南京航空航天大学
研究生系列精品教材

数据挖掘

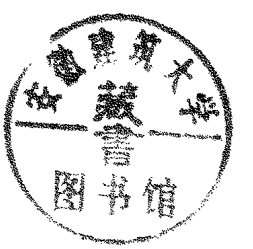
张道强　李　静　蔡昕烨　编著

科学出版社

南京航空航天大学研究生系列精品教材

数 据 挖 掘

张道强　李　静　蔡昕烨　编著



科 学 出 版 社

北　京

内 容 简 介

本书较全面地介绍了数据挖掘的基本理论、算法及应用。首先介绍数据挖掘的基本概念，随后重点讲述关联规则、分类、聚类等模式的挖掘技术并介绍相关的经典算法，同时注重数据挖掘技术的应用实例讲解，包括多模态脑影像挖掘、脑网络分析及其在生物信息学和软件工程中的应用。最后，对近年来发展迅猛的领域，如使用进化计算作为主要方法的数据挖掘技术也用了一定篇幅讲述其基本内容。

本书结构简明，内容丰富，可作为从事数据分析和挖掘研究的一部入门书籍。本书适用于计算机类专业的本科生或研究生，也可作为数据挖掘、机器智能等领域的科技人员和高校师生以及相关专业工程技术人员的参考书。

图书在版编目(CIP)数据

数据挖掘/张道强，李静，蔡昕烨编著. —北京：科学出版社，2018.6
南京航空航天大学研究生系列精品教材
ISBN 978-7-03-057390-2

Ⅰ. ①数… Ⅱ. ①张…②李… ③蔡… Ⅲ. ①数据采集-研究生-教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 095864 号

责任编辑：潘斯斯 张丽花 / 责任校对：郭瑞芝
责任印制：吴兆东 / 封面设计：迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京九州逐驰传媒文化有限公司 印刷
科学出版社发行 各地新华书店经销
*
2018 年 6 月第 一 版 开本：787×1092 1/16
2018 年 6 月第一次印刷 印张：9 1/4
字数：204 000

定价：59.00 元
（如有印装质量问题，我社负责调换）

前　言

随着计算机硬件资源成本的持续下降，软件开发技术的不断进步，基于不同领域的大数据研究与应用性研发工作正在如火如荼地开展。如今大数据成为热门话题，“数据就是财富”的观念为众人所熟知。然而，随着数据量的剧增，信息过量就成为人们不得不面对的问题。如何才能不被信息的汪洋大海所淹没，从中发现及时有用的知识，提高信息的利用率呢？正是基于对此类问题的思考，“数据挖掘”作为大数据挖掘、分析与处理的一个强有力的工具应运而生。它帮助我们从“数据矿山”中找到蕴藏的“知识金块”，以便将其在包括商务管理、生产控制、市场分析、工程设计和科学探索等各种应用中加以使用。

本书在内容安排上遵循先理论后实际应用的基本原则，首先通过讲解数据挖掘相关理论使得读者掌握必备的基础知识，然后通过几个具体的专题应用使得读者对数据挖掘有更加直观和全面的认识。

本书主要内容安排如下：第 1 章概述数据挖掘的基本含义和应用。第 2～5 章讲解数据挖掘的相关理论，包括数据类型和属性、数据质量和数据的预处理、关联规则相关的定义和概念、Apriori 算法和 FP-Growth 算法、分类相关的概念及经典的分类方法、聚类分析相关的概念和算法。第 6～9 章讲述一些数据挖掘的实例，包括多模态脑影像挖掘、脑网络分析、数据挖掘在生物信息学中的应用、软件数据挖掘。第 10 章主要介绍使用进化计算作为主要方法的数据挖掘技术。

