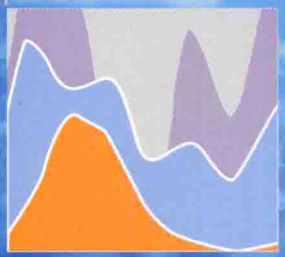
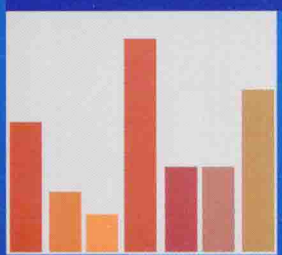
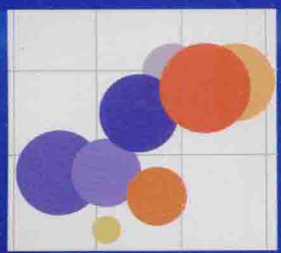


DATA MINING TECHNOLOGY

数据挖掘技术





王小妮 著

北京航空航天大学出版社
BEIHANG UNIVERSITY PRESS

数据挖掘技术

王小妮　著

北京航空航天大学出版社

内 容 简 介

本书是基于数据挖掘经典算法及数据挖掘领域最新研究技术进行数据分析的教材。全书内容包括数据挖掘概述、分类算法、聚类算法、关联规则算法及相应典型算法的算法描述及分析等。对当前数据挖掘的新技术——流数据挖掘技术、高维聚类算法、分布式数据挖掘、物联网数据挖掘进行了详细的介绍。该部分在讲述基本概念及典型算法的基础上配有新研究的算法模型及分析，并有实验数据分析及结果显示。最后对其他数据挖掘新技术，包括业务活动监控挖掘技术、云计算平台架构和数据挖掘方法及思维流程数据挖掘技术进行了描述。

本书可以作为高等院校信息管理、数理统计等专业有关数据挖掘教学的本科生或者研究生的专业课教材，也可以作为各类相关培训班的教材，还可以作为从事数据分析、智能产品软件开发人员的参考书及数据挖掘爱好者的自学用书。

图书在版编目(CIP)数据

数据挖掘技术 / 王小妮著. —北京：北京航空航天大学出版社，2014. 8

ISBN 978-7-5124-1376-4

Ⅰ. ①数… Ⅱ. ①王… Ⅲ. ①数据采集—教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2014)第 199323 号

数据挖掘技术

王小妮 著

责任编辑：史 东 史世芬

*

北京航空航天大学出版社出版发行

北京市海淀区学院路 37 号(邮编 100191) http://www.buaapress.com.cn

发行部电话：(010)82317024 传真：(010)82328026

读者信箱：bhpress@263.net 邮购电话：(010)82316524

北京时代华都印刷有限公司印装 各地书店经销

*

开本：787×960 1/16 印张：15 字数：336 千字

2014 年 8 月第 1 版 2014 年 8 月第 1 次印刷 印数：4 000 册

ISBN 978-7-5124-1376-4 定价：29.00 元

前　言

数据挖掘是从大量的数据中通过算法搜索隐藏于其中的信息的过程。当今社会，随着计算机网络、分布式系统及物联网的发展，产生了海量数据、高维数据及遍布在不同地方的分布式数据。快速、有效地获取有价值的信息并及时捕捉到敏感信息的变化，对于企业决策者迅速作出决策判断、拓展集团客户市场、在新兴市场中立于不败之地起着重要作用。

