软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(Suppl.(2)):178−187

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于图模型的医学图像聚类算法*

潘海为, 谷井子, 韩启龙, 谢晓芹, 张志强, 荣晶施

(哈尔滨工程大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

通讯作者: 潘海为, E-mail: panhaiwei@hrbeu.edu.cn

摘 要: 医学图像聚类算法的研究是面向特殊领域图像挖掘的重要组成部分,由于存在很多技术和特定领域方面的问题,使得这个方向的研究非常具有挑战性.已有的聚类算法对数据对象的形状和密度有要求,应用到医学图像聚类方面不能取得很好的结果.针对以上问题,在领域知识指导下,首先对图像进行了纹理检测,提出了面向纹理的T-LBP方法;然后对预处理之后的图像进行了空间划分,并对每个空间内的纹理求取LBP值,建立按空间序列排序的LBP直方图;最后,将以LBP直方图作为特征,提出了基于图模型的医学图像聚类算法.实验结果表明,该算法在时间复杂度和聚类结果方面具有良好的效果.

关键词: 医学图像;聚类;LBP;纹理;MCST;图模型

中文引用格式: 潘海为,谷井子,韩启龙,谢晓芹,张志强,荣晶施.基于图模型的医学图像聚类算法.软件学报,2013,24(Suppl.(2)):178−187. http://www.jos.org.cn/1000-9825/13035.htm

英文引用格式: Pan HW, Gu JZ, Han QL, Xie XQ, Zhang ZQ, Rong JS. Medical image clustering algorithm based on graph model. Ruan Jian Xue Bao/Journal of Software, 2013,24(Suppl.(2)):178−187 (in Chinese). http://www.jos.org.cn/1000-9825/13035.htm

Medical Image Clustering Algorithm Based on Graph Model

PAN Hai-Wei, GU Jing-Zi, HAN Qi-Long, XIE Xiao-Qin, ZHANG Zhi-Qiang, RONG Jing-Shi

(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Corresponding author: PAN Hai-Wei, E-mail: panhaiwei@hrbeu.edu.cn

Abstract: Clustering algorithm of medical image is a significant part of special field image clustering. Due to technical limit and many problems in specific area, the study in this direction has been very challenging. The exiting algorithms of clustering require shape and density of data object, which imply that there won't be a good outcome for the application of medical image clustering. To solve the problem above, this paper firstly detects texture from image, proposes T-LBP method, divides the preprocessed image into multiple spaces, calculates the value of LBP spaces, and then builds a spatial sequence LBP histogram. In the end, the clustering method of MCST is proposed based on the created LBP histogram. The outcome of this experiment indicates that the algorithm presented in this paper achieved good results in terms of time complexity and clustering function.

Key words: medical image; clustering; LBP; texture; MCST; graph model

近年来,随着计算机技术的发展,人体医学成像技术的应用越来越广泛,通过CT等医学图像进行医学诊断是非常重要的诊断方法之一.医学图像具有很大的数据量和丰富的特征信息,人工处理难以实现,因此针对医学图像领域的研究开始兴起,其中包括医学图像的聚类[1]、分类、关联规则挖掘与检索[2−7]等技术.图像的数据挖掘[8]可以发现图像数据中隐含的知识,包括图像内部的信息和图像之间的信息,并且根据这些信息来帮助医生

* 基金项目: 国家自然科学基金(61272184, 61202090, 61100007, 60803037); 黑龙江省自然科学基金(F200903, F201016, F201024, F201130); 中央高校自由探索计划(HEUCF100609, HEUCFT1202); 教育部新世纪人才支持计划(NCET-11-0829); 哈尔滨市青年科技创新人才研究专项基金(RC2010QN010024)

收稿时间: 2013-03-15; 定稿时间: 2013-07-11

进行诊断.目前医学图像的数据挖掘还是一个新兴领域,处于研究起步阶段,但却是一个非常有潜力,并且非常重要的研究领域.这个方向的研究可以辅助不同地域的医生进行高质量的诊断,提高医疗水平.

传统的数据挖掘与图像数据挖掘有很大的区别,传统的数据挖掘需要进行数据清理、数据融合等预处理操作,在图像数据挖掘中不仅进行这些操作,还要对图像进行分割、目标识别、特征提取等操作.在一系列预处理操作后才能对图像进行数据挖掘,此外,还要结合图像信息以及数据挖掘的方法,找到符合图像特征的数据挖掘方法.

图像的预处理操作是图像聚类的重要前提,预处理效果直接影响聚类的效果.根据不同的需要对图像进行不同的预处理变换,其中,感兴趣区域的检测、纹理检测[9−12]、小波变换等操作是图像预处理的重要方法.本文使用了 Canny 纹理检测算法对图像进行预处理.Simonyan 等人提出了一种医学图像的实时查询算法[13],用户可以选择感兴趣区域(ROI[14]),通过 ROI 来查询返回相似区域,返回的图像根据相似情况来排名.在图像检索的研究中,有很多图像特征提取的方法可以借鉴,并应用到聚类中.文献[15]提出一种图像查询方法,这种方法是基于手绘草图进行查询,然后实时地返回符合查询条件的图像,并且相似度越大的图像,排名越靠前.文献[16]介绍了一种基于图的图像分割聚类算法,把图像的边界区域看作子图,对像素聚类,进而对图像进行分割,这种算法的一个重要特征是保存了低变化区域的细节,忽略了高变化区域的细节,所以得到了较好的分割结果.并且,文献[16]中的聚类的特征是图像的光强度和颜色等,主要对图像进行分割,本文把这种聚类方法应用到多张医学图像的聚类过程中.

