数据挖掘理论与实例

SHUJU WAJUE LILUN YU SHILI

王 晖 王 琪 何 琼/著

经济科学出版社
Economic Science Press

数据挖掘理论与实例

王晖 王琪 何琼 著

经济科学出版社

图书在版编目（CIP）数据

数据挖掘理论与实例/王晖、王琪、何琼著. —北京：经济科学出版社，2012.7

ISBN 978-7-5141-1814-8

Ⅰ.①数… Ⅱ.①王…②王…③何…… Ⅲ.①数据采集 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2012）第 070330 号

责任编辑：初少磊 孙 偲

责任校对：徐领柱

版式设计：代小卫

责任印制：李 鹏

数据挖掘理论与实例

王 晖 王 琪 何 琼 著

经济科学出版社出版、发行 新华书店经销

社址：北京市海淀区阜成路甲 28 号 邮编：100142

总编部电话：88191217 发行部电话：88191537

网址：www.esp.com.cn

电子邮件：esp@esp.com.cn

北京中科印刷有限公司印装

880×1230 32 开 7.5 印张 200000 字

2012 年 7 月第 1 版 2012 年 7 月第 1 次印刷

ISBN 978-7-5141-1814-8 定价：25.00 元

（图书出现印装问题，本社负责调换。电话：88191502）

前　言

数据挖掘是一种技术，它将传统的数据分析方法与处理大量数据的复杂算法相结合。《数据挖掘理论与实例》的主要目标是，通过不同领域的应用案例来说明数据挖掘在实际应用中的具体操作方法。将数据库管理系统 MySQL 和统计软件 R 结合，利用数据挖掘技术帮助陷入海量数据中的组织和个人提取有用的信息。

本书的主要内容包括理论和实例两部分：第一、第二章介绍数据挖掘、数据仓库和数据挖掘的常用技术等基本理论；实例部分是以作者的两个研究课题为基础，第三、第四章介绍呼叫中心数据仓库的构建和数据挖掘模型与实现（MS SQL Server 2000、决策树）；第五、第六章介绍数据挖掘在 QFII 投资理念与持股偏好研究中的应用（Rsoftware、MySQL、多元逐步线性回归、因子分析、聚类分析等）。

本书可作为管理者、信息分析人员、数据统计人员等的参考资料，也可作为高等院校数据挖掘的教材和参考书。全书共分六章：何琼撰写第一、第二章；王琪撰写第三、第四章；王晖撰写第五、第六章。

本书得到北京市属高等学校科技创新平台、人才强教计划、北京市重点建设学科、北京知识管理研究基地项目资助。

本书出版之际，感谢经济科学出版社编辑初少磊同志为本书所付出的辛勤劳动，感谢为本书写作提供资料和意见的同志。由于作者水平有限，本书错漏之处，恳请广大师生及其他读者朋友对本教材给予批评指正。

目　录

第一章

绪　论

随着计算机技术、信息技术和网络技术的发展，越来越多的人发现，我们正在被数据所淹没，从海量的数据中吸取有用的数据越来越困难，我们对数据的掌握、了解和处理的速度远远赶不上数据升级的速度。在此大背景之下，数据挖掘逐渐成为未来信息处理的骨干之一，它以一种全新的概念改变着人类利用数据的方式。在本章中我们主要介绍一些数据挖掘的基本概念、应用现状以及数据挖掘技术的发展。

第一节　什么是数据挖掘

数据挖掘（Data Mining，DM）就是从大量的数据中挖掘出有用的信息。一般人们认为它是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的以及最终可理解的模式的过程。它也可理解为是在一些事实或观察数据的集合中找模式的决策支持过程。数据挖掘是一门涉及面很广的交叉学科，包括机器学习、数理统计、神经网络、数据库、模式识别、粗糙集、模糊数学等相关技术。

数据挖掘的主要步骤是：数据准备、数据挖掘、结果的解释评估。数据挖掘的主要功能有：分类或预测模型、数据总结、数据聚类、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。

