◎工程与应用◎

商集统计 Rough sets 及其医学辅助诊断模式

聂 斌,王命延,邱桃荣,于海雯,方 娜
NIE Bin,WANG Ming-yan,QIU Tao-rong,YU Hai-wen,FANG Na

南昌大学 计算机系,南昌 330031
Department of Computer Science,Nanchang University,Nanchang 330031,China
E-mail:ncunb@163.com

NIE Bin,WANG Ming-yan,QIU Tao-rong,et al.Model to assistant medical diagnosis based on quotient set and statistics. Computer Engineering and Applications,2008,44(35):197-199.

Abstract: Without a correct diagnosis,a correct treatment may not be made in medical science.The paper put forward a model to assistant medical diagnosis based on quotient set and statistics,in connection with the phenomenon that the work of collect numerous of confirmed cases,false positive and false negative,and the request of diagnosis.It is proved to be feasible and effective after tested with a database,and a good proposal,which could assist medical diagnosis in accordance with diagnostic needs.
Key words: quotient set and statistics;rough exclusive;tolerance;diagnosis model

摘 要:在医学上没有正确的诊断,就没有正确的治疗。针对收集的确诊病例多,因假阳性或假阴性等造成误诊和漏诊的情况,以及根据诊断匹配时的要求不同,提出一种基于商集统计 Rough sets 的医学辅助诊断模式。通过对数据测试表明,该模式是可行有效的,是一种能根据临床需要进行辅助诊断的方案。
关键词:商集统计;粗糙排斥;容假度;诊断模式
DOI:10.3778/j.issn.1002-8331.2008.35.059 **文章编号**:1002-8331(2008)35-0197-03 **文献标识码**:A **中图分类号**:TP399

1 引言

Rough 集理论是由波兰科学家 Z.Pawlak[1]在 20 世纪 80 年代初提出,近年来得到了迅速的发展和完善,是一种用于处理不确定性、不分明性和模糊性知识的软计算方法,其基本思想是在保持分类能力不变的前提下,通过知识约简,导出概念的分类规则,并从中发现隐含的知识,揭示出潜在的新知识和规律。目前,Rough 集理论在许多[2-10]方面都有成功的研究和应用。

文中基于商集统计 Rough sets 的医学辅助诊断模式采用的方法分 4 步:第 1 步,分层或分组,即根据商集统计法预处理, 再根据症状特征化后的属性-属性值对求出各决策属性两两之间的排斥域,然后根据排斥度从大到小依次层层分出两个组别;第 2 步,每组依次提取规则;第 3 步,合并规则,即根据需要合并第 2 步中提取的规则, 得到每一种疾病的诊断规则;第 4 步,在给定容假度阈值条件下,对实例进行匹配,得到诊断结果,辅助临床诊断。

2 Rough 集基本理论

2.1 信息系统

一个信息系统可以定义为 $S=(U,A\cup D)$,其中 U 是一个非空、有穷、被称为全域的个体的集合,A 是非空、有穷的属性集合,即对于属性 $a\in A$,有 $a:U\to V_a$,其中 V_a 被称作属性 a 的值集 $V=\bigvee_{a\in A}V_a$;集合被说成是属性集 A 的值区域[11-12],D 被称作决策属性集。例:表 1 为糖尿病数据库中随意截取的子集并用简化分明矩阵法约简后的信息表系统,U={1,2,3,4,6},条件属性集 A={a,b,c,e}, 决策属性集 D={Eye diabetes,Kidney diabetes,Angiocarpy diabetes}。属性 $a\in A$,V_a属性值有{0,1},Eye diabetes 是决策属性。

表 1 信息表

No	a	b	c	e	d
1	0	1	1	0	Eye diabetes
2	0	1	1	0	Eye diabetes
3	0	1	0	0	Eye diabetes
4	1	0	1	0	Kidney
5	1	0	0	0	Kidney
6	0	0	0	1	Angiocarpy

注:a 为 albuminuria,b为retina,c为hand and foot,e 为 angina,d 为diabetes type,1为abnormal,0为normal