图像的 LBP 特征提取是一种重要的图像特征提取方法.Unay 等人介绍了旋转不变的 LBP 特征提取方法[17],中心点 gc 的灰度值作为一个阈值,周围的点根据这个阈值形成一串二进制的编码,灰度值大于 gc 的点,编码为 1,小于 gc 的点,编码为 0.然后,不断围绕中心点旋转邻域点即可得到一系列不同的二进制编码,取其最小值作为最后的 LBP 值.通过引入旋转不变的定义,LBP 算子不仅对图像的旋转表现得更为鲁棒,而且 LBP 的模式会进一步减少.Ojala 等人提出了一种特征提取的方法,关于灰度图像的简单而有效的基于 LBP 的旋转不变特征分析方法[18],介绍了使用不同半径、不同点的个数来求 LBP 的算法.例如当半径 R=1,且点的个数 P=8 时,求一个点的相邻 8 个点的 LBP,求得 LBP 后,对 LBP 的值进行循环移位 8 次,把最小值作为最后的 LBP 值.邻域点的个数不同,LBP 值不同,半径范围越大,包含的信息越多.Unay 等人介绍了对医学图像进行空间划分的图像的 LBP 值[19],以图像中心为圆心,可以划分不同的区域,并且以图像中心作为坐标原点,对图像进行等分,这样就把图像进行了分块,把分得的块进行编号,每一块得到一个直方图,按照块的顺序把每一块的直方图合并成一个整张图像的直方图.把这个大的直方图作为图像的特征,并且使用整张图像的直方图进行相似度度量,把整张图像的直方图对应块作差,求得的值就是相似度.Unay 等人提出的方法需要计算整张图像的 LBP 值,该方法计算量很大,花费时间较多.本文针对上述问题提出了面向纹理的 T-LBP 方法,本方法只需计算纹理点的 LBP 值,大大减少了计算点的个数,降低了计算量并缩短了所花费的时间.

已存在的聚类算法对数据的形状和密度有要求,有的对球状簇聚类结果比较好,有的对非球状簇聚类结果比较好,有的只对一种松散程度的数据聚类结果比较好.本文提出的 MCST(medical clustering method of minimum spanning tree)算法不仅对聚类的形状没有要求,而且对同一数据集中的不同松散程度的数据聚类结果也很好.在同一数据集中同时存在松散的点和密集的点,使用 MCST 聚类算法,可以把属于同一类中密集的点聚在一起,并且可以把属于同一类中松散的点聚在一起.

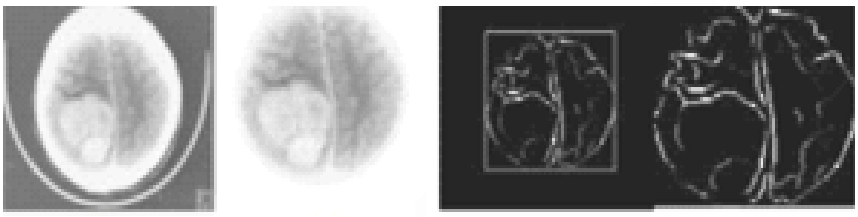
本文开始部分介绍背景知识及相关工作.第 1 节介绍面向纹理的 T-LBP 方法.第 2 节介绍基于图模型的聚类方法.第 3 节介绍所进行的实验.第 4 节进行全文总结.

1 面向纹理的 T-LBP 方法

本文针对文献[15]给出的特征提取方法中需要计算整张图像的 LBP 值的缺点,提出了面向纹理的 T-LBP 方法,此方法只需计算纹理点的 LBP 值,减少了计算量,降低了时间复杂度.

1.1 预处理过程

医学图像的数据挖掘需要用到图像处理领域的知识对图像进行预处理.本文对图像的预处理包括提取感兴趣区域,去掉大脑皮层,然后使用 Canny 检测算法提取图像的纹理信息.例如图 1 是脑部 CT 图像预处理的例子,图 1(a)所示是脑部 CT 图像,使用原图直接处理效果不好,因为图 1(a)中底下白色半圆圈不是脑部图层的信息,并且大脑皮层这部分纹理是干扰信息,对聚类结果影响较严重,所以先对图像提取感兴趣区域,只保存有用的脑部信息,如图 1(b)所示.然后对图 1(b)进行纹理检测得到图 1(c),再对图 1(c)矫正得到图 1(d),进行这些操作后得到了重要的脑部纹理信息.



