数据挖掘

及其在金融信息处理中的应用

刘彦保　乔克林　著

中国水利水电出版社
www.waterpub.com.cn

数据挖掘

及其在金融信息处理中的应用

刘彦保　乔克林　著

中国水利水电出版社
www.waterpub.com.cn
·北京·

内 容 提 要

本书从理论、应用实例以及数据挖掘的发展趋势等几个方面，对数据挖掘技术进行了详细探讨。在介绍数据挖掘技术理论和算法的基础上，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，以期为读者提供一个更为广阔的视角。

本书重点对数据预处理、关联规则、聚类分析等内容进行了详尽的阐述。

全书内容丰富新颖，具有较强的可读性，可供从事数据挖掘工作以及其他相关工程技术的工作人员参考使用。

图书在版编目（CIP）数据

数据挖掘及其在金融信息处理中的应用 / 刘彦保，乔克林著. -- 北京 : 中国水利水电出版社，2018.11
ISBN 978-7-5170-7082-5

Ⅰ. ①数… Ⅱ. ①刘… ②乔… Ⅲ. ①数据采集－应用－金融－信息处理－研究 Ⅳ. ①F830.49

中国版本图书馆CIP数据核字(2018)第248522号

书　　名	数据挖掘及其在金融信息处理中的应用 SHUJU WAJUE JI QI ZAI JINRONG XINXI CHULI ZHONG DE YINGYONG
作　　者	刘彦保　乔克林　著
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路1号D座 100038） 网址：www.waterpub.com.cn E-mail：sales@waterpub.com.cn 电话：(010)68367658(营销中心)
经　　售	北京科水图书销售中心(零售) 电话：(010)88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排　　版	北京亚吉飞数码科技有限公司
印　　刷	三河市元兴印务有限公司
规　　格	170mm×240mm　16开本　22.75印张　408千字
版　　次	2019年2月第1版　2019年2月第1次印刷
印　　数	0001—2000册
定　　价	110.00元

前　言

自20世纪80年代后期出现数据挖掘以来，随着信息技术尤其是计算机及互联网技术的飞速发展，数据挖掘技术、理论及应用得到了快速发展。在当今大数据时代，金融行业每天都在产生着海量的数据。对这些数据进行统计、分析并挖掘出隐藏在数据内部有价值的信息，为金融行业的决策提供指导，已经成为具有挑战性的新课题。金融行业尤其是银行业对数据挖掘与分析技术的需求已经迫在眉睫。

在这种背景下，本书从数据挖掘与分析技术的基本理论出发，紧紧把握金融数据挖掘与分析的最新动向，对数据挖掘与分析技术及其在金融行业中的应用进行了详细介绍，并对未来金融数据挖掘与分析的发展进行了展望。

本书具有如下特点：一是理论与应用相呼应。从数据挖掘算法理论与方法、工具和应用两个方面进行阐述，既注重理论，同时贴近实战，解行相应，希望学习者既能很快将理论应用于实际领域的数据分析中，同时也具备厚积薄发的能力；二是基础与发展一脉相承。大数据新常态下经典数据挖掘的基本原理仍然适用，不同之处在于，根据现有分布式、并行环境，对原有算法进行优化。本书循序渐进地介绍经典数据挖掘算法，以及大数据环境下数据挖掘算法的新特点和新延展，有助于学习者全面掌握数据挖掘理论。

本书梳理了数据挖掘的多种研究方法，注重领域核心方法的论述，知识点比较广泛，叙述简明、语言准确。全书共11章，第1章导论，介绍了数据挖掘的起源、概念和分类、功能、典型的应用领域等；第2章介绍了数据预处理；第3章～第5章分别从关联规则、聚类分析、分类与预测等方面讲述了算法和概念。第6章～第9章介绍了数据挖掘技术，包括Web数据挖掘、复杂类型数据挖掘及应用、流数据挖掘技术和其他相关技术。第10章和第11章阐述了金融数据挖掘的理论及实际应用。

