高等院校计算机教育系列教材

数据挖掘导论

戴 红 常子冠 于 宁 主编

- 结构清晰，知识完整
- 入门快速，易教易学
- 实例丰富，实用性强
- 学以致用，注重能力

赠送实例代码、电子课件和课后习题答案

清华大学出版社

高等院校计算机教育系列教材

数据挖掘导论

戴 红 常子冠 于 宁 主 编

清华大学出版社
北 京

内 容 简 介

本书为数据挖掘入门级教材，共分 8 章，主要内容分为三个专题：技术、数据和评估。技术专题包括决策树技术、K-means 算法、关联分析技术、神经网络技术、回归分析技术、贝叶斯分析、凝聚聚类、概念分层聚类、混合模型聚类技术的 EM 算法、时间序列分析和基于 Web 的数据挖掘等常用的机器学习方法和统计技术。数据专题包括数据库中的知识发现处理模型和数据仓库及 OLAP 技术。评估专题包括利用检验集分类正确率和混淆矩阵，并结合检验集置信区间评估有指导学习模型，使用无指导聚类技术评估有指导模型，利用 Lift 和假设检验比较两个有指导学习模型，使用 MS Excel 2010 和经典的假设检验模型评估属性，使用簇质量度量方法和有指导学习技术评估无指导聚类模型。

本书秉承教材风格，强调广度讲解。注重成熟模型和开源工具的使用，以提高学习者的应用能力为目标；注重结合实例和实验，加强基本概念和原理的理解和运用；注重实例的趣味性和生活性，提高学习者学习的积极性。使用章后练习、计算和实验作业巩固和检验所学内容；使用词汇表附录，解释和规范数据挖掘学科专业术语；使用适合教学的简单易用开源的 Weka 和通用的 MS Excel 软件工具实施数据挖掘验证和体验数据挖掘的精妙。

本书可作为普通高等院校计算机科学、信息科学、数学和统计学专业的入门教材，也可作为如经济学、管理学、档案学等对数据管理、数据分析与数据挖掘有教学需求的其他相关专业的基础教材。同时，对数据挖掘技术和方法感兴趣，致力于相关方面的研究和应用的其他读者，也可以从本书中获取基本的指导和体验。

本书配有教学幻灯片、大部分章后习题和实验的参考答案以及课程大纲。

图书在版编目(CIP)数据

数据挖掘导论/戴红，常子冠，于宁主编. --北京：清华大学出版社，2014
(高等院校计算机教育系列教材)
ISBN 978-7-302-38104-4

Ⅰ. ①数…　Ⅱ. ①戴…　②常…　③于…　Ⅲ. ①数据采集—高等学校—教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2014)第 224385 号

责任编辑：章忆文　陈立静
装帧设计：杨玉兰
责任校对：周剑云
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn,010-62791865

印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张**：13.75　　**字　　数**：331 千字
版　　次：2015 年 1 月第 1 版　　**印　　次**：2015 年 1 月第1 次印刷
印　　数：1～2500
定　　价：28.00 元

产品编号：061066-01

前　言

未来学家约翰·奈斯比特(John Naisbitt)惊呼："人类正被数据淹没，却饥渴于信息。"从浩瀚无际的数据海洋中发现潜在的、有价值的信息，是这个大数据时代的一个标志性工作。

数据挖掘(Data Mining)是利用一种或多种计算机学习技术，从数据中自动分析并提取信息的处理过程，其目的是发现数据中潜在的和有价值的信息、知识、规律、联系、模式，从而为解释当前行为和预测未来结果提供支持。数据挖掘一般使用机器学习、统计学、联机分析处理、专家系统和模式识别等多种方法来实现，是一门交叉学科，涉及数据库技术、人工智能技术、统计学方法、可视化技术、并行计算等。数据挖掘是一种商业智能信息处理技术，其围绕商业目标，对大量商业数据进行抽取、转换、分析和处理，从中提取辅助商业决策的关键性数据，揭示隐藏的、未知的或验证已知的规律性，是一种深层次的商业数据分析方法。

