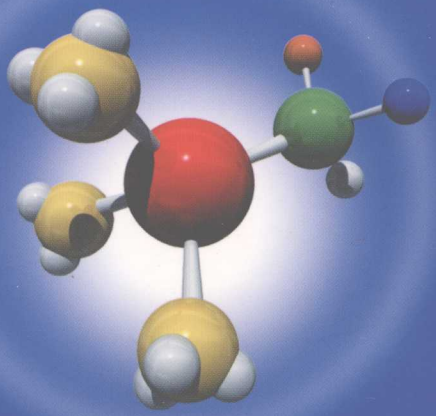
医学图像数据

挖掘关键技术的研究

Yixue Tuxiang Shuju Wajue Guanjian Jishu de Yanjiu

潘海为 著



黑龙江人民出版社

医学图像数据挖掘关键技术的研究

潘海为 著

黑龙江人民出版社

图书在版编目(CIP)数据

医学图像数据挖掘关键技术的研究/潘海为著. —哈尔滨：黑龙江人民出版社，2007.9
ISBN 978-7-207-07502-4

Ⅰ.图… Ⅱ.潘… Ⅲ.医学图像—数据采集
Ⅳ.R445-39

中国版本图书馆 CIP 数据核字(2007)第 151289 号

责任编辑：姜海霞
封面设计：张 涛

医学图像数据挖掘关键技术的研究
潘海为 著

出版发行 黑龙江人民出版社
通讯地址 哈尔滨市南岗区宣庆小区 1 号楼
邮 编 150008
网 址 www.longpress.com
电子邮箱 E-mail hljrmcbs@yeah.net
印 刷 黑龙江省地质测绘印制中心印刷厂
开 本 880×1230 毫米 1/32
印 张 5.5
字 数 13 万
版 次 2007 年 9 月第 1 版 2007 年 9 月第 1 次印刷
书 号 ISBN 978-7-207-07502-4/TP·13
定 价 12.00 元

(如发现本书有印制质量问题，印刷厂负责调换)
本社常年法律顾问：北京市大成律师事务所哈尔滨分所律师赵学利、赵景波

前　言

多媒体数据获取和存储技术的飞速发展导致了大规模多媒体数据库的出现。多媒体数据的类型多种多样，包括图像、文本、视频、音频等等。对这些多媒体数据进行挖掘分析，能够揭示出很多有意义的信息和知识。在多媒体数据库上的挖掘称为多媒体数据挖掘。图像挖掘是多媒体挖掘的一个重要组成部分，可以提取出那些隐含在图像数据库中的知识或模式。最近，以医学图像为对象的图像挖掘形成了一个重要研究领域 - -医学图像挖掘。医学图像挖掘是一个多学科交叉的研究方向，涉及到医学、计算机视觉、图像处理、图像检索、机器学习、模式识别、人工智能、数据库和数据挖掘等等。传统数据挖掘算法很少考虑图像数据和医学领域知识，不适于医学图像的数据挖掘。

本书针对医学图像的特点，结合医学领域知识，以脑部医学图像（包括与图像相关的文字）为例，集中研究了医学图像挖掘问题，提出了一系列医学图像关联规则、分类、相似性搜索和聚类等挖掘算法。本书的创新之处主要表现在如下五个方面：

第一，针对已有关联规则算法中存在的只注重优化频繁项集生成过程，不关注规则的生成过程，只注重优化算法的执行速度，不关注挖掘知识的质量，尤其没有考虑应用领域知识提高算法效率和挖掘知识的质量等问题，本书提出了在领域知识指导下的医学图像关联规则挖掘算法。在频繁项集的生成算法中，本书根据

领域知识给出了一个约束函数，对同时出现在同一个项集的对象进行了限制，大大减少了频繁项集的生成。在规则生成算法中，本书给出另一个约束函数，对出现在规则前后项的对象进行限制，降低了规则的数量，同时提高了规则的质量。

