一种X线征象的智能数据分析方法

吴大岳[1],谢福鼎[1,2]
WU Da-yue[1],XIE Fu-ding[1,2]

1.辽宁师范大学 计算机与信息技术学院,辽宁 大连 116029
2.中国科学院 数学机械化重点实验室,北京 100080
1.School of Computer and Information Technology,Liaoning Normal University,Dalian,Liaoning 116029,China
2.Key Laboratory of Mathematics Mechanization,AMSS,Chinese Academy of Sciences,Beijing 100080,China
E-mail:pjwdy@sohu.com

WU Da-yue,XIE Fu-ding.Approach of intelligent data analysis for X-ray sign.Computer Engineering and Applications,2007,43(28):193-195.

Abstract: With the development of the information times,the method of intelligent data analysis has been widely employed in the field of medical.During the process of medical image diagnosis,the using of more efficient methods to analyze the data is just what people are looking forward to.The theory of the formal concept analysis will be applied to the X-ray diagnosis of expertise database in this paper.The purpose of the intelligent analysis is achieved for the syndrome of radiograph holders through the calculation of similarity between X-ray signs.

Key words: formal concept analysis;expert knowledge database;X-ray sign;intelligent data analysis

摘 要:随着信息时代的发展,智能数据分析方法在医学领域中的应用地位不断提升。其中在医学影像诊断过程中,高效数据分析方法的介入更是人们所期待的。将形式概念分析理论应用到X线诊断的专家经验知识库中,通过X线征象间相似度的计算,达到对持有X光片的病人进行病症的智能分析目的。

关键词:形式概念分析;专家知识库;X线征象;智能数据分析
文章编号:1002-8331(2007)28-0193-03 **文献标识码**:A **中图分类号**:TP182

1 前言

医学领域内的医学信息作为医疗机构不可分割的重要资源,随着信息化的迅速发展而急剧增长,所以传统的手工数据分析方法已经无法满足要求,而寻求一种高效的基于计算机的智能数据分析方法恰是医疗领域目前所期盼的。医学IDA(Intelligent Data Analysis)是医护人员的智能化助手,使他们在恰当的时间拥有恰当的信息,帮助他们在有限的时间内做出恰当的决定、采取恰当的行动[1]。在医疗领域,医学影像技术的发展大大提高了疾病诊断的正确性与准确率,但由于其对各种疾病的分析很大程度上依赖于专家的主观经验知识,而对专家经验知识在领域内的共享、学习与维护尤为困难,本文通过应用形式概念分析理论使专家经验知识快速推广与应用成为可能。

自从形式概念分析理论[2]于1982年提出以来,已经在知识发现领域、软件工程领域、知识工程领域、经济分析领域及信息检索等众多领域得到了广泛的应用[3-8]。概念格作为其一种核心数据结构也越来越受到了人们的重视。

本文主要以医学影像技术中的X线诊断方式为基础,通过建立各种病症的X线诊断专家经验知识库,应用形式概念分析理论,实现对病人X线征像与库中标准病症X线征像的匹配,最终匹配出的结果即可作为医疗者诊断病症时的辨别依据,进而提高临床诊断的准确性,降低误诊率。

2 形式概念分析

下面简单地介绍一下形式概念分析的基本理论,详细内容见参考文献[2]。

定义形式背景为三元组$K=(U,A,R)$,其中U是对象集,A是属性集,R是U和A之间的二元关系,即$R\subseteq U\times A$,uRa表示"对象u具有属性a"。在形式背景K中,在U的幂集$P(U)$和A的幂集$P(A)$之间可定义两个映射f和g如下:

$$\forall u_1\subseteq U:f(u_1)=\{a|\forall x\in u_1(xRa)\}$$

$$\forall a_1\subseteq A:g(a_1)=\{x|\forall a\in a_1(xRa)\}$$

称其为U的幂集和A的幂集之间的Galois联接。对二元组$C=(u_1,a_1)\in P(U)\times P(A)$如满足两个条件:$u_1=g(a_1)$及$a_1=f(u_1)$,则称其为形式背景$K$的一个形式概念,其中$u_1$和$a_1$分别被称为概念$(u_1,a_1)$的外延和内涵,其外延和内涵也可以分别用

基金项目:国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318000);辽宁省教育厅科技项目(the Science & Technology Project of the Education Department of Liaoning Province)。

