基于 Rough Sets 的中医指症挖掘研究与应用

丁卫平[1],管致锦[1,2],顾春华[3]
DING Wei-ping[1],GUAN Zhi-jin[1,2],GU Chun-hua[3]

1.南通大学 计算机科学与技术学院,江苏 南通 226019
2.南京航空航天大学 信息科学与计算机学院,南京 210003
3.南通市中医院,江苏 南通 226006
1.School of Computer Science and Technology,Nantong University,Nantong,Jiangsu 226019,China
2.College of Information Science and Technology,Nanhang University,Nanjing 210003,China
3.Nantong China Medicine Hospitial,Nantong,Jiangsu 226006,China
E-mail:dwp9988@163.com

DING Wei-ping,GUAN Zhi-jin,GU Chun-hua.Research and application on TCM diagnosis mining base on Rough Sets. Computer Engineering and Applications,2008,44(7):234-237.

Abstract: According to the problems existing which include the much larger numbers of dimensions of the sample,much heavier data features and redundant attribution of TCM (Traditional Chinese Medicine) diagnosis,the paper provid the GENRED_GROWTH TCM diagnosis algorithm combining the frequency with the importance of attribution on the basic of research and study of the basic theory and attribution reduction of Rough Sets.Meanwhile the algorithm is applied in the prototype data mining experiment on TCM diagnosis.The results show that the algorithm can be used to reduce redundant attributes in data mining on TCM diagnosis,and the classification accuracy is higher.And it performs accurately on TCM diagnosis datasets.The useful information for the diagnosis can be extracted in order to help to provide some decision-making for the assistant diagnoses.
Key words: Rough Sets;attribute reduction;TCM (Traditional Chinese Medical) diagnosis;data mining

摘 要:针对中医病历数据库中指症样本维数较大、数据特征和属性冗余量较多等特征,在对 Rough Sets 基本理论和属性约简算法研究的基础上,提出了将属性频度和属性重要性相结合的 GENRED_GROWTH 中医指症挖掘算法,并进行了基于 GENRED_GROWTH 的中医指症挖掘原型系统设计与实现。通过分析和实验结果表明:该算法能较好地进行中医指症属性约简,分类精度较高,并且能抽取中医指症相关诊断规则以辅助医生的诊断和治疗。
关键词:Rough Sets;属性约简;中医指症;数据挖掘
文章编号:1002-8331(2008)07-0234-04 **文献标识码**:A **中图分类号**:TP301

中医作为中华民族特有的文化和科学遗产,对世界有着重大的贡献。中医指症研究是中医理论研究的难点,中医诊断主要根据病人的症状来辨证判断,即所谓的“望、闻、问、切”等复杂诊断过程,其证型之间、症状之间错综交叉,辨证认治十分困难。然而病症是可以依症进行分类的,症型划分是否准确,对于中医疾病的诊断结局和转归有着极为重要的意义。由于中医的模糊性和不确定性等特点,要想建立一个数据库进行症型智能化辅助划分相对比较困难,因为中医病历数据库中存在大量冗余数据,对这些冗余的数据若不进行一定的处理、分类,将很难对实际医疗起到实质性的作用,有时甚至还会干扰医生的正常判断,严重时还有可能导致误诊等[2-4]。所以在不影响精度的前提下,必须寻求一种有效的处理中医指症病历分类和挖掘方法。

针对这上述问题,目前有很多种方法开展了这方面的研究,如文献[2,3,7,8]中所述,但这些方法或多或少存在一些不足:如神经网络方法不能自动地选择合适的属性集,知识表达和解释能力很差,网络参数缺乏物理意义,并且在学习过程中容易陷入局部极值;模糊集合论的方法只能处理模糊类数据,不能处理不完整数据,而且还要提供隶属函数这种先验知识。这两种方法均过多的依赖于专家知识,缺乏学习和适应的能力[5]。Rough Sets(粗糙集)方法是一种新型的用于处理模糊和不精确性数据的数学工具。近年来,由于其在处理不确定和不完备信息方面的独特优势,赢得了越来越多研究者的关注,并在不同的领域中得到了广泛的应用。它不需要先验知识,既能处理模糊类的数据和不完整数据,而且也能从数据库中发现异