分类、聚类和关联规则是数据挖掘常用的方法。本书在介绍这些方法的基本概念及分类的基础上，描述了几种经典的数据挖掘算法——ID3、C4.5、K-means、K-medoid、BIRCH、CURE、DBSCAN、OPTICS、Apriori、FP-Growth、CluStream、STREAM 等，并针对现在数据的特点，在流数据、高维数据、分布式数据、物联网数据、业务活动监控、云计算等方面提出了现有算法的缺点及改进方法。高维聚类方法针对维度对数据对象间距离和聚类算法精度的影响进行了降维；分布式数据挖掘在分布式 K-means 聚类算法的基础上，针对资源的约束情况对分布式 K-means 算法进行了改进，完成了 DRA-Kmeans 局部和全局数据挖掘；物联网数据挖掘在 RA-Cluster 聚类算法的基础上，针对资源的约束情况对 RA-Cluster 聚类算法及 AODVjr 路由算法进行了改进，完成了 RA-AODVjr 算法。另外，还搭建了云计算平台及相对应的数据挖掘方法。

本书适合作为“数据挖掘”、“分布式数据挖掘”、“流数据挖掘”和“数据挖掘算法”课程的教材。

本书由王小妮撰写，北京科技大学的高学东、武森、魏桂英老师和郝媛、陈学昌、谷淑娟、陈敏、刘燕驰、白尘等博士给予指导和帮助。该书作者从事数据挖掘方面的教学和研究工作多年，先后发表了多篇数据挖掘方面的 EI 期刊、EI 会议及中文核心期刊论文，主持并参与了多项数据挖掘方面的学校及市教委课题。

在本书的编写过程中，得到北京航空航天大学出版社、北京信息科技大学理学院及北京科技大学经济管理学院的大力支持，在此表示衷心感谢！本书由北京市教委科技计划面上项目（KM201110772018）支持编写。

由于编者水平有限，错误及不当之处在所难免，敬请读者批评指正。

作　者

北京信息科技大学

2014 年 3 月

目　录

第1章 数据挖掘概述

针对商业、工业、信息检索和金融等各种应用所产生的巨大数据集而进行的算法开发，是数据挖掘研究的主要动力。在商业中使用数据挖掘，可以在当今全球化市场竞争中获得明显的优势。比如：零售业使用数据挖掘技术来分析顾客的购买模式，邮购商利用这种技术来选择和定位市场，电信业用其尽快出台网络报警分析和预测，信用卡业用其检测欺诈行为。此外，电子商务的日益增长也产生了大量的在线数据，这些数据急需成熟而复杂的数据挖掘技术。

1.1 数据挖掘的概念

1.1.1 KDD 与数据挖掘

1. KDD 概念

KDD(Knowledge Discovery in Database，数据库知识发现)一词于 1989 年首次出现在美国底特律市举行的第 11 届国际人工智能联合学术会议(International Joint Conference on Artificial Intelligence，IJCAI)上，其本义为“从数据中发现隐含的、先前不知道的、潜在有用的信息的非平凡过程”。1993 年，IEEE 的 Knowledge and Data Engineering 率先出版了 KDD 专刊，随后各类 KDD 会议、研讨会纷纷出现。在 1996 年出版的总结该领域进展的权威论文集《知识发现与数据挖掘研究进展》中，Fayyad 等人对 KDD 和数据挖掘加以区分：KDD 是从数据中辨别有效的、新颖的、潜在有用的、最终可理解的模式的过程；数据挖掘是 KDD 中通过特定的算法在可接受的计算效率限制内生成特定模式的一个步骤。数据挖掘利用某些特定的知识发现算法，在数据中搜索发现隐含在这些数据中的知识。正是由于数据挖掘在数据库知识发现中的重要地位，很多文献对数据挖掘与数据库知识发现不加区别，统称为数据挖掘。KDD 是一个包括数据选择(Data Selection)、数据预处理(Data Preprocessing)、数据变换(Data Transformation)、数据挖掘(Data Mining)、模式评估(Pattern Evaluation)等步骤，最终得到知识的全过程，如图 1.1 所示；而数据挖掘只是其中的一个关键步骤。

