挖掘脑部医学图像序列相似模式*

潘海为+, 李建中, 张 炜

(哈尔滨工业大学 计算机科学与工程系,黑龙江 哈尔滨 150001)

Mining Image Sequence Similarity Patterns in Brain Images

PAN Hai-Wei+, LI Jian-Zhong, ZHANG Wei

(Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: +86-451-86415872, E-mail: heaven_007cn@yahoo.com.cn, http://www.hit.edu.cn

Received 2003-05-10; Accepted 2004-07-16

Pan HW, Li JZ, Zhang W. Mining image sequence similarity patterns in brain images. *Journal of Software*, 2004,15(Suppl.):1~12.

Abstract: The high incidence of brain disease, especially brain tumor, has increased significantly in recent years. It is becoming more and more concernful to discover knowledge through mining medical brain image to aid doctors' diagnosis. Image mining is the important branch of data mining. It is more than just an extension of data mining to image domain but an interdisciplinary endeavor. Image clustering and similarity retrieval are two basilic parts of image mining. In this paper, we introduce a notion of image sequence similarity patterns (ISSP) for medical image database. ISSP refer to the longest similar and continuous sub-patterns hidden in two objects each of which contains an image sequence. These patterns are significant in medical images because the similarity for two medical images is not important, but rather, it is the similarity of objects each of which has an image sequence that is meaningful. We design the new algorithms with the guidance of the domain knowledge to discover the possible space occupying (PSO) in brain images and ISSP for similarity retrieval. The experimental results demonstrate that the results of similarity retrieval are meaningful and interesting to medical doctors.

Key words: database; data mining; image mining; clustering; similarity retrieval

摘 要: 随着脑部疾病(尤其是脑癌)发生率的逐年上升,通过挖掘脑部医学图像来发现知识对辅助医生的诊断变得越来越重要.图像挖掘是数据挖掘的重要分支,它不仅仅是数据挖掘简单的扩展到图像领域,而是一个多学科交叉的研究方向.图像的聚类和相似性搜索是图像挖掘的两个非常重要的领域.针对医学图像数据库引入了图像序列相似模式(ISSP)的概念,对于各自包含一个图像序列的两个对象,ISSP 是指隐藏在他们中的最长相似连续子模式.这些模式在医学图像中是很有意义的,因为对医生来说两个对象相似要比两个图像相似更有意义.设计了新的基于领域知识指导下的算法来发现可能性占位(PSO)和 ISSP 以支持相似性索引.实验表明,该研究对医生的辅助诊断有比较

* Supported by the National Natural Science Foundation of China under Grant No.60273082 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2002AA444110 (国家高技术研究发展计划(863)); the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704 (国家重点基础研究发展规划(973))

作者简介: 潘海为(1974－),男,黑龙江哈尔滨人,博士生,主要研究领域为数据库系统,数据挖掘;李建中(1950－),男,教授,博士生导师,主要研究领域为数据库系统,并行计算;张炜(1975－),男,博士生,主要研究领域为数据库系统,移动对象数据库.

好的效果.

关键词: 数据库;数据挖掘;图像挖掘;聚类;相似性搜索

图像获取和存储技术的发展已经促进了大规模图像数据库的飞速发展[1].我们的日常生活和各个领域每天都有大量的图像产生,例如医学图像(CT 图像,ECT 图像,核磁共振图像),人造卫星图像和各类数字照片等等,这些图像包含了大量的对人们有用的信息,但是对于用户来说,发现这些潜在的知识是十分困难的.

图像挖掘可以自动的从大量图像中发现这些隐含的知识或者模式,它在数据挖掘领域正受到越来越多的重视.图像挖掘不仅仅是数据挖掘简单的扩展到图像领域,它是一个多学科交叉的研究方向,包括计算机视觉,图像处理,图像检索,数据挖掘,机器学习,人工智能和数据库等等.尽管以上各领域都有很多成熟的技术,但是图像挖掘仍然处于起步阶段[2].当前的图像挖掘的研究分为两个方向:(1) 面向特殊领域的图像挖掘;(2) 通用的图像挖掘[2].面向特殊领域的图像挖掘的研究主要集中在如何从图像中提取最相关的特征,并组织成适合数据挖掘的形式[3~5];通用的图像挖掘的研究主要集中在如何产生图像模式,以帮助人们更好的理解图像表达的意义[1,6,7].医学图像上的数据挖掘就属于第 1 个方向.

脑组织是人体的高级神经中枢,它的功能格外重要,发生在脑组织的疾病历来都受到医学界的高度重视.在中国,每年大约有 40 000 的新增脑瘤患者,其中 16%的患者为儿童.尤其近年来,脑部疾病(尤其是脑瘤)的发生率呈上升趋势,对人们的生活质量甚至生命都构成了很大的威胁,因此,脑部疾病的早期诊断就变得特别重要,并且直接关系到患者的治疗效果,所以,在脑部医学图像上研究数据挖掘具有非常重要的意义,而且由于它的很强的领域性,使得这方面的研究具有更大的挑战性.

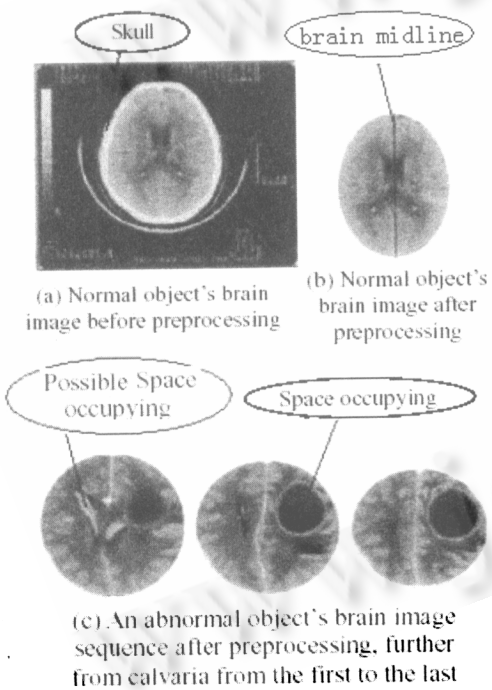


图 1 一个正常和患病对象的脑部图像示例

CT 影像是医生在临床上进行诊断最常用的技术之一.脑部 CT 是指从患者的头顶向下每隔一段间距(通常是几个毫米)生成一个图像,因此对于每个患者(以下称为对象)的脑部 CT 就是一个图像的序列,这个序列中的图像之间存在着很强的位置关系,我们将在这样的数据集合上利用数据挖掘技术来发现知识,辅助医生的诊断.

