全国高等医药类院校计算机规划教材

医学数据挖掘
——SQL Server 2005案例分析

周 怡 王世伟 主编 刘建平 彭 勇 副主编

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

全国高等医药类院校计算机规划教材

医学数据挖掘
——SQL Server 2005 案例分析

周　怡　王世伟　主　编
刘建平　彭　勇　副主编

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

本书是一本很值得一读的医学案例数据挖掘教材，较全面地介绍了医学数据挖掘的基本任务、方法及数据挖掘技术及基于 SQL Server 2005 的医学实践。

全书共 7 章，内容涵盖核心的数据挖掘技术，包括：决策树算法、回归与时序算法、人工神经网络、关联规则和聚类分析。此外，还提供了医学数据挖掘的最佳实践方法论，介绍了 SQL Server 2005 中数据挖掘的功能，并且对这些功能结合医学实例作了较为详细的讲述。在附录中还提供了 SQL Server 2005 的安装方法。通过学习本书，读者能理解数据挖掘的重要性并学会如何实践数据挖掘。

本书适合作为高等院校相关专业高年级本科生、研究生的教材或参考书，也适合当前医学数据挖掘实践者学习和参考。

图书在版编目（CIP）数据

医学数据挖掘——SQL Server 2005 案例分析/周怡，王世伟主编. —北京：中国铁道出版社，2008.5
全国高等医药类院校计算机规划教材
ISBN 978-7-113-08799-9

Ⅰ. 医… Ⅱ. ①周… ②王… Ⅲ. ①医学—数据采集—医学院校－教材 ②关系数据库—数据库管理系统，SQL Server 2005—医学院校—教材 Ⅳ. R—05 TP311.138

中国版本图书馆 CIP 数据核字（2008）第 071044 号

书　　名：医学数据挖掘——SQL Server 2005 案例分析
作　　者：周　怡　王世伟　主编

策划编辑： 严晓舟　秦绪好
责任编辑： 王占清　　**编辑部电话：**（010）63583215
封面制作： 白　雪
责任印制： 李　佳　　**责任校对：** 侯　颖　张国成

出版发行： 中国铁道出版社（北京市宣武区右安门西街 8 号　　邮政编码：100054）
印　　刷： 北京京海印刷厂
版　　次： 2008 年 6 月第 1 版　2008 年 6 月第 1 次印刷
开　　本： 787mm×1092mm　1/16　**印张：** 10　**字数：** 221 千
印　　数： 5 000 册
书　　号： ISBN 978-7-113-08799-9/TP・2830
定　　价： 18.00 元

编审委员会主任、主审简介

王世伟，1950 年 12 月生，现任教育部高等学校医药类计算机基础课程教学指导分委员会委员、中国医科大学计算机中心主任，教授，硕士研究生导师，理工学部副主任。现担任中华医院管理学会信息管理专业委员会委员，全国高等院校计算机基础教育研究会医学专业委员会副主任委员，辽宁省高等院校计算机基础教育研究会副理事长，辽宁省卫生信息化建设专家组首席专家，辽宁省医学影像技术学会常务理事等学术职务。20 年来一直从事计算机基础教育与科研工作。在国家级核心期刊发表论文 30 余篇，主持国家、省级科研课题 3 项，其中“构建医学特色的《大学计算机基础》课程体系”系辽宁省“十一五”规划课题，在此基础上主编出版了《医学信息系统教程》、《现代医学影像技术》、《网站的规划与建设》等 20 余册全国高等医药院校计算机规划系列教材。

周怡，1954 年 4 月生，广东药学院医药信息工程学院院长，医药信息学教授，东南大学本科及硕士研究生毕业。
主要研究方向有：计算机科学与网络技术在医药中的应用、医药信息整合与数据挖掘、智能化医药信息系统工程等。
教育部高等学校（本科）计算机基础教学指导委员会医药类计算机基础课程教学指导分委员会委员（2006~2010）。
全国高等院校计算机基础教育研究会医学专业委员会副主任委员（2005~2008）。
中国卫生信息学会卫生信息技术应用专业委员会委员（2004~2007）。
广东省“南粤教书育人优秀教师”。

