



由资深数据挖掘技术专家撰写，深入学习数据挖掘技术并进行工程实践的必读之作
从基本概念到数据挖掘应用系统开发，包含数据挖掘实践的全过程与经验总结



技术丛书



Data Mining Technology and Practice

数据挖掘技术与 工程实践

(加) 洪松林 (Hong Song Lin) 著
(中) 庄映辉 李堃



机械工业出版社
China Machine Press

大数据技术丛书

数据挖掘技术与工程实践

（加）洪松林等 著

ISBN: 978-7-111-48076-1

本书纸版由机械工业出版社于2014年出版，电子版由华章分社（北京华章图文信息有限公司，北京奥维博世图书发行有限公司）全球范围内制作与发行。

版权所有，侵权必究

客服热线：+ 86-10-68995265

客服信箱：service@bbbvip.com

官方网址：www.hzmedia.com.cn

新浪微博 @研发书局

腾讯微博 @yanfabook

目录

前言

第1章 数据挖掘应用绪论

1.1 认识数据挖掘

1.1.1 数据挖掘概念

1.1.2 数据挖掘与生活

1.1.3 数据挖掘与知识

1.2 数据挖掘应用基础

1.2.1 事物与维度

1.2.2 分布与关系

1.2.3 描绘与预测

1.2.4 现象和知识

1.2.5 规律与因果

1.3 数据挖掘应用系统工程

1.3.1 数据层

1.3.2 算法层

1.3.3 应用层

1.4 数据挖掘应用体会

1.4.1 项目关键点

1.4.2 技术与应用创新

1.4.3 经验积累与应用

1.5 无限三维嵌套空间假说

1.5.1 一维空间

1.5.2 二维空间

1.5.3 三维空间

1.5.4 突破三维空间

1.5.5 五维空间

1.5.6 六维空间

1.6 本章小结

第2章 数据探索与准备

2.1 数据关系探索

2.1.1 业务发现

2.1.2 关系发现

2.1.3 数据质量探索

2.1.4 数据整合

2.2 数据特征探索

2.2.1 数据的统计学特征

2.2.2 统计学特征应用

2.3 数据选择

2.3.1 适当的数据规模

2.3.2 数据的代表性

2.3.3 数据的选取

2.4 数据处理

2.4.1 数据标准化

2.4.2 数据离散化

2.5 统计学算法的数量条件

2.5.1 样本量估计概念

2.5.2 单样本总体均值比较的样本量估计 (T-Test)

2.5.3 两样本总体均值比较的样本量估计 (T-Test)

2.5.4 多样本总体均值比较的样本量估计 (F-Test)

2.5.5 区组设计多样本总体均值比较的样本量估计 (F-Test)

2.5.6 直线回归与相关的样本量估计

2.5.7 对照分析的样本量估计

2.6 数据探索应用

2.6.1 检验项的疾病分布

2.6.2 疾病中检验项的分布

2.6.3 成对检验项的相关分析

2.6.4 两种药物的应用分析

2.7 本章小结

第3章 数据挖掘应用算法

3.1 聚类分析

3.1.1 划分聚类算法 (K均值)

3.1.2 层次聚类算法 (组平均)

3.1.3 密度聚类算法

3.2 特性选择

3.2.1 特性选择概念

3.2.2 线性相关算法

3.2.3 相关因子SRCF算法

3.3 特征抽取

3.3.1 主成分分析算法

3.3.2 因子分析算法

3.3.3 非负矩阵因子分解NMF算法

3.4 关联规则

3.4.1 关联规则概念

3.4.2 Apriori算法

3.4.3 FP树频集算法

3.4.4 提升Lift

3.5 分类和预测

3.5.1 支持向量机

3.5.2 Logistic回归算法

3.5.3 朴素贝叶斯分类算法

3.5.4 决策树

3.5.5 人工神经网络

3.5.6 分类与聚类的关系

3.6 时间序列

3.6.1 灰色系统预测模型

3.6.2 ARIMA模型预测

3.7 本章小结

第4章 数据挖掘应用案例

4.1 特性选择的应用

4.1.1 数据整合

4.1.2 数据描绘

4.1.3 数据标准化

4.1.4 特性选择探索

4.2 分类模型的应用——算法比较

4.2.1 数据整合

4.2.2 数据描绘

4.2.3 数据标准化

4.2.4 特性选择探索

4.2.5 分类模型

4.3 分类模型的应用——网络异常侦测

4.3.1 计算机网络异常行为

4.3.2 网络异常数据模型

4.3.3 分类模型算法应用

4.4 算法的综合应用——肿瘤标志物的研究

4.4.1 样本选取

4.4.2 癌胚抗原临床特征主题分析

4.4.3 癌胚抗原临床特征规则分析

4.4.4 癌胚抗原临床特征规则的比较分析

4.4.5 癌胚抗原相关因子分析

4.4.6 不同等级癌胚抗原组差异分析

4.5 数据挖掘在其他领域中的应用

4.6 本章小结

第5章 数据挖掘行业应用原理

- 5.1 传统医学科研方法的现状
 - 5.1.1 传统医学科研的命题与假说
 - 5.1.2 传统医学科研的数据应用
 - 5.1.3 传统的医学科研的统计学应用
 - 5.1.4 传统医学科研的流程
- 5.2 智能医学科研系统的需求
 - 5.2.1 临床医学科研的问题
 - 5.2.2 临床医学科研的解决思路
- 5.3 智能医学科研系统的设计思想
 - 5.3.1 科研立题
 - 5.3.2 科研设计与统计分析
 - 5.3.3 样本数据收集与分析
- 5.4 智能医学科研系统的核心技术方法
- 5.5 智能医学科研系统的科研数据仓库建设
 - 5.5.1 医学科研数据仓库建设的技術方法
 - 5.5.2 医学科研数据仓库的建设过程
 - 5.5.3 科研数据仓库的数据安全
- 5.6 智能医学科研系统的核心功能设计
- 5.7 智能医学科研系统的整体功能设计
 - 5.7.1 智能医学科研系统主要功能
 - 5.7.2 智能医学科研系统的模块设计 and 应用实现
 - 5.7.3 智能医学科研系统的评估方法
- 5.8 智能医学科研系统的应用价值

5.9 本章小结

第6章 数据挖掘应用系统的开发

6.1 数据挖掘应用系统的意义

6.2 IMRS系统设计

6.2.1 对数据源的分析

6.2.2 数据挖掘应用系统IMRS的总体设计

6.3 IMRS异常侦测模型的开发

6.3.1 异常侦测模型的功能展示

6.3.2 数据挖掘技术开发要点

6.4 IMRS特征抽取模型的开发

6.4.1 特征抽取模型的功能展示

6.4.2 数据挖掘技术开发要点

6.5 IMRS智能统计模型的开发

6.5.1 回归模型的开发实现

6.5.2 线性相关模型的开发实现

6.6 IMRS的算法开发

6.6.1 相关因子算法SRCF的实现

6.6.2 朴素贝叶斯分类算法的实现

6.7 本章小结

第7章 数据挖掘应用

7.1 分布探索

7.1.1 两维度聚类模型应用

7.1.2 高维度聚类模型应用

7.2 关系探索

7.2.1 关联规则的应用

7.2.2 特性选择的应用

7.3 特征探索

7.3.1 不稳定心绞痛的特征总结

7.3.2 动脉硬化心脏病的临床特征

7.4 异常探索

7.4.1 生理指标的异常侦测

7.4.2 异常侦测模型的比较

7.5 推测探索

7.6 应用系统的高级应用

7.6.1 异常侦测的高级用法

7.6.2 关联规则的高级应用

7.7 本章小结

第8章 数据挖掘工具的应用

8.1 应用Oracle Data Mining

8.1.1 ODM数据挖掘流程

8.1.2 ODM算法模型

8.1.3 ODM算法应用

8.2 应用IBM SPSS Modeler

8.2.1 IBM SPSS Modeler介绍

8.2.2 SPSS Modeler独立应用

8.2.3 SPSS Modeler与应用系统的联合应用

8.3 本章小结

参考文献

献给我亲爱的Eddie!

前言

我们仍在山脚下

很多人将数据挖掘归类为IT技术的一个分支。而在IT业，历史上有一种夸大其词的习惯，经常将一些高新技术的出现说成是“革命性的”、“里程碑式的”，近些年又常常使用“颠覆性的”等词汇。但回顾一下，我们一次次地发现我们的生产方式和生活方式并没有被革命、被颠覆，IT技术的发展与提高基本上是渐进性的，只是我们一次又一次地被忽悠了。那些所谓的高新技术几乎无一例外地也有很多的缺陷和毛病。许多单位在付出不少的“学费”之后，那些应用了高新技术的IT系统才慢慢成熟起来。

从现代科技发展的历史长河来看，那些高新技术也只能算是历史发展中必然的、一个需要不断改进的新技术而已。数据挖掘技术也是这样。不同的是，由于现有的科学理论和方法还比较初级，以及人类的大脑还不足够发达，以至于像数据挖掘这样的工作，对人类来讲还是具有很大挑战性的。同时也说明，我们现有的数据挖掘技术和方法还很有限，很多问题我们还解决不了或解决不好。因此可以说，面对数据挖掘这座高峰，我们还只是在山脚下攀爬！

实践与应用

有很多人在没有经历过数据挖掘成熟应用项目的情况下，投入到这方面的技术探索与实践工作中来，他们的客户也没有过数据挖掘应用项目的经验，于是他们双方对于数据挖掘应用的认识就基于他们共同完成的项目的效果上了。于是有的人认为数据挖掘是虚的、是炒作，对于数据挖掘的许多方面都在打问号。实际上，数据挖掘与很多技术（如IT技术）一样是一项很实用的技术，它必然确而无疑地要为各行各业的工作带来高效益和高效率。否则，要么数据挖掘是没用的技术，要么数据挖掘应用的项目是失败的。因此，能带来高效益和高效率的数据挖掘实践是数据挖掘技术和相关商业应用项目的检验标准。比方说，一个数据挖掘应用项目的结果给出了一个与实际相反的指导结论，那这个项目无疑是失败的；一个数据挖掘应用项目的结果没有达到业务需要的精度，或精度达到了但模型不稳定，业务尚不能稳定、可靠地依赖于它的指导结果，那项目也是不成功的；一个数据挖掘应用项目的结果精度和可靠性都能满足业务需求了，但投入数据挖掘应用项目的成本高于业务上的收益，或者即使有净收益，但并不高于不用数据挖掘技术带来的收益，这也不算成功的数据挖掘商业应用。总之，数据挖掘不是用于体现技术高尚的一个工具，而是一个实实在在的能给业务带来可观收益的实用工具！

大数据与数据挖掘

大数据是时下的时髦用词。与过去相比，数据是变大了，而且是在不断地变大，但并没有“爆炸”。现在的大数据与过去讲的数据相

比，在内涵和外延上看都有了拓展，但并没有发生本质的变化。我们早几年将数据说成是大数据也可以，按照现在的发展态势，晚几年将数据说成大数据也未尝不可。实际上，大数据与我们广义说的数据并没有形成明显的界限划分。对于长期从事数据分析的人来说，不论你是否提出大数据的概念，我们都要脚踏实地地不断解决新问题、满足新需求，因此基于大数据的数据挖掘实在是一项平常的工作。但即使你不提出大数据的概念，我们也要面对数据增大带来的新问题，要研发出数据挖掘的新技术来开发大数据，这是我们从业人员的职责，也是极富挑战性的一项工作。

本书写作方法

在数据挖掘的技术中有一个重要的方法，就是被称为最小描述长度（MDL）原理的技术，说的是对某一事物有很多不同的描述方法，但最简单的描述是最好的方法，也称为最佳描述模型。本书的写作也试图应用这一原理，即将复杂问题简单化，对于很多复杂的数据挖掘算法和应用，通过我们自己的深入理解，用最简单、最容易理解的方式将其核心内容展示出来。例如，经典的神经网络模型，很多书已经介绍得很详细了，但不少人对我说还是很难理解，尤其是有些语言、用词本身就不易理解，更不用说其复杂的内涵了。我们用基于我们的理解凝练出的方法，给一些人讲授神经网络原理，这些人都是第一次接触数据挖掘的，讲完之后他们说完全听得明白，而且还提出一些很专业的问题。这使我们很兴奋，原来数据挖掘本应是如此简单的！

本书导读

本书内容的跨度是比较大的，涵盖的内容比较广泛，既有对数据挖掘概念的探讨，也有对数据挖掘技术和原理的介绍，还有对数据挖掘应用实践的体会和总结。其中包括数据挖掘定制化项目案例，也涵盖了数据挖掘应用系统的开发及详细技术介绍，还有通过数据挖掘通用工具开展的应用案例展示。从技术上，涉及数据挖掘、数理统计、数据库技术，以及更广泛的各种IT技术。具体体现在如下几个方面。

首先，通过我们多年在数据挖掘商用项目中的实践经历和体会，提出了我们对数据挖掘概念的理解，即数据挖掘是一个更广义的有目的地探索数据中隐含的规律和知识的活动。

其次，大量的实践工作也培养了我们形成了一个数据挖掘的思维模式：即通过现象看本质的思维，突破传统、不断创新的思维，几乎穷尽事物的所有维度来认知事物的高维度思维，以及一个普遍联系、不断扩大认知的思维。

另外，在本书的算法介绍中，基于我们实践应用方法和自身的理解讲述了一些经典数据挖掘算法原理，如K-Means、SVM、MDL、神经网络等，也有我们自己研创的用于商用项目的算法，如SRCF算法就是首次在本书中完整公开发表。

在数据挖掘的应用章节，我们例举了由多种数据挖掘算法（包括聚类分析、特性选择、特征抽取，关联规则等算法）联合应用的典型

案例。在深入一个行业的应用中，我们较深入地探讨了数据挖掘在该行业中应用的原理和方法论以及具体实现，使数据挖掘在行业中的应用上升到一个新的高度。这样才有可能最大限度地发挥数据挖掘技术在行业应用中的作用，为业务工作带来显著效益。同时，通过一个行业的应用来说明数据挖掘应用的普遍原理。

不仅如此，我们还在本书中详细剖析了数据挖掘在商用应用系统的技术实现，并首次展示了数据挖掘应用软件产品的实现，并较深入地讨论了数据挖掘在行业中的应用意义。书中重点剖析了数据挖掘在临床医学、健康管理、信息安全、证券预测等领域的应用，还简要分析了数据挖掘在金融、电信等行业的应用。

最后，再说明一下所谓“基于我们的理解”的含义。举个简单的例子，数据挖掘中有两种技术，一种叫Feature Extraction，另一种叫Feature Selection。文献中通常把前者称为“特征抽取”，后者称为“特征选择”，英文“Feature”在中文中用了同一个词“特征”。但是许多人对两者的理解容易混淆，不易弄清它们的本质区别。在本书中，我们把Feature Extraction还称为“特征抽取”，而把Feature Selection称为“特性选择”。一字之差，却有本质区别。“特征抽取”是指将大量的原有变量进行整合与重组，并生成了较少的、更具有特征代表性的新变量。“特性选择”是指从大量的原有变量中选择出对于目标变量相关性更大的几个，选择后的变量本身没有变化，这时我们将这些变量称为“特性”。“特征”与“特性”已不是同一概念，这样，从字面上就将二者的本质区分开了，新手也很容易理解。因此，本书中有一些词汇和

用语是基于我们对数据挖掘的理解和认识提出的，可能不大大同于已有的一些文献。这些词汇和用语在我们公司内部和外部的培训中证明是有效的。

本书的读者对象

从前面的介绍中，大家对本书的内容已经有了一定的了解。我们认为本书非常适合如下几类人士阅读：

- 数据挖掘的初学者。由于我们将数据挖掘从生活到工作、从理论到实践，采用复杂问题简单化的方法，对数据挖掘的概念、主要技术和典型应用加以介绍，对于初学者来说更易于理解和快速入门。可帮助读者加快、加深对数据挖掘算法的理解。

- 对数据挖掘的理论知识基本掌握但希望在实践中不断提高的技术人员。本书可帮助读者在商用项目中尽快走上良性发展之路。

- 在商业项目中需要更多启示和更多解决方案的人士，可从数据挖掘定制化项目案例和数据挖掘应用系统的开发案例中得到启发和提示。

- 希望在医学数据挖掘或健康大数据分析方面借鉴经验的人士。本书从始至终贯穿了医疗、健康数据方面的数据挖掘和探索，从原理到方法、从设计到实现、从技术到讨论等各方面阐述了数据挖掘的应用，提炼了大量实际经验。

· 从事大数据分析的技术人员、高校师生、科研人员，以及公司的管理者和决策者，均可从本书中有所收益。

致谢

由于本书写作的内容主要涉及一线的商业化的数据挖掘应用，是作者对自己近20年间参与的国内外很多数据挖掘商业项目的成果思考，因此作者在此对与自己共同战斗过的同事、同行表示感谢。

本书的正式写作大约始于一年以前，集中式的写作阶段大约有半年时间左右，其他半年时间主要为间歇性的修改和完善。除本人是主创外，参与本书编写的还有我的两个助手，一位是庄映辉，另一位是李堃，他们均是我（福安易数据技术）公司数据挖掘应用项目的主要设计者和参与者，有着比较丰富的实践经验。其中庄映辉主要参与了第1章、第2章、第7章的部分写作，李堃主要参与了第2章、第3章的部分写作。另外，除此之外，我公司的张双、肖芄、白丽娜、李飞翔也参与了本书编写中大量的整理、编辑、校对工作，在此，对以上人员一并表示感谢！机械工业出版社的吴怡编辑不仅促成了本书，还在本书的策划、编辑、审校等方面做了大量的工作，在此向吴编辑表示衷心的感谢！由于写书是件比较耗费时间的事，本书的写作不仅花费了我不少的工作时间，而且还占用了我很多业余时间，因此特别感谢我的家人，是他们的支持，才使我得以在较短的时间内将书写成，尤其是我那还在上小学的儿子Eddie，写书不仅占去了不少本属于我陪

他的时间，而且他还时常提醒我“少看球，快写书”。因此，我想将本书献给我亲爱的Eddie！

个人的能力永远是有限的，我和我的团队的认识也是有限的。书中的错误和不当之处在所难免，敬请广大读者指正，不胜感谢！联系邮箱为：hong.forest@hotmail.com。

洪松林（Hong Song Lin）

2014年7月15日

本书涉及的数据挖掘算法应用索引

序号	算法名称	算法类型	原理章节	应用章节	
1	K-Means (划分聚类算法)	聚类分析	3.1.1.1	4.4.1.2	7.1.1/7.1.2/8.1.3.3/ 8.2.3.2
2	凝聚的方法 (层次聚类算法)	聚类分析	3.1.2.1		
3	分裂的方法 (层次聚类算法)	聚类分析	3.1.2.2		
4	BIRCH 算法 (综合层次聚类方法)	聚类分析	3.1.2.3		
5	CURE 算法 (综合层次聚类方法)	聚类分析	3.1.2.3		
6	DBSCAN 算法 (密度聚类算法)	聚类分析	3.1.3.2		
7	MDL 算法 (最小描述长度)	特性选择	3.2	4.1.4/4.2.4.2/4.4.5.2	5.7.2.4/7.2.2/8.1.3.1
8	Pearson 相关 (线性相关算法)	特性选择	3.2.2.1		
9	Spearman 秩相关 (线性相关算法)	特性选择	3.2.2.1		
10	SRCF1 算法 (相关因子 SRCF 算法)	特性选择	3.2.3.2	4.1.4/4.2.4.1/6.6.1	
11	SRCF2 算法 (相关因子 SRCF 算法)	特性选择	3.2.3.3		
12	主成分分析算法	特征抽取	3.3.1		5.7.2.3/7.3.1/6.4.2.2/ 6.4.2.3
13	因子分析算法	特征抽取	3.3.2		
14	非负矩阵因子分解 NMF 算法	特征抽取	3.3.3	4.4.2.3	
15	Apriori 算法	关联规则	3.4.2		1.3.2.3/4.4.3/5.7.2.2/ 7.2.1/7.6.2/8.2.2.2
16	FP-树 频繁集算法	关联规则	3.4.3		
17	支持向量机	分类和预测	3.5.1	4.2.5/4.3.3.1	8.1.3.2/8.2.3.1
18	logistic 回归算法	分类和预测	3.5.2		
19	ID3 算法 (决策树)	分类和预测	3.5.4.3		
20	CHAID	分类与预测		8.2.3.2	
21	朴素贝叶斯分类算法	分类和预测	3.5.3	4.2.5/4.3.3.1/6.6.2	
22	BP 神经网络原理 (人工神经网络)	分类和预测	3.5.5.2		
23	回归预测	分类与预测		5.7.2.7/7.5.1/ 8.1.3.4/6.5.1	
24	灰色系统预测模型	时间序列	3.6.1		
25	ARIMA 模型预测	时间序列	3.6.2		

