北京大学信息管理系

《数据挖掘导论》讲义

第一章 绪论

北京大学信息管理系

2016 年秋

目录

第一章 绪论

1.1 什么是数据挖掘

数据挖掘涉及多学科领域，根据内容侧重点的不同，其可以用多种方法定义，也因此据有多个术语名称，例如：

数据挖掘（data mining）

数据库中的知识发现（KDD, knowledge discovery in databases）

知识抽取(knowledge extraction)

信息发现(information discovery)

智能数据分析(intelligent data analysis)

探索式数据分析(exploratory data analysis)

信息收获 (information harvesting)

数据考古(data archeology)等。

简单来说，数据挖掘是从大量数据中提取或发现（挖掘）知识的过程。许多人将数据挖掘视为数据库中知识发现（Knowledge discovery in database, KDD）的同义词，或 KDD 过程中不可缺少的一部分。

所谓知识发现，是指从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。下面将对以上定义进行更详细的解释：

- 数据集：一组事实 F，如关系数据库中的记录；
- 模式：一个用语言 L 表示的一个表达式 E，它可以用来描述数据集 F 的一个子集 F_E ，E 作为一个模式要求它比对数据子集 F_E 的枚举要简单（所用的描述信息量要少）。如 y=f(x)就是一个一元线性函数模式；
- 过程：KDD 是一个多阶段的过程，需要多阶段的处理，涉及数据准备、模式搜索、知识评价以及反复的修改求精；
- 非平凡：有一定的智能性和自动性，例如仅仅计算数据总和或平均值都不能算做一个发现过程；
- 有效性：所发现的模式对新的数据仍保持一定的可信度；
- 新颖性：所发现的模式应该是新的、用户未知的或未预料到的；
- 潜在有用性：所发现的模式将来有实际的效用，例如用户可根据发现的模式进行商业决策从而产生一定的经济效益；
- 最终可理解性：要求所发现的模式容易被用户理解。

1.2 数据挖掘的步骤

上节提到，许多人把数据挖掘（准确的说是数据挖掘的算法）视为知识发现过程的一个基本步骤，而知识发现是将未加工的数据转换为有用信息的整个过程，其粗略过程如图 1-1 所示。该过程包括三个转换步骤，从数据准备到数据挖掘结果的解释与评估。

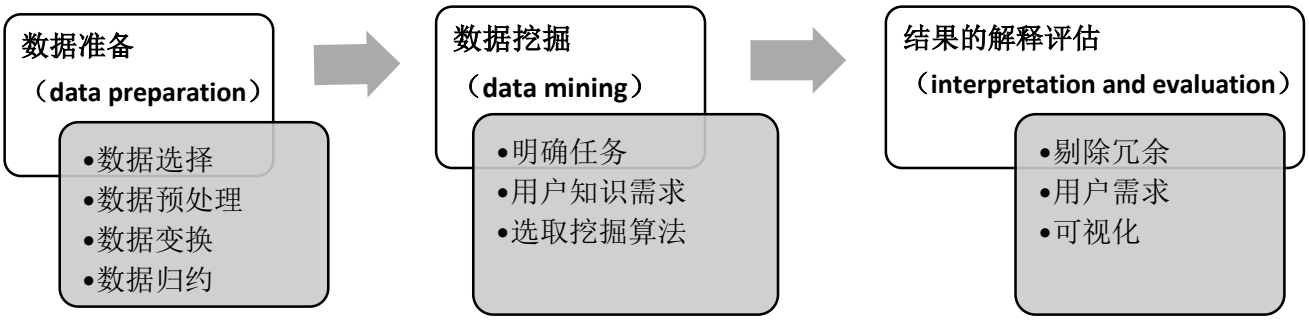

图 1-1 知识发现（KDD）的过程

（1）数据准备。

包括数据选择、数据预处理、数据变换和数据归约等。其中，数据选择是指从数据库中提取与待分析任务相关的数据，即目标数据；数据预处理的目的是将未加工的输入数据转换成适合分析的形式，涉及的步骤包括数据集成（融合来自多个数据源的数据）、数据清理（消除噪声并删除重复的观测值）；数据变换指通过汇总或聚集操作把数据变换和统一成适合挖掘的形式（如连续数据离散化）；数据归约的主要目的在于消减数据维数，根据用户需求对数据的特征进行相应的选择或抽取，从而在尽量不牺牲数据完整性的基础之上得到原始数据的较小表示。