KDD 过程的 5 个部分的作用：

- 数据选择：从数据库中提取与分析人物相关的数据。
- 数据预处理：在主要的处理以前对数据进行的一些处理。

- 数据变换：将数据从一种表现形式变为另一种表现形式的过程。
- 数据挖掘：从大量的数据中，通过算法搜索隐藏于其中信息的过程。
- 模式评估：根据某种兴趣度度量，从挖掘结果中识别表示知识的真正有趣的模式。

图 1.1　KDD 过程

2. 数据挖掘概念

数据挖掘(Data Mining，DM)源于 KDD。第一届知识发现和数据挖掘国际学术会议于 1995 年在加拿大召开，由于与会者把数据库中的“数据”比喻成矿床，“数据挖掘”一词很快就流行开来，自此“数据挖掘”一词被广泛使用。数据挖掘是从大量的数据中，提取人们事先不知道的、有价值的信息和知识的过程。这些数据可能是大量的、不完全的、有噪声的、模糊的、随机的、动态的实际数据；信息和知识包括研究对象间的关系、模式、类别和发展趋势等方面。也有一些文献把数据挖掘叫做数据抽取、数据考古学、数据捕捞。

数据挖掘是一门新兴的边缘学科，涉及的学科领域和方法很多，汇集了来自数据库技术、机器学习、模式识别、人工智能以及管理信息系统等各学科的成果。多学科的相互交融和相互促进，使得数据挖掘这一学科得以蓬勃发展，而且已经初具规模。目前，数据挖掘技术及知识发现被认为是数据库和人工智能领域中研究、开发和应用最活跃的分支之一，是计算机科学界的研究热点。随着全球一体化进程的推进、信息技术的迅速发展和广泛应用，越来越多的企业认识到，要实现组织目标、提高组织效率、提升竞争力，要求企业的业务流程更加柔软和灵活，这也使业务流程中的决策问题被人们所关注。决策的效率和效果直接影响整个业务流程最终目标的实现。数据挖掘技术的引入，极大地改变了业务流程的决策环境，为解决业务流程中的决策问题提供了新的思路。国外许多公司，如通用电器公司、IBM 等，非常重视数据挖掘技术的开发应用，已经提出了基于数据挖掘的商业智能解决方案，相关软件也开始销售。

数据挖掘的对象有很多，如数据仓库(Data Warehouse)、文本、多媒体、WEB 网页等，其中应用最多的是数据仓库。数据仓库和数据挖掘是数据库研究、开发和应用最活跃的分支之一，也是决策支持系统(Decision Support System，DSS)的关键因素。数据仓库是一个支持管理决策过程的、面向主题的、随时间而变的数据集合，它是集成的，也是稳定的。数据挖掘是采用人工智能的方法对数据库或数据仓库中的数据进行分析、获取知识的过程。它们的结合能更好地为企业或有关部门不同范围的决策分析提供有力的依据。

按照驱动的方法，通常把数据挖掘分为自主数据挖掘、数据驱动挖掘、查询数据挖掘以及交互式数据挖掘。如果按用户的活动角度，大体可分三类：模式识别、预测建模和分析评价。数据挖掘的方法有很多，如 Han 等人把概念层次引入数据挖掘，从而使面向属性归纳(Attribute-Oriented Induction，AOI)成为最有效的数据挖掘技术；Michalski 等人提出关联规则及挖

掘算法等。

1.1.2 数据挖掘过程

数据挖掘过程(Data Mining Process)一般可以分为3个阶段,包括数据准备(Data Preparation)、模式发现(Pattern Discovery)与挖掘结果(Mining Result)。

(1)数据准备阶段用于为后续的模式发现提供高质量的输入数据。主要包括数据净化(Data Cleaning)、数据集成(Data Integration)、数据变换(Data Transformation)和数据归约(Data Reduction)。数据净化是清除数据源中不正确、不完整、不一致或其他方面不符合要求的数据;数据集成是将多个数据源的数据进行统一的存储;数据变换是对数据进行转换,使其满足分析需求;数据归约是通过消减数据量或降低数据维数来提高挖掘算法的效率和质量。数据准备也就是数据预处理,是数据挖掘的瓶颈问题之一。