第二，本书针对医学图像集合以及医学图像与相关文字混合的数据集合，构建了一个通用的分类器。对于医学图像集合，本书提出了关联分类方法。该方法首先应用基于三级粒度表结构的关联规则挖掘算法，将上一级粒度表挖掘得到的结果用于组成下一级粒度表，大大压缩了每次扫描的数据表的规模。由于生成规则的后项限制为类标识，因此降低了挖掘规则的数量。本书将挖掘得到的关联规则作为分类规则构造了关联分类模型，同时给出了一个用于分支选择的判定规则，对关联分类模型无法给出类标识的图像，使用基于神经网络的方法对医学图像进行分类，取得很好的效果。对于图像和文字混合的数据集合，本书在医学领域知识的指导下提出了基于决策树的分类算法，给出了与图像相关的文字的泛化规则和分类度量属性的优先级，避免了选择分类度量属性的复杂计算过程。在复杂混合数据分类过程中，采用联机获取领域知识的方法，增加了分类的准确率和可靠性。

第三，本书针对医学图像的序列形式，提出了图像序列相似模式（Image Sequence Similarity Patterns，记为 ISSP）的概念。对于各自包含一个图像序列的两个病患，ISSP 是指隐藏在他们中的最长相似连续子模式。这些模式在医学上具有很重要的意义，因为对医生来说两个病患的相似（即图像序列相似）要比两个单一的图像相似更有实际意义。本书设计了基于领域知识的算法来发现每个病患的图像序列模式 ISP 和病患之间的 ISSP，以支持医学图像

此为试读，需要完整PDF请访问：www.ertongbook.com

序列相似性搜索，提高了检索的准确率。

第四，本书以具有诊断意义的关键像素区域（Region of interest，记为ROI）为核心，提出了基于ROI的两级聚类策略对医学图像进行聚类。在医学领域知识的指导下，本书首先从ROI中提取最相关的特征，基于这些特征定义了ROI的相似性度量，提出ROI聚类算法。接下来，本书应用信息检索中的理论定义了ROI在图像中的权值和基于ROI权值的图像相似性度量，给出了图像聚类算法，获得很好的聚类效果。

第五，目前还没有一个完整的针对医学图像挖掘的系统框架。本书提出了一个多级知识驱动的医学图像挖掘框架，由五级组成：像素级挖掘器、ROI级挖掘器、图像级挖掘器、语义级挖掘器和描述性文字挖掘器。每一级的挖掘工作都可以向用户和更高级别的挖掘过程提供知识。这个框架具有较好的通用性。

本书是在作者博士学位论文的基础上完成的。在写作过程中，我的导师李建中教授给予了我极大帮助和无微不至的关怀，为我的论文能够最终顺利完成付出了大量的心血。在此，我要衷心地感谢他引领将我走上了数据库研究的道路，教会了我从事科学研究的方法，并为我树立了今后从事科研工作效仿的楷模。

十分感谢黑龙江人民出版社的编审张晔明先生为本书的出版所给予的大量帮助和悉心指导。

限于作者的实践经验和学术水平，书中疏漏与不当之处在所难免，恳请专家和读者批评指正。

潘海为

2007年9月

目 录

Contents

此为试读,需要完整PDF请访问: www.ertongbook.com

第一章 绪 论

1.1 医学图像挖掘技术简介及意义

二十世纪90年代以来，随着信息技术的发展，信息系统在迅速变化。社会信息化水平正在成为影响一个国家科技及经济发展的重要因素，信息的占有量和利用率被公认为是衡量一个国家综合实力的重要指标。一场关于信息技术的竞争正在世界范围内激烈进行，谁的收集、处理、存储、传播、利用信息的能力强，谁就能够在竞争中占领制高点，掌握主动权，从而变得更强大。