本文在脑部医学图像集合上,提出了一个新的方法来支持相似对象的搜索,我们的创新点主要表现在以下 3 个方面:(1) 我们提出了两个不同的基于像素的聚类算法来发现可能的占位(PSO),如图 1(c)所示;(2) 提出了一个新的概念—图像序列相似模式(ISSP)来支持相似性搜索;(3) 我们有效的将领域知识引入到算法中,提出了基于领域知识指导的挖掘算法.

本文第 1 节是问题的提出.第 2 节是预处理阶段.第 3 节为基于像素聚类的 PSO 检测.第 4 节为基于 ISSP 的相似性搜索.第 5 节是实验.第 6 节是结论和未来的工作.

1 问题的提出

目前,医学图像挖掘的主要工作[3,8~11]有以下两个特点:(1) 研究的内容是医学图像数据库中的图像,而不是具有这些图像的人(以下称为对象).例如,在图 1(c)中,这 3 个图像是同一个人的脑部图像,由于是对脑部不同层面的反映,所以表现有所不同.如果按照已有方法(例如,直方图的方法),第 1 个图像与其他两个图像将被区分开,属于两类图像,这将决定被挖掘的知识的类型;(2) 研究的方法通常是从图像中提取特征组成特征属性,然后在

这些特征属性上应用数据挖掘的方法进行知识的发现,而没有很好的考虑图像的基本构成元素——像素的意义,但是医生的诊断却主要是根据医学领域知识和像素的明暗.图 2 是已有工作的数据组织形式,其中 IM_i 是图像数据库中的图像,F_{ij} 是从 IM_i 中提取的特征,然后被组织成图像的一个属性.

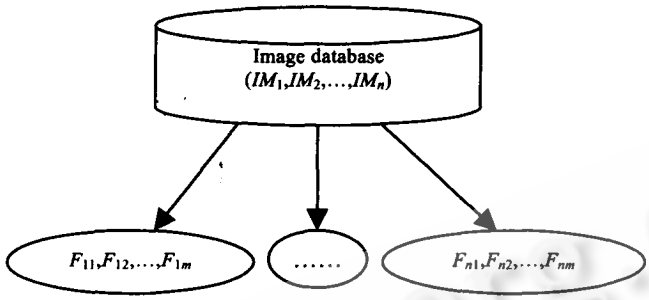


图 2 已有工作的数据组织框架 IM_n 表示图像,$F_{n\,m}$ 表示从 IM_n 中提取的特征

我们的研究内容是具有图像序列的对象,如图 3 所示.一个对象 S_i 的所有图像 $IM_{i1},IM_{i2},\ldots,IM_{ij}$ 反映了这个对象脑部的不同层面的表现,这些图像有着相同的结构和相似的明暗对比度和像素密度分布,这与视频中的帧的序列是不同的,如图 1(c)所示.对于两个不同的对象,他们图像的明暗对比度和像素密度分布可能相差很大,但是他们却可能有相同的病症,如图 9 所示.在脑部图像中有很多特殊的区域,这些区域中既包括了占位(即脑瘤,记为 SO),也包括了由占位引起的脑部其他组织的密度的变化,我们称之为可能性占位(PSO).SO 和 PSO 在形态上十分相似,我们很难用一个统一的标准来区分它们,因此,在不至于混淆的情况下,我们后面统称为 PSO.但是,SO 和 PSO 的出现通常是有因果关系的,也就是 PSO 是由 SO 导致的,而且他们在图像中又存在一定的空间关系,例如,在图 1(c)中,脑部左半球的 PSO 与右半球的 SO 位于大脑两侧,且是相邻的.因此,我们首先在每一个图像中检测这些 PSO 并找到它们的空间关系,然后发现每个对象的图像序列模式(ISP),最后我们检索具有相似 ISP 的对象来辅助医生的诊断.

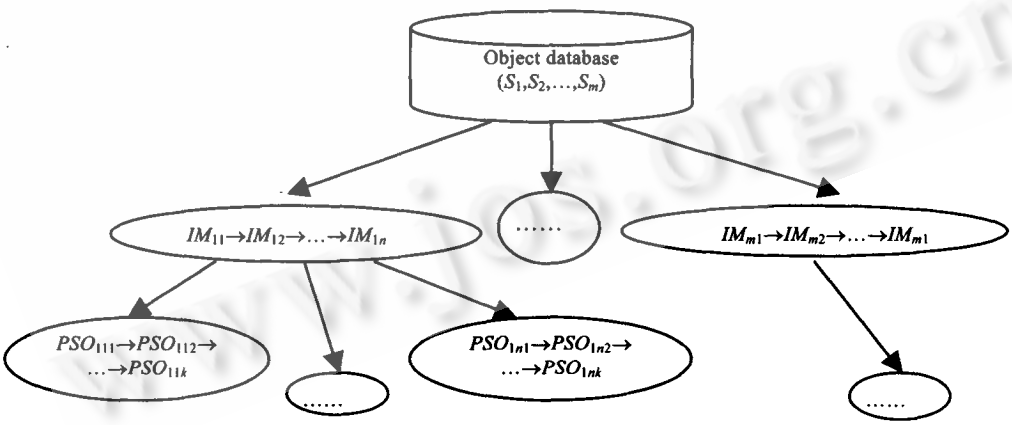


图 3 我们的数据组织框架

2 预处理阶段

在预处理阶段,已有的方法都是进行特征提取或者是对图像进行二值化,忽略了图像的最基本元素——像素本身的灰度所具有的意义,而且也没有很好的利用领域知识对图像进行预处理.