作者在多年教学、科学研究的基础上，广泛吸收了国内外学者在数据挖掘方面的研究成果，在此向相关内容的原作者表示诚挚的敬意和谢意。

本书的出版，也得到了我校（延安大学计算机软件与理论校级重点学科）的支助与支持，在此一并表示感谢。全书1～6章由刘彦保撰写，7～11章由乔克林撰写。

由于作者水平有限，加之时间仓促，错误和遗漏在所难免，恳请读者批评指正。

作　者

2018年5月

目　录

第1章　导论

随着计算机技术、数据库技术和传感器技术的飞速发展，人们获取数据和存储数据变得越来越容易。社会信息化水平的不断提高和数据库应用的日益普及，使人类积累的数据量正在以指数方式增长。与日趋成熟的数据管理技术和软件工具相比，数据分析技术和工具所提供的功能，无法有效地为决策者提供为其决策所需的有效知识，从而形成了一种“丰富的数据、贫乏的知识”的现象。为有效解决这一问题，自20世纪80年代开始，数据挖掘技术逐步发展起来，人们迫切希望能对海量数据进行更加深入的分析，发现并提取隐藏在其中的有价值信息，以便更好地利用这些数据。数据挖掘技术的迅速发展，得益于目前全世界所拥有的巨大数据资源，以及对其中有价值的信息和知识的巨大需求。在这种背景下，数据挖掘的理论和方法获得了飞速的发展，其技术和工具已经广泛应用到互联网、金融、电商、管理、生产等各个领域。

1.1　数据挖掘的起源

数据挖掘技术出现于20世纪80年代末，是在多个学科发展的基础上发展起来的。随着数据库技术的发展应用，数据的积累不断膨胀，简单的查询和统计已经无法满足企业的商业需求，急需一些革命性的技术去挖掘数据背后的信息。与此同时，计算机领域的人工智能(Artificial Intelligence，简称AI)也取得了巨大进展，进入了机器学习的阶段。基于此，人们将两者结合起来，用数据库管理系统存储数据，用计算机分析数据，并且尝试挖掘数据背后的信息，这两者的结合促生了一门新的学科，即数据挖掘。

1989年8月于美国底特律市召开的第十一届国际联合人工智能学术会议上首次提到“知识发现”这一概念，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。1995年在加拿大蒙特利尔召开的首届“知识发现和数据挖掘”国际学术会议上，首次提出了“数据挖掘”这一学科的名称，并把数据挖掘技术分为科研领域的知识发现与工程领域的数据挖掘。

数据挖掘可以在任何类型的存储信息上进行，如关系数据库、数据仓库、文本和多媒体数据库、事务数据库、等。目前，数据挖掘技术在购物篮分析、金融风险预测、分子生物学、基因工程研究、Internet 站点访问模式发现以及信息搜索等领域得到了广泛的应用。因此数据挖掘技术具有极其重要的研究意义，也给各个领域的研究人员提供了一种新的认识数据、使用数据的智能手段。

大部分数据挖掘问题和相应的解决方法都起源于传统的数据分析。数据挖掘起源于多种学科，其中最重要的两门是统计学和机器学习。统计学起源于数学，因此，数据挖掘强调数学上的精确。在实践测试之前，要求在理论上得到验证；相比之下，机器学习更多地起源于计算机实践。如果说数据挖掘的统计学方法和机器学习方法之间的主要区别在于模型和算法规则之间侧重点的不同。现代统计学几乎完全是由模型概念驱动的，是一个假定的结构，或者说是一个结构的近似，这个结构能够产生数据。统计学强调模型，而机器学习倾向于强调算法。数据挖掘中的基本模型法则也起源于控制理论，控制理论主要应用于工程系统和工业过程。通过观察一个未知系统（也被称为目标系统）的输入输出信息，决定其数学模型的问题通常被称为系统识别。系统识别的目标是多样化的，并且是从数据挖掘的立场出发的。最重要的是预测系统的行为，并解释系统变量之间的相互作用和关系。

对数据挖掘而言，哪里有数据哪里就能挖掘到“金子”，但是，随着物联网、云计算和大数据时代的来临，要在急剧膨胀的数据中挖掘“金子”，无疑给数据挖掘技术的实施提出了挑战。

物联网就是物物相连的网络，是数字世界与物理世界的高度融合。物联网底层的大量传感器为信息的获取提供了一种新的方式，这些传感器不断地产生着新的数据，随着各种各样的异构终端设备的接入，物联网采集的数据量也就越来越大，其数据类型和数据格式也会越来越复杂。这些数据与时间和空间相关联，有着动态、异构和分布的特性，也为数据挖掘任务带来了新的挑战。