本书作为一本数据挖掘的入门级教材，关注于数据挖掘的基本概念、基本原理和基本技术的介绍和实践应用。全书围绕知识发现过程中的数据专题、技术专题和评估专题展开，包含大量实例和实验。实验采用 Weka 开源数据挖掘工具和 MS Excel 2010，两者作为教学软件，具有很好的通用性和易学易用性。本书最后附有词汇表和数据挖掘数据集，包括了书中涉及的数据挖掘的最基本词汇、例子及实验所用数据集。其中数据集有来自 UCI 的共享数据集，也有为了举例和实验而设计的假想数据集。

本书分为 8 章和两个附录，其中戴红编写了 8 章中的大部分内容，常子冠和于宁编写了附录 A 和附录 B，以及前 8 章的部分内容。

本书目标

本书希望帮助读者达到以下学习目标。

- 了解数据挖掘的技术定义和商业定义、作用和应用领域。
- 了解数据挖掘与知识发现、数据查询、专家系统的关系。
- 掌握数据挖掘和知识发现的处理过程。
- 掌握数据挖掘的基本技术和方法，包括有指导的学习技术——决策树技术、产生式规则、神经网络技术和统计分析方法，以及无指导聚类技术和关联分析方法。
- 掌握数据挖掘的评估技术，包括数据评估和模型评估方法。
- 了解数据仓库的设计目标和结构。
- 了解联机分析处理(OLAP)的目标和数据分析方法。
- 掌握时间序列分析方法，了解基于 Web 的数据挖掘目标、方法和技术。
- 能够使用 Weka 软件工具，应用各种数据挖掘算法，建立分类和聚类模型并进行

关联分析，尝试解决实际问题。

- 能够使用 MS Excel 进行数据相关性分析，建立回归模型，以及使用 Excel 的数据透视表和数据透视图进行 OLAP 分析。

本书读者

本书既可作为计算机科学、信息科学、数学和统计学专业的入门教材，也可作为如经济学、管理学、档案学等，对数据管理、数据分析与数据挖掘有教学需求的其他相关专业的基础教材。同时，对数据挖掘技术和方法感兴趣，致力于相关方面的研究和应用的其他读者，也可以从本书中获取基本的指导和体验。

本书特点

本书强调基本概念、基本原理、基本技术的广度讲解。注重成熟模型和开源工具的介绍和使用；注重对数据挖掘经典算法过程的可理解性描述，而非聚焦细节的剖析，以提高授课学生的应用能力；注重结合基础实用案例，通过案例加强基本概念和原理的理解和运用；同时注重提高实例的趣味性和生活性，以提高学生的学习积极性。

本书秉承教材风格，使用实例和实验来描述和验证概念、原理和技术；使用章后练习、计算和实验作业巩固和检验所学内容；使用词汇表附录，解释和规范数据挖掘学科专业术语；使用适合教学的简单易用开源的 Weka 和通用的 MS Excel 软件工具实施数据挖掘，验证和体验数据挖掘的精妙。

本书内容

第 1 章 认识数据挖掘。主要是对数据挖掘作全面的概述，包括数据挖掘的基本概念、作用、过程、方法、技术和应用。同时介绍了本书使用的开源数据挖掘软件 Weka。

从第 2 章到第 8 章，可分为三个专题：技术专题、数据专题和评估专题。

技术专题

第 2 章 基本数据挖掘技术。介绍有指导学习技术中的决策树算法、无指导聚类和 K-means 算法，重点讨论生成关联规则技术和针对不同问题如何考虑选择不同的数据挖掘技术和算法。

第 6 章 神经网络技术。介绍神经网络的基本概念、结构模型、反向传播学习、自组织学习方法和神经网络技术的优势和缺点，讨论神经网络的输入和输出数据的要求，详细描述反向传播学习算法和自组织学习方法的一次迭代过程，并通过两个实验，介绍了使用 Weka 软件实现 BP 前馈神经网络模型的过程。

第 7 章 统计技术。介绍数据挖掘中几种常用的统计技术，包括线性回归、非线性回归和树回归，贝叶斯分类器，聚类技术中的凝聚聚类、概念分层聚类和混合模型聚类技术的 EM 算法，对比了统计技术和机器学习方法的不同之处，为针对不同的问题和数据情况选择不同的数据挖掘技术提供参考。

