# 交通领域中的聚类分析方法研究

李桃迎/著

Research on Clustering Algorithms in Traffic Domain

科学出版社

本书由大连市人民政府资助出版

# 交通领域中的聚类分析方法研究

# Research on Clustering Algorithms in Traffic Domain

李桃迎　著

科学出版社
北京

## 内 容 简 介

本书系统而详细地阐述了聚类分析的多种相关方法、技术及具体应用，主要内容包括绪论、复杂多源异构数据整合方法研究、数据预处理技术、常用聚类分析方法、面向混合特征的权熵模糊 C-均值优化方法研究、面向混合属性数据的聚类融合方法研究、基于聚类融合的混合属性数据增量聚类方法研究、聚类分析方法在交通领域中的应用。

本书可作为管理科学与工程、信息科学与技术、计算机应用、应用数学等相关专业高年级本科生和研究生的教材或参考资料，也可帮助相关领域研究人员提升聚类分析的技巧。

**图书在版编目(CIP)数据**

交通领域中的聚类分析方法研究/李桃迎著. —北京：科学出版社，2013
ISBN 978-7-03-039918-2

Ⅰ.①交… Ⅱ.①李… Ⅲ.①聚类分析-分析方法-研究 Ⅳ.①O212.4-34
中国版本图书馆 CIP 数据核字(2014)第 038581 号

责任编辑：李 莉 / 责任校对：赵桂芬
责任印制：阎 磊 / 封面设计：无极书装

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

北京市文林印务有限公司 印刷
科学出版社发行 各地新华书店经销
*
2014 年 3 月第 一 版 开本：720×1000 B5
2014 年 3 月第一次印刷 印张：16
字数：300 000

**定价：62.00 元**
(如有印装质量问题，我社负责调换)

# 作 者 简 介

李桃迎，女，博士，大连海事大学交通运输管理学院副教授，曾作为第二主编撰写《管理信息系统开发教程》、《数据挖掘与聚类分析》。主持辽宁省教育厅一般项目1项，企业委托项目1项；参与国家级、省（部、市）级项目多项；博士论文“交通领域中的聚类分析方法研究”获得2012年辽宁省优秀博士学位论文；获得省部级科技进步奖励5项，发表相关学术论文40余篇。

# 前　言

随着信息化技术的发展，各领域系统中积累的数据越来越多，简单的查询统计功能已经满足不了实际需求，运用数据挖掘方法从现有数据中发现潜在、有意义的规律，获取有价值的知识，为高层管理与辅助决策提供依据已经成为解决问题的关键。

鉴于交通领域各系统的建设时期、开发部门、使用设备、技术发展阶段以及能力水平都存在差异，导致“信息孤岛”现象的出现，降低了工作的效率，妨碍了管理决策。笔者在聚类分析方面潜心研究多年，尤其是近年来，通过参加国家自然科学基金委员会、科学技术部和交通运输部及多个省（市）主办的多项科研课题，深入研究了数据挖掘特别是聚类分析的理论、技术与方法，获得多项科研成果。特别是面向交通运输、物流管理等特色领域，开展基于聚类分析的创新性研究，在提高管理效率和挖掘效率方面取得了良好的社会效益与经济效益。

聚类分析因其应用非常广泛而成为数据挖掘研究的重要子领域，可为探索未知的数据结构提供帮助，并可作为一系列数据分析的起点，因此聚类分析成为本书的重要内容。本书采用逐步演算和流程运行相结合的方式，力争使广大读者通过本书的学习快速掌握聚类分析的理论、技术与方法。

陈燕教授对于本书的撰写、章节安排等提出了很多指导性的意见，孙骏雄、王任远、李鹏辉等同学参与完成部分章节的校对工作。

本书旨在涵盖典型和有代表性的聚类分析算法，但由于聚类分析方法多样，还有许多聚类模型需要进一步探讨。在编写过程中，笔者查阅了国内外大量文献资料，谨向书中提到的和参考文献中列出的学者表示感谢。如果本书中某处内容所参考的文献没有列出，那么在此向所涉及的作者深表歉意。本书获得大连市人民政府（或重点）资助出版，在此表示感谢。