云计算是一种基于互联网相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态、易扩展且经常是虚拟化的资源（包括硬件、平台和软件），实现了设备之间的数据应用和共享。随着物联网的发展，感知的信息不断增加，需要不断地增加服务器的数目来满足需求，但由于服务器的承载能力是有限的，使得服务器在节点上出现混乱和错误的概率大大增加。为了更好地服务，基于云计算的系统能有效地解决物联网分布式数据挖掘中所遇到的问题，在进行相关数据挖掘时能够显著地提高性能。

1.2 数据挖掘的概念和分类

1.2.1 数据挖掘的概念

数据挖掘(Data Mining,DM),是从大量的、有噪声的、不完全的、模糊和随机的数据中,提取出隐含在其中的、人们事先不知道的、具有潜在利用价值的信息和知识的过程。所提取到的知识的表示形式可以是概念、规律、规则与模式等。数据挖掘能够对将来的趋势和行为进行预测,从而帮助决策者做出科学合理的决策。比如,通过对公司数据库系统的分析,数据挖掘可以回答诸如"哪些客户最有可能购买我们公司的什么产品?""客户有哪些常见的消费模式和消费习惯?"等类似问题。

数据挖掘是一门交叉学科,涉及数据库技术、人工智能、数理统计、机器学习、模式识别、高性能计算、知识工程、神经网络、信息检索、信息的可视化等众多领域,其中数据库技术、机器学习、统计学对数据挖掘的影响最大。对数据挖掘而言,数据库为其提供数据管理技术,机器学习和统计学提供数据分析技术。数据挖掘所采用的算法,一部分是机器学习的理论和方法,如神经网络、决策树等;另一部分是基于统计学习理论,如支持向量机、分类回归树和关联分析等。但传统的机器学习和统计学研究往往并不把海量数据作为处理对象,而数据挖掘要把这两类技术用于海量数据中的知识发现,需要对算法进行改造,使得算法性能和空间占用达到实用的地步。

常见的数据挖掘对象有以下七大类:

(1)关系型数据库、事务型数据库、面向对象的数据库。

(2)数据仓库/多维数据库。

(3)空间数据(如地图信息)。

(4)工程数据(如建筑、集成电路信息)。

(5)文本和多媒体数据(如文本、图像、音频、视频数据)。

(6)时间相关的数据(如历史数据或股票交换数据)。

(7)万维网(如半结构化的 HTML、结构化的 XML 以及其他网络信息)。

1.2.2 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。按照具体的研发工作任务,可以将数

据挖掘所讨论的内容分为两大任务类型：描述型的数据挖掘任务和预测型的数据挖掘任务。描述型数据挖掘主要是根据数据仓库中的数据，分析其中隐含的规律性描述，例如频繁模式挖掘、聚类、关联规则的挖掘等都属于描述型数据挖掘的范畴。预测型数据挖掘主要是根据数据仓库中的数据，开展对于未知规律和知识的预测研究，例如分类、回归等方面的研究工作就属于预测型的研究。

1. 描述型数据挖掘

(1)广义知识。在建立模型之前，首先要了解数据，获得广义知识，即类别特征的概括性描述知识。根据数据的微观特性发现其表征的、带有普遍性的、较高层次概念的、中观和宏观的知识，反映同类事物的共同性质，是对数据的概括、精炼和抽象。

(2)聚类。聚类的目的是把数据对象分成各个聚类、各个簇，但聚类与分类也有显著的不同。分类的训练样本集的类标号是已知的，通过学习对训练数据集得出一个分类规则，再利用分类规则判定某个未知数据的类标号，分类是有指导的学习。进行聚类时，不存在类标号已知的训练数据集，没有什么模型可参考，聚类算法必须自己总结出各个聚类或簇之间的区别，根据某种规则对数据对象进行聚类或分类，从这个角度上讲，聚类是无指导的学习，它的算法本身远比分类的复杂度要高。

聚类分析是数据挖掘领域的一项非常重要的研究课题，基于聚类分析的算法也不断被人们提出来，如基于划分方法的 K-means 算法、K-medians 算法以及针对数据流的 Stream 算法；基于层次方法的 Birch 算法、Cure 算法以及针对数据流的 CluStream 算法、HPStream 算法；基于密度方法的 DBSCAN 算法、DENCLUE 算法以及针对数据流的 DenStream 算法；基于网格的 D-Stream 算法、STING 算法；基于子空间的 GSCDS 算法；基于混合属性的 HCluStream 算法。聚类分析内容非常丰富，有系统聚类法、有序样品聚类法、动态聚类法、模糊聚类法、图论聚类法、聚类预报法等。