邹赛德，中国医药信息学会常务理事、中华医院管理协会信息专业委员会委员、广东医院管理学会信息管理专业委员会主任委员、广东医药信息学会理事长、中国电子学会医药信息学分会常务委员、广州市卫生局的咨询小组成员；是《医药信息技术基础》、《计算机应用基础》、《医学计算机应用基础》、《医学计算机应用基础实验教程》四本本科规划教材的主编，《医学计算机实用教程》研究生规划教材的副主编和《医院管理学信息管理分册》专著的编委。在全国医药院校计算机教学中具有表率和影响的作用，同时在广东医学信息学领域内居学术领导地位。在国内同行中具有较高的威信和影响力。

全国高等医药类院校计算机规划教材

编审委员会

本书编委会

主　编：周　怡　王世伟

副主编：刘建平　彭　勇

编　委：（按姓氏拼音排序）

蔡永铭　杜珠英　蒋世忠　刘建平

彭文青　彭　勇　王世伟　杨　进

易　珺　查　涛　周　怡

序

随着 21 世纪科教兴国战略的实施及信息化社会进程的加速，形成了以信息化带动医药卫生事业现代化的整体发展趋势，并深刻地影响与改变着传统的医药科学，使今天的医学工作者和医药院校的师生们同样面临着 IT 知识更新的机遇和挑战。

我国医疗卫生信息化建设同时面临着两个十分紧迫又必须解决的重要问题：对高素质医学人才不断快速增长的需求；现行医药院校的计算机与信息技术基础教学体系中的 IT 知识结构不够全面以及专业领域的 IT 操作技能不适用。

本着“社会的进步靠科学，科学的进步靠人才，人才的培养靠教育，教育的发展靠理念”的精神，我们必须努力构建具有医学特色的“大学计算机基础与应用”课程体系，紧密结合本学科信息化建设与应用的发展方向，用科学发展观来培养能进行自主学习，且适应未来社会发展及医药信息化建设需求的合格医学人才。为了推进全国高等医学院校计算机基础课程体系的教学改革，做好教材建设先行的工作，我们编写了“全国高等医药类院校计算机规划教材”丛书。

“全国高等医药类院校计算机规划教材”丛书，先期包括计算机与信息技术基础类、程序设计基础类及医学 IT 实用技术基础类，共三大类 12 种教材。涵盖了全国高等医药院校本、专科各专业的计算机与信息技术应用基础课程的教学用书。教材内容覆盖面和知识点的取舍本着全面系统、科学合理、结合专业、注重实用、知识宽泛、关注发展的六项原则，比较完整地构建了具有医学特色的计算机与信息技术基础课程教材体系。其中：

- 计算机与信息技术基础类：包括《医学计算机与信息技术应用基础》、《医学计算机与信息技术应用基础实验指导》、《医学多媒体实用技术教程》、《医学网络实用技术教程》；
- 程序设计基础类：包括《Visual FoxPro 程序设计教程》、《Visual FoxPro 程序设计上机指导与习题集》、《Visual Basic 程序设计教程》、《Visual Basic 程序设计上机指导与习题集》；
- 医学 IT 实用技术基础类：包括《医学统计实用技术教程》、《医学信息系统教程》、《医学影像实用技术教程》、《医学数据挖掘——SQL Server 2005 案例分析》。

“全国高等医药类院校计算机规划教材”丛书的编写与出版，得到了国内许多著名医药院校的鼎力支持与合作，教材的编者包括国内医学院校知名教育专家、教育部医药类计算机基础课程教学指导分委员会委员，大量从事一线教学工作、具备丰富教学经验的教师。他们