同时，由于时间仓促和笔者能力有限，书中难免存在一些不足之处，敬请广大读者批评指正。

作　者

2014 年 1 月

# 目　录

# 第1章 绪 论

## 1.1 本书的撰写目的及意义

随着信息技术的高速发展，交通领域各行业都加快了信息化建设的步伐，信息化应用水平有了明显的提高，伴随数据库技术的迅速发展，以及信息系统的广泛普及，交通领域现行系统中积累的数据越来越多。大量积累的数据背后隐藏着许多非常重要的信息，人们希望可以对其进行更深层次的分析，以便更好地利用这些数据。数据挖掘促使数据库技术进入了更为高级的阶段，不仅能满足对过去的数据进行查询和遍历的需要，而且能找出各种数据之间潜在的联系和规律，促进信息的传递。

采用数据挖掘技术实现对交通领域各行业现有数据进行挖掘的任务，其前提是实现数据的全面整合，建立数据仓库或数据中心，才能在此基础上建立各相应的数据挖掘模型。但是，就目前交通领域各单位、处室现有的应用系统来说，有的是根据业务需求自行开发研制，有的是由部、省、厅等上级管理部门下发，由于各系统的建设时期、开发部门、使用设备、技术发展阶段以及能力水平都不同，数据分散在不同的应用系统中，无法实现数据的共享和互联互通。交通领域各部门间信息的不通畅导致上下级部门之间、单位内业务处室之间存在“信息孤岛”的现象，数据、信息在单系统内封闭运行，系统化监控机制无法实施，严重妨碍了交通领域管理的高效决策，不仅增加了业务处理的复杂程度，也降低了工作的效率。

有的系统虽然实现了数据的集中管理，但是只提供查询访问功能，大量的数据存储在各系统的数据库中，不能被充分利用。无法预测未来的发展趋势，充分发挥数据、信息的价值，缺乏挖掘数据背后隐藏知识的手段和功能，出现“数据爆炸但知识贫乏”的问题。

聚类分析作为数据挖掘与统计分析研究的一个重要方面，为探索未知的数据结构提供帮助，并能成为一系列数据分析的起点，已被广泛应用于道路交通、港口、航空、航务海事及海运等领域，并形成了系统的方法体系。许多文献已经对聚类分析进行了深入的研究，并结合实际问题提出了许多改进方法。然而综合国内外对聚类分析的研究，同时考虑到交通领域的现状，仍有很多极具挑战性的问题亟待解决：

（1）很多聚类算法存在自身的缺陷和不足，需要对算法本身进行改进和完善；

（2）针对聚类有效性和稳定性方面的研究有待深入，同时缺乏对增量聚类的稳定性、精确性和高效性的研究，对聚类的精度和效率的研究也有待提高；

（3）缺少针对交通领域高层管理方面进行辅助决策的研究，特别是还未建立面向交通领域的增量聚类模型，因此提高交通管理部门的管理效率是一项艰巨的任务。

本书提出的交通领域中的聚类分析方法研究的意义在于：采用数据、信息的集成整合技术实现交通领域各行业系统中数据、信息的有效共享和互联互通；采用数据挖掘方法中的聚类分析方法进行研究，提高算法的精度、效率、稳定性；对现行系统中的数据进行分析、加工，建立面向交通领域的聚类分析模型，从中找出潜在的规律，以提高相关管理部门的管理效率，使管理模式更加科学、合理、高效。

## 1.2 国内外研究现状

近年来，随着信息化建设的发展，交通领域各部门都有了数据库管理系统，并且积累了大量的数据，但是由于各系统的建设时期、开发部门、使用设备、技术发展阶段以及能力水平的不同，纵然系统之间、业务功能之间有交叉，也不能共享数据，导致“信息孤岛”现象的出现，降低了交通领域管理部门的管理水平，因此，当前信息化建设迫切需要实现对现有系统的集成整合，消除“信息孤岛”。

