龚著琳
陈　瑛
章　鲁
顾顺德
等编著

生物医学数据挖掘

（第二版）

上海科学技术出版社


生物医学
数据挖掘

生物医学数据挖掘

（第二版）

龚著琳　陈瑛　章鲁　顾顺德　等编著

上海科学技术出版社

图书在版编目(CIP)数据

生物医学数据挖掘/龚著琳等编著. —2版. —上海：上海科学技术出版社，2011.8
ISBN 978—7—5478—0918—1

Ⅰ.①生… Ⅱ.①龚… Ⅲ.①生物工程：医学工程—数据采集 Ⅳ.①R318—39

中国版本图书馆CIP数据核字(2011)第131662号

上海世纪出版股份有限公司
上海科学技术出版社 出版、发行
(上海钦州南路71号 邮政编码200235)
新华书店上海发行所经销
苏州望电印刷有限公司印刷
开本 787×1092 1/16 印张 8
字数：160千字
2008年2月第1版
2011年8月第2版 2011年8月第2次印刷
ISBN 978—7—5478—0918—1/R·304
定价：24.00元

前　言

当前，人类社会面临的是一个信息爆炸的社会。信息是事物运动状态和特征的反映，它和材料及能量一起构成社会的三个要素。但是，信息具有一些不同于材料和能量的特征。例如，信息具有普遍性，即任何事物都具有信息；信息具有无损性，即它会随着事物的发展而不断产生和增长，不会被消耗，而且可以被复制；信息具有时空独立性，即可脱离其载体而存在或传播；等等。正是由于信息具有这些特征，因此，它和人类文明及社会发展的各个阶段都有密切的联系。尤其是近50年来发生的信息技术（即获取信息、传输信息、处理信息和应用信息的技术）革命，更进一步促成和显现信息对各学科、各行业乃至整个社会发展的巨大影响。

和其他学科一样，生物医学受到信息科学及技术的影响，进而相互渗透、融合。由此产生的对生物和医学学科发展的促进作用也是显而易见的。

例如，X射线早在1895年已被发现，并被应用于医学；核磁共振现象在1946年已被发现，发现者因此而获得了1952年度诺贝尔物理学奖。然而，只有在计算机技术快速发展、图像重建及处理方法因而得以实现的基础上，G.N. Hounsfield和A.M. Cormack才发明了X射线计算机断层扫描成像术（X-CT）并应用于医学，他们因此而获得了1979年度诺贝尔生理学或医学奖；P.C. Lauterbur和P. Mansfield也才发明了核磁共振成像术（MRI）并应用于医学，他们因此而获得了2003年度诺贝尔生理学或医学奖。现在，X-CT和MRI已成为医学临床及基础研究的常规检测手段，研究者能借此获取用传统方法难以获得的被检者的解剖及生理或病理信息。

又如，随着计算机网络的迅速发展，特别是因特网（Internet）及其下一代计算机网格技术的广泛应用，生物医学信息的共享突破了时空的局限。共享资源的形式也不仅限于专业文献资料，已拓展到开放式的实验及临床数据库、经验知识甚至实验设施等。在网格基础上建立的虚拟实验（Virtual Lab）环境是一种全新的工作模式，它为集聚各种资源、极大地提高医学基础科研及临床试验的效率和质量提供了条件。

再如，计算机性能的指数式增长（摩尔定律），使研究人员有可能利用计算机实施各种数学方法来分析复杂的生物系统和生命现象，以数学方式描述生命现象的过程和规律并预测其发展趋势（生物医学建模），或在看似互不相干的众多数据中挖掘和发现事先未知的规则和联系（生物医学数据挖掘），使“数据→信息→知识”的认知链更臻完善。在这一方面最成功的实例之一就是用信息技术来处理生物学数据和理解生物系统，并在此基础上形成新的交叉学科：生物信息学（Bioinformatics）和计算生物学（Computational Biology），为生物学及相关学科的发展开辟了一个全新的领域，等等。

