动态知识库和概念格在病症智能诊断中的应用

刘 玲,张 永,李 明,杨德三
LIU Ling,ZHANG Yong,LI Ming,YANG De-san

兰州理工大学 计算机与通信学院,兰州 730050
School of Computer and Communication Lanzhou University of Technology,Lanzhou 730050,China
E-mail:linglin8778@sina.com

LIU Ling,ZHANG Yong,LI Ming,et al.Study of dynamic knowledge bases and concept lattices applied to intelligence disease diagnosis.Computer Engineering and Applications,2007,43(28):233-236.

Abstract:The case-base and patients' symptom are described based on formal concept analysis.With the aid of automatic study strategy,system forms knowledge bases from case-base,then gets the best diagnose project by computing the similarity degrees among the concept lattices.Consequently,disease is intelligently diagnosed according to the dynamic knowledge base.
Key words:concept lattice;dynamic knowledge bases;similarity degree;study strategy

摘 要:用形式概念分析理论描述了案例库和待诊病人症状,系统借助自动学习策略由实际的案例库生成知识库,然后通过计算概念格间的相似度获得最佳诊断方案,从而实现了病症的智能诊断。
关键词 概念格;动态知识库;相似度;学习策略
文章编号:1002-8331(2007)28-0233-04 **文献标识码**:A **中图分类号**:TP182

1 引言

1982年,Wille教授提出了一种基于形式背景(formal context)[1]表示形式概念(formal concept)的新模型,即概念格(concept lattice)。形式概念是现实世界中各种概念的抽象,通过概念外延与内涵之间的关系形式化刻画一般的抽象概念,这种概念层次结构是数据分析与规则提取的有效工具,基于概念格的各种规则提取方法在数据库知识发现领域,如信息检索、数字图书馆、软件工程等方面已经获得广泛应用[2-4]。

智能型计算机辅助诊疗专家系统是计算机科学、工程数学、认知科学、逻辑学、心理学等学科与医学相结合的产物。对于传统的智能诊断系统,人们已经发现它们至少存在如下问题[6]:

(1)脆弱性。对于给定的目标,一名专家能够在不同的条件下,采用不同的方法去运用他的丰富知识,而传统的智能诊断系统则在这方面表面出很大的脆弱性,当它的知识库中以规则形式表示的知识不能覆盖当前的情况(即知识不够完备)时,该诊断系统的性能将会有一个大幅度降低,而不是适度的性能降低。

(2)知识获取困难。在传统的智能诊断系统中,知识的获取工作主要都是通过知识工程师与领域专家之间的交互来完成的。但实践证明,这种获取方法也存在很多问题,如'瓶颈'现象。领域专家很不习惯用规则等形式来表达其所拥有的领域知识,即使愿意,所给出的规则常常是不完全的,如某个条件被"遗忘",太具体化或太一般化,重要的规则被忽略,多个专家之间叙述方式不同产生的理解差别,不同地区同种病症的症状表现不同等。另一方面,大多数领域专家在形式化他们的问题求解策略方面存在差异。

本文方法不同于传统的基于知识的系统,其所依赖的知识主要是系统所存储的相关领域中以前解决问题的具体记录,缓解了在常规的知识库中知识获取的瓶颈。它的最大优点在于动态知识库,即通过增量学习而不断增加知识的案例库。

此方法首先根据相关部门具体的医学病例构成案例库,并为每一个案例构建相应的概念格,根据已有的经验为每一个单位概念和关键字赋权值,形成权重表。这是整个智能诊断过程的载体。相当于一般专家系统的知识库和模型库,体现了原理性知识和专家的经验性知识的结合。其推理过程就是在案例库中匹配检索与当前求解问题相似的案例,并作适当调整,使之适应于所求解的问题。一旦匹配或调整后匹配成功,则该病症中解决问题的措施即可作为当前问题的解。