为了满足信息化建设发展的需要，各级管理部门、高等院校、研究机构和软件开发公司都展开了对交通领域中系统数据、信息整合与集成技术的研究。根据研究内容的不同，现有研究主要分成两个方向：一个是针对数据的集成和整合问题；另一个是采用数据挖掘方法挖掘数据中的潜在规律和知识，辅助管理者的决策。

下面将对数据整合、数据挖掘、聚类分析、聚类分析在交通领域中的应用四个方面的研究现状进行详细介绍。

### 1.2.1 数据整合的研究现状

在信息化建设过程中，大规模企业的下属公司或行政管理部门的下属单位往往各自负责局部的信息系统选型、建设和维护，导致“信息孤岛”问题的出现。同时，由于数据来源广泛，数据格式多样，数据主要集中存储在文件系统、数据库和消息队列之中。如何把具有不同业务语义，不同格式的数据整合起来是信息化建设中一项具有挑战性的工作。但是，由于数据源的不同，数据的内容、格式和质量千差万别，数据的准确性、真实性和完整性存在差异，数据共享和数据分析的实施就需要对数据进行整理，另外对数据进行有效的整合也是必不可少的步骤。

数据整合的目的是运用一些技术手段把分布的、异构的系统中的数据按特定的方式组织成一个整体，使用户能有效地对其进行共享、分析。数据整合是在逻辑上、物理上把不同来源、格式、特点的数据有机地整合，从而全面、有效地实现部门、企业之间的数据共享。

鉴于数据集成整合要考虑数据的范围和逻辑关系、完整性和约束性、访问权限、语义冲突等，但是领域、行业之间往往存在差异，目前尚没有统一的集成整合标准，大多采用联邦数据库、数据仓库、中间件等方法，这些技术为解决数据的共享问题提供了不同的方式。联邦数据库（federation）方法的优点在于数据依然保留在原来的存储位置，不必构建一个集中式数据仓库，节省了空间。但其不足之处是查询速度较慢，不适用于频繁查询，且易出现死锁等问题，而且各个数据库系统需要开放访问接口，当集成的系统较多、较大时，为实际开发带来巨大的困难。中间件方法通过统一的全局数据模型访问异构数据库、原有系统和Web资源等。通常中间件处在异构数据源系统（数据层）和应用程序（应用层）之间，向下协调各数据源系统，向上为访问集成整合数据的应用提供相应的数据模式和通用的数据访问接口（API），但是各数据源的应用仍然停留在完成各自的任务，中间件主要为异构数据源提供检索服务，优点是用户可以不必知道数据源的位置、结构及访问方法而实现对该数据源的访问。数据仓库的数据集成整合主要是数据的提取、转换和装载（ETL）过程。数据仓库的优点是可提供简单、方便、频繁、高效的查询，进而可以直接进行数据挖掘、联机分析处理（OLAP）等智能分析。数据仓库可以满足决策的分析需求，对已集中的数据进行深度分析以挖掘历史数据中隐含的规律、趋势。但是数据仓库通常花费的时间较长，而且数据的更新往往不及时，不能满足一些对数据要求实时性的项目（董燕，2009）。三种集成整合方法的特点及适用性，如表1.1所示。

**表1.1 数据集成整合特点及适用性**

| 数据特点及业务需求 | 适用的技术 | 可结合的数据处理技术 |
|---|---|---|
| 异构数据源 | 联邦数据库、数据仓库、中间件 | XML技术、Web Service技术 |
| 异地分布式 | 联邦数据库、中间件 | XML技术、Web Service技术 |
| 大数据量 | 联邦数据库、数据仓库 | 数据库复制技术 |
| 自治性 | 数据仓库、中间件 | XML技术、Web Service技术、Message技术 |
| 实时性 | 中间件 | XML技术、Web Service技术、Message技术 |
| 海量多媒体信息 | | Web Service技术 |
| 综合、全局的决策分析 | 联邦数据库、数据仓库 | Web Service技术 |