信息技术不仅仅是“工具”和“方法”，也会促使人们转变其某些固有的思维模式和行为方式。信息技术的发展不仅促使知识的快速增长（知识“爆炸”），也促成了知识的高度集成。新知识的形成和新成果的产生往往是吸取各学科知识、对原有知识系统进行重组和整合的结果。因此，面对信息社会之际，包括生物医学在内的各种专业技术人员迫切需要充实自己原有的知识结构成为顺理成章的结果。

正是基于这样的背景，我们在向医学院校的本科生、研究生开设相关课程长达10余年的基础上，编写了一系列有关信息技术在生物医学领域中应用的参考书籍。编写这套参考书籍的主要目的在于通过介绍信息技术的基本知识及其在生物医学中的应用和实例，使读者能够做到以下几点：

（1）对信息技术和理论及相关的数学工具在医学科学研究和临床实践中的应用保持高度的敏感。

（2）在理论的指导下，理性而非盲目地运用现有信息技术工具，解决一些生物医学领域的问题。

（3）建立和工程技术人员的共同语言，为开展多学科合作，进一步拓展信息技术在生物医学领域的应用奠定基础。

正是基于这一目的和定位，在编写原则上，这套参考书既维持了相关信息技术本身应有的系统性和理论性，更着重体现其在医学学科中应用的实用性和针对性。在内容的取舍上，既选取了具有代表性的经典内容，也结合了作者多年来在此领域的一些研究工作和教案。在叙述方法上，力求简明扼要，着重于应用的意识和方法，淡化或省略了对应用方法直接影响较少的数学推导和论证过程。

这套参考书籍适合于作为医学院校本科生及研究生的选修课教材，也可以作为医学基础及临床科学工作者继续教育的教材及参考资料。由于信息技术所涉及的理论和知识非常广泛，并且又经历着日新月异的发展，这套书籍中许多章节所涉及的内容都足可自成一书也不为过。因此，这套书籍的叙述方式只能是提纲挈领式的，供读者在实际应用中作为参考，并为进一步的深入学习和研究奠定基础。

上海交通大学医学院的王成、刘雅琴、朱浩栋、邵建兵、张方、张剑戈、徐立钧、苏懿、陈瑛、顾顺德、黄昕、崔茂龙、龚著琳和章鲁等教师参与了本套丛书的编写。在编写及酝酿过程中，还得到了原上海第二医科大学金正均教授的指导和帮助，采用了组织胚胎学教研室徐晨、冯京生以及解剖教研室黄耀德等教师提供的部分图片，采用了王军、司京玉、安建福、岑康、余晨光、张骊峰、张毓敏、周妮、聂生东和黄永锋等研究生的部分研究内容作为应用实例，采用了上海瑞金医院陈克敏和柴维敏等医生、上海仁济医院王家东、华佳和柴伟民等医生、上海肿瘤医院顾雅佳和柳光宇等医生、上海新华医院潘曙明和汤璐佳等医生、上海交通大学公共卫生学院程琦教授、上海中医药大学余安胜等老师以及上海市疾病预防控制中心郑莹等医生提供的部分资料作为应用实例，并且在和清华大学李三立院士、中国电子学会生物医学电子学分会王保华、庄天戈、方祖祥、陈俊强、陈明进、罗立明、陆祖宏、严壮志等教授、上海交通大学 Med-X 研究所徐学敏、赵俊、刘萍和孙建奇等教授以及上海大学徐纬民等教授的学术讨论中得到许多有益的启发。对此，一并表示感谢。限于作者的水平，本丛书的不足之处在所难免，敬请读者不吝指正。

