数据挖掘技术及工程实践系列丛书之

数据挖掘应用 20 个案例分析

(V01.00.000)

www.tipdm.com

广州太普信息技术有限公司

地址： 广州市经济技术开发区科学城 232 号
网址： http://www.tipdm.com
邮箱： 5iai2008@gmail.com
热线： 40068-40020
邮编： 510663
电话： （020）85661483/85666585

前　言

数据挖掘技术近年来越来越受到人们的重视，被认为是21世纪最具有发展前途的领域，当前数据挖掘技术已广泛应用在电信、零售、农业、网络、银行、电力、生物、天体、化工、医药等方面。随着中国企业信息化水平的不断提高，商务智能作为支持企业决策的重要手段，逐渐被很多企业所采用，社会对数据挖掘人才的需求持续增长。特别是有实践经验的数据挖掘人才更是各企业争夺的热门。为了满足日益增长的数据挖掘人才需求，很多大学都在开设不同程度的商务智能课程。“数据挖掘”作为商务智能的核心技术，已经成为许多高校计算机、软件工程等专业的重要课程之一。由于条件和学时所限，这些内容的教学主要限于理论介绍。因为抽象，学生理解起来往往比较困难，教学效果不甚理想。

目前，数据挖掘有关的书籍已经很多，但数据挖掘这门课程既强调深入的算法理论，又强调实际应用，因此，要真正讲好这门课，让学生既能掌握必要的理论基础，又能将数据挖掘理论和方法用于解决实际问题，对授课教师而言都有一定的难度。虽然国外已经有数据挖掘的商业软件，但通常是面向大型企业应用的，而且价格都比较昂贵，使很多大学不能建立有效的教学实习环境，缺乏相应的实验教材和讲义，导致数据挖掘课程的教学大多只能以讲解原理为主，实验环节非常薄弱，造成学生缺少数据挖掘实践经验、缺乏实践应用能力。因此高校迫切需要一款适合自己的，具有丰富数据挖掘实际应用案例的，以提高学员数据挖掘实际应用和动手能力为目的，寓教于学的高校数据挖掘教学套件。

教材是课程设计的载体，体现课程的设计思路。本课程的总体设计思路是：以数据挖掘技术在电信、电力、政府、金融、保险、制造等行业的应用作为教学内容，针对案例的数据挖掘目标，从分析方法与过程来详尽展开组织教学过程，以具备数据挖掘建模体验室作为上课场所，采用教、学、用三者结合以用为主的教学方式，使用学生在轻松的教学环境中快速熟习和掌握数据挖掘建模过程和方法。

本教材根据数据挖掘建模过程确定的教学任务主要包括数据探索、数据预处理、分类与回归、聚类分析、时序预测、关联规则挖掘、偏差检测，以及基于云计算的海量数据挖掘。教学内容的编排以解决某个应用的挖掘目标为前提；在教学过程组织上，按照先介绍案例背景，再阐述分析方法与过程，最后完成模型构建的顺序进行，在讲授建模过程的同时进行动手操作训练，把相关的知识点嵌入相应的操作过程中。与本教材配套，在线数据挖掘平台（TipDM）是本课程的辅助教学工具，该工具支持 CRISP-DM 流程所需的主要过程，完成包括对数据进行预处理，包括空值处理、降维处理、离散处理，因子分析、主成分分析、抽样、过滤等，创建、训练、评估模型，预测，修改模型参数，误差分析等一系列功能。

通过本课程的学习，学生能够学习到：

1、数据挖掘技术的基本知识及理论

2、数据挖掘技术的主要应用场景及常用算法

3、数据挖掘技术针对具体应用的技术实现路线

www.tipdm.com

4、基于云计算的海量数据挖掘技术

本教材具有以下特点：

第一，教材充分体现了以数据挖掘技术在行业中的应用做引领、以挖掘建模方法与过程分析为导向的设计思想。按照挖掘目标的完成顺序将数据挖掘过程分解成五个典型的工作任务，按完成工作任务的顺序组织教学内容。

第二，教材充分体现了高校教育的创新性、实践性和开放性的要求。将企业进行数据挖掘应用涉及的新方法和新工具及时地纳入了教材。教材中的案例设计内容具体，较好地模拟了数据挖掘建模的具体实施过程，在实训室环境下具有可操作性。

第三，教材与目前广泛采用的在线数据挖掘平台（TipDM）相结合，引导教学场所的设备和软件的配备。教材表达通俗易懂、图文并茂，既是教科书又是软件操作说明书，还是引导教学过程进行的脚本。

在课程开发和教材建设的同时，本例为本课程教学配备了大量视频教学资源，可供教师备课参考和学生自主学习。网址是：http://www.tipdm.com　打开网址后进入“视频教学”栏目。其中，教学资源主要集中在该栏目中。

编者　张小波　于广州

2012 年 10 月

www.tipdm.com

目　录

第 1 章 基础知识 7

1.1. 什么是数据挖掘 7

1.2. 数据挖掘在企业商务智能应用中的定位 8

1.2.1. 数据挖掘给企业带来最大的投资收益 8

1.2.2. 数据挖掘从本质上提升商务智能平台的价值 8

1.2.3. 数据挖掘让商务智能流程真正的形成闭环 9

1.3. 信息类 BI 与知识类 BI 11

1.4. 数据挖掘现状及应用前景 12

1.5. 数据挖掘应用 14

1.5.1. 分类与回归 14

1.5.2. 聚类分析 16

1.5.3. 关联规则 17

1.5.4. 时序模式 18

1.5.5. 偏差检测 20

1.6. 数据挖掘建模过程 21

1.6.1. 定义挖掘目标 21

1.6.2. 数据取样 22

1.6.3. 数据探索 22

1.6.4. 预处理 23

1.6.5. 模式发现 26

1.6.6. 预测建模 26

1.6.7. 模型评价 27

第 2 章 案例分析 28

2.1. 金融电信 28

2.1.1. 案例一 基于公司价值评价的证券策略投资 28

2.1.2. 案例二 基于拐点变化的股票趋势预测 35

2.1.3. 案例三 电信 3G 客户识别系统 40

www.tipdm.com

2.1.4. 案例四 基于客户分群的精准智能营销 49
2.2. 电力行业 65
2.2.1. 案例五 电力负荷预测 65
2.2.2. 案例六 防窃漏电自动诊断 75
2.3. 互联网行业 84
2.3.1. 案例七 商业零售行业中的购物蓝分析 84
2.3.2. 案例八 电子商务网站中的商品推荐 88
2.3.3. 案例九 网络入侵智能检测 92
2.3.4. 案例十 基于 RFM 的客户细分 96
2.3.5. 案例十一 信息系统客户服务感知评价 108
2.4. 生产制造 137
2.4.1. 案例十二 基于小波变换的桩基完整性检测 137
2.4.2. 案例十三 基于水色图像的水质评价 145
2.4.3. 案例十四 生物质废物混合厌氧消化在线建模 153
2.4.4. 案例十五 水产养殖投入产出多目标优化仿真 160
2.5. 公共服务 169
2.5.1. 案例十六 乳腺癌证素变化规律及截断疗法 169
2.5.2. 案例十七 卷烟消费者购买行为分析 187
2.5.3. 案例十八 纳税人偷漏税评估 196
2.5.4. 案例十九 城市综合环境质量评价 207
2.5.5. 案例二十 企业信用评级 211
第 3 章 建模工具 219
3.1. ENTERPRISE MINER 219
3.2. CLEMENTINE 219
3.3. INTELLIGENT MINER 219
3.4. SQL SERVER 219
3.5. WEKA 219
3.6. TIPDM 219
第 4 章 顶尖数据挖掘平台（TIPDM） 221

www.tipdm.com

4.1. 产品简介 221
4.2. 产品功能 221
4.2.1. 数据探索及预处理 221
4.2.2. 分类与回归 222
4.2.3. 时序模式 224
4.2.4. 聚类分析 225
4.2.5. 关联规则 225
4.3. 使用说明 226
4.4. 产品特点 242
第 5 章 上机实践 249
实验一 数据探索及数据预处理 249
实验二 神经网络模型的构建与使用 250
实验三 决策树模型的构建与使用 253
实验四 聚类算法的构建与使用 255
实验五 关联规则模型的构建与使用 257
实验六 时间序列模型的构建与使用 259
附录三 参考文献 261

www.tipdm.com

第1章 基础知识

1.1. 什么是数据挖掘

随着计算机技术、网络技术、通讯技术和 Internet 技术的发展和各行各业业务操作流程的自动化，企业内产生了大量业务数据，这些数据动辄以 TB 计。但这些数据和由此产生的信息是企业的财富，它如实的记录着企业运作的本质状况。但是面对大量的数据，迫使人们不断寻找新的工具，来对企业的运营规律进行探索，为商业决策提供有价值的信息，使企业获得利润。能满足企业这一迫切需求的强有力的工具就是数据挖掘。对于企业而言，数据挖掘有助于发现业务的趋势，揭示已知的事实， 预测未知的结果。从这个意义上讲，知识是力量，数据挖掘是财富。

数据挖掘（Data Mining）就是从大量数据（包括文本）中挖掘出隐含的、先前未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程。是利用各种分析工具在海量数据中发现模型和数据之间关系的过程，这些模型和关系可以被企业用来分析风险、进行预测。

数据挖掘的目的就是从数据中淘金，就是从数据中获取智能的过程。

“数据挖掘是通过仔细分析大量数据来揭示有意义的新的关系、模式和趋势的过程。它使用模式认知技术、统计技术和数学技术。”（Gartner Group）。

“数据挖掘是一个从大型数据库中提取以前不知道的可操作性信息的知识挖掘过程。”（Aaron Zornes，The META Group）。

报表和商业智能解决方案对于了解过去和现在的状况是非常有用的。但是，预测分析解决方案还能使企业预见未来的发展状况，让企业能够先发制人，而不是处于被动。预测分析解决方案将复杂的统计方法和机器学习技术应用到数据之中，通过使用预测分析技术来揭示隐藏在交易系统或企业资源计划（ERP）、结构数据库和普通文件中的模式和趋势。

总之，由于企业内产生了大量的业务数据，这些数据和由此产生的信息是企业的财富，它如实记录了企业运作的本质状况。通过数据挖掘分析，能帮助企业发现业务的趋势，揭示已知的事实， 预测未知的结果。“数据挖掘”已成为企业保持竞争力的必要方法。

www.itpub.com

1.2. 数据挖掘在企业商务智能应用中的定位

1.2.1. 数据挖掘给企业带来最大的投资收益

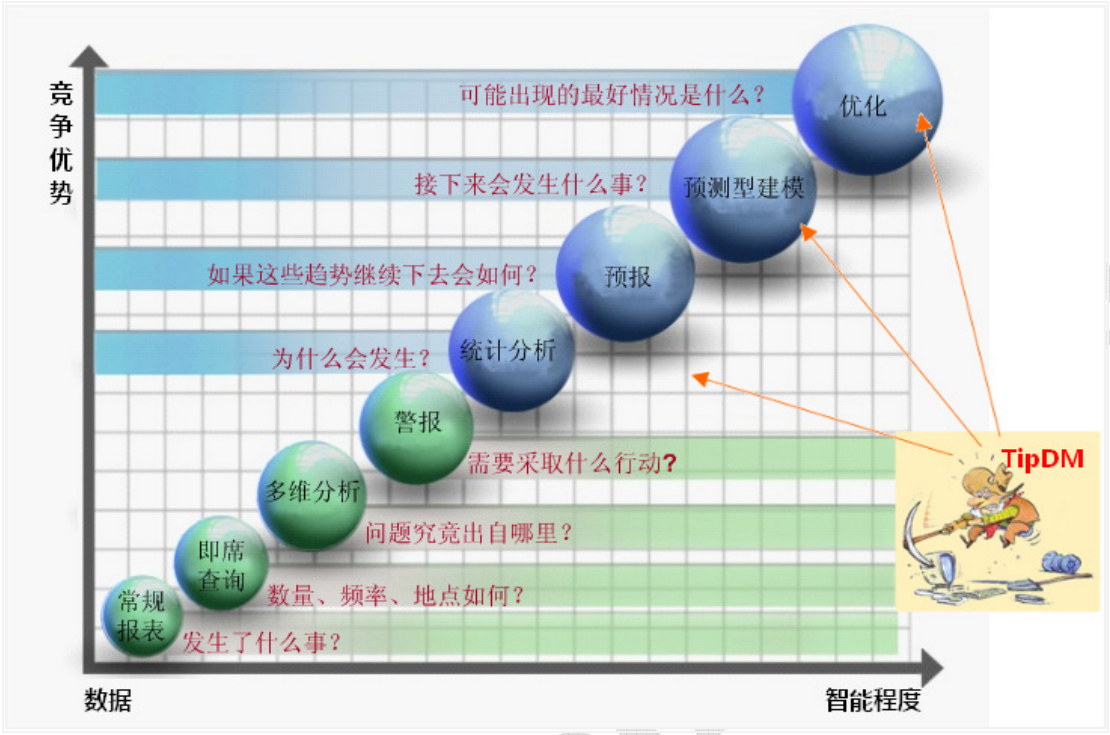


图 1-1 数据分析能力的演进

数据挖掘的主要功能是从大量数据（包括文本）中挖掘出隐含的、先前未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程。随着计算机技术、网络技术、通讯技术和Internet技术的发展和各行各业业务操作流程的自动化，企业内产生了大量业务数据。这些数据和由此产生的信息是企业的财富。它如实的记录着企业运作的本质状况。但是面对大量的数据，迫使人们不断寻找新的工具，来对企业的运营规律进行探索，为商业决策提供有价值的信息，使企业获得利润。能满足企业这一迫切需求的强有力的工具就是数据挖掘。对于企业而言，数据挖掘有助于发现业务的趋势，揭示已知的事实，预测未知的结果。从这个意义上讲，知识是力量，数据挖掘是财富。

报告和商业智能解决方案对于了解过去和现在的状况是非常有用的。但是，预测分析解决方案还能使您预见未来的发展状况，让您的机构能够先发制人，而不是处于被动。预测分析解决方案将复杂的统计方法和机器学习技术应用到数据之中，通过使用预测分析技术来揭示隐藏在交易系统或企业资源计划（ERP）、结构数据库和普通文件中的模式和趋势。

1.2.2. 数据挖掘从本质上提升商务智能平台的价值

传统的商务智能应用都局限在数据仓库、OLAP和各种类型报表的展现上。传统的OLAP分析和报表分析只能反映企业过去以及现在企业营运的问题，而不能明确、可量化的回答企业问题的原因、问题可能的解决方式以及预测企业的经营活动。另外，伴随着业务系统的不断发展以及数据仓库数据的不断积累和补充，企业分析人员无法应对如此多的数据，无法从繁缛的报表中得到有益的信息，从而陷入信息爆炸的困境。

数据挖掘功能是现有商务智能平台的核心组成部分，TipDM数据挖掘工具能够从本质上提升企业商务智能平台的价值，真正的把企业积累的数据转化为企业营运和管理的知识。TipDM数据挖掘工具不但能够帮助企业能够定性、定量的去了解各种业务问题的本质，而且能够帮助企业能够定性、定量的了解各种市场活动和企业内部运作可能带来的收益，从而不断的发现新的收益增长点。

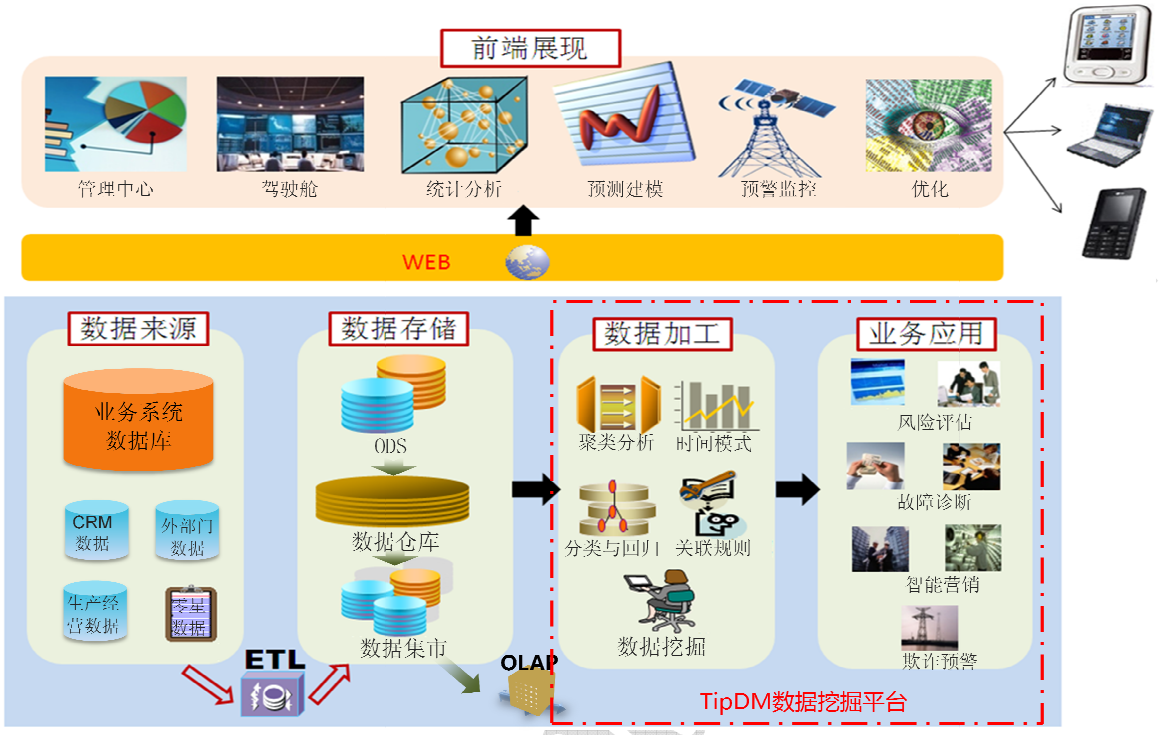


图 1-2 商务智能分析中的数据挖掘建模

1.2.3. 数据挖掘让商务智能流程真正的形成闭环

www.tipdm.com

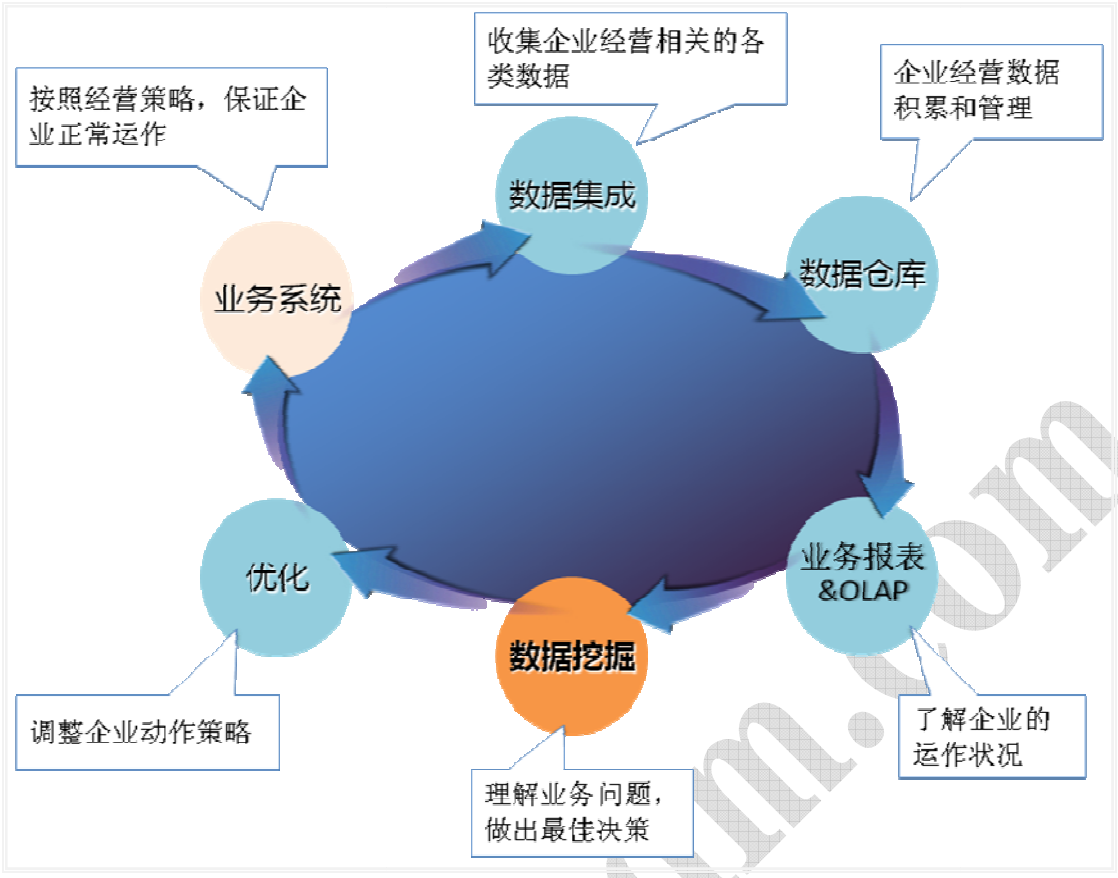


图 1-3 数据挖掘让商务智能流程形成闭环

数据挖掘应用让企业的商务智能流程真正形成了闭环。它帮助企业不断了解自身运作中的各种问题、发现新的市场机会，并适时调整企业经营的策略，从而螺旋式地提高企业的经营情况和管理水平。

www.tipdm.com

1.3. 信息类 BI 与知识类 BI

BI(Business Intelligence) 是一种运用了数据仓库、在线分析和数据挖掘等技术来处理和分析数据的崭新技术，目的是为企业决策者提供决策支持。

BI 是一个数据加工工厂。BI 的原材料是海量的数据；BI 的产品是由数据加工而来的信息和知识；BI 将这些产品推送给企业决策者；企业决策者利用 BI 工厂的产品做出正确的决策，促进企业的发展。这就是 Business Intelligence，即商业智能——连接数据与决策者，变数据为价值。

BI 应用的两大类别是信息类应用和知识类应用，两者的主要区别见表 1-1：

表 1-1 信息类 BI 与知识类 BI 的区别

信息类 BI 应用	知识类 BI 应用（TipDM）
指由原始数据加工而来的数据查询、报表图表、多维分析、数据可视化等应用，这些应用的共同特点是：将数据转换为决策者可接受的信息，展现给决策者。 例如将银行交易数据加工为银行财务报表。	指通过数据挖掘技术和工具，将数据中隐含的关系发掘出来，利用计算机直接将数据加工为知识，展现给决策者。
仅负责提供信息，而不会主动去分析数据。 例如，银行财务报表工具没有深入分析客户流失和银行利率之间关系的能力，而只能靠决策者结合信息，通过人的思考，得出知识。	会主动去数据中探查数据关联关系，发掘那些决策者人脑无法迅速发掘的隐含知识，并将其以可理解的形式呈现在决策者面前。

信息类 BI 应用都局限在数据仓库、OLAP 和各种类型报表的展现上。她只能反映企业过去以及现在企业营运的问题，而不能明确、可量化的回答企业问题的原因、问题可能的解决方式以及预测企业的经营活动。另外，伴随着业务系统的不断发展以及数据仓库数据的不断积累和补充，企业分析人员无法应对如此多的数据，无法从繁缛的报表中得到有益的信息，从而陷入信息爆炸的困境。

知识类 BI 是现有商务智能平台的关键组成部分，她能从本质上提升企业商务智能平台的价值，真正的把企业积累的数据转化为企业营运和管理的知识。她不但能够帮助企业定性、定量的去了解各种业务问题的本质，而且能够帮助企业定性、定量的了解各种市场活动和企业内部运作可能带来的收益，从而不断的发现新的收益增长点。

www.tipdm.com

1.4. 数据挖掘现状及应用前景

“数据海量，信息缺乏”是大多数金融企业所面对的尴尬。目前金融业实施的大多数数据库只能实现数据的录入、查询、统计等较低层次的功能，却无法发现数据中存在的各种有用的信息，譬如对这些数据进行分析，发现其数据模式及特征，然后可能发现某个客户、消费群体或组织的金融和商业兴趣，并可观察金融市场的变化趋势。

作为一门应用技术，数据挖掘可谓涵盖广泛，尤其在发达国家，数据挖掘技术的触角已经伸向了各行各业。只要某产业拥有具分析价值与需求的数据仓储或数据库， 皆可利用挖掘工具进行有目的的挖掘分析。一般较常见的应用案例多发生在零售业、直效行销界、制造业、财务金融保险、通讯业以及医疗服务等。数据挖掘的应用领域具体可以分为三类：第一类，商业与电子商务数据。银行、管理部门、网络应用在商业运作过程中产生大量数据，这些行业需要通过数据分析做出有效的决策。第二类，科学、工程学和卫生保健数据。 工程领域的数据往往比商业数据更复杂，此外，科学家和工程师越来越多地使用模拟系统。第三类：网络数据。网络上的数据不仅在数量上日益膨胀，在内容上也越来越复杂。网络数据已经不仅仅包括图象、文本，还包括数据流和数值数据。

可以说，未来若干年，数据挖掘都将是极为重要的成长领域，数据挖掘的应用越来越广泛。研究结果显示，企业所处理的数据每五年就会呈现倍数增长。大部分的企业并没有数据不足的问题，过度的数据重复与不一致才是大问题，这使得企业无论在使用、有效管理、以及将这些数据用于决策过程方面都遭遇到了问题。因此市场需要的是能够将数据转变成可靠与可用信息的系统。不同领域的专家对数据挖掘都表现出了极大的兴趣，例如在信息服务业中出现了一些应用，在 Internet 之数据仓储和线上服务中也给企业增加了许多生机。同时在产学合作下，又发展出了许多实用的系统，例如 MDT、Coverstory and Spotlight、Nich work visualization system LBS 、FALCON 、FAIS 、NYNEX、 TASA 等等。目前已被许多研究者视为结合数据库系统和机器学习技术的重要领域。对于研究者来说，数据挖掘是个充满潜力和机遇无限的研究领域。

由于数据挖掘能分析出数据中的有用信息，给企业带来显著的经济效益，这使得数据挖掘技术越来越普及。据 IDC 预测，目前数据挖掘市场规模大约为 140 亿美元，并对欧洲和北美 62 家采用了数据挖掘技术的企业进行调查分析，发现这些企业的 3 年平均投资回报率为 401%，其中 25%的企业的投资回报率超过 600%。美国 Palo Alto 管理集团公司对欧洲、北美和日本 375 家大中型企业的数据挖掘技术的采用情况进行了调查。结果显示，在金融领域，数据挖掘技术的应用水平已经达到或接近 70%，在营销领域也达到 50%，并且在未来的数年中，各个应用领域对该技术的采纳水平都将提高约 50%。近年来，甲骨文、IBM、微软和 SAP 耗费超过 150 亿美元收购专门从事数据挖掘的软件公司。这个产业预计有 1000 亿美元市场容量，并以每年 10%的速度增长，是软件产业整体增速的两倍。英国电信需要发布一种新的产品，需要通过直邮的方式向客户推荐这种产品，使用数据挖掘技术使直邮的回应率提高了 100%;GUS 日用品零售商店需要准确的预测未来的商品销售量，降低库存成本，使用数据挖掘技术使库存成本比原来减少了 3.8%；汇丰银行需要对不断增长的客户群进行分类，对每种产品找出最有价值的客户，使用数据挖掘技术营销费用减少了 30%；美国国防财务部需要从每年上百万笔的军火交易中发现可能存在的欺诈现象，使用数据挖掘技术发现可能存在欺诈的交易，进行深入调查，节约了大量的调查成本。

业内专家认为，从国外多年的发展经验来看，数据挖掘技术的应用对传统产业的升级具有重要意义，对我国企业竞争力的提升有显著拉动效果。未来几年，数据挖掘将是极为重要的成长领域，数据挖掘的应用越来越广泛。研究结果显示，企业所处理的数据每五年就会呈现倍数增长。大部分的企业并没有数据不足的问题，过度的数据重复与不一致才是大问题，这使得企业无论在使用、有

效管理、以及将这些数据用于决策过程方面都遭遇到了问题。该项目研制的平台技术和标准，能极大地方便企业将数据转变成可靠与可用信息，为做大做强我国自主的信息服务产业，具有很大的意义。

www.tipdm.com

1.5. 数据挖掘应用

由于数据挖掘能分析出数据中的有用信息，给企业带来显著的经济效益，这使得数据挖掘技术越来越普及。

英国电信需要发布一种新的产品，需要通过直邮的方式向客户推荐这种产品，使用数据挖掘技术使直邮的回应率提高了 100%；GUS 日用品零售商店需要准确的预测未来的商品销售量，降低库存成本，使用数据挖掘技术使库存成本比原来减少了 3.8%；汇丰银行需要对不断增长的客户群进行分类，对每种产品找出最有价值的客户，使用数据挖掘技术营销费用减少了 30%；美国国防财务部需要从每年上百万笔的军火交易中发现可能存在的欺诈现象，使用数据挖掘技术发现可能存在欺诈的交易，进行深入调查，节约了大量的调查成本。

可以说，数据挖掘在各领域的应用非常广泛，只要该产业拥有具分析价值与需求的数据仓储或数据库，皆可利用挖掘工具进行有目的的挖掘分析。一般较常见的应用案例多发生在零售业、制造业、财务金融保险、通讯业以及医疗服务等。

在销售数据中发掘顾客的消费习性，并可藉由交易纪录找出顾客偏好的产品组合，其它包括找出流失顾客的特征与推出新产品的时机点等等都是零售业常见的实例；直效行销强调的分众概念与数据库行销方式在导入数据挖掘的技术后，使直效行销的发展性更为强大，例如利用数据挖掘分析顾客群之消费行为与交易纪录，结合基本数据，并依其对品牌价值等级的高低来区隔顾客，进而达到差异化行销的目的；制造业对数据挖掘的需求多运用在品质控管方面，由制造过程中找出影响产品品质最重要的因素，以期提高作业流程的效率。

近来电话公司、信用卡公司、保险公司以及股票交易商对于诈欺行为的侦测（Fraud Detection）都很有兴趣，这些行业每年因为诈欺行为而造成的损失都非常可观，数据挖掘可以从一些信用不良的客户数据中找出相似特征并预测可能的诈欺交易，达到减少损失的目的。财务金融业可以利用数据挖掘来分析市场动向，并预测个别公司的营运以及股价走向。数据挖掘的另一个独特的用法是在医疗业，用来预测手术、用药、诊断、或是流程控制的效率。

归纳起来，数据挖掘技术的基本任务主要体现在分类与回归、聚类、关联规则发现、时序模式、偏差检测等五个方面。图 1-4 是一例典型的数据挖掘应用场景。

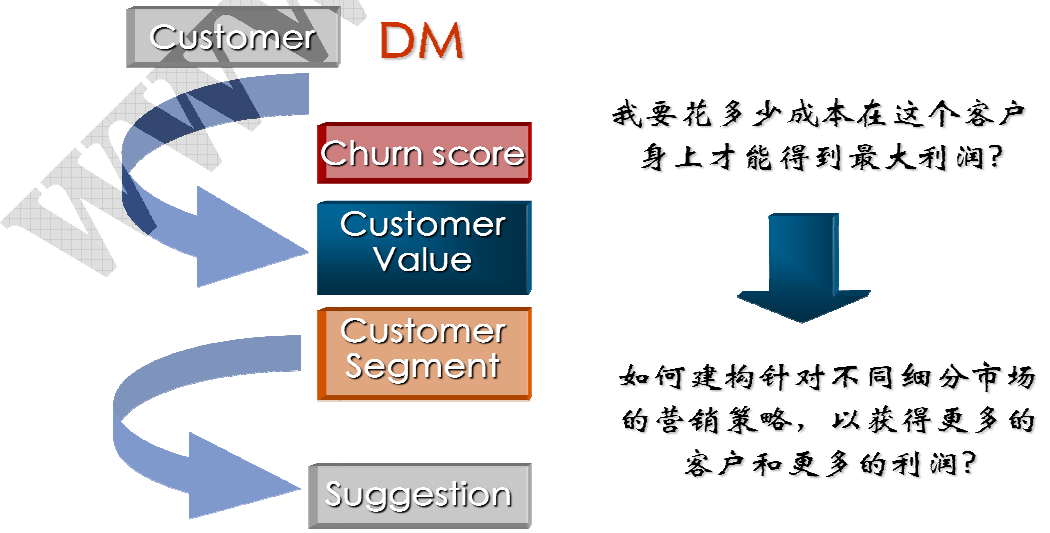


图 1-4 数据挖掘应用示例

1.5.1. 分类与回归

我们经常会碰到这样的问题：

1、如何将信用卡申请人分为低、中、高风险群？

2、如何预测哪些顾客在未来半年内会取消该公司服务，哪些电话用户会申请增值服务？

3、如何预测具有某些特征的顾客是否会购买一台新的计算机？

4、如何预测病人应当接受三种具体治疗方案的哪一种？

5、如何预测一位顾客在一次销售期间将花多少钱？

6、如何预测银行可以安全地贷给贷款人的贷款量？

7、使用 2G 通信网络的手机用户哪些有可能转换到 3G 通信网络？

8、如何有效预测房地产开发中存在的风险？

除此之外，市场经理需要数据分析，以便帮助他来猜测具有某些特征的顾客是否会购买一台新的计算机；医学研究者希望分析乳腺癌数据，预测病人应当接受三种具体治疗方案的哪一种；市场经理希望预测一位顾客在一次销售期间将花多少钱；预测银行可以安全地贷给贷款人的贷款量，这些都是分类与回归的例子。

分类（Classification）：分类是指将数据映射到预先定义好的群组或类。因为在分析测试数据之前，类别就已经确定了，所以分类通常被称为有监督的学习。分类算法要求基于数据属性值来定义类别，通常通过已知所属类别的数据的特征来描述类别。

分类就是构造一个分类函数(分类模型)，把具有某些特征的数据项映射到某个给定的类别上。该过程由两步构成：模型创建和模型使用。模型创建是指通过对训练数据集的学习来建立分类模型；模型使用是指使用分类模型对测试数据和新的数据进行分类。其中的训练数据集是带有类标号的，也就是说在分类之前，要划分的类别是已经确定的。通常分类模型是以分类规则，决策树或数学表达式的形式给出，图 1-5 就是一个三分类问题。



图 1-5 分类问题

回归（Regression）：用属性的历史数据预测未来趋势。回归首先假设一些已知类型的函数（例如线性函数、Logistic 函数等）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，图 1-6 就是一个非线性回归问题。

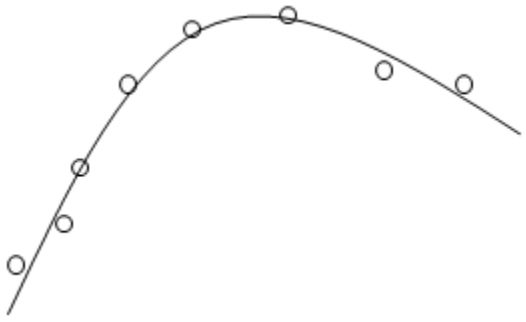


图 1-6 回归问题

回归模式的函数定义与分类模式相似，主要差别在于分类模式采用离散预测值(例如类标号)，

而回归模式采用连续的预测值。在这种观点下，分类和回归都是预测问题。但在数据挖掘业界，大家普遍认为：用预测法预测类标号为分类，预测连续值(例如使用回归方法)为预测。许多问题可以用线性回归解决，对于许多非线性问题可以通过对变量进行变化，从而转换为线性问题来解决。

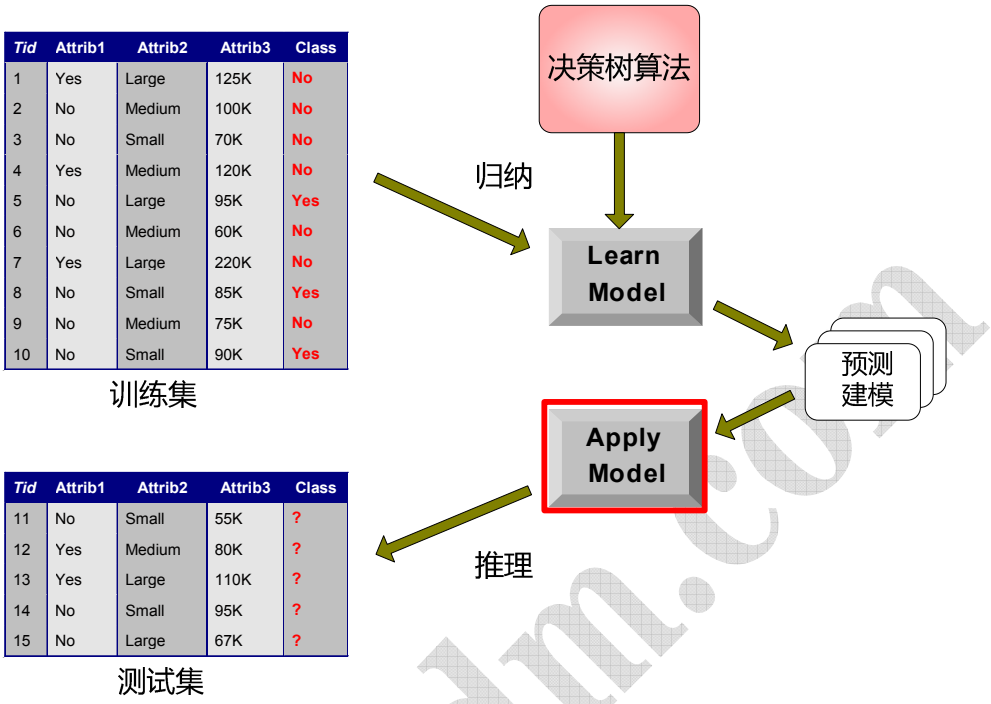


图 1-7 分类与回归建模原理

常用分类与回归算法见表 1-2。

表 1-2 主要分类与回归算法分类

类别	包括的主要算法
回归分析	线性回归、逐步回归、逻辑回归、保序回归
神经网络	BP 神经网络、LM 神经网络、RBF 神经网络、FNN 神经网络、ANFIS 神经网络、WNN 神经网络
决策树	CART 决策树、ID3 决策树、C4.5 决策树、决策表
贝叶斯网络	朴素贝叶斯网络、贝叶斯信念网络
组合算法	AdaBoostM1 算法、KStar 算法、SVM 支持向量机、K-最近邻分类
数据降维	主成分分析、因子分析

1.5.2. 聚类分析

我们经常会碰到这样的问题：

1、如何通过一些特定的症状归纳某类特定的疾病？

2、谁是银行信用卡的黄金客户？

3、谁喜欢打国际长途，在什么时间，打到那里？

4、对住宅区进行聚类，确定自动提款机 ATM 的安放位置。

5、如何对用户 WAP 上网行为进行分析，通过客户分群，进行精确营销？

除此之外，促销应该针对哪一类客户，这类客户具有哪些特征？这类问题往往是在促销前首要

www.tipdm.com

解决的问题，对整个客户做分群，将客户分组在各自的群组里，然后对每个不同的群组，采取不同的营销策略。这些都是聚类分析的例子。

聚类（Clustering）：聚类与分类不同，聚类分析是在没有给定划分类的情况下，根据信息相似度进行信息聚类的一种方法，故聚类又称为无指导的学习。而分类则需要先定义类别和训练样本，是有指导的学习。聚类就是将数据划分或分割成相交或者不相交的群组的过程，通过确定数据之间在预先指定的属性上的相似性就可以完成聚类任务。聚类的输入是一组未被标记的数据，根据数据自身的距离或相似度进行划分。划分的原则是保持最大的组内相似性和最小的组间相似性，也就是使得不同聚类中的数据尽可能地不同，而同一聚类中的数据尽可能地相似。比如根据股票价格的波动情况，可以将股票分成不同的类，总共可以分成几类，各类包含哪些股票，每一类的特征是什么，这对投资者、尤其对投资基金来说，可能就是很重要的信息。当然，聚类除了将样本分类外，它还可以完成孤立点挖掘，如将其应用于网络入侵检测或金融风险欺诈探测中。

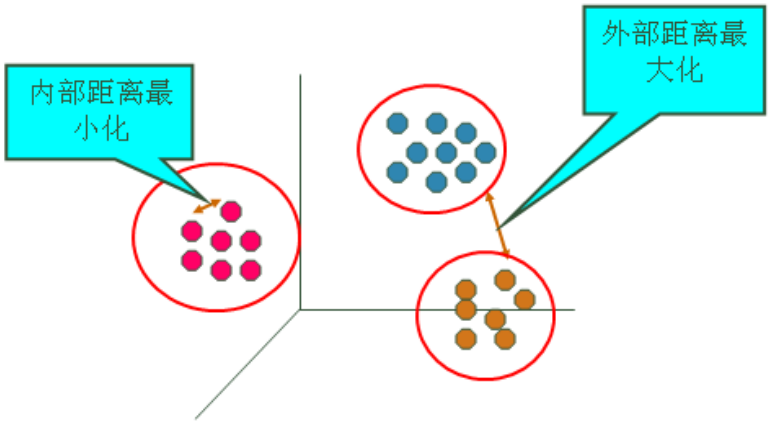


图 1-8 聚类分析建模原理

常用聚类算法见表 1-3。

表 1-3 主要聚类算法分类

类别	包括的主要算法
划分(分裂)方法	K-MEANS 算法(K-平均)、K-MEDOIDS 算法(K-中心点)、CLARANS 算法(基于选择的算法)
层次方法	BIRCH 算法(平衡迭代规约和聚类)、CURE 算法(代表点聚类)、CHAMELEON 算法(动态模型)
基于密度的方法	DBSCAN 算法(基于高密度连接区域)、DENCLUE 算法(密度分布函数)、OPTICS 算法(对象排序识别)
基于网格的方法	STING 算法(统计信息网络)、CLIQUE 算法(聚类高维空间)、WAVE-CLUSTER 算法(小波变换)
基于模型的方法	统计学方法、神经网络方法

1.5.3. 关联规则

我们经常会碰到这样的问题：

1、商业销售上，如何通过交叉销售，以得到更大的收入？

2、保险方面，如何分析索赔要求，发现潜在的欺诈行为？

3、银行方面，如何分析顾客消费行业，以便有针对性的向其推荐感兴趣的服务？

4、哪些制造零件和设备设置与故障事件关联？

5、哪些病人和药物属性与结果关联？

6、哪些商品是已经购买商品 A 的人最有可能购买的？

除此之外，人们希望从大量的商业交易记录中发现有价值的关联知识以帮助进行商品目录的设计、交叉营销或帮助进行其它有关的商业决策。在商业销售上，关联规则可用于交叉销售，以得到更大的收入；在保险业务方面，如果出现了不常见的索赔要求组合，则可能为欺诈，需要作进一步的调查；在医疗方面，可找出可能的治疗组合；在银行方面，对顾客进行分析，可以推荐感兴趣的服务等等。这些都属于关联规则挖掘问题，关联规则挖掘的目的就在于在一个数据集中找出项之间的关系，从大量的数据中挖掘出有价值描述数据项之间相互联系的有关知识。随着收集和存储在数据库中的数据规模越来越大，人们对从这些数据中挖掘相应的关联知识越来越有兴趣。

关联规则(Association)：揭示数据之间的相互关系，而这种关系没有在数据中直接表示出来。关联分析的任务就是发现事物间的关联规则，或称相关程度。关联规则的一般形式是:

如果 A 发生，则 B 有百分之 C 的可能发生。C 称为关联规则的置信度（Confidence)。

关联分析能寻找数据库中大量数据的相关联系，常用的 2 种技术为关联规则和序列模式。关联规则是发现一个事物与其他事物间的相互关联性或相互依赖性，可用于如分析客户在超市买牙刷的同时又买牙膏的可能性；序列模式分析将重点放在分析数据之间的前后因果关系，如买了电脑的顾客则会在 3 个月内买杀毒软件。

关于“尿布与啤酒”的故事：在一家超市里，有一个有趣的现象：尿布和啤酒赫然摆在一起出售。但是这个奇怪的举措却使尿布和啤酒的销量双双增加了。这不是一个笑话，而是发生在美国沃尔玛连锁店超市的真实案例，并一直为商家所津津乐道。沃尔玛拥有世界上最大的数据仓库系统，为了能够准确了解顾客在其门店的购买习惯，沃尔玛对其顾客的购物行为进行购物篮分析，想知道顾客经常一起购买的商品有哪些。沃尔玛数据仓库里集中了其各门店的详细原始交易数据。在这些原始交易数据的基础上，沃尔玛利用数据挖掘方法对这些数据进行分析和挖掘。一个意外的发现是：跟尿布一起购买最多的商品竟是啤酒！经过大量实际调查和分析，揭示了一个隐藏在“尿布与啤酒”背后的美国人的一种行为模式：在美国，一些年轻的父亲下班后经常要到超市去买婴儿尿布，而他们中有 30%～40%的人同时也为自己买一些啤酒。产生这一现象的原因是：美国的太太们常叮嘱她们的丈夫下班后为小孩买尿布，而丈夫们在买尿布后又随手带回了他们喜欢的啤酒。

按常规思维，尿布与啤酒风马牛不相及，若不是借助数据挖掘技术对大量交易数据进行挖掘分析，沃尔玛是不可能发现数据内在这一有价值的规律的。

下面对常用关联规则算法进行简要介绍。

1.5.4. 时序模式

我们经常会碰到这样的问题：

1、下个月的商品销量、销售额或库存量是多少？

2、明天广州市的最高用电负荷是多少？

时序模式是描述基于时间或其他序列的经常发生的规律或趋势，并对其建模。与回归一样，它也是用已知的数据预测未来的值，但这些数据的区别是变量所处时间的不同。序列模式将关联模式和时间序列模式结合起来，重点考虑数据之间在时间维度上的关联性。时序模式包含时间序列分析和序列发现。

时间序列分析（Time Series)：用已有的数据序列预测未来。在时间序列分析中，数据的属性

值是随着时间不断变化的。回归不强调数据间的先后顺序，而时间序列要考虑时间特性，尤其要考虑时间周期的层次，如天、周、月、年等，有时还要考虑日历的影响，如节假日等。

序列发现：序列发现用于确定数据之间与时间相关的序列模式。这些模式与在数据（或者事件）中发现的相关的关联规则很相似，只是这些序列是与时间相关的。

时间序列的组合成分

时间序列的变化主要受到长期趋势、季节变动和不规则变动这四个因素的影响。其中：

✧ 长期趋势因素（T）

长期趋势因素（T）反映了经济现象在一个较长时间内的发展方向，它可以在一个相当长的时间内表现为一种近似直线的持续向上或持续向下或平稳的趋势。

✧ 季节变动因素（S）

季节变动因素（S）是经济现象受季节变动影响所形成的一种长度和幅度固定的周期波动。

✧ 周期变动因素（C）

周期变动因素也称循环变动因素，它是受各种经济因素影响形成的上下起伏不定的波动。

✧ 不规则变动因素（I）

不规则变动又称随机变动，它是受各种偶然因素影响所形成的不规则变动。

时间序列的组合模型

时间序列 Y 可以表示为以上四个因素的函数，即：

$$y_t = f(T_t, S_t, C_t, I_t) \tag{1-1}$$

时间序列分解的方法有很多，较常用的模型有加法模型和乘法模型。

✧ 加法模型

假定时间序列是基于 4 种成份相加而成的。长期趋势并不影响季节变动。若以 Y 表示时间序列，则加法模型为：

$$y_t = T_t + S_t + C_t + I_t \tag{1-2}$$

✧ 乘法模型

假定时间序列是基于 4 种成份相乘而成的。假定季节变动与循环变动为长期趋势的函数。该模型的方程式为：

$$y_t = T_t \times S_t \times C_t \times I_t \tag{1-3}$$

乘法模型分解的基本步骤是：

① 运用移动平均法剔除长期趋势和周期变化，得到序列 TC。然后再用按月（季）平均法求出季节指数 S。

② 作散点图，选择适合的曲线模型拟合序列的长期趋势，得到长期趋势 T。

③ 计算周期因素 C。用序列 TC 除以 T 即可得到周期变动因素 C。

④ 将时间序列的 T、S、C 分解出来后，剩余的即为不规则变动，即：

$$I = \frac{Y}{TSC} \tag{1-4}$$

常用时序算法及适应范围见表 1-4。

表 1-4 常用时序算法及适应范围

方法	时间范围	适用情况
一元线性回归预测法	短、中期	自变量与因变量之间存在线性关系
多元线性回归预测法	短、中期	因变量与两个或两个以上自变量之间存在线性关系
非线性回归预测法	短、中期	因变量与一个自变量或多个其它自变量之间存在某种非线性关系
趋势外推法	中期到长期	当被预测项目的有关变量用时间表示时，用非线性回归
移动平均法	短期	不带季节变动的反复预测
指数平滑法	短期	具有或不具有季节变动的反复预测
平稳时间序列预测法	短期	适用于任何序列的发展型态的一种高级预测方法
灰色预测法	短、中期	适用于时间序列的发展呈指数型趋势

1.5.5. 偏差检测

偏差是对差异和极端特例的表述，如分类中的反常实例、聚类外的离群值、不满足规则的特例等，见图 1-9。大部分数据挖掘方法都将这种差异信息视为噪声而丢弃，然而在一些应用中，罕见的数据可能比正常的数据更有用。

偏差检测是用来发现与正常情况不同的异常和变化，并进一步分析这种变化是否是有意的诈骗行为，还是正常的变化。如果是异常行为，则需提示预防措施，尽早防范。

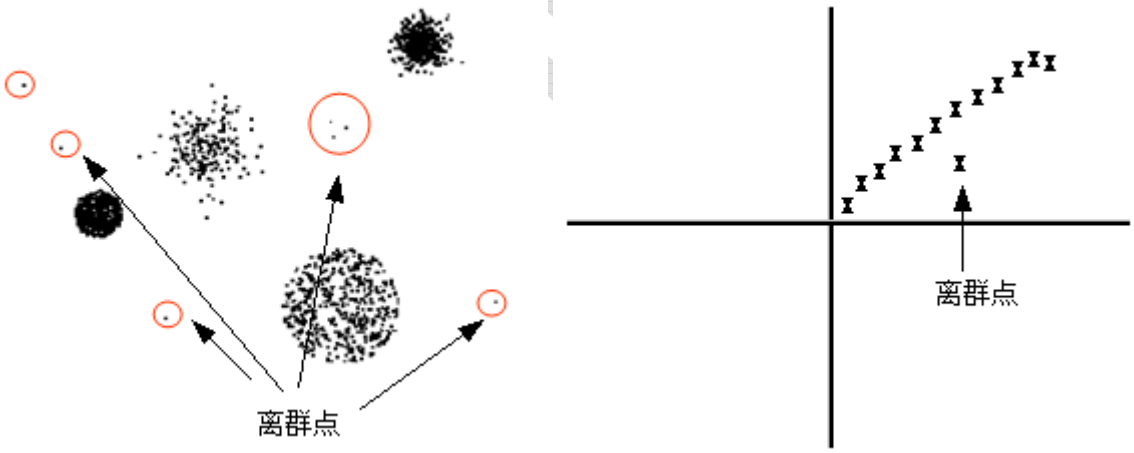


图 1-9 偏差检测示意图

1.6. 数据挖掘建模过程

数据挖掘是从海量数据中提取隐含在其中的有用信息和知识的过程。各运营支撑系统所积累的海量历史数据是企业的一笔宝贵财富，谁能正确地挖掘与分析隐含在数据中的知识，谁就能更好地向用户提供产品与服务，从而在竞争中脱颖而出。数据挖掘提供了从数据到价值的解决方案："数据 + 工具 + 方法 + 目标 + 行动 = 价值"。数据挖掘目前已有一系列应用：分类分析、聚类分析、预测分析、偏差分析、关联分析和时序模式等，这些应用涉及到的技术和工具各不相同，然而却可以依据统一的方法论来实行，并可以协同作战，解决许多有价值的商业问题，图 1-10 是进行数据挖掘的一般过程。

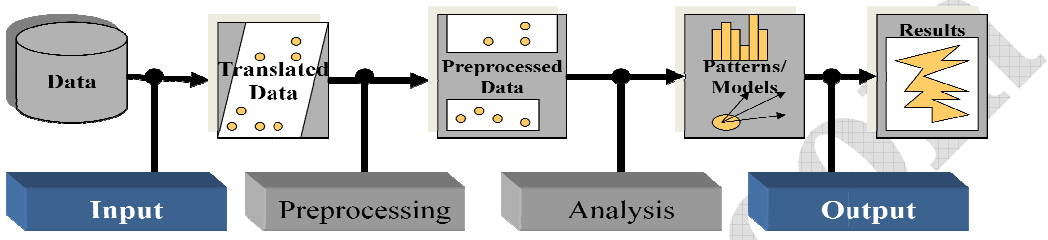


图 1-10 数据挖掘过程

太普公司在多年的数据挖掘项目实施过程中，积累了一套行之有效的数据挖掘方法论，数据挖掘建模过程如图 1-11。

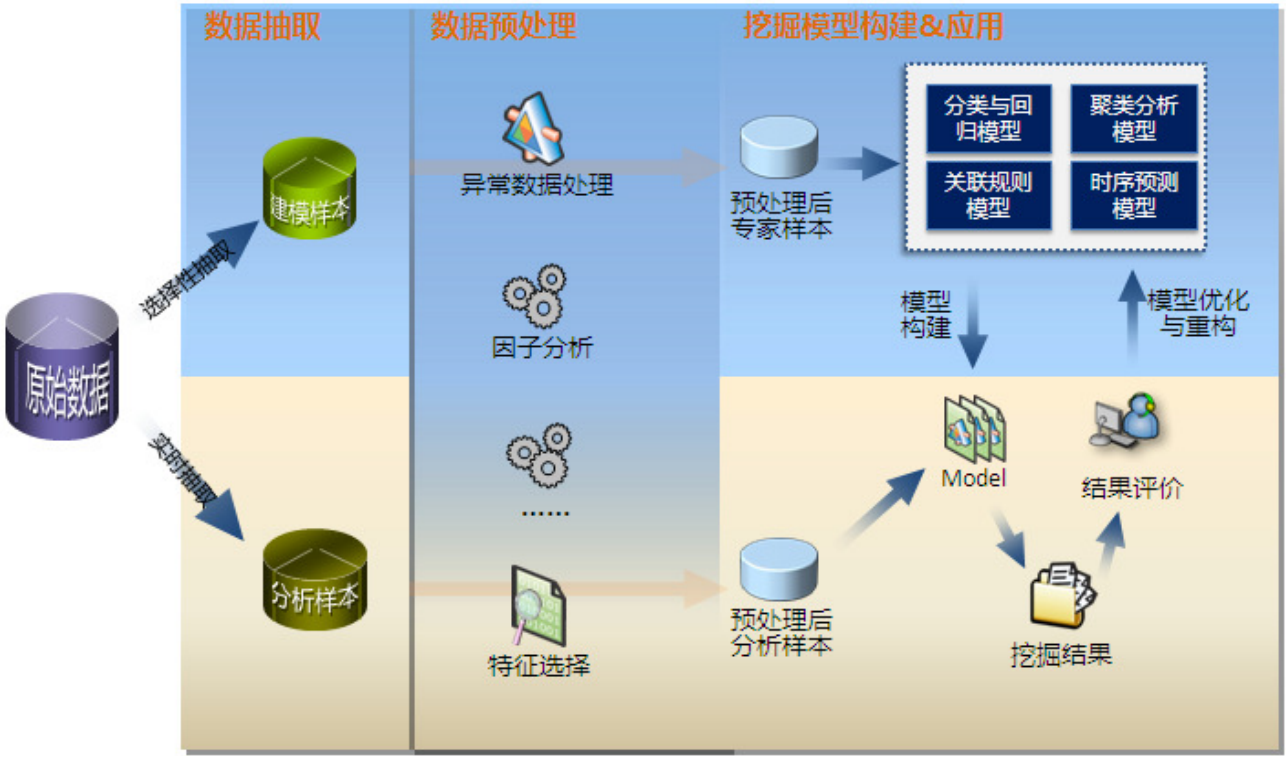


图 1-11 数据挖掘建模过程

1.6.1. 定义挖掘目标

进行数据挖掘，首先必须分析应用领域，包括应用中的各种知识和应用目标。了解相关领域的有关情况，熟悉背景知识，弄清用户要求。要想充分发挥数据挖掘的价值，必须要对目标有一个清

晰明确的定义，即决定到底想干什么。否则，很难得到正确的结果。

1.6.2. 数据取样

1、哪些数据源可用，哪些数据与当前问题相关？

2、取样数据是否满足本次数据挖掘目标？

3、如何保证取样数据的质量？是否在足够范围内有代表性？

4、数据样本取多少合适？如何分类（训练集、验证集、测试集）？

在明确了需要进行数据挖掘的预测目标后，接下来就需要从企业大量数据中抽取出一个与预测目标相关的样本数据子集。抽取数据的标准，一是相关性，二是可靠性，三是最新性。而不是动用全部企业数据。这就象在对开采出来矿石首先要进行选矿一样。通过数据样本的精选，不仅能减少数据处理量，节省系统资源，而且能通过数据的筛选，使你想要它反映的规律性更加凸现出来。

通过数据取样，要把好数据的质量关。在任何时候都不要忽视数据的质量，即使是从一个数据仓库中进行数据取样，也不要忘记检查其质量如何。因为通过数据挖掘是要探索企业运作的规律性的，原始数据有误，就很难从中探索规律性。若真的从中还探索出来了什么“规律性”，再依此去指导工作，则很可能是在进行误导。若从正在运行着的系统中进行数据取样，则更要注意数据的完整性和有效性。

衡量取样数据质量的标准包括：

✧ 资料完整无缺，各类指标项齐全。

✧ 数据准确无误，反映的都是正常（而不是反常）状态下的水平。

对获取的数据，可再从中作抽样操作。抽样的方式是多种多样的，有：随机抽样、等距抽样、分层抽样、从起始顺序抽样和分类抽样等方式。

✧ 随机抽样

在采用随机抽样方式时，数据集中的每一组观测值都有相同的被抽样的概率。如按 10%的比例对一个数据集进行随机抽样，则每一组观测值都有 10%的机会被取到。

✧ 等距抽样

如按 5%的比例对一个有 100 组观测值的数据集进行等距抽样，则有：100 / 5 = 20，等距抽样方式是取第 20、40、60、80 和第 100 等五组观测值。

✧ 分层抽样

在这种抽样操作时，首先将样本总体分成若干层次（或者说分成若干个子集）。在每个层次中的观测值都具有相同的被选用的概率，但对不同的层次您可设定不同的概率。这样的抽样结果可能具有更好的代表性，进而使模型具有更好的拟合精度。

✧ 从起始顺序抽样

这种抽样方式是从输入数据集的起始处开始抽样。抽样的数量可以给定一个百分比，或者就直接给定选取观测值的组数。

✧ 分类抽样

在前述几种抽样方式中，抽样的单位都是一组观测值。分类抽样的单位是一类观测值。这里的分类是按观测值的某种属性进行区分。如按客户名称分类、按地址区域分类等。显然在同一类中可能会有多组观测值。分类抽样的选取方式就是前面所述的几种方式，只是抽样以类为单位。

1.6.3. 数据探索

1、数据中是否存在残缺值、异常值？

2、抽样数据是否存在明显的规律和趋势？数据是否存在周期性？

3、样本数据是否按时间顺序排列？不同时间粒度(年、月、日)数据是否存在明显分组？

4、样本属性与挖掘目标之间是否存在相关性？

前面所叙述的数据取样，多少是带着人们对如何达到数据挖掘目的的先验的认识进行操作的。当我们拿到了一个样本数据集后，它是否达到我们原来设想的要求；其中有没有什么明显的规律和趋势；有没有出现你所从未设想过的数据状态；因素之间有什么相关性；它们可区分成怎样一些类别……这都是要首先探索的内容。

这里的数据探索，就是我们通常所进行的深入调查的过程。你最终要达到的目的可能是要搞清多因素相互影响的，十分复杂的关系。但是，这种复杂的关系不可能一下子建立起来。一开始，可以先观察众多因素之间的相关性；再按其相关的程度，以了解它们之间相互作用的情况。这些探索、分析，并没有一成不变操作规律性；相反，是要有耐心的反复的试探，仔细的观察。在此过程中，你原来的专业技术知识是非常有用的，它会帮助你进行有效的观察。但是，你也要注意，不要让你的专业知识束缚了你对数据特征观察的敏锐性。可能实际存在着你的先验知识认为不存在的关系。假如你的数据是真实可靠的话，那末你绝对不要轻易地否定数据呈现给你的新关系。很可能这里就是发现的新知识！有了它，也许会导引你在此后的分析中，得出比你原有的认识更加符合实际的规律性知识。假如在你的操作中出现了这种情况，应当说，你的数据挖掘已挖到了有效的矿脉。

对所抽取的样本数据进行探索、审核和必要的加工处理，是保证预测质量所必需的。可以说，预测的质量不会超过抽取样本的质量。数据探索和预处理的目的是为了保证样本数据的质量，从而为保证预测质量打下基础。

1.6.4. 预处理

1、当采样数据维度过大，如何进行降维处理？

2、采样数据中的缺失值如何处理？

由于采样数据中常常包含许多含有噪声、不完整、甚至是不一致的数据。显然对数据挖掘所涉及的数据对象必须进行预处理。那么如何对数据进行预处理以改善数据质量，并最终达到完善最终的数据挖掘结果之目的呢？

数据预处理主要包括:数据清洗、数据集成、数据转换、数据消减、坏数据处理、降维处理等。

✧ **数据筛选**

通过数据筛选可从观测值样本中筛选掉不希望包括进来的观测值。对于离散变量可给定某一类的类值说明此类观测值是要排除于抽样范围之外的。对于连续变量可指定其值大于或小于某值时的这些组观测值是要排除于抽样范围之外的。

✧ **数据变量转换**

将某一个数据进行某种转换操作，然后将转换后的值作为新的变量存放在样本数据中。转换的目的是为了使数据和将来要建立的模型拟合的更好。例如，原来的非线性模型线性化、加强变量的稳定性等。可进行取幂、对数、开方等转换。当然，也可给定一个公式进行转换。

✧ **缺失值处理**

造成数据缺失的原因主要有：

1) 有些信息暂时无法获取，或者获取信息的代价太大。
2) 有些信息是被遗漏的。可能是因为输入时认为不重要、忘记填写了或对数据理解错误而遗漏，也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障、一些人为因素等原因而丢失了。

3) 有些对象的某个或某些属性是不可用的。也就是说，对于这个对象来说，该属性值是不存在的，如一个未婚者的配偶姓名、一个儿童的固定收入状况等。

数据缺失在许多研究领域都是一个复杂的问题。对数据挖掘来说，空值的存在，造成的影响主要有：首先，系统丢失了大量的有用信息；第二，系统中所表现出的不确定性更加显著，系统中蕴涵的确定性成分更难把握；第三，包含空值的数据会使挖掘过程陷入混乱，导致不可靠的输出。

数据挖掘算法本身更致力于避免数据过分适合所建的模型，这一特性使得它难以通过自身的算法去很好地处理不完整数据。因此，空缺的数据需要通过专门的方法进行推导、填充等，以减少数据挖掘算法与实际应用之间的差距。

处理不完备数据集的方法主要有以下三大类：

1、删除元组

也就是将存在遗漏信息属性值的对象（元组，记录）删除，从而得到一个完备的信息表。这种方法简单易行，在对象有多个属性缺失值、被删除的含缺失值的对象与信息表中的数据量相比非常小的情况下是非常有效的，类标号（假设是分类任务）缺少时通常使用。然而，这种方法却有很大的局限性。它是以减少历史数据来换取信息的完备，会造成资源的大量浪费，丢弃了大量隐藏在这些对象中的信息。在信息表中本来包含的对象很少的情况下，删除少量对象就足以严重影响到信息表信息的客观性和结果的正确性；当每个属性空值的百分比变化很大时，它的性能非常差。因此，当遗漏数据所占比例较大，特别当遗漏数据非随机分布时，这种方法可能导致数据发生偏离，从而引出错误的结论。

2、数据补齐

这类方法是用一定的值去填充空值，从而使信息表完备化。通常基于统计学原理，根据决策表中其余对象取值的分布情况来对一个空值进行填充，譬如用其余属性的平均值来进行补充等。数据挖掘中常用的有以下几种补齐方法：

（1）人工填写

由于最了解数据的还是用户自己，因此这个方法产生数据偏离最小，可能是填充效果最好的一种。然而一般来说，该方法很费时，当数据规模很大、空值很多的时候，该方法是不可行的。

（2）特殊值填充

将空值作为一种特殊的属性值来处理，它不同于其他的任何属性值。如所有的空值都用“unknown”填充。这样将形成另一个有趣的概念，可能导致严重的数据偏离，一般不推荐使用。

（3）平均值填充

将信息表中的属性分为数值属性和非数值属性来分别进行处理。如果空值是数值型的，就根据该属性在其他所有对象的取值的平均值来填充该缺失的属性值；如果空值是非数值型的，就根据统计学中的众数原理，用该属性在其他所有对象的取值次数最多的值(即出现频率最高的值)来补齐该缺失的属性值。另外有一种与其相似的方法叫条件平均值填充法，在该方法中，缺失属性值的补齐同样是靠该属性在其他对象中的取值求平均得到，但不同的是用于求平均的值并不是从信息表所有对象中取，而是从与该对象具有相同决策属性值的对象中取得。这两种数据的补齐方法，其基本的出发点都是一样的，以最大概率可能的取值来补充缺失的属性值，只是在具体方法上有一点不同。与其他方法相比，它是用现存数据的多数信息来推测缺失值。

（4）热卡填充（或就近补齐）

对于一个包含空值的对象，热卡填充法在完整数据中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似进行判定。该方法概念上很简单，且利用了数据间的关系来进行空值估计。这个方法的缺点在于难以定义相似标准，主观因素

较多。

（5）K 最近距离邻法

先根据欧式距离或相关分析来确定距离具有缺失数据样本最近的 K 个样本，将这 K 个值加权平均来估计该样本的缺失数据。

（6）使用所有可能的值填充

这种方法是用空缺属性值的所有可能的属性取值来填充，能够得到较好的补齐效果。但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大，可能的测试方案很多。另有一种方法，填补遗漏属性值的原则是一样的，不同的只是从决策相同的对象中尝试所有的属性值的可能情况，而不是根据信息表中所有对象进行尝试，这样能够在一定程度上减小原方法的代价。

（7）组合完整化方法

这种方法是用空缺属性值的所有可能的属性取值来试，并从最终属性的约简结果中选择最好的一个作为填补的属性值。这是以约简为目的的数据补齐方法，能够得到好的约简结果；但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大。另一种称为条件组合完整化方法，填补遗漏属性值的原则是一样的，不同的只是从决策相同的对象中尝试所有的属性值的可能情况，而不是根据信息表中所有对象进行尝试。条件组合完整化方法能够在一定程度上减小组合完整化方法的代价。在信息表包含不完整数据较多的情况下，可能的测试方案将巨增。

（8）回归

基于完整的数据集，建立回归方程（模型）。对于包含空值的对象，将已知属性值代入方程来估计未知属性值，以此估计值来进行填充。当变量不是线性相关或预测变量高度相关时会导致有偏差的估计。

（9）期望值最大化方法

即 EM 算法，是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。在每一迭代循环过程中交替执行两个步骤：E 步（Excepctaion step,期望步），在给定完全数据和前一次迭代所得到的参数估计的情况下计算完全数据对应的对数似然函数的条件期望；M 步（Maximzation step，极大化步），用极大化对数似然函数以确定参数的值，并用于下步的迭代。算法在 E 步和 M 步之间不断迭代直至收敛，即两次迭代之间的参数变化小于一个预先给定的阈值时结束。该方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。

（10）多重填补

多重填补方法分为三个步骤：①为每个空值产生一套可能的填补值，这些值反映了无响应模型的不确定性；每个值都被用来填补数据集中的缺失值，产生若干个完整数据集合。②每个填补数据集合都用针对完整数据集的统计方法进行统计分析。③对来自各个填补数据集的结果进行综合，产生最终的统计推断，这一推断考虑到了由于数据填补而产生的不确定性。该方法将空缺值视为随机样本，这样计算出来的统计推断可能受到空缺值的不确定性的影响。该方法的计算也很复杂。

（11）C4.5 方法

通过寻找属性间的关系来对遗失值填充。它寻找之间具有最大相关性的两个属性，其中没有遗失值的一个称为代理属性，另一个称为原始属性，用代理属性决定原始属性中的遗失值。这种基于规则归纳的方法只能处理基数较小的名词型属性。

补齐处理只是将未知值补以我们的主观估计值，不一定完全符合客观事实，在对不完备信息进行补齐处理的同时，我们或多或少地改变了原始的信息系统。而且，对空值不正确的填充往往将新的噪声引入数据中，使挖掘任务产生错误的结果。因此，在许多情况下，我们还是希望在保持原始信息不发生变化的前提下对信息系统进行处理。这就是第三种方法：

3、不处理

直接在包含空值的数据上进行数据挖掘。这类方法包括贝叶斯网络和人工神经网络等。

贝叶斯网络是用来表示变量间连接概率的图形模式，它提供了一种自然的表示因果信息的方法，用来发现数据间的潜在关系。在这个网络中，用节点表示变量，有向边表示变量间的依赖关系。贝叶斯网络仅适合于对领域知识具有一定了解的情况，至少对变量间的依赖关系较清楚的情况。否则直接从数据中学习贝叶斯网的结构不但复杂性较高（随着变量的增加，指数级增加），网络维护代价昂贵，而且它的估计参数较多，为系统带来了高方差，影响了它的预测精度。当在任何一个对象中的缺失值数量很大时，存在指数爆炸的危险。

✧ **坏数据处理**

如果抽取数据中存在坏数据（脏数据），则需要对坏数据进行预处理。通常的做法是采用绝对均值法或莱因达法等对样本中的坏点数据进行剔除处理。

✧ **属性选择**

属性选择是数据预处理的一部分，因采集的数据中的每一个属性对于整个数据的挖掘结果的作用不是完全对等的，一些属性对结果的影响占主导地位，一些属性对结果的影响不大，甚至没有影响，采用相应的算法，对数据的属性值进行评估，如去掉某个属性后对挖掘结果无影响，从而减少后续挖掘算法的运行时间，同时也能有效地去除数据中含有的噪声数据。

如果建模数据集的维度较高，或输入属性与输出属性的相关性不明确时，对其进行属性选择是必要的步骤。综合考虑应用实现的复杂性，可使用标准属性选择方法，用一个评估标准对属性的有用性进行度量，并选择其最有用的一部分属性作为下一部分算法的输入。为了确定属性选择的标准，可选用多种方法进行属性评价，选出 5-10 个不同的属性，然后对于处理后的数据集进行测试，综合评价维度约减前后的预测模型的性能及效果。

属性选择是用于对属性进行筛选，搜索数据集中全部属性的所有可能组合，找出预测效果最好的那一组属性，一般在预处理阶段进行。为实现这一目标，必须设定两个东西：属性评估器和搜索策略。属性评估器是对属性/属性子集进行评估确定，决定了怎样给一组属性安排一个表示它们好坏的值。搜索策略确定搜索算法，决定了要怎样进行搜索。

✧ **数据规约**

把繁杂的样本数据信息进行数据规约，简化以后存储在数据表中，避免数据的不一致性。

1.6.5. 模式发现

1、本次数据挖掘属于应用中的哪类问题（分类？聚类？关联规则？时序模式？）

2、用哪种算法进行预测建模？

预测建模的前提是在数据集中发现模式，比如关联规则，分类模型，序列模式等。

在问题进一步明确化的基础上，你就可以按照问题的具体要求来审视你的数据集了，看它是否适应你的问题的需要。

针对问题的需要可能要对数据进行增删；也可能按照你对整个数据挖掘过程的新认识，要组合或者生成一些新的变量，以体现对状态的有效的描述。

在问题进一步明确；数据结构和内容进一步调整的基础上，下一步数据挖掘应采用的技术手段就更加清晰、明确了。

1.6.6. 预测建模

1、针对本些数据挖掘该选用哪个模型算法？

2、预测模型的构建步骤是怎样的？

这一步是数据挖掘工作的核心环节。预测模型是对采样数据轨迹的概括，它反映的是采样数据内部结构的一般特征，与该采样数据的具体结构并不完全吻合。模型的具体化就是预测公式，公式可以产生与观察值有相似结构的输出，这就是预测值。预测模型是多种多样的，可以适用于不同结构的样本数据，因此，对一个具体采样数据，就有选择适当预测模型的问题。正确选择预测模型在数据挖掘过程中是具有关键性的一步。有时由于模型选择不当，造成预测误差过大，就需要改换模型。必要时，可同时采用几种预测模型进行运算，以便对比、选择。对建立模型来说，要记住的最重要的事就是它是一个反复的过程。需要仔细考察不同的模型以判断哪个模型对问题最有用。

模型的构建主要包括模型建立、模型训练、模型预测 3 个步骤。

1.6.7. 模型评价

1、 模型评估的目的是什么？

2、 如何评价预测模型的效果？通过什么评价指标来衡量？

从前面的预测建模过程中将会得出一系列的分析结果、模式或模型。同一个采样数据可以利用多种数据分析方法和模型进行分析，模型评估的目的之一就是从这些模型中自动找出一个最好的模型出来，另外就是要对模型进行针对业务的解释和应用。

模型效果评价的办法之一就是直接使用原来建立模型的样本数据来进行检验。假如这一关都通不过的话，那末所建立的决策支持信息价值就不太大了。一般来说，在这一步应得到较好的评价。这说明你确实从这批数据样本中挖掘出了符合实际的规律性。

另一种办法是另外找一批数据，已知这些数据是反映客观实际的规律性的。这次的检验效果可能会比前一种差。差多少是要注意的，若是差到所不能容忍程度，那就要考虑第一次构建的样本数据是否具有充分的代表性；或是模型本身是否够完善。这时候可能要对前面的工作进行反思了。若这一步也得到了肯定的结果时，那所建立的数据挖掘模型应得到很好的评价了。

练习题

1、什么是数据挖掘，数据挖掘有哪些主要应用？

2、请列举 3 例数据挖掘应用的例子。

www.cnpn.com

第2章 案例分析

2.1. 金融电信

2.1.1. 案例一 基于公司价值评价的证券策略投资

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

随着我国市场经济的进一步发展，证券投资已成为企业与个人投资的热点，而证券投资是收益与风险并存的一大投资方向。由于受到国内外政治经济环境以及企业自身等各种因素的影响，股票价格总是处在不断波动之中，各种因素的影响方式又极其复杂，所以股票价格涨跌无序，有很大的随机游走性。一些经济学家已经用实证的方法证明我国股市基本属于弱有效市场。因此，对股票价格的各种分析，预测都存在一定的局限性，其结果也可能存在较大误差。

在众多上市公司中，如何选择行业股票，是股票投资者的热门话题，而在同一行业上市公司中，也是良莠不齐，投资者应该如何分析上市公司的财务状况和发展潜力？如何才能选择收益大而风险小的上市公司进行投资？

本案例的挖掘目标是：

1） 构建一个上市公司内在价值评价模型

2） 遍历所有股票得到最优投资股清单

3） 对所选股票提供最优投资策略

分析方法与过程

股票自动推荐建模过程主要由模型构建、模型修正和策略投资三部分构成，见图 2-1。

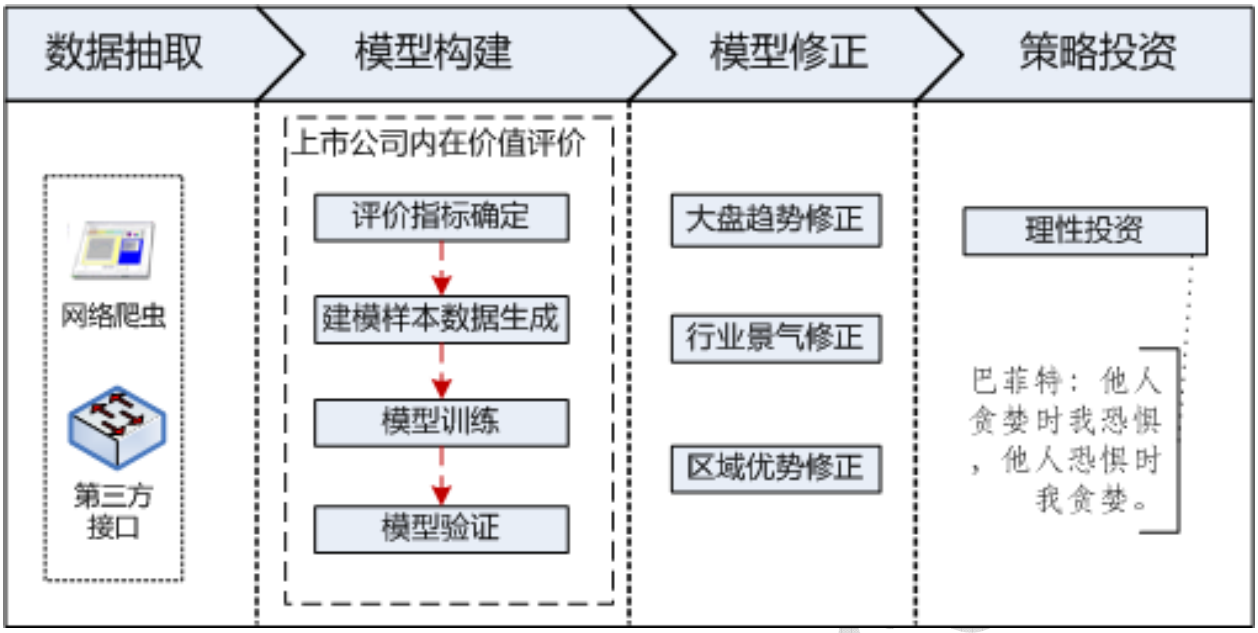


图 2-1 股票自动推荐建模过程

对于股票投资选择这类经济决策问题，单纯地运用统计分析又难于达到决策的目的，为此，本文将数据挖掘与多指标决策结合起来，首先运用因子分析法对上市公司投资价值的多项指标进行特征选择，将上市公司繁杂的高维价值指标浓缩为低维指标来处理，并进而构建基于上市公司内在价值的评价模型，从而为机构管理者和决策者提供了一种科学而合理的决策依据和决策方法。

股票推荐系数评价需要从上市公司内在价值、股票大盘趋势、行业景气程度、区域优势政策等方面进行综合评定，股票推荐系数计算公式：

$$Y_{股票推荐} = Y_{内在价值} \times \gamma_{大盘趋势} \times \gamma_{行业景气} \times \gamma_{区域优势}$$

式中：

$Y_{股票推荐}$为股票推荐综合指数

$Y_{内在价值}$为公司内在价值指数

$\gamma_{大盘趋势}$为大盘总体趋势修正系数

$\gamma_{行业景气}$为行业景气修正系数

$\gamma_{区域优势}$为区域优势修正系数

各指标具体描述见表 2-1。

表 2-1 股票推荐综合评价因素构成

序号	指标类别	指标描述	取值范围
1	公司内在价值指数	体现上市公司综合盈利能力	0.001~0.999
2	大盘总体趋势修正系数	体现大盘总体走势情况	0.5~1.5，默认为 1
3	行业景气修正系数	体现行业（板块）景气情况	0.5~1.5，默认为 1
4	区域优势修正系数	体现政策对区域经济扶持情况	0.5~1.5，默认为 1

1、上市公司内在价值综合评价

要分析上市公司的内在价值，确定公司是否值得投资，首先要确定要分析上市公司的哪些指标。如何选择正确而且有效的指标？根据多年的投资经验，一个企业是否值得投资，应从以下几方面进行考虑:

1）公司主营业务突出。即公司将资金集中在专门的市场、专门的产品或提供特别的服务和技术，

而此种产品和技术基本被公司垄断，其他公司很难插足投入。

2）产品和技术投入的市场具有美好灿烂的发展前景。比如，世界电脑市场的庞大需求使美国股票“IBM”和“苹果”成为当时的主力股；香港因为人多地少，加上地产发展商的眼光，使地产上市公司利润和股价上升数倍。

3）上市公司为小规模或股本结构不大的公司。由于股本小，扩张的需求强烈，适宜大比例地送配股，也使股价走高后因送配而除权，在除权后能填权并为下一次送配做准备。如此周而复始，股价看似不高，却已经翻了数倍。

4）上市公司所经营的行业不容易受经济周期所影响。即不但在经济增长的大环境中能保持较高的增长率，而且也不会因为宏观经济衰退而轻易遭淘汰。

5）公司经营灵活。生产工艺创新，尤其是在高科技范围，它们比其他的公司有高人一筹的生产技巧，容易适应社会转变所带来的影响。

6）公司肯花大力气进行新技术研究和创新。肯花钱或有能力不断发展新产品，始终把研究和发展新技术和新产品放在公司重要地位，如此，公司才有能力抵御潜在的竞争对手，使自己产品不断更新而立于不败之地。

7）公司利润较高或暂不高。但由于发展前景可观而有可能带来增长性很高的潜在利润。因为利润成长性比高利润的现实性更为重要，动态市盈率的持续预期降低比低企的市盈率更为吸引人。

8）公司的股本结构应具有流通盘小、总股本尚不太大。且总股本中持有部分股份的大股东应保持稳定，大股东实力较强对公司发展具有促进作用，甚至有带来优质资产的可能。

9）公司仍未被大多数人所认识。虽然有人听说过公司名称，但大多数投资者都不知道该公司的股本结构、潜质、经营手法、生产技巧和利润潜力等，太出名或已经被投资者广为知道的上市公司不可能成为潜力股。

10）公司股在行情初起时股价应不高。至少应在中、低价股的范围内寻找，因为从几元的股票翻倍到数十元较容易，而几十元股票要涨到数百元则对大众心理压力较大

基于以上观点，我们主要从公司的成长性、盈利能力、股本扩张能力三项上层指标来反映上市公司的投资价值，而这三项上层指标又分别由三项下层指标来具体体现。为此本例选取了如下的指标：总资产增长率、主营业务收入增长率、净利润增长率、主营业务利润率、净资产利润率、每股收益、每股净资产、每股公积金、每股未分配利润。

1）评价指标体系

其基本框架如下图所示：

- 公司内在价值指数
  - 成长性
    - 总资产增长率
    - 主营业务收入增长率
    - 净利润增长率
  - 盈利能力
    - 主营业务利润率
    - 净资产收益率
    - 每股收益
  - 股本扩张能力
    - 每股净资产
    - 每股公积金
    - 每股未分配利润

图 2-2 公司内在价值指数构成

表 2-2 公司内在价值指数指标定义

序号	指标名称	指标描述
1	总资产增长率	输入指标，反应企业本期资产规模增长情况
2	主营业务收入增长率	输入指标，用来衡量公司产品的生命周期，判断公司发展所处的阶段
3	净利润增长率	输入指标，反映公司的经营效益
4	主营业务利润率	输入指标，反映公司主营业务的获利水平
5	净资产收益率	输入指标，反映公司的盈利能力
6	每股收益	输入指标，反映公司的获利能力
7	每股净资产	输入指标，反映每股股票所拥有的资产现值
8	每股公积金	输入指标，是公司未来扩张的物质基础
9	每股未分配利润	输入指标，是企业留待以后年度进行分配的结存利润
10	公司内在价值指数	输出指标，反映上市公司的投资价值，值越大说明越具投资价值

各指标详细描述如下：

总资产增长率：总资产增长率是企业本年总资产增长额同年初资产总额的比率，反应企业本期资产规模增长的情况。

主营业务收入增长率:主营业务收入增长率可以用来衡量公司产品的生命周期,判断公司发展所处的阶段。一般来说：如果公司的主营业务收入增长率超过 10%，说明公司产品处于成长期。如果主营业务收入增长率在 5%至 10%之间，说明公司已经进入稳定期。如果公司的主营业务收入增长率低于 5%，说明公司已经进入衰退期。

净利润增长率：净利润是一个企业最终经营的成果，净利润多，企业效益就好，净利润少，企业经营效益就差，它是衡量一个企业经营效益的重要指标。其计算公式为：净利润增长率=（本期净利润增长额/上期净利润）*100%。

主营业务利润率：主营业务利润率即主营业务利润与主营业务收入的比率。该指标反映公司主营业务的获利水平。

净资产利润率：净资产利润率即净资产收益率。定义公式为：净利润*2/（期初净资产+期末净资产）。

每股收益：它是测定股票价值的重要指标之一，是分析每股价值的一个基础性指标，是综合反映公司获利能力的重要指标，它是公司某一时期净收益与股份数的比率。

每股净资产：每股净资产是指股东权益与股本总额的比率。这一指标反映每股股票所拥有的资产现值。每股净资产越高，股东拥有的资产现值越多；每股净资产越少，股东拥有的资产现值越少。

每股公积金：公积金是公司的“最后储备”，它既是公司未来扩张的物质基础，也可以是股东未来转赠红股的希望之所在。

每股未分配利润：未分配利润是企业留待以后年度进行分配的结存利润。未分配利润有两个方面的含义：一是留待以后年度分配的利润；二是尚未指定特定用途的利润。

2）构建模型训练样本集

为了从复杂的问题中，剥离出事物本质，需要对多个指标分开并计算其对输出指标的贡献度，

最后将结果汇总。对于单个指标对输出指标的关系如利润增长率指标（以下用x代替）对公司内在价值指数（以下用y代替）的影响可以用表 2-3 表述：

表 2-3 净利润增长率指标对输出的贡献率

序号	规则	数学式：$y = S(x)$
1	公司内在价值指数最小为 0，最大为 1	$S(0) = 0$，$S(max(y)) = 1$
2	净利润增长率越大，公司内在价值指数越大	$S(x1) > S(x2)$，当$x1 > x2$
3	净利润增长率为 0，对公司内在价值指数没影响，净利润增长率影响不会使公司内在价值指数超过 1	$0 \le S(x) \le 1$
4	由于噪声影响导致三相不平衡率是一个非常小的数，公司内在价值指数基本不受影响。但随着不平衡率变大，公司内在价值指数将迅速加大。	$S'(0) = 0$ $S'(x) > 0$且$S''(x) > 0$，当$x < mid$，mid是$F(x) = 0.5$ 时，x 的取值
5	公司内在价值指数不会一直快速增长，最终会收敛到 1	$S'(x) > 0$且$S''(x) < 0$，当$x > mid$，mid是$S(x) = 0.5$ 时，x 的取值

基于表 2-3，选择种群增长S型曲线对规则进行模拟是非常恰当的。

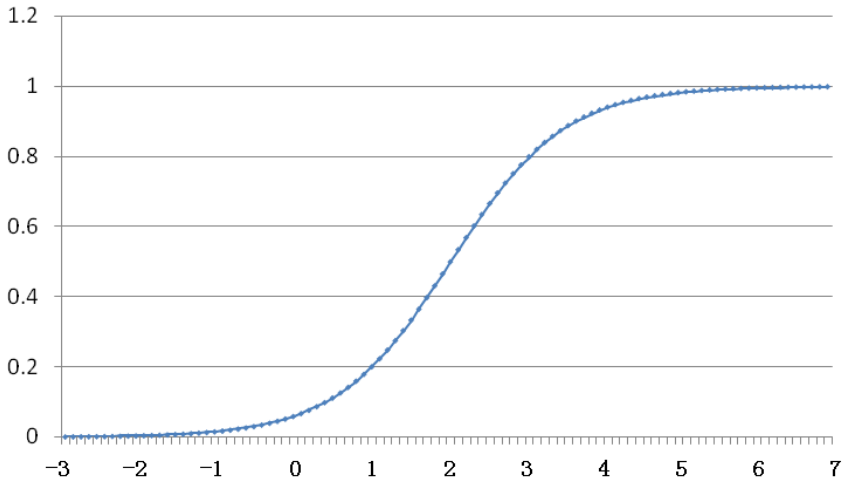


图 2-3 S 型增长率曲线

图 2-3 所示为种群增长曲线的形状。生物学中，在理想环境下种群有过度繁衍趋势，随着基数不断增大，繁衍速度也加快，呈现 J 型增长；但在实际环境中，受到面积，食物和竞争等因素的影响，增长曲线图会呈现如图 2-3 的先剧增后平衡的S型，该曲线并不一定对称。

公式定义：

$$y = \frac{K}{1 + (\frac{K}{N_0} - 1)} * Math.\exp(-\gamma * t)$$

其中，N_0为初始个体，大于 0，K为总体容量，γ为自然增长率.将该曲线应用于公司内在价值评价的净利润增长率影响中，满足表 2-3 所述的每项规则。于是 N_0取-3.00 时，可以认为公司内在价值最小，K为 0.001；N_0取 7.00 时，公司内在价值最大，K取值为 0.999，γ取值可以通过mid计算得到，假设当$t = mid$时，$y = 0.5$。

则可以解方程：$Math.\log((K-N_0)/N_0)/mid$ 求得γ的值。

本次求解过程需要mid值参数，也就是当净利润增长率等于多少时，公司内在价值指数等于0.5.这个参数需由有股票投资经验丰富的专家才能准确给出。

表 2-4 公司内在价值评价建模配置参数

指标名称	y=0.5 时 t 值	最小边界 t 值	最大边界 t 值	t 的增长步长
总资产增长率	0.45	-0.10	1.00	0.22
主营业务收入增长率	0.25	-0.50	1.00	0.30
净利润增长率	2.00	-3.00	7.00	2.00
主营业务利润率	0.23	-0.15	0.60	0.15
净资产收益率	0.07	-0.05	0.20	0.05
每股收益	0.23	-0.05	0.50	0.11
每股净资产	2.80	-0.40	6.00	1.28
每股公积金	2.27	0.05	4.50	0.89
每股未分配利润	0.00	-2.00	2.00	0.80

表中 t 值确定是通过对 2011 年 100 个农业模块的上市公司财务数据分析整理得到。

由表 2-4，可以构建若干评价公司内在价值的输入序列，但序列对应的输出应该如何确定？

公司内在价值评价的每个输入指标与输出指标的作用关系都可定义为图 2-3 所示曲线关系。单个指标贡献度皆回归为 0,1 问题，而对贡献度的汇总应遵循表 2-5 所示规则（令各指标贡献度为y_i，总输出为$Y_{内在价值}$）：

表 2-5 公司内在价值评价各指标贡献度汇总规则

序号	规则	数学式：$y=V(x)$）
1	各项贡献度取值介于 0 到 1 之间	$0 \le y_i \le 1$
2	汇总结果介于 0 到 1 之间	$0 \le Y_{内在价值} \le 1$
3	任意一项贡献度变大，将导致汇总结果变大	y_i增大→$Y_{内在价值}$增大

由以上分析，可以使用(2-1)式计算上市公司内在价值指数：

$$Y_{内在价值} = 1-(1-y_1)\times(1-y_2)\times\cdots\cdots\times(1-y_n) \quad (2\text{-}1)$$

3）模型构建

在准备好了专家样本数据后，就可以通过分类与回归模型，完成模型构建。

4、策略投资

股市中最难做的经常不是正确的分析判断，而是保持良好的心态，或者说保持理性。什么时候该贪婪？什么时候该恐惧？如果用简单逻辑，无疑是在高位应该恐惧，而在低位应该贪婪。但是，这看似简单的道理，却实际做起来很难，事实上绝大多数投资者都是正好相反。

在市场中，危险和机会总是并存的，明白此理的人在危急中以不变应万变，往往能够转危为安，转败为胜，鹤立鸡群，更多的人看到危险来临，则会恐慌，绝望，感性战胜理智，错失危急中的大机会，与暴利和辉煌绝缘。因为在股市中，绝大多数投资者都不能理性去判断市场，其对市场的判断主要是受过去一段时间的市场趋势和自己的盈亏情况左右，认为过去的趋势还会持续下去。因此，其结果就是，股市越涨，投资者参与热情越高涨，点位越高投资者的信心越足。反之，股市越跌，投资者参与热情越低落，点位越低投资者越悲观。巴菲特有句名言：“他人贪婪时我恐惧，他人恐

惧时我贪婪。”这也正是巴菲特能够长期战胜市场，成为“股神”的重要秘诀！

作为长线投资者，要精确把握市场的顶部或底部是很困难的，没人能做到那么完美，总是在最高点离开市场，在最低点进入市场。但事实上只要能把握个大概，在相对高位就离开市场，在相对低位进入市场，也就能成为这个市场的少数赢家。可是，绝大多数投资者却总是不切实际地期望精确地把握市场的转折，从而为了踏准小趋势的节奏，却反而错过了大趋势。

2012年10月25日，星石投资杨玲发表了一篇题为《楼市暴涨时代已去 股票市场长线乐观》的文章，她对未来股票型产品的投资观点是——2000点以下宝塔式建仓。以下是她的观点：

从投资的风险预算角度看，要控制理财的风险，就必须把最差的情况估计到。假设上证综指还会在2000点的基础上再跌20%，那就已经到了前期低点——1600点。从大概率情况看，这已经可以算作很差的情况了。如果我们2000点以下采取宝塔式建仓，即每跌100点加仓20%，就算指数跌到1600点，建仓成本也能控制在1800点左右的水平，最大下行风险在10%左右，如果你选择的标的能够稍微超越大盘，那么几乎不用冒什么风险。当然，另一种情况是，大盘跌不到1600点就涨上去了，那么大家就完全可以从容的做右侧交易，风险就更小了。

但不管怎么样，我国股票市场已经连跌了5年（图 2-4为上证指数趋势图），即使明年经济遭遇低谷，股票市场也不会特别差。因为只有经济真正回落，央行才能像美联储一样，毫无顾虑的提供低利率环境以刺激经济。从美股的例子看，流动性好的股票市场表现肯定是远好于房地产的。



图 2-4 上证指数走势

什么是宝塔式建仓？

先买少量，以试盘、验证自己的判断是否正确，如若正确接下来就一次次的增量去买，比如按1-2-3-4把资金分成10等份，根据不同价位按比例买入。价格越低，买入越多。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的相关性分析、主成分分析和 LM 神经网络优化等算法进行模型构建。

2003年9月下旬，上海股市经过长期阴跌，已跌进历史的底部区域。一些绩优股价格严重超跌，非常低廉。通过本案例模型，对钢铁模板所有公司价值进行综合评价，凌钢股份、酒钢宏信两只股票的综合推荐指数分别为：0.956和0.943，故我们备选了这两只股票。

我们同时从另外一个角度来分析一下这两只股票。以凌钢股份为例，我们认为它极具投资价值：

① 该公司稳定的赢利能力很强。连续五年来，它每年每股赢利达0.5 元以上，是沪市少有的持续发展的绩优股。而且公开信息显示，它的当年每股赢利将比前一年增长50% 以上。

② 它的股价的动态市盈率只有7倍，是中国股市中最低的，甚至远低于国际成熟市场的钢铁股平均市盈率(25倍)。我们坚信，当该企业的年报发布时，其优良的业绩将刺激市场大大抬高它的现行股价(6.6 元)。

③ 它以前多年的平均年派息率远远高于当年的银行无风险利率。该年应一样。④2003年9 月之前，该股的历史最低价为5.9 元，最高价为14.15元。现价6.6 元已处历史低位。

⑤ 该股价格如此低廉是因为市场人士认为钢铁企业是夕阳企业，从而远离它的股票。而我们认为，对于中国这样的发展中国家，钢铁企业是朝阳企业，发展前景远没有终结。

以凌钢股份的投资价值分析和我们多年来对上海股市的市场习性的了解，我们认为，该股在2003年9 月至2004年6 月(上一年度上市公司业绩公布截止期) 这段时间里的股价可能波动的区间为[5,10.5]，最可能波动的区间为[6,8.5]，并由风险厌恶型的证券投资数学模型[1]，计算得知，如果按6.6元买人， 预期赢利有保障。

按照以上结果，2012年9月我们分别以现价每股6.6 元、6.7元购买了2万元的凌钢股份、1万元的酒钢宏信股票。2013年1月上旬，当钢铁股业绩普遍大幅增长的信息被市场了解后，钢铁股股价普遍大涨。我们分别以每股9.05元、8.78元的股价卖出凌钢股份和酒钢宏信，平均获利率达34 % 左右。

练习题

1、除了本例提到了基于公司价值评价的方式指导用户选择股票，还有哪些方法可以实现股票预测？

2、请说明基于基本面和技术面进行股票预测的原理。

2.1.2. 案例二 基于拐点变化的股票趋势预测

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

股票价格时间序列是一个受政治、经济、心理等多方面因素影响的离散时间序列，使得股票价格常表现出包括混沌在内的各种复杂现象与行为。这些现象与行为若采用传统的统计学的方法处理往往难以得到令人满意的结果。

但股票趋势预测代表经济预测中的一种特定类型，它又具有其独有的特点。因为实际操作中我们不需要准确地知道下一天的收盘价，只需知道被观测市场的走向（升或降），从而用来决定我们的交易操作（长期/短期—买入/卖出）。我们所希望的是利用历史价格数据预测将来中短期（从 2 到 10 或 15 天）内的价格走势。因而预测模型也不需要预测下一个交易日的准确收盘价，我们只需从预测的趋势上去发现我们的交易策略[13]。

换句话说，我们需要预测的是市场的拐点，如图 2-5。

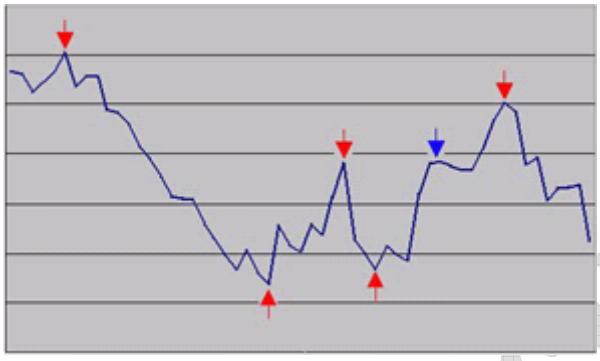


图 2-5 股票趋势拐点示意图

这是一只股票的日 K 曲线，我们要成功进行股市操作就要根据红色箭头来交易，在最低点买进，在最高点卖出。一个好的交易系统应当只是真正的拐点出现时给出提示，避免生成虚假信号，例如，象篮色箭头所指部分，市场只下降了几个百分点就马上又继续上升。

可见，我们所要设计的预测模型不需要预测市场的日准确收盘价，我们只需对正确预测拐点感兴趣，那么如何将下一个时刻的拐点准确地预测出来，并判断其有效性?

