# 变精度粗糙集挖掘技术的应用研究

陈 健[1,2], 赵跃龙[1,3]

(1. 中南大学 信息科学与工程学院,湖南 长沙 410083;
2. 福建省商业高等专科学校, 福建 福州 350012;
3. 华南理工大学 计算机科学与工程学院,广东 广州 510640)

**摘 要**:粗糙集理论是一种处理模糊和不确定知识的一种新型数学工具,在很多领域取得了成功的应用。但是经典粗糙集理论处理的分类必须是完全正确的,在实际应用中,缺乏对噪声数据的适应能力,为了克服这个缺点,提出一种变精度的粗糙集模型,以适应实际应用的需要。对变精度粗糙集理论的数据预处理、属性约简、值约简和规则提取等问题进行了分析和研究,提出属性约简算法和基于求核值属性的归纳值约简算法,并将其运用于医疗系统的手术诊断数据表的数据挖掘分析过程中,所得到的实验结果与专家诊断结果基本吻合,取得了较好的实际应用效果。

**关键词**:变精度粗糙集;属性约简;手术诊断;决策表;数据挖掘

**中图分类号**:TP182 **文献标识码**:A **文章编号**:1673-629X(2008)03-0203-04

# Research of Application of Variable Precision Rough Set

CHEN Jian[1,2], ZHAO Yue-long[1,3]

(1. School of Information Science and Technology, Central South University, Changsha 410083, China;
2. Fujian Commercial College, Fuzhou 350012, China;
3. School of Computer Science and Technology, South China University of Technology, Guangzhou 510640, China)

**Abstract**: Rough sets theory is a new mathematical tool to deal with problems on vagueness and uncertainty. It obtained many achievements in many fields in recent years. Unfortunately, it requires accurate classification in practice, it is lack of the capability of noise data processing. To deal with inconsistency in decision tables, the variable precision rough set model(VPRS) was developed by Prof. W. Ziarko in 1990s. This article introduces data mining process including data pre-process, attribute reduction, value reduction and rule generation based on the VPRS theory. These steps are used to deal with the data mining of medical diagnosis system. The theoretical result is basically identical with that of the qualitative experiment. Finally, the experimental results are used to show the application of VPRS theory and proves the rationality of the theory.

**Key words**: variable precision rough set; attribute reduction; operation diagnosis; strategy table; data mining

## 0 引 言

粗糙集[1](Rough Set, RS)理论由波兰逻辑学家Pawlak[2]教授于1982年提出,由于它能有效地分析和处理不精确、不一致、不完整等各种不完备的信息,并能从中揭示出潜在的规律,所以近年来在机器学习、数据挖掘等多个领域得到了广泛应用。但是传统的粗糙集模型假设全域是已知的,所推出的结论仅适用于全域中的对象,而在实际应用中能够满足此条件是非常困难的。为解决这个问题,Ziarko[3]提出了一种称之为变精度的粗糙集模型,它是粗糙集的直接扩展,该模型给出了错误率低于预先给定值的分类策略,克服了传统粗糙集模型边界区域刻画过于简单的缺点。

文中在介绍知识发现及数据挖掘有关概念的基础上,着重研究将变精度粗糙集方法运用于手术诊断的数据挖掘中。

## 1 变精度粗糙集模型

### 1.1 变精度粗糙集模型简介

变精度粗糙集模型[4]是粗糙集模型的扩充,它是在传统粗糙集模型的基础上引入了 $\beta$,$1-\beta$ 是依赖于数据中噪音程度的一个取值在[0,0.5)上的数,允许一定程度的错误分类率存在。

定义1(变精度粗糙集的下近似集、上近似集、正

**收稿日期**:2007-06-16
**基金项目**:湖南省自然科学基金资助项目(05JJ30120)
**作者简介**:陈 健(1970-),女,福建福州人,硕士研究生,讲师,研究方向为人工智能、数据挖掘和计算机应用等。

域、边界域和负区域)设 $U$ 是论域,$R$ 为 $U$ 上的等价关系,对于 $\beta \in (0.5,1]$,$\forall X \subseteq U$,则 $X$ 关于 $R$ 的下近似集和上近似集的定义为:

(1)$\underline{R}^{\beta}(X) = \{x \in U : D(X/[x]_R) \geqslant \beta\}$,其中 $D(X/[x]_R) = |X \cap [x]_R| / |[x]_R|$

(2)$\overline{R}^{\beta}(X) = \{x \in U : D(X/[x]_R) \geqslant 1-\beta\}$

(3)$\beta$ 正域的定义为 $\mathrm{POS}_R^{\beta}(D) = \bigcup_{Y_i \in U/D} \underline{R}^{\beta}(D_i)$

(4)$X$ 的 $\beta$ 边界域的定义为:$\mathrm{BNR}^{\beta}(X) = \{x \in U : \beta > D(X/[x]_R) > 1-\beta\}$

(5)$X$ 的 $\beta$ 负区域的定义为:$\mathrm{NEG}^{\beta}(X) = \{x \in U : 1-\beta > D(X/[x]_R)\}$

(6)设属性集 $R \subseteq C$,$U/\mathrm{IND}(D) = \{D_1, D_1, \cdots, D_r\}$ 为由决策属性 $D$ 所决定的 $U$ 的划分,$R$ 的 $\beta$— 依赖度可以用如下公式表示:

$$\gamma_R^{\beta}(D) = \frac{|\mathrm{POS}_R^{\beta}(D)|}{|U|}$$

其中 $\mathrm{POS}_R^{\beta}(D) = \bigcup_{Y_i \in U/D} \underline{R}^{\beta}(D_i)$

### 1.2 变精度模型的属性约简

#### 1.2.1 变精度模型的属性约简定义

定义 2[5](属性约简)设属性集合 $R \subset C$,如果满足条件:

1) 对于任意的 $D_j$ 都有 $\mathrm{POS}_R^{\beta}(D_j) = \mathrm{POS}_C^{\beta}(D_j)$;

2) 对于 $\forall P(P \subset R)$,$\exists D_i$ 使得 $\mathrm{POS}_P^{\beta}(D_i) \neq \mathrm{POS}_C^{\beta}(D_i)$(其中:$i = 1,2,\cdots,l$),则 $R$ 为条件属性集 $C$ 关于决策属性集 $D_j$ 的 $\beta$ 约简。

#### 1.2.2 求全部属性约简算法

设 $R$ 的 $\beta$— 依赖度

$\gamma^{\beta}(R,D) = \sum \mathrm{card}(\underline{PY_i}^{\beta})/\mathrm{card}(U)$

输入:已知一个信息系统 $S = (U, A = C \cup D, V, f)$,给定精度 $\beta(0 < \beta < 1)$;

输出:$S$ 的变精度 $\beta$ 下的全部约简。

约简算法如下:

procedure

设 $|C| = m$,则 $C$ 的幂子集 $T_i(C)$ 个数为 $2^m - 1$ 个,$1 \leqslant i \leqslant 2^m - 1$

step1 置 $M = \varnothing$

step2 for $i = 1$ to $2^m - 1$ do

begin

对每个幂子集 $T_i(C)$ 求其 $\beta$— 依赖度 $\gamma^{\beta}(T_i(C), D)$;

if $\gamma^{\beta}(C,D) - \varepsilon \leqslant \gamma^{\beta}(T_i(C),D) \leqslant \gamma^{\beta}(C,D) + \varepsilon$ then 置 $M = M \cup \{T_i(C)\}$;

end for

step3 对 $M$ 中每个幂子集进行反向剔除,保证没有冗余属性。

step4 输出 $M$。

end procedure

#### 1.2.3 值约简与规则提取

对经属性约简后的信息表再进行值约简,删除所有对提取规则无关的属性值,就得到学习后的规则知识。这里介绍基于求核值属性的归纳值约简算法[6]。

定义 $\mathrm{core}(x) = \{a \in C \mid x \in U - \mathrm{POS}_{C-\{a\}}(D)\}$ 表示对象 $x$ 的核值属性集。

用 $\mathrm{core}(D_i)$,$\forall D_i \in U/\mathrm{IND}(D)$ 表示决策类 $D_i$ 的核值属性集。

$\mathrm{DRC}(D_i) = \{x \mid x \in U \wedge f(x,D) = D_i\}$,$\forall D_i \in U/\mathrm{IND}(D)$ 表示决策值为 $D_i$ 的对象集。

算法如下:

1) 任取 $x \in \mathrm{DRC}(D_i)$;

2) if $[x]_{\mathrm{core}(x)} \subseteq \mathrm{DRC}(D_i)$ then 输出决策规则 $\mathrm{des}([x]_{\mathrm{core}(x)}) \Rightarrow \mathrm{des}(f(x,D))$,$\mathrm{DRC}(D_i) = \mathrm{DRC}(D_i) - [x]_{\mathrm{core}(x)}$,转步骤 7);

3) $A_1 = \mathrm{core}(D_i) - \mathrm{core}(x)$,$A_2 = U - \mathrm{core}(D_i)$,用 $\gamma_{C-\{a\}}(D) = |\mathrm{POS}_{C-\{a\}}(D)| / |U|$ 对 $A_1, A_2$ 中的元素由大到小排序,得有序集 $OA$,设 $|OA| = m$,按照组合数由 1 到 $m$ 产生 $OA$ 的有序幂子集 $T_i(OA)$,$i \in 1 \cdots 2^m - 1$;

4) 令 $i = 1$;

5) $B = \mathrm{core}(x) \cup T_i(OA)$,if $[x]_B \subseteq D_i$ then 输出决策规则 $\mathrm{des}([x]_B) \Rightarrow \mathrm{des}(f(x,D))$,$\mathrm{DRC}(D_i) = \mathrm{DRC}(D_i) - [x]_B$,转步骤 7);

6) $i = i + 1$;if $i \leqslant 2^m - 1$ then 转步骤 5);

7) if $\mathrm{DRC}(D_i) \neq \varnothing$ then 转步骤 1);

8) 算法结束。

根据以上步骤,依次求得各决策类 $D_i \in U/\mathrm{IND}(D)$ 的最小决策规则集,可以得到整个决策表的最小决策规则集。

## 2 应用实例

变精度粗糙集理论被广泛应用于知识发现的各阶段,包括数据准备、数据预处理、属性约简、值约简、生成决策规则及提取等内容。下面将从这几个方面对它在手术诊断中的实际应用进行介绍,分析如何将变精度粗糙集理论模型应用于手术诊断。

### 2.1 数据准备

数据准备阶段主要是病人数据的收集准备工作。病人在住院期间有大量的相关数据,如病情诊治信息、手术诊断信息、病人基本信息、费用等,其中有些是与

手术诊断无关的信息。数据准备阶段是要搜集并整理与手术诊断相关的病人数据,如手术诊断、手术类别、麻醉情况、手术环境等。

### 2.2 数据预处理

数据预处理主要包括数据的理解、属性选择、连续属性离散化、数据中的噪声及丢失值处理、实例选择等。数据预处理的目的是提供可以进行数据挖掘和知识发现的数据。粗糙集预处理的任务是确定条件属性和决策属性,建立信息表。

(1)属性选择。数据挖掘对象是院内病人的各项病情实测数据,挖掘目标是发现与手术诊断相关的观测知识和预测知识。将手术诊断信息表、病人基本信息表、主刀医师基本信息表通过属性键值联表查询获得手术诊断决策信息表(文中约70000条记录),这样可以得到与手术诊断相关的一些属性,如手术类别、病人性别、病人年龄、麻醉情况、手术环境、诊断类别等。

(2)连续属性离散化。一般病人信息的部分属性需要离散化,以便作定量分析。如疾病代码(BH)按照《国际疾病代码标准》分类有大约100个大类,小数点后四位按照《中国疾病代码标准》细分为小类(比如疾病代码55对应手术大类泌尿生殖系统,疾病代码55.7001对应手术名称是肾固定术)。考虑手术大类有100个左右,可以按照感官功能进一步泛化成12类(比如疾病代码为55.0000－75.9999可泛化成第10类,对应手术大类是泌尿生殖系统)。

(3)噪声及丢失值处理。数据库中的噪声和缺值是一种常见现象。对病人数据中存在的噪声和缺失值的处理比较简单:删除噪声数据,以正常值代替缺失值。

(4)属性编码。每个属性用一个数据库字段表示,不同的病情属性取值不同。在已获得手术诊断决策信息表(约70000条记录)中有部分记录缺损得比较严重,考虑把这些记录删除,同时删除部分对诊断决策不产生影响的属性和冗余的属性后,将获得一个完备的手术诊断决策信息表(共53112条记录,10个条件属性为{手术类别, 病人性别, 病人年龄, 麻醉情况,诊断类别, 主刀职称, 手术环境,住院日期, 出院日期, 手术日期},1个决策属性为{术后情况})。