编　者

2011.5

目 录

第一章 概 论

1.1 什么是数据挖掘

1.1.1 数据、信息和知识

数据（data）是对客观事物特征状态的记录。例如，商店内某种商品的销售量或销售额、医院内某类药物的使用数量、某临床科室的床位周转率、患者的心率和血压等生理参数……这些都是数据。客观事物某些特征状态的记录还受制于技术。例如，在 X 线及其医学应用被认识以前，人体内各种组织对 X 线的不同衰减特征就无法记录。因此，随着人类生产能力和科学技术的进步以及人类社会活动的发展，数据的种类、形式和数量日益增多。同时，以计算机和网络为代表的信息技术的发展，使数据的采集、存储、管理和重用更为简便和规范（例如，通过数据库管理系统以一定的格式或结构来存储和组织数据），使数据的流通和共享性增大（例如，通过因特网共享数据）。在这样的背景下，“人类被淹没在日益增长的数据之中”正成为当前社会的特征之一。

另一方面，数据是信息（information）及知识（knowledge）的载体。信息和知识才是真正有意义的。然而，相对于数据的急速增长，人类从数据中提取有用的信息，并将这些信息归纳上升为知识的能力却极大地滞后（见图 1.1）。

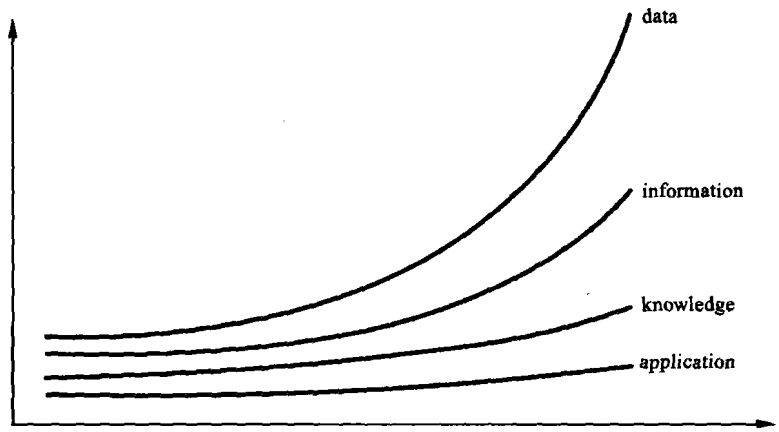

图 1.1 数据、信息和知识的不同增长率

因此，如何从大量数据中发现和找出以隐含方式存在于其中的、有意义的信息和知识是一个迫切要解决的问题。

1.1.2 数据挖掘的定义

数据挖掘（data mining，DM）是对大量观察到的数据进行分析，以便从中发现事先未知的联系和规律的过程。其目的是让数据拥有者得到非常清晰而有用的结果（即信息和知识）。

通常，被分析的数据是以一定的格式存储在数据库中（database，DB）并加以组织的。因此，在不十分严格的意义上，有时又可将数据挖掘等同于“数据库中的知识发现”（knowledge discovery in database，KDD）。而严格地说，数据挖掘只是 KDD 过程中的一个环节。从数据挖掘的定义可看到：其目的和人类的学习相类似，而其过程通常是在计算机上实施的。因此，有时也将数据挖掘看作是一种“机器学习”（machine learning）的技术。“机器学习”的许多方法和技术（如分类、人工神经网络、聚类分析、关联规则等）也为数据挖掘所采用。

统计分析（如线性回归、logistic 回归等）是数据挖掘可采用的方法之一，但通常的统计处理和数据挖掘是有区别的。一个主要差别在于数据集的尺度：数据挖掘经常会面对 GB 甚至 TB 数量级的数据库（如人类基因工程要完成整个人体基因的测序可能产生超过 3.3×10^9 个核苷酸的数据集），而用传统的统计方法很难处理这么大尺度的数据集。另一个差别在于：传统的统计处理往往是本位分析（primary analysis），即针对特定的问题采集数据（甚至通过试验设计加以优化），分析数据来回答特定问题；而数据挖掘却往往是数据分析的次级过程，其所用的数据原本可能是为其他目的采集的，因而可能不一定适用于原来目的以外的数据分析。基于上述原因，数据挖掘往往还需要应用传统统计学以外的其他方法和技术。

