数据挖掘

—原理与算法—

Principle and Algorithm of Data Mining

邵峰晶　于忠清　编著

中国水利水电出版社
www.waterpub.com.cn

数据挖掘原理与算法

邵峰晶　于忠清　编著

中国水利水电出版社

内 容 提 要

本书在介绍了数据挖掘原理的基础上，从实用的角度出发，详细地介绍了数据挖掘的经典算法。本书是国内第一本对数据挖掘技术基础算法进行详细描述的实用性教材。

第 1 章从不同的角度对数据挖掘进行了介绍。第 2 章介绍了数据仓库技术的概念并给出了数据立方体的理论基础。第 3 章讲述了数据挖掘的数据预处理所涉及到的概念及算法。第 4 章～第 8 章详细介绍了数据挖掘的经典领域的算法，其中第 6 章简单介绍了数据可视化的内容。第 9 章介绍了开放的数据挖掘平台。

本书的使用对象是在校高年级的本科生、研究生及各个领域的高级软件开发人员。

图书在版编目（CIP）数据

数据挖掘原理与算法/邵峰晶等编著. －北京：中国水利水电出版社，2003
ISBN 7-5084-1653-8

Ⅰ. 数… Ⅱ. 邵… Ⅲ. ① 数据采集 ② 算法分析 Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2003）第 067964 号

书　　名	**数据挖掘原理与算法**
作　　者	邵峰晶　于忠清　编著
出版、发行	中国水利水电出版社（北京市三里河路 6 号　100044） 网址：www.waterpub.com.cn E-mail：mchannel@public3.bta.net.cn（万水） sale@waterpub.com.cn 电话：（010）63202266（总机）、68331835（营销中心）、82562819（万水）
经　　售	全国各地新华书店和相关出版物销售网点
排　　版	北京万水电子信息有限公司
印　　刷	北京市天竺颖华印刷厂
规　　格	787×1092 毫米　16 开本　20.75 印张　470 千字
版　　次	2003 年 8 月第一版　2003 年 8 月北京第一次印刷
印　　数	0001—5000 册
定　　价	32.00 元

凡购买我社图书，如有缺页、倒页、脱页的，本社营销中心负责调换

前　言

数据挖掘技术是近几年国内外迅速发展起来的一门交叉学科，涉及到数据库、统计学、人工智能与机器学习等多个领域。计算机的应用普及产生了大量的数据，数据挖掘就是利用上述学科的技术进行大数据量的处理。数据挖掘的应用领域非常宽广，从农业生产的预测到基因分类，从化学分子结构的识别到 NBA 教练临场更换队员，从信用卡欺诈到税务稽查，数据挖掘技术对未来社会的各个领域将起到越来越主要的作用。

我国的数据挖掘技术一方面是科研机构停留在学术研究上，另一方面是利用国外公司的软件产品解决具体问题。为了提高学术水平，科研人员只得进行高水平但很难实用的算法研究；为了提高经济效益，销售国外软件公司的产品最稳健。但是，数据挖掘技术在解决实际问题的过程中需要的是成熟技术加针对具体问题的修正，因此，国内迫切需要对国外十余年的数据挖掘具体技术进行剖析，在掌握核心技术的前提下才能真正赶超。本书的背景是在我们三年前开始开发数据仓库产品及对数据挖掘技术进行了将近两年的跟踪的基础上，根据大量参考文献及内部技术报告，结合研究生的教学工作完成的。目前，我们已完成了开放式的数据挖掘平台及部分算法的实现。

本书的使用对象是在校高年级的本科生、研究生及各个领域的高级软件开发人员，书中介绍了大量的数据挖掘算法，各个算法具有很强的实用性。本书是国内第一本对数据挖掘技术基础算法进行详细描述的实用性书籍。

本书共分 9 章。第 1 章对数据挖掘从各个角度进行了剖析，从社会需求开始对数据挖掘的概念、数据挖掘的数据来源、数据挖掘的分类、体系结构、运行过程、数据挖掘与其他领域之间的关系、评价标准及未来的发展方向进行了全面的介绍。