第1章 数据挖掘应用绪论

我们从本章开始一直到第4章结束，按照先后顺序，分别讲解数据挖掘应用的一些基础内容（第1章）、数据探索与准备阶段的工作和方法（第2章）、数据挖掘应用算法阶段的有关技术（第3章），以及基于前三章内容基础上的数据挖掘应用的各种案例（第4章）。这四章构成了数据挖掘应用的一个较完整的讲解，可以算作数据挖掘应用的基础篇。本章又是这个基础篇的基础，希望能给读者一个数据挖掘应用的快速引领和轮廓印象。我们从第5章开始直到第7章结束，讲述了数据挖掘在行业的实践应用，即数据挖掘在医学（科研）领域的一个完整应用，首先较系统地阐述了医学数据挖掘应用的原理（第5章），之后较详细地介绍了医学数据挖掘应用的开发实现（第6章），最后较全面地探讨了医学数据挖掘的实践应用（第7章）。这三章可以看作是数据挖掘应用的高级篇。最后一章（第8章），我们介绍了利用数据挖掘通用工具软件开展数据挖掘商业应用的案例。

我们在本书中尽可能少涉及各种公式和数学推导（除非必要时，如第3章），而深入原理部分。理论是灰色的，而实践则是最鲜活的。对于一本数据挖掘应用方面的书籍，我们希望尽可能多讲些实践和案例，并多用图画、图表说明大部分的数据挖掘原理和应用，让读者更能贴近实际。

1.1 认识数据挖掘

什么是数据挖掘？不同的人会给出不同的答案，很多人也会给出相似的答案，因为有很多经典的数据挖掘论著已经给出较为公认的定义。作为常年工作于数据挖掘应用项目的人来说，我们也有自己的一点认识，可能与理论书籍的概念稍有不同。本节我们就从这个话题开始，介绍数据挖掘概念、数据挖掘与生活和数据挖掘与知识等内容。

1.1.1 数据挖掘概念

我们认为，数据挖掘应是一个更加广义的概念，甚至可以说不是一个传统意义上的定义，而是一类活动的集合，凡是有目的的探索数据中隐含的规律和知识的活动都可称作数据挖掘。在这里，我们重点强调的要素是：

- 有目的
- 探索性地获取
- 数据中隐含的规律和知识

我们稍后会详细讨论这三个要素。在这个定义中，我们并没有提及应用什么方法和手段获取数据中隐含的规律和知识，这就意味着不限任何方法和手段，无论是数学的还是非数学的，无论是复杂的还是简单的，只要能揭示数据中隐含的规律和知识，都可以被称为数据挖掘。

1.1.1.1 数据的“形状”

从字面而言，数据挖掘包含数据和挖掘两部分，二者同样重要，缺一不可。数据是数据挖掘的基础素材，我们首先谈一谈数据的形状。

数据的“形状”之一，大数据。经典意义上的数据挖掘，通常是指

对海量数据进行分析。怎么样才算是海量数据？目前还没有明确的标准。而近几年，类似于海量数据，又产生了大数据的提法，其概念无论从内涵和外延上都有了扩展。但从本质上，我们认为，大数据和海量数据是相似的。在实践中，不单单是记录数多的就称为大数据，通常大数据是指数据量和数据维度均很大，数据形式很广泛，如数字、文本、图像、声音等。而大数据往往可能蕴含着丰富的规律和知识，所以在大数据之上应用数据挖掘就成了理所当然的活动了。

数据的“形状”之二，小数据。相对于大数据，在实践中还会存在不少特殊情况。例如在医学上有些疾病极为少见，只出现几百例，甚至几十例就几乎是该病的总体了，我们称之为小数据。业务中需要对这些小数据进行深入分析和探索，以便挖掘出罕见疾病的特征，并为相应的临床应对提出依据。对于这样规模的数据进行分析，如果按照记录数，依照传统数据挖掘观念、方法和技术，无法开展探索性的分析工作。我们认为，需求引领观念和技术，数据挖掘的一个发展分支应该是从规模较小的、有限的数据中探索其中的规律和知识，尽管目前的技术还很有限。

数据的“形状”之三，宽数据。还有一种情况是小数据高维度，小样本大信息，我们称之为宽数据。如某些基因组信息，数据量很少，通常只有几十例到几百例，但维度很高，通常有几百个到几千个。更极端情况的是个人大信息，即单个记录下的高维信息，如从宽带、移动支付、物联网、手机等媒介收集个人信息。在不远的将来会出现单独个体的高维数据，并需要解决此类数据挖掘的新理论和新算法。

数据的“形状”之四，深数据。我们还会遇到一种数据，涉及维度不是很宽，但是数据在某几个维度上跨度非常大，历史数据非常多，或者数据量的增长速度非常快，我们称之为深数据。如医学检查中24小时心电图监测、较长时段（如一小时以上）的脑电图监测，每小时会产生几十万至几百万条数据；再如，互联网服务商的DNS服务器对互联网访问事件的日志记录，也是每小时会产生几十万至几百万条数据。这类数据，我们有时也称为流数据。对这些深数据的挖掘也是非常具有挑战性的，一方面由于它的数据量非常大，另一方面也由于对这类数据进行挖掘的实时性要求较高。

这些随着数据收集手段的进步而形成的各有特色的数据，正在逐步进入数据挖掘研究的视野。所以说，这门科学叫做数据挖掘，它应包括大数据挖掘、小数据挖掘、宽数据挖掘和深数据挖掘。我们需要做的是处理好各类数据来获取知识，研究解决各类型数据的挖掘的新理论和新算法，这些数据的分析算法不完全与经典大数据挖掘相同。例如医学上的个性化精确治疗，就离不开涉及个人的宽数据和深数据。

1.1.1.2 挖掘的思维

数据挖掘的目的是为了获得知识，很多书中也将数据挖掘称作KDD，数据库中的知识发现，也是数据挖掘的一个别名。只要是为了获得知识，那么用什么工具并不重要，重要的是从数据里面获取知识。至于用了什么手段获得，那只是从愿望到目的的桥梁，重要的是

结果。此外，我们说在数据挖掘应用中，不是处理方法越复杂就越好，其实即使是非常简单的方法也可以睿智地理解数据。例如，世界大战中，统计学家沃德在被咨询飞机上什么部位的钢板需要加强时，他画了飞机的轮廓，标出返航战斗机上受敌军创伤的弹孔位置。统计积累一段时间后，机身各部位几乎都被标满了。最后，沃德建议，把剩下少数几个没有弹孔的位置加强，因为被击中这些位置的飞机都没有返航。最后实践验证了沃德对飞机改进的良好效果。

我们认为，不要被数据挖掘的传统概念限制思维。很多从业者或希望从事数据挖掘的人，把数据挖掘这个概念狭义化了。数据挖掘不是有限的几种工具或算法，例如聚类、分类和预测等，它是一个目的性导向的学科，目的是从数据中获取知识、规则，或其他可直接、间接用以产生效益的信息。广义上的数据挖掘是和概率统计、高等数学、数学分析、离散数学等数学分支无法清楚分割的，也是和数据库、网络、大数据等技术无法分割的，更是和各行各业的专业知识和业务需求无法分割的。

1.1.1.3 数据挖掘要素

下面说说我们提到的数据挖掘概念的要素。先说数据中隐含的规律或知识。我们知道数据是用以描述和反映人类社会中所发生的各种人文活动和事件及自然活动和事件的载体，而大量的人文事件和自然事件中通常蕴含着某些特点和规律。因此，我们利用各种形式的数据（包括数字形式或数据库形式，也包括书籍，图案，声音等形式）将

这些活动和事件如实地描述和记录下来，然后应用各种技术手段来研究和挖掘这些数据中所隐含的东西，这些隐含的东西反映了人文或自然活动和事件的本质特征，这些本质特征通常又不是体现在人文或自然活动和事件的表面或较肤浅的层面。我们得到的这些本质特征可能表现为与某些事物相关的一些规律或知识，从而延伸了事物表面所展现的规律或知识，但又绝不类似于事物表面所展现的规律或知识。隐含知识与表象知识通常是完全两码事，隐含知识比表象知识具有更大的价值。所以说，数据挖掘要的不是事物的表面现象，而是事物所隐藏的东西。反过来说，展现事物表象的知识不属于数据挖掘。

再说数据挖掘是一种探索性的活动。我们认为，由数据所表达的大量事物中通常可能蕴含了一些规律或知识，但谁也不敢保证一定有。另外，挖掘大量数据中所隐含的知识本身，无论从技术上还是从专业上都是一项极富挑战性的工作。因此，我们说数据挖掘是一种探索性质的活动。探索性质的活动意味着过程可能会很艰辛，结果可能不可预料。所以，如果数据挖掘的结果达不到我们的预期，一种可能是我们的技术、方法不行，一种可能是数据没有能够真实描绘、反映事物，还有一种可能是事物中没有蕴含我们想要的东西。但是，由于隐含知识通常比表象知识具有更大的价值，而需求引导我们不断地去追求，因此，我们会不停地探索。

最后，再说数据挖掘是有目的的活动。数据挖掘的方向是由业务需求所引领的，知识发现是一项目的性很强的工作。不同的数据挖掘目的所涉及的技术、方法，甚至投入的人力、物力都大不相同，因

此，要选择恰当的目的，使得数据挖掘工作可控、成本可控。在这里，我们着重讲的是数据挖掘的商业应用，而不是诸如有关数据挖掘基础研究、教学实验方面的工作。既然是商业应用，更加需要讲效益、投入产出比。因此，商业数据挖掘的目的性非常强，数据挖掘通常分为评估性初探、计划、评估、实施、再评估、部署、维护等过程。如果数据挖掘目的不明确、缺乏效果评估和风险评估，则项目的失败实在是在所难免。另一方面，即使是数据挖掘基础研究和教学实验，确定一个科学的工作目标也是必须的。我们说数据挖掘是有目的的，大家很容易理解，但数据挖掘本身作为具有探索性的一项工作，其工作目标的制订，尤其在商业项目中决定着项目的成败，因此，我们将其作为一个重点要素在此提出。

1.1.2 数据挖掘与生活

很多人一提到数据挖掘，觉得它一定很高深。的确，数据挖掘不是一件很容易的事，尤其经过数据挖掘应用项目的人感受更深，往往不得不跟一些复杂的算法、繁琐的数据处理打交道。但实际上，数据挖掘并不都是深奥和难懂的，很多原理就是生活的一部分。现实生活中，很多情况下，人们是在不知不觉中使用了一些数据挖掘的方法或思想。例如聚类，就是将具有相似特征的事物划分为一类，把特征不同的事物区分开来。实际上，聚类这个动作贯穿了我们生活的很多方面。例如，我们对事物分类的基础和分类的规则，大部分就是通过聚类活动实现的。例如，原始人面临猛兽和猎物，必须会区分这两者，知道自己可以吃掉谁，谁可以吃掉自己（这个是最早的分类之一）。可以说，不懂得这个分类的原始人基本上都灭绝了。而聚类提供的是什么呢？原始人经过大量的实践（数据的积累），如图1-1所示，在吃和被吃的两个分类下（那时候还没有家畜和宠物），如图1-2所示，对大量事例进行思考、总结、聚类（类内相似最小，类间相异最大），得到规则并实施，然后幸福地进化着。

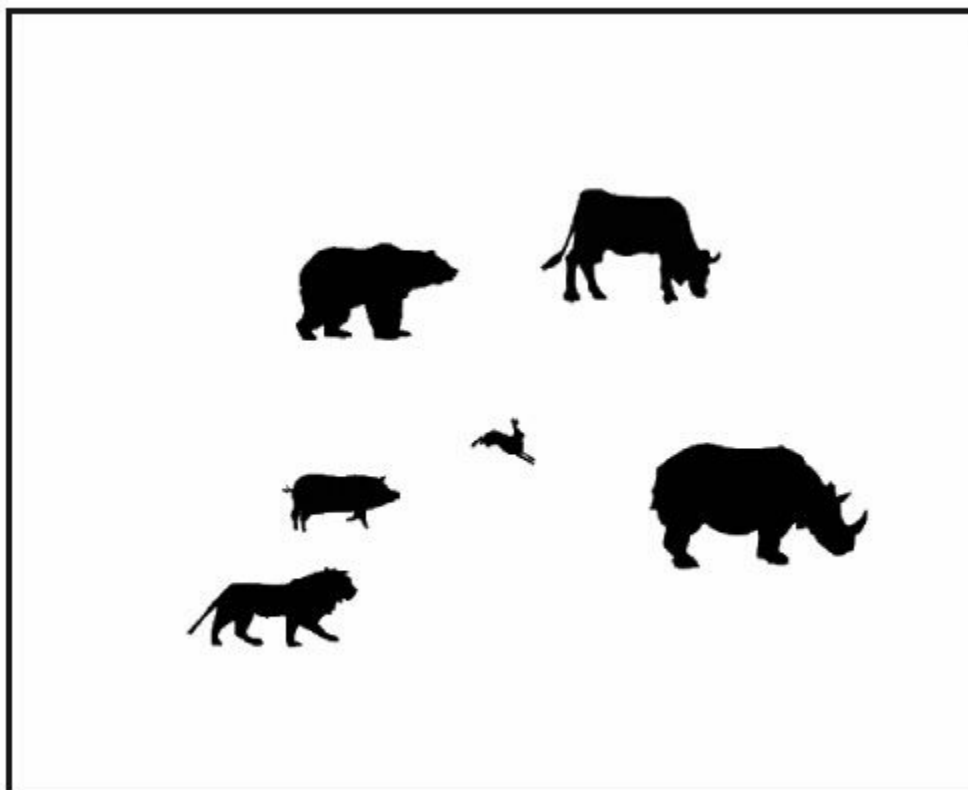


图 1-1

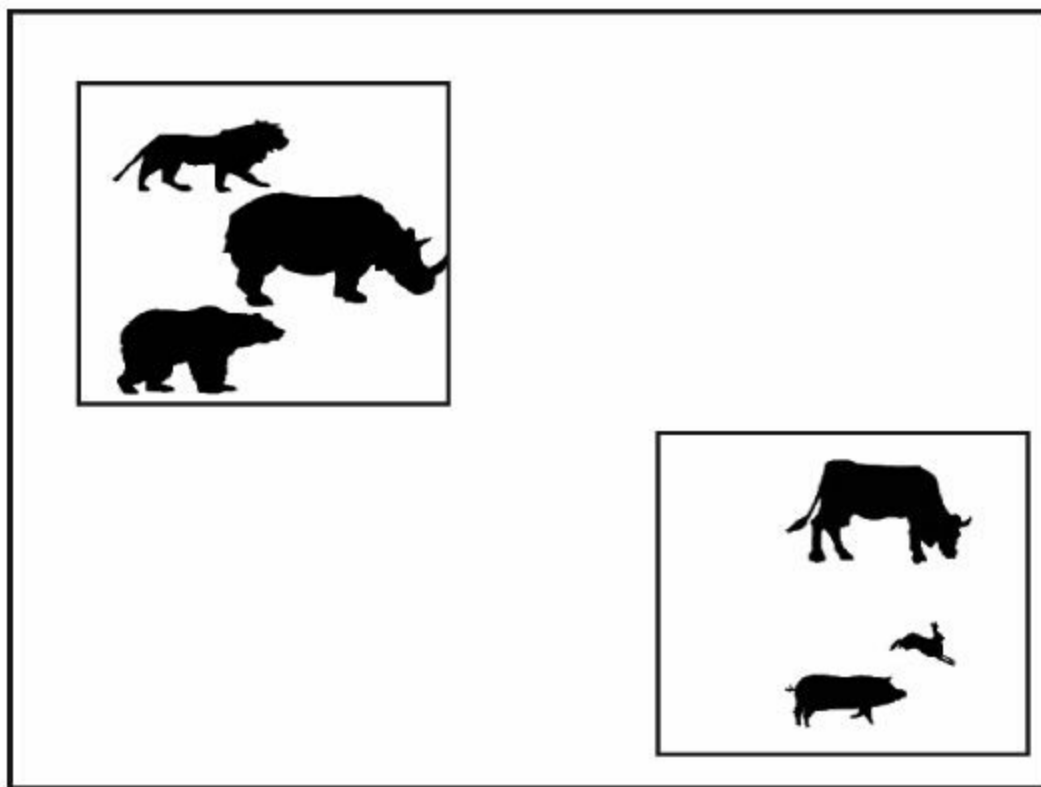


图 1-2

再例如，关联规则的应用。所谓关联规则，就是在前件事件发生的条件下寻找高概率发生的后件事件。例如，原始人看见了前件——猎物，马上会关联后件——烤肉（原始人学会了使用火以后），于是他们实施的行动就是捕猎与钻木取火，如图1-3所示。原始人看见了前件——猛兽，马上会关联后件——被吃，于是他们实施的行动就是狂奔与逃窜，如图1-4所示。