(a) 原始图像 (b) 感兴趣区域灰度图像 (c) 纹理图像 (d) 矫正后的纹理图像

Fig.1 The brain CT

图 1 脑部 CT 图

定义 1. 集合 $M=\{g_c(x,y)|g_c(x,y)$是位于灰度图像中(x,y)的像素$\}$称为灰度图像点集;集合 $N=\{g_c'(x,y)|g_c'(x,y)$是位于纹理检测后的图像中(x,y)的像素$\}$称为纹理图像点集.我们用 $h(g_c(x,y))$表示集合 M 中任意像素点 $g_c(x,y)$的灰度值.

定义 2. 集合 $P=\{g_p(x',y')||x'-x|\leqslant 1\cap|y'-y|\leqslant 1,(x,y)\in M\}$称为集合 M 任一点 $g_c(x,y)$的八邻域点集.

给预处理后的纹理图像建立索引,每张图像对应一个存储纹理坐标的队列,队列中存储纹理图像点集 N 中的纹理点的横、纵坐标位置.每张图像都对应一个索引队列,通过索引找到纹理点的队列进行计算,根据图像对应的纹理队列找到图像所有纹理点的坐标,队列每次弹出一对坐标,直到队列为空.通过建立索引,可以避免遍历纹理图像,只需在图像对应队列中查找纹理点的坐标,从而极大地降低了时间复杂度.

1.2 面向纹理的T-LBP

面向纹理的 T-LBP 方法只计算纹理点的 LBP 值,相对于计算整个图像的 LBP 方法,大大减少了计算点的个数,降低了时间复杂度.下面具体介绍 T-LBP 方法.

T-LBP 方法的主要目的是计算纹理点的 LBP 值,对图像进行纹理检测后得到纹理图像,此时,纹理图像中只有纹理点的灰度值,其他点的值都为 0(即图像中显示为黑色),所以对于纹理图像不能直接计算 LBP 值.由于周围点的灰度值为 0,得到的 LBP 值不准确,因此,需要把纹理图像的纹理点坐标进行存储.每张图像的索引对应一个存储纹理的队列,T-LBP 方法通过纹理图像索引队列找到感兴趣区域灰度图像中对应位置的纹理点坐标,并且计算纹理点的 LBP 值.已知图像集合 M,N,对于集合 N 内的任意点 $g_c'(x,y)$,集合 M 内的点 $g_c(x,y),g_c'(x,y)$与 $g_c(x,y)$在横纵坐标都相等的情况下,在图像中的位置是相对应的.计算方法如公式(1)所示:

$$LBP=\sum_{p=0}^{p=7}s(h(g_p(x',y'))-h(g_c'(x,y)))2^p, g_p(x',y')\in M, g_c'(x,y)\in N \tag{1}$$

函数 s 的计算方法见公式(2):

$$s(e)=\begin{cases}1, & e\geqslant 0\\ 0, & e<0\end{cases} \tag{2}$$

当 e 的值大于等于 0 时,s 的值为 1;当 e 的值小于 0 时,s 的值为 0.如果集合 P 内的点的灰度值 $h(g_p(x',y'))$与点 $g_c'(x,y)$的灰度值 $h(g_c'(x,y))$的差大于 0,则 s 的值为 1;如果差小于 0,则 s 的值为 0.由此得到一串二进制编码,这一二进制编码就是 $g_c'(x,y)$的 LBP 值.

本方法通过纹理图像和感兴趣区域灰度图像之间的关系,求得集合 N 中纹理点的 LBP 值,如图 2 所示.T-LBP 方法只需计算黑色纹理处的 LBP 值,而文献[19]中所提方法需要计算整张灰度图像的 LBP 值,相比之下,T-LBP 方法计算量小,花费时间少.计算整张图像的 LBP 值的时间复杂度为 $O(8n)$,其中,n 为灰度图像的像素点的个数.T-LBP 方法的时间复杂度为 $O(8t)$,其中,t 是纹理点的个数.通过对数据进行分析可以看出,纹理点的平均数量约为感兴趣区域灰度图像的 1/10,由此可见,T-LBP 方法所需要计算的点的个数大为降低,因此,T-LBP 方法的时间复杂度为 $O(4/5n)$.

Fig.2 Grayscale texture image
图 2 灰度纹理图

由于纹理点个数比整个灰度空间点的个数少很多,所以纹理点的位置和数量稍有变化,就会对图像对比的结果有明显的影响.使用纹理的 LBP 作为图像的特征,更能准确地描述图像,因为纹理两侧的像素点的灰度值变化更加剧烈,面向纹理的 T-LBP 只描绘了变化部分的 LBP,而灰度空间 LBP 描绘了所有点的 LBP 值.所以,纹理空间 LBP 的值更加准确,而灰度空间 LBP 描绘了所有点的信息,其中很多点的 LBP 值没有体现图像的变化,这些信息不仅会造成干扰,而且浪费了时间和空间信息.普通图像通常包含比较明确的语义信息,例如水果、动物、人物等,而医学图像中的语义信息比较模糊,无法给出一个灰度区域具体的语义,只能通过对纹理的提取和分析来做出判断,因此,T-LBP 方法更适用于医学图像.

2 基于图模型的聚类算法

上述预处理过程执行之后,本节将对纹理图像进行划分,划分的规则是以图像中心为圆心划分成圆形区域,并且以图像中心为基准划分成扇形区域.划分之后利用 T-LBP 方法对每个划分区域进行计算,得到其 LBP 直方图.