注：XML技术又称可扩展标记语言技术；Web Service技术又称Web服务技术；Message技术又称消息技术

资料来源：董燕，2009

由表 1.1 可知，根据不同的数据特点可以选择适用的技术，如果选择的方法不当不仅达不到数据整合的真正目的，反而会影响系统的整体效率。

现在存在很多有关数据集成、整合的研究及应用，如李宝玲（2010）对我国的档案资源整合进行研究，分析现有档案资源整合的理论研究滞后性，指出现有研究对相关领域借鉴研究不够、对特色资源整合研究不够等，并提出了其个人的见解。张晓娟和张洁丽（2009）总结我国信息资源整合研究的特点，分析信息资源整合中存在的问题，如缺乏辩证的态度、缺乏对信息资源整合对象个体差异性与特色的研究、有关信息资源整合标准的研究不足、缺乏相关法律保障的研究、对相关领域及国外已有的成果借鉴不够等，并认为目前相关领域并未形成完善的理论体系。

Yuan 等（2010）针对虚拟数据库系统中的现存算法效率较低的问题，利用 Map Reduce（MR）可以有效地并行处理大数据集的特征，提出 VDB-MR 模型，该模型基于 MR 技术，可以有效地整合异构数据源的数据。尽管结构化查询语言（SQL）操作可以完成分组和连接的功能，但是不适用于属性值在一定的准则下不相同但是相似的情况。为此，Schallehn 等（2004）提出用一种基于相似度的分组和连接算子，扩展的分组算子可以将相似元组划分成组，而扩展的连接算组可以将满足给定相似性条件的元组结合起来。Olga Brazhnik 和 John F. Jones（2007）认为数据处理的最终目的是产生可利用的信息，随着科技的进步，产生的数据越来越多，需要对这些由于目的、商业规则、潜在模型、有效技术等驱动下产生的不同数据源的数据进行集成。参考模型、语义网、标准、本体和其他相关技术促使数据快速地出现多样性，产生的数据的可靠性很大程度上被认为是由数据呈现现实的程度决定的。因此，他们提出一个框架，用于估计数据信息的值，包括数据的维度，衡量实际工作中数据的质量、识别数据源和集成的关键技术、融合模型、统一变化频率的修改、重写数据集等。Juraj Bartok 等为了探测重要的大气现象，描述了数据挖掘气象项目对参数模型和方法的计划分布，特别是雾和低云覆盖的问题。这个项目期望包含分布式大气数据的集成方法，包括运行预测模型、训练模型，然后挖掘数据，以此快速有效地预测随机现象的发生。他们对分布在不同服务器的数据进行集成，对大气探测模型则采用基于统计学和大气模型的理论，并结合知识发现-数据挖掘模型进行构建。

综上所述，现在已经存在很多有关数据集成和整合方法的研究，但是现有研究中缺少对交通领域数据整合的应用研究，多数都是局限于理论研究本身，或者只是简单的统计汇总，而结合具体应用领域进行数据整合的研究将会是一项应用价值极高的研究。

### 1.2.2 数据挖掘的研究现状

数据挖掘是可用于数据分析和理解、揭示数据内部蕴藏知识的技术。随着计

算机技术的发展、信息化水平的提高，数据库中存储的数据规模越来越大，如何从数据库中发现有用的信息，并采用一定的方法和技术挖掘出数据中的潜在规律，辅助管理者进行决策，显得非常重要。

对于数据挖掘（data ming，DM）的定义，至今尚没有形成统一的观点，不同的领域和文献中的概念也存在差异，但是被普遍接受的定义是：从大量的、不完全的、随机的、有噪声的实际应用数据中发现隐含的、规律性的、人们事先未知的但又是潜在有用的并且最终可以被理解的信息和知识的非平凡过程（Han and Kamber，2006）。

众所周知，数据挖掘是一门交叉学科，它以强化应用为主，把人们对数据的应用从简单的查询提升到从数据中挖掘知识，辅助决策。在这种需求的牵引下，汇聚了数据库、人工智能、机器学习、数理统计、可视化、并行计算技术等方面的学者和研究人员，针对数据挖掘这一新兴的领域展开研究。在这一领域中数据库、人工智能和数理统计技术是数据挖掘技术的强大技术支撑。