基金项目:江苏省高校自然科学基础研究项目(No.02KJB520004,No.05KJB520107);南通大学自然科学基金项目(No.05Z061);南通大学通信与信息系统学科科技创新资助。
作者简介:丁卫平(1979-),男,讲师,研究方向为粗糙集、概念格和电子病历挖掘等;管致锦(1962-),男,教授,博士,硕导,研究方向为量子可逆计算等;顾春华(1976-),男,工程师,主要研究方向中医电子病历。
收稿日期:2007-07-02 **修回日期**:2007-09-24

常，排除噪声干扰，较好地进行知识发现[1,6]。

本论文对 Rough Sets 基本理论和属性约简算法研究的基础上，提出了将属性重要性和属性依赖性相结合的 GENRED_GROWTH 中医指症挖掘算法，通过分析和实验结果表明该算法能较好地去除中医指症中冗余属性，分类精度较高，简化抽取出中医指症病历中诊断规则，能较好地辅助医生相关诊断和治疗。

1　Rough Sets 基本理论

Pawlak 提出的 Rough Sets 理论是智能数据分析和数据挖掘中一种新的数学方法[1]，它从新的视觉出发对知识进行了定义，它把知识看作是关于论域的划分，并引入代数学中的等价关系来讨论知识。该理论已渗透到各个领域，如机器学习、决策分析、模式识别等，并且取得了较好的应用效果。

1.1　基本 Pawlak 模型[6]

给定一个有限的非空对象集合 U，称为论域。$R\subseteq U\times U$ 表示 U 上的一个等价关系，代表关于 U 的知识。等价关系 R 将集合 U 划分成不相交的子集，记作 U/R，其表示 R 的所有等价类族。

二元偶对 $apr=(U,R)$称为近似空间。如果 U 上两个元素 x 和 y 属于相同的等价类，则称 x 和 y 是不可分辩的。R 的等价类和空集 Ø 称为近似空间 $apr=(U,R)$的原子集合。多个原子集合的并集称作复合集合，所有复合集合(包括空集)的族，表示为 $Com=(apr)$。

任意给定一个集合 $X\subseteq U$，是一个等价关系，如果使用 R 等价关系无法精确描述 X，则 X 就是 R 的粗糙集；反之 X 是的 R 精确集。

一般来说，粗糙集可以用两个精确集合：下近似(集)和上近似(集)来加以定义：

$\underline{R}X=\{x|x\in U,且[x]_R\subseteq X\}$为集合 X 的 R 下近似集；

$\overline{R}X=\{x|x\in U,且[x]_R\cap X\neq\Phi\}$为集合 X 的 R 上近似集。

1.2　属性约简和核方法

在对象识别或决策系统等实际应用中，众多属性进行约简或规约是数据挖掘的关键，其主要目的是要找出最相关的(最重要的)属性，去除不相关的(或不重要的)属性，同时又不丢失原始信息。进行属性约简中主要用到两个概念：规约和核心。属性的规约是属性的必不可少的子集，它足够用来对出现在所考虑的知识中所有基本概念进行定义，而核心是属性最重要的部分[7]。

一般通过以下步骤来进行属性约简：

(1)通过分辨矩阵求出属性规约集的核心

核心与矩阵有着紧密的联系，核心是由分辨矩阵内仅含单个元素的单元中的元素构成的集合，根据下面这个定理，可以非常容易地从分辨矩阵中计算出核心。

(2)运用一定的属性约简算法计算规约集，并根据某种评判准则确定最佳规约集或者用户定义最小属性集。

1.3　规则获取方法

在多值信息系统 $M=(U,A=C\cup\{d\},D)$中，对于某个 $r\in U$，某些 $B\subseteq C$，如果有$[r]_B\subseteq[r]_d$，则存在规则 $\wedge_{a\in B}x(a)\cap s(a)=\Phi\Rightarrow x(d)\cap r(d)\neq\Phi$。如果 B 是满足$[r]_B\subseteq[r]_d$的最小属性集，那么得到的规则是最小规则。