2 概念格理论

定义1 一个形式背景K是一个三元组:$K=(U,D,R)$,其中U为对象集合,D为属性集合,$R\subseteq U\times D$为U和D之间的关系。对于$x\in U,y\in D,(x,y)\in R$,表示"对象x具有属性y"。

定义2 设$K=(U,D,R)$为一形式背景。对于集合$X\subseteq U$,

基金项目:甘肃省自然科学基金(the Natural Science Foundation of Gansu Province of China under Grant No.3ZS042-B25-007,No.2ZS042-B25-014)。
作者简介:刘玲(1982-),女,硕士研究生,主要研究方向:人工智能、数据挖掘等;张永(1963-),男,副教授,硕士研究生导师,主要研究领域:人工智能,智能信息处理技术;李明(1959-),男,教授,硕士研究生导师,主要研究方向:粗糙集理论,数据挖掘与人工智能;杨德三(1982-),男,硕士研究生,主要研究方向:概念格,数据挖掘。

记 $f(X)=\{y\in Y\mid(x,y)\in R,\forall x\in X\}$,相应地,对于集合 $Y\in D$,记 $g(Y)=\{x\in U\mid(x,y)\in R,\forall y\in Y\}$。

定义 3 设 $K=(U,D,R)$ 为一形式背景,$X\subseteq U$,$Y\subseteq D$,称 $H=(X,Y)$ 为 K 的一个概念,如果 $f(X)=Y$ 且 $g(Y)=X$。此时称 X 为 H 的外延,Y 为 H 的内涵。用 $B(K)$ 记 K 的所有概念组成的集合。

定义 4 设 $K=(U,D,R)$ 为一形式背景,$H_1=(X_1,Y_1)$,$H_2=(X_2,Y_2)$ 是 K 的两个概念,规定:$H_1\leq H_2\Leftrightarrow X_1\subseteq X_2$ 此时,H_2 称为 H_1 的超概念,H_1 称为 H_2 的子概念。关系"$\leq$"诱导出 $B(K)$ 上的一个完备格结构,格中元素满足:

$$H_1\cup H_2=(g(Y_1\cap Y_2),Y_1\cap Y_2) \quad (1)$$

$$H_1\cup H_2=(X_1\cap Y_2,f(X_1\cap X_2)) \quad (2)$$

其中 $H_1=(X_1,Y_1)$,$H_2=(X_2,Y_2)$ 是 K 的两个概念,称此完备格为 K 的概念格。

3 基于病症的形式概念

为了得到关于病症的形式背景,首先应该将病症形式化。将病症的某种症状(如发烧)作为相应的对象,症状的程度(如发烧中的高烧、低烧)作为属性,构建概念格。对象——属性对(即症状——程度)定义为单位概念。每一个单位概念都有由权重值形成的权重表。

图 1 为表 1 对应的概念格。表 2 是单位概念的权值分配。此外,形式背景中的每一个对象(a、b、c、d)都看作一个关键字,并且有各自的权重(W_a等)。

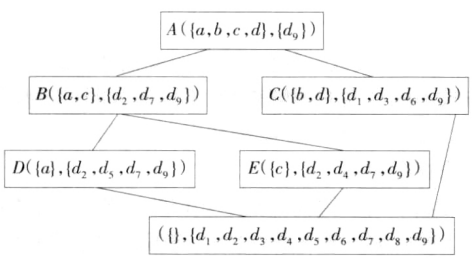

图 1 病症一:伤风的概念格

表 1 病症一 伤风的形式背景

	程度		频度			持续时间		是否带血	
	厉害	一般	高	中	低	长	短	是	否
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
a	x		x			x			x
b		x		x			x		x
c	x			x		x			x
d		x			x		x		x
e		x		x			x		x

其中:a 为发烧、b 为流鼻涕、c 为头痛、d 为咳嗽。

表 2 $\{a,c\}\rightarrow\{d_2,d_7,d_9\}$的权值分配

	d_2 (W_{d_2})	d_7 (W_{d_7})	d_9 (W_{d_9})
a (W_a)	W_{a,d_2}	W_{a,d_7}	W_{a,d_9}
c (W_c)	W_{c,d_2}	W_{c,d_7}	W_{c,d_9}