(3)关联分析。关联分析是一种探索数据的描述性方法，这些数据可以帮助识别数据库中数值之间的关系，它反映一个事件和其他事件之间依赖或关联的信息。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. 预测型数据挖掘

预测型数据挖掘的目的是通过分析建立一个或一组模型，并试图预测新数据的行为。在从多种来源搜集数据的基础上，它通过构建现实世界的模型来实现，这些来源可包括企业交易、顾客历史和人口统计信息、过程控制数据，以及相关的外部数据库，例如银行交易信息或气象数据。模型建立的结果是对那些能用来进行有效预测的数据中的模式和关系的描述。

确定了预测目标之后，下一步是决定最合适的预测类型：预测行为属于什么类别或等级，或预测变量会有什么数值（如果它是随时间变化的变量，这就是所谓的时间序列预测）。之后选择模型类型：用神经网络来进行回归分析，以及可能用决策树来进行分类。也有传统的统计模型可供选择，如逻辑回归、判别分析，或一般线性模型。数据挖掘中最重要的模型类型将在后续章节中进行描述。

在预测模型中，我们的预测值或类型被称为响应、相关或目标变量。用于建立或者训练预测模型使用的数据是已知变量响应的数值。这种训练有时被称为监督学习，因为被计算值或估计值会与已知的结果进行比较（相反，在前面的描述性技术，如聚类，有时被称为无监督学习，因为没有已知的结果来引导算法）。

(1)分类。分类是预测分类标号。什么是分类标号呢？属性值有两种基本的属性值：一种是分类属性；另一种是量化属性。分类属性也叫离散属性，它的值是分成固定的区间之内的，是离散的值，而量化属性对应的是连续的值，根据分类时所对应的是离散的属性还是量化的属性，就可以把分类挖掘分成分类和预测两种类型。分类预测的是分类编号，根据训练数据集和类标号属性构建模型来分类新数据，这里主要包括两个过程：一个是构建模型来分类现有的数据；另一个是利用已有的模型对新数据进行分类。

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。

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。

(2)回归。回归分析是应用现有的数值来预测其他数值是什么。在最简单的情况下,回归分析应用的是诸如一元线性回归、多元线性回归等标准统计技术。不幸的是,很多现实世界的问题不是对原值简单的线性预测。例如,销售量、股票价格及产品的次品率都难以预测,因为它们可能要依赖于多个预测变量非线性的相互作用。因此,用更复杂的技术(如逻辑回归、决策树或神经网络)来预测未来值可能性是十分有必要的。

相同的模型类别,通常是可以被用于回归和分类的。例如,CART(分类和回归树)决策树算法可以被用来建立分类树(区分分类响应变量)和回归树(预测连续响应变量)。神经网络也可以用来创建分类和回归模型。

1.3 数据挖掘的过程

1999 年,欧盟创建了跨行业的数据挖掘标准流程,即 CRISP-DM (Cross Industry Standard Process for Data Mining),提供了一个数据挖掘生命周期的全面评述,包括业务理解、数据理解、数据准备、数据建模、模型评估和部署 6 个阶段,如图 1-1 所示。

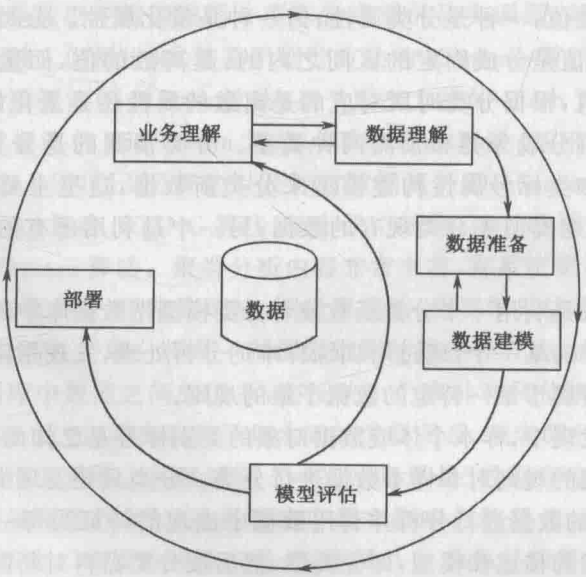

图 1-1 CRISP-DM 数据挖掘流程示意图

第1阶段：业务理解，主要任务是深刻理解业务需求，在需求的基础上制订数据挖掘的目标和实现目标的计划。

第2阶段：数据理解，主要收集数据、熟悉数据、识别数据的质量问题，并探索引起兴趣的子集。

第3阶段：数据准备，根据所要解决的问题，确定待挖掘的目标，搜索所有与业务对象有关的内部和外部数据信息，从收集来的数据集中选择必要的属性(因素)，并按关联关系将它们连接成一个数据集，然后进行数据清洗，即空值和异常值处理、离群值剔除和数据标准化等。