Rough Sets 规则获取一般采用数据预处理、数据约简、规则生成、数据依赖关系获取等步骤。首先经过属性约简，去除了一些不必要的、多余的属性后，复杂繁多的数据集已变得简单易懂，再通过使用面向属性集的方法从规约后的数据集中可以进行规则的获取，删除数据集中不属于最佳规约集或用户定义最小属性子集的属性，最后将简化后的数据集中元组合并相同或相似的元组，对每一个决策类，可将约简后的数据集中相关元组转化为决策规则，进行有关规则的获取。

2　一种改进的 GENRED_GROWTH 中医指症挖掘算法

中医诊断是一个复杂的过程，证型之间、症状之间错综复杂，数据冗余量较大，辨证认治非常困难。如何利用现代化的科学技术手段抽取挖掘出中医指症诊断规律是中医诊断现代化关键问题。目前 Rough Sets 方法在中医诊断方面取得了较好的应用，其中经典的 GENRED 属性约简算法就是其中较好的例子[10]，该算法以核为起点，基于属性重要性来进行排序，每次选择重要性最大的属性加入预置的属性集中，然后再通过反馈过程进行斟酌筛选，剔掉多余属性，保留必要属性，最终得到最佳规约集或用户定义最小属性集[8]。但是从实际应用中可发现该算法存在一些缺点：

(1)利用该算法所求得的中医指症属性约简往往不能保证是最小属性约简，存在较多的冗余量，这对中医指症的实际挖掘结果带来较大的不确定性；

(2)对于大规模的中医诊断系统，该算法需存储一个较大的可辨识矩阵，这样将占用了大量的计算机内存，运行速度较慢，不易于中医诊断计算机智能化快速实现。

本文在分析上述问题的基础上，提出了一种改进中医指症挖掘算法，即 GENRED_GROWTH 算法，该算法将属性重要性和属性依赖性相结合，属性约简和分类效果较好。

2.1　基本定义

(1)属性的重要性

不同属性对于决定条件属性和决策属性之间的依赖关系起着不同的作用。属性 a 加入 R，对于分类 $U/IND(P)$的重要性定义如下：

$$SGF(A,R,P)=\gamma_R(P)-\gamma_{R-\{a\}}(P)$$

其基本含义为将属性 a 加入到属性集 R 中，体现属性集 R 对属性集 P 的依赖程度，以此来体现属性 a 的重要性。可以得出属性 a 的重要性是相对而言的，它依赖于属性集 P 和 R。因此在不同的背景下，属性的重要性可能不同。

(2)属性的依赖性

在数据规约中，利用两个属性集合 $P,R\subseteq Q$ 之间的相互依赖程度，可以确定一个属性 a 的重要性。属性集 P 对 R 的依赖性用 $\gamma_R(P)$表示，定义如下：

$$\gamma_R(P)=\frac{card(POS_R(P))}{card(U)}$$

$$POS_R(P)=\bigcup_{X\in U/IND(P)}\underline{apr}_R(X)$$

其中 $card(\cdot)$表示集合的基数。$POS_R(P)$是属性集 R 在 $U/IND(P)$中的正区域，即下近似。

2.2 算法基本思想

在本文中提出的改进 GENRED_GROWTH 算法主要依据属性在可辨识矩阵中出现的频率和长短来定义属性的重要性和依赖性作为启发信息,主要基于两个重要的准则:

(1)属性在构造的可辨识矩阵中出现的次数越多,则属性的重要性越大;