作者简介:吴大岳(1979-),硕士研究生,主要研究方向:人工智能,数据挖掘;谢福鼎(1965-),博士,教授,主要研究方向:人工智能,数据挖掘。

$Int(C)$和$Ext(C)$来表示。K的所有形式概念的集合被标记为$FC(K)$。在每个概念中,每个对象和每个属性组成的一个序偶称作单位概念。

$FC(K)$上最重要的结构是由子概念——父概念关系产生的,其定义如下:给定形式概念c_1和c_2,如果$Ext(c_1)\subset Ext(c_2)$或者$Int(c_2)\subset Int(c_1)$,则称c_1是c_2的子概念,c_2是c_1的父概念,记为$c_1<c_2$。通过这个关系,可以得到偏序集$\underline{FC(K)}=(FC(K),<)$。因为对于$FC(K)$任意非空子集$B$中任意两个形式概念都有最小上界和最大下界,所以偏序集$\underline{FC(K)}$是一个完全格,称为形式背景K的概念格,记为$L(K)$。概念格中的两个不同的节点$c_1=(u_1,a_1)$和$c_2=(u_2,a_2)$,如果c_1是c_2的一个子概念且不存在其它的节点c_3,满足$c_1<c_3<c_2$,则称c_1为c_2的子节点或直接后继,而c_2为c_1的父节点或直接前趋。由概念的前趋和后继关系得到的一个图,为Hasse图。概念格通过Hasse图生动、简洁地体现了这些概念之间的泛化和特化关系(见图1)。关于概念格的构造有多种方式,见文献[6,8]。

表1　原发性肺结核形式背景

	密度		面积		位置			形状	
	高	低	大	小	上叶	中叶	下叶	规则	不规则
	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
阴影 a		X		X	X			X	
纹理 b		X	X			X			X
空洞 d	X			X			X	X	
钙化 g	X			X	X				X

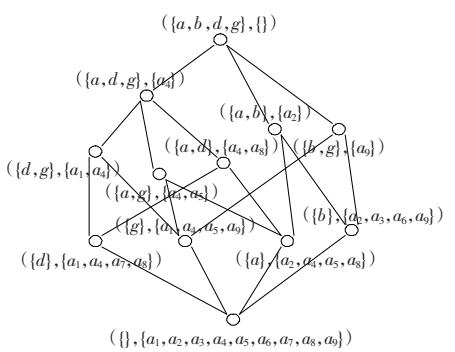

图1　表1所对应的Hasse图

3　概念格的相似度

设概念格L_1与L_2中的概念集分别为$L_1c=\{c_1,c_2,\cdots,c_n\}$,$L_2c=\{d_1,d_2,\cdots,d_m\}$,$(n,m$分别为对应格中的概念总数$)$则$L_1c$中的任一概念$c_i$与$L_2c$中的任一概念$d_j$的匹配,称为概念间的匹配。

定义1　设两个形式概念$p=(A,B)$,$q=(C,D)$,如$A\subset C$且$B\subset D$,则称$p\subset q$。

定义2　设两个形式概念$p=(A,B)$,$q=(C,D)$,p,q的匹配是一个概念c,$c=(A\cap B,C\cap D)$如给定的单位概念与格中的点完全相同,那么两个概念叫做完全概念匹配;如$c=p\subset q$,那么两个概念就叫概念匹配(或节点匹配);$c\neq p$,$c\neq q$且$A\cap C\neq\phi$,$B\cap D\neq\phi$,则这两个概念叫做部分匹配。

有时候两个概念有相同的对象,或者有相同的属性,但有相同的对象和有相同的属性这两个条件并不同时存在。即:$A\cap C\neq\phi$,$B\cap D=\phi$或$A\cap C=\phi$,$B\cap D\neq\phi$这样就会出现没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。

在匹配过程中由于会出现单位概念与关键字重复匹配的情况,所以在匹配时应对其进行必要的处理,如:进行概念匹配时,出现了一个单位概念,经过与概念集1(见表2)的匹配,匹配成$(\{d,g\},\{a_1,a_4\})$,$(\{d\},\{a_1,a_4,a_7,a_8\})$节点,但在关键字$d$的匹配却匹配成$(\{a,d,g\},\{a_4\})$,$(\{d,g\},\{a_1,a_4\})$,$(\{a,d\},\{a_4,a_8\})$,$(\{d\},\{a_1,a_4,a_7,a_8\})$节点,因为已存在了$(d,a_1)$单位概念的匹配,那么该单位概念中对象$d$若再作为关键字进行匹配,就认为是重复匹配现象发生,这时应将对象d匹配到的节点去掉。同样若概念中某一属性出现了类似情况,也应做删除匹配处理。