数据挖掘具有以下几个基本特点：（1）数据挖掘的数据量巨大；（2）数据挖掘在不确定的查询需求情况下为用户寻找他可能感兴趣的信息；（3）数据挖掘分析大量的原始数据，挖掘内在的有价值的知识，用于描述过去的趋势和预测未来的趋势；（4）数据挖掘应对数据量的快速增长及时快速地做出响应，提供决策支持信息。

数据挖掘与统计学关系密切，主要表现在：

（1）它们的目标相同，都是发现数据中的结构。

（2）数据挖掘的出现为统计学提供了一个新的应用领域，也给统计学的理论研究提出了新的课题，它将推动统计学的发展。

（3）数据挖掘并不是统计学的分支，因为它还应用了其他领域的思想、工具、方法。

（4）统计作为数据挖掘中的一种成熟、应用广泛的技术，在数据挖掘中地位显著。

第二节　基本数据挖掘任务

基本的数据挖掘任务包括：分类、聚类、回归、时间序列分析、偏差分析等。

一、分类

分类就是把一些新的数据项映射到给定类别中的某一个类别，如发表一篇文章的时候，就可以自动地把这篇文章划分到某一个文章类别。一般的过程是根据样本数据利用一定的分类算法得到分类规则，新获得的数据就依据该规则进行类别的划分。

分类在数据挖掘中是一项非常重要的任务，有很多用途，如预测，即从历史的样本数据中推算出未来数据的趋向，有一个比较著名的预测的例子就是大豆学习。再如分析用户行为，我们常称之为受众分析，通过这种分类，我们可以得知某一商品的用户群，对销

售工作大有帮助。

分类器的构造方法有统计方法、机器学习方法、神经网络方法等。

二、聚类[①]

聚类分析（Cluster Analysis）又称群分析，是研究分类问题的一种多元统计分析方法，所谓类是指相似元素的集合，是指它最早应用于生物学的分析，近来已广泛应用于各学科。聚类结果体现了数据的分布特征，聚类方法多种多样，针对不同的问题应该采取不同的方法。聚类分析又可以分为系统聚类分析和模糊聚类分析。

由于类与类之间的距离计算方法不同，从而形成了不同的系数聚类方法。一般的有最短距离法、模糊聚类关系。模糊聚类分析是在模糊分类关系的基础上进行分类，用相似性尺度衡量事物之间的亲疏程度来实现分类，而模糊聚类分析的实现就是根据研究对象本身的属性而构造模糊矩阵，在此基础上根据一定的隶属度来确定其分类关系。

三、关联[②]

自然界的事物都是相互关联、互相影响和依存的。统计分析的目的就是探索事物之间、变量之间的关系，从而说明关系的性质、密切程度等。

变量与变量之间的关系有两种：一种是相关关系，另一种是函数关系。函数关系是一种确定的关系，它是指在一个变化的过程中，两个变量按照某一个规律变化的不确定的依存关系，即一个变量的取值不能有另一个变量来唯一确定，变量之间不存在一一对应的确定性关系。函数关系和相关关系虽然是两种不同类型的变量关系，

① 葛新权、王斌：《应用统计》，社会科学文献出版社 2006 年版，第 138 页。
② 杨孝海：《统计学原理》，西南经济大学出版社 2008 年版，第 321～323 页。

但它们之间并没有绝对的界限，在一定条件下是可以相互转化的。

连接分析也称做亲和力分析或关联分析，是指揭示数据直接相互联系的一项数据挖掘任务，而这些关系在数据中没有直接的表示。这项数据挖掘任务的最佳应用例子就是确定关联规则。关联规则是可以识别出特殊类型的数据关联的模型。这些关联通常用于零售业以了解哪些商品频繁的被顾客同时购买。

四、回归[①]

研究现象间在数量上关系的统计分析方法有关联分析和回归分析。关联分析主要分析变量之间关系的密切程度；回归分析是在确定现象相关关系密切程度的基础上，进一步用数学模型模拟它们之间的变动规律性及进行预测的统计方法。