(2)属性在构造的可辨识矩阵中的项长度越短,则属性的依赖性越大。

利用求核属性的方法,先计算核属性的频率;然后将属性中包含核属性的项删除,再重新计算属性项的依赖度,重复以上操作。这时,利用属性频率为启发知识来判断属性间的重要性时,同时可以约简属性间的依赖关系,使其属性频率更能反映单个属性重要性和属性间的依赖关系。

2.3 算法实现步骤

一个信息系统 $S=(U,Q,V,f)$,$\boldsymbol{M}$ 是 S 的分辨矩阵,$\boldsymbol{M}$ 的元素为 U_{ij},U_{ij} 是 S 中的第 i 个对象和第 j 个对象的有差别的所属性的集合。$Q=\{A_1,A_2,A_3,\cdots,A_n\}$是 S 的所有属性的集合,C 为条件属性集,D 为决策属性集,$U_{ij}\subseteq Q$,$A_k\in U_{ij}$。

令 $SGF(A_k)$是属性 A_k 的重要性,$P(A_k)$是 $\boldsymbol{M}$ 中属性 A_k 与其它属性的依赖函数。算法描述如下:

Co 是 $\boldsymbol{M}$ 的核心集合,并且 $R=Co$(R 为属性约简后的属性集)。

(1)在分辨矩阵中找出 Co;

(2)$B=\{U_{ij}:U_{ij}\cap R\neq\varnothing,i\neq j,i,j=1,2,\cdots,n\}$,$E=\boldsymbol{M}-B$;

(3)对所有 $A_k\in E$,计算其在 E 中的 $P(A_k)$,且令 $P(A_k)=\underset{k}{\mathrm{MAX}}\,|P(A_k)|$;

(4)$R\leftarrow R\cup\{A_q\}$,如果 $P(A_q)=P(Ap)=\underset{k}{\mathrm{MAX}}\,|P(A_k)|$,则计算各自的重要性与依赖性,若 $SGF(A_q)>SGF(A_p)$,则 $R\leftarrow R\cup\{A_q\}$,反之,$R\leftarrow R\cup\{A_p\}$。若 $SGF(A_q)=SGF(A_p)$,则选择属性组合数最少的一个加入到 R 中。

(5)重复上述过程,直到 $E=\varnothing$。

2.4 算法比较分析

下面用实验来进行 GENRED_GROWTH 算法和 GENRED 算法在分类精度比较。实验在一台 Pentium2.4 G,内存 256 M的 PC 机上运行,系统是 Microsoft Windows 2000 Professional,所有程序用 Microsoft/Visual C#.net 编写。实验数据来源于南通中医院电子病历数据库,经过数据预处理中医指症样本集 11 906 条记录,包含 100 多个症状。

在逐步增加记录数时比较 GENRED_GROWTH 算法和 GENRED 算法的分类精度,实验结果如下图 1 所示。可见随着中医指症记录数的增加,GENRED_GROWTH 算法分类精度要明显高于 GENRED 算法。

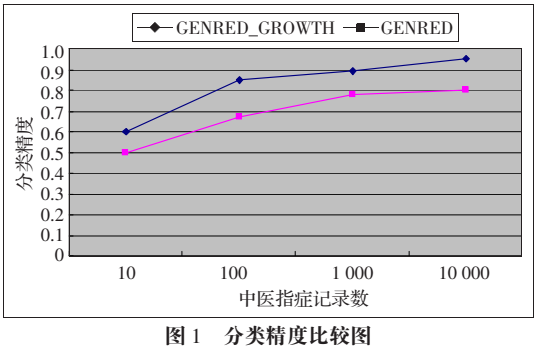

图 1 分类精度比较图

3 中医指症病历挖掘原型系统设计与实现

为了进一步验证 GENRED_GROWTH 算法在中医指症挖掘方面的有效性,设计了一个中医指症病历挖掘原型系统。这个系统主要根据用户收集的中医病历,在数据预处理后,进行相关属性约简,得出相关中医病例的诊断知识,对实际诊断提供一定的辅助诊断作用[9]。