2.2 商集统计

商集统计,是按商集[13]思想把复杂的数据进行简化,再用统计法进行分类汇总,用此法得出规则训练集中确诊病例各属

作者简介:聂斌(1972),男,硕士,主要研究领域为人工智能、数据挖掘、粗糙集、粒计算;王命延(1959),男,教授,主要研究领域为人工智能、数据挖掘、软件工程等;邱桃荣,男,博士,副教授,主要研究领域为数据挖掘、粗糙集、粒计算等;于海雯,女,讲师;方娜,女,助教。
收稿日期:2008-09-10 **修回日期**:2008-10-16

性-属性值对的基数。

定义 1 集合 A 上的等价关系 R,其等价类集合$\{[a]_R|a\in A\}$称作 A 关于 R 的商集,记作 A/R。

表 1 数据按商集统计法可得到表 2 数据。

表 2 商集统计信息表

No	a	b	c	e	d
1	03 10	00 13	01 12	03 10	3 Eye diabetes
2	00 12	02 10	01 11	02 10	2 Kidney diabetes
3	01 10	01 10	01 10	00 11	1 Angiocarpy

注:当 Eye diabetes 基为 3 时,a=0 时的基为 3,a=1 的基为 0

2.3 准确度和覆盖

属性商集统计后,就可以对属性进行分类了,在分类之前,先引入两个量度:准确度和覆盖,用准确度来表明分类的充分性,用覆盖来表明分类的必要性。

定义 2 给定 F 为条件属性集和 D 为决策集在信息表 $S=(U,A\cup D)$中,那么分类准确度和覆盖可以定义为:

$\alpha_F(D)=|F_A\cap D|/|F_A|\quad \beta_F(D)=|F_A\cap D|/|D|$

其中$|D|$、$|F_A|$相应是集合 D,集合 F_A 的基数,$\alpha_F(D)$、$\beta_F(D)$是相应于分类的准确度和覆盖。对于表 1 例子中,F 取[c=1],D 取[d=eye diabetes]时,$|F|$=2,$|D|$=3,所以,$\alpha_F(D)$=2/3,$\beta_F(D)$=2/3。

2.4 可能性规则

用准确度和覆盖来表示一种可能性建议,即在 η_α、η_β,预先给定的情况下,满足条件 $\alpha_F(D)\geqslant\eta_\alpha$ 和 $\beta_F(D)\geqslant\eta_\beta$ 时,某种病症在这程度上可能推导出某种疾病。表 2 中,假如给定 η_α=0.6、η_β=0.6,取病症[c=1],满足条件 $\alpha_F(D)$=0.67≥0.6,$\beta_F(D)$=0.67≥0.6,表明[c=1]是可能推导出 Eye diabetes 的。

2.5 症状特征化

医学专家在诊断时,通常会把这些可能推导出某类疾病的症状收集起来,从而进一步做出诊断,可称为症状特征化。用覆盖来定义。

定义 3 每对条件属性-属性值满足覆盖大于或等于给定阈值 η_β 时,得到的基本集 $P(D)$,都有可能推导出某种决策属性 D,称为症状特征化,正如文献[14]可表示为:

$P_{\eta\beta}(D)=\{[[a_i=v_j]|\beta_{[a_i=v_j]}(D)\geqslant\eta_\beta\}$

比如在表 2 中,先给定覆盖的阈值 η_β=0.6,用 D_1、D_2、D_3 分别代表 Eye diabetes、Kidney diabetes、Angiocarpy diabetes。则:

$P_{0.6}(eye)=\{[a=0],[b=1],[c=1],[e=0]\}$

表明覆盖大于等于 0.6 的情况下,取[a=0],[b=1],[c=1],[e=0]中的任一个都可能推导出眼疾性糖尿病。以下同理可得:

$P_{0.6}(kidney)=\{[a=1],[b=0],[e=0]\}$

$P_{0.6}(angiocarpy)=\{[a=0],[b=0],[c=0],[e=1]\}$,在这 3 个属性基本集 $P(D)$之间,前两者用 c 的属性-属性值对可以区分,第 1 和第 3 用 b 和 e 的属性-属性值对可以区分,这将在区分两种疾病提取规则时会用到。