(2)模式发现阶段是数据挖掘过程的核心阶段。首要工作是确定挖掘的任务,然后根据挖掘的任务选择合适的挖掘算法,例如关联规则(Association Rule)、聚类(Cluster)、分类(Classification)等。通过对历史数据的分析,结合用户需求、数据特点等因素全面考虑后,得到供决策使用的各种模式与规则,从该任务的众多算法中选择合适算法进行实际的挖掘操作,得出挖掘结果,即相应的模式。这是数据挖掘研究中最核心、难度最大的领域。

(3)挖掘结果阶段关注于规则和模式的可视化(Visualization)表示,即如何将挖掘出来的模式与规则以一种直观、容易理解的方式呈现给用户。数据挖掘得到的模式可能出现不理想甚至不满足用户要求的情况,因此需要对挖掘结果进行评估。对于无关模式或模式中存在的冗余,将其删除;对于不满足要求的模式,重新选择数据,重新进行数据准备和数据挖掘工作,直到符合用户需求。最后得到的数据挖掘结果应解释成为用户可以理解的形式,例如对其进行可视化操作,使其一目了然。

一般来说,数据挖掘技术应用过程分为问题描述、分析主题确定、分析任务确定、数据准备、数据分析、方案提出、方案评估和方案实施8个阶段,如图1.2所示。

在问题描述阶段,专家和数据挖掘应用人员将问题分解为条件、约束、当前状态和最终目标。专家和数据挖掘应用人员根据问题分解结果对问题进行分析。通过抽象问题的具体条件、约束和目标,识别问题本质,对问题进行分类。成功描述问题之后,根据问题的本质描述和抽象分类,专家和数据分析人员确定数据挖掘的分析问题,然后根据分析问题,进一步确定详细的分析任务。数据准备阶段的主要任务是数据的预处理工作。原始数据集中存在大量不完整的、还有可能是噪声的、甚至是错误的数据。直接使用原始数据进行数据挖掘,数据挖掘结果的质量难以保证。数据预处理通过数据的净化和填补等操作,清除错误数据,填补缺失数据,保证数据质量,提高数据挖掘的效果。通过数据集成、变换和归约的方式转变数据形式、集成多格式数据、降低数据规模,保证数据挖掘的效率。数据准备阶段工作量大、耗时长,是数据挖掘最基础的工作之一。数据分析阶段目的在于为问题解决方案的提出提供数据支持。数据

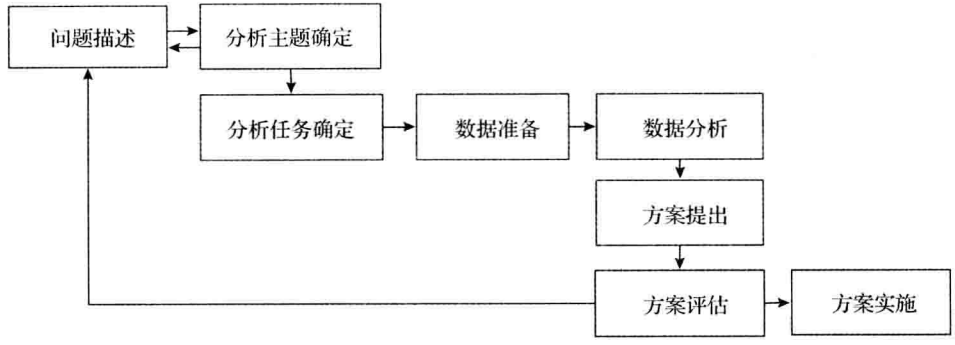


图 1.2　数据挖掘技术应用过程

分析阶段首先确定数据挖掘的任务。数据挖掘任务确定后，根据不同的挖掘任务和实际的环境条件选取合适的挖掘算法，进行实际的挖掘操作，发现可支持问题解决方案的模式、规则或趋势等。方案提出后，需要根据问题的描述对方案的可行性和整体效果做出评估。如果方案符合问题描述中的约束，任务完成度高，方案可行性强，那么方案就被实施；如果方案与问题描述中的条件或约束相悖，任务完成度不符合要求，可行性存在疑问，那么方案就被搁置，新的方案设计过程将启动，问题将被重新分析。