3.1 中医指症病历预处理

在这个系统中,主要中医指症病历数据源是关于“寒证”的属性,要从一系列的症状(条件属性)中提取出有价值的属性信息,去除冗余的信息,进行行属性约简,分类和相关症状训练[6]。其主要数据如下:

(1)寒证(决策属性)

(2)各个症状:

口渴(不渴:1),(渴:0);

四肢(怕冷:1),(不怕冷:0);

脉象(迟/紧:1),(非迟/紧:0);

发热(发热:1),(不发热:0);

倦怠乏力(倦怠乏力:1),(不倦怠乏力:0);

咳嗽(咳嗽:1),(不咳嗽:0);

头晕耳鸣(头晕耳鸣:1),(不头晕耳鸣:0);

胸闷(胸闷:1),(不胸闷:0)。

上述症状中 1:表示病例表现出该症状;

0:表示病例没有表现出该症状。

该实验是从上述症状中提取出“寒证”必要的症状表现,剔除其冗余属性。

3.2 系统设计步骤

(1)构造原始决策表

将上述数据源转换为实验所需的原始决策表,如表 1 所示。

表 1 “寒证”原始决策信息表

对象	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$	$A8$	$A9$
$U1$	1	1	1	1	0	0	0	1	1
$U2$	0	1	0	0	1	1	0	0	1
$U3$	1	1	1	1	1	0	0	1	1
$U4$	1	1	1	0	1	0	0	1	1
$U5$	0	0	0	0	0	0	0	1	0
$U6$	0	0	0	0	1	1	0	0	0
$U7$	0	1	1	0	1	1	0	0	0
$U8$	0	0	1	1	0	1	0	0	0
$U9$	0	1	0	0	1	0	1	1	0

(2)属性和对象抽象表示

用 $U1\sim U9$ 表示 9 个对象(病例);

$A1\sim A9$ 表示 9 个属性,每个属性对应一个症状名,即 $A1$:发热;$A2$:口渴;$A3$:四肢;$A4$:倦怠乏力;$A5$:咳嗽;$A6$:脉象迟/紧;$A7$:头晕耳鸣;$A8$:胸闷;$A9$:寒证。其中 $A1\sim A8$ 是条件属性,$A9$ 是决策属性。

(3)等价类划分

将 $A9$ 属性划分:$\{U1,U2,U3,U4\}$,$\{U5,U6,U7,U8,U9\}$,其中集合$\{U1,U2,U3,U4\}$是患有“寒证”,集合$\{U5,U6,U7,U8,U9\}$是没患“寒证”。

(4)构造分辨矩阵(表 2)

(5)属性约简

运用改进的 GENRED_GROWTH 算法进行属性约简,得到初步属性约简表,如表 3 所示。

表 2　分辨矩阵 M

对象	$U1$	$U2$	$U3$	$U4$
$U5$	A1A2A3A4	A2A5A6A8	A1A2A3A4A5	A1A2A3A5
$U6$	A1A2A3A4A5A6A8	A2	A1A2A3A4A6A8	A1A2A3A6A8
$U7$	A1A4A5A6A8	A3	A1A4A6A8	A1A6A8
$U8$	A1A2A6A8	A2A3A4A5	A1A2A5A6A8	A1A2A4A5A6
$U9$	A1A3A4A5A7	A6A7A8	A1A3A4A7	A1A3A7

表 3　属性约简表

对象	$A2$(口渴)	$A3$(四肢)	$A6$(脉象迟/紧)	$A9$(寒证)
$U1$	1	1	0	1
$U2$	1	0	1	1
$U3$	1	1	0	1
$U4$	1	1	0	1
$U5$	0	0	0	0
$U6$	0	0	1	0
$U7$	1	1	1	0
$U8$	0	1	1	0
$U9$	1	0	0	0

(6)优化属性约简

通过合并相同元组，进一步优化属性，其最优约简表如表 4 所示。

表 4　最优属性约简表

对象	$A2$(口渴)	$A3$(四肢)	$A6$(脉象迟/紧)	$A9$(寒证)
$U1U3U4$	1	1	0	1
$U2$	1	0	1	1
$U5$	0	0	0	0
$U6$	0	0	1	0
$U7$	1	1	1	0
$U8$	0	1	1	0
$U9$	1	0	0	0