在医学专家诊断时,常根据一些症状来区分出疾病,或者排除一些疾病,得到诊断结果。如果还要继续深入,再从排斥度次小的一类疾病中进行区分。这里使用粗糙排斥域和粗糙排斥度来区分。

2.6 粗糙排斥域和粗糙排斥度

粗糙排斥域和粗糙排斥度可以这样定义。

定义 4 y 和 z 表示两个由属性-属性值对组成的集合,E 表示粗糙排斥域,E^d 表示粗糙排斥度,则 y 对 z 的粗糙排斥域可表示为:$E(y,z)=y-(y\cap z)$,y 对 z 的粗糙排斥度可表示为:$E^d(y,z)=|y-(y\cap z)|/|y|$。

如表 1 中:

$E(P_{0.6}(eye),P_{0.6}(kidney))=\{[a=0],[b=1],c=[1]\}$

$E^d(P_{0.6}(eye),P_{0.6}(kidney))=|\{[b=1],\cdots\}|/|\{[a=0],\cdots\}|=3/4$

3 容假度

定义 5 给定 F 为实例条件属性集在信息表 $S=(U,A\cup D)$中,SL 为排斥规则库中一条诊断规则属性-属性值对集,则诊断匹配的容假度可表示为:$C^F=|F_A|/|SL|$, 其中$|F_A|$、$|SL|$相应是集合 F_A,集合 SL 的基数。

引理 1 在诊断时, 如果匹配时的容假度 C^F 小于等于给定的阈值 ε 时,则可认为匹配成功。

4 基于商集统计 Rough sets 的医学辅助诊断模式

4.1 算法描述

采用的方法分 4 步:第 1 步,分层或分组,即根据商集统计法预处理,再根据症状特征化后的属性-属性值对求出各决策属性两两之间的排斥域,然后根据排斥度从大到小依次层层分出两个组别;第 2 步,每组依次提取规则;第 3 步,合并规则,即根据需要合并第 2 步中提取的规则, 得到每一种疾病的诊断规则;第 4 步,在给定容假度阈值条件下,对实例进行匹配,得到诊断结果,辅助临床诊断。

具体的生成和匹配规则算法见算法 1~4。在第 4 步中,如果只得到一种匹配结果,则直接给出;如果得到两种或两种以上的匹配结果,则给出参考结果。

算法 1 分层算法

```
Procedure rules //(main process of generating diagnostic rules)
//i,j:integer;P:attribute-value sets,W:a new list for P,L:a list for //rough-exclusive,LD:classificaion
//Di:diagnostic rules
//input:attribute-value sets,threshold ηβ
//output:rules
begin
Calculate αF(D) and βF(D) for each elementary relation F and each class D;
   Make a list W(D):={F∈P|βF(D)≥ηβ}for each class D;//ηβ could adjustment
 for i:=1 to n-1 do//n:number of decision attributes
 begin
  for j:=i+1 to n do
   begin
      L:=L∪E^d(W(i),W(j))
   end
   while L≠Ø do
   begin
      m←MAX(L)
      LD=LD∪W(j)   //m=E^d(W(i),W(j))
      L=L-{m}
   end
```

```
    Induce classification rules from LD(算法 2)
    Integrate rules for disease(算法 3)
    output Di
  end
```

算法 2 提取子规则

```
Procedure extracting sub-rules
//input:sub-list
//output:Conjunctive formulae
begin
  for i:=1 to n do //Mi is a list from LD for each disease
    Mi=E(LD(i))//(a list of exclusive Disease);
     for j:=1 to m do// m:total number of attributes in a database
    begin
     Select one pair F=∧[aj=vk] from Mi;
       if(βF(D)≥ηβ)then do
         if(αF(D)≥ηα)then do
           SLi=SLi{F};//select one attribute-value from exclusive region
           end
       end// Conjunctive and disjunctive
      SLi:=(A list of the whole combination of the conjunction formulae in M);
      end
  end
end
```