1.1.3　数据挖掘任务

数据挖掘的目标是从数据库中发现隐含的、有意义的知识，主要有以下 6 类基本任务：

(1) 概念描述(Concept Description)：数据库中通常存放大量的细节数据；然而，用户通常希望以简洁的描述形式观察汇总的数据集。要分析一个数据库，获取其中隐藏的信息，往往需要将算法运行于数据库的每一个子集上。所以，如何快速、准确地搜索并标识出相应的数据库子集并将其装入内存，也是一个重要的步骤。这种数据描述可以提供一类数据的概貌，或将它与对比类相区别。此外，用户希望方便、灵活地以不同的粒度和从不同的角度描述数据集。这种描述性数据挖掘称为概念描述。

(2) 关联分析(Association Analysis)：关联分析就是从大量数据中发现项集(Item Set)之间有趣的关联。随着大量数据不停地收集和存储，许多业界人士对于从其数据库中挖掘关联规则越来越感兴趣。从大量商务事务记录中发现有趣的关联关系，可以有助于许多商务决策的制定。

(3) 分类和预测(Classification and Prediction)：分类和预测是两种数据分析形式，可以用于提取描述重要数据类的模型或预测数据未来的趋势。分类和预测的应用十分广泛，例如，可以建立一个分类模型，对银行的贷款客户进行分类，以降低贷款的风险；也可以通过建立分类模型，对工厂的机器运转情况进行分类，用来预测机器故障的发生。

(4) 聚类分析(Cluster Analysis):根据最大化类内相似性(Similarity)、最小化类间相似性的原则进行聚类,使得在同一个类中的对象具有很高的相似性,而与其他类中的对象很不相似。聚类形成的每个类可以看作一个对象类,由它可以导出规则。聚类也便于将观察到的内容组织成分层结构,把类似的事件组织在一起。

(5) 孤立点分析(Outlier Analysis):数据库中可能包含一些数据对象,它们与数据的一般行为或模式不一致。这些数据对象就是孤立点。许多数据挖掘算法试图使孤立点的影响最小化,或者排除它们;但在一些应用中,孤立点本身可能是非常重要的信息。例如在欺诈探测中,孤立点可能预示着欺诈行为。

(6) 演变分析(Evolution Analysis):数据演变分析描述行为随时间变化的规律和趋势,并对其建模(Modeling)。可以从股票交易数据中挖掘出整个股票市场和特定公司的股票演变规律,以帮助预测股票市场的未来走向,帮助对股票投资做出决策。

数据挖掘任务的完成,一般都可以分为两个阶段,第一阶段是数据预处理,为数据挖掘准备数据;另一个阶段是运行挖掘算法。其中,数据预处理是数据挖掘的瓶颈问题之一。要分析一个数据库,获取其中隐藏的信息,往往需要将算法运行于数据库的每一个子集上。所以,如何快速、准确地搜索并标识出相应的数据库子集并将其装入内存,也是一个重要的步骤。

关于如何组织数据已经有许多研究,如采用算法 B^+ 树、哈希索引技术等。

1.2 数据挖掘的发展历程

计算机的应用发展大致可归结为 3 个阶段:数值计算(Numerical Calculation),数据处理(Data Processing)和知识处理(Knowledge Processing)。数值计算属于算法研究,用 FORTRAN、PASCAL、C 语言等实现数值计算。数据处理是对大量数据的处理,用数据库语言对信息进行收集、传递、存储、加工、维护和使用。知识处理是从大量积累的历史数据中提取有用的信息,这就是知识发现和数据挖掘的任务。