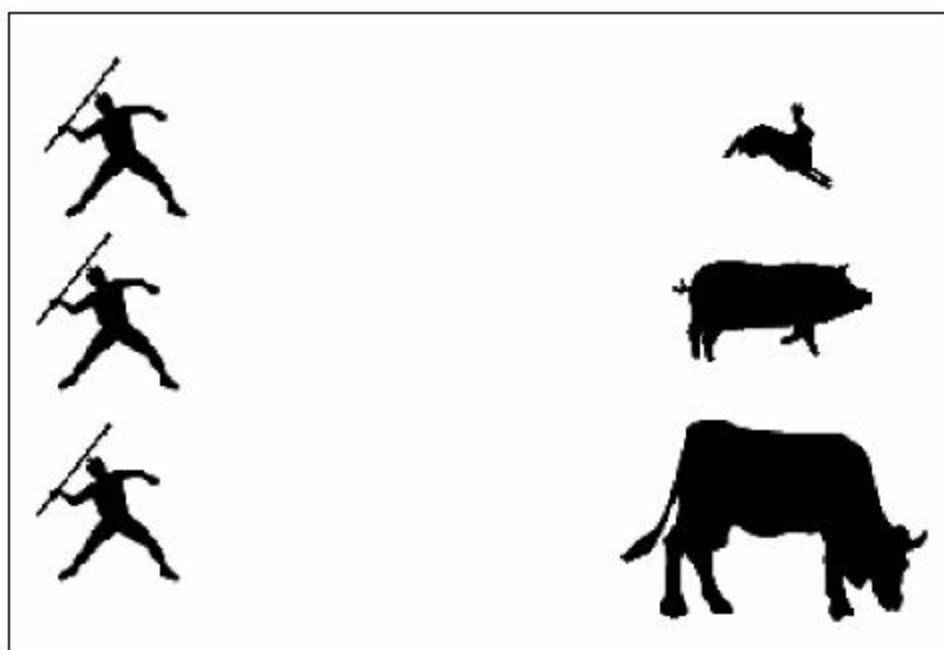


图 1-3

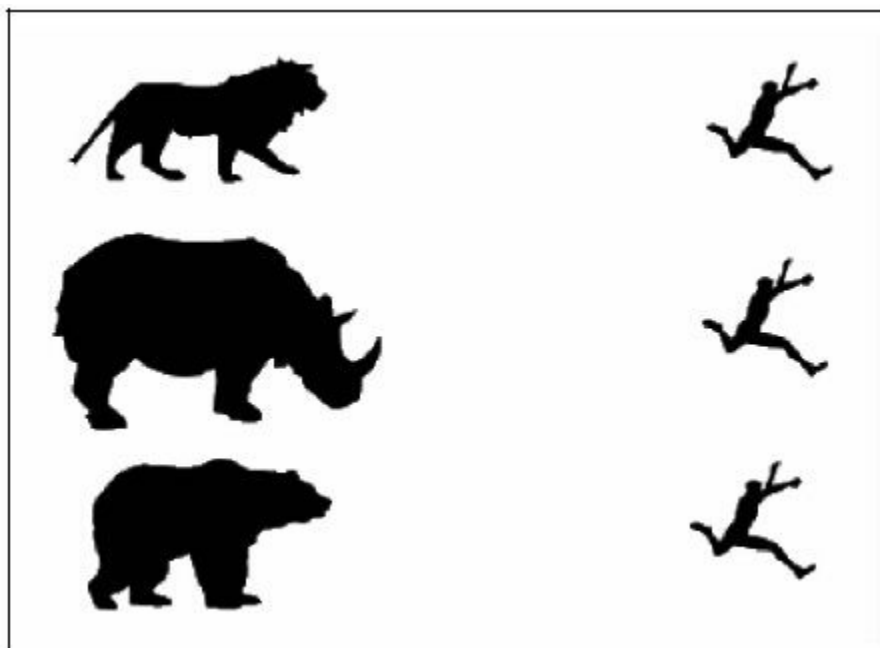


图 1-4

数据挖掘就是伴随着社会事件或自然活动的大量产生（数据的海量增长）孕育而生的。所以说，数据挖掘不是孤立存在的，而是和社会生活及其衍生的需求有机结合的。

数据挖掘来源于生活，还要服务于生活。但很多时候，困扰数据挖掘学习者和应用者的是数据挖掘复杂的理论、算法和公式。如果使用数据挖掘解决实践问题，我们希望就像医生不需要明白药品的合成制造原理和过程而能科学地使用药物、手机用户不必搞懂通信协议就能熟练地使用手机的各种功能一样，我们也不必学习数据挖掘的算法理论与开发技术，也能做好生活中的数据挖掘。要实现数据挖掘的大众化，一个方法就需要将数据挖掘应用面向各个专业、面向生活的各个方面，实现产品化和工具化，尽量使技术细节对使用者透明化，而使用户专注于数据挖掘技术在行业领域、生活领域的应用。

一般来说，数据挖掘的应用者需要知道数据挖掘各个算法的基本目的、基本原理。这里说的基本原理是高层面（high level）原理，而不是具体技术性的原理。例如聚类，我们只要知道它的原理就是“物以类聚，人以群分”就足够了。数据挖掘技术的一个方向就是让每个需要使用的人都可以方便地使用，从而享受新技术带来的好处。这样数据挖掘才可以真正融入生活的各个方面，成为一门成功的科学技术。数据挖掘的产品希望做到和手机一样，不需要用户懂得什么GSM、CDMA的原理之类（当然，名称还是要知道，不然购买手机后发现SIM卡不支持就Game Over了）就能通话，会打字就可以用手机发短信。

1.1.3 数据挖掘与知识

当然，数据挖掘现在还远没有走入人们的生活，大多数情况下是应用于一些商务场合，即数据挖掘是以商业项目形式存在。根据经验，我们把数据挖掘项目分成几个不同的层面，如决策层面、设计层面、技术层面和应用层面。数据挖掘项目不同的层面需要不同的知识结构。

决策层面需要知道数据挖掘能干什么，能为本单位带来什么效益；设计层面需要行业领域相关知识和数据挖掘技术相关的知识；技术层面需要高等数学、概率学、统计学、数据库原理、分布计算、编程语言等知识，还需要掌握具体算法的原理；应用层面需要知道数据挖掘怎样结合行业领域的需求，怎样应用数据挖掘的结果解决业务问题。

所以，数据挖掘的应用所需要的知识与人们在项目的职责与角色密切相关。使用数据挖掘需要从高等数学学起吗？还是懂得原理就可以了？决策者在行业目的的选择上需要懂得多深的数据挖掘知识？不同的目的，需要人们掌握的知识不一样。即使是都是技术人员，不同的层面，其要求也不一样。数据挖掘应用技术人员需要掌握调节算法、算法的适用性和结果的合适表现形式；数据挖掘研究者需要探索新的理论，具备创新、改进新算法的知识和能力。

要把数据挖掘应用讲得很明白，其实需要各个方面的很多知识。

我们既要讲解数据挖掘原理，还要讨论数据挖掘具体的技术、数据挖掘应用的行业知识、数据挖掘具体技术与行业知识的结合、数据挖掘结果在行业领域的含义和使用等。涉及数据挖掘原理，其实原理和理论并不是同一个事物，原理可以不借助各种公式而存在，并且可能相对简单。

数据挖掘应用有自身的特色，就是我们下面需要重点讲解的。而目前，对数据挖掘结果的分析还无法实现全自动化，现在更多的只是把现象（数据知识）而不是深入的专业知识（行业知识）展现给用户。这就需要应用者掌握深入的专业知识和技术知识，把数据挖掘自主发现的规律总结成专业知识，并应用于工作实践。

1.2 数据挖掘应用基础

数据挖掘是一种获得知识的技术。它的基础是数据，手段是各种算法，目的是获得数据中蕴含的知识。发现知识并非易事，人们总是受到各种各样的局限。在信息时代之前，数据的缺乏成为发现知识的限制。在天文学的早期，能有所发现不仅是需要有敏锐的头脑和先进的观测设备，前人留下的大量观测数据也是必不可少的条件。实际上，就算是目前，数据的缺乏仍然是发现知识的主要瓶颈。随着数据采集和存储技术的发展，对大量数据的分析和使用成为了一个新的难题。数据挖掘是一门处理大数据的应用科学，它是随着对大数据分析处理的需要而诞生的新兴学科，出现至今只有短短二、三十年的时间，在各方面还在不断发展和完善之中。对数据挖掘应用而言，知识的发现存在两个极限，一个是前面提到的数据极限，即数据要么非常非常庞大，要么数据量足够的小，或数据数量小但维度非常大；另一个是算法极限，即我们针对很多数据（不同的性质、不同的形式）和很多需求，目前所有的算法尚不能很好地解决某些问题。

做一个比喻，要想获得知识的甜美果实，需要穿过的两层硬壳，如图1-5所示，一层是穿越处理足够数量和质量的数据硬壳，另一层是穿越足以处理、发现并展示蕴含于数据中知识的算法的硬壳。但同时，壳的核心有没有果实是最重要的。

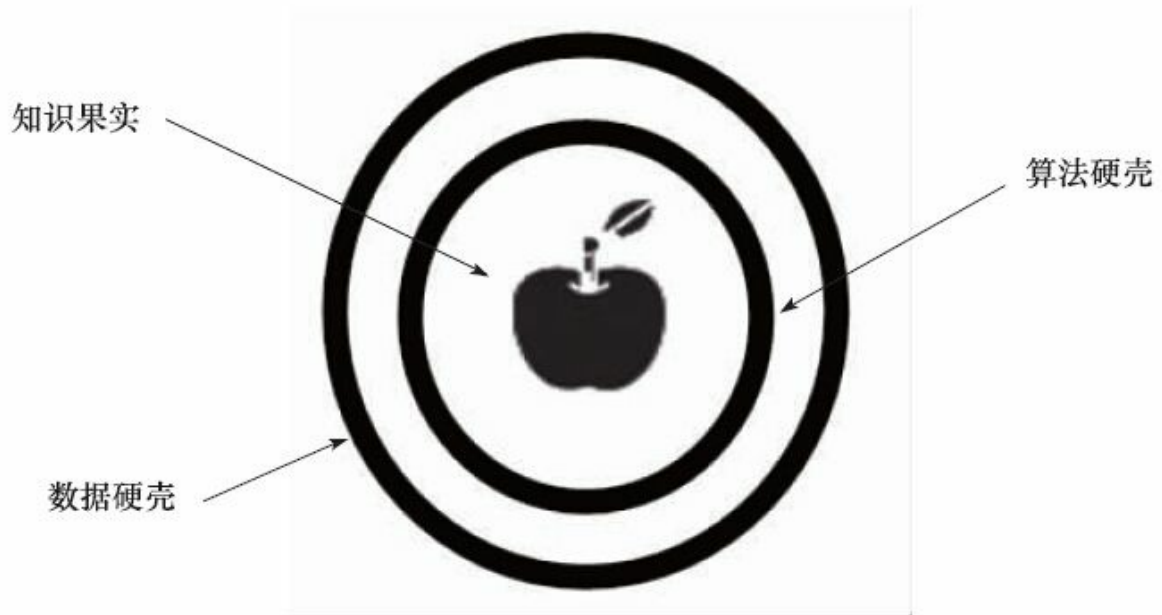


图 1-5

因此，数据挖掘应用具有三个要素：数据、算法、知识。了解、学习数据挖掘，我们需要从数据的源头讲起，从大千世界的事物讲起。

本节将讨论以下几个方面的内容：事物与维度、分布与关系、描绘与预测、现象和知识，以及规律与因果。

1.2.1 事物与维度

事物概念是指客观存在的一切事情和物体，也许是具体的事物，例如人、动物、植物、电脑、面包等；或者是抽象的事物，例如购买、就医等。它是一组数据描绘的对象。维度，某种事物的任意特征形成的一个变量，对于不同个体可能有不同的值，如身高、血压等。事物的属性就是事物的本质在各个方面的表现，维度和属性可以互换。表1-1、图1-6和图1-7列举了几种事物和其常见的维度。

表1-1 维度

事物	维度
人	身高、体重、职业、肤色
鱼	背鳍、尾鳍、侧线
花	花瓣、花蕊
购买	单价、数量、折扣、卖方
就医	时间、病种、预后

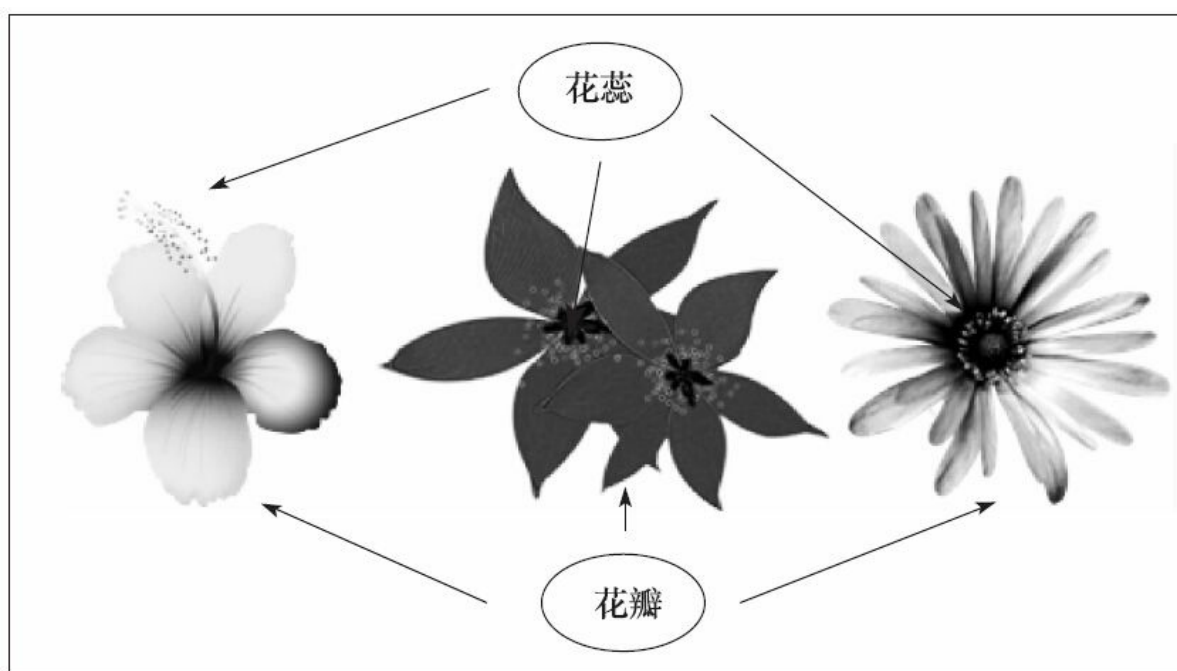


图 1-6

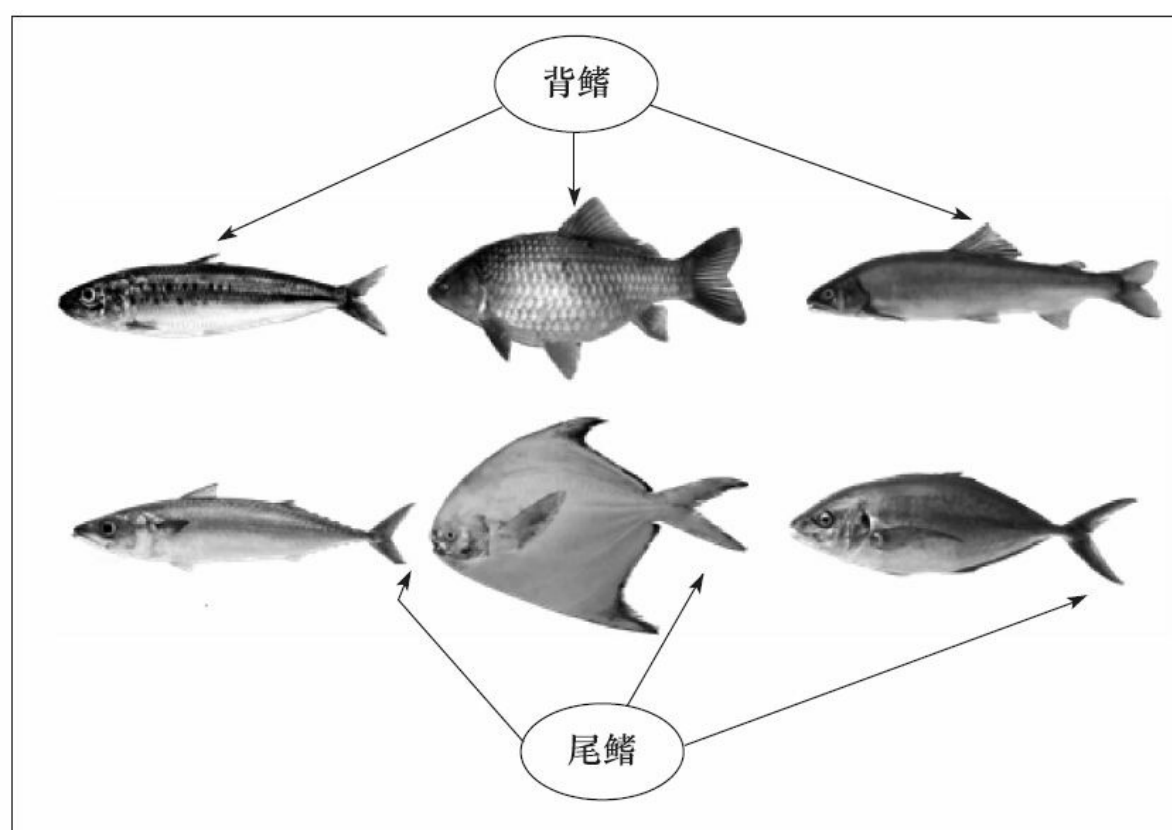


图 1-7

事物就是各个维度的总合。要描绘事物或区分不同的事物，就需

要从事物的属性入手，而且是描绘的事物特殊属性。例如，描绘鲸鱼就会说巨大、鱼（其实是哺乳类的）、会喷水等。描绘大象会说鼻子长、陆生、巨大等。描绘人类自己也是一样，我们去某公司问哪位是经理时，会说是那位高个的、穿深蓝西装的、坐在转椅上的那位男士等（如图1-8所示）。这些描述并没有把事物的每个维度都告诉我们，而是重点说明了几个维度，就足够使我们了解或区分事物。这些维度我们称为该事物的特征维度，它们是可以区别（划分）于其他事物的特殊属性。事物的特征不会体现在所有维度上，正如区分两个人不需要对比每个细胞一样，只要比较一下指纹或DNA就可以了。指纹和DNA就是区分人类个体之间的特征维度。

特征维度可以是长期的，例如大象的鼻子；也可以是暂时的，例如经理的座位、穿着等。特征维度也随着人们对事物不同的兴趣而不同，例如医生要研究心血管疾病的血脂指标（医学维度）；商场的经理要分析客户购物的年消费指标（购物维度）等。除了这些不同之外，事物之间还有彼此相似的维度。例如带鱼和鲸都生活在水中，从这个维度人们无法获知两者的区别。所以说，针对不同事物的比较，某个事物的特征维度也可能是不同的。

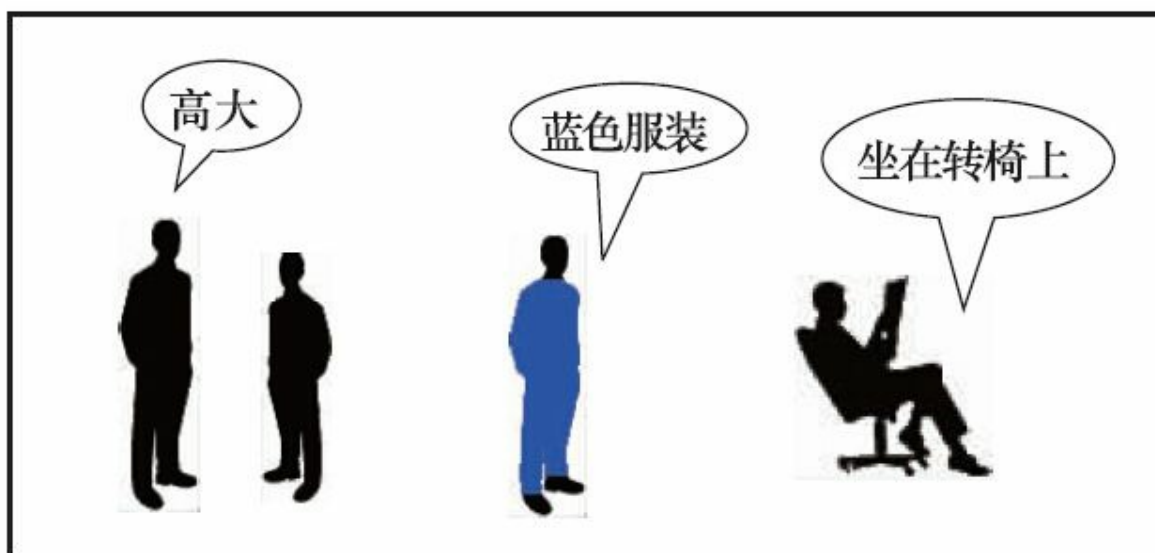


图 1-8

在数据技术上，我们使用不同的数据类型来描述事物不同的维度。维度在数据上有四种常见的取值类型：

- 标称类型，如生物名称：抹香鲸、大白鲨、蓝鲸等，可以用于区分。

- 有序类型，如生物身长：大（蓝鲸）、中（抹香鲸）、小（大白鲨）等，可以排序。

- 区间类型，如海水温度：摄氏温度10度至摄氏温度30度等，可以计算差值。区间类型有个特点：就是可以计算30摄氏度比10摄氏度高了20度，但是说30摄氏度是10摄氏度的3倍则是错误的。