视角独特，洞见非凡，匠心独运地将目前开展得如火如荼的IT与医疗卫生信息化建设及实践以这套全国高等医药类院校计算机规划教材丛书的形式表现了出来。中国铁道出版社对本系列教材进行了远见卓识的精心策划、科学论证、倾力帮助、编辑发行等大量认真而卓有成效的工作。此间，还有全国许多从事计算机基础教学方面的领导、专家、老师和同学们给了我们很大的支持和帮助，在此一并表示衷心的感谢。

在此，我们愿把这套规划教材丛书奉献给全国医药院校的师生们，为构建我国具有医学特色的计算机与信息技术应用基础课程体系，共同走出一条新路，在深化教学改革及先行教材建设中做出贡献。

王世伟　周怡

2006年7月

前　言

医学数据挖掘是当前研究的热门领域，它是从大量医学数据中提取出可信的、新颖的、有效的，并最终能被人理解的模式的处理过程，涉及数据库、人工智能、统计学、模式识别、可视化技术、并行计算等众多领域知识。

为了更好地利用这一技术，许多医药院校的相关专业已经开设或者准备开设这一门课程。在作者多年教学过程中，发现医药院校的师生们急需具有如下特点的教材：一是内容较为全面、系统，并具有开放性；二是符合医药院校课程体系特点，既要讲清楚数据挖掘技术的基本工作原理，又不能讲得太深、太多；三是让学生可以结合实例跟着做，使学生在实践中理解数据挖掘技术的基本原理、应用领域与方法。

正是基于这一原因，《医学数据挖掘—— SQL Server 2005 案例分析》这本教材出现了。本书将医学实例、数据挖掘技术、计算机应用程序三者结合起来。首先介绍了数据挖掘任务；然后介绍了完成数据挖掘任务的方法论基础；最后根据数据挖掘任务，依据数据挖掘的方法论，借助于 SQL Server 2005 详细介绍了医学数据挖掘领域的重要问题——决策树算法、回归与时序算法、人工神经网络、关联规则和聚类分析。

总体来说，本书主要具有如下 3 个特点：

一是注重方法论。因为掌握良好的方法是学习数据挖掘的关键和基础。

二是强调基本概念和基本方法，并注重利用现有计算机应用程序。因为医学数据挖掘是一个新兴领域，学生只有在正确理解基本概念，掌握基本方法的基础上才有可能在以后的工作中使用数据挖掘技术。同时，考虑到学生的基础知识与课程安排，不可能对所有数据挖掘技术都讲得非常全面和深入。本书对数据挖掘算法的广度和深度做了如下处理：所讲述的数据挖掘算法具有系统性，尽量全面，当论述比较简单的算法的时候，尽量深入介绍算法的基本原理；当论述比较难的算法的时候，只是详细介绍了算法的概念以及使用方法、条件，然后借助于现有的计算机软件（如 SQL Server 2005）来讲解算法，这样处理有利于学生理解数据挖掘的基本原理和掌握数据挖掘技术。

三是强调实践性。虽然数据挖掘很有用，但如果学生不知道如何使用数据挖掘，数据挖掘的教学也就失去了意义，本书每章后面都安排了一个数据挖掘实例。

由于本书主要目的是让读者学会结合 SQL Server 2005 做数据挖掘，最佳教学学时为 54 学时（理论知识 36 学时，实验 18 学时）。因此，对各种数据挖掘模型的对比以及 SQL Server 2005 参数的设置都未作深入讨论。鉴于此，我们编写了《数据挖掘模型深入探讨及实验讲义》，结合本书使用，可用于 72 学时和 90 学时的教学安排。选用本书作为教材的教师，可以发邮件向出版社索取本实验讲义和课件。

虽然我们为本书能够尽早为读者服务付出了努力，同时也相信本书会是一本可读的数据挖掘技术及其实践的初级教程，但由于水平有限，目前医学数据挖掘并不是很成熟，再加上时间有限，书中会有不妥之处，我们期待着您的批评和建议。在使用过程中，如有任何意见或建议，请发邮件到 zhouzhouyi@163.com 或 zhouyi@gdpu.edu.cn 或 py.china@126.com，谢谢！