韩家炜(Han Jiawei)教授在《数据挖掘:概念与技术》一书中介绍过数据挖掘一词的来源。在科研界,最初一直沿用"数据库中的知识发现",即 KDD(Knowledge Discovery in Database)。在第一届 KDD 国际会议中,委员会曾经展开讨论,是继续沿用 KDD,还是改名为 Data Mining(数据挖掘)。最后大家决定投票表决,采纳票数多的一方的选择。投票结果颇有戏剧性,一共 14 名委员,其中 7 位投票赞成 KDD,另 7 位赞成 Data Mining。最后一位元老提出"数据挖掘这个术语过于含糊,做科研应该要有知识",于是在科研界便继续沿用 KDD 这个术语。而在商用领域,因为"数据库中的知识发现"显得过于冗长,就普遍采用了更加通俗简单的术语——"数据挖掘"。

到 1993 年,美国电气电子工程师学会(IEEE)的《知识与数据工程》(*Knowledge and Data Engineering*)会刊出版了 KDD 技术专刊,发表的论文和摘要体现了当时 KDD 的最新研究

成果和动态。随着来自各个领域的研究人员和应用开发者不断增多，1995 年在加拿大蒙特利尔召开了首届 KDD 国际学术年会，会上把数据挖掘技术分为工程领域的数据挖掘与科研领域的知识发现。此后，此类会议每年召开一次，数量和规模逐渐扩大，从专题研讨会一直发展到国际学术大会，并成为当前计算机领域的研究方向和研究热点。目前对 KDD 的研究主要围绕理论、技术和应用这三个方面展开。1997 年，首届蒙特利尔 KDD 国际学术大会召开。两年后，PAKDD 学术会议(Pacific-Asia Conference on KDD)在亚太地区顺利召开，这标志着亚太地区数据挖掘研究进入发展时期。PAKDD 会议每年召开一次，其中，新加坡第十届 PAKDD 会议除了进行数据挖掘学术研究外，还与新加坡统计协会(SIS)、新加坡模式识别和机器智能协会(PREMIA)共同组织了一场基于解决电信运营商问题的数据挖掘竞赛。其内容为“如何区分移动通讯网客户中使用第二代(2G)和第三代(3G)服务的用户”，旨在明确目前 2G 网络用户中哪些使用者具有巨大的潜在可能性转移到使用移动运营商的 3G 移动网络和服务上。与 KDD 国际学术会议(ACMSIGKDD International Conference on Knowledge Discovery and Data Mining)或 ECML/PKDD 学术会议(European Conference on Machine Learning & European Conference on Principles and Practice of Knowledge Discovery in Databases)定期举办竞赛模式不同，新加坡 PAKDD 会议是继 2000 年第四届京都 PAKDD 会议后，第二次举办类似的比赛。2001—2007 年共 7 年时间中，PAKDD 会议依次由香港、台北、首尔、悉尼、河内、新加坡和南京主办。

由于数据挖掘技术在各领域被广泛应用，其软件市场需求量也变得很大，因此，包括国际知名公司在内的软件公司纷纷加入数据挖掘工具研发的行列中来。根据 National Center for Data Mining at UIC(University of Illinois at Chicago)的 R. Grossman 观点，数据挖掘软件的发展经历了 4 个时代：

第一代：数据挖掘软件支持少数几个用于商业系统数据挖掘算法，这些算法用于数据向量挖掘。Salford Systems 公司早期的 CART 系统就属于这种系统。新加坡国立大学研制的 CBA 是基于关联规则的分类算法，能从关系数据或者交易数据中挖掘关联规则，利用关联规则进行分类和预测。