4 概念格匹配

概念格中的每个点都代表一个概念。对于一个给定的概念,把它和另外一个格中的点进行匹配,也就是格中的点与给定概念的相似度。

定义 5 设两个形式概念 $p=(A_1,B_1)$,$q=(A_2,B_2)$,如果 $A_1\subset A_2$ 并且 $B_1\subset B_2$,则称 $p\subset q$。

定义 6 设两个形式概念 $p=(A_1,B_1)$,$q=(A_2,B_2)$,p,q 的匹配是一个概念 c,$c=(A_1\cap A_2,B_1\cap B_2)$。

c 中的单位概念决定 p,q 之间的相似程度。

(1) 如果给定概念与格中的点完全相同(理想情况),即 $c=p=q=(A_1,B_1)=(A_2,B_2)$,则这两个概念为完全概念匹配;

(2) 如果 $c=p\subset q$,则这两个概念为概念匹配,又称节点匹配;

(3) 如果 $c\neq p$,$c\neq q$ 且 $A_1\cap A_2\neq\phi$,$B_1\cap B_2\neq\phi$,这两个概念为部分匹配。

在实际中,由于病情的症状有很大程度的不确定性。例如患有同一种病症的不同患者的症状不相同;不同地区患同一种病症的症状表现有差异;即使是同一位病人在患病的不同时期所表现出的症状也不同。我们不能等到节点匹配或完全概念匹配成功后才在对病人病情做出判断。那样会耽误治疗的最佳时期。所以在病情匹配的过程中十分需要部分匹配。

(4) 如果两个概念有共同的对象,或者有共同的属性,但是不可能同时有共同的概念和共同的属性,即 $A_1\cap A_2\neq\phi$,$B_1\cap B_2=\phi$ 或者 $A_1\cap A_2=\phi$,$B_1\cap B_2\neq\phi$。那样就不可能有任何一个单位概念能匹配。这种单独的对象或者单独的属性的匹配称作关键字匹配。

但在实际智能诊断中,单独的只有相同的属性,没有相同对象的情况是不能构成关键字匹配的。因为在没有症状的前提下讨论程度是没有意义的。所以本文所说的关键字匹配只有一种情况,即 $A_1\cap A_2\neq\phi$,$B_1\cap B_2=\phi$。

由于在实际匹配病情的时候,希望有单位概念匹配,所以当单位概念能匹配成功时,就不再进行关键字匹配。比如给定的概念的对象集中包括词"a",属性集中包括"d_9"。即假设待诊断的概念格中包含一个节点表示概念 $\{a\}\rightarrow\{d_9\}$。经过概念匹配,将概念$\{a\}\rightarrow\{d_9\}$匹配成图 1 中的 A、B、D 点所表示的概念。但在"d_9"的关键词匹配时匹配成 C 点,因为可以进行单位概念 $\{a\}\rightarrow\{d_9\}$匹配,那么关键字"d_9"的匹配就被删掉。

5 学习策略

学习策略主要是根据系统查询者的"是"(相关)与"不是"(不相关)形式的反馈,来改变案例中各个部分的权值。

具体的工作流程为:如果查询的新案例与系统给出的案例相关,则查询概念格中那些在输出案例中没有的单位概念将增加到案例库中相应案例的概念格中,并且对应的概念赋予初始权值。在输出案例中已经存在的查询概念格中的单位概念则认为是比较重要的概念,因此它的权重值将会增加 ΔW。相反,如果:如果查询者认为查询的新案例与系统输出的案例不相关,那么在输出案例中存在的那些查询概念格中的单位概念则认为是在检索过程中容易导致检索错误的概念,它们相应的权重值将会减少 ΔW。

5.1 权值及权值改变量的确定

关键字和单位概念的初始权值都定在 5.0(这个值也可以根据具体的情况由相应的人员确定)。权值的范围是任意选的,本文选为(0.1~10.0)。权值的最小值选为 0.1 而不是 0,是为了更直观的表示这个案例是有此属性的。