编　者

2008 年 3 月于羊城

目 录

第 1 章　医学数据挖掘概述

近年来，随着电子信息技术的迅速发展，医院信息系统（Hospital Information System，HIS）、数字医疗设备和医药企事业单位信息系统的广泛应用，各医疗卫生单位计算机中的数据容量不断膨胀。数据库技术的发展在不断地解决海量数据的存储和数据检索的效率问题，但无法改变“数据爆炸但知识贫乏”的现象。如何充分利用这些宝贵的医学数据资源来为疾病的诊断和治疗提供科学的决策，促进医学研究，已成为人们关注的焦点。数据挖掘（Data Mining，DM）是一个近些年才发展起来的信息处理技术，它是从大量数据中提取出可信的、新颖的、有效的并最终能被人理解的信息模式处理过程，它涉及数据库、人工智能、统计学、模式识别、可视化技术、并行计算等众多领域知识。医学数据挖掘是一门涉及面广、技术难度大的新兴交叉学科，它需要从事智能信息处理、计算机、应用数学的科研人员与医务工作者通力合作，将数据挖掘技术应用到医学数据库中，用以发现其中的医学诊断规则和模式，从而辅助医生进行疾病诊断，帮助管理者发现并创造新的管理方法和手段。

本章将对医学领域的数据挖掘作简要介绍，读者在学习本章以后能够清楚为什么要做数据挖掘，用什么方法进行数据挖掘。

1.1　数据挖掘概念

本节将对数据挖掘的产生、基本概念、数据挖掘与其相关领域的关系作详细介绍。

1.1.1　数据挖掘的产生

在这被称之为信息爆炸的时代，信息成几何倍数增长，比如，《纽约时报》由 20 世纪 60 年代的 10 ~ 20 版扩张至现在的 100 ~ 200 版，最高曾达 1 572 版；《北京青年报》是 16 ~ 40 版；中医药是中华民族的瑰宝，几千年来积累了数十万首中药处方，建立了众多中医药处方数据库。大量数据在给人们带来方便的同时引出了一系列问题：第一是数据过量，难以消化；第二是数据真假难以辨识；第三是数据安全难以保证；第四是数据形式不一致，难以统一处理。人们开始提出一个新的口号：“要学会抛弃数据。”人们开始考虑“如何才能不被数据淹没，而是从中及时发现有用的信息和知识，提高数据和信息的利用率”，面对这一挑战，数据挖掘和知识发现（DMKD）技术应运而生，并显示出其强大的生命力。

另一方面，随着数据库技术的迅速发展以及数据库管理系统的广泛应用，积累的数据越来越多。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但不能主动地发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象，如何使人们能够快速有效地获取

自己所需的知识，成为广大信息工作者的重要研究课题。正是这种需求催生了一门目前在信息领域里最为活跃、最令人激动的领域——数据挖掘和知识发现。

1989 年 8 月，在美国底特律召开的第 11 届国际联合人工智能学术会议（International Joint Conference on Artificial Intelligence, IJCAI）上，有关专家专门组织了关于知识发现（Knowledge Discovery in Databases, KDD）的专题讨论会，首次提出了从数据库中发现知识的概念。随后引起国际人工智能和数据库等领域专家的广泛关注，1991 年麻省理工学院（Massachusetts Institute of Technology，MIT）出版社出版了《Knowledge Discovery in Databases》一书。1995 年，在加拿大蒙特利尔召开了首届数据挖掘与知识发现国际学术会议（KDD'95），这次会议是数据挖掘与知识发现领域中一次具有里程碑意义的会议。此后，知识发现与数据挖掘国际学术会议每年召开一次。1996 年 MIT 出版社又出版了《Advance in Knowledge Discovery and Data Mining》一书。1997 年，《Knowledge Discovery and Data Mining》杂志创刊。