数据挖掘的主要技术包括：①预测技术；②关联规则技术；③聚类分析技术；④分类分析技术；⑤粗糙集技术；⑥进化计算技术；⑦灰色系统技术；⑧模糊逻辑技术；⑨人工智能与机器学习技术；⑩决策树技术；⑪统计分析方法；⑫知识获取、知识表示、知识推理和知识搜索技术；⑬决策与控制理论；⑭可视化技术；⑮并行计算技术和海量存储技术（陈燕，2010）。

1）预测（forecast）技术

为了科学、详细地了解某企业（某生产部门）的业务发展情况和今后的走势，可采用预测技术对其有利于生产的条件进行科学论证和判断。一般在预测过程中，可以根据目标范围的不同，将其分为宏观预测和微观预测。例如，宏观经济预测是指对整个国民经济或一个地区、一个部门的经济发展前景的预测；而微观经济预测是以单个经济单位的经济活动前景作为考察的对象。按预测期限长短不同，可分为长期预测、中期预测和短期预测。按预测结果的性质不同，可分为定性预测与定量预测，有的时候也采用混合预测方法。

2）关联规则（association rules）技术

数据之间的关联规则指的是在数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性，就称为关联。关联分析的目的是找出数据库中隐藏的关联网。关联规则技术主要应用在从大型数据库中找出潜在的相关属性的知识上。例如，通过调研发现，在大多数的汽车修理部门，在修理汽车的同时，也存在着购买汽车椅垫和其他零部件的可能，如果将这些相关的物品和零部件都放在汽车修理部门中，则会发现三者的效益会同时上升，从数据挖掘的角度来认识此类问题，则认为是关联知识挖掘的问题。目前，利用关联规则技术进行数据挖掘的研究非常盛行，著名的 Apriori 算法属于目前关联规则挖掘

较好的算法模型之一，已经被应用在不同的研究领域中。

3）聚类分析（clustering analysis）技术

聚类分析是根据事物的特征对其进行聚类或分类，通过聚类或分类可以发现其中的规律和模式。聚类或分类以后，样本数据集就转化为类集。同一类的样本数据具有相似的变量值，不同类的样本数据的变量值不具有相似性。

4）分类分析（classification analysis）技术

分类就是找出一个类别的概念特征，用以表示这类数据的整体信息，即该类的内涵描述，进而用这种描述来构造模型，通常用规则模型或决策树模型表示。分类通常利用训练数据集，采用一定的算法来求得分类规则。

5）粗糙集（rough sets）技术

粗糙集技术采用的理论是粗糙集理论，将约简技术应用在不确定数据的范化和数据挖掘中。粗糙集理论是波兰教授 Z. Pawlak 在 1982 年提出的一种智能决策分析理论，它是一种刻画不完整性和不确定性的数学工具，能有效地分析不精确、不一致、不完整等各种不完备的信息，并且能够将其不确定数据分析的结果，即不确定和不精确的知识用已知的知识库来近似刻画和处理。利用粗糙集理论可以解决的实际问题有：不确定（不精确）数据的简化、不确定（不精确）数据的关联性发现、不确定（不精确）数据所产生的决策模型、不确定（不精确）数据所产生的范化、基于不确定（不精确）数据的知识发现等。目前粗糙集理论与方法已被广泛应用于不精确、不确定、不完全的信息分类和知识获取。

6）进化计算（evolutionary computation，EC）技术

基于生物界的自然选择和自然遗传机制的计算方法，如遗传算法（genetic algorithm，GA）、进化策略（evolution strategies，ES）和进化规则（evolutionary programming，EP）等方法，在科研和实际问题中的应用越来越广泛，并取得了较好的成果。这些方法都是基于生物进化的基本思想来设计、控制和优化的人工系统，一般将这类计算方法统称为进化计算，而将相应的算法统称为“进化算法”或者“进化程序”。这些方法能够在可以接受的计算时间内，很好地解决复杂的非线性优化问题，解决具有多个局部极值的非线性最优化问题，找到全局最优解，也可以解决复杂的组合规划或者整数规划问题。