虽然数据挖掘的对象往往是被存储于数据库中的，数据库也是采集和组织数据的有效手段，但数据挖掘和数据库处理还是有区别的。在数据库处理中，常采用结构化查询语言（structure query language，SQL），其处理的结果总是原数据库中数据的一个子集或数据的简单汇总，而数据挖掘并不能简单地采用 SQL 来完成任务，其输出也并非原数据库中数据的一个子集或汇总，而是这些数据之间的某种联系或蕴涵的某种规律。

例如，有一个乳腺癌患者资料的数据库，包含各个患者的姓名、住院号、年龄、性别、婚姻状况、乳腺癌临床分级（非浸润性、0、Ⅰ、Ⅱ、Ⅲ_a、Ⅲ_b 和Ⅳ）、乳腺癌组织学分级（c_1、c_2、c_3 和 c_4）、治疗方案（手术、放疗、化疗、内分泌治疗和联合治疗）和 5 年生存状态（生存和死亡）等数据。如果以通常的 SQL 处理该数据库，可以得到满足某种条件的输出（如所有 40 岁以上乳腺癌患者的记录、曾接受手术治疗的乳腺癌患者记录、所有接受放射治疗且生存期超过 5 年的乳腺癌患者记录等），这些输出都是原数据库中全体数据的一个子集。如果对这个数据库的数据进行数据挖掘，则很可能得到这样的结论：“大部分患非浸润性、高分化原位乳腺癌的患者在接受手术治疗后，其生存期一般都超过五年”。这样的输出结果并非是原数据库中数据的一个子集，而是蕴涵在原数据库数

据中的一种联系。医生在为特定的患者选择治疗方案及评估疗效时可参考这种联系的涵义。

1.2 数据挖掘的应用及方法

1.2.1 应用

以下通过若干个具有一定代表性（但并非全面）的应用实例来说明数据挖掘的一些基本功能，若将一些基本功能组合起来，还可完成更为复杂的数据挖掘应用。

【例 1.1】某医疗机构数据库内有一批与乳腺疾病相关的记录，每个记录的相关数据包括：乳腺钼靶 X 线影像（包括钙化点、乳腺结构扭曲和团块组织等影像学特征）、病历记录（包括既往乳腺病史和体检发现等）和流行病学调查报告（包括乳腺癌家族史、生育史和哺乳史等），每个记录都有已经证实的明确诊断结论（正常、良性乳腺疾病或乳腺癌）。现有一个受检者记录（具有和上述历史记录相同结构的相关数据），要求根据历史记录的相关数据和受检者的数据之间的关系，判断该受检者可能的诊断结论。

这是一个分类（classification）的应用实例。所谓分类是指：基于已知所属类别的历史记录的特征数据描述预先定义好的类别（在该例中共有“正常”、“良性乳腺疾病”和“乳腺癌”3 类），再根据待查记录相关特征与这些类别相应特征之间的相似程度，确定该待查记录应划归入哪个类别。

【例 1.2】利用美国国家癌症研究所（National Cancer Institute，NCI）提供的流行病学调查（the Surveillance, Epidemiology, and End Results，SEER）报告中的癌症发病率公用数据库（Cancer Incidence Public-Use Database，CIPUD），从 433272 例癌症患者病例记录选取子宫颈癌病例（每个病例具有 72 项数据），研究预测子宫颈癌患者存活率及其受各危险因子影响的模型。

这是一个回归（regression）的应用实例。所谓回归是指：确定响应变量和一个或多个自变量之间的依赖关系以构建预测模型。回归可以是一元回归（即只有一个自变量），也可以是多元回归（即可有多个自变量）。一般可假设以线性函数、logistic 函数等一些已知类型的函数（不一定是显性的解析函数）来拟合已有数据，再利用误差分析方法分析各种拟合误差，最终确定拟合程度最好（即拟合误差最小）的一个函数。一元线性回归是最简单的回归方法。