对于图像中噪声数据的消除,我们有效的利用了领域知识(记为 DK1)——人脑的颅骨的密度很高,在 CT 影像中表现为白色,因此它将我们研究感兴趣的的部分(大脑)和噪声数据(图像的背景和上面的符号)完全分离开来.如图 1(a)所示.我们利用图像处理中的剪切技术来消除这些噪声数据,并且给出了领域知识对预处理的影响

因子(factor):

$$\text{Factor}=\frac{\text{在DK1指导下剪切后的图像中的像素数}}{\text{原始图像的像素数}} \tag{1}$$

对于脑部医学图像,医生最关注的就是像素的位置和灰度的不同所反映出的信息,因此我们在预处理过程中完整的保留了图像中有意义像素的原始状态,避免了对图像进行二值化等处理可能造成的信息丢失,也避免了特征提取之后,忽略了像素本身所具有的意义.

图像经预处理后,我们将所有对象的数据组织成以下形式,见表 1.其中每个对象都有唯一的标识(ID),每个对象的图像部分(IM)是一个由预处理后的图像组成的序列,每个编号的格式是 ID.序号,表示该对象某一层面的图像,箭头表示了图像从头顶向下的位置关系.

表 1 数据组织形式

ID	IM
001	001.01→001.02→001.03→…→001.07
002	002.01→002.02→002.03→…→002.13
…	…
n	*n*.01→*n*.02→*n*.03→…→*n*.10

3 基于像素聚类的 PSO 检测

我们用以下的领域知识(记为 DK2)来指导聚类算法:(1) 正常人的大脑结构基本相同,具有明显的对称性.也就是说,相对脑中线(如图 1(b)所示)的大脑左右半球的密度分布基本相同,如果有占位,则密度会发生明显变化,而且破坏对称性;(2) 如果大脑出现占位,则它在若干个连续层面的图像上表现出来的可能性非常大,如图 1(c)所示.基于以上领域知识,我们引入如下概念来描述问题.

3.1 基本概念

定义 1. $S=\{S_i|i=1\ldots m\}$称为对象集;$S_i=\{IM_{i1},IM_{i2},\ldots,IM_{in}\}$称为对象或有序图像集,其中

(1) IM_{i1} 和 IM_{in} 是距离头顶最近和最远的两个图像;

(2) 对于任意前 p 个图像,IM_{ip} 必定是距离头顶最远的图像.

对于任意预处理后的图像 IM_p,被脑中线垂直平分后,由两部分组成:$IM_p(L)$表示大脑左半球图像,$IM_p(R)$表示大脑右半球图像.

定义 2. 对于任意的 IM_p 和 IM_j,如果

(1) $IM_p\in S_i$ 且 $IM_j\in S_i$;

(2) $p=j+1$ 或者 $p=j-1$;

则称 IM_p 和 IM_j 是相邻的.

定义 3. 我们称 $P=\{p_i|p_i$ 是位于图像中(x_i,y_i)的像素$\}$为 IM_p 的像素集.$P(L)$和 $P(R)$分别表示 $IM_p(L)$和 $IM_p(R)$的像素集.

依据 DK2 中大脑结构的对称性,我们假定 $IM_p(L)$和 $IM_p(R)$包含的像素数相同,即$|P(L)|=|P(R)|=|P|/2$.

定义 4. 对于任意的 $p_{li}\in P(L),p_{rj}\in P(R)$,如果 p_{li} 和 p_{rj} 的连线被脑中线垂直平分,则称 p_{li} 和 p_{rj} 互为对称点,以下记为 p_{li} 和 p_{ri}.

定义 5. 将像素集 P 中的像素划分为 m 块,每块中的像素具有相同的灰度,不同块中的像素灰度不同.如果第 i 块中的像素的灰度为 g_i',那么将 $g'_1,g'_2,\ldots,g'_m$ 按递增排序后得到的集合 $G(P)=\{g_1,g_2,\ldots,g_m\}$称为 P 的灰阶集合,其中,$g_i(i=1\ldots m)$称为第 i 级灰阶,g_1 和 g_m 分别为 P 的最小和最大灰阶.像素 p_i 的灰阶记为 $g(p_i)$.

定义 6. 如果 $g_{mean}(P)=\sum_{i=1}^{|P|}g(p_i)/|P|$,

则称 $g_{mean}(P)$为均值灰阶.

定义 7. 对于任意的 P 和距离函数 $DisA=|g_k-g_{mean}(P)|$,使 $DisA$ 最小的灰阶集合的中间值称为中值灰阶,由

中值灰阶组成的集合称为中值灰阶集合.

结论. 中值灰阶集合至多包含两个值.

证明:对于 *DisA* 的值由两种可能的情况:

(1) 如果 $g_k-g_{mean}(P)<0$,

那么 $DisA=g_{mean}(P)-g_k$;

(2) 如果 $g_k-g_{mean}(P)\geq 0$,

那么 $DisA=g_k-g_{mean}(P)$.

在情况(1)中,如果使 *DisA* 等于最小值 γ 的灰阶多于两个:$g'_1,g'_2,\ldots,g'_k(k>2)$,那么必有 $g'_1=g'_2=\ldots=g'_k$,这与灰阶集合的定义不符,因此在情况(1)中使 *DisA* 最小的灰阶只能有一个,记为 g_{mida}.

同理,在情况(2)中使 *DisA* 最小的灰阶也只能有一个,记为 g_{midb}.

如果 $g_{mean}(P)-g_{mida}=g_{midb}-g_{mean}(P)$,那么中值集合就包含两个元素;否则的话,中值集合就只包含 g_{mida} 和 g_{midb} 中使 *DisA* 最小的元素. □

定义 8. 对于任意 *P*,如果

(1) 中值集合中包含一个元素 g_{mid},g_s 是 g_{mean} 和 g_{mid} 中最小的灰阶;

(2) 中值集合中包含两个元素,g_s 是两个元素中最小的灰阶;

则称 g_s 为基准灰阶,而另一个灰阶记为 g'_s.

定义 9. 在像素集 *P* 中,我们称

$g^{(l)}=\{g_i|g_1\leq g_i\leq g_1+|g_1-g_s|/2\}$为低边缘灰阶;

$g^{(h)}=\{g_i|g_m-|g_m-g'_s|/2\leq g_i\leq g_m\}$为高边缘灰阶;

$g^{(b)}=g^{(l)}\cup g^{(h)}$为边缘灰阶.