3.3　实验结果与分析

用 C# 高级编程语言实现的中医“寒证”指症数据挖掘系统如图 2 所示。首先从数据库中提取出相关的中医“寒证”相关属性：口渴，四肢，脉象，发热，倦怠乏力，咳嗽，头晕耳鸣，胸闷等，并进行计算机处理的中医数据数字化加载到系统，通过等价划分得到分辨矩阵，从这个矩阵中能够很容易的得出核，通过核使得易于进一步进行属性约简，再进行相同元组的合并分类，进一步简化得到最终结果，如图 2 所示。

图 2　属性约简训练结果图

在上述详细设计过程中，原始的决策表中含有 9 个属性，其中“寒证”为决策属性，其余 8 个属性为条件属性，但是这 8 个条件属性中有些症状属性对于“寒证”来说是冗余的，使用改进的 GENRED_GROUP 算法，从图可以看出属性约简结果相比原始的决策信息表提取出了有用的属性信息，即口渴、四肢、脉象迟/紧 3 个属性是“寒证”的必然表现症状，其余的症状属性对于“寒证”来说是冗余属性，这对实际诊断具有一定的指导价值。并且在实际样本训练“寒证”分类精度较高，样本训练速度得到了明显的提高，当然可以对上述数据进行扩充，加大中医指症数据量和属性数，从实验中得出中医“寒证”指症的决策属性判别分类所得结果和临床医生常规判断所得出结果进行比较，实践证明结果基本达到一致，从而证实了中医指症挖掘的客观性和科学性。

4　结束语

近年来 Rough Sets 方法迅速发展，它在许多方面有着不可替代的优越性和广泛的应用，而目前将 Rough Sets 方法应用到中医指症的研究还很少。本文在对 Rough Sets 基本理论和属性约简算法研究的基础上，提出了 GENRED_GROWTH 中医指症挖掘算法，并进行了相关原型系统的设计，结果表明该算法能够较好地进行中医指症数据属性约简，分类精度较高，能较好挖掘出中医病历中相关症状，为中医诊断提供相关辅助性决策。

目前 Rough Sets 方法在中医指症方面的应用研究国内尚处于起步阶段，很多方面还有待计算机研究人员和医务工作者共同研究，进一步促进我国中医诊断向智能化方向发展。

参考文献：

[1] Pawlak Z.Rough sets[J].International Journal of Computes and Information Science，1982，11(5)：341-356.

[2] Kononenko I.Machine learning for medical diagnosis：history，state of the art and perspective[J].Artificial Intelligence in Medicine，2002，23：89-109.

[3] Tsumoto S.Automated discovery of positive and negative knowledge in clinical databases[J].IEEE Engineering in Medicine Biology，2000，4(19)：56-62.

[4] 姚美村，袁月梅，艾路，等.数据挖掘及其在中医药现代化研究中的应用[J].北京中医药大学学报，2002，25(5)：20-23.

[5] 秦中广，毛宗源.粗糙神经网络及其在中医智能诊断系统中的应用[J].计算机工程与应用，2001，37(18)：34-35.

[6] 陈文伟，黄金才，赵新昱.数据挖掘技术[M].北京：北京工业大学出版社，2002.

[7] 胡彧，李智玲，李春伟.一种基于区分矩阵的属性约简算法[J].计算机工程与应用，2007，43(9)：178-180.

[8] 朱金伟，鞠时光，辛燕.基于数据挖掘的中医药数据预处理方法[J].计算机工程，2006，32(15)：280-282.

[9] 张召.支持向量机在中医指症数据挖掘中的应用研究[D].上海：华东师范大学，2003.

[10] 秦中广.基于粗糙集的交叉研究及其在中医诊断的应用[D].广州：华南理工大学，2002.