【例 1.3】胰岛素依赖型糖尿病（Ⅰ型糖尿病）患者需要定期注射胰岛素来控制血糖水平。血糖水平过高或过低都会危害患者的健康，因此必须严格控制胰岛素剂量，并根据影响因素的改变而作适时调节。为此，在一段时间内，每天间隔一定时间采集患者的相关数据，包括血糖水平、胰岛素剂量、饮食摄入、体力运动和可能影响葡萄糖代谢的事件（如发热）等。分析这些数据，目的在于了解该患者的血糖水平在一天内的变化周期、变

化趋势以及各种因素影响血糖水平变化的方式和程度，以便指导该患者的胰岛素用药方式和剂量。

这是一个时间序列分析（time series analysis）的应用实例。时间序列数据是指一个或多个数据属性和时间相关（随时间而改变）。分析时间序列的结果往往是得到预测性的数值输出。利用时间序列分析可以解决：比较或确定不同时间序列数据的相似性；确定时间序列数据的变化方式或规律；利用历史时间序列数据分析结果预测该序列数据在将来的某个时刻的值。显然，当回归分析中至少有一个自变量与时间相关时，这种回归分析也成了一种时间序列分析方法。

【例 1.4】肾透析是治疗肾功能衰竭患者经常采用的措施之一，如何提高接受肾透析治疗患者的生存率是一个有待解决的问题。医生根据患者病情开出的肾透析处方包括透析次数、透析间隔时间、透析液的组成成分和为控制发生诸如高血压之类副作用的附加药物等要素。在进行肾透析时，除了要考虑如年龄、病情等患者数据外，还需要监测水和电解质平衡等 50 多项参数。所有这些因素以及它们之间的相互作用都会影响肾透析的疗效和患者的生存。根据已有的肾透析患者治疗的所有相关数据，找出各因素影响疗效和生存率的规律。按此规律预测对特定患者的疗效，从而可针对性地制订治疗方案。

这是一个预测（prediction）的应用实例。所谓预测是指：基于对历史数据和当前数据的分析，判断未来数据的可能状态。预测也可看作是一种分类，两者不同之处在于预测是判断未来数据（当前尚未发生）的状态，而分类是判断已发生的数据的状态。上述的回归分析、时间序列分析和其他一些方法都可用来进行预测。

以上 4 个应用实例都属于预测型（predictive）数据挖掘，即利用从历史数据中发现的已知结果，预测数据的值。这些历史数据和被预测数据可以属于同一个个体，也可来自于不同个体。

【例 1.5】为了研究一些流行病学因素对肺癌患者临床医学状况的影响，从 SEER 的数据库中选取 217558 例肺癌病例，每个病例数据包含 23 个流行病学特征属性（如年龄、性别、吸烟史和家族肺癌病史等）和 22 个临床医学状态特征属性（如肺癌的种类、病程等）。根据流行病学特征属性的相似程度将所有病例数据划分成 20 类（属于同类的数据相似程度高，而不同类别之间的数据相似程度低），比较两两类别之间的临床医学状态特征属性。有 2 对类别之间只有 1 个临床医学特征属性有明显差异；有 4 对类别之间只有 2 个临床医学特征属性有明显差异；其他各对类别之间至少有 3 个临床医学特征属性有明显差异。在此基础上可进一步分析各类流行病学因素对肺癌患者临床医学状况的不同影响。

这是一个聚类（clustering）的应用实例。所谓聚类是指：根据原始数据之间的相似性将这些数据划分成若干类别，其结果是属于同类的数据相似程度远大于不同类别之间的数据相似程度。既可以预先确定划分成多少类，也可以事先并不确定划分类别的数量。聚类和分类很相似，不同的是在完成聚类之前对所有数据的类别归属并不确定，而在分类过程中是根据已知类别归属的数据特征来判断未知数据归属于何种类别。