- 比例类型，如质量：大白鲨2000千克，抹香鲸20000千克等，可以计算倍率。我们可以计算出抹香鲸的体重是大白鲨的10倍。

在四种类型中，标称类型和有序类型是定性的，区间类型和比例类型是定量的。

数据挖掘中的连续型数据只可以是区间类型和比例类型，离散型数据可以是上面四种类型任意一种（对区间类型和比例类型进行离散化后）。

1.2.2 分布与关系

某事物与其他事物一定在某些维度上有所不同，否则我们就无法清楚地将该事物与其他事物区分开来。维度从两个方面描绘事物：维度的分布和维度的关系。正如所有维度一样，特征维度也包括分布特征和关系特征，我们常见的分布特征更容易被发现。维度的分布一方面可以从维度的分布类型上来考量，另一方面可以从维度的分布范围上来考量，如图1-9所示。

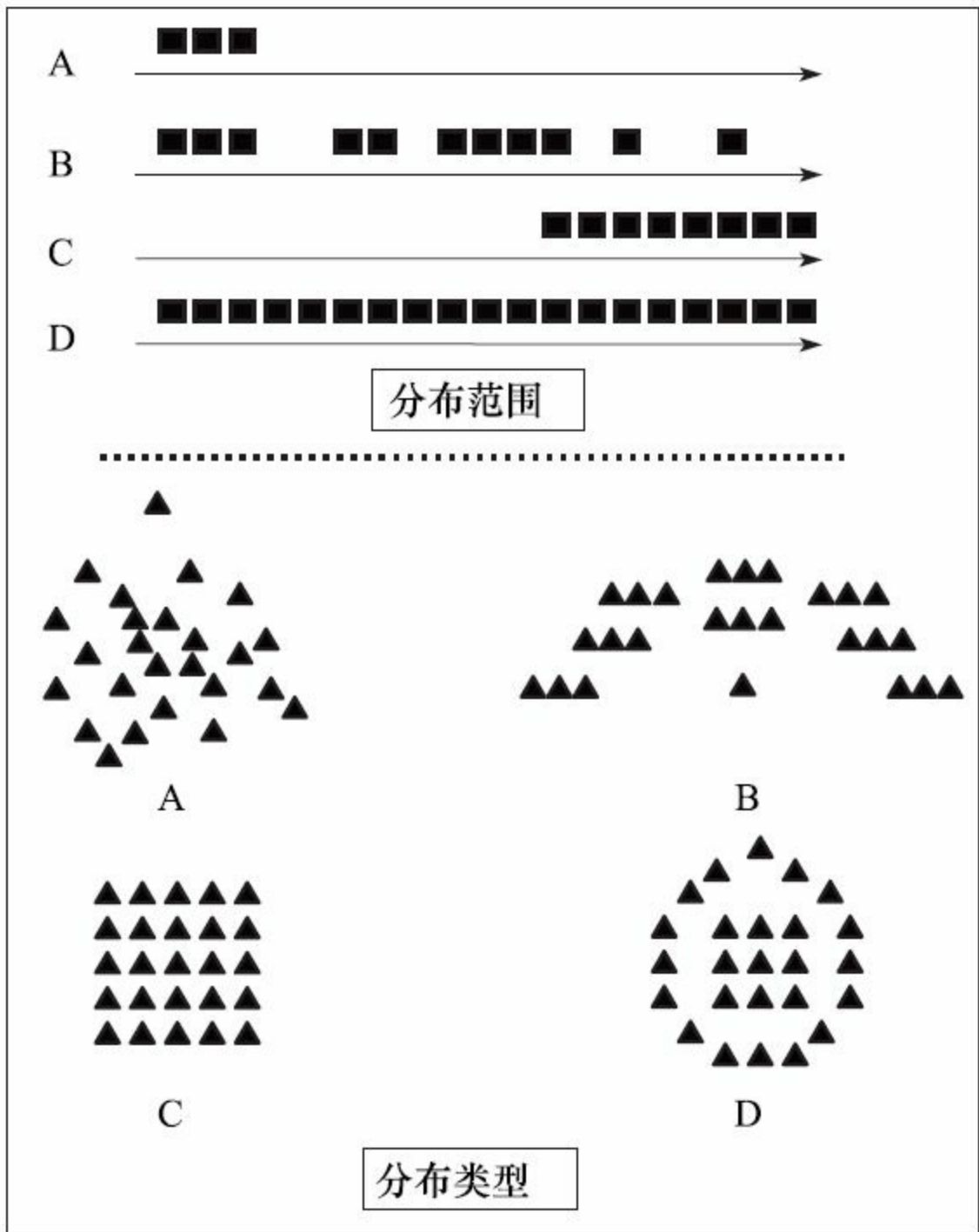


图 1-9

维度的分布范围，例如，人类的血钾生理指标的正常范围是3.5至5.5mmol每升，低血钾的范围是小于3.5mmol每升，反之，高血钾的范围是大于5.5mmol每升。血钾的上限极值或下限极值可能是个隐含

值，高血钾再高也不会到10000mmol，所以应该是5.5至某个值（致死量的值）。除了维度的分布范围之外，维度的分布类型也参与描述事物。例如，博弈游戏中使用的骰子，最简单的是六面骰子。正常骰子是均匀分布（图1-10），各面以六分之一的概率出现，那些有问题的骰子是非均匀分布（图1-11）。

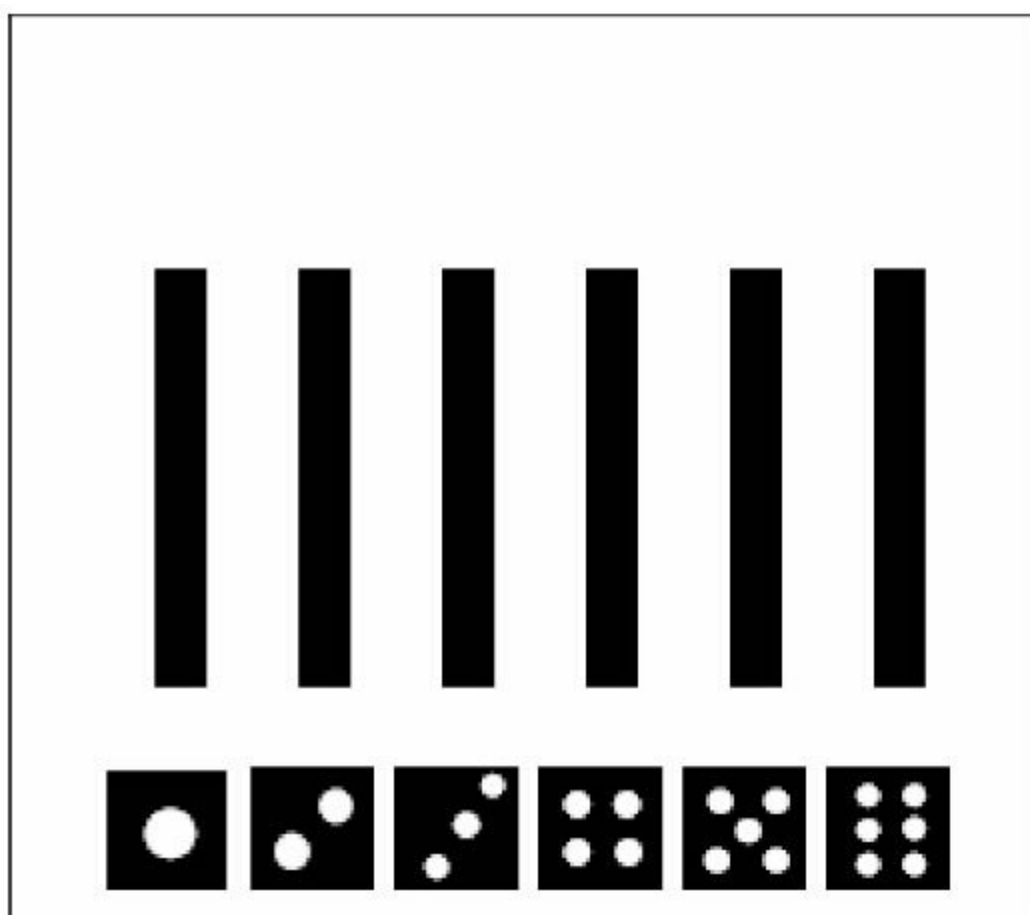


图 1-10

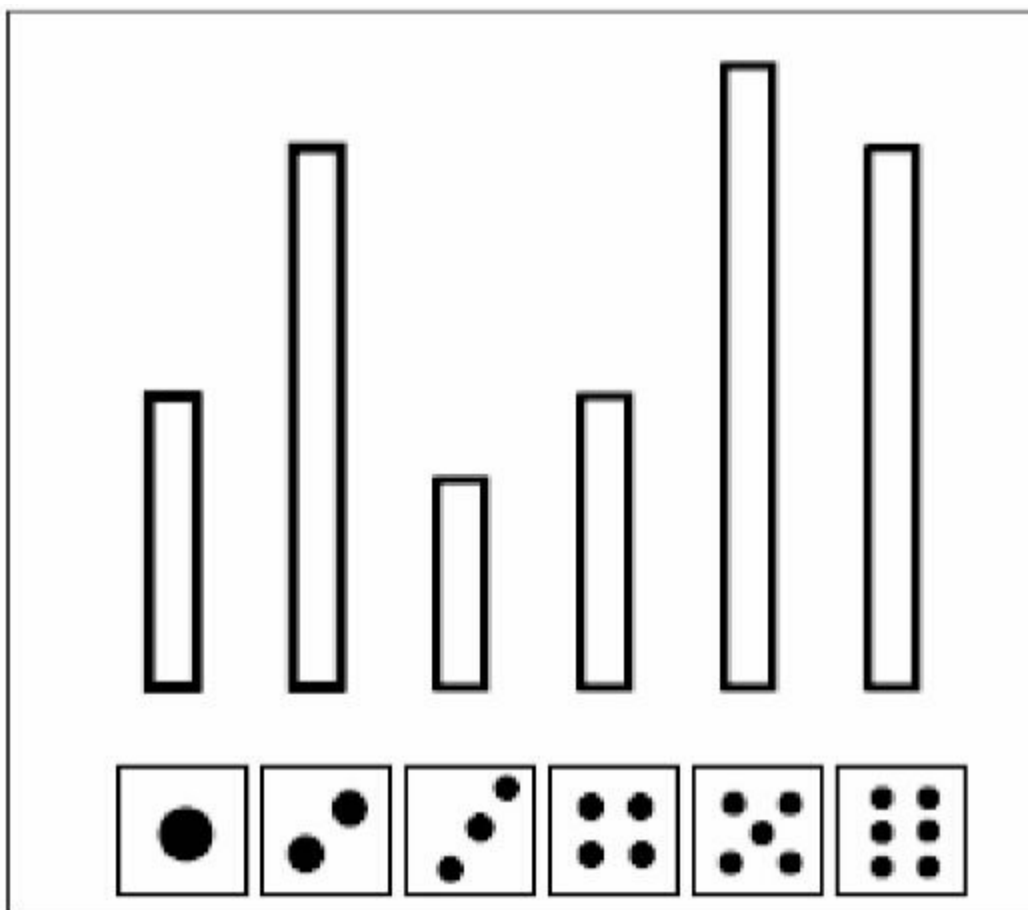


图 1-11

除了维度的分布，维度的关系也是事物的特征。例如，图1-12描述了正常人和某种代谢病人的血糖维度和胰岛素维度的关系曲线。

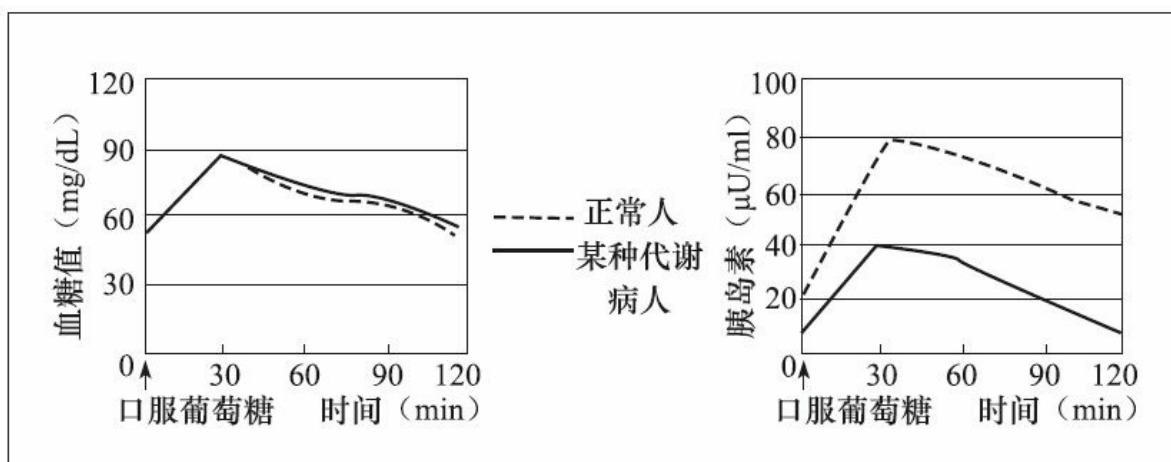


图 1-12

可以看出该病患者的血糖和胰岛素的关系特征与正常人的不同，它显示该病患者胰岛素分泌出现特殊病理性改变。

维度的关系特征常见的表现有伴随关系和影响关系。维度的伴随关系特征，例如，在北半球，秋天来临时，树木落叶。这个自然现象维度伴随季节维度的关系特征是我们常见的，当季节维度的一个变化值秋天出现时，自然现象维度的一个变化值落叶也伴随出现。读者要注意的是，伴随关系存在着方向性，即秋天带来了落叶，而不是落叶带来了秋天。

常见的维度的关系特征还有维度的影响特征。例如，在冠心病患者的生理指标中，影响血甘油三酯的维度有哪些？它们对血甘油三酯的影响强度是怎样的？图1-13所示的是甘油三酯维度的关系影响特征，高密度脂蛋白胆固醇对甘油三酯的影响度最高，其次是总胆固醇、低密度脂蛋白胆固醇、A/G等指标维度。

研究疾病：冠状动脉粥样硬化性心脏病
研究检验：★甘油三酯

结果分析图

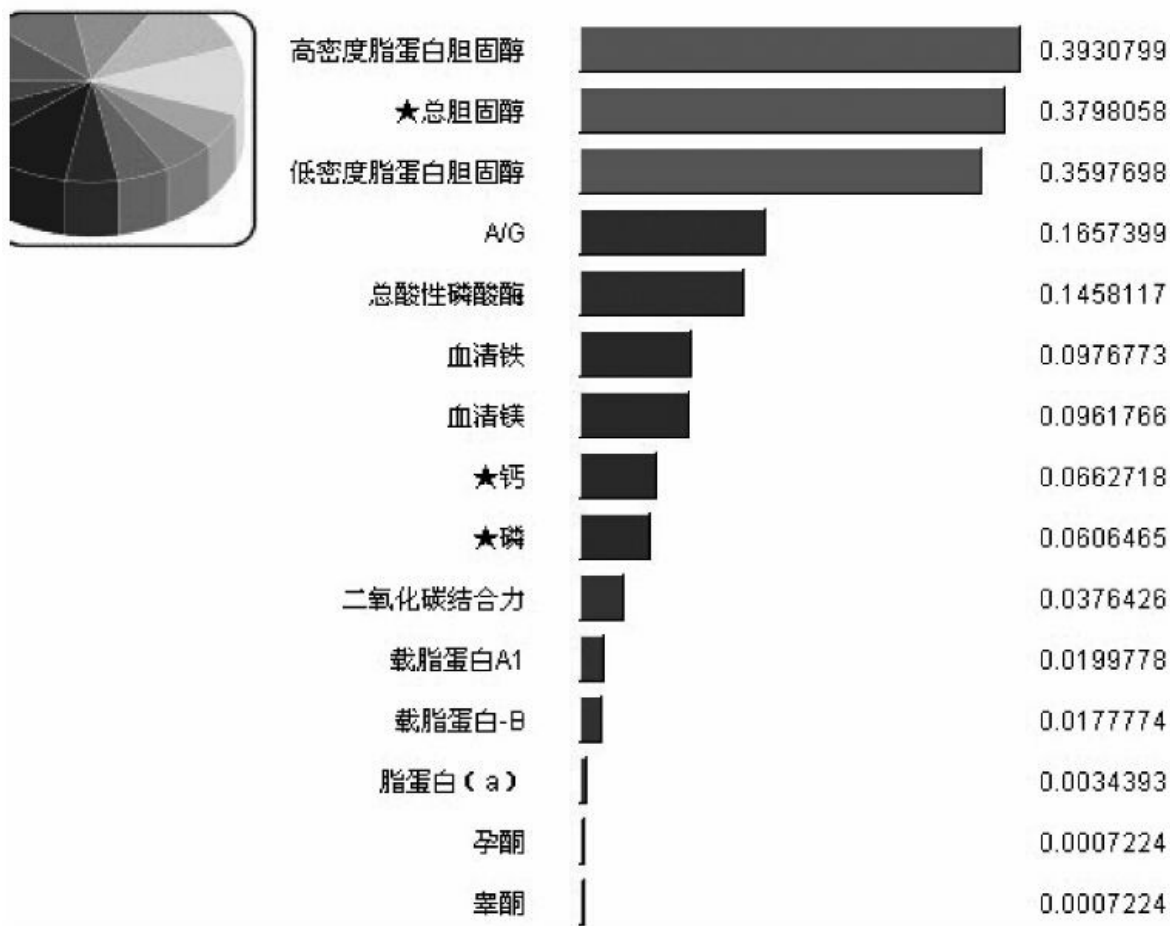


图 1-13

可以看出，维度的分布和关系都可以描绘事物，事物的特征也蕴含其中。发现事物维度的分布与维度的关系，就是数据挖掘算法的基础，如图1-14所示。

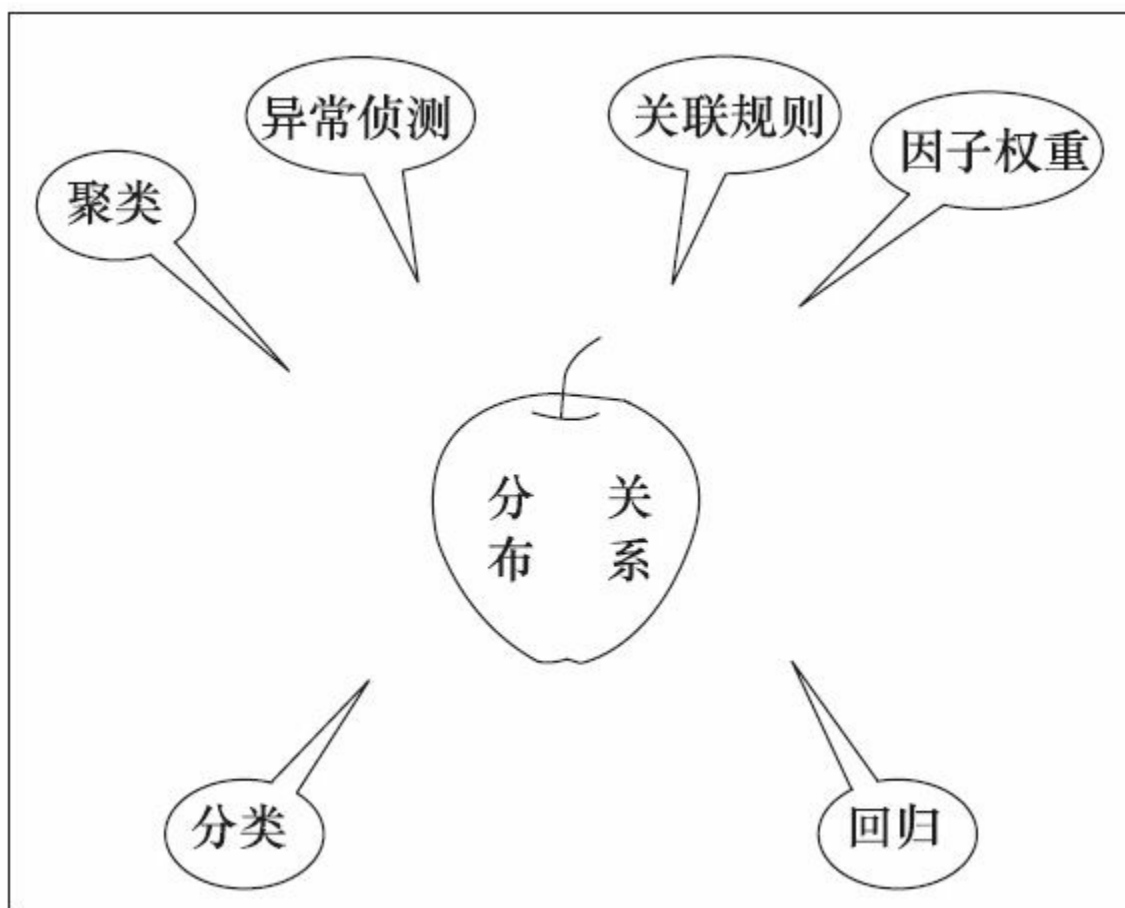


图 1-14

1.2.3 描绘与预测

事物维度的分布和关系可以用来描绘事物特征，也可以用来区分彼此。不同的事物之间有着不同的特征维度。如果事物内部分有亚型或子类，每个子类之间也必然存在某特定的维度来区分彼此，这个特征维度可能是母类的特征维度，也可能不是。如下图1-15所示，每种鸟类在特征维度上将它与其他鸟类区分开来，而鸟类的各个亚型既具有母型所拥有的维度特征，例如翅膀和双腿，同时具有自身的特点。