数据的爆炸性增长（全球拥有的数据量每20个月翻一番[1]）和数据库应用的普及，使人们正逐步陷入“数据丰富，知识贫乏”的尴尬境地。数据挖掘（Data Mining，简称DM）是20世纪末兴起的数据智能分析技术，由于其具有广阔的应用前景，因而备受关注。数据挖掘也称知识发现（Knowledge Discovery from Dataset，简记KDD），被定义为从数据中发现隐含的、具有潜在用途的、人类可理解的知识[2]。数据挖掘通过发现有用的新规律和新概念，提高了数据拥有者对大量原始数据的深层次理解、认识和应用[3]。

多媒体数据是一类重要的数据资源，包括图像、音频、视频、文本等各种类型的数据。目前，多媒体数据的获取和存储技术的发展已经促进了大规模多媒体数据库的飞速发展。各个领域每天都有大量的多媒体数据产生，例如医学图像（CT图像，ECT图像，核

磁共振图像）、人造卫星图像以及各种数字设备采集的图像、音频、视频和文本等等。这些多媒体数据包含了大量对人们有用的信息。但是，在多媒体数据中发现这些潜在的知识和模式是十分困难的。

多媒体挖掘是数据挖掘领域的重要分支，它可以自动地从大量多媒体数据中发现隐含的知识或者模式。根据数据类型的不同，多媒体挖掘又分为图像挖掘、视频挖掘、音频挖掘、文本挖掘等等。图像挖掘是多媒体挖掘领域的重要分支，目前多媒体挖掘的大部分工作都集中在图像挖掘方面。图像挖掘旨在自动地从大量图像中发现隐含的知识或者模式，这项工作引起了数据挖掘工作者越来越多的重视。

图像挖掘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、神经网络、统计学、模式识别、知识获取、信息检索、人工智能、数据库和数据挖掘等。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。

医学图像是图像挖掘的一个重要应用领域。医生每天都要看大量的影像，往往会因为疲劳或者其它个人原因而导致诊断的准确率降低。如果几个不同医院的医生（尤其是有经验的教授或者专家）联合进行诊断，会大大提高诊断的准确率，但是这种做法的代价是非常大的，大多数情况下不太可能，往往只是由一个医生来对病人的影像做出诊断。判断病人是否患有疾病主要是看影像上是否有占位。如果有，病人应当尽早的进行治疗，以防止病情恶化而加大治疗的难度。但是有些时候占位并不明显，通常医生很难看出来，只有结合病人的病史，反复分析影像，凭借丰富的经验才能做出正确的诊断。这就是说，一旦诊断失误，那就会延误病人的

早期治疗，导致病情恶化，甚至会危及到病人的生命或者将来的生活，例如，造成视觉障碍，偏瘫等等。正因为如此，在医学图像和相关文字上研究数据挖掘来发现知识，对辅助医生的诊断和提高医学经验的共享都有极其重要的意义，而且由于它具有很强的领域性，使得这方面的研究具有非常大的挑战性。

本课题的目的是在医学领域知识的指导下，研究具有高效率和高准确率的医学图像（包括图像和相关文字）数据挖掘算法，以获得有用的知识和模式，辅助医生进行高质量的诊断。

1.2 图像挖掘技术的研究进展

本书将从以下三个方面来介绍医学图像挖掘技术的国内外现状和研究进展：数据挖掘、图像挖掘和医学图像挖掘。

1.2.1 数据挖掘的研究进展

数据挖掘是一个新兴的领域，从 90 年代初开始兴起，在短短几年内得到了迅速的发展。多数人把数据挖掘看作是知识发现的同义词[4]，也有人把数据挖掘看作知识发现中必不可少的一个步骤[2,5]，如图 1－1 所示[6]。图中包括 3 个主要阶段：数据准备、挖掘操作、结果表达和解释。知识发现可以描述为这 3 个阶段的反复过程[2]。

一般地，数据挖掘的任务可以分为两大类：描述性数据挖掘和预测性数据挖掘。前者用简单总结的方式描述数据，表示出人们感兴趣的一般数据特点；后者则构造一个或一组模型，完成在有效数据集上的推论，并且能够预测新的数据集合的行为。