第 8 章 时间序列分析和基于 Web 的挖掘。介绍如何使用神经网络技术和线性回归方法建立预测模型，解决时间序列预测问题，使用数据挖掘对 Web 站点进行自动化评估和提供个性化服务，并就 Web 站点的自适应调整和改善进行了简单阐述，同时针对多模型应用中的两种著名方法装袋和推进进行了简单介绍。

数据专题

第 3 章 数据库中的知识发现。介绍了知识发现的基本概念、基本过程和典型模型，重点剖析知识发现过程中的每个步骤的任务和方法，并通过一个案例说明知识发现的整个过程。

第 4 章 数据仓库。概括性地阐述了数据库和数据仓库的基本概念和特点，介绍了数据仓库模型的设计，重点讨论最常用的星型模型、雪花模型和星座模型的设计，并解释了数据集市和决策支持系统的基本概念。通过一个实验，描述了从决策支持的角度，对数据仓库中的数据进行多维分析的方法。最后介绍了利用 MS Excel 数据透视表和数据透视图建立多维数据分析模型的方法。

评估专题

第 5 章 评估技术。概述了数据挖掘过程中评估的内容和工具，介绍了具有分类输出的有指导学习模型的最基本评估工具——检验集分类正确率和混淆矩阵、数值型输出模型的评估、检验置信区间的计算以及无指导聚类技术对于有指导学习模型的评估作用、有指导学习模型的比较方法，重点讨论了利用 Lift 和假设检验对两个有指导学习模型的性能进行比较。同时，讨论了属性评估，使用 MS Excel 的函数和散点图进行属性相关性分析，以及在属性选择中，如何通过应用经典的假设检验模型来确定数值属性的重要性。本章最后给出了两种无指导聚类模型的评估方法。

附录 本书有两个附录：附录 A 为词汇表，包含了各章以及 Weka 软件中出现的主要词汇和关键术语；附录 B 为本书各章实例、实验、章后习题中涉及的数据集的相关描述，有来自 UCI 的网络共享数据集，也有假想的数据集。

本书资源

- 教学幻灯片，包括所有章节的 PowerPoint 教学幻灯片。
- 习题答案，包括大部分章后习题和实验的参考答案。
- 课程大纲，包括学时建议和各学时的授课内容、讨论议题、习题和实验选择以及阶段测验的建议。

推荐资源如下。

(1) 全球最大的数据挖掘信息网站——http://www.kdnuggets.com/。Data Mining Community's Top Resource for Data Mining and Analytics Software, Jobs, Consulting, Courses, and more。

(2) 机器学习领域的 UCI 数据集——http://archive.ics.uci.edu/ml/。UCI 数据库是加州大

学欧文分校(University of California Irvine)提出的用于机器学习的数据库，目前拥有 200 多个数据集，并且数目还在不断增加。UCI 数据集在数据挖掘领域被认为是标准测试数据集。

欢迎读者来函，对书中不妥之处批评、指正。联系邮箱：daihong@buu.edu.cn。

编　者

目　录

第 1 章　认识数据挖掘

本章要点提示

千百年来，人类总是从自然界和人类社会中不断地寻找和发现信息、知识、规律、联系和模式来发展自己，推进人类的进步。如农民在耕种中寻找着庄稼生长的规律，猎人在动物活动行为中寻找猎物的生活习性，教师在教学中寻找着教学规律，医生在患者病例中寻找疾病之间的联系，商人在消费行为中寻找模式等。数据挖掘就是在数据中发现潜在的和有用的信息、知识、规律、联系和模式的过程。

从本章开始，我们将进入数据挖掘和知识发现的神奇之旅。本章为全书的导入，在本章中将对数据挖掘的基本概念、作用、过程、方法、技术和应用作全面的概述。本章 1.1 节给出了数据挖掘的定义。1.2 节将对与数据挖掘有着密切关系的机器学习进行探讨。1.3 节介绍了数据查询与数据挖掘之间的关系。1.4 节介绍了专家系统和数据挖掘方法解决问题的不同。1.5 节描述了数据挖掘的过程。1.6 节对数据挖掘的作用进行了全面阐述。1.7 节介绍了几种常见的数据挖掘方法和技术。1.8 节对数据挖掘的应用领域和经典案例进行了简单介绍。1.9 节介绍了本书使用的一种开源数据挖掘软件 Weka。