本书由张道强、李静和蔡昕烨等编著。具体责任分工如下：张道强负责编写第 1 章、第 6～9 章；李静负责编写第 2～5 章；蔡昕烨负责编写第 10 章；另外，程波、接标、刘明霞、郝小可、祖辰、邵伟、光俊叶、林华锋、梁大川、叶婷婷、杜俊强、屠黎阳、孙亮、张丹、路子祥、李蝉秀等也部分参与了上述各章内容的编写工作。在本书编写过程中，作者力求精益求精，其顺利成文得益于很多朋友、同事的帮助，感谢他们的贡献；感谢国内外的同行，你们在网络上发表了众多优秀的文章，作者从中学习到很多知识。

由于作者水平和时间的限制，书中疏漏和不足之处在所难免，敬请广大读者批评指正。

作　者

2018 年 2 月

目　录

第1章 绪　论

数据挖掘(Data Mining)一般也被译为资料探勘、数据采矿。顾名思义，数据挖掘一般是指从大量的数据中通过算法挖掘出隐藏于其中的信息的过程，是数据库知识发现中的一个重要步骤。

近年来，随着海量数据的产生，我们需要从这些数量庞大的数据中获取有用的信息和知识，以便将其在包括商务管理、生产控制、市场分析、工程设计和科学探索等各种应用中加以使用，这种需求是迫切的也是具有挑战性的。然而数据挖掘正好可以通过统计、在线分析处理、情报检索、机器学习、专家系统(依靠过去的经验法则)和模式识别等诸多方法来解决上述需求问题。此外，数据挖掘是一个多学科的交叉性研究领域，有着十分丰富的内涵。本章概述数据挖掘，并简单列举书中所涵盖的关键性主题。首先介绍数据挖掘中一些需要具备的基础知识。

1.1 什么是数据挖掘

数据挖掘是采用数学的、统计的、人工智能和神经网络等领域的方法，从大量数据中提取出隐藏于其中的知识和信息。它是一个将未加工的数据转换为有用的信息的过程，这就好比从未被开采的矿山中提炼和挖掘出对人类有用的资源。该过程需要一定的转换步骤，包括定义问题、收集数据、数据预处理、生成模型、结果可视化与验证以及模型更新，如图 1-1 所示。

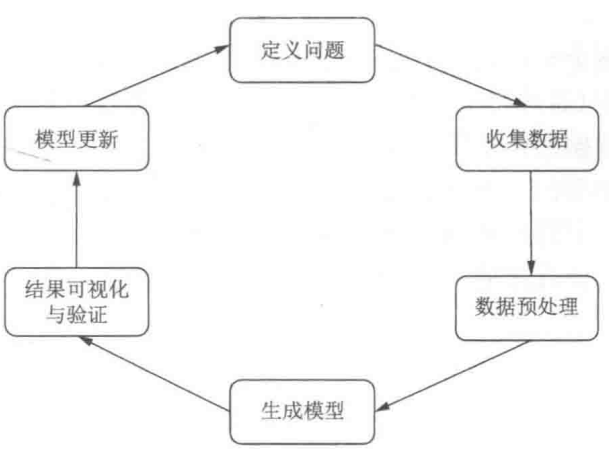


图 1-1　数据挖掘的基本过程

1989 年 8 月在美国底特律召开的第十一届国际人工智能联合会议的专题讨论会上，首次提出了数据库中的知识发现(Knowledge Discovery in Database，KDD)。KDD 是数据库、机器学习、统计学、模式识别、神经网络、信息检索等众多学科技术的集成，它

应用了这些领域中不同的理论和方法，并形成了独立的研究方向。1995 年在加拿大蒙特利尔召开的第一届知识发现和数据挖掘国际学术会议上首次提出数据挖掘这一学科名称。自 1995 年美国计算机年会开始，数据挖掘被看成 KDD 的一个重要过程。近年来，数据挖掘引起了整个社会的广泛关注，其主要原因是在这个大数据时代，它可以把大量数据转换成有用的信息，在商业中带来了巨大的价值。例如，银行可以通过数据挖掘对个人或者企业的信用提供一个客观、准确的评估，从而降低信贷业务的风险。利用数据挖掘技术还可以进行客户分析，把一个庞大的消费群体划分为一个个细分的客户群体，发掘同一客户群体之间的相似之处，如背景资料、盈利能力、消费水平等，针对不同客户群体制定不同的营销策略，从而获取更大的利益。在零售业中，数据挖掘也有着不可替代的巨大作用。各大超市可以通过每日营业数据调查顾客的最大需求，安排货物摆放位置，从而扩大市场。总体来说，利用数据挖掘技术来支持商业决策是一种基于数据分析的科学决策方式，所以这一技术在实际应用中将会得到更加广泛的认可与推广。