第 2 章对数据挖掘的孪生兄弟——数据仓库技术进行了简单的介绍，由于数据挖掘技术的一个重要发展方向就是嵌入到数据仓库中，即数据挖掘所使用的大数据集直接来自于数据仓库。在简单地回顾了数据仓库技术之后，给出了一种多维数据的模型，这是实施联机分析处理（OLAP）的一种关键技术，同时简单介绍了我们自行开发的 OLAP 展示工具的体系结构，并介绍了数据仓库在银行的应用案例。

第 3 章讲述的是数据挖掘的数据预处理所涉及到的概念及算法。干净而合乎要求的数据是数据挖掘成功应用的基础，对数据进行整理是一项枯燥而艰苦的工作。本章在介绍了数据挖掘的数据准备工作之后，给出了一种常用的数字属性的离散化及属性选择算法。数据挖掘虽然可以解决大数据集的问题，但在分布完全相同的前提下，算法处理十万条记录与百万条记录的时间代价是完全不同的。数据采样技术同样有多种方法，每种方法适合解决的问题是不同的。本章最后一部分介绍了数据抽象问题，即如何将大量的数据进行概念提升。

第 4 章对关联分析给出了详细的算法。无论是在国内还是在国外，关联分析是数据挖

掘发展的先行者，并且几乎与其他学科没有交叉。Apriori 算法是关联分析的基础，多值属性的关联分析所关心的问题是如何将连续数值的关联分析转化为布尔值，多层关联分析与约束性的关联分析都是解决实用问题的算法，本章最后给出了增量的关联分析解决算法。

第 5 章讲述了数据分类，给出了分类的各种基本算法，包括国外数据挖掘最早的 ID3 算法及 C4.5 算法。对来自统计学的 CART 算法给出了详细的描述，同时对如何解决大数据集问题的 SLIQ 算法及并行问题的 SPRINT 分类器也给出了详细的说明。

第 6 章讲述了多维方向与数据可视化。它虽然不是数据挖掘的直接内容，但聚类的多种算法都用到了多维数据访问的技术。而空间数据挖掘的基础则是多维访问。数据可视化技术中对数据的观察进行了阐述。

第 7 章给出了聚类的多种实用算法及基础算法。聚类算法采用了多种技术，用途非常广泛，本章给出了大量的详细的算法。分层的聚类来自于统计学，虽然不能解决大数据量问题，但作为基础还是进行了详细的说明。分区算法介绍了 PAM、CLARA 及 CLARANS 算法，其中对 CLARANS 算法进行扩充，可以用于空间数据挖掘。k-means 算法是最常见也是最实用的算法，特别介绍了处理离散数据的聚类算法 k-modes。OPTICS 是一种复杂的算法，用途也最广泛。BIRCH 的特色是只需访问一次数据库，对该算法给出了详细的描述。最后，对用途广泛的孤立点问题给出了最先进的算法。

第 8 章介绍了序列模式及时间序列。序列模式给出了最早也是最实用的算法。时间序列只是介绍了概貌，没有给出具体的算法，因为时间序列本身就是一门交叉学科。

第 9 章介绍了我们开发的开放式的数据挖掘平台，限于篇幅只是给出了体系结构，对数据挖掘平台中所用的 OLE DB For DataMining 及可预测模型描述语言 PMML 也进行了简单的介绍。

书中的第 1 章、第 2 章和第 7 章由邵峰晶教授编写，其余章节由于忠清研究员编写。在本书的编写过程中得到了南京大学徐洁磐教授、北京大学的邵维忠教授及青岛市副市长马论业教授的多次指导，在此表示感谢。青岛海尔青大海威软件公司的刘志强、林永及贾胜中三位工程师在海威数据仓库与数据挖掘软件及资料方面给予了大力支持，李洁小姐在文字及图形的整理方面做了大量的工作，在此一并表示谢意。