【例 1.6】病人住院期间感染是影响患者健康的重要因素之一。目前，由耐药性引起的感染数量在不断增长。例如，对于万古霉素（Vancomycin）有耐药性的肠道球菌会产生格兰氏阴性杆菌，就是一种可引起高发病率和高死亡率院内感染的常见原因。大规模院内感染的爆发往往和细菌出现耐药性有关。细菌的耐药性总是起源于一个小环境（特别是医院的监护病房），当条件成熟后即传播到一个更大的环境。因此，监测耐药性状态和早期发现其异常是有效控制院内感染的必要措施之一。某医院为了设计一个控制院内感染的监测系统，采集了历史上 15 个月内所有住院病人的细菌培养数据及相关信息，每种分离细菌的相关信息包括：细菌名称、格兰氏染色、形态学特征、标本采集日期、来源（如血液、痰或尿等）、病人住院地点（外科监护病房还是内科监护病房等）以及耐药性检验结果等。分析这些数据，发现细菌耐药性产生和院内感染发生相联系的规律，从而为监测细菌耐药性，进而控制院内感染提供依据。

这是一个关联规则（association rule）分析的应用实例。关联规则分析简称关联分析，又称亲和力（affinity）分析，它是描述数据之间隐含的特定关系（即事件一起发生）的可能性。这些事件一起发生的原因在于其内在的关联，而并不一定是因果关系。

【例 1.7】有机体的遗传和功能信息存储于 DNA、RNA 和蛋白质中，而这些大分子又分别是由确定的组织成分构成的线性结构，因此可用符号序列来表示。果蝇中有一个基因 eyeless，若将其敲掉，则导致果蝇失去眼睛。人类有一个基因 aniridia，若这个基因丧失或发生突变，则会导致人眼没有虹膜。实验发现：将基因 aniridia 插入因被敲掉基因 eyeless 而没有眼睛的果蝇中，果蝇可以形成正常的眼睛。基因 aniridia 和基因 eyeless 是否具有相似的功能？为了研究这两个基因作用原理中的共性，需要比对构成它们的核酸或氨基酸的序列。

这是一个序列发现（sequence discovery）或序列分析（sequence analysis）的应用实例。序列是指按一定顺序或规律排列构成的一系列符号、数值或事件。所谓序列发现是指：分析和发现序列构成的规律或模式（pattern）、两个序列组成部分的相同性（identity）或相似性（similarity）、或者一个序列组成部分的片段（segment）之间的相同性或相似性。

以上 3 个应用实例都是属于描述型（descriptive）的数据挖掘，即识别数据中的模式（pattern）或关系。与预测型数据挖掘不同，描述型数据挖掘旨在探索被分析数据的内在性质，而不是预测新的性质。

1.2.2 方法

数据挖掘是一种机器学习的技术，它的目的及过程和人类的学习有类似之处。从学习对象和过程的角度来看，数据挖掘方法可分为“有指导（监督）学习”（supervised learning）和“无指导（监督）学习”（unsupervised learning）两类。不论哪种类型的学习，都需要有学习样本集（training data set），通过对这些样本的“学习”，发现隐含于

其中的规律。学习样本也称训练样本。所谓有指导学习是指学习样本的归属都是已知的、确定的。

例如，在“分类”中，每个学习样本的类别归属都是事先确切已知的；而所谓无指导学习是指学习样本的归属事先并不确定或已知。又如，在“聚类”中，学习样本的类别归属事先并不明确，甚至有几个类别也可能是不确定的。只有在聚类分析结束后，原来隐含在学习样本集合中的分类信息（共有几个类别以及每个样本分别归属于哪个类别）才能得以明确地显现和表示。

从学习方法的角度来看，有多种数学工具可用。例如，回归分析学统计学方法、人工神经网络（artificial neural network）、决策树（decision tree）等。一些数学工具可以适用于不同类型的应用。例如，人工神经网络方法既能用于分类，也能用于回归分析。另一方面，要解决某一种类型的应用，往往也有多种数学工具可供选择。例如，统计方法、决策树、人工神经网络和其他一些方法都可用于分类。究竟哪个方法能最好地解决问题，这不仅取决于应用的类型和要求，还常常受到被分析数据本身的特点影响。因此，根据待解决问题的类型、要求，以及数据本身的特点，确定最佳的方法，是数据挖掘研究的重要内容，也是影响数据挖掘结果的关键之一。