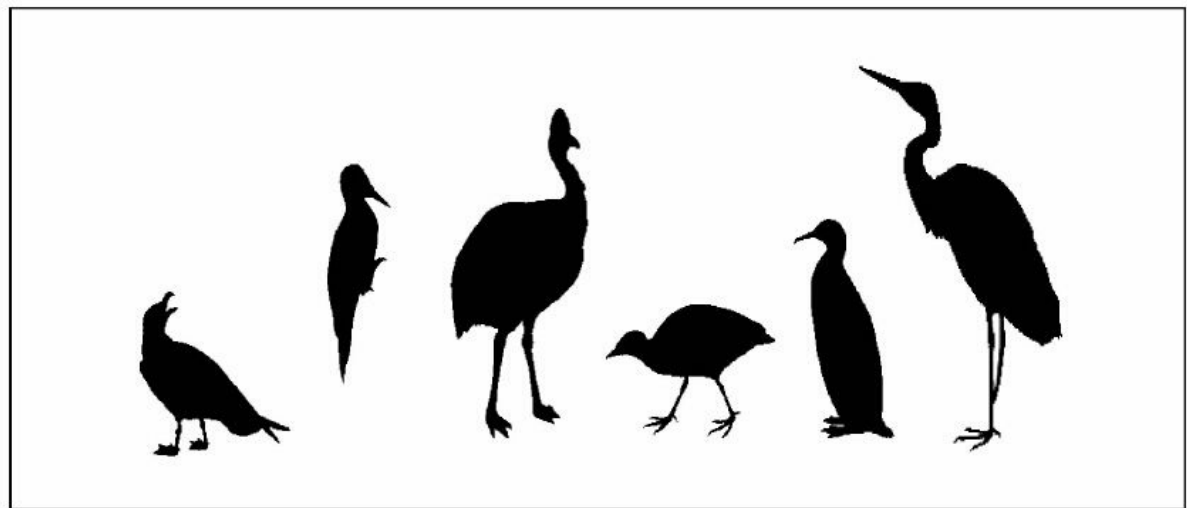


图 1-15

再如，高血压的特征维度是血压，各个亚型从血压指标维度上将它与其他高血压亚型区别开来，而各个高血压亚型同样也具有母型所拥有的维度特点，同时具有自身的特点，如表1-2高血压分型表。

表1-2 高血压分型表

分类	收缩压 (mmHg)	舒张压 (mmHg)
正常血压	<120 和	<80
正常高值	120 – 139 和/或	80 – 89
高血压	≥ 140 和/或	≥ 90
1 级高血压 (轻度)	140 – 159 和/或	90 – 99
2 级高血压 (中度)	160 – 179 和/或	100 – 109
3 级高血压 (重度)	≥ 180 和/或	≥ 110
单纯收缩期高血压	≥ 140 和	<90

另外，我们还可以按照高血压另外的一个特征维度，即高血压的性质进行分型，如可以将高血压分为原发性高血压和继发性高血压。

数据挖掘的很多的技术是用来发现特征维度的，例如后面章节讲的特性选择技术和特征抽取技术。发现事物的特征维度就可以准确描绘事物、区分事物或其子类。反过来，如果一个或一组维度可以良好地区分事物或其子类，那么这些维度也就很有可能是这个事物或其子类的特征维度。

通过维度的分布和关系，我们可以对事物进行有效地预测。离散型数据的推测一般叫分类，也就是标称类型和有序类型。例如通过哺乳这个维度，我们把鲸鱼归类于哺乳类，尽管它长得像鱼。再例如，预测药品的研发结果，也就是在二元数据（成功、失败）中对药品实验进行分类。预测政策的出台也是一个分类推测，例如银根是收紧还是放松？政策是鼓励还是抑制？还有预测经济危机爆发、自然灾害发生、地区爆发战争等，都是分类推测的范畴。又例如，通过重量维度可以推测盒子是否是空的，这样不必一个个打开盒子查看，而是使用风扇吹落较轻的空盒子就可以了。对于推测连续型维度的未来值，我们把它叫回归预测。例如，通过各种交易维度预测股票的价格。再例

如，很多医学检查花费高昂，或属于破坏性检查，这些检查的数据不容易获得，我们就可以使用数据上容易获得的医学检查来推测出那些数据不容易获得的检查值。

1.2.4 现象和知识

知识的获得是从思路被启发开始的。对于数据挖掘，数据探索之前的思路启发才是整个挖掘的开始。思路被启发是从看到特定的现象开始的。无论是传说中的苹果砸牛顿，还是现实中的知识发现，都是被现象启发了思路。数据中虽然存在各种现象，但是很多都难以直接观察到。数据挖掘技术从海量的数据中挖掘出现象并展示给使用者，让使用者从现象中寻找灵感和启发，如图1-16所示。这样，使用者可以接触到大量的现象，更多也更容易受到启发。

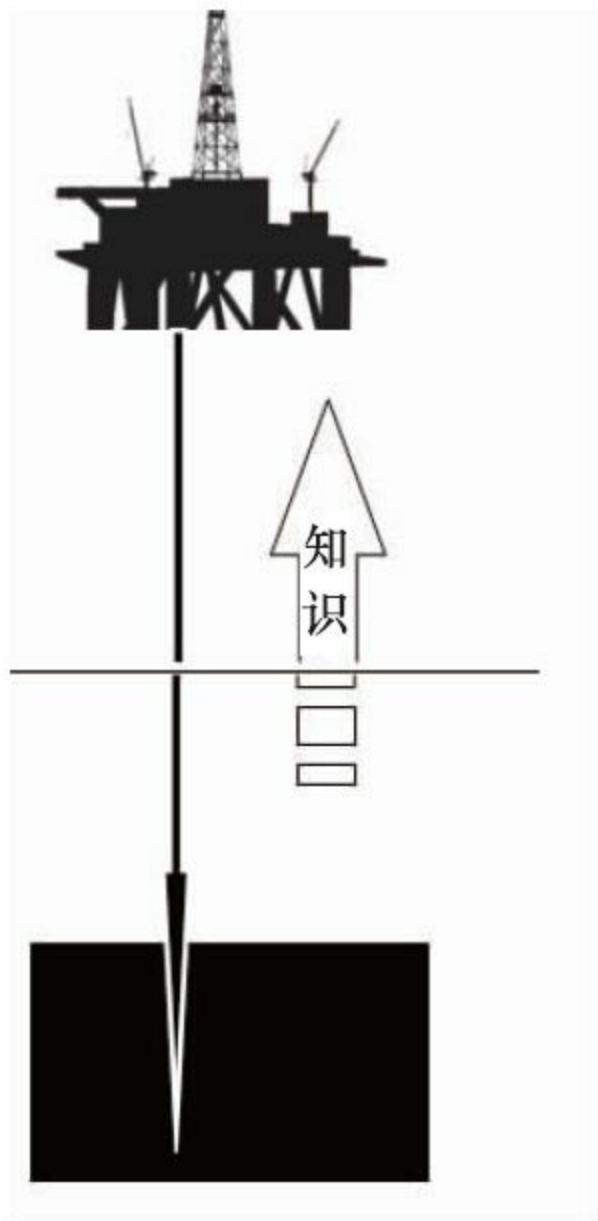


图 1-16

数据挖掘技术的应用，特别是数据挖掘应用产品化之后，被产品包装的、方便易用的数据挖掘技术甚至可以自动进行数据探索，进而能够提供大量的现象来给人以启发。数据挖掘的结果很大一部分是现象，或仅能称作数据知识，而不是人们更需要的业务知识，还需要人脑归纳加工成专业领域的知识。

1.2.5 规律与因果

知识的获得不是没有目的的乱撞，而是从一个现象开始。大部分时候是某种现象给了我们启发，随后经过一系列的分析发现了规律。规律只是回答了What，而没有回答Why，还需要发现规律背后的原因，也就是因果。从规律开始，规律和因果都是可以利用的，都可以称为广义的知识。数据挖掘发现的这些维度间的分布、维度间的关系，通常都是一种规律。数据挖掘目前还是一种工具，观察事物维度的工具。如同通过显微镜可以观察肉眼看不见的细菌、病毒一样，数据挖掘可以发现人脑难以发现的维度分布和维度关系，并且以适当的方式展现出来。但如何展示某些高维度的关系也不是一件容易的事，而且，关系过于复杂的情况下，人脑也难以理解这些结果。

算法把这些分布与关系发现并展示出来，但没有进一步说明为什么这样。也就是说，大部分的输出结果还是一种现象级的规律。当然，规律并非要等到知道为什么才可以使用。随着数据越来越丰富，算法越来越发展，我们获得的规律会越来越多，来不及知道什么的规律也会越来越多。一边使用What一边发现Why，很可能是大数据时代知识的使用方式。

凭借因果关系来预测和推测在目前也有一些问题，因为事物的因果关系并不是都是简明的、线性的。面对复杂的高维度的事物，探索因果关系往往会捉襟见肘。有时候，我们实际上不可能等到通晓运行机制之后再返回头来应对眼下正在发生的种种问题，不可能在完美解

释之后再来解决。更多的时候，我们只能根据有限的数据来做出反应。例如，我们应对高致病性流感，无论从时间上，还是从数据上，都不能等到一切完美了再行动应对。

1.3 数据挖掘应用系统工程

数据挖掘不是凭空存在的空中楼阁，而是一个大系统中的一个组成部分。谈及数据挖掘的实践，就不能把数据挖掘从行业应用中分割开来。数据挖掘这个词汇是一个复合词，由数据和挖掘两个部分组成。数据不是凭空出现的，而是来自各种不同的业务。数据挖掘工作于业务数据之上，自然要与业务产生联系。只不过，数据挖掘可以把一个行业的知识挖掘出来后抽象化和概括化，甚至使之有可以跨行业使用的能力。当然，更多的时候获取的知识首先是应用于本行业的。

数据挖掘既是整个业务系统的一部分，但自身也是一个体系。数据挖掘应用是一个多层次、流程化的工程任务，开展数据挖掘应用也需要从每个层面加以处理，才能保证整个数据挖掘系统的成功运行。数据挖掘应用从上到下可分为三个大层面，每个大层面下又有若干子层面，具体结构如下：

- 应用层：把数据挖掘结果应用于实践。
- 算法层：提供算法、引擎和界面。
- 数据层：提供数据源、数据探索、数据准备。

层次从低到高，是一个蕴含于事物中的知识发现和使用的过程，如图1-17所示。

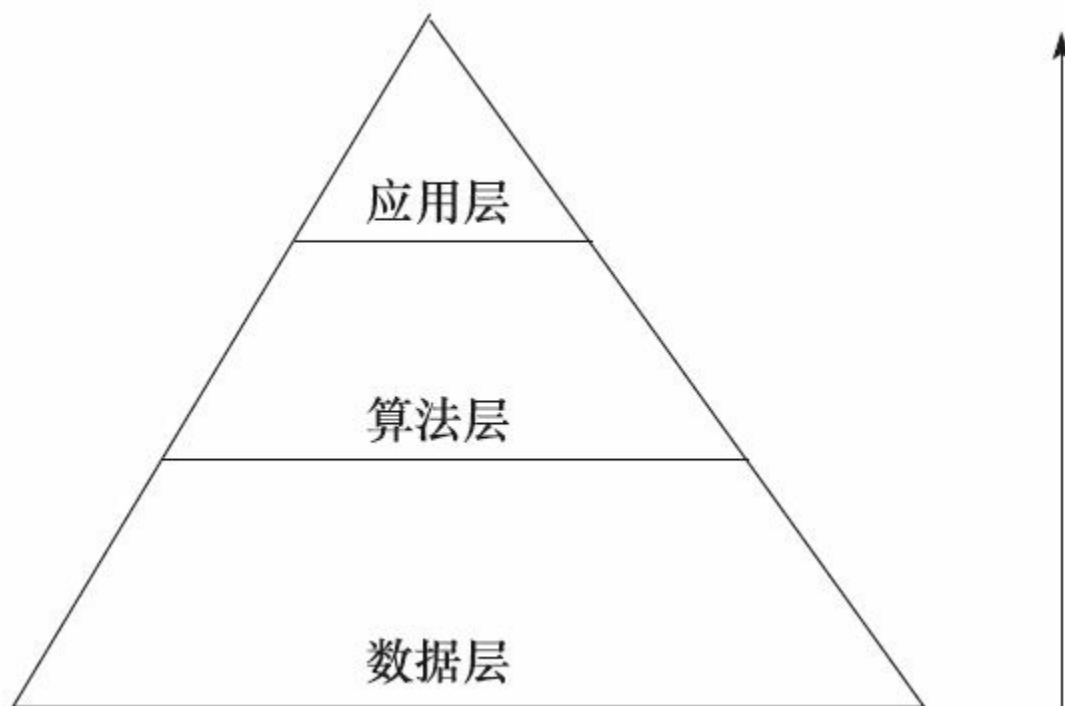


图 1-17

本节将对以上的三个层面做个详细探讨：首先是数据层，包括数据来源、数据流、数据的提取与利用、数据的先验知识等话题；然后是算法层，主要包括分布探索、异常探索、关系探索等内容；最后是应用层，包括数据挖掘的专业应用和面向行业的应用产品两个方面。

1.3.1 数据层

我们首先看一下数据层，这是数据挖掘的基础所在。知识是蕴含于数据层的，数据层如果不存在知识，那么最后也就没有知识可以发现并应用。

1.3.1.1 数据来源

1.接触

数据来源是事物中可以接触到的部分，例如病人来医院就诊，我们可以发现发生的病症，例行体检的人员也可以被查出某种疾病的发生。顾客访问商店，我们可以了解顾客的购买意向。数据来源有可能是来自事物的全部，也有可能是来自事物的一部分。事物中可以接触到的部分，我们称为数据来源。接触不到的，例如深海的未知生物，遥远星球的生命，无法成为数据来源。不同的事物有着不同的可接触性。例如，对于人们来说，池塘的鱼类的可接触性就大于深海鱼类。因此，在考察事物的特征时，我们要随时提醒自己，是不是还有事物的某些部分没有接触到？这些没有接触到的信息是不是会影响挖掘结果？数据来源有可能提供事物的全部维度，也有可能只提供事物的部分维度，如糖尿病患者的血糖、血胰岛素；风湿患者的风湿因子；顾客购买的商品。

2.采集

接触事物并不代表事物可以被记录，事物的各维度数据被采集到后，才是可以使用的数据。并且，数据采集的方法会影响对数据来源的真实反映，从而直接影响数据挖掘的最终效果。例如己糖激酶法（Hexokinase method）测量血糖的精度就高于葡萄糖氧化酶法（Glucose oxidase method），而更古老的福林-吴氏法（Folin-Wu method）测血糖含量时，因并非全部为葡萄糖，有不少是非糖的还原物质，因而测得的数值比实际高。此外，还要注意的是数据采集的目的，在多数情况下是支持业务的运行，而不是为了数据挖掘而直接准备的。因此，这样的数据在进行挖掘前需要进行大量的预处理。

1.3.1.2 数据流

事物的各个维度信息不断地流入业务系统，构成了大量的业务数据。而大量的业务数据再源源不断地流入用于数据挖掘的数据库，形成一个历史的、不断增长的大数据。一个可分析的大数据需要考虑它的流入，也就是从数据来源采集的高维度信息，也要考虑流出，也就是数据的过期和损坏。如图1-18所示，显示的是数据源的存量和流量。

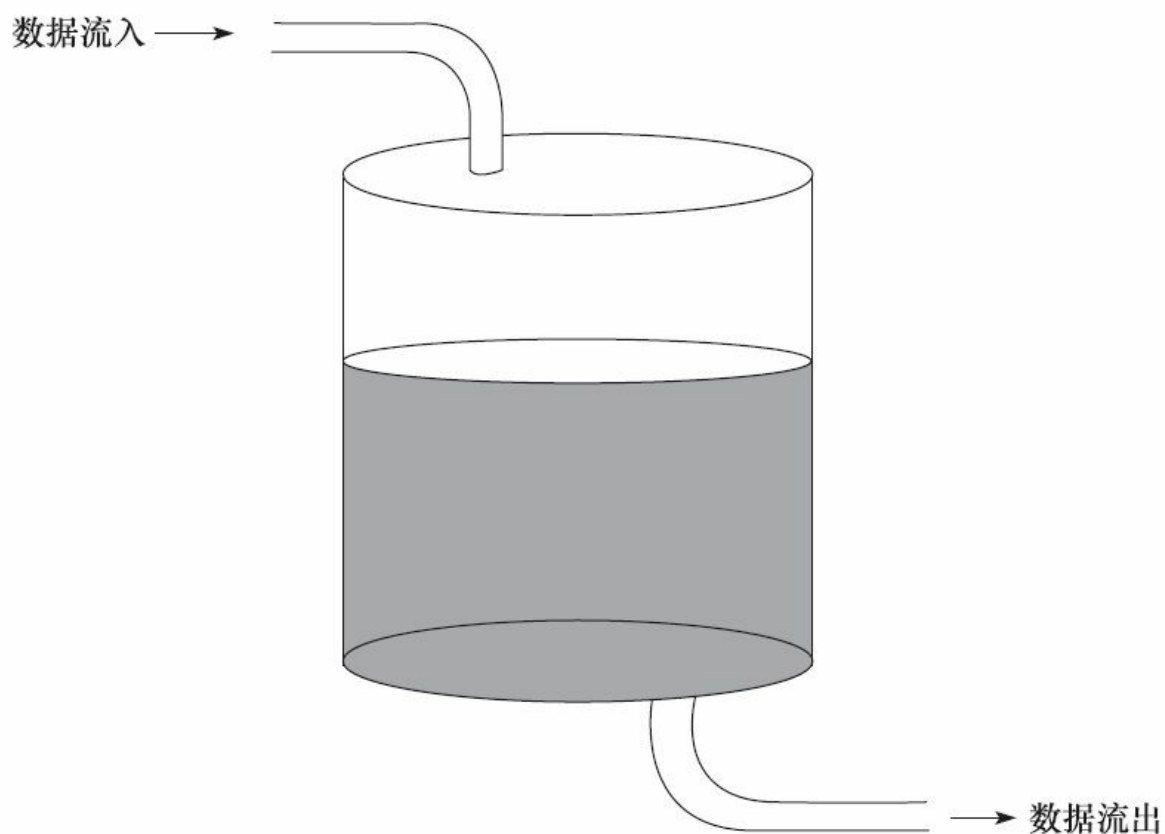


图 1-18

1.数据流入

对大数据的分析，首先就需要观察各个子领域的数据增长量和增长模式。还是以医学为例，图1-19为医疗机构的数据增长量。读者可以注意到收费数据的增长量大于医嘱数据的增长量，原因是绝大部分医嘱需要记账，账面收费和实际收费还涉及医疗保险和社会保险的问题。