现有的聚类方法不能在输入较少参数的情况下得到任意形状和不同密度的结果,本文针对上述问题,提出了基于最小生成树的医学图像聚类方法(MCST 方法),此方法对聚类形状没有要求,可以得到任意形状的聚类,并且在聚类过程中,并没有局限聚类的密度,在聚类结果中可以将稀疏点聚为一类,密集点聚为一类,得到不同松散程度的聚类.另外,由于离群点对聚类结果的影响较小,当数据集中存在离群点时,它与其他数据之间的距离比较大,也就是说,相似度较小,MCST 方法可以区分出相似度较小的点,使它不能聚到某一类中,成为孤立点.

2.1 特征提取

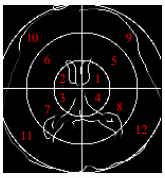


Fig.3 Partition texture image
图 3 空间纹理图

对图像进行空间划分,可以对图像的细节进行更好的表达,根据对应区域的特征比较图像之间的相似度.图像的空间划分是将一张 CT 图像分成不同的区域,基本思想如下:把图像的中心点作为圆心可以划分出不同圆形区域,并且以中心点作为中心把图像进行等分.我们规定将纹理图像按照从内到外的顺序划分成圆形区域,再以图像中心为圆心划分扇形区域,图像的标号顺序为从内圈到外圈,并且从 x 轴正半轴上方逆时针依次标号.

例如:图 3 中划分成 3 个圆形区域和 4 个扇形区域,图像转换为集合 $S(s1,\ldots,s12)$,得到一个 12 维的向量.利用 T-LBP 方法对 12 个区域进行计算,可以得到 12 个 LBP 直方图.

假设经过空间划分后的区域数为 num,$Hist[i]$代表 num 个区域中的第 i 个区域的 LBP 直方图,$Hist[i]$由 256 个灰度作为横坐标,纹理 LBP 值在某一灰度值的个数作为纵坐标,如图 4 所示.$Hist[1],\ldots,Hist[mn]$代表从区域 1 到区域 $m\times n$ 的直方图.由于 CT 图像是灰度图像,灰度范围从 0~255,所以直方图的横坐标是 0~255 的灰度值,纵坐标 r 是 T-LBP 在对应灰度值下的数量.

为了使计算结果更准确,本文对直方图进行归一化处理,见公式(3):

$$Hist[i]=\frac{value(r)}{sum(number)} \tag{3}$$

其中,*value*(*r*)是直方图中对应每个灰度级别的纹理点的个数,*sum*(*number*)的值为总的纹理点的个数.

图像的空间划分将图像划分为不同的小区域,这样可以使提取的特征表达更准确,不会造成相互干扰,如果对一整张图像提取 LBP,那么其中的灰度信息是整张图像的灰度信息,对图像结构和细节的表达不够准确.对图像进行空间划分后,只需比较对应区域的 T-LBP 特征.

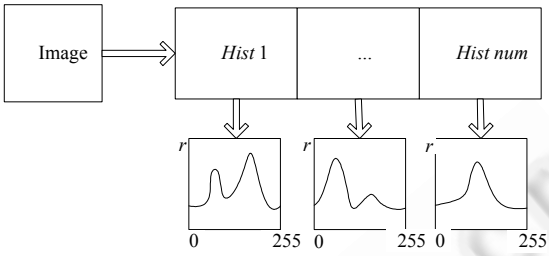


Fig.4 Histogram constitution
图 4 直方图构成图

2.2 图模型

本文通过基于图的最小生成树方法把数据集合 G 划分为不同集合,每个集合内相似度较高,不同集合内相似度较低.划分后的集合用 S 表示,即 $S(C_1,\ldots,C_r)$.

基于图论的方法是解决聚类问题的有效途径,将每张医学图像看作是图的一个顶点,我们构建了无向完全图 $G=(V,E)$,其中,$V=\{v_1\ldots v_n\}$为顶点集合,v_i 表示一张医学图像,$E=\{e_{ij}|1\leqq i\leqq n,1\leqq j\leqq n$ 且 $i\in V,j\in V\}$为边的集合,e_{ij} 为顶点 v_i 和 v_j 构成的边.对于每条边,v_i 和 v_j 之间都有一个相似性度量的权值 $w(v_i,v_j)$.权值 $w(v_i,v_j)$的计算方法是对第 i 张图像和第 j 张图像对应的划分区域求距离,划分的每个小区域的距离是小区域的直方图的差.计算方法见公式(4):

$$w(v_i,v_j)=\sqrt{\sum_{x,y=1}^{x,y=num}(G_i(Hist[x])-G_j(Hist[y]))^2} \tag{4}$$

其中,变量 x,y 是区域的标号,取值从 1 到 num.图像之间的距离是相应直方图的距离之和,直方图之间距离的计算方法见公式(5):

$$Dis(G_i(Hist[x]),G_j(Hist[y]))=\sqrt{(G_i(Hist[x])-G_j(Hist[y]))^2}=\sqrt{\sum_{index=0}^{index=255}(r_{i,x}[index]-r_{j,y}[index])^2} \tag{5}$$

其中,$Dis(G_i(Hist[x])),G_j(Hist[y]))$是图像 G_i 第 x 个区域的直方图和图像 G_j 第 y 个区域的直方图的距离.变量 $index$ 代表灰度值的取值范围(0~255),r 是在相应像素点对应的纵坐标的值,这个值为 T-LBP 值归一化后的值.