1.2　数据挖掘的任务

通常，数据挖掘任务分为以下几类。

1. 关联规则

关联规则挖掘是数据挖掘领域的一个很重要的课题，顾名思义，它是从数据背后发现大量项集间可能存在的有趣的关联或相互联系。关联规则挖掘(Association Rule Mining)由 Agrawal 等提出，最早提出时主要针对购物篮分析问题，其目的是从事务数据库中发现顾客购买的不同商品之间的联系规则。

2. 分类

分类是一种数据分析形式，可从预定的数据集或概念集中提取重要数据的类别信息或建立数据分类模型(通常称作分类器)。该模型能把数据库中的数据记录映射到某一具体的类别，并预测数据未来的发展趋势，这是数据挖掘领域一个重要的研究方向。决策树分类方法是一种典型的分类方法，通常包括特征选择、决策树的生成和决策树的修剪。决策树分类方法的主要思想来源于 Quinlan 提出的 ID3 算法和 C4.5 算法以及 Breiman 等提出的 CART 算法。除此之外，还有贝叶斯分类、神经网络、支持向量机等经典的分类方法。

3. 回归

回归是一种数据分析形式，也是数据挖掘中一个重要的研究方向。回归与分类的区别是，分类的输出是一个离散的结果，而回归是建立一个连续的函数值模型，函数的输出是连续的。回归使用的是回归分析，包括线性回归和非线性回归。经典算法有最小二乘估计、矩阵分解、局部加权线性回归等。

4. 聚类

聚类分析是数据挖掘的一种重要手段和工具，其主要任务是根据数据对象的属性，将数据对象划分为不同的簇，从而使得同一簇中对象之间相似度较高而不同簇之间相似度较低，进而可标识出人们感兴趣的分布或模式。通过聚类分析可识别出对象空间中的稠密与稀疏区域，从而发现全局分布模式与数据属性之间的关联。聚类属于无监督学习，即不需要类别信息。经典的聚类方法包括：k 均值聚类算法、基于划分的聚类算法、基于层次的聚类算法、基于网格的聚类算法、基于密度的聚类算法与基于模型的聚类算法。

与同类书籍相比，本书的一大特色是不但介绍了数据挖掘的基本算法，还介绍了这些算法在脑疾病诊断以及生物信息学中的具体应用，以下将对其进行简要介绍。

1.3 数据挖掘在脑疾病诊断以及生物信息学中的应用

人脑的结构和功能极其复杂，理解大脑的运转机制是 21 世纪人类面临的最大挑战之一。为此世界各国均投入了大量的人力和物力进行研究。其中以老年痴呆症(又称阿尔茨海默病，Alzheimer's Disease，AD)为代表的脑疾病诊断是脑科学研究的热点。AD 的特点是发病年龄较早，病程进展缓慢。因此准确诊断 AD，尤其是它的早期阶段，对尽早治疗和推迟疾病的恶化是非常重要的。

数据挖掘在脑科学领域可以起到重要的作用，主要包括多模态脑影像挖掘和脑网络分析。脑图像作为现代医学的一种重要工具，可以客观地揭示与脑疾病相关的脑结构和脑功能的变化。借助脑影像数据可以进一步构建脑网络，而脑网络能从脑连接层面刻画大脑功能或者结构的交互，脑网络分析已经成为近年来脑科学研究中的一个热点。

生物信息学是一门新兴的交叉学科，是采用计算机技术和信息论方法研究蛋白质及核苷酸等各种生物信息采集、存储、传递、检索、分析和解读的学科。在本书中我们将介绍数据挖掘技术在基因影像关联分析以及蛋白质亚细胞定位这两个生物信息学课题中的应用。