1.1　数据挖掘的定义

数据挖掘(Data Mining)是利用一种或多种计算机学习技术，从数据中自动分析并提取信息的处理过程。数据挖掘的目的是寻找和发现数据中潜在的有价值的信息、知识、规律、联系和模式。数据挖掘与计算机科学有关，一般使用机器学习、统计学、联机分析处理、专家系统和模式识别等多种方法来实现。从学科的角度上看，数据挖掘是一门交叉学科，涉及数据库技术、人工智能技术、统计学、可视化技术、并行计算等多种技术。

以上是从技术角度给出的数据挖掘定义。从商业角度上来描述数据挖掘的定义为：数据挖掘是一种商业智能信息处理技术，是围绕商业目标开展的，对大量商业数据进行抽取、转换、分析和处理，从中提取辅助商业决策的关键性数据，揭示隐藏的、未知的或验证已知的规律性，是一种深层次的商业数据分析方法。

以下是对定义中的几个概念进行的进一步解释。

(1) 数据。数据挖掘使用的数据一般是真实的、大量的、可能具有噪声的数据，数据的质量很大程度上影响着数据挖掘的质量。目前随着计算机硬件技术和数据库、数据仓库数据管理等软件技术的发展，计算机能够收集和分析并处理大量的、结构复杂的、异构的数据。同时大量的数据中，可能真正有价值的信息很少，数据挖掘就是要在这些数据中发现有价值的信息。“人类正被数据淹没，却饥渴于信息”——约翰·奈斯比特(John Naisbitt, 未来学家)。

(2) 潜在的有价值的信息、知识、规律、联系、模式。一般从数据中发现的不是浅知识(Shallow Knowledge)，即不是通过查询和搜索就能够获取的信息，而是隐含的、潜在的

规律和模式。并且发现的知识是可被用户接受和理解的，往往可用于解决某个特定问题或进行特定领域的决策支持。

(3) 数据挖掘与知识发现的关系。数据库中的知识发现(Knowledge Discovery in Database，KDD)是一个经常与数据挖掘互换使用的术语。KDD 是一个处理过程和方法体系，它包括目标定义、数据准备、数据挖掘和解释、模型检验和评估、模型应用等阶段。尽管在很多场合下，数据挖掘和知识发现之间的界限并不明显，看到不加区分地使用，但严格来说，数据挖掘其实仅仅是 KDD 过程中的一个阶段。第 3 章中将详细讨论 KDD 过程和方法。

除了知识发现，与数据挖掘相关的词汇还有机器学习和人工智能、商务智能、模式识别、数据查询和数据分析、决策支持和专家系统等。下面对与数据挖掘相关的机器学习、数据查询和专家系统进行简单解释，并给出它们与数据挖掘之间的关系。在第 4 章中对数据分析、决策支持作进一步的阐述。

1.2 机 器 学 习

机器学习(Machine Learning，ML)是模拟人类的学习方法来解决计算机获取知识问题的方法。通过机器学习，可以利用大量的经验积累来改善系统的性能。机器学习是人工智能(Artificial Intelligence)的核心，是使计算机具有智能的根本途径，在商业智能分析等领域具有广泛的应用。

1.2.1 概念学习

机器学习是通过对大量的实例进行训练，从中发现经验化规律的过程。机器学习结果的通常表现形式为概念，即机器最擅长的是学习概念。概念(Concept)是具有某些共同特征的对象、符号或事件的集合。概念可以从三个不同的角度来看待，分别为概念定义的传统角度、概率角度和样本角度。

1. 传统角度

在传统角度(Classical View)中，所有概念都有明确的定义，某个实例是否属于一个概念，需要按照这个明确的定义来确定。如“优秀学生”若使用经典概念观点，则可定义为：每学期平均成绩 85 分(含)以上、参加社会工作 1 项及以上的学生。这个定义中存在两个条件，一为平均成绩的条件，二是参加社会工作情况。若将平均成绩和参加社会工作作为两个属性，≥85 分和 1 项作为属性的值，这个定义可以写成如下形式。

