_j|}{|E_i|}$，规则的覆盖度为：$fg = \frac{|E_i \cap Y_j|}{|Y_j|}$。

3 挖掘应用

本算法系统在 Window 2000 以上版本调试通过。并应用于中山大学肿瘤防治中心病人病案资料的“TNM 分期与 5 年生存率”进行挖掘尝试。

在这个挖掘模型中，选择了“TNM 分期”即总分期，作为挖掘的条件字段，“生存年限”作为决策字段，此表由 5989 条记录经算法程序得出如

图 2 所示的 8 条规则, 从挖掘算法的规则结果可见, 第一分期的 5 年存活率的可信度达到 35. 58%, 第二分期的 5 年生成率次之 34. 04%, 而相对来说, 第四期的 5 年生存期只有 18. 57%的规则可信度。另一方面, 从规则覆盖度可以看到, 在所有诊治鼻咽癌病人当中, Ⅲ期病人最多, 其覆盖率分别为 50. 86%和 44. 37%, Ⅱ期次之, Ⅳ期病人较少, Ⅰ期病人最少。Ⅳ期 (最晚期) 病人的“5 年以上生存期”的规则可信度最低。这也揭示了鼻咽癌诊治的早发现、早诊断、早治疗的必要性和提高鼻咽癌病人带瘤生存时间的重要性, 具有一定的临床意义。

图 2 基于粗糙集算法的“疾病分期与生存期关系”的规则表

Fig. 2 The rules table of stages of disease and living base on Rough sets arithmetic

4 结 语

粗糙集知识规则的挖掘算法具有广泛的应用前景, 目前数据挖掘技术在各个领域基本处于起步阶段, 有等我们进一步推广应用。要建立一个更完善、更高效的粗糙集数据挖掘系统, 有待进一步设法计算属性的重要度, 以及属性之间的依赖度评估, 找出最小属性集, 以及对规则进一步约简。

参考文献:

[1] PAWLAK Z. Rough sets[J]. International Journal of Information and Computer Science, 1982(11): 341—356

[2] 韩祯祥, 张琦, 文福拴. 粗糙集理论及其应用综述[J]. 控制理论与应用, 1999(2): 153—157.

[3] 王国胤. Rough 集理论与知识获取[M]. 第 3 版. 西安: 西安交通大学出版社, 2003.

[4] (美) MEHMED Kantardzic, 数据挖掘——概念、模型、方法和算法[M]. 闪四清, 陈茵, 程雁等译, 北京: 清华大学出版社, 2003.

Implementation of a Rough Sets Knowledge Mining Algorithm

XIAO Ping, ZHANG Tong, JIANG Di-gang, DENG Jing

(Biomedical Engineering Department of Sun Yat-sen University North Campus, Guangzhou 510080, China)

Abstract: To put forward a data mining system based on the rule of rough set. It is a common system that can mining knowledge from various domain. The data mining system no to limit the number of condition field and decision field in theory. It can mining knowledge from the discrete data and form the rule base.

Key words: data ming; rough sets; rule; confidence