经过十多年的努力，数据挖掘技术的研究已经取得了丰硕的成果，发表了大量的研究论文，并且出现了许多成功的应用案例。不少软件公司已研制出数据挖掘软件产品，并在北美、欧洲等国家得到应用。如 IBM 公司开发的 QUEST 和 Intelligent Miner，以及本书将重点介绍的 Microsoft 公司开发的 SQL Server Business Intelligence Development Studio 等。

随着国外知识发现的兴起，我国也很快跟上了国际步伐。自 1997 年在新加坡召开了第一次亚太知识发现与数据挖掘会议（Pacific-Asia Conference on Knowledge Discovery and Data Mining，PAKDD），此后每年召开一次，1999 年 4 月在北京召开了第三届亚太知识发现与数据挖掘会议（PAKDD'99）。国内的许多科研单位和高等院校竞相开展知识发现与数据挖掘的基础理论及其应用研究。我国各大科研资助项目（如“国家自然科学基金”、“973”、“863”及“攻关”等）都设立了 KDD 的研究课题。一些企业也开展了此类项目的研究和开发。

近年来，数据挖掘技术在医学领域中的应用越来越广泛。在疾病诊断、治疗、器官移植、基因研究、图像分析、康复、药物开发、科学研究等方面都获得了可喜的成果。如南加州大学脊椎病医院利用 Information Discovery 进行数据挖掘，该技术已应用到肿瘤学、肝脏病理学、肝炎的生存几率预测、泌尿学、甲状腺病例诊断、风湿病学、皮肤病诊断、心脏病学、神经心理学、妇科学、产科学等医学领域。数据挖掘在医学上的应用有其自身的优势，因为医学上收集到的数据大多是实际诊断和运作数据真实可靠、不受其他因素影响的，而且数据集的稳定性较强。这些对挖掘结果的维护、不断提高挖掘模式的质量都是非常有利的条件。随着电子病历的推广，用计算机存储病案在医院已经比较普遍。如果各医院将收集的数据进一步汇总，数据总量是相当大的，而且都是病人的真实数据。从这样的数据集中运用各种数据挖掘技术了解各种疾病之间的相互关系、各种疾病的发展规律，总结各种治疗方案的治疗效果，以及对疾病的诊断、治疗和医学研究都是非常有价值的。

1.1.2 数据挖掘的定义

数据挖掘是商务智能（Business Intelligence, BI）中的关键成员之一，BI 中的关键成员还包括联机分析处理（Online Analytical Processing, OLAP）、企业报表和 ETL（Extract-Transform-Load 的缩写，即数据抽取、转换、装载的过程）。数据挖掘的定义比较多，要理解数据挖掘的定义，

本节从下面几个方面给予介绍。

1．商业企业角度的定义

数据挖掘是一种新的商业企业数据处理技术，其主要特点是对商业企业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业企业决策的关键性信息。

简而言之，数据挖掘其实是一类深层次的数据分析方法。数据分析本身已经有很多年的历史，只不过在过去数据收集和分析的目的是用于科学研究，另外，由于当时计算能力的限制，对大数据量进行分析的复杂数据分析方法受到很大限制。现在，由于各行业业务自动化的实现，商业企业领域产生了大量的业务数据，这些数据不再是为了分析的目的而收集的，而是由于纯机会的（opportunistic）商业企业运作而产生的。分析这些数据也不再是单纯为了研究的需要，更主要是为企业决策提供真正有价值的信息，进而获取利润。但所有企业面临的一个共同问题是：企业数据量非常大，而其中真正有价值的信息却很少，因此从大量的数据中经过深层分析，获得有利于企业运作、提高竞争力的信息，就像从矿石中淘金一样，数据挖掘也因此而得名。

因此，商业企业角度的数据挖掘可以描述为：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的、先进有效的方法。

2．技术角度的定义

数据挖掘（data mining）是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

与数据挖掘相近的同义词有数据融合、数据分析和决策支持等。这个定义包括几层含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用；并不要求发现放之四海皆准的知识，仅支持特定的发现问题。