(1) 平均成绩≥85。

(2) 承担社会工作≥1。

传统概念定义中，概念的特征是定义明确的，不允许出现模棱两可的情况。以上两个条件必须同时满足，这样的学生才是优秀学生。

2. 概率角度

对个别样本实例进行概括性描述，这些概括性说明就构成了概率角度(Probabilistic

View)概念。如“优秀学生”的概率角度的概念定义如下。

(1) 一贯表现较好、成绩优良的学生，大多数都是优秀学生。

(2) 承担过社会工作，平均成绩在80分以上的学生，80%都是优秀学生。

以上两条是通过观察大量“优秀学生”实例得出的概括性描述。概率的观点并未给出优秀学生的确切定义，只是提供了优秀学生判定的一个参考。通过概率观点所定义的概念，不能直接得出判断结论。如一个参加过社会工作、平均成绩为85分的学生，不能肯定其就是“优秀学生”，他作为“优秀学生”的概率为80%。

3. 样本角度(Exemplar View)

样本角度(Exemplar View)概念定义既不是传统定义明确的条件，也不是概括性描述，而是将某个概念中的典型实例组成一个集合，使用该集合来描述概念定义。判断一个新实例是否属于某个概念分类，就将其与该集合中的典型实例进行比较，符合其中的某个实例，它就是这个概念类中的一员。如“优秀学生”的样本角度的概念定义如下。

(1) 承担过1项社会工作，平均成绩85分。

(2) 承担过2项社会工作，平均成绩83分。

(3) 没有承担过社会工作，平均成绩90分。

以上仅仅列出了三个样本组成的集合，实际中，为了更好地覆盖所有概念类的实例情况，样本除了能够正确描述概念类，具有典型性之外，还需要具有一定的覆盖度。从样本角度上，是将概念通过概念样本来表达，并用概念样本分类新的实例。若一名学生平均成绩为91分，若其与该概念样本充分地相似，则可以认为他是“优秀学生”。

在机器学习中，机器学习工具的不同决定了所学概念的不同表达形式。一般的概念结构如树、规则、网络和数学方程等。其中树结构和规则是人类容易解释和理解的概念形式，被称为白盒子结构，而网络和数学方程是人类不容易解释和理解的概念结构，被称为黑盒子结构。

1.2.2 归纳学习

机器学习的方式是基于归纳的学习。归纳学习(Induction-Based Learning)方法是人类学习的最重要方式之一。人类通过对事物的特定实例的观察，对所掌握的已有经验材料的研究，从归纳中获取和探索新知识，并以概念的形式表现出来。如小时候，我们在认知这个世界时，通过各种事物的典型实例，如动物中的老虎、狮子、大象等，植物中的松树、玫瑰花、兰花草等，在大脑中形成个别实例的记忆，通过大脑的加工抽象出表达这些事物的典型特征(属性)，如外观、形状、颜色、声音、动作等，最终形成动物和植物的概念分类模型。模型建立完成后，在对世界的进一步认知过程中，就会自然地使用这些模型来区分具有相似特征的更多的事物或实例。在应用概念分类模型进行未知实例分类的过程中，还在使用新的实例进行模型的进一步修正，这个过程使得我们大脑中对于事物的认识进一步地准确和完整。这种学习就是归纳学习。

以下是几个归纳学习的例子。

(1) 通过分析信用卡持卡人的消费行为，归纳出他的信用卡消费模式(模型)。当信用卡

被盗刷时，信用卡公司可以利用这个消费模式(模型)判断出该消费行为是异常的，从而提醒持卡人该卡被盗用。

(2) 零售商经常通过分析顾客的购买行为，找出行为中的规律，如经典的啤酒和尿布案例，归纳出一般性规律，从而指导货架的摆放和商品的促销。

(3) 通过鸢尾花的花瓣和花萼的长度和宽度的特点，归纳出鸢尾花的类别，用该分类模型来判断未知种类鸢尾花的类别。

数据挖掘中使用了大量的机器学习方法，一般分为两大类：有指导(监督)的学习和无指导(监督)的聚类。有指导的学习就是上述的基于归纳的学习，是通过对大量已知分类或输出结果的实例进行训练，建立分类或预测模型，用来分类未知实例或预测输出结果的未来值。