第4阶段：数据建模，选择应用不同的数据挖掘技术，并确定模型最佳的参数。如果初步分析发现模型的效果不太满意，要再跳回数据准备阶段，甚至数据理解阶段。

第5阶段：模型评估，对建立的模型进行可靠性评估和合理性解释，未经过评估的模型不能直接应用。彻底地评估模型，检查构造模型的步骤，确保模型可以完成业务目标。如果评估结果没有达到预想的业务目标，要再跳回业务理解阶段。

第6阶段：部署阶段，根据评估后认为合理的模型，制定将其应用于实际工作的策略，形成应用部署报告。

1.4 数据挖掘的功能

对一个数据挖掘系统而言，它应该能同时搜索发现多种模式的知识，以满足用户的期望和实际需要。此外，数据挖掘系统应能够挖掘出多种层次(抽象水平)的模式知识，允许用户来指导挖掘搜索有价值的模式知识。由于有些模式并非对数据库中的所有数据都成立，通常每个被发现的模式带上一个确定性或“可信性”度量。数据挖掘的功能以及它们可以发现的模式类型介绍如下：

(1)概念描述。被分析的数据称为目标数据集，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，一般分为定性概念描述和对比概念描述。

(2)关联分析。关联分析就是从给定的数据集中发现频繁出现的项集模式知识，即发现各属性之间的关联关系并用关联规则描述出来(又称关联规则)。

(3)分类和预测。根据一系列已知数据，分类找出一组能够描述并区分数据或概念的模型，以便能够使用模型预测未知的对象类。导出模型是基

于训练数据集的分析。例如，指纹识别、人脸识别、工业上故障诊断和商业中的客户识别分类等都是分类问题。

(4)聚类分析。根据物以类聚原则，利用属性特征将数据集合分为由类似的数据组成的多个类的过程称为聚类。即对象的聚类(簇)这样形成，使得在一个聚类中的对象具有很高的相似性，而不在一个聚类中的对象具有很高的非相似性。

(5)趋势分析。对于上面提到的四种功能，事件产生的顺序信息都被忽略，被简化地作为一条静态的记录来对待。而趋势分析是对随时间变化的数据对象的变化规律和趋势进行建模描述，根据前一段时间的运动预测下一个时间点的状态。解决的问题一般可分为两类：总结数据的序列或者变化趋势，如期货交易/预测股票，网页点击顺序记录等；检测数据随时间变化的变化，如自来水厂用水量的日、周、月、年等周期变化。

(6)异类(孤立点)分析。数据集中那些不符合大多数数据对象所构成的规律(模型)的数据对象被称为异类(outlier)。大部分数据挖掘方法将异类视为噪声或异常丢弃。然而，在某些特定应用场合(如商业欺诈行为的自动检测)，小概率发生的事件(数据)比经常发生的事件(数据)更有挖掘价值。

(7)演化分析。演化分析是对随时间变化的数据对象的变化规律和趋势进行建模描述，根据前一段时间的运动预测下一个时间点的状态。

1.5 数据挖掘的典型应用领域

数据挖掘已经在很多领域得到了应用，虽然这些应用可能是初步的，但是它们反映了数据挖掘技术的应用趋势。

1.5.1 数据挖掘在电信业中的应用

随着4G时代的到来，电信业发展面临着前所未有的机遇和挑战，电信业务从原来单纯的通话业务，扩展到了数字通信等多种不同的业务类型，所以客户服务的质量是关系到电信运营商发展的主要因素。数据挖掘广泛应用于国内电信行业中，对企业日常经营数据进行数据分析与挖掘，从海量数据中寻找数据相互之间的关系或模式。特别是当前电信业的激烈竞争推动了数据挖掘技术在该行业的深入应用，数据挖掘技术在电信领域的多方面发挥着作用。