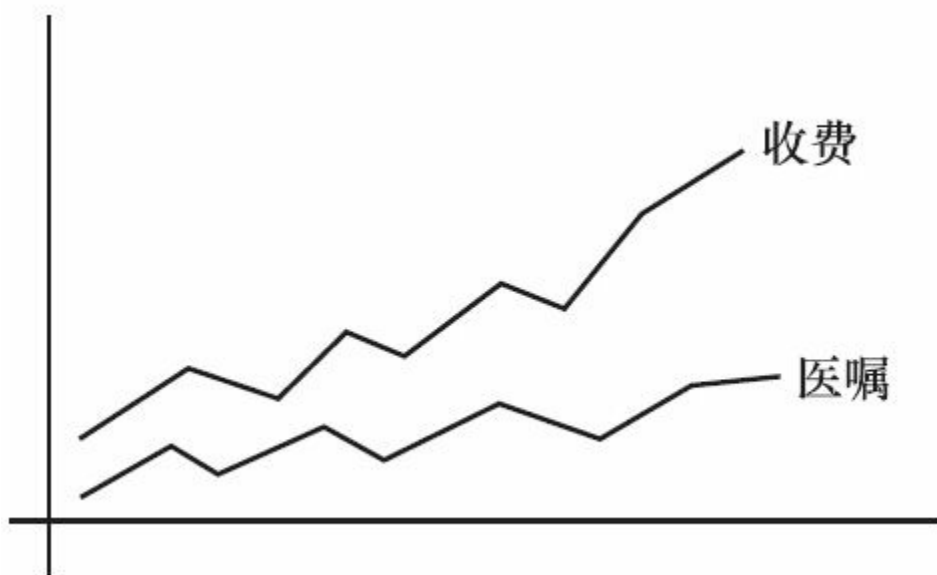


图 1-19

我们看一下收费表（图1-20、图1-21），其中包括治疗、化验、检查、药品、材料和其他，而对应的医嘱表的材料类几乎没有。化验结果表是最后真正做了化验的记录，医嘱表是开单，未必成功执行，收费表是计费，标本未必上机。只有最后的化验结果表是患者真正做了化验的标记（注意标本采集时间，后面解释）。对于用药来说，没有一个药品结果表，所以要分析其流程（见第2章）。

药品	钙尔奇D
药品	注射用血塞通
药品	葡萄糖注射液250ml
药品	丁咯地尔针
药品	安定片
药品	尼莫地平
药品	脑复康片
检验	二氧化碳结合力
检验	钾-离子
检验	氯化物
检验	钠-离子选
检验	尿素测定
检验	葡萄糖测定
检验	血肌酐测定
检验	肌酸激酶
检验	肌酸激酶同功酶
检验	血气酸碱分析
材料	一次性头皮针
材料	一次性头皮针
材料	一次性头皮针
材料	一次性头皮针
材料	一次性头皮针
材料	一次性头皮针
材料	一次性头皮针

图1-20 收费表-明细

2床	钙尔奇D	口服
2床	注射用血塞通	注射
2床	葡萄糖注射液250ml	静脉
2床	丁咯地尔针	注射
2床	安定片	口服
2床	尼莫地平	口服
2床	脑复康片	口服
2床	二氧化碳结合力	普管
2床	钾-离子	普管
2床	氯化物	普管
2床	钠-离子选	普管
2床	尿素测定	普管
2床	葡萄糖测定	普管
2床	血肌酐测定	普管
2床	肌酸激酶	普管
2床	肌酸激酶同功酶	普管
2床	血气酸碱分析	普管

图1-21 收费-数量

2.数据流出

数据可能丢失或损坏。例如业务系统切换的时候，或者存储空间紧张的时候，都可能丢失部分数据。还有就是数据的时效性，数据可能会过期，数据蕴含的过去知识现在可能已经不再需要了。数据的形式也可以多种多样，如数据可以来自关系数据库，也可以来自图片库，还可以来自音频库等。再如，医院的x线胶片组成的实体胶片库，纸质病历组成的文本病案库也是数据的一种形式。各种形式的数据都可能由于各种原因而损坏或丢失。如数据纸质品的风化，存储影像数据的胶片老化导致影像数据的丢失，都是数据损坏的一种。而数据也存在过期问题，例如，随着人们购物行为的变化，很多以前的购物记录就难以预测新顾客群体的购物习惯。而数据使用目的不同，也

会使同一数据的有效期不同。

1.3.1.3 数据的提取与利用

1.数据提取

上面我们提到了数据有各种形式，如可以是关系数据、可以是文本数据、可以是影像数据或是音频数据。从技术上讲，数据可以分为结构化数据和非结构化数据。

数据还可以按照行业划分为，如商业数据、金融数据、医学数据等。还可以按照存储数据的数据平台划分，例如常见的DB2数据、Oracle数据、SQL Server数据、MySQL数据等，一个企业往往采用不同的系统记录业务数据。

描述数据数量主要有两个方面：记录数和维度数。一个数据源的总记录数多，不能说明需要研究的目标的记录数多。而足够的可用于分析的记录是数据挖掘引擎正常工作的需要。数据的实际数量和事物的实际数量有所不同。维度数越多，可以探索的方面就越多，可能发现的分布和关系就越多。所以大数据的“大”，一个是记录数量巨大，一个就是维度涵盖广阔。

从数据源中获取合适的数据，才有可能获得数据中蕴含的知识。从存储在数据库里的大数据中提取数据的常见方法有查询索引，使用筛选查询语句。SQL语言是经常采用的针对数据库的数据获取工具，

还有更方便的、为应用定制开发的数据筛选器。不管哪种方法，获取的数据需符合以下要求：一是要符合算法的数据需要，例如，有的算法只能处理离散型数据，连续型数据就需要进行加工处理。二是要符合业务目的，例如，服用某种药物会影响观测结果，就需要筛选没有服用过某种药物的该病种人群。

2.数据利用

数据是经过各种方式采集的，是事物的维度集合和管理对象，如果无法有效管理数据，那么数据是难以使用的。例如实体胶片源的索引建立不好，就无法找到需要的胶片，那么拥有再多的胶片也难以有效利用。病案的索引不好，也无法找到需要的病历并加以利用。

多数据源的利用有很多优势。例如多中心数据源，可以有效克服单数据中心产生的偏倚。多行业数据源的利用，例如把医疗行业的数据与工业企业的数据同时应用，可能会发现疾病高发区域和工业分布的联系。再例如，将医疗数据和气象数据整合，就可以研究气候和疾病的关系。如研究某种心血管疾病和气候关系时，有趣的是，发现了风速对发病的影响，风大时心血管疾病发病率就低。其实进一步发现的是空气污染对该病的影响，风速高时，小于2.5的微粒（著名的PM2.5）含量减低，于是该病的发病率降低。这个PM2.5，除了大家都认为会影响人们的呼吸系统的健康，其实还对心血管系统产生影响。

1.3.1.4 数据的先验知识

可将数据挖掘视为辅助人脑进行信息加工的技术和过程。在一个完整的数据挖掘过程中，第一步是先验知识的预分析和数据准备，这一步非常重要。数据挖掘工程一定要注意行业知识，用好这些有助于提供预分析数据质量的信息，对评估数据分析的结果是否可用起着很大的作用。例如，图1-22为医院某种代谢病的月份分布情况。

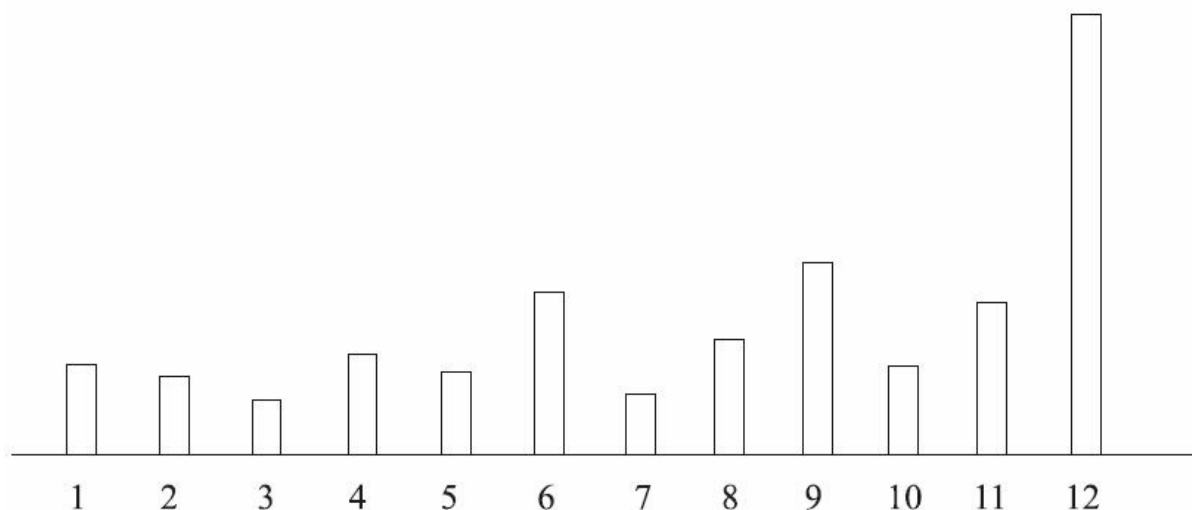


图1-22 医院某种代谢病的月份分布

数据分析显示，该病发生有明显的时间性，高发于12月。而医学专业方面的知识难以解释这种现象。实际上，该病的时间性与该病的基金有关。每到年底基金就会清零重置，如果不用完剩余基金就会净损失。所以在年底前用完就是大多数人的选择，而这种选择的表现就是这种时间分布。数据的分析一定要结合多方面的先验知识，否则就会得出错误结论。当在行业内无法解释结果的情况下，需要考虑来自行业外的影响。

先验知识对于数据挖掘的用处，一是限制了事物的变化，使得我们从更少的选项中选择，提高正确率。二是提供了更多的信息，让我们有更多的可能进行正确的选择。

1.3.2 算法层

算法研究的是如何把数据中蕴含的规律找到，并展现成为人脑可以理解的形式。我们把算法分为六大类，分布探索、关系探索、特征探索、异常探索、推测探索和趋势探索。

1.分布探索。所谓分布探索就是探索数据的客观分布，数据挖掘的主要技术是聚类分析。

2.关系探索。关系探索是探索事物与事物间的伴随关系，以及变量与变量之间的影响关系。采用数据挖掘的主要技术是关联规则和特性选择。

3.特征探索。特征探索是探索由高维变量表达的事物的主题特征，数据挖掘的主要技术是特征抽取。

4.异常探索。异常探索是探索由高维变量表达的事物的离群个案，数据挖掘的主要技术是异常侦测。

5.推测探索。所谓推测探索就是根据有限变量的已知数据推测目标变量的未知值，通常有两种类型，一种是目标变量为离散型，我们将这种推测探索称之为分类模型；另一种是目标变量为连续型，我们将这种推测探索称之为预测模型，也称回归预测。我们把用于推测目标变量的有限已知变量称之为分类因子（分类模型）或预测因子（回归预测）。

6.趋势探索。所谓趋势探索是按照某种次序或序列，通常是时间次序，考察并推测事物发生的变化，数据挖掘的主要技术是时间序列。

详细内容会在后面的章节中阐述，下面我们仅仅举几个例子。

1.3.2.1 分布探索

这类算法是发现分布规律的。图1-23和图1-24显示了事物的分布特点，这类算法也叫分布算法，也可以理解为基于分布的不同而区分事物或子类。

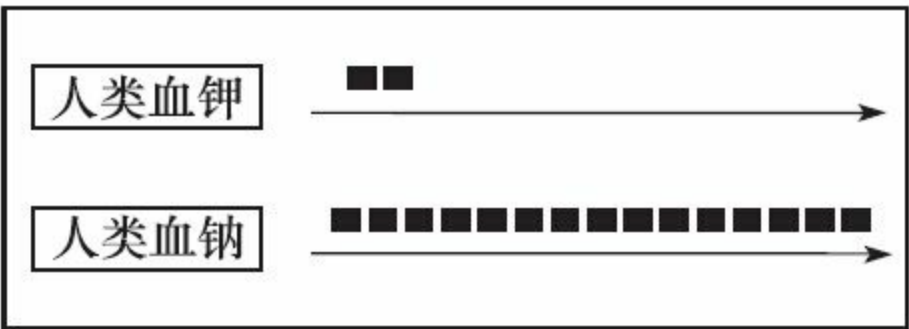


图1-23 分布范围

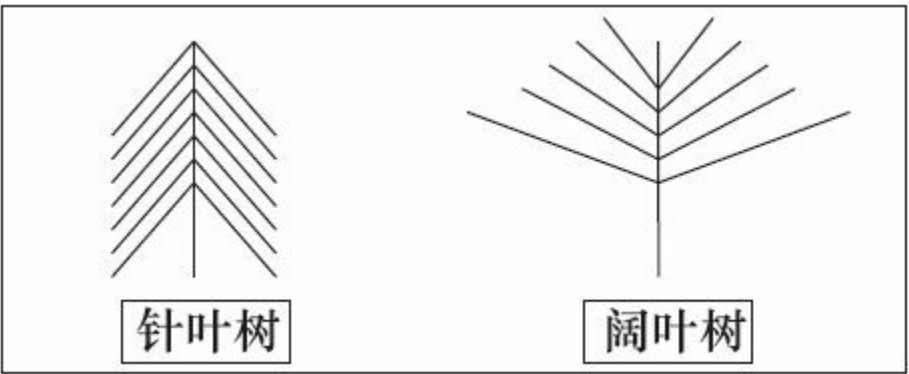


图1-24 分布类型

下面介绍典型的分布算法——聚类。物以类聚，这句话是对聚类算法的最好阐释。正是因为物以类聚，不同类之间的差别和同类之间的相似才可以形成簇或子类。良好的算法形成的簇体现了簇内各个个体的同质性，而且是在特定维度上的同质性。例如，高血压病人在血压上的同质性，生活在海水里的鱼在海洋中的同质性，体检人群中健康人群的体检指标的同质性。簇与簇之间在特定维度上会表现为异质性，例如，高血压病人和低血压病人在血压维度存在异质性，海水鱼和淡水鱼在生活环境维度上存在异质性，健康人和疾病患者在体检指标上存在异质性。

分布的取值范围和分布的形态是聚类的基础。聚类发现维度分布的不均一性，并以簇的形式展现。如图1-25所示，显示了不同的数据呈现出不同的疏密分布。

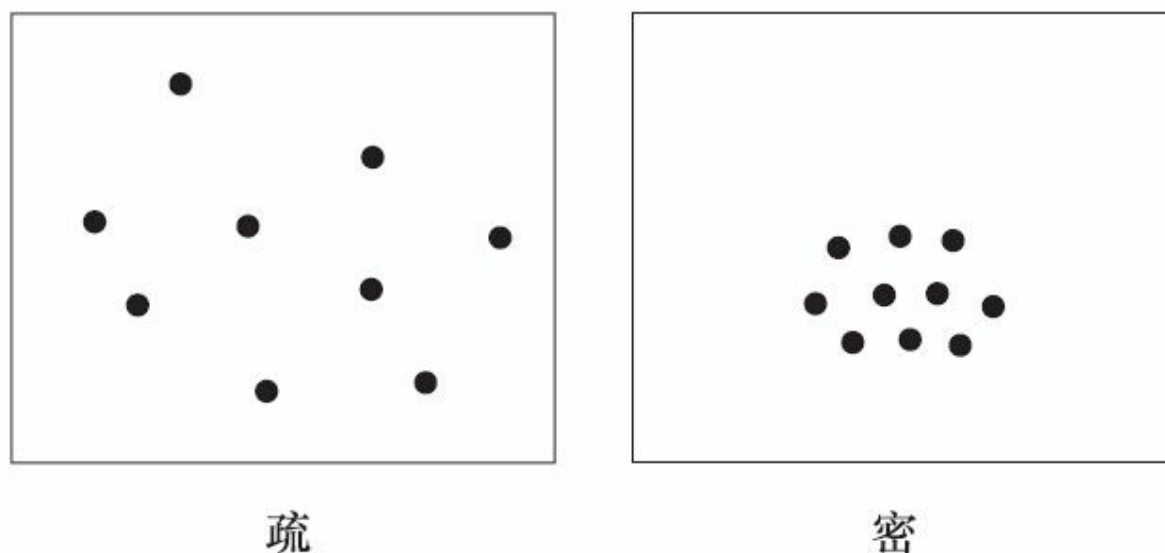


图 1-25

按照簇间差异最大和簇内差异最小，在空间上表现为密度的不

同。还有很多情况，事物之间在特征维度上的分布也存在重叠的现象，如图1-26和图1-27所示。

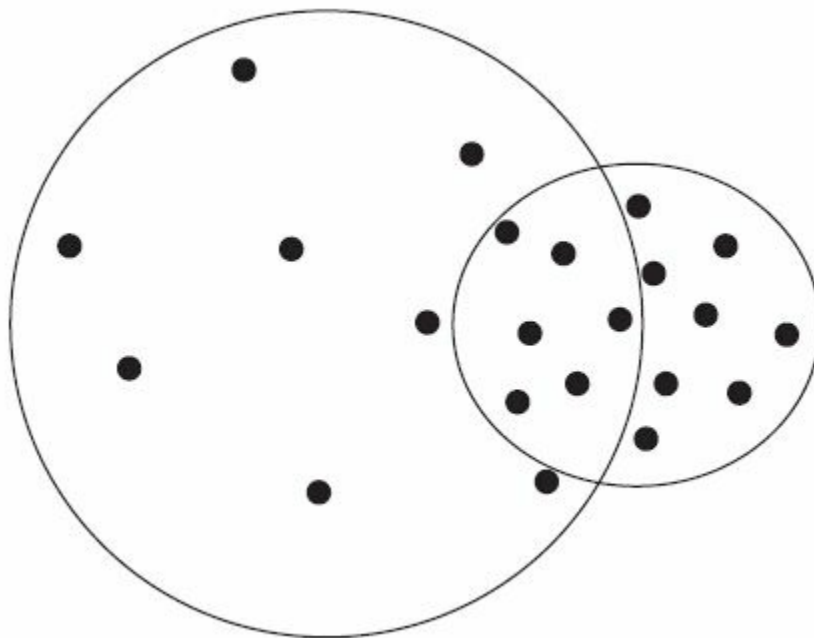


图 1-26

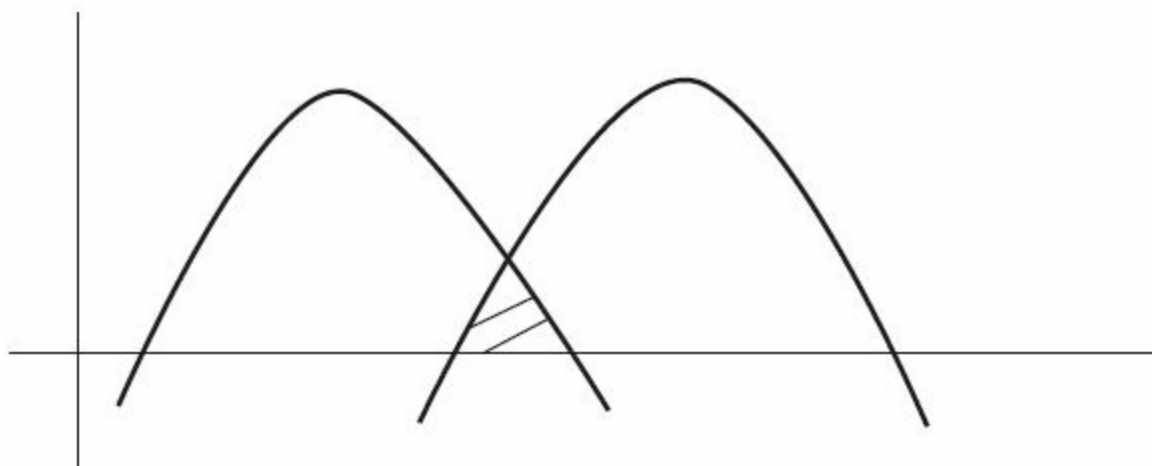


图 1-27

针对以上情况，有三个策略可以选择：

- 如果业务需要必需将重叠部分分开，则根据业务的选择和优化

将重叠部分分配给两个簇中的一个，或者平均分配给两个簇。

- 如果业务不需要必需将重叠部分分开，则保留当前置信度下的划分方法。

- 如果业务对以上两种方案不满意，而且允许使用其他维度来划分数据，则尝试其他维度重新划分。如果获得满意的分类效果，则该维度就是数据的特征维度。

1.3.2.2 异常探索

异常侦测是数据挖掘中异常探索的主要技术，异常侦测是指这样一种工具，它把所有数据尽可能地划分为一类，分进这一类的数据称为正常点，分不进这一类的数据称为异常点。如图1-28所示，不是圈子里面的点都被认定是不正常的，它是一种特殊的分类算法。虽然异常侦测把数据划分为两个部分，一部分认为是正常，另一部分认为是异常，但不同于分类模型中的二分类，异常侦测划分的这两部分从数量上通常有很大的差距，正常的部分应该远远大于异常的部分，也就是说，异常侦测的核心通常是寻找少数、甚至是极少数的异常点。但在实际应用项目中，有时我们也可以扩展异常侦测模型更广泛的应用。