1.2.3 有指导的学习

归纳学习是为了建立一个用于分类或预测的模型，而通过对大量已知分类或输出结果值的实例进行训练，调整分类模型的结构，达到建立能够准确分类或预测未知的模型的目的。这种基于归纳的概念学习过程被称为有指导(监督)的学习(Supervised Learning)。其中，用于有指导学习的样本数据被称为数据实例(Instance)，用于训练的实例被称为训练实例(Training Instance)。除此之外，分类模型建立完成后，通常需要经过检验实例(Test Instance)进行检验，判断模型是否能够很好地应用在未知实例的分类或预测中。

模型的训练过程是从个体实例归纳出概念类，属于归纳学习，但利用分类模型对未知实例进行分类判断的过程则是演绎的过程。下面通过一个例子来说明有指导的学习过程。

【例 1.1】 给定如表 1.1 所示的数据集 T，使用有指导的学习方法建立分类模型，对未知类别的实例进行分类。

表 1.1 感冒诊断假想数据集

序号	Increased -lym 淋巴细胞升高	Leukocytosis 白细胞升高	Fever 发烧	Acute-onset 起病急	Sore-throat 咽痛	Cooling-effect 退热效果	Group 群体发病	Cold-type 感冒类型
1	Yes	No	Yes	Yes	No	Good	Yes	Viral
2	No	Yes	Yes	No	Yes	Not good	Yes	Bacterial
3	Yes	No	Yes	Yes	Yes	Good	Yes	Viral
4	Yes	No	No	Yes	No	Unknown	No	Viral
5	No	No	No	No	Yes	Unknown	No	Bacterial
6	No	Yes	Yes	Yes	Yes	Not good	No	Bacterial
7	No	Yes	Yes	No	Yes	Not good	No	Viral
8	Yes	No	Yes	No	No	Good	Yes	Viral
9	Yes	Yes	Yes	Yes	Yes	Good	Yes	Viral
10	Yes	Yes	Yes	No	Yes	Not good	No	Bacterial

表 1.1 是一个关于感冒类型诊断的小型假想数据集，数据集的格式为“属性-值”格式

(Attribute-Value Format)，表中第一行显示了属性的名称。数据集共有 8 个属性，前 7 个属性表达了病人患感冒的临床症状，分别为 Increased-lym(淋巴细胞是否升高)、Leukocytosis(白细胞是否升高)、Fever(是否发烧)、Acute-onset(是否起病急)、Sore-throat(是否有咽痛症状)、Cooling-effect(服用退烧药的退热效果如何)、Group(是否有群体发病情况)。这些属性在有指导的学习中被称为输入属性(Input Attribute)，是用来表示分类特征的属性。第 8 个属性为 Cold-type(感冒类型)，它有两个取值：Viral(病毒性的)和 Bacterial(细菌性的)，是有指导学习中的输出结果，被称为类或输出属性(Output Attribute)。

数据集中有 10 个实例，每个实例显示一位感冒患者的症状和类型。例如，第一个实例表示感冒患者淋巴细胞升高、白细胞未升高、发烧、起病急、咽部不疼痛、使用退烧药效果较好、有群体发病情况，最后该患者被诊断为病毒性感冒。

机器学习中的有指导学习方法和技术很多，常用的有决策树、产生式规则、神经网络等。下面使用最常用的决策树方法建立表 1.1 的分类模型，用于对一个未知感冒类型的患者进行诊断。

决策树(Decision Tree)是一种简单的、易于解释和理解的概念结构。决策树是一个倒立的树，树的非叶子节点表示在一个属性上的分类检查，叶子节点表示决策判断的结果，该结果选择了正确分类较多实例的分类。决策树有很多算法，在第 2 章中将对此进行详细介绍，这里使用决策树的经典算法 C4.5。决策树如图 1.1 所示。

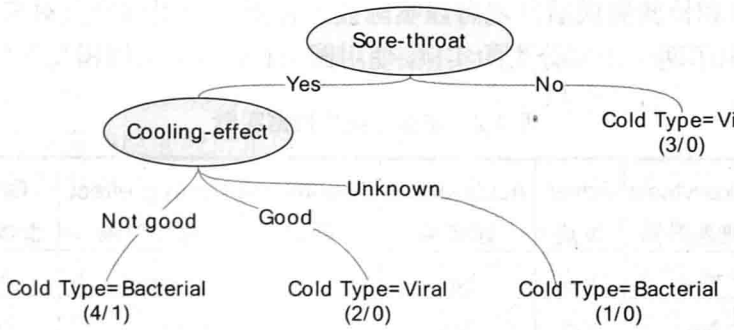

图 1.1　感冒类型诊断决策树

从这棵决策树中得出以下几点结论。

(1) 当患者没有咽痛症状(Sore-throat=No)，可以认为其患有病毒性感冒(Cold-type=Viral)。

(2) 当患者有咽痛症状(Sore-throat=Yes)，并且使用了退烧药，用药效果不好(Cooling-effect=Not good)，则可以判断其得了细菌性感冒(Cold-type = Bacterial)。