2.3 MCST聚类算法

本文使用基于图的最小生成树方法进行聚类,在聚类过程中,S 是 G 的分类,集合 $C\in S$ 是 G 的子图,符合图 $G'=(V,E')$,并且 $E'\in E$.相同集合内的元素相似,不同集合内的元素不相似.也就是说,同一子图内顶点之间的边权值低,不同子图之间顶点的边有高权值.度量标准是聚类的关键所在,本文用 $Dif(C_1,C_2)$来表示两个类的相似度,$Dif(C_1,C_2)$为两个子图顶点之间的最小距离,见公式(6):

$$Dif(C_1,C_2)=\min_{v_i\in C_1,v_j\in C_2}w(v_i,v_j) \tag{6}$$

如果两个子图之间没有边,那么 $Dif(C_1,C_2)$的值是∞. 当 $Dif(C_1,C_2)\leqq MInt(C_1,C_2)$时,我们将两个子图合并为一个图,即聚为一类.对 $MInt(C_1,C_2)$的计算见公式(7):

$$MInt(C_1,C_2)=\mathrm{MIN}(Int(C_1)+\tau(C_1),Int(C_2)+\tau(C_2)) \tag{7}$$

其中,$\tau(C)=k/|C|$,k 是给定的参数,$|C|$是子图内元素的个数.对于每个子图 C,$Int(C)$是子图 C 的最小生成树(MST)的最大值,用来表示子图的内相似度,可见公式(8):

$$Int(C)=\max_{v_i\in C,v_j\in C} w(v_i,v_j) \tag{8}$$

其中,$w(v_i,v_j)$是权值,v_i和v_j属于子图C.在公式(7)中添加$\tau(C)$的目的是使算法更健壮,因为当集合C内只有一个元素时,$Int(C)$的值为0,此时,只要C_1,C_2两子图之间有边就会合并,所有离群点都会分配到其他集合中,也就是说,聚类结果将没有离群点.引入$\tau(C)$很好地克服了这一缺点,并且对聚类算法有很好的调整,在后面将会提到这一点.对于参数k的取值,可以根据需要进行调整,当k值大的时候,说明聚类的条件是宽松的,此时,类内元素的个数相对较多,类内距离也较大,当k值较小时说明聚类条件是严格的,此时,类内元素较少,类内距离也小.我们可以根据实际需要来调整k的值,使聚类结果满足要求.基于图模型的聚类算法如下所示:

MCST 算法.
输入:$G=(V,E)$;
输出:划分$S=(C_1,\ldots,C_r)$.

1.首先把图G中的边按照递增顺序排列,得到集合$O=(o_1,\ldots,o_m)$.
2.以S^0作为初始聚类,每个顶点v_i是一个初始集合.
3.for(q=0; q=1 to m; $q<m$) {
4. 按顺序排列的第q条边由v_i和v_j两个顶点连接,$o_q=(v_i,v_j)$.
5. 通过S^{q-1}构造S^q,过程如下:
If (v_i和v_j不在分割S^{q-1}中,并且$w(o_q)$比两个点的子图内部的相似度要小)
then {合并两个子图}
If (C_i^{q-1}是包含v_i的S^{q-1}的子图,并且v_j所在的子图是C_j^{q-1})
then If ($C_i^{q-1}\neq C_j^{q-1}$ and $w(o_q)\leqslant MInt(C_i^{q-1},C_j^{q-1})$)
then {从S^{q-1}通过合并得到S^q}
else {$S^q=S^{q-1}$}
6. If(合并后的S^q不存在环)
then {合并C_i^{q-1}和C_j^{q-1}以及$o_q=(v_i,v_j)$,形成新的子图集合}
If(合并后的S^q存在环)
then {将C_i^{q-1}和C_j^{q-1}包含的对象重新生成最小生成树,形成新的子图}
}
7.返回$S=S^m$.

MCST 算法中,G是一个图,S是对图G的划分.首先将图G的边按照升序排列,得到集合$O=(o_1,\ldots,o_m)$.边按照升序排列说明两个子图内部的最小生成树的值小于两子图之间的边,此时判断D为true.S^0作为初始聚类,S^0中的点是G中的顶点,每个顶点是一个初始集合.G中排好序的边作为循环条件,$o_q=(v_i,v_j)$代表按顺序排列的第q条边由v_i和v_j两个顶点连接.然后通过S^{q-1}构造S^q:首先判断v_i和v_j是否在分割S^{q-1}中,并且$w(o_q)$比两个点的子图内部的相似度要小,如果不在,合并v_i和v_j所在的子图;再判断C_i^{q-1}是否包含v_i的S^{q-1}的子图,并且v_j所在的子图是否为C_j^{q-1},如果不是,则$S^q=S^{q-1}$,构造完成;如果是,继续判断$C_i^{q-1}\neq C_j^{q-1}$和$w(o_q)\leqslant MInt(C_i^{q-1},C_j^{q-1})$,如果是,从$S^{q-1}$通过合并得到$S^q$.接下来判断合并后的$S^q$是否存在环,如果不存在,则合并子图$C_i^{q-1}$和$C_j^{q-1}$,以及边$o_q=(v_i,v_j)$,形成新的子图集合;如果存在,则将$C_i^{q-1}$和$C_j^{q-1}$包含的点新生成最小生成树,形成新的子图.反复进行上述过程,直到循环结束.