回归首先假设一些已知类型的函数可以拟合目标数据，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。回归分析根据变量之间的关系，可以分为线性回归和非线性回归；按回归方程所涉及变量个数的不同，分为一元回归和多元回归。

五、序列分析

时间序列分析包括一般统计分析（如自相关分析、谱分析等），统计模型的建立与推断，以及关于时间序列的最优预测、控制与滤波等内容。经典的统计分析都假定数据序列具有独立性，而时间序列分析则侧重研究数据序列的互相依赖关系。后者实际上是对离散指标的随机过程的统计分析，所以又可看做是随机过程统计的一个组成部分。例如，记录了某地区第一个月、第二个月、……第 N 个月的降雨量，利用时间序列分析方法，可以对未来各月的雨量进行预报。

---

① 张建同、孙昌言、王世进：《应用统计学》，清华大学出版社 2010 年版，第 192～195 页。

时间序列是按时间顺序的一组数字序列。时间序列分析就是利用这组数列，应用数理统计方法加以处理，以预测未来事物的发展。时间序列分析是定量预测方法之一，它的基本原理是：（1）承认事物发展的延续性。应用过去数据，就能推测事物的发展趋势。（2）考虑到事物发展的随机性。任何事物发展都可能受偶然因素影响，为此要利用统计分析中的加权平均法对历史数据进行处理。该方法简单易行，便于掌握，但准确性差，一般只适用于短期预测。时间序列预测一般反映三种实际变化规律：趋势变化、周期性变化、随机性变化。

时间序列分析是根据系统观测得到的时间序列数据，通过曲线拟合和参数估计来建立数学模型的理论和方法。它一般采用曲线拟合和参数估计方法（如非线性最小二乘法）进行。时间序列分析常用在国民经济宏观控制、区域综合发展规划、企业经营管理、市场潜量预测、气象预报、水文预报、地震前兆预报、农作物病虫灾害预报、环境污染控制、生态平衡、天文学和海洋学等方面。

时间序列主要应用于以下几个方面：（1）系统描述。根据对系统进行观测得到的时间序列数据，用曲线拟合方法对系统进行客观的描述。（2）系统分析。当观测值取自两个以上变量时，可用一个时间序列中的变化去说明另一个时间序列中的变化，从而深入了解给定时间序列产生的机理。（3）预测未来。一般用 ARMA（Auto-Regressive and Moving Average Model）模型拟合时间序列，预测该时间序列未来值。（4）决策和控制。根据时间可调整输入变量使系统发展过程保持在目标值上，即预测到过程要偏离目标时便可进行必要的控制。

六、偏差分析[①]

偏差可以分为语言偏差、搜索偏差和避免过度拟合偏差。

---

① ［新西兰］威滕（Ian H. Witten）：《数据挖掘实用机器学习技术》，机械工业出版社 2006 年版，第 20～22 页。

对于语言偏差来说，最重要的问题是概念描述语言是否具有普遍性，或者它是否在能够被学到的概念上加了约束条件。如果考虑一个包含所有可能的样本集，那么概念就是将这个集合划分为多个子集的分界线。如果概念描述语言允许许多包括逻辑，也就是析取，那么任何一个子集都是能表示的。然而，如果不允许使用逻辑或的原则，一些可能的概念将不可能被表达出来。

搜索偏差。在实际的数据挖掘问题中，存在一些可选的概念描述，它们都与数据相匹配。问题是要根据一些标准从中找出“最佳”的一个。我们使用统计学的术语拟合，意味着寻找一个与数据合理拟合的最佳描述。但是要在整个空间进行搜索并保证所有找到的描述是最好的一个，通常从计算上是不可行的。所以搜索过程是启发式的，并且不能做出有关最终结果是最优的保证。这些偏差的产生创造了空间：不同的搜索引导方式以不同的方式在搜索中产生偏差。