(3) 当患者有咽痛症状，并且使用了退烧药，用药效果较好(Cooling-effect = Good)，则可以判断其得了病毒性感冒。

(4) 当患者有咽痛症状，未用退烧药(Cooling-effect = Unknown)，则可以判断其得了细菌性感冒。

从决策树中可以看到，决策树中仅出现了两个输入属性：Sore-throat(是否有咽痛症状)和 Cooling-effect(服用退烧药的退热效果如何)，其他属性如 Increased-lym(淋巴细胞是否升高)、Leukocytosis(白细胞是否升高)、Fever(是否发烧)、Acute-onset(是否起病急)、Group(是否有群体发病情况)对于诊断感冒类型没有起到任何作用。因数据集数据量太少，此结论仅

供参考。

决策树叶子节点中的数字格式(m/n)表示沿着这条树的路径(分支)达到叶子节点的实例数共 m 个，其中 n 个实例被分类错误。例如，当患者没有咽痛症状(Sore-throat = No)，该分支的决策结果是认为患者患有病毒性感冒(Cold-type = Viral)。该分支的叶子节点中数字为(3/0)，表示符合这条分支判断条件的实例共有 3 条，全部被分类为患有病毒性感冒，并且与实际情况相比，全部被分类正确。而当患者有咽痛症状(Sore-throat = Yes)，并且使用了退烧药，用药效果不好(Cooling-effect = Not good)，该分支的决策结果为判断患者得了细菌性感冒(Cold-type = Bacterial)。该分支的叶子节点中数字为(4/1)，表示符合这条分支判断条件的实例共有 4 条，全部被分类为患有细菌性感冒，但实际上，其中有一人是患有病毒性感冒的，即有一条实例被分类错误。

表 1.1 中的 10 条实例作为训练数据用于创建决策树模型，这棵决策树能够正确分类这 10 条实例中的 9 条，分类正确率达到 9/10=90%。但是这个分类正确率仅仅说明对于训练数据的分类正确程度，没有检验对于未参与训练的其他实例的分类正确程度，所以还应该使用检验实例来检验模型分类未参与训练的未知实例的分类正确率，从而确定模型在后续使用中的效果。检验集中实例的分类也是已知的，这样才能对模型计算的分类结果与实际分类结果进行比较，计算出检验数据上的分类正确率，这个检验集分类正确率将预示着模型未来的性能。

分类模型建立和检验完成后，就可以实际投入使用，即用该模型对未知分类的实例进行分类。表 1.2 给出了两个未知分类的实例，使用图 1.1 中的决策树模型对它他们进行分类。

表 1.2　未知分类的数据实例

Increased -lym 淋巴细胞升高	Leukocytosis 白细胞升高	Fever 发烧	Acute-onset 起病急	Sore-throat 咽痛	Cooling-effect 退热效果	Group 群体发病	Cold-type 感冒类型
No	Yes	Yes	No	No	Not good	No	?
Yes	No	Yes	No	Yes	Good	No	?

(1) 对于第一条实例，患者没有咽痛症状(Sore-throat = No)，则可以诊断为患有病毒性感冒(Cold-type = Viral)。

(2) 对于第二条实例，患者有咽痛症状(Sore-throat = Yes)，并且使用了退烧药，用药效果较好(Cooling-effect = Good)，则可以诊断为也患有病毒性感冒(Cold-type = Viral)。