1.3 生物医学数据挖掘的特殊性

医学的研究和处理对象是人类，具有许多特殊性。例如：

- 人类是地球上最自我关注的物种，具有许多可供观察的表现（如视觉、听觉、痛觉、幻觉、不舒服的感觉、对既往相关经历的回忆等），是其他学科实验所不完全具备的。
- 可以跟踪观察医学上感兴趣的疾病（如癌前病变和动脉粥样硬化等）的长期进程，这不可能通过大多数是短期的动物实验来完成。
- 人类的医学资料（病史记录等）数量巨大、个体差异显著，这是其他学科实验数据无可比拟的。
- 采集人类医学数据又要受到不同于其他学科的伦理、法律和社会等因素的制约等。

因此，在进行生物医学数据挖掘时会发生一些在其他学科领域内未必会遇到的特殊问题。重视这些问题，对患者、医学研究人员和从事数据挖掘的研究人员都是有益的。

生物医学数据挖掘的特殊性主要表现在以下几个方面。

1.3.1 医学数据的特殊性

从数据挖掘的角度看，生物医学数据具有以下一些主要的特殊性：

- 生物医学领域中原始数据的数量大，且呈多样性或异质性（heterogeneity）。医学数据的形式可以是二维信号（如超声、X-CT、MRI 和 PET 等）、一维信号（如 ECG 等）及其参数、临床化验或生理参数（一般为数值，如血液中高、低密度脂蛋白及甘油三

酯含量、血压、心率等）、医生的问诊、观察和解释（一般为非结构化的描述语言）等。这些数据都和诊断、预后及治疗有关，在数据挖掘中都应予以考虑。

正是由于原始医学数据的数量大，且呈多样性，一般不会直接对所有原始数据进行数据挖掘，而是从原始数据中先抽取一个有代表性的数据子集（subset）进行挖掘。通常有两种方法可降低数据集的维度（dimension），其一是选取部分病例（一般是随机选取）；其二是针对性地选取部分特征。

● 除了数据形式各异、数据量极大之外，与物理学的许多领域相比，医学数据的另一个特点在于有时很难以数学方式来表达其结构及特征。物理学家采集数据，并将这些数据输入能合理反映数据间关系的公式和模型。而医学概念有时由文字描述构成，并且对于用词和基本概念之间的关系缺乏规范和限制。炎症、缺血和肿瘤等医学基本事件，对医生而言是实在的，就如质量、长度、力等概念对物理学家是实在的一样。但是，数据挖掘研究者一般难以将这些医学信息加以结构化，来进行聚类、回归或序列分析等处理。

医生对医学影像、信号和其他临床数据的解释含有丰富的知识和经验，是生物医学数据挖掘必须依赖的数据形式之一。但这些解释大多是以非结构化的语言或文字形式自由口述或书写的，很难标准化（因此，也无法直接进行数据挖掘处理）。即使是同一专业的专科医生也很难一致地采用明确的词语描述某一特定病情（不仅使用的词句或术语不同，甚至用不同的语法结构描述），这使问题更加复杂化。例如，有学者研究了不同的放射科医生观察和描述乳腺钼靶 X 线影像中钙化点时所采用的术语，11 位资深放射科医生观察并描述了相同的 20 幅乳腺钼靶 X 线影像中的钙化点（其中 4 幅影像中无钙化点，8 幅影像中有钙化点、且患者为良性肿瘤，另 8 幅影像中有钙化点、且患者为恶性肿瘤）。他们在描述钙化点亮斑的形状、边界、大小、亮度、指向和分布等 12 个方面的特征时，所采用的“术语”竟多达 159 个。

● 随着病程改变，医学数据经常需要更新（例如，需要对随访患者复查一次血液化验或 X-CT 检查等）。因此，需要建立一种方法，以此相应地更新或修正根据既往数据挖掘而得到的知识。