本案例的挖掘目标是：

1） 构建股票拐点预测模型

2） 拐点有效性判断

分析方法与过程

基于拐点变化的股票趋势预测建模过程主要由数据抽取、数据预处理、建模&预测、结果&反馈等几部分构成，见图 2-6。

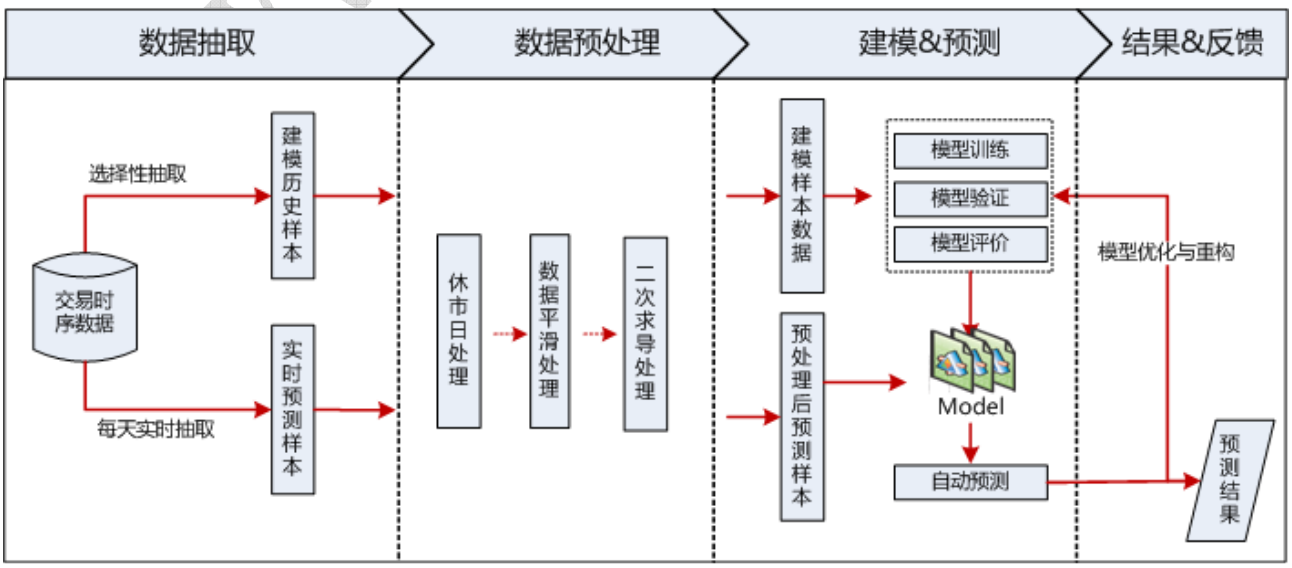


图 2-6 股票拐点趋势预测建模流程

1. 样本数据生成

我们现在需要决定如何训练神经网络，首先必须获得训练阶段的监督信号即期望值数据。我们所设计的模型使用时间序列转折指标量$\phi(t)$作为监督信号。令θ为我们期望的转折变化率，$\phi(t)$算法如下：

1) 从时间序列第一个数据项开始，依次计算相邻数据 x(t+1)与 x(t)之间的微增量并将其累计。
2) 当市场价格上升超过期望的变化率θ时，前面一个低点就被标注为‘买’信号，相应的输出值$\phi(t)$设为 0.05。
3) 当市场价格下降超过期望的变化率θ时，前面一个高点就被标注为‘卖’信号，相应的输出值$\phi(t)$设为 0.95。
4) 上述两点之间的日期对应的输出值$\phi(t)$以插值法规范化到 0.05 与 0.95 之间。

期望的转折变化率θ不能设置太小值，否则会生成太多的信号（其中很多都是错误信号）。下面两张图显示对于一个给定的时间序列，当它们的期望的变化率θ相应地分别设为 5%和 8%时，转折信号量$\phi(t)$的计算情况。

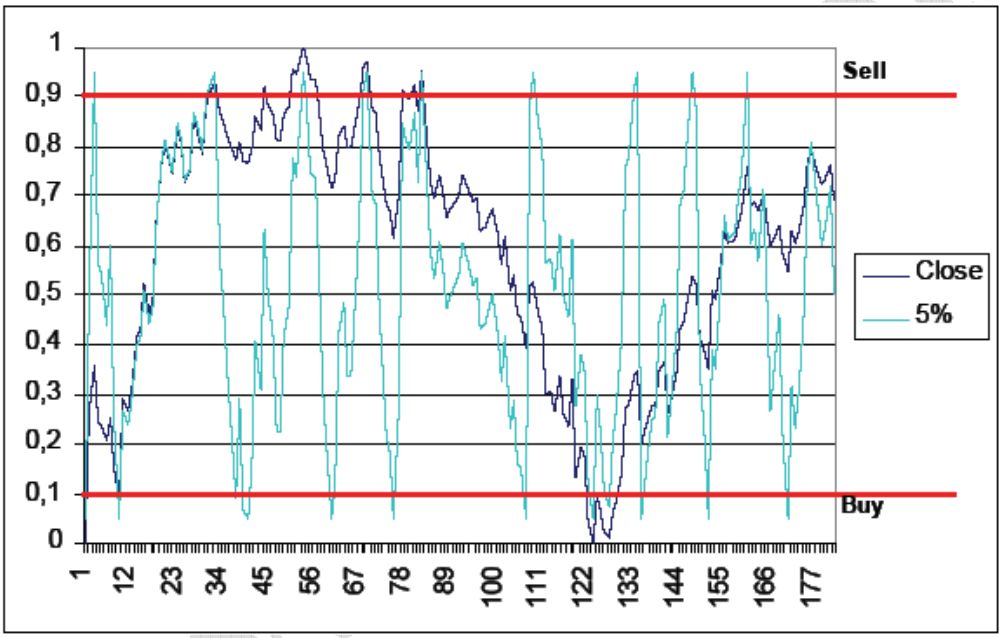


图 2-7 θ为 5 时转折信号量$\phi(t)$与转折变化率θ的关系图

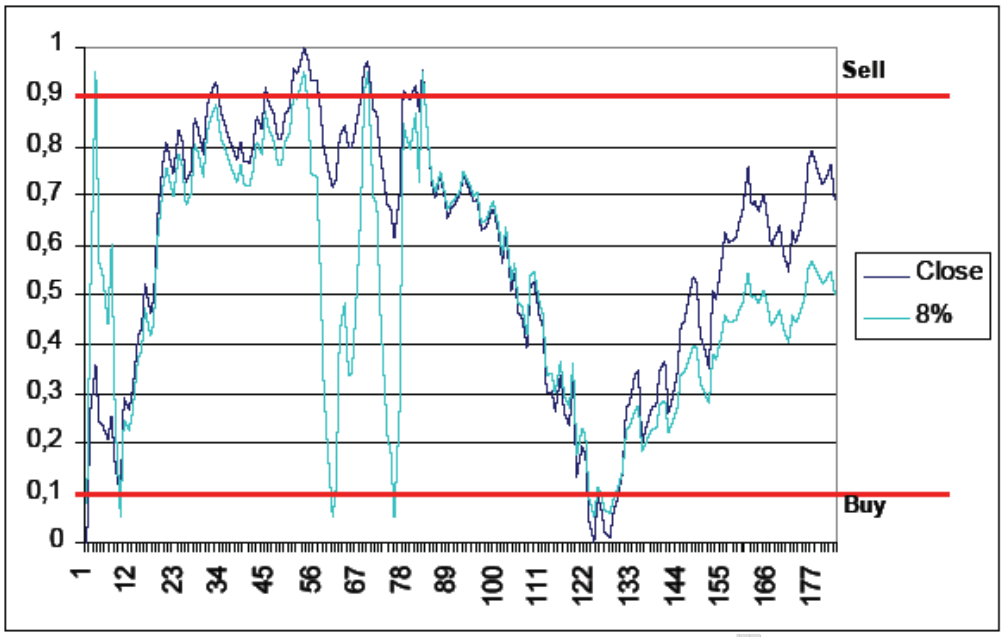


图 2-8 θ 为 8 时转折信号量 $\phi(t)$ 与转折变化率 θ 的关系图

得到转折信号量 $\phi(t)$ 后，根据 Takens 定理，构建出在微分同胚意义下与原动力系统“动力学等价”的混沌动力系统，从而由重构相空间中的轨迹预测，进行基于转折信号量 $\phi(t)$ 的股票价格拐点预测。

2. 股票的混沌识别及关联维确定

以 2004 年 1 月 1 日为起点，至 2005 年 4 月 1 日，共得到实测数据样本 357 点，首先按下式对数据进行归一化处理

$$x(i)=(y(i)-\frac{1}{N}\sum_{i=1}^{N}y(i))\Big/\left\{\frac{1}{N}\sum_{j=1}^{N}(y(j)-\frac{1}{N}\sum_{i=1}^{N}y(i))^2\right\}^{1/2} \tag{2-2}$$

利用上面混沌吸引子特征量测定的方法，首先判断时间序列的混沌性。

用归一化的时间序列样本数据构造时间延迟为 τ 的 m 维相空间，相空间的点 $y(t)=(x(t),x(t+\tau),\cdots,x(t+(m-1)\tau)$，考虑到证券交易中每周有五个交易日，可让 τ=5，而嵌入维数 m 分别取 $m=3,4,\cdots 10$ 等正整数，按前述理论，利用计算机软件编程可算出关联积分 $C_{m_0}(r)=\frac{1}{N^2}\sum_{i,j=1}^{N}\theta(r-|y_i-y_j|)$，该关联积分值与给定距离 r 的取值有关. 理论上存在 r 的一个取值范围，吸引子的维数 d 与关联积分函数 $C(r)$ 满足对数线性关系，即 $d(m_0)=\ln C_{m_0}(r)/\ln r$。在不同的嵌入维下，对一系列的 r 值可计算出一系列的关联积分值 $C_m(r)$。以 $\ln r$ 为横轴，以 $\ln C_m(r)$ 为纵轴绘成 $\ln C(r)-\ln r$ 图如图 2-9。

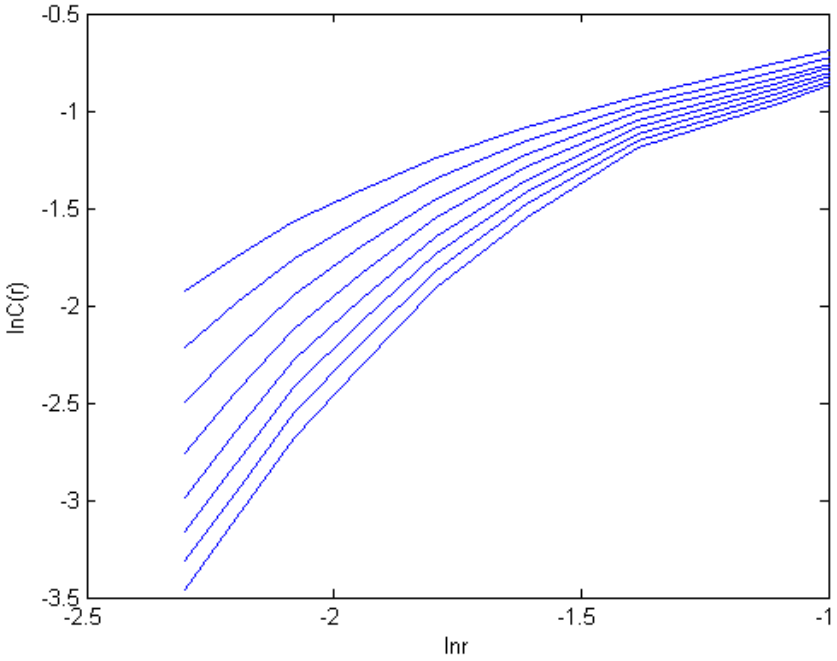


图 2-9 混沌时间序列的 $\ln C(r)-\ln r$ 图

注：图中由上到下的曲线依次表示嵌入维 $m=3,4,\cdots 10$ 的 $\ln C(r)-\ln r$ 。

从图中可以看出不断增加嵌入维数 m 的取值，随着 m 的增大，直线变得平行，直线的斜率值变得稳定，用最小二乘法估计这些直线的斜率，得到的趋于稳定的值就是关联维数的估计值，见表 2-6。

表 2-6 关联维数的估计值

m	3	4	5	6	7	8	9	10
$d(m)$	1.6033	2.0523	2.5033	2.894	3.1583	3.3414	3.4281	3.4985

从表 2-6 可以看出，当嵌入维增加到 8 后，关联维数 $d(m)$ 趋于稳定，大约为 3.4，最佳嵌入维是 9。这说明该时间序列具备混沌动力学特性，其混沌吸引子的分数维约为 3.4。且嵌入维数 $9>2*3.4281$，满足 *Takens* 定理中恢复微分同胚动力系统的条件，本例可基于嵌入维数 9 重构相空间。

3. 拐点预测

基于延迟时间序列进行拐点时序预测。

建模仿真

本例取英国证券市场 FTSE-100 指数成分股-英国 Tesco 公司股票(TSCO.L)实测数据对本文的理论进行实证研究。

取 Tesco 股票 2005 年 4 月 1 日至 2005 年 10 月 1 日的实测数据为样本，以 2005 年 7 月 1 日为界分为两段，前段作为训练样本后段为预测样本。由计算得到的重构相空间最佳嵌入维数建立径向基函数(RBF)网络，根据上面提出的拐点值计算方法确定网络训练的监督信号，经过训练确定各神经元参数。然后将预测样本加载到训练好的 RBF 网络上，通过计算机仿真实验，可得到图 2-10 所示的实验结果。

从图中可以看到，根据本例提出的基于混沌分析的拐点预测方法，预测样本中大部分的有效拐

点都预测到了，这证明了本例提出的预测模型的有效性。如图中箭头所标出的部分，预测转折指标量大于 0.8 或小于 0.2 的点经实际价格走势验证均为有效拐点，可以以此转折指标量作为卖出或买入交易信号对用户的交易操作进行辅助指导。

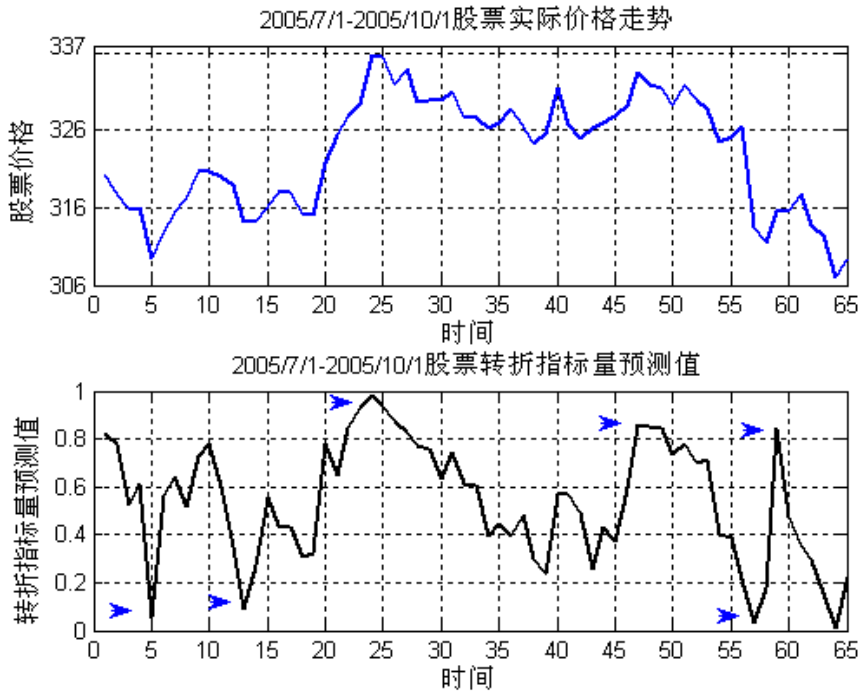


图 2-10 股票实际价格走势及预测走势

练习题

1、除了本例提到了基于公司价值评价的方式指导用户选择股票，还有哪些方法可以实现股票预测？

2、请说明基于基本面和技术面进行股票预测的原理。

3、尝试不同的参数设置对模型的影响，思考参数如何设置才最合理，并思考其原因？

2.1.3. 案例三 电信 3G 客户识别系统

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

有专家预计，目前我国 3G 普及率约 14%，2012 年或能达到 20%以上，中国已经步入进入 3G 快速增长期。3G 已经是我国电信运营市场毫无争议的主角。三大运营商都纷纷采取各种方法争夺 3G

市场，竞争趋于白热化。移动借由 2G 时代积累的优势，在 3G 用户总数上暂时领先，但优势已不明显，目前 3G 市场已成“三分天下”之势，未来运营商角力的关键还在于各自的网络质量和发展客户的手段。其中，发展客户的手段与对客户信息的掌握有很大的关系[19]。

如何在海量的数据中识别出 3G 客户？并进而认知 3G 客户的特征特点，为相关部门对这部分客户进行重点宣传、发送 E-mail 提供决策支持，以提升预期的宣传效果？

本案例的挖掘目标是：

1） 构建 3G 客户识别分类模型，并进行特征分析

建模方法与过程

图 2-11 和图 2-12 分别是 3G 客户识别数据挖掘建模过程和数据流程。

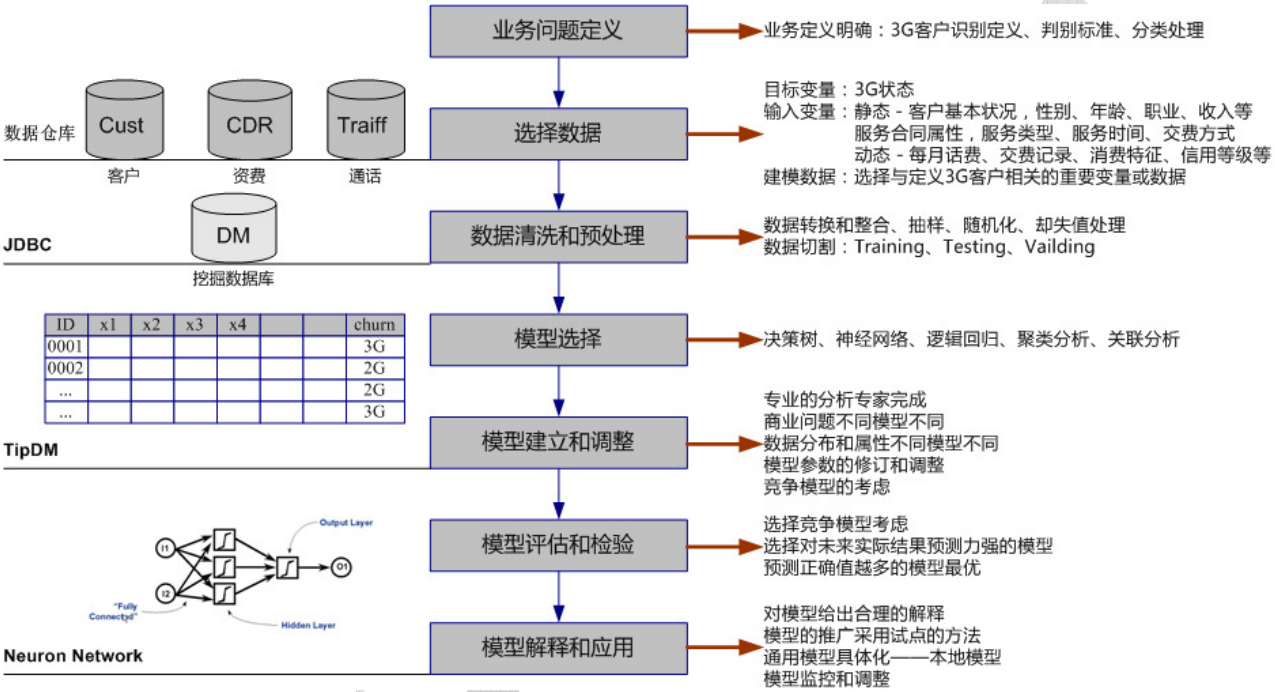


图 2-11 3G 客户识别数据挖掘建模过程

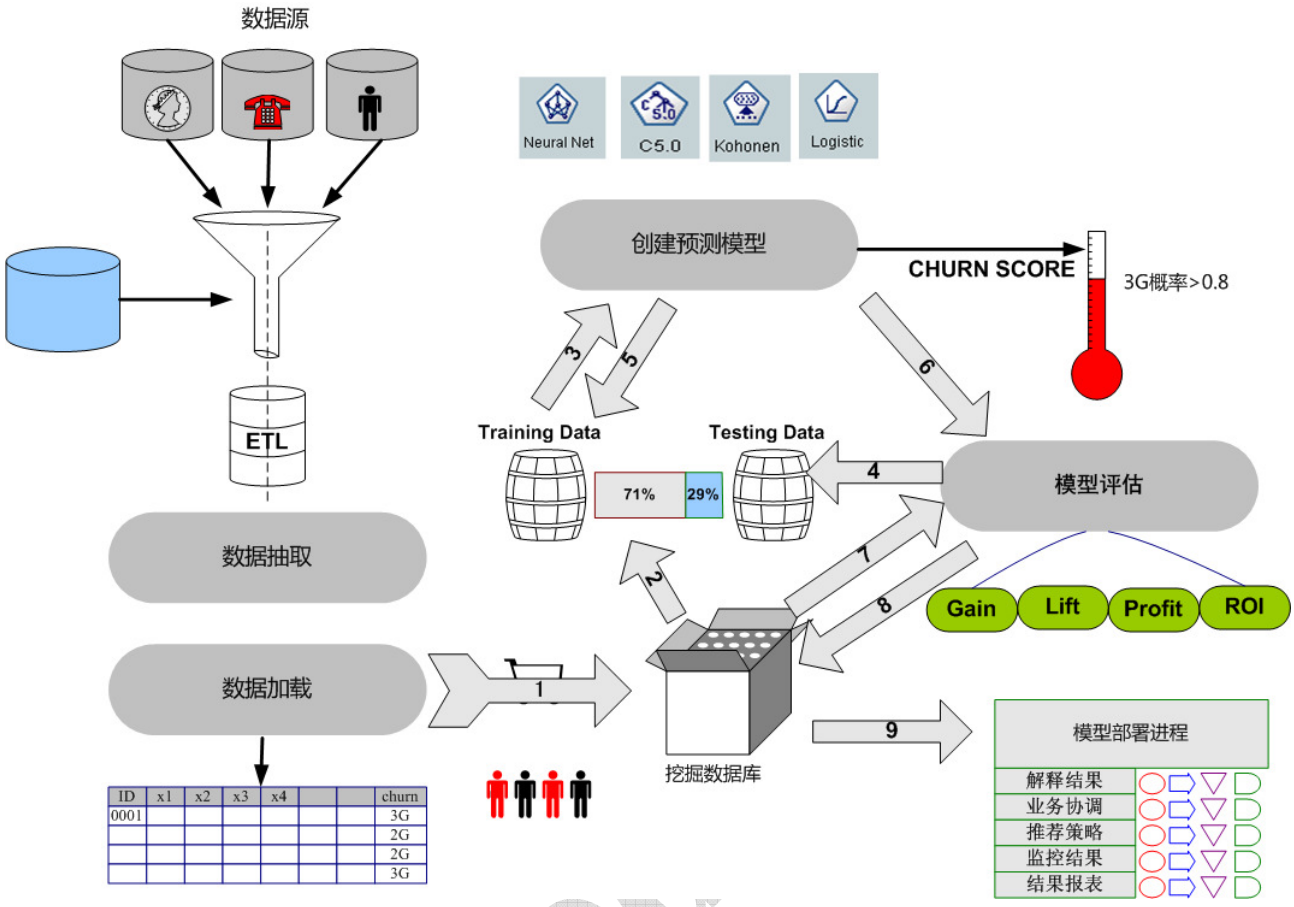


图 2-12 3G 客户识别建模数据流程

下面对 3G 客户识别主要过程加以说明。

1. 数据抽取

本案例数据来源于某电信运营公司，该数据总量为 24000 条。每一条对应一个客户近六个月的统计信息。其中已知分类（2G、3G）的为 18000 条，未知分类 6000 条。其中包括客户年龄，月平均消费额，月平均通话时长等属性 250 个（包括客户类型）。

本例利用 18000 条已知分类的数据进行分类模型的建立，随机抽取训练数据 10000 条，测试数据 8000 条。

2. 数据预处理

有效的数据预处理能使数据挖掘事半功倍。经统计，该数据有缺失数据多，数据取值范围广，数据取值类型多样的特点。因此建模前的预处理是必不可少的。比如采集的样本数据存在一些属性值的缺省或空值，如果不做处理，将直接影响后续算法的挖掘效果，严重时甚至得到错误的结果。采用相关的算法，如用该属性值的平均值、固定值或出现最多属性值，对缺省的属性值进行补充，也可以去除属性值缺省的那个实例，为后续的数据挖掘提供高效可用的数据。

1）数据清洗

数据清洗分为类型转换和缺失数据填补两部分。具体操作如表 2-7 和表 2-8：

表 2-7 数据类型转换

属性	取值	类型转换
CUSTOMER_TYPE	2G、3G	2G-0, 3G-1

GENDER	M、F	M-0，F-1
MARITAL_STATUS	S、M	S-0，M-1
TOP1_INT_CD	NONE、正整数	NONE-0
TOP2_INT_CD	NONE、正整数	NONE-0
TOP3_INT_CD	NONE、正整数	NONE-0

表 2-8 缺失数据填补

属性	缺失数目	填补方法
GENDER	3	随机填补 0、1
MARITAL_STATUS	1060	随机填补 0、1
AGE	500	用均值填补
CONTRACT_FLAG	918	随机填补 0、1
DAYS_TO_CONTRACT_EXPIRY	918	用均值填补
TOT_DEBIT_SHARE	288	用均值填补
其余	2127	由于分散在各变量，因此统一用 0 填补

2）属性筛选

属性筛选分为人工筛选和通过相关系数分析实现属性选择两部分。人工筛选结果见表 2-9。

表 2-9 属性人工筛选

剔除属性		剔除理由
NATIONALITY	SUBPLAN	取值离散、且各种取值没有明确说明含义
HS_MODEL	SUBPLAN_PREVIOUS	
HS_MANUFACTURER	REVPAY_PREV_CD	
PAY_METD	TOP1_INT_CD	
PAY_METD_PREV	TOP2_INT_CD	
OCCUP_CD	TOP3_INT_CD	

除了表 2-9 所示 12 个属性，剩下还有 237 个属性（不包括客户类别）。若将剩下的属性都作为客户识别模型的输入，算法的计算时间明显增大，而且部分不相关属性会成为建模过程中不必要的噪声。本案例采用特征相关分析法对剩下的 237 个属性作了分析。将所有的属性逐个对目标属性（客户类别）计算相关系数，计算公式如下：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2-3}$$

相关系数γ介于（-1,1），各特征的重要性可用相关系数的绝对值大小作评估。

3）数据平衡

由于作为训练的 10000 条数据中只有 1227 条是 3G 用户，数据偏向 2G 用户将导致模型输出结果偏向判别为 2G 客户。因此本文将训练数据中的 3G 用户数据多次复制，直到数量与 2G 用户数量近似 1:1。

www.tiedm.com

4）数据归一化

为了提高 BP 神经网络的性能，需要对数据进行归一化。

输入 BP 神经网络的数据进行归一化，输出神经网络的结果必须按输入归一化的方法进行反归一化，才能得到正确的输出结果。

5）离散化

对于决策树算法，需要对数据进行离散化，否则生产的决策树将会过于茂盛，以至于无法分析。而且离散化能使输出的决策树通俗易懂，方便人们从中发现规律。本文对输入决策树分类模型的连续变量进行了离散化，将各属性最大最小值的区间平均分成 8 份，每 1 份对应 1 个离散值。

3. 分类识别

本案例分别采用 BP 神经网络和决策树进行建模，实现对 3G 客户的分类识别。

建模仿真

本案例运用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中相关性分析进行数据探索，再运用 BP 神经网络和 C4.5 决策树进行客户识别。

1. 模型输入

模型的输入包括两部分，一部分是建模专家样本数据的输入，另一部分是建模参数的输入。

本案例中，样本数据涉及客户年龄，月平均消费额，月平均通话时长等 250 个属性（包括客户类型），各指标描述见表 2-10。

表 2-10 3G 客户识别样本数据属性

序号	变量名称	变量描述
1	AGE	年龄
2	GENDER	性别
3	MARITAL_STATUS	婚姻状态
4	HS_AGE	现在使用的手机机龄
5	VAS_AR_FLAG	自动漫游业务
6	AVG_USAGE_DAYS	月平均使用天数
7	AVG_BILL_AMT	月平均消费额
8	AVG_MINS	月平均通话时间
9	AVG_CALL	月平均通话次数
10	AVG_VAS_GAMES	月平均上网游戏流量
	……	

表 2-11 相关分析建模参数

序号	参数名称	参数描述
1	计算方法	选择相关分析计算方法，如 Pearson、Sperarman、Kendall
2		

表 2-12 C4.5 决策树建模参数

序号	参数名称	参数描述

www.tipdm.com

1	二元分割	构建树时是否使用二元分割
2	置信因子	用于树修剪的置信因子，值越小，树修剪耗费越大
3	最少叶片数	最少决策树叶片数
4	减少错差剪枝	是否用减少误差修剪来替代 C4.5 修剪
5	提升子树	修剪时是否使用提升子树操作
6	使用拉普拉斯算子	是否使用拉普拉斯算子进行平滑

表 2-13 CART 决策树建模参数

序号	参数名称	参数描述
1	最小数	节点的分支数
2	种子数	初始随机选取的样本数

2. 仿真过程

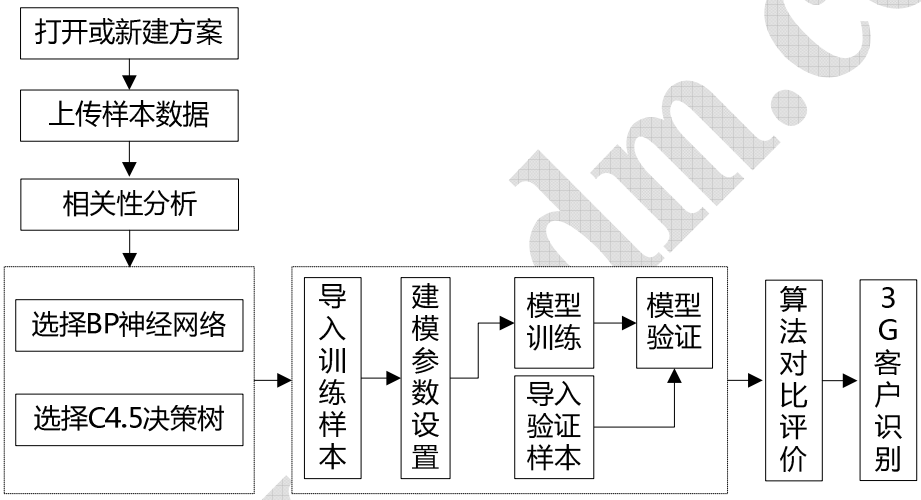


图 2-13 仿真过程

建模仿真过程说明如下：

1） 登录 TipDM 平台，在方案管理页面中，新建方案或者打开一个已建方案。

2） 切换到数据管理页面，上传经预处理后的专家样本数据文件。

3） 选择相关性分析功能，导入样本数据进行相关性分析。

4） 分别选择 BP 神经网络算法和 C4.5 决策树算法，进行模型构建。

5） 对比 BP 神经网络和 C4.5 决策树的建模结果，并选择最优算法。

6） 用最优法对测试样本进行 3G 客户识别。

3. 结果及分析

1）电信 3G 客户识别样本相关性分析

相关性分析后，各属性的相关系数计算结果由大到小排列如表 2-14 所示（由于属性数量较多，因此只列出相关系数大于 0.1 的属性）。

表 2-14 属性相关系数计算结果

属性	相关系数	属性	相关系数

AVG_BILL_AMT	0.2294	STD_VAS_AR	0.1479
HS_AGE	0.2241	AVG_BILL_VOICEI	0.1475
AVG_MINS_OBPK	0.2182	TOT_USAGE_DAYS	0.1417
AVG_NO_CALLED	0.2174	AVG_USAGE_DAYS	0.1417
AVG_CALL_OB	0.2121	AVG_MINS_IB	0.1409
AVG_CALL	0.2109	AVG_MINS_OP	0.1379
AVG_MINS_OB	0.2095	STD_MINS_MOB	0.1353
AVG_MINS_PK	0.2087	AVG_CALL_T1	0.1316
AVG_MINS_LOCAL	0.2087	STD_MINS_OB	0.1284
AVG_CALL_PK	0.206	STD_CALL_PK	0.1251
HIGHEND_PROGRAM_FLAG	0.2023	STD_CALL_MOB	0.1246
AVG_CALL_MOB	0.2013	AVG_CALL_LOCAL	0.1228
AVG_MINS	0.1997	STD_CALL_OB	0.1225
AVG_CALL_EXTRAN	0.1952	STD_CALL_EXTRAN	0.122
AVG_MINS_INTRAN	0.1947	STD_CALL_INTRAN	0.1204
AVG_MINS_EXTRAN	0.1933	STD_MINS_INTRAN	0.1203
AVG_CALL_IB	0.1856	STD_MINS_INTT3	0.1191
AVG_MINS_MOB	0.1818	STD_MINS_EXTRAN	0.1187
AVG_CALL_OP	0.1798	STD_CALL	0.1186
AVG_NO_RECV	0.1792	AVG_MINS_FIX	0.1183
AVG_CALL_INTRAN	0.1769	AVG_VAS_SMS	0.1174
AVG_CALL_FIX	0.1675	STD_MINS_PK	0.1174
VAS_AR_FLAG	0.1661	STD_CALL_OP	0.1169
STD_VAS_ARC	0.1647	STD_CALL_LOCAL	0.1166
AVG_MINS_IBPK	0.1588	STD_VAS_GAMES	0.1149
AVG_VAS_ARC	0.1577	AVG_BILL_SMS	0.1146
TOT_DIS_INT	0.157	STD_BILL_VOICE	0.1122
AVG_BILL_VOICE	0.1547	STD_CALL_T1	0.11
STD_BILL_AMT	0.153	AVG_VAS_GAMES	0.1059
AVG_MINS_T1	0.1526	AVG_CALL_EXTRANT1	0.1007
AVG_VAS_AR	0.1516		
AVG_MINS_OBOP	0.1502		

结合对电信数据的理解，最后选择进行分类模型的输入属性如表 2-15：

表 2-15 最终选择属性

序号	最终选择的属性	含义
1	HS_AGE	现在使用的手机机龄
2	VAS_AR_FLAG	自动漫游业务
3	AVG_USAGE_DAYS	月平均使用天数
4	AVG_BILL_AMT	月平均消费额

www.ilem.com

5	AVG_MINS	月平均通话时间
6	AVG_CALL	月平均通话次数
7	AVG_VAS_GAMES	月平均上网游戏流量

2）基于 BP 神经网络的模型构建

由于神经网络算法输出结果受到训练次数影响，并伴随一定的随机性，多次实验得到的分类正确率如下：

表 2-16 BP 神经网络分类正确率统计

项目	百分比
3G 客户分类正确率	78%
2G 客户分类正确率	74%
总体正确率	75%

3）基于 C4.5 决策树的模型构建

表 2-17 CART 决策树分类正确率统计

项目	百分比
3G 客户分类正确率	77%
2G 客户分类正确率	63%
总体正确率	65%

4）模型评价

BP 神经网络与决策树对于 3G 客户的识别正确率都接近 80%,说明本用例建立的分类模型对 3G 客户的敏感度比较高，基本能识别出 3G 用户，能达到预期目标。但是只看 3G 客户的识别正确率是不科学的，还要看 2G 客户的识别正确率和总体识别正确率。

2G 客户的识别正确率方面，决策树对 2G 客户识别的正确率只有 63%。说明决策树对 2G 用户的敏感度并不高。相比之下， BP 神经网络仍有 74%的正确率。在这个方面 BP 神经网络明显占优。

从总体正确率看，BP 神经网络的正确率仍然比决策树高近 10%。BP 神经网络无论是总体性能还是对局部分类的敏感度都表现不错。而决策树分类模型性能还有待提高。

综上所述，本例所设计的 3G 客户识别模型能达到基本的要求。但性能上仍有很大的提升空间。在数据预处理和模型参数选择上或许能有进一步的优化。

5）决策建议

虽然本例的客户识别未能达到百分百准确。但从另一个角度看，一味追求正确率并没有太多意义。因为本来运营商对各个用户的类别就已经作了登记。反而，我们或许能从客户的误识别中获得更多信息。

本来是 2G 客户，但被识别为 3G 客户的用户很有可能就是因为它的各项属性与 3G 客户的相似，如月均网络游戏游流量比较大，月均消费额较高，说明这批人是有机会被发展成 3G 客户，电信运营商应对这批客户大力推销 3G 应用。

而部分 3G 客户被识别为 2G 客户，很有可能是因为对目前的 3G 自费或 3G 服务不满意的客户，他们对 3G 应用并不感兴趣，因此和 3G 用户的特征不相符。这批客户有可能在未来流失，所以，电

www.tipdm.com

信运营商应更加关注这批客户的情况，争取减少流失。

练习题

1. 尝试不同的参数设置对模型的影响，思考参数如何设置才最合理，并思考其原因？
2. 当采样数据维度过大，如何进行降维处理？
3. 采样数据中的缺失值如何处理？
4. 由于采样数据中常常包含许多含有噪声、不完整、甚至是不一致的数据。显然对数据挖掘所涉及的数据对象必须进行预处理。那么如何对数据进行预处理以改善数据质量，并最终达到完善最终的数据挖掘结果之目的呢？
5. 数据预处理主要包括哪些功能？

www.tipdm.com

2.1.4. 案例四 基于客户分群的精准智能营销

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

随着国内电信市场竞争的日趋激烈，电信运营商的经营模式逐渐从“技术驱动”向“市场驱动”、“客户驱动”转化。面对客户的多样化、层次化、个性化的需求，大众化营销已经失去了其优势，基于客户信息、客户价值和行为，深入数据分析的洞察力营销、精确化营销的理念逐渐被各大电信运营商所接受，并渴望通过从数据中挖掘价值来减少营销成本、提高营销效益。

如何从电信运营系统中所积累的数据，发掘出有价值的商业规则，建立客户细分、流失预警、新业务响应等模型，并进而设计出符合用户需求、拉动业务量增长的业务经营目标和盈利点？

本案例的挖掘目标是：

1) 争取更多的客户：根据对现有客户分析，识别潜在客户，提高市场反应速度，优化销售渠道结构，提供差异化产品。
2) 减少客户流失率：了解流失率较高的客户群特征，特别是获利较多客户的个性特征，通过市场细分监控具有类似个性特征的客户发展动态，提高在线客户流失率预测准确率，提前做好预防客户流失措施。
3) 降低服务成本，提高业务收入,提高企业的运营效率。通过细致地分析用于各个用户群的服务成本、用较少的成本定位在目标市场，最优化投资，设计出有吸引力的且节省成本的的服务组合提供给各个细分市场。例如在移动领域，将呼叫转移功能与非高峰时段服务进行捆绑就可以满足那些相对固定在办公室上班的用户的基本业务消费需求。
4) 增加 ARPU 值：更准确的价格细分，更精准的营销信息传递，更好的渠道策略，销售渠道结构，捆绑服务及打包服务，及时获取业务需求变化，更好地合作策略，服务策略及其管理，增加第三方业务收入。
5) 优化服务：监控每个细分市场的业务使用和获利情况，建立不同的销售渠道来满足不同电信需求市场，根据客户需求定制个性化服务产品，及时洞悉客户的业务或者产品的使用情况，提高用户业务使用满意度。
6) 精准的市场营销策略的制定：通过熟悉各个细分市场消费特征，为各个细分市场定制专门的价格、渠道和促销和个性化产品。

分析方法与过程

准确的细分市场和差异化的营销策略是目前电信企业市场营销所必须面对的难题。所谓市场细分是指营销者根据顾客之间需求的差异性，把一个整体市场划分为若干个消费者群的市场分类过程。由于顾客对电信产品需求的多样性、变动性以及电信企业资源的有限性，电信企业在进行市场营销过程中，必须进行市场分析，选择目标市场，做出市场定位，并结合目标市场的特点和结构制定有针对性的市场营销策略。客户分群是了解客户进行市场细分和目标市场营销的前提。数据挖掘的分

类和聚类的方法都可应用于客户分群。本案例将对用聚类方法进行客户分群的实现模式进行详细阐述，为电信目标市场营销的客户分群提供完整的解决方案，并以实际案例验证其可行性。



图 2-14 基于客户分群的精准智能营销建模过程

本案例主要采用聚类分析实现基于客户分群的精准智能营销，建模过程详见图 2-14。

聚类（clustering）是指把一组个体按照相似性归成若干类别，即“物以类聚”，其目的是使得属于同一类别的个体之间的相似度尽可能大，而不同类别的个体之间的相似度尽可能小。相似度是根据描述对象的属性来测算的，距离是经常采用的度量方式。聚类和分类有着很大的区别：分类时，我们总是事先知道哪些属性是重要的，运营商总是将重要的、有影响力的属性作为分类的依据；而聚类时，运营商事先根本不知道哪些属性起作用，任务之一就是要找到那些起关键作用的属性。聚类分析能够帮助我们发现特征迥异的不同客户群，和对客户分群起关健作用的指标变量，并辅助运营商对各客户群的特征进行深刻洞察。

1. 客户分群的商业理解

依据 CRISP-DM 流程首先要进行客户分群的“商业理解”，这一初始阶段集中在从商业角度理解项目的目标和要求，然后把理解转化为数据挖掘问题的定义和一个旨在实现目标的初步计划。本例必须明确项目的商业目标，这个目标应该是适于用基于聚类分析的客户分群方法去达到的。比如某电信运营商定义的客户分群的商业目标是“对某市数十万公众客户，从价值和行为的分析维度进行客户分群，以了解不同客户群的消费行为特征，为发展新业务、流失客户保有、他网用户争夺的针对性营销策略的制订提供分析依据，并实现企业保存量、激增量的战略目标”。电信客户从营销属性分为三类：公众客户、商业客户和大客户，其中公众客户消费行为有较大的随机性，客户分布难有规律可寻，比较适于聚类分析。本例将此商业目标转化为数据挖掘的可行性方案：从价值和行为维度，考察客户业务拥有与使用、消费行为变化、他网业务渗透等方面属性，采用聚类分析的数据挖掘技术对研究的目标客户（公众客户、入网时长、地域属性、产品拥有类型等方面限定)进行客户分群，对各客户群进行特征刻画和属性分析，为针对性营销确定目标客户群，并根据客户群属性和营销目标量体裁衣制订恰当的营销方案。由于客户的特性是不断变化的，数据挖掘的分析结果具有一定时效性，因此数据挖掘必须以项目来实施，在目标、进度和资源安排上明确要求。

针对各业务部门的需求访谈及策略分析见表 2-18。

表 2-18 需求访谈及策略分析

访谈部门	需求点	策略分析
计费业务中心	主要通过计费业务中心了解客户细分项目系统平台的需求和系统环境	
市场经营部	关注全球通品牌整合，希望通过品牌整合提高服务品质，实现在业务、回馈、渠道上的差异化	通过客户细分，分析品牌整合高中端人群行为消费特征，为其设计差异化服务。并对品牌整合的效果从客户细分角度进行监测。
	如何更好服务于中端用户	在客户细分后准确定位全球通中端用户，研究其行为方式和特征，进行有针对性的服务设计和活动策划。

www.tpdn.com

	通过提高服务来挽留有离网倾向用户	研究分析有价值的离网用户行为特征，策划有针对性的活动或服务进行挽留。
营销中心	关注双机倾向用户的挽留。希望通过发现有双机倾向的用户，并在其购买之前进行挽留	分析以往双机用户的行为特征及关注点，针对有价值用户设计服务项目及策划活动。
	预警离网用户，通过提高服务来挽留有离网倾向用户	研究分析有价值的离网用户行为特征，策划有针对性的活动或服务进行挽留。
	关注大众卡的效果	通过客户细分了解大众卡的受众群，向有流失可能的用户进行宣传，并避免直接影响全球通其他用户。
客户服务中心 客户关系管理室	目前的外呼服务多为服务宣传和用户调查。外呼用户多为随机抽取，缺乏策略性。希望通过客户细分节约外呼成本，提高外呼效率和效果，并深入挖掘外呼应用。	对客户进行细分后，在外呼时可提供有针对性的服务。在推荐产品或进行调查前便了解用户的行为特征、生活方式等，有助于提高亲和力，节省人员和资金成本，并提升效果。
数据部	对于目标用户只了解其大致特征，不清楚其更细致的行为特征、生活方式等，因此在推各项业务时定位不准。希望通过客户细分模型了解各种数据业务的使用人群以及使用行为方式。	通过客户细分，寻找其重要业务的目标人群，并详细分析特征，对业务的规则制定、推广提供依据。
市场部	小灵通和联通竞争激烈，希望了解有转网倾向的用户的特征，并进行及时挽留	分析有可能转至小灵通、联通的用户，由于整个样本空间规模较小，可以适当对中高端人群制定一对一的服务策略
	目前所有产品都针对所有人群进行推广，希望找到目标人群有的放矢地进行推广，提高效率，节约成本 。	分析各种产品的目标用户，研究其行为特征，并通过交叉销售带动整个产品的销售增量。
	希望找到短信密集使用人群，并以此为导入点，制定有效的方式刺激整个短信市场	通过对短信密集人群的行为分析，找到此类人群，通过他们带动周围人群的短信使用量，并通过他们对移动产品及业务进行软性宣传

基于访谈中了解的客户需求，本案例中采用用户行为特征作为细分变量（用于进行客户细分的变量），用户人口统计信息和客户价值作为描述变量（将客户细分成各个群体后，各群体的基本特征），从而定位人群特征，并制定了图 2-15 所示的客户价值图。

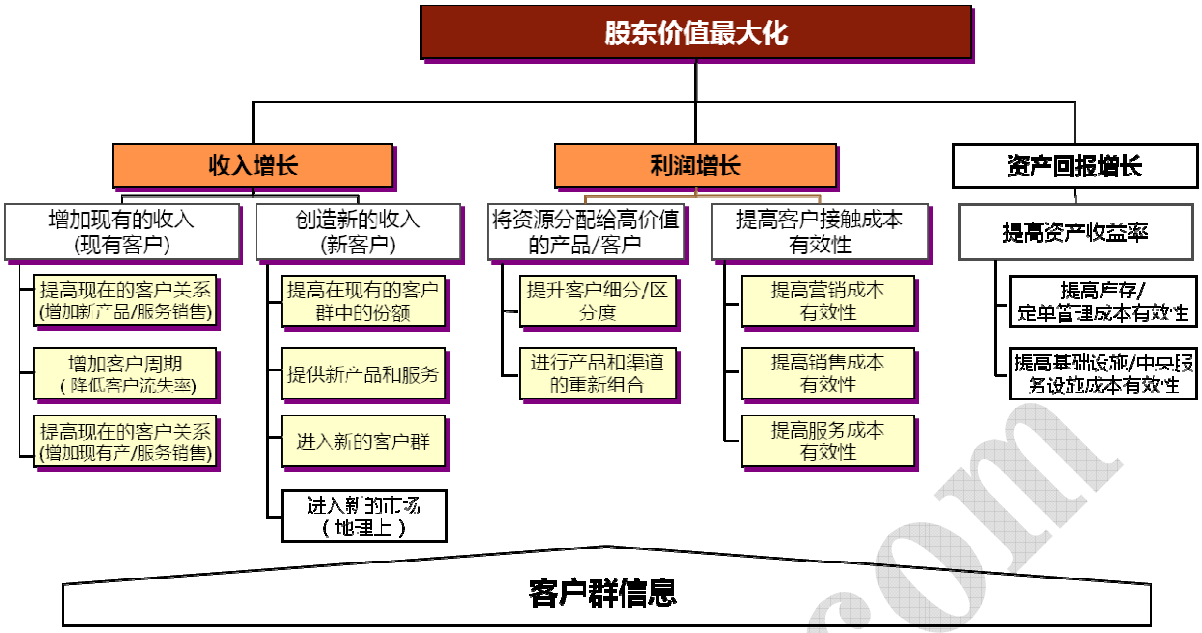


图 2-15 客户价值图

2. 客户分群的数据理解

“巧妇难为无米之炊”，数据是挖掘的基础，在确定目标和方案后需要进行“数据理解”，以确定要支持本例的分析目标需要哪些方面的数据，数据基础是否已经具备，数据质量是否能满足要求，如果不能得到肯定的答复，建议推迟项目实施直至条件成熟，因为“进去的是垃圾出来的仍是垃圾”，错误的分析结果可能会给本例带来重大的损失。本案例分析了各业务系统及企业数据仓库中客户信息、客户消费及购买使用行为三个方面最近六个月的历史数据。电信企业拥有业务受理开通的 CRM 系统，进行计费、帐务及欠费处理的计费系统，卡类业务的智能网系统，客户服务的 10000 号系统，营销服务的渠道系统，还有结算系统、宽带、窄带及小灵通系统等等，这些业务系统储蓄了企业运营的海量客户数据。有些电信企业还建立了数据仓库系统，对这些数据进行了清洗、整合和集中，建立了企业统一数据视图，并确保了数据来源的广泛性、完整性和准确性，为数据挖掘创造了更好的条件。本例从企业数据仓库中确定了以下数据的来源：

- 客户基本信息（DW_USR_DCUSTM）
- 用户帐务信息（DW_FEE_SHOULDDM）
- CDR 信息 （CDR:Calling detail records，即呼叫详细记录 ）
  - 语音 CDR（DW_CALL_CDR）
  - IP 业务 CDR（DW_NEWBUSI_CDR）
  - 短信业务 CDR（DW_NEWBUSI_SMSCDR）
  - 梦网业务 CDR（DW_NEWBUSI_MESGCDR）
- 客服信息

除了数据源分析，数据理解阶段的另一些重要任务是进行数据探索和检验数据质量。本例可以借助一些可视化工具或统计分析进行数据探索，以明确数据的分布状况和重要的属性及其关系。比如通过值分析对数据进行基本的探查，包括空值、唯一值、空字符串、零值、正值、负值的统计；统计分析计算各数值型变量的最小值、最大值、均值、标准差、众数等，有利于发现一些异常值，便于更好地理解数据；频次分析、直方图分析有助于更准确地了解数据的分布，从中发现有价值的点，其中频次分析主要面向离散型变量，而直方图分析主要面向连续型变量。数据探索有助于提炼

数据描述和质量报告，还能发现数据异常，并为进一步的数据转换和数据准备打基础，甚至可能直接指向数据挖掘的目标。数据探索的同时可以全面检验数据质量，对数据的完整性、正确性、一致性有更细致的认识。除了统计分析，抽样核查比较也是数据检验常用的方法，在进行数据质量核查时，需要有对数据意义和取值范围敏感的业务人员参与。

3. 客户分群的数据准备

“数据准备”包括所有从原始的未加工的数据构造最终分析数据集的活动，是数据挖掘过程中最耗时的环节，甚至要占据整个数据挖掘项目一半以上的工作量。数据准备工作的流程如图 2-16 所示。

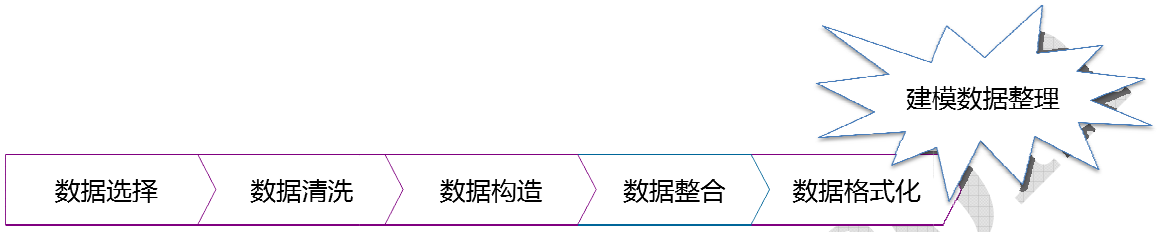


图 2-16 数据准备流程

1）选择数据 决定用来分析的数据。选择标准包括与数据挖掘目标的相关性，数据质量和工具技术的限制，如对数据容量或数据类型的限制。数据选择包括数据表格中属性（列）和记录（行）的选择。本例可以分主题在企业数据仓库中选择本例需要的各类数据，并进行按月汇总，生成月粒度数据宽表（宽表是将数据经过组合所形成的包括细分变量和描述变量的记录表）。宽表中每个用户每个帐务月的信息汇总成一条记录，宽表数据生成流程见图 2-17。

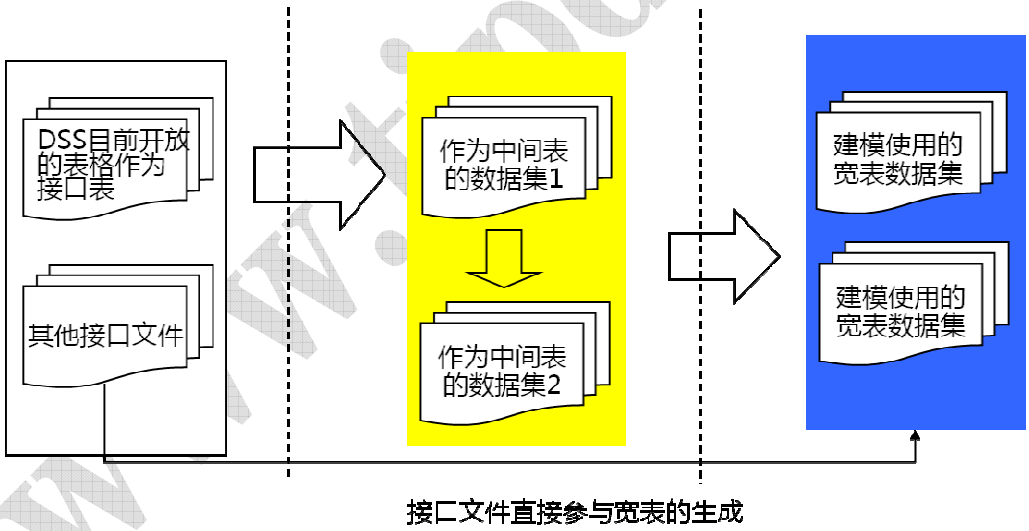


图 2-17 宽表数据生成流程

客户行为特征从表 2-19 所列的几个方面来获取信息。

表 2-19 客户行为特征信息获取

客户数据	导出客户需求种类
SMS 使用次数 Monternet 使用次数 WAP 上网时间 GPRS 数据流量	对方便性及信息实时性的需求

IP 使用情况 优惠时段通话情况 拨打 1861 次数	对资费的敏感程度
本地、长途、漫游呼叫时长 本地、长途、漫游呼叫次数 呼叫时间、次数（繁忙/非繁忙时段，工作/休息时段） 呼叫类型（主叫、被叫、呼转）	对通话的多层次需求
服务类型	对多元化服务的需求程度

宽表中数据属性粒度要求尽量细，以便于在后期灵活构建数据。本例将以上价值和行为变量从专业（区内、区间、本地移动、本地异商固网、传统国内长途、传统国际长途、传统港澳台长途、传统异地行动、IP 国内长途、IP 国际长途、IP 港澳台长途、IP 异地行动）和时段（白天:7:00-18:0、晚上：18:00-7:00、特殊时段：0:00-7:00）两个维度进行了划分。

2）清洗数据 将数据质量提高到所选分析技术和分析目标要求的水平。这包括选择需要进行数据清洗的子集，插入适当的默认值或者通过更加复杂的技术如建模来估计缺失值，比如某月份的数据缺失本例可以用前后月份数据的平均值来填充，本例将拥有产品较多（可能不是公众客户）及公免的客户数据剔除。此阶段需要对宽表的数据进一步探索和检验。

3）构建数据 该任务包括构建数据的准备操作，如进行变量设计生成派生属性、生成完整的新记录或者已存属性的转换值（将字符型字段转化为数值型字段）。在宽表基础上通过抽取、合并、衍生得到分主题汇集的价值变量和行为变量形成的中间表，中间表中每个用户 6 个月的信息汇总成一条记录。抽取指直接从宽表中选择变量进行汇总；合并指将宽表中多个字段合并成一个字段汇总，如将区内通话时长和区间通话时长合并成本地通话时长；衍生指根据各种逻辑关系派生出新的变量，如均值变量、占比变量、趋势变量、波动变量等。在进行变量设计时建议技术人员和业务人员密切配合讨论，根据业务需要、挖掘目标及数据源的实际状况确定数据选择，确定基础变量和数据源的映射关系，确定衍生变量的数据转换逻辑。中间表变量命名的规范化有助于用户理解、记忆和应用，便利于将来数据挖掘的应用分析。变量命名采用对变量属性进行描述的分段的英文缩写（首字母大字）用下划线连接来实现，字段按属性意义从大到小的范围排列，从而在对变量进行排序时可以把同一类型的变量聚集在一起，便于分析。

4）整合数据 对各中间表的数据进行联合，生成最终的分析数据集也称宽表。本例以客户标识为主键，串联起客户相关的所有信息数据，建立起统一客户视图。在整合数据时，本例要回顾商业理解阶段对数据挖掘目标的定义，以使本例的分析数据集能满足需求。有些电信企业由于历史数据以用户为中心，难以建立起统一客户数据视图；这时可使数据挖掘的目标限定于对某类用户进行分析；或以分析某类用户为主，通过企业数据库中用户之关的关系数据，去关联其它产品信息辅助分析。在完成数据整合后，本例需要再次对数据质量进行核查。

5）格式数据 格式化转换是指根据建模的要求，对数据表现形式进行变换，如用 K 均值算法做聚类分析时，需要先将数据进行标准化处理，对数据进行 Z 变换，以消除量纲不同可能引起的数据差异。

4. 客户分群的模型建立

在生成最终分析的数据集后，就可以建立模型进行聚类分析了。“模型建立”阶段主要是通过因子分析找到变量之间的关系，并优化变量组合。在对模型结果的分析中，根据标准——群间差距最大，群内差距最小的原则进行分析，同时调整变量组合，以尽量接近标准。以此方式循环，逐步

使模型得到优化，建模过程见图 2-18。

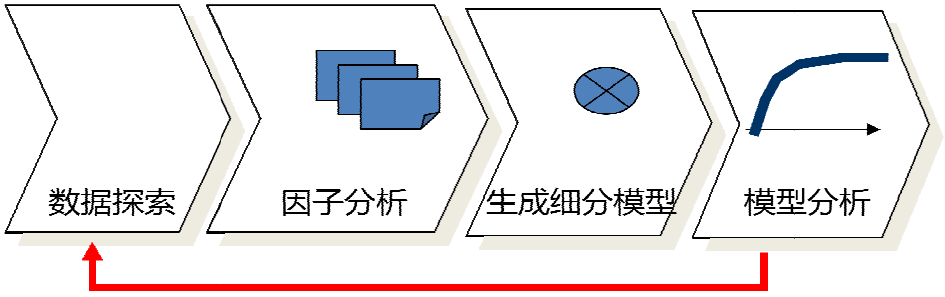


图 2-18 建模过程

1）数据探索

前面所叙述的数据取样，多少是带着人们对如何达到数据挖掘目的的先验的认识进行操作的。当我们拿到了一个样本数据集后，它是否达到我们原来设想的要求；其中有没有什么明显的规律和趋势；有没有出现你所从未设想过的数据状态；因素之间有什么相关性；它们可区分成怎样一些类别……这都是要首先探索的内容。

这里的数据探索，就是我们通常所进行的深入调查的过程。你最终要达到的目的可能是要搞清多因素相互影响的，十分复杂的关系。但是，这种复杂的关系不可能一下子建立起来。一开始，可以先观察众多因素之间的相关性；再按其相关的程度，以了解它们之间相互作用的情况。这些探索、分析，并没有一成不变操作规律性；相反，是要有耐心的反复的试探，仔细的观察。在此过程中，你原来的专业技术知识是非常有用的，它会帮助你进行有效的观察。但是，你也要注意，不要让你的专业知识束缚了你对数据特征观察的敏锐性。可能实际存在着你的先验知识认为不存在的关系。假如你的数据是真实可靠的话，那末你绝对不要轻易地否定数据呈现给你的新关系。很可能这里就是发现的新知识！有了它，也许会导引你在此后的分析中，得出比你原有的认识更加符合实际的规律性知识。假如在你的操作中出现了这种情况，应当说，你的数据挖掘已挖到了有效的矿脉。

对所抽取的样本数据进行探索、审核和必要的加工处理，是保证预测质量所必需的。可以说，预测的质量不会超过抽取样本的质量。数据探索和预处理的目的是为了保证样本数据的质量，从而为保证预测质量打下基础。

表 2-20 至表 2-22，从客户类型、客户状态和在网天数展示了客户的分布情况，结合对电信业务的理解及本此挖掘建模目标，认为公免不能代表普遍用户行为，容易对在聚类形成噪音；在用户状态中仅选择正常用户；选择入网时间 90 天以上用户，保证研究样本拥有完整的研究期间数据。故表中底色部分为需去掉的数据。

表 2-20 客户类型

项目	百分比	数量
用户类型不详	0.01	7
B 普通	73.44	95758
B 重要客户 A1	0.14	185
B 重要客户 A2	0.10	133
B 重要客户 A3	0.18	241
B 重要客户 A4	0.21	276
B 集团客户 B1	0.22	282

B 集团客户 B2	1.13	1473
B 集团客户 B3	2.25	2930
B 集团客户 B4	6.28	8193
B 信誉客户 C1	0.01	11
B 信誉客户 C2	0.03	40
B 信誉客户 C3	1.38	1793
B 信誉客户 C4	1.01	1321
B 信誉客户 C5	1.51	1968
B 信誉客户 C6	1.54	2006
B 潜在大用户	10.29	13411
B 老客户	0.05	70
B 原重要	0.02	21
B 公免	0.21	269

表 2-21 客户状态

项目	百分比	数量
正常	79.05	103066
冒高	17.43	22721
欠停	0.34	448
报停	1.69	2206
挂失	0.02	23
预销	1.28	1669
强开	0.05	62
强关	0.15	193

表 2-22 在网天数

项目	百分比	数量
在网时间 90 天以上	0.79	102942
在网时间 90 天以内	0.21	27446

在记录中发现 should_fee（用户应收费）=favor_fee（优惠费）的用户记录，详细查看其记录，比较异常。在与乐山移动分公司访谈后了解此类记录多为用于测试的号码。于是去掉该部分记录

表 2-23 部分 should_fee=favor_fee 的记录

customer_id	phone_no	should_fee	favor_fee
320035348	13708130041	381.76	381.76
320035349	13708130042	375.02	375.02
320035350	13708130043	797.27	797.27
320035351	13708130044	1090.68	1090.68
320035352	13708130045	6.66	6.66
320035353	13708130046	432.07	432.07
320035354	13708130047	1478.76	1478.76

在通过变量的标记直接去掉非正常值外，还通过查看数据去掉极值，以免极少数用户的行为对

细分结果形成噪音，如短信发送人数极多号码有可能为电信统一向用户发送。另外，本例还从用户的 ARPU 值分布、用户年龄分布、用户性别比例、用户区域分布等方面进行全面探索分析。

2）因子分析

宽表中包含了大量客户数据变量，参与建模的变量太多会削弱主要业务属性的影响，并给理解分群结果带来困难；太少则不能全面覆盖需要考察的各方面属性，可能会遗漏一些重要的属性关系。输入变量的选择对建立满意的模型至关重要。本例结合商业理解，选择有重要业务意义并与数据挖掘目标密切相关的变量；被选择的变量应具备较好的数据质量；被选变量之间相关性不宜太强，如在总量与分量之间只挑选一类参与建模。

3）生成客户细分模型

由于价值变量和行为变量有较强的相关性，本例可以只挑选客户业务收入变量进行客户价值分群，也可只挑选客户消费行为变量进行客户行为分群，根据数据挖掘的商业目标选择一种分群模式。也可以同时用两种分群模式对同一批客户作两次分群，然后根据两次分群的结果进行组合，如先分成 7 个价值分群，再分成 9 个行为分群，组合后会有 63 个子群，见图 3。由于组合后子群数目较多不便分析和管理，本例可以借助透视图分析将特征相似的子群进行归并，建议最终归并成 7 到 9 个分群。进行价值和行为组合分群的好处是，能同时兼顾考虑价值和行为两方面因素对客户分群的影响，更利于对各分群特征的深刻刻画，并能有效消除单次分群产生的偏差，但过程较为复杂并且不能做到对参与分群客户的全覆盖。

4）模型分析

细分前整个用户群数据落差较大。细分后，各组的强势变量的分布趋于平缓，聚类模型将具有相似特征的记录聚在一组。

模型建立是一个螺旋上升，不断优化的过程，在每一次分群结束后，需要判断分群结果在业务上是否有意义，其各群特征是否明显。如果结果不理想，则需要调整分群模型，对模型进行优化，称之为分群调优。分群调优可通过调整分群个数及调整分群变量输入来实现，也可以通过多次运行，选择满意的结果。通常可以依据以下原则判断分群结果是否理想：

✧ 群间特征差异是否明显

各分群之间有明显特性差异；各分群主要的特征各不相同；决定各分群主要特征的变量各不相同或变量的取值属性各不相同。

✧ 群内特征是否相似

各分群有各自明显的特性；各分群有决定其主要特性贡献度最大的变量；决定各分群主要特征的变量在此群中的分布特性与在全体样本中的分布特性有明显差异；

✧ 分群是否易于管理及是否具有业务指导意义

分群的个数及各群人数的分布应相对合理，分群结果能从业务上作出合理理解和解释，并能切合业务需要，实现对客户的深刻洞察，帮助制订合适的营销措施。

5. 客户分群的模型评估

在分群调优过程中本例已经需要对模型进行合理评估。在完成模型建立后，从数据分析的角度来看，模型似乎有很高的质量，然而在模型最后发布前仍有必要更为彻底地评估模型和检查建立模型的各个步骤，从而确保它真正地达到了商业目标。“模型评估”阶段需要对数据挖掘过程进行一次全面的回顾，从而决定是否存在重要的因素或任务由于某些原因而被忽视，此阶段关键目的是决定是否还存在一些重要的商业问题仍未得到充分地考虑。这种回顾也包括质量保证问题，如：过程的每一步是否必要？是否被恰当地执行？是否可以改进？有什么不足及不确定的地方及会产生何种影响？根据评估结果和过程回顾，决定是完成该项目并在适当的时候进行发布，还是开始进一步的反

复或建立新的数据挖掘项目。

6. 客户分群的模型发布

模型的创建通常并不是项目的结尾。即使建模的目的是增加对数据的了解，所获得的了解也需要进行组织并以一种客户能够使用的方式呈现出来。根据需要，发布过程可以简单到产生一个报告，也可以复杂到在整个企业中执行一个可重复的数据挖掘过程。客户分群的结果发布是通过客户群特征刻画和客户群属性分析来展现的。特征刻画是对单个客户群特征的详细描述，属性分析是对客户群之间的属性进行比较分析。

形成客户分群后，对客户群的特征描述直接影响到营销活动的策划和执行。客户群的特征描述是把很多枯燥无味的数据变成生动形象的客户体现，以帮助市场营销人员更好地理解客户群。参与分群的变量决定了各分群的主要特性，除了对这些变量的统计及分布特性进行深入刻画外，对未参加分群的变量也可在特征刻画阶段来考察其统计特性。特征刻画首先进行客户群特征粗略定性比较分析，然后可利用透视图等工具对各客户群宽表变量分类进行详细的定量刻画，见表 2-29 和表 2-30。表中不同底色分别反映了最强特征、次强特征、最弱特征和次弱特征。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 k-Means 改进聚类、DBScan 聚类、EM 聚类等算法进行模型构建。

1. 模型输入

模型的输入包括两部分，一部分是建模专家样本数据（包括训练样本和验证样本）的输入，另一部分是建模参数的输入。

基于本案例的挖掘目标，并结合用户行为特征选取的方向，案例中定义了几组数据作为细分变量，见表 2-24。

表 2-24 细分变量定义

细分变量来源	细分变量	细分变量描述
通话范围	d_local	本地通话
	t_local	
	d_toll_inprn	省内长途通话
	t_toll_inprn	
	d_toll_btwprn	省间长途通话
	t_toll_btwprn	
	d_toll_htm+idd	国际、港澳台长途通话
	t_toll_htm+idd	
活动范围	d_rm_inprn	省内漫游通话
	t_rm_inprn	
	d_rm_btwprn	省际漫游通话
	t_rm_btwprn	
	d_rm_idd	国际漫游通话
	t_rm_idd	

www.tipdm.com

IP 使用情况	d_ip	IP 通话
	t_ip	
与各运营商联系程度	d_mob_vs_Ttl	网内通话比例
	t_mob_vs_Ttl	
	d_uni_vs_Ttl	联通通话比例
	t_uni_vs_Ttl	
	d_phs_vs_Ttl	小灵通通话比例
	t_phs_vs_Ttl	
	d_fix_vs_Ttl	固话通话比例
	t_fix_vs_Ttl	
呼转行为	d_fwd_total	呼转使用
	t_fwd_total	
数据业务使用情况	d_wap	WAP 使用
	t_wap	
	t_sms_total	点对点及网间短信发送次数
	t_sms_mont	梦网短信发送次数
客　服	t_service	到营业厅次数
	t_1860	拨打 1860 次数
	t_1861	拨打 1861 次数

注：其中 d_X 代表时长，t_X 代表次数。

通过细分变量将客户进行细分，再通过表 2-25 所列描述变量定位人群，进一步分析人群特征。

表 2-25 描述变量定义

描述变量来源	描述变量	描述变量来源	描述变量
客户基本信息	age	主叫被叫情况	d_outgoing_total
	male		t_outgoing_total
	female		d_incoming_total
	GenderMissing		t_incoming_total
	CashPay	联系人群范围	distinct_out_total
	PrePay		distinct_in_total
	vip_mark	与联系紧密人群的主被叫情况	d_outgoing_topN
	due_mark		t_outgoing_topN
	tenure_in_days		d_incoming_topN
费用信息	ARPU		t_incoming_topN
	should_fee	各项呼转情况	d_fwd_mob
	favor_fee		t_fwd_mob
	rent_fee		d_fwd_srv
	local_fee		t_fwd_srv

www.tpdm.com

	inprn_fee		d_fwd_uni
	btwprn_fee		t_fwd_uni
	htm_fee		d_fwd_fix
	idd_fee		t_fwd_fix
	rm_inprn_fee	闲时忙时通话情况	d_pkon_total
	rm_btwprn_fee		t_pkon_total
	rm_idd_fee		d_pkoff_total
	rm_toll_fee		t_pkoff_total
	special_fee	工作时间、生活时间通话情况	d_worktime_total
	sms_fee		t_worktime_total
	wap_fee		d_resttime_total
	ip_fee		t_resttime_total
	gprs_fee	各项短信使用情况	t_sms_p2p
	message_fee		t_sms_link
	other_fee		t_sms_topN

表 2-26 表 2-27 表 2-28 为本案例要用到的建模算法参数设置信息。

表 2-26 改进聚类算法建模参数

序号	参数名称	参数描述
1	聚类数	设定的聚类个数
2	种子数	初始随机选取的样本数

表 2-27 DBScan 聚类算法建模参数

序号	参数名称	参数描述
1	半径	聚类区域的半径
2	最小样本数	最小建模样本数

表 2-28 EM 聚类算法建模参数

序号	参数名称	参数描述
1	最大迭代次数	最大循环迭代的次数
2	最小标准误差	设定允许误差最小值
3	聚类数	聚类的个数
4	种子数	初始随机选取的样本数

2. 结果分析

是各组相对强弱势情况比较，表 2-32 和表 2-33 分别是客户细分后针对低使用率组和短信高使用组的细分组信息、客户群分析和客户策略，表 2-34 是对短信高使用组进行人群特征分析后采取的相应市场策略。

表 2-29 细分结果数据表—细分变量

	组 1	组 2	组 3		组 4		组 5	组 6	组 7	组
客户细分群	cluster-1	cluster-8	cluster-12	cluster-9	cluster-7	cluster-5	cluster-10	cluster-3	cluster-11	clus
客户数										
占总客户比例	0.065	0.090	0.197	0.193	0.102	0.158	0.125	0.011	0.035	0.02
ARPU	68.53	98.87	126.41	135.44	145.52	162.74	153.84	194.58	467.68	545.
年龄	30.99	35.39	34.19	34.32	32.79	32.92	32.74	24.18	33.79	33.7
男性比例	0.59	0.64	0.67	0.66	0.69	0.72	0.66	0.47	0.75	0.74
女性比例	0.29	0.27	0.25	0.25	0.22	0.20	0.25	0.44	0.16	0.19
性别不详比例	0.11	0.09	0.09	0.09	0.08	0.08	0.09	0.09	0.08	0.07
现金付费比例	0.09	0.19	0.19	0.21	0.16	0.18	0.18	0.08	0.51	0.51
预付费比例	0.91	0.81	0.81	0.79	0.84	0.82	0.82	0.92	0.49	0.49
VIP 比例	0.07	0.13	0.18	0.20	0.18	0.23	0.20	0.18	0.76	0.81
欠费比例	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.01	0.09	0.09
在网时长（天）	775.00	823.20	852.02	888.43	760.20	827.97	794.64	424.44	1214.86	1152.
应收费	88.89	128.22	164.29	179.07	194.14	220.03	215.79	258.79	692.31	697.
优惠费	20.35	29.35	37.88	43.63	48.61	57.29	61.95	64.21	224.63	151.
月租费	21.56	35.10	38.07	39.20	39.74	42.81	40.71	18.98	48.90	49.3
基本通话费	42.48	59.87	87.67	92.93	115.20	128.86	113.00	97.73	500.66	329.
本地通话费比例	0.48	0.47	0.53	0.52	0.59	0.59	0.52	0.38	0.72	0.47
省内长途费	0.85	2.34	2.47	2.94	2.28	3.10	3.74	3.51	14.29	22.7
省际长途费	1.74	4.52	5.20	6.85	5.10	6.74	9.30	7.90	27.07	65.8
港澳台长途费	0.00	0.02	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.03
国际长途费	0.01	0.34	0.08	0.11	0.03	0.01	0.08	0.00	0.51	0.21
长途费比例	0.03	0.06	0.05	0.06	0.04	0.04	0.06	0.04	0.06	0.13
省内漫游费	3.48	6.71	8.74	11.67	8.67	12.08	17.24	14.85	48.02	127.
省际漫游费	0.43	2.79	2.22	3.33	1.55	2.29	4.02	3.36	6.51	23.3
国际漫游费	0.02	0.22	0.12	0.26	0.14	0.13	0.50	0.00	2.20	0.94
漫游费比例	0.04	0.08	0.07	0.09	0.05	0.07	0.10	0.07	0.08	0.22
漫游长途费	0.24	1.89	1.39	2.05	0.90	1.35	2.34	1.79	4.74	12.0
特服费	6.63	6.75	7.34	7.58	7.74	8.40	8.15	6.17	13.79	14.3
SMS 费	4.59	1.38	2.11	2.24	2.40	2.63	3.71	81.09	3.38	4.32
WAP 费	0.03	0.01	0.01	0.01	0.02	0.02	0.02	0.12	0.04	0.03
IP 费	1.62	2.10	1.99	2.32	2.08	2.42	3.70	9.60	4.19	27.8
IP 费比例	0.57	0.23	0.22	0.19	0.25	0.22	0.24	0.73	0.09	0.28
GPRS 费	0.45	0.10	0.15	0.19	0.14	0.15	0.31	0.36	1.71	2.28
代收信息费	4.34	3.57	4.68	5.12	4.36	5.36	5.28	12.35	10.35	11.2
其它费	0.36	0.55	2.02	2.20	3.71	3.65	3.66	0.60	5.88	6.11

表 2-30 细分结果数据表—描述变量

	组 1	组 2	组 3		组 4		组 5	组 6	组 7	组 8
客户细分群	cluster-1	cluster-8	cluster-12	cluster-9	cluster-7	cluster-5	cluster-10	cluster-3	cluster-11	clus
客户数										
占总客户比例	0.065	0.090	0.197	0.193	0.102	0.158	0.125	0.011	0.035	0.020
ARPU	68.53	98.87	126.41	135.44	145.52	162.74	153.84	194.58	467.68	545.8
年龄	30.99	35.39	34.19	34.32	32.79	32.92	32.74	24.18	33.79	33.72
男性比例	0.59	0.64	0.67	0.66	0.69	0.72	0.66	0.47	0.75	0.74
女性比例	0.29	0.27	0.25	0.25	0.22	0.20	0.25	0.44	0.16	0.19
性别不详比例	0.11	0.09	0.09	0.09	0.08	0.08	0.09	0.09	0.08	0.07
现金付费比例	0.09	0.19	0.19	0.21	0.16	0.18	0.18	0.08	0.51	0.51
预付费比例	0.91	0.81	0.81	0.79	0.84	0.82	0.82	0.92	0.49	0.49
VIP 比例	0.07	0.13	0.18	0.20	0.18	0.23	0.20	0.18	0.76	0.81
欠费比例	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.01	0.09	0.09
在网时长（天）	775.00	823.20	852.02	888.43	760.20	827.97	794.64	424.44	1214.86	1152.
应收费	88.89	128.22	164.29	179.07	194.14	220.03	215.79	258.79	692.31	697.
优惠费	20.35	29.35	37.88	43.63	48.61	57.29	61.95	64.21	224.63	151.
月租费	21.56	35.10	38.07	39.20	39.74	42.81	40.71	18.98	48.90	49.3
基本通话费	42.48	59.87	87.67	92.93	115.20	128.86	113.00	97.73	500.66	329.
本地通话费比例	0.48	0.47	0.53	0.52	0.59	0.59	0.52	0.38	0.72	0.47
省内长途费	0.85	2.34	2.47	2.94	2.28	3.10	3.74	3.51	14.29	22.7
省际长途费	1.74	4.52	5.20	6.85	5.10	6.74	9.30	7.90	27.07	65.8
港澳台长途费	0.00	0.02	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.03
国际长途费	0.01	0.34	0.08	0.11	0.03	0.01	0.08	0.00	0.51	0.21
长途费比例	0.03	0.06	0.05	0.06	0.04	0.04	0.06	0.04	0.06	0.13
省内漫游费	3.48	6.71	8.74	11.67	8.67	12.08	17.24	14.85	48.02	127.
省际漫游费	0.43	2.79	2.22	3.33	1.55	2.29	4.02	3.36	6.51	23.3
国际漫游费	0.02	0.22	0.12	0.26	0.14	0.13	0.50	0.00	2.20	0.94
漫游费比例	0.04	0.08	0.07	0.09	0.05	0.07	0.10	0.07	0.08	0.22
漫游长途费	0.24	1.89	1.39	2.05	0.90	1.35	2.34	1.79	4.74	12.0
特服费	6.63	6.75	7.34	7.58	7.74	8.40	8.15	6.17	13.79	14.36
SMS 费	4.59	1.38	2.11	2.24	2.40	2.63	3.71	81.09	3.38	4.32
WAP 费	0.03	0.01	0.01	0.01	0.02	0.02	0.02	0.12	0.04	0.03
IP 费	1.62	2.10	1.99	2.32	2.08	2.42	3.70	9.60	4.19	27.82
IP 费比例	0.57	0.23	0.22	0.19	0.25	0.22	0.24	0.73	0.09	0.28
GPRS 费	0.45	0.10	0.15	0.19	0.14	0.15	0.31	0.36	1.71	2.28
代收信息费	4.34	3.57	4.68	5.12	4.36	5.36	5.28	12.35	10.35	11.22
其它费	0.36	0.55	2.02	2.20	3.71	3.65	3.66	0.60	5.88	6.11

表 2-31 各组相对强弱势比较

分组号	细分编号	强势特征	弱势特征
组 1：低使用率组	1	与小灵通通话多，拨打 1860、1861 次数	本地、省内长途漫游、省间长途、短信、IP、与联通用户通话
组 2：固话联系紧密组	8	与固定电话通话多	本地、省内长途漫游、省间长途、短信、IP、与联通用户通话
组 3：中低使用率组	12	与固定电话通话多	省际长途、IP 电话
	9		与联通通话次数
组 4：联通联系紧密组	7	与联通用户通话时长、次数	省际漫游
	5	与联通用户通话时长、次数	
组 5：移动联系紧密组	10	与移动用户通话时长、次数	与联通和固话的通话时长、次数
组 6：短信高使用组	3	短信、梦网短信、WAP、拨打 1860、1861	港澳台、国际长途漫游。
组 7：大量本地通话组	11	本地通话时长、次数。呼叫转移。与移动用户通话时长、次数	
组 8：高 IP&省内漫游组	4	IP 通话时长、次数。省内漫游	

表 2-32 低使用率组

细分组信息	数量：　　比例：6% 年龄：30.99　ARPU：68　在网时间(月)：27　女性比例：29% 预付费比例：81%　　VIP 比例：7%
客户群分析	✧ 本组用户对手机基本使用通话功能，并且通话需求低。平均每天 2.8 次 ✧ 主要在本地活动，并且也主要与市内用户沟通。联系圈子小。并且与小灵通联系较多。 ✧ 每个市内、长途电话的时长在 12 组人群中居第一、二位，但次数却最少。可看出此类人群一般是有事情才打电话，所以谈话时间会长，但次数却少。 ✧ 优惠时段通话次数在所有组中最高，占了整个通话次数的 5%。 ✧ 短信使用量相对其他低端人群较多，结合其低 ARPU 值的特征，估计因为资费敏感而用会采用短信替代一定通话行为 ✧ IP 通话电话次数占整个通话次数比例较高，有 28%，可看出其对资费非常敏感。 ✧ 主叫电话次数是被叫电话次数的 1.6 倍，主叫时长是被叫时长的 1.79 倍，是 12 组中最高的一组。并且与经常联系的人群也是主叫居多。结合其资费敏感的特征，分析其被叫行为可能因为资费因素而被压抑。 ✧ 访问营业厅次数少。拨打 1860、1861 次数多。由于 1861 为话费查询电话，可见此客户群对资费比较敏感。
客户策略	产品/服务： ✧ 此类人群对资费非常敏感，可能会对省钱类产品感兴趣。 ✧ 此类人群被叫行为可能有所压抑，可通过被叫包月类产品释放其被叫需求。 ✧ 由于人群圈子窄，并且主要为市内通话，容易成为小灵通的抢夺对象，可设计亲情号码类产品。 推广： ✧ 估计此类人群中包括低工资人群、家庭妇女、老人、学生等人群、社交需求少的技术人员，所以可选择超市、商场、学校等地进行海报宣传或安排人员发送宣传单。 ✧ 在人群聚集地进行一定促销活动，实物奖励会对该组人群有一定吸引力 ✧ 1860、1861 接听宣传 销售渠道：

www.ctom.com

	✧ 除基本代销点外，可挑选此客户群体的居民楼附近超市、小商品店等作为代销点。

表 2-33 短信高使用组

细分组信息	数量： 比例：1% 年龄：24 ARPU：194 在网时间(月)： 14 女性比例：44% 预付费比例：92% VIP 比例：18%
客户群分析	✧ 其最明显特征是点对点短信、网间短信、梦网短信、WAP 等数据业务使用量大，估计此类人群对新技术比较敏感，对活动参与性高，信息渠道广泛。 ✧ 通话次数居中，平均每天 5 次左右。每个电话通话时间在 12 组中最高。优惠时段通话次数、时长最高。IP 使用占整个通话的 30%，并且每个 IP 电话时间最长。综合考虑，估计此类人群有较强的长聊需求，对资费比较敏感。 ✧ 通话联系人群居中，但短信联系人群最高，有 27.78 人。与联系最密切的交往圈短信发送量也最高，平均每天 2.2 条。可见其短信联系面广且密切。可能一方面是资费敏感，另一方面是该组人群喜欢并习惯短信渠道进行交流，还有一部分可能喜欢传播短信。 ✧ 拨打 1860、1861 次数最高，特别是 1861 资费查询号码，平均每月拨打 11.52 次。可见该人群对资费比较敏感。
客户策略	产品： ✧ 该组用户对资费比较敏感，且有沟通和长聊的需求，可设计随心聊等进一步拉动长聊行为的产品。 ✧ 此类客户应对新推出的增值业务特别是基于短信应用的增值业务产品很感兴趣。 ✧ 由于此群体平均年龄在 24 岁，是用户组中最年轻的人群，并且女性用户比例最高，可考虑一些适用年轻人群或女性用户的信息类产品。 推广： ✧ 该群体中可能包括大部分学生，所以可选择学校、电影院、商场作为宣传渠道。可进行海报，发送传单，活动促销等宣传。 ✧ 短信渠道。 ✧ 网络渠道宣传。 ✧ 时尚杂志广告。 ✧ 销售渠道： ✧ 学校附近的代销点。 ✧ 电子城、手机市场代销点等。 其他： ✧ 该群体入网时间最短，平均 14 个月，需从培养其忠实度角度设计产品和服务。

表 2-34 针对短信高使用人群的策略

组号	人群特征分析	市场策略
普通短信使用组	通话需求一般。对短信使用率较低。	该组用户量大，可通过促销或市场活动拉动该组用户短信需求量。可推荐实用性梦网短信。如笑话、菜谱等。
点对点短信热衷组	热衷于短信交流的年轻人群。短信交流范围广，对资费有一定考虑	可利用该组人群转播短信，如笑话类。组建短信俱乐部等。活动对这部分用户具有一定吸引力

www.cidm.com

网间短信高使用组	无论短信还是通话，都与联通用户联系紧密。	定期发送关怀短信，发送移动近期活动等，与用户保持联系，防止流失。
技术敏感组	梦网使用次数、托收信息费。对新业务新技术较敏感	有新的数据业务可先让该组人群尝试。
高端本地商务组	大量本地通话、呼转，估计包括本地商务人士、政府机关人员等	体现关怀。如节假日发送关怀短信等。
中端移动型商务组	有大量长途、漫游通话需求，并且IP使用量大。估计包括业务员、中端商旅人士等	节假日发送段子，供其传递。可推荐信息类梦网短信，如新闻、天气预报
高端移动型商务组	大量长途漫游，对资费不敏感。社交人群广泛	节假日发送礼节性短信，供商用。可推荐信息类梦网短信，如新闻、天气预报

实践是检验真理的唯一标准，模型发布后也需要在营销实践中验证调整。另外，数据挖掘结果已经成为日常业务和其环境的一部分后，在运用模型的过程中，模型的监测和维护也是十分重要的事情。周密的维护策略将有助于避免不必要地长期误用数据挖掘结果。为了监测数据挖掘结果的发布，项目需要根据应用类型制定一个关于监测过程的详细计划，例如定期察看各分群的主要特性是否已产生较大的偏移和变迁。这些监测一个方面可以对营销效果进行评估反馈，另一方面也为模型的维护和调整提供决策依据，如是否需要运行模型重新生成分群？是否应对模型作出调整或重新生成模型？是否应终止模型的使用？

在获取了客户分群结果，了解了客户之后，应该针对所掌握的情况，有所行动。通过分群本例有了对客户的整体感觉：哪些是最有价值的，哪些是最需要关注的，哪些可能是对企业贡献度最低等等。通过客户分群，得到了的客户的不同贡献及不同的消费特征，进而得到的整体优先级考虑是进行下一步工作的基础和起点，也是客户分群最为重要的收获之一。在客户分群后，接下来就可以进行营销活动的设计与执行了。

2.2. 电力行业

2.2.1. 案例五 电力负荷预测

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

由于电力工业与一般的其它产业不同，其产品即电能无法大量储存，电力的生产和消费必须在同一瞬间进行，电网结构示意图见图 2-19，因此电力负荷预测成了电力系统运行调度、生产规划、电力市场竞价决策的重要组成部分。做好电力负荷预测管理工作可以有效降低电网公司运行成本和提高电力设备运行效率，其预测精度不仅影响到电网安全可靠供电，而且直接影响到电网经营企业的生产经营决策及经营效益。随着电力改革的深化、电力市场的开放，进一步提高短期负荷预测管理水平愈显得重要和迫切。

目前国内外的预测应用软件大多基于特定的少数的几种模型，而选择模型单一造成的后果是预

测结果往往只对某种发展规律有效，当事物发展规律改变时，如果仍然采用原有的单一模型，就会造成预测结果偏差过大，从而失去了预测的实际意义。尤其是对于使用系统的各个供电公司，由于发展水平不同，用电结构不同，负荷特性差异很大，特定的某种预测方法将很难在各地都发挥出良好的效果。另外，在电力市场环境下，影响电力发展的因素非常多，包括经济发展、能源消费、气象条件等众多影响因素，加之不同系统间数据共享性差，很难考虑相关因素的影响等，导致目前基于用电侧的电力负荷预测效果不甚理想。

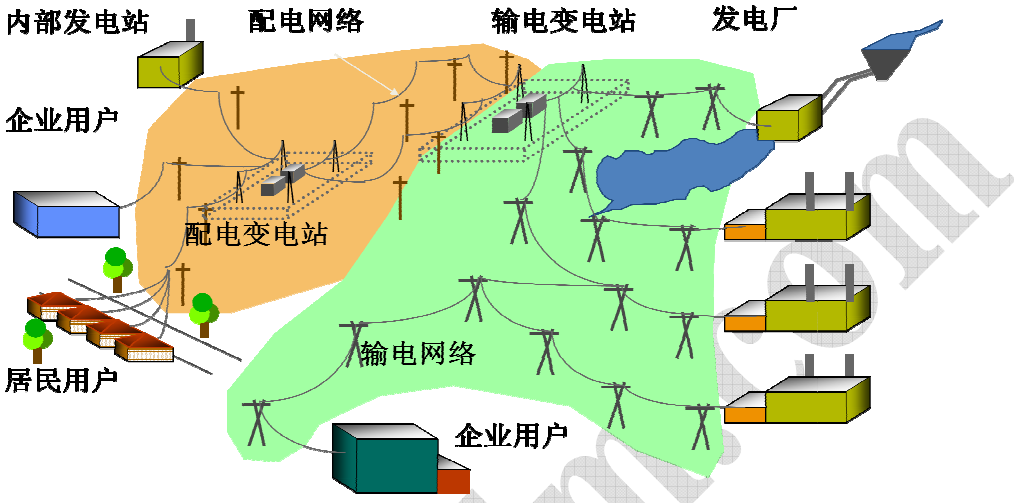


图 2-19 电网结构示意图

基于用电侧数据如何进行有效的电力负荷预测，为供电部门进行安全监视、预防性控制和紧急状态处理提供依据？

本案例的挖掘目标是：

1） 基于用电特征对用户进行分群

2） 采用基于客户分群的方法进行用电侧负荷预测，并评价其效果

分析方法与过程

电力系统负荷一般可以分为城市民用负荷、商业负荷、农村负荷、工业负荷以及其他负荷等，不同类型的负荷具有不同的特点和规律。城市民用负荷主要是城市居民的家用负荷，它具有经常的年增长以及明显的季节性波动特点，而且民用负荷的特点还与居民的日常生活和工作的规律紧密相关。商业负荷主要是指商业部门里的照明、空调、动力等用电负荷，覆盖面积大，且用电增长迅速，同样具有季节性变动的特性。商业部门由于商业行为在节假日会增加营业时间，从而成为节假日中影响电力负荷的重要因素之一。工业负荷是指工业企业用于工业生产的用电，其负荷比重在用电构成中一般居于首位，它不仅取决于工业用户的工作方式（包括设备利用情况、企业的工作班制等），而且与各工业行业的行业特点、季节因素都有紧密的联系。农村用电负荷则是指农村居民用电和农业生产用电，受气候、季节等自然条件的影响很大，这是由于农业生产的特点所决定的。我国地域广阔，由于地理位置的原因，各地降雨季节有较大差异，一场大雨对北方而言可造成农业负荷骤降，但却可能造成南方地区排涝用电负荷剧增。

在实际分类过程中，由于大量的用户并未明确其类型特征，而且每个用户的具体特性也并不完全符合几个大类的特征，因此大量的用户细分工作还需要通过聚类分析来自动地进行聚合。用户细分中需考虑的因素包括行业分类、电价类别、电压等级、功率因数，日总用电量、峰用电量、谷用电量、平用电量，以及计量自动化系统中收集到的每 15 分钟的用电负荷数据等等。