经过专家对概念集进行赋值后,将具有权值的概念格存入知识库中,两个概念格的相似度通过下面方法计算:

$$S=\sum_i\sum_j(W_j^c+W_j^k)$$

其中,i是准备与知识库中的格进行匹配的格中所有概念的个数,j是知识库中某一个格的全部概念的个数,W^c是所有匹配出的单位概念的权值和。W^k是所有匹配出的关键字的权值和。

下面描述了具体的相似度匹配过程:

步骤1　根据病人X线征象生成的形式背景,生成概念格C,形成概念集$D=\{C_1,C_2,\cdots,C_n\}$其中C_i为概念集D中的第i个概念。

步骤2　从D中任取一未匹配的概念C_i,与标准病症概念集$Q=\{q_1,q_2,\cdots,q_m\}$中的每个概念q_j(这里q_j为概念集Q中的第j个概念)进行匹配,将概念$C_i=(A,B)$与标准病症中任一未匹配过的概念$q_j(C,D)$进行匹配

(1)若$A\cap C\neq\phi$,$B\cap D\neq\phi$,则将C_i与q_j中匹配的单位概念记录在集合T中;

(2)若$A\cap C=\phi$,$B\cap D\neq\phi$,则将$B\cap D$得到的公共属性记录在集合K中;

(3)若$A\cap C\neq\phi$,$B\cap D=\phi$,则将$A\cap C$得到的公共对象记录在集合K中;

(4)若$A\cap C=\phi$,$B\cap D=\phi$,则没有匹配的单位概念或关键字。

步骤3　重复步骤2,直至集合Q为空。

步骤4　删除K中在T中出现过的对象和属性后,根据医学专家对每个单位概念和关键字所赋予的权值,计算T中全部单位概念的权值与K中全部关键字的权值之和记为W_i。

步骤5　清空K集与T集中的内容。

步骤6　重复步骤2至步骤5直至集合D为空,$S=\sum_{i=1}^{n}W_i$。

最终所计算的S即为病人X线征象概念格与某一病症概念格的相似度。

4　例子

在这一部分,以肺部X线的诊断为例,说明我们所给出算法在医学领域中的应用。

对某一病症将其X线征象特征作为相应的对象,将其征象特征的影象表示作为属性,建立了3个X线诊断病症的简单形式背景,(由于篇幅有限这里所建立的形式背景仅仅可用来描述病症,若要在实际中作为诊断依据,还应当加入如病灶的边缘清晰度、病人的已往病史等属性,同时若要提高准确度

可对所列出属性进行更为详细的划分。)由此生成了相应的概念集(见表 2、3、4),由专家赋予权值后将它们存入到专家经验知识库中(见表 5、6、7)。

表 2　原发性肺结核概念集

外延	内涵	外延	内涵
d,g	a_1,a_4	a,g	a_4,a_5
a,b	a_2	g	a_1,a_4,a_5,a_9
b	a_2,a_3,a_6,a_9	d	a_1,a_4,a_7,a_8
a,d,g	a_4	a,d	a_4,a_8
a	a_2,a_4,a_5,a_8	b,g	a_9

表 3　中心型肺癌概念集

外延	内涵
a,d,e,f	a_4,a_6
d,e,f	a_1,a_4,a_6
a	a_2,a_4,a_6,a_8
a,d	a_4,a_6,a_8
d	a_1,a_4,a_6,a_8
e,f	a_1,a_4,a_6,a_9

表 4　肺脓肿概念集

外延	内涵
a,c,d	a_7,a_9
d	a_1,a_4,a_7,a_9
a,c	a_2,a_3,a_7,a_9

表 5　原发性肺结核 X 线征象形式背景及概念集上相应权值

	密度		面积		位置			形状	
	0.3	0.2	0.3	0.2	0.1	0.2	0.2	0.3	0.2
	a_1 高	a_2 低	a_3 大	a_4 小	a_5 上叶	a_6 中叶	a_7 下叶	a_8 规则	a_9 不规则
0.1 阴影 a		X 0.5		X 0.3	X 0.6			X 0.5	
0.6 纹理 b		X 0.9	X 0.5			X 0.5			X 0.2
0.1 空洞 d	X 0.6			X 0.3			X 0.4	X 0.4	
0.2 钙化 g	X 0.7			X 0.3	X 0.4				X 0.5