定义 10. 在像素集 *P* 中,我们称

$P^{(l)}=\{p_i|g(p_i)\in g^{(l)}\}$为低边缘像素集;

$P^{(h)}=\{p_i|g(p_i)\in g^{(h)}\}$为高边缘像素集;

$P^{(b)}=P^{(l)}\cup P^{(h)}$为边缘像素集,

其中的像素称为边缘像素.

定义 11. 对于 *P* 中的任意对称点 p_{li} 和 p_{ri},设$\Delta g_i=g(p_{li})-g(p_{ri})$,则称$\Delta g(P)=\{\Delta g_i|i=1,2,\ldots,|p|/2\}$为 IM_p 的差值集合.

定义 12. 对于任意图像序列$\langle IM_{ij},\ldots,IM_{ik}\rangle$,如果

(1) 只有序列两端的 IM_{ij} 和 IM_{ik} 有一个相邻的 *IM*;

(2) 其他 IM_{ip}(如果存在的话)都有两个相邻的 *IM*;

则称此图像序列具有连续性.

定义 13. 对于任意图像序列$\langle IM_{ij},\ldots,IM_{ik}\rangle$,如果不满足连续性,则称此图像序列具有间断性.

定义 14. 对于任意像素 p_i 和某一给定正数ε,与像素 p_i 的距离小于等于ε的像素的全体称为ε-邻域.

定义 15. 若一个像素 p_i 的ε-邻域至少包含 *MP* 个符合条件的像素,则称 p_i 为核心像素;

定义 16. 如果像素 p_i 是另一个像素 p_j 的ε-邻域且 p_j 为核心像素,则称 p_i 是从 p_j 直接密度可达的.

定义 17. 对于一系列像素 $p_1,p_2,\ldots,p_k$,如果任意 p_{i+1} 都是从 p_i 直接密度可达的,则称 p_1 和 p_k 是密度可达的.

定义 18. 如果存在像素 p_k,它既与像素 p_i 是密度可达的,又与 p_j 是密度可达的,则称 p_i 和 p_j 是密度可连接的.

3.2 DK2指导下的聚类算法

对于医生来说,确定脑部图像中是否存在占位是非常关键的.我们将在图像的像素上使用聚类的方法来判断占位的可能性.首先,对于任意 IM_p,我们求出 IM_p 差值集合$\Delta g(P)$,对$\Delta g(P)$中的每个元素取绝对值并排序后得到集合$\Delta g'(P)=\{|\Delta g_i||\Delta g_i\in\Delta g(P)$且对于任意$|\Delta g_i|$必定是前 i 个元素中最大的$\}$.我们将灰阶等于$\Delta g'(P)$中每个元素的像素个数作为一个原子聚类,根据下面的像素灰阶间差值的相似性函数:

similarity(C_i,C_{i+1})=min$|T_i-T_{i+1}|$,其中 T_i 和 T_{i+1} 分别是聚类 C_i 和 C_{i+1} 中像素个数的均值进行自下而上的层次聚类,直到满足指定的聚类个数 k 为止.算法如下:

聚类算法 I.

输入:集合$\Delta g'(P)$和聚类个数 k.

输出:满足相似性函数 similarity(C_i,C_{i+1})的 k 个聚类.

1. 将灰阶等于$\Delta g'(P)$中每个元素的像素个数作为一个原子聚类,计算相邻聚类的$|T_i-T_{i+1}|$;
2. 根据 similarity(C_i,C_{i+1})进行聚类;
3. while (聚类个数不为 k) {
4. 计算相邻聚类的$|T_i-T_{i+1}|$;
5. 根据 similarity(C_i,C_{i+1})进行聚类;}

根据 DK2 中的"如果有占位,则密度发生明显变化,并可能破坏对称性",我们可以推断出假如存在占位,那么占位位置的像素灰阶会发生改变,在集合$\Delta g'(P)$中对应这些像素的元素将远大于 0,否则的话,在集合$\Delta g'(P)$中对应的元素将接近 0,因此我们把聚类个数设置为 2 作为终止条件.聚类算法 I 的第 1 步对集合$\Delta g'(P)$的$|P|/2$ 个元素扫描一遍即可,时间复杂性为 $O(|P|/2)$,第 2 步是选出所有最小值,因此也可以用 $O(|P|/2)$.第 3 步循环的次数与聚类的聚合速度有关,在最坏情况下,每次循环只有两个聚类聚合为一个更大的聚类,则循环次数为$|P|/2-3$,而第 4 步和第 5 步的时间复杂性则为 $O(|P|/2-i)$,i 为第 i 次循环,因此第 3~5 步最坏情况下需要$(n-3)(n+2)/2$ 次计算,时间复杂性为 $O(n^2)$.

按照 DK2 中大脑结构的对称性,我们选择聚类算法 I 中具有较大$|\Delta g_i|$的聚类(以下称为高差值聚类)作为下一步的主要研究对象,这意味着要处理的数据量可能会大大的降低.对于每个Δg_i 都有两个对称点 p_{li} 和 p_{ri} 与之相对应,我们将判断所有高差值聚类中对称点的 $g(p_{li})$和 $g(p_{ri})$是否属于边缘灰阶,生成若干边缘像素集.

接下来,我们使用基于密度的聚类方法对这些边缘像素集进行再聚类,确定每个脑部图像中占位的可能位置和大小.算法如下:

聚类算法 II.

输入:所有边缘像素集,ε和 mp.

输出:k 个聚类.

1. 假设所有边缘像素集的像素数为 bn,检查 bn 个像素的ε-邻域;
2. if 像素 p_i 的ε-邻域包含了多于 mp 的边缘像素
3. then 标记 p_i 为核心像素
4. while (所有核心像素) {
5. 聚合所有密度可达的像素;}

对于聚类算法 II 中的ε和 mp,我们通过对健康对象的脑部图像进行学习来指定.学习的过程如下:第 1 步,按照聚类算法 II 之前的步骤,对健康对象的 IM_p 进行第 1 次聚类,求出边缘像素集;第 2 步,对每个边缘像素集(这实际上属于噪声数据,而不是占位)进行统计(而不是聚类),计算出所有边缘像素集的半径和边缘像素数的最大值,已此来作为ε和 mp 的最大下界.此方法的计算复杂性为 $O(n\log n)$.通过此算法产生的 k 个聚类,就是 k 个可能的占位.