在算法 MCST 步骤 6,如果不进行环的判断,那么每次都需要重新生成最小生成树,时间复杂度较高,合并后的子图如果不存在环,则仍然是最小生成树,此时不需要重新生成.下面的定理证明了合并后不存在环的子图仍然是最小生成树.

定理. C_1,C_2代表V的一个划分,且都是最小生成树,假设C_1,C_2之间有一条边,这条边满足两个集合合并的条件,并且合并后的集合不形成环,那么合并后的新集合C'一定是最小生成树.

证明:假设集合C'不是最小生成树,此时把C'内的点重新生成最小生成树,至少不包含原来C'中的其中一条边,因为本算法是每次选择最小的边进行合并,所以重新生成的最小生成树的边的权值之和一定大于原来C'的权值之和,所以假设不成立. □

MCST 算法的初始数据集是图像集合,不限制数据集中对象间的松散程度和形状,并且能够很好地克服离

群点的影响.如果在初始数据集中既有松散的数据,又有密集的数据,此聚类算法可以把密集的聚在一起,松散的对象聚在一起,因此不限制对象的密度.由于此方法基于最小生成树来实现,只计算每个划分内的最小生成树,点之间的联系是点之间的距离,也就是相似程度,所以对原始数据集的形状没有要求.根据 k 值的不同来调节聚类的严格程度,当聚类要求松散时,k 值相对较大,当聚类要求严格时,k 的值较小.此方法对离群点的处理也非常好,如果数据集中存在离群点,那么说明这个离群点与其他数据的距离较远,相似度低,根据本文的算法可以区别出离群点.

3 实验结果与分析

实验数据包括 300 张医学图像.实验分为两部分,第 1 部分使用 100 张图像进行聚类结果时间的比较;第 2 部分使用 200 张图片进行实验误差率的比较.

医学图像聚类的结果以误差率来评判结果的好坏,误差率计算方法见公式(9),其中,p 是分类错误的图像数,q 是总的图像数.

$$e=\frac{p}{q} \tag{9}$$

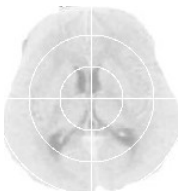


Fig.5 Space grayscale graph
图 5 空间灰度图

文献[19]把一张 CT 图分成不同的区域,图像的中心点作为圆心可以划分出不同区域,并且以中心点作为中心把图像进行等分,如图 5 所示.该方法把图像限定为不同的区域,根据划分出的区域来计算每个区域对象的灰度直方图,这个灰度直方图是由此图像的灰度 LBP 值得到的.根据每张图像的对应区域的灰度直方图进行比较,计算相似度.

下面的实验通过灰度图像的 LBP 直方图和纹理图像的 LBP 直方图使用 MCST 方法进行聚类.首先比较 T-LBP 特征提取方法和灰度图像的 LBP 特征提取方法所花费的时间,由于 MCST 方法对聚类形状和松紧度没有要求,聚类结果较好,所以,使用 MCST 方法对提取的特征进行聚类,与将 5 层大脑图片混合到一起的 200 张两组图像的聚类结果的误差率进行对比,最后再分别比较同一脑部图层的使用 MCST 方法的图像的聚类结果的误差率.

3.1 聚类时间结果分析

图 6 是对 100 张图像分别使用 T-LBP 方法和传统的整张图像 LBP 方法进行特征提取的时间比较结果,实验结果表明,纹理图像中纹理点的个数平均值为感兴趣区域灰度图像点的个数的十分之一,所以,T-LBP 特征提取所计算的点的个数为整张图像点的个数的十分之一.由此可知,T-LBP 特征提取时的计算时间也比整张图像 LBP 特征提取的时间要短,如图 6 所示为提取两种特征的时间比较图.其中横坐标是图像的序号,一共 100 张图像,序号从 1~100;纵坐标是时间,单位是 ms.其中方框点是 T-LBP 特征提取所用的时间分布,菱形点是整张图像的 LBP 特征提取的时间分布.从图中可以看出,AllLBP 特征提取的时间大概在 10ms~12ms 之间,而 T-LBP 特征提取的时间在 0~2ms 之间.实验说明,使用 T-LBP 方法可将特征提取的时间复杂度降低一个数量级.

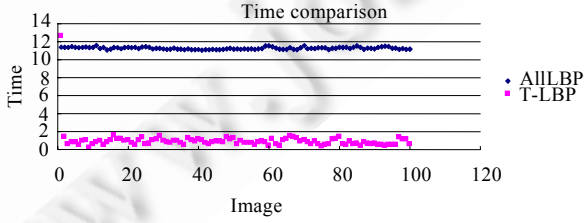


Fig.6 Time comparison
图 6 时间比较

3.2 聚类结果误差率分析

以下所有实验结果图中菱形框连出的线是使用 T-LBP 进行特征提取后,并且使用 MCST 算法进行聚类的误差率曲线;方框连出的线是对整张图像进行特征提取后,并且使用 MCST 算法进行聚类的误差率曲线图.其中,横坐标是 MCST 算法中 k 的取值,纵坐标是误差率.