权重改变量是与它现有的权值成比例的。这个改变值主要是根据现有权值与边界的距离和它改变的方向(即如果权值增加就和现在的权值与上边界(10.0)的距离有关,如果权值减少

就和现在的权值与下边界(0.1)的距离有关)确定。

权重改变值计算公式为:

$$\Delta W=\begin{cases} W_{max}-W_{old} & W_{new}\text{将增加} \\ W_{old}-W_{min} & W_{new}\text{将减少} \end{cases} \quad (1)$$

其中 $W_{max}=10.0$,$W_{min}=0.1$

新权值计算公式为:

$$W_{new}=W_{old}+\mu\Delta W \quad (2)$$

其中 μ 为学习速率。

5.2 学习速率的确定

学习速率是一个常数,它是由权值与实际 ΔW 的比例决定的。在实际病情诊断过程中,只有很少的病症案例真正与待查询案例相似,因此,只有这些很少的案例是需要相应权值增加的。其它所有经过系统检索出的相似案例的相应权值是要减少的。这样一来,关键字和单位概念的权值的趋势是减少幅度大于增加幅度。如果不能很好的处理这个增减的不平衡,长此以往,就会导致所有的权值将最终降为最小权值(0.1)。解决这个不平衡的一个办法就是为权值的增加和减少设立不同的学习速率。增加(μ_i)和减少(μ_d)的学习速率的精确的值是很难确定的。它需要根据特征的数目、待诊概念格中概念的数目、系统训练集的数量、查询者的反馈的准确程度等等确定。

5.3 信息量系数

4.1 节中论述的权值的改变是对所有形式的单位概念和关键字都同样适用的。因此,不管单位概念、关键字匹配对概念格匹配的重要程度是多少,每一个单位概念、关键字的权值都是根据上面所述的现有的权值和学习速率进行改变。但是,并不是任何一个概念、关键字在匹配过程中对概念格匹配成功所提供信息的重要程度都是相同的。所以,在相似度计算的时候,还考虑到"信息量因素"。根据信息提供的重要程度把信息量因素分为如下四个等级。括号中的数字表示表示各等级信息量系数的经验值(不是最优值,可以根据实际情况确定):

(1) 只有一个关键字匹配(0.13);

(2) 多个关键字匹配(0.22);

(3) 只有一个对象和一个属性的单位概念匹配(0.26);

(4) 多个对象和属性的单位概念匹配(至少对象和属性中有一个包含多个关键字)(0.39)。

相似度计算公式为:

$$RSV=\alpha\sum W_i^c+\beta W_j^k \quad (3)$$

其中 i 的取值为所有匹配的单位概念,j 的取值为所有匹配的关键词概念,W_c 是单位概念的权值,W_k 是关键词的权值,α β 为信息量系数。

6 诊断过程

首先根据相关的具体医学病例构成案例库。然后预处理案例库,即为每一个案例构建相应的概念格,根据已有的经验为每一个单位概念和关键字赋权值,形成权重表。这是整个智能诊断过程的载体。相当于一般专家系统的知识库和模型库,体现了原理性知识和专家的经验性知识的结合。并设置一个相似度的阈值 ψ。

步骤 1 根据输入的症状生成待诊断的形式背景,生成相应的概念格 C。作集合 $S=\{C_1,C_2,\cdots,C_n\}$,其中 C_i 为概念格中的第 i 个概念,n 为概念格中的概念总数;

步骤 2 从集合 S 中取出概念 C_1,然后将 C_1 分别与案例库中各个病症概念格中每一个节点进行概念间的匹配;

步骤 3 重复步骤 2,直至集合 S 为空;

步骤 4 计算待诊断格病症一标准格的相似度 RSV_1;