由于时间仓促，书中的错误与不足之处在所难免，敬请读者批评指正。

作者

2003 年 6 月

目　　录

第 1 章　导论

1.1　数据挖掘的社会需求

一切新事物的产生都是由需求驱动的。希望能让计算机自动智能地分析数据库中的大量数据以获取信息，是推动挖掘型工具产生并发展的强大动力。从生产成本的角度来看，公司的人工费用在不断提升，产品与服务的价格持续下降，激烈的市场竞争迫使决策者想办法降低成本并扩大产品与服务的销售量来提高公司的竞争力。从计算机的应用角度来看，无论是硬件与网络在性能方面的提高，还是软件的技术与功能的提高，都要求软件从单纯的管理功能向综合的分析功能转变。从数据管理的角度来看，历史的数据是一笔宝贵的财富，而且这些数据正以几何级数或指数级数增长。从软件技术的发展方面来看，大数据量的分析对原来各个领域的技术带来了极大的挑战，需要采用综合性的技术来迎接这些挑战。

随着数据库技术的飞速发展以及人们获取数据手段的多样化，人类所拥有的数据急剧增加，可是目前用于对这些数据进行分析处理的工具却很少。数据库系统所能做到的只是对数据库中已有的数据进行存取和简单的操作，人们通过这些数据所获得的信息量仅仅是整个数据库所包含的信息量的很少一部分，隐藏在这些数据之后的更重要的信息是关于这些数据的整体特征的描述及对其发展趋势的预测，这些信息在决策制定的过程中具有重要的参考价值。

例如，股票经纪人需要从日积月累的大量股票行情变化的历史记录中发现其变化规律，以供预测未来趋势；超级市场的经理人员希望能够从过去几年的销售记录中分析出顾客的消费习惯和行为，以便及时变换营销策略；地质学家想通过分析地球资源卫星发回的大量数据和照片来发现有开采价值的矿物资源等。有一个很普通却很能说明数据挖掘如何产生效益的例子：美国加州某个超级连锁店通过数据挖掘，从记录着每天销售信息和顾客基本情况的数据库中发现，在下班后前来购买婴儿尿布的顾客多数是男性，而且往往也同时购买啤酒。于是这个连锁店的经理当机立断，重新布置了货架，把啤酒类商品布置在婴儿尿布货架附近，并在二者之间放上土豆之类的佐酒小食品，同时把男士们需要的日常生活用品也就近布置。这样一来，上述几种商品的销量大增。

通过上面的例子可以看出，数据挖掘能为决策者提供重要的、极有价值的信息或知识，从而产生不可估量的效益。因此，虽然数据挖掘产品尚不成熟，但其市场份额却正日益扩大，越来越多的大中型企业开始利用数据挖掘来分析公司的数据以辅助决策，数据挖掘正逐渐成为在市场竞争中立于不败之地的法宝。

在经过十几年的技术发展之后，国外在数据挖掘技术上取得了丰富的经验。不但在研

究方面使各个学科的经验向该领域集中，而且出现了大量的软件产品，在社会的各个领域的应用也取得了丰硕的成果。

在国内，数据挖掘已经从单纯的研究走向产品的开发及技术的应用，随着市场经济的不断完善，数据挖掘的市场需求正在高速增长。数据挖掘与其他软件不同，由于需要不断地实验与评估，不懂原理或没有核心软件技术，其应用效果将大打折扣。在数据挖掘领域，我国的国产商品软件刚刚起步，但发展速度很快，随着市场的成熟与应用水平的提高，将会出现大量的国产软件产品。

1.2 什么是数据挖掘

数据挖掘（DM，Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在的有用信息和知识的过程。还有很多和这一术语相近的术语，如从数据库中发现知识（KDD）、数据分析、知识抽取、模式分析、数据考古、数据采集、信息收割、商业智能、数据融合以及决策支持等。国内的学者也把 Data Mining 译为数据采掘或数据开采。

人们把原始数据看作是形成知识的源泉，就像从矿石中采矿一样。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、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、过程控制等，还可以用于数据自身的维护。因此，数据挖掘是一门很广义的交叉学科，它汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。