（2）数据挖掘算法的选择。

在选择合适的算法之前，首先需要明确数据挖掘的任务，换而言之，用数据挖掘来解决什么问题。一般来说，数据挖掘的基本任务包括数据总结、分类、聚类、关联规则分析、序列模式发现等。其次需要考虑用户的知识需求，用户是希望得到描述性的知识（用来说明事物的性质、特征和状态以区别和辨别事物），还是预测型知识（由历史数据和当前数据产生的并能推测未来数据趋势的知识）。然后再根据具体的数据集合，选择有效的挖掘算法。

（3）结果的解释评估。

一般来说，并不是所有挖掘出的结果（模式）都是有用的。首先需要经用户或机器评价，剔除冗余或无关的模式；若模式无法满足用户需求，应重新返回某一步重新挖掘，例如重新选择数据、采用新的数据变换方法、设定新的数据挖掘参数，或者换一种挖掘算法；此外，由于挖掘的结果是面向用户的，因此应对挖掘结果进行可视化或者转化为用户易于理解的形式表示。

具体来看，数据库中的知识发现过程如图 1-2 所示，由以下步骤的迭代序列组成：

- 数据清理，消除噪声和删除不一致数据，可能要占全过程 60％以上的工作量；
- 数据集成，多种数据源可组合在一起；
- 数据选择，从数据库中提取与待分析任务相关的数据；
- 数据变换，通过汇总或聚集操作，把数据变换和统一成适合挖掘的形式；
- 数据挖掘，基本步骤，选择适当的算法来找到感兴趣的模式；
- 模式评估，根据某种兴趣度度量，识别代表知识的真正有趣的模式；
- 知识表示，使用可视化和知识表示技术，向用户提供挖掘的知识。

以上观点把数据挖掘看作知识发现过程中的一个步骤，尽管是最重要的一个步骤，但其实在产业界、媒体和研究界，“数据挖掘”通常用来表示整个知识发现的过程。因此，我们采用广义的数据挖掘功能的观点，即前文提到的“知识发现”的定义与步骤同时也是“数据挖掘”的定义和步骤。