本案例中，进行负荷预测时，采用先对用户进行细分，再分别进行预测的方法。即先对用户按

行业进行分类，对于占比较大的行业，通过因子分析找出用于聚类分群的指标变量，并利用聚类算法将用户分为不同用电特征的几类，然后将群组内计量点负荷叠加后再分别利用时序预测算法进行负荷预测，最后将各种预测算法下的负载预测结果进行加权求和，形成最终的预测结果。另外还将实际预测结果与历史实际数据进行比较，对预测结果评价，并反馈至预测模型，通过调整建模参数，提升预测模型精度[20]。

1. 用户分群

1）分群数据准备

电力用户（其实是针对计量点，若一个用户对应有多个计量点，则应视为多个用户）分群的数据准备，其目的是生成用于用户分群的样本数据集，主要包括电力用户档案类（属性）数据、计量类（行为）数据，并对这两类数据进行联合，生成最终便于聚类分析的宽表数据集，宽表数据生成流程见图 2-20。

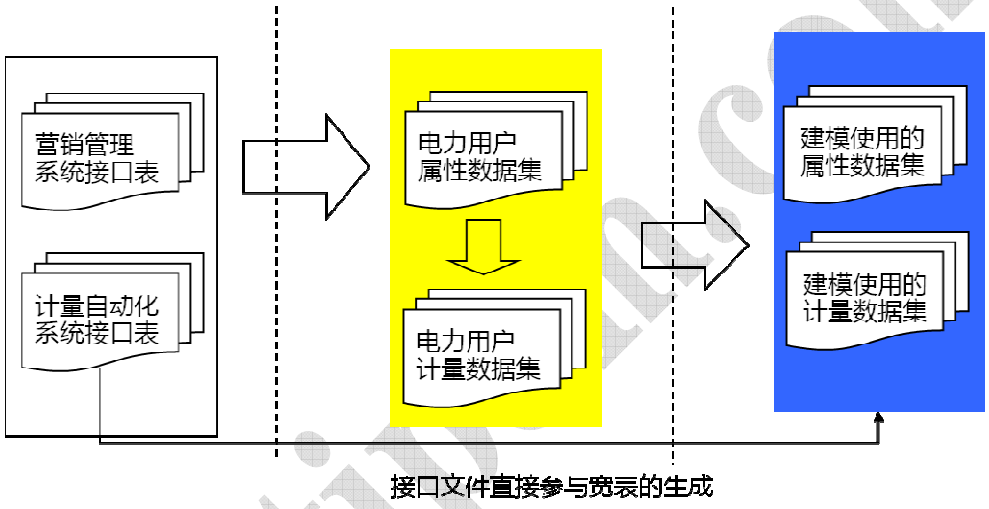


图 2-20 电力用户宽表数据生成流程

电力用户分群的数据准备，从以下 3 个方面来准备数据：

- ✧ 用户档案信息（例如：用电类别、行业类别、电压等级、倍率等）；
- ✧ 用户电量信息（例如：总用电量、峰用电量、平用电量、谷用电量、尖用电量、无功总电量等）；
- ✧ 用户负荷信息（电流、电压、功率因数、有功功率等）。

电力负荷预测，主要考虑对负荷波动的影响较大的影响因子，故除了用电类别、行业类别、最大功率外，增加几个用来衡量在规定时间内负荷变动情况的计算指标：负荷率（平均负荷/最大负荷）、峰总比（峰电量/总电量）、平总比（平电量/总电量）、谷总比（谷电量/总电量）。

2）用户分群

案例采用 K-means 聚类算法，对电力用户进行分群，K-均值聚类(K-means clustering)是 Mac Queen 提出的一种非监督实时聚类算法，在最小化误差函数的基础上将数据划分为预定的类数 K。该算法原理简单并便于处理大量数据。在 K-means 算法运行前必须先指定聚类数目 K 和迭代次数或收敛条件，并指定 K 个初始聚类中心，根据一定的相似性度量准则，将每一个用户分配到最近或“相似”的聚类中心，形成类，然后以每一类的平均矢量作为这一类的聚类中心，重新分配，反复迭代直到类收敛或达到最大的迭代次数。

2. 时序预测

1）电力负荷趋势特征

以某供电局 2010 年至 2011 年的用电负荷为例，我们来分析其用电负荷趋势特征。

✧ 实时负荷趋势分析

实时负荷走势及相关数据分别见图 2-21 和表 2-35。

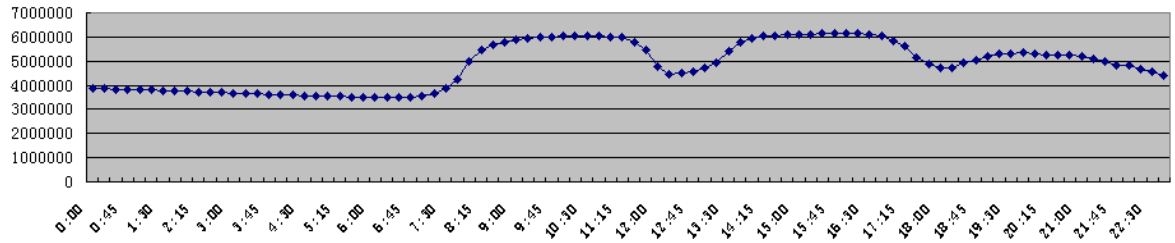


图 2-21 实时负荷走势图

表 2-35 实时负荷数据

时间点	实时负荷	时间点	实时负荷
0:00	3878945.9	12:00	4757384.0
0:15	3846909.1	12:15	4438244.4
0:30	3833372.8	12:30	4494877.1
0:45	3822145.2	12:45	4550353.9
1:00	3794444.5	13:00	4725130.9
1:15	3792574.2	13:15	4937103.3
1:30	3783239.0	13:30	5424533.6
1:45	3782948.9	13:45	5767906.5
2:00	3765410.1	14:00	5961429.4
2:15	3735782.9	14:15	6045112.1
2:30	3713090.4	14:30	6067456.9
2:45	3687288.7	14:45	6094653.9
3:00	3669258.5	15:00	6092863.3
3:15	3640069.3	15:15	6113014.9
3:30	3640213.3	15:30	6132568.6
3:45	3620153.8	15:45	6141525.0
4:00	3584586.3	16:00	6130849.6
4:15	3585294.1	16:15	6126057.3
4:30	3567898.7	16:30	6094743.7
4:45	3551999.5	16:45	6038652.7
5:00	3541774.6	17:00	5827923.7
5:15	3529791.6	17:15	5638672.2
5:30	3498463.7	17:30	5132126.4
5:45	3499502.6	17:45	4870687.3
6:00	3491580.8	18:00	4706205.3
6:15	3500498.1	18:15	4720383.9
6:30	3513732.4	18:30	4917195.0
6:45	3522376.5	18:45	5031562.0
7:00	3559071.4	19:00	5180512.0
7:15	3643313.7	19:15	5283401.2
7:30	3891367.4	19:30	5315447.8

www.ppt.com

7:45	4233821.7	19:45	5332899.1
8:00	4981080.7	20:00	5280510.7
8:15	5471625.8	20:15	5263271.4
8:30	5674677.2	20:30	5243938.5
8:45	5777193.0	20:45	5229307.2
9:00	5874952.1	21:00	5185996.3
9:15	5928102.5	21:15	5075564.9
9:30	5973342.7	21:30	4997823.9
9:45	5998919.8	21:45	4850054.2
10:00	6025379.1	22:00	4840681.2
10:15	6025378.2	22:15	4652426.1
10:30	6035454.3	22:30	4562117.4
10:45	6020803.3	22:45	4411943.0
11:00	5995270.2	……	……
11:15	5982486.3		
11:30	5771786.2		
11:45	5485443.3		

✧ 日最大负荷分析

日最大负荷走势及相关数据分别见图 2-22 和表 2-36。

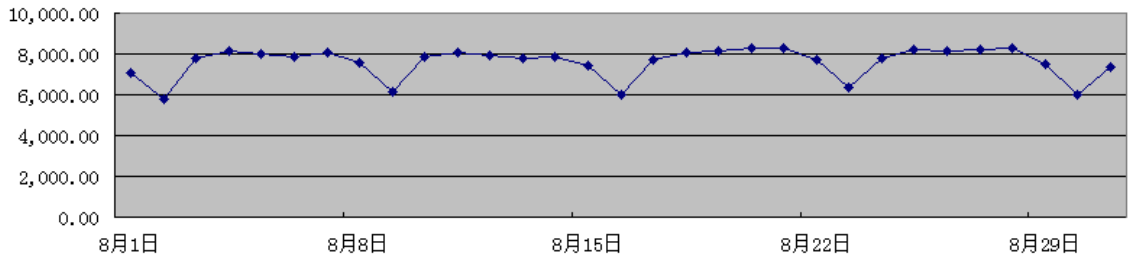


图 2-22 日最大负荷走势

表 2-36 日最大负荷数据

日期	负荷极值	日期	负荷极值
8月1日	7,048.33	8月17日	7,701.56
8月2日	5,811.08	8月18日	8,083.37
8月3日	7,794.47	8月19日	8,158.03
8月4日	8,172.99	8月20日	8,258.37
8月5日	7,996.86	8月21日	8,277.68
8月6日	7,855.03	8月22日	7,729.60
8月7日	8,043.59	8月23日	6,324.36
8月8日	7,559.65	8月24日	7,750.98
8月9日	6,145.04	8月25日	8,211.27
8月10日	7,873.33	8月26日	8,138.19
8月11日	8,080.94	8月27日	8,231.13
8月12日	7,900.36	8月28日	8,263.95
8月13日	7,785.64	8月29日	7,502.56
8月14日	7,834.43	8月30日	6,022.70
8月15日	7,437.19	8月31日	7,347.94
8月16日	6,018.17	……	……

✧ 月最大负荷分析

月最大负荷走势及相关数据分别见图 2-23 和表 2-37。

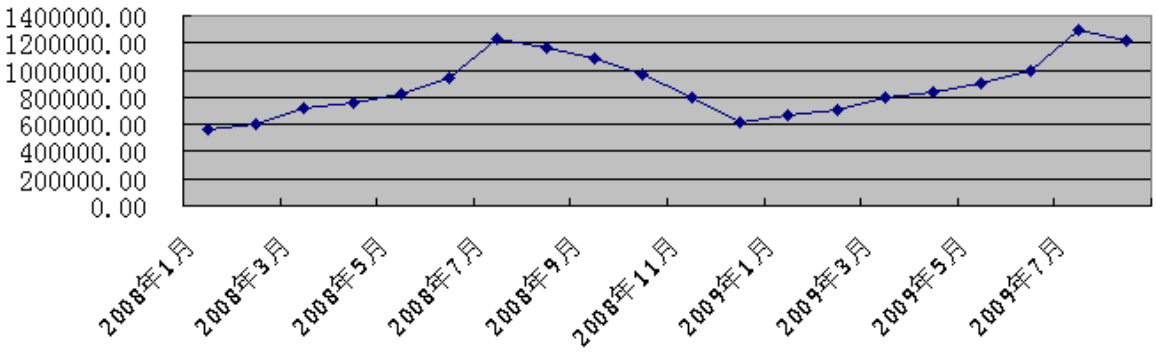


图 2-23 月最大负荷走势

表 2-37 月最大负荷数据

月份	月负荷极大值	月份	月负荷极大值
2008 年 1 月	562461.77	2009 年 1 月	663461.27
2008 年 2 月	602330.65	2009 年 2 月	712330.65
2008 年 3 月	715244.44	2009 年 3 月	795245.27
2008 年 4 月	753920.59	2009 年 4 月	833924.53
2008 年 5 月	824777.25	2009 年 5 月	898477.21
2008 年 6 月	945914.52	2009 年 6 月	988914.56
2008 年 7 月	1233605.93	2009 年 7 月	1299605.03
2008 年 8 月	1160289.66	2009 年 8 月	1210289.74
2008 年 9 月	1084564.78	……	……
2008 年 10 月	974567.32		
2008 年 11 月	798394.67		
2008 年 12 月	619425.48		

由电力负荷趋势有关图表可以看出，电力负荷走势有一明显的特性就是记忆性(指时间数列中的任一观测值的表现皆受到过去观测值影响)。故在进行负荷预测时主要考虑的因素是：

✧ 长期趋势(Long-term trend)
- ➢ 时间序列可能相当稳定或随时间呈现某种趋势。
- ➢ 时间序列趋势一般为线性的(linear)，二次方程式的 (quadratic)或指数函数(exponential function)。

✧ 季节性变动(Seasonal variation)
- ➢ 按时间变动，呈现重复性行为的序列。
- ➢ 季节性变动通常和日期或气候有关。
- ➢ 季节性变动通常和年周期有关。

✧ 周期性变动(Cyclical variation)
- ➢ 相对于季节性变动，时间序列可能经历“周期性变动”。
- ➢ 周期性变动通常是因为经济变动。

✧ 随机影响(Random effects)

负荷预测分为超短期预测、短期预测和中长期预测，不同的预测周期对样本数据的要求和数据结构各不相同。比如超期负荷预测，取样数据可以最好为实时数据；短期负荷预测，取样数据最好为日数据；中期负荷预测，取样数据最好为日数据或月数据。在明确了预测目标后，除了实时负荷外，进行日最大负荷、月最大负荷预测时，需要对采样的实时负荷数据进行重新分组。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的基于分布式计算的 k-Means 聚类和时间序列等算法进行模型构建。首先采用聚类算法对用户进行分群，然后对群组进行负荷叠加并对叠加后的负荷通过时序预测算法，完成负荷预测。

1. 模型输入

模型的输入包括两部分，一部分是建模样本数据的输入，另一部分是建模参数的输入。

表 2-38 用户分群输入变量

序号	变量名称	变量描述
1	计量点号	描述变量，不参与建模
2	用户名称	描述变量，不参与建模
3	电压等级	描述变量，不参与建模
4	总用电量	描述变量，不直接参与建模
5	峰用电量	描述变量，不直接参与建模
6	平用电量	描述变量，不直接参与建模
7	谷用电量	描述变量，不直接参与建模
8	尖用电量	描述变量，不直接参与建模
9	无功总电量	描述变量，不直接参与建模
10	……	
11	用电类别	描述变量，不参与建模
12	行业类别	描述变量，不参与建模
13	最大功率	细分变量
14	负荷率	细分变量，平均负荷/最大负荷
15	峰总比	细分变量，峰电量/总电量
16	平总比	细分变量，平电量/总电量
17	谷总比	细分变量，谷电量/总电量

表 2-39 时序预测输入变量

序号	变量名称	变量描述
1	时间	1）对于实时负荷预测，时间为整 15 分钟 2）对于日最大负荷预测，时间为日 3）对于月最大负荷预测，时间为月
2	负荷值	不同时间对应的采集（或计算）负荷值

www.tipdm.com

表 2-40 k-Means 均值聚类算法建模参数

序号	参数名称	参数描述
1	显示标准误差	是否显示出标准误差
2	距离函数	计算样本距离的公式
3	缺失值不处理	如果缺失值处理的话，就用平均值代替，或则不处理
4	最大迭代次数	最大循环迭代的次数
5	聚类数	设定的聚类个数
6	保留实例序列	保留样本原来的顺序或序列
7	种子数	初始随机选取的样本数。

表 2-41 SVM 支持向量机建模参数

序号	参数名称	参数描述
1	SVM 类型	设定 SVM 的类型， SVM 主要解决三类问题：分类问题（包括 C-SVC、nu-SVC）、回归问题（包括 epsilon-SVR、nu-SVR）以及分布估计（one-class SVM）等问题
2	核函数类型	选择核函数，目前采用的核函数，从特征空间的基来看，主要分为两类：其一，将核函数使用泰勒级数展开，特征空间的基是多项式类；其二，将核函数使用傅立叶级数展开，特征空间的基是三角函数
3	终止误差	循环终止的设定误差
4	标准化数据	输入数据是否需要标准化处理
5	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果

表 2-42 线性回归建模参数

序号	参数名称	参数描述
1	属性选择策略	如 M5 方法、Greedy 方法
2	岭参数	对数似然值的岭参数
3	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果

表 2-43 逐步回归建模参数

序号	参数名称	参数描述
1	评估器	评估函数
2	阈值	评估函数对应的阈值
3	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果

www.cdinfo.com

表 2-44 保序回归建模参数

序号	参数名称	参数描述
1	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果
2		

2. 仿真过程

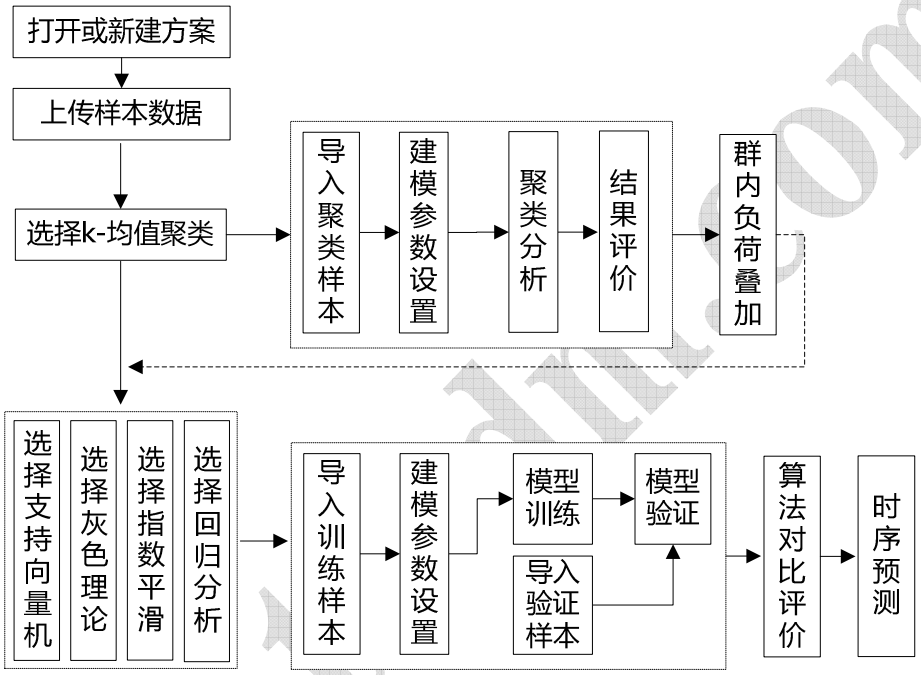


图 2-24 仿真过程

建模仿真过程说明如下：

1） 登录 TipDM 平台，在方案管理页面中，创建一个新方案：基于 k-Means 的电力用户分群。

2） 切换到数据管理页面，上传经预处理后的用于聚类分群的样本数据文件。

3） 选择 k-Means 聚类算法对导入的样本数据进行聚类分析。

4） 对聚类后的计量点负荷按分群结果进行实时负荷叠加。

5） 分别选择时序模式->GM 灰色理论、指数平滑、回归分析、SVM 支持向量机等时序算法。

6） 分别用 GM 灰色理论、指数平滑、回归分析、SVM 支持向量机等时序算法对导入的分组时序数据进行模型构建。

7） 对 GM 灰色理论、指数平滑、回归分析、SVM 支持向量机等时序算法的建模结果进行比较。

8） 选择最优算法进行时序预测（实时负荷预测、日最大负荷预测和月最大负荷预测要分别建模）。

3. 结果及分析

1）电力客户分群

选取某供电局 2000 个大客户 2012 年 5 月运行数据的日平均值，选择 k-Means 均值聚类，经过反复试验，最后选定聚类数为 20，经过 21 次迭代后，最后集群内误差平方和：2.973，各组样本数

量、占比等分类结果如表 2-45 所示。

表 2-45 电力用户聚类后的分类结果

	样本数量	占比	最大功率	峰总比	平总比	谷总比	负荷率
组 1	372	31.00%	2.2985	0.4131	0.4935	0.0934	0.2509
组 2	1	0.08%	9.6964	0.3040	0.4646	0.2314	0.0037
组 3	3	0.25%	0.1226	0.6491	0.0496	0.3013	0.2792
组 4	86	7.17%	0.1810	0.0000	0.5619	0.0722	0.1602
组 5	20	1.67%	1.2065	0.2289	0.3001	0.4709	0.2167
组 6	8	0.67%	0.6676	0.3332	0.6457	0.0211	0.0633
组 7	6	0.50%	0.5139	0.3558	0.3384	0.3059	0.6090
组 8	23	1.92%	0.4467	0.3160	0.4791	0.2049	0.5352
组 9	18	1.50%	0.6037	0.3826	0.4999	0.7074	0.8237
组 10	4	0.33%	1.5637	0.1110	0.1548	0.7342	0.2102
组 11	38	3.17%	0.5137	0.4223	0.4802	0.0975	0.1051
组 12	7	0.58%	0.2956	0.0000	0.5168	0.1945	0.2440
组 13	52	4.33%	0.5140	0.0000	0.3016	0.4786	0.9634
组 14	68	5.67%	1.0448	0.4415	0.5364	0.0221	0.8076
组 15	35	2.92%	0.4406	0.2965	0.3994	0.3041	0.2258
组 16	50	4.17%	0.7783	0.2999	0.6821	0.0180	0.2534
组 17	23	1.92%	1.5314	0.3497	0.1270	0.5233	0.4867
组 18	301	25.08%	3.3815	0.2143	0.3732	0.4124	0.9002
组 19	51	4.25%	0.0908	0.1900	0.6095	0.2005	0.1122
组 20	34	2.83%	1.9184	0.0000	0.3003	0.5019	0.2340

2）时序预测

在本例中，首先对聚类后的群组进行负荷叠加，然后针对三类不同的负荷特征（实时负荷、日最大负荷、月最大负荷），分别群组内计量点负荷进行叠加，采用不同的时序算法，完成负荷预测。

图 2-25、图 2-26 和图 2-27 分别为实时负荷（样本数据为第 4 组电力用户实时负荷叠加结果，采用 2012 年 5 月 1 日至 7 月 31 日的 15 分钟实时负荷建模，预测 2012 年 8 月 1 日 0:00 后 100 个时间点的实时负荷）的时序曲线图和参数选择结果图。

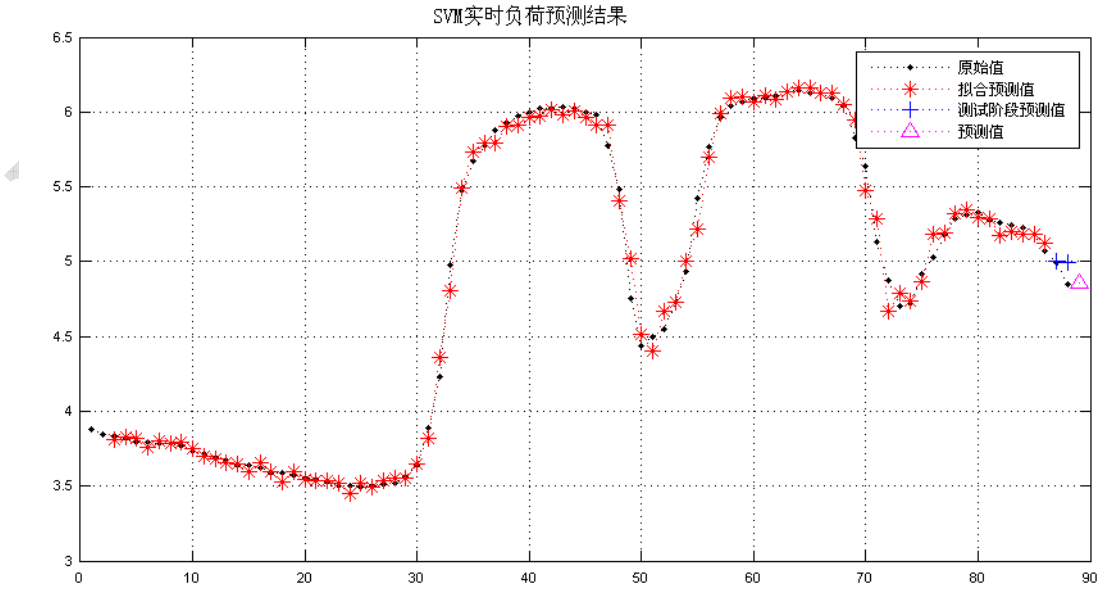


图 2-25 SVM 实时负荷预测结果

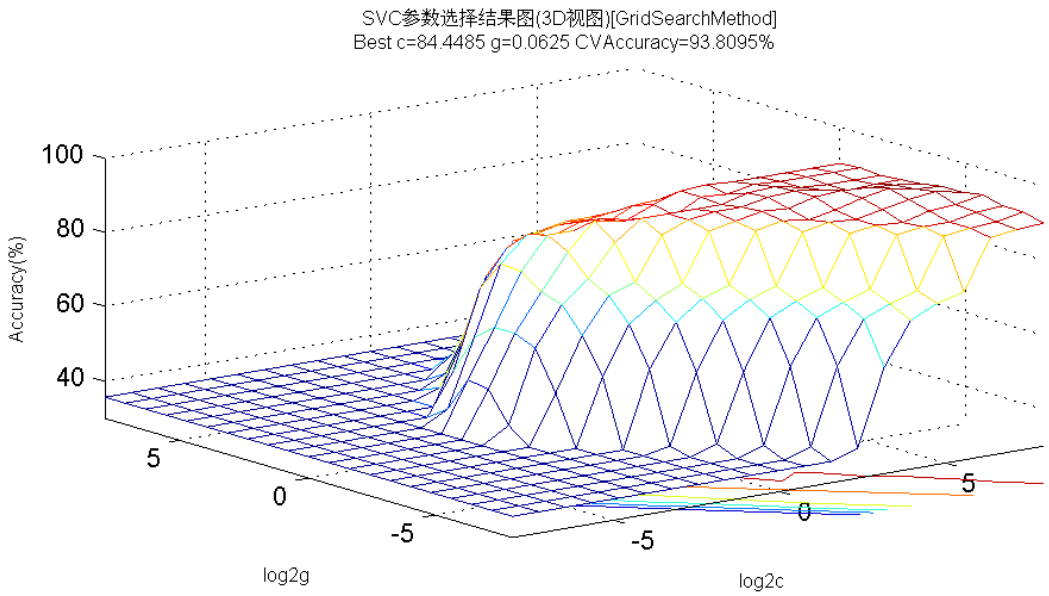


图 2-26 SVM 建模时参数选择结果 3D 视图

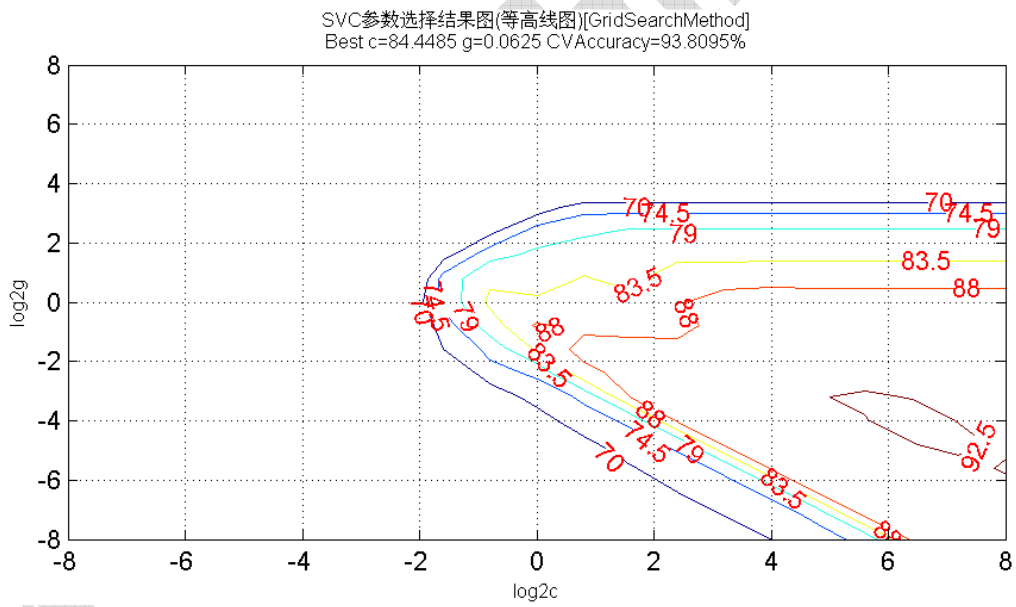


图 2-27 SVM 建模时参数选择结果等高线图

2.2.2. 案例六 防窃漏电自动诊断

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

窃电和计量装置故障造成漏收、少收电费使电力系统利益受损。传统方法主要通过定期巡检、定期校验电表、用户举报窃电等手段来发现窃电或计量装置故障。对人的依赖性太强，抓窃查漏的目标不明确。为了更好的发挥计量自动化系统的远程在线监测平台，目前很多供电局主要通过营销稽查人员、用电检查人员和计量工作人员利用计量异常报警功能和电能量数据查询功能开展用户用电情况的在线监控工作，通过采集电量异常、负荷异常、终端报警、主站报警、线损异常等信息，建立数据分析模型，来实时监测窃漏电情况和发现计量装置的故障。并通过 “失压失流”、“超合同容量用电”、“电流过负荷”等用电报警信息，以及根据报警事件发生前后客户计量点有关的电流、电压、负荷数据情况等，通过构建基于指标加权的用电异常分析模型，实现检查客户是否存在窃电、违章用电及计量装置故障等目的。

以上防窃漏电的诊断方法，虽然能获得用电异常的某些信息，由于终端误报或无用信息太多，无法达到真正快速精确定位窃漏电嫌疑用户和计量故障的目的，往往令稽查工作人员无所适从。而且在采用这种方法建模时，模型各输入指标权重的确定需要用专家的知识和经验，具有很大的主观性，存在明显的缺陷，所以实施效果往往不尽如人意。

目前各供电局已基本建成了涵盖各种计量点及采集终端，集信息采集、监控、分析和计量管理于一体的计量自动化应用平台，完成了对电厂、变电站、公变、专变、低压集抄等发电侧、供电侧、配电侧、售电侧的综合性统一数据的自动采集监控，电力计量数据采集过程示意图见图 2-28。已实现对居民用电的远程集中抄表功能，计量自动化系统与营销管理系统、生产调度系统等均不同程度的通过数据接口实现了数据共享，现在应用系统中已存储了大量用户用电信息。

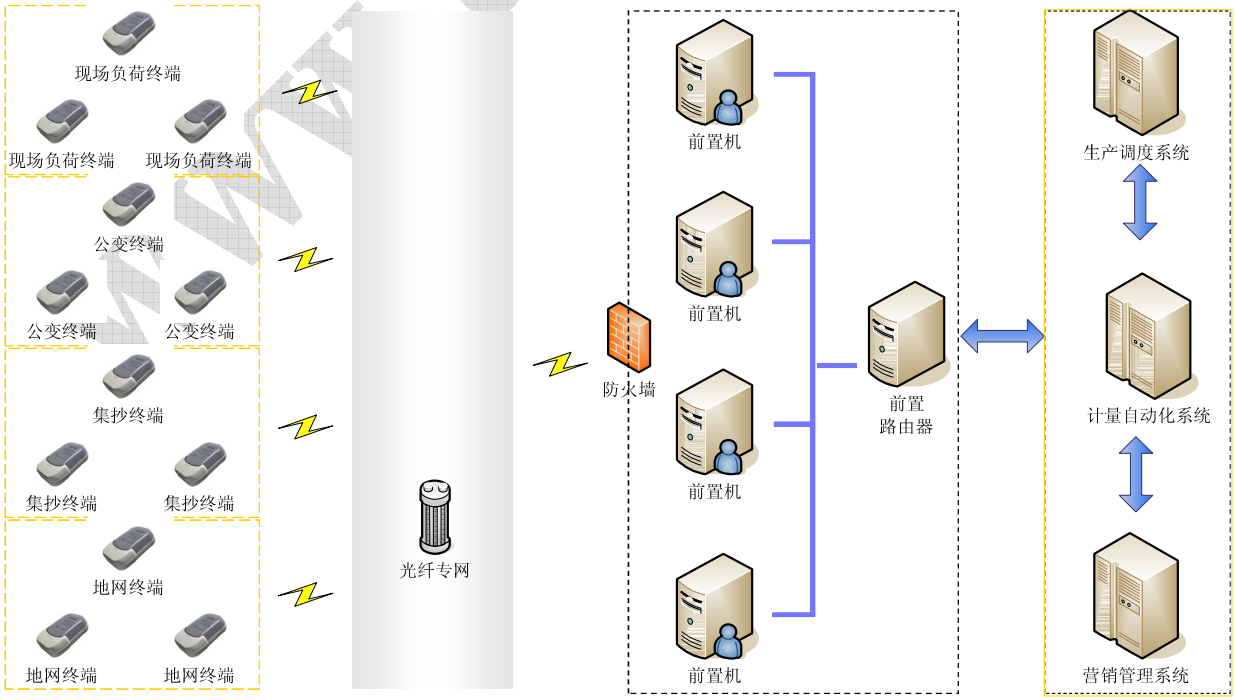


图 2-28 电力计量数据采集过程示意图

如何对这些历史数据进行科学的分析，归纳出不同类型用户的用电特征和规律，真正实现快速精确定位窃漏电嫌疑用户的目的？

分析方法与过程

防窃漏电诊断系统是一个分析型系统，为开展窃电侦查工作和计量装置检查工作提供依据和锁定范围。首先，系统需要提供漏窃电嫌疑户列表，并给出每户的漏窃电嫌疑系数。市场部和计量部就可以根据这份嫌疑户列表到现场排查，判断是窃电还是计量装置故障。接下来如果需要计算追补电量，系统应提供参考值。

本案例重点介绍如何进行窃漏电用电特征提取及自适应窃漏电诊断模型的构建。

窃电者为了达到窃电目的，最常见的是从电能计量原理入手。一个电能表计量电量的多少，主要决定于电压、电流、功率因数三素和时间的乘积，因此，只要想办法改变三要素中的任何一个要素都可以使电表慢转、停转甚至反转，从而达到窃电的目的；另外，通过采用改变电表本身的结构性能的手法，使电表慢转，也可以达到窃电的目的。

1. 窃漏电用电特征分析

窃漏电虽然手法五花八门，但万变不离其宗，即都会反映到电量及负荷的变化上，通过采集和分析某供电局近三年来所有的窃漏电大用户及随机选取 8%正常用户的用电特征（用电负荷具有类似的规律，在此不作详述），总结如下：

1）正常用电电量特征分析

正常用电量特征表现见图 2-29 至图 2-30，由图 2-31 窃漏电用电量特征表现见知，电量移动差分最大为 2，表示连续两天上升趋势，大部分为 0，表示几乎不变，这就是用户正常用量的电量指标特征。

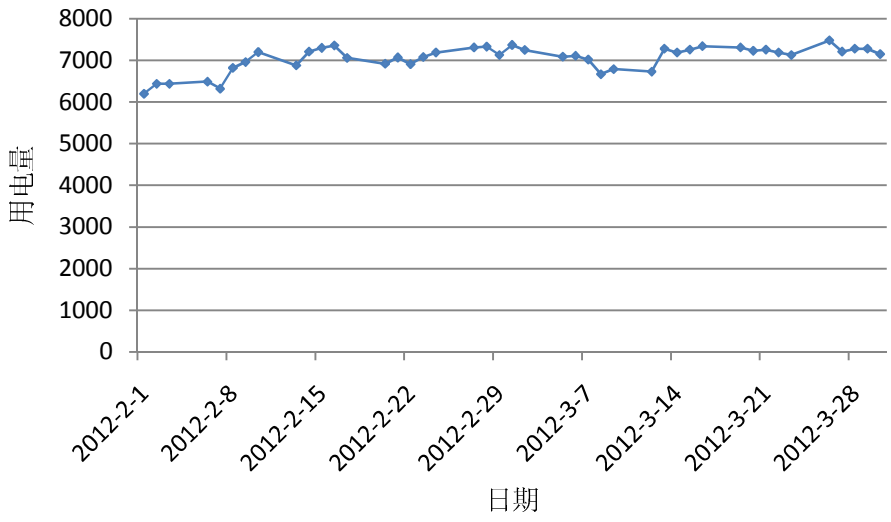


图 2-29 正常用电电量趋势

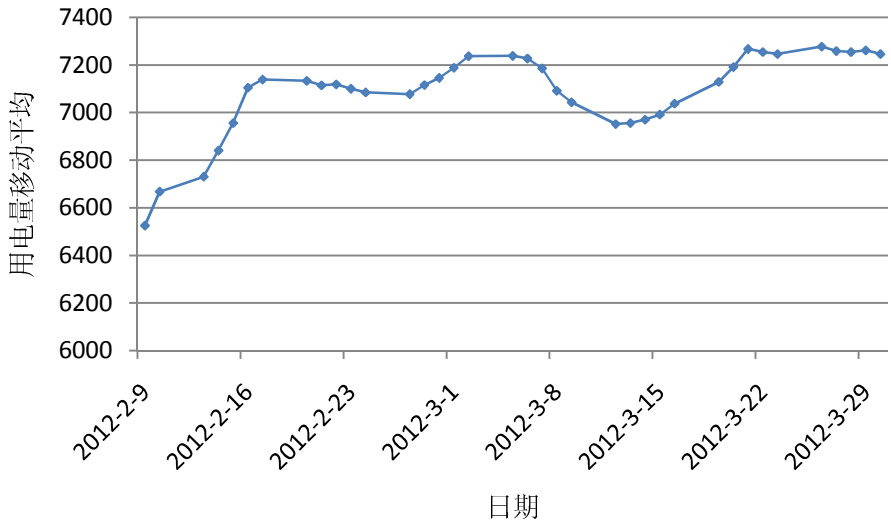


图 2-30 正常用电电量移动平均走势

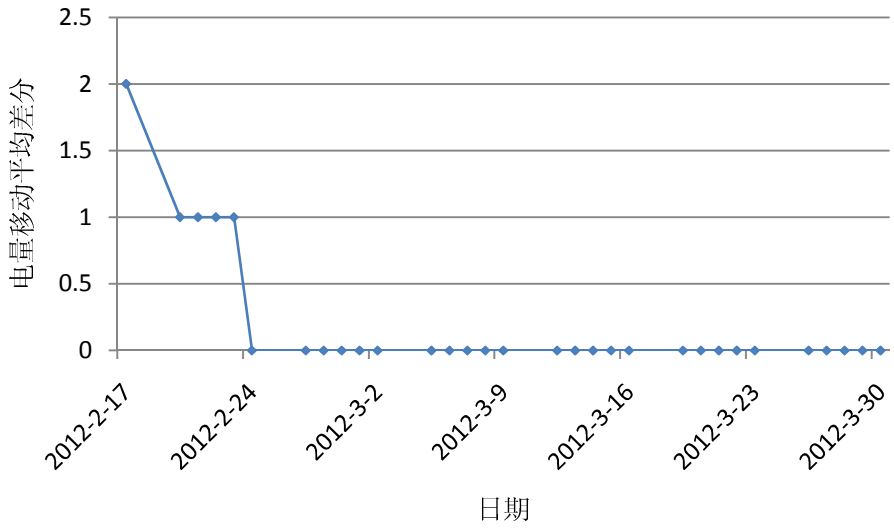


图 2-31 正常用电电量移动平均差分

2）从正常用电到窃电电量特征分析

窃漏电用电量特征表现见至图 2-32 和图 2-34，由图 2-34 知，正常用电到窃电过程是用电量下降的过程，用电移动差分数据下降到一个大负数，这就是用户开始窃电时所表现出来的重要特征。

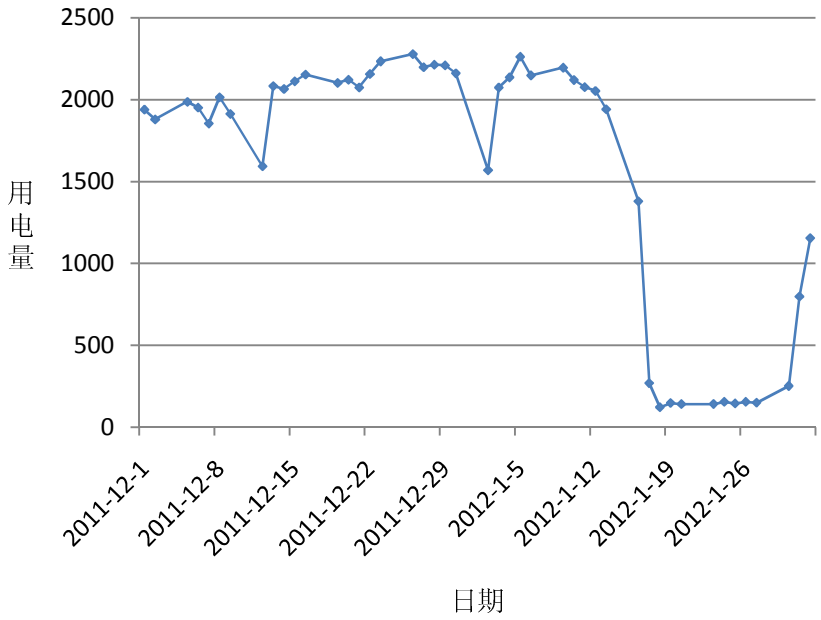


图 2-32 从正常用电到窃电电量走势

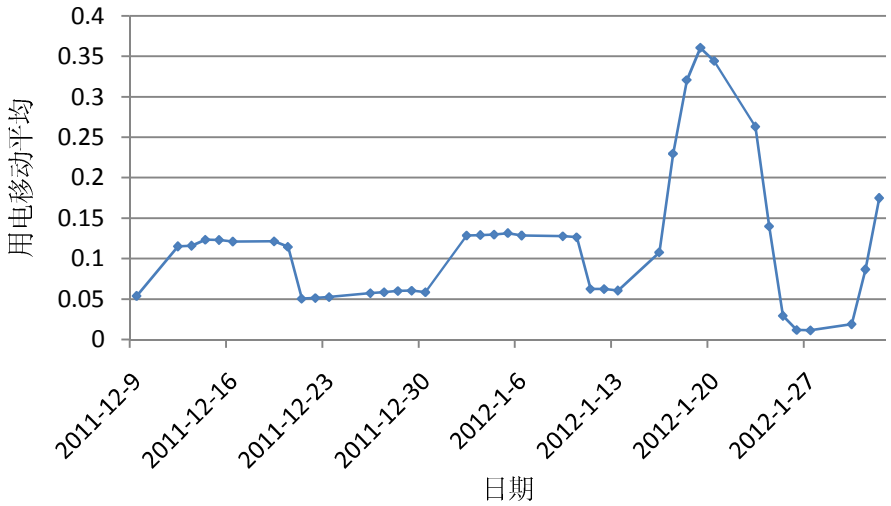


图 2-33 从正常用电到窃电电量移动平均走势

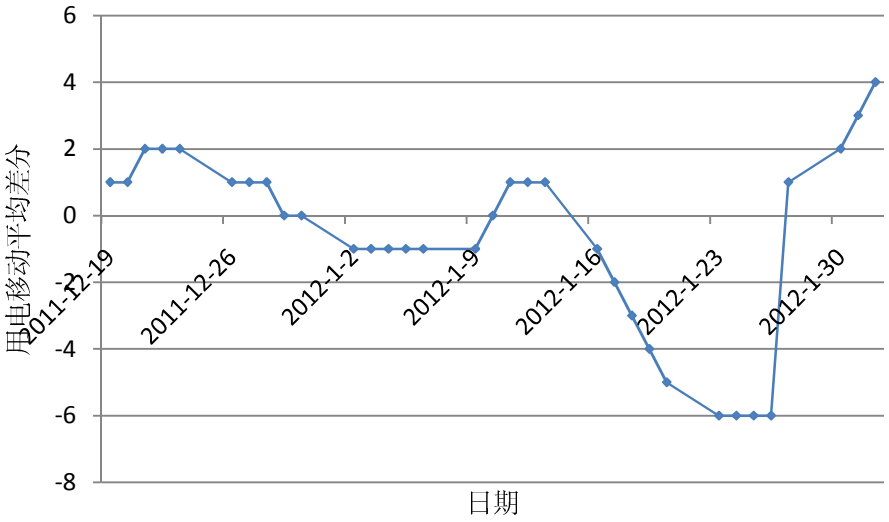


图 2-34 从正常用电到窃电电量移动平均差分

2. 窃漏电评价指标体系

通过电力计量系统采集的电量、负荷虽然在一定程度上能反映用户窃漏电行为的某些规律，但要作为构建模型的专家样本，特征不明显，需要进行重新构造。基于前面的分析，总结若干新的评价指标来表征窃漏电行为所具有的规律，并确立了如图 2-35 所示的窃漏电评价指标模型。

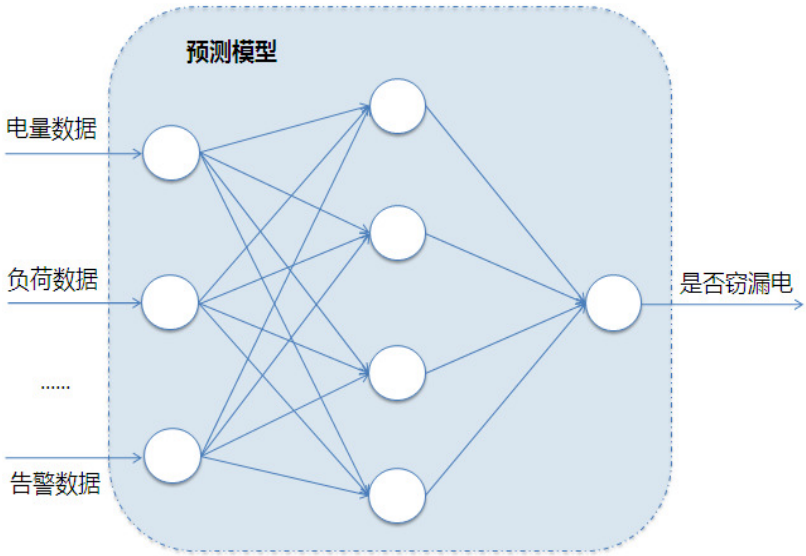


图 2-35 窃漏电评价指标模型

3. 模型构建

防窃漏电自动诊断主要包括以下步骤：

步骤 1：样本数据抽取　由图 2-35 知，窃漏电评价影响因子主要有电量类指标、负荷类指标、线损类指标和报警类指标等，故进行窃漏电诊断时需根据建模要求有选择性地从营销系统及计量系统中抽取电量、负荷、报警及线损等数据。

为了尽可能全面覆盖各种窃漏电方式，建模样本包含近三年来所有的窃漏电用户及部分正常用

户。

窃漏电用户的窃漏电开始时间和结束时间是表征其窃漏电的关键时间节点，见表 2-46《违约、窃电处理通知书》，在这些时间节点上，电量、负荷、线损、报警和六角图等指标也会有明显的特征变化，故样本数据抽取时务必要包含关键时间节点前后各两个月的数据。

表 2-46 违约、窃电处理通知书

确定依据：（1）、你户提供的检讨书变代时间（2011年9月至10月期间窃电）；（2）、计费电能表厂家检测报告反映，你户计费电能表2011年9月2日计费电能表存在失压失流记录，同时失压后电流数值比失压前电流数值明显减少；（3）、抄表用电记录（2011年10月1日及之后抄见电量均比2011年9月1日及之前抄见电量明显减少）.

现确定你户从2011年9月2日起开始窃电至2011年10月20日止，共48天。

步骤 2：样本数据预处理　数据预处理主要包括缺失值处理、异常值处理、节假日数据处理、负荷数据平滑处理等。

✧　缺失值处理

在原始计量数据，特别是实时负荷数据抽取过程中，发现存在缺失的现象，为确保建模数据的有效性，需要对这些缺失数据进行补齐处理。规则主要为通过同类型日数据结合插值算法进行处理。

✧　异常值处理

在原始样本数据中，存在大量异常值的情况，如-10000，以及一些超出指标阀值范围的数据，对于这类异常数据，同样需要通过同类型日数据结合插值算法进行修正处理。

✧　节假日数据修正

节假日的用电量及实时负荷数据同工作日比起来，会明显偏低，为了确保不同日期类型的计量数据在一定时间周期内具有可比性和连贯性，需要对这些特殊类型日期的数据进行修正处理。

✧　负荷数据平滑处理

在进行数据（5L 指标）格式化时，考虑到样本数据的一些边际效应，有必要对实时负荷数据进行平滑处理。通过反复试算，本案例中采用最高次数为 20 的多项式拟合，可以达到最佳的处理效果。

步骤 3：构建专家样本库　基于前一步骤的预处理数据，结合窃漏电评价指标体系，构建专家样本库。专家样本以日为单位表征各计量点的用电情况，输入项包括 5E 指标、5L 指标、报警指标、线损指标和六角图指标，输出项为窃漏电嫌疑系数。

步骤 4：构建窃漏电诊断模型　在样本数据准备完成后，即可进行模型构建，即分析电力系统积累下来的大量窃漏电历史数据，对各类用电异常的状况分类整理，并给出相应的判定结果，形成专家样库，最后利用数据挖掘技术将这些数据中蕴藏着的许多潜在的重要因素、事实和关联等有价值的信息提炼出来，分析窃漏电事件同其它因素（如负荷、电量、报警等）的相关性，并进而构建智能诊断模型，真正做到高效、实时地进行智能诊断。

防窃漏电评价可通过构建分类预测模型来实现，比较常用的分类预测模型有人工神经网络，由于在窃漏电诊断中，涉及到大量的复杂现象和多种因素的相互作用，而模糊神经网络结合了模糊评价法与神经网络评价法的优点，在解决这类问题时具有明显优势，因此本案例主要基于 FNN 模糊神

www.cpdm.com

经网络进行防窃漏电建模，模型构建流程如图 2-36。

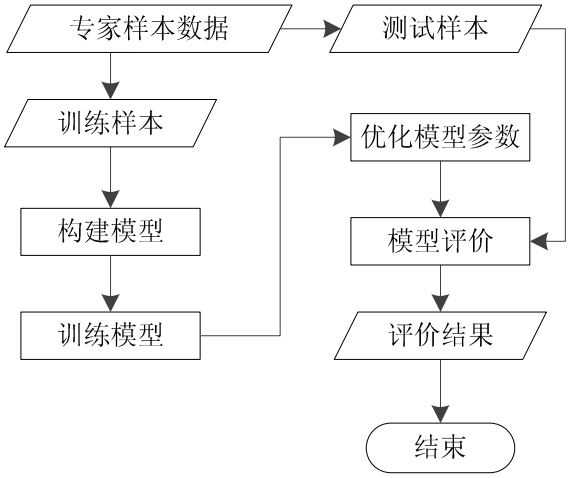


图 2-36 窃漏电诊断模型构建

步骤 5：进行窃漏电诊断　在模型构建完成后，就可以基于实时采集的计量数据调用训练好的模型实现对用户的窃漏电诊断。

步骤 6：优化模型参数，并重构模型。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 BP 神经网络和 LM 神经网络优化算法进行模型构建。

1.　模型输入

模型的输入包括两部分，一部分是建模专家样本数据（包括训练样本和验证样本）的输入，另一部分是建模参数的输入。

表 2-47 预测模型输入变量

序号	变量名称	变量描述	取值范围
1	OBJ_ID	计量点编号	0~1
2	DATA_DT	计量日期	0~1
3	X1	5E 指标斜率突变量	-5~+5
4	X2	5E 指标截距突变量	-5~+5
5	X3	5L 指标斜率突变量	-5~+5
6	X4	5L 指标截距突变量	-5~+5
7	X5	线损同比超差率	0~1
8	X6	线损偏差率	0~1
9	X7	相位角异常次数	>=0
10	X8	报警次数	>=0
11	X10	上一日窃漏电嫌疑系数	0~1
12	Y	诊断日窃漏电嫌疑系数	0~1

说明：本案例中未对指标 X5~X8 进行采集及参与建模。

表 2-48 相关分析建模参数

序号	参数名称	参数描述

www.tipdm.com

1	计算方法	选择相关分析计算方法，如 Pearson、Sperarman、Kendall
2		

表 2-49 属性选择建模参数

序号	参数名称	参数描述
1	属性评估器	对属性/属性子集进行评估确定
2	搜索策略	确定搜索算法

表 2-50 逻辑回归建模参数

序号	参数名称	参数描述
1	岭参数	对数似然值的岭参数
2	最大迭代次数	最大训练（迭代）次数
3	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果

表 2-51 LM 神经网络建模参数

序号	参数名称	参数描述
1	显示间隔次数	两次显示时间的间隔
2	最大循环次数	模型训练时最大循环（迭代）次数，当训练时循环次数达到此参数时，网络训练将会终止
3	目标误差	网络的训练目标误差
4	学习动量	权重更新量与上次权重更新量之间的关系，$\Delta\omega_{ji}(n)=\eta\delta_j x_{ji}+\alpha\Delta\omega_{ji}(n-1)$，其中$\eta$为学习速率，$\alpha$为学习动量
5	学习速率	Marquart 调整参数，即网络权重更新的加权因子。学习速率设得太小，网络收敛速度慢，而且可能会收敛于局部最小点，学习速率设得太大的话，可能会导致网络不稳定
6	学习速率增加比率	mu 的上升因子
7	学习速率减少比率	mu 的下降因子
8	最大误差比率	mu 的最大值
9	隐层节点数	隐层节点个数
10	输入层至隐含层传递函数	输入层至隐含层传递函数，通常为双曲正切 S 型函数
11	隐含层至输出层传递函数	隐含层至输出层传递函数，通常为线性函数

表 2-52 FNN 神经网络建模参数

序号	参数名称	参数描述
1	进化次数	最大进化（迭代）次数

www.bdtm.com

2	隐层节点数	隐含层节点数
3	修正系数	神经网络修正系数
4	学习率	神经网络学习率

2. 仿真过程

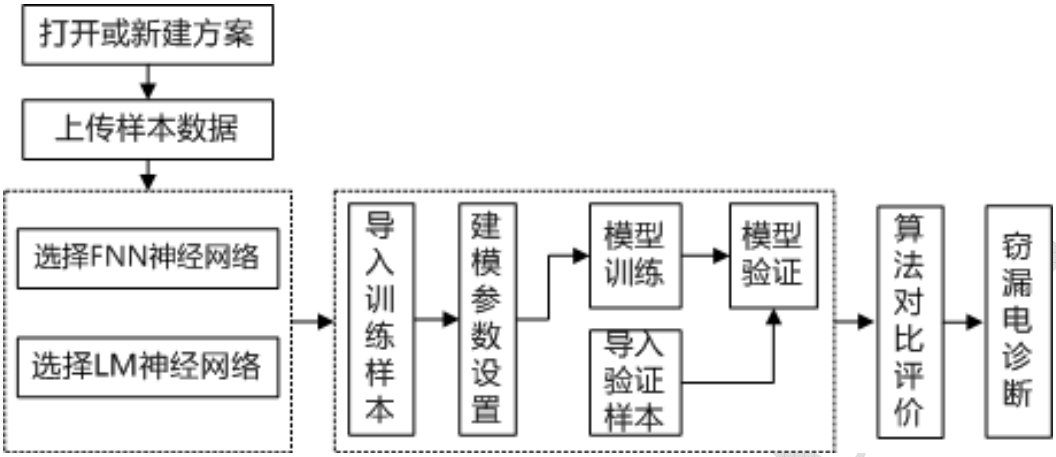


图 2-37 仿真过程

建模仿真过程说明如下：

1) 登录 TipDM 平台，在方案管理页面中，新建方案或者打开一个已建方案。
2) 切换到数据管理页面，上传经预处理后的专家样本数据文件。
3) 分别选择分类与回归->神经网络菜单中的 BP 神经网络及 LM 神经网络算法。
4) 分别用 FNN 神经网络及 LM 神经网络算法对导入的专家样本数据进行模型构建。
5) 对 FNN 神经网络和 LM 神经网络的建模结果进行比较。
6) 选择最优算法进行窃漏电诊断。

2.3. 互联网行业

2.3.1. 案例七 商业零售行业中的购物蓝分析

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

商业零售行业，商场以获得最大的销售利润为目的，零售商都在考虑如下问题：销售什么样的商品？采用什么样的促销策略？商品在货架上如何摆放？了解顾客的购买习惯和偏爱，会使他们对以上问题做出正确的决策具有指导意义。如何对商场销售数据进行分析，从而得到顾客的购买特性，并根据发现的规律而采取有效的行动。如何分析商场销售商品的构成，从而发现不同类别商品的共同特征及其规则，并进而通过这些规则对商场的市场定位、商品定价、新商品采购等进行决策？

分析方法与过程

近年来，关联规则挖掘技术已经被广泛应用在金融行业企业中，它可以成功预测银行客户需求。一旦获得了这些信息，银行就可以改善自身营销。现在银行天天都在开发新的沟通客户的方法。各银行在自己的 ATM 机上就捆绑了顾客可能感兴趣的本行产品信息，供使用本行 ATM 机的用户了解。如果数据库中显示，某个高信用限额的客户更换了地址，这个客户很有可能新近购买了一栋更大的住宅，因此会有可能需要更高信用限额，更高端的新信用卡，或者需要一个住房改善贷款，这些产品都可以通过信用卡账单邮寄给客户。当客户打电话咨询的时候，数据库可以有力地帮助电话销售代表。销售代表的电脑屏幕上可以显示出客户的特点，同时也可以显示出顾客会对什么产品感兴趣。

“如何从大型数据库中挖掘关联规则呢？”关联规则的挖掘是一个两步的过程：

（1）根据最小支持度找出事务数据库 D 中所有的频繁项目集。

（2）由频繁项目集和最小支持度产生强关联规则，也可以使用附加的兴趣度来对规则进行量度。

第一个子问题的任务是关联规则挖掘的中心问题，如何迅速高效的找出 D 中所有频繁项目集。第二个子问题求解比较容易、直接，但也需要根据不用的应用领域设置求解指标（如兴趣度的度量标准），关联规则挖掘的基本模型如图 2-38 所示。

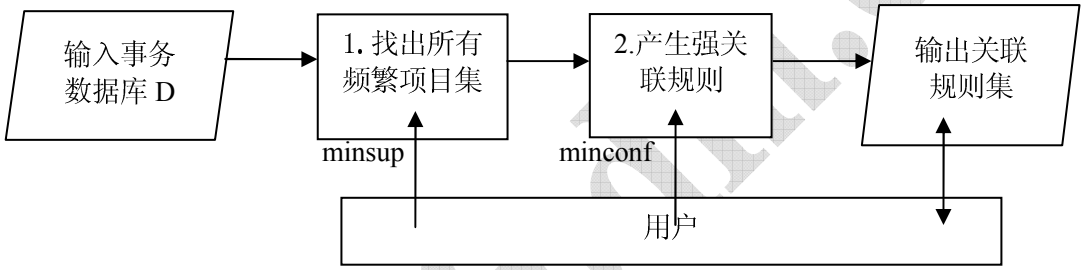


图 2-38 关联规则挖掘的基本模型

图 2-38 中 D 为输入的事务数据库，第一步根据搜索算法找出频繁项目集，第二步的算法从频繁项目集中产生有用的关联规则，最后输出挖掘出的关联规则集合。用户可以通过指定 minsup 和 minconf 分别与第一步和第二步交互，并通过与输出集合的交互对结果进行解释和评估。

如何来度量一个规则是否够好？有两个量，置信度(Confidence)和支持度(Support)。假设有如下表的购买记录。

表 2-53 购买记录表

顾客	项目
1	orange juice, coke
2	milk, orange juice, window cleaner
3	orange juice, detergent
4	orange juice, detergent, coke
5	window cleaner

将上表整理一下，得到如表 2-54购买记录经转换后的2维表的一个购买记录转换后的2维表。

表 2-54 购买记录经转换后的 2 维表

	Orange	Win Cl	Milk	Coke	Detergent
Orange	4	1	1	2	2
WinCl	1	2	1	0	0
Milk	1	1	1	0	0
Coke	2	0	0	2	1

Detergent	1	0	0	0	2

表中横栏和纵栏的数字表示同时购买这两种商品的交易条数。如购买有Orange的交易数为4，而同时购买Orange和Coke的交易数为2。

置信度表示了这条规则有多大程度上值得可信。设条件的项的集合为A,结果的集合为B。置信度计算在A中，同时也含有B的概率。即 Confidence(A==>B)=P(B|A)。例如计算"如果Orange则Coke"的置信度。由于在含有Orange的4条交易中，仅有2条交易含有Coke.其置信度为0.5。

支持度计算在所有的交易集中，既有A又有B的概率。例如在5条记录中，既有Orange又有Coke的记录有2条。则此条规则的支持度为 2/5=0.4。现在这条规则可表述为，如果一个顾客购买了Orange,则有50%的可能购买Coke。而这样的情况（即买了Orange会再买 Coke）会有40%的可能发生。

再来考虑下述情况。

项	支持度
A	0.45
B	0.42
C	0.4
A and B	0.25
A and C	0.2
B and C	0.15
A,B， and C	0.05

可得到下述规则

规则	置信度
If B and C then A	0.05/0.15*100%=33.33%
If A and C then B	0.05/0.20*100%=25%
If A and B then C	0.05/0.25*100%=20%

上述的三条规则，哪一条规则有用呢？

对于规则" If B and C then A"，同时购买B和C的人中，有33.33%会购买A。而单项A的支持度有0.45,也就是说在所有交易中,会有45%的人购买A.看来使用这条规则来进行推荐,还不如不推荐，随机对顾客进荐好了。

为此引入另外一个量，即提升度(Lift)，以度量此规则是否可用。描述的是相对于不用规则，使用规则可以提高多少。有用的规则的提升度大于1。计算方式为 Lift(A==>B)=Confidence(A==>B)/Support(B)=Support(A==>B) /(Support(A)*Support(B))。在上例中，Lift(If B and C The A)=0.05/(0.15*0.45)=0.74。而Lift(If A then B)=0.25/(0.45*0.42)=1.32。也就是说对买了A的人进行推荐B,购买概率是随机推荐B的1.32倍。

如何产生规则呢。可以分两步走。

首先找出频繁集(frequent itemset)。所谓频繁集指满足最小支持度或置信度的集合。其次从频繁集中找出强规则(strong rules)。强规则指既满足最小支持度又满足最小置信度的规则。

我们来看如何产生频繁集。

这其中有一个定理。即频繁集的子集也一定是频繁集。比如，如果{A,B,C}是一个3项的频繁集，则其子集{A,B},{B,C},{A,C}也一定是2项的频繁集。为方便，可以把含有k项的集合称之为 k-itemsets.

下面以迭代的方式找出频繁集。首先找出1-itemsets的频繁集，然后使用这个1-itemsets，进行组合，找出2-itemsets的频繁集。如此下去，直到不再满足最小支持度或置信度的条件为止。这其中重要的两步骤分别是连接(join)和剪枝(prune)。即从(k-1)-itemsets中的项进行组合，产生备选集(Candidate itemsets)。再从备选集中，将不符合最小支持度或置信度的项删去。例如

www.cndm.com

Frequent 2-itemsets		Candidate 3-itemsets		Frqquent 3-itemsets
I1,I2	==>	I1,I2,I4	==>	I1,I2,I4
I1,I4		I2,I3,I4		
I2,I3				
I2,I4				

下面我们再来看一个详细的例子。

设最小支持度为2，以Ck表示k-itemsets备选集,以Lk表示k-itemsets频繁集。

ID	Items		Itemset	Sup. count		Itemset		Itemset
100	I1,I2,I5		I1	6		I1		I1,I2
200	I2,I4	==>C_1:	I2	7	==>L_1:	I2	==>C_2	I1,I3
300	I2,I3		I3	6		I3		I1,I4
400	I1,I2,I4		I4	2		I4		I1,I5
500	I1,I3		I5	2		I5		I2,I3
600	I2,I3							I2,I4
700	I1,I3							I2,I5
800	I1,I2,I3,I5							I3,I4
900	I1,I2,I3							I3,I5
								I4,I5

对 C_2 进行扫描，计算支持度。

Itemset	Sup. count		Itemset		Itemset	Sup. count		Itemset
I1,I2	4	==> L_2:	I1,I2	==> C_3	I1,I2,I3	2	==> L_3:	I1,I2,I3
I1,I3	4		I1,I3		I1,I2,I5	2		I1,I2,I5
I1,I4	1		I1,I5					
I1,I5	2		I2,I3					
I2,I3	4		I2,I4					
I2,I4	2		I2,I5					
I2,I5	2							
I3,I4	0							
I3,I5	1							
I4,I5	0							

对于频繁集中的每一项k-itemset，可以产生非空子集，对每一个子集，可以得到满足最小置信度的规则了。例如考虑{I1,I2,I5}。其子集有 {I1,I2}, {I1,I5}, {I2,I5}, {I1}, {I2}, {I5}。可以产生规则，{I1,I2} => {I5} (50%), {I1,I5} => {I2} (100%), {I2,I5} =>{I1} (100%),{I1} => {I2,I5} (33%), {I2} =>{I1,I5} (29%), {I5} =>{I1,I2} (100%)。

也不是每个数据集都有产生强规则。例如"Thinkpad notebook" 和"Canon printer"一起可能很难产生有效规则。因为恰好一起买这两个牌子的产品的顾客太少。但不妨将Thinkpad notebook放到Notebook这一层次上考虑，而Canon printer放到printer这一去层次上考虑。这样的话，一起买notebook和printer的顾客就较多了。也即Multilevel association rules。

挖掘建模

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 Apriori 关联规则挖掘算法进行模型构建，表 2-55 为 Apriori 关联规则算法建模参数说明。

表 2-55 Apriori 关联规则建模参数

序号	参数名称	参数描述
1	列索引	类别属性的选择，-1 代表最后一个属性为类别属性

www.tipdm.com

2	增量	支持度的变化量
3	最小置信度	设定最低的置信度值
4	规则条数	关联规则的条数
5	显著性水平	估计错误的概率
6	最小支持度下界	最小支持度的范围下界
7	最小支持度上界	最小支持度的范围上界

2.3.2. 案例八 电子商务网站中的商品推荐

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

当今的商业竞争日趋激烈，获得一个新客户的成本越来越高，保持原有顾客也就显得越来越重要。营销实践表明：争取一个新客户的花费常常可以达到留住一个老客户花费的 5 到 10 倍。客户忠诚是客户在较长的一段时间内，对于企业产品或服务保持的选择偏好与重复性购买。忠诚的客户不仅会增加购买量，而且往往会为企业介绍新客户。与传统的商务相比较，电子商务的客户忠诚度更重要。

影响客户忠诚度的因素非常多，有客户自身方面的原因，企业方面的原因，还有客户和企业以外的其他因素如社会文化、国家政策等。但除了企业自身外，其他都属于不可控因素。从这点出发，企业可以从自身寻找一下影响客户忠诚度的原因。比如某个客户的忠诚度下降是因为他常买的某类商品的质量出现问题或价格过高，导致该客户转向了企业的竞争对手。对于这种情况，企业需要一种方法来对客户信息和营销数据的分析，找出哪些原因导致了客户的忠诚度下降，并且针对这些原因采取措施，挽回那些即将变为不忠诚的客户．数据挖掘技术可以建立客户忠诚度分析模型，了解哪些因素对客户的忠诚度有较大影响，从而采取相应措施。因此基于数据挖掘技术的客户忠诚度分析具有重要的应用价值。

电子商务网站实现了一个网上超市，用户可以通过网站进行在线购物，实现电子商务方便快捷的优势。网站的整个操作流程如图所示。

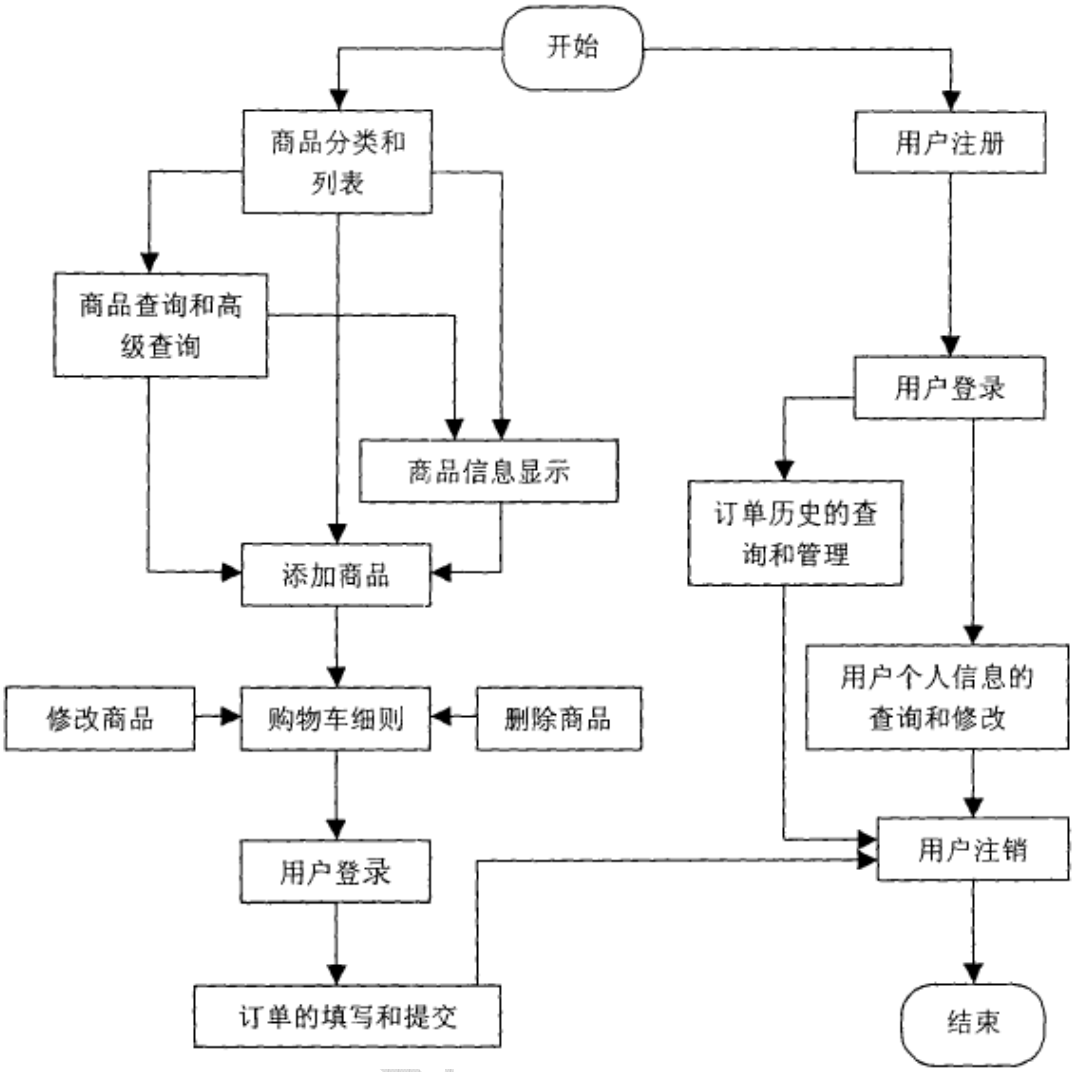


图 2-39 电子商务网站操作流程

分析方法与过程

在电子商务系统中，忠诚度分析所需要的客户信息和交易信息分别存放在网站数据库的客户表，订单表及订单明细表中。因此必须去除这些表中不需要的信息（例如用户电话、传真、身份证号码、联系方式之类的信息），抽取需要的信息。抽取信息时应注重抽取能够反映客户个人身份背景、学历等方面的信息以及反映其交易心理的相关信息。并将抽取出来的数据整理成为能被挖掘算法所利用的表格，放入数据仓库中。在计算客户忠诚度的时候，将客户的忠诚度分为四个等级：0，忠诚；1，由忠诚变为不忠诚：2，由不忠诚变为忠诚：3，不忠诚。如果客户本月的消费额比本月之前半年内每月平均消费额减少达到 50%以上，则忠诚度等级降低一级；如果客户本月的消费额比本月之前半年内平均消费额增加达到 20%以上，则忠诚度等级升高一级。最后生成的新表格如表 2-56。

表 2-56 经抽取而成的客户信息表 Loyalty

客户编号	性别	年龄	教育程度	…	距最近一次购买时间（天）	月均购买频率	已消费金额	忠诚度级别
20120001	男	40	大专	…	5	3.4	801.6	0

20120002	女	28	本科	…	11	1.9	246.3	1
…	…	…	…	…	…	…	…	…

所得到的用户数据很难做到完整全面，用户在注册时可能选择不填注册信息的几项，造成数据项空缺。对于空缺的数据项，要视情况排除或填入缺省值，例如对于数值型数据来说可以取平均值作为缺省值。抽取得到的表中数据的类型和挖掘算法需要的类型不一定一致。此时还需要做一些转换工作。例如 ID3 分类算法需要离散的源数据；C4.5 等算法虽可由程序自动寻找离散化方法，但是有时数据自动分段的边界显得不够自然，不符合人们的一般习惯。这里由分析人员按照一般的统计划分经验来对属性值进行分段，实现离散化。下表为离散化变换后的结果如表 2-57。

表 2-57 经离散变换后的客户信息表 Loyalty

客户编号	性别	年龄	教育程度	…	距最近一次购买时间	月均购买频率	已消费金额	忠诚度级别
20120001	男	30-40	大专	…	0-10	2-4	800-1000	0
20120002	女	20-30	本科	…	10-20	0-2	0-500	1
…	…	…	…	…	…	…	…	…

本案例采用基于信息论的 ID3 决策树分类算法进行客户忠诚度分析。该算法根据离散属性集的集合来做出一系列判断将数据分类。它的输入数据是已分好类的样本数据，输出一棵代表分类规则的二叉树或多叉树。

客户群细分是根据公共属性将客户划分成为同类群体的过程，细分的目的是按照客户之间的密切关系或相似程度将客户划分到事先已经定义好的各个客户群中，为营销人员与客户之间的交流提供了一个有效的平台，从而使得公司可以更好地识别不同的客户群体，区别对待不同客户，采取不同的客户战略，达到最优化配置客户资源的目的。

在客户群细分的基础上，通过建立客户行为模型，可以作为营销人员进行一对一营销的依据。一对一的营销思想，要求企业能够了解每个客户的爱好、需求，针对客户的个人特点进行营销，和客户建立起长久稳定的关系。长久以来，这一策略只能依靠营销人员与用户个人保持联系而完成。辅助统计分析工具只能了解客户群体宏观层次上表现出来的一些特性。现在，基于数据挖掘工具，可以把客户划分成更加细小的、其消费行为存在较大相似性的微小群体。虽然，还不可能细化到表现每个人的全部个性的程度；而且由于客户群体的宠大，每个细分群包括的客户数目事实上也相当可观，离一对一营销还很远。但是这样的细分，能够表现某些群体的消费行为共性，对企业制定营销策略已经具有很大的指导意义。

客户群细分变量可以采用一般人口统计学变量（如年龄、性别、收入、教育背景和职业等），也可以采用客户的购买行为特征变量（如客户购买量、购买的产品类型结构和购买频率等）。在本子系统中，我们采用了后者。通过分析客户的购买行为，我们使用数据挖掘技术将具有相似消费特征的顾客归为同一类。当某顾客在购买商品时，网站可以利用挖掘结果向该客户推荐他所在客户群的其他客户购买的商品。例如：客户甲经常光顾网站购买录音带，而客户乙经常光顾网站购买 CD ，可见甲和乙对音像制品都很感兴趣，都是音乐爱好者。那么通过客户群细分，可以认为甲和乙是同一类客户。当甲再次进入网站购物时，我们可以向他推荐购买 CO ，从而为客户提供个性化服务。客户群细分可以使用分类或聚类来实现。区别如前所述，分类需要已经由营销人员分好类的样本，聚类则自主的对客户群体进行分类。决策树等分类算法易于理解，但受样本划分准确度的影响；聚类算法有时也可以发现营销人员没有发现的一些事实。所以，在本案例中我们使用聚类算法进行客户群的细分。系统客户群细分所需要的客户信息和交易信息与客户忠诚度分析大致相仿，分别存放在

www.tipdm.com

客户表、商品类别表、订单表以及订单明细表等多个表内。数据预处理过程主要将这些表内反映客户身份背景、购买兴趣度等相关信息提取出来，并加以清理，除去噪声数据，对信息不完全的数据填入缺省值或舍去，进行必要的离散化变换。购买兴趣度信息是根据客户对各个商品的购买情况统计得出的，记录了客户对系统提供的 49 个商品类别的购买量。最终形成的表包含的属性如表 2-58 所示。

表 2-58 客户兴趣度表 interest_matrix

客户编号	性别	年龄	教育程度	类别 1 购买量	类别 2 购买量	…	类别 49 购买量
20120001	男	30-40	大专	0	17	…	61
20120002	女	20-30	本科	23	1	…	0
…	…	…	…	…	…	…	…

然后对表使用聚类算法进行挖掘。聚类算法分为基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法、基于模型的方法等几大类，本案例中选择基于划分的 K-Means 算法。

客户群细分主要是为下面将介绍的商品推荐做准备的，它的结果将被写入数据仓库的 user cluster 和 cluster info 表中。user cluster 表记录客户属于哪个类，共有 2 个字段，分别为客户编号和类编号。cluster info 表记录每一个客户类别中所有顾客的商品购买统计信息，共有 3 个字段，分别为类编号，商品编号和购买量。

商品推荐是电子商务网站用来向访问网站的顾客提供商品信息和建议，并模拟销售人员帮助顾客完成购买过程。它是利用数据挖掘技术在电子商务网站中来帮助顾客访问有兴趣的产品信息。推荐可以是根据其他客户的信息或是此客户的信息，参照该顾客以往的购买行为预测未来的购买行为，帮助用户从庞大的商品目录中挑选真正适合自己需要的商品。推荐技术在帮助了客户的同时也提高了顾客对网站的满意度，换来对商务网站的进一步支持。

商品推荐的主要任务是回答这样一个问题：当前访问网站的这位客户最可能想要的是哪些商品？对于推荐任务的实现，首先要求结果的准确性，总是向客户推荐其不想要的商品只会导致客户不满而转向其他网站；其次推荐的商品应尽可能多地覆盖用户实际喜欢的范围，以最大限度地提高推荐效果。另外，与客户的实时交互也对算法的效率提出了较高的要求。

有许多方法可以实现推荐任务。最简单的就是以编辑推荐或专家推荐的形式，比如定期推出的专题，汇集一系列围绕某主题的商品目录，这些目录都是由编辑手工编写的。一些简单的统计数据也可以作为推荐的手段，如销售排行榜，放在网页的醒目位置，对于新的来访者相当有效。另一些推荐方式则较为复杂，大部分工作需要计算机来完成。通常前者被称为“人工式推荐系统”，而后者称为“自动式推荐系统”。要真正地实现针对每一个客户的个性化服务必须借助于自动式推荐系统，它可以充分考虑每位客户的特点，在与用户的实时交互过程中动态地产生推荐结果。但两者并不相互排斥，实际的系统经常会综合多种推荐方法，互补长短。

我们为“易购 365 ”设计了结合多种方法的商品推荐方案：首先，利用统计方法在网站的首页醒目位置列出销售量处于前 10 名的热销商品，为访问者和新注册的用户提供最普通的推荐服务；其次，对于已注册并有购买纪录的顾客，当他登录网站的时候，将享受到级别更高的推荐服务。该推荐分为两部分：利用客户群细分的结果，将同一个类中其他用户购买最多的 N 个商品或与这些商品同类的新商品推荐给顾客；利用数据挖掘中的关联规则技术，列出目标客户最感兴趣的 N 个商品的推荐列表。相比以往的商品推荐，这样的方案既弥补了系统无法为新客户提供有效推荐服务的缺点，同时也弥补了未获得足够销售量的新商品不易被推荐出去的缺陷，有效地提高了对客户的推荐精度。

比较常用的与推荐相关的数据挖掘技术有关联规则、贝叶斯网络技术、聚类技术和最邻近技术等。本系统采用的是关联规则中效率较高的 FP-Growth 算法来得到满足最小支持度和置信度要求的关联规则。由 Han Jiawei 等人提出的 FP-Growth 算法，没有采用 Apriori 算法的框架，而是采取了分而治之的策略：在经过了第一次扫描之后，把数据库中的频繁集压缩进一棵频繁模式树（FP-tree），同时依然保留其中的关联信息。随后再将 FP-tree 分化成一些条件库，每个库和一个长度为 1 的频繁集相关。然后再对这些条件库分别进行挖掘。当原始数据库很大的时候，也可以结合划分的方法，使得一个 FP-tree 可以放入主存中。

在网站的商品推荐中，关联规则部分所需要的客户交易数据分别存放在网站数据库的订单表和订单明细表中。表中我们关心的只有订单编号、商品编号等少数几个属性。我们根据订单号到订单明细表中去寻找一次交易购买商品的编号，对于空缺的值将其排除。

仿真过程

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 ID3 决策树、k-Means 聚类分析等算法进行模型构建。

表 2-59 ID3 决策树建模参数

序号	参数名称	参数描述
1	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果
2		

表 2-60 k-Means 均值聚类算法建模参数

序号	参数名称	参数描述
1	显示标准误差	是否显示出标准误差
2	距离函数	计算样本距离的公式
3	缺失值不处理	如果缺失值处理的话，就用平均值代替，或则不处理
4	最大迭代次数	最大循环迭代的次数
5	聚类数	设定的聚类个数
6	保留实例序列	保留样本原来的顺序或序列
7	种子数	初始随机选取的样本数。

2.3.3. 案例九 网络入侵智能检测

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

随着计算机网络的不断普及以及相关技术的不断发展，越来越多的企业、政府和个人在开放的互联网上开展业务、获取信息。但是互联网在给公众带来巨大方便的同时，其安全问题已成为其发展的一个瓶颈问题。虽然近年来网络安全越来越得到关注，但网络入侵的种类也越来越多，技巧也越来越复杂，始终在明处抵挡外来攻击的防火墙难以抵挡黑客针对防火墙的防不胜防的手段翻新，以及很多来源于网络内部攻击，使得原有的防火墙难以单独保障网络的安全，这构成了极大的安全威胁。

目前，计算机网络的安全已成为一个国际化的问题，据统计，几乎每 20 秒全球就有一起黑客事件发生，仅美国每年由于黑客攻击所造成的经济损失就超过 100 亿美元。在过去 5 年中信息攻击以 250%速度增长，大多数的大公司都发生入侵事件。世界著名的商业网站，如 Yahoo， Buy， EBay，Amazon， CNN 都曾遭受到黑客的入侵，造成了巨大的经济损失。更有甚者，有的专门从事网络安全防护的 RSA 网站也曾遭受到黑客的攻击[18]。

在我国，计算机网络的安全状况也不容乐观，据经济专刊《华尔街日报》2002 年 3 月 18 日的报道，中国已成为世界上黑客活动最多的国家，仅次于美国和韩国。在 2006 年我国全国信息网络安全状况调查分析报告显示：2005 年 5 月至 2006 年 5 月，在 13824 家被调查单位中，有 85%的信息网络已接入互联网，其中有 54%的被调查单位遭受过网络攻击，比去年上升了 5%；其中有 22%的发生过 3 次以上的攻击，比去年上升了 7%。

然而在近年来，网络攻击产生了新的发展趋势，主要表现在以下几个方面：

(1) 攻击速度和自动化程度不断提高；

(2) 攻击工具越来越复杂；

(3) 从出现安全漏洞到被利用形成攻击时间越来越短；

(4) 网络攻击将攻击手段和病毒体结合在一起；

(5) 攻击不断更新，导致原有的防火墙防御能力不断下降；

(6) 攻击对基础设施将产生更大的威胁；

(7) 攻击的动机已由过去的引起注意或名气转变为追求利益；

(8) 攻击目的为窃取用户机密数据的攻击将具有更大的破坏力；

(9) 连接到 Internet 上其他系统的安全状态将决定可能遭受到攻击的可能。

出现的攻击这些新特点将使借助于互联网运行业务的机构面临前所未有的考验。为了保护网络免于攻击，入侵检测系统(Intrusion Detection System)在信息安全领域将发挥出不可替代的作用。

分析方法与过程

从技术层面来看，传统的安全技术，如加密技术、防火墙、口令认证、安全审计等，都是属于静态的防御技术。静态防御中过于严格的安全策略牺牲不利于用户的方便性，与当前网络的开放性、共享性不相容，很难做到一个好的利弊平衡。而且，现有的各种安全防御机制都有无法克服的局限，例如加密技术、防火墙技术的安全系数设置的非常高，也很难防止密码失窃和来自内部人员的攻击；一个安全的系统也很难保证内部人员的误操作，以及由于设置错误而导致系统漏洞的产生。因此，网络安全系统不能只依靠单一的安全防御机制和防御技术。在对充分研究网络安全防御的体系和各种网络安全技术和工具的基础上，制定详细的系统安全策略，通过设立多层的安全防线、集成各种可靠的安全机制以建立完善的多层安全防御体系[2][3]才能够有效地抵抗来自系统内部、外部的入侵攻击，从而达到有效地维护网络系统安全的目的。

入侵检测技术是信息网络安全的重要保障。然而，入侵检测技术经过了二十多年的发展，到目前为止，其应用现状却并不乐观。网络技术在时间和空间上的延伸，使得目前的网络环境变的越来

越复杂，网络基础设施不断升级所带来的爆炸式的网络数据流量，层出不穷与日趋多样化的攻击手段，使得传统的仅靠单一的基于模式匹配方法的入侵检测技术难以适用于目前的网络状况。与其他网络安全产品不同的是，入侵检测系统需要智能化，它必须能分析得到的数据，并给出有用的结果。一个合格的入侵检测系统能大大地简化管理员的工作，保证网络系统安全的运行。

基于多层次聚类分析的入侵检测过程可以描述为：首先搜集来自网络上的数据包，再利用抓包工具（如 tcpdump 它能够监听和接收网络中所有正在传输的数据包）将这些来自网络上的数据包中的数据记录到文件中。为了对这些数据进行聚类分析，需要将原始的网络数据包恢复成 TCP/IP 层的连接记录，其中每个连接记录代表一次 TCP/IP 层的连接，并包含一个网络连接的多个属性，包括网络协议、连接起始时间、连接结束时间、服务（端口号）、源 IP 地址、目的 IP 地址、连接终止状态、TCP 标志等。

经过数据标准化、聚类分析、标记实例数据、实时入侵检测等过程来对数据进行聚类分析。

(1) 对数据进行标准化处理

由于测量值属性值之间的差别可能很大，如果使用不同度量方法或者采用不同的单位来度量，就会聚类准确率受到影响。为了达到理想的聚类效果，使数据能够在统一的标准下聚类，在聚类前需要对实例数据的测量值进行标准化处理。对于同类型数据，应该将数据集转换成一个标准的单位，再进行聚类分析，然后采用基于距离度量来计算两两之间的距离。

(2) 对数据进行聚类

搜集来自网络上的数据包，再利用抓包工具（如 tcpdump 它能够监听和接收网络中所有正在传输的数据包）将这些来自网络上的数据包中的数据记录到文件中。对其进行标准化处理成为能够进行聚类分析的初始数据集，然后利用改进的 K-means 聚类算法对这些数据实例进行聚类分析及分类。

(3) 对类进行划分

接下来需要对聚类结果进行分析，由于在一般情况下正常记录远远大于异常记录，即正常行为的类及其子类数据的数量，远远大于各种体现攻击行为的类及其子类的数量。通过对聚类的数据进行分析，从而区分出那些是正常行为那些是异常行为。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的多层次聚类算法进行模型构建。

本例采用《系统安全与入侵检测》中的实验数据作为原始数据，如样本表所示。表中列出了 20 条网络连接记录的特征数据，显示了用户登陆的一些特征。主要针对三维数组（Count，Srv_serror，Srv_count）对数据进行分析。这个三维数组较好的描述了数据是否具有攻击特征。主要对这三个参数考察来区别攻击、异常、正常行为。

- 攻击：在一个时间窗口内，目标主机与当前连接相同的次数大于等于 15；同一主机的连接中出现 SYN 错误的百分比大于等于 88%，且目标端口与当前连接相同次数大于等于 25。
- 异常：在一个时间窗口内，目标主机与当前连接相同次数大于等于 6；同一主机连接中出现 SYN 错误的百分比大于 75%，且目标端口与当前连接相同次数大于等于 6。
- 正常：如果不满足上述两个条件。

样本数据属性如表 2-61，多层次聚类算法建模参数说明见表 2-62。

表 2-61 预测模型输入变量

序号	变量名称	变量描述

1	Count	在一个时间窗口内目标主机是与当前连接相同的连接次数
2	Serror	出现 SYN 错误的连接百分比
3	Same_srv	目标端口(service)相同的连接所占的百分比
4	Diff_srv	目标端口(service)不同的连接所占的百分比
5	Srv_count	目标端口(service)与当前连接相同的连接次数
6	Srv_serror	出现 SYN 错误的连接百分比
7	Rv_diff_host	目标主机不同的连接所占的百分比

表 2-62 多层次聚类算法建模参数

序号	参数名称	参数描述
1	聚类数	设定的聚类个数
2	距离计算方法	如欧氏距离、布洛克距离、马氏距离等
3	聚类树的算法	如最短距离法、最长距离法、加权平均法等

✧ **结果分析**

选择 Count，Srv_serror，Srv_count 属性，并导入 20 条网络连接记录的特征数据进行聚类分析。

聚类分析结果表明：当聚类数设置为 3，距离计算方法选择欧氏距离，聚类树算法选择最短距离法时，系统将 20 条记录分为以下三类，聚类结果参数及聚类结果分别见表 2-63 和表 2-64。

表 2-63 聚类结构参数

序号	重心	中位数	最小方差
1	15	1	11
2	10	7	2
3	4	6	12
4	22	24	26
5	9	25	5
6	28	35	34
7	3	16	14
8	18	17	8
9	21	20	13
10	27	23	30
11	31	32	33
12	19	29	36
13	37	38	0
14	0	0	0
15	0	0	0
16	0	0	1
17	1	1	1
18	1	1	1
19	1	4	19

www.ipm.com

表 2-64 聚类数为 3 时的聚类结果

聚类号	类别说明	样本数及百分比	样本点
0	正常	12(60%)	1，2，7，8，9，10，11，14，15，16，17，18
1	异常	7(35%)	4，5，6，12，13，19，20
3	攻击	1(5%)	3

多层次聚类结构如图 2-40。

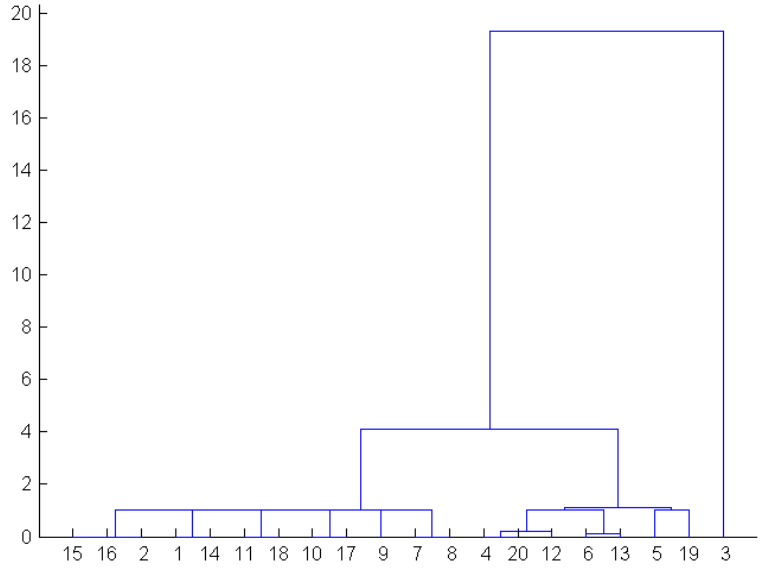


图 2-40 多层次聚类结构图

当聚类数设置为 4 时，系统将 20 个样本点分为以下四类，结果见表 2-65。

表 2-65 聚类数为 4 时的聚类结果

聚类号	类别说明	样本数及百分比	样本点
0	正常	2(10%)	1，2，7，8，9，10，11，14，15，16，17，18
1	异常	5(25%)	4，6，12，13，20
2	异常 (为需重点监视的异常行为)	12(60%)	5，19
3	攻击	1(5%)	3

结果分析：当聚类数由 3 调整为 4 时，聚类结果将异常样本点进一步细分为两类，其中一类为需重点监视的异常行为。

2.3.4. 案例十 基于 RFM 的客户细分

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn

- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

客户决定着企业存在的价值，能否满足客户需求是公司商业运作成功与否的关键所在。在复杂多变的市场情形中，只有那些了解客户，把握市场变化，能够及时吸收先进管理理论，利用先进的信息技术和数量分析方法，对市场信息进行科学分析和科学决策的企业才可能成为市场的赢家。因此，愈来愈多的企业将关注的重点从以产品为中心的商业模式向以客户为中心的新型商业模式转移。随着信息技术的迅猛发展，人们积累的数据量也急剧增长，动辄以 TB 级计，因此，如何从海量的数据中发掘、提取出有用的规则与知识，也就成为企业的当务之急。

C-company 位于中国上海，其主要产品是各类型电流稳压器，是国家稳压器定点生产厂家之一，具备全系列交流稳压器的生产经验。同众多企业一样，迫切需解决的一个问题也是其业务过程中累积的大量数据需分析处理。C-company 在数据分析处理上大致经历了如下三个阶段，第一个阶段是统计报表阶段，业务人员利用一些常用的计算机工具或统计软件，分类统计出企业所感兴趣的静态数据，从而为企业领导决策提供参考依据；第二个阶段是 OLTP 或 OLAP 分析阶段，它比统计报表更进了一步，会做一些多维分析及原因分析，如分析调查今年以来，哪些产品是最有利润的？最有利润产品是不是和去年一样？同时，一些简单的预测功能也能,如销售量的预测等等；第三个阶段则是数据挖掘技术的应用，公司利用一些数据挖掘工具对历史经营活动中累积的业务数据等建立数学模型，如分类、聚类模型、预测模型等。

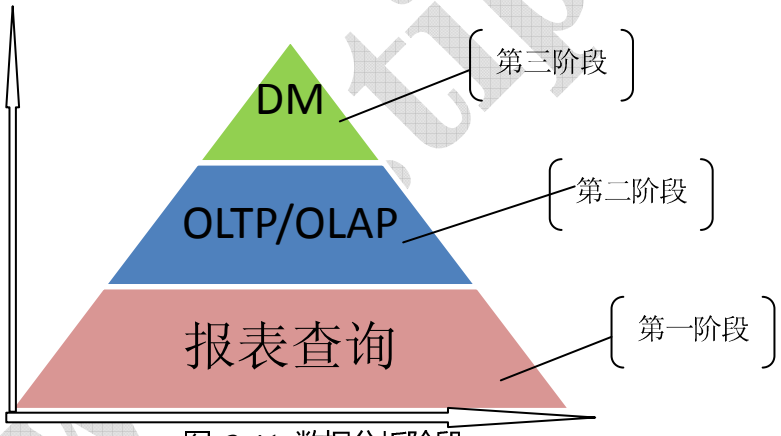


图 2-41 数据分析阶段

在世界经济一体化进程的大背景下，C-company 在努力开拓国内市场的同时，在国际上也创造了自己的品牌地位。虽自从 2006 年以来，C-company 即已开始应用数据挖掘技术来整理分析其业务数据，但由于世界经济一体化进程的加速，特别是加入 WTO 以来，市场竞争的日趋激烈，原有的 IT 系统单一的统计分析技术、数据挖掘工具来已不能满足公司的业务需要。C-company 逐渐意识到建立一个强大有效的决策支持系统，增强数据仓库（DW）特别是数据挖掘（DM）能力，已是公司迫切需要解决的问题。