接下来分别对图像的 T-LBP 特征和整张图像的 LBP 特征使用 MCST 方法进行聚类,并且进行误差率比较.通过对图像库进行模拟,得到了 200 张图像,把 200 张图像分为 2 组进行聚类,每组是 5 层大脑图像.聚类结果如图 7 所示.从图 7 中可以看出,本组图像使用 T-LBP 方法进行特征提取,当 $0<k<0.048$ 时,误差率较高并且逐渐降低,此时,k 值较小,松散度高,聚类个数多,同一类内的距离较小,随着 k 值的增加,聚类内部的距离增大,类内对象的个数增多;当 $k=0.048$ 时,误差率最小为 0.04,远小于 AllLBP 聚类的误差率,此时,k 值的松散度对此组图像最合适;当 $k>0.048$ 时,误差率逐渐增大,此时,聚类个数减小,类内距离逐渐增大,把不是同一类的图像也聚到了一起,所以误差率逐渐增大.对整张图像进行 LBP 特征提取的 AllLBP 曲线大概一直维持在 0.7 左右.所以,当 $k=0.048$ 时,T-LBP 特征提取方法的误差率远远低于对整张图像进行特征提取后的误差率,此时聚类效果最好.

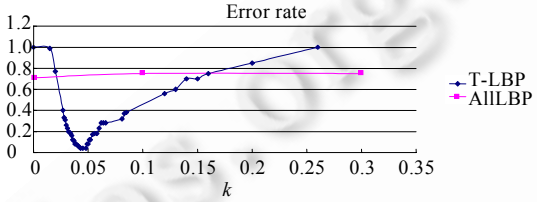


Fig.7 Error rate 1

图 7 误差率比较 1

在图 8 中,当 $0<k<0.056$ 时,误差率较高并且逐渐降低,此时,k 值较小,松散度高,聚类个数多,同一类内的距离较小,随着 k 值的增加,聚类内部的距离增大,类内对象的个数增多;当 $k=0.056$ 时,聚类结果的误差率最小,为 0.09,远小于 AllLBP 聚类的误差率;当 $k>0.056$ 时,聚类个数减小,类内距离逐渐增大,把不是同一类的图像也聚到了一起,所以误差率逐渐增大.对整张图像进行 LBP 特征提取的 AllLBP 曲线大概一直维持在 0.7~0.9 之间,所以 T-LBP 特征提取方法的误差率远远低于对整张图像进行特征提取后的误差率.

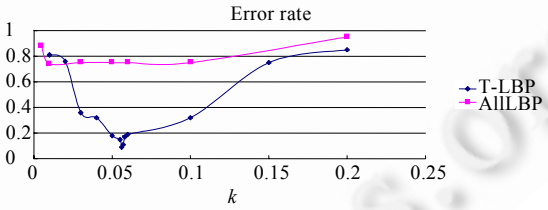


Fig.8 Error rate 2

图 8 误差率比较 2

上面混合 5 层大脑图像的两组实验结果表明,T-LBP 的 MCST 方法有较好的聚类结果,误差率远低于对整张图像提取 LBP 的方法.

接下来是同一脑部图层聚类结果图,图中对两种特征进行了误差率比较.在图 9 中,当 $0<k<0.05$ 时,误差率较高并且逐渐降低;当 $0.05<k<0.15$ 时,聚类结果的误差率最小为 0,远小于 AllLBP 聚类的误差率;当 $k>0.15$ 时,T-LBP 方法的误差率逐渐升高.对整张图像进行 LBP 特征提取的 AllLBP 曲线大概一直维持在 0.7~0.8 之间,所以,当 $0.05<k<0.15$ 时,T-LBP 特征提取方法的误差率远远低于对整张图像进行特征提取后的误差率,此时,类间的差异较大,所以,在这个范围内误差率都为 0.

如图 9 所示的实验中,当误差率从高到低下降时,随着 k 值的增加,聚类内部的距离增大,类内对象的个数增

多,当误差率降到最低时,说明 k 取此时的值对聚类的结果最好.当误差率由低到高逐渐升高时,类内距离逐渐增大,把不是同一类的图片也聚到了一起,所以误差率逐渐增大.从实验结果可以看出,对于同一层脑部图像,T-LBP 特征提取方法的最小误差率远小于整个图像的 LBP 特征提取方法的误差率.