步骤 5 重复步骤 2、步骤 3、步骤 4 直至案例库为空。得到待诊断格与各种病症标准格的相似度 RSV_1,RSV_1,…,RSV_m,其中 RSV_i 为待诊断症状格与第 i 中标准病症格的相似度,m 为标准病症格的总数;

步骤 6 比较 RSV_i,i=1,2,…,m 的大小,得到最大的 RSV_{max};

步骤 7 比较 RSV_{max} 与 ψ:

If $RSV_{max}<\psi$ Then 认为此查询案例为一新案例,存入案例库,并建立相应的概念格,为相应的单位概念和关键字赋予初始权重值,形成权重表;

步骤 8 根据第 4 章中所介绍的,改变相应案例的属性及权值,并更新案例库;

步骤 9 输出 RSV_{max} 所对应的病症,此病症即为待诊病人最可能患有的病症。并根据以往此案例的处理方案,为病人提出参考意见。

本文创建一个简单的案例集,其中包含三种案例,如表 3,相应的概念格中的概念如表 4 所示,权重表如表 5(为说明问题此权值是作者取的),设 ψ=10。

表 3 各案例的形式背景

案例一 伤风的形式背景如表 1

案例二 流行性感冒

	程度		频度			持续时间		是否带血	
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
a	x		x			x			x
b		x		x			x		x
c	x			x		x			x
d		x			x		x		x
e		x		x			x		x

案例三 大叶肺炎

	程度		频度			持续时间		是否带血	
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
a	x		x			x			x
b		x		x			x		x
c	x			x		x			x
d		x			x		x		x
e		x		x			x		x

其中:e 为痰、a- d、d_1- d_9 含意见表 1。

表 4 各案例的概念表

案例一 伤风案例

概念集	属性集
{a,b,c,d}	{d_9}
{a,c}	{d_2,d_7,d_9}
{b,d}	{d_1,d_5,d_6,d_9}
{a}	{d_2,d_5,d_7,d_9}
{c}	{d_2,d_4,d_7,d_9}

案例三 大叶肺炎

概念集	属性集
{a,d,e}	{d_1,d_3,d_8}
{a,d}	{d_1,d_3,d_4,d_8}
{e}	{d_1,d_3,d_6,d_8}

案例二 流行性感冒

概念集	属性集
{a,b,c,d,e}	{d_9}
{a,c}	{d_1,d_9}
{b,c,e}	{d_4,d_9}
{b,d,e}	{d_2,d_7,d_9}
{a}	{d_1,d_3,d_9}
{b,c}	{d_4,d_6,d_9}
{b,e}	{d_2,d_4,d_9}
{d}	{d_2,d_5,d_7,d_9}
{c}	{d_4,d_7,d_9}
{b}	{d_2,d_4,d_5,d_6,d_7,d_9}

表 5 各案例的权重表

案例一 伤风

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
a(1.9)		2.5			2.1		3.2		0.1
b(7.6)	8.8		4.9			6.7			0.1
c(2.4)		1.0		4.7			2.9		0.1
d(5.5)	7.5		5.1			6.7			0.1

案例二 流行性感冒

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
a(8.7)	9.8		9.5			8.9			0.1
b(2.5)		4.2		4.0			2.9		0.1
c(4.3)	7.9			4.4		5.0			0.1
d(1.0)		4.9			1.6		1.3		0.1
e(1.4)		3.0		2.9			1.8		0.1

案例三 大叶肺炎

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
a(9.0)	9.9		9.9			9.8			0.1
b(7.2)	8.8		4.9			8.5			0.1
e(5.9)	3.9		8.9			6.4		4.6	

如果新得到一位病人的症状如表 6 所示，相应的概念如表 7 所示。

表 6 病人症状的形式背景

	程度		频度			持续时间		是否带血	
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
a	x		x			x			x
b		x		x			x		x
c	x			x		x			x
d		x			x		x		x
e		x		x			x		x