决策树一般都可以被翻译为一个产生式规则集合。产生式规则的格式为：

IF 前提条件 THEN 结论

前提条件描述输入属性的值，结论说明输出属性的结果。将决策树翻译为产生式规则的方法是从根节点出发，沿着树的一条路径到叶子节点来创建规则。规则的前提条件由这条路径中的所有属性值组成，规则的结论是叶子节点的输出值。图 1.1 的感冒类型诊断决策树可以翻译为以下 4 条产生式规则。

```
(1) IF Sore-throat = No  THEN  Cold-type = Viral
(2) IF Sore-throat = Yes & Cooling-effect = Good  THEN Cold-type = Viral
(3) IF Sore-throat = Yes & Cooling-effect = Not good THEN Cold-type =Bacterial
(4) IF Sore-throat = Yes & Cooling-effect = Unknown  THEN Cold-type = Bacterial
```

现在可以使用产生式规则对表 1.2 中的未知实例进行分类。

(1) 对于第一条实例，患者没有咽痛症状(Sore-throat = No)，适用第一条规则，则可以诊断为患有病毒性感冒(Cold-type = Viral)。

(2) 对于第二条实例，患者有咽痛症状(Sore-throat = Yes)，并且使用了退烧药，用药效果较好(Cooling-effect = Good)，适用第二条规则，则可以诊断为也患有病毒性感冒(Cold-type = Viral)。

1.2.4 无指导的聚类

无指导(监督)聚类(Unsupervised Clustering)是一种无指导(无教师)的学习，在学习训练之前，没有预先定义好分类的实例，数据实例按照某种相似性度量方法，计算实例之间的相似程度，将最为相似的实例聚类在一个组——簇(Cluster)中，再解释和理解每个簇的含义，从中发现聚类的意义。

【例 1.2】 给定如表 1.1 所示的数据集 T，使用无指导聚类方法，对所有实例进行分类，解释每个簇的含义。

对于表 1.1 中的数据先进行简单处理，删除 Cold-type(感冒类型)属性，这样表中数据仅为患者的患病症状，没有诊断结果，即没有任何有指导性的分类信息。现在我们希望通过无指导聚类方法，从这些数据中挖掘出潜在的有价值的信息或模式。

与有指导学习不同，在无指导聚类之前，不能确定数据挖掘的目标，即我们希望找到有价值的信息，但具体找什么，没有明确的目标。一般情况下，可以在评估了无指导聚类模型的质量后，对于将实例聚类为质量较好的几个簇的属性进行评估，评估哪些属性能够较好地聚类簇，哪些属性能够较好地区分不同簇的实例。

无指导聚类有很多种算法，如 K-means(K-均值)算法、凝聚聚类方法、概念分层 Cobweb 算法、EM 算法等，其中 K-means 算法是一种最为常用和易用的算法。算法需要在聚类前指定一个初始簇的个数，本例中，可以将初始簇个数指定为 2，应用 K-means 算法，将去掉感冒类型后的表 1.1 中的实例聚类为两个簇，每个簇有 5 个实例，分别为 Cluster 0 = {1,3,4,8,9}和 Cluster 1 = {2,5,6,7,10}(其中的数字为实例在表 1.1 中的序号)。通过观察这两个簇实例的感冒类型属性，发现实际上两个簇分别表达了病毒性感冒(Cold-type = Viral)和细菌性感冒(Cold-type = Bacterial)两种感冒类型。每个簇的概念结构可以表示为一个产生式规则，其规则如下。

```
(1)  IF Increased -lym = Yes & Cooling-effect =Good  THEN Cluster = 0
(rule accuracy = 4/4 = 100%, rule coverage = 4/5 = 80%)
(2)  IF Sore-throat = Yes & Cooling-effect = Not good  THEN Cluster = 1
(rule accuracy = 4/4 = 100%, rule coverage = 4/5 = 80%)
```

每条规则结论的后面的数字表示规则的准确率和覆盖率，分别表示了规则的置信度和有效性。Cluster 0 和 Cluster 1 的规则准确率分别为 100%，表示这两条规则在满足前提条件的情况下，100%是正确的。Cluster 0 和 Cluster 1 的规则覆盖率分别为 80%，表示在 Cluster 0 和 Cluster 1 的实例中的 80%满足规则的前提条件。

Cluster 0 规则显示出当某人淋巴细胞升高且用了退烧药后效果较好，则他一定患有病毒性感冒，在患病毒性感冒的人里，有 80%淋巴细胞升高且用了退烧药后效果较好。