特别要指出的是，数据挖掘技术从一开始就是面向应用的。它不仅是面向特定数据库的简单检索查询调用，而且要对这些数据进行微观或宏观的统计、分析、综合和推理，以指导实际问题的求解，企图发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测。例如，加拿大 BC 省电话公司要求加拿大 Simon Fraser 大学数据挖掘研究组，根据其拥有的十多年的客户数据，总结、分析并提出新的电话收费和管理办法，制定既有利于公司又有利于客户的优惠政策。美国 NBA 的著名篮球队的教练，利用 IBM 公司提供的数据挖掘技术，临场决定替换队员，一度在数据库界被传为佳话。这样一来，就把人们对数据的应用，从低层次的末端查询操作，提高到为各级经营决策者提供决策支持。需要指出的是，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，更不是什么机器的定理证明。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发现结果。因此数据挖掘的研究成果很讲求实际。1997 年第 3 届 KDD 国际学术大会上进行的实实在在的数据挖掘工具的竞赛评奖活动，就是一个生动的证明。最近，还有不少数据挖掘产品用来筛选 Internet 上的新闻，保护用户不受无聊电子邮件的干扰和商业推销，受到极大的欢迎。

当今数据库的容量已经达到上万亿的水平。在这些大量数据的背后隐藏了很多具有决策意义的信息，那么怎么得到这些“知识”呢？也就是怎样通过一棵棵的树木了解到整个森林的情况呢？

计算机科学对这个问题给出的最新回答就是数据挖掘，在“数据矿山”中找到蕴藏的“知识金块”，帮助企业减少不必要投资的同时提高资金回报。数据挖掘给企业带来的潜在的投资回报几乎是无止境的。世界范围内具有创新性的公司都开始采用数据挖掘技术来判断哪些是最有价值客户，重新制定产品推广策略（把产品推广给最需要它们的人），以用最小的花费得到最好的回报。

数据挖掘是一个利用各种分析工具在海量数据中发现模型和数据间关系的过程，这些模型和关系可以用来做出预测。

数据挖掘的第一步是描述数据——计算统计变量（比如平均值、均方差等），再用图表或图片直观地表示出来，进而可以看出一些变量之间的相关性（比如有一些值经常同时出现）。选择正确的数据源对整个数据挖掘项目的成败至关重要，在后面数据挖掘的步骤中会着重强调这一点。

单单是数据描述并不能为人们制定行动计划提供足够的依据，必须用这些历史数据建立一个预言模型，然后再用另外一些数据对这个模型进行测试，一个好的模型没必要与数据库中的数据 100%的相符（城市交通图也不是完全的实际交通线路的等比缩小），但它在做决策时是一个很好的指南和依据。

最后一步是验证模型。比如用所有对产品推广计划做出回应的人的数据库做了一个模型，来预测什么样的人会对产品感兴趣。是在得到这个模型后就直接利用这个模型做出决策或采取行动呢？还是更稳妥一点先对一小部分客户做一个实际的测试，然后再决定？

数据挖掘是一个工具，而不是有魔力的权杖。它不会坐在数据库上一直监视数据库，然后当它发现有意义的模型时发一封电子邮件。它仍然需要了解业务，理解数据，弄清分析方法。数据挖掘只是帮助商业人士更深入、更容易地分析数据，它无法告诉某个模型对企业的实际价值，而且数据挖掘中得到的模型必须要在现实生活中进行验证。

注意数据挖掘中得到的预言模型并不会告诉一个人为什么会做一件事或采取某个行动，数据挖掘只会告诉使用者它会这样做，至于为什么它这样做则需要人去考虑。比如，数据挖掘可能会告诉你，如果这个人是男的、年收入在 5 万到 6 万之间，那么他可能会买你的商品或服务。你可能会利用这条规则，集中向这类人推销商品并从中获益，但是数据挖掘工具不会告诉你他们为什么会买你的东西，也不能保证所有符合这条规则的人都会买。

为了保证数据挖掘结果的价值，必须了解数据，这一点至关重要。输入数据库中的异常数据、不相关的字段或互相冲突的字段（比如身份证号码和生日不一致）、数据的编码方式等都会对数据挖掘输出结果的质量产生影响。虽然一些算法自身会对上面提到的这些问题做一些考虑，但让算法自己做所有这些决定是不明智的。