目前，电信市场的竞争也变得越来越激烈和全方位化，不管是住宅电话还是移动电话，每天的使用量都很大。对电信公司来讲，如何充分使用这些数据为自己赢得更多的利润就成了主要问题。利用数据挖掘来帮助理解商业行为、对电信数据的多维分析、检测非典型的使用模式以寻找潜在的盗用者、分析用户一系列的电信服务使用模式来改进服务、根据地域分布疏密性找出最急需建立网点的位置、确定电信模式、捕捉盗用行为、更好地利用资源和提高服务质量是非常必要的。借助数据挖掘，可以减少很多损失，保住顾客。

数据挖掘在电信业的应用包括以下方面：

(1)对电信数据的多维分析。

(2)检测非典型的使用模式以寻找潜在的盗用者。

(3)分析用户一系列的电信服务使用模式来改进服务。

(4)搅拌分析等。

这些应用可以帮助电信企业制定合理的电话收费和服务标准，针对客户群的优惠政策、防止欺诈费用等行为。

1. 客户细分

客户细分就是将客户划分为不同的群体，采用数据挖掘中的聚类和分类算法对数据集进行划分，以便企业制定适宜的营销策略、广告策略、促销策略等来实现更好的客户服务，增加企业的语音业务和各项增值业务的收入。例如，中国移动针对不同客户群体推出全球通、神州行和动感地带三大客户品牌。全球通的资费标准最高，主要针对高端用户，如经常出差的商务人士；神州行适合低端预付费用户；动感地带适合年轻群体，在短信包月方面有很大优势，同时还提供多种迎合年轻人喜好的定制服务。

2. 客户流失分析

客户流失分析是一种预测客户流失的重要技术，它通过预测可能流失的客户，帮助公司针对这些客户制定一些挽留策略，如降价或提供特殊服务吸引客户留下。决策树是最常用的一种分类预测方法，建立实用模型预测哪些客户具有流失倾向。

3. 客户个性化分析

通过建立面向多维(包括通信时间、通信者的位置、通信的途径类型)电信数据的数据仓库或数据立方体系统，帮助分析人员对客户的行为进行统计分析，同时也可以对这些数据进行序列分析、聚类等挖掘操作，分析客户

的联系网络构成，以更好地了解客户行为和习惯等。这些信息有助于电信企业制订面向客户的个性化策略，以吸引客户，同时也可以方便设备管理人员对通信设备进行合理的配置（包括通信能力、缓存等），为客户提供更良好的使用感受。

4. 恶意通信行为分析

电信业有一些恶意的通信行为，如盗用号码、发送垃圾短信、拨打推销电话等。这些行为严重地影响了用户的使用体验。为了阻止这些恶意行为对客户的影响，可以利用频繁模式挖掘、孤立点分析等方法对这些恶意行为的模式进行挖掘，找出其行为的特点，并利用预测挖掘方法进行预测，从而及时发现并采取阻止措施。

1.5.2 数据挖掘在生物信息学和医学领域中的应用

近20年来，生物信息学技术得到了长足的发展，而且生物信息学已经和医学紧密地结合在一起，逐渐形成了新的医疗和制药体系。在生物信息学和医学领域，数据挖掘算法与传统的结构化数据分析有很多的差别，产生了一系列新的技术。

生物信息学研究的基础是基因序列数据分析。根据“中心法则”，基因序列包括DNA序列、RNA序列等，它们构成了所有活生物体的基因代码基础。下面以DNA序列为例进行介绍。DNA序列由四种脱氧核苷酸构成，分别是腺嘌呤（A）、胞嘧啶（C）、鸟嘌呤（G）、胸腺嘧啶（T）。这四种脱氧核苷酸构成的序列或链，形成一个双绞旋梯。生物的基因序列都很长，一个基因通常由成百个脱氧核苷酸构成。脱氧核苷酸按不同的次序形成不同的基因，不同的基因导致生物体呈现不同的性状，因此，研究基因序列中的模式与病症、性状的模式之间的关联关系成为这方面研究的核心技术，这就需要相应的数据挖掘技术支撑。

在生物信息学和医学数据分析中包含以下技术。

1. DNA序列间的相似搜索和比较

DNA序列相似性比较是生物信息学和现代医学的基础，但DNA序列的相似性度量方法与传统的序列相似性度量方法有很大的不同。DNA序列之间的距离往往是通过编辑距离来描述的，并且不同的脱氧核糖核苷酸之间的距离也是不同的，这与数字序列和一般的离散数字序列分析技术是不同的，因此产生了一系列针对DNA序列数据的挖掘方法。