算法 3 合并提取规则

```
Procedure Integration()
//input sub-list
//output rules
//D:diagnostic rules
begin
    D:=(SL1+…+SLn-1)+SLn  //n:the cardinality of SL
end
```

算法 4 诊断结果

```
Procedure diagnostic()
  //input condition-list
  //output diagnostic result
  begin
    for i:=1 to n do //i:the sum of exclusive disease
      for j:=1to m do //j:the sum of each disease's exclusive rules
        begin
        if C^F≤ε
      r:=D    //r:the diagnostic result
      end
  end
```

4.2　诊断模式图

采用的辅助诊断模式,是先用商集统计法对临床数据进行预处理,然后采用粗糙排斥法进行训练和测试,得到满意结果后作为规则库,再对实例进行容假度匹配,得出诊断结果,以辅助临床医生诊断,如图 1 所示。

4.3　实验结果

从《中国典型病例大全》中提取 3 696 例糖尿病数据进行训练学习,并按医学诊断模式[15-17]提取诊断规则,用 84 例确诊病例进行测试,测试结果如图 2 所示。

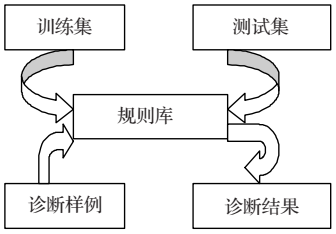

图 1　诊断模式

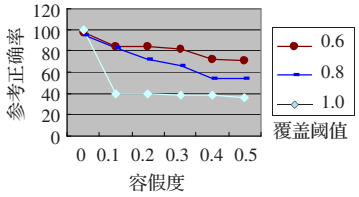

图 2　诊断测试图

图 2 显示:覆盖阈值从 0.6~1.0,阈值越大,准确度相对越小;容假度越大,准确度越小;当容假度为 0 时,覆盖阈值从 0.6~1.0,准确度基本持平。以上结果与医学临床诊断模式正好吻合。

5　结语

Rough 集以及 Rough 集理论在处理不分明问题时,非常接近人类的思维方式。将 Rough 集理论用于医学诊断中,许多教授和学者已经作了深入和成功的研究。针对收集的确诊病例多且杂,以及临床上常出现假阳性或假阴性而造成误诊和漏诊的情况,提出一种基于商集统计 Rough sets 的医学辅助诊断模式。通过对数据测试表明,该模式是可行有效的,是一种能根据临床需要进行辅助诊断的方案。

但提出的医学辅助诊断模式,是否适合多类病种,以及是否适合故障诊断等,还需要进一步研究和测试。

参考文献:

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

[2] 李男,邱天爽.基于粗糙集的数据挖掘技术及其在临床医学诊断中的应用[J].上海生物医学工程,2002,2(23):3-7.

[3] 晏峻峰,朱文锋.粗糙集理论在中医证素辨证研究中的应用[J].中国中医基础医学杂志,2006,2(12):90-93.

[4] 周珂,彭宏,胡劲松,等.粗糙集在心电图分类诊断中的应用[J].计算机工程与应用,2006,42(13):206-208.

[5] 张瑜,何俊民.基于 Rough set 理论的小儿常见病诊断智能超媒体系统[J].微型电脑应用,2005,7(21):20-22.

[6] 王小凤,周明全,耿国华.一种基于模糊粗糙集理论的算法及其在医学影像中的应用[J].计算机应用研究,2005,11:222-224.

[7] 蒋芸,李战怀,王勇,等.基于粗糙神经网络的医学图像分类新方法[J].计算机科学,2006,33(11):151.

[8] Midelfart H,Komorowski J.Learning rough set classifiers from gene expressions and clinical data[J].Fundamental Information,2002,53(2):155-183.

[9] Tsumoto S,Hirano S.Automated discovery of chronological patterns in long time-series medical datasets[J].Research Articles,International Journal of Intelligent Systems,2005,20(7):737-757.

(下转 205 页)