基于之前系统存在的问题，本案例结合 RFM 理论，运用组合数据挖掘的方法，先对客户进行聚类分析，从而有效地实现对客户的分类，再在此基础上提取出规则。实践证明，结合 RFM 模型，应用组合数据挖掘方法，能有效地挖掘出客户有价值的信息，识别客户的最大价值。基于 RFM 模型的

组合数据挖掘技术进行客户细分及规则挖掘是有效的。

尽管对客户状况的理解有所不同，但提供产品或服务给客户以满足他们的需求对于每个公司是相同的。因此，运用有效的数据挖掘技术，对其客户进行理解和分析，无疑对于每个企业都有重要的实际意义和指导意义。

分析过程与方法

结合 RFM 思想，我们按以下步骤进行相应客户分析工作：（1）选择数据集，并进行预处理；（2）将 RFM 模型中的近度(R)、频度(F)、值度(M)作为输入变量，进行聚类分析，输出为客户价值，根据具体问题不同可以定义输出为 3、5 或 7 类等；（3）运用数据挖掘算法，提取每个类别的规则特征；（4）结果描述，并对比列出不同类型的客户规则。

详细的分析流程如图 2-42 分析模式。

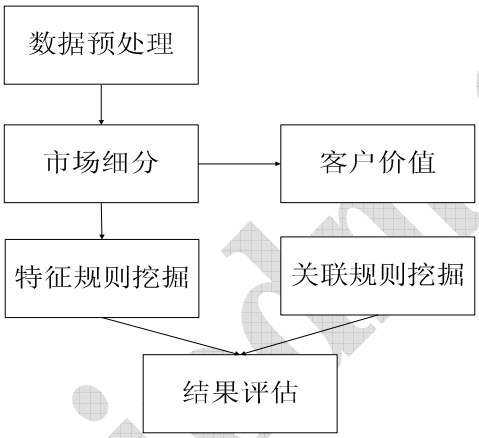


图 2-42 分析模式

根据美国数据库营销研究所 Arthur Hughes 的研究，客户数据库中有三个神奇的要素，这三个要素构成了数据分析最好的指标，它们分别是：最近购买时间 R(Re-cency)、购买频率 F(Frequency)和总购买金额 M(Monetary Value)，各指标具体含义如下表 2-66 RFM MODEL 所示：

表 2-66 RFM MODEL

	R（近度）	F（频度）	M（值度）
RFM 模型	客户最近一次购买距离分析点的时间	客户一定时期内购企业产品的次数	客户一定时期内购买该产品的总金额

首先，近度是衡量顾客最近一次购买距离现今时点的日子。一般而言，最近购买日越小，即表示顾客最近一次购买完产品的时间距现今不久，所以被认为是较有可能再度购买的顾客，这群顾客的得分比最近购买日长的顾客高。频度指的是一个顾客在一段时期内购买产品的次数。频度愈高，则顾客对产品的忠诚度相对较高。值度指的是在一段时间内，顾客购买产品金额的总和。一般而言，金额总和愈高的顾客，其相对于企业的重要性也愈高；反之，总和价值愈低的顾客得分就相对不高。因此，通过对以上三个指标分析加总，则可得到各个顾客对企业的价值。

详细的数据挖掘步骤如下所示。

www.tidm.com

1．数据预处理

首先选择数据集，将有遗漏或不准确信息的记录删除，去除冗余属性并将数据转换成聚类方法容易并有效处理的格式。

对近度数据值进行预处理，衡量数据库记录中整个的时间日期记录范围，对其均分为五等份，如从 2008 年到 2012 年的划分，可取 2011 年到 2012 年为 5,2010 年到 2011 年为 4，其余以此类推。则可将数据库中日期近度值均匀地划分为 5 个等分，将每个顾客的购买日期分别归入不同等分里面（如下表 2-67 日期记录编号）。

表 2-67 日期记录编号

分值	1	2	3	4	5
时间	08 年以下	08 年-09 年	09 年-10 年	10 年-11 年	11 年-12 年
客户					

然后，同样按照上面的方法分别对频度（购买频率）、值度（消费金额）分别进行排序，如表 2-68 和表 2-69。

表 2-68 频率记录编号

	1	2	3	4	5
频率	1 … N				
客户					

表 2-69 金额记录编号

	1	2	3	4
消费金额	1 … N			
客户				

完成上面步骤后，即完成了对 RFM 数据的预处理。数据库中每个客户都有一个对应的简单 3 位数来描述，如 221，222，223 等。这样，每个客户就有共 5*5*5=125 种 RFM 状态值。

2．市场细分

美国著名的市场学家温德尔·史密斯（Wendell R·Smith）在其 1958 年发表的《市场营销策略中的产品差异化与客户细分》一文中指出：市场细分是指企业按照客户属性将客户群体分为若干个子客户群体，细分后客户群体之间的差异最大化，每个客户群体尽可能的相似。正确的客户细分能够有效地降低成本，保证投入的资源得到回报，同时获得更强、更有利可图的市场渗透。管理实践表明，企业 80%的利润往往来自于其 20%的客户。因此，区分这两类客户，保持和发展这 20%的最大利润客户，从而实现永续经营，对企业来说极为关键。

通过对客户数据的聚类分析，可以挖掘出不同类别客户所隐藏的一些信息。这样企业就可以针对不同客户采取不同的营销策略，重点开发出对企业贡献度大的客户类别，同时吸引某些行为类似

的贡献度较低类的客户发展为较高贡献的客户。

基于 RFM 的聚类分析，可以按以下步骤进行：

第一步：先确定一种聚类算法用于待聚类的客户数据集聚类分析，将数据集 D 划分为若干个类 $C=\{C_1,C_2,...,C_p\}$。其中，$D=C_1\cup C_2\cup...\cup_p$；

第二步：选取：近度 R(Re-cency)、频度 F(Frequency)和值度 M(Monetary Value)，三个指标作为输入变量，应用聚类算法进行聚类分析。

第三步：对每一个 $C_i,i=1,2,...,p$，根据其中成员的特点，归纳出能描述其特征的一条或几条规则。评判规则，若聚类结果具有较高的可信度，则将聚类结果应用于实际中，否则，选择其它的聚类算法再进行分析。

这样，就实现了对客户的细分，从而就可以对不同类别客户采取有针对性的营销策略了。

3．规则挖掘

在进行完聚类分析之后，我们便可提取出每一类别客户的相关规则了，主要包括特征规则和关联规则。利用数据挖掘工具对数据库中的潜在规则进行深层次挖掘，这是数据挖掘最关键的步骤，也是技术难关所在。

（1）特征规则

特征规则是数据挖掘的一个重要内容，在企业的决策支持系统中，如营销领域的针对性营销、精准营销等，特征规则都有着广泛应用。

特征规则是指原始数据的一些描述规则，这些规则概括了数据集中不同概念的特征，规则通常被写成 IF-THEN 形式，是一些专家知识的汇集。其中，IF 部份是规则前件，它给出了每个规则的先决条件，由命题的逻辑组合来完成；THEN 部份是规则后件，为一组结论或判断。这样，就保证了在一定精度的前提下提取得到尽可能容易理解的符号规则。

（2）关联规则

两个或两个以上变量的取值之间存在某种规律性，就称为关联。数据关联是数据库中存在的一类重要的、可被发现的知识，关联分析的目的就是找出数据库中隐藏的关联网。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 k-Means 聚类分析和 Apriori 关联规则挖掘等算法进行模型构建。

1．数据预处理

在 C-company 业务数据表中共收集了 2003 年到 2007 年的客户交易记录，删除包含遗漏和不正确值的数据，并去掉冗余的属性。然后，将数据按照上文提到的预处理方法转变成合适的格式。最后，保存了 275 条规范的业务记录，具体包括了：最近购买时间 R(Re-cency)、购买频次 F(Frequency)和总购预买金额 M(Monetary Value)，三个关键指标。获得的原始记录如表 2-70 所示：

表 2-70 Sales Dataset

单　　位	地　　址	合同时间	型号	价格 (元)	数量(台)
上海华新丽华工程塑料有限公司	上海市嘉定区南翔镇工业开发区	2005.1.7	SBW-15	10,000.00	1
			SBW-30	12,500.00	1
			SBW-50	13,500.00	1
上海光联通讯技术有限公司	上海嘉定叶城路 925 号 A4 栋	2004.12.15	SBW-30	13,400.00	1
上海光联通讯技术有限公司			SG-200	48,000.00	1
			SBW-100	27,000.00	1
上海瑞卡电子有限公司	虹梅南路 1755 号 C 区 3 号	2006.12.2	SGB-50	12,500.00	1

www.tipdm.com

上海市党委系统网管理中心	华阳路 330 号	2007.12.3	DJA-10	6,500.00	1
上海市党委系统网管理中心			SVC-10	4,200.00	1

按上文中的提到的预处理方法，对原始数据表进行预处理以得到数据挖掘工具所合适的格式，详细步骤如下：

（1）定义 RFM 的取值范围，如划分为 5 等份，分别为 5、4、3、2、1 等。对于 R 值,如最近合同时间是 2007 年 1 月以后的定为 5，2006 年 1 月到 2006 年 12 月的定义为 4，依此类推。同样的方法（表二），对 F-Frequency、M-Monetary 也分别类似处理，详见表 2-71 RFM 取值范围。

表 2-71 RFM 取值范围

Scaling	Scaling name	R-Recency	F-Frequency	M-Monetary
5 Score	Very High	2007/1—	Over 6 records	Over 100K
4 Score	High	2006/1—2006/12	4-5 records	75K-100K
3 Score	Medium	2005/1—2006/12	3 records	50K-75K
2 Score	Low	2004/1—2005/12	2 records	25K-50K
1 Score	Very Low	—2003/12	One record	Under 25K

（2)对客户数据表中的每一条客户记录按上述②的方式量化以后，得到表 2-72 RFM 量化结果。

表 2-72 RFM 量化结果

区域	客户 ID	R-Recency	F-Frequency	M-Monetary
华中	⋮	⋮	⋮	⋮
	0205	2	1	1
	0206	1	1	2
	0207	4	1	1
	0208	4	1	4
	0209	3	1	1
	0210	3	1	1
	0211	5	3	4
	⋮	⋮	⋮	⋮

（3）分别评估 R(近度)、F(频度)、M（值度）的权重，如 1:1:1 等。对于 RFM 模型三个变量的重要程度，Hughes（1994）[10]认为 RFM 对于衡量一个顾客的重要性程度是一致的，事实上它们的权重应该是相同的。

2．K-means 聚类分析

根据预处理后的 RFM 数值，利用基于距离的初始聚类方法对华东地区的 186 个客户分 5 类得到的初始聚类中心值为：

Initial Cluster center： C1 （2.29, 2.13, 3.78） C2 （3.21, 1.38, 1.76）

C3 （1.89, 1.22, 1.08） C4 （1.84, 1.24, 1.70）

C5 （1.95, 1.27, 1.11）

然后，运用 TIPDM 中的 K-均值聚类算法，完成数据集的聚类分析，基本步骤如下：

（1）导入数据

将预处理好的数据按 TIPDM 所支持的 xls 格式存储，选中“数据管理”，点击“浏览”找到存储的数据文件，点击“上传”后将数据导入方案中。如果数据未及时显示，点击“刷新“则显示出

www.tipdm.com

导入的数据。

（2）选择聚类的列

在系统菜单栏选择“聚类分析”其下有本次应用的算法“K-均值算法”，点击该菜单，右边出现该算法页面。

（3）参数设置

在“参数设置”按钮框中设置好相应参数，这里重点考虑聚类类别数，定义聚为 5 类。

如下，运用 K-means 法，得到的最终聚类结果中心值如表 2-73 所示。

表 2-73 聚类输出

Cluster Center	C_1	C_2	C_3	C_4	C_5
C_1_R	1.72	4.67	1.54	4.59	1.44
C_2_F	1.18	2.92	1.23	1.17	3.22
C_3_M	1.31	4.75	3.54	1.58	5.00
D_i	2.46	7.27	4.05	4.99	6.12
客户价值(输出)	非常低	非常高	低	中等	高
客户数量	93	12	13	59	9

聚类结果表明，华东地区 186 个客户中，第一类包括 93 个客户数，第二类包括 12 个客户数，第三类包括 13 个客户数，第四类包括 59 个客户数，第五类包括 9 个客户数。第一类主要为低价值或旧客户，最近与公司无业务往来已 4 年以上，累计交易额在 2.5 万元以下，平均交易次数小于 2 次。第二类为高价值客户，与公司最近有业务来往，总频次 3 次以上，累计交易额一般超过 7.5 万元，是贸易能力较强的客户。第三类跟第一类客户情况类似，不过交易额较大。第四类，与公司最近有业务往来，但交易额及频次不是很大，应有较大挖掘潜力，是重点关注的对象。第五类与第二类相似，为高价值客户，不过与公司无业务往来时间比较长了，须增强促销策略并分析原因。

这样，我们就得到了每一个客户所属的聚类类别，如 ID 为 0127 的客户归于 C_2类，即为高价值客户，ID 为 0130 的客户归于 C_1类，即属低价值的客户。通过聚类方法我们找到了对企业具有重要价值的客户，根据 20/80 法则以及开发新客户的费用是保留老客户费用的 5 倍等营销原则，就可以针对不同客户采取不同的营销策略了，正如 Arthur Hughes 所说的那样，“如果你的预算不多，你会将信息邮寄给贡献 40%收入的顾客，还是那些不到 1%的顾客？数据库营销有时候就是这么简单，这样的营销所节省下来的成本是相当可观的。”，同时，通过对客户进行聚类，我们不但可以更方便地为不同的客户提供不同的营销策略，也可以为其它分析提供依据，如在聚类基础上进行规则提取等。在聚类后再进行其它的数据挖掘将更有针对性，可以得到更有意义的结果。

下一节中将详细说明。同样，对非华东地区客户聚类结果如表 2-74 所示。

表 2-74 非华东地区聚类结果

Cluster Center	C_1	C_2	C_3	C_4	C_5
C_1_R	1.32	3.44	3.97	3.62	5.00
C_2_F	1.18	3.67	1.18	1.31	5.00
C_3_M	1.89	4.67	1.29	4.00	2.00
D_i	2.59	6.86	4.34	5.55	7.35
忠诚度(输出)	非常低	高	低	中等	非常高
客户数量	28	9	38	13	1

3．特征规则提取

多种数据挖掘方法可以用来提取特征规则，在进行特征规则挖掘之前，我们可以通过对几种常用

的特征规则挖掘方法进行比较，如实验对比分析等，从中选择出相对较优的方法来提取规则。如对于数据集的一个样本类别结果，其前件如果符合提取的规则特征，则判断该样本为正确，否则为错误，统计出一个数据集中正确的样本数，从而计算出该算法在此数据集中的准确率。这样，我们可以选择相对较优的算法来进行特征规则挖掘。

本例中，选取 UCI 数据集中 balance scale , congressional voting records, hepatitis, iris plant , statlog australian credit approval 及 statlog german credit 共六个数据集作为实验数据集，分别应用 Single NN (single neural network)方法、GMDH 网络方法（Knowledgeminer），神经网络方法（Clementine），C4.5 决策树方法(weka)，进行对比分析。将每一个数据集分成 5 部份，以每一个为检测集，其它四个为学习集，五次运行后的平均结果为最终结果。

如对于 balance-scale 数据集，按顺序选取 500 个样本构成学习集，125 个样本构成检测集，运用 GMDH 网络分类得到的结果是：学习集中 47 个错误，错误率 9.4% ，在检测集中 8 个错误，错误率 6.4%, 再分别任选其它一部份作检测集，其余四部份作为学习集重复运行五次，对检测集中分类错误平均，得到检测集中平均错误率为 13.44%（5 次共 84 个错误）。

提取的 L 类特征规则是：

$$IF \quad NOT-A_1 \ \& \ C_1 \ \ OR \ \ NOT-B_1 \ \& \ D_1$$
$$THEN \ \ L \qquad (1)$$

对于规则(1)的解释为：如果左、右两边重量都不是 1，或者左、右两边距离都不为 1，则属于 L 类。

对其它数据集，不同方法同样处理，得到分类错误率如表 2-75 所示。对于连续型属性，同样，将某个连续属性的取值分成若干个区间，将连续属性离散化，再建立规则输入输出模型。

表 2-75 Comparison of generalization error

Data set	Single NN	GMDH NET	ANN	C4.5 Rule
balance	9.12%	13.44%	9.00%	17.00%
voting	4.32%	3.45%	4.45%	6.67%
hepatitis	16.25%	7.70%	22.20%	27.78%
iris	6.00%	0%	9.30%	23.26%
credit-a	16.23%	3.95%	7.00%	3.95%
credit-g	28.00%	7.80%	12.00%	14.56%
average	*13.32%*	*6.06%*	*10.66%*	*15.54%*

上面一共列出了 4 种算法对 6 个数据集的错误情况对比，可用 Friedman 统计检验方法对各算法之间是否有显著差异进行判断。设 r_i^j 是第 j 个算法在第 i 个数据集上的排序，Firedman 检验比较各种算法的平均排序 $R_j = \frac{1}{N}\sum_i r_i^j$ 。Firedman 检验的零假设是各算法表现是一样的，也即它们的平均排序相同。

Firedman 检验：

$x_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$，服从自由度为 $k-1$ 的卡方分布。

在此基础上，Iman 和 Davenport 指出 Firedman 过于保守，他们提出了另一种更好的统计检验：

$F_F = \frac{(N-1)x_F^2}{N(k-1)-x_F^2}$ 服从自由度为 $k-1$ 和 $(k-1)(N-1)$ 的 F 分布.

www.tipdm.com

引入秩后表如下所示：

表 2-76 续表 2-37

Data set	Single NN	GMDH NET	ANN	C4.5 Rule
balance	9.12%（3）	13.44%（2）	9.00%（4）	17.00%（1）
voting	4.32%（3）	3.45%（4）	4.45%（2）	6.67%（1）
hepatitis	16.25%（3）	7.70%（4）	22.20%（2）	27.78%（1）
iris	6.00%（3）	0%　（4）	9.30%（2）	23.26%（1）
credit-a	16.23%（1）	3.95%（4）	7.00%（2）	3.95%（3）
credit-g	28.00%（1）	7.80%（4）	12.00%（3）	14.56%（2）
average rank	*2.33*	*3.67*	*2.5*	*1.5*

$$x_F^2=\frac{12*6}{4(4+1)}\left[(2.33^2+3.67^2+2.5^2+1.5^2)-\frac{4*5^2}{4}\right]=8.63$$

$$F_F=\frac{(6-1)*8.63}{6(4-1)-8.63}=4.61$$

4 种算法，6 个数据集，F_F 服从自由度为 4-1=3 和(4-1)*(6-1)=15 的 F 分布。查表得当 $\alpha=0.05$ 时，$F(3,15)=2.49$，由于 $F_F=4.61>2.49$，所以可以拒绝零假设，即算法之间存在差异，由上表可以看出，GMDH 网络方法要好于其它三种方法。

因 GMDH 的平均秩 3.67 为最大，所以可用 Holm 方法对算法进行进一步的检验。Holm 检验从最显著的 p_1 值开始，如果 $p_1<\alpha/(k-1)$，拒绝相应的零假设，同时进一步比较 p_2 和 $\alpha/(k-2)$，如果第 2 个假设被拒，则继续第 3 个比较，直至被接受为止。令 $z=(R_i-R_j)/\sqrt{\frac{k(k+1)}{6N}}$，z 值用于从正态分布表中查找相应的概率值（$P(|x|\geq z)=2*(1-p(x\leq z))$），然后将概率值与相应的 α (0.05)进行比较。

表 2-77 Holm 检验

i	Classifier	*z=(R0-Ri)/SE*	*p*	α/i
3	C4.5 Rule	(3.67-1.5)/0.75=2.89	0.004	0.017
2	Single NN	(3.67-2.33)/0.75=1.80	0.072	0.025
1	ANN	(3.67-2.5)/0.75=1.57	0.116	0.05

从上表可以看到，0.004<0.017，显然， C4.5 Rule 要劣于 GMDH 方法。但由于 0.072>0.025，0.116>0.05，因此，在 95%的置信度下不能拒绝零假设，即神经网络方法劣于 GMDH 方法并不明显。

由以上实验设计可知，GMDH 网络方法具备较好的特征提取能力，由于需要预先知道样本的类型，所以在企业面对众多客户的情况下，结合聚类和 GMDH 网络进行特征规则挖掘是一种有效而实用的方法。在本例中通过 TIPDM 完成对客户进行聚类分析后，我们就知道了每个客户分别属于哪类客户（客户价值），然后在此基础上运用 GMDH 网络方法提取出一些规则特征。如针对华东市场（共 186 个客户），随机选取 160 个样本构成学习集， 26 个样本为检测集，提取规则如下：

$$IF\ \ SBW\&SG\ \ OR\ \ SG\&SGA$$
$$THEN\ \ L \qquad (1)$$

对于规则（1）在学习集中有 15 个错误，正确率为 90.63%， 在检测集中 2 个分类错误，准确率为 92.31%，规则的解释可以描述为：在华东地区，高价值客户主要体现在购买 SBW 和 SG 或 SG 和

www.tipdm.com

SGA 这三种产品品种的，即 SBW 和 SG 或 SG 和 SGA 这三种产品的客户表现为忠诚度较高。

同样，对其它市场(共 89 个客户)，随机选取 60 个样本构成学习集， 29 个样本为检测集，运用 KnowledgeMiner 提取规则如下：

$$IF \ \ SBW \& SVC$$
$$THEN \ \ L \qquad (2)$$

对于规则（2）在学习集中有 6 个错误，正确率为 90.00%， 在检测集中 3 个分类错误，准确率为 89.66%，规则的解释可以描述为：在其它地区，SBW 和 SVC 用户表现为高价值客户忠。

这样，聚类后通过提取规则。我们就从客户关系数据库中得到了一些描述规则，这些规则概括了数据集中不同概念的特征，从而使公司的营销活动更有针对性。有了这些特征规则，决策者可以作出一个正确的销售和广告宣传决策。同样，销售人员也可以缩小受众群体，尽量对最有可能的销售业务进行营销活动，减少拒绝，提高成功率。

4.Apriori 算法产生强关联规则

1. 算法过程

Apriori 算法是通过对数据库的多次扫描来发现所有的频繁项目集，在每一次扫描中只考虑具有同一长度（即项目集中所含项目的个数）的所有项目集，在第一次扫描中计算中所有单个项目的支持度，生成所有长度为 1 的频繁项目集。在后续的每一次扫描中，首先以 K-1 次扫描所生成的所有频繁项目集为基础产生新的候选项目集，然后扫描数据库，计算这些候选项目集的支持度，删除其支持度低于用户给定的最小支持的项目集。最后，生成所有长度为 K 的频繁项目集。重复过程，直至再也找不到新的频繁项目集为止。

下面选取数据表中几条客户的交易记录，对 Apriori 算法过程说明如下。其中，事务标识为客户 ID，Item 项为客户购买的产品项目。

表 2-78 产品项目表

事务标识（客户 ID）	Item（购买产品）
0001	SBW-30 SBW-50 SBW-100
0002	SBW-30 SBW-50 SG-200 SBW-100
…	…
0049	SBW-50 SBW-100
0051	SBW-100 SG-100 SG-200

由 Apriori 算法寻找频繁项集，设最小支持数为 2。首先扫描原始数据表，列出各个产品型号的客户支持数，支持数小于 2 的不参与下一频繁项集的生成，大于等于 2 的再两两组合，构成候选 2 项集。再扫描原始数据表，保留支持数大于等于最小支持数 2 的项集，以次类推，直至不能找到频繁项集为止，过程如下图所示：

事务标识 （客户 ID）	Item （购买产品）
0001	SBW-30 SBW-50 SBW-100
0002	SBW-30 SBW-50 SG-200 SBW-100
…	…
0049	SBW-50 SBW-100
0051	SBW-100 SG-100 SG-200

Itemset	Support
SBW-30	2
SBW-50	3
SBW-100	4
SG-100	1
SG-200	2

Itemset	Support
SBW-30 SBW-50	2
SBW-30 SBW-100	2
SBW-30 SG-200	1
SBW-50 SBW-100	3
SBW-50 SG-200	1
SBW-100 SG-200	2

Itemset	Support
SBW-30	2
SBW-50	3
SBW-100	4
SG-200	2

Itemset	Support
SBW-30 SBW-50	2
SBW-30 SBW-100	2
SBW-50 SBW-100	3
SBW-100 SG-200	2

Itemset	Support
SBW-30 SBW-50 SBW-100	2

图 2-43 Apriori 算法寻找频繁项集过程示意图

这样，就得到一个频繁 3 项集 $\{SBW-30, SBW-50, SBW-100\}$，它的所有非空真子集有：$\{SBW-30\}, \{SBW-50\}, \{SBW-100\}, \{SBW-30, SBW-50\}, \{SBW-30, SBW-100\}, \{SBW-50, SBW-100\}$，设最小置信度为 50%，则可以输出强关联规则如：$SBW-50 \Rightarrow SBW-30 \wedge SBW-100$ ，置信度（2/3=67%），支持度（2/4=50%）。挖掘到这些规则，业务部门就可以在营销活动中制定出合理的销售匹配方式，如由上述置信度 67%可以知道，购买 SBW50 的客户很可能（67%）会同时购买 SBW30 和 SBW100。公司就可以据此展开交叉营销活动，或将三种产品库存维持相当水平。

2. TIPDM 挖掘关联规则

打开 TIPDM 后，点击“方案管理”创建一个关联规则挖掘方案，点击“浏览”、“上传”加载所需挖掘的数据集。关联规则挖掘中涉及到大量的对数据项的处理，因此在输入数据时，应先注意对各输入属性字段的设置。TIPDM 中导入的数据表结构及内容如下图所示：（注：TIPDM 可参考附录二相关知识介绍）

www.tipdm.com

TID	ItemList
0001	SBW-30,SBW-50...
0002	SBW-30,SBW-50...
0003	SGB-50
0004	DJA-10,SVC-10
0005	SBW-250

图 2-44 事务数据表

将以上表结构的数据库表导入 TIPDM 后，在“参数设置”中设计好关联规则挖掘所需的参数后，即可开始调用 Apriori 算法。关联规则挖掘思想及步骤在前面相关知识介绍中已讨论过，即第一步找出所有频繁项集，就是找出所有支持度不小于给定的最小支持度的项集；第二步，则通过频繁项集推出关联规则。

这样，通过 TIPDM 对 Apriori 算法的实现，就可以从销售数据库中挖掘出关联规则了。如计算得到的{*SBW*−30 *SBW*−50 *SBW*−100}及{*SBW*−100 *SBW*−180 *SBW*−400}，均为支持度大于 2 的频繁 3 项集，同样可以计算出各个产品间的置信度。例：$SBW-100 \Rightarrow SBW-180 \wedge SBW-400$，因 $SBW-100$ 的支持度计数为 39，而 $SBW-180 \wedge SBW-400$ 的支持度计数为 2，即 $SBW-100 \Rightarrow SBW-180 \wedge SBW-400$ 的置信度（2/39=5.13%），购买 SBW100 的客户 5.13%的可能性会同时购买 SBW180 和 SBW400，但由 $SBW-180 \Rightarrow SBW-400$ 置信度(10/19=52.63%)，即购买了 SBW180 的客户很可能(52.63%的可能性)会同时购买 SBW400。这样，企业就可以据此采取相应的促销措施，展开交叉销售活动，从而促进更多产品的销售，更好地满足客户需求。

结论

随着世界经济一体化进程的加速，企业面临更加激烈的市场竞争，利用数据挖掘技术，对企业数据仓库中收集的大量业务数据及客户数据进行分析和处理，已成为面对全球化竞争的重要举措。结合 RFM 模型，将改进的 K-means 方法用于客户细分，在运用统计的方法对多种数据挖掘方法进行比较分析后，文章进行了客户特征规则提取及关联规则挖掘。本案例有如下结论：

结合 RFM 模型，K-means 聚类方法，对企业客户进行分类，分析挖掘出对企业具有高价值、高效用的客户，从而为客户关系管理活动中正确的客户细分提供了有益的思路。

通过对神经网络方法、C4.5 方法及 GMDH 网络方法在六个 UCI 数据集上的比较实验，确认几种算法之间是有差异的，并指出结合聚类（客户细分）与 GMDH 网络方法建立模型，从而提取出特征规则是比较理想与可行的一种方法。

综上，运用组合数据挖掘技术（将改进的 K-means 方法用于客户细分，结合聚类与 GMDH 网络提取客户特征规则及 Apriori 挖掘出产品的关联规则）进行客户细分和规则挖掘是有效的。

www.tipdm.com

2.3.5. 案例十一 信息系统客户服务感知评价

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

信息系统客户服务感知是指客户对信息系统及其运维服务的体验和感受，反映当前信息系统及其运维服务的质量与用户期望的差距。信息系统客户服务感知通过感知度来衡量，它是客户对信息系统及其运维服务质量在主观感受上的综合满意程度。其评价主体是客户，评价对象是信息系统。

通过多年信息化建设，某电信企业建成了众多业务信息系统，每个业务系统运行过程中是否正常、是否存在问题隐患？信息系统是否真正减轻业务人员的日常工作量提高工作效率？如何从提供“被动”服务转变为根据客户感知提供“主动”服务，真正实现某电信企业对信息系统服务的有效管理？如何构建一套适合企业的信息系统客户服务感知模型，通过模型准确定位信息系统客户服务过程中存在的问题，并通过建立信息系统客户服务管控体系，不断完善和优化运维服务，提高客户服务水平，提升信息系统客户服务满意度？已成为企业有效促进信息化工作水平提高的重要工作。

本案例的挖掘目标：

1） 构建客户服务感知评价指标体系

2） 建立信息系统客户服务感知评价模型

分析方法与过程

信息系统客户服务感知是指客户对信息系统的体验和感受，反映当前信息系统的质量与用户期望的差距。

信息系统客户服务感知模型的构建，与 CRISP-DM 定义的数据挖掘项目的标准化流程基本一致，主要分为 6 个步骤，如图 2-45 所示：

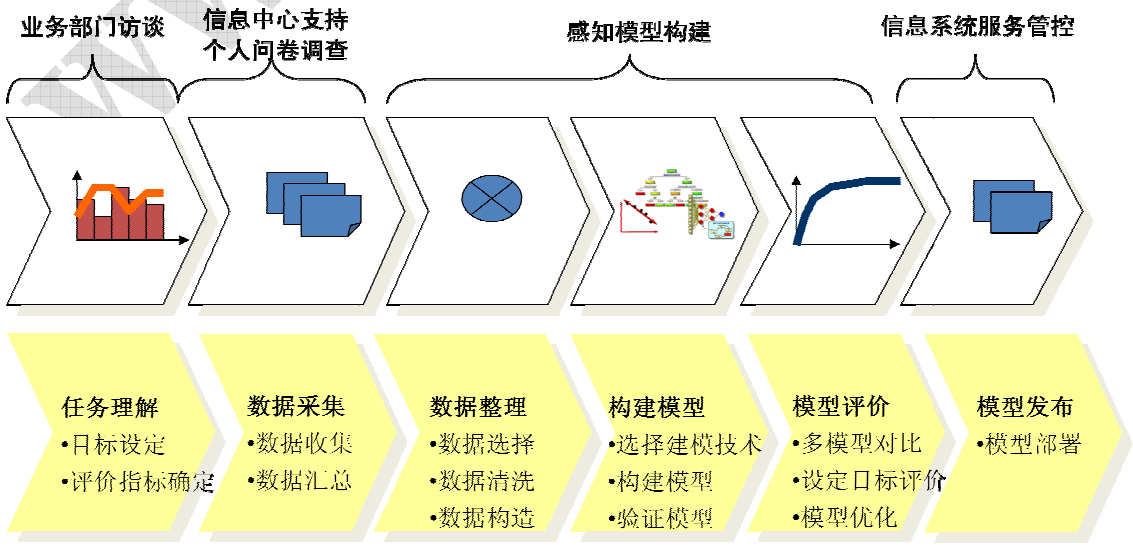


图 2-45 信息系统客户服务感知建模流程

1. 任务理解

建立信息系统客户服务感知模型，主要是通过设置若干评价指标，由用户对这些评价指标评分，以便准确定位信息系统服务过程中存在的问题，并通过服务管控，不断完善和优化运维服务，提高客户服务水平，信息系统客户服务感知及管控结构见图 2-46。

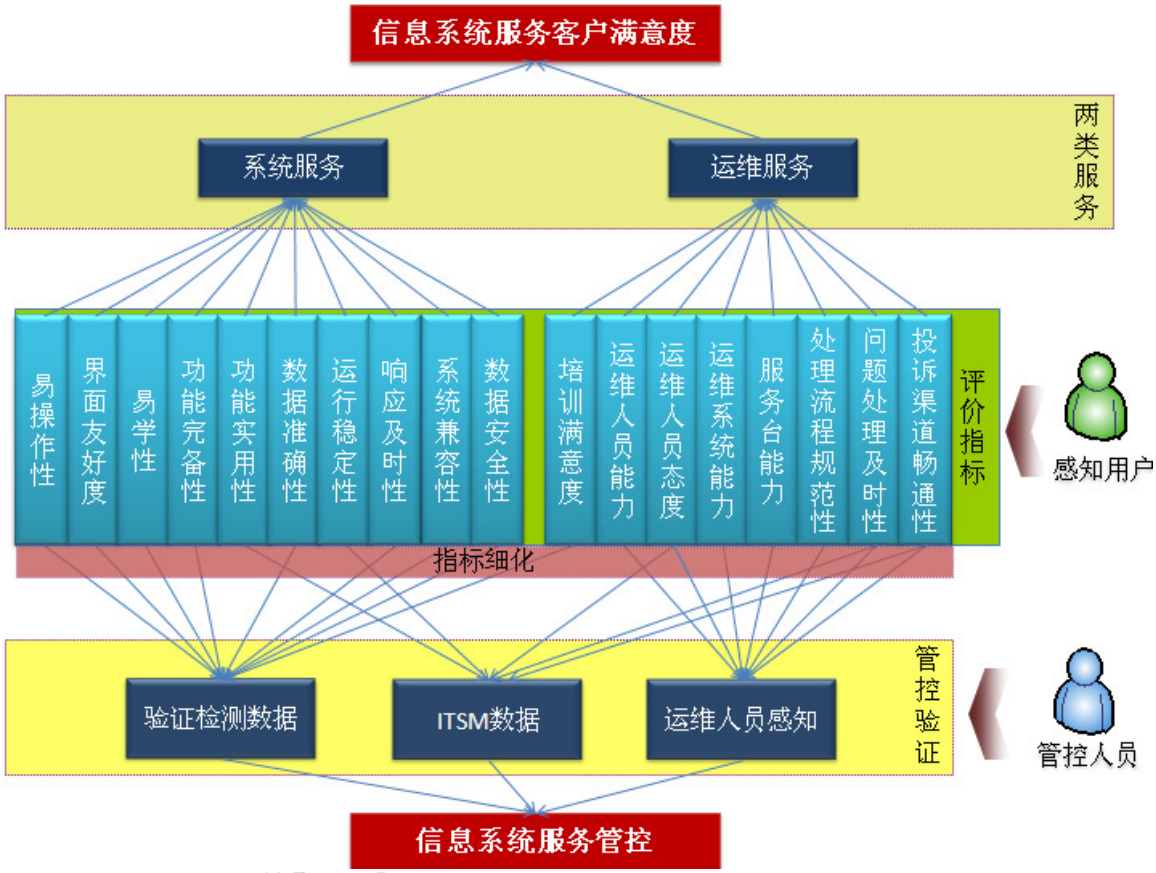


图 2-46 信息系统客户服务感知及管控结构

通过对项目需求的深入理解，结合前期部门访谈结果，分析出用户能够感知到的服务分为两类：系统服务和运维服务，相应地归纳总结出两大类指标：

- ✧ 系统服务指标
- ✧ 运维服务指标

通过对信息系统满意度评价用户进行分析，信息系统服务感知用户主要分为两类：

- ✧ 业务人员（感知对象：业务系统和系统运维服务）
- ✧ 运维人员（感知对象：运维系统和运维服务）

其中，运维系统是一类特殊的应用系统（具有提供服务和被感知的双重身份）；运维服务人员既是运维系统的感知人员，又是业务人员感知运维服务的对象，即具有感知系统和提供服务的双重身份。研究发现仅靠业务人员感知运维服务不能准确发现运维服务各方面存在的问题，业务人员对运维服务的感知停留在运维服务的表层，难以感知信息系统运维服务深层面的问题（特别是运维部门内部存在的问题），要有效提升信息系统运维服务质量，需要充分考虑参与运维服务人员对运维服务提出的意见和建议。评价指标体系与感知用户关系如图 2-47 所示：

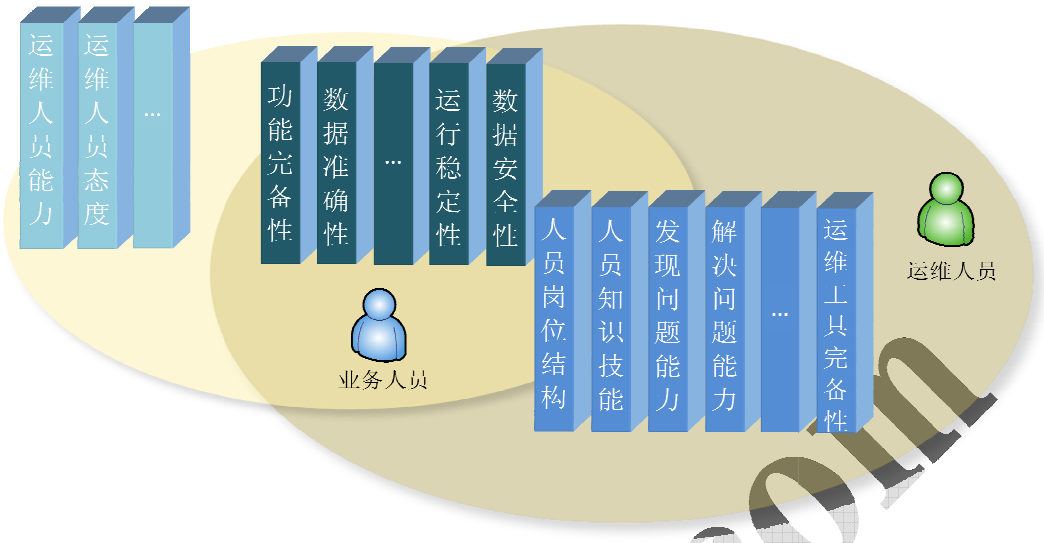


图 2-47 评价指标体系与感知用户关系

基于此评价指标体系，客户服务感知将分两步进行：

第一步，由业务系统人员做感知评价，定位业务系统服务过程中存在的问题（对运维服务做出表层评价）；

第二步，由运维人员做感知评价，定位运维服务过程存在的具体问题（其评价指标为第一步运维服务指标的细化，对运维服务做出深层评价）。

评价指标体系一：系统客户感知评价指标体系

通过对某电信企业多个业务部门访谈调研，综合分析信息系统客户服务感知在其它行业的应用情况，归纳整理出基于供电行业的信息系统客户服务感知指标评价体系，一级指标包括系统易用性、系统功能、系统性能、运维人员、运维资源、运维过程等 6 个方面，二级指标共 18 个，指标体系结构如图 2-48。

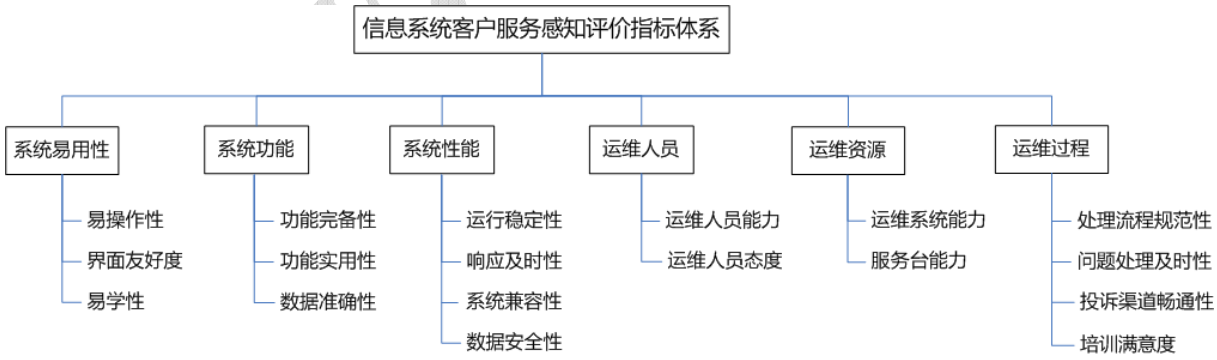


图 2-48 信息系统客户服务感知评价指标体系

评价指标体系二：运维服务感知指标体系

针对运维人员的感知评价指标，主要包括系统功能、系统性能、运维人员、运维资源、运维技术和运维过程等 6 个一级指标及 18 个二级指标，指标体系结构如图 2-49。

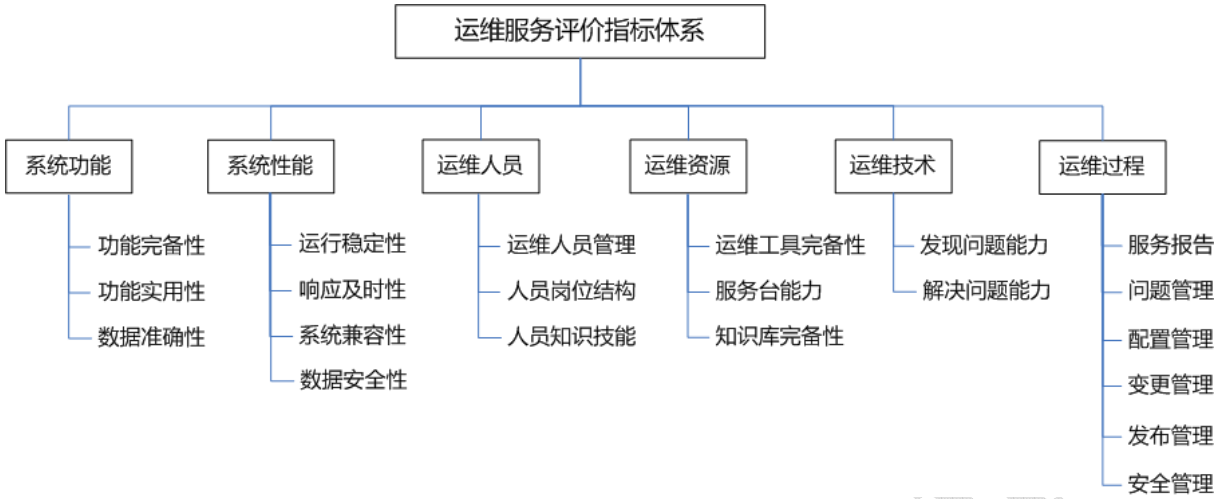


图 2-49 运维服务感知评价指标体系

两套评价指标体系的关联关系如图 2-50。

www.tipdm.com

图 2-50 两套评价指标体系关系图

如图 2-50，以“故障处理及时性”为例，说明两个评价指标体系间的关系。第一层的“故障处理及时性”评价指标主要与第二层的功能完备性(运维系统)、运行稳定性(运维系统)、响应及时性(运维系统)、人员岗位结构(运维人员)、人员知识技能(运维人员)、运维工具完备性、服务台能力、发现故障能力(运维人员)、解决故障能力(运维人员)等评价指标有关。当业务人员对信息系统进行感知评价时，发现“故障处理及时性”评价指标偏差，可以由运维人员对功能完备性(运维系统)、运行稳定性(运维系统)等相关运维评价指标进行感知，找出具体问题所在。

进行样本数据采集（个人问卷调查）时，以 5 分制评分，分别为 5（很好）、4（好）、3（一般）、2（差）、1（很差）。

当样本融合后，综合评价结果分为五级，取值范围分别为：优[4.5~5]、良[3.5~4.5)、中[2.5~3.5)、

差[1.5~2.5)、很差[1~1.5)。

2. 数据采集

本例中，样本数据采集经历了问卷准备、问卷测试、试点调查、批量调查等几个步骤，如图 2-51。

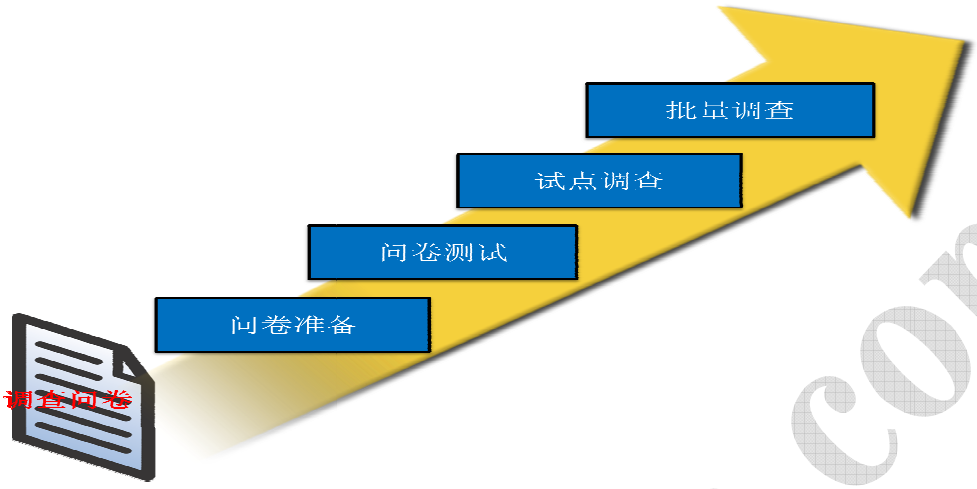


图 2-51 样本数据采集流程

1）问卷准备

建模的样本数据主要是通过调查问卷的方式收集，由某电信企业的业务人员从 18 个感知角度及总体满意度对业务系统进行评价。参与调查的业务人员来自各部门、各岗位，业务系统包括在用的主要系统。

为方便问卷调查，项目组按不同业务部门个性化定制《客户服务感知个人调查问卷》，即不同业务部门调查的业务系统不同。调查问卷采用单选题的方式，对 18 个评价指标项进行评分，同时收集各评价指标项的改进建议。

2）问卷测试

由于本例采用数据挖掘技术对信息系统进行服务感知评价，样本数据至关重要，而调查问卷质量高低对调查结果的真实性、适用性等具有决定性的作用。为了确保问卷具有较高的可靠性和有效性，在形成正式问卷之前，需对问卷进行试测，并对试测结果进行信度和效度分析，根据分析结果筛选问卷题项，调整问卷结构，从而提高问卷的信度和效度。

问卷测试工作主要在太普公司内部进行，从不同岗位抽取 8 名员工进行问卷测试（调查对象为公司目前使用的业务系统：财务报销系统和办公 OA 系统），并对回收问卷进行信度分析。

信度分析能对回收问卷的调查结果的正确性和一致性进行评价，从而得到问卷调查结果的可靠度和可信度。本项目研究中对公司内部测试回收的问卷应用“α 信度系数法”进行信度分析。再根据分析结果筛选问卷题项，调整问卷结构，从而有效保证了调查问卷的可信度和可靠度。

问卷测试结果表 2-79。

www.tpom.com

表 2-79 问卷测试结果

姓名	应用系统	岗位	易操作性	界面美好度	易学性	功能完备性	功能实用性	数据准确性	运行稳定性	响应及时性	系统兼容性	数据安全性	运维人员能力	运维人员态度	运维
李**	办公OA系统	测试工程师	一般	一般	一般	差	一般	好	一般	好	一般	好	一般	好	一
刘**	办公OA系统	技术总监	一般	一般	一般	一般	一般	好	一般	一般	好	好	一般	好	一
陆**	办公OA系统	项目主管	差	一般	差	一般	一般	好	一般	好	好	好	一般	一般	一
王**	办公OA系统	开发工程师	一般	一般	一般	差	一般	一般	一般	一般	一般	一般	差	一般	一
尹**	办公OA系统	财务总监	一般	一般	好	一般	一般	好	一般	一般	好	好	一般	好	一
张**	办公OA系统	平面设计师	一般	差	一般	好	一般	一般	好	一般	一般	一般	一般	差	差
赵**	办公OA系统	开发工程师	好	好	一般	一般	好	一般	好	一般	好	好	一般	一般	一
龙**	办公OA系统	财务助理	一般	好	一般	一般	好	一般	好	一般	好	好	一般	一般	好
李**	财务报销系统	测试工程师	差	一般	差	一般	一般	好	一般	一般	一般	好	一般	一般	一
刘**	财务报销系统	技术总监	一般	差	一般	差	差	一般	一般	一般	一般	好	一般	差	差
陆**	财务报销系统	项目主管	差	一般	好	差	差	好	一般	一般	一般	好	一般	好	一
王**	财务报销系统	开发工程师	差	差	一般	一般	一般	一般	一般	一般	差	一般	差	一般	差
尹**	财务报销系统	财务总监	一般	差	一般	差	差	一般	一般	一般	一般	好	一般	差	差
张**	财务报销系统	平面设计师	差	差	一般	差	一般	一般	一般	差	差	一般	差	一般	差
赵**	财务报销系统	开发工程师	一般	好	一般	一般	好	一般	好	一般	好	一般	好	好	好
龙**	财务报销系统	财务助理	好	一般	一般	一般	好	一般	好	一般	好	一般	好	一般	好

www.ijqdm.com

将测试的办公 OA 系统问卷结果应用“α信度系数法”进行问卷信度分析，各评价指标情况统计如表 2-80。

表 2-80 各评价指标统计情况

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
X1	58.3750	20.839	0.234	0.000	0.791
X2	58.2500	18.500	0.609	0.000	0.764
X3	58.3750	20.839	0.234	0.000	0.791
X4	58.5000	22.571	-0.117	0.000	0.818
X5	58.1250	20.125	0.464	0.000	0.778
X6	57.8750	20.982	0.204	0.000	0.793
X7	58.0000	22.000	0.000	0.000	0.804
X8	58.1250	22.696	-0.146	0.000	0.810
X9	57.7500	18.786	0.716	0.000	0.761
X10	57.6250	18.554	0.878	0.000	0.754
X11	58.5000	20.286	0.583	0.000	0.775
X12	58.1250	18.982	0.452	0.000	0.777
X13	58.3750	19.696	0.482	0.000	0.776
X14	58.2500	18.786	0.553	0.000	0.769
X15	58.2500	21.071	0.330	0.000	0.786
X16	58.0000	21.143	0.084	0.000	0.808
X17	58.6250	20.268	0.428	0.000	0.780
X18	58.0000	19.429	0.564	0.000	0.771
X19	57.6250	18.554	0.878	0.000	0.754

信度分析结果如表 2-81。

表 2-81 信度统计结果

Cronbach’s Alpha	Cronbach’s Alpha Based Standardized Items	N of Items
0.792	0.811	19

表中α=0.792，显著相关性说明该套问卷的信度值较好，这也证明这份调查问卷能够稳定地反映实际情况。

同样将测试的财务报销系统问卷结果导入信度分析工具中，并应用“α信度系数法”进行问卷信度分析。得到α=0.896，同样也证明该套问卷的信度值较好，能够稳定地反映出问卷调查的实际情况。

通过这次试卷的信度检验合格后，证明问卷设计是科学可行的，为后期正式发放调查问卷，保证问卷效度奠定了基础。

3）试点调查

在某电信企业选择了两个试点业务部门（客服部和基建部），两个部门试点情况分析：回收周期在 7 个工作日内，除了少数问卷不合格（主要表现为必填项未填写），大部分回收问卷有效。如基建部问卷总数为 17 份，有 16 份基本符合要求，有效率为 94.12%。通过对回收问卷进行统计分析，分析问卷设计不足并对问卷进行二次优化。

4）批量调查

根据前两个阶段的试用情况，优化个人问卷调查方案，并在全公司其它业务部门进行批量调查。

回收问卷有效性分析

对于大批量的问卷调查，通常会存在回收问卷质量低下的情况，为保证后续感知评价模型的构建，需对回收的调查问卷有效性进行判断，以确保样本数据的真实可用性。

所有问卷是按部门个性化定制的，即不同部门虽然问题一致，但是因为应用系统的不同，调研的问卷是有差别的。回收问卷的有效性主要基于以下几点判断：

5）问卷中某些指标项未填或漏填；

指标项得分与总体评价不一致，即总体评价结果超出了指标项的边界范围，如 18 个输入指标的结果均介于“好”与“差”之间（输入范围为：很好、好、一般、差、很差），而总体评价结果为“优”（总体评价范围：优、良、中、差、很差）。

本次调研，共涉及 19 个业务部门，参与调研人员 1216 人，共回收对信息系统的感知评价 3111 份（每个用户感知评价 2~4 个应用系统），经过上述规则的筛选之后，剩余 2356 份，有效率 75.73%。

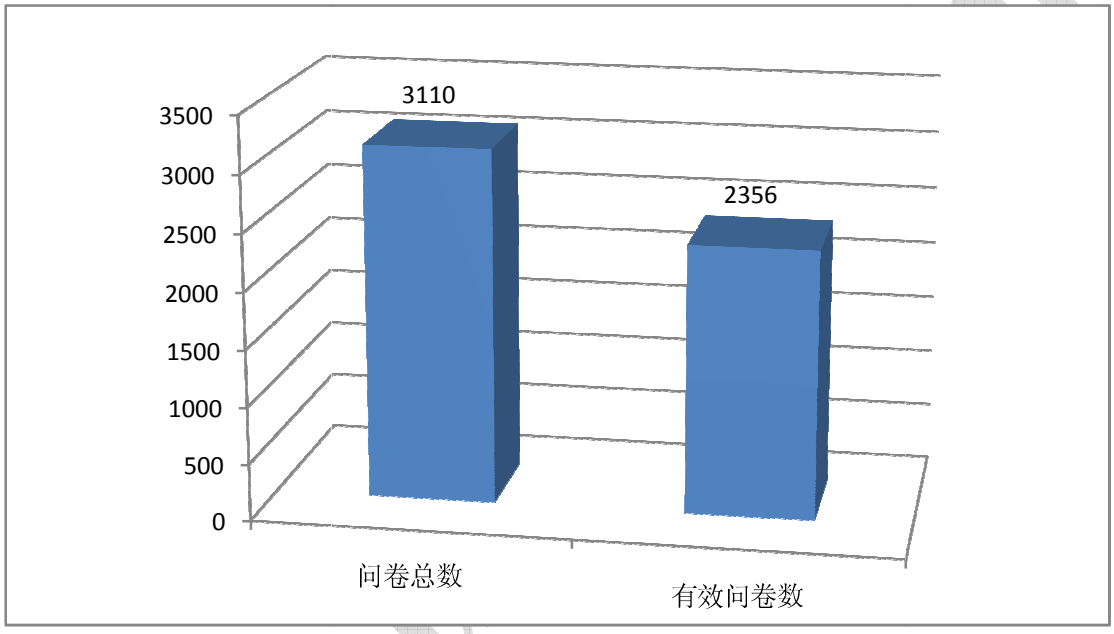


图 2-52 样本数据采集情况

对回收的有效问卷，按部门、岗位、系统等维度整理，并进行分布分析。

有效问卷的部门分布情况如图 2-53。

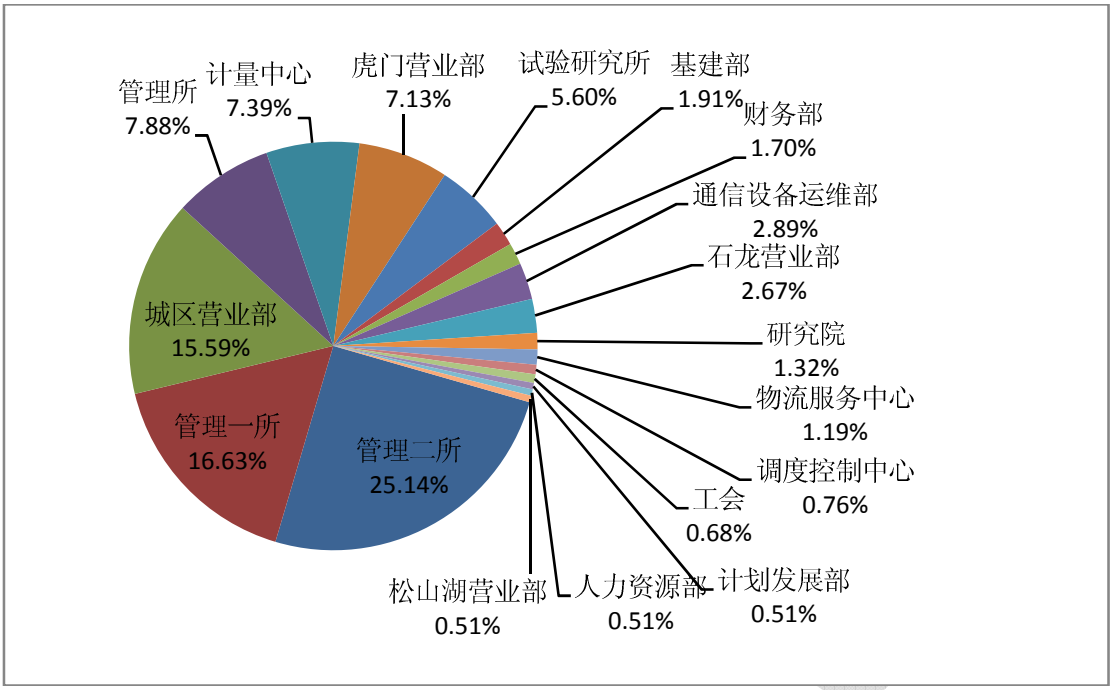


图 2-53 按部门维度分布情况

岗位分布情况如图 2-54。

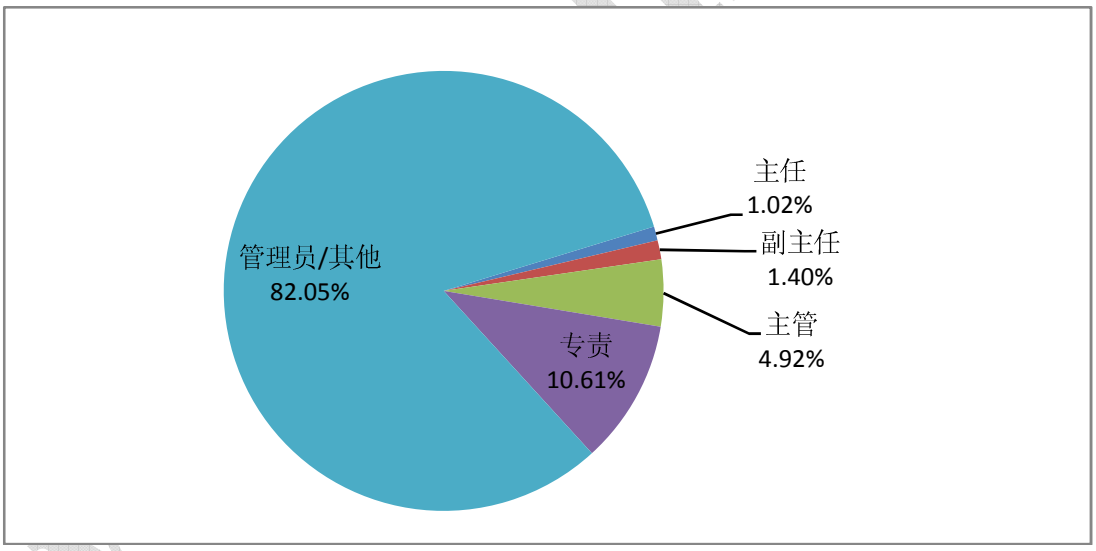


图 2-54 按岗位维度分布情况

业务系统分布情况如图 2-55。

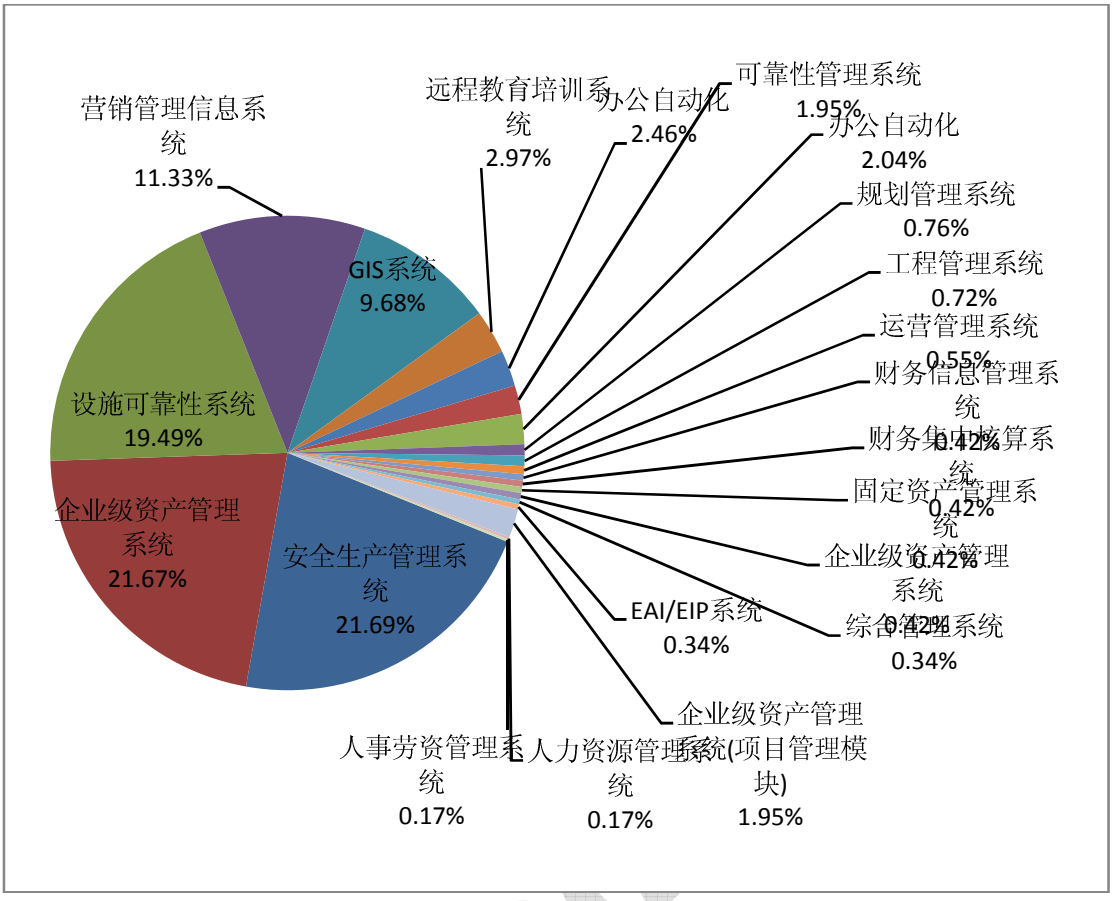


图 2-55 按系统维度分布情况

3. 数据整理

“数据整理”包括从原始的未加工的回收调查问卷到构造最终用于建模分析的数据集的所有活动，是构建客户服务感知模型过程中最耗时的环节，数据整理工作的流程如图 2-56 所示。



图 2-56 样本数据整理过程

1）数据选择

由于回收的样本数据集中可能存在和学习任务不相关的属性/变量，或者是高度相关而冗余属性/变量，它们会导致无效的归纳和降低学习的效率。

数据选择包括数据集中属性（列）和记录（行）的选择，通过对样本数据的相关性分析，见后续相关章节，18 项评价指标（属性）将全部保留，本研究中只对记录（行）进行选择。

针对数据记录（行），选择标准包括与数据挖掘目标的相关性，数据质量（如问卷中过多指标项未填或漏填、指标项得分与总体评价明显不一致）。

对回收的感知样本 3111 份经过上述规则筛选后，剩余 2356 份，进入数据清洗环节。

2）数据清洗

数据清洗主要是对异常数据（离群值或异常值）进行检测，并将数据质量提高到所选分析技术和分析目标要求的水平。本项目研究中清洗数据工作主要包括：

✧ 插入适当的默认值来估计缺失值

✧ 过滤掉重复的样本数据

3）数据构造

由于问卷调查是针对某电信企业的业务部门，加之回收的问卷有部分无效。采集的样本数据非常有限，有必要对缺失边界的属性进行样本数据构建，以确保样本数据达到具备构建感知模型的有效样本集的条件。构建数据工作主要包括：

✧ 对缺失边界的属性，可通过趋势外延法增补样本

✧ 对数据过度稀疏的属性，可通过内插法增补样本

✧ 增加新的指标因子时，通过对应用系统用户访谈及根据建模专家经验增加样本

经过数据清洗和数据构造后，样本数据剩余 2346 条。

4）数据格式化

将样本数据标准化（如样本数值化、规范化）为构建感知模型的专家样本集。

5）样本数据总体情况分析

本项目研究中基于模糊神经网络，采用 10-折交叉验证（10-fold cross validation），对经预处理后的总体样本（2346 条）进行综合评估。10 折交叉验证结果如下：

指标	值
正确分类样本及百分比	2042　　87.0418 %
错误分类样本及百分比	304　　12.9582 %
Kappa 统计	0.7914
平均绝对误差	0.0646
均方根误差	0.202
样本总数	2346

混淆矩阵结果见表 2-82。

表 2-82 混淆矩阵结果

a	b	c	d	e	<— 分类为
11	1	1	0	0	a = 很差
6	172	52	2	0	b = 差
0	3	663	26	10	c = 中
0	3	74	1091	27	d = 良
0	0	12	80	115	e = 优

表 2-82 中，以第 3 行为例，有 663 个样本正确分类(属于“中”)，有 26 个样本本应该属于“中”，却错误分到了“良”，有 10 个样本本应属于“中”，却错误分到了“优”， 有 3 个样本本应属于“中”，而错误分到了“差”。

总体来说，10 折交叉验证的正确分类百分比为 87.04%，可见经抽选的专家样本数据能满足建模要求。

www.tipdm.com

4. 构建模型

本次针对某电信企业的个人问卷调查，共涉及 19 个业务部门、5 类岗位、21 个应用系统，经数据预处理后，将其中的 19 个业务系统的样本数据用于预测建模，任意保留两个业务系统（本项目研究中选择生产管理系统和营销管理系统）的数据用于模型验证。

1）专家样本维护

专家样本库数据来源于某电信企业用户对 19 个应用系统的问卷调查，经过数据清洗、数据构造，初步整理出专家样本数据集，当应用环境发生变化时，如需增加或调整感知评价指标，可通过对应用系统用户进行访谈，分析增加或调整后的评价指标对总体评价结果的影响，在专家样本集中构建相应评价指标的专家样本数据。

本项目研究中共整理出专家样本数据 2346 条。

2）建立模型

本项目研究中，主要基于模糊神经网络(FNN)进行客户服务感知建模，建立模型流程如图 2-57。

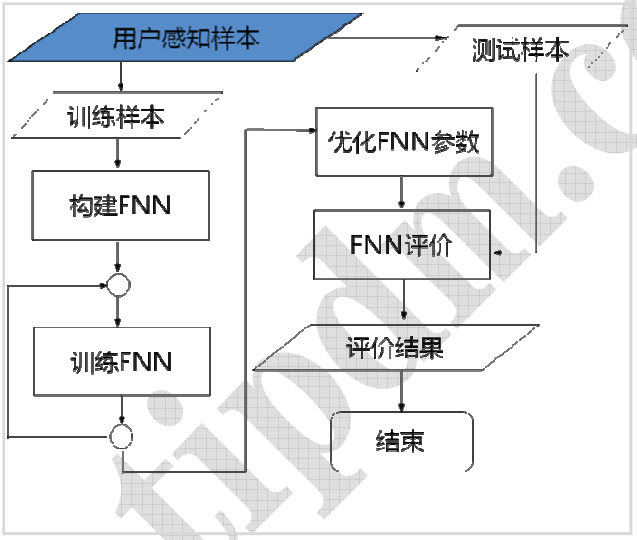


图 2-57 模糊神经网络建模流程

针对本项目的实际情况，模糊神经网络各层说明如下：

第 1 层：输入层，本项目中输入层节点数为 18，即对应 18 个输入节点，分别对应影响客户服务感知满意度的 18 个指标——系统易用性、界面美观性、页面帮助完备性、功能完备性、功能实用性、数据准确性、查询便利性、运行稳定性、响应及时性、数据安全性、数据服务支撑的稳定性和及时性、投诉渠道畅通性、故障处理及时性、问题处理方便性、培训满意度、期望目标实现度、建议采纳度。

第 2 层：模糊化层，本项目中，输入变量共有 18 个：（X1，X2，……X18），将每个输入变量划分为 3 个模糊度(即模糊集合 Ai，当 i =1, 2, 3 时)，18 个变量，共划分为 18*3＝54 个模糊度，每个节点的激活函数分别为输入变量 Xi 的各个模糊度的隶属度函数(即 $\mu_A(x)$，当 i =1, 2, …, m 时)，其输出分别为输入变量 Xi 的各个模糊度的隶属度函数值，是[0, 1]之间的值。共 54 个神经元。

$$O_i^2 = \mu_A(x_1) \qquad i = 1,2,3$$

$$O_j^2 = \mu_A(x_2) \qquad j = 4,5,6$$

……

$$O_k^2 = \mu_A(x_{18}) \qquad k = 52,53,54$$

www.ijdm.com

本项目中，隶属度函数$\mu_A(x)$采用高斯函数：

$$g(x;c,\sigma)=e^{-\frac{1}{2}(\frac{x-c}{\sigma})^2}$$

其中c为隶属度函数的中心；σ为隶属度函数的宽度。建模开始时，c和σ通过随机函数初始化，之后在网络训练过程中不断修正调整，直至稳定。

第 3 层：模糊推理层，该层每个节点只与第二层中每个节点中的一个等相连，共有节点数为 18 个 3 相乘，也就是有（18 个 3 相乘）条规则。

第 4 层：去模糊化层，节点数为输出变量模糊度划分的个数,本项目为 5，即对应总体评价的 5 个等级：优、良、中、差、很差（数值化表示为 5、4、3、2、1）。该层与第三层的连接为全互连，连接权值为 Wkj。

第 5 层：输出层，该层将第四层各个节点的输出，转换为输出变量的精确值。该层节点数为输出变量的个数，本项目为 1 个，即总体评价值。

3）模型验证

本例中，选择两个有代表性的业务系统（生产管理系统和营销管理系统）对训练好的模型进行验证，生产管理系统感知样本数据见表 2-83。

表 2-83 生产管理系统感知样本数据

部门	岗位	易操作性	界面友好度	易学性	功能完备性	功能实用性	数据准确性	运行稳定性	响应及时性	系统兼容性	数据安全性	运维人员能力	运维人员态度	运维系统能力	服务台能力	处理流程规范性	故障处理及时性	投诉渠道畅通性	培训满意度	总评结果
基建部	主任	4	3	3	4	4	4	3	3	3	4	4	3	3	3	4	4	4	4	3
电力调度	主任	4	3	3	3	3	4	3	4	2	4	2	3	2	3	4	3	2	4	3
试验研究	主任	3	2	3	2	3	2	1	2	1	3	2	4	4	2	2	2	3	3	2
基建部	副主任	4	3	3	4	4	4	3	3	3	4	4	3	3	3	4	4	4	4	3
通信设备	副主任	4	3	4	3	3	3	3	4	4	4	4	4	4	4	4	3	3	4	4
计量中心	副主任	4	5	4	4	4	4	4	5	4	4	4	4	4	4	4	4	4	4	4
变电管理	副主任	4	4	3	3	3	4	4	4	4	4	2	4	3	3	4	4	4	4	4
基建部	管理员	4	5	4	3	3	3	3	4	4	5	3	4	4	4	3	3	4	3	3
基建部	管理员	4	4	4	3	3	3	3	4	4	5	3	4	4	3	3	3	4	3	3
基建部	管理员	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
基建部	管理员	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
电力调度	管理员	4	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4
电力调度	管理员	3	4	3	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3
通信设备	管理员	4	4	4	4	3	4	4	5	5	5	4	4	5	4	4	4	4	5	5
通信设备	管理员	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
通信设备	管理员	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4
通信设备	管理员	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4
通信设备	管理员	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
通信设备	管理员	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4
通信设备	管理员	4	4	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4	4	4
通信设备	管理员	3	4	4	3	3	4	3	4	3	4	4	3	3	4	4	4	4	4	4
通信设备	管理员	3	4	4	3	3	4	3	4	3	4	4	3	3	4	4	4	4	4	4

营销管理系统感知样本数据见表 2-84。

表 2-84 营销管理系统感知样本数据

部门	岗位	易操作性	界面友好度	易学性	功能完备性	功能实用性	数据准确性	运行稳定性	响应及时性	系统兼容性	数据安全性	运维人员能力	运维人员态度	运维系统能力	服务台能力	处理流程规范性	故障处理及时性	投诉渠道畅通性	培训满意度	总评结果
石龙配电营	主任	4	4	4	5	4	4	4	5	4	4	4	5	5	5	5	5	5	5	4
城区配电营	主任	4	3	4	4	4	4	4	4	3	4	4	3	3	4	4	4	4	4	4
松山湖配电	主任	4	5	5	4	5	4	4	3	4	5	3	3	2	2	4	3	3	4	4
计量中心	副主任	5	4	4	3	3	3	3	3	3	4	3	3	3	3	4	3	3	3	3
石龙配电营	副主任	5	4	5	5	5	5	4	4	5	5	5	5	5	5	5	4	5	5	5
石龙配电营	副主任	5	5	5	4	5	5	5	4	5	5	4	5	4	5	5	5	5	5	5
虎门配电营	副主任	3	3	3	3	3	3	4	4	4	3	3	3	3	3	4	4	4	3	3
城区配电营	副主任	4	4	4	4	5	3	4	4	4	4	4	4	4	4	4	3	4	4	4
城区配电营	副主任	4	3	3	3	3	3	4	3	4	4	3	4	4	4	4	4	3	4	4
计量中心	管理员/	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
计量中心	管理员/	2	2	2	2	3	3	1	2	2	2	2	2	1	2	2	2	2	1	3
计量中心	管理员/	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
计量中心	管理员/	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
计量中心	管理员/	4	4	3	3	4	3	3	3	3	4	3	4	3	4	3	3	4	3	4
计量中心	管理员/	4	4	3	3	5	5	4	2	4	4	4	4	3	4	4	4	3	3	4
计量中心	管理员/	4	4	3	3	5	5	4	2	4	4	4	4	3	4	4	4	3	3	4
计量中心	管理员/	4	3	3	3	3	4	3	4	3	3	3	3	4	4	3	4	4	3	4
计量中心	管理员/	4	4	4	3	4	4	3	3	3	3	4	3	3	4	4	3	3	4	4
计量中心	管理员/	4	4	3	3	5	5	4	2	4	4	4	4	3	4	4	4	3	3	4
计量中心	管理员/	4	3	4	4	4	4	4	4	4	5	4	3	3	3	4	4	4	4	4
计量中心	管理员/	4	4	3	4	4	4	4	4	4	4	4	4	4	3	3	4	4	4	4
计量中心	管理员/	3	4	3	3	2	3	2	3	2	3	3	3	3	2	3	3	3	3	3
计量中心	管理员/	4	3	5	5	5	4	5	3	4	5	4	5	4	4	4	4	5	4	4
计量中心	管理员/	4	4	3	3	5	5	4	2	4	4	4	4	3	4	4	4	3	3	4
计量中心	管理员/	4	3	4	4	4	4	4	4	4	5	4	3	3	3	4	4	4	4	4
计量中心	管理员/	4	3	4	4	4	4	4	4	4	5	4	3	3	3	4	4	4	4	4
计量中心	管理员/	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

以下以生产管理系统为例进行后继分析。

✧ 分布分析

分别按用户的岗位维度和部门维度，对样本进行分布分析，确保用于验证的样本集合比较合理。生产管理系统的问卷调查样本数共 511 条，按岗位维度和部门维度分布如图 2-58、图 2-59。

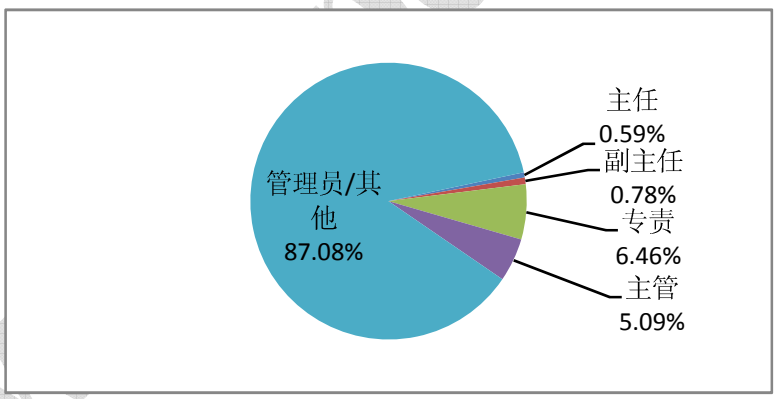


图 2-58 按岗位维度分布图

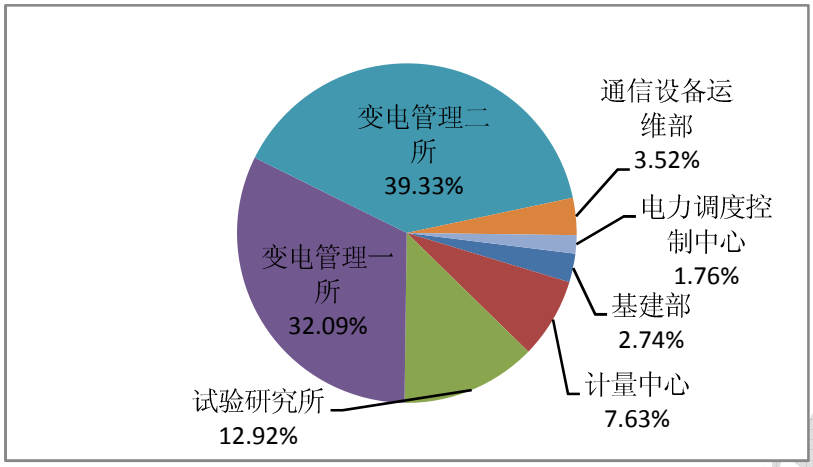


图 2-59 按部门维度分布图

通过以上分布分析，管理员、专责、主管所占比重较大，由于专责、主管和管理员等岗位在信息系统感知评价中起主要作用，同时样本数据也覆盖了主要业务部门，因此用生产管理系统进行模型验证是可行的。

✧ 样本融合

对用于验证的样本数据进行融合处理，即对多个用户的 18 个输入和 1 个输出指标进行融合处理，融合后结果如表 2-85。

表 2-85 验证数据融合结果

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	易操作性	界面友好度	易学性	功能完备性	功能实用性	数据准确性	运行稳定性	响应及时性	系统兼容性	数据安全性	运维人员能力	运维人员态度	运维系统能力	服务台能力	处理流程规范性	故障处理及时性	投诉渠道畅通性	培训满意度	总评结果
494	3	3	4	3	3	3	3	3	3	4	4	3	4	4	4	2	3	4	4
495	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
496	3	3	3	2	3	4	3	3	4	4	3	4	3	4	4	3	3	4	4
497	3	3	3	3	3	4	3	3	4	3	3	3	3	3	3	3	3	3	3
498	2	3	3	3	2	3	3	3	3	3	3	3	3	3	2	2	2	3	3
499	3	3	3	3	3	3	3	3	4	4	3	3	4	3	3	3	3	3	3
500	3	3	3	3	3	3	3	3	4	4	3	3	4	3	3	3	3	3	3
501	3	3	3	3	2	3	2	3	3	3	3	4	2	2	3	3	2	2	3
502	5	4	4	4	4	4	3	3	3	3	4	3	3	4	4	4	3	3	3
503	3	3	3	3	2	3	2	3	3	3	3	4	2	2	3	3	2	2	3
504	3	3	3	3	2	3	2	3	3	3	3	4	2	2	3	3	2	2	3
505	3	3	3	3	3	3	4	3	3	4	4	4	3	3	3	3	3	3	3
506	4	4	4	3	4	3	3	3	3	4	4	4	3	4	4	3	4	4	4
507	5	4	4	4	4	4	3	3	3	3	4	3	3	4	4	4	3	3	3
508	3	3	3	3	2	3	2	3	3	3	3	4	2	2	3	3	2	2	3
509	3	2	3	2	3	2	1	2	1	3	2	4	4	2	2	2	3	3	2
510	4	4	4	3	4	4	4	4	4	4	3	4	4	4	4	3	4	4	4
511	3	3	3	3	3	3	3	4	3	4	4	2	2	2	2	3	3	3	3
512	3	3	3	3	4	4	3	4	4	4	3	3	2	3	2	3	3	4	4
513	3.5702	3.5843	3.3836	3.5722	3.6010	3.5754	3.4740	3.4894	3.3817	3.7377	3.6523	3.3419	3.2059	3.1539	3.3182	3.3842	3.4298	3.5151	3.5690

✧ 模型验证

将融合后的样本数据导入训练好的模型，分析预测结果与用户实际评价结果。下面为对验证样本的预测输出如表 2-86。

表 2-86 模型预测结果

输入项	易操作性	3.5702
	界面友好度	3.5843

	易学性	3.3836
	功能完备性	3.5722
	功能实用性	3.6010
	数据准确性	3.5754
	运行稳定性	3.4740
	响应及时性	3.4894
	系统兼容性	3.3817
	数据安全性	3.7377
	运维人员能力	3.6523
	运维人员态度	3.3419
	运维系统能力	3.2059
	服务台能力	3.1539
	处理流程规范性	3.3182
	故障处理及时性	3.3842
	投诉渠道畅通性	3.4298
	培训满意度	3.5151
预测项	总评结果	3.5736

由表 2-86 知，模型预测结果为：3.5736，相对误差为 1.29%，

用同样方法对另一系统（营销管理系统）进行模型验证，模型预测结果为 3.6386，相对误差为 0.98%。

通过生产管理系统和营销管理系统对感知评价模型进行验证，预测结果与实际用户总体感知评价结果基本吻合，说明基于模糊神经网络的感知评价模型应用于本项目是可行的。

5. 模型评价

模型评价是对同一采样数据可以用多种数据分析方法和模型进行分析，从多个模型中找出最优模型。

本例中，分别用生产管理系统、营销管理系统对已构建的感知评价模型进行验证，对两系统的评价指标进行数据融合后的结果如表 2-87。

表 2-87 评价指标融合后结果

序号	评价指标	评价指标融合后结果	
		生产管理系统	营销管理系统
1	易操作性	3.5702	3.7037
2	界面友好度	3.5843	3.6622
3	易学性	3.3836	3.6634
4	功能完备性	3.5722	3.5829
5	功能实用性	3.6010	3.7366
6	数据准确性	3.5754	3.6720
7	运行稳定性	3.4740	3.5000
8	响应及时性	3.4894	3.3061
9	系统兼容性	3.3817	3.4317
10	数据安全性	3.7377	3.8146

www.ipqm.com

11	运维人员能力	3.6523	3.5878
12	运维人员态度	3.3419	3.4585
13	运维系统能力	3.2059	3.4183
14	服务台能力	3.1539	3.3866
15	处理流程规范性	3.3182	3.5829
16	故障处理及时性	3.3842	3.4317
17	投诉渠道畅通性	3.4298	3.4024
18	培训满意度	3.5151	3.5390
19	总评结果	3.5690	3.6707

本例中，除了用模糊神经网络完成信息系统客户服务感知评价建模外，同时通过回归分析、BP 神经网络、RBF 神经网络建模，对不同模型算法的建模结果进行对比分析。不同模型算法的预测评价结果见表 2-88。

表 2-88 多个模型评价结果

序号	感知评价模型	模型评价结果	
		生产管理系统	营销管理系统
1	回归分析	3.7586	3.9299
2	BP 神经网络	3.9367	3.9644
3	RBF 神经网络	3.8080	3.8104
4	FNN 神经网络	3.5736	3.6386

由上表知，不同感知模型的评价结果基本都能反映用户对应用系统的满意度评价情况，具体哪个算法最优，可通过不同算法在模型验证时的均方根误差来衡量，对比分析结果如表 2-89。

表 2-89 多个模型评价结果误差分析

感知评价模型	生产管理系统		均方根误差
	实际值	预测值	
回归分析	3.5690	3.7586	0.3297
BP 神经网络	3.5690	3.9367	0.3328
RBF 神经网络	3.5690	3.8080	0.1958
FNN 神经网络	3.5690	3.5736	0.0229

从建模过程及验证结果来看，FNN 神经网络虽然比回归分析、BP 神经网络、RBF 神经网络等算法建模速度稍慢，但总的来说 FNN 神经网络比回归分析、BP 神经网络、RBF 神经网络等算法的预测精度均要高，这也体现了模糊神经网络用于信息系统客户服务感知建模的优势。

6. 模型应用

模型的创建通常并不是目的，在模型创建并经验证满足实际应用后，需发布以完成实际的数据挖掘过程。

1）用户感知

由用户对指定的应用系统进行客户服务感知评分，构成感知评价样本集，为后续模型评价提供数据源，调查问卷见图 2-60。

信息系统客户服务感知调查问卷			
序号	问题	答案选项	
		安全生产管理系统	创先指标管理系统
1	**易操作性**（是否通俗易懂，容易使用？是否符合操作习惯，不用培训或简单培训就会使用？）	○很好 ○好 ○一般 ○差 ○很差	○很好 ○好 ○一般 ○差 ○很差
	您的改进建议：		
2	**页面友好度**（页面布局是否合理，内容主次是否分明？是否不易产生视觉疲劳？）	○很好 ○好 ○一般 ○差 ○很差	○很好 ○好 ○一般 ○差 ○很差
	您的改进建议：		
3	**易学性**（专业术语是否有说明？操作是否有提示？必填项是否有标明？）	○很好 ○好 ○一般 ○差 ○很差	○很好 ○好 ○一般 ○差 ○很差
	您的改进建议：		
4	**功能完备性**（系统功能模块是否齐全？有没有需要增加的功能模块？）	○很好 ○好 ○一般 ○差 ○很差	○很好 ○好 ○一般 ○差 ○很差
	您的改进建议：		
5	**功能实用性**（各功能模块是否符合实际应用需要？与实际工作的结合度高不高？）	○很好 ○好 ○一般 ○差 ○很差	○很好 ○好 ○一般 ○差 ○很差
	您的改进建议：		
6	**数据准确性**（系统数据是否完全准确[不包含用户误操作所造成的数据偏差]？）	○很好 ○好 ○一般 ○差 ○很差	○很好 ○好 ○一般 ○差 ○很差
	您的改进建议：		
7	**查询便利性**（查询条件是否足够？能否满足各种查询、统计需要？常用查询是否做成专用查询按钮？）	○很好 ○好 ○一般 ○差 ○很差	○很好 ○好 ○一般 ○差 ○很差
	您的改进建议：		

图 2-60 信息系统客户服务感知调查问卷

2）分布分析

为了保证感知用户对信息系统评价的合理性，需对采集的数据进行分布分析（主要针对岗位维度，参见 4.2.4.2 节），确保不同岗位的感知评价对象分布在一个相对合理的范围。

3）样本融合

信息系统的感知评价是对多个采集样本的一个综合性满意度评价，在确保采集的感知评价样本分布在一个合理的范围后，需对样本数据进行融合处理，具体规则：

✧ 确定不同人员岗位的权系数（由管控人员维护）

序号	岗位	权系数（ζ）
1	主任	1.0
2	副主任	1.0
3	主管	1.5
4	专责	2.0
5	管理员/其他	1.5

✧ 对多个用户的评价指标进行融合计算

多个用户对某个指标融合计算规则：

$$\frac{\sum(\sum v * \zeta)}{\sum \zeta}$$

其中：v 为感知评价人员不同岗位的评分，ζ 为不同岗位的权系数。

假如某组采集样本的单个评价指标（如培训满意度）的评价结果见表 2-90。

表 2-90 不同岗位用户评价结果

样本号	用户姓名	岗位	指标评价结果	指标数值映射
1	N1	主任	好	4
2	N2	副主任	很好	5
3	N3	主管	一般	3
4	N4	专责	好	4
5	N5	专责	好	4
6	N6	专责	很好	5
7	N7	管理员/其他	好	4
8	N8	管理员/其他	一般	3
9	N9	管理员/其他	一般	3

则对该指标的综合评价结果为：

$$=\frac{4*1.0+5*1.0+3*1.5+4*2.0+4*2.0+5*2.0+4*1.5+3*1.5+3*1.5}{1.0+1.0+1.5+2.0+2.0+2.0+1.5+1.5+1.5}$$

$=3.893$

即该组用户对“培训满意度”的综合评价为 3.893。

其它 17 个指标采用同样方法进行融合计算，最后得到一条针对所有人员的综合评价感知数据。

也可以从不同的角度，对生产管理系统样本数据分别按综合（所有岗位和所有部门）、按部门维度、按岗位维度分别进行融合处理结果见表 2-91、表 2-92。

www.cnki.com

表 2-91 综合（所有岗位和所有部门）融合结果

1	部门	岗位	易操作性	界面友好度	易学性	功能完备性	功能实用性	数据准确性	运行稳定性	响应及时性	系统兼容性	数据安全性	运维人员能力	运维人员态度	运维系统能力
487	电力调度	专责	4	4	3	3	4	4	4	4	3	4	3	3	4
488	通信设备	专责	4	4	4	4	3	4	4	5	5	5	4	4	5
489	通信设备	专责	4	4	3	4	4	5	4	4	3	4	3	4	5
490	通信设备	专责	4	4	4	4	4	4	4	3	4	4	4	4	4
491	通信设备	专责	4	4	4	4	3	4	4	5	5	5	4	4	5
492	计量中心	专责	4	4	5	4	4	5	3	4	4	5	4	4	5
493	计量中心	专责	4	4	4	4	4	4	4	4	4	4	4	4	4
494	计量中心	专责	3	3	4	3	3	3	3	3	3	4	4	3	4
495	计量中心	专责	3	3	3	3	3	3	3	3	3	3	3	3	3
496	计量中心	专责	3	3	3	2	3	4	3	3	4	4	3	4	3
497	变电管理	专责	3	3	3	3	3	4	3	3	4	3	3	3	3
498	变电管理	专责	2	3	3	3	2	3	3	3	3	3	3	3	3
499	变电管理	专责	3	3	3	3	3	3	3	3	4	4	3	3	4
500	变电管理	专责	3	3	3	3	3	3	3	3	4	4	3	3	4
501	试验研究	专责	3	3	3	3	2	3	2	3	3	3	3	4	2
502	试验研究	专责	5	4	4	4	4	4	3	3	3	3	4	3	3
503	试验研究	专责	3	3	3	3	2	3	2	3	3	3	3	4	2
504	试验研究	专责	3	3	3	3	2	3	2	3	3	3	3	4	2
505	试验研究	专责	3	3	3	3	3	3	4	3	3	4	4	4	3
506	试验研究	专责	4	4	4	3	4	3	3	3	3	4	4	4	3
507	试验研究	专责	5	4	4	4	4	4	3	3	3	3	4	3	3
508	试验研究	专责	3	3	3	3	2	3	2	3	3	3	3	4	2
509	试验研究	专责	3	2	3	2	3	2	1	2	1	3	2	4	4
510	变电管理	专责	4	4	4	3	4	4	4	4	4	4	3	4	4
511	变电管理	专责	3	3	3	3	3	3	3	4	3	4	4	2	2
512	变电管理	专责	3	3	3	3	4	4	3	4	4	4	3	3	2
513		样本融合	3.5702	3.5843	3.3836	3.5722	3.6010	3.5754	3.4740	3.4894	3.3817	3.7377	3.6523	3.3419	3.2059

www.