近年来，基因与疾病关联分析作为生物医学这一交叉领域的重要研究课题而备受关注。与此同时，随着神经影像技术的发展，基因关联分析衍生出了“基因-影像关联”(Imaging Genetics)这一新的研究领域，其目标是检测并发现影响人脑结构或功能的基因；相对于传统的风险基因与诊断量表得出关联分析，借助多模态脑影像定量性状作为基因与诊断的中间介质的辅助参考，可以方便我们对复杂的致病生物机制有更加清晰和完整的认识。

与此同时，随着测序技术的不断发展，蛋白质数据呈现爆炸式增长趋势，如何找到这些蛋白质所处的亚细胞位置并分析它们的动态变化是一个难题，利用蛋白质图像预测其所处的亚细胞位置是解决该问题的一个新的方向。本书讲解了一种新的基于图像的蛋白质亚细胞定位算法，利用与细胞器层次结构相关的先验信息预测蛋白质所处的亚细胞位置。

1.4　数据挖掘在软件设计和应用领域的应用

尽管软件开发是系统化并且严格遵循约束的，但是在开发过程中仍然有众多的因素难以掌握，这导致软件缺陷在软件开发过程中难以避免。然而软件缺陷对应用软件在生产和生活中造成的危害是极大甚至是无法估量的，因此，我们需要一些技术来检测软件的缺陷。软件缺陷预测，即利用软件度量数据、已发现的缺陷历史数据以及数据挖掘技术来预测软件系统中缺陷的分布、类型或数目。目前已存在很多比较成熟的软件度量方法，它们能很好地指导我们对软件特征进行提取。基于提取的特征，软件领域的分类预测可以对软件开发过程中可能存在的缺陷作出预测，从而合理分配测试资源、辅助开发过程、提高程序的健壮性及可维护性等。因此，利用数据挖掘技术对软件缺陷进行预测是一项非常有意义的研究工作。

1.5　基于进化计算的数据挖掘技术

进化计算是借助(生物界)自然进化和演变的启示，根据其规律设计出的各种计算方法的总称。同时它又是一个快速发展的多学科交叉领域：数学、生物学、物理学、化学、心理学、神经科学和计算机科学等学科的规律都可能成为进化计算的基础和思想来源。

在最近十几年，进化计算无论在理论分析还是在工业应用上都取得了显著的发展。它涉及的领域已由最初的生物计算发展到各种类型的自然计算算法和技术，这其中包括进化计算、神经计算、生态计算、社会和经济计算等。作为一种智能优化算法，进化计算已经被广泛应用，包括商务管理、生产控制、市场分析、工程设计和科学探索等。在本书中我们会简要介绍基于进化计算的数据挖掘技术。

1.6　本书的内容与组织

为了更好地理解数据挖掘技术如何在各种类型中的应用，我们从第 2 章开始详细介绍数据挖掘涉及的概念、方法以及应用。第 2 章讨论数据类型和属性、数据质量和数据的预处理。第 3 章主要讲述关联规则，包括关联规则的定义和相关概念、Apriori 算法和 FP-Growth 算法，并进一步拓展了相关知识。第 4 章详细描述分类的相关概念以及一些经典的分类方法，包括决策树分类、基于规则的分类、贝叶斯分类、神经网络分类器、支持向量机分类器、最近邻分类器以及集成学习，并在本章最后介绍了分类器的评估和度量。第 5 章讲述聚类分析的相关概念和算法，主要内容包括基于划分的聚类算法、局域层次的聚类算法、基于网格与密度的聚类算法、基于模型的聚类等方法，并在最后讲述如何评价一个聚类算法的有效性。第 6～9 章讲述一些数据挖掘的实例，包括多模态脑影像挖掘、脑网络分析、数据挖掘在生物信息学中的应用、软件数据挖掘。第 10 章