表 7 病人症状的概念表

概念集	属性集	概念集	属性集
{a,c,d}	$\{d_9\}$	{a}	$\{d_1,d_3,d_6,d_9\}$
{a,c}	$\{d_1,d_9\}$	{c}	$\{d_1,d_4,d_7,d_9\}$
{c,d}	$\{d_7,d_9\}$	{d}	$\{d_2,d_5,d_7,d_9\}$

根据上面描述的方法将病人的概念格与病症一、二、三所对应的概念格分别进行单位概念匹配和关键字匹配，计算相应的相似度分别为 $RSV_1=8.018$，$RSV_2=25.272$，$RSV_3=20.267$，$RSV_2=25.272>RSV_3=20.267>\psi=10>RSV_1=8.018$，相似度最大的格所对应的案例为案例二，所以这位病人最可能患有案例二：流行性感冒。

7 小结

在医学诊断预测领域，专家的优势在于丰富的临床经验和与众不同的创造性思维方式。本文综合运用概念格的知识，根据临床医学诊断的一般过程，提出了智能诊断的一个方法，为概念格在生命科学中的应用作了一些探索。本方法还需要进一步研究，比如：先将案例库中的案例进行聚类操作，在某一类中进行检索，增加系统的准确性等。（收稿日期：2007 年 1 月）

参考文献：

[1] Wille R.Restructuring lattice theory:an approach based on hierarchies of concepts[M]//Ordered Sets.Boston:Springer,1982:445-470.

[2] Godin R,Milih,Mineaugw,et al.Design of class hierarchices based on concept(Galois) lattices[J].Theory and Application of Object System,1998,4(2):117-134.

[3] Sahamim.Learning classification rules using lattices(Extended Abstract)[C]//Proc of the Eighth European Conf on Machine Learning.Berlin:Springer-Verlag,1995:343-346.

[4] Godin R,Missaoni R,Aprila.Experimental comparison of navigation in a galois lattices with conventional information retrieval methods[J].International Journal of Man-Machine Studies,1993,38:747-767.

[5] Rajapakse R K,Denham M.Text retrieval with more realistic concept matching and reinforcement learning[J].Information Processing & Management,2006,42(5):1260-1275.

[6] 杨叔子，丁洪，史铁林，等.基于知识的诊断推理[M].北京：清华大学出版社，1993:11-12.

[7] 李锋刚，倪志伟，郜峦.案例推理技术在医学诊断专家系统中的设计思路探讨[J].中医药临床杂志，2005,17(2).

（上接 202 页）

设计理论，降低了信息系统设计的复杂性，可在设计过程中判断设计是否满足独立性公理，是否为合理的设计，进而改进设计，降低模块间的耦合度。在考虑模块之间的相互影响和操作的前提下，按照信息系统流程图进行设计，缩短了信息系统编码时间和调试时间，提高了工作效率，保证了信息系统较高的可靠性和模块重用性。（收稿日期：2006 年 12 月）

参考文献：

[1] 季绍波.中国信息系统(IS)研究现状和国际比较[J].管理科学学报，2006,2:76-80.

[2] 姚莉.基于多智能体的复杂信息系统开发方法研究[J].管理科学学报，2002,10:44-45.

[3] Nam P S.Axiomatic design—advances and applications[M].[S.l.]:Oxford University Press,2001.

[4] 程贤福.基于公理性设计的企业电子商务策略研究[J].商业研究，2005,14:196-199.

[5] 江屏.公理设计应用软件研究[J],计算机集成制造系统(CIMS)，2004,10:1199-1206.

[6] 杨德林.一种基于公理设计的产品知识表达方法研究[J].管理工程学报，2004,4:21-24.

[7] 程贤福.基于公理化设计理论的供应链设计[J].设计与研究，2005,2:19-23.

[8] 劳动和社会保障部信息中心.劳动和社会保障信息化建设文件资料集[M].北京：中国劳动社会保障出版社，2003.

[9] 高远忠.数字就业和社会保障系统的规划与实施[J].福建电脑，2002,6:1-2.

[10] 程晓雷.社会保险系统特点及架构设计[J].电子政务，2005,8:73-75.