7）灰色系统（grey system）技术

灰色系统是通过对原始数据的搜集与整理来寻求其发展变化的规律。客观系统所表现出来的现象尽管纷繁复杂，但其发展变化有着自己的客观逻辑规律，体现系统整体各功能间的协调统一。因此，如何通过散乱的数据序列去寻找其内在的发展规律就显得特别重要。灰色系统理论认为，一切灰色序列都能通过某种方式生成弱化其随机性而呈现本来的规律，认为微分方程能较准确地反映事件的客观规律，也就是通过灰色数据序列建立系统反应模型，并通过该模型预测系统的

可能变化状态。

8）模糊逻辑（fuzzy logic）技术

模糊数学是继经典数学、统计数学之后，在数学上的又一新的发展。在数据挖掘领域，基于模糊逻辑可以实现模糊综合判别、模糊聚类分析等多种数据挖掘模型。

9）人工智能（artificial intelligence，AI）技术

人工智能研究计算和知识之间的关系。用机器去模拟人的智能，使机器具有类似于人的智能，其实质是研究如何构造智能机器或智能系统，以模拟、延伸、扩展人类的智能。AI技术是在计算机科学、控制论、信息论、神经心理学、哲学、语言学等多种学科研究的基础上发展起来的。早期的研究领域有专家系统、机器学习、模式识别、自然语言理解、自动定理证明、自动程序设计、机器人学、博弈、人工神经网络等；目前已涉及数据挖掘、智能决策支持系统、知识工程、分布式人工智能等。人工智能技术包括推理技术、搜索技术、知识表示与知识库技术、归纳技术、联想技术、分类技术、聚类技术等，其中最基本的三种技术，即知识表示、推理和搜索都在数据挖掘中得到了体现。

人工智能有许多研究领域，本书主要介绍以下几个领域。

（1）专家系统（expert system）。专家系统是依靠人类专家已有的知识建立起来的知识系统。目前专家系统是人工智能研究中开展较早、最活跃、成果最多的领域，广泛应用于医疗诊断、地质勘探、石油化工、军事、文化教育等方面。它是在特定的领域内具有相应的知识和经验的程序系统，应用人工智能技术，模拟人类专家解决问题时的思维过程来求解领域内的各种问题，达到或接近专家的水平。

（2）机器学习（machine learning）。要使计算机具有知识一般有两种方法：一种方法是由知识工程师将有关的知识归纳、整理，并且用计算机可以接受、处理的方式输入计算机；另一种方法是使计算机本身有获得知识的能力，它可以学习人类已有的知识，并且在实践过程中总结、完善，这种方式称为机器学习。机器学习的研究主要体现在以下三个方面：一是研究人类学习的机理、人脑思维的过程；二是机器学习的方法；三是建立针对具体任务的学习系统。

（3）模式识别（pattern recognition）。模式识别是研究如何使机器具有感知能力，主要研究视觉模式和听觉模式的识别，如识别物体、地形、图像、字体（如签字）等。在日常生活中的各个方面以及军事上都有广泛的用途。近年来迅速发展起来的应用模糊数学模式、人工神经网络模式的方法逐渐取代了传统的基于统计模式和结构模式的识别方法。

（4）自然语言理解。计算机如能“听懂”人的语言（如汉语、英语等），便可以直接用口语操作计算机，这将给人们带来极大的便利。计算机理解自然语言的研究有以下三个目标：一是计算机能正确理解人类的自然语言输入的信息，并

能正确答复（或响应）输入的信息；二是计算机对输入的信息能产生相应的摘要，而且复述输入的内容；三是计算机能把输入的自然语言翻译成所要求的另一种语言，如将汉语译成英语或将英语译成汉语等。目前，人们做了大量的尝试，研究如何利用计算机进行文字或语言的自动翻译，但还没有找到最佳的方法，有待于深入探索。

（5）机器人学。机器人是一种能模拟人类行为的机械，研究经历了三代：第一代（程序控制）机器人；第二代（自适应）机器人；第三代（智能）机器人。智能机器人具有类似于人的智能，装备了高灵敏度的传感器，具有超过一般人的视觉、听觉、嗅觉、触觉的能力，能对感知的信息进行分析，控制自己的行为，处理环境发生的变化，完成各种复杂而困难的任务，而且具有自我学习、归纳、总结、提高已掌握知识的能力。目前研制的智能机器人大多只具有部分智能，和真正意义上的智能机器人相差甚远。