4 基于 ISSP 的相似性搜索

在这一节,我们将介绍两部分内容:(1) 发现一个对象的图像序列模式(ISP);(2) 发现对象之间的图像序列相似模式(ISSP).

4.1 发现图像序列模式(ISP)

通过以上两个聚类算法我们将得到所有的 PSO,我们把这些 PSO 记为如下形式:$\langle H(L),(x_i,y_i),(x_{a1},x_{a2}),(y_{b1},y_{b2})\rangle$,其中 $H(L)$表示高(低)边缘灰阶,(x_i,y_i)是 PSO 的中心坐标,(x_{a1},x_{a2})和(y_{b1},y_{b2})是 PSO 的最大和最小 x 坐标和 y

坐标.我们可以通过下面公式得到上面的参数:

$$x_i=\frac{1}{k}\sum_{j=1}^{k}x_j \tag{2}$$

$$y_i=\frac{1}{k}\sum_{j=1}^{k}y_j \tag{3}$$

$$x_{a1}=\max_{j=1}^{k}(x_j) \tag{4}$$

$$x_{a2}=\min_{j=1}^{k}(x_{ij}) \tag{5}$$

$$y_{b1}=\max_{j=1}^{k}(y_j) \tag{6}$$

$$y_{b2}=\min_{j=1}^{k}(y_j) \tag{7}$$

我们把大脑的中心作为坐标原点,如果 $x_i\leq 0$,那么这个 PSO 就属于 $IM(L)$,否则,这个 PSO 属于 $IM(R)$.

定义 19. 对于任意 $PSO_k=\langle H(L),(x_k,y_k),(x_{ka1},x_{ka2}),(y_{kb1},y_{kb2})\rangle$和 $PSO_j=\langle H(L),(x_j,y_j),(x_{ja1},x_{ja2}),(y_{jb1},y_{jb2})\rangle$,如果它们满足如下条件:

(a) $x_k<x_j$;(b) $x_k=x_j$ and $y_k>y_j$;(c) $x_{ka1}\leq x_{ja1}$ and $x_{ka2}\geq x_{ja2}$;(d) $y_{kb1}\leq y_{jb1}$ and $y_{kb2}\geq y_{jb2}$;

那么我们称 PSO_k 优先于 PSO_j,记为 $PSO_k>>PSO_j$.

按照这种优先性定义,我们将每个图像的 PSO 模式(PSOP)表示成如下形式:

$$PSOP(IM_i)=\langle L_{i1},L_{i2},\ldots,L_{im};R_{i1},R_{i2},\ldots,R_{in}\rangle,$$

其中,L_{im} 和 R_{in} 分别表示在 $IM_i(L)$和 $IM_i(R)$中的 PSO.

定义 20. 如果$|PSOP(IM_i)|=|PSOP(IM_j)|$,并且相应的 L_{ik} 和 L_{jk}(或者 R_{ik} 和 R_{jk})满足下面条件:

(a) 属于相同的边缘灰阶;(b) 同时属于 $IM(L)$或者 $IM(R)$;(c) 满足相同的优先条件;

那么我们称两个图像的 PSOP 是完全相似的.

定义 21. 如果从一个图像或这两个图像中移去一个或者多个不连续的 PSO 后使得$|PSOP(IM_i)|=|PSOP(IM_j)|$,并且相应的 L_{ik} 和 L_{jk}(或者 R_{ik} 和 R_{jk})满足下面条件:

(a) 属于相同的边缘灰阶;(b) 同时属于 $IM(L)$或者 $IM(R)$;(c) 满足相同的优先条件;

那么我们称两个图像的 PSOP 是不完全相似的.

一个对象的每一幅图像都具有一个 PSOP,它可能与其邻近图像的 PSOP 相差很大.我们处理一个对象的所有图像,得到这个对象的一个模式序列,我们称为图像序列模式(ISP).发现 ISP 的算法描述如下:

DISP 算法.

输入:一个对象的 m 个图像.

输出:图像序列模式(ISP).

(1) 初始化:$j=1,n_j=1,k=1$;

(2) For $i=1$ to m {

(3) 计算得到第 i 和$(i+1)$个图像的 PSO 模式:$PSOP(IM_i)$和 $PSOP(IM_{i+1})$;

(4) If 完全相似

(5) Then 记录 PSOP 为$\langle PSOP(IM_k),n_j=n_j+1\rangle$;

(6) Else if $j=i$

(7) Then $k=i$ 并且记录 PSOP 为$\langle PSOP(IM_k),n_j=1\rangle;j=j+1$;

(8) Else $k=i+1$ 并且记录 PSOP 为$\langle PSOP(IM_k),n_j=1\rangle;j=j+1;\}$

我们用一个例子来说明 DISP 算法,如图 4 所示.为了方便描述,我们将 ISP 记为如下形式:

$$ISP(S_i)=\{\langle M_i(IM_{i1}),n_1\rangle,\langle M_i(IM_{i2}),n_2\rangle,\ldots,\langle M_i(IM_{ik}),n_k\rangle\},\text{ or}$$

$$\{\langle M_i,n_1\rangle,\langle M_i,n_2\rangle,\ldots,\langle M_i,n_k\rangle\}$$

其中对于 $j=1,\ldots k$,M_j 和 M_{j+1} 是两个完全不同的 PSOP.IM_{ij} 是具有 PSO 模式 M_j 的第 1 个图像,n_j 是具有相同 PSO 模式 M_j 的相邻图像的个数.