何为知识？从广义上理解，数据、信息也是知识的表现形式，但是人们更把概念、规则、模式、规律和约束等看作知识。人们把数据看作是形成知识的源泉，好像从矿石中采矿或淘金一样。原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形和图像数据；甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现的知识可以被用于信息管理、查询优化、决策支持和过程控制等，还可以用于数据自身的维护。因此，数据挖掘是一门交叉学科，它把人们对数据的应用从低层次的简单查询，提升到从数据中挖掘知识，提供决策支持。在这种需求牵引下，汇聚了不同领域的研究者，尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和工程技术人员，他们投身到数据挖掘这一新兴的研究领域，形成新的技术热点。

这里所说的知识发现，不是要求发现放之四海而皆准的真理，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明。实际上，所有发现的知识都是相对的，是有特定前提和约束条件，面向特定领域的，同时还要能够易于被用户理解，最好能用自然语言表达所发现的结果。

3．数据挖掘与传统分析方法的区别

数据挖掘与传统数据分析（如查询、报表、联机应用分析）的本质区别是数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识。数据挖掘所得到的信息应具有预先未知、有效和可实

用三个特征。

预先未知的信息是指该信息是预先未曾预料到的，即数据挖掘是要发现那些不能靠直觉发现的信息或知识，甚至是违背直觉的信息或知识，挖掘出的信息越是出乎意料，就可能越有价值。在商业应用中最典型的例子就是一家连锁店通过数据挖掘发现了小孩尿布和啤酒之间有着惊人的联系。

4．数据挖掘和数据仓库

大部分情况下，数据挖掘都要先把数据从数据仓库中导入到数据挖掘库或数据集中（见图 1-1）。从数据仓库中直接得到进行数据挖掘的数据有许多好处。比如，数据仓库的数据清理和数据挖掘的数据清理差不多，如果数据在导入数据仓库时已经清理过，那很可能在做数据挖掘时就没必要再清理一次，而且所有的数据不一致的问题都已经被解决了。

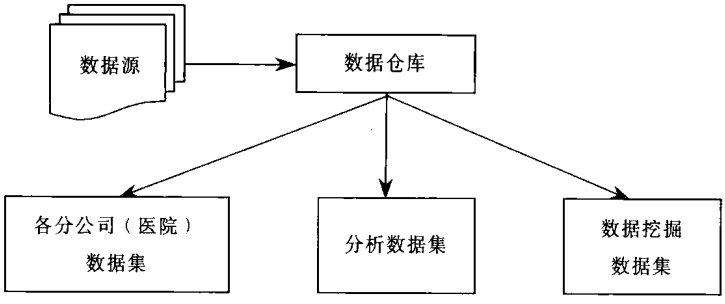

图 1-1　数据挖掘库从数据仓库中得出

数据挖掘库可能是数据仓库的一个逻辑上的子集，而不一定非得是物理上单独的数据库。但如果数据仓库的计算资源已经很紧张，那最好还是建立一个单独的数据挖掘库。

当然为了数据挖掘也不必非得建立一个数据仓库，数据仓库不是必需的。建立一个巨大的数据仓库，把各个不同源的数据统一在一起，解决所有的数据冲突问题，然后把所有的数据导入到一个数据仓库内，是一项巨大的工程，可能要用几年的时间，花费上百万的资金才能完成。只是为了数据挖掘，你可以把一个或几个事务数据库导入到一个只读的数据库中，就把它当作数据集，然后在它上面进行数据挖掘，如图 1-2 所示。基于这样的原因，本书将不对数据仓库作专门介绍。

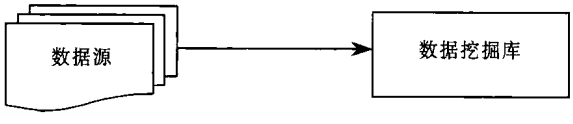

图 1-2　数据挖掘库从事务数据库中得出

5．数据挖掘和 OLAP

经常遇到一个问题：数据挖掘和联机分析处理（On-Line Analytical Processing，OLAP）到底有何不同？总体来说，两者是完全不同的工具，基于的技术也大相径庭。