数据挖掘不会在缺乏指导的情况下自动地发现模型。不能让数据挖掘工具帮我们提高直接邮件推销的响应率，而是应该让数据挖掘工具找对推销回应的人，或既回应又做了大

量订单的人的特征。在数据挖掘中，寻找这两种模型是很不相同的。

虽然数据挖掘工具可让使用者不必再掌握高深的统计分析技术，但使用者仍然需要知道所选用的数据挖掘工具是如何工作的，它所采用的算法的原理是什么。所选用的技术和优化方法会对模型的准确度和生成速度产生很大影响。

数据挖掘永远不会替代有经验的商业分析师或管理人员所起的作用，它只是提供一个强大的工具。每个成熟的、了解市场的公司都已经具有一些重要的、能产生高回报的模型，这些模型可能是管理人员花了很长时间，作了很多调查，甚至是经过很多失误之后得来的。数据挖掘工具要做的就是使模型得到得更容易、更方便，而且有根据。

1.3 数据挖掘的数据来源

数据挖掘所依赖的数据来源多种多样，可以是常用的关系数据库、事务数据库、文本数据库、多媒体数据库等，主要取决于用户的目的及所处的领域。目前，数据挖掘的数据主要取自关系数据库与数据仓库。

1. 关系数据库

日常运行的业务系统拥有大量的数据库，如保险公司的客户记录、交通运管部门的车辆数据库，但随着业务的变化及时间的推移，这些数据库的数据格式会发生变化，需要对这些数据先进行整理及清洗。

2. 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拿到数据挖掘库或数据集市中（见图 1-1）。从数据仓库中直接得到数据挖掘的数据有许多好处。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不必要再清理一次了，而且所有的数据不一致问题都已经被解决了。

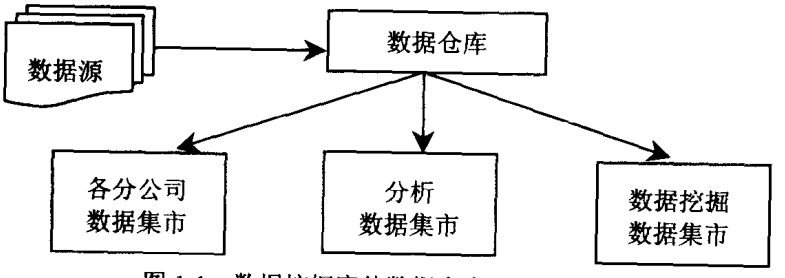

图 1-1 数据挖掘库从数据仓库中得出

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 。

3. 事务数据库

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、把所有的数据导入一个数据仓库内，是一项巨大的工程，可能要用几年的时间并花费数百万的资金才能完成。若只是为了数据挖掘，可以是把一个或几个事务数据库集中到一个只读的数据挖掘库（见图 1-2）中，就把它当作数据集市，然后

在它上面进行数据挖掘。

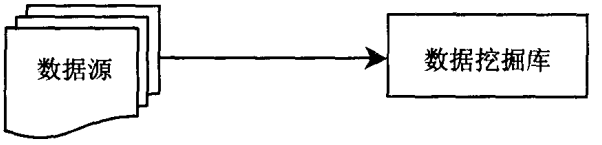

图 1-2　数据挖掘库从事务数据库中得出

4. 高级数据库及高级数据库的应用

近年来，数据库技术已发生了很大的变化，数据库的应用在 CAD、软件工程及办公信息系统等领域已得到运用。由原来的单一关系数据库发展到面向对象数据库、事务（Transaction）数据库、空间（Spatial）数据库、对象-关系数据库、文本数据库、多媒体数据库等新的数据库系统，同样，数据挖掘的数据来源也可取自这些类型的数据库系统。

面向对象的数据库是符合面向对象编程规范的数据库系统，即基于对象封装的概念之上，具有对象的继承特点等，面向对象概念均适用于这种数据库系统。

面向对象数据库的主要构造方法是在关系数据库上增加面向对象的概念，称为对象-关系数据库；另一种方法是利用永久对象编程语言构造，即将数据库功能集成到永久对象编程语言中，具有较好的性能。