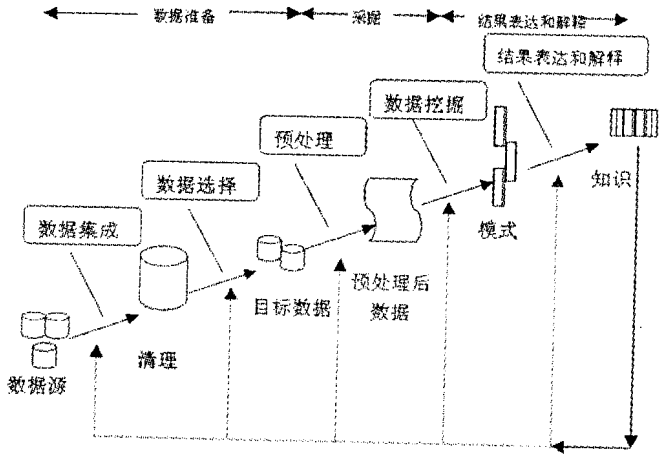


图 1-1 数据挖掘在知识发现过程中的位置

Figure 1-1 Data mining as a step in the process of knowledge discovery

数据挖掘的研究工作主要包括如下几个方面:关联规则、分类、聚类、相似性搜索、时间序列分析和预测算法。下边综述数据挖掘的主要研究结果。

(1)关联规则(Association rule)挖掘算法的研究

关联规则的挖掘是数据挖掘的一个重要研究领域。自从Agrawal等人于1993年提出关联规则[7]的概念,并给出在大数据量上的有效算法以来,已经出现了很多关联规则的挖掘算法。较为经典的算法是Apriori、AprioriHybrid[8],后来的改进算法有DHP[9]、DIC[10]、抽样算法[11]、分割算法[12]、多层次挖掘算法[13]以及一些并行算法等。关联规则挖掘算法的核心是频繁项集(Frequent Itemset)的求解。在频繁项集的求解过程中,需要计算指定的候选项集(Candidate Itemset)中的所有数据项同时出现在同一个记录中的次数。以下称为计数计算。关联规则挖掘算法可以分

为四类。

第一类关联规则挖掘算法称为经典算法。经典算法以文献[14]提出的 Apriori 算法为基础,分两步发现关联规则。第一步,确定频繁项集[14]。第二步,在频繁项集之间发现关联规则。算法第二步相对简单,时间复杂性较小。经典算法的研究重点集中在算法第一步的改进上。利用频繁项集具有向下封闭性(即频繁项集的子集仍然是频繁项集)[14]的特点,研究者提出了很多减少求解频繁项集时间复杂性的算法[3,7,8,9,10,11,12,14,15,16],如 FUP 算法[3]、DHP 算法等。经典算法需要扫描数据库许多遍,扫描遍数不小于最长频繁项集的长度,一般大于 5。

第二类算法是基于抽样和分割的算法。这类算法是针对第一类算法 I/O 时间复杂性高的问题提出来的,包括基于抽样的算法[11,14,15,16]、基于分割的算法[12]、多方法相结合的算法[3,10,15]。基于抽样的算法包括两类。一类是基于抽样的基本算法[15,16],另一类是扩展的抽样算法[11,14]。基于抽样的基本算法首先对大型数据集合进行抽样,然后在抽样数据上计算频繁项集,最后对原始数据集合进行一次扫描,完成关联规则的挖掘。虽然基于抽样的基本算法具有很高的效率,但不能保证算法的完整性,即可能丢失频繁项集,进而遗漏一些关联规则。为了解决基本抽样方法的完整性问题,人们又提出了扩展的抽样算法。基于分割的算法[12]分三步来计算频繁项集。第一步,把数据集合分割成很多小块,使得每块的所有局部频繁项集可以在内存中求出;第二步,把所有块的局部频繁项集并在一起,得到局部频繁项集集族(原始数据集合的频繁项集集族的超集);第三步,对整个数据集合进行一次扫描,加工处理局部频繁项集集族,计算出原始数据集合的全部频繁项

此为试读,需要完整PDF请访问: www.ertongbook.com