主要介绍将进化计算作为主要方法的数据挖掘技术。

数据挖掘是一个年轻的跨学科领域，随着计算机计算能力的发展和业务复杂性的提高，数据类型越来越多且越来越复杂，数据挖掘必然会发挥越来越大的作用。然而数据挖掘所包含的内容范围很大，我们很难通过本书来全面地介绍。本书只将一些我们所熟悉的或者与相关的研究方向和主题的内容与读者分享。

第2章 数 据 准 备

数据挖掘的目的是从大量数据中发现其隐含的模式，但目前的数据挖掘工作大多着眼于对数据挖掘算法的探讨，而对数据预处理的研究比较滞后。一些成熟的算法对输入数据的完整性、一致性及属性之间的相关性都有一定的要求。直接从现实世界中收集数据往往具有难以预知的多样性、不确定性及复杂性，导致原始输入数据集的不规范性，因此，无法直接满足数据挖掘算法的要求。原始数据主要存在以下几方面的问题[1,2]。

(1) 杂乱性。原始数据的来源有可能是不同的应用系统，而由于各应用系统的数据缺乏统一标准的定义，数据结构往往存在一定的差异，造成数据不一致，因此，不能直接被现有算法使用。

(2) 不完整性。不完整性是指数据记录中一些属性值为空、无法确定或必要数据的缺失。数据采集系统的缺陷或一些人为因素都可能造成数据不完整。有些数据的不完整性对结果并没有多大的影响，可以排除，但绝大部分情况下，数据的不完整性会影响后续知识发现的准确性。

(3) 重复性。系统实际使用过程中，数据库可能出现对同一事物有两种或两种以上完全相同的描述，造成数据重复和信息冗余。这是几乎所有应用系统中都普遍存在的实际问题。

(4) 噪声数据。噪声数据是指原始数据中人为因素或数据收集设备故障等原因产生的一些不正确属性值或与期望值不对应的离群值。

以上问题不仅大大降低了算法的执行效率，还会造成挖掘结果的偏差。因此，对无法满足算法需求的原始数据进行有效的分析和预处理是数据挖掘任务的一项不可缺少的必要工作。

2.1 数 据

数据是能够被识别和处理的符号集合，包含数字、字母、图像、影视信息等。数据对象是数据所描述的事物，代表一个实体。不同的数据对象有其特有的属性，属性之间的差异是区分不同数据对象的依据。

2.1.1 数据集类型

随着数据挖掘技术的发展与成熟，越来越多的数据类型被用于分析。目前常用的数据类型主要包括记录型数据、图形化数据、序列型数据和非记录型数据[3]。

(1) 记录型数据。目前许多数据挖掘任务都假定数据集是记录(数据对象)的集合，每个记录包含固定的数据字段(属性)集。记录之间或字段之间没有明显的联系，记录型数据通常存放在平台文件或数据库中。记录型数据主要包括事务数据、数据矩阵与

稀疏矩阵。

(2)图形化数据。图形化数据是一种可以方便而有效地表示数据特征的形式，其可视化的显示效果给人一种直观的感觉。这种类型的数据主要包括带有对象之间联系的数据及具有图形对象的数据。其中，带有对象之间联系的数据指的是对象之间的联系常常携带重要信息。因此，当用图形表示数据时，一般把数据对象映射到图的节点，而对象之间的联系用链的方向、权值等性质表示。而具有图形对象的数据是指一个数据对象与其子对象之间具有结构化的联系。例如，化合物的数据对象可以用图形表示，其中节点是原子，节点之间的链是化学键。

(3)序列型数据。对于某些数据对象而言，属性所包含的值与时间或空间具有紧密联系，由此形成了序列型数据类型数据集，主要包括时序数据、序列数据、时间序列数据及空间数据。时序数据是记录数据的一种扩充，它的每条记录都包含一个与之相关联的时间标记，所以也称时间数据。序列数据是包含各个实体的序列的一个数据集合，常见的例子如词或字母的序列。时间序列数据是一种特殊的时序数据，它的每个记录都是一个时间序列，如股票价格的时间序列信息。空间数据除了包含一些常见数据类型的属性之外，还包含空间属性，如位置或区域信息。空间数据的典型例子是不同地理位置的气象数据(如降水量、气温、气压)。