表 2-92 按部门维度融合结果

界面友好度	易学性	功能完备性	功能实用性	数据准确性	运行稳定性	响应及时性	系统兼容性	数据安全性	运维人员能力	运维人员态度	运维系统能力	服务台能力	处理流程规范性	故障处理及时性
3.817	3.248	3.606	3.787	3.616	3.548	3.345	3.441	3.566	3.417	3.528	3.508	3.358	3.343	3.4
3.566	3.505	3.648	3.695	3.640	3.663	3.976	3.556	4.091	4.160	3.103	2.889	2.915	3.234	3.4
3.030	3.061	3.061	3.303	3.545	3.061	3.152	2.576	3.242	2.939	3.242	3.212	3.091	3.242	2.9
3.267	3.311	3.378	3.533	3.467	3.022	3.311	3.111	3.778	3.356	3.311	3.311	3.178	3.311	3.2
3.298	3.669	3.521	3.479	3.545	3.372	3.446	3.438	3.851	3.537	3.388	3.372	3.165	3.488	3.4
3.189	3.233	3.432	3.010	3.243	2.971	2.738	2.825	3.257	3.282	3.248	2.733	2.908	3.146	3.0
3.912	3.930	3.649	3.456	3.982	3.649	3.912	3.860	4.193	3.930	3.737	4.000	4.000	4.070	3.9

3、按岗位维度融合结果

界面友好度	易学性	功能完备性	功能实用性	数据准确性	运行稳定性	响应及时性	系统兼容性	数据安全性	运维人员能力	运维人员态度	运维系统能力	服务台能力	处理流程规范性	故障处理及时性
2.667	3.000	3.000	3.333	3.333	2.333	3.000	2.000	3.667	2.667	3.333	3.000	2.667	3.333	3.00
3.750	3.500	3.500	3.500	3.750	3.500	4.000	3.750	4.000	3.500	3.750	3.500	3.500	4.000	3.75
3.303	3.455	3.364	3.394	3.667	3.121	3.424	3.303	3.758	3.394	3.545	3.424	3.303	3.455	3.24
3.615	3.462	3.500	3.577	3.538	3.500	3.615	3.500	3.654	3.462	3.615	3.538	3.269	3.385	3.42
3.613	3.373	3.600	3.625	3.569	3.512	3.488	3.387	3.739	3.694	3.303	3.164	3.133	3.297	3.39

www.zindm.com

4）感知评价

将综合评价感知记录，载入感知评价模型，即可由系统自动得到评价结果。

本例中，针对所有感知用户，基于模糊神经网络的综合评价结果为：3.5736

基于模糊神经网络各部门的综合评价结果见表 2-93。

表 2-93 各部门综合评价结果

业务部门	模型评价
管理二所	3.6054
管理一所	3.6928
调度控制中心	3.2577
基建部	3.3205
计量中心	3.5981
试验研究所	3.1388
通信设备运维部	4.0820

基于模糊神经网络各岗位的综合评价结果见表 2-94。

表 2-94 各岗位综合评价结果

岗位	模型评价
主任	3.0659
副主任	3.8789
专责	3.5322
主管	3.6630
管理员/其他	3.5727

www.tipdm.com

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的相关性分析、属性选择、RBF 径向基神经网络、FNN 模糊神经网络等算法进行模型构建。

1. 模型输入

模型的输入包括两部分，一部分是建模样本数据的输入，另一部分是建模参数的输入。

表 2-95 信息系统客户服务感知评价模型输入变量

变量分类	变量名称	变量描述
A *系统易用性*	a_1 易操作性	✧ 用户操作和使用软件的难易程度，如系统是否通俗易懂，容易使用？是否符合操作习惯？是否无需培训就能上手使用？
	a_2 界面友好度	✧ 用户界面是否美观方便，如页面布局是否清晰合理？内容主次是否分明？配色和结构是否合理？
	a_3 易学性	✧ 用户使用手册的详细程度和易理解性，以及帮助机制的完备性，如专业术语是否有说明？操作是否有提示？必填项是否有标注？
B *系统功能*	b_1 功能完备性	✧ 系统功能或功能模块是否齐全？有没有需要增加新的功能或功能模块？
	b_2 功能实用性	✧ 各功能模块是否符合实际应用需要？与实际工作的结合度高不高？
	b_3 数据准确性	✧ 系统数据是否完全准确（不包含用户误操作所造成的数据偏差）？
C *系统性能*	c_1 运行稳定性	✧ 系统是否发生过服务中断，或由于系统原因无法登陆的情况？
	c_2 响应及时性	✧ 访问系统时响应速度快不快？页面反馈结果是否及时？
	c_3 系统兼容性	✧ 系统在不同环境（操作系统、地点、客户端）下运行时是否具有较好的适应能力？
	c_4 数据安全性	✧ 使用过程中是否发生过数据丢失、数据泄露现象？有没有采取足够的防范措施？
D *运维人员*	d_1 运维人员能力	✧ 运维人员的知识、技能、经验、安全意识以及对自身业务的熟练程度如何？能否让人感到放心可靠？
	d_2 运维人员态度	✧ 运维人员的工作热情和态度如何？能否主动发现问题并积极处理？
E	e_1 运维系统能力	✧ 运维系统（如 ITSM）是否起到预想的运维效果？

www.tipdm.com

运维资源	e_2 服务台能力	✧ 服务台（如 1600 客服）是否及时受理运维服务请求，及时跟踪服务请求的处理进展？
F *运维过程*	f_1 处理流程规范性	✧ 是否有严格的规定及管理制度，以确保服务请求的接收、记录、执行、跟踪和反馈符合流程规范？
	f_2 投诉渠道畅通性	✧ 系统出现问题时，反映渠道是否多样、通畅？问题能否及时反映到相关处理人员？
	f_3 故障处理及时性	✧ 出现问题时是否能得到及时处理？处理结果是否能及时知会相关人员？
	f_4 培训满意度	✧ 培训是否达到预期效果？培训目标是否实现？对环境、时间安排、讲师技巧、培训内容等是否满意？

表 2-96 属性选择建模参数

序号	参数名称	参数描述
1	属性评估器	对属性/属性子集进行评估确定
2	搜索策略	确定搜索算法

表 2-97 FNN 神经网络建模参数

序号	参数名称	参数描述
1	进化次数	最大进化（迭代）次数
2	隐层节点数	隐含层节点数
3	修正系数	神经网络修正系数
4	学习率	神经网络学习率

表 2-98 RBF 神经网络建模参数

序号	参数名称	参数描述
1	聚类种子数	K-均值聚类种子数
2	最大迭代次数	最大训练（迭代）次数
3	最小标准差	设定允许误差最小值
4	聚类数	样本最后聚类的个数
5	岭参数	对数似然值的岭参数

2. 仿真过程

www.cnpn.com

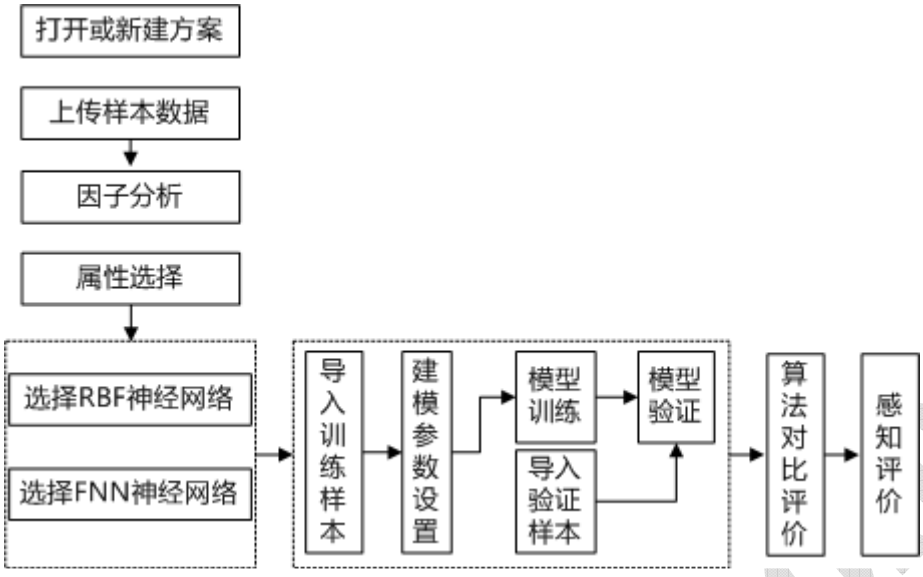


图 2-61 仿真过程

建模仿真过程说明如下：

1) 登录 TipDM 平台，在方案管理页面中，创建信息系统客户服务感知评价模型。

2) 切换到数据管理页面，上传专家样本数据文件。

3) 选择相关性分析功能，导入样本数据进行因子分析。

4) 选择属性选择功能，导入样本数据，尝试选择不同的属性评估器和搜索策略组合进行属性选择。

5) 分别选择 RBF 神经网络算法和 FNN 神经网络算法，进行模型构建。

6) 对比 RBF 神经网络算法和 FNN 神经网络算法的建模结果，并选择最优算法。

7) 用最优法对测试样本进行信息系统客户服务感知评价。

3. 结果及分析

1）属性选择

由于最近邻算法（KNN）对不相关属性非常敏感，本例中采用最近邻算法，判断客户感知（满意度评价）与系统易用性、页面美观性、帮助完整性、系统功能完备性等感知评价指标之间的关联程度。各指标相关程度见表 2-99。

表 2-99 各指标相关程度

序号	相关程度	指标名称
1	0.15911	运行稳定性
2	0.05672	响应及时性
3	0.03216	故障处理及时性
4	0.02398	功能实用性
5	0.01498	投诉渠道畅通性
6	0.01149	数据准确性
7	0.01083	数据安全性
8	0.01024	易操作性

www.tipdm.com

9	0.00791	运维人员态度
10	0.00721	运维人员能力
11	0.00721	功能完备性
12	0.00695	培训满意度
13	0.00648	服务台能力
14	0.00510	运维系统能力
15	0.00460	易学性
16	0.00371	界面友好度
17	0.00365	处理流程规范性
18	0.00251	系统兼容性

表中的排序结果也反映了各评价指标与总体评价结果的相关程度。

数据选择包括数据集中属性（列）和记录（行）的选择，通过对样本数据的相关性分析，本例将 18 项评价指标（属性）全部保留。

2）FNN 神经网络建模

模型神经网络训练后的网络结构参数：

- ✧ c：训练后的隶属度函数中心
- ✧ b：训练后的隶属度函数宽度
- ✧ mp：训练后的神经网络权系数

当 FNN 模糊神经网络的输入节点数为 18，输出节点数为 1（网络结构图见图 2-62），进化次数为 100，隐层节点数为 6，修正系统为 0.001，学习率为 0.05 时，训练步数为 100 时的均方差最小，最小值均方误差为：0.0229，建模过程训练误差图见图 2-63，训练样本和测试样本预测结果分别见图 2-64 和图 2-65。

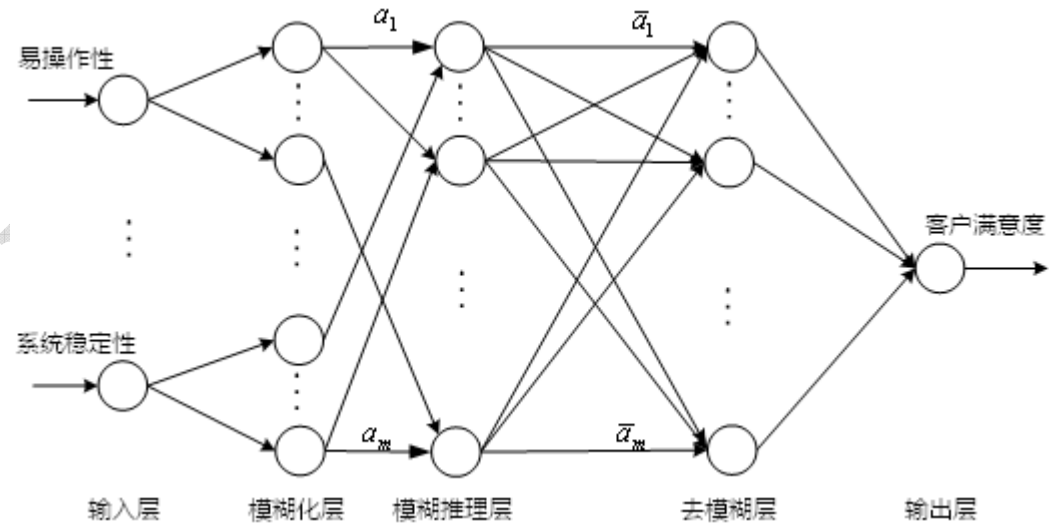


图 2-62FNN 模糊神经网络结构图

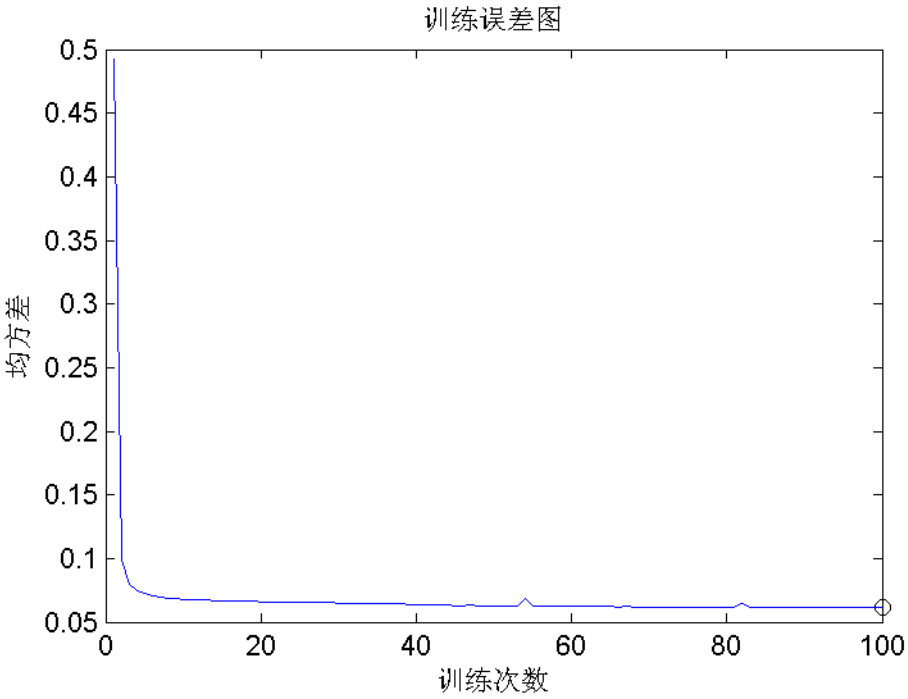


图 2-63 建模过程训练误差

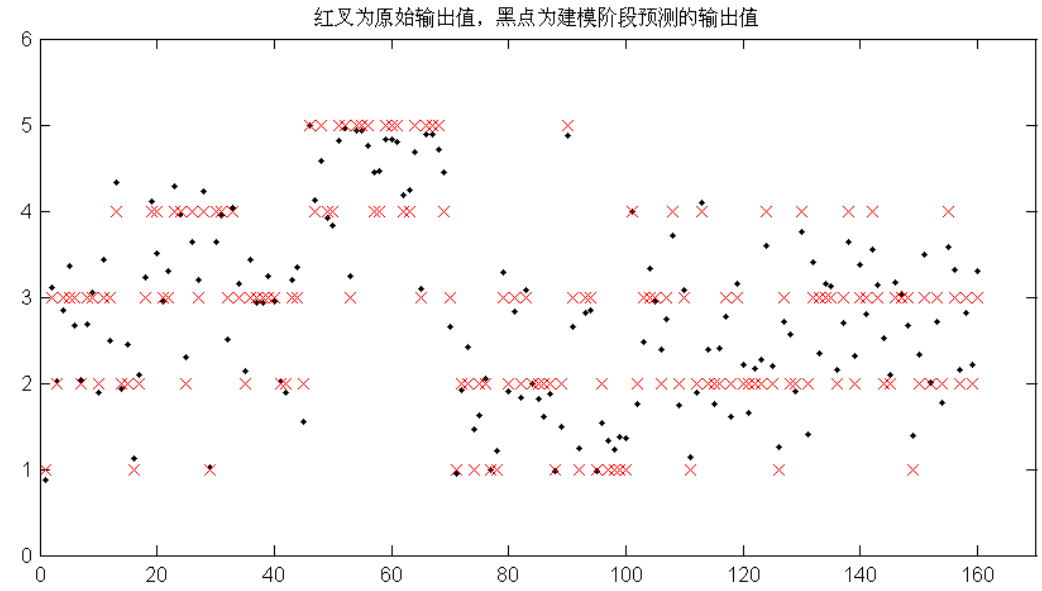


图 2-64 训练样本预测结果图

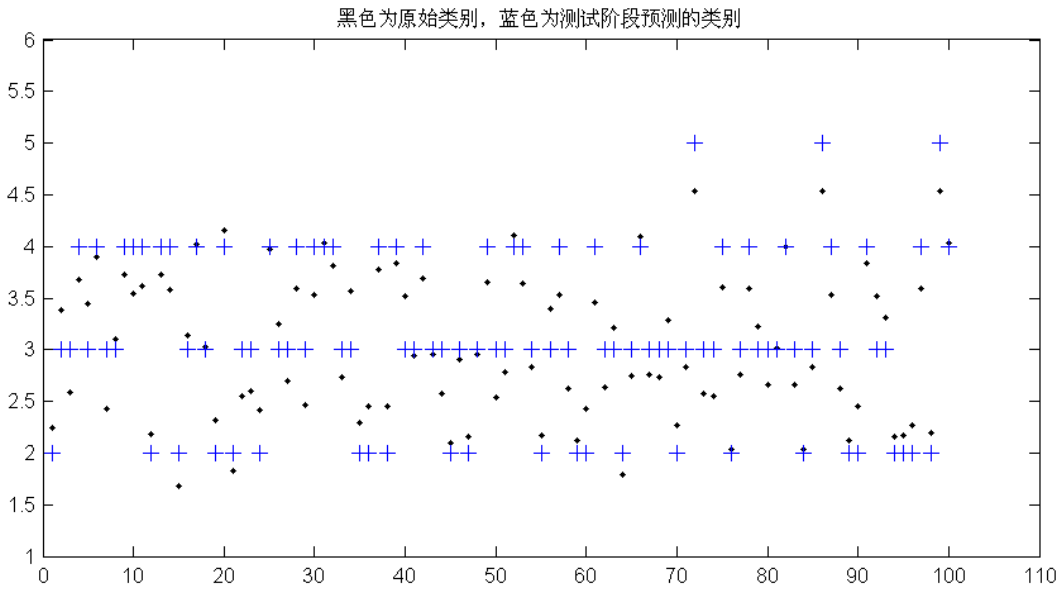


图 2-65 测试样本预测结果图

建模参数输出结果：

✧ 训练后的隶属度函数中心

c_{3*18} =

1.8001	1.5605	1.7386	0.2899	0.7020	0.1519
0.7311	0.7635	1.1550	1.6784	1.0117	0.4776
0.2202	0.4877	1.1060	1.2747	0.8108	0.2456
0.3678	0.0993	0.9817	0.8077	1.8843	0.1195
0.4789	1.7930	0.9706	0.2035	1.8896	0.4679
0.8443	1.9110	0.6870	0.2596	1.1866	0.7098
1.6427	0.3380	1.2955	0.5929	1.3736	1.2512
-0.0153	1.2711	0.8977	1.4322	0.3691	1.5496
0.1123	1.5496	1.1152	0.4100	0.7621	0.2459

✧ 训练后的隶属度函数宽度

b_{3*18} =

1.8588	0.8718	1.0170	1.5897	1.6232	1.8780
1.5527	0.9189	1.0307	1.2897	1.0670	1.7570
0.9706	0.5805	1.6327	0.7544	0.6954	1.0851
1.2450	0.6027	1.6886	0.3414	0.6219	0.3696
1.1768	0.9452	0.4168	0.4661	1.8630	1.8106
0.4050	0.4539	0.4355	0.8688	0.8189	1.9585
0.8773	0.8174	1.2057	0.2341	0.8483	0.5250
0.2901	1.2131	1.4234	0.6969	1.0161	1.6063
0.4942	0.2826	0.4256	0.6049	0.1442	-0.1829

www.tpdm.c

✧ 训练后的神经网络权系数

mp3*19=

0.3000	0.3000	0.3000	0.3000	0.2999	0.3000	0.3000
0.4467	-0.2487	0.1617	0.0041	0.0455	-0.0581	-0.0083
0.2683	0.2621	0.3082	0.2780	0.2709	0.2680	0.2783
0.3000	0.3000	0.3000	0.3000	0.3000	0.3000	0.3001
-0.0416	-0.0575	-0.0441	-0.0160	-0.0231	0.0389	0.0177
0.2723	0.2765	0.2532	0.2660	0.3217	0.2760	0.3199
0.3000	0.3000	0.3000	0.3000	0.3000		
0.0753	-0.0188	0.1407	-0.0685	0.0458		
0.3540	0.2774	0.3260	0.2799	0.2723		

www.tipdm.com

2.4. 生产制造

2.4.1. 案例十二 基于小波变换的桩基完整性检测

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

随着高层建筑、大型工程的蓬勃兴起，桩基础已在我国得到普遍采用。由于地质条件、施工技术等因素，部分桩基可能存在不同程度的缺陷，如图 2-66，图 2-67。由于桩基属于隐蔽工程，肉眼不易察觉，但它的质量直接影响地面建筑物。因此，采用恰当的方法检测桩基的施工质量，防止工程事故的发生，就成为当今桩基工程中的重要问题。



图 2-66 桩基缺陷



图 2-67 断桩处钻取的芯样图片

桩基分为钻孔灌注桩和预制桩。对于钻孔灌注桩，由于必须在地下或水下施工，且存在施工技术复杂，工艺流程相互衔接紧密，施工质量影响因素较多的实际情况，因此在施工过程中容易出现各类缺陷，影响桩基的质量完整性和单桩的承载力。资料表明，国外在现场灌注桩施工中桩身出现缺陷的概率约为 15%～20%，国内这一概率约为 20%。为了切实掌握桩基的工程质量状况，保证结构的可靠性，对尽可能多的桩基进行质量检测非常必要，以往的钻孔取芯及静力加载方法由于费用、时间等原因，只能对极少数桩基检测，已远不能满足工程发展的需求。近几年发展起来的反射波法桩基完整性测试技术以其现场测试方便快捷，试验费用低，适用情况广等优点而在工程施工中得到广泛的推广与应用。为保证工程建设质量及速度发挥了较大的作用。

反射波法桩基质量检测技术应用弹性波理论。反射波法检测桩基质量时，用重锤或其它重物敲击桩头表面，在桩顶上将产生一个类似点振源的激发信号，该振动将在桩身中以球面状辐射传播。由于桩径是有限的，由几何学知，在 1.5 倍桩径以下，波阵面将几乎与桩身垂直。因此，当振源沿桩轴向激发时，桩身中的波动可以近似为一维波动。另外，由于波沿桩轴向传播，为法向入射，因此传播过程中不会产生转换波[12]。

从弹性波理论知，当波传播至波阻抗界面时，将发生能量再分配，产生反射和透射，波的反射系数为:

$$R_{\mathrm{f}} = (\mathrm{Z}_1 - \mathrm{Z}_2)/(\mathrm{Z}_1 + \mathrm{Z}_2)$$

式中：$Z = \rho \cdot V_{\mathrm{P}} \cdot \mathrm{A}$ 为广义波阻抗，下标 1、2 分别代表界面上下介质；ρ为混凝土质量密度；V_{P}为混凝土弹性波速；A 为桩身截面积。

测试时安装在桩顶一传感器，用一重锤敲击桩顶，敲击时产生的应力波沿桩身向下传播，当下传的应力波遇到桩底或桩身缺陷界面（即波阻抗分界面）处，即有部分应力波被反射，反射波再被桩顶的传感器接收，经过对接收波形进行一定的计算分析，即可判定桩基的质量完整性。反射波法桩基质量检测现场如图 2-68。



图 2-68 反射波法桩基质量检测现场

采集到反射波法信号后，如何对信号波形进行分析，有效判断出桩身缺陷位置和类型？

分析方法与过程

桩身中的缺陷大致分为两种类型：一种是结构上的缺陷，如裂纹，缩颈，断裂，扩底等；另一种是材料性质的变化引起的缺陷，如混凝土离析，夹泥，桩底松散等。这些缺陷的检测统称为桩基完整性检测。检测桩基完整性的方法很多，一般采用无损检测方法，例如声脉冲反射波法。该方法基于桩身中的缺陷会引起波阻抗的改变这一原理，用一手锤在桩顶激发出一声脉冲，利用安装在桩顶的加速度计接收反射波信号，在此基础上，通过小波变换把反射波信号分解到不同层次不同频带的序列，并总结出特征量与桩基缺陷类型对照表。然后，将模糊综合评判和径向基函数网络有机结合在一起，建立基于小波变换特征提取的模糊神经网络诊断模型，最终达到识别桩身缺陷出现的位

www.ibdn.com

置和类型的目的，图 2-69 和图 2-70 分别为完整桩和局部存在缺陷桩的反射波形特征。

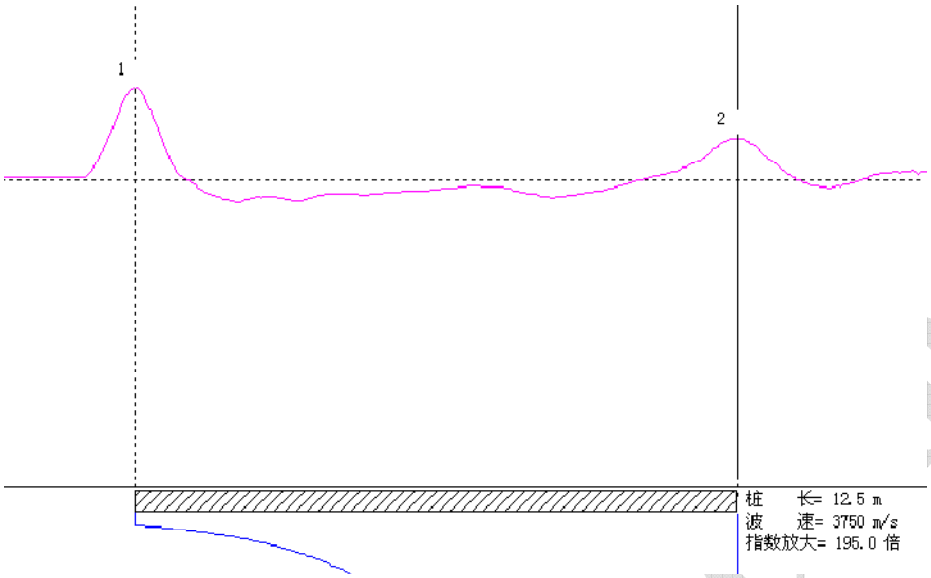


图 2-69 完整桩反射波形特征

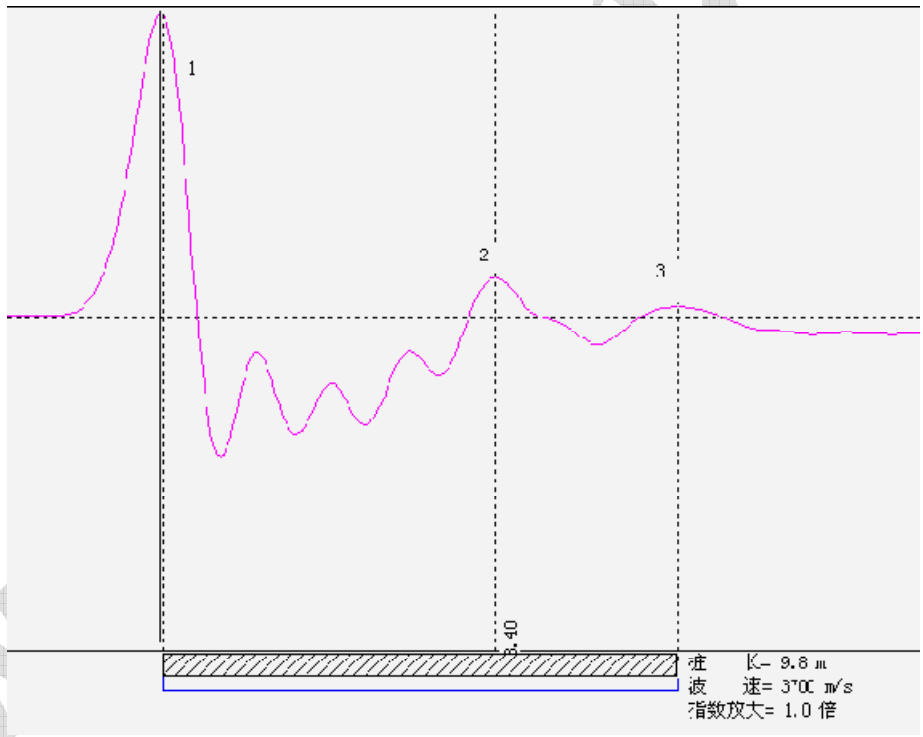


图 2-70 局部存在缺陷桩反射波形特征

图 2-71 为桩基完整性检测判别流程图。

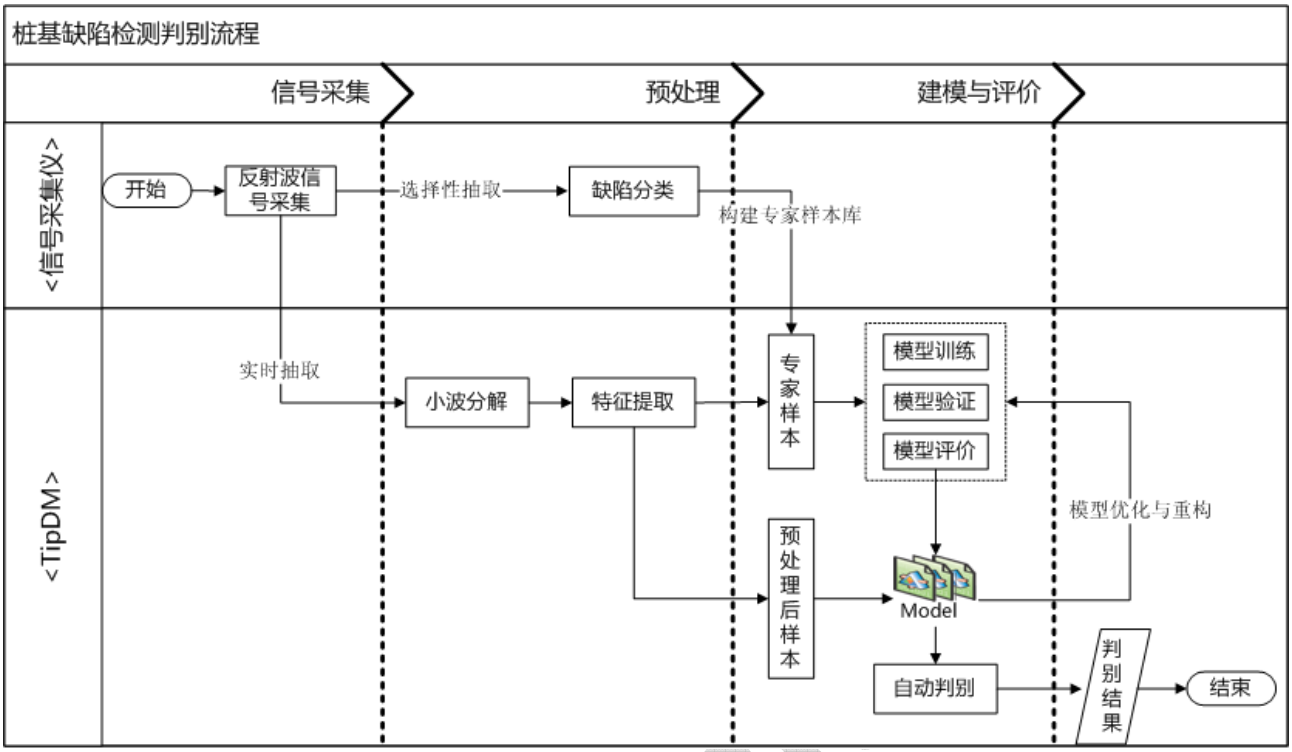


图 2-71 桩基完整性检测判别流程

1. 建模样本数据采集

本案例采用声脉冲反射波法，信号采集装置见图 2-72。试验时使用的手锤重量为 0.6kg，利用加速度计得到桩顶加速度响应曲线。



图 2-72 声脉冲反射波法信号采集装置

试验桩共 26 根，尺寸为 1.5×0.15×0.15m^3,分别模拟缺陷如下：

- ✧ 正常桩：桩身完整，无缺陷。
- ✧ 扩底桩：桩底部形成一个扩大头，上部突出截面高的 30%，两边扩大部分近一根桩厚。
- ✧ 离析：在离析处采用低强混凝土，按 C10 配制。
- ✧ 缩颈：缩颈处混凝土下凹达 30%截面高。
- ✧ 裂缝：在离桩头 20cm 位置，预制达 30%截面高的裂缝。
- ✧ 两扩：将扩颈处混凝土上部突起达 30%截面高，缩颈处下凹达 30%截面高。
- ✧ 桩底松散：桩底松散处采用强度低的混凝土，按 C10 配制且振捣不够密实。

试验时用力锤在桩头处施加一瞬态脉冲激励，使桩身产生压缩应力波 ，应力波沿着桩身自上而下传播，利用在桩头安装的传感器接收由初始信号和桩身缺陷或桩底产生的反射信号组合的时程曲线。

对时域内桩基检测结果进行波形的分析和比对是反射波法低应变动载检测中最重要的一环，它直接影响到对桩身完整性的判定和缺陷性质位置的确定。因此，在分析过程中要结合基桩施工纪录、地质情况对波形特征进行全面的分析，反复比对，认真研判，以确定基桩缺陷的类型和位置，判定桩身完整性。

本案例通过试验桩进行数据采集，从模型试验桩的外观很容易将不同缺陷的类型区分开来。然而由常规的声脉冲反射波法得到的速度时程曲线却较难进行区分。图 2-73 分别显示各种不同缺陷的桩顶反射波信号(速度响应)。

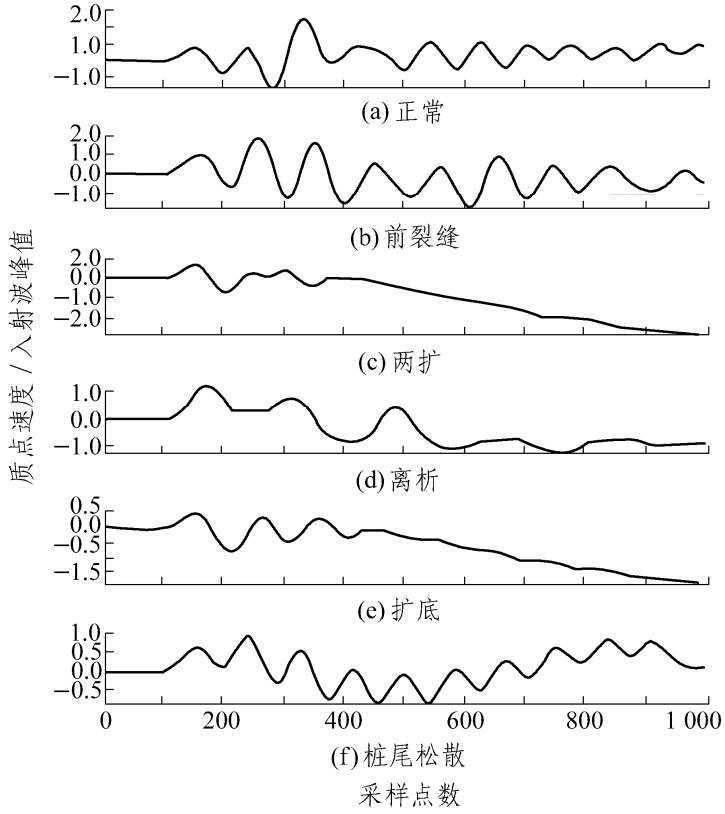


图 2-73 不同缺陷桩顶反射波信号

2. 小波分解及特征提取

傅立叶变换是研究超声回波信号频域特性的常用方法。然而，经典的傅立叶分析只能提供时间窗内信号总体频谱分析的信息。对如波阻抗变化不太大而引起的微弱反射信号，回波信号的傅立叶变换较难体现缺陷的特征信息。小波分析方法对信号的高频部分使用短的时间窗，低频部分使用长的时间窗，利用这种时-频分析工具来研究桩中缺陷回波信号的频域分布特征可望得到良好的结果。

对于一个多分辨分析$\{V_k\}$，W_k 是V_k 关于V_{k+1} 的补空间，对于 $f_N(x)\in L^2(R)$ 都有唯一分解 $f_N(x)=g_{N-1}(X)+g_{N-2}(X)+\cdots+g_{N-M}(X)+f_{N-M}(X)$。因此，分解技术能够把非平稳振动信号映射到由一个小波伸缩而成的一组基函数上，信息量完整无缺，在通频范围内得到分布在不同层次不同频带内的分解序列。小波分解过程如图 2-74 所示，F0 代表低频系数，F1 代表高频系数。根据不同类型缺陷在不同频带内能量分布的特点，选取图 2-74 中 x0, x1, x2, x3, x4, x5 作为特征量。

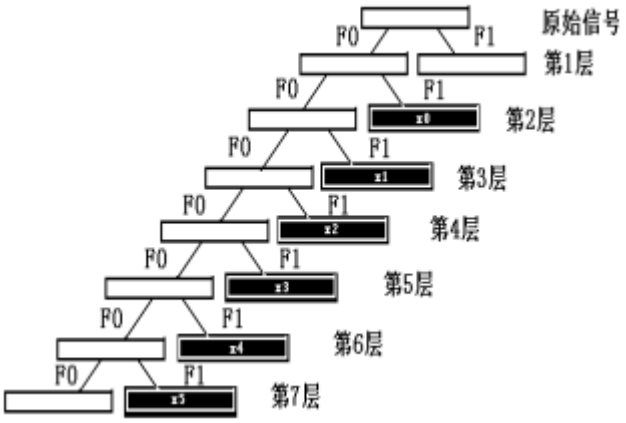


图 2-74 桩顶应力波信号的小波分解

脉冲信号 $f(t)$ 的积分小波变换定义为：$W_f(a,b)=<f,\varphi_{a,b}>=|a|^{-\frac{1}{2}}\int_{-\infty}^{+\infty}f(t)\varphi\left(\frac{t-b}{a}\right)dt$，其中 a 是尺度函数，b 是位置系数。小波函数 $\varphi_{a,b}(t)=|a|^{-\frac{1}{2}}\varphi\left(\frac{t-b}{a}\right)$ 由母小波函数 $\varphi(t)$ 平移、伸缩而得到。小波变换值 $W_f(a,b)$ 反映的是信号 $f(t)$ 在一个与 a、b 有关的局部时区内，关于一个与 a 有关的频带上的局部时-频信息特征，当参数 a、b 取定各种不同值，就可以用对小波变换 $W_f(a,b)$ 的分析来取代对信号 $f(t)$ 直接的分析，从而能够提取出信号 $f(t)$ 的更深刻的性质与特征。可见，小波变换的结果反映了 $f(t)$ 在尺度 a(频率)和位置 b(时间)的状态，所以它是一种时-频变换。

图 2-75 给出了图 2-73 中前裂缝缺陷加速度信号的时间-尺度平面上的信息(考虑到桩长及应力波在桩中传播的特性，在此仅取前 500 个点进行分析)。模型试验信号的采样频率为 100 kHz，因此可以认为该信号中包含了 0～50 kHz 频率范围内的振动成分。小波分解采用 db10，分解到第 7 层，得到 8 个频率通道 w1，w2，w3，w4，w5，w6，w7 和 s7，它们大体包含频率范围[25000, 50000]，[12500, 25000]，[6250, 12500]，[3125, 6250]，[1562.5, 3125]，[781.25, 1562.5]，[390.62, 781.25]和[0, 390.62]的成分。从图 2-75 可以看出，无论哪种缺陷，桩顶处的反射信号能量主要分布在大尺度。但是，值得注意的是不同类型缺陷的反射波信号在各尺度上的分布是不同的，即在频域上分布不同。因此 w5，w6 和 w7 应是我们分析的主要通道。表 2-100 列出了各缺陷在不同尺度上的能量分布，其具体数值是能量与入射波能量的比值。从表 2-100 的分析结果不难发现，不同缺陷类型的桩在不同尺度(频带)内的能量分布存在差异，根据这一特征量即可准确地确定缺陷的类别。

www.tjpu.com

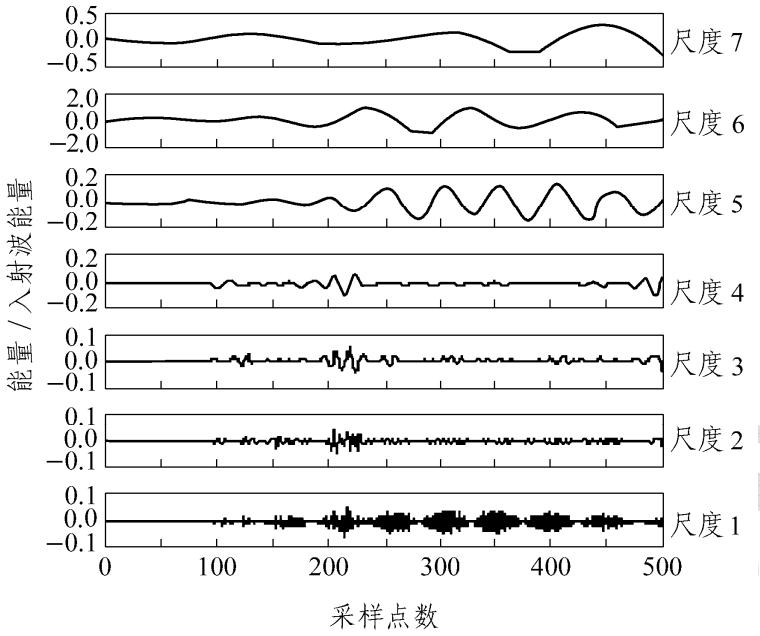


图 2-75 前裂缝缺陷的反射波信号的时-频细节信息

表 2-100 不同缺陷在不同尺度上的特征量分布范围

缺陷类型	X5	X4	X3	X2	X1	X0
正常	0.28~0.32	0.92~0.93	0.18~0.22	0.02~0.05	0.03~0.065	0.02~0.031
扩底	0.35~0.43	0.87~0.9	0.19~0.22	0.06~0.08	0.05~0.07	0.04~0.06
离析	0.6~0.79	0.46~0.65	0.09~0.18	0.01~0.07	0.01~0.07	0.04~0.07
缩颈	0.68~0.76	0.05~0.061	0.11~0.37	0.02~0.09	0.02~0.07	0.02~0.05
裂缝	0.47~0.63	0.54~0.83	0.16~0.4	0.04~0.09	0.03~0.09	0.01~0.06
两扩	0.5~0.6	0.75~0.85	0.16~0.26	0.03~0.10	0.02~0.06	0.02~0.05
桩底松散	0.10~0.12	0.94~0.95	0.22~0.26	0.03~0.045	0.03~0.04	0.03~0.04

仿真过程

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 RBF 径向基神经网络、贝叶斯分类等算法进行模型构建。

表 2-101 朴素贝叶斯建模参数

序号	参数名称	参数描述
1	核估计	核密度估计
2	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果

表 2-102 贝叶斯网络建模参数

序号	参数名称	参数描述
1	评估器	评估函数
2	搜索策略	搜索网络结构的方法
3	ADTree 算法	是否选用 ADTree 算法
4	预测输出	选择“True”，当交叉验证和模型训练时，将会输

		出样本的预测结果

表 2-103 RBF 神经网络建模参数

序号	参数名称	参数描述
1	聚类种子数	K-均值聚类种子数
2	最大迭代次数	最大训练（迭代）次数
3	最小标准差	设定允许误差最小值
4	聚类数	样本最后聚类的个数
5	岭参数	对数似然值的岭参数

下面主要介绍基于 RBF 径向基神经网络实现建模过程，贝叶斯分类算法建模请用户尝试构建并比较建模效果。

1. 模糊径向基函数网络模型的建立

设 a，b 为相应特征分量 x_i 的允许范围，那么 x_i 通过下式的隶属函数取其隶属度[5]：$A(x)=\frac{1}{2}+\frac{1}{2}\sin\frac{\pi}{a_2-a_1}\left(x-\frac{a_1+a_2}{2}\right)$，$a_1<x\le a_2$。这样就构成了缺陷模糊向量 $A=(a_0,a_1,a_2,a_3,a_4,a_5)^T$。根据小波分解原理，总结出特征量分布与桩基缺陷对照表，建立了桩基缺陷的模糊关系矩阵 $\widetilde{R}$，如表 2 所示。将缺陷模糊向量 A 与模糊关系矩阵 $\widetilde{R}$ 进行模糊运算，得出反映缺陷严重程度的缺陷隶属度 B：$B=\widetilde{R}A$。根据模糊评判要求，综合考虑各特征分量对某一缺陷的影响，因此采用“与 1 环和”计算模型进行计算，则有：

$$b_i=\min\left(1,\sum_{j=0}^{5}a_j r_{ij}\right)$$

表 2-104 小波分解特征量与桩基缺陷对照表

缺陷类型	X5	X4	X3	X2	X1	X0
正常	0.2899	0.9276	0.2148	0.0488	0.0358	0.0308
扩底	0.3823	0.8939	0.1972	0.0652	0.0572	0.0489
离析	0.6704	0.6336	0.1707	0.0698	0.0682	0.0629
缩颈	0.7180	0.5920	0.3014	0.0283	0.0544	0.0383
裂缝	0.6269	0.6388	0.2230	0.0693	0.0442	0.0357
两扩	0.5077	0.8403	0.1645	0.0355	0.0226	0.0216
桩底松散	0.1059	0.9496	0.2267	0.0435	0.038	0.0394

基于模糊径向基函数网络原理，建立了一个五层模糊网络，如图 2-76 所示。网络的输入输出值为：(X_i,B_i,Y_i)，其中对应于第 j 个缺陷 Y_i 的第 j 个分量为 1，即 $Y_i=(0,0\ldots,1,\ldots,0)$。该网络由两个径向基函数子网络组成，第一层为模糊评判子网络，第二层为诊断子网络。选取值最接近于 1 的 y_{ij} 所对应的缺陷隶属度为该缺陷的置信度。

www.ijdm.com

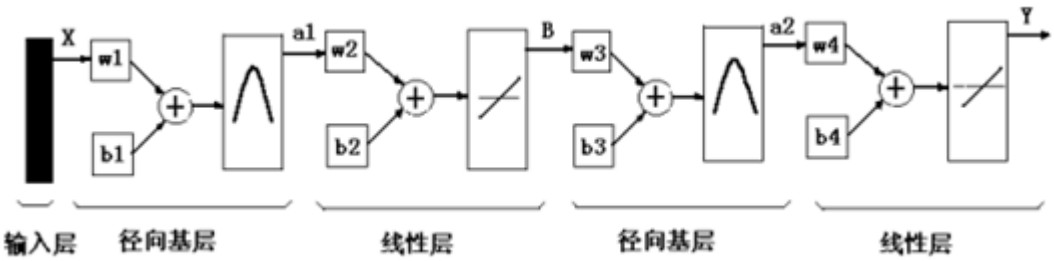


图 2-76 模糊径向基函数网络结构图

2. 结果仿真

用实验获取的典型缺陷样本对该网络进行训练，网络不仅能准确识别已训练的缺陷，还能在很短的时间内对未知缺陷做出预测，并给出某种缺陷的置信度。例如，缺陷为桩底松散的实验桩的脉冲信号经小波分解后，将其特征值作为一组样本训练网络。因为 y_{ij} =0.9642，接近于 1，故网络识别出缺陷为第七类—桩底松散，置信度为 97.85%。实验仿真结果见表 2-105 实验仿真结果。

表 2-105 实验仿真结果

	正常	扩底	离析	缩颈	裂缝	两扩	桩底松散
置信度 B	1.0000	0.9491	1.0000	1.0000	0.4389	0.9286	0.9785
输出 Y	0.0817	0.0485	0.1748	0.1147	0.2180	0.0503	0.9642

2.4.2. 案例十三 基于水色图像的水质评价

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

水质评价是根据养殖水体当前的水质状况、污染物的迁移特性以及流域内污染物的排放情况来预测水质未来的变化趋势，以起到防患于未然的作用。在水产养殖中，由于池塘和湖泊中的水质环境影响因素很多，影响因子与水质评价之间关系复杂且是非线性的，因此应从满足水产养殖的需求出发，借鉴环保领域水质智能监测技术，构建适宜的水质智能监测方法和评价技术，影响水质状况的指标主要有：

1） 溶解氧：衡量水体的自净能力（传感器网络自动采集）
2） 温度：水温随着天气的变化（传感器网络自动采集）
3） PH 值：反映水质的酸碱程度（传感器网络自动采集）
4） 氨氮：代表水中营养性污染物的含量（手持传感器手动采集）
5） 氧化还原电位：水溶液氧化还原能力的测量指标（手持传感器手动采集）
6） 当前环境因素（天气，水体环境状况）

通过对温度、pH 值、氨氮、溶解氧和氧化还原电位等通过在线监测得到的水质指标，结合现有水产养殖的水质标准和不同水质条件对水生生物生理活动影响程度的基础研究，研究各个水质指标

对水质总体评价的影响关系，建立水体环境质量评价量化模型，实现池塘水质智能化识别。

由于常规量化指标采集有一定难度，因此水质评价除通过以上常规量指标进行评价外，通过水色判别也是一种常用的方法[16]。

由于水色能反映水中浮游植物的种类和多少，所以利用图像处理技术对水色分析后，可以得出水质类型。目前这种方法广泛应用在海洋和湖泊的水质预测上，通过拍摄图像，根据图像的颜色，可快速反演推出水质的好坏。图 2-77 为原始水样图像。



图 2-77 原始水样图像

如何通过水色图像实现对水质的自动评价？

分析方法与过程

通过对拍摄的水样图像，选取其中有代表性的一部分，再利用图像特征提取算法，提取出水样图像的特征值，即建立水样图像与反映该图像特征的数据信息关系，同时利用专家对水样图像的分类处理结果，建立水样/数据信息与水质类别的专家样本库，通过分类算法进行预测建模，得到水样图像与水质类别的映射关系，最后利用训练好的预测模型，用户就能方便地通过水样图像，自动判别出该水样的水质类别。图 2-78 为基于水色图像特征提取的水质评价流程。

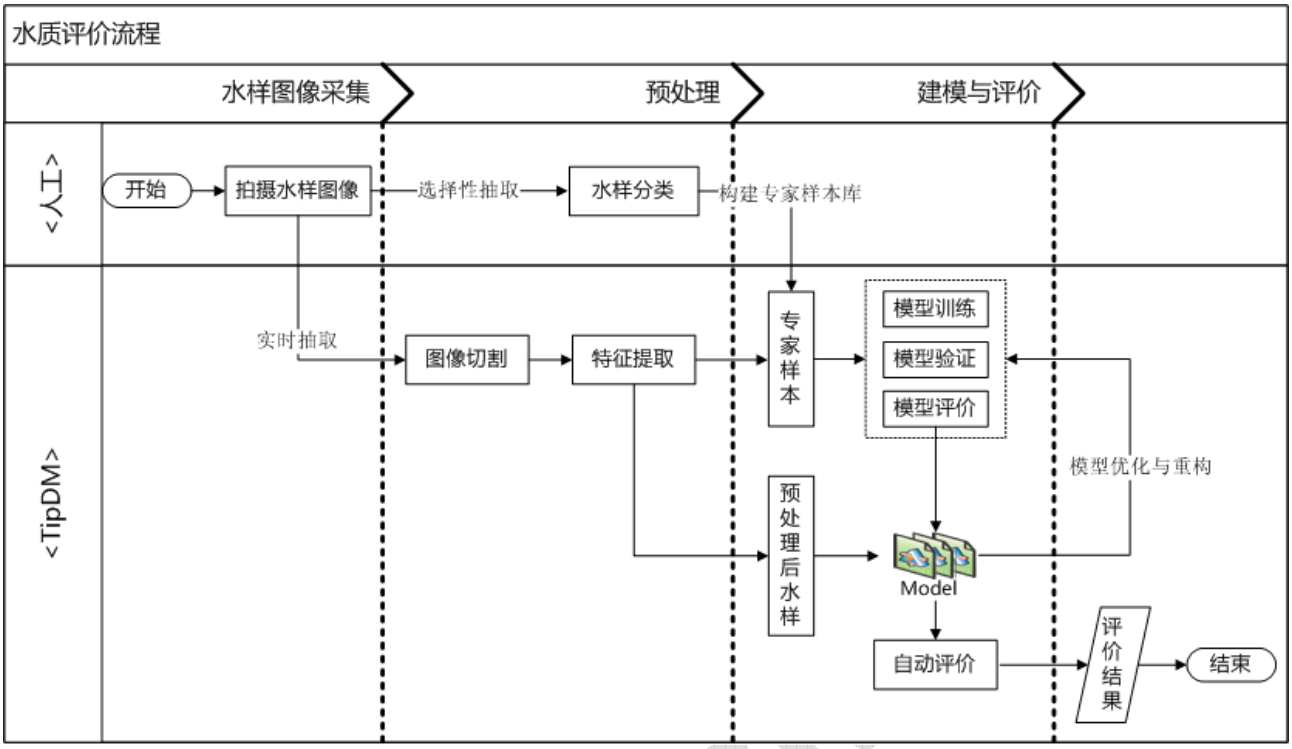


图 2-78 基于水色图像特征提取的水质评价流程

基于水色判别的水质评价流程，主要包括以下几个步骤：

4. 水样图像采集

在标准光照条件下，用标准像素的数码相机对养殖水体进行拍摄，图 2-79 为在标准条件下拍摄的养殖水体图像。

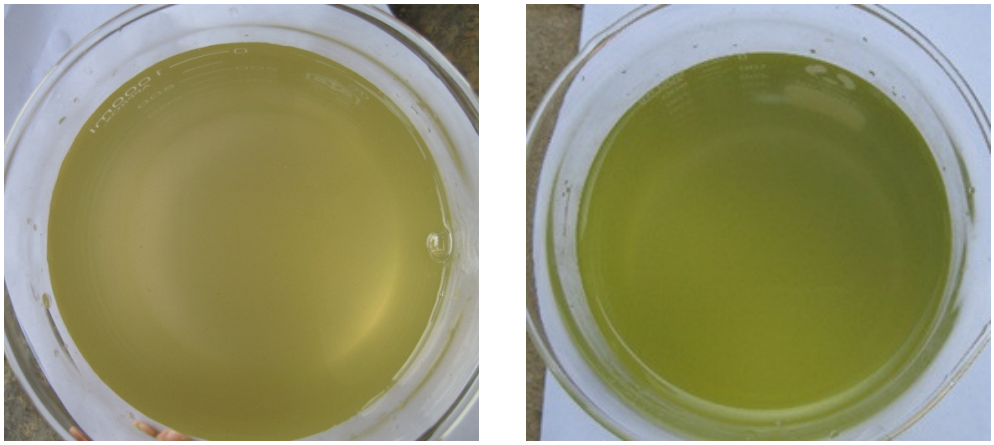


图 2-79 标准条件下拍摄的采样图像

5. 水样分类

浮游生物的量反映池塘的水色和水的肥瘦程度，同时也是衡量水质的定性的标准。表 2-106 为水色分类及需采取的措施。

表 2-106 水色分类

水色	浅绿色（清水或	黄褐色	灰蓝色	茶褐色	绿色水（黄绿、

	浊水）			（姜黄、茶褐、红褐、褐中带绿等）	油绿、蓝绿、墨绿、绿中带褐等）
水质类型	瘦水	肥水	老水	优质水Ⅰ硅藻水华	优质水Ⅱ：绿藻水华
采取的措施	对于清水的措施：引入藻种、施肥，改善水质。对于浊水的措施：净水、引入藻种、施肥，改善水质	经常加注新水，控制水质，使鞭毛藻占绝对优势。或者利用净水剂净水、引入藻种、施肥，改善水质	这种水质既不利于鱼类生活，也无法为鱼虾提供优质的饵料生物，必须及时更换新水	主要含硅藻、圆筛藻、舟形藻等，养虾的最佳水色，尽量保持水色持续稳定。	应控制藻类的过度繁殖。少量生石灰、熟石灰调节水质。

水样图像采集回来后，可由有经验的专家对水样图像根据经验进行分类。

6. 图像切割

由于采集的水样图像并非由均匀的水体构成，本系统的水色图像中包含了盛水的容器，在进行水色特征提取前，要对水色图像进行预处理，也就是需要通过图像切割方法，从图像中取出有代表意义的一小块水样图像作为真实的水色样本。

7. 特征提取

特征提取是指根据图像特征，提取反映图像本质的一些关键指标，以达到自动进行图像识别或分类的目的。显然，图像特征提取是图像识别或分类的关键步骤，图像特征提取的效果如何直接影响到图像识别和分类的好坏。

图像特征主要包括有颜色特征、纹理特征、形状特征、空间关系特征等。同几何特征相比，颜色特征更为稳健，对于物体的大小和方向均不敏感，表现出较强的鲁棒性。本系统中由于水色图像是均匀的，故主要关注颜色特征。

颜色特征是一种全局特征,描述了图像或图像区域所对应的景物的表面性质。一般颜色特征是基于像素点的特征，此时所有属于图像或图像区域的像素都有各自的贡献。在利用图像的颜色信息进行图像处理、识别、分类的研究中，在实现方法上已有大量的研究成果，主要采用颜色处理常用的直方图法和颜色矩方法等。

颜色直方图是最基本的颜色特征表示方法，它反映的是图像中颜色的组成分布,即出现了哪些颜色以及各种颜色出现的概率。其优点在于它能简单描述一幅图像中颜色的全局分布，即不同色彩在整幅图像中所占的比例，特别适用于描述那些难以自动分割的图像和不需要考虑物体空间位置的图像。其缺点在于它无法描述图像中颜色的局部分布及每种色彩所处的空间位置，即无法描述图像中的某一具体的对象或物体。

基于颜色矩提取图像特征的数学基础在于图像中任何的颜色分布均可以用它的矩来表示。根据概率论，一个概率分布可以唯一的由矩来表征。一副图像的色彩分布也可认为是一种概率分布，所以颜色分布也可由矩来决定。存储一副图像的每个颜色通道的前三阶矩的方法来提取物体特征，对于一副 RGB 或 HSI 格式的图像，仅需计算 9 个矩值。每个颜色通道的第一阶矩是图像的色彩均值，第二阶中心矩是图像的方差，第三阶矩是图像每个通道的偏度。为了让矩值具有相同的单位，可以使用图像的标准差和偏度的三次方根。

颜色模型是颜色的表示方法，可以分为面向硬件和面向视觉感知两种。面向硬件的颜色空间包括 RGB，CMY 等。其中 RGB 是现在数字图像广泛采用的颜色空间. 这些颜色空间适合于诸如彩色显示器或打印机之类的硬件设备。面向视觉感知的颜色空间包括 HSI，HCV，HSV，HSB 等，这些颜色空间与人的感觉相符合。本系统采用了 HSI 颜色空间，它由 3 个分量组成，分别代表色调 H(hue) 、饱和度 S(saturation) 、密度 I(intensity)。

本例由于水色图像是均匀的，故主要关注颜色特征。并采用颜色处理方法中常用的颜色矩方法，提取颜色特征值，表 2-107 为经特征提取后的水色图像特征与相应的水色类别。

表 2-107 水色图像类别与特征值

特征值 1	特征值 2	特征值 3	类别
0.6059	0.1378	123.5384	Ⅰ类
0.4821	0.5243	101.1183	Ⅰ类
0.4544	0.5489	121.3445	Ⅱ类
0.4713	0.5073	104.6833	Ⅱ类
0.6882	0.4891	56.12670	Ⅲ类
0.5142	0.5746	111.0206	Ⅲ类
0.5096	0.7725	104.4392	Ⅳ类
0.4988	0.2264	124.9185	Ⅳ类
0.475	0.5439	108.4268	Ⅰ类
0.4792	0.4080	113.0219	Ⅰ类
0.4951	0.3574	103.6713	Ⅱ类
0.5425	0.2249	107.4110	Ⅱ类
0.4896	0.4443	118.5889	Ⅲ类
0.4802	0.3722	135.6017	Ⅲ类
0.4576	0.5362	112.7714	Ⅳ类
0.4610	0.5787	100.1307	Ⅳ类
0.5315	0.4340	86.3454	Ⅰ类

水色图像特征提取过程如图 2-80。

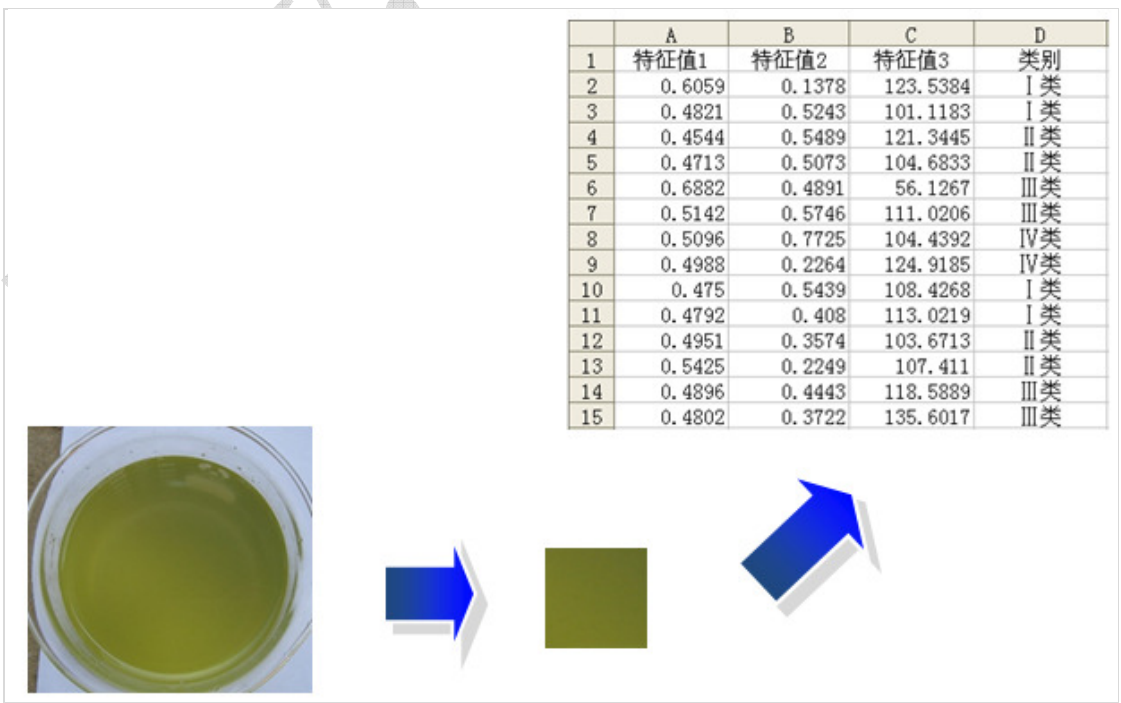


图 2-80 水色图像特征提取过程示意图

8. 预测建模

利用第（1）、（2）步生成的水色图片的特征值及对应水质类别作为训练样本，建立水质预测模型。

9. 水质评价

取需要预测的水色图片的特征值作为输入样本，送入到第（3）步中所获得的预测模型，得到输出结果，即预测的水质类型。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 BP 神经网络和 LM 神经网络优化算法进行模型构建。

1. 模型输入

模型的输入包括两部分，一部分是建模专家样本数据（包括训练样本和验证样本）的输入，另一部分是建模参数的输入。

表 2-108 预测模型输入变量

序号	变量名称	变量描述	取值范围
1	特征值 1	图像特征提取特征值 1	0~1
2	特征值 2	图像特征提取特征值 2	0~1
3	特征值 3	图像特征提取特征值 3	>1
4	水质类别	不同类别能表征水中浮游植物的种类和多少	Ⅰ类~Ⅳ类

表 2-109 BP 神经网络建模参数

序号	参数名称	参数描述
1	隐含层数	一个正整数序列，对每一个隐含层用逗号分隔的数字来表示，如果没有隐含层，则设置为 0。理论证明任意函数可以被一个有三层单元的网络以任意精度逼近，所以隐含层数为 1~2 层就足够了，一般为 1 层。默认值 a 表示由系统自动给定。
2	学习速率	网络权重更新的加权因子。学习速率设得太小，网络收敛速度慢，而且可能会收敛于局部最小点，学习速率设得太大的话，可能会导致网络不稳定
3	学习动量	权重更新量与上次权重更新量之间的关系，$\Delta\omega_{ji}(n)=\eta\delta_j x_{ji}+\alpha\Delta\omega_{ji}(n-1)$，其中$\eta$为学习速率，$\alpha$为学习动量
4	属性归一化	是否对属性（包括名词性属性和数值型属性）进行归一化处理，归一化之后的范围为[-1,1]
5	数值归一化	是否对数值型属性归一化至[-1,1]
6	种子数	设置随机函数的种子，用于随机产生初始权值和设定训练数据的顺序
7	训练次数	模型训练时最大循环（迭代）次数，当训练时循环次数达到此参数时，网络训练将会终止

www.tipdm.com

表 2-110 LM 神经网络建模参数

序号	参数名称	参数描述
1	显示间隔次数	两次显示时间的间隔
2	最大循环次数	模型训练时最大循环（迭代）次数，当训练时循环次数达到此参数时，网络训练将会终止
3	目标误差	网络的训练目标误差
4	学习动量	权重更新量与上次权重更新量之间的关系，$\Delta\omega_{ji}(n)=\eta\delta_j x_{ji}+\alpha\Delta\omega_{ji}(n-1)$，其中$\eta$为学习速率，$\alpha$为学习动量
5	学习速率	Marquart 调整参数，即网络权重更新的加权因子。学习速率设得太小，网络收敛速度慢，而且可能会收敛于局部最小点，学习速率设得太大的话，可能会导致网络不稳定
6	学习速率增加比率	mu 的上升因子
7	学习速率减少比率	mu 的下降因子
8	最大误差比率	mu 的最大值
9	隐层节点数	隐层节点个数
10	输入层至隐含层传递函数	输入层至隐含层传递函数，通常为双曲正切 S 型函数
11	隐含层至输出层传递函数	隐含层至输出层传递函数，通常为线性函数

2. 仿真过程

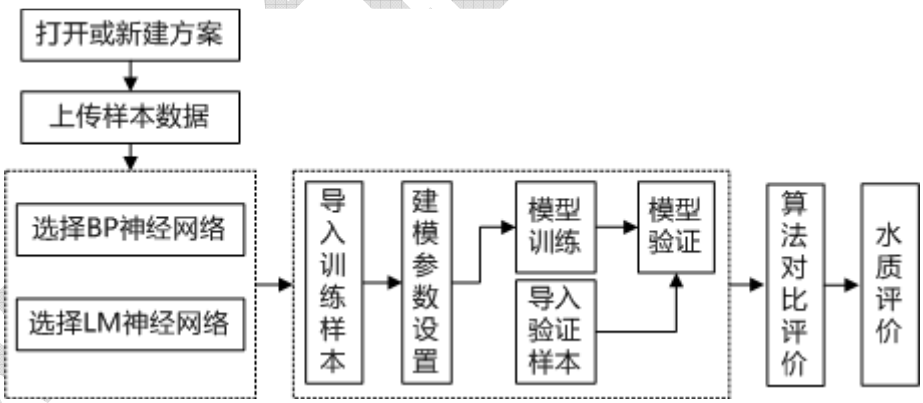


图 2-81 仿真过程

建模仿真过程说明如下：

1） 登录 TipDM 平台，在方案管理页面中，新建方案或者打开一个已建方案。

2） 切换到数据管理页面，上传经水色特征提取得到的专家样本数据文件。

3） 分别选择分类与回归->神经网络菜单中的 BP 神经网络及 LM 神经网络算法。

4） 分别用 BP 神经网络及 LM 神经网络算法对导入的专家样本数据进行模型构建。

5） 对 BP 神经网络和 LM 神经网络的建模结果进行比较。

6） 选择最优算法进行水质评价。

3. 结果及分析

当 LM 神经网络的输入节点数为 3，输出节点数为 1，隐层节点为 4，隐层节点数为 6，显示间隔次数为 25，最大循环次数为 100，目标误差为 0.02，学习速率为 0.001，学习动量为 0.001，学习速率增加比率为 10，学习速率减少比率为 0.1，最大误差比率为 10000000 时，训练完成后最小值均方误差为：0.09812，建模过程训练误差图见图 2-82，训练样本和测试样本预测结果分别见图 2-83 和图 2-84。

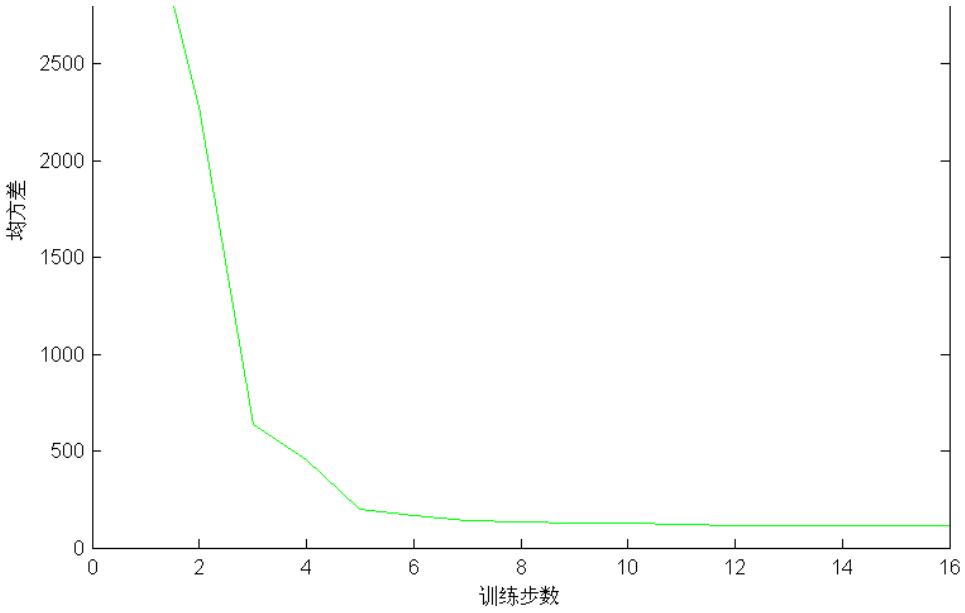


图 2-82 训练误差图

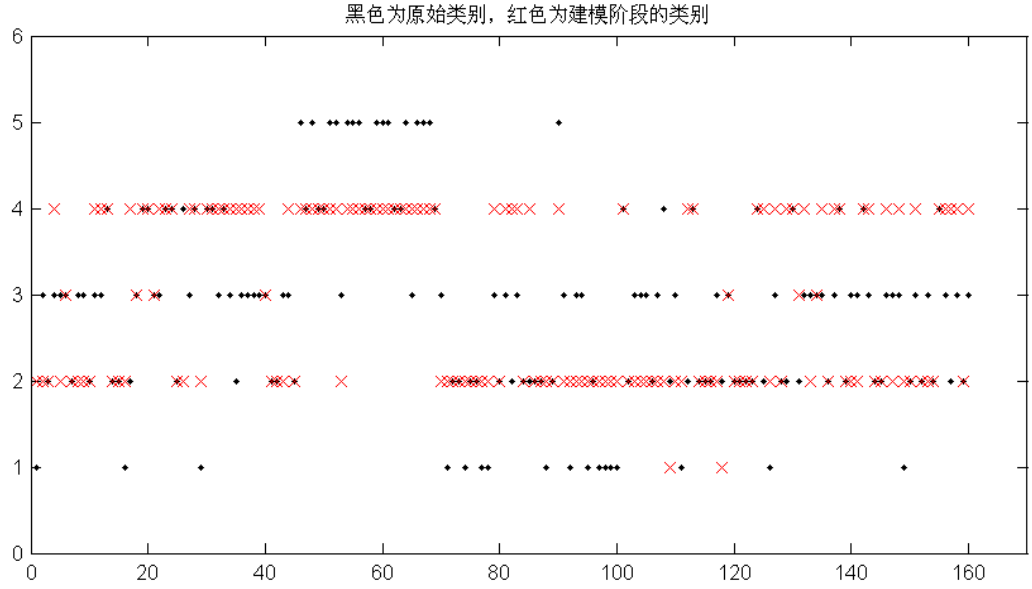


图 2-83 训练样本预测结果图

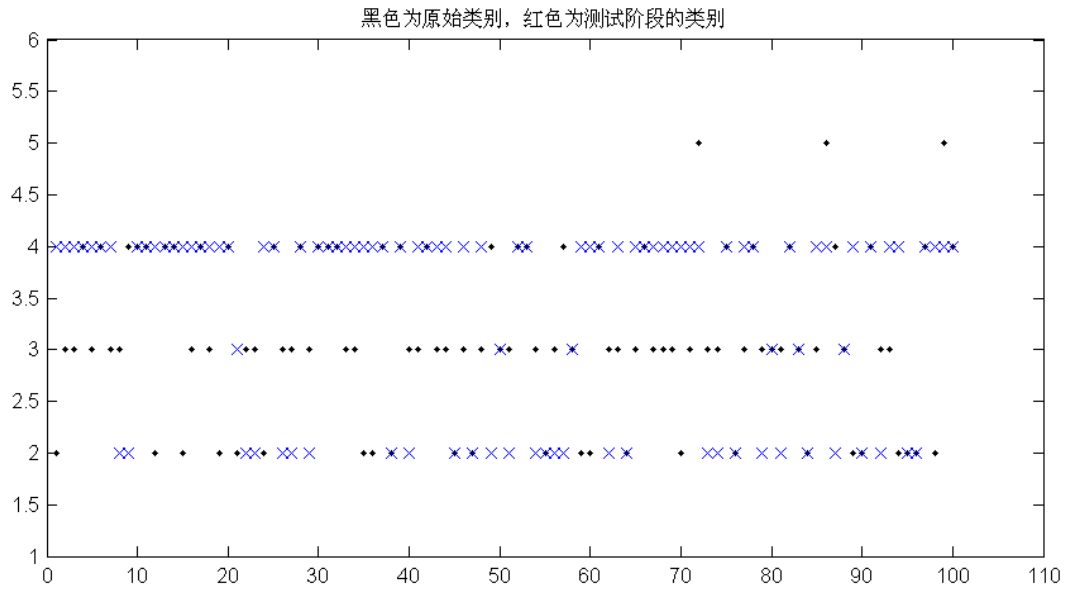


图 2-84 测试样本预测结果图

2.4.3. 案例十四 生物质废物混合厌氧消化在线建模

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

城市化进程加快，这是我国目前经济迅猛发展的显著标志，城市崛起需要消耗大量的钢材、水泥、电力等资源，加上目前我国能源储量逐步下降的现状，能源短缺成为必然。能源需求的持续增长对能源供给造成了很大压力；而可再生能源、清洁能源、替代能源等技术的开发相对滞后，节能降耗、污染治理等技术的应用还不广泛，这将导致我国在相当长一段时间内处于能源短缺中。因此，在能源短缺的今天，采用厌氧消化技术处理城市生活垃圾以及畜禽养殖废物等生物质废弃物，利用其高含水率与高有机质含量的特性，将产生的沼气等能源加以有效利用，不仅能够缓解能源短缺的压力，同时也减少了 CO2、CH4 等温室气体的排放[15]。

在厌氧消化技术中，两种或者两种以上有机废物的混合厌氧消化，可以更好地调节微生物所需的营养，显著提高消化器的容积产气率。

如果采用厌氧消化技术来处理单一种类的可生化废弃物，要调节到适宜的固含量，就必须要加入水分。然而，如果要采用厌氧消化技术，来处理污水处理厂的剩余污泥，或者是畜禽养殖废物，

由于它们含水率比较高，为了得到适宜的固含量，又要加入可生化废弃物。那么，把可生化废弃物、畜禽养殖废物、污水处理厂的剩余污泥这三类物质混合，再进行厌氧消化的处理，就刚好达到了废物尽其用的优势互补目的。这种做法，不需要额外投资,又解决了厨余垃圾带来的环境污染问题。为了提高产气量，只要搭配合理，这些不同的物料的基本营养基质具有互补性，就可以使厌氧消化过程达到很好的平衡。

如何预测和控制对混合厌氧消化系统的最佳进料搭配和相关最佳工艺条件，自动实现不同种类生物质混合厌氧消化过程中组分的搭配优化？

本案例的目标：

1）以混合厌氧消化试验获取的产气阶段数据为样本，根据实际预测控制需要，建立混合厌氧消化系统的在线预测模型，并进行在线预测。

2）在线预测模型能够实时反映系统当前状态，在线更新训练样本，滚动优化模型结构并预测系统下一步输出。

分析方法与过程

本案例中，系列试验以污水处理厂剩余污泥为基本底物，将其与不同类型含有某一优势营养组分的生物质废物的相互搭配混合，以及与餐厨垃圾混合，进行中温和高温混合厌氧消化试验。其中，涉及的富含某一优势营养组分的典型生物质废弃物有以下几类：基本成分富含氮的城市污水处理厂剩余污泥；基本成分富含碳水化合物菜蔬、植物、谷类废物（来自居民厨余，学校或大型企业食堂，餐饮业，园林绿化等）；基本成分富含蛋白质废物（来自豆类加工厂）； 基本成分富含脂肪类废物（来自学校或大型企业食堂，餐饮业）。试验对相关物料优势营养组分进行搭配正交试验，分析互补机制及其对过程产物和相关工艺参数的影响，通过控制对混合厌氧消化系统的最佳进料搭配和相关最佳工艺条件，自动实现不同种类生物质混合厌氧消化过程中组分的搭配优化。

1. 试验装置

本例基于单基质生物质与剩余污泥混合试验数据进行仿真分析。选取在中温且控制 PH 值条件下，对单基质生物质与剩余污泥进行混合厌氧消化，试验装置如图 2-85 所示。厌氧反应容器为 500ml 的抽滤瓶，对物料定期进行人工搅拌，产生的气体由出气孔导出，用装满饱和食盐水的集气瓶收集，对气体的体积每天记录，对气体成分定期监测。监测消化过程的方法是取消化液分析，每三天对出料消化液进行一次分析，分析其碱度、VFA、氨氮等参数。

www.ipdm.com

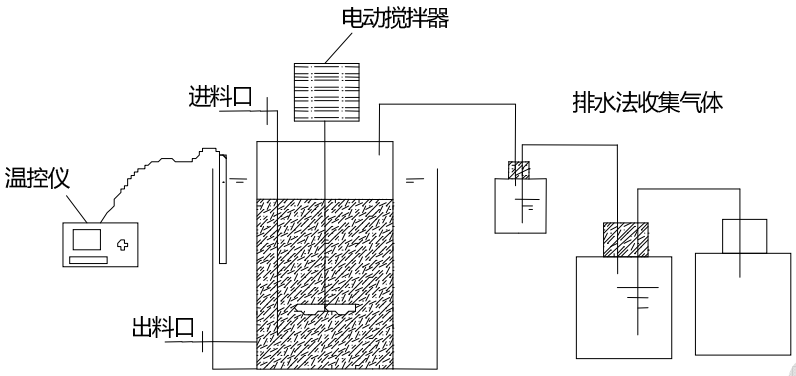


图 2-85 试验装置图

混合厌氧消化过程主要包括水解阶段、发酵阶段、产乙酸阶段和产甲烷阶段，这四个阶段是依次连续发生的。前几个阶段模型相对简单，而最后一个阶段则较为复杂，本例主要建立产甲烷阶段（共 82 组数据）的模型，并采用泛化均方根误差（RMSE）来判断模型预测效果。

$$RMSE=\sqrt{\frac{1}{l}\sum_{i=1}^{l}(y_i-\overline{y_i})^2},\ i=1,2,\cdots,l \tag{2-4}$$

2. 数据探索

这里选取在中温条件下，以肥肉为基质，和剩余污泥混合的控制 PH 值的厌氧消化过程的试验数据来进行分析。

图 2-86是PH值和VFA的变化曲线。可以看出，肥肉在厌氧消化过程中，PH值呈下降趋势，投料后第1天其pH值几乎没什么变化，在第2天和第3天持续调节PH两天，第4天保持在6.5，第5天上升为6.82，但第6天却又降至了6.2，将其调回6.5后，第8天再调节一次，此后PH保持在该水平，直至第26天开始上升至7以上，直至消化结束。从以上分析中看出，在没有调节肥肉发酵的PH的情况下，仍能有一定程度的回升，说明脂肪分解菌也具有较大的抗酸能力。

VFA 浓度呈现出先增后降的变化趋势，在投料后的第 1 天 VFA 浓度较低，为 120mg/L,这可能是由于肥肉其易降解物质较少，而且进一步的水解受到已产生的 LCFA 的抑制，之后便开始增长，在第 14 天就到达最高值 2430mg/L，直至消化结束一直呈下降趋势，但在第 14 天到第 68 天这段时间内的下降幅度较大，此后基本维持在 350mg/L 左右。肥肉发酵产生的 VFA 浓度在四种基质中最小，这是由于脂肪酶本身的特殊性，使得脂肪的水解成为其降解的限速步骤，进而产生的 VFA 不会在短期内积聚较大的浓度。

www.tipdm.com

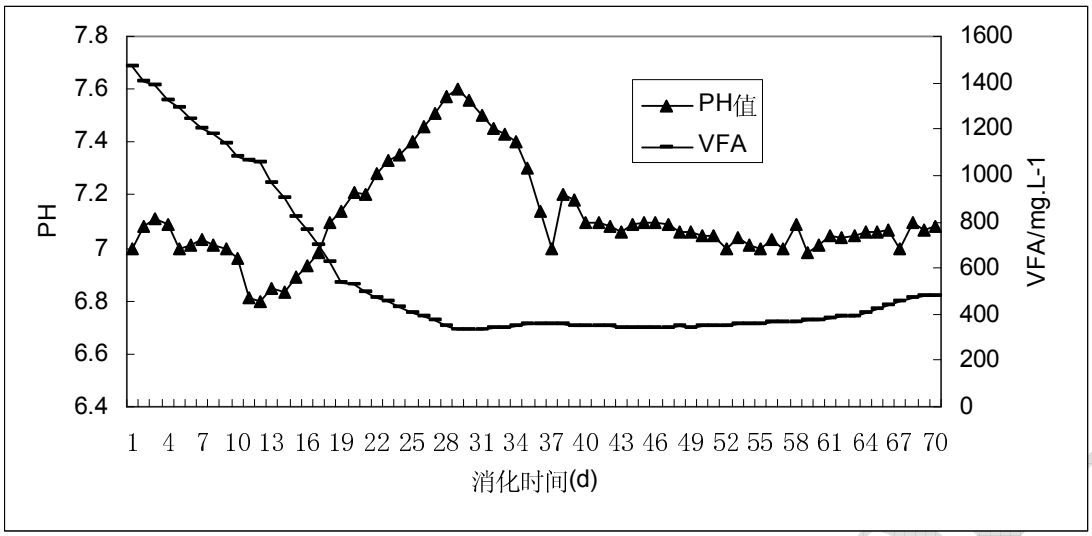


图 2-86 肥肉中温厌氧消化 PH 和 VFA 变化曲线

图 2-87 是日产气量变化曲线。可以看出，PH 的控制对肥肉消化的产气量产生了较大的影响，在投料初期产气量较小，虽然在此阶段持续产气 7 天，但日均产气量只有 35.7ml，这跟脂肪降解的机理有关，脂肪的水不溶性导致其降解时倾向于形成悬浮聚合体，这些聚合体较低的比表面积使得其被微生物降解的速度减慢，第 8 天停止产气，这是由于油脂在厌氧环境中首先被水解为甘油和自由长链脂肪酸（LCFA），LCFA 抑制了油脂的进一步降解，但这种抑制并不是不可逆转的，在 LCFA 降解后，短时间内，产甲烷速率就开始增加，因此在第 19 天开始产气，但在第 22 天又开始下降了，可能此时产甲烷菌还没有形成优势菌种，这种较少量的产气持续到第 43 天，第 44 天开始日产气量上升至 100ml 以上，此后产甲烷速率稳定，持续产气直至消化结束，累计产气量达 14700ml，此阶段甲烷含量均在 60%以上，最大值为 72.5471%，直至消化结束也都保持在 65%以上。

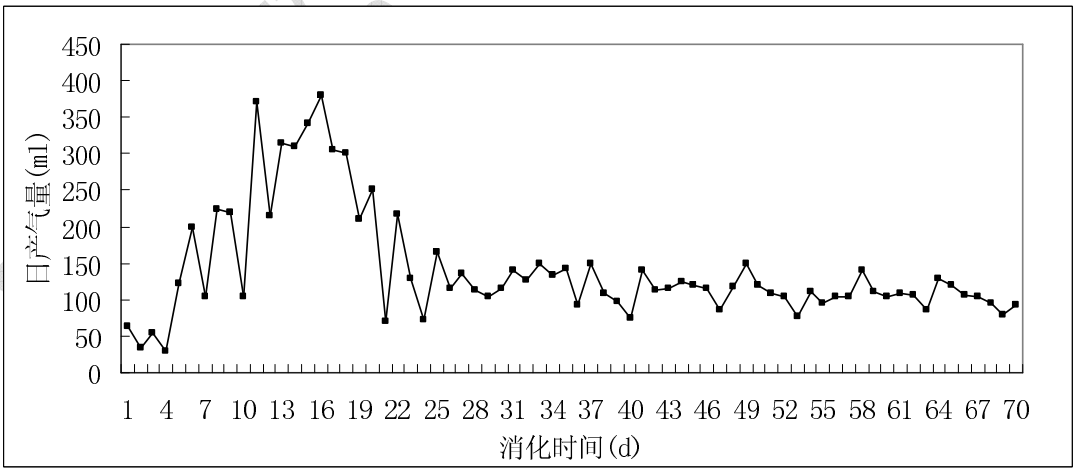


图 2-87 肥肉中温厌氧消化日产气量变化曲线

3. 在线建模

单基质生物质与剩余污泥混合厌氧消化系统是一个动态系统，在混合厌氧消化过程中，产气量可以看作是一个非线性递增的非平稳时间序列，厌氧消化过程中当前时刻产气量不仅与容器内环境

干预（如温度、有机负荷、PH 值等）有关，还与历史时刻产气量有关，因此，构造混合厌氧消化模型如图 2-88 所示。

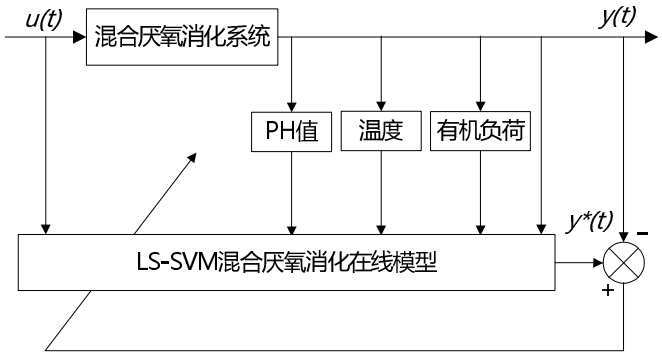


图 2-88 混合厌氧消化在线模型结构

其中，k 时刻模型输出可表示如下：

$$y^*(k)=f(y(k-1),\cdots,y(k-n),u(k),\cdots,u(k-m),\theta) \tag{2-5}$$

式中：θ为模型参数集，n、m为时间延迟。时间延迟的确定特别重要，时间延迟过大，模型结构复杂，计算量增大；时间延迟过小，则无法反映系统动态特性，建模精度变差。本文根据仿真实验时模型的拟合精度来确定时间延迟的大小。

传统的离线建模方法是将训练样本一次性加入模型进行训练，然后预测全部未来输出，虽然预测结果是全局最优解，却不能反映系统的动态特性。由于实际的系统在运行中，随着时间的推移，其状态是不断发生改变的。对于离线模型，由于模型结构已经确定，当系统受到干扰或者运行状态发生变化时，极易导致模型失配，从而难以应用于在线预测控制。对于在线建模，模型将最新采集的系统数据补充到训练样本中，在线更新训练样本，根据 k 时刻的训练样本训练模型，在有限时域内完成优化，以预测 k+1 时刻的系统状态。因此，在线建模是动态的，模型能够根据系统的不断变化而改变。在线建模的优化目标不是全局最优，而是使 k+1 时刻预测结果为最优的模型参数，对模型参数的优化是在每一步预测输出前反复进行的滚动优化。因此，在线建模能够实时的反映系统当前状态，准确的预测系统下一时刻的输出。

在线建模时，需要不断将最新采集的数据加入到训练样本中。当系统的训练样本增加到一定的规模时，早期的历史数据对训练模型的作用就不断的减小[9]。多余的训练样本不仅会影响计算速度，也会使模型预测受到干扰。为了减少历史数据对建模的影响，使训练样本始终保持一个合适的规模而不至于过于庞大，需要在不断补充最新采集的样本数据同时，“舍弃”早期的训练样本，使训练样本始终保持不断向前滚动更新的状态。建立基于 LS-SVM 混合厌氧消化在线模型的步骤如下：

1）设定模型初始参数；

2）建立初始状态模型训练样本集。系统开始运行后，当采集到第 n+1（n 为时间延迟，当系统数据采集到第 n 步后，才能构成一个完整的样本）个数据后，即可以建立初始训练样本；

3）采集当前时刻系统数据，将其加入训练样本集中，在线更新训练样本；

4）以当前模型参数及建模样本对参数c和σ^2进行寻优，找到最优值；

5）设定最优参数$(c,\sigma^2)^2_{MMSE}$作回归训练，计算出权系数向量α和偏置量b，建立量测量与状态量的估计模型式。

6）把t时刻标准化后的量测值z_t^{pre}代入式(7)，得到状态值z_t^{pre}，再对z_t^{pre}作反标准化处理，即得到t时刻的系统输出值$\hat{x}_t$。

7）比较模型预测输出与采集到的系统输出间的误差是否满足要求；

8）若预测误差满足小于 δ（δ 为设定的最大允许误差），则保持当前模型参数组合c、σ^2不变，删除样本集内最早的样本，转步骤 3）重新建立模型；

9）若预测误差大于 δ，不能满足系统要求，则转步骤 4），重新搜索当前样本对应的最优参数组合。

实际应用时在线模型优化方法分为两种情况：以训练样本规模设定为 50 个为例，在训练样本数小于 50 的时候，在步骤 3）中更新训练集时，只增加当前时刻的样本，而不删除历史样本。当样本集达到 50 个样本后，每增加一个样本的同时就要删除一个最早的历史样本。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的基于网格搜索支持向量机（LS-SVM）方法进行模型构建，以单基质生物质与剩余污泥进行混合厌氧消化开始产甲烷时采集的数据作为初始训练样本进行离线建模并优化，取时间延迟 n＝7，根据上节所述建模步骤，从第 n+1 步起进行在线预测，预测输出泛化均方差为 0.072139。图 2-89 和图 2-90 分别显示了混合厌氧消化的预测输出结果和预测误差。从图中可以看到原始样本数据点共 82 个，拟合预测点 74 个（去除 7 个时间延迟点和 1 个验证点），第 83 个点为预测点；图 2-91 和图 2-92 显示了基于网格搜索的 LS-SVM 参数选择结果，图中 log2c 为以 2 为底对数的正则化参数，log2g 为以 2 为底对数的核参数 σ 2，CVmse 为最佳参数下的交叉验证均方误差。

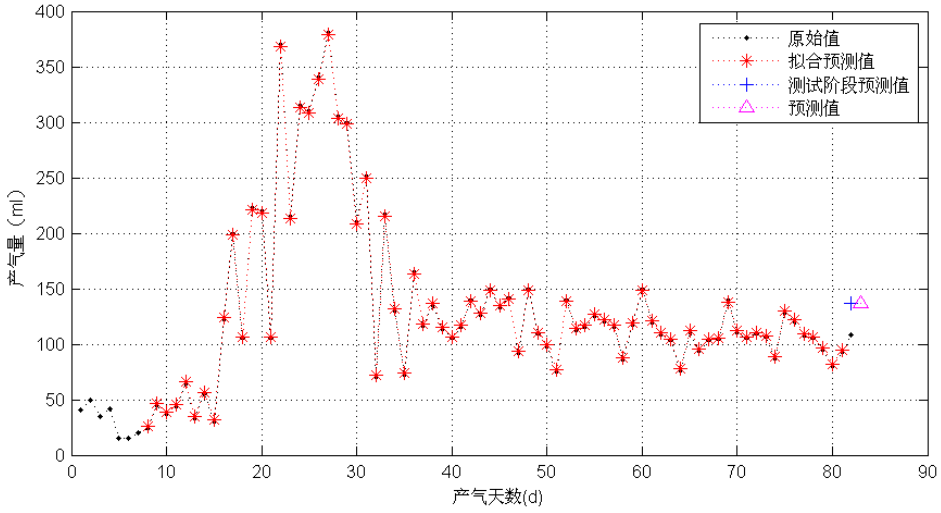


图 2-89 基于 LS-SVM 的混合厌氧消化在线预测结果

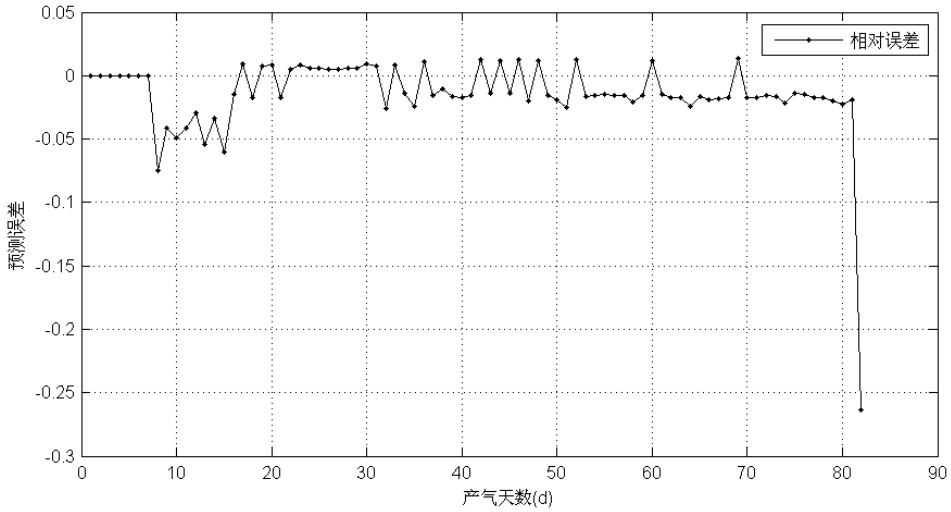


图 2-90 基于 LS-SVM 的混合厌氧消化在线预测误差

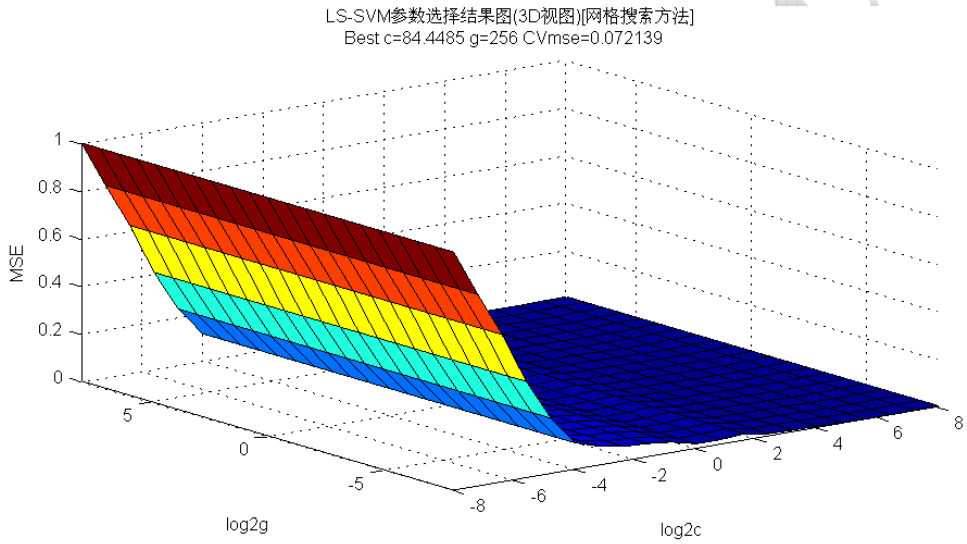


图 2-91 基于网格搜索的 LS-SVM 参数选择结果 3D 视图

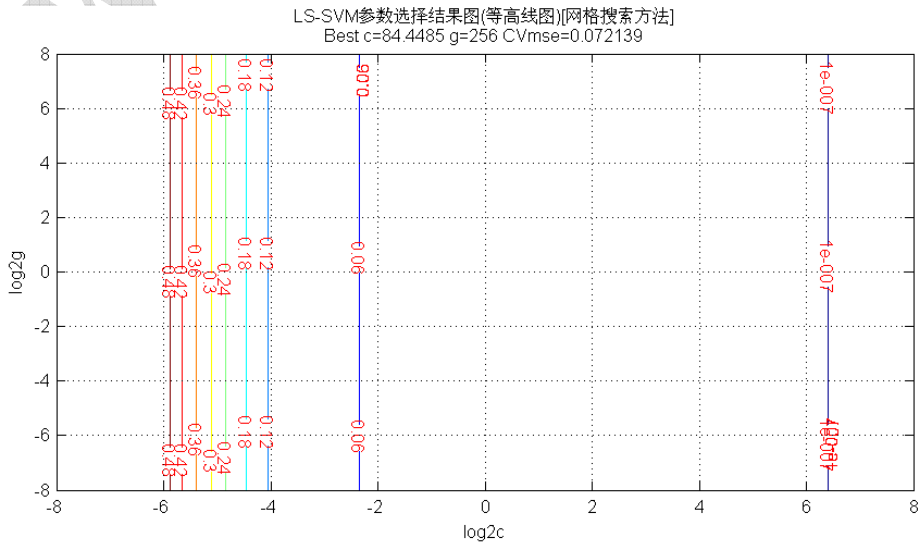


图 2-92 基于网格搜索的 LS-SVM 参数选择结果等高线图

由于构成样本数据时采用了时间延迟，因此在采集的系统数据小于 n 时无法构成一个完整的训练样本，模型实际的预测步数要小于采集的系统数据数量。模型预测输出产生较大跳跃，如图 5 第 8d 至 15d，产生较大误差。误差曲线在持续几步的振荡后，当扰动点处移出训练样本覆盖范围后，模型预测输出误差显著减小，能够以较高的精度拟合实际产气量曲线，说明模型泛化能力较强，具有一定鲁棒性，预测误差均在允许的范围内。

2.4.4. 案例十五 水产养殖投入产出多目标优化仿真

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

水产养殖业是我国渔业的重要组成部分，也是渔业发展的主要增长点。我国的渔业发展重心由“捕捞为主”向“养殖为主”的转移，促使水产养殖业发生了巨大变化。2009 年我国水产品总产量为 5116 万吨，其中养殖产量 3622 万吨，连续多年居世界第一位。工厂化养鱼、网箱养鱼、流水养鱼等各种高产养殖方式，立体利用水域、水陆复合生产的生态渔业以及能量充分利用等各种高效利用模式得以较为广泛的应用，保持渔业资源和水域资源可持续利用的生产技术已越来越被生产者接受和掌握。同时，以生物技术（细胞工程及基因工程应用）、信息技术为主的渔业高新技术也有了较大发展，有些已在生产中发挥了作用。这些技术的应用和日益普及，加快了渔业现代化的步伐。分析我国的资源现状，进一步发展水产养殖业生产仍具有相当大的潜力。据统计，全国有适宜养殖的浅海滩涂 260 万 hm^2，目前仅利用约 1/3；可养内陆水域 670 万 hm^2，也仅利用 2/3；还有未开发的宜渔盐碱低洼荒地 230 万 hm2。同时，已利用的养殖水面，提高单产的潜力也很大。因此，发展我国的水产养殖业，无论从经济发展的需要，还是从资源的潜力来讲都具有较好的条件。但我国水产养殖业的发展目前还面临很多新情况、新问题。

养殖户整体计算产量和效益时，往往发现饲料系数偏高，养殖成本高，产量近几年逐渐增加，而效益却没有增加多少，有时甚至下降，收效甚微。近几年来，鱼价低迷，而塘租、饲料、人工、水电等成本逐年增加，许多养殖户面临着上述同样的困惑和窘境，养殖的出路到底在哪里已经成为较多养殖户关注的热点[17]。

在一些水产养殖区，经常会听到一些渔民朋友抱怨：“现在养鱼，鱼价低、鱼料贵、鱼病多，本大利小，弄不好就要赔钱。”

俗话说：有同行无同利。这句话用在当前的水产养殖业中再恰当不过。同样是从事水产养殖，为什么有人赚钱，有人保本，而有的人却赔钱呢?