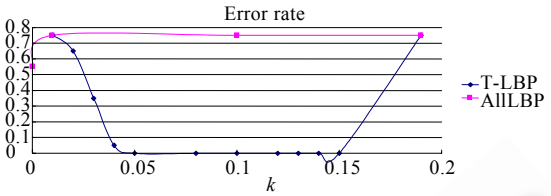


Fig.9 Error rate 3

图 9 误差率比较图 3

4 总 结

本文总结了我们对医学图像进行特征提取,然后根据提取的图像特征进行聚类的工作过程.本文采用 T-LBP 特征提取方法,提出了一种 MCST 聚类方法.MCST 算法是一种基于最小生成树的聚类算法,通过调整参数 k 的值,可以实现对任意形状、任意密度的数据进行很好的处理,使得相对密集且相似度较高的点聚在一起,相对松散且相似度较高的点也能聚在一起.医学图像的数据挖掘相关研究是为了辅助医生进行临床诊断,通过对医学图像的聚类可以帮助医生分辨不同的脑部图层的扫描结果,也可以对某一类相似疾病进行判断.本文提出的基于图模型的聚类方法可以作为自动医学诊断系统的一部分,辅助医生判断病人的病变类型以及病变情况.

References:

[1] Pan HW, Li JZ, Zhang W. Incorporating domain knowledge into medical image clustering. Applied Mathematics and Computation, 2007,2(185):844−856.

[2] Datta R, Joshi D, Li J, Wang JZ. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, 2008, 40(2):5−60.

[3] Swain MJ, Ballard DH. Color indexing. Int'l Journal of Computer Vision, 1991,7:11−22.

[4] Liu GH, Yang JY. Image retrieval based on the texton co-occurrence matrix. Pattern Recognition, 2008,41(12):3521−3527.

[5] Quellec G, Lamard M, Cazuguel G, Cochener B, Roux C. Fast wavelet-based image characterization for highly adaptive image retrieval. IEEE Trans. on Image Processing, 2012,21(4):1613−1623.

[6] Xu XQ, Lee DJ, Antani S, *et al*. Spine *x*-ray image retrieval using partial vertebral boundaries. IEEE Trans. on Information Technology in Biomedicine, 2008,12(1):100−108.

[7] Hsu W, Antani S, Long LR, *et al*. SPIRS: A Web-based image retrieval system for large biomedical databases. Int'l Journal of Medical Informatics, 2009,78(1):S13−S24.

[8] Han JW, Kamber M. Data Mining: Concepts and Techniques. 2nd ed., Beijing: China Machine Press, 2006. 396−399.

[9] Grigorescu C, Petkov N. Contour detection based on nonclassical receptive field inhibition. IEEE Trans. on Image Processing, 2003,12(7):729−739.

[10] Arbelaez P, Maire M, Fowlkes C, *et al*. Contour detection and hierarchical image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,33(5):898−916.

[11] Catanzaro B, Su BY, Sundaram N, *et al*. Efficient, high-quality image contour detection. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2009. 2381−2388.

[12] Wang B, Fan SS. An improved CANNY edge detection algorithm. Computer Science and Engineering, 2009, 497−500.

[13] Simonyan K, Zisserman A, Criminisi A. Immediate structured visual search for medical images. In: Proc. of the MICCAI. 2011. 288−296.

[14] Bradski G, Kaebler A, Wrote; Yu SQ, Liu RZ, Trans. Learning OpenCV. Beijing: Tsinghua University Press, 2009 (in Chinese).

[15] Cao Y, Wang CH, Zhang LQ, *et al*. Edgel index for large-scale sketch-based image search. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2011. 761−768.

[16] Felzenszwalb P, Huttenlocher D. Efficient graph-based image segmentation. Int'l Journal of Computer, 2004,59(2):167−181.

[17] Unay D, Ekin A, Cetin M, Jasinschi, Ecil A. Robustness of local binary patterns in brain MR image analysis. In: Proc. of the 29th Annual Int'l Conf. of the IEEE EMBS. 2007. 2098−2011.

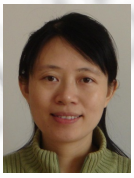
[18] Ojala T, Pietikäinen M. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002, 971−987.

[19] Unay D, Ekin A, Jasinschi RS. Local structure-based region-of-interest retrieval in brain MR images. IEEE Trans. on Information Technology in Biomedicine, 2010, 897−903.

附中文参考文献:

[14] Bradski G, Kaebler A,著;于仕琪,刘瑞祯,译.学习 OpenCV.北京:清华大学出版社,2009.

潘海为(1974－),男,辽宁瓦房店人,博士,副教授,CCF 高级会员,主要研究领域为数据库,图像挖掘.
E-mail: panhaiwei@hrbeu.edu.cn

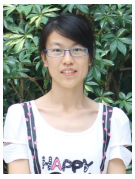


谢晓芹(1973－),女,博士,副教授,CCF 高级会员,主要研究领域为社会网络分析挖掘,数据挖掘.
E-mail: Xiexiaoqin@hrbeu.edu.cn

谷井子(1986－),女,硕士生,主要研究领域为数据库,图像挖掘.
E-mail: jingzigu@live.cn

张志强(1973－),男,博士,教授,CCF 高级会员,主要研究领域为数据库,信息检索.
E-mail: zqzhang@hrbeu.edu.cn

韩启龙(1974－),男,博士,副教授,CCF 高级会员,主要研究领域为空间数据库.
E-mail: hanqilong@hrbeu.edu.cn



荣晶施(1989－),女,硕士生,主要研究领域为数据库,图像挖掘.
E-mail: rongmeng1988@gmail.com