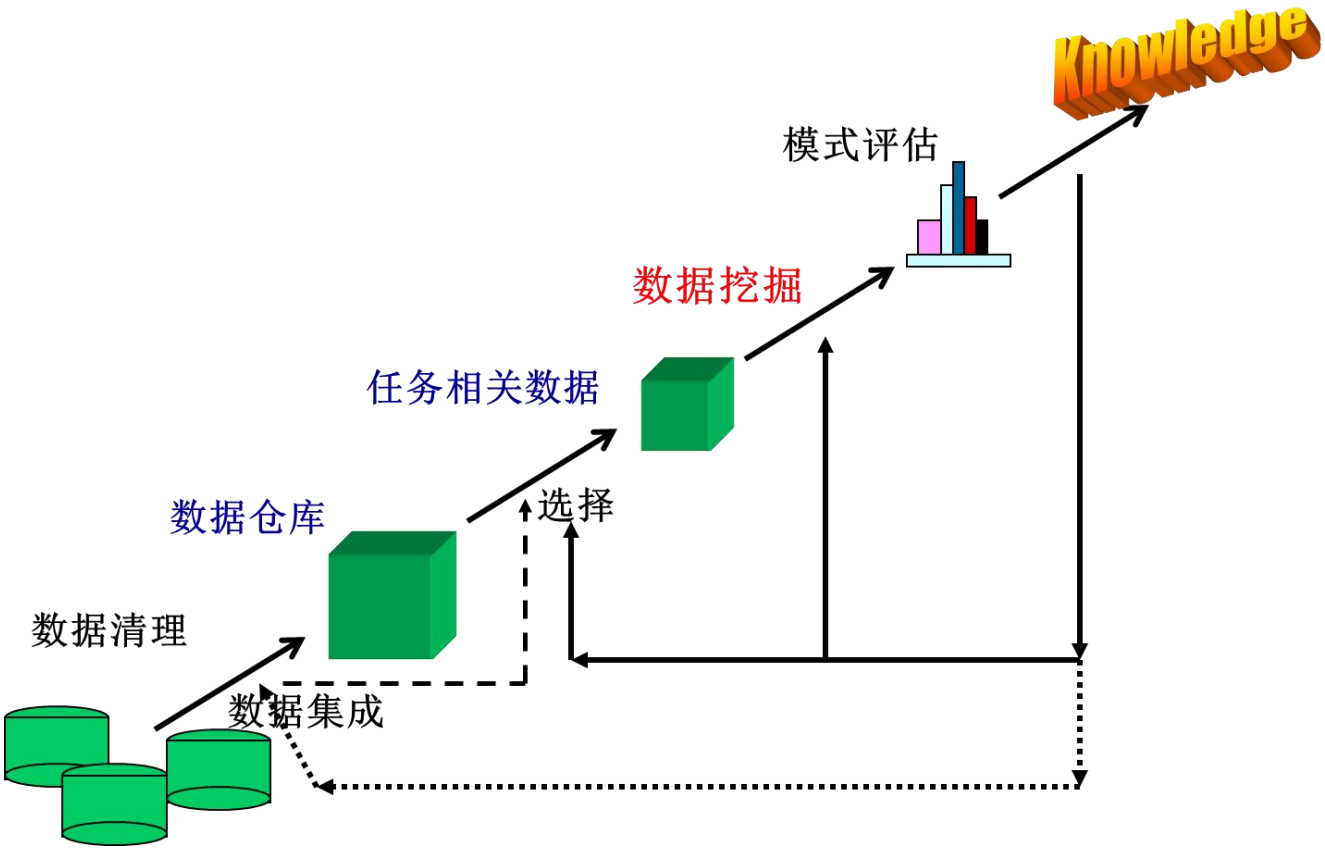

图 1-2 数据库中的知识发现过程

此外，需要注意的是：

（1）影响数据挖掘结果质量的因素有很多，包括采用的算法、数据本身的质量与数量等；

（2）数据挖掘的过程是一个不断反馈的过程。

（3）可视化在数据挖掘过程的各个阶段都扮演着重要角色，如用散点图或直方图等统计可视化技术来显示有关数据，以期对数据有一个初步的了解。

1.3 数据挖掘的主要功能

数据挖掘的主要任务大致可以分为描述性任务（descriptive）和预测性任务（predictive）。描述性任务的目标是概括数据中潜在的联系模式（包括简单汇总、关联、聚类、异常等）；预测性任务是在当前数据上进行归纳和推断，以便做出预测。被预测的属性一般称为目标变量（target variable）或因变量（dependent variable），而用来做预测的属性为说明变量（explanatory variable）或自变量（independent variable）。

1.3.1 预测性

预测建模任务主要包括分类（预测离散型的目标变量）和回归分析（预测连续型的目标变量）。例如，预测一个 Web 用户是否会在网上书店买书是分类任务，因为该目标变量是二值的，而预测某股票的未来价格则是回归任务，因为价格具有连续值属性。两项任务目标都是训练一个模型，使目标变量预测值与实际值之间的误差达到最小。预测建模可以用来确定顾客对产品促销活动的反应，预测地球生态系统的扰动，或根据检查结果判断病人是否患有某种疾病。

（1）分类

分类（classification）是这样的过程：通过对训练数据集（类标号已知的数据对象）进行分析，找出描述和区分数据类或概念的模型/函数，以便能够使用模型预测类标号未知的对象的类标号。例如自动文档分类系统（Automatic text categorization，ATC），它在给定的分类体系下根据文本的内容用计算机程序来确定文本所属类别。

常见的构建分类器的方法包括决策树、Rocchio 方法、朴素贝叶斯、k-近邻法和支持向量机等。其中，决策树是一种类似于流程图的树结构，其中每个结点代表在一个属性值上的测试，每个分支代表测试的一个结果，而树叶代表类或类分布。图 1-3 分别说明了用 IF-THEN 规则和决策树两种不同形式表示的同一分类模型。