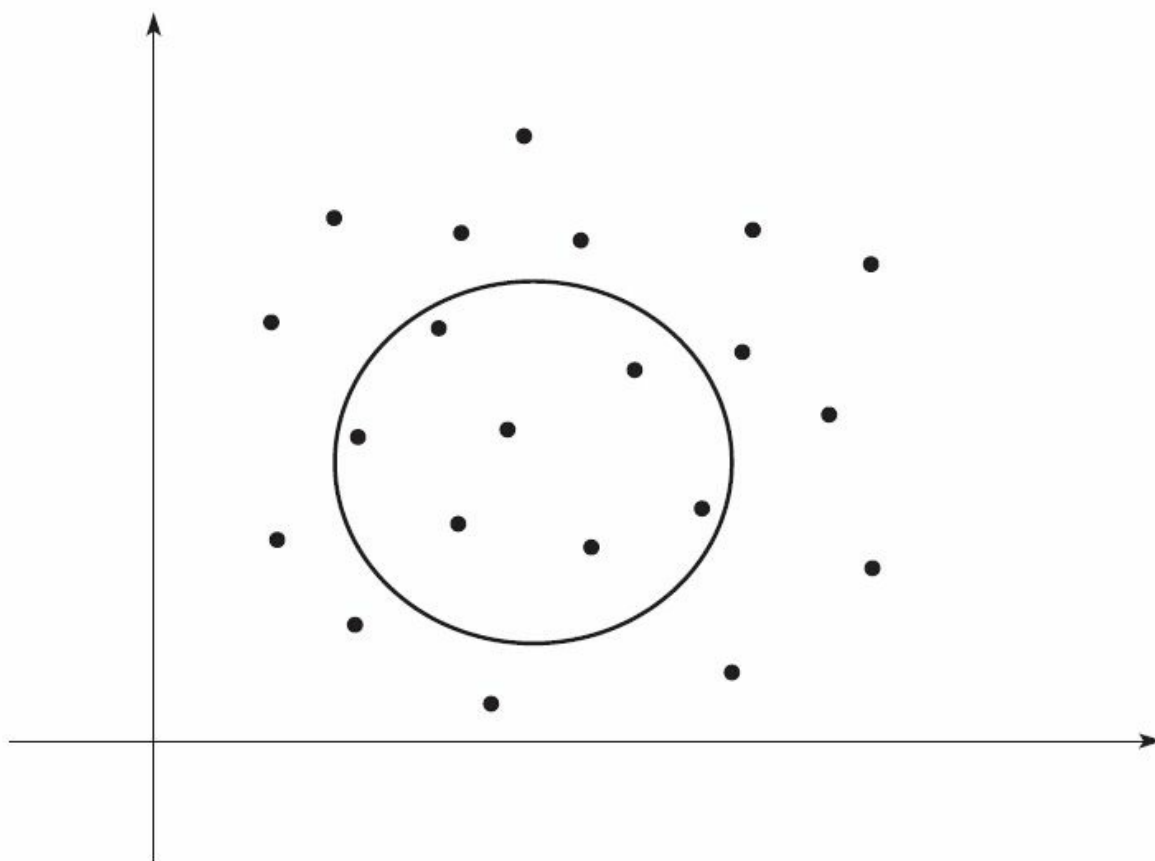


图 1-28

异常侦测模型中的正常和异常是相对而言的。例如，按照不同的维度（如体型、衣着）将一些人群分为男和女，是男人异常？还是女人异常？实际男人、女人都很正常。所以，正常、异常只是一个技术上的标称变量而已，还要根据业务知识进行实际含义的解释。

异常侦测将数据划分为正常和异常，也可以说分为同和异。同的部分，彼此具有一定的同质性；而异的部分，除了与同的部分具有异质性外，彼此之间也可能存在异质性。行业领域中的异常侦测包括发现疾病、发现信用卡盗用等等这些发现异常的应用。

例如，我们将大量的体检人群，根据他们做的所有生理指标检查

结果来进行异常侦测分析。如果我们用两种不同参数设定（异常侦测标准）的异常侦测模型做了两次分析，第一次分析得到的簇A称为健康人群，第二次分析得到的簇B减去簇A的部分称为亚健康人群，簇B之外的部分称为疾病人群，那么可以得到一个同质性提高的簇A和一个异质性更明显的簇B，如图1-29所示。

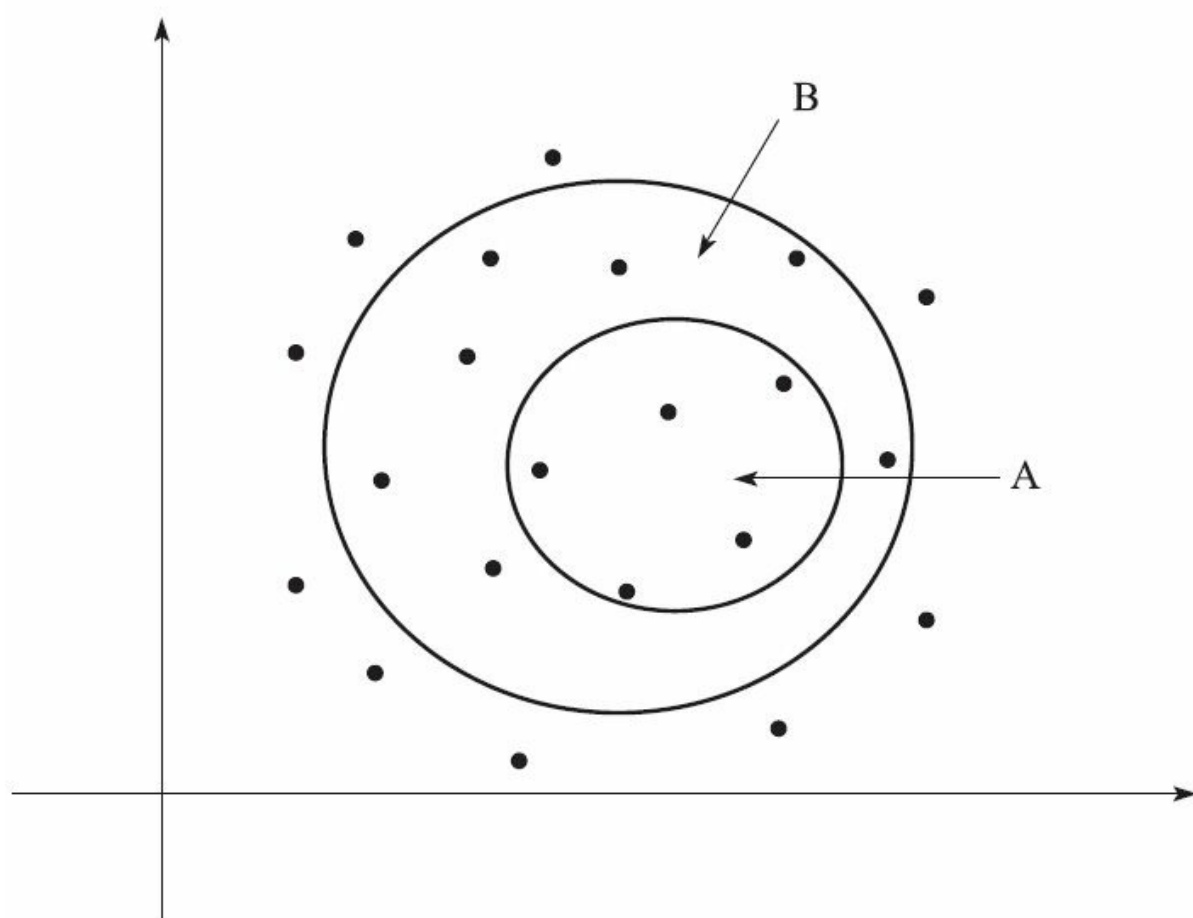


图 1-29

簇A是由各项体检指标均在正常范围的个体组成，他们的同就表现在所有指标维度（例如血压、血脂、转氨酶、心电图、X片等）均在正常范围内，这个簇或群体一般称为健康人群。簇B我们称为异常簇或非健康人群，他们的某一项或某几项指标在正常范围之外。例如

有的血压高，有的胆固醇高，有的高密度脂蛋白低，有的X片发现阴影，有的血压低等。即可能是这几项异常，也可能是那几项异常。这之间并没有同质性，从总体看该簇的异质性显著提高了。健康有一定范围，但疾病多种多样，这就像欢乐者有共同的欢乐、而悲伤者有各自的悲伤一样。

还有一种特殊情况，就是各项体检维度单独看均在正常范围，但是综合分析却有异常，在高维度下的这种异常性表现复杂，往往难以发现，而异常侦测算法将这些个体也分入异常。如图1-30所示，红圈内的个案就是每一项都正常、而整体却落在正常之外。这个例子在实践上还属于低维度案例，高维度案例有几十个上百个维度。体检行业也是在上百个检查维度中分析非健康人群。这种超高维度的分析，人脑是望尘莫及了，需要数据挖掘工具来辅助完成。

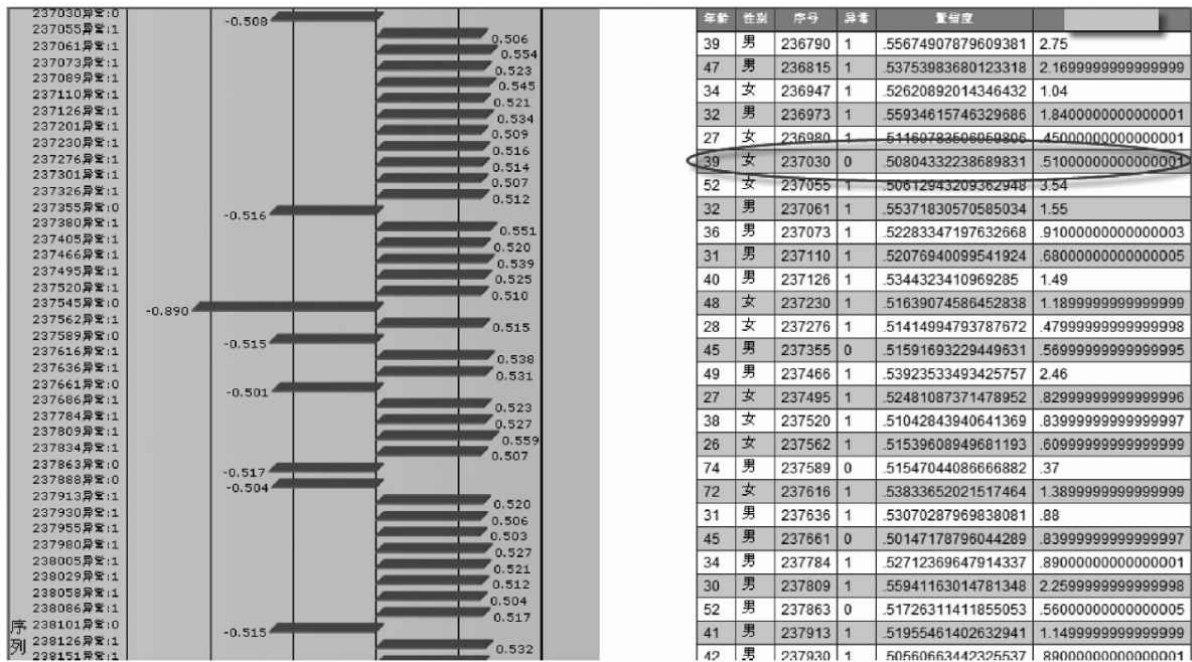


图1-30 体检疾病发现

使用正常组同质性的特点，如图1-29中的簇A，我们可以建立一个不断趋同的过程来实现同质组（图1-31）。在这个过程中，异质性的部分不断排除，剩下的部分同质性越来越高。同质性的过程在科研上可以提高观察组和对照组的均衡性。异常侦测还可以进行同质性分组的逆过程，即异质性的发现。同质性过程的逆过程正好与其相反，如图1-29中的簇B，通过建立一个不断异化的过程来实现（图1-32）。在一步一步的过程中，同质性的部分不断排除，剩下的异质性就越来越高。

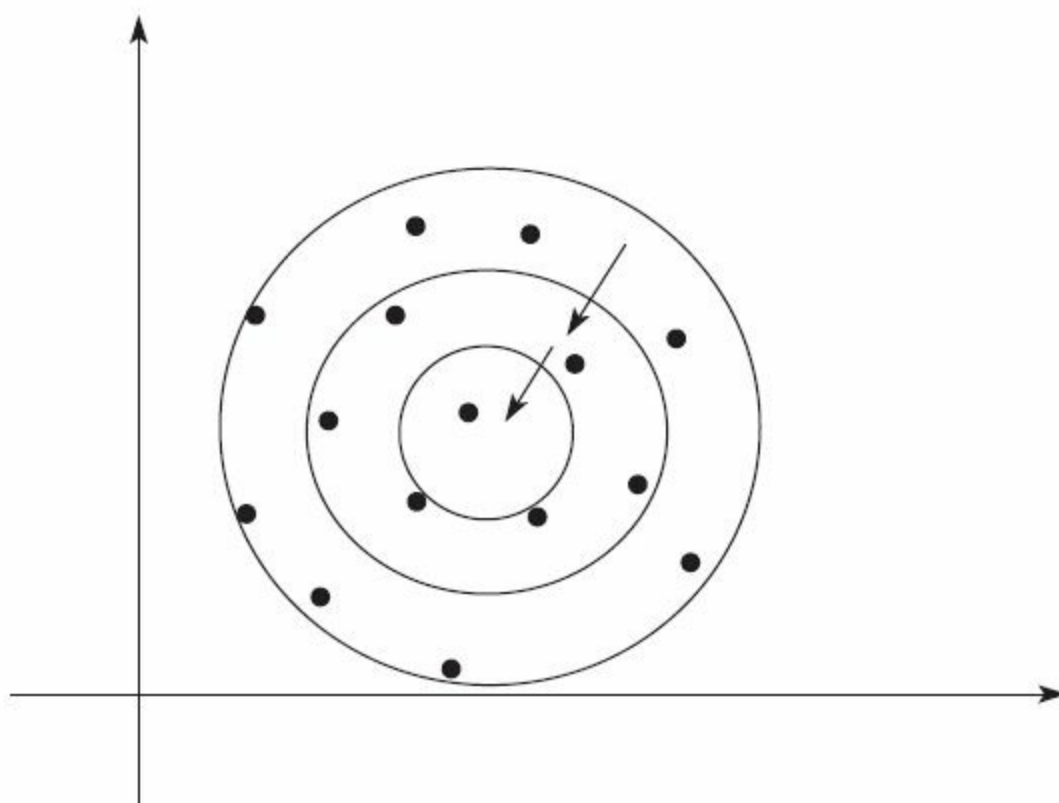


图 1-31

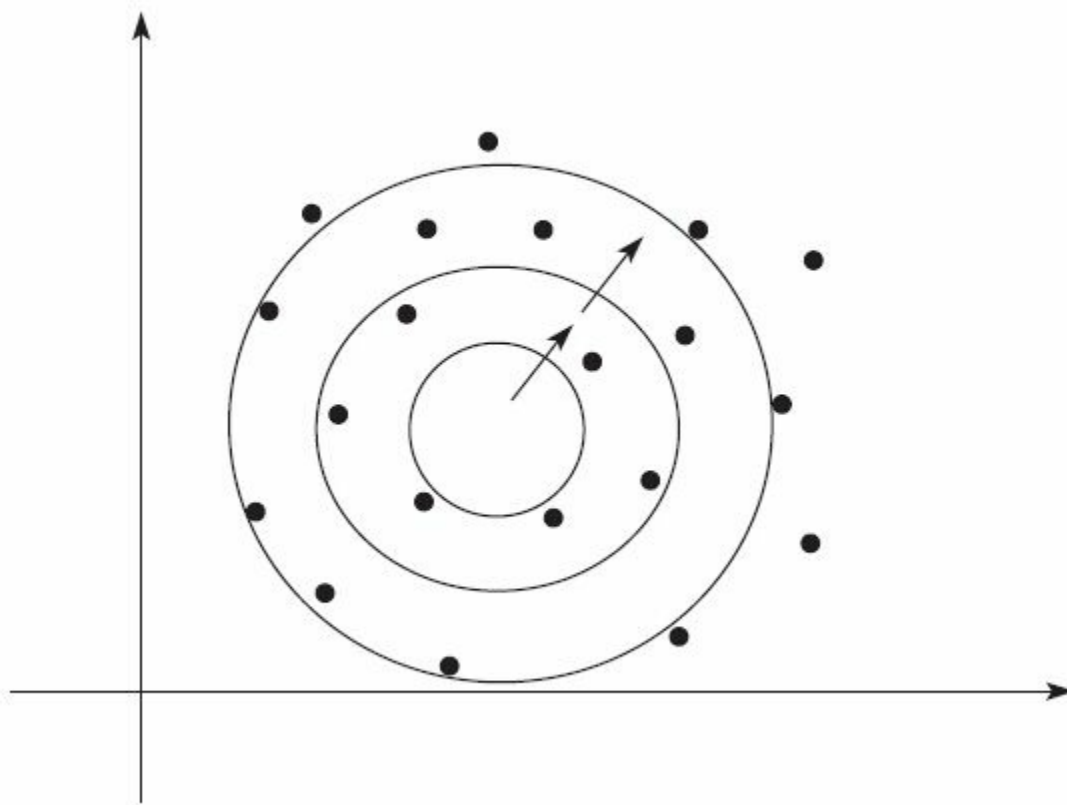


图 1-32

数据处理中常常发现一些逻辑错误，例如，医学数据中的成人身高2米、体重30公斤之类的错误。这类错误不同于数据错误，例如空值，不是很容易被发现，有时候需要行业专家来协助进行。而逐条进行这项工作，在大数据量的情况下难以实现。这些逻辑错误之所以被这样称呼，是因为它们与常识、特别是专业常识不符合。这种不符合常识有各种各样的形式，前面说的2米大汉30公斤，还有血钾20mmol每升、血压300mmhg之类，我们都可以把这些看作是一种异，不同于常识的同。常识的同之外，就是逻辑错误的异。这样可以减少发现和处理数据逻辑错误的消耗。异质性的发现用于探索特殊事物规律，例如某只股票的暴涨或暴跌，体检人群中的非健康人群等。