● 采集医学数据很难完全避免噪声（noise）干扰。因此，应选择对噪声敏感性低的方法进行数据挖掘，并要确保以后采集数据时的噪声不高于当前水平，以免影响对未来数据的分析结果。

● 生物医学数据难免会发生丢失（可能是由于疏忽而缺失，也可能是出于技术、经济或伦理等原因故意遗漏）。用于数据挖掘的医学数据往往是医学行为的副产品，而并非为进行研究而专门采集的，因此，许多病例记录都是不完整的（缺少某些特征数据）。为此，应采取相应的方法弥补或代替缺失数据后才能进行数据挖掘。

● 医学数据可能包含冗余的、没有意义的或不一致的属性（当同一个数据被归属为一个以上互相排斥的类别时，即发生数据的不一致性）。例如，一个患者的血钾浓度异常高，与其实际健康状况不相符合。对这种矛盾现象的一个可能的解释是：血样标本在送检

过程中过度摇晃导致血球中的钾离子进入血清。但是，在未经实验验证（这在医学数据挖掘过程中往往是无法进行的）之前，无法断定这样的解释是否符合实际。

● 如何将生物医学学科领域的知识和经验加入数据挖掘的机制是一个值得关注的问题。同样，生物医学领域的研究人员希望数据挖掘的过程和结果的形式有助于表达和理解医学概念，而不太愿意接受“黑箱”（black box）形式（这种形式在工程领域中经常能有效地应用）。

1.3.2 伦理、法律和社会等方面对私密敏感的问题

在医学数据挖掘中，数据的所有权是尚未解决的问题。在法律意义上，所有权属于被授权处置（包括出售）某项财产的自然人或法人。生物医学数据（除动物实验数据之外）大多和患者相联系，但这些和患者相关信息的所有权问题是不确定或模糊的，属于患者本人？医生？医疗机构？还是医疗保险提供者？这种模糊性也容易导致患者针对医生和医疗机构的诉讼。在这样一种紧张的环境下，医生和其他医学数据管理者都不愿将医学数据提供于数据挖掘，这也就可以理解了。

医学数据的另一特点是私密性。患者是绝对不希望医生将其身份以及相关的医学信息公诸于众的。如果根据患者的数据发现重要的诊断信息，就会涉及相关的私密问题。一旦泄露私密信息，不仅会使患者产生潜在的不信任感以及由此引发的司法行为，也会损害医患关系。有的国家的法律还规定不得泄露患者的身份。

鉴于上述原因，在提供医学数据进行数据挖掘之前，必须对原始数据作“匿名”处理，隐去患者身份和其他有关信息（究竟应该隐去什么信息，还取决于研究目的和授权）。考虑到传输数据时的安全性问题（通过网络传输数据可能不是十分安全），在隐去患者身份之前，只有得到授权的人员才能访问患者的数据。即使在同一机构的不同部门之间传输数据，也要隐去患者的身份。

另一方面，有时出于研究的目的，需要在已隐去患者身份的数据上重新恢复身份信息。例如，为了防止同一患者的记录重复并由此歪曲研究结论，可能有强制性的需要提及原始（未隐去患者身份）的医疗记录来证实正确性及真实性，或得到特定患者的附加信息等。这些特殊的要求应由合适的管理机构来处理，但如果数据是完全匿名的，则这些要求就难以满足。例如，从因某种疾病死亡的患者尸体上取一块组织作为组织学实验中的对照标本。在采集标本时患者的身份没有被记录，因此也就无法恢复。

所有医学行为的首要目的都只能是维护患者的健康、治疗其疾病，而不能是仅仅为了科研。决定是否采集某项医学数据的唯一准则是：这项措施对该患者是否直接有利。有些患者在充分理解他们并不能从中直接获益的情况下，可能仍然会同意参与某项科研项目。即使如此，这样的数据采集也只能是小规模的，有高度针对性的，且应严格遵守法律和伦理上的制约。