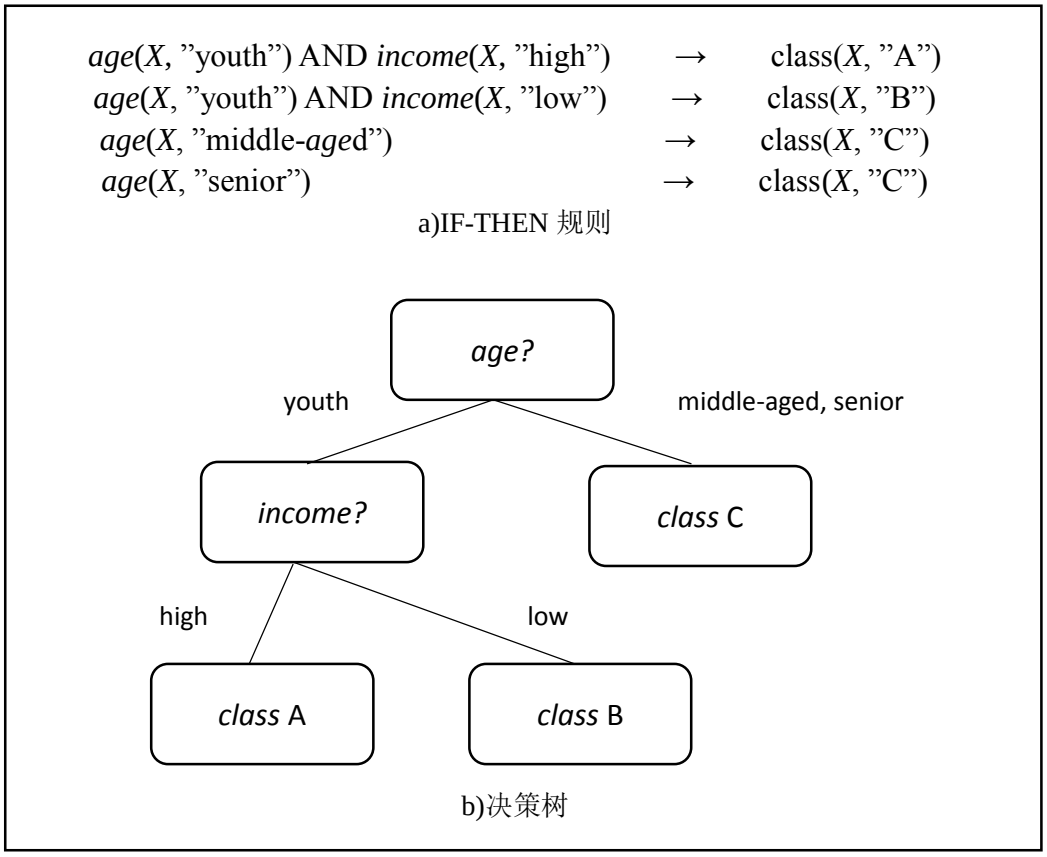

图 1-3 分类模型的 IF-THEN 规则表示和决策树表示

（2）回归分析

回归分析（regression analysis）是一种最常用的数值预测的统计学方法。该方法首先假设一些已知类型的函数（例如线性函数、Logistic 函数等）可以拟合目标数据，然后利用某种误差分析确定一个与目标数据拟合程度最好的函数。

1.3.2 描述性

（1）概念/类描述：特征化和区分

概念描述（concept description）是指对含有大量数据的数据集合进行概述性的总结并获得简明、准确的描述。例如。这种描述可通过两种方式得到：

- 数据特征化（data characterization），是目标类数据的一般特性或特征的汇总。例如，为研究上一年销售增加 10%的软件产品的特征，可以通过在销售数据库上执行一个 SQL 查询来收集关于这些产品的数据。数据特征化的输出可以用多种形式提供，例如饼图、条形图、曲线、多维数据立方体和包括交叉表在内的多维表。
- 数据区分（data discrimination），是将目标类数据对象的一般特性与一个或多个对比

类对象的一般特性进行比较。例如，用户可能希望将上一年销售增加 10%的软件产品与同一时期销售至少下降 30%的软件产品进行比较。区分结果一般以表格或对比规则形式给出。