(4)非记录型数据。记录型数据具有一定的结构组织，如数据库、数据表等，易于被计算机理解与查找，而非记录型数据没有一个事先定义好的数据模型，也没有符合一定的结构组织格式便于计算机辨识与处理。这类数据主要包括文本数据、网络数据和多媒体数据(如声音、图像、视频数据等)。相对于记录型数据存储在数据库中且用二维表结构来表达实现的行数据而言，不方便用数据库二维表来表示的数据即为非记录型数据。要对非记录型数据进行挖掘，需要事先将其转化为记录型数据，再对其进行数据挖掘。

2.1.2 数据属性及类型

属性，在有些情况下也称为维(Dimension)、特征(Feature)和变量(Variable)，它们之间可以相互使用。在数据挖掘和数据库领域一般使用术语“属性”。属性并非数字或符号，而是为了讨论和精细地分析对象的特性，给它们赋予了一定数字或符号，同时为了明确定义这一过程，引入了测量标度的概念[3]。

1. 数据属性

属性(Attribute)是一个数据字段，表示数据对象的性质或特性，它因对象而异或随时间而变化[3]。例如，眼球颜色因人而异，而物体的温度随时间而变。其中，眼球颜色是一种符号属性，具有少量可能的值（棕色、黑色、蓝色、绿色……），而温度是数值属性，可以取无穷多个值。

测量标度(Measurement Scale)是数值或符号值与属性之间相关联的规则(函数)。测量标度按类型进行划分的情况如图 2-1 所示。测量过程是使用测量标度将一个值与某一特定对象的特定属性相关联。

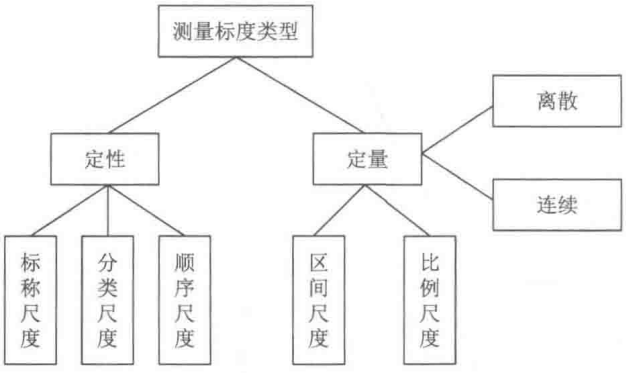


图 2-1　测量标度类型

2. 属性类型

属性类型是由该属性可能具有的值的类型所决定的，属性类型一般可分为标称、序数、区间、比率四种[3]。标称和序数统称分类或定性属性。区间和比率属性统称定量或数值属性。定量属性用数表示，并且具有数的大部分性质。定量属性所能赋予的值可以是离散值或连续值。

标称属性(Nominal Attribute)的值是一些符号或事物的名称。每个值代表某种类别、编码或状态。因此，标称属性又可称为分类的属性。它不要求具有一定的顺序，如人的肤色(黑、白、黄等)、婚姻状况(未婚、已婚)等，这些都属于标称属性。

序数属性(Ordinal Attribute)是一种有序属性，其可能的值之间具有前后关联的性质，但是彼此之间的差异是未知的，如成绩(A、B、C、D)、职位(助教、讲师、副教授及教授)等。

区间标度(Interval Scaled)属性用相等的单位尺度度量。区间属性的值有序，可以为正、零或负。同时，区间标度属性可以比较和定量评估值之间的差。例如，人的年龄 20 岁比 15 岁大 5 岁。

比率标度(Ratio Scaled)属性是具有一个固定零点的数值属性。假如用比率标度作为度量，那么可以说一个属性值是另一个的多少倍。同样，这些属性值是有序的且可以计算值之间的差，也能计算均值、中位数和众数。