表 6　中心型肺癌 X 线征象的形式背景及概念集上相应权值

	密度		面积		位置			形状	
	0.3	0.2	0.3	0.2	0.1	0.2	0.2	0.3	0.2
	a_1 高	a_2 低	a_3 大	a_4 小	a_5 上叶	a_6 中叶	a_7 下叶	a_8 规则	a_9 不规则
0.1 阴影 a		X 0.4		X 0.3		X 0.4		X 0.5	
0.1 空洞 d	X 0.4			X 0.3		X 0.5		X 0.4	
0.1 结节 e	X 0.5			X 0.3		X 0.3			X 0.4
0.7 毛刺 f	X 0.9			X 0.5		X 0.2			X 0.2

表 7　肺脓肿 X 线征象的形式背景及概念集上相应权值

	密度		面积		位置			形状	
	0.3	0.2	0.3	0	0.1	0.2	0	0.3	0.2
	a_1 高	a_2 低	a_3 大	a_4 小	a_5 上叶	a_6 中叶	a_7 下叶	a_8 规则	a_9 不规则
0.2 阴影 a		X 0.5	X 0.5				X 1		X 0.4
0.6 液平 c		X 0.8	X 0.6				X 1		X 0.9
0.2 空洞 d	X 0.5			X 0			X 1		X 0.3

当医疗者对一个病人的 X 线征象求助于知识库时，根据病人 X 线征象的形式背景(表 8)生成概念集(表 9)，将该概念集与专家经验知识库中的病症概念格进行相似计算。

表 8　病人 X 线征象形式背景

	密度		面积		位置			形状	
	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
	高	低	大	小	上叶	中叶	下叶	规则	不规则
阴影 a		X	X	X		X			X
液平面 c	X	X					X		X
结点 e				X		X			X

表 9　病人 X 线征象概念集

外延	内涵	外延	内涵
a,c,e	a_9	c	a_2,a_3,a_7,a_9
e	a_1,a_4,a_6,a_9	a,e	a_4,a_6,a_9
a,c	a_2,a_9	a	a_2,a_4,a_6,a_9

需要注意的是,为方便描述,表 5、6、7 中每项小方块里的数值为单位概念和关键字的权值,而非形式背景的内容。

根据本文的算法,通过对病人 X 线征象概念格与所给的 3 种病症的 X 线征象概念格相似度的计算，得到与原发性肺结核概念集上的相似度 S_1=12.5；与中心型肺癌概念集上的相似度 S_2=11.9;与肺脓肿概念集上的相似度 S_3=15.1。这里相似度比较的原则为选取最大的 S_n(n 为与第 n 个病症的相似度),所以 S_3 对应的肺脓肿为专家经验知识库对病人 X 线征象所做出的病症分析。

5　结束语

基于形式概念分析,本文提出了一个概念格之间的匹配算法,并把该算法运用到医学领域中去,取得了比较好的结果。所给出的基于专家经验知识库分析的 X 线征象诊断方法，具备较高的准确度,可在实际应用过程中作为参考的依据。为了提高其准确度,需要增加更多的属性,如病灶的边缘清晰度、病人的已往病史等。当然,要确诊一个病人的病症,还需要其它方法,这里只是给出了关于 X 线征象部分的诊断结果。

(收稿日期:2007 年 5 月)

参考文献:

[1] 姚卫新,黄丽华.智能数据分析在医学领域的应用综述[J].计算机工程,2004,30(7):3-5.

[2] Wille R.Restructuring lattice theory:An approach based on hierarchies of concepts[C]//RivalI.Ordered Sets Dordrecht:Reidel,1982.

[3] Godin R,Missaoui R.An incremental concept formation approach for learning from databasees[J].Theoretical Computer Science,1994,133:387-419.

[4] Zupa B,Bohance M.Learning by discovering concept hierarchies[J].Artificial Intelligence,1999,109(1-2):211-242.

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

[6] 胡学钢,张玉红,唐志军,等.一种新的概念格并行构造方法[J].合肥工业大学学报:自然科学版,2005,28(12).

[7] Barahona P,Christensen J P.Knowledge and decisions in health telematics[M].[S.l.]:IOS Press,2001.

[8] Nourine L,Raynaud O.A fast algorithm for building lattices[J].Information Processing Letters,1999,71(5/6):199-204.