如何通过市场预测，选择合适的养殖模式，不断调整生产结构，学会“灵活应变”，从而获得理想的经济效益？

在有限的资源和技术参数条件下，如何统筹安排，合理规划，达到最大的经济效益？

分析方法与过程

随着水产养殖模式的复杂多变性，以及行业内竞争的日趋激烈，传统的水产养殖效益评价方法已经很难满足水产养殖当前生存和发展的需要，实现水产养殖投入产出的仿真试验势在必行。在计算机上建立水产养殖投入产出各种仿真的数学模型，能够迅速、灵活的进行虚拟试验，方便对水产养殖模式的调整和再规划，从而带来巨大的社会效益与经济效益。仿真技术现在已经成为各类应用系统分析、研究、预演中不可缺少的重要手段，在降低项目实施风险等方面起着越来越重要的作用。

本案例构建水产养殖投入产出多目标优化仿真平台，通过养殖水域生态系统分析、生态可持续发展分析，与经济模型进行整合，采用多目标决策分析技术，实现了经济环境模型的动态协同运算和多目标优化求解，从而较为充分地体现了经济系统与环境系统间相互影响、相互制约的动态关系。

水产养殖投入产出多目标优化仿真平台，是以人工智能仿真技术和养殖工程技术为支撑，帮助用户在实施水产养殖项目前，通过模拟水产健康养殖所涉及的对象和全养殖过程进行可视化表达、设计、控制和管理，使生产方式从传统的以物流为核心转换为以信息流为核心，通过仿真平台模拟，实现以最少的资源耗费而获得最大的优质产出和高效益的事前预演，最大限度地推动养殖增长方式的转变。

建立一个实用的水产优化养殖投入产出仿真平台，实质上就是要求仿真系统支持水产养殖灵活建模数据结构，而多养殖模式要求仿真系统具有更多的产业结构优化算法支持，因此构建一个灵活的水产养殖仿真系统需要在生产模型灵活性和可操作性之间寻求合适的平衡点。

有效地模拟水产优化养殖的投入产出行为需要运用具有学习能力的智能代理根据市场的变化，进行资源的合理利用，这就对算法的“智能”水平提出了很高的要求，一方面要适应本地的气候条件，另一方面，又要对市场进行预测，不断调整生产结构。

水产养殖仿真完成后需要对养殖结果进行分析和评估，确定一系列分析指标，对试验结果和预期结果进行比较。

上述问题是构建水产养殖仿真系统的关键性问题。为了更有效、直观地进行仿真实验，还要求系统具有良好地交互性能、直观地实验展示能力和方便地实验管理功能。

本项目主要应用灰色系统理论等方法建立水产养殖的投入产出模型，通过仿真方式模拟选择合适的养殖模式、养殖技术、投入成本，同时利用灰色线性规划对市场进行预测，优化调整产业结构，从而获得理想的经济效益。应用该仿真系统进行水产养殖投入产出模拟实验，能在一定程度上还原在指定养殖周期和各种资源（水体、资金、种苗等）都有限的情况下，如何统筹安排，合理规划，达到最大经济效益。

水产养殖投入产出多目标优化仿真平台的主要内容包括：1）养殖水域生态系统分析及生态系统

结构优化分析；2）基于动态扩展的养殖水域产量因素分析；3）基于可持续发展的水产养殖投入产出分析；4）生态系统水平的水产养殖效益分析；5）生态环境下的水产养殖投入产出动态仿真分析。水产养殖投入产出多目标优化仿真流程见图 2-93。

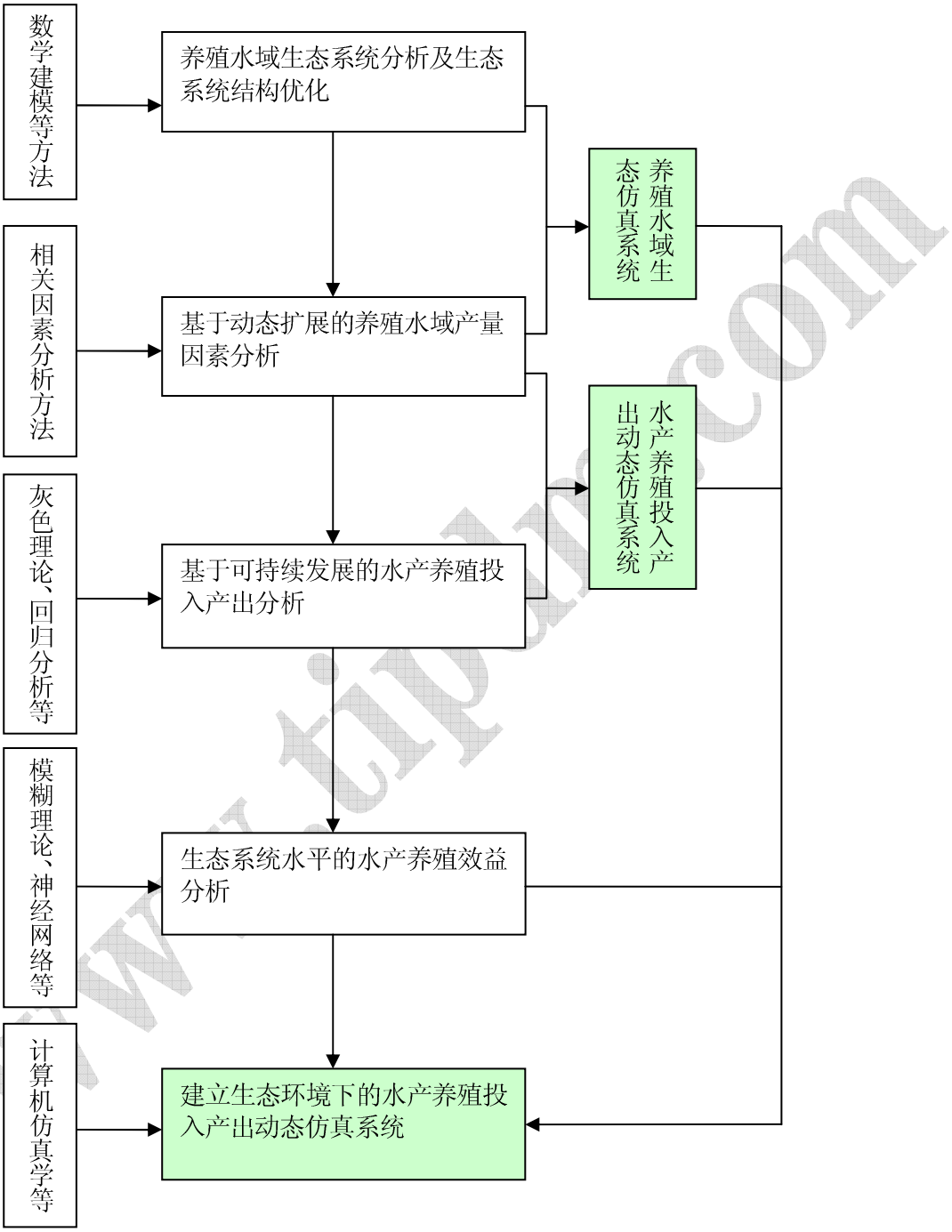


图 2-93 水产养殖投入产出多目标优化仿真流程

1．养殖水域生态系统分析

养殖水域生态系统通常是一个封闭或半封闭的生态系，与自然海区的能量传递和物质循环经常受到阻滞。与自然海域的生态系相比，具有生物种类组成简单，养殖生物在群落中占绝对优势，主

要食物链短且简单，营养层次少，缺少由碎屑食物链和腐食食物链组成的微型食物网，生态系统稳定性差，对病害和其它灾害的抵抗能力弱等特点。天气变化和人工调控都能在短时间内大幅度改变系统中的水化因子以及细菌、浮游生物、原生动物等的生物量和种类组成，使其生态结构和功能发生很大变化。

养殖水产品的生物量优势是在人工扶持下形成的，其庞大的生物量造成了生态金字塔畸形，它是经济产品较大输出量的保证，但也决定了该系统的低生态缓冲能力和脆弱性。养殖水域生态系统的低生物多样性和高生产力输出是现实所追求的经济效益所决定的，这也是这一生态系统存在的根本目的。

2．养殖水域生态系统结构优化分析

我国的水产养殖业发展迅速，养殖产量在世界水产养殖业和我国渔业中所占的比重也逐年加大。由于各养殖户分散经营和管理，对科学养殖方面的知识了解不多，缺乏总体规划，使一些科技性的指导贯彻不到生产中去，致使养殖布局出现不合理的现象，养殖容量超过了水域的负荷力，导致养殖水域污染加重，底质恶化，水产品质量下降(种质退化），赤潮频发，养殖病害蔓延，有时还会发生已被污染的海水直接返回的二次污染等严重问题。各种病原微生物只是养殖病害的直接诱发因子，而促使养殖品种发病的间接因素同样重要也更为复杂。根据生态学上耐受性定律的原理可知，当环境因子不是处于养殖生物最适度的状态时，养殖生物对于其它生态因子的耐受限度也会下降，如对病原菌或病毒的抗发病能力。Fegand1991 年认为，保持养殖水环境的生态平衡可以降低养殖生物的发病率，即使有病毒感染，只要处于最适养殖环境中，也可以不发病，并且在以后的生长过程中，病毒的数量会逐渐减少以至难以检出。因此保持养殖环境的适宜性和稳定性，优化养殖水域生态结构和管理，是使养殖活动得到良性循环发展，取得最大的经济、生态和社会效益的好出路。

由于养殖水域生态系统本身的性质决定了它的结构不稳定，为维持其生态平衡仅靠自然的调节是远不够的，必须引入人为的调控和管理。

3．基于动态扩展的养殖水域产量因素分析

由于多品种混养、养殖技术、价格的变化，使得水产养殖的模型建立的难度大、效果差。建立模型如选择多个因素，反而效果差，只有选择其主因子进行分析，才能提高模型的精确度。

养殖水域往往由于混养品种多，影响产量的因素也多，还有些因素不了解；虽然数据量大(样本大），但不一定具有典型分布，以往对养殖水域生态的分析法主要是统计方法，如回归分析等。一般来说，回归分析大多用于少因素、线性的情况，对于多因素的、非线性的则难以处理；且回归分析要求数据有典型分布，才能用其相应公式进行运算。但许多数据不一定具备以上的条件，此时如用回归分析往往出现与事实相反的结果。灰色理论的关联分析中考虑了以上不足，以曲线的发展态势的接近程度来比较曲线的关联大小，适应于多因素、非线性、非典型分布的数据。灰色理论的五步建模方法中的动态 GM（1，N）模型，它是建立在原始数据的基础上，通过生成数建立微分方程，避免了用线性方法时存在的问题，并初步建立了主体鱼的投入产出模型，为养殖结构的调整、改进提供了理论依据。

4．基于可持续发展的水产养殖投入产出模型分析

投入产出模型是一种确定性的模型，反映的经济关系是投入和产出及其相互间的一般平衡关系，它只是对现实经济联系的反映，不能应用模型本身来求最优。优化技术是帮助解决生产和经济规划中某些实际问题，并使之发挥最大效益的一门科学。应用最优化技术可以帮助人们较快地选择出最优方案，做出最优决策，因此，最优化方法在各种工程技术、自动控制、系统工程、运筹学以及经济计划、企业管理等方面都有广泛应用。在生产计划管理、社会经济问题中，人们总希望采取种种措施，以便在有限的资源条件下，或规定的约束条件下得到最满意的效果，希望制定一个既满足投入与产出的平衡关系，又能使经济系统取得最优效果的最佳方案，并且希望投入产出分析能为优化产业结构、提高经济效益和社会效益、最优配置人力资源以及自然资源和财力资源、保护环境资源等提供决策依据。因此，投入产出分析必须与优化方法相结合，建立投入产出优化模型，以实现其优化功能。

水产养殖的投入产出模型的建立逐渐引起人们的重视。已有过应用线性规划对养殖虹鳟、对虾等鱼的养殖经营进行过线性规划分析。有人对肥育池的产量进行多元线性建模。国内史洪芳建立过池塘养鱼利润的一元回归分析方程；陈大庆应用通径系数分析了各因素对产量的影响；张展等应用线性规划对池塘的资金、水体等进行规划。以上的方法均为线性方法，我国池塘养鱼是多品种、多规格鱼类混养，数据一般为非线性、多因子的，这样的因子不能简单套用线性方法。由于池塘的投入、产出是清楚的，而太阳能的转化效率、固氮水生作用、地表径流代入的营养等是难以准确计算的，这种系统属于灰色系统，本项目主要应用灰色系统理论方法建立养殖水域的投入产出模型。

建模仿真

甲鱼是种名贵的水产动物，具有主富的营养价值，亦是我国传统的滋补佳品。由于目然资源的减少，人们对甲鱼需求增加，一度曾使甲鱼身价百倍，甲鱼养殖有着巨额利润，在全国形成一个甲鱼养殖高潮，养殖面积和产量迅速提高。

但随着甲鱼养殖产量的提高，进口甲鱼对我国市场的冲击等因素，1996 年甲鱼价格一路下滑，致使许多单位出现亏本。养殖甲鱼是否有前途，该产业能否继续发展，成为养殖者关心的问题。我们知道，甲鱼毕竟是人们喜爱的滋补佳品，市场价格合理时，消费需求又会增加。今后的甲鱼养殖业是在竞争中求发展，而竞争主要是规模和技术的竞争。选择合适的养殖模式，提高技术，降低成本，从而获得理想的经济效益。养殖模式的选择，一方面要适应本地的气候条件，另一方面，又要对市场进行预测，不断调整。我们利用灰色线性规划对市场进行预测，并提出产业结构调整的方法。

甲鱼近年价格变化大，种在变、成甲鱼价格也在变。而饲料人工等基本上无大的变动。因而选用甲鱼种苗和成鱼价格进行分析，提供一些预测值，做为养殖规划的参考，为近期的调查结果。

应用灰色理论对甲鱼的苗种价格、成鱼价格进行了预测，在这基础上对甲鱼生产进行了灰色线性规划优化。对一个场的优化结果表明，同样投入资金 400 万，不同的生产安排结果表明：一个产出为 533 万，另一个可产出 775 万。

本案例将分别采用灰色理论进行价格时序预测和灰色线性规划进行投入产出模型构建。

4. 模型输入

模型的输入包括两部分，一部分是建模样本数据的输入，另一部分是建模参数的输入。

表 2-111 两种规格甲鱼成本核算

指标	大规格	小规格
重量(克)	50~150	5~50
成活率(%)	90%	70%
饵料系数	2	2.5
人工、水电		
养殖周期(月)	6	12
苗价格(元)	20,18,15,12	10,8,5,4
成品价格(元)	170,150,120,100,80,60,55,50	150,130,100,80,60,55,50,48

表 2-112 灰色理论时序预测建模参数

序号	参数名称	参数描述
1	预测步长	预测周期
2		

表 2-113 灰色线性规划建模参数

序号	参数名称	参数描述
1	变量个数	目标函数中变量个数
2	约束条件个数	建模约束条件个数
3	目标函数	建模目标函数
4	约束条件	目标函数的约束条件

5. 仿真过程

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 GM 灰色时序预测和灰色线性规则等算法进行模型构建，建模仿真过程见图 2-94。

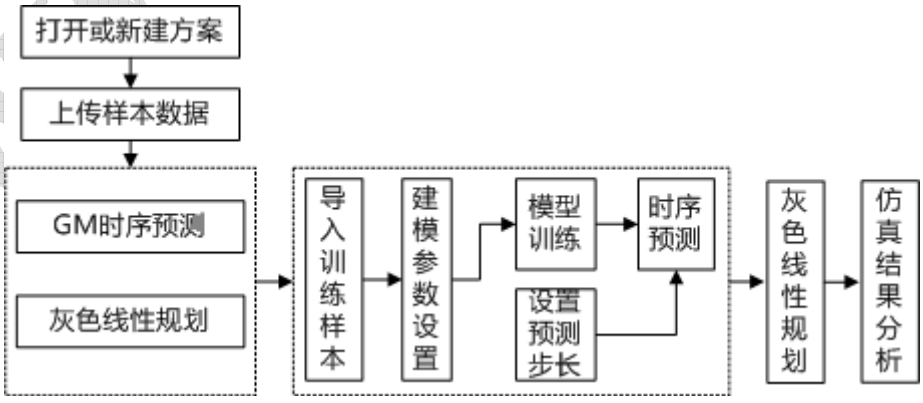


图 2-94 仿真过程

建模仿真过程说明如下：

1） 大规格苗种价格预测；

2） 小规格苗种价格预测；

3） 大规格成鱼价格预测；

4） 小规格成鱼价格预测；

5） 计算符合约束条件的最优解。

6. 结果及分析

（1）采用GM灰色时序预测算法对苗种及成鱼价格进行预测

以大规格苗种为例，GM时序预测结果见表 2-114。

表 2-114 对于大规格苗种价格预测结果

年份序列	实际值	一次累加值	预测值
1	20.0	20.0	20.0000
2	18.0	38.0	18.0273
3	15.0	53.0	14.7693
4	12.0	65.0	12.1001
5			9.9133
6			8.1218

一次累加数列的时间响应函数：

x1f[k] = (x0[0] - b/a) * e^(-a*k) + b/a = -99.75000000000053*e^(-0.19933554817275673*k)+119.75000000000053

（2）对两种甲鱼养殖面积进行灰色线性规划，计算符合约束条件的最优解。

所谓的灰色线性规划就是把某些约束值用 GM（1,1）模型进行预测，然后把预测值当做某个约束值再做线性规划，实现动态的线性规划。

注意：在使用单纯形法求解线性规划时必须先把全部条件化成标准形式。

1）某养殖场，现有资金400万，水面150亩，大规格苗种养成成鱼时价格为12000元/吨，小规格苗种养成成鱼时价格为9000元/吨，大规格、小规格各养多少才能使养殖效益最高?

```
目标函数：
    Z = 12.0x1 + 9.0x2
约束条件：
    0.0010x1 + 0.0007x2 <= 150.0 （亩，面积）
    9.0x1 + 8.0x2 <= 400.0 （万元，资金）
    2222.0x1 + 2500.0x2 <= 150000.0 （万只，种苗）
```

说明：x1、x2分别对应大规格甲鱼和不规格甲鱼的放养量。x3、x4、x5是把问题转化成标准形式需要而另外加的变量。

```
约束条件的系数矩阵A：
    m_row = 3
    m_col = 5
    0.0010  0.0007  1.0 0.0 0.0
    9.0 8.0 0.0 1.0 0.0
    2222.0  2500.0  0.0 0.0 1.0
约束条件的系数矩阵中对应基变量的部分B：
    m_row = 3
    m_col = 3
    0.0010  1.0 0.0
    9.0 0.0 0.0
```

www.ddjm.com

```
      2222.0   0.0  1.0
约束条件的系数矩阵中对应非基变量的部分 N:
      m_row = 3
      m_col = 2
      7.0E-4   0.0
      8.0  1.0
      2500.0   0.0
目标函数的系数矩阵 C:
      m_row = 5
      m_col = 1
      12.0
      9.0
      0.0
      0.0
      0.0
目标函数的系数矩阵中对应基变量的部分 CB:
      m_row = 3
      m_col = 1
      12.0
      0.0
      0.0
目标函数的系数矩阵中对应基变量的部分 CN:
      m_row = 2
      m_col = 1
      9.0
      0.0
约束值矩阵 b:
      m_row = 3
      m_col = 1
      150.0
      400.0
      150000.0
目标函数的最优解 Z = 533.3333333333333
取得最优解时各基变量的值为:
      x1 = 44.44444444444444
      x3 = 149.95555555555555
      x5 = 51244.44444444445
```

由上可知，全部养殖大规格甲鱼利润最大。

2) 当其它条件同 1)，种苗无限制，只限制水面和资金时：

```
目标函数:
    Z = 12.0x1 + 9.0x2
约束条件:
    0.0010x1 + 0.0007X2 <= 150.0 （亩，面积）
    5.5x1 + 3.2x2 <= 400.0 （万元，资金）
```

说明：x1、x2 分别对应大规格甲鱼和不规格甲鱼的放养量。

程序输出：

```
约束条件的系数矩阵 A:
```

www.tipdm.com

```
    m_row = 2
    m_col = 4
    0.0010   7.0E-4   1.0 0.0
    4.5 3.2 0.0 1.0
约束条件的系数矩阵中对应基变量的部分 B:
    m_row = 2
    m_col = 2
    7.0E-4   1.0
    3.2 0.0
约束条件的系数矩阵中对应非基变量的部分 N:
    m_row = 2
    m_col = 2
    0.0010   0.0
    4.5 1.0
目标函数的系数矩阵 C:
    m_row = 4
    m_col = 1
    7.2
    6.2
    0.0
    0.0
目标函数的系数矩阵中对应基变量的部分 CB:
    m_row = 2
    m_col = 1
    6.2
    0.0
目标函数的系数矩阵中对应基变量的部分 CN:
    m_row = 2
    m_col = 1
    7.2
    0.0
约束值矩阵 b:
    m_row = 2
    m_col = 1
    150.0
    400.0
目标函数的最优解 Z = 775.0
取得最优解时各基变量的值为:
    x2 = 125.0
    x3 = 149.9125
```

可见苗种充足时利润最大。

由上面的几种情况看出，资金、水面、种苗的不同应用，效益差别悬殊。

www.tpdm.com

2.5. 公共服务

2.5.1. 案例十六 乳腺癌证素变化规律及截断疗法

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

目前，癌症已经成为威胁我国居民生命健康的主要杀手，我国每年新发癌症病例 200 万人，因癌症死亡人数为 140 万；我国居民每死亡 5 人中，即有 1 人死于癌症；在全国不少大城市，恶性肿瘤已经超越心脑血管疾病，成为第一死亡原因。近年来癌症死亡约占我国城乡居民总死亡构成的 22.32%，癌症死亡率比 70 年代中期增加了 83.1%，比 90 年代初期增加了 22.5%；其中，乳腺癌、肝癌、胃癌、食管癌、结直肠癌、乳腺癌、宫颈癌及鼻咽癌合计占癌症死因的 80%以上。癌症发病年龄提前了 15-20 年，35-55 岁发病群体比率趋于上升，乳腺癌发病年龄提前到了 20 岁。据卫生部《第三次全国人口死因调查》报告显示：与环境、生活方式有关的乳腺癌、肝癌、结直肠癌、乳腺癌等癌症死亡率及其构成呈明显上升趋势，其中乳腺癌和肺癌上升幅度最大，过去 30 年分别上升了 46.5%和 96% 。

乳腺癌是妇女常见恶性肿瘤之一，现全世界每年发病数已达 130 万例。我国发病率逐年增高，在部分大城市，如北京、上海等，已占妇女恶性肿瘤的首位，因此乳腺癌又有“城市女性第一杀手”的称号，其每年死亡人数约占发病人数的 1/3，严重威胁广大妇女的生命健康。乳腺癌的发病病因还不完全清楚，尚无有效预防措施，也未能理想地控制其复发转移，在我国死亡率仍在持续上升阶段。

控制癌症肆虐的根本出路在于预防。要搭建三级预防癌症的“屏障”。一级预防即消除或减少可能致癌的因素，防止癌症的发生；二级预防利用早期发现、早期诊断和早期治疗的有效手段防止初发疾病的发展；三级预防目标是防止病情恶化，预防残疾，采取多学科综合诊断和治疗，正确选择合理的诊疗方案，为能够治愈的患者提供根治性治疗，以达到治愈的目的；为已无法治愈的患者提供姑息治疗和临终治疗，以消除痛苦、恢复体力、延长生存时间、改善生活质量。

乳腺癌是一种异质性较强的疾病，不同分子表型的乳腺癌生物学行为差异明显，对化疗、内分泌治疗、分子靶向药物治疗的反应各异，预后也相差甚远。所谓三阴乳腺癌（triple negative breast cancer，TNBC）是指雌激素受体(Estrogen receptor，ER)、孕激素受体(progesterone receptor，PR)和人表皮生长因子受体 2(human epidermal growth factor receptor2， HER-2)均为阴性的乳腺癌。三阴乳腺癌治疗手段有限，总体生存期较短，可谓是乳腺癌中最难驯服的异类。

恶性肿瘤已成为危害人类健康的第一位杀手，应用中医药治疗恶性肿瘤已成为公认的综合治疗的方法之一，研究新的思路和方法是提高临床疗效的关键。已有的研究证明，肿瘤的转移复发成为肿瘤治疗的最大难点，约 70%的患者均会转移复发，图 2-95 为三阴乳腺癌局部复发和远处转移率情况。中医药干预后，能使肿瘤转移复发率大幅度下降。上海中医药大学附属龙华医院研究的结论是：治疗组（西医＋中医）乳腺癌 5 年转移复发率为 11.1%，对照组（单纯用西医）乳腺癌 5 年转

移复发率 32.1%。说明中医药在抗乳腺癌转移复发方面作用明显。

中医药治疗乳腺癌有着广泛的适应证和独特的优势，从整体出发，调整机体气血、阴阳、脏腑功能的平衡，根据不同的临床证候进行辨证论治。尤其对乳腺癌术后并发症的治疗有着明显疗效，对减轻乳腺癌放化疗的毒副反应、内分泌治疗所致类更年期综合征的干预、增进乳腺癌术后患者体质恢复、改善患者生存质量、提高生存率、降低发病率都具有重要的临床意义和广泛的应用前景[1][14]。

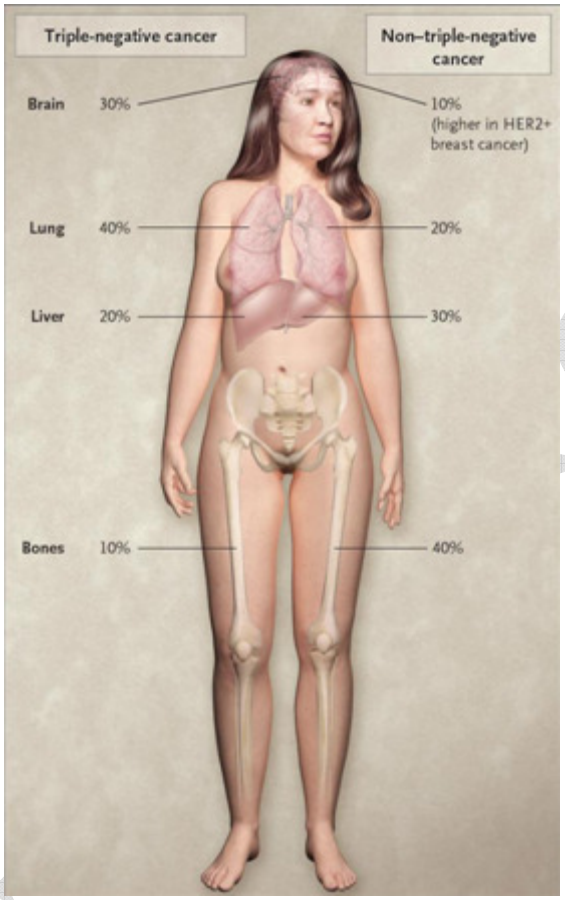


图 2-95 三阴乳腺癌局部复发和远处转移率

截断疗法治疗乳腺癌的难点是挖掘潜性证素、如何确定先证而治的方向，即后续证尚未出我们要截断什么？不同病理类型、不同病程、不同治疗阶段乳腺癌证型发展规律究竟有何异同？其截断疗法具体实施方式有何异同？

分析方法与过程

本案例提取乳腺癌中医证素并拟定乳腺癌中医证素调查问卷，选择不同病程阶段的三阴乳腺癌患者开展问卷调查，运用数据挖掘技术对问卷进行统计分析，挖掘各中医证素与临床病理及肿瘤生物学特征指标之间的对应关系，探寻三阴乳腺从发病到复发转移过程中的中医证素分布规律，以及截断病变发展、先期干预的治疗思路，帮助指导三阴乳腺癌的中医临床治疗。图 2-96 为乳腺癌证素变化规律分析过程图。

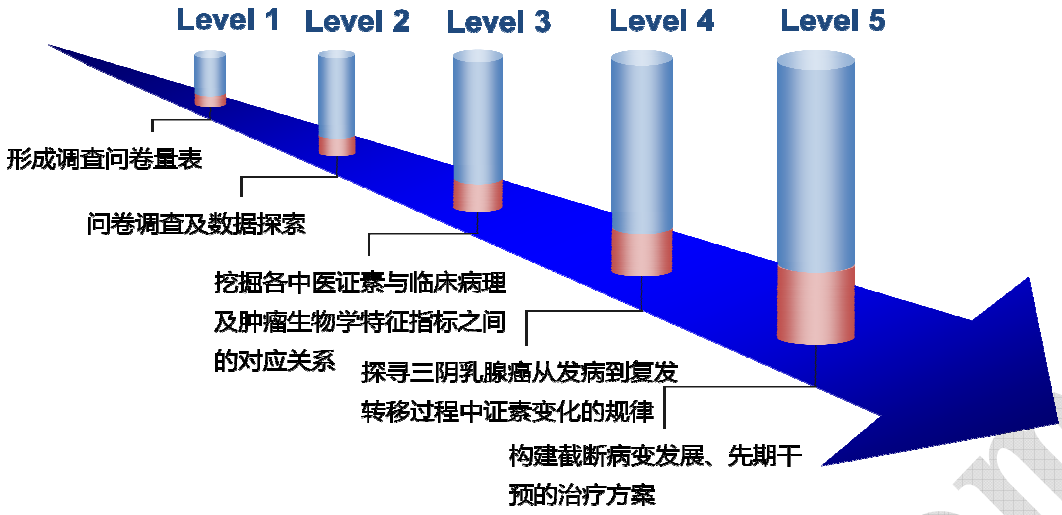


图 2-96 乳腺癌证素变化规律分析过程

1. 形成调查问卷量表

三阴乳腺癌问卷调查的临床及病理分期标准：参照乳腺癌 2002 年国际抗癌联盟 UICC（TNM）分类及分期。

1）纳入标准

✧ 病理诊断为乳腺癌。

✧ 病历完整，能提供既往接受检查、治疗等相关信息，包括发病年龄、月经状态、原发肿瘤大小、区域淋巴结状态、组织学类型、组织学分级、P53 表达、VEGF 表达等，作为临床病理及肿瘤生物学的特征指标。

✧ 没有精神类疾病，能自主回答问卷调查者。

2）排除标准

✧ 本研究中临床、病理、肿瘤生物学指标不齐全者。

✧ 存在第二肿瘤（非乳腺癌转移）。

✧ 精神病患者或不能自主回答问卷调查者。

✧ 不愿意参加本次调查者。

3）剔除标准

✧ 填写的资料无法根据诊疗标准进行分析者。

✧ 资料填写不全或中途退出本次调查者。

乳腺癌中医证素调查问卷：

表 2-115 三阴乳腺癌中医证素调查问卷（样表）

编号			**填表日期**	年　　月　　日			
姓名		**性别**		**年龄**		**确诊为乳腺癌的年龄**	
婚姻状况	☐已婚　☐未婚　☐离异　☐丧偶						
文化程度	☐小学　☐初中　☐高中　☐中专　☐大学及以上　☐其他						
职业	☐工人　☐农民　☐知识分子　☐干部　☐个体经商户　☐无职业						
工作单位/家庭住址							
联系方式				**病人种类**	☐门诊　☐住院		
月	初潮_____岁；月经（☐规律　☐不规律）；　持续_____天；间隔_____天；						

www.tjpdm.com

经史	痛经	□有 □无	末次月经时间	
	闭经	□是 □否【若是，则：闭经于____岁；闭经症状（□有 □无）】		
婚育史	婚否	□未婚 □已婚【若已婚，则：结婚年龄为____岁】		
	生育状况	□未生育 □已生育【若已生育，则育____胎，生产____胎，流产____胎；首胎生于____岁，末胎生于____岁】		
哺乳史	是否哺乳	□是 □否【若是，则哺乳____个孩子；最长哺乳____年____月，最短哺乳____年____月】		
	乳汁量	□少 □一般 □多	哺乳部位	□双侧 □左侧 □右侧
乳腺肿块	部位	□外上 □内上 □外下 □内下 □乳头后		
	发生时间及经过			
乳腺疼痛	有无疼痛	□有 □无	性质	□刺痛 □胀痛 □隐痛 □灼痛
	与月经来潮关系	□有 □无		
乳头溢液		□有 □无	性质	□水样 □乳汁样 □血样 □脓性 □浆液性
皮肤水肿		□有 □无	腋下肿块	□有 □无
乳头乳晕糜烂		□有 □无	其他症状	
曾经治疗	新辅助治疗	化疗：方案（剂量）_____已进行___周期 内分泌治疗：方案（剂量）_____使用时间_____		
	术前放疗	部位：□乳房 □内乳区 □锁骨区 剂量：_____ 次数：_____		
	辅助治疗	化疗：方案（剂量）_____已进行___周期 内分泌治疗：方案（剂量）_____使用时间_____		
	中医药治疗	治疗时间：_____ 效果：_____		

[术后病理及免疫组化资料]

原发肿瘤直径	区域淋巴结状态	TNM 分期	组织学类型	组织学分级

P-Gp	GSTπ	TOPOⅡ	Ki-67	VEGF 表达	P53 表达

[病程阶段分期]

围手术期	围化疗期	围放疗期	巩固期

[三阴乳腺癌中医证素诊断量表]（附使用说明）

Ⅰ．肝气郁结证			
定义	肝失疏泄，气机郁滞，所表现的情志抑郁，胁胀，胁痛等证候		
必备证素	肝，气滞	或兼证素	心神[脑]，(胆)，胞宫

常见证候及计量值					
3 分		2 分		1 分	
抑郁或忧虑//喜叹气	□	情志有关	□	排便不爽	□
胁胀	□	烦躁//急躁易怒	□	嗳气	□
乳房胀	□	胸闷	□	咽部异物感	□
乳房痛	□	腹胀//脘痞胀	□	口苦	□
胁痛//右上腹痛	□	胀痛或窜痛	□		
		大便时溏时结	□		

		痛经	□		
		月经错乱	□		
		乳房结块	□		
		肝大//胆囊肿大//脾大	□		
		脉弦	□		
小计(A)	___×3 分=___分	小计(B)	___×2 分=___分	小计(C)	___×1 分=___分
总分(A+B+C)		_____分			

Ⅱ. 热毒蕴结证									
定义	热毒蕴结肌肤，所表现的痈疖疮疡，患处灼热肿痛、溃烂、流脓等证候								
必备证素	肌肤，热[火]，毒			**或兼证素**	（血热），血瘀，（阴虚），（脓）				
常见证候及计量值									
5 分		**4 分**		**3 分**		**2 分**		**1 分**	
痈疖痱子	□	患部红肿	□	出疹	□	灼痛	□	新病便秘	□
患部溃烂	□			患处痛	□	局部灼热感	□	尿短黄	□
脓肿或流脓	□					皮肤瘙痒	□	脉滑	□
						发热	□		
						口渴	□		
						舌赤	□		
						舌苔黄	□		
						脉数	□		
小计(A)		小计(B)		小计(C)		小计(D)		小计(E)	
___×5 分=___分		___×4 分=___分		___×3 分=___分		___×2 分=___分		___×1 分=___分	
总分(A+B+C+D+E)		_____分							

Ⅲ. 冲任失调证					
定义	肝郁肾虚，冲任二脉功能失调，常表现为月经紊乱、乳房胀痛、腰膝酸软、烦躁易怒等证侯。				
必备证素	冲任，气郁，阴虚		**或兼证素**	肝，肾，阳虚	
常见证候及计量值					
3 分		**2 分**		**1 分**	
乳胀	□	乳房结块	□	舌淡	□
乳痛	□	腰膝酸软	□	苔少	□
月经错乱（推迟或超过 7 天）	□	五心烦热	□	黄褐斑	□
大龄未育（>30 岁）	□	口干咽燥	□	脉细	□
多次流产（>3 次）	□	体倦乏力	□		
烘热汗出	□	烦躁易怒	□		
		眩晕	□		
		胁胀	□		
		耳鸣	□		

小计(A)	___×3 分=___分	小计(B)	___×2 分=___分	小计(C)	___×1 分=___分
总分(A+B+C)		_____分			

Ⅳ．气血两虚证					
定义	气虚血亏，所表现的头晕、神疲、气短、唇淡、舌淡、脉细无力等证候				
必备证素	气虚，血虚		**或兼证素**	脾，阳虚，动血	
常见证候及计量值					
3 分		**2 分**		**1 分**	
头晕	□	神疲	□	失眠	□
月经量少//月经稀淡	□	多梦	□	懒言	□
唇淡	□	健忘	□	月经推迟	□
面色少华//面色萎黄//面色淡白//面色苍白//面色白光白	□	眼花//视物模糊	□	月经量多	□
舌淡	□	心悸//心慌	□	经闭	□
		气短	□	形体消瘦	□
		倦怠乏力	□		
		声低	□		
		口淡//久不欲食	□		
		脉细	□		
		脉虚	□		
小计(A)	___×3 分=___分	小计(B)	___×2 分=___分	小计(C)	___×1 分=___分
总分(A+B+C)		_____分			

Ⅴ．脾胃虚弱证					
定义	中气虚弱，脾胃失运，所表现的口淡、食少，脘腹痞胀、疼痛，呕吐，经常腹泄等证候				
必备证素	脾，胃，气虚		**或兼证素**	气滞，阳虚	
常见证候及计量值					
3 分		**2 分**		**1 分**	
口淡//久不欲食//长期食少//进食无味	□	得食痛缓	□	恶心	□
食后痞胀	□	脘痞胀	□	神疲	□
经常腹泄//经常便溏	□	腹胀	□	形体消瘦	□
		脘腹痛//腹痛	□	舌苔白	□
		隐痛	□		
		痛喜按或按之舒	□		
		嗳气	□		
		呕吐//呕吐清水//呕吐酸水	□		
		气短	□		
		懒言	□		

www.tcmjm.com

		倦怠乏力	□		
		内脏下垂	□		
		面色萎黄	□		
		舌淡	□		
		脉虚	□		
小计(A)	___×3 分=___分	小计(B)	___×2 分=___分	小计(C)	___×1 分=___分
总分(A+B+C)		_____分			

Ⅵ. 肝肾阴虚证			
定义	肝肾阴液亏虚，虚热内扰，所表现的头晕、眼花、耳鸣，腰酸、腰痛，盗汗、手足心烧，遗精等证候		
必备证素	肝，肾，阴虚	**或兼证素**	(热[火])，血虚，动风

常见证候及计量值					
3 分		**2 分**		**1 分**	
耳久鸣//耳失聪=3；	□	视物模糊//眼花	□	胁痛	□
眼干涩	□	腰痛	□	形体消瘦	□
		腰膝酸软	□	多食易饥	□
		口渴//咽干//无热饮多	□	心烦	□
		盗汗	□	失眠	□
		遗精	□	多梦	□
		月经量少	□	健忘	□
		血压高	□	便干结	□
		手足心烧	□	颧红	□
				舌赤	□
				苔剥//苔少//无苔	□
				脉弦	□
				脉细	□
				脉数	□
小计(A) ___×3 分=___分		小计(B) ___×2 分=___分		小计(C) ___×1 分=___分	
总分(A+B+C)		_____分			

附：使用说明

1 常见证的诊断，应具备该证型的主要证候（症状、体征）；总计量值≥诊断阈值；含必备证素。

2 本标准中列有部分同义词，临床诊断时，应使用正名，避免使用同义词。标准中用方括号[ ]所括起的部分，表示放在括号中间的词可以代替部分前面的词； 用圆括号（ ）所括起的部分，表示可以省略；//表示或者，一般只选择其中 1 项。

3 中医辨证应遵循“从症辨证”的原则。症状的全面、真实、客观、规范，是准确辨证的前提。主诉症或多个（≥3 个）症状不能纳入时，不宜按本标准进行辨证。

4 使用时， 首先将患者的每一症状、体征等病情资料，按提示的证型分别进行加权求和（症状轻重的计量：主诉症或症状重，计量值×1.5；症状中等，计量值×1；症状轻，计量值×0.7）。然后对证型作出判断。证型判断一般以 20 作为通用阈值。

5 证型轻重判断标准：总计量值＜14，该证型的诊断不能成立；总计量值 14～20，该证型属Ⅰ（一级，较轻）；总计量值 21～30，该证型属Ⅱ（二级，明显）；总计量值＞30，该证型属Ⅲ（三级，严重）。

6 临床辨证时， 可有多个近似证的计量值达到或超过诊断阈值，一般应取最高计量值的证型作为首选诊断，或将计量值超过通用阈值的几个证型有机整合成一个证名诊断。

7 “常见证”有限，临床上的实际证型则很多。因此，临床运用时，虽然提示某常见证的计量值最高，但若该证中有任 1 项“必备证素”未达到诊断阈值，则仍不得诊断为该常见证，而可能是次高计量值的近似证，或其它非常

见证。

8 辨证提示某常见证的计量值最高， 且其所含必备证素均具备，但若还有其它证素存在，一般属该常见证的“或兼证素”，即兼夹证，一般可在该常见证中加“兼//夹××”；若另有某证型达到诊断阈值，且包含了该证素，则以该常见证作为诊断。

[专科检查]

两侧乳房	□对称 □不对称【若不对称，则：____侧乳头高，____侧乳头缺如】 可见长约_____cm（□纵 □横 □斜 ）手术疤痕。
乳房发育	□正常 □小乳房 □大乳房 □萎缩型 □下垂型 □肥胖型
患侧乳房胸壁皮肤	□正常 □橘皮样改变：部位_____范围_____ □静脉曲张： 部位_____范围_____ □发 红： 部位_____范围_____ □破 溃： 部位_____范围_____ □卫星结节： 部位_____范围_____个数_____
患侧乳头	□正常 □粘连凹陷 □固定 □糜烂 □溃烂 □缺如
肿块	个数_____，位于_____侧_____点钟；瘤缘距乳头缘_____cm； **性质描述：**【□硬 □软 □实性 □囊性】【□边界清楚 □边界不清楚】 【□活动度好 □活动度一般 □活动度差】 【□与胸大肌粘连固定 □与胸壁粘连固定】 【□表面无溃疡 □表面有溃疡，大小_____cm】
同侧腋下淋巴结肿大	□无 □有【若有：大小_____cm，□散在□融合□粘连固定】
同侧锁骨下淋巴结肿大	□无 □有【若有：大小_____cm，□散在□融合□粘连固定】
对侧乳房及腋下淋巴结情况	

2. 问卷调查及数据探索

本用例研究病例来源于 2011 年 6 月至 2012 年 2 月期间于江苏省中医院、江苏省肿瘤医院、南京妇幼保健医院、马鞍山医药集团门诊或住院的乳腺癌患者，各病程阶段三阴乳腺癌样本量 250 例。问卷信息采集者均要求有中医诊断学基础，能准确识别病人的舌苔脉象，用通俗的语言解释医学术语，并确保患者信息填写准确。

1）回收问卷有效性分析

本案例面向各大医疗机构及南京市、区各级癌友协会共发放问卷 250 份，回收 213 份，经剔除标准筛选后符合统计要求问卷 186 份，有效问卷比例为 87.3%，样本数据采集情况见图 2-97。

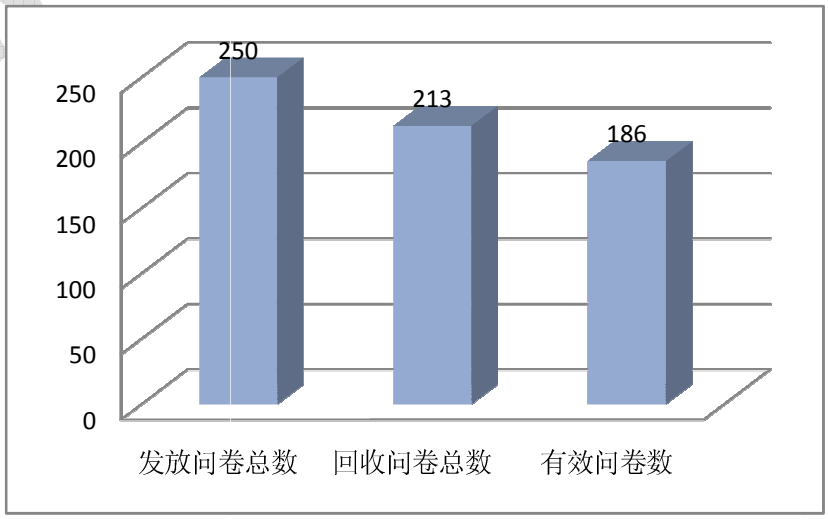


图 2-97 样本数据采集情况

有效问卷分布分析如
图 2-98 所示。

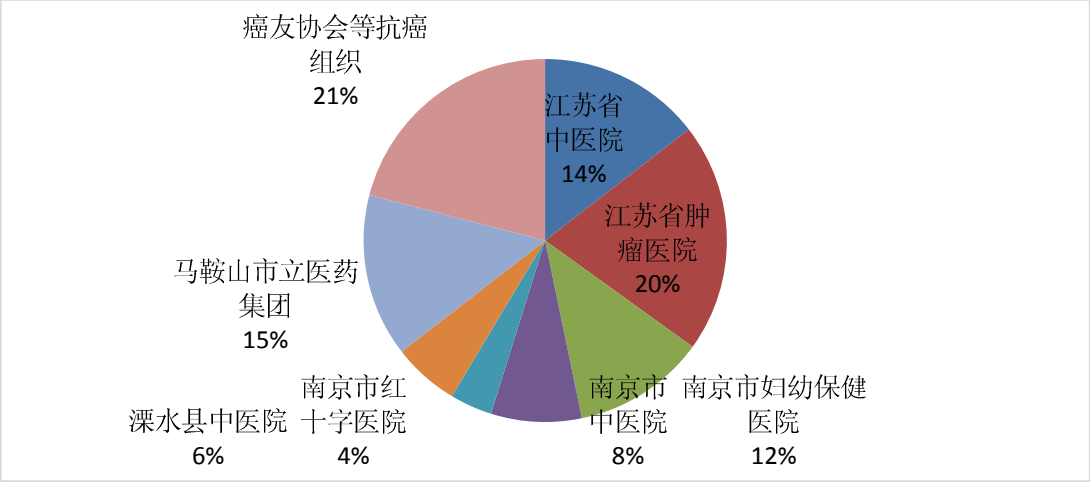


图 2-98 有效问卷来源分布情况

2）数据整理

数据整理是数据挖掘建模过程中最耗时的环节，需要把纸质的原始问卷构造最终可以用于建模分析的数据集。样本数据整理过程包括数据选择、数据清洗、数据构造和数据格式化四个环节，见图 2-99。



图 2-99 样本数据整理过程

回收样本数据集中存在与学习任务不相关的属性或冗余属性，这些属性会导致无效的归纳并降低神经网络的学习效率。数据选择就是将不相关属性和冗余属性剔除的过程。患者基本信息与病程变化规律之间关联性不强，且数据复杂，故样本数据未选择患者基本信息的内容。

数据清洗主要是对异常数据进行检测，并根据所选分析技术和分析目标的要求来提高数据质量。数据清洗主要包括插入适当的默认值来估计缺失值，过滤掉重复的样本数据等。

由于问卷调查采取多中心调研的形式进行搜集，三阴乳腺癌患者群体较小，且回收的问卷有部分无效，采集的样本数据非常有限，有必要对缺失边界的属性进行样本数据构建，以确保样本数据达到具备构建感知模型的有效样本集的条件。构建数据工作主要包括：对缺失边界的属性，可通过

趋势外延法增补样本；对数据过度稀疏的属性，可通过内插法增补样本；增加新的指标因子时，通过肿瘤的访谈及根据建模专家的经验增加样本。经过数据清洗和数据构造后，样本数据剩余 186 条。

数据的格式化是指将样本数据标准化，如样本数值化、规范化，以便构建感知模型的专家样本集。在本课题设计之初，拟采取以证素为单元探讨单一证素与病程变化之间的联系，但由于涉及证素过多，每种证素仅存在是或否两种情况，且病例数相对较少，建模预测时出现了过拟合情况，故在统计学专家及肿瘤学专家的帮助下，课题改用证型相关系数代替具体个别证素进行数据挖掘研究，将单证型的证素评分加权求和除以该证型总分，计算出单一证型的相关系数，然后形成以六种证型相关系数、病程阶段、TNM 分期、转移部位、确诊几年后发生转移为指标的样本集，进行数据挖掘。

3）三阴乳腺癌证素数据探索

本案例所搜集三阴乳腺癌病例 186 例，样本分布情况见图 2-100 至图 2-104。

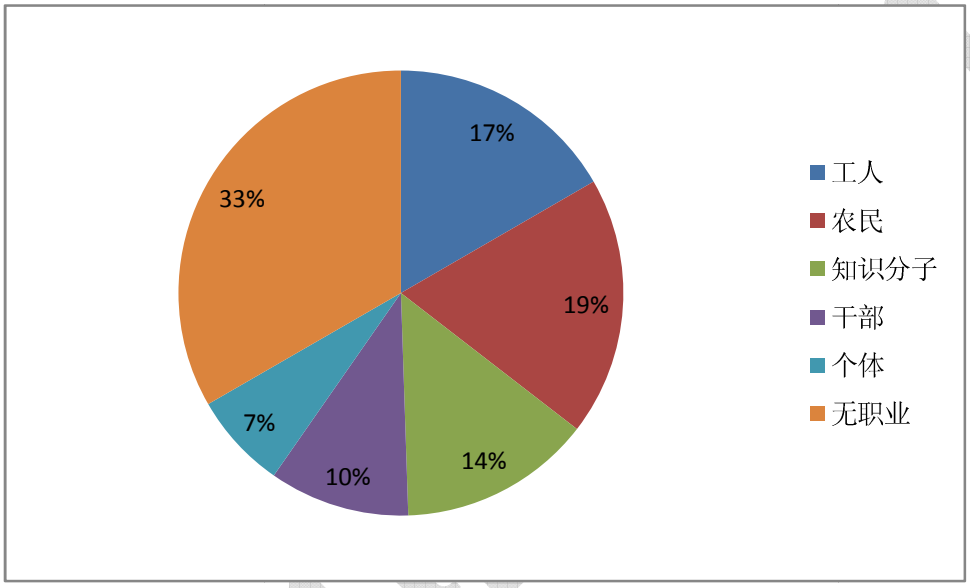


图 2-100 患者职业分布情况

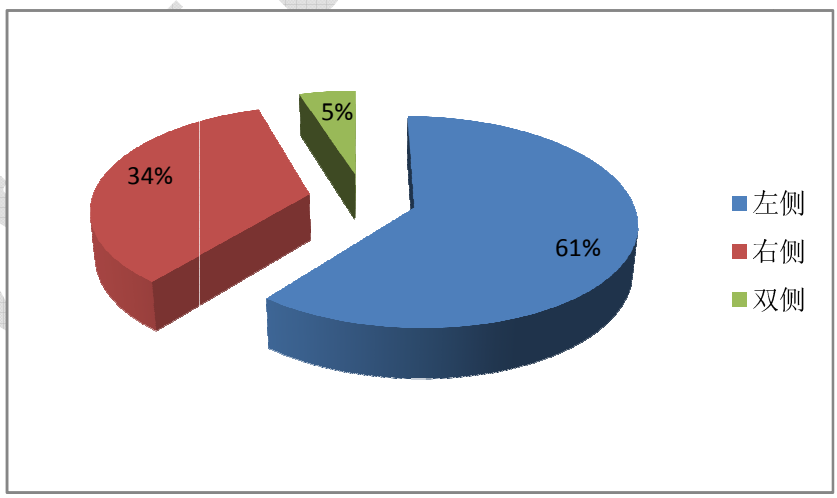


图 2-101 患者发病乳房

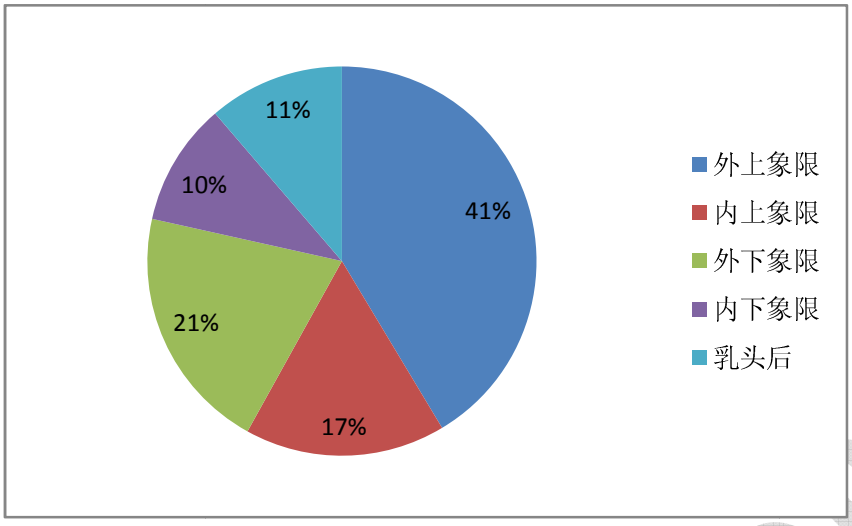


图 2-102 患者肿块部位分布情况

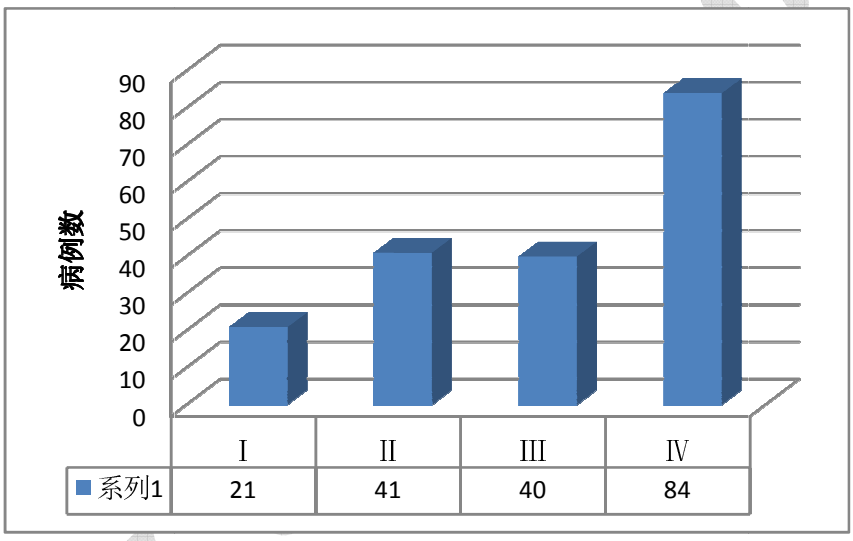


图 2-103 患者 TNM 分期情况

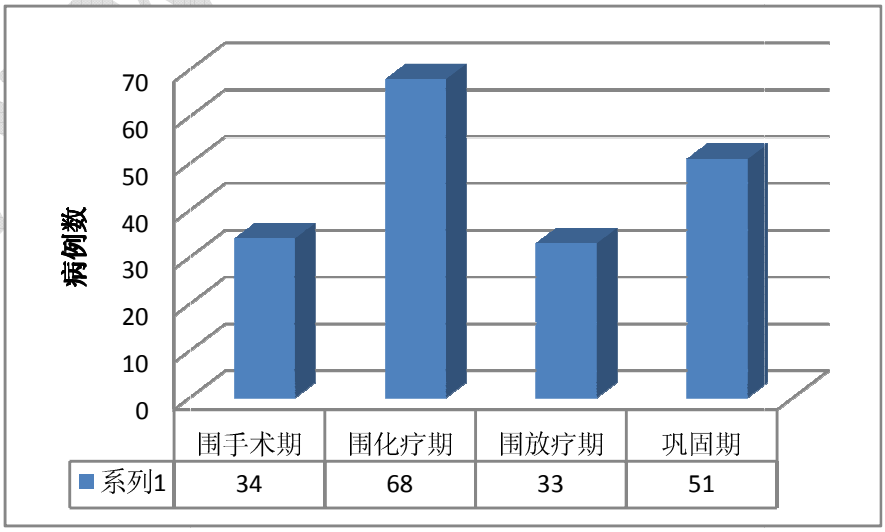


图 2-104 患者病程阶段分布情况

3. 乳腺癌中医证素挖掘分析

（1）关联规则分析：基于样本的统计规律，以进行关联规则挖掘，当置信度达到某一阈值时，就可以认为规则成立。本项目中，通过对中医证素与临床病理及肿瘤生物学特征指标数据进行分析，挖掘出有意义的不同类型、不同病程阶段中医证素的最佳关联模式。

（2）聚类分析：通过数据聚类，分析乳腺癌不同类型、不同病程阶段中医证素分布特点。

（3）分类预测：通过分类与回归预测建模，系统能自动推导乳腺癌从发病到复发转移证素变化规律，提出截断病变，先期干预的治疗思路。

数据挖掘处理过程见图 2-105。

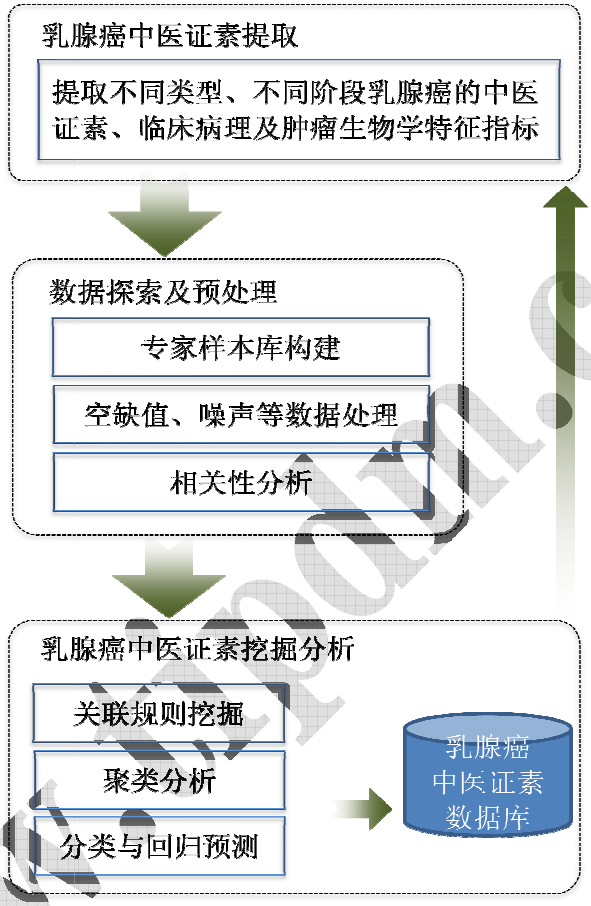


图 2-105 基于数据挖掘技术的乳腺癌中医证素分析流程

仿真建模

本案例采用TipDM数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的相关性分析及HOTSPOT算法探索全部患者TNM分期与中医证型系数之间的关联规则。

4. 模型输入

模型的输入包括两部分，一部分是建模专家样本数据的输入，另一部分是建模参数的输入。

本案例中，通过调查问卷采集的原始样本数据包括实际年龄、发病年龄、初潮年龄、既往月经是否规律、是否痛经、是否绝经、是否有更年期症状、产几胎、流几胎、生育年龄、是否哺乳、哺乳时间、TNM 分期、确诊后几年发现转移、转移部位、病程阶段等 73 项指标，各指标描述见表 2-116。

表 2-116 调查问卷原始指标

序号	指标名称	指标描述

www.tipdm.com

1	实际年龄	A1:≤30 岁；A2：31-40 岁；A3：41-50 岁 A4:51-60 岁；A5：61-70 岁；A6：≥71 岁
2	发病年龄	B1:≤30 岁 B2：31-40 岁 B3：41-50 岁 B4:51-60 岁 B5：61-70 岁 B6：≥71 岁
3	初潮年龄	C1：≤12 岁 C2:13-15 岁 C3：≥16 岁
4	既往月经是否规律	D1：月经规律 D2：月经先期 D3：月经后期 D4：月经先后不定期
5	是否痛经	Y：是 N：否
6	是否绝经	Y：是 N：否
7	是否有更年期症状	Y：是 N：否
8	产儿胎	K1：未生育 K2：单胎 K3：多胎
	……	
70	TNM 分期	I1：I；I2：II I3：III；I4：IV
71	确诊后几年发现转移	1.无转移：BU0 2.小于等于三年：BU1 3.大于三年小于等于五年：BU2 4.大于五年：BU3
72	转移部位	R1：骨 R2：肺 R3：脑 R4：肝 R5：胸膜 R6：纵膈 R0：无转移
73	病程阶段	S1：围手术期 S2：围化疗期 S3：围放疗期 S4：巩固期

对原始指标进行归纳，整理出表 2-117 所示的模型输入变量。

表 2-117 关联规则模型输入变量

序号	变量名称	变量描述/取值范围
1	肝气郁结证型系数	0~1
2	热毒蕴结证型系数	0~1
3	冲任失调证型系数	0~1
4	气血两虚证型系数	0~1

www.tjdm.com

5	脾胃虚弱证型系数	0~1
6	肝肾阴虚证型系数	0~1
7	病程阶段	S1：围手术期；S2：围化疗期； S3：围放疗期；S4：巩固期
8	TNM 分期	H1：I 期；H2：II 期 H3：III 期；H4：IV 期
9	转移部位	R1：骨；R2：肺；R3：脑 R4：肝；R5：其他；R0：无转移
10	确诊后几年发现转移	J0：未转移；J1：小于等于 3 年 J2：3 年以上，小于等于 5 年；J3：5 年以上

表 2-118 HotSpot 关联规则建模参数

序号	参数名称	参数描述
1	最大分支数	最大分支数量
2	最小改进量	增加一个新的分支时最小改进量
3	目标最小化	选择 True 为目标最小化；选择/False 为目标最大化
4	支持度	最小支持度，当值介于 0 和 1 之间时为相对支持度（百分比），当值大于 1 时为绝对支持度（个数）
5	目标	关注的目标属性列，默认为 last
6	目标索引	关注的目标值（仅针对分类型），默认为 first

5. 仿真过程

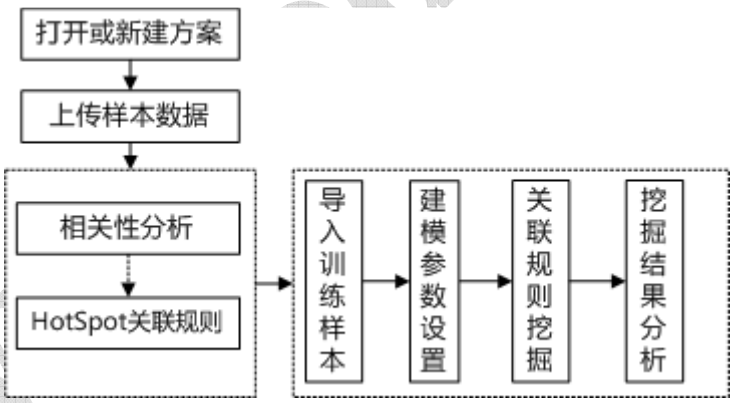


图 2-106 仿真过程

建模仿真过程说明如下：

1） 登录 TipDM 平台，在方案管理页面中，新建方案或者打开一个已建方案。

2） 切换到数据管理页面，上传经预处理后的专家样本数据文件。

3） 选择关联规则菜单中的 HotSpot 关联规则算法。

4） 设置有关参数后，用 HotSpot 关联规则算法对导入的样本数据进行关联规则挖掘。

5） 对挖掘结果进行分析。

6. 结果及分析

1）患者 TNM 分期与中医证型的关联规则

当支持度设定为 0.33 时，用于分段的最小个数为 61，取最大分支系数为 4，得到以下关联规则：

总数： 186 样本数

```
目标属性: TNM分期
目标值: H4 [个数:  84  样本数  (45.16%)]
用于分段的最小个数: 61 instances (33% 占总数)
Maximum branching factor: 4
Minimum improvement in target: 10%

TNM分期=H4 (45.16% [84/186])
  肝肾阴虚证型系数 > 0.276 (66.3% [61/92])
  脾胃虚弱证型系数 > 0.173 (53.98% [61/113])
  肝气郁结证型系数 > 0.207 (52.99% [62/117])
  气血两虚证型系数 > 0.172 (51.16% [66/129])
| 肝肾阴虚证型系数 > 0.153 (57.27% [63/110])
```

186例患者中TNM分期诊断为Ⅳ期的患者有84例，占总样本量的45.16%。186例患者满足肝肾阴虚证型系数＞0.276的有92例，其中有61例TNM分期诊断为Ⅳ期，置信度为66.3%。脾胃虚弱证型系数＞0.173的有113例，其中61例TNM分期诊断为Ⅳ期，置信度为53.98%。肝气郁结证型系数＞0.207的有117例，其中62例TNM分期诊断为Ⅳ期，置信度为52.99%。气血两虚证型系数＞0.172的有129例，其中66例TNM分期诊断为Ⅳ期，置信度为51.16%，在这129例患者中，有110例患者肝肾阴虚证型系数＞0.153，其中的63例TNM分期诊断为Ⅳ期，置信度为57.27%。

以上关联规则提示，TNM 分期为Ⅳ期的三阴乳腺癌患者证型主要为肝肾阴虚证、脾胃虚弱证、肝气郁结证和气血两虚证，Ⅳ期患者肝肾阴虚证和肝气郁结证的临床表现较为突出。此外，通过关联规则挖掘发现，肝肾阴虚证与气血两虚证的置信度达57.27%，提示肝肾阴虚证的Ⅳ期患者有发展为气血两虚证的可能，故在对三阴乳腺癌Ⅳ期肝肾阴虚证的患者施行放化疗等存在骨髓抑制风险的治疗时要注重保护患者的造血功能，中药治疗可先期予以补髓养血之品，减少骨髓抑制的发生，降低骨髓抑制的程度。

1）患者病程与中医证型的关联规则

当支持度设定为0.25时，用于分段的最小个数为47，取最大分支系数为2，得到以下关联规则：

```
总数:  186 样本数
目标属性: 病程阶段
目标值: S4 [个数:  51  样本数  (27.42%)]
用于分段的最小个数: 47 instances (25% 占总数)
Maximum branching factor: 2
Minimum improvement in target: 10%

病程阶段=S4 (27.42% [51/186])
  脾胃虚弱证型系数 <= 0.293 (37.9% [47/124])
  热毒蕴结证型系数 <= 0.351 (31.41% [49/156])
  |  脾胃虚弱证型系数 <= 0.372 (34.78% [48/138])
```

当支持度设定为0.25时，用于分段的最小个数为47，取最大分支系数为2，得到以下关联规则：186例患者中处于巩固期的患者有51例，占总样本数的27.42%。186例患者满足脾胃虚弱证型系数≤0.293的有124例，其中有47例处于巩固期，置信度为37.9%。热毒蕴结证型系数≤0.351的有

156 例，其中有 49 例处于巩固期，置信度为 31.41%。138 例脾胃虚弱证型系数≤0.372，其中有 48 例处于巩固期，且同时满足热毒蕴结证型系数≤0.351，置信度为 34.78%。

以上关联规则提示，病程阶段处于巩固期的患者，较少表现为脾胃虚弱证和热毒蕴结证。一般，若患者出现患处红肿热痛等明显的热毒蕴结症状会主动求医，在巩固期患者因化疗等引起的胃肠道反应也会逐渐减轻，故以上关联规则符合临床实际。

3）患者复发转移时间与中医证型的关联规则

当支持度设定为 0.25 时，用于分段的最小个数为 47，取最大分支系数为 2，得到以下关联规则：

```
总数:  186 样本数
目标属性: 确诊后几年发现转移
目标值: J1 [个数:  53  样本数  (28.49%)]
用于分段的最小个数: 47 instances (25% 占总数)
Maximum branching factor: 2
Minimum improvement in target: 10%

确诊后几年发现转移=J1 (28.49% [53/186])
  肝肾阴虚证型系数 > 0.217 (38.84% [47/121])
  肝气郁结证型系数 > 0.166 (32% [48/150])
  肝肾阴虚证型系数 > 0.166 (37.6% [47/125])
```

186 例患者中在 3 年内发生转移复发的有 53 例，占总样本数的 28.49%。186 例患者满足肝肾阴虚证型系数＞0.217 的有 121 例，其中有 47 例 3 年内出现肿瘤转移复发，置信度为 38.84%。肝气郁结证型系数＞0.166 的有 150 例，其中有 48 例 3 年内出现肿瘤转移复发，置信度为 32%。125 例患者肝肾阴虚证型系数＞0.166，3 年内出现肿瘤复发转移，且同时满足肝气郁结证型系数＞0.166，置信度为 37.6%。

以上关联规则提示，表现为肝肾阴虚和肝气郁结证的三阴乳腺癌患者 3 年内肿瘤转移复发的风险较高，以肝肾阴虚证的临床表现更为明显，且肝肾阴虚证与肝气郁结证可并行出现。肝、肾为预防三阴乳腺癌 3 年内转移复发的主要病位证素。

4）Ⅳ期患者各中医证型的关联规则

提取Ⅳ期复发转移患者共 84 例作为样本用 HOTSPOT 算法挖掘各中医证型系数之间的关联规则。

以肝肾阴虚证型系数为目标属性，探索Ⅳ期患者肝肾阴虚证型系数与其他中医证型的关联规则如下：

```
总数:  84 样本数
目标属性: 肝肾阴虚证型系数
Target average in total population: 0.317
Minimum segment size: 28 instances (33% 占总数)
Maximum branching factor: 2
Minimum improvement in target: 10%

肝肾阴虚证型系数 (0.3171)
热毒蕴结证型系数 > 0.277 (0.3551 [28])
```

www.ilqm.com

当支持度设定为 0.33，最大分支系数为 2，探索各证型相关系数之间的关联时，取目标属性为肝肾阴虚证，得到以下关联规则：84 例转移复发患者中，有 28 例满足肝肾阴虚证型系数＞0.317，且热毒蕴结证型系数＞0.277，置信度为 35.51%，提示热毒蕴结证的三阴乳腺癌患者可能存在肝肾阴虚证，故中医治疗上施以清热解毒治法的同时当兼顾滋养肝肾，顾护阴液。

以冲任失调证型系数为目标属性，探索Ⅳ期患者冲任失调证型系数与其他中医证型的关联规则如下：

```
总数:  84 样本数
目标属性: 冲任失调证型系数
Target average in total population: 0.252
Minimum segment size: 28 instances (33% 占总数)
Maximum branching factor: 2
Minimum improvement in target: 10%

冲任失调证型系数 (0.252)
  肝肾阴虚证型系数 > 0.365 (0.2873 [29])
  肝气郁结证型系数 > 0.268 (0.2821 [29])
```

当支持度设定为0.33，最大分支系数为2，探索各证型相关系数之间的关联时，取目标属性为冲任失调证，得到以下关联规则：84例转移复发患者中，有29例满足冲任失调证型系数＞0.252，且肝肾阴虚证型系数＞0.365，置信度为28.73%；有29例满足冲任失调证型系数＞0.252，且肝肾阴虚证型系数＞0.268，置信度为28.21%。提示冲任失调与肝肾阴虚、肝气郁结相关程度高，患者肝肾阴虚证的临床表现更为明显。

以气血两虚证型系数为目标属性，探索Ⅳ期患者气血两虚证型系数与其他中医证型的关联规则如下：

```
总数:  84 样本数
目标属性: 气血两虚证型系数
Target average in total population: 0.227
Minimum segment size: 28 instances (33% 占总数)
Maximum branching factor: 4
Minimum improvement in target: 10%

气血两虚证型系数 (0.2271)
  脾胃虚弱证型系数 > 0.306 (0.2709 [28])
  肝气郁结证型系数 <= 0.232 (0.2516 [31])
  热毒蕴结证型系数 <= 0.152 (0.251 [28])
  肝肾阴虚证型系数 <= 0.289 (0.2506 [28])
```

当支持度设定为0.33，最大分支系数为4，目标属性为气血两虚证时，探索各证型相关系数之间的关联，得到以下关联规则：84例转移复发患者中，有28例满足气血两虚证型系数＞0.227，且脾胃虚弱证型系数＞0.306，置信度为27.09%；有31例满足气血两虚证型系数＞0.227，且肝气郁结证型系数≤0.232，置信度为25.16%；有28例满足气血两虚证型系数＞0.227，且热毒蕴结证型系数≤0.152，

置信度为25.1%；有28例满足气血两虚证型系数＞0.227，且肝肾阴虚证型系数≤0.289，置信度为25.06%。提示三阴乳腺癌转移患者中，气血两虚证于脾胃虚弱证关联程度较高，而表现为气血两虚证的患者其肝气郁结证、热毒蕴结证、肝肾阴虚证的临床表现常不明显。

以热毒蕴结证型系数为目标属性，探索Ⅳ期患者热毒蕴结证型系数与其他中医证型的关联规则如下：

```
总数:  84 样本数
目标属性: 热毒蕴结证型系数
Target average in total population: 0.235
Minimum segment size: 28 instances (33% 占总数)
Maximum branching factor: 4
Minimum improvement in target: 10%

热毒蕴结证型系数 (0.2348)
  肝肾阴虚证型系数 > 0.357 (0.2939 [33])
  脾胃虚弱证型系数 <= 0.231 (0.2755 [42])
  |   肝肾阴虚证型系数 > 0.311 (0.3238 [29])
  |   脾胃虚弱证型系数 > 0.131 (0.3038 [31])
  冲任失调证型系数 > 0.272 (0.2696 [38])
```

当支持度设定为 0.33，最大分支系数为 2，目标属性为热毒蕴结证时，探索各证型相关系数之间的关联，得到以下关联规则：84 例转移复发患者中，有 33 例满足热毒蕴结证型系数＞0.235，且肝肾阴虚证型系数＞0.357，置信度为 29.39%；有 42 例满足热毒蕴结证型系数＞0.235，且脾胃虚弱证型系数≤0.231，置信度为 27.55%；有 38 例满足热毒蕴结证型系数＞0.235，且冲任失调证型系数＞0.272，置信度为 26.96%。提示热毒蕴结证与肝肾阴虚证、冲任失调证的出现关联程度高，热毒蕴结证的患者脾胃虚弱症状不明显。

5）Ⅳ期患者复发转移时间与各中医证型的关联规则

将所有出现复发转移的患者单独提取作为一个集合，运用 HOTSPOT 算法探索Ⅳ期患者复发转移时间与中医证型系数之间的关联规则，规则挖掘结果如下：

```
总数:  84 样本数
目标属性: 确诊后几年发现转移
目标值: J1 [个数:  53  样本数  (63.1%)]
用于分段的最小个数: 28 instances (33% 占总数)
Maximum branching factor: 4
Minimum improvement in target: 10%

确诊后几年发现转移=J1 (63.1% [53/84])
  热毒蕴结证型系数 > 0.213 (84.85% [28/33])
  脾胃虚弱证型系数 <= 0.231 (76.19% [32/42])
  冲任失调证型系数 > 0.235 (70.45% [31/44])
  肝肾阴虚证型系数 > 0.294 (70.37% [38/54])
```

当支持度设定为0.33，最大分支系数为4，目标属性为转移复发时间时，探索84例已经出现肿瘤转移复发的三阴乳腺癌患者各证型相关系数与转移复发时间之间的关联，得到以下关联规则：84例

转移复发患者中，有53例出现在确诊后3年内，占转移复发者总数的63.1%，84例患者中有33例热毒蕴结证型系数＞0.213，其中28例同时满足转移复发时间在3年内，热毒蕴结证型系数＞0.213与3年内出现转移复发的置信度为84.85%；84例患者中有42例脾胃虚弱证型系数≤0.231，其中32例同时满足转移复发时间在3年内，脾胃虚弱证型系数≤0.231与3年内出现转移复发的置信度为76.19%；84例患者中有44例冲任失调证型系数＞0.235，其中31例同时满足转移复发时间在3年内，冲任失调证型系数＞0.235与3年内出现转移复发的置信度为70.45%；84例患者中有54例肝肾阴虚证型系数＞0.294，其中38例同时满足转移复发时间在3年内，肝肾阴虚证型系数＞0.294与3年内出现转移复发的置信度为70.37%。

以上关联规则提示，三阴乳腺癌转移复发的高峰在确诊后的3年内，3年内转移复发的患者的中医证型主要为热毒蕴结证、冲任失调证和肝肾阴虚证。对比186例患者各证型系数与转移复发时间的关联规则，可以看出，热毒蕴结证和冲任失调证在186例三阴乳腺癌患者中分布较少，但集中出现在3年内转移复发的患者群里，因此热毒蕴结、冲任失调、肝肾阴虚是治疗早期转移复发患者的三个主要靶点，而肝肾阴虚和肝气郁结又是预防三阴乳腺癌患者出现早期转移复发的关键。

2.5.2. 案例十七 卷烟消费者购买行为分析

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

长期以来，烟草企业一直都是国家的利税大户，在各级政府的可支配收入中，约有 8%的份额是由烟草行业提供的。2012 年上半年，全国共实现财政收入 63795 亿元，其中税收收入 54931.63 亿元，而中国烟草共实现工商税利 4807.16 亿元，其中税收 3286.79 亿元(不含烟叶特产税及企业所得税)，占全国的 5.98%,另外，在全国国有及规模以上非国有工业企业(年销售收入在 500 万元以上)中，其中销售收入的近 2%、利税总额的近 4%都是来自于中国烟草工业。因此，可以说，卷烟企业一直是各地区及国家的纳税大户。

我国是世界上人口最多的国家，同时烟草消费量也是世界上最大的国家，有着近 3.2 亿的烟民大军，无疑这是一个非常庞大的市场。但随着中国市场的逐步开放，中国卷烟的体制性保护和地方壁垒也正逐渐消减，特别是为适应世贸组织的要求，如自 2004 年开始我国即已取消特种烟草专卖零售许可证，所有卷烟零售商既可经营国产卷烟，也可经营外国卷烟。卷烟行业的竞争可以说是已经开始销烟四起，众多卷烟厂商不断面临跨区域和跨国际的市场况争。在这样的大背景下，作为割据的地方卷烟企业该何去何从?作为不算强大的本土卷烟产业又该如何应对激烈的外来竞争?“工欲善其事，必先利其器”，在新的形势下，必须要有新的方法来应对。

本案例的挖掘目标在于：

1）在细分市场差异不明显的情况下，通过数据挖掘的方法实现市场细分；

2）基于组合数据挖掘技术，提取出客户特征规则，为有针对性的营销活动提供依据；

3）应用先进的建模工具和统计的方法对卷烟消费者行为进行研究分析。

分析过程与方法

消费者行为分析，可以根据卷烟行业消费者行为特征，设计出问卷，进行问卷调查，通过问卷调查分析，应用据挖掘工具建立模型，对所收集而来的信息和知识进行挖掘建模，进而提取出有意义的特征或规律。如图 2-107 所示，具体包括如指标的建立、模型构建和结果评估三部份。

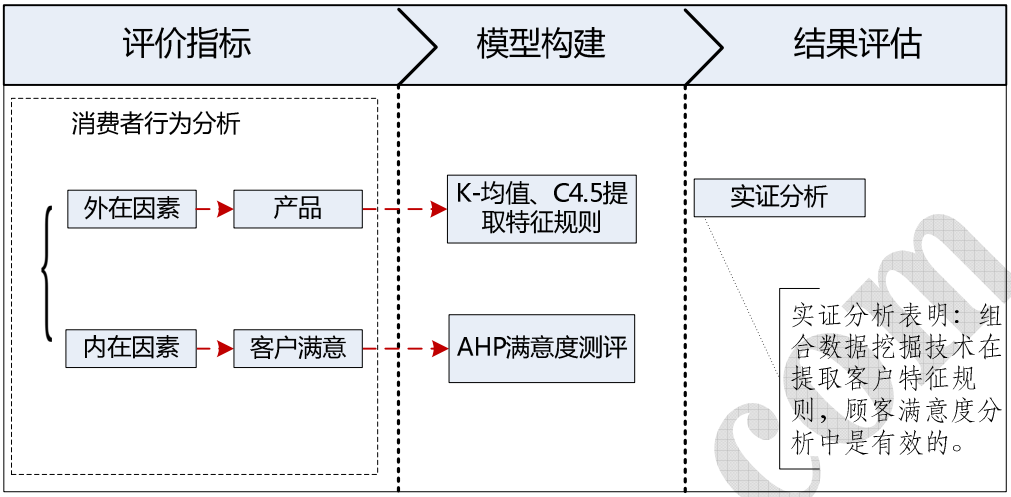


图 2-107 分析过程与方法

1. 确定分析指标、方法

在商品品牌多样化和市场化的今天，消费者作为理性的主体，不可能凭空做出自己对某商品的购买决策，他们的购买决策在很大程度上受到文化、社会、个人和心理等各种因素的影响，如图 2-108 所示。一个人的购买行为是文化、社会、个人和心理因素之间相互影响、相互作用的结果。

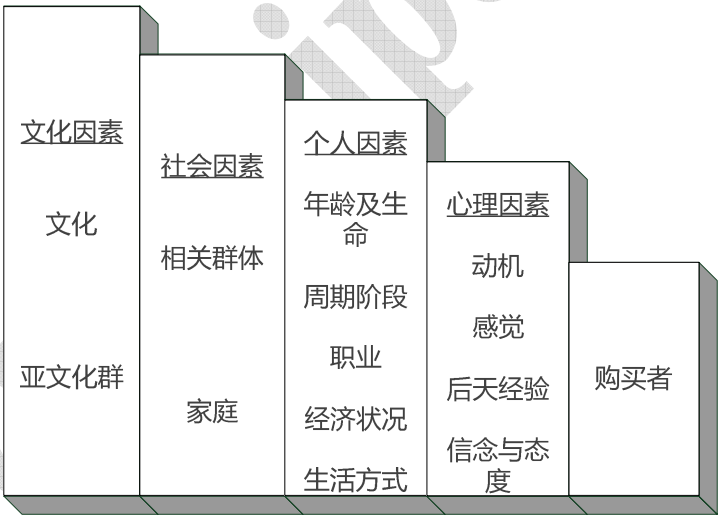


图 2-108 购买决策影响因素

文化、社会、个人和心理等因素影响着消费者的行为，而这些因素又由很多的子因素组成，因此，众多因素之间构成消费者行为的复杂性客观问题。

针对具体的卷烟市场，我们从影响消费者行为的两种刺激即内部刺激和外部刺激，分别加以探讨。

（1）外在因素

一般来说，消费者的购买决策随其购买决策类型的不同而变化。较为复杂和花钱多的决策往往凝结着购买者的反复权衡和众多人的参与决策，这种购买过程相对繁琐，需要购买者收集信息，评

价方案等。较为简单和花钱少的决策，消费者通常购买频繁，并且只花最少的精力和最少的时间去比较品牌、价格。香烟作为一种便利品，消费者的购买决策相对简单，选择一种品牌或放弃一种品牌都相对容易。因此，卷烟企业对顾客需求认识上的细微偏差都有可能改变顾客的选择，而这种偏差会产生剧烈的马太效应，则能根本改变企业的命运。对于影响到消费者是否对某品牌卷烟购买的影响因素，可以将其细分如下表 2-119 所示:

表 2-119 影响购买决策因素细分

序号	购买影响因素
1	自己习惯的牌子
2	追求新鲜，尝试新口味
3	同价位比较，品牌名气大，名烟
4	包装精美
5	喜欢硬盒装
6	喜欢软盒
7	烟丝色泽金黄、有油润感、烟梗少
8	烟丝包裹松紧度适中
9	朋友、家人影响
10	烟劲大、味浓烟味清淡
11	醇和，不燥
12	卖烟人推荐
13	广告影响
14	焦油含量、烟碱量低

那么这些因素对卷烟消费者的影响程度如何?不同的卷烟消费者是否存在共有的主要影响因素?目前国际上先进的数据挖掘方法，从输入输出关系上对规则进行探讨，能较好地处理这一类的复杂性客观问题。

（2）内在因素

内在因素主要从顾客对产品的满意程度来分析，ISO9000:2000 版对“顾客满意”的定义为:顾客对其要求已被满足的程度感知。通过对一个产品或一项服务的可感知效果（感知质量）与他的期望（认知质量）相比较后形成的感觉状态，即顾客对其要求已被满足的程度的感受。顾客如果对此产品满意，则在下次购买活动中可能继续购买该产品，并向其他人宣传该产品的优点；如果对产品不满意，则有可能放弃继续购买此产品，而转向别的品牌。

不同的企业或产品，影响顾客满意度的关键指标是不同的，这些关键指标通常是顾客对该产品期望较高或抱怨较多的，也是企业应特别重视的测评指标。香烟作为一种便利品，消费者的购买决策往往相对简单，通常来说，口感、烟丝质量、包装、价格，四个指标基本上是顾客重点考虑的对象。目前，国际上流行的顾客满意度模型是费纳理论模型，如下图 2-109 所示。

www.tipm.com

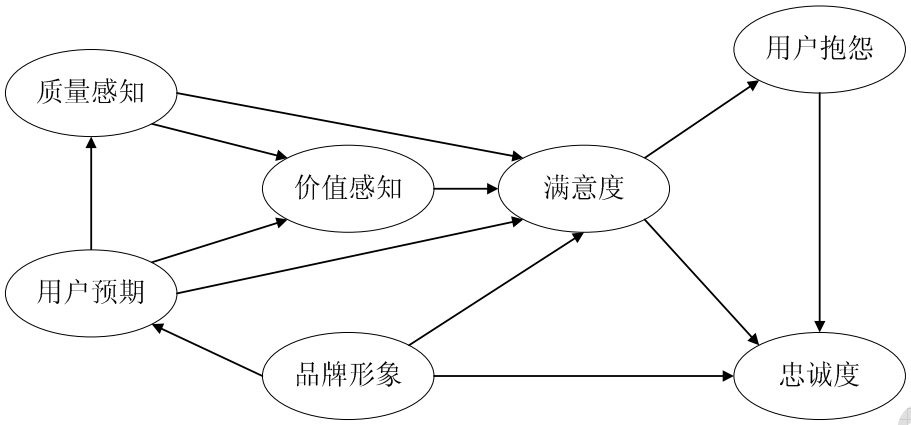


图 2-109 费纳理论模型

该模型是由 6 个结构变量构成的因果关系模型。其中，用户满意度是最终追求的目标变量:品牌形象、质量感知、用户预期和价值感知是影响客户满意度的四个前置因素；用户抱怨和顾客忠诚则是顾客满意度的结果变量。顾客满意度指标的分析必须是从顾客的角度，从顾客实际使用的感受中，从顾客的需求和期望中去判断。同时，消费者也是一个复杂的群体，单一的、定性的评价指标体系往往不能很好地了解消费者。因此有必要建立一套科学的评价体系，采用科学的评价方法对不同特征的消费者进行客观评价。

2. 问卷设计、调查

问卷设计是根据调查的目的和要求，将需要调查的问题具体化，从而使研究者能顺利地取得必要的数据、信息，以便于分析决策。

通常，问卷设计首先需列出要收集信息的清单、确定问卷的类型与实施方法，然后设计出每个问题的内容及回答的先后顺序，接着确定问题的格式及排版，试答并作必要的修正，最后付印。我们在本次卷烟市场的调查中设计的问题如下(因篇幅关系，不一一列出)：

Q1.受访者性别（访问员根据观察填写:［1]男 ［2]女）

Q2.年龄段:（[1] 18 岁以下；[2] 18—29 岁；[3] 30—39 岁；[4] 40—49 岁；[5] 50—59 岁 ；[6] 60 岁以上）

Q3.文化程度:（[1] 大专以下；[2] 大专；[3]本科；[4]研究生及以上）

Q4.职业:（[1]务农；[2]个体户；[3]机关事业单位；[4]企业管理人员；[5]工人；[6]自由职业者；[7]学生；[8]待业、失业 ； [9]其他（请注明；））

Q5.您为什么最经常买 XX 牌子的烟呢？**(多选题，限选 3 项)**（[1]卷烟口味适合；[2]一直习惯抽；[3]品牌高档/名气大；[4]价格合理；[5]身边很多人抽；[6]卷烟包装精美；[7]容易买到；[8]广告多；[9]个性特别；[10]其他 1（请注明：_____）；[11]其他 2（请注明：_____））

Q5.1. 其中最主要的原因是：_____

设计好问卷后，我们便可以对卷烟消费者进行问卷调查了，通过问卷调查的方式，收取到我们所要了解掌握的信息和知识。

3. 数据预处理

www.cndm.com

由于现实世界的数据一般是脏的、不完整的和不一致的，因此，对问卷调查所收集起来的数据资料，有必要进行预处理，转换成数据挖掘工具所能分析处理的数据类型。数据预处理涉及的范围比较繁杂，一般包括：数据清理、数据集成、数据变换、归约等相关工作。

4. 建立模型

利用第（2）、（3）步所得到的数据资料，结合相应的消费者研究需求，构建数学模型，从而挖掘出所需的消费者规则。

5. 结果评估

根据数据挖掘模型得到的输出结果，从实证中分析检验，进行结果评估。

挖掘建模

1. 特征规则挖掘

下面说明通过两个模型（K-均值、C4.5 决策树）对消费者行为分析的挖掘建模。模型结果为将客户细分为多个类别，每个类别有不同的特征，细分后即可以根据实际情况对感兴趣的类别提取出主要特征，从而可对感兴趣的客户进行市场策划，制定不同的管理目标和执行方案。

模型的建立过程如下图所示：

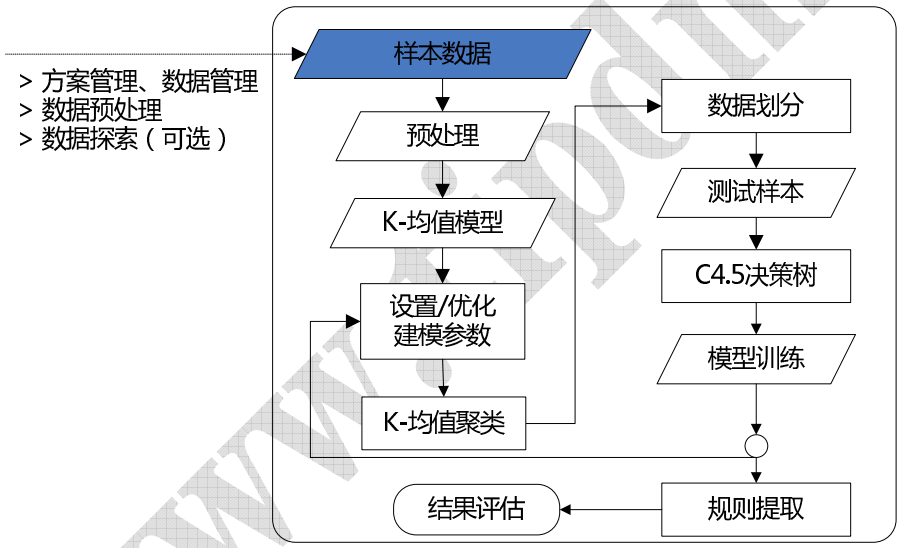


图 2-110 建模过程

由于调查收集到的数据信息中会存在一些噪声或缺损，因此在建模前，我们首先需对要分析的数据进行预处理，包括如一些异常值的处理、缺损数据的填补等。然后就是对不同群体进行划分，即市场细分了，市场细分可根据固有的条件属性如区域、性别等对客户进行分类，也可应用数据挖掘的方法对客户实现分类，这里主要应用 K-均值的方法进行市场细分，最后，根据有分类标记的数据应用如决策树等方法提取出每一类别的特征。下面简要介绍。

随机选取问卷调查得到的 100 个样本数据，对消费者按特征进行聚类，采用欧氏距离作为相似性度量，运用 K-均值算法将样本聚成了两类，第一类（0 类）有 33 个样本，第二类（1 类）共 65 个样本，得到的聚类结果下表（注：未全部列出，另两份问卷因缺损数据太多，未加入分析。）

表 2-120 K-均值聚类

id	region	S1	S2	D1	Q1.1	Q3_Fixed	Q9	Q9.2	class

A0622079	6	1	6	5	1	2	1	4	0
080082	8	1	3	5	1	2	1	22	0
A0422033	4	1	6	5	4	10	4	4	0
A0206058	2	1	3	5	1	1	1	3	1
A0310243	3	1	4	5	1	10	2	2	1
A0310242	3	1	4	5	1	10	2	2	1
A0734010	7	1	3	5	1	1	6	5	1
A0205061	2	1	4	5	1	1	2	3	1
A0424006	4	1	6	5	1	30	1	2	1
A0914013	9	1	3	5	1	2	2	3	0
A0736008	7	1	2	5	1	1	2	3	1
080029	8	1	3	5	1	1	1	2	1
080027	8	2	2	5	1	1	1	2	1
A0312095	1	1	5	5	1	1	3	3	1
A0734012	7	1	2	5	1	1	1	2	0
A0207029	2	1	4	5	1	1	1	2	1
A0916085	9	1	2	5	1	1	1	2	1
A0916060	9	1	3	5	1	2	1	2	1

下一步，根据已有形成类别数的数据表，将经过概括的数据转变为特征规则，多种数据挖掘方法如模糊规则归纳方法、神经网络方法及决策树的方法等均可以实现。这里我们主要采用决策树的方法来实现客户的特征规则提取。