例 1.1 数据特征化与数据区分。某电子商务的客户关系经理提出了如下数据挖掘任务："汇总一年内在本网站花费 5000 元以上的顾客特征"。得出的结果可能是顾客的概况，如年龄在 40~50 岁、有工作、有较好的信用等级。这些对顾客简单明了的描述就是数据特征化。该客户经理又提出想比较两组顾客——经常来购买计算机产品的顾客（每月多于 2 次）和不经常购买计算机产品的顾客（每年少于 2 次）的区别所在。挖掘结果可提供这些顾客比较的概况，例如经常购买计算机产品的顾客 80%在 20~30 岁之间，受过大学教育；而不经常购买这种产品的顾客 60%或者年龄太大或者太年轻，没有大学学位。这种对两个或以上数据汇集进行比较的结果就是数据区分。

（2）关联分析

在概述关联分析之前，有必要先介绍频繁模式（frequent pattern）。正如名称所示，频繁模式是在数据中频繁出现的模式，其具体包括频繁项集、频繁子序列和频繁子结构。频繁项集一般是指频繁地在事务数据集中一起出现的商品的集合，如超市中被许多顾客频繁地一起购买的牛奶和面包；频繁子序列强调商品出现的先后顺序，如顾客倾向于先买手机，再购买耳机，然后再购买充电宝这样的模式就是一个序列模式；子结构可能涉及不同的结构形式（例如图、树、表格等），可以与项集或子序列结合在一起。

关联分析（association analysis）用来发现大量数据中项集之间有趣的关联。典型案例是"购物篮问题"。假设在某商场中有大量的商品（牛奶、面包等），用户将想购买的商品放入自己的购物篮中，我们可以通过发现顾客放入购物篮中的不同商品之间的联系，分析顾客的购买习惯。比如，哪些物品经常被顾客购买（即求频繁项集）？ 同一次购买中，哪些商品经常会被一起购买，购买过程中是否存在一定的购买时间顺序（即求频繁子序列）？

例 1.2 关联规则分析。假设你是某超市的市场部经理，你想知道哪些商品经常一起被购买。从超市的事务数据库中挖掘出来这样一条规则：

$$buys(\mathrm{X}, \text{“}computer\text{”}) \rightarrow buys(\mathrm{X}, \text{“}software\text{”})[support=1\%, confidence=50\%]$$

其中，X 是变量，代表顾客；50%是置信度，表示如果一位顾客购买计算机，则购买软件的可能性是 50%；1%是支持度，代表分析的所有事务中，有 1%的事务显示计算机与软件一起被购买。这个关联规则涉及一个重复的属性（即 buys），这种仅包含单个属性的关联规则又称为单维关联规则（single-dimensional association rule）。

假设你同时具有和用户信息相关的关系数据库，并挖掘出了如下形式的规则：

$$age(\mathrm{X},\text{“}20\ldots29\text{”})\cap income(\mathrm{X},\text{“}4k\ldots9k\text{”})\rightarrow buys(\mathrm{X},\text{“}laptop\text{”})[support=2\%, confidence=60\%]$$

该规则表示，在所有被研究的超市顾客中，符合年龄是 20~29 岁、月收入为 4000~9000，并且在超市购买了笔记本电脑的顾客占 2%，这个年龄和收入组合的顾客购买笔记本电脑的概率为 60%。在这里，该规则涉及多个属性（即 age、income 和 buys），因此可被称为多维关联规则（multi-dimensional association rule）。

关联规则分析的应用较为广泛，例如商品货架设计，设计出更加适合客户的购物路径；货存安排，实现超市的零库存管理；以及用户分类，为用户提供个性化的服务等。

（3）聚类分析

聚类（clustering）是对数据对象进行划分的一种过程。与分类不同的是，它所划分的类是未知的，故此是一个"无指导的学习"（unsupervised learning）过程。在进行聚类时，不需要提供训练数据，聚类算法更倾向于数据的自然划分。例如聚类在文本处理领域中的应用—

—文本聚类（Text clustering）。其原理在于，将文本集合分组成多个类或簇，使得在同一个簇中的文本内容具有较高的相似度，而不同簇中的文本内容差别较大。

例 1.3 聚类分析。我们可以在某超市的顾客位置数据上进行聚类分析，识别顾客的同类子群。图 1-4 显示了某城市内该超市顾客居住位置的二维图，数据点的三个簇是显而易见的。