2.1.3　数据相似性与相异性

相似性和相异性在聚类、最近邻分类和异常检测等数据挖掘任务中都有使用，是重要的概念。在进行数据挖掘前，先将原始数据变换到一个相似性或相异性的空间。邻近度是数据挖掘中用来表示相似性或相异性的度量，两个对象之间的邻近度是度量对象属性之间的邻近函数。相似度(Similarity)是表示两个对象之间相似程度的一种数值度量[3]。两个对象相似度越高，表示两个对象越相似。一般地，相似度都是非负的且在 0(不相似)和 1 (完全相似)之间取值。类似地，相异度(Dissimilarity)表示两个对象之间的差异程度。两个对象相似度越高，它们的相异度就越低[3]。

相似度与相异度之间可以相互转换，也可以变换到一个特定区间。通常，邻近度度

量(特别是相似度)被定义为或变换到区间[0,1]中的值。一般来说，相似度到[0,1]区间的变换可用 $s'=(s-\min\_s)/(\max\_s-\min\_s)$ 来实现，其中 $\max\_s$ 和 $\min\_s$ 分别是相似度的最大值和最小值，s 和 s' 分别是相似度的原始值和变换后的值。例如，存在一个对象之间的相似度度量，且在 1(完全不相似)和 10(完全相似)之间变化，则可以用公式 $s'=(s-1)/9$ 将它变换到[0,1]区间。

如果相似度(相异度)落在[0,1]区间，则相异度(相似度)可以定义为 $d=1-s$ (或 $s=1-d$)。另一种简单的方法是定义相似度为负的相异度。例如，相异度 0、1、10 和 100 可以分别变换成相似度 0、-1、-10 和-100。负变换产生的相似度结果不要求局限于[0,1]区间，经常使用的变换有

$$s=d+1,\quad s=\mathrm{e}^{-d}\quad 或\quad s=1-\frac{d-\min\_d}{\max\_d-\min\_d}$$

对象之间的邻近度可以由它们单个属性的邻近度的组合来确定，因此，首先考虑具有一个标称属性描述的对象的情况。由于标称属性是分类属性，所以只具有对象的相异性信息，只能说两个对象是否有相同的值。在这种情况下，如果属性值相同，则其相似度为 1，否则为 0。同样，相异度可以用相反的方法确定，若属性值相同，则相异度为 0，否则为 1。

下面介绍涉及多个属性的对象之间的邻近性度量：数据对象之间的相异度和数据对象之间的相似度[3]。

1. 数据对象之间的相异度

距离度量可用来衡量数据对象在空间上的差异，距离越大表明数据对象间的相异度也越大。

假设 $d(x,y)$ 是两个点 x 和 y 之间的距离，则距离满足以下性质[4]。

(1)非负性。对于所有 x 和 y，$d(x,y)\geqslant 0$；当且仅当 $x=y$ 时，$d(x,y)=0$。

(2)对称性。对于所有 x 和 y，$d(x,y)=d(y,x)$。

(3)三角不等式。对于所有 x、y 和 z，$d(x,z)\leqslant d(x,y)+d(y,z)$。

常用的距离度量方法[3]有欧几里得距离、闵可夫斯基距离、曼哈顿距离、切比雪夫距离。假设需要计算 $x=\{x_1,x_2,x_3,\cdots,x_n\}$ 与 $y=\{y_1,y_2,y_3,\cdots,y_n\}$ 间的各种距离。

(1)欧几里得距离。欧几里得距离(Euclidean Distance)可以用来衡量多维空间中各个数据对象点之间的绝对距离，是最常用的一种距离度量，简称欧氏距离。公式定义如下

$$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2} \tag{2-1-1}$$

其中，n 是维数；x_k 和 y_k 分别是 x 和 y 的第 k 个属性值(分量)。

(2)闵可夫斯基距离。闵可夫斯基距离(Minkowski Distance)是欧氏距离的一种推广形式，可对多个距离度量公式进行概括性表述，简称闵氏距离，其公式如下

$$d(x,y)=\left(\sum_{k=1}^{n}|x_k-y_k|^r\right)^{\frac{1}{r}} \tag{2-1-2}$$