（6）智能决策支持系统（IDSS）。属于管理科学的范畴，它与“知识—智能”有着极其密切的关系。将人工智能中的技术，特别是智能和知识处理技术应用于决策支持系统，扩大了决策支持系统的应用范围，提高了系统解决问题的能力，逐渐形成智能决策支持系统。

（7）人工神经网络（artificial neural network）。人工神经网络从研究人脑的奥秘中得到启发，试图用大量的处理单元（人工神经元、处理元件、电子元件等）模仿人脑神经系统工程结构和工作机理。一般可分为三种网络模型：①前馈式网络，以感知机、误差反向传播模型、函数型网络为代表，可用于预测、模式识别等；②反馈式网络，它以霍普菲尔德（Hopfield）的离散模型和连续模型为代表，分别用于联想记忆和优化计算；③自组织网络，它以 ART 模型、Koholon 模型为代表，用于聚类分析等。

10）决策树（decision tree）技术

决策树技术主要是指针对给定的一组样本数据及其对应的规则，最终选取相应的一组动作。决策树方法是利用训练集生成一个测试函数，根据不同的取值建立树的分支；在每个分支子集中重复建立下层节点和分支。这样便生成一棵决策树，然后对决策树进行剪枝处理，最后把决策树转化为规则，决策树方法主要用于分类挖掘。决策树方法是利用信息论中的互信息（信息增益）寻找数据库中具有最大信息量的属性字段，从而建立决策树的一个节点，再根据该属性字段的不同取值建立树的分支，最后在每个分支子集中再重复建立树的下层节点和分支的过程。国际上最早、也是最有影响的决策树方法是在 1986 年由 Quinlan 提出的迭代二叉树三代方法（ID3 方法）。ID3 方法是基于信息熵的决策树分类算法，根据属性集的取值选择实例的类别，要解决的核心问题是在决策树中各层节点上选择属性。用信息增益率作为属性选择的标准，使得在每个非叶节点测试时，能获

得关于被测试例子最大的类别信息。使用该属性将例子集分成子集后，系统的熵值最小，使得该非叶子节点到其对应的后代叶子节点的平均路径最短，从而使得所生成的决策树的平均深度较小，进一步提高分类的速度和准确率。

11）统计分析（statistical analysis）方法

统计学是“数据科学”，即搜集、分析、展示及解释数据的科学。统计学在数据样本选择、数据预处理、数据挖掘过程及评价抽取知识的步骤中有着非常重要的作用。许多统计学的工作是针对数据和假设检验的模型进行评价，也包括评价数据挖掘的结果。在数据预处理步骤中，统计学提出了估计噪声参数过程中要用的平滑处理技术，一定程度上弥补丢失数据和消除奇异值对结果的负面影响作用。数据总结的最简单方法就是传统的统计方法，计算出数据库中各个数据项的总和、均值、方差、最大值、最小值、百分位数等基本描述统计量，还可利用图形工具，制作总体的频率直方图、饼状图、盒形图、茎叶图、散点图及拟合概率分布图等，将结果直观地提供给分析者。多元统计分析中的聚类分析、判别分析、回归分析、主分量分析、因子分析、典型相关分析、偏最小二乘回归等方法都能在一定程度上达到数据挖掘的目的，在数据挖掘的数据搜集、清理环节发挥作用。多元分析与其他挖掘技术相结合，使之成为数据挖掘中不可或缺的工具。

12）知识获取（knowledge acquisition）、知识表示（knowledge representation）、知识推理（knowledge reasoning）和知识搜索（knowledge search）技术