一个更通用的和更高层的查询偏差需要调查查询时是由一般性的数据描述开始，再对它进行提炼；还是由一个特殊的样本出发，然后对它进行推广。前者称为从一般到具体的搜索偏差，后者是从具体到一般的搜索偏差。

避免过度拟合偏差是另一种搜索形式的偏差。但是因为它涉及一个比较特殊的问题，所以我们将它加以区别对待。通常解决这个问题的方法是从一个最简单的概念出发，逐渐将它复杂化：由简到繁的顺序。这是为了拟合简单的概念描述，从而使查询产生偏差。

第三节　数据挖掘的过程

一个完整的数据挖掘包括四个过程：数据准备、数据选择、数据预处理、数据挖掘及模型评价。

一、数据准备

数据准备有四个基本的原则。第一个原则：面向主题。第二个

原则：将数据处理成有意义的。如身份证中相同的数字是没有意义的，我们应该把重点放在不同的数字上。第三个原则：不起坏作用的。如果某个数据项在挖掘过程中具有显著意义，但在处理过程中又无法恢复其原有面貌，应该采取措施让其保持“中庸”，即不干扰其他信息的分析。第四个原则：效率原则。尽管对挖掘型数据库的响应时间没有对操作型数据库的要求高，效率还是必需的。处理的周期要尽可能短，数据挖掘的速度要尽可能快，这一原则将体现在数据分割、数据约简等一些处理过程中。

数据准备的方法有：（1）基于契比雪夫定理的统计学方法。这种方法随机抽取样本数据进行分析，检测速度快，但准确性降低。（2）模式识别的方法。基于数据挖掘和机器学习算法来查找异常数据，主要涉及关联规则。基于距离的聚类方法评测标准为欧几里德距离或者 Edit 距离，用来发现数据集中的重复记录。

二、数据选择

数据选择是数据处理过程中关键的环节，所选择的数据不同，得出的结果就会出现很大的差别。所以在数据选择的过程中要根据挖掘的主题对数据进行选择，不能将所有数据全部交给计算机处理，应该事先对数据进行处理和甄别。

三、数据预处理

数据预处理是数据整理的前期步骤，它是对数据分组前所做的必要处理，内容包括数据的审核、排序等工作。

数据的审核是指对原始数据的审查与核对。按照数据质量保准的要求，对于通过直接调查取得的原始数据，其审核的内容应主要包含以下四个方面。

1. 准确性审核：准确性审核主要从数据的真实性与精确性角度检查资料，其审核的重点是检查调查过程中所发生的误差。准确性审查可以包括下面几个方面：逻辑性审查，它是用于论及理论检查

数据之间有无矛盾；比较审查法，它是在数据之间进行比较方式检查；设置疑问框架审查，一般情况下，数据之间存在一定的取值范围与比例关系，利用这种范围和比例关系可以设置疑问框，从而审查数据是否有疑问。

2. 适用性审查：主要是根据数据的用途，检查数据解释说明问题的程度。具体包括数据与调查主题、与目标总体的界定、与调查项目的揭示等是否匹配。

3. 及时性审查：审查数据的及时性主要是检查数据是否按规定时间报送，如未按规定时间报送，就要检查未按规定时间报送的原因。

4. 一致性：审查数据的一致性主要是检查数据在不同地区或国家、在不同的时间段是否具有可比性。

数据排序时需按照一定规则，如高矮、优劣等次序将数据排列，以便研究者通过浏览数据发现一些明显的特征和趋势，找到解决问题的线索。排序还有助于对数据检查纠错，以及为更新归类或分组等提供方便。无论是数值型数据，还是非数值型数据的排序，都可以方便地使用各种计算机软件来实现，Excel 就具有很强的数据排序功能。