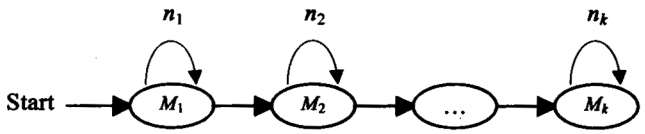


图 4 由 DISP 算法得到的模式

4.2 发现图像序列相似模式(ISSP)

对于任意 $ISP(S_i)$和 $ISP(S_j)$,两个对象的图像序列相似模式(ISSP)是指它们的最长相似连续子模式.例如,假设 $ISP(S_i)=\{\langle M_1,1\rangle,\langle M_2,2\rangle,\langle M_3,2\rangle,\langle M_4,1\rangle,\langle M_5,1\rangle\}$,$ISP(S_j)=\{\langle M'_1,1\rangle,\langle M'_2,3\rangle,\langle M'_3,3\rangle,\langle M'_4,2\rangle,\langle M'_5,1\rangle,\langle M'_6,1\rangle\}$,如果 M_1 同 M'_1 相似,M_3 同 M'_3 相似,M_4 同 M'_4 相似,M_5 同 M'_6 相似,那么两个对象的 ISSP 就是$\langle M'_1,M'_3,M'_4,M'_6\rangle$.这里,$M'_3$ 和 M'_4 是连续子模式,M'_1 和 M'_3 是间断子模式.

由于任意对象 S_i 中的图像都是具有空间位置关系的,也就是,对于 $IM_{i1},IM_{i2},\ldots,IM_{in},IM_{i(j+1)}$一定要比 IM_{ij} 远离头顶,因此没有必要检索对象所有的 PSOP 来找到与一个给定的 M_i 相似的模式.例如,M_1 是对象中离头顶最远的图像的 PSOP,而 M_p 是另一个对象中离头顶最近的图像的 PSOP,那么比较 M_1 和 M_p 的相似性就是没有意义的,因为他们反映的是大脑的不同部位.据此,我们引入两个规则来减小搜索空间以提高效率.对于两个对象 S_i 和 S_j,假设 $ISP(S_i)=m,ISP(S_j)=n$,

(1) 如果 $m=n$,也就是,两个对象的 PSOP 的规模相同,那么,我们只要检索 $ISP(S_j)$中的 M'_{i-1},M'_i 和 M'_{i+1} 来发现与 $ISP(S_i)$中 M_i 相似的模式即可;

(2) 如果 $m<n$,那么,我们只要检索 $ISP(S_j)$中的 $M'_i,M'_{i+1},\ldots,M'_{i+n-m}$ 来发现与 $ISP(S_i)$中 M_i 相似的模式即可.

下面是发现 ISSP 的算法.

DISSP 算法.

输入:两个对象的 ISP.

输出:图像序列相似模式(ISSP).

1. 初始化:C=NULL,NC=NULL;
2. 假设 $ISP(S_i)=m,ISP(S_j)=n$,如果 $m\le n$,开始下面的步骤:
3. for i=1 to m {
4. if $m=n$, then goto step (5); if $m<n$, then goto step (7);
5. 比较 $ISP(S_i)$中的 M_i 和 $ISP(S_j)$中的 M'_{i-1},M'_i 和 M'_{i+1};
   if 发现一个完全相似的 PSOP, then $C=C\cup(M_i\rightarrow M_j)$;
   if $ISP(S_j)$中没有 PSOP 可比较,then 返后 step (3)开始下一个迭代;
   if 没有发现完全相似的 PSOP,then goto step (6);
6. 比较 $ISP(S_i)$中的 M_i 和 $ISP(S_j)$中的 M'_{i-1},M'_i 和 M'_{i+1};
   if 发现不完全相似的 PSOP, then $NC=NC\cup(M_i\rightarrow M_j)$;
   if $ISP(S_j)$中没有 PSOP 可比较,then 返后 step (3)开始下一个迭代;
7. 比较 $ISP(S_i)$中的 M_i 和 $ISP(S_j)$中的 $M'_i,M'_{i+1},\ldots,M'_{i+n-m}$;
   if 发现不完全相似的 PSOP, then $C=C\cup(M_i\rightarrow M_j)$;
   if $ISP(S_j)$中没有 PSOP 可比较,then 返后 step (3)开始下一个迭代;
   if 没有发现完全相似的 PSOP,then goto step (8);
8. 比较 $ISP(S_i)$中的 M_i 和 $ISP(S_j)$中的 $M'_i,M'_{i+1},\ldots,M'_{i+n-m}$;
   if 发现不完全相似的 PSOP, then $NC=NC\cup(M_i\rightarrow M_j)$;
   if $ISP(S_j)$中没有 PSOP 可比较,then 返后 step (3)开始下一个迭代;}
9. 按照 M_i 的空间关系将 $C\cup NC$ 中的所有模式进行排序,然后利用穷举法来发现 ISSP;

由于以上的规则减少了被检索的PSOP的个数,算法的时间代价主要依赖于两个PSOP的相似性比较过程,而这个过程又是由第4.2节中的两个算法生成的PSO的个数决定的,因此,算法的时间复杂性为$O(km)$,其中k是PSO的个数.

定义22. 如果两个对象存在图像相似序列模式ISSP,那么,我们程两个对象是相似的.

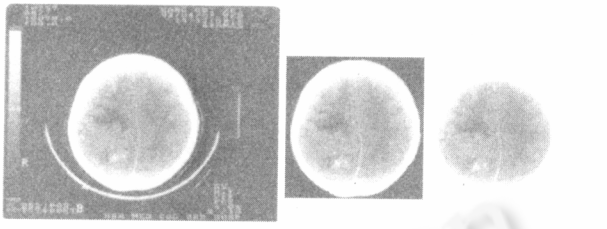
我们给出一个例子来说明这个算法.对于两个对象S_i和S_j,假设$ISP(S_i)=\{\langle M_1,1\rangle,\langle M_2,2\rangle,\langle M_3,2\rangle,\langle M_4,1\rangle,\langle M_5,1\rangle,\langle M_6,1\rangle\}$,$ISP(S_j)=\{\langle M'_1,1\rangle,\langle M'_2,3\rangle,\langle M'_3,1\rangle\}$,我们用DISSP算法得到如下结果:$C=(M'_2\rightarrow M_2,M_4)$和$NC=(M'_1\rightarrow M_1,M_3)\cup(M'_3\rightarrow M_6)$,其中,$C$和$NC$分别包含了完全的和非完全的相似模式.注意到$|ISP(S_j)|<|ISP(S_i)|$,我们将$C\cup NC$按照$ISP(S_j)$中$M_j$的空间关系进行排序.结果是$(M'_1\rightarrow M_1,M_3)\cup(M'_2\rightarrow M_2,M_4)\cup(M'_3\rightarrow M_6)$.利用穷举方法,最后的ISSP是$\langle M_1,M_2,M_6\rangle$和$\langle M_3,M_4,M_6\rangle$.因此,$S_i$和$S_j$是相似的.