第二代：数据挖掘软件系统与数据库管理系统(DBMS)集成，支持数据库和数据仓库，具有高性能的接口和较高的可扩展性。这些软件能够挖掘大数据集以及更复杂的数据集和高维数据。但其只注重模型的生成，典型代表有 DBMiner 和 SAS Enterprise Miner。

第三代：数据挖掘系统的特点是和预言模型系统之间能够实现无缝的集成，使得由数据挖掘软件产生的模型的变化能够及时反映到语言模型系统中。由数据挖掘软件产生的预言模型能够自动地被操作型系统吸收，从而与操作型系统中的语言模型相联合提供决策支持的功能。它能够挖掘网络环境下的分布式和高度异质的数据，并且能够有效地和操作型系统集成。其缺点是不能支持移动环境。这一代数据挖掘系统关键的技术之一，是提供对建立在异质系统上的多个预言模型以及管理这些预言模型的元数据提供第一级别的支持。SPSS Clemen-

tine 就是属于这一代的产品。

第四代：数据挖掘软件能够挖掘嵌入式系统、移动系统和普遍存在的计算设备产生的各种类型的数据。2001—2006 年，马里兰巴尔的摩州立大学的 Kargupta 正在研制 CAREER 数据挖掘项目，其目的是开发挖掘分布式和异质数据的第四代数据挖掘系统。

1.3 数据挖掘的分类

确定数据挖掘的任务并选择挖掘算法是数据挖掘的核心工作，针对同一个挖掘任务又存在多种挖掘算法。目前数据挖掘算法按功能主要分为如下几类：

1）关联规则算法

大量数据项之间有时会有一定的关联性，关联规则算法就是用来发现这种关联的。关联规则是指搜索业务系统中的所有细节或事务，从中寻找重复出现概率很高的模式。用于关联规则的主要对象是事务型数据库，其中每个事物被定义为一系列相关数据项，要求找出所有能把一组事件或数据项与另一组事件或数据项联系起来的规则。而关联分析则是从给定的数据集寻找发现频繁出现的项集模式。比如，寻找数据子集之间的关联关系，或者某些数据与其他数据之间的派生关系等。关联规则 $X \Rightarrow Y$ 的意思是“数据库中满足条件 X 的记录也一定满足条件 Y”。此类算法中最有影响力的是 Agrawal 等人提出的 Apriori 算法。该算法的特点是，频繁项中 K 项集是频繁的性质，则其所有 $K-1$ 子集都是频繁的性质。其他常用的关联规则算法有 FP-Growth、H-mine 和 OP 算法等。

2）分类算法

分类算法是利用一个分类函数或者分类模型把数据库中的数据项映射到给定的类别中的某一个，通过对训练样本的分析处理，发现指定的某一商品类或事件是否属于某一特定的数据子集的规则。分类是数据挖掘中非常重要的一项任务，分类的目的是利用一个分类函数或分类模型把数据库中的数据项映射到给定类别中的某一个，通过对训练样本集的分析处理，发现指定的某一商品类或事件是否属于某一特定数据子集的规则。在分类发现中，样本个体或数据对象的类别标号是已知的，根据从已知的样本中发现的规则对非样本数据进行分类。分类只是发现的一个基本任务，它对输入的数据进行分析并利用数据中出现的特征为每一个类别构造一个较为精确的描述和模型，即分类器，然后按分类器再对新的数据集进行分类预测。通常构造分类器需要有训练样本数据集作为输入。训练集由一定数量的例子组成，每个例子具有多个属性或特征。大家经常见到并且使用的分类算法主要包括：决策树算法、贝叶斯分类、粗糙集方法、神经网络、朴素贝叶斯、支持向量机、K 紧邻算法、基于案例的推理和遗传算法等。

3）聚类算法

数据聚类是用于发现在数据库中未知的数据类。这种数据类划分的依据是“物以类聚”，即考察个体或数据对象间的相似性，满足相似性条件的个体或数据对象划分在一组内，不满足