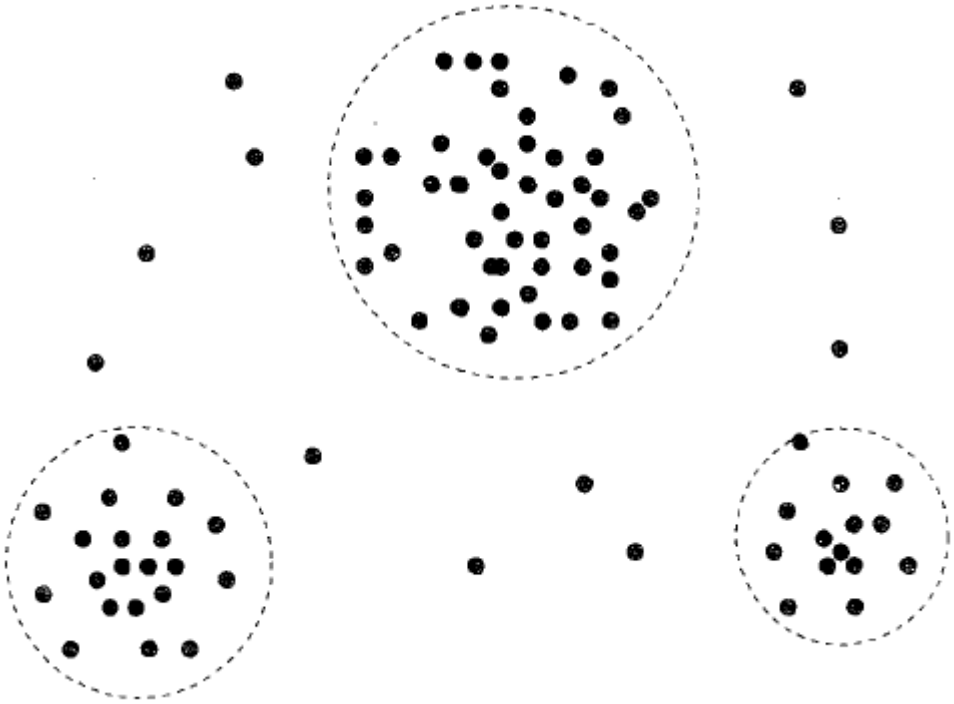

图 1-4 某城市内该超市顾客居住位置图

（4）孤立点分析

数据集中可能包含一些数据对象，它们与数据的一般行为或模型不一致，这些数据对象就是孤立点（outlier）。在大部分数据挖掘过程中，这些孤立点被当做噪声或异常数据而被剔除。然而在一些应用中，孤立点数据更为有趣，例如银行诈骗、洗黑钱、恐怖行为等。

孤立点有多种可能的分析方法，可以假定一个数据分布或概率模型，使用统计检验来检测孤立点；或者使用距离度量，将远离任何簇的对象视为孤立点；亦或不使用统计或距离度量，基于密度的方法也可以识别区域中的孤立点，等。

例 1.4 孤立点分析。信用卡公司可记录每个持卡人所做的交易，同时也记录信用限度、年龄、年薪和地址等个人信息。由于与合法交易相比，欺诈行为的数目相对较少，因此异常检测技术可以用来构造用户的合法交易模型。当每接受一个新的交易申请时就与该模型相比，如果新交易的特性与先前所构造的模型很不相同，就认为该交易很有可能是欺诈。

1.4 在何种数据上进行数据挖掘

数据的最基本形式是数据库数据、数据仓库数据和事务数据。随着科学技术的发展，一些高级数据库系统和信息库逐渐出现，例如空间数据库、时间数据库、时间序列数据库、流数据、多媒体数据库、面向对象数据库和对象-关系数据库、异种数据库和历史数据库、文本数据库、万维网(WWW)数据等。本节主要对常见数据形式进行介绍。

1.4.1 数据库数据

数据库系统，也称数据库管理系统（DBMS），由一组内部相关的数据（称作数据库）和一组管理和存取数据的软件程序组成。