四、数据挖掘及模型评价

对所得到的经过转换的数据进行挖掘。除了完善和选择合适的挖掘算法外，其余一切工作都能自动地完成。

在数据挖掘中，评价模型的主要标准有：基于损失函数的标准、基于统计检验的标准、基于评分函数的标准、贝叶斯标准，以及计算标准。具体的评价方法可以查看参考文献［8］。

第四节　数据仓库与 OLAP 技术概述

随着数据库应用技术的发展，数据仓库、OLAP（Online Ana-

lytical Processing，联机分析处理）技术已成为近年来数据库界研究的热点。数据仓库与 OLAP 技术为企业的分析决策提供了强大的支持，正确及时的决策是企业生存和发展的最重要环节。现在，愈来愈多的企业认识到，要想在竞争中取胜，获得更大的收益，必须利用网络、数据仓库等计算机技术，深层次地挖掘、分析当前和历史的生产业务数据，以及相关的环境数据，自动快速地获取其中有用的决策信息，为企业提供快速、准确和方便的决策支持。

一、什么是数据仓库

数据仓库（Data Warehouse，DW）是一个面向主题的、集成的、时变的和非易失数据集合，支持管理部门的决策过程。数据仓库的构建是一个处理过程，数据仓库是一个从多个数据源收集的信息存储库，存放在一个一致的模式下并且通常驻留在单个站点。数据仓库通过数据清理、数据变换、数据集成、数据装入和定期数据刷新过程来构造。数据仓库系统由数据仓库、数据仓库管理系统、数据仓库工具三个部分组成。在整个系统中，DW 居于核心地位，是信息挖掘的基础；数据仓库管理系统负责管理整个系统的运作；数据仓库工具则是整个系统发挥作用的关键，包含用于完成实际决策问题所需的各种查询检索工具、多维数据的 OLAP 分析工具、数据挖掘 DM 工具等，以实现决策支持的各种要求。①

数据仓库具有以下几个关键特征：

1. 数据仓库是面向主题的。数据仓库围绕一些主题，如顾客、供应商、产品和销售来组织。

2. 数据仓库是集成的。数据仓库的数据来自于分散的操作型数据，即将所需数据从原来的数据中抽取出来，进行加工与集成，统一与综合之后才能进入数据仓库。

---

① 叶得学、韩如冰：《浅谈数据仓库与 OLAP 技术》，载于《甘肃科技纵横》2009 年第 2 期。

3. 数据仓库是不可更新的。数据仓库主要是为决策分析提供数据，所涉及的操作主要是数据的查询。

4. 数据仓库是随时间而变化的。传统的关系数据库系统比较适合处理格式化的数据，能够较好地满足商业商务处理的需求，它在商业领域取得了巨大的成功。

二、多维数据模型

多维数据模型是为了满足用户从多角度多层次进行数据查询和分析的需要而建立起来的基于事实和维的数据库模型，其基本的应用是为了实现 OLAP。数据仓库和 OLAP 工具都基于多维数据模型。在多维数据库中，数据以多维方式组织，经综合汇总后，存放在多位数组中，以提高系统响应速度；在前端展现工具中，用表或图的形式通过维展现度量的值，并提供灵活的分析方式：

1. 超立方结构（Hypercube）：是指用三维或更多的维数来描述一个对象，每个维彼此垂直，数据的测量值发生在维的交叉点上，数据空间的各个部分都有相同的维属性。这种结构可应用在多维数据库和面向关系数据库的 OLAP 系统中，其主要特点是简化终端用户的操作。

2. 多立方结构（Multicube）：该结构是将大的数据结构分成多个多维结构，这些多维结构是大数据维数的子集，面向某一特定应用对维进行分割，即将超立方结构变为子立方结构。它具有很强的灵活性，提高了数据（特别是稀疏数据）的分析效率。多立方结构是存储稀疏矩阵的一个有效方法，并能减少计算量。复杂的系统和预先建立的通用应用倾向于使用多立方结构，以使数据结构能更好地得到调整，满足常用的应用需求。许多产品结合了上述两种结构，它们的数据物理结构是多立方结构，但却利用超立方结构来进行计算，结合了超立方结构的简化性和多立方结构的旋转存储特性。

另外，OLAP 多维数据还提供了多种分析操作，常用的有以下