知识表示是指在计算机中对知识的一种描述，是一种计算机可以接受的用于描述知识的数据结构。表示方法可分为符号表示法和连接表示法。符号表示法使用各种包含具体含义的符号，以各种不同的方式和次序组合起来表示知识，它主要用来表示逻辑性知识。连接表示法是把各种物理对象以不同的方式及次序连接起来，并在其间相互传递及加工各种包含具体意义的信息。在数据挖掘中关联规则的挖掘用到了符号表示法。知识推理是从已知的事实出发，运用已掌握的知识，找出其中蕴含的事实，或归纳出新的事实。推理可分为经典推理和非经典推理，前者包括自然演绎推理、归纳演绎推理、与/或形演绎推理等，后者主要包括多值逻辑推理、模态逻辑推理、非单调推理等。知识搜索是根据问题的实际情况不断寻找可利用的知识，从而构造一条代价较小的推理路线。搜索分为盲目搜索和启发式搜索，盲目搜索是按预定的控制策略进行搜索，在搜索过程中获得的中间信息不用来改进控制策略。启发式搜索是在搜索过程中加入与问题有关的启发性信息，用于指导搜索朝着最有希望的方向前进，加速问题的求解过程，并找到最优解。

13）决策与控制理论（decision and control）

传统的储存决策支持系统通常是在某个假设的前提下通过数据查询和分析来验证或否定这个假设，而数据挖掘技术则能够自动分析数据，进行归纳整理，从中发现潜在的模式，或产生联想，建立新的业务模型，帮助决策者调整市场策略

并找出正确的决策。数据挖掘的出现使决策支持工具跨入了一个新阶段。数据挖掘技术的兴起为 IDSS 研究指明新的方向，即基于数据挖掘的 IDSS。

14）可视化技术（visual technology）

该方法采用直观的图形、图表方式将挖掘出来的模式加以表现，数据可视化大大扩展了数据的表达能力，从而也便于用户的理解。因此，数据挖掘中的可视化技术得到数据挖掘研究人员日益广泛的重视。

15）并行计算技术（parallel computing technologies）和海量存储（mass storage）

强大的并行处理计算机可以提高数据挖掘的应用，因为并行处理技术可以将一个复杂的查询分解成多个子查询，每个子查询交给不同的处理器处理，这一处理过程是并行执行的。因此，并行处理技术可以大大加速数据挖掘的过程。

国内外专家学者对数据挖掘领域的研究热点可以归纳为预测分析、聚类分析、分类分析、关联规则等技术。与国外相比，国内对数据挖掘与知识发现的研究稍晚。目前，国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用的研究，相关单位包括清华大学、中国科学院计算技术研究所、空军第三研究所、海军装备论证中心等。北京系统工程研究所对模糊方法在知识发现中的应用进行了较深入的研究；北京大学也在开展对数据立方体代数的研究；华中理工大学、复旦大学、浙江大学、中国科技大学、中国科学院数学研究所、吉林大学等单位开展了对关联规则开采算法的优化和改造；南京大学和上海交通大学等单位探讨并研究非结构化数据的知识发现以及 Web 数据挖掘。这些研究的目的在于解决数据丰富与知识匮乏的矛盾，主要任务是对大型、分散数据库中的数据资源进行重新规划和重新组织，营造出新的、容易利用的企业信息资源库，以达到对信息流、物流、资金流等资源的统一管理和分析，挖掘出有价值的信息和知识，给企业的管理者和决策者提供有力的决策支持。

除了上述数据挖掘的主要技术外，数据挖掘中还存在许多问题有待解决：

（1）算法效率和可伸缩性。目前，数据库的规模呈指数增长。Mb 字节规模的数据库已经非常普遍。在商业数据库中，Gb 字节规模和 Tb 字节规模的数据库也已经在使用中。当把万维网包括进来的时候，Pb 字节规模的数据库正在出现。例如，美国国家航天局轨道卫星上的地球观测系统（EOS）每小时会向地面发回大量图像数据；大型天文望远镜每年会产生不少于 10 Tb 的数据，等等。据统计，数据和计算资源的增长速度符合摩尔定理，每 18 个月翻一番。因此，海量数据挖掘的最大挑战不仅仅在于数据库的绝对规模，还在于数据挖掘系统能够处理这些持续增长的数据集合。

（2）处理不同类型的数据和数据源。目前数据挖掘系统处理的数据库大多是关系数据库。随着数据库应用范围和规模的日益扩大、功能的日益完善，数据库