### 2.3 数据挖掘

粗糙集理论的知识发现通过对决策表的知识约简来完成,在数据挖掘阶段,先用前述求全部属性约简算法对决策表进行属性约简,取 $\varepsilon = 0.0002, \beta = 0.75$,可获得手术诊断决策表的全部约简。决策属性为SHQK,$POS_R^{\beta}(SHQK) = 52058$,$\gamma_C^{\beta}(SHQK) = 98.02\%$。表1为一种手术诊断决策表的约简表,图1为实验所得17个属性分布情况。

表1　一种手术诊断决策表的约简表(部分)

| 条件属性集合 | $POS_R^{\beta}(D)$ | $\gamma_R^{\beta}(D)$ |
|---|---|---|
| {手术类别,病人性别,麻醉情况,手术环境} | 52055 | 98.01% |
| {手术类别,麻醉情况,主刀职称,手术环境} | 52066 | 98.03% |
| {手术类别,病人性别,麻醉情况,诊断类别,手术环境} | 52055 | 98.01% |
| {手术类别,病人性别,麻醉情况,手术环境,出院日期} | 52062 | 98.02% |
| {手术类别,病人性别,主刀职称,手术环境,住院日期} | 52049 | 98.00% |
| {手术类别,麻醉情况,诊断类别,主刀职称,手术环境} | 52066 | 98.03% |
| {手术类别,病人性别,病人年龄,麻醉情况,手术环境,手术日期} | 52048 | 98.00% |
| {手术类别,病人性别,麻醉情况,诊断类别,手术环境,出院日期} | 52062 | 98.02% |
| {手术类别,病人性别,麻醉情况,手术环境,住院日期,手术日期} | 52064 | 98.03% |
| {手术类别,病人性别,诊断类别,主刀职称,手术环境,住院日期} | 52049 | 98.00% |
| {手术类别,病人年龄,主刀职称,手术环境,住院日期,出院日期} | 52051 | 98.00% |

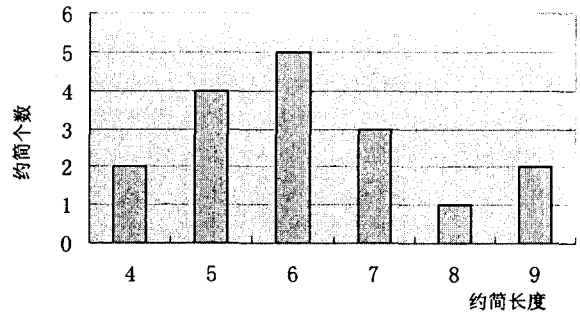


图1　实验所得17个属性分布情况

注:约简长度表示组成该约简的条件属性集合中条件属性的个数。

然后用基于求核值属性的归纳值约简算法进行值约简。由经约简后的决策表导出规则知识,最后,对挖掘结果加以解释并转换成易于理解的显式知识。在实例中,采用 $\beta = 0.75$ 下近似约简得出的最小约简{手术类别,病人性别,麻醉情况,手术环境}来浓缩决策表,最后得到的规则集如下:

手术环境=1→术后情况=1

手术类别=1 and 手术环境=2→术后情况=1

手术类别=10 and 手术环境=2→术后情况=1

手术类别=10 and 麻醉情况=1→术后情况=1

手术类别=10 and 麻醉情况=2→术后情况=1

手术类别=11→术后情况=1

手术类别=12 and 手术环境=2→术后情况=1

手术类别=2→术后情况=1

手术类别=3→术后情况=1

麻醉情况=2 and 手术环境=2→术后情况=1

麻醉情况=3 and 手术类别=4→术后情况=1

手术类别=5 and 手术环境=2→术后情况=1

手术类别=6→术后情况=1

手术类别=7→术后情况=1

手术类别=8→术后情况=1

手术类别=9→术后情况=1

手术类别=12 and 手术环境=3→术后情况=2

手术类别=4 and 手术环境=3 and 病人性别=2→术后情况=2

手术类别=10 and 手术环境=3 and 麻醉情况=5→术后情况=3

手术类别=4 and 手术环境=2 and 麻醉情况=1→术后情况=3

手术类别=5 and 手术环境=3→术后情况=3

## 3 结 论

将变精度粗糙集理论的数据挖掘技术应用于手术诊断的知识发现,提出属性约简算法和基于求核值属性的归纳值约简算法,在实际应用中,在智能诊断的知识自动获取方面取得新的进展,具有良好应用前景。因为粗糙集基于约简的思想,能将知识高度浓缩,所以最后得到的规则知识相对较少,也更有效,而且由最终约简后的信息表得到的规则知识非常直观。

**参考文献:**


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

[2] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science,1982,11(5):314-356.

[3] Ziarko W. Variable Precision Rough Set Model[J]. Journal of Computer and System Sciences,1993,46(1):39-59.

[4] 黎东英,王应明. 基于可变精度粗集理论的规则挖掘模型[J]. 计算机测量与控制, 2005,13(8):833-839.

[5] 米据生,吴伟志,张文修. 基于变精度粗糙集理论的知识约简方法[J]. 系统工程理论与实践,2004(1):76-82.

[6] 代建华,李元香. 一种基于粗糙集的决策系统属性约简算法[J]. 小型微型计算机系统,2003,24(3):523-526.


(上接第 153 页)


究与发展,2001,38(4):405-414.

[3] Brin S,Page L. The Anatomy of a Large-scale Hypertextual Web Search Engine[C]//In:Proceedings of the 7th International World Wide Web Conference. Australia:[s.n.],1998.

[4] Kleinberg J M. Authoriative Sources in a Hyperlinked Environment[J]. Journal of the ACM,1999,46(5):604-632.

[5] Pawlak Z. Rough Set Theory and Its Application to Data Analysis[J]. Cybernetics and Systems,1998,29(9):661-668.

[6] 曾黄麟. 粗集理论及其应用[M]. 重庆:重庆大学出版社,1998.

[7] Pitkow J E. Characterizing World Wide Web Ecologies[D]. Georgia:Georgia Institute of Technology,1997.

[8] Weise R,Veles B. HyPursuit:A Hierarchical Network Search Engine that Exploits Content-link Hypertext Clustering [C]//In Proceedings of the 7th ACM Conference on Hypertext. Bethesda,Maryland, United States:[s.n.],1996.

[9] Spertus E. Parasite:Mining Structural Information on the Web [C]//In:Proceedings of the Sixth International World Wide Web Conference. Santa Clara, California, United States:[s.n.],1997.


(上接第 202 页)

从测试结果可以看出,文中提出的算法对远距离车辆的检测有较好的效果,并且降低了误检率。下一步工作,将要进一步探索和研究多种类型车辆和车辆遮挡的检测以及车辆 ROI 的精确检测。

**参考文献:**


[1] Sun Zehang, Bebis G. On-road vehicle detection using Gabor filters and support vector machines[C]// International Conference on Digital Signal Processing. Greece:[s.n.], 2002.

[2] Betke M. Multiple Vehicle Detection and Tracking in Hard Real-Time[C]// IEEE Symposium on Intelligent Vehicles. France:[s.n.],2002.

[3] Bensrhair A, Broggi A. Stereo Vision-based Feature Extraction for Vehicle Detection[C]// IEEE Symposium on Intelligent Vehicles. France:[s.n.]2002.

[4] Srinivasa N. A vision-based vehicle detection and tracking method for forward collision warning[C]// IEEE Intelligent Vehicle Symposium. France:[s.n.],2002.

[5] Bensrhair A, Broggi A. A cooperative approach to vision-based vehicle detection[C]// Proceedings IEEE Intelligent Transportation Conference. Oakland,CA,USA:[s.n.],2001: 209-214.

[6] Graefe V,Efenberger W. A Novel Approach for the Detection of Vehicles on Freeways by Real-time Vision[C]//in Procs. IEEE Intelligent Vehicles Symposium '96. Tokyo,Japan:[s.n.],1996:363-368.

[7] Shannon C E. A mathematical theory of communication[J]. Bell System Technical Journal,1948,27(3-4):379-423; 623-656.