将聚类后的数据输入后，选取对应的决策树方法，分别选取好训练集与检测集的样本数据，设置好相关参数值后执行，即可生成输出报表。如考察样本数最多的类别 2 生成的规则，第 2 类中一共有 65 条样本，选中输出报表窗口并拖动滚动条，可看到 Class 2 的规则列表。交换重设好数据样本的训练集与检测集，重复执行，可挖掘出列表所给出了其所有规则，随机选取第 2 类的二条特征规则如下：

规则 1. $Q3\_BuyBox=1$ *and* $D1=5$: *rule accuracy* 97.12% : *rule* coverage 93.80%

规则 2. $Q3\_BuyBox=1$ *and* $Q9.2=2$: *rule accuracy* 95.68% : *rule* coverage 50.51%

如对规则 1 的完整表述应为：

$$IF\ Q_3\_BuyBox=1$$
$$and\ D1=5$$
$$Then\ \ class=2$$
$$rule\ accuracy\ 97.12\%$$
$$rule\ \ coverage\ 93.80\%$$

对规则 1 的解释为：如果通常每次购买 1 包卷烟，且本人或其家人从事的是在“烟草生产”、“ 香烟配送”、“ 香烟专卖”、“ 广告、市场调研、营销策划等行业”之外的其它行业，属于第 2 类消费者，该规则 97.12%是正确的，规则覆盖度说明该规则 93.8%适用于第 2 类消费者。

对规则 2 的解释为：如果通常每次购买 1 包卷烟，且购买品牌卷烟的原因是“一直习惯抽”， 属

www.tpdm.com

于第 2 类消费者，该规则 95.68%是正确的，规则覆盖度说明该规则 50.51%适用于第 2 类消费者。

这样，通过特征规则挖掘可以看出“0”、“1”两类消费者特征规则是不同的，对于第一类消费者（“0”类，占 33.67%）,他们认为烟草特性方面如口味浓、劲大等对于吸烟比较重要；对于第 2 类消费者（“1”类,占 66.33%）则对于焦油含量、烟碱量、口味等内在品质不太重视，他们对价格也不敏感而且一直习惯自己的品牌。这说明，对于第一类消费者，可以在产品口味上注意，适当偏重，提升内在品质；而对于第二类消费者，可以从香烟外在如包装等方面入手，突出香烟的品味，包装上讲究精美，并且在产品质量上着重如烟灰等这些影响外观感受的因素上，促销手段上多下功夫，多进行一些如文艺促销活动等。

通过上面分析，我们可以看到数据挖掘方法能很好地用于提取消费者群体的特征，直接使用单一的数据挖掘方法提取特征规则会碰到一些问题，即在细分市场差异不明显的情况下不能保证模型有好的效果，模型精度较低。因此，有必要改变消费者群体的划分方法，K-均值聚类正好提供了这样一个平台，通过聚类方法客观地细分消费者，这样大大提高了 C4.5 决策树方法的实用性与有效性。

2. 满意度测评

（1）AHP 法确定指标权重

这里采用层次分析法（AHP 方法）对顾客的满意程度进行分析。层次分析法可以通过测评指标两两比较，使复杂的无序的定性问题量化。表 2-121 为层次分析法对重要性程度的划分情况。

表 2-121 重要性程度划分表

标度	含义：i 方案(因素)比 j 方案(因素)
1	同等重要
3	稍重要
5	重要
7	明显重要
9	绝对重要
2,4,6,8	介于上述两个相邻等级之间
1/2,1/3,…,1/9	I 方案(因素)比 j 方案(因素)重要性地位互换

该表反映了测评指标之间两两比较相对重要程度的得分，如设测评指标 i 相对测评指标 j 的比较得分为 a_{ij}，则指标 j 相对 i 的比较得分为 $a_{ji}=1/a_{ij}$

通过对消费者调查数据的统计，口感、烟丝质量、包装、价格四个指标的重要性程度如下表 2-122 所示：

表 2-122 测评指标比较矩阵

指标	价格	口感	包装	烟丝质量
价格	1	1	3	1
口感	1	1	3	1
包装	1/3	1/3	1	1
烟丝质量	1	1	1	1

根据 AHP 原理计算出矩阵权重系数：

$$w_i = \frac{\sqrt[n]{\prod_j a_{ij}}}{\sum \sqrt[n]{\prod_j a_{ij}}}$$

（2-6）

运算过程如下表 2-123 所示：

表 2-123 运算过程

指标	相乘	开方	权重
价格	1*1*3*1	$\sqrt[4]{1*1*3*1}$ =1.316	0.3127
口感	1*1*3*1	$\sqrt[4]{1*1*3*1}$ =1.316	0.3127
包装	1/3*1/3*1*1	$\sqrt[4]{1/3*1/3*1*1}$ =0.577	0.1371
烟丝质量	1*1*1*1	$\sqrt[4]{1*1*1*1}$ =1	0.2376

比较矩阵的一致性检验：

$$AW=\begin{bmatrix}1.2743\\1.2743\\0.5832\\1.0000\end{bmatrix} \qquad \lambda_{\max}=\frac{1}{n}\sum_i\left[\frac{(AW)_i}{w_i}\right]=4.1532$$

利用$\lambda_{\max}$与阶数 n 之差检验一致性$C.I.$，考虑到 n 的影响，使用随机一致性比值：

$$C.R.=\frac{C.I.}{R.I.} \tag{2-7}$$

其中，$R.I.$为平均随机一致性指标，可以由相应指标表查得；$C.I.$可以通过公式：

$$C.I.=\frac{\lambda_{\max}-n}{n-1} \tag{2-8}$$

求出。当$C.R.<0.1$时，比较矩阵的一致性是可以接受的。

以上计算出 *C.I.*=0.0511；*C.R.*=0.0574<0.1，说明比较矩阵的一致性是可以接受的。即通过层次分析法所确定的各测评指标的权重是客观、可接受的。反映出顾客对卷烟商品满意的几个综合指标价格、口感、包装、烟丝质量权重分别为 0.3127、0.3127、0.1371、0.2376。

（2）综合满意度计算

通过获得的统计数据资料，我们就可以计算出每一个指标的顾客满意度指数，并结合指标权重计算出此品牌卷烟的综合满意度指数了。

首先，依据加权平均法分别计算出每个指标的满意度：

$$S_i=\frac{1}{n}\times\sum A_{ij} \tag{2-9}$$

其中，S_i表示第i（i=1,2,3,4,5)个指标的顾客满意度指数；n表示回收的顾客调查数；A_{ij}表示第i个指标第j个顾客的评分。

然后 ，根据每个指标的顾客满意度及权重，计算综合的顾客满意度指数，公式如下：

$$S=\sum \lambda_i S_i \tag{2-10}$$

其中，λ_i 表示第 i （i=1,2,3,4,5)个指标的权重；S_i 表示第 i 个指标的顾客满意度指数；S 表示产品综合满意度指数。

根据上面的算法步骤及统计调查数据，计算得到的卷烟品牌顾客满意度指数如下表所示。

表 2-124 满意度计算

指标	价格	口感	包装	烟丝质量
权重	0.3127	0.3127	0.1371	0.2376
S_i	76.2	79.8	72.5	75.1

因此，综合满意度指数 S 为：

$$\begin{aligned} S &= \sum \lambda_i S_i = \lambda_1 S_1 + \lambda_2 S_2 + \lambda_3 S_3 + \lambda_4 S_4 \\ &= 0.3127\times 76.2 + 0.3127\times 79.8 + 0.1372\times 72.5 \\ &\quad + 0.2376\times 75.1 \\ &= 76.6 \end{aligned}$$

AHP 分析问题大体经过以下四个步骤：(1)分析问题，确定系统中各因素之间的关系，建立系统的递阶层次结构模型；(2)构造两两比较的判断矩阵；(3)计算各层次相对权重的单排序；(4)计算各层元素对系统的合成权重，并进行总排序。如图 2-111 所示：

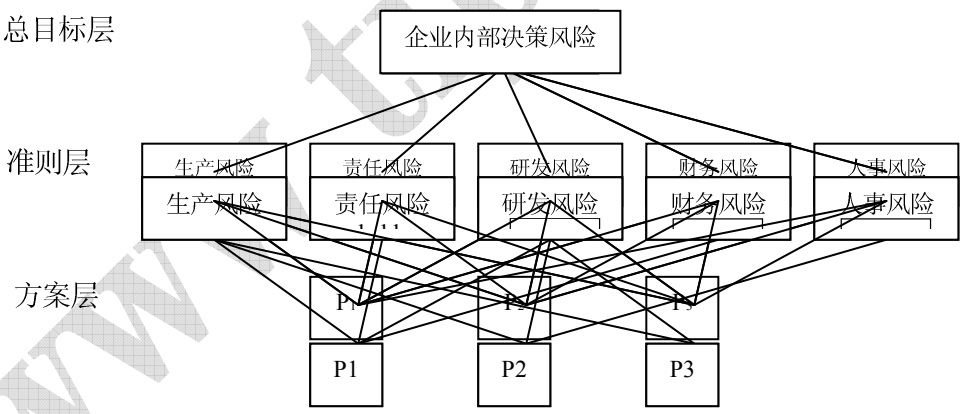


图 2-111 企业内部决策风险权重合成示意图

模型目的在于从企业内部不同的决策方案中找出最优，从而使风险最小而收益最大。企业内部决策方案的制定大概要基于以下三方面分析，即企业资源分析、核心能力分析和业务外包分析。

(1)企业资源分析：指找出企业的战略资源，制定资源的重组方案，对未来潜在的战略资源进行激发；从而确立企业的长远规划。企业资源包括物质资源（如工厂位置、设备的精良程度、原材料获取途径等)、人力资源（员工培训、判断力、承担义务及忠诚等)、技术资源（专利、商业秘密等）等，可分为有形资源和无形资源。

(2)核心能力分析：又称专长、核心竞争力，也是企业最根本、最关键的经营能力，或者说，也只有在本企业中，这种竞争能力与优势才能得到最充分的发挥。凭借这种最根本、最关键的经营能

力，企业才拥有自己的市场与效益。核心能力分析主要是明确企业的核心优势，具体包括为：企业现有优势和专长的保护和发展，对市场的反映，竞争者情报、动向等分析。

(3)业务外包分析：企业业务外包符合经济学中的“木筒理论”，也是近十年来国外兴起的管理新理论。企业竞争力是由产、供、销等多个环节共同决定的，企业要将每个环节都做得一样好，不太可能，于是，外包现象由运而生。企业能否成功经营很大程度取决于合作伙伴即协力厂商的经营能力。欧美企业“外包”规模年增长率达 35%。高技术企业，特别是信息技术企业“外包”比例是最大的，几乎占总“外包”的 30%，人们预测几年以后高科技企业将会达到 50%。业务外包已成为一种趋势，因此，哪些业务由自己做，哪些业务“外包”出去？经常会成为企业经营决策考虑的对象。

3. 基于风险最小的方案确定

根据上述计算结果，CR 值均小于 0.1,通过一致性检验。可将表 2 中计算出的各风险排序值看作风险权重，应用加权平均，计算出三个方案各自的综合风险排序值，从而找出风险最小的决策方案，详见表 2-125 所示：

表 2-125 方案综合评价

内部风险	a	b	c	d	e	综合排序值
风险权重	0.118	0.056	0.348	0.323	0.155	
P1	0.188	0.701	0.672	0.637	0.353	0.556
P2	0.731	0.106	0.063	0.258	0.061	0.207
P3	0.081	0.193	0.265	0.105	0.586	0.237

由表 2-125 的计算结果易知，综合排序值最小的是方案 P2，为 0.207，即基于风险最小考量的决策方案是 P2。

为进一步掌握市场规律，展开有针对性的营销活动，某烟草公司在市场调研活动中，重点研究不同消费者群体的一些特征规则。通过设计问卷，对卷烟消费者展开调查，同时，在收集到相关消费者信息、数据后，应用一种多用途的数据挖掘工具 ESX 提取市场中不同类别消费者的一些特征，从而为公司的市场决策提供了参考。

2.5.3. 案例十八 纳税人偷漏税评估

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

某国税纳税评估部门每年必须定期开展纳税评估，评估部门年初根据当年的工作情况、工作重点，制定当年纳税评估计划，确定当年纳税评估的重点行业和重点企业。同时根据其它职能部门的工作需要制定专项纳税评估，纳税评估实施流程见图 2-112，具体纳税评估工作程序如下：

1、资料复核

每月、每季、每年结束后，对当期的申报资料进行复核，对申报资料不完整的，通过资料纠正

程序加以纠正，确保当期纳税评估资料的完整。

2、制定纳税评估方案

根据年初纳税评估计划的工作重点和要求，对全体纳税人制定评估方案，产生各行各业的评估基本方案、基础的纳税人疑点底册，供各个职能部门开展工作使用。

做好其它职能部门要求开展的专项评估。

3、对由纳税评估部门产生的疑点纳税人，根据底册情况，选择一定数量的疑点纳税人，通过评估系统传送到税收管理员系统或稽查系统，作为税收管理员开展质疑约谈对象和税务稽查选案案源。

4、对移送税收管理员和稽查的疑点纳税人，做好跟踪落实，质疑约谈或检查完成的，做好纳税评估信息反馈工作。

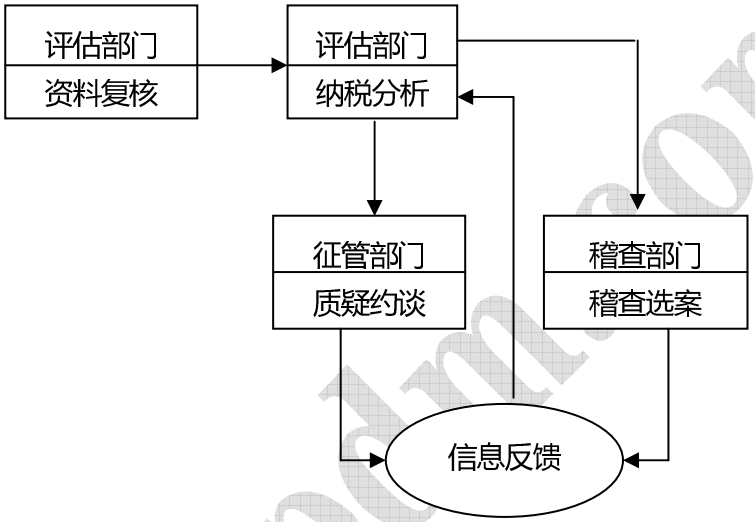


图 2-112 纳税评估实施流程图

随着国家对金税工程的大量投入及网络技术的大量运用，税务系统建立了全国的网络运行平台，每天都产生 TG 级的数据。这些数据的产生对于税务部门评估税收质量及发现纳税人欺诈提供了可能。

但税收分析，目前还停留在以现有知识逻辑为基础的简单分析，主要是通过历史数据的纵向比较和地区数据的横向比较，或者是一些全局性的税收或经济指标与局部数据的比较以发现问题。近年来，很多税务部门通过引进风险管理理论和流程再造理论，建立了风险特征库和优化了管理流程。

这些分析方法与管理措施虽然对发现问题和改进管理发挥了一定的积极作用，但由于是建立在已有经验上的管理，传统的数据库技术及数据库管理系统的传统方法虽然可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。以某国税局为例，2012 年税收总量将超过 200 亿，一般纳税人达到两万多，管理总量接近十万户，每天都会产生大量的数据。并且这些数据仅仅只是管理对象在申报纳税过程中申报或者在征管过程中形成的，申报的数据具有很多欺诈、隐瞒的成份，录入的数据存在许多的杂音、杂质。并且这样的数据与纳税人所拥有的全部可以证明纳税真实程度的信息存在着巨大的信息不对称，虽然政府也成立了协护税网络及情报交换平台，但由于规则的不健全及缺乏制约机制，信息的获取极不正常也极其困难。

由于数据的缺失和失真，这些数据不经清洗整理的简单使用，会导致税务部门的判断失误。因此税收征管人员不仅要知道数据库中都有什么（What happened），这些，传统的查询和报表工具就会告诉他们，他们还想知道下一步会怎么样（What next）、和如果采取这样的措施又会怎么样（What if）。不仅如此，相关人员更需要更进一步知道这些海量数据背后所隐藏的秘密，不仅要用已有的知识去管理，还需要向数据学习，由数据提供新的知识。

如何有效评估税源质量，提高评估效率，做到应收尽收？对纳税人的纳税申报情况之真实性和准确性做出的定性和定量判断，为纳税补查和更正补税提供依据？

本案例的挖掘目标是：

1) 基于不同行业形成有效的纳税评估体系，并确立纳税评估模型；
2) 借助纳税评估管理，挽回因复杂的偷税行为而导致的经济损失：偷税行为作为纳税评估问题的主要形式之一，导致国家税收蒙受巨大的经济损失。基于数据挖掘技术的纳税评估系统可为发现偷税行为提供自动化、智能化的手段，也为追回税收提供坚实的数据支撑。
3) 借助纳税评估管理，降低征管纳税人员成本的投入：目前的税收稽查模式严重依赖于技术骨干的经验判断，人员需求大、工作效率低、准确性不高。“专家库诊断”模式的引入可将专家经验存储在计算机数据仓库中并“学以致用”，利用计算机程序替代人为经验判断，可极大提高纳税评估效率，降低人员成本。

分析方法与过程

本案例提出的纳税评估流程如图 2-113。

www.tipdm.com

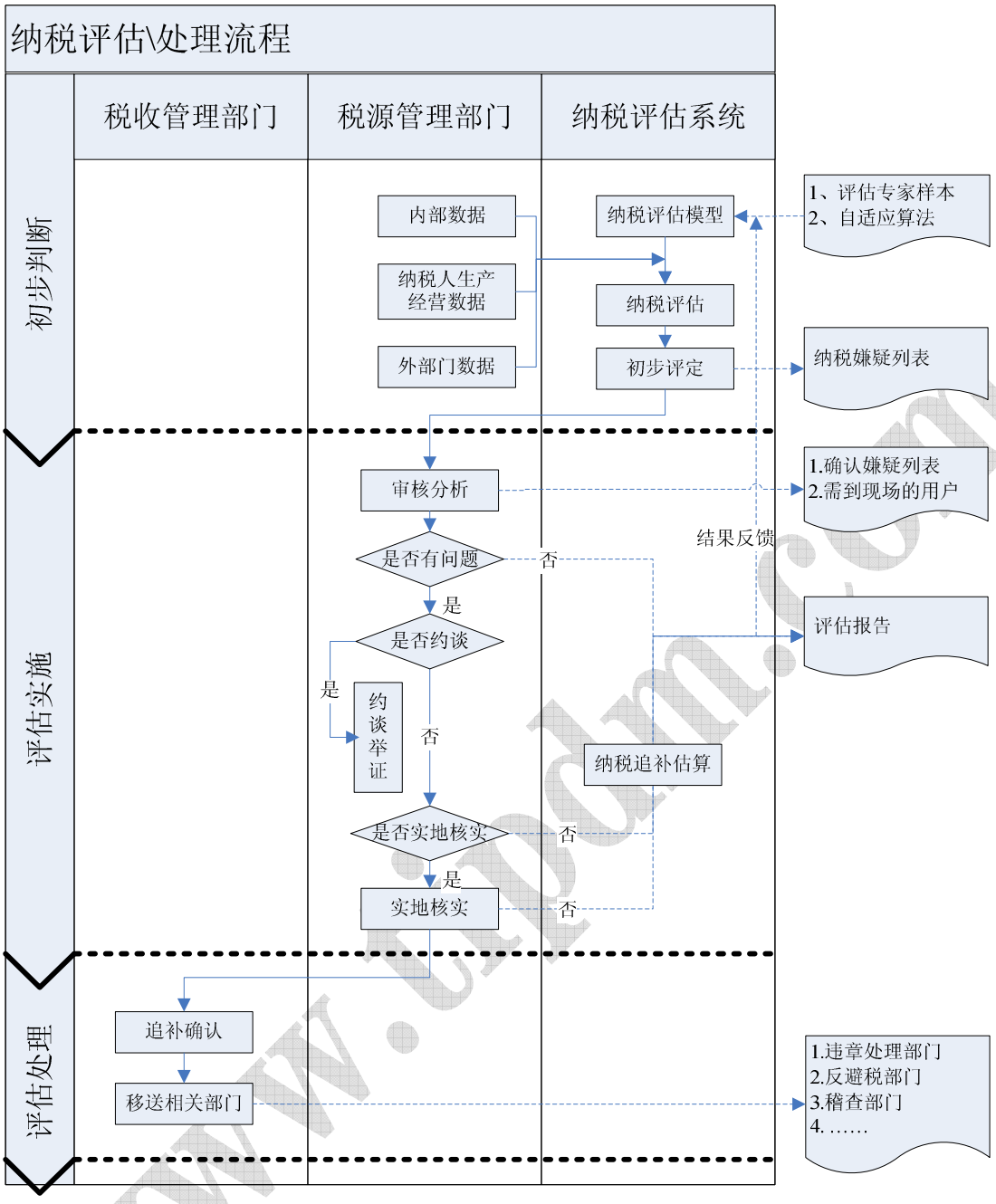


图 2-113 纳税评估流程

1. 纳税评估指标

纳税评估工作并不是简单的计算工作，开展纳税评估，首先需要明确评估的目的，根据工作要求制定纳税情况分析方案，税务机关筛选评估对象、进行重点分析的主要指标就是纳税评估指标。

通常纳税评估指标分为通用分析指标和特定分析指标两大类，通用分析指标是指各税种的评估可选用的指标；特定分析指标是指根据各税种的特点选取的，适合特定税种评估分析的指标。在实际评估工作中，通常会将有关指标进行配比分析。

在与业务人员深入沟通的基础上，初步确立了纳税评估指标体系，指标体系涵盖了不同的行业类别，共 7 个一级指标和 19 个二级指标，如图 2-114。

www.updm.com

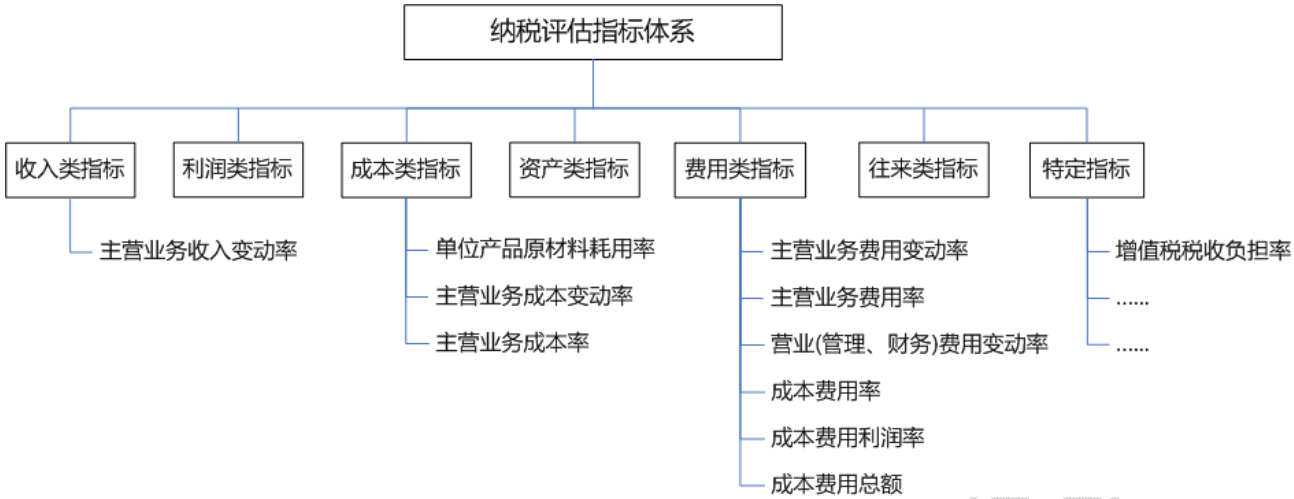


图 2-114 纳税评估指标体系

针对具体行业，评价指标根据实际情况可能会有所调整。另外，在很多时候，纳税评估因为缺乏直接证据，无法证明纳税人申报材料有虚假，而不能对它进行追缴，过往的很多个案表明，往往需要从第三方获得数据支持，如：获得对疑点纳税人举报线索，疑点纳税人在国税的纳税情况，通过银行协助获取疑点纳税人的资金往来状况，通过调查疑点纳税人所在行业的行业发展状况等等，对疑点纳税人的经营状况合理性进行综合评判，才能够发现纳税人疑点。如果条件许可，在构建纳税评估体系时，可考虑将第三方数据纳入考虑范围内。

2. 数据来源及建模专家样本库

纳税评估的主要数据来源包括：“一户式”存储的纳税人各类纳税信息资料，主要有：纳税人税务登记的基本情况，各项核定、认定、减免缓抵退税审批事项的结果，纳税人申报纳税资料，财务会计报表以及税务机关要求纳税人提供的其他相关资料，增值税交叉稽核系统各类票证比对结果等；

税收管理员通过日常管理所掌握的纳税人生产经营实际情况，主要包括：生产经营规模、产销量、工艺流程、成本、费用、能耗、物耗情况等各类与税收相关的数据信息；

上级税务机关发布的宏观税收分析数据，行业税负的监控数据，各类评估指标的预警值；

本地区的主要经济指标、产业和行业的相关指标数据，外部交换信息，以及与纳税人申报纳税相关的其他信息。

本例采用分类回归模型实现纳税评估模型的构建，这就要求每条专家样本数据除了有输入，还必须有已知的输出，即该纳税人在当前输入条件下，是否偷漏税？偷漏税嫌疑系数是多少？这些数据需要由征管稽查人员（人员）给出，而实际情况往往是纳税评估因为缺乏直接证据，无法证明纳税人申报材料有虚假，这就使得专家样本数据准备成了模型构建最困难的环节，具体操作过程中，可通过以下技术手段对历史数据进行初步评定：

1) 一般同类企业的纳税率，在正常经营的情况下，是相同或偏差不大的，考察有无偷税漏税情况。如果企业税负太低，则存在偷漏税嫌疑。
2) 企业的经营都是按步就班进行的，因此可建立自回归模型对企业的收入进行预测，考察有无偷税漏税情况。如果企业的营业收入低于模型的预测值过多，则存在偷漏税嫌疑。
3) 可建立多元回归模型对企业的成本进行预测，考察有无偷税漏税情况。如果企业的成本高于模型的预测值过多，特别是管理费过高、运费过高，则存在偷漏税嫌疑。
4) 根据企业规模、生产经营的特点，建立存货预测模型，考察有无偷税漏税情况。如果企业

存货过多，则存在偷漏税嫌疑。

5) 用单因子分析方法分别将销项税额、进项税额、应纳税额、运费发票、一般固定资产与其他各个指标进行相关性分析，然后对每个年度的相关性作对比，考察有无偷税漏税情况。

6) 建立纳税额与资产、负债和利润之间的回归模型，同时考察存在交互效用的多因素影响效应分析，并建立剔除交互效应显著的因素后的回归模型，从实际数据和理论上对模型进行检验，考察有无偷税漏税情况。如果企业的税收低于模型的预测值过多，则存在偷漏税嫌疑。

7) 如果怀疑企业有偷税漏税情形，可采用偏相关分析方法对影响纳税额的因子进行分析或典型相关分析方法对影响纳税额的因子进行分析，从而找出可能偷税漏税的因子。

8) 如果怀疑企业有偷税漏税情形，可采用主成分分析方法对影响纳税额的因子主成分进行分析，从而找出可能偷税漏税的成分。

9) 如果怀疑企业有偷税漏税情形，可采用数据删除的统计诊断的方法，检验过去可能有偷税漏税年份。

通过技术手段对历史数据进行初步评定后，再由富有经验的税收征管人员对数据逐条进行审核，即可形成建模用的专家样本数据集。

3. 纳税评估模型

1）纳税评估特点

在纳税评估中，由于税种多样、行业众多，需要考虑地区的主要经济指标、产业和行业的相关指标数据，外部交换信息，以及与纳税人申报纳税相关的其他信息等多种因素，但百变不离其宗，所有因素信息都会反映到与其相关的纳税评估指标上，建立完善的纳税评估指标后，通过对这些指标进行归纳并构建智能决策模型，对于解决纳税判别问题具有明显的优势。

2）纳税智能诊断

根据纳税评估结果辨识，对各类纳税的状况进行统计分析,给出相应的指标，标准，参数参考，同时，对于辨识系统积累的历史数据，可以利用数据挖掘技术将这些数据中蕴藏着的许多潜在的重要因素、事实和关联等有价值的信息提炼出来，分析纳税个案中指标的相关性，并进而构建智能诊断模型，真正做到高效、实时地进行智能诊断。

为了确保纳税的准确性和可应用性，本案例采用基于数据挖掘技术的纳税诊断模型。根据宏观税收分析和行业税负监控结果以及相关数据设立评估指标及其预警值；综合运用各类对比分析方法筛选评估对象；对所筛选出的异常情况进行深入分析并作出定性和定量的判断。

3）模型构建

纳税评估样本数据来自《广州市国家税务局纳税评估优秀案例集》中近 50 个用例，共涉及工业企业（建材企业、服装纺织品企业、石油化工企业）、商业企业、出口企业等行业，经数据预处理后，将其中的 45 个样本数据用于预测建模，保留 5 个用例用于模型验证。本例采用 ANFIS 自适应神经模糊推理系统来构建纳税评估模型，ANFIS 自适应神经模糊推理系统模型构建流程如图 2-115。

www.tibm.com

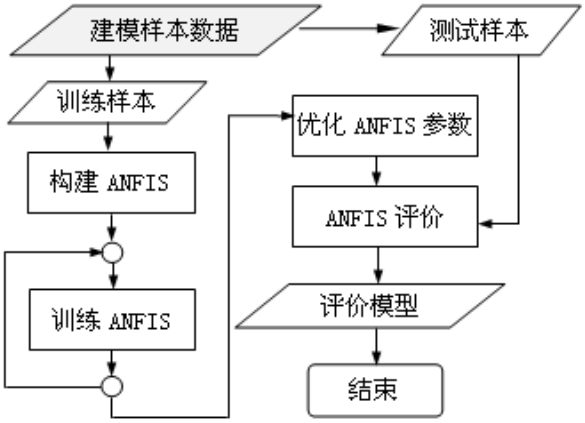
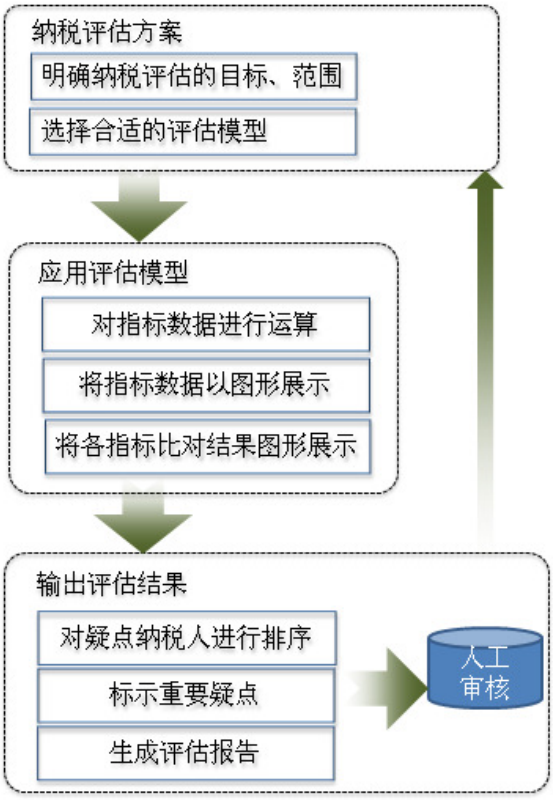


图 2-115 模型构建流程图

实际建模时，可通过多种模型建模方法进行效果对比，以寻找一个适合特定样本数据的评价模型。

4. 纳税评估模型应用



当纳税评估模型建立之后，可将模型应用到纳税评估实际操作中。模型实现与“大集中”的数据仓库对接，从该数据仓库中提取数据，模型自动对指标数据按照国家税法规定的公式进行计算，自动验证企业数据的合理性，得出纳税人嫌疑指数，对不合理的结果进行提示。自动将疑点纳税人与同行业、同规模、同区域的其它纳税人的各项指标进行对比，并以图像展示，方便评估操作人员及时发现疑点，明确下一步工作方向和工作重点。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的属性选择、决策表和 CART 决策树完成模型构建。由于不同行业的特殊性，本例中针对汽车销售行业，构建纳税

评估模型，样本数据集共有 124 个实例，每个实例包含 15 个属性。

1. 模型输入

模型的输入包括两部分，一部分是建模样本数据的输入，另一部分是建模参数的输入。

表 2-126 评估模型输入变量

序号	变量名称	变量描述
1	纳税人名称	不参与建模
2	年度	不参与建模
3	销售模式	输入变量，4S 店、一级代理商、二级及二级以下代理商
4	销售类型	输入变量，国产轿车、进口轿车、大客车、商用货车、卡车及轻卡、工程车等
5	维修毛利	输入变量
6	企业维修收入占销售收入比重	输入变量
7	增值税税负	输入变量
8	存货周转率	输入变量
9	成本费用利润率	输入变量
10	整体理论税负	输入变量
11	整体税负控制数	输入变量
12	办牌率	输入变量
13	单台办牌手续费收入	输入变量
14	代办保险率	输入变量
17	保费返还率	输入变量
18	是否偷漏税	输出变量，偷漏税、不偷漏税

表 2-127 CART 决策树建模参数

序号	参数名称	参数描述
1	最小数	节点的分支数
2	种子数	初始随机选取的样本数

表 2-128 决策表建模参数

序号	参数名称	参数描述
1	交叉验证折数	设计交叉验证折数，为 1 时表示弃一法交叉验证
2	显示规则	设置是否显示规则输出
3	评价措施	用来评估决策表中使用的属性组合性能的措施
4	搜索方法	在决策表中寻找最佳属性组合的搜索方法
5	用 IBk 代替多数分类	设定是否 IBK 代替多数分类

2. 仿真过程

www.ilpan.com

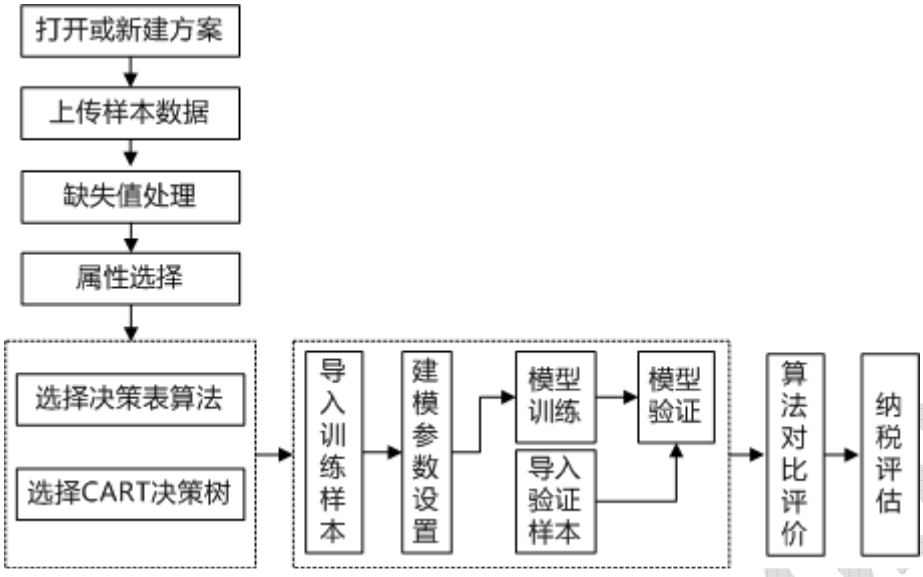


图 2-116 仿真过程

建模仿真过程说明如下：

1） 登录 TipDM 平台，在方案管理页面中，创建针对汽车销售行业的纳税评估模型。

2） 切换到数据管理页面，上传专家样本数据文件。

3） 选择缺失值处理功能，导入样本数据进行缺失值处理，并替换原专家样本数据。

4） 选择属性选择功能，导入样本数据，尝试选择不同的属性评估器和搜索策略组合进行属性选择。

5） 分别选择 CART 决策树算法和决策表算法，进行模型构建。

6） 对比 CART 决策树算法和决策表算法的建模结果，并选择最优算法。

7） 用最优法对测试样本进行纳税评估。

3. 结果及分析

1）缺失值处理

从抽取数据来看，样本数据集中存在着很多残缺值，某些属性的缺失率高达 24%，因此在进行相关数据挖掘方案前，需对该数据集进行预处理，以保证样本数据的质量，从而为后续的纳税评估建模打下基础。

2）特征属性的选择

从税源管理部门收集的数据会涉及到很多的属性，他们对纳税评估的作用不是完全对等的，一些属性对结果的影响占主导地位，一些属性对结果的影响不大，甚至没有影响，故需要采用相应的算法，对数据的属性值进行评估，并去掉那些对评估无太大影响的属性，从而减少后续挖掘算法的运行时间，同时也能有效地去除数据中含有的噪声数据。

表 2-129 基于信息增益选出的纳税评估特征子集结果

属性 评估 (supervised, Class (nominal): 15 输出): Gain Ratio feature evaluator
Ranked attributes: 0.5359 10 整体税负控制数 0.3755 4 维修毛利 0.3661 5 企业维修收入占销售收入比重

0.1996	7 存货周转率
0.1818	13 代办保险率
0.1562	3 汽车销售平均毛利
0.1162	9 整体理论税负
0.0894	2 销售模式
0.037	1 主要汽车销售类型
0	14 保费返还率
0	12 单台办牌手续费收入
0	8 成本费用利润率
0	11 办牌率
0	6 增值税税负
选择的属性: 10,4,5,7,13,3,9,2,1,14,12,8,11,6 : 14	

表 2-130 基于卡方算法选出的纳税评估特征子集结果

属性 评估 (supervised, Class (nominal): 15 输出): Chi-squared Ranking Filter	
Ranked attributes:	
79.814	10 整体税负控制数
58.404	4 维修毛利
56.239	5 企业维修收入占销售收入比重
29.808	7 存货周转率
25.253	3 汽车销售平均毛利
24.754	2 销售模式
16.52	9 整体理论税负
13.784	1 主要汽车销售类型
9.918	13 代办保险率
0	8 成本费用利润率
0	11 办牌率
0	6 增值税税负
0	12 单台办牌手续费收入
0	14 保费返还率
选择的属性: 10,4,5,7,3,2,9,1,13,8,11,6,12,14 : 14	

表 2-131 基于 FCBF 算法选出的纳税评估特征子集结果

属性 评估 (supervised, Class (nominal): 15 输出): Symmetrical Uncertainty Ranking Filter	
Ranked attributes:	
0.5359	10 整体税负控制数
0.3749	4 维修毛利
0.3687	5 企业维修收入占销售收入比重
0.1967	7 存货周转率
0.1544	3 汽车销售平均毛利
0.1155	13 代办保险率
0.1122	2 销售模式
0.1067	9 整体理论税负
0.0515	1 主要汽车销售类型
0	11 办牌率
0	12 单台办牌手续费收入

www.tipdm.com

0	8 成本费用利润率
0	14 保费返还率
0	6 增值税税负
选择的属性：10,4,5,7,3,13,2,9,1,11,12,8,14,6 ： 14	

综合表 2-129~表 2-131 的特征子集选择结果，最后选取整体税负控制数、维修毛利、企业维修收入占销售收入比重、存货周转率、汽车销售平均毛利、代办保险率、整体理论税负等指标作为纳税评估的评价指标。

2）决策表

运用决策表算法根据整体税负控制数、维修毛利、企业维修收入占销售收入比重、存货周转率、汽车销售平均毛利、代办保险率、整体理论税负的数据信息，对纳税评估进行预测建模。采用 10 折交叉验证对样本数据进行总体评价，结果。

表 2-132 决策表分层交叉验证误差

评价项目	评价结果
正确分类样本及百分比	109 87.9032 %
错误分类样本及百分比	15 12.0968 %
Kappa 统计	0.7534
平均绝对误差	0.2009
均方根误差	0.3279
相对绝对误差	41.0171 %
相对平方根误差	66.2534 %

表 2-133 决策表分层交叉验证详细信息

正确肯定率	错误肯定率	精确率	反馈率	ROC Area	类别
0.887	0.132	0.9	0.887	0.849	正常
0.868	0.113	0.852	0.868	0.849	异常
Weighted Avg.					
0.879	0.124	0.879	0.879	0.849	

表 2-134 决策表分层交叉验证混淆矩阵

a b	← 分类为
63 8	a = 正常
7 6	b = 异常

2）CART 决策树

运用 CART 决策树算法根据整体税负控制数、维修毛利、企业维修收入占销售收入比重、存货周转率、汽车销售平均毛利、代办保险率、整体理论税负的数据信息，对纳税评估进行预测建模。采用 10 折交叉验证对样本数据进行总体评价，结果。

表 2-135 CART 决策树分层交叉验证误差

评价项目	评价结果
正确分类样本及百分比	109　87.9032 %
错误分类样本及百分比	15　12.0968 %
Kappa 统计	0.7534
平均绝对误差	0.1928
均方根误差	0.3288
相对绝对误差	39.359 %
相对平方根误差	66.4421%

表 2-136 CART 决策树分层交叉验证详细信息

正确肯定率	错误肯定率	精确率	反馈率	ROC Area	类别
0.887	0.132	0.9	0.887	0.841	正常
0.868	0.113	0.852	0.868	0.841	异常
Weighted Avg.					
0.879	0.124	0.879	0.879	0.841	

表 2-137 CART 决策树分层交叉验证混淆矩阵

a　b	← 分类为
63　8	a = 正常
7　46	b = 异常

4）建模结果比较

比较决策表和 CART 决策树针对总体样本数据的误差信息，总体评价效果相近，但决策表的 ROC Area 稍大，说明决策表算法评价效果稍好。

2.5.4. 案例十九 城市综合环境质量评价

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

环境质量评价是我国实施的重要的环境管理手段之一。环境质量评价就是按照一定的标准和方法对某一区域环境质量的优劣进行定量的或定性描述的一种方法理论，其评价对象是环境质量及其价值。环境质量评价是一个理论与实践相结合的适用性强的学科，是人们认识环境的本质和进一步保护与改善环境质量的手段与工具。

环境质量评价的根本目的就是为各级政府和有关部门制定经济发展计划，制定能源政策，确定大型工程项目及为区域规划提供环境保护的依据，并为各级环境部门制定环境规则，贯彻以管促治方针，为实现全面、科学的环境管理服务。因此，环境质量评价是帮助本例协调经济发展和保护环境的一项有效措施，也是强化环境管理的有效手段，它为环境管理、环境工程、环境标准制订、环境规划、环境污染综合防治、生态环境建设等提供科学依据[10]。

如何选择城市综合环境质量评价指标？如何进行评价？

本案例的挖掘目标是：

1) 选择合适的城市综合环境质量评价指标；

2) 选择合适的模型算法对城市综合环境质量进行评价。

分析方法与过程

我国环境质量评价工作是年代后才逐步发展起来的。发展至今，在评价指标体系及评价理论探索等方面均有较大进展。但目前我国环境评价实际工作中，所采用的方法通常是一些比较传统的评价方法，往往是从单个污染因子的角度对其进行简单评价。然而对某区域的环境质量如水质、大气质量等的综合评价一般涉及到较多的评价因素，且各因素与区域环境整体质量关系复杂，因而采用单项污染指数评价法无法客观准确地反映各污染因子之间相互作用对环境质量的影响。基于上述原因，要客观评价一个区域的环境质量状况，需要综合考虑各种因素之间以及影响因素与环境质量之间错综复杂的关系，采用传统的方法存在着一定的局限性和不合理性。因此，从学术研究的角度对环境评价的技术方法及其理论进行探讨，寻求能更全面、客观、准确反映环境质量的新的理论方法具有重要的现实意义。

根据空气中 SO_2、NO、NO_2、NO_X、PM_{10} 和 $PM_{2.5}$ 值的含量，建立 C4.5 决策树的分类预测模型，实现对空气质量进行评价。其实质是：运用 C4.5 算法进行数据挖掘，获得分类规律，即空气污染物与空气等级之间的关系；推导出分类规则，即空气质量智能评价模型。分 4 个步骤实现：数据预处理、生成决策树、分类规则生成及化简、模型准确性评估，如图 2-117。

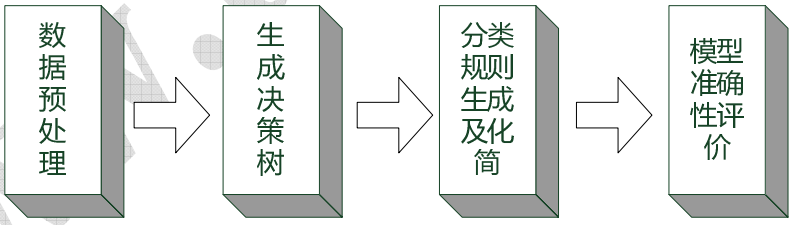


图 2-117 城市综合环境质量评价建模流程图

对于采集到的空气污染物浓度的数据（SO_2、NO、NO_2、NO_X、PM_{10} 和 $PM_{2.5}$ 值的含量），首先根据我国城市空气质量分级标准，分为优（I）、良（II）、轻微污染（III）、轻度污染（V）、中度污染（VI）、中度重污染（VII）、重污染（VIII）七个等级。这样经过预处理的数据包括 1 个类别项(空气等级)和 6 个影响分类的属性项(空气污染物)。随机抽取预处理数据（共 320 个数据）中 2/3 的数据，即 240 个数据样本，作为训练集构造决策树并生成决策规则。

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 C4.5 决策树进行模型构建。

1. 模型输入

模型的输入包括两部分，一部分是建模专家样本数据（包括训练样本和验证样本）的输入，另一部分是建模参数的输入。

部分原始样本数据经过预处理后如表 2-138 所示。

表 2-138 建模样本数据

SO2	NO	NO2	NOx	PM10	PM2-5	空气等级
0.031	0	0.046	0.047	0.085	0.058	I
0.022	0	0.053	0.053	0.07	0.048	II
0.017	0	0.029	0.029	0.057	0.04	I
0.026	0	0.026	0.026	0.049	0.034	I
0.018	0	0.027	0.027	0.051	0.035	I
0.019	0	0.052	0.053	0.06	0.04	II
0.022	0	0.059	0.06	0.064	0.042	II
0.023	0.01	0.085	0.099	0.07	0.044	II
0.022	0.012	0.066	0.084	0.073	0.042	II
0.017	0.007	0.037	0.048	0.069	0.04	I

随机抽取预处理数据（共 320 个数据）中 2/3 的数据，即 240 个数据样本，作为训练集构造决策树并生成决策规则。

C4.5 决策树和 K-最近邻分类算法的建模参数如表 2-139，本案例主要介绍 C4.5 决策树的建模过程，读者可尝试用另一种算法进行建模并比较建模效果。

表 2-139 C4.5 决策树建模参数

序号	参数名称	参数描述
1	二元分割	构建树时是否使用二元分割
2	置信因子	用于树修剪的置信因子，值越小，树修剪耗费越大
3	最少叶片数	最少决策树叶片数
4	减少错差剪枝	是否用减少误差修剪来替代 C4.5 修剪
5	提升子树	修剪时是否使用提升子树操作
6	使用拉普拉斯算子	是否使用拉普拉斯算子进行平滑

表 2-140 K-最近邻分类算法建模参数

序号	参数名称	参数描述
1	使用近邻个数	设置使用近邻个数
2	交叉验证	是否进行交叉验证
3	距离权重	设置距离权重
4	搜索最近邻方法	设置搜索最近邻方法，如 LinearNNSearch
5	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果

2. 结果评价

本案例中基于 C4.5 决策树，采用 10-折交叉验证（10-fold cross validation），对经预处理后的总体样本（240 条）进行综合评估。10 折交叉验证结果如表 2-141。

表 2-141 10 折交叉验证结果

指标	值
正确分类样本及百分比	229　95.4167%
错误分类样本及百分比	11　4.5833 %
Kappa 统计	0.9315
平均绝对误差	0.0127
均方根误差	0.1089
相对绝对误差	6.5971 %
相对平方根误差	35.2377 %
样本总数	240

交叉验证混淆矩阵结果见表 2-142。

表 2-142 混淆矩阵结果

a	b	c	d	e	f	g	<— 分类为
39	1	0	0	0	0	0	a = I
0	107	3	0	0	0	0	b = II
0	2	69	2	0	0	0	c = III
0	0	0	8	0	0	0	d = IV
0	0	0	0	6	1	0	e = V
0	0	0	0	2	0	0	f = VI
0	0	0	0	0	0	0	g = VII

表 2-142 中，以第 3 行为例，有 107 个样本正确分类（属于“II”），有 3 个样本本应该属于“II”，却错误分到了“III”。

总体来说，10 折交叉验证的正确分类百分比为 95.4167%，可见经抽选的专家样本数据能满足建模要求。

随机抽取 240 条样本进行模型训练，再对训练好的网络模型，通过模型验证（预测值与实际值比对）来评价其有效性，以确保所生成的预测模型满足实际应用要求。将预处理好的数据划分成两个独立的集合：240 条样本数据随机分配到训练集，其余 80 条数据随机分配到测试集。测试分类结果正确率如表 2-143 所示，该模型总的预测分类正确率达到 97.5%，混淆矩阵结果见表 2-144。这一评估结果说明，通过 C4.5 算法进行数据挖掘所得到的分类规则集合对未来数据样本的分类正确率能够达到较高的要求。

表 2-143 模型验证结果

指标	值
正确分类样本及百分比	78　97.5 %
错误分类样本及百分比	2　2.5 %
Kappa 统计	0.9639
平均绝对误差	0.0107
均方根误差	0.0772

www.tiodm.com

样本总数	80

表 2-144 混淆矩阵结果

a	b	c	d	e	f	g	<— 分类为
23	1	0	0	0	0	0	a = I
00	36	3	0	0	0	0	b = II
0	0	10	2	0	0	0	c = III
0	0	0	3	0	0	0	d = IV
0	0	0	0	2	1	0	e = V
0	0	0	0	4	0	0	f = VI
0	0	0	0	0	1	0	g = VII

由于城市空气质量与各种空气污染物之间的错综复杂关系，需要按照一定的标准和方法对某一区域空气质量的优劣进行定量的或定性描述。通过仿真和实验表明，在随机抽取了 320 个空气污染物浓度数据的情况下进行评估，基于 C4.5 决策树算法的空气质量智能评价模型可以达到很好的分类预测效果。

2.5.5. 案例二十 企业信用评级

温馨提示：
- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com
- 寻合作（科技申报、高校教学、企业咨询），请联系：40068-40020

挖掘目标的提出

在我国的经济社会活动中，判断一个企业的信用级别，牵涉到多个部门，包括质量技术监督、商检、工商、税务、海关、银行、消费者协会等。其中，每个职能部门都有一些关于企业信用的统计数据，如何从这些数据中判定企业的信用，进而将企业明确地标记为某个信用级别，是一个非常复杂的问题。本例不能仅根据企业财务状况指标，而应根据影响企业信用状况的多维指标来评估企业的信用状况。因此，企业信用判别方法的关键步骤和难点在于指标体系的确立和评估模型的选择，也即如何确立一种更为科学、有效的信用评级衡量标准。

国内外在对企业信用评价中，广泛采用了基于统计判别方法的预测模型。总的来说，这些模型都被表述为一类分类系统，它们接受定义在已选变量集合上的一个随机观测值样本，建立判别函数，进行分类。常用的模型有：回归分析法、多元判别分析法、Logit 法、Probit 法等，这些模型已经得到广泛的应用。这些方法的最大优点是使用简单，但也存在很大的局限性，如企业的实际信用评级指标的各个因素之间有些相互影响，有些相互独立，呈现出复杂的非线性关系，故用上述方法做出的评估比较粗糙。

作为研究复杂性的有力工具，神经网络技术近年来在模式识别与分类、识别滤波、自动控制、预测等方面已展示了其非凡的优越性，特别是能处理任意类型的数据，这是许多传统方法所无法比拟的。通过不断学习，能够从未知模式的大量的复杂数据中发现其规律。神经网络方法克服了传统分析过程的复杂性及选择适当模型函数形式的困难，它是一种自然的非线性建模过程，毋需分清存在何种非线性关系，给建模与分析带来极大的方便。它不依赖于变量之间必须线性相关或是相互独立的假设。变量之间存有微妙联系，如同数据不连续或不完全一样，均可被系统辨识并自动实现分类。

具体应用为：首先，由输入大量的企业信用资料及相应信用等级数据并对其进行分析，此过程称为“训练”。在“训练”期间，神经网络系统研究它收到的原始的随机性的数据，重建它们的数学关系，将之转换为连续性的有数学规律性的形式，结果是得到一个适合这些数据的模型（称为建模）。即神经网络系统依据输入输出的数据对，通过学习得到一个描述系统输入输出关系的非线性映射。然后神经网络系统还会自动对它本身的模型进行调整，所以神经网络系统还有“推断”功能。

建模的关键，就在于每个处理单元之间建立适当的连结，这就像脑神经细胞间彼此用突触做沟通的渠道一样。本例可用一般的电脑来模拟人工的神经元，每个神经元都可同时接受许多输入讯号，根据内部简单的运算，送出一个单一的讯号，作为另一神经元的输入讯号。这些神经元的连结并不象想像的那样杂乱无章，而是被组织成层次分明的神经元，每一层内的神经元并不相互连结，它们只接受下层神经元的讯号，并向上送出结果。其中，输入层负责接收资料，而输出层则送出最终的结果，并且它们之间还存在着隐含层，夹杂在两者之间。

一旦模型生成，用户就可在输入仿真数据的基础上，调入对应的网络结构与建模时生成模型参数，得到实际的评级结果。

本例在分析企业信用评价指标体系的基础上，总结出企业信用评价指标与信用等级的对应关系。然后，将模糊综合评判和径向基函数网络有机结合在一起，建立了基于模糊径向基函数的神经网络模型。分析表明，该模型除可进行信息的快速并行处理以实现映射变换之外，还具有较强的联想记忆和联想映射能力。若输入一个实际样本，网络可通过非线性映射实现快速准确地判断。因此，模糊径向基函数诊断网络在企业信用评级中是一种有效的智能分类器[8]。

分析过程与方法

1．企业信用评价指标体系

要进行企业信用评级，首先要考虑的问题是：企业的哪些信息可以作为企业信用评级的指标？如何将这些指标进行分类及量化处理？

某市正式投入使用的企业信用信息系统中，由市工商局、国税局、地税局、质量技术监督局（打假办）、外经贸局、银行和海关等 7 个职能部门组成的企业征信机构披露的企业信用信息包括：企业基本情况：企业名称、法定代表人姓名、企业住所、注册资本、经营范围，税务登记证号、核算方式、行业、税务登记验证和换证情况、纳税人性质和税务管理状态，企业联合年检情况，进出口经营资格和企业类型，企业的全国组织机构代码等；企业经营和财务状况：主要产品（业务）、年销售（营业）收入、年纳税总额、年纳税入库总额；企业荣誉纪录：重大奖励，驰名、著名和重点保护商标资料，法定代表人荣誉记录以及企业其他荣誉资料；企业不良纪录：企业走私、逃骗套汇、偷逃骗抗税、制假贩假、恶意逃废债务、利用合同诈骗等违法情况，以及违法受吊销企业营业执照、较大数额罚款等行政处罚情况等。经反复分析和筛选，最后本例得到如图 2-118 所示的企业信用评价指标体系：

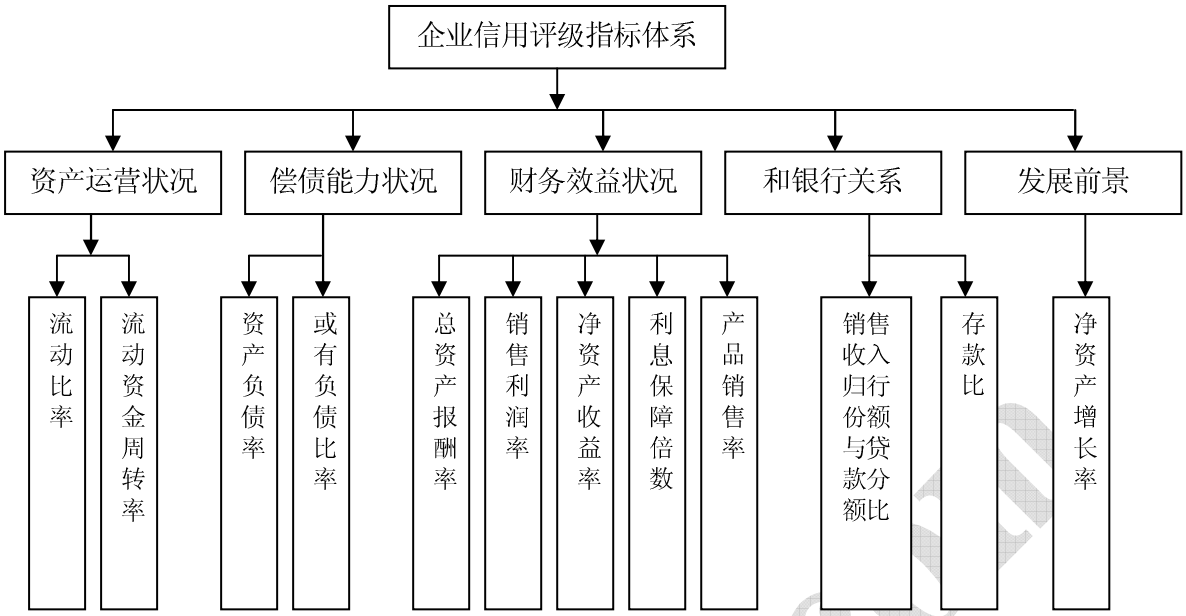


图 2-118 企业信用评价指标体系

如图 2-118 所示，上面的 12 个指标构成了企业信用评级的指标体系，另外，在此基础上，可把领导者素质、管理素质、技术素质和市场定位等作为定性指标，予以考虑。

作为企业信用等级的划分采用目前国际上通用的四等十级制，即将企业信用等级分为 A 级、B 级、C 级、D 级、E 级等共 5 级，也有评估机构将信用等级分为 AAA 级、AA 级、A 级、BBB 级、BB 级、B 级、CCC 级、CC 级、C 级、D 级等共 10 级，本案例采用 5 级（分别对应企业信用状况从信用极好到没有信用）。

2．有效模式和样本集的确定

模式是具有较强可比性和相似性的样本的集合，模式样本数量满足建立一个评估模型要求的模式称为有效模式，否则称为无效模式。

样本集是按照一定的规则划分的具有某种相同属性的样本的集合，样本集在样本的某一属性上具有同一性。在企业信用评级中，如按所有制的不同，可以划分出不同的样本集，按行业的不同也可以划分出不同的样本集。所以有实际操作中，有必要对企业数据进行分类，筛选出不同的样本集。

因为样本跨越不同的行业、所有制和规模，每个财务指标在数值上的差异非常大，如果不剔除掉那些偏离中心太远的样本点，整个样本集的数据势必会受到这些奇异样本的影响，使整个样本集的信息受到弱化。所以在选取样本时，必须考虑样本中奇异点的剔除。

在确定了企业信用评级指标体系后，本例从某市的企业信用信息系统中筛选出所有制形式为私营，行业为工业的企业数据，并将样本中的奇异点剔除，组成一个符合有效模式条件的样本集，以此作为后续训练神经网络模型的专家样本数据，图 2-119 为构建专家样本库的具体操作步骤。

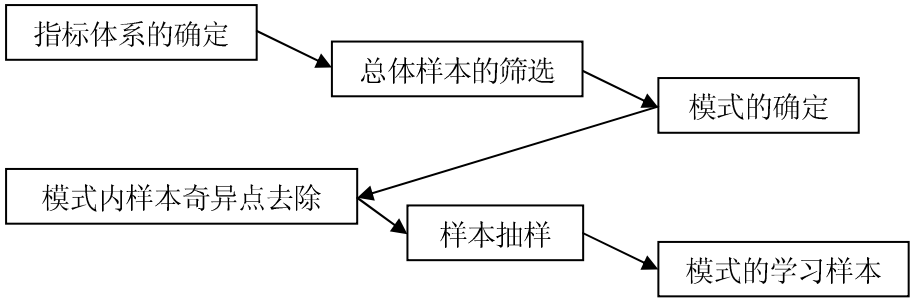


图 2-119 构建专家样本库流程

建模仿真

本案例采用 TipDM 数据挖掘在线建模平台（www.tipdm.cn；www.tipdm.com.cn）中的 ANFIS 自适应模糊神经网络推理系统、KStar 懒惰算法、AdaBoost 分类算法等进行模型构建。

1. 模型输入

模型的输入包括两部分，一部分是建模专家样本数据（包括训练样本和验证样本）的输入，另一部分是建模参数的输入。

表 2-145 预测模型输入变量

序号	变量名称	变量描述	取值范围
1	X1	流动比率	0~1
2	X2	流动资金周转率	0~1
3	X3	资产负债率	0~1
4	X4	或有负债比率	0~1
5	X5	总资产报酬率	0~1
6	X6	销售利润率	
7	X7	净资产收益率	0~1
8	X8	利息保障倍数	0~1
9	X9	产品销售率	0~1
10	X10	销售收入归行份额与贷款分额比	0~1
11	X11	存款比	0~1
12	X12	净资产增长率	0~1
13	Y	信用级别	A 级、B 级、C 级、D 级、E 级

说明：表中 X 为输入变量，Y 为输出变量。

表 2-146 ANFIS 神经网络建模参数

序号	参数名称	参数描述
1	训练次数	最大训练（迭代）次数
2	期望误差	期望目标均方根误差（RMSE）
3	初始步长	初始步长，默认 0.01
4	步长递减率	步长递减率，默认 0.09
5	步长递增率	步长递增率，默认 1.1
6	隶属度函数类型	三角形、梯形、钟型、高斯型、高斯 2 型、S 型、S 和型、S 积型、π型
7	隶属度函数个数	隶属度函数个数，一般不超过5个

表 2-147 KStar 懒惰算法建模参数

序号	参数名称	参数描述
1	混合信息熵	是否使用混合信息熵
2	全局混合值	全局混合的参数，参数值在[0，100]选取
3	缺失处理模式	缺失值的处理方式
4	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果

www.tipdm.com

表 2-148 AdaBoost 分类算法建模参数

序号	参数名称	参数描述
1	搜索策略	搜索网络结构的方法
2	迭代次数	循环迭代的次数
3	种子数	初始随机选取的样本数
4	重采样	是否需要对样本重采样
5	权阈值	权值改变的阈值
6	预测输出	选择“True”，当交叉验证和模型训练时，将会输出样本的预测结果

2. 仿真过程

下面详细介绍 ANFIS 算法的建模过程，另外两种算法的模型构建请读者自行尝试，并比较建模效果。

ANFIS 建模构建流程如图 2-120 所示。

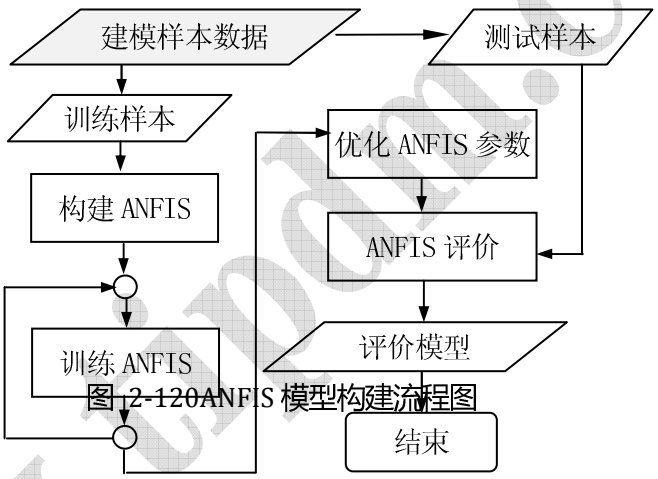


图 2-120ANFIS 模型构建流程图

建模过程说明如下：

1）选取训练样本和测试样本。

2）初始化模糊推理系统（FIS）。确定初始的参数化 FIS 模型结构，旨在为输入与隶属度函数、隶属度函数与规则、规则与输出、输出与隶属度函数间建立联系。ANFIS 提供了人工方法、网格分割、减法聚类三种方法。人工方法依赖领域专家经验确定隶属度函数，而网格分割、减法聚类采用特定的方法由系统自动实现，其中减法聚类能确保模糊规则数最少，适用于样本数较多的情形。

3）调用 ANFIS 函数对模糊推理系统（FIS）结构进行训练和参数优化；

4）使用测试样本，调用 evalfis、plot 函数对训练好的 ANFIS 进行验证。

3. 结果评价

当设置 ANFIS 自适应模糊神经推系统的训练次数为 100，期望误差为 0.001，初始步长为 0.01，步长递减率为 0.09，步长递增率为 1.1，隶属度函数为钟型，隶属度函数个数为 4 时，系统输出训练前后的隶属度函数如图 2-121 图 2-122，建模过程训练误差如图 2-123。经对建模参数进行反复优化后，建模过程训练误差如图 2-124，训练样本和测试样本预测结果分别见图 2-125 和图 2-126。

www.ilpdm.com

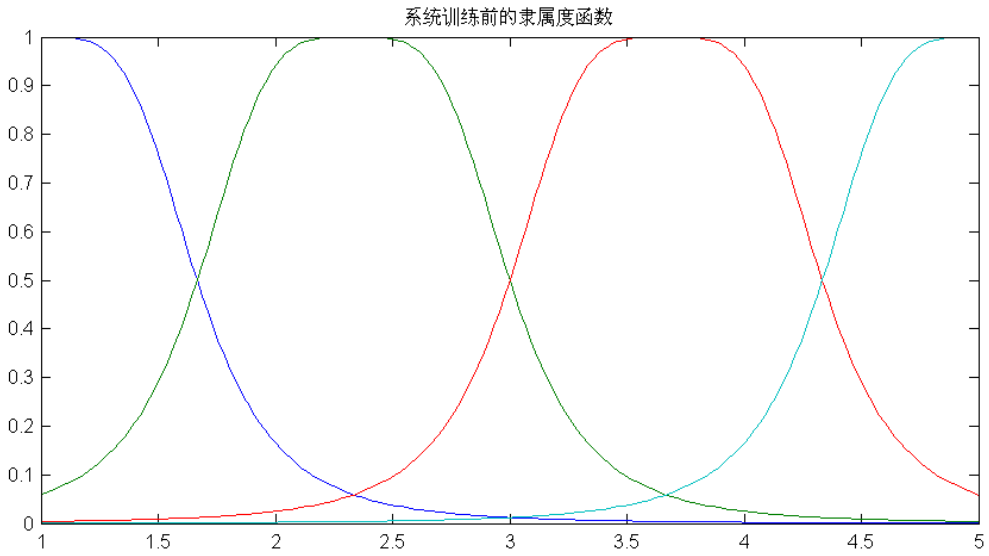


图 2-121 系统训练前的隶属度函数

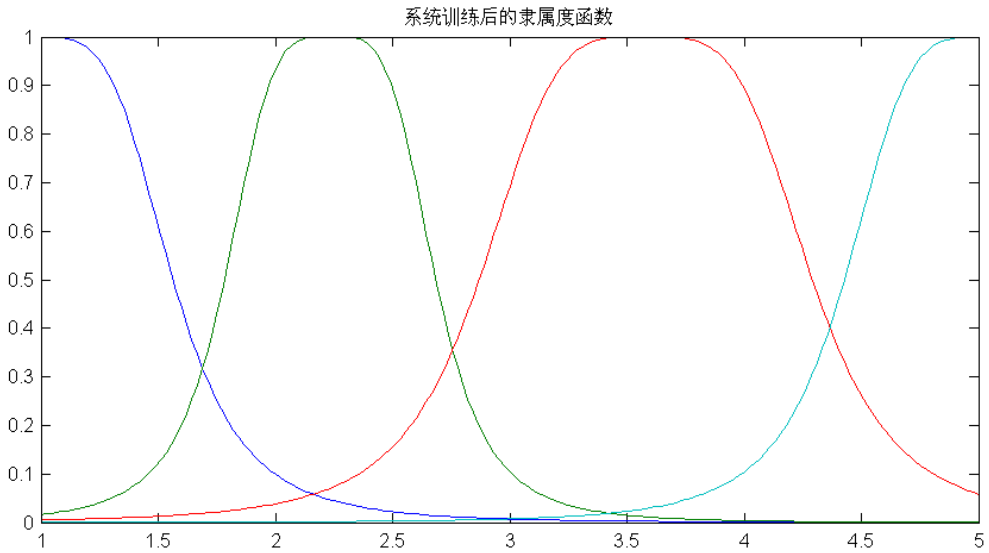


图 2-122 系统训练后的隶属度函数

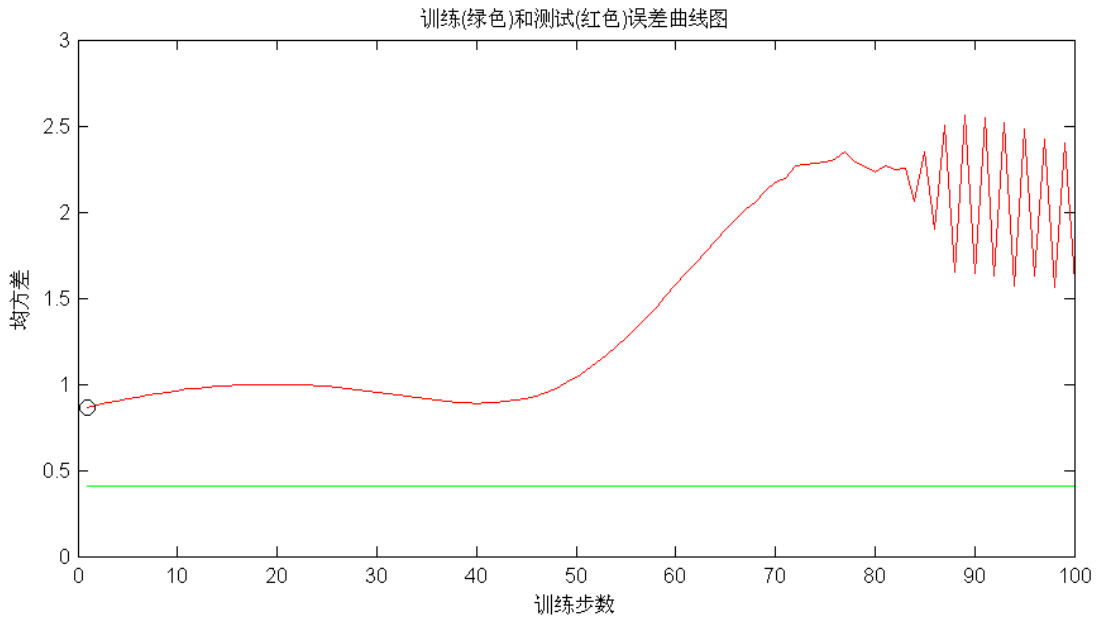


图 2-123 模型训练误差图

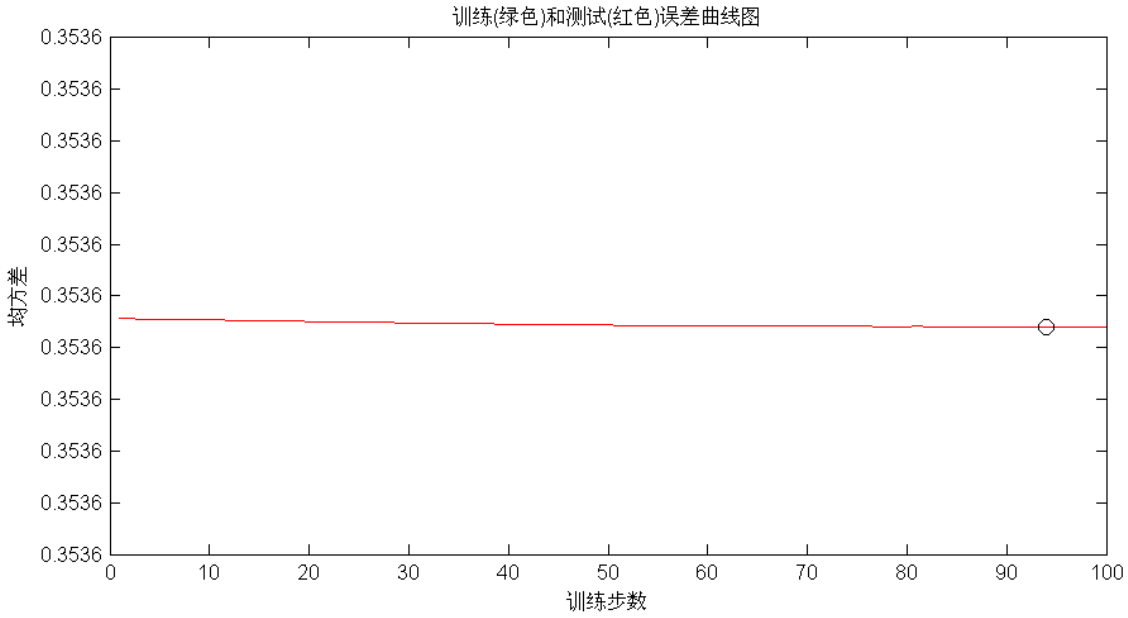


图 2-124 优化建模参数后的训练误差图

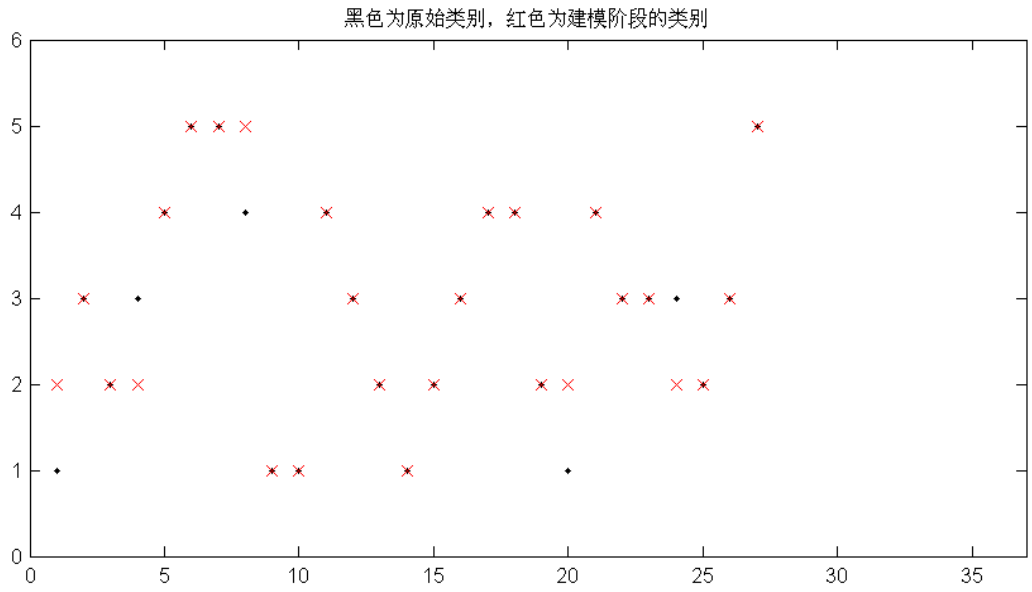


图 2-125 训练样本预测结果图

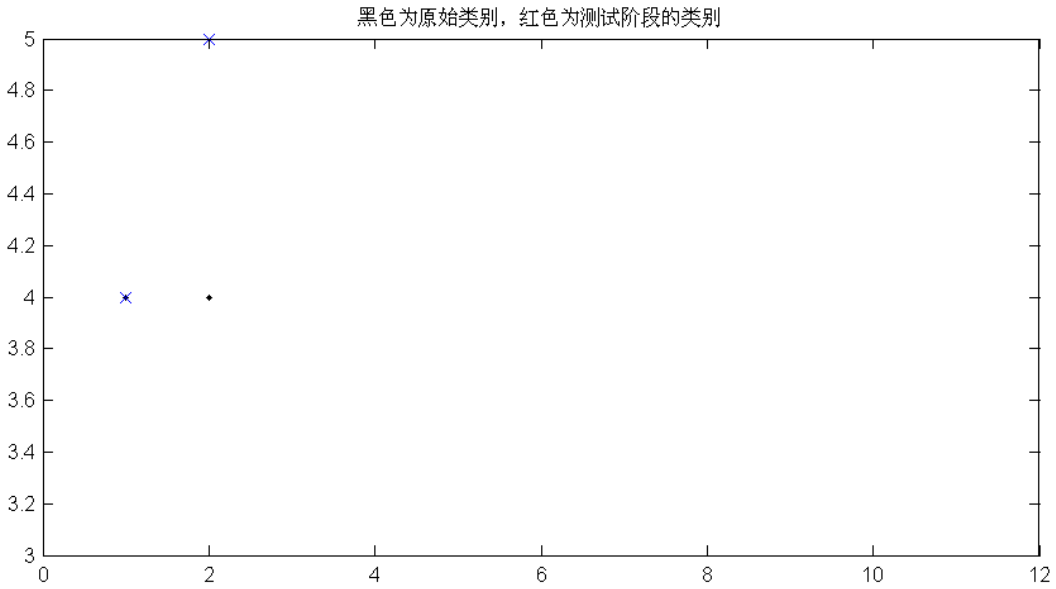


图 2-126 测试样本预测结果图

第3章 建模工具

3.1. Enterprise Miner

Enterprise Miner(EM) 是SAS推出的一个集成的数据挖掘系统，允许使用和比较不同的技术，同时还集成了复杂的数据库管理软件。它的运行方式是通过在一个工作空间（workspace）中按照一定的顺序添加各种可以实现不同功能的节点，然后对不同节点进行相应的设置，最后运行整个工作流程(workflow)，便可以得到相应的结果。

3.2. Clementine

SPSS 的核心挖掘产品 Clementine 提供了一个可视化的快速建立模型的环境。作为业内领先的数据挖掘平台，允许企业利用业务专业技能快速开发预测模型，并迅速部署提升决策水平。Clementine 能够帮助企业达成目标：①提升客户获取能力和维持能力；②提升单位客户的收益和利润；③最小化成本。Clementine 虽然在数据挖掘算法上做得很好，但是没有充分考虑综合各种预测方法来优化预测结果，另外就是其可测量性较差。

3.3. Intelligent Miner

IBM 的 Intelligent Miner 正在竞争数据挖掘工具市场的领导地位，竞争力主要集中在资料探勘上。

3.4. SQL SERVER

Microsoft 的 SQL Server 中集成了数据挖掘工具，借助 SQL Server 的数据库管理功能，用户可以实现数据挖掘建模。在 SQL Server2008 中提供了决策树算法、聚类分析算法、Naive Bayes 算法、关联规则算法、时序算法、神经网络算法、线性回归算法等 9 种常用的数据挖掘算法。但是其预测建模的实现是基于 SQL Server 平台的，平台移植性相对较差，也没有考虑综合各种预测方法来优化预测结果。

3.5. Weka

Weka (Waikato Environment for Knowledge Analysis)，这是名气最大的开源机器学习和数据挖掘软件。高级用户可以通过 Java 编程和命令行来调用其分析组件。同时，Weka 也为普通用户提供了图形化界面，称为 Weka Knowledge Flow Environment 和 Weka Explorer，可以实现预处理、分类、聚类、关联规则、文本挖掘、可视化等等。

3.6. TipDM

顶尖数据挖掘平台（TipDM）是广州太普软件科技有限公司花费数年时间自主研发的一个数据挖掘平台，基于云计算和 SOA 架构，使用 JAVA 语言开发，能从各种数据源获取数据，建立各种不同的数据挖掘模型。TipDM 支持数据挖掘流程所需的主要过程，并提供开放的应用接口和常用算法，能够满足各种复杂的应用需求。支持 CRISP-DM 流程所需的主要过程，完成包括对数据进行预处理，包括空值处理、降维处理、离散处理，因子分析、主成分分析、抽样、过滤等，创建、训练、评估

www.tipdm.com

模型，预测，修改模型参数，误差分析等一系列功能。

TipDM 典型应用场景见图 3-1

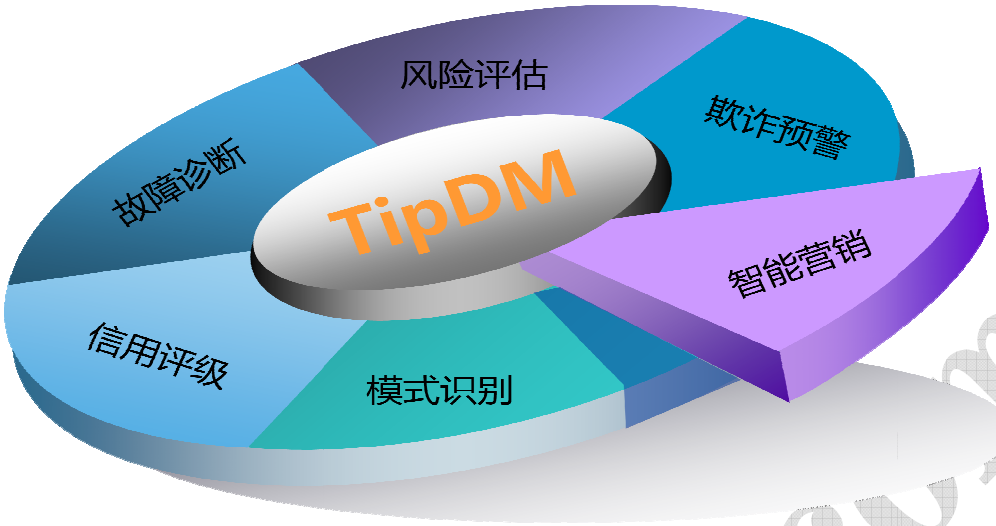


图 3-1 TipDM 典型应用场景

www.tipdm.com

第4章 顶尖数据挖掘平台（TipDM）

温馨提示：

- TipDM 数据挖掘平台（www.tipdm.cn,www.tipdm.com.cn）可免费注册
- TipDM 数据挖掘平台操作讲解视频下载网址：http://www.tipdm.com/teach.php?act=list
- 购买了 TipDM 产品的用户可以使用平台的所有算法功能

4.1. 产品简介

顶尖数据挖掘平台（TipDM）是广州太普软件科技有限公司花费数年时间自主研发的一个数据挖掘平台，基于云计算和 SOA 架构，使用 JAVA 语言开发，能从各种数据源获取数据，建立各种不同的数据挖掘模型。系统支持数据挖掘流程所需的主要过程，并提供开放的应用接口和常用算法，能够满足各种复杂的应用需求。2010 年初“顶尖数据挖掘平台（TipDM）”通过了由广州赛宝软件评测中心的功能和性能测试。现产品在广东省电信规划设计院、珠江水产研究所、华南师范大学、广东中医药大学、武汉理工大学、交通运输部公路科学研究院等单位成功试用，受到用户的赞许与肯定。

TipDM 以智能预测技术为核心，并提供开放的应用接口，用户也可以嵌入其它自己开发的任何算法，平台支持 CRISP-DM 流程所需的主要过程，完成包括对数据进行预处理，包括空值处理、降维处理、离散处理，因子分析、主成分分析、抽样、过滤等，创建、训练、评估模型，预测，修改模型参数，误差分析等一系列功能。随着企业信息化建设的深入，TipDM 必将成为企业数据分析的得力助手，为企业带来很好的经济效益和社会效益。

TipDM 的底层算法，主要基于 Matlab、R 和 Weka 等通过封装形成，所以建模输出也与这三个工具的输出结果类同。

4.2. 产品功能

4.2.1. 数据探索及预处理

数据探索是对导入系统中的数据进行初步研究，以便更好地理解它的特殊性质，有助于选择合适的数据预处理和数据分析技术。

模型预测的质量不会超过抽取样本的质量。数据探索和预处理的目的是为了保证样本数据的质量，从而为保证预测质量打下基础。

算法名称	算法描述
相关性分析	相关性分析是考察两个变量之间相互关系的方向和程度的一种统计分析方法。更精确地说，当一个变量发生变化时，另一个变量如何变化，此时就需要通过计算相关系数来做深入的定量考察。当然，还有其他类型的相关系数用于测量两个以上变量之间的关系，如多元回归等。
主成分分析	主成分分析分析（PCA）是指用几个较少的综合指标来代替原来较多的指标，而这些较少的综合指标既能尽可能多地反映原来较多指标的有用信息，且相互之间又是无关的。

www.tipdm.com

周期性分析	分析数据变化过程中，某些特征重复出现，连续的两次出现的客观规律。在时序预测时，经常要对样本数据进行周期性分析，以更好地理解样本数据变化的特点，为时序预测分析提供指导依据。
缺失值处理	样本数据中经常存在一些缺失值，在进行预测建模前，需要对缺失值按照某些规则进行处理。处理方法主要有：1）删除；2）数据补齐（人工填写、平均值填充、特殊值填充、热卡填充、K 最近距离邻法、回归、期望值最大化方法等）。
坏数据处理	如果抽取数据中存在坏数据（脏数据），则需要对坏数据进行预处理。通常的做法是采用绝对均值法或莱因达法等对样本中的坏点数据进行剔除处理。
属性选择	由给定的属性构造或添加新的属性，以帮助提高精度和对高维数据结构的理解，是具有最佳预测准确率的子集。
数据规约	将属性数据按比例缩放，使之落入到一个小的特定区间。找出最小属性集合，使得数据类的概率分布尽可能地接近使用所有属性的原分布。
离散处理	离散化技术用来减少给定连续属性的个数，这个过程通常是递归的，而且大量时间花在排序上。对于给定的数值属性，概念分层定义了该属性的一个离散化的值。
特征提取	主要对图像、声音、信号等数据源，通过图像处理、小波变换等建立一组新的、更紧凑的属性来表示数据的过程。如图像特征提取是指根据图像特征，提取反映图像本质的一些关键指标，以达到自动进行图像识别或分类的目的。

说明：版本持续更新，具体以网站上发布的为准。

4.2.2. 分类与回归

分类是数据挖掘中应用得最多的方法。分类就是找出一个类别的概念描述，它代表了这类数据的整体信息，即该类的内涵描述，并用这种描述来构造模型，一般用规则或决策树模式表示。分类是利用训练数据集通过一定的算法而求得分类规则。分类可被用于规则描述和预测。

回归是利用历史数据找出变化规律，建立模型，并由此模型对未来数据的种类及特征进行预测。预测关心的是精度和不确定性，通常用预测方差来度量。

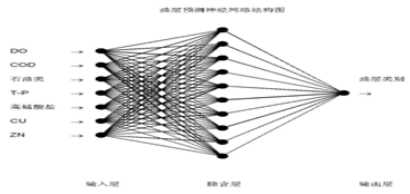
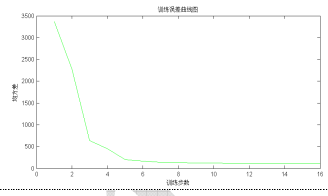
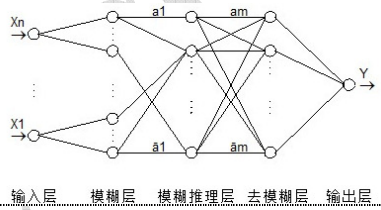
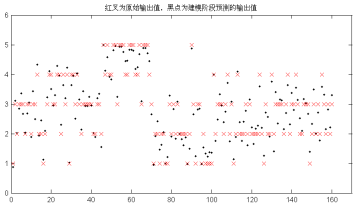
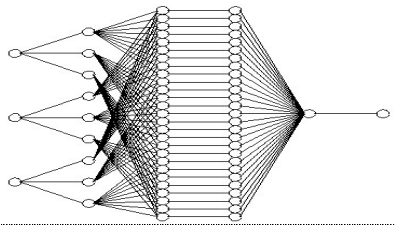
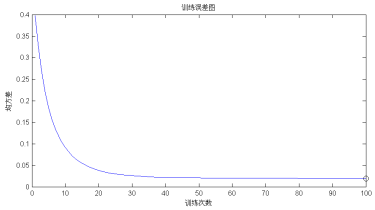
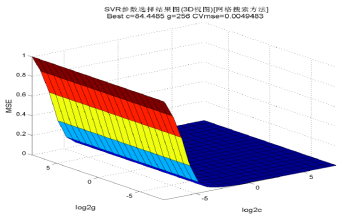
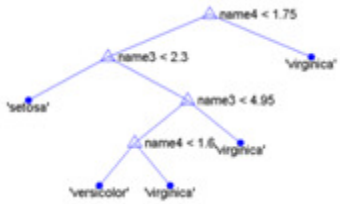
分类与回归算法主要有：

- 贝叶斯：朴素贝叶斯网络、贝叶斯信念网络
- 决策树及表：决策表、CART决策树、ID3决策树、C4.5决策树
- 神经网络：BP神经网络、LM神经网络、RBF神经网络、FNN神经网络、ANFIS神经网络、WNN神经网络
- 回归分析：线性回归、逐步回归、逻辑回归、保序回归
- 其它算法：AdaBoostM1算法、KStar算法、SVM支持向量机、K-最近邻分类

主要算法介绍：

算法名称	算法描述	输出示例

www.tipdm.com

BP 神经网络	BP（Back Propagation）神经网络是是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。	
LM 神经网络	Levenberg-Marquardt 是基于梯度下降法和牛顿法结合的优化算法，特点：迭代次数少，收敛速度快，精确度高。	
FNN 模糊神经网络	模糊神经网络（Fuzzy Neural Network,FNN）是具有模糊权系数或者输入信号是模糊量的神经网络，是模糊系统与神经网络相结合的产物，它汇聚了神经网络与模糊系统的优点	
RBF 径向基神经网络	径向基函数(Radial Basis Function, RBF)神经网络是具有单隐层的三层前馈网络。它是一种局部逼近网络，能够以任意精度逼近任意连续函数，特别适合于解决分类问题。	
ANFIS 自适应神经模糊推理系统	自适应神经模糊推理系统（Adaptive Neural Fuzzy Inference System,ANFIS）是功能上与一阶 T-S 模糊推理系统等价的自适应网络，它是将神经网络的学习机制引入模糊系统，构成一个带有人类感觉和认知成分的自适应系统。	
WNN 小波神经网络	小波神经网络（Wavelet Neural Network,WNN）是基于小波变换而构成的神经网络模型,即用非线性小波基取代通常的神经元非线性激励函数(如 Sigmoid 函数),把小波变换与神经网络有机地结合起来,充分继承了两者的优点。	
SVM 支持向量机	支持向量机是 V.Vapnik 等人在研究统计学习理论的基础上发展起来的一种新的机器学习算法，它在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势。支持向量机根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷，以获得最好的推广能力。	
CART 决策树	分类与回归树（classification and regression tree,CART），是一种十分有效的非参数分类和回归方法，通过构建树、修剪树、评估树来构建一个二叉树。	

说明：版本持续更新，具体以网站上发布的为准。

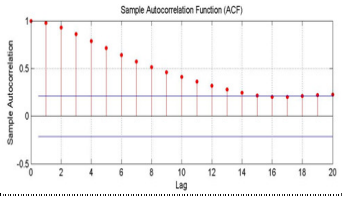
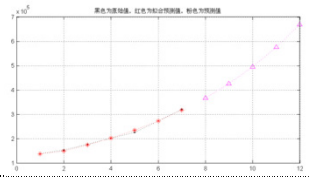
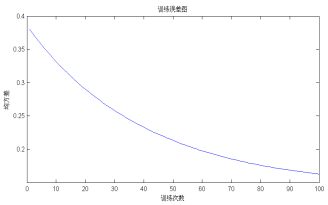
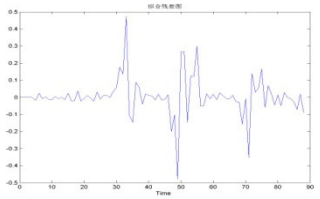
4.2.3. **时序模式**

时序模式是指通过时间序列搜索出的重复发生概率较高的模式。与回归一样，它也是用已知的数据预测未来的值，但这些数据的区别是变量所处时间的不同。

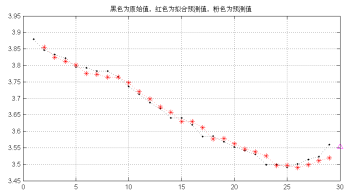
时序模式类算法主要有：

- 指数平滑
- 多元回归
- GM灰色理论
- RBF神经网络
- ANFIS神经网络
- SVM支持向量机

主要算法介绍：

算法名称	算法描述	输出示例
指数平滑	使用指数平滑（Exponential Smoothing）方法进行时间序列预测	
GM 灰色理论	使用灰色 GM(1,1)模型进行时间序列预测	
RBF 神经网络	使用 RBF 神经网络方法进行时间序列预测	
多元回归	使用回归方法进行时间序列预测	
SVM 支持向量机	使用 SVM 支持向量机方法进行时间序列预测	

www.ipom.com

ANFIS 自适应神经模糊推理系统	使用 ANFIS 自适应神经模糊推理系统方法进行时间序列预测	

说明：版本持续更新，具体以网站上发布的为准。

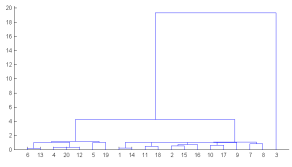
4.2.4. 聚类分析

聚类是把数据按照相似性归纳成若干类别，同一类中的数据彼此相似，不同类中的数据相异。聚类分析可以建立宏观的概念，发现数据的分布模式，以及可能的数据属性之间的相互关系。

聚类分析类算法主要有：

- K-均值算法
- EM最大期望算法
- DBScan密度算法
- 改进K-均值算法
- 多层次聚类

主要算法介绍：

算法名称	算法描述	输出示例
K-均值算法	K-均值聚类(K-means clustering)是Mac Queen提出的一种非监督实时聚类算法，在最小化误差函数的基础上将数据划分为预定的类数K。	
EM最大期望算法	最大期望（Expectation－Maximization，EM）算法是Dempster，Laind，Rubin提出的求参数极大似然估计的一种方法，它可以从非完整数据集中对参数进行MLE 估计。	
DBScan密度算法	DBScan（Density Based Spatial Clustering of Applications with Noise），是基于密度的聚类方法，它根据对象周围的密度不断增长聚类。它能从含有噪声的空间数据库中发现任意形状的聚类。	
多层次聚类	这种方法对给定的数据集合进行层次的分解，根据层次的分解如何形成，它又可分为凝聚法(也称自底向上方法)和分裂法(也称为从上向下方法)	

说明：版本持续更新，具体以网站上发布的为准。

4.2.5. 关联规则

关联规则挖掘是由Rakesh Apwal等人首先提出的。两个或两个以上变量的取值之间存在某种规律性，就称为关联。数据关联是数据库中存在的一类重要的、可被发现的知识。关联分为简单关联、时序关联和因果关联。关联分析的目的是找出数据库中隐藏的关联网。一般用支持度和可信度两个阀值来度量关联规则的相关性，还不断引入兴趣度、相关性等参数，使得所挖掘的规则更符合需求。

关联分析类算法主要有：

- HotSpot关联规则
- Apriori关联规则
- FP-Tree关联规则
- HOTSPOT关联规则
- 灰色关联法

4.3. 使用说明

TipDM 平台总体操作流程、数据探索操作流程、数据预处理操作流程、分类与回归操作、聚类分析操作流程、时序预测操作流程及关联规则挖掘操作流程分别见图 4-1 至图 4-7。