一个事务具有原子性、连续性、互斥性及永久性，典型的事务处理是商场中的 POS 机售货，不但输出事务处理的结果，还将后台的库存、收款等数据库进行了永久性的更新，在现实生活中的计算机系统中有大量的事务处理数据库，如财务软件中的凭证处理、银行中的存款取款等处理均为事务处理。

空间（Spatial）数据库存储空间位置的信息并提供基于空间位置的有效查询及索引支持。例如，数据库中存储一个多边形的集合，可通过查询找到其中的一个多边形。空间数据库一般使用特殊的索引结构。空间数据库主要有用于 CAD 的设计数据库及地理信息系统（GIS）数据库。

多媒体（Multimedia）数据库存储多媒体数据，如图像、声频数据及视频数据。数据挖掘技术用于该领域时一般采用模式识别技术。

近几年，因特网的大规模普及使人们接受的信息无法承受，利用数据挖掘技术在 Web 上进行信息搜索成为高速增长的一个领域。

1.4　数据挖掘的分类

数据挖掘系统利用的技术越多，得出的结果精确性就越高。原因很简单，对于某一种技术不适用的问题，其他方法却可能奏效。这主要取决于问题的类型以及数据的类型和规模。

数据挖掘涉及的学科领域和方法很多，有多种分类法。根据挖掘任务，可分为分类或预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等。根据挖掘对象分，有关系数据库、面向对象数据库、空间数据库、

时态数据库、文本数据库、多媒体数据库、异构数据库、遗产数据库以及 Web。根据挖掘方法，可分为机器学习方法、统计方法、神经网络方法和数据库方法。机器学习包含归纳学习方法、基于案例学习、遗传算法等。统计方法包含回归分析、判别分析、聚类分析、探索性分析等。神经网络方法包含前向神经网络、自组织神经网络等。数据库方法主要是多维数据分析方法，另外还有面向属性的归纳方法。

数据挖掘所能发现的知识有如下几种：广义型知识，反映同类事物共同性质的知识；特征型知识，反映事物各方面的特征知识；差异型知识，反映不同事物之间属性差别的知识；关联型知识，反映事物之间依赖或关联关系的知识；预测型知识，根据历史的和当前的数据推测未来数据；偏离型知识，揭示事物偏离常规的异常现象。所有这些知识都可以在不同的概念层次上被发现，随着概念树的提升，从微观到宏观，以满足不同用户、不同层次决策的需要。例如，从一家超市的数据仓库中，可以发现的一条典型关联规则可能是“买面包和黄油的顾客十有八九也买牛奶”，也可能是“买食品的顾客几乎都用信用卡”，这种规则对于商家开发和实施客户化的销售计划和策略是非常有用的。至于发现工具和方法，常用的有分类、聚类、关联、模式识别、可视化、决策树、遗传算法、不确定性处理等。

1.4.1 分类分析（Classification Analysis）

预言模型以通过数据库中的某些数据得到另外的数据为目标。若预测的变量是离散的（如批准或者否决一项贷款），这类问题就称为分类（Classification）；如果预测的变量是连续的（如预测盈亏情况），这种问题称之为回归（Regression）。分类一直为人们所关注。数据挖掘广泛使用的方法有决策树、神经网络、径向基础函数（Radial Basis Functions）等。

基于债务水平、收入水平和工作情况，可对给定用户进行信用风险分析。分类算法通过判断以上属性与已知训练数据中风险程度的关系给出预言结果。决策树是一种常见且有用的预言模型。图 1-3 是一个可用于判断信用风险的训练数据集。客户的债务情况、收入情况、工作情况及信用情况被收集其中。图 1-4 显示了一个由图 1-3 中原始数据生成的决策树。

客户编号	债务情况	收入情况	工作类型	信用风险
1	High	High	SelfEmployed	Bad
2	High	High	Salaried	Bad
3	High	Low	Salaried	Bad
4	Low	Low	Salaried	Good
5	Low	Low	SelfEmployed	Bad
6	Low	High	SelfEmployed	Good
7	Low	High	Salaried	Good

图 1-3 原始数据

在这个普通的例子中，一个决策树算法对于信用风险预测来说，最重要的属性是债务情况。因此决策树中的第一个分支点设在债务情况。叶子“Debt=High”包含三条“Credit