5 实 验

在我们的实验中使用的数据集合完全是从医院中得到的真实的数据,没有任何模拟数据,主要是为了避免模拟数据可能导致的聚类和相似性搜索的可信度和准确率下降.在数据获取的过程中,我们遇到很多关于医院管理和个人隐私等方面的问题,同时也得到了有关专家的大力帮助和支持,使得我们已经获得了103份宝贵的资料,其中包括11份健康对象的脑部CT影像和92份患病对象的病史和脑部CT影像.在收集的数据中,所有图像都具有随机性,也就是说,图像中包含有什么样的潜在规律对于专家和我们来说是不知道的.

5.1 DK1指导下的预处理

对于没有领域知识指导的预处理,通常得方法是沿着图像的横向和纵向进行切割,如图5(b)所示.图5(c)是在领域知识指导下预处理后的图像.由于图像有不同的大小,每个图像的像素个数是不同的,因此,我们计算所有原始图像,没有领域知识指导的预处理后的图像和领域知识指导下的预处理后的图像得像素个数的平均值,来评估我们的预处理方法的结果,如图6所示.通过式(1)我们可以计算得到DK1指导下的预处理的影响因子.我们可以从结果中看到,在DK1指导下预处理后的所有图像的平均大小是所有原始图像平均大小的57.53%,是没有DK1指导的预处理图像平均大小的32.85%.很显然,在领域知识的指导下,计算复杂性大大降低了.



(a) Original image (b) Cropped image without DK1 (c) Cropped image with DK1

图5 在有无领域知识指导的图像预处理示例

5.2 DK1指导下的相似性搜索

我们从数据集合中随机抽取10%的正常对象和10%的患病对象作为目标,在整个数据集合里检索与每一个目标对象相似的对象.下面的公式用来评估我们算法的准确率和查全率.

$$P(p)=\frac{TP}{TP+FP},\qquad R(p)=\frac{TP}{TP+FN},$$

$$P(n)=\frac{TN}{TN+FN},\qquad R(n)=\frac{TN}{TN+FP},$$

其中,TP表示被正确分类到患病对象类里面的患病对象个数,FP表示通过算法被错误分类到患病对象类里面的正常对象个数,FN表示被分类到正常对象类里面的患病对象个数,TN表示被正确分类到正常对象类里面的正常对象个数.$P(p)$和$R(p)$表示对患病对象进行相似性搜索的准确率和查全率,$P(n)$和$R(n)$表示对正常对象进行相似性搜索的准确率和查全率.图7和图8是按照上述的方法进行5次随机抽取目标后进行相似性搜索的实验结果.我们可以看到,对于正常对象进行相似性搜索的时候,准确率和查全率都很高.实际上,每次只有一个患病对象被检索出来同其他的正常对象相似,因为这个患者的脑瘤形态非常对称,以致于我们的聚类算法在计算差

值集合的时候忽略了这个肿瘤.每一次都有一个正常对象没有被检索出来,因为这个对象的所有图像都非常亮,一些较暗的噪音像素点导致了错误的结果.

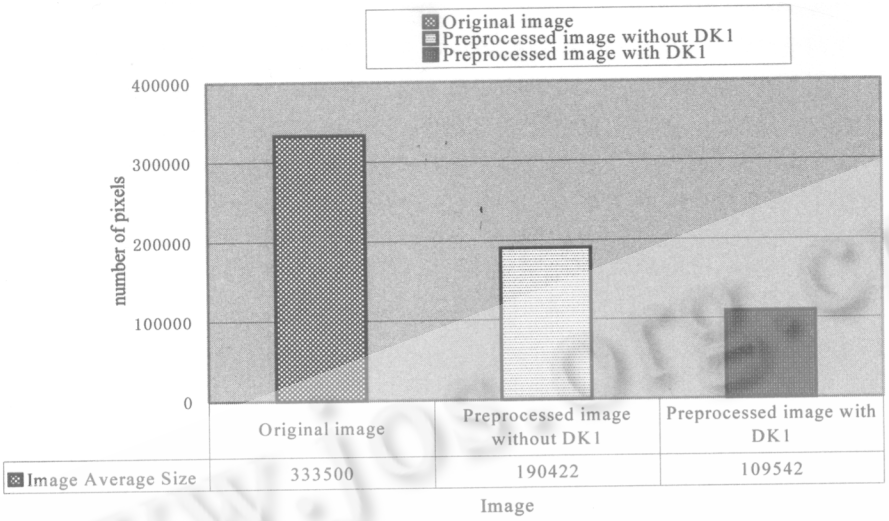


图 6 图像预处理效果的比较

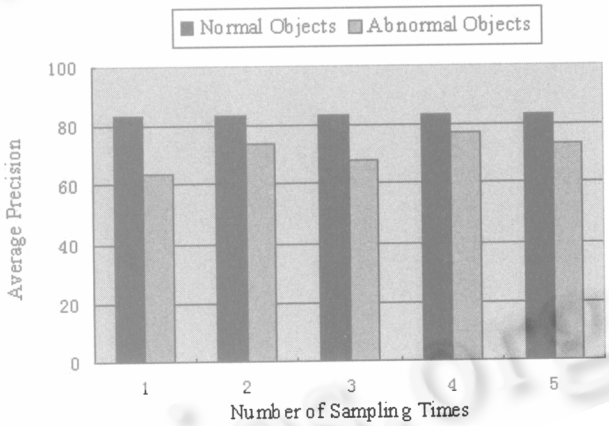


图 7 以正常对象和患病对象为目标进行相似性搜索的平均准确度

患病对象的相似性搜索准确率不是很高,大于 60%.在我们的结果中,所有被检索出来的对象都是脑部有肿瘤的患病对象,他们的图像序列与目标对象的很相似,而准确率不高的原因是由于他们的脑部肿瘤的具体类型与目标对象的不同,因此,医生做出了不同的诊断.但是,这可以表明我们提出的基于 ISSP 的相似性搜索算法可以获得与目标相似的图像序列.另一方面,患病对象的平均查全率却很高,这说明基于 ISSP 的相似性搜索算法能够从数据集合中检索出大部分与目标对象相似的对象.例如,在数据集合中实际上有 7 个对象(包括目标对象自身)与目标对象相似,我们的方法检索到 10 个相似的对象,其中有 6 个对象属于上面的 7 个对象.