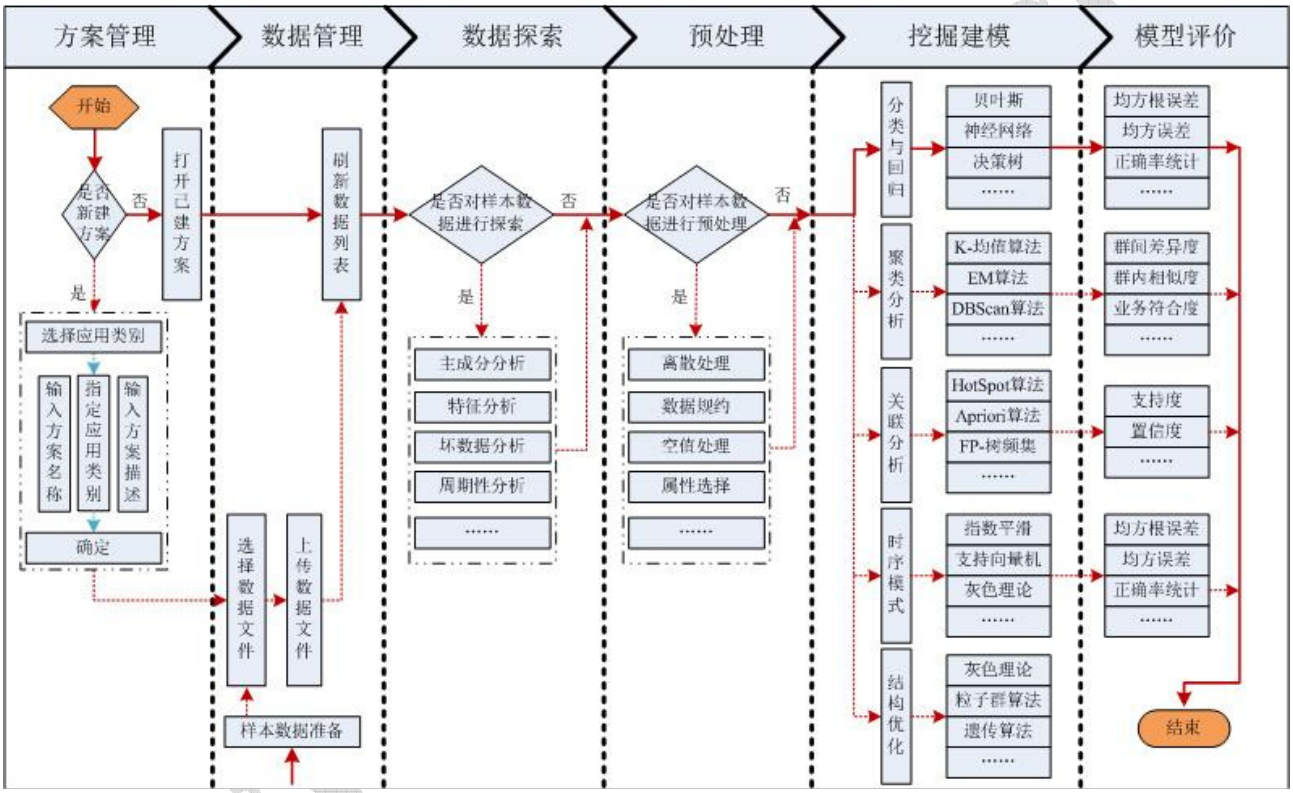


图 4-1TipDM 总体操作流程

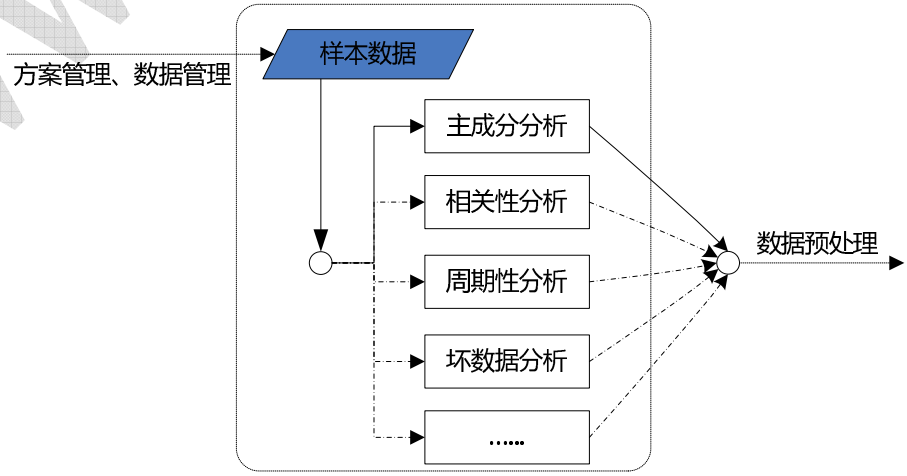


图 4-2 数据探索操作流程

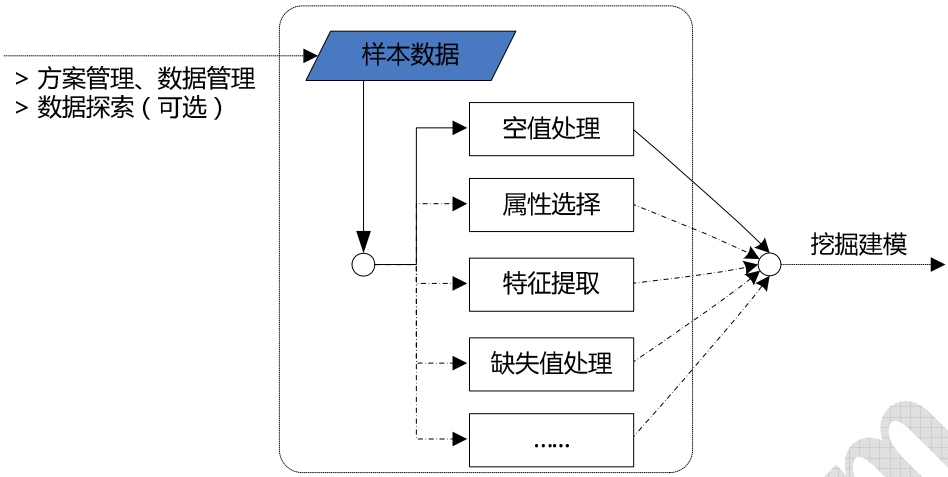


图 4-3 数据预处理操作流程

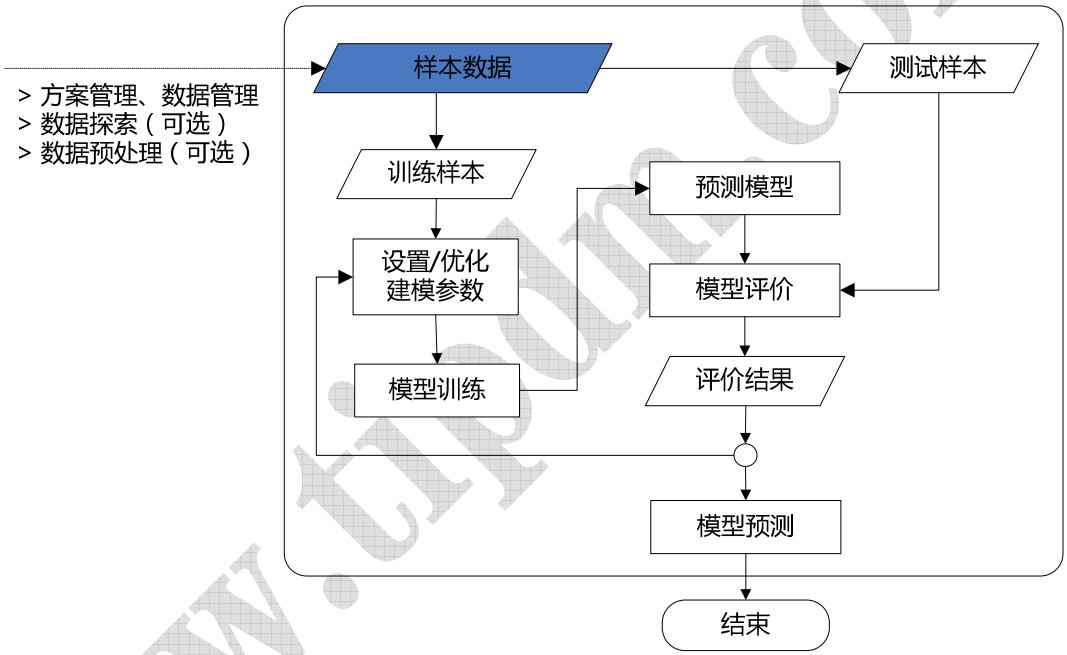
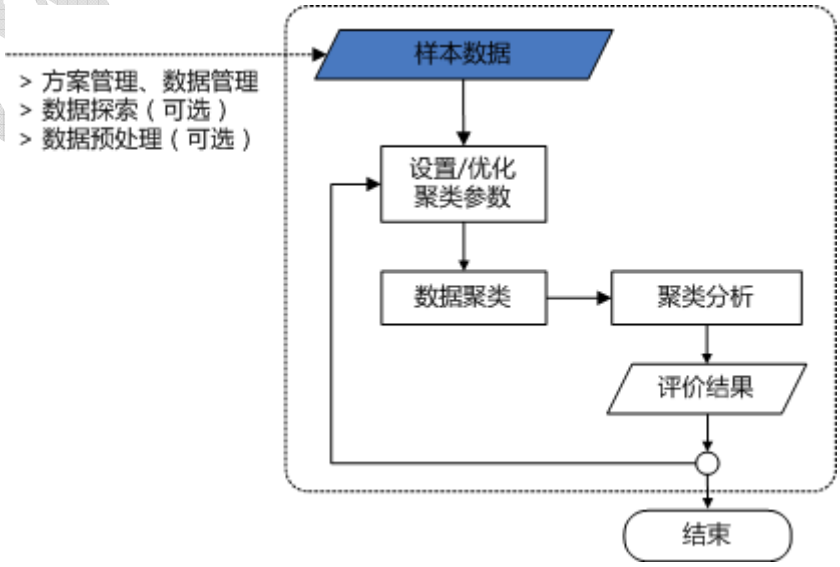


图 4-4 分类与回归操作流程



www.dpm.com

图 4-5 聚类分析操作流程

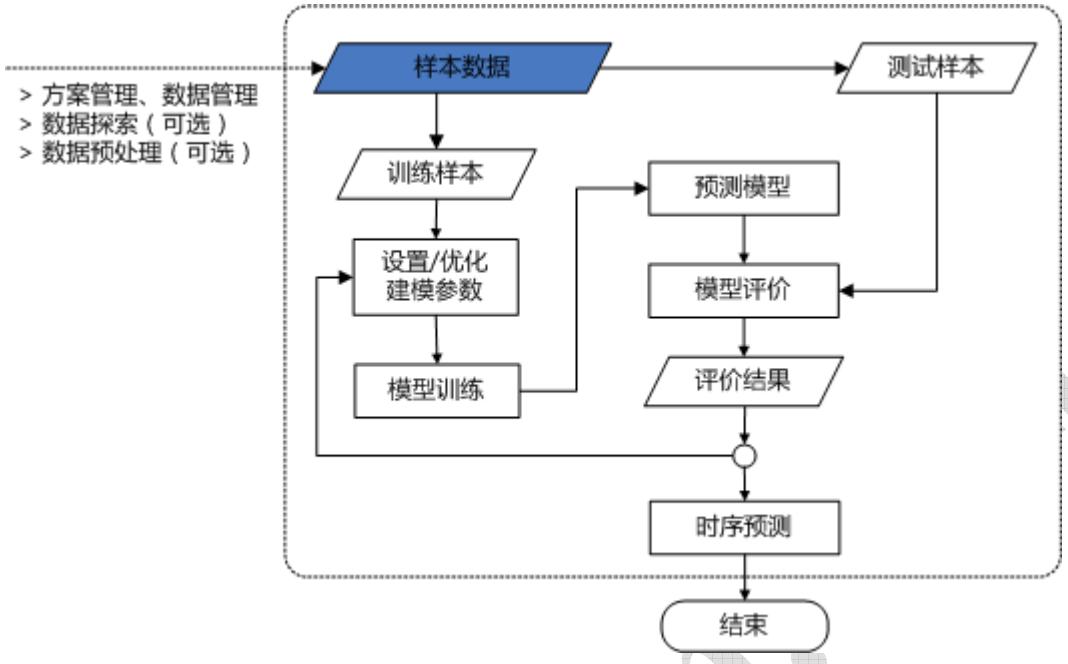


图 4-6 时序预测操作流程

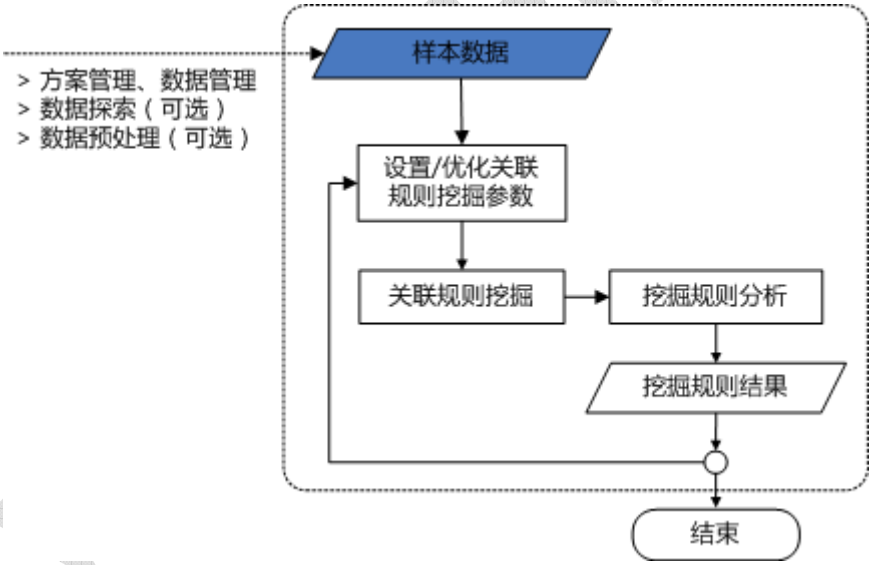


图 4-7 关联规则挖掘操作流程

1）创建一个简单的数据挖掘模型

这部分一起来创建一个简单的 BP 神经网络模型，网络结构为：具有两个输入节点包含四个样本的输入文件，样本数据如下：

	A	B	C
1	输入1	输入2	输出
2	0	0	0
3	0	1	1
4	1	0	1
5	1	1	0

创建模型，主要基于以下几个步骤：

第一步：创建方案



第二步：加载数据

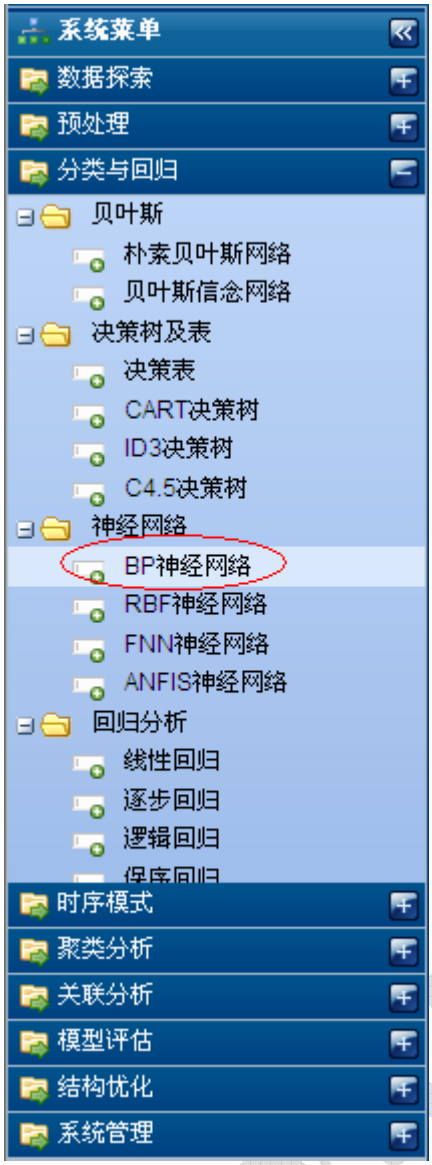
www.tipdm



第三步：预测建模

从菜单中选择 BP 神经网络：

www.tipc



① 导入数据

www.clpdm.com


方案管理
数据管理
BP神经网络
分类与回归(BP神经网络)
训练数据
index
输入1
输入2
输出
1 0 0 0
2 0 1 1
3 1 0 1
4 1 1 0
未选择列
已选择列(最后一列为输出)
>
<
>>
<<
输入1
输入2
输出
开始行号
1
结束行号
4
导入数据
导入数据
参数设置
模型训练
模型验证
模型预测

② 参数设置


方案管理
数据管理
LM神经网络
分类与回归(BP神经网络)
训练数据
index
输入1
输入2
输出
显示间隔次数
25
最大循环次数
100
目标误差
0.02
动量常数
0.001
学习速率
0.001
学习速率增加比率
10
学习速率减少比率
0.1
最大误差比率
100000000
隐层节点数
6
输入函数
线性型函数
输出函数
双曲正切S型函数
导入数据
参数设置
模型训练
模型验证
模型预测
当前位置 | 分类与回归--LM神经网络
操作流程

③ 模型训练


方案管理
数据管理
LM神经网络
分类与回归(BP神经网络)
训练数据
index
输入1
输入2
输出
=== 模型训练信息 ===
样本总数：4，训练耗时：16.1579秒
1、BP神经网络结构见图1，具体信息如下：
2个输入层节点，6.0个隐含层节点，1个输出层节点
输入属性：
输入1
输入2
输出属性：
输出
2、BP神经网络建模参数：
显示间隔次数：25.0
最大循环次数：100.0
目标均方误差：0.02
学习速度：0.0010
学习速率增加比率：10.0
学习速率减少比率：0.1
最大误差比率：1.0E8
导入数据
参数设置
模型训练
模型验证
模型预测
当前位置 | 分类与回归--LM神经网络
操作流程

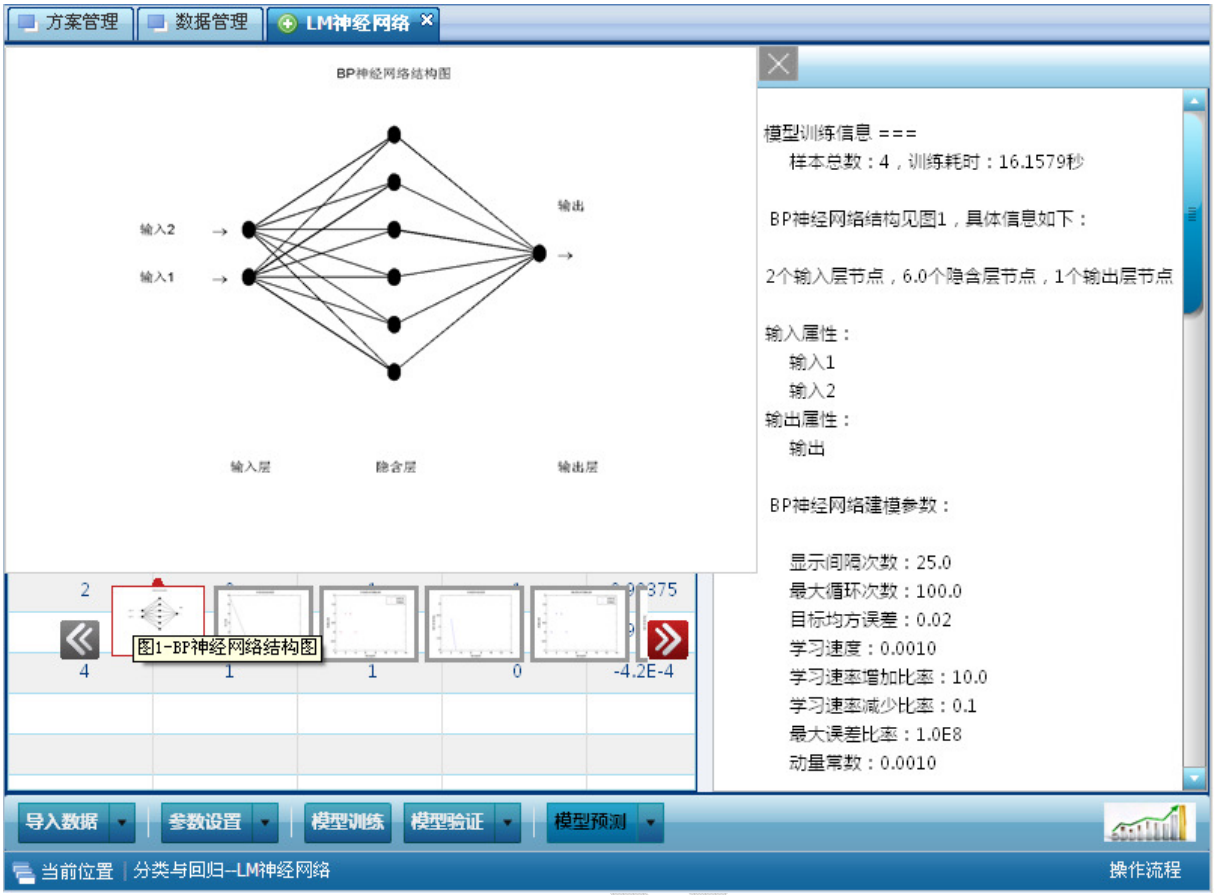
④ 模型验证



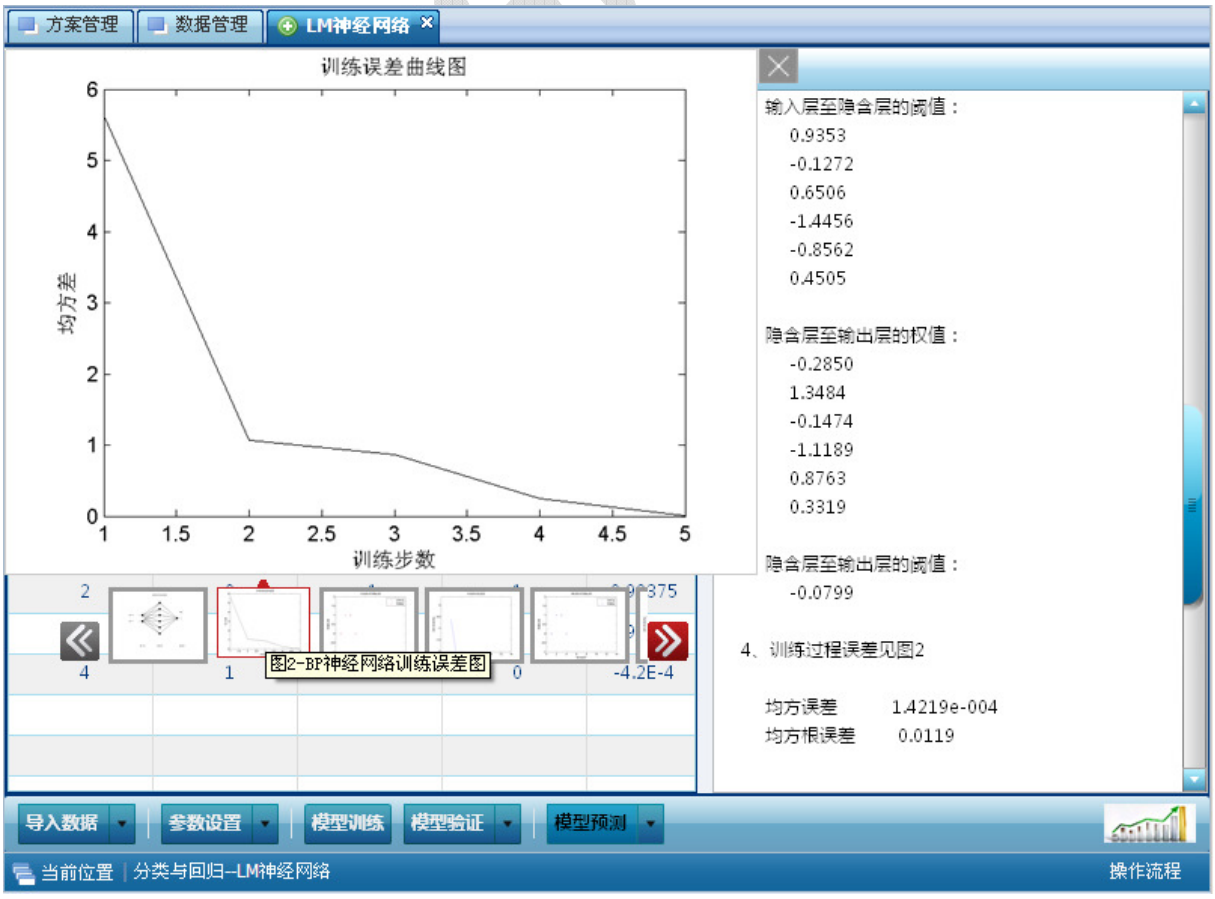
⑤ 建模输出

✧ 神经网络结构图

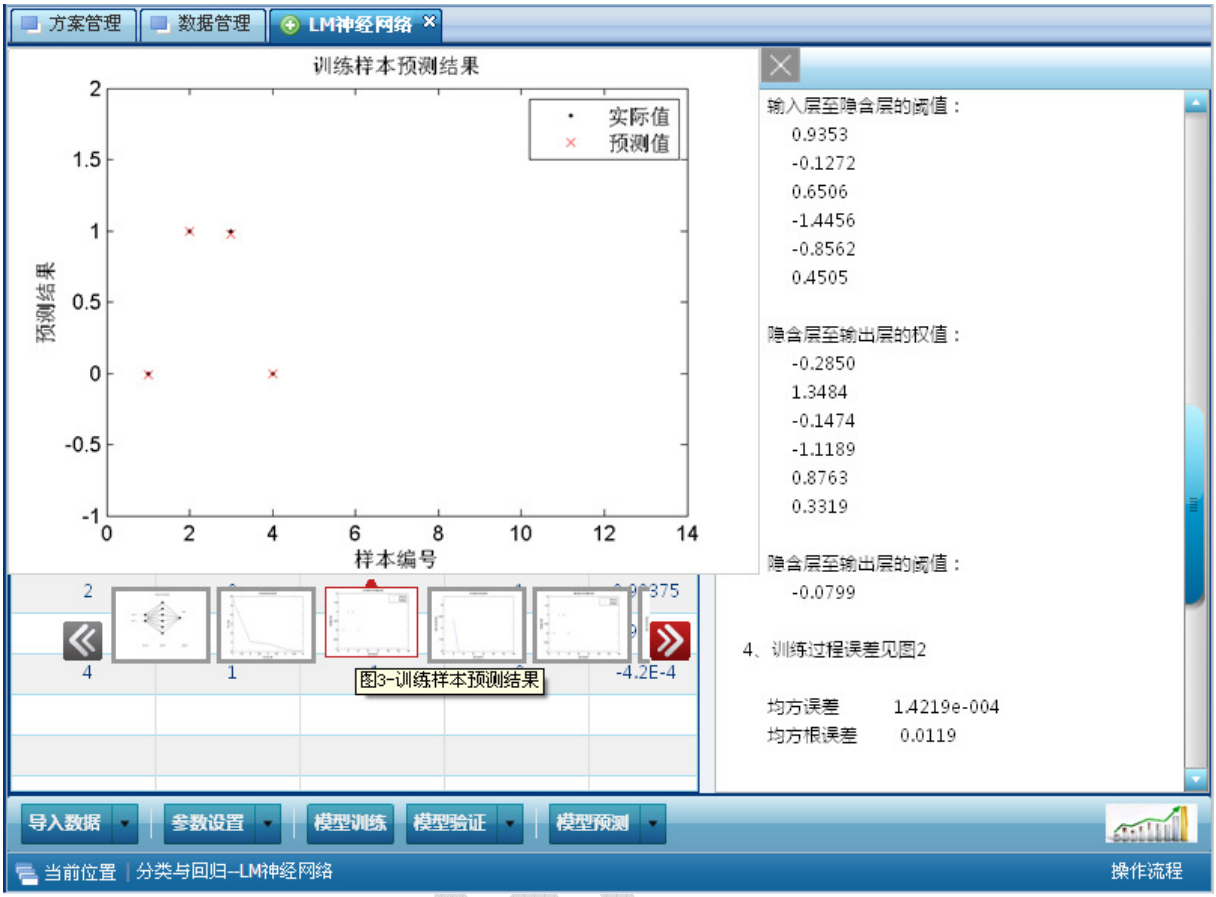
www.dip


方案管理
数据管理
LM神经网络
BP神经网络结构图
输入2
输入1
输出
输入层
隐含层
输出层
模型训练信息 ===
样本总数：4，训练耗时：16.1579秒
BP神经网络结构见图1，具体信息如下：
2个输入层节点，6.0个隐含层节点，1个输出层节点
输入属性：
输入1
输入2
输出属性：
输出
BP神经网络建模参数：
显示间隔次数：25.0
最大循环次数：100.0
目标均方误差：0.02
学习速度：0.0010
学习速率增加比率：10.0
学习速率减少比率：0.1
最大误差比率：1.0E8
动量常数：0.0010
图1-BP神经网络结构图
导入数据
参数设置
模型训练
模型验证
模型预测
当前位置 | 分类与回归--LM神经网络
操作流程

✧ 训练误差

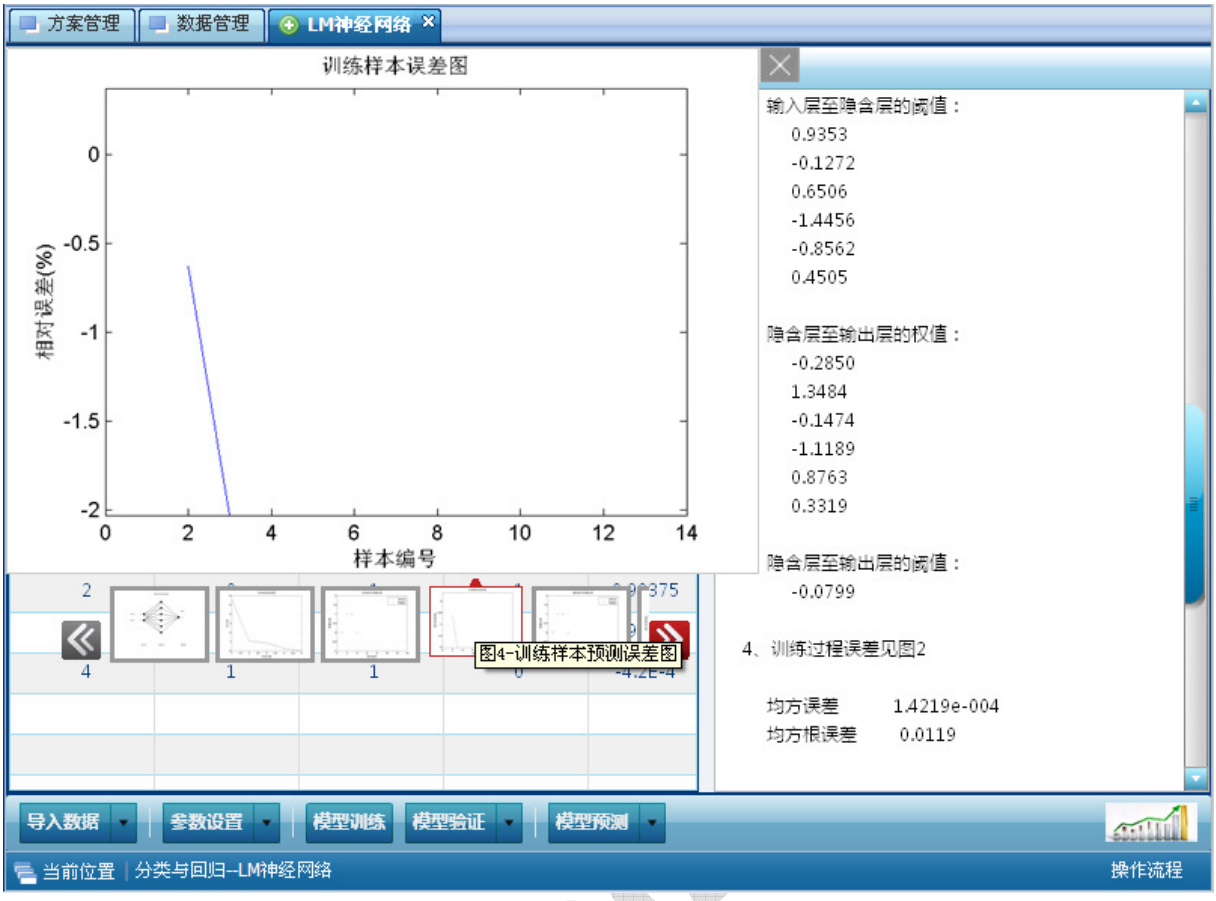

方案管理
数据管理
LM神经网络
训练误差曲线图
均方差
训练步数
输入层至隐含层的阈值：
0.9353
-0.1272
0.6506
-1.4456
-0.8562
0.4505
隐含层至输出层的权值：
-0.2850
1.3484
-0.1474
-1.1189
0.8763
0.3319
隐含层至输出层的阈值：
-0.0799
4、训练过程误差见图2
均方误差 1.4219e-004
均方根误差 0.0119
图2-BP神经网络训练误差图
导入数据
参数设置
模型训练
模型验证
模型预测
当前位置 | 分类与回归--LM神经网络
操作流程

✧ 训练样本预测结果

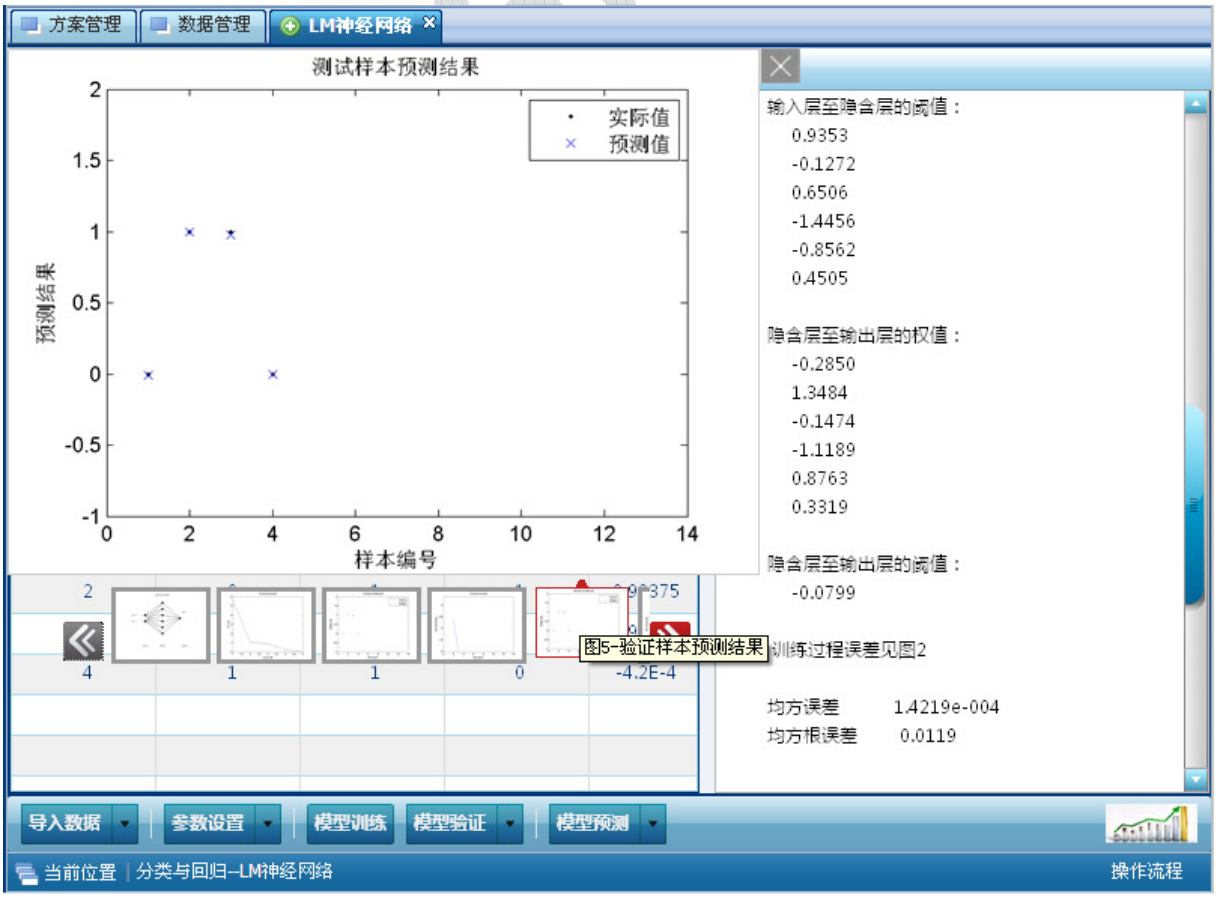


✧ 训练样本误差图

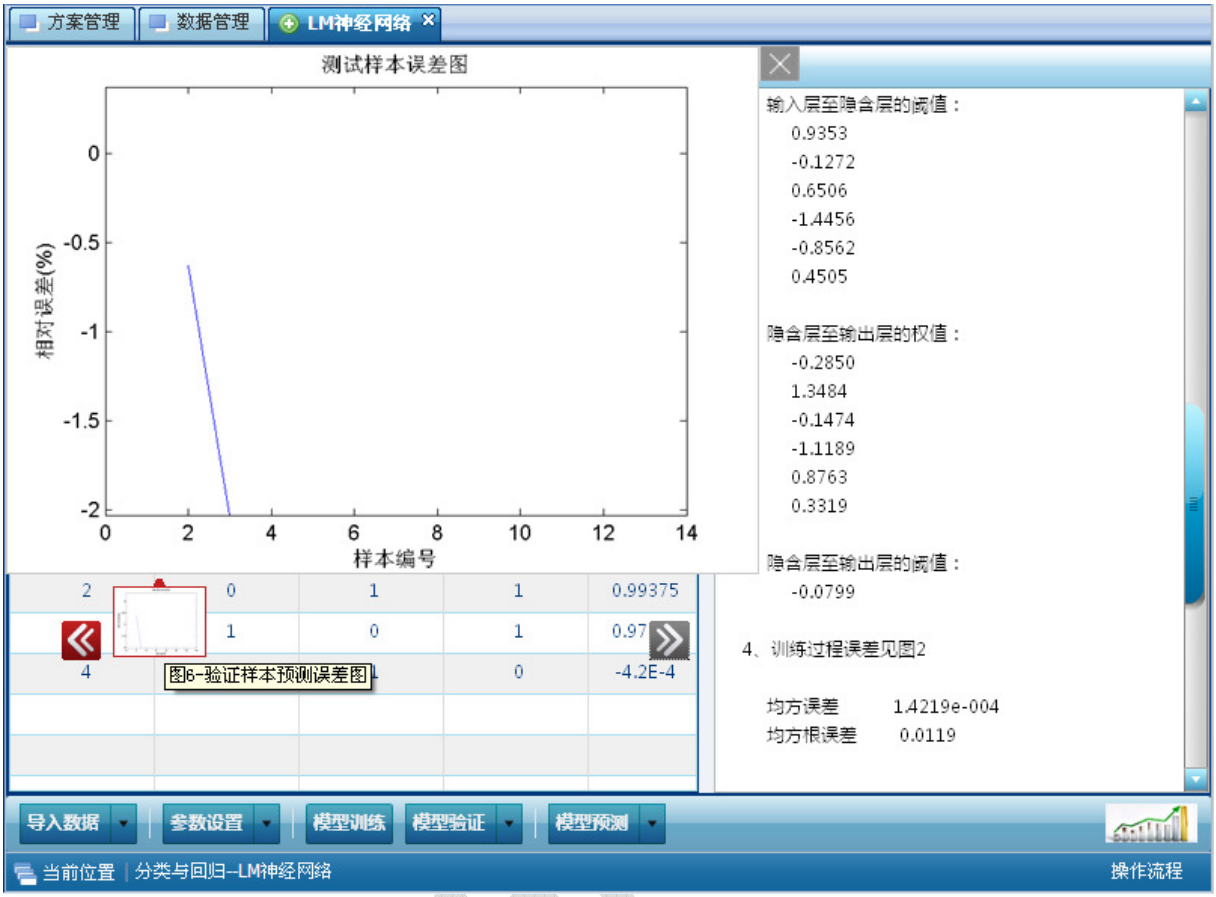
www.dip


方案管理
数据管理
LM神经网络
训练样本误差图
相对误差(%)
样本编号
输入层至隐含层的阈值：
0.9353
-0.1272
0.6506
-1.4456
-0.8562
0.4505
隐含层至输出层的权值：
-0.2850
1.3484
-0.1474
-1.1189
0.8763
0.3319
隐含层至输出层的阈值：
-0.0799
4、训练过程误差见图2
均方误差 1.4219e-004
均方根误差 0.0119
图4-训练样本预测误差图
导入数据
参数设置
模型训练
模型验证
模型预测
当前位置 | 分类与回归--LM神经网络
操作流程

✧ 验证样本预测结果图


方案管理
数据管理
LM神经网络
测试样本预测结果
实际值
预测值
预测结果
样本编号
输入层至隐含层的阈值：
0.9353
-0.1272
0.6506
-1.4456
-0.8562
0.4505
隐含层至输出层的权值：
-0.2850
1.3484
-0.1474
-1.1189
0.8763
0.3319
隐含层至输出层的阈值：
-0.0799
训练过程误差见图2
均方误差 1.4219e-004
均方根误差 0.0119
图5-验证样本预测结果
导入数据
参数设置
模型训练
模型验证
模型预测
当前位置 | 分类与回归--LM神经网络
操作流程

✧ 验证样本误差图



⑥ 模型预测

www.dip



2）如何创建一个数据挖掘模型

最经典的一个网络模型是 XOR（eXclusive-OR）问题。其问题实质是：如果两个输入项中的一个为 true，其输出为 true，但如果两个输入项均为 false 或 true，则输出为 false。

其关系如下图：

Input 1	Input 2	Output
false	false	false
false	true	true
true	false	true
true	true	false

用神经网络算法处理这个问题，需要有三个神经元。

这个问题使用 true 和 false 数据，本例一般的做法是，用 1 和 0 来表示，网络模型结构如下：

- ✧ 两个节点的输入层
- ✧ 一个节点的输出层

下一步骤，“如何创建和打开一个方案” 将演示如何创建一个数据挖掘模型。

3）如何创建和打开一个方案

用户登录后，默认进入方案主界面，如图：



TipDM 允许用户创建一个新的数据挖掘方案，也可以打开一个已有的数据挖掘方案。

① 选择某个应用类别，并点“创建”按钮，弹出创建新方案对话框，如下：



输入方案名称、方案描述，并选择应用类别后，点“确定”保存方案。

② 双击“最近方案”列表中的某条记录，即可打开选中方案，用户可以对当前方案进行编辑。

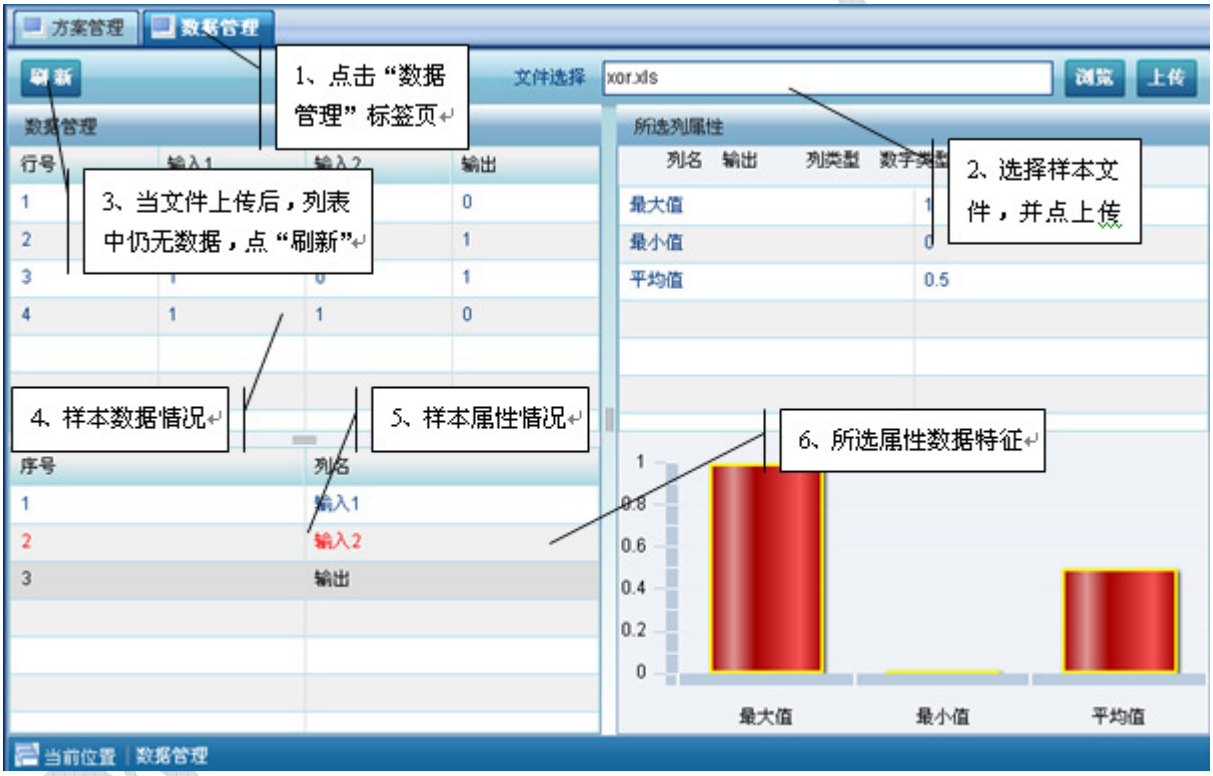
③ 用户也可以点击“更多”按钮，从弹出对话框中选中一条记录，作为当前方案，如下图：



说明：用户当前的所有操作都是针对当前方案的。

4）如何在方案中加载数据

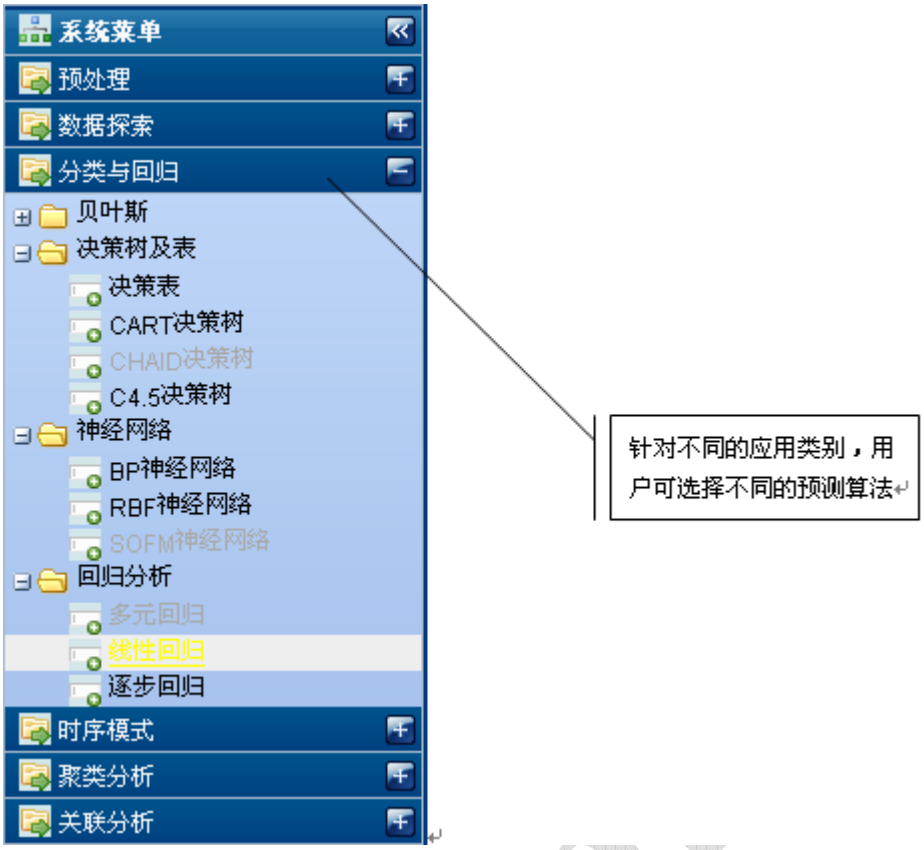
切换到“数据管理”标签页，上传并加载样本数据：



说明：用户当前的所有操作都是针对当前数据体的。

5）当前方案该选用哪种算法

针对当前方案的应用类别，用户可选择菜单中不同的数据挖掘算法进行预测建模。



4.4. 产品特点

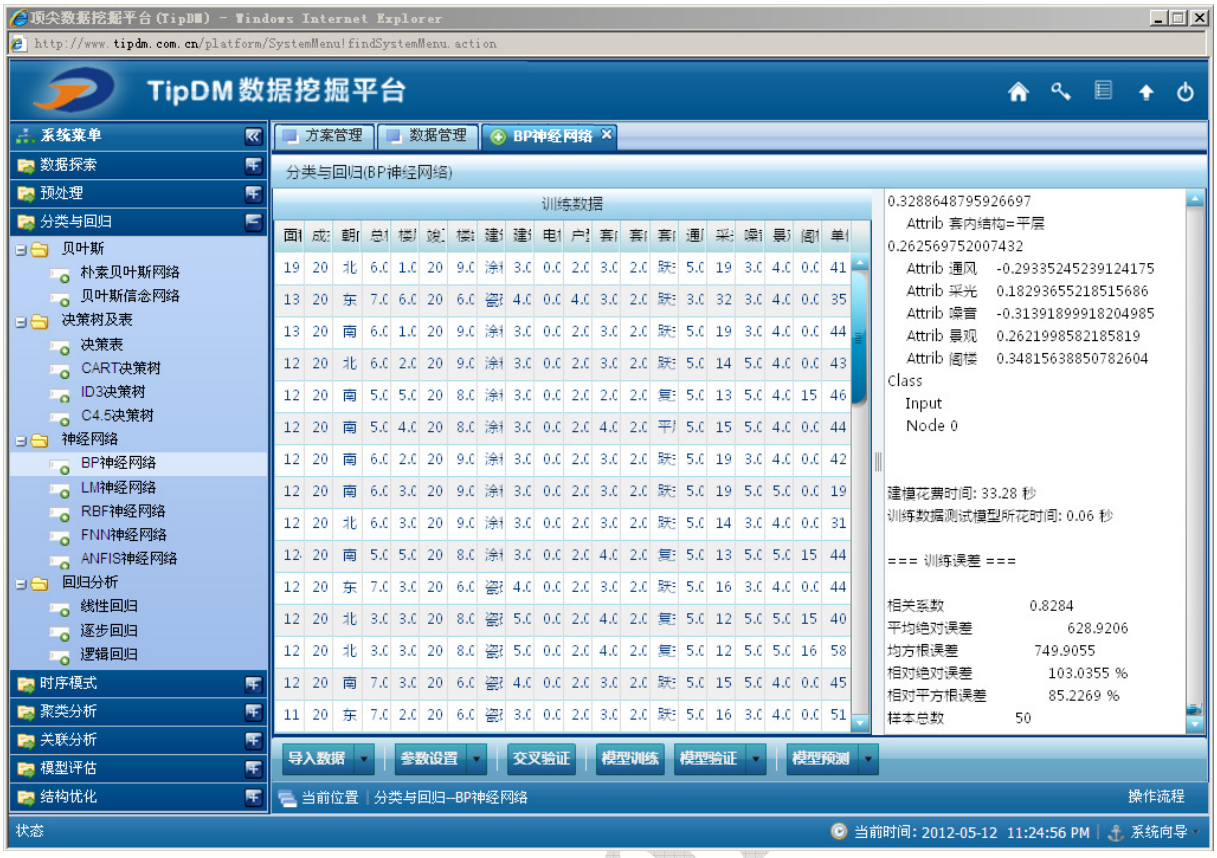
✧ 支持CRISP-DM数据挖掘标准流程

CRISP-DM是数据挖掘的标准商业流程，与仅仅局限在技术层面上的数据挖掘方法论不同，CRISP-DM 把数据挖掘看作一个商业过程，并将其具体的商业目标映射为数据挖掘目标。有调查结果显示，目前绝大部分数据挖掘工具均采用CRISP-DM 的数据挖掘流程，它已经成为事实上的行业标准。

TipDM 完全支持CRISP-DM 标准，这不但规避了许多常规错误，而且其显著的智能预测模型有助于快速解决出现的问题。

✧ 提供丰富的数据挖掘模型和灵活算法

TipDM提供多种数据挖掘建模方法，这些方法分别来自于机器学习、人工智能和统计学。每种方法都有自己的长处，并且可以整合在一起使用，可以灵活的解决各种类型的问题。这些模型算法分成五类：数据探索与预处理、分类与回归、聚类分析、关联规则挖掘、时序模式。



- **数据探索与预处理算法：**

1） 数据探索：相关性分析、主成分分析、周期性分析、缺失值处理

2） 数据预处理：平滑处理、坏数据处理、属性选择、数据规约、离散处理、特征提取

- **分类与回归算法：**

1） 贝叶斯：朴素贝叶斯网络、贝叶斯信念网络

2） 决策树及表：决策表、CART 决策树、ID3 决策树、C4.5 决策树

3） 神经网络：BP 神经网络、LM 神经网络、RBF 神经网络、FNN 神经网络、ANFIS 神经网络、WNN 神经网络

4） 回归分析：线性回归、逐步回归、逻辑回归、保序回归

5） 其它：AdaBoostM1 算法、KStar 算法、SVM 支持向量机、K-最近邻分类

- **聚类分析算法：**

1） K-Means 均值聚类

2） EM 最大期望聚类

3） DBScan 密度聚类

4） 改进 K-均值聚类

5） 多层次聚类（系统聚类）

- **关联分析类算法：**

1） HotSpot 关联规则

2） Apriori 关联规则

3） FP-Tree 关联规则

4） HOTSPOT 关联规则

5） 灰色关联法

- **时序模式算法：**

1） 指数平滑

2） 多元回归

3） GM 灰色理论

4） RBF 神经网络

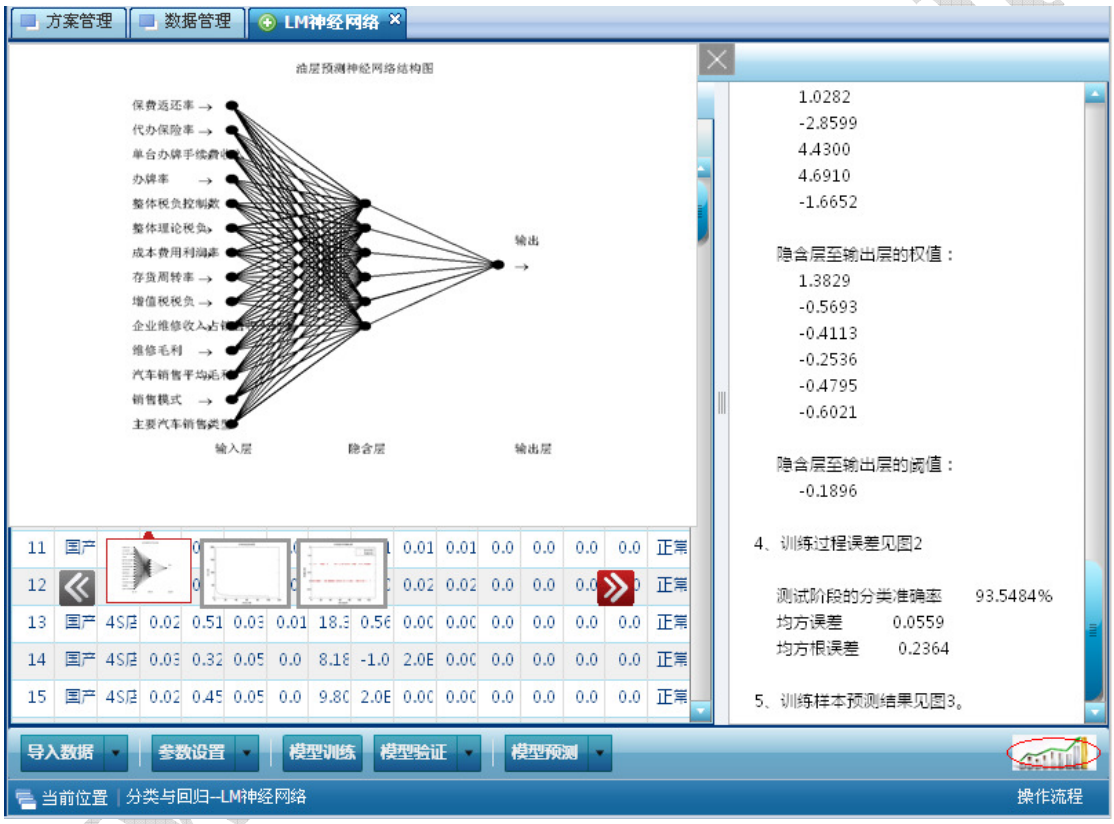
5） ANFIS 神经网络

6） SVM 支持向量机

✧ 提供丰富的图表输出效果

TipDM提供丰富的建模过程图表，示例如下：

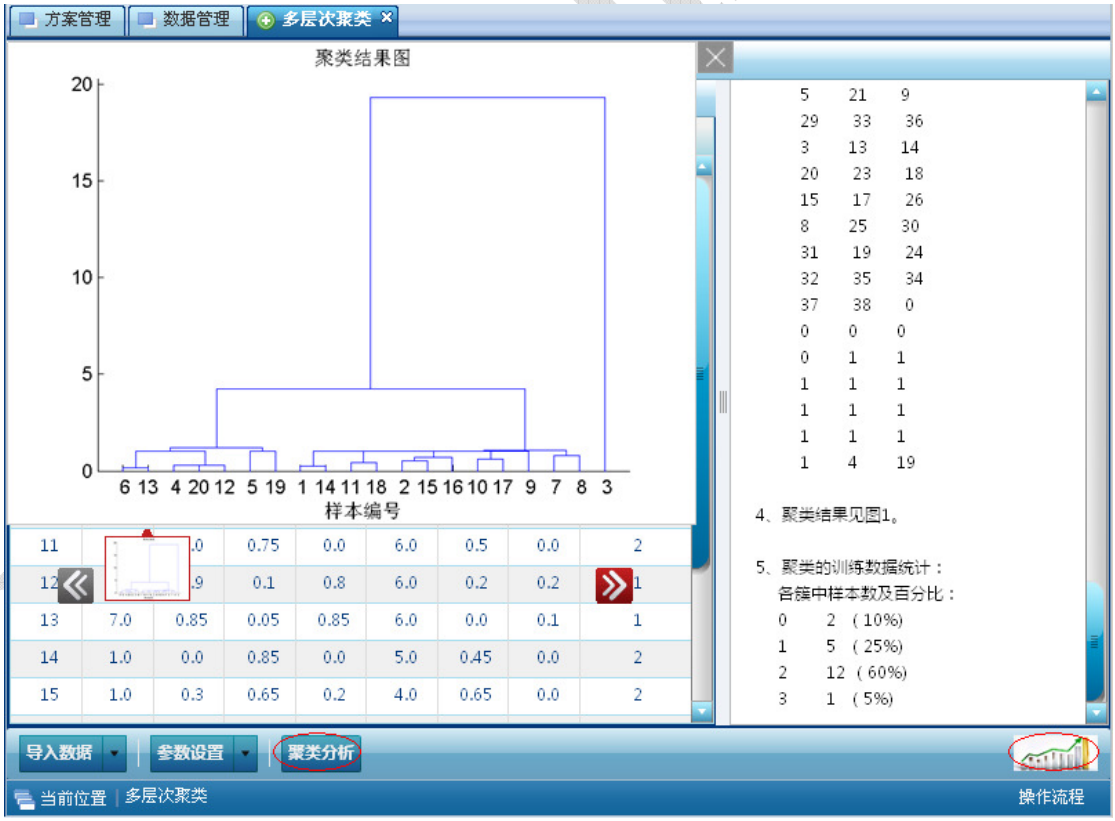
1） 神经网络结构图



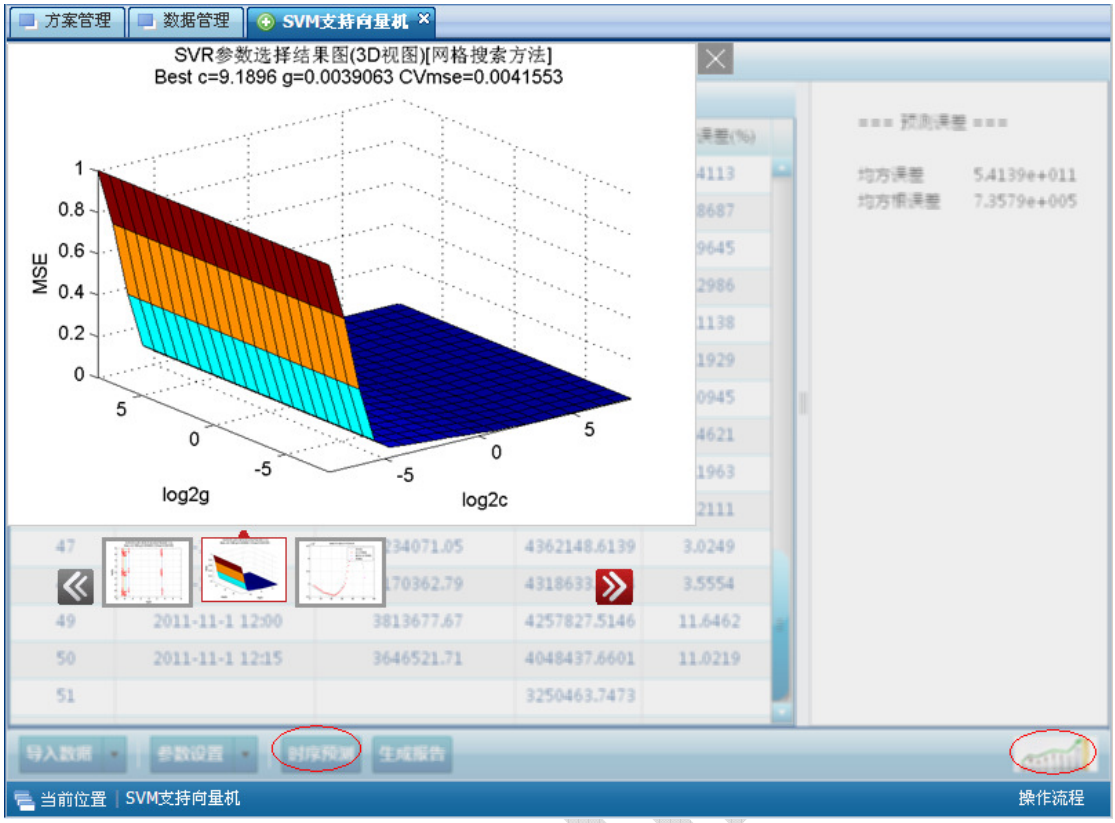
2） 建模过程输出



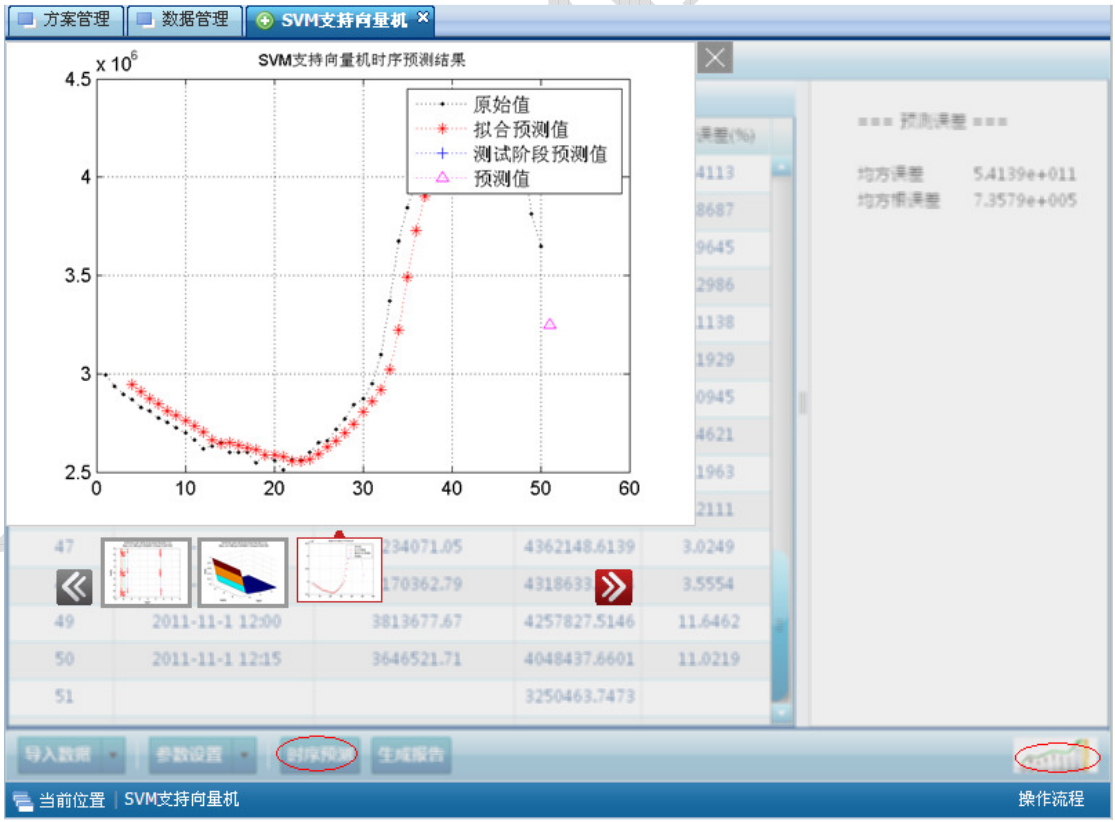
3) 多层次聚类图



4) SVM 参数选择结果图



5) 时序预测结果图



✧ 具有多模型的整合能力，使得生成的模型更加稳定和高效

TipDM通过数据流的方式构建数据挖掘模型，用户可以把不同模型按照需求顺次连接就可以达到整合多个模型的目的。举例来说：建模前用户可以选择先对样本属性进行主成分分析或属性选择进

行降维处理，然后再指定某个算法进行建模。另外TipDM还提供了多种模型效果的评估技术和思路（例如收益图表、投资回报图表、利润图表、响应图表及各类表格等）来检验模型的效果，从而使用户可以选择最稳定、高效的模型进行发布。

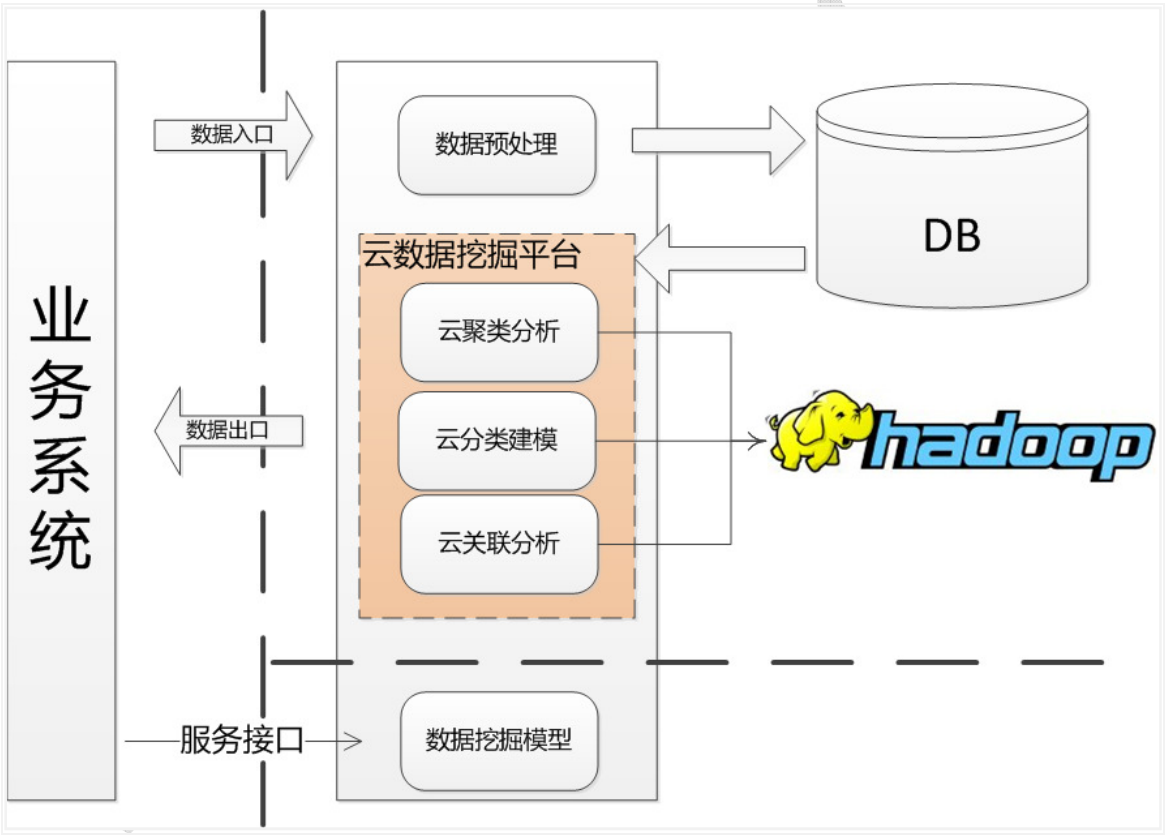
✧ 提供灵活多样的应用开发接口

TipDM提供一套基于行业标准的编程接口及常用的数据挖掘算法。它可用于开发各类数据挖掘应用程序，从简单的预测建模到庞大的集成系统。数据引擎可由JDBC和XML访问分析行业标准数据挖掘API。

TipDM提供Web Service、DLL 或是 JAR三种不同的使用接口，方便第三方软件商集成开发，快速构建出大型企业级海量数据挖掘应用系统。

✧ 海量数据的处理能力

TipDM高级版本是基于云计算平台分布式文件系统HDFS、并行计算框架MapReduce和MPP数据仓库基础上搭建的。由于架构在云计算平台之上，因此TipDM克服了传统工具的问题，能够处理TB级的海量数据挖掘，具备了双向扩展、高容错性、易于部署等特点。



TipDM具有并行、多线程处理能力，并能提供优化机制以保障在海量数据和大规模计算时的性能。

✧ 适应不同类型层次人员需求

针对不同用户类型，TipDM推出四个版本：个人版、企业版、嵌入版和教研版，以满足不同应用层次人员的需求。

1）个人版：即在线数据挖掘（www.tipdm.cn,www.tipdm.com.cn），用户不用安装系统，而是通过浏览器直接在线使用，适合有数据挖掘研究需要的单位及个人。利用系统提供的方案管理、数据管理、数据探索、数据预处理、预测建模、模型评价等功能，快速引导用户实现数据挖掘建模，同

时结合系统自带的数据挖掘教程及参考案例，通过模板，可快速输出用户所需的预测分析报告。

2）企业版：适合有数据挖掘分析需要的企业，项目框架来源于目前基于云计算和SOA架构的海量数据挖掘平台的B/S版本，即在此平台框架下，保持现有应用框架、核心算法、基本功能模块，仅增加业务功能模块，进行快速二次开发。

3）嵌入版：适合第三方集成商有数据挖掘应用需要，但本身缺乏这方面的技术资源，在投入较少成本的前提下，利用TipDM提供的各类数据挖掘算法标准接口，快速实现数据挖掘应用的开发。

4）教研版：主要适合高校及科研院所教学及科学研究，数据挖掘技术近年来越来越受到人们的重视，她被认为是21世纪最具有发展前途的领域。因此国内不少高校将此内容引入本科教学中，在自动化、电子信息、金融等专业开设了数据挖掘技术相关的课程。由于条件和学时所限，这些内容的教学主要限于理论介绍。因为抽象，学生理解起来往往比较困难，教学效果不甚理想。基于以上原因，TipDM推出了适合高校教学的数据挖掘辅助教学套件，教学套件包括：辅助教学建模软件、辅助教学多媒体软件、辅助教学PPT、辅助教学实训教程。

www.tipdm.com

第5章 上机实践

温馨提示：

- 需支持，找广州太普技术：www.tipdm.com
- 要建模，上数据挖掘在线平台：www.tipdm.cn；www.tipdm.com.cn
- 有问题，找太普问库：wenku.tipdm.comhttp://www.tipdm.com/teach.php?act=list
- 想兼职，找数据挖掘创业联盟：518.tipdm.com

实验（实践）环节内容及其要求：采用 TipDM 数据挖掘工具软件，学生根据要求通过数据挖掘工具实现典型数据挖掘算法。

教师在课程开始前完成实验环境的准备工作；实验过程中对学员出现的问题给予及时指导，帮助学员完成实验内容；实验结束后，对学员的实验情况进行及时整理分析，并提出相应建议，以提高后续实验质量。

学员在实验前根据老师布置的内容和要求认真准备本次实验操作的内容，包括认真阅读参考资料、熟悉实验主要内容、步骤和预期实验结果；认真完成实验内容，出现问题及时和指导教师进行沟通；实验结束后认真总结，巩固课堂学习内容，提高运用所学理论知识和方法的能力。

实验一 数据探索及数据预处理

1. 实验目的

- ✧ 加深对主成分分析、相关性分析及属性选择等功能的理解及使用；
- ✧ 掌握主成分分析、相关性分析及属性选择的应用范围；
- ✧ 掌握使用主成分分析、相关性分析及属性选择解决实际问题的方法。

2. 实验内容

1）熟悉 TipDM 平台；

2）掌握相关性分析及主成分分析方法并对输出结果进行分析；

3）掌握属性选择功能及不同参数条件下对输出结果的影响。

3. 实验方法与步骤

- ✧ 数据准备

① 使用银行对资产进行评估的数据 bank-data.xls，数据里有 12 个属性，分别是 id,age,sex,region,income,married,children, car,save_act, current_act, mortgage,pep。共 600 个实例。

② 将含有 600 个实例的“bank-data.xls”文件中取出 550 个实例用做训练集,在另外 300 个实例中取 50 个作为待预测的实例。

③ 本测试集中将 pep 属性作为分类的目标类别。

- ✧ 创建方案：登录 TipDM 平台后的默认页面即为“方案管理”，在此页面，选择“分类与回归”创建一个新方案，方案名称：数据探索及数据预处理实验。方案描述：使用银行对资产进行评估的数据，进行数据探索及数据预处理。信息输入完成后点“确定”保存方案。
- ✧ 上传数据：进入“数据管理”标签页，选择《样本数据.xls》，并上传，上传的数据将自动显示在列表框中，若不能正确显示，点“刷新”按钮。
- ✧ 主成分分析

  “系统菜单”中选择“数据探索→主成分分析”。

  a）导入数据：除 id 外，选择 age,sex,region,income,married,children, car,save_act, current_act, mortgage,pep 等属性（最后一列为输出项），以及样本数据范围（从第 1 行至 550 行），导入样本数据到当前算法操作区。

  b）参数设置：可设置“核估计”、“预测输出”等参数。

  c）进行主成分分析，并对输出结果进行解释。
- ✧ 相关性分析

  “系统菜单”中选择“数据探索→相关性分析”。

  a）导入数据：除 id 外，选择 age,sex,region,income,married,children, car,save_act, current_act, mortgage,pep 等属性（最后一列为输出项），以及样本数据范围（从第 1 行至 550 行），导入样本数据到当前算法操作区。

  b）参数设置：可设置“计算方法”等参数。

  c）进行相关性分析，并对输出结果进行解释。
- ✧ 属性选择

  “系统菜单”中选择“数据预处理→属性选择”。

  a）导入数据：除 id 外，选择 age,sex,region,income,married,children, car,save_act, current_act, mortgage,pep 等属性（最后一列为输出项），以及样本数据范围（从第 1 行至 550 行），导入样本数据到当前算法操作区。

  b）参数设置：可设置“属性评估器”、“搜索策略”等参数。

  c）进行属性选择，并对选择结果进行解释。

4. 思考与实验总结

- ✧ 进行相关性分析时，计算方法分别选择：Pearson、Sperarman、Kendall 时，分析输出结果的变化。
- ✧ 当采样数据维度过大，如何进行降维处理？

实验二 神经网络模型的构建与使用

www.tipdm.com

1. 实验目的

- ✧ 加深对 LM 神经网络原理的理解及使用；
- ✧ 掌握 LM 神经网络的特性和应用范围；
- ✧ 掌握使用 LM 神经网络解决实际问题的方法。

2. 实验内容

本次实验使用 BP 神经网络中的 Levenberg-Marquardt 优化算法进行水质分类建模，并观察 BP 神经网络的各个参数对 BP 神经网络的影响：

- ✧ 水质评价中，常规重要的量化指标主要有 DO，pH，三态氮，磷，水色，饲料投喂技术，浮游植物等，本实验中通过选取 DO、PH、nh、NO2-N、SD 等指标，来建立水体环境质量评价量化模型，实现池塘水质智能化识别，实验数据见《样本数据.xls》。
- ✧ 在初始参数下，观测 BP 神经网络的训练过程及使用测试数据进行仿真时的输出，然后将最大训练次数设置为 100、500、1000 等等，分析网络的输出效果，以便确定合适的训练次数。
- ✧ 在保持其它参数不变的情况下，改变隐含层神经网络的个数，例如取 2 个、8 个、10 个、20 个等等，观测 BP 神经网络的训练过程和输出，以便最终确定隐含层神经网络的个数。
- ✧ 在 BP 人工神经网络已经训练好的情况下，从样本数据中抽取若干数据进行模型测试，测试训练好的 BP 人工神经网络能否正确逼近它们。

注意：模型训练时可选取 80%样本用于建模，另 20%用作模型验证。

3. 实验方法与步骤

- ✧ 数据准备

  ① 准备样本数据文件：实验二样本数据.xls。

  ② 将“实验二样本数据.xls”文件的 350 条数据中前 300 条数据作为训练数据集，后 50 条数据作为测试数据集。

- ✧ 创建方案：登录 TipDM 平台后的默认页面即为“方案管理”，在此页面，选择“分类与回归”创建一个新方案，方案名称：基于 BP 神经网络算法的分类预测模型。方案描述：基于水质评价量化指标 DO、PH、nh、NO2-N、SD，建立水体环境质量评价量化模型，实现池塘水质智能化识别。信息输入完成后点“确定”保存方案。
- ✧ 上传数据：进入“数据管理”标签页，选择《样本数据.xls》，并上传，上传的数据将自动显示在列表框中，若不能正确显示，点“刷新”按钮。
- ✧ 预测建模

  “系统菜单”中选择“分类与回归→神经网络→LM 神经网络”。

www.tipdm.com

a）导入数据：选择 DO、PH、nh、NO2-N、SD 等属性（最后一列为输出项），以及样本数据范围（从第 1 行至 300 行），导入样本数据到当前算法操作区。

b）参数设置：可设置“显示间隔次数”、“最大循环次数”、“目标误差”、“动量常数”、“学习速率”、“学习速率增加比率”、“学习速率减少比率”、“最大误差比率”、“输入层至隐含层传递函数”、“隐含层至输出层传递函数”，以及隐层节点数等参数。

c）模型训练：对导入的专家样本数据，进行模型训练。分析训练过程中输出的相关信息。

d）模型验证：导入 301 至 350 行共 50 条样本，对训练好的模型进行验证。分析验证过程中输出的相关信息。

e）模型预测：对 DO、PH、nh、NO2-N、SD 等属性输入数据，并点“模型预测”。分析输出的有关预测信息。

4. 思考与实验总结

- ✧ 本次实验中构建的 LM 神经网络的合适的训练数据大致为多少（可以给出一个范围）？并说明理由。
- ✧ 本次实验中构建的 LM 神经网络的合适的隐含层神经元的个数为多少（可以给出一个范围）？并说明理由。
- ✧ 分析训练样本验证误差与测试样本验证误差。模型验证时样本数据应该如何选取？
- ✧ 总结 LM 神经网络建模的步骤。
- ✧ 试用 BP 神经网络、RBF 径向基神经网络、FNN 模糊神经网络、ANFIS 神经模糊推理系统分别对本实验数据进行建模，并比较建模效果。

www.ilpdm.com

实验三 决策树模型的构建与使用

1. 实验目的

- ✧ 加深对决策树 ID3 算法原理的理解及使用；
- ✧ 掌握决策树 ID3 算法的特性和应用范围；
- ✧ 掌握使用决策树 ID3 算法解决实际问题的方法。

2.实验内容

在肝病评价中，常规重要的量化指标主要有 x1（食道经脉曲张），x2（门脉癌栓），x3（Hbsag），x4（Anti-HCV），x5（肿瘤部位），x6（肿瘤大小），x7（肿瘤生长方式），x8（肿瘤包膜），x9（肿旁的微小子灶） x10（术后腹水），DECISION（是否复发）等，本实验中通过选取 x1,x2,x3,x4,x5,x6,x7,x8,x9 x10,DECISION 等指标，来建立肝病的评价，分析是否复发。实验数据见《样本数据》。

注意：模型训练时可选取 80%样本用于建模，另 20%用作模型验证。

3.实验方法与步骤

- ✧ 数据准备

  ① 准备样本数据文件：《样本数据.xls》。

  ② 将“《样本数据.xls》”文件的 20 条数据中前 16 条数据作为训练数据集，后 4 条数据作为测试数据集，它们的“DECISION”属性都设为缺失值，并保存为文件。

- ✧ 创建方案：登录 TipDM 平台后的默认页面即为“方案管理”，在此页面，选择“数据分类”创建一个新方案。方案名称：决策树 ID3 算法在医院数据挖掘中的应用探索。方案描述：以肝癌为例对医院临床数据进行挖掘，经过数据预处理，应用 ID3 算法建立完整的决策树。用测试数据测试，该决策树具有很高的准确率，指出决策树算法在医院信息系统中具有广泛的应用前景。信息输入完成后点“确定”保存方案。
- ✧ 上传数据：进入“数据管理”标签页，选择《样本数据.xls》，并上传，上传的数据将自动显示在列表框中，若不能正确显示，点“刷新”按钮。
- ✧ 预测建模

  从“系统菜单”中选择“分类与回归→决策树→ID3 决策树”。

  a）导入数据：选择 x1,x2,x3,x4,x5,x6,x7,x8,x9 x10,DECISION 等属性（最后一列为输出项），以及样本数据范围（从第 1 行至 16 行），导入样本数据到当前算法操作区。

  b）参数设置：“预测输出”参数。

  c） 交叉验证：为了保证生成的模型的准确性而不至于出现过拟合(overfitting)的现象，

www.tipdm.com

有必要采用交叉验证来评估模型。这里采用 10-折交叉验证(10-fold cross validation)。分析交叉验证时输出的相关信息。

d）模型训练：在通过交叉验证对模型进行初步评估后，接下来可对导入的专家样本数据，进行模型训练。分析训练过程中输出的相关信息。

e）模型验证：导入 17 至 20 行共 4 条样本，对训练好的模型进行验证。分析验证过程中输出的相关信息。

f）模型预测：对 x1,x2,x3,x4,x5,x6,x7, x8,x9 x10 等属性输入数据，并点“模型预测”。分析输出的有关预测信息。

4.思考与实验总结

- ✧ 在进行交叉验证时，请对比分析“训练误差”与“分层交叉验证误差”的有关输出信息。
- ✧ 请说明“正确肯定率”、“错误肯定率”、“精确率”、“反馈率”、“ROC Area”等名词所代表的意义。
- ✧ 请解释说明模型验证时，输出的混淆矩阵信息。
- ✧ 试用 C4.5 决策树、CART 决策树和决策表分别对本实验数据进行建模，并比较建模效果。

www.tipdm.com

实验四 聚类算法的构建与使用

1. 实验目的

- ✧ 加深对 K-均值聚类算法原理的理解及使用；
- ✧ 掌握 K-均值聚类算法的特性和应用范围；
- ✧ 掌握使用 K-均值聚类算法解决实际问题的方法。

2. 实验内容

本次实验通过选择具有攻击特征的指标，利用 K-均值聚类分析来进行网络入侵检测，以区别攻击、异常、正常等行为，并观察 K-均值聚类算法的各个参数对聚类结果的影响。

- ✧ 入侵检测指通过对系统或网络资源进行实时检测，及时发现闯入系统或网络的入侵者。入侵检测的数据采集内容包括系统日志、网络数据包、重要文件以及用户活动的状态与行为等。本实验选择具有攻击特征的若干指标，利用聚类分析来进行网络入侵检测，以区别攻击、异常、正常等行为，实验数据见《样本数据.xls》。
- ✧ 在初始参数下，观测 K-均值聚类算法的输出结果，然后将最聚类数设置为 3、4 等等，分析聚类的输出效果，以便确定合适的聚类数。

3.实验方法与步骤

- ✧ 数据准备

  ① 准备样本数据文件：样本数据.xls。

  ② 样本数据各列属性：

  - ✓ Count：在一个时间窗口内目标主机是与当前连接相同的连接次数。
  - ✓ Serror：出现 SYN 错误的连接百分比。
  - ✓ Same_srv：目标端口(service)相同的连接所占的百分比。
  - ✓ Diff_srv：目标端口(service)不同的连接所占的百分比。
  - ✓ Srv_count：目标端(service)与当前连接相同的连接次数。
  - ✓ Srv_serror：出现 SYN 错误的连接百分比。
  - ✓ Rv_diff_host：目标主机不同的连接所占的百分比。

- ✧ 创建方案：登录 TipDM 平台后的默认页面即为“方案管理”，在此页面，选择“聚类分析”创建一个新方案，方案名称：基于聚类分析的网络入侵检测。方案描述：通过选择具有攻击特征的指标，利用 K-均值聚类分析来进行网络入侵检测，以区别攻击、异常、正常等行为。信息输入完成后点“确定”保存方案。
- ✧ 上传数据：进入“数据管理”标签页，选择《样本数据.xls》，并上传，上传的数据将自动显示在列表框中，若不能正确显示，点“刷新”按钮。

www.tipdm.com

✧ 聚类分析

“系统菜单”中选择“聚类分析→K-均值算法”。

a）导入数据：选择 Count、Serror、Same_srv、Diff_srv、Srv_count、Srv_serror、Rv_diff_host 等属性，以及样本数据范围（从第 1 行至 20 行），导入样本数据到当前算法操作区。

b）参数设置：可设置“距离函数”、“最大迭代次数”、“聚类数”、“种子数”等参数。

c）聚类分析：对导入的样本数据，进行聚类分析。分析聚类过程中输出的相关信息。

4.思考与实验总结

✧ 本次实验中用 K-均值聚类进行网络入侵检测，聚类数设置为多少比较合适？请说明理由。

✧ 比较 K-均值聚类、DBScan 聚类、多层次聚类在聚类参数和聚类输出上有什么区别？

✧ 总结 K-均值聚类的步骤。

✧ 改用其它聚类算法，如改进 K-均值聚类、DBScan 聚类、多层次聚类等，观察聚类参数及聚类输出结果。

www.tipdm.com

实验五 关联规则模型的构建与使用

1. 实验目的

- ✧ 加深对 Apriori 关联规则算法基本原理的理解；
- ✧ 掌握 Apriori 关联规则算法的特性和应用范围；
- ✧ 掌握使用 Apriori 关联规则算法解决实际问题的方法。

2.实验内容

本次实验使用 Apriori 关联规则分析某工程技术研究中心的投入产出比与研发人员占比的关系，观察样本集中的实例，进行关联信息获取，更好地理解和掌握关联规则算法的基本原理和解决实际问题的方法。

1、对某工程技术研究中心的 30 家数据（研发人员、总人数、总投入、总收入），利用关联规则分析投入产出比与研发人员占比的关系，实验数据见《样本数据.xls》。

2、为方便关联规则挖掘，对样本数据进行预处理：增加便于关联信息获取的数据列。

3、设置不同的参数（最小置信度、最小支持度下界、最小支持度上界、增量、规则条数）时，观测 Apriori 关联规则挖掘结果。

3.实验方法与步骤

- ✧ 数据准备

  ① 准备样本数据文件：样本数据.xls。

  ②在《样本数据.xls》中新增研发人员比例范围、投入产出比率范围数据列：

  - ✓ A1 表示研发人员比例范围<50%
  - ✓ A2 表示研发人员比例范围 50%~70%
  - ✓ A3 表示研发人员比例范围 70%~85%
  - ✓ A4 表示研发人员比例范围 85%~95%
  - ✓ M1 表示投入产出比率大于 80%
  - ✓ M2 表示投入产出比率 40%~80%
  - ✓ M3 表示投入产出比率 15%~40%
  - ✓ M4 表示投入产出比率<15%

- ✧ 创建方案：登录 TipDM 平台后的默认页面即为“方案管理”，在此页面，选择“关联规则”创建一个新方案，方案名称：基于 Apriori 算法的关联规则挖掘模型。方案描述：对某工程技术研究中心的 30 家数据（研发人员、总人数、总投入、总收入），利用关联规则分析投入产出比与研发人员占比的关系。信息输入完成后点“确定”保存方案。

www.tipdm.com

- ✧ 上传数据：进入“数据管理”标签页，选择《样本数据.xls》，并上传，上传的数据将自动显示在列表框中，若不能正确显示，点“刷新”按钮。
- ✧ 关联规则挖掘

  “系统菜单”中选择“关联分析→Apriori 关联规则”。

  a）导入数据：选择数据列：研发人员比例范围、投入产出比率范围，以及样本数据范围（从第 1 行至 30 行），导入样本数据到当前算法操作区。

  b）参数设置：主要设置“最小置信度”、“最小支持度下界”、“最小支持度上界”、“增量”、“规则条数”等参数。

  c）规则挖掘：对导入的样本数据，进行关联规则挖掘。分析关联规则挖掘过程中输出的相关信息。

4.思考与实验总结

- ✧ Apriori 算法进行关联规则挖掘时，对样本数据有什么样的要求？
- ✧ 实验中，“最小置信度”、“最小支持度下界”、“最小支持度上界”、“增量”、“规则条数”等参数应该怎样设置才合理？请说明理由。
- ✧ 对关联规则挖掘输出的项集和规则进行解释。针对本实验案例，结论是什么？

www.tipdm.com

实验六 时间序列模型的构建与使用

1.实验目的

- ✧ 加深对指数平滑原理的理解及使用；
- ✧ 掌握指数平滑的特性和应用范围；
- ✧ 掌握指数平滑算法解决实际问题的的方法

2.实验内容

本次实验把指数平滑应用在股票的预测中，将采用时间序列的指数平法法来对股票走势进行预测分析。已达到预期的效果。

3.实验方法与步骤

- ✧ 数据准备

  ① 准备样本数据文件: stack-data.xls。

  ② 本设计数据来源于大智慧，数据内容为开盘价、最高价、最底价、收盘价、成交量、成交额六项，由于本实验只需收盘价即可实现预测。故在表中只有时间和收盘价两列属性值。

  ③ 以亚泰集团(600881)为例，我们要预测 2007 年 11 月的亚泰集团走势，数据来源为长财证券大智慧。时间序列我们选取从 2007 年 5 月 8 日～2007 年 9 月 31 日的收盘价，周期为半年，共计 104 个数据。选取这段数据的理由在于：

  (1)这段期间亚泰集团没有发生派股行为。

  (2)这段期间亚泰集团走势即有平稳又有波动。

- ✧ 创建方案：登录 TipDM 平台后的默认页面即为“方案管理”，在此页面，选择“时序模式”创建一个新方案。方案名称：基于数时序模式的股票预测研究。方案描述：以大智慧从 2007 年 5 月 8 日～2007 年 9 月 31 日的收盘价数据为例对该股票的收盘价进行预测。信息输入完成后点“确定”保存方案。
- ✧ 上传数据：进入“数据管理”标签页，选择 stack-data.xls，并上传，上传的数据将自动显示在列表框中，若不能正确显示，点“刷新”按钮。
- ✧ 预测建模

  从“系统菜单”中选择“时序模式→指数平滑（其他算法)”。

  a）导入数据:，选择时间、收盘价等属性，以及样本数据范围（从第 1 行至 110 行），导入样本数据到当前算法操作区。

  b）参数设置:“apha”参数。

  c）时序预测：执行得到预测结果。

www.tipdm.com

4. 思考与实验总结

- ✧ 针对本例，请对指数平滑、多元回归、以及其他相关的时序模式算法的预测效果进行评价。

www.tipdm.com

附录三 参考文献

[1] 李建新,胡刚.风险厌恶型的证券投资数学模型. 数量经济技术经济研究 2005 年 第 3 期

[2] Hesieh D.Testing for Bonlinear Dependence in Daily Foreign Exchange Rates[J].Journal Business,1989,62(3):339-359.

[3] Eldridge R M,Coleman M P.The British FTSE-100 index:Chaotically Deterministic or Random?[J]. Working Paper,School of Business,Fairfied Univ.,1993.

[4] Chen Feng,Guangrong Ji,Wencang Zhao,Rui Nian.The Prediction of the Financial Time Series Based on Correlation Dimension[J]. Lecture Notes in Computer Science,2005, Vol.3610:1256-1265.

[5] Hirotaka Inoue,Yoshinobu Fukunaga,Hiroyuki Narihisa.E□cient Hybrid Neural Network for Chaotic Time Series Prediction[J].Lecture Notes in Computer Science,2001, Vol.2130: 712-718.

[6] Carmen Fierascu.Counterpropagation with Delays with Applications in Time Series Prediction[J]. Lecture Notes in Computer Science, 2005,Vol.3697:533-539.

[7] F Takens.Determing Strange Acttractors in Turbulence[J].Lecture notes in Math,1981,898:361-381.

[8] 邓华丽,张良均.模糊神经网络在企业信用评级中的应用. 统计与决策 2007 年 第 07 期

[9] 张良均,曹晶,蒋世忠.神经网络实用教程[M].北京:机械工业出版社,2008.2

[10] 宋晖,张良均. C4.5 决策树法在空气质量评价中的应用[J] .科学技术与工程,2011.8

[11] 宋晖,薛云,张良均. 基于 SVM 分类问题的核函数选择仿真研究.计算机与现代化,2011.9

[12] 张良均,王靖涛,李国成等.小波变换在桩基完整性检测中的应用. 岩石力学与工程学报,2002 年

[13] 邓华丽,李修全.基于混沌时间序列分析的股票价格拐点预测方法. 统计与决策 2007 年 第 09 期

[14] 鹿竞文,徐力.三阴乳腺癌证素变化规律及截断疗法研究. 硕士论文 2011 年 4 月

[15] 李娜,张文阳.不同种类生物质废物混合厌氧消化产气预测模型的研究. 硕士论文 2010 年 4 月

[16] 王海英,曹晶.基于 L-M 神经网络优化算法的池塘水色判别系统的初步建立.渔业现代化 2010 年

[17] 王广军,张良均. 基于动态扩展的水产养殖投入产出多目标优化仿真平台,农业攻关项目 2012 年

[18] 张萧木,赵合计. 基于主机入侵检测系统的设计与实现. 硕士论文 2007 年 4 月

[19] 梁立昌,胡晓辉. 基于数据挖掘技术的电信客户识别. 毕业论文 2011 年 5 月

[20] 刘涛,张良均. 大规模智能用电系统海量数据处理与数据挖掘技术研究及应用. 博士后重大专项

www.tipdm.com