另外,在我们的数据集合中,有两个对象在医生的第一次诊断中被误诊,最后的诊断不得不借助活组织切片技术,这种技术需要昂贵的开销,而且还需要从患者脑部提取活组织,我们的算法成功的检索到了与这两个对象诊断结果相似的对象.图 9(a)是一个被误诊对象的图像,图 9(b)是与之相似的对象的图像,这个结果将对医生的诊断提供很有意义的信息.

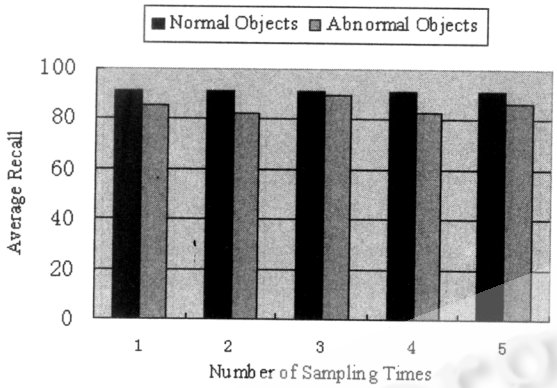


图 8 以正常对象和患病对象为目标进行相似性搜索的平均查全率

6 结论和未来的工作

随着脑部疾病(尤其是脑瘤)发生率的逐年上升,通过挖掘脑部医学图像来发现知识对辅助医生的诊断变得越来越重要.本文首先使用两个聚类算法来产生 PSO 集合,然后在这样的集合上来发现所有对象的 ISP.我们在文章中引入了图像序列相似模式(ISSP)的概念.对于各自包含一个图像序列的两个对象,ISSP 是指隐藏在他们中的最长相似连续子模式.这些模式在医学图像中是很有意义的,因为对医生来说两个对象相似要比两个图像相似更有意义.我们设计了新的基于领域知识指导下的算法来产生 ISSP,用于相似性搜索.我们的实验表明领域知识的指导是非常有意义的,而且相似性搜索的结果也对医生很有帮助.

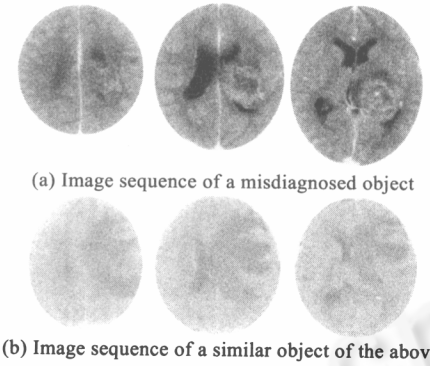


图9 具有相同临床诊断的两个对象的脑部图像序列

未来的工作包括进一步了解医学领域知识,并将其更好的融入到医学图像挖掘的其他算法(分类算法,聚类算法,关联规则挖掘等)的研究中.在医学图像和医学文本的混合集合上的数据挖掘具有更大的挑战性,而且在这样的集合上发现的知识更加丰富,更有利于辅助医生的诊断.

References:

[1] Osmar RZ, Jiawei H, Ze-Nian L, Jean H. Mining MultiMedia data. In: CASCON'98: Meeting of Minds. Toronto, 1998. 83~96.

[2] Wynne H, Mong LL, Ji Z. Image mining: Trends and developments. Journal of Intelligent Information Systems, 2002,19(1):7~23.

[3] Wynne H, Mong LL, Kheng GG. Image mining in IRIS: Integrated retinal information system. In: Proc. of the ACM SIGMOD. Dellas, 2000. 593.

[4] Chabane D. Relationship extraction from large image databases. In: Proc. of the 2nd Int'l Workshop on Multimedia Data Mining (MDM/KDD 2001). San Francisco, 2001. 44~49.

[5] Kitamoto A. Data mining for typhoon image collection. In: Proc. of the 2nd Int'l Workshop on Multimedia Data Mining (MDM/KDD 2001). San Francisco, 2001. 68~77.

[6] Ordonez C. Omiecinski E. Discovering association rules based on image content. In: IEEE Advances in Digital Libraries Conf. 1999. 38~49.

[7] Michael CB, Charless F, Joseph R. Mining for image content. In: Systems, Cybernetics, and Informatics/Information Systems: Analysis and Synthesis. Orlando, 1999.

[8] Vasileios M, Christos D, Edward HH. Mining Lesion-Deficit associations in a brain image database. In: KDD'99. San Diego, 1999. 347~351.

[9] Liu Y, Dellaert F, Rothfus WE, Moore A, Schneider J, Kanade T. Classification-Driven pathological neuroimage retrieval using statistical asymmetry measures. In: Proc. of the Medical Imaging Computing and Computer Assisted Intervention Conference (MICCAI 2001). Utrecht, 2001.

[10] Antonie M-L, Zaiane OR, Coman A. Application of data mining techniques for medical image classification. In: Proc. of the 2nd Int'l Workshop on Multimedia Data Mining (MDM/KDD 2001). San Francisco, 2001. 94~101.

[11] Zaiane OR, Antonie ML, Coman A. Mammography classification by an association rule-based classifier. In: Proc. of the 3rd Int'l Workshop on Multimedia Data Mining (MDM/KDD 2002). Alberta, 2002. 62~69.