1.3.2.3 关系探索

关系探索是探索事物与事物间的伴随关系，以及变量（维度）与变量（维度）之间的影响关系。我们这里作为一个概念性介绍，仅简单介绍一下伴随关系的探索。探索事物与事物间的伴随关系，常见的方法有关联规则算法，用于发现多个事物是否会同时出现，以及出现的数量和可能的概率。如图1-33所示，圆圈的出现伴随着三角的出现，方形的出现伴随着星星的出现。

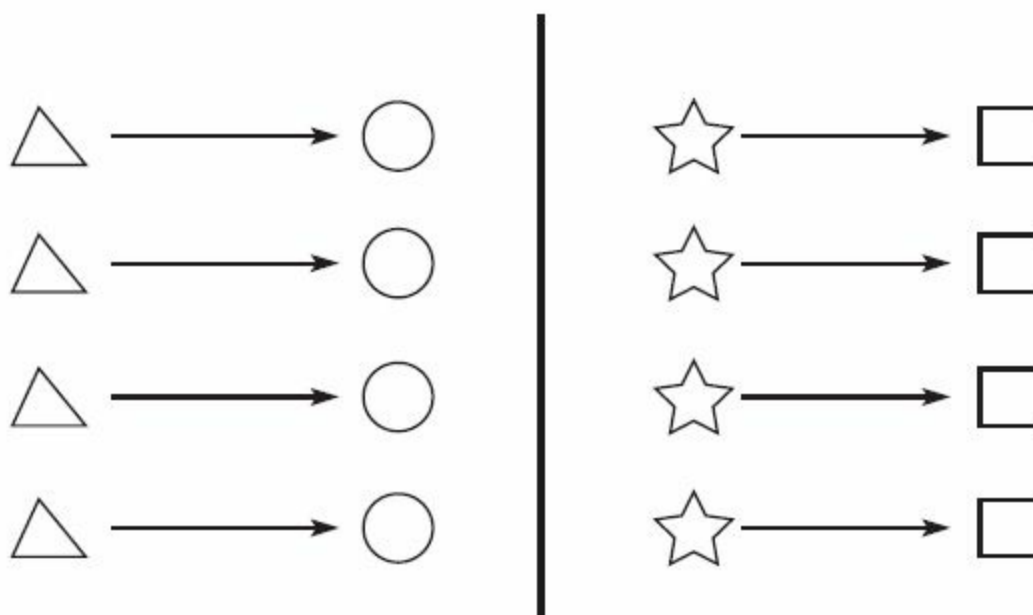


图1-33 出现联系：三角和星星

对于疾病而言，特殊事件的关联出现，也是一种关联模型的例证。例如，有病人患上了伤寒症，也同时伴随着相对缓脉症状的发生，图1-34显示了伤寒与相对缓脉的关系图。

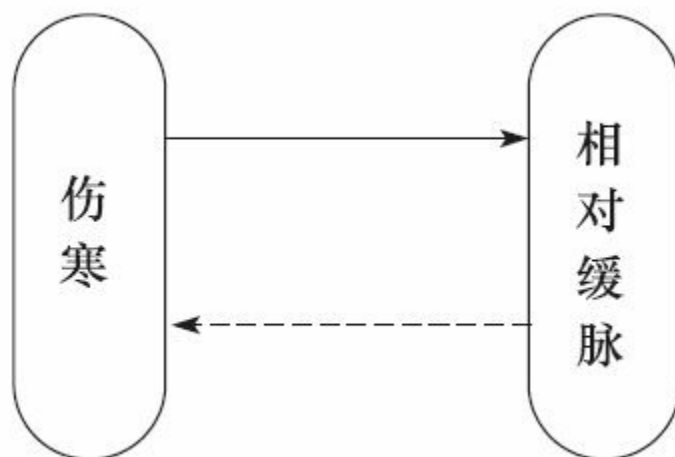


图 1-34

再如，肝外胆管结石继发胆管炎的病人会出现腹痛、寒战高热和黄疸三个症状（夏科氏三联征）先后发生，如图1-35所示，这种病人首先出现腹痛，接着出现寒战高热，最后发生黄疸。数据挖掘的关联规则算法可以发现事物的这些关联特征。

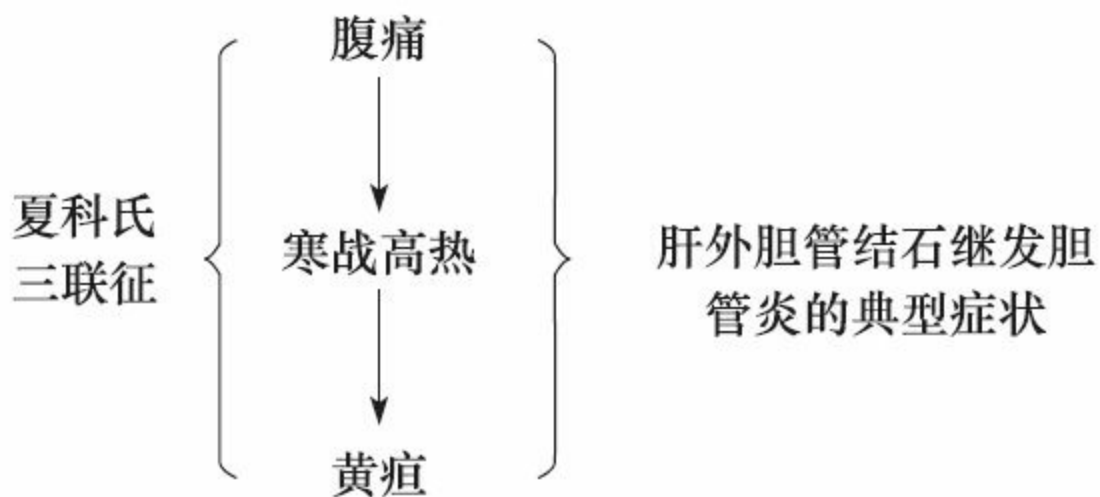


图1-35 胆囊夏科氏三联征

关联规则模型的特点，就是从大量随机发生的并发事件中，找到

强关联的现象，使得在某事件发生的前提下，另一事件的发生具有很高的概率，并且是一种具有业务意义的强规则。例如，在图1-34显示的案例中，我们就是要从大量的各种疾病症状现象中发现是否具有这样一条强规则（详见第3章），即“如果病人患上伤寒症，大多同时伴随着相对缓脉症状的发生”。同时，医学家可能也很关心：“如果病人出现了相对缓脉的症状，（再根据病人其他的临床表现），怀疑患者是否患上了伤寒症？”。再例如，在图1-35显示的案例中，我们就是要从大量的各种疾病症状现象中发现是否具有这样一条强规则，“如果病人同时出现了腹痛、寒战高热和黄疸的症状，怀疑患者是否患上了肝外胆管结石继发胆管炎？”。显然，关联规则模型对各个专业的业务分析是非常重要的。

我们对本小节提到的几个方面的探索分析算法，做个简单的小结，即：

- 分布探索主要用于事物分类，发现子类，并后期用于预测归类。
- 异常探索主要用于寻找个别分子，也可以扩展用于数据的同质性分析和异质性分析。
- 关系探索主要用于发现事物之间的伴随规则，也可以用于特征维度的发现并预测未知维度（参见后面章节的特性选择部分）。
- 另外，只要有特征就可以区分，不论是分布特征还是关系特

征。

1.3.3 应用层

数据挖掘应用是一个多层次、流程化的工程任务。我们前面简要介绍了有关数据层和算法层是如何工作的，接下来，我们需要基于前两个层面所做的工作，做好数据挖掘项目的最后一项任务，也是最重要的一项工作，就是将数据挖掘的结果进行理解，并部署和应用于业务活动中。这就是数据挖掘工程应用层的内容。

1.3.3.1 数据挖掘的专业应用

挖掘的方法和结果如何应用于实际工作，是数据挖掘应用的一个至关重要的环节。数据挖掘的方法应用和结果应用是一个数据挖掘技术和行业知识结合的过程，需要数据挖掘专家和行业领域专家共同完成。不同的行业对结果感兴趣的部分可能不同。另外，数据挖掘的实践应用不仅要考虑速度和准确性，还要考虑投入和收益。例如在实践中经常会遇到明明知道某个维度更适合划分事物，但是因为数据采集花费过高而放弃使用，退而求其次，采用另一个更具性价比的维度。下面，我们举个例子来说明如何将数据挖掘的方法应用于专业项目中。

在聚类模型的实际应用中，数据分布经常存在叠加的情况（如图1-26和图1-27所示）。处理这种叠加分布，正如前面我们谈到的，可以考虑使用两种方法，一种方法是增加维度或更换维度，使用更多的或其他的消息量，尝试从其他角度划分叠加部分。另一种方法是根据

专业知识区分叠加部分。对于第一种增加维度的方法还有两种做法，一是直接增加维度到算法中，用更高的维度划分所有数据，处理叠加部分；另一种是首先用原来的维度划分全部数据，出现分布重叠时，使用筛选机收集这部分数据，然后以新增维度为划分标准进行划分。这种双段式划分方法，在实践中早已经被采用。比如医学上，病人就诊时出现如图1-27所示的分布重叠，容易诊断的病人会形成两边的波峰，而叠加部分形成波谷。容易诊断是临床意义，它的数学意义就是被划分得清楚。不容易诊断的病人会被要求进一步检查。临床上，并不是一开始就让每个就诊患者把诊断实验检查和鉴别实验检查一起做，而是分为几个阶段：先把可以使用诊断实验检查的患者诊断了，对于剩下需要鉴别的患者（这部分患者也是做过诊断实验检查的）再进行鉴别诊断，即需要特殊的鉴别实验检查来对叠加的部分进行区分。所以，第一步划分不但是划分也是数据的纯化，以纯化数据为基础，第二次划分才有意义。聚类的双段式划分方法，就可以应用到类似本例病人诊断划分的实际项目中。更进一步还有多段式划分方法的应用，与此类似，可以处理更细微的叠加分布。

聚类一般被认为是划分算法，虽然它有时也可以用于其他方面，如用于连续变量离散化和异常侦测等。现在，我们看一个年龄和性别的双维度聚类的例子（详细的分析我们将在后面的章节中介绍）。在我们这个例子里，聚类模型是两个维度。聚类模型也可能是更多维度，但是一般不会高于五个维度，因为理解过高维度的结果是非常困难的。如图1-36所示，年龄和性别双维度聚类模型输出的结果为一个

树状结构，我们可以很清楚地看到划分的过程。需要注意的是，两个维度对每一次划分的影响是不同的，有的划分簇中男女没有区分开（簇中男女均有），年龄被划分开了；有的划分则相反，性别对划分的影响更大。对于特定的数据分布，虽然最后簇的划分是一定的，但是划分的过程仍然能告诉我们很多额外的消息。

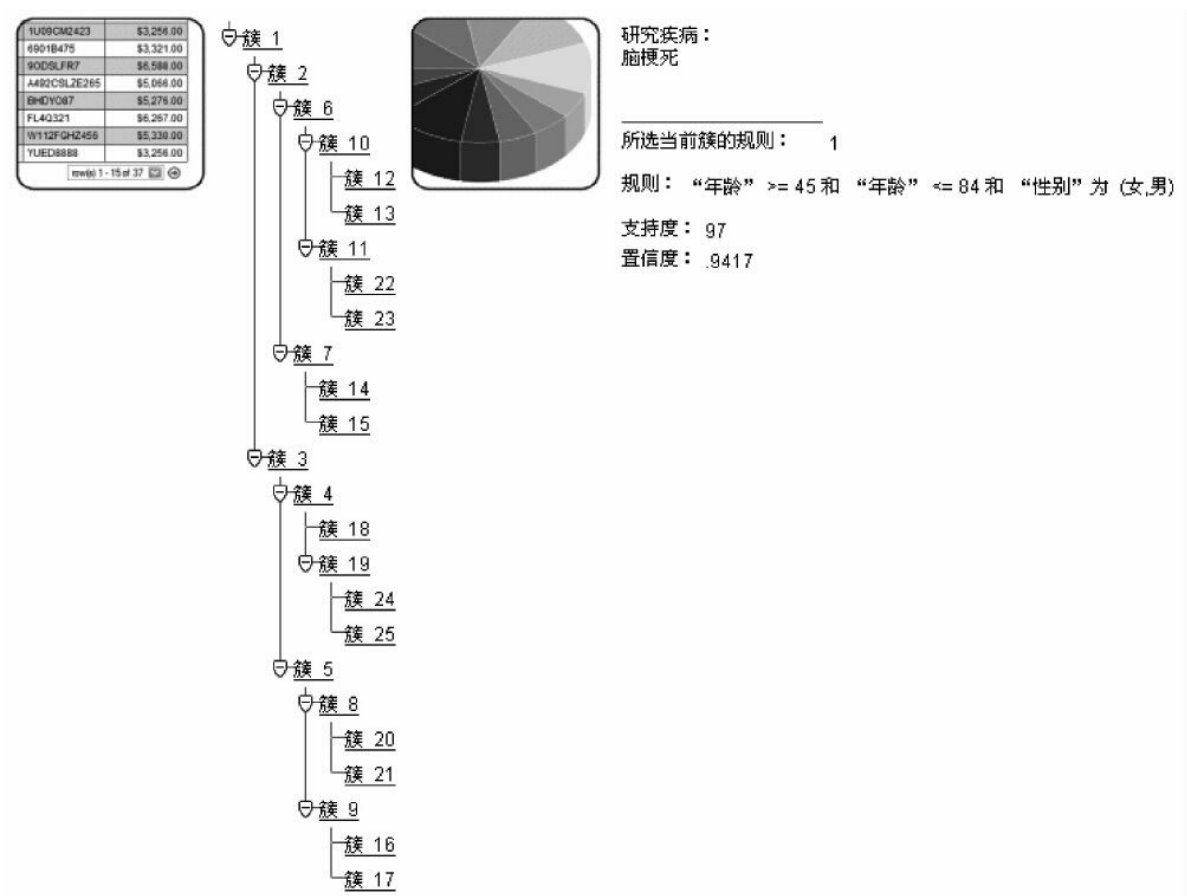


图 1-36

观察和分析每次划分中各个维度对划分的影响，是对聚类输出结果的理解和使用的重要部分。聚类模型的结果主要体现在两个方面：簇的划分和维度的影响，它们受到算法、数据和业务三个方面的影响。以聚类而言，在业务上，并不是每个簇都有专业意义，这样就需

要根据业务来合并簇，特别要注意根据业务来合并簇而不是按照数据的分布情况来合并簇。这也从一个侧面说明，聚类模型输出结果的形式很重要。在低维度聚类引擎上，树状结构在聚类模型的输出上有独特的优势，可方便观察每一次划分和不同维度对划分的影响。

算法处理数据，而输出结果决定了如何被使用。我们既要理解算法，也要精通输出结果的解释。在实践工程或项目上更是如此，对数据挖掘的输出结果的解释和理解最为重要。

1.3.3.2 面向行业的应用产品

在数据挖掘应用工程项目中，我们有时也会遇到一些半工程化、半产品化的工作。因此，我们这里也简要谈一谈有关面向行业的应用产品的话题。

数据挖掘应用主要有三种形式：通用工具（参见第8章）、专业应用系统（参见第6、7章）和定制化工程（参见第4章）。目前定制化工程是最为多见的形式，通用工具也不少，但专业应用系统不太多。定制化工程具有很强的针对性和目的性，通用工具则以通用性为特点，不针对专业用户设计，专业应用系统则介于两者之间。从数据挖掘的用途来看，数据挖掘可以挖掘、探索出通用性的知识，而数据挖掘的对象却有着很强的行业性和领域性，专业应用系统和定制化工程可根据数据特点做算法优化，相对于通用工具在专业应用上更具有优势。从最终用户角度看，通用工具的门槛最高，使用者对数据挖掘

的原理和各种功能的了解直接决定了挖掘结果的好坏。专业应用系统和定制化工程门槛较低，引擎的采用使得数据挖掘技术对于使用者而言透明化。因为很多最终用户不精通数据挖掘技术，他们可以专心处理本领域的专业事务，专业应用系统和定制化工程能帮助用户解决技术性问题。专业应用系统相对于定制化工程更具有较宽的适用性和通用性，它是数据挖掘技术和用户的桥梁，把复杂隐藏在简单的后面，使用者可以更加专注于行业需求，可以了解原理而不是执着于过程。用户使用专业应用系统不需要从基础学起。当然，知其所以然会用得更好。

专业应用系统是数据挖掘从专家走向大众的桥梁，引擎是数据挖掘产品化的基础。数据挖掘的广泛应用需要经由产品化过程。计算产品从商用计算机到个人电脑、再到智能终端的发展也证明了这一点。数据挖掘的应用也是要从写算法走向产品发展。产品化是技术的趋势，一项技术能否被大多数人使用，方便大多数人使用，是它能否发展的重要因素。如同使用手机的人，谁真正会制造手机、真正了解通讯协议？数据挖掘也需要把用户从写算法中解放出来，使用户把更多的精力集中在行业应用上。

一种新事物如果具有真实的生命力和长远的影响力，那它一定不是全新的，而是隐含带有旧的特性。它是旧的新事物，而不是新的新事物。如同电子商务，电子是新的，而商务确是旧的。数据挖掘其实不算是一种行业，至少不完全是。而是一种渗透和影响到许多行业的技术和应用。数据挖掘早晚也会和电子商务、移动互联网一样，转变

为几乎所有行业都能利用的一种手段、一种工具。

1.4 数据挖掘应用体会

在这一节里，我们再简要谈一下有关数据挖掘应用方法方面的内容。我们团队多年从事数据挖掘方面的商业应用，其中经历了许多大大小小的项目。我们有过不少比较成功的案例，遇到过很多的困难，也有一些失败的教训。总结起来有不少体会，就与数据挖掘商业应用关系比较紧密的一些方面，与大家讨论并分享一下。

首先说明一下，在本书中提到的数据挖掘应用的含义是什么。我们讲的数据挖掘应用也称为数据挖掘的商务应用或商业数据挖掘应用，是指在各行各业中用于直接或间接产生经济效益或社会效益的、24小时（日常性或经常性）业务依赖的、长期运行稳定可靠的数据挖掘项目，不包括用于研究、学术、教学、探索为目的的实验性数据挖掘项目，这部分项目我们称之为非商业数据挖掘应用。即使有些数据挖掘应用从理论和技术性上证明都是不错的，但由于目前尚未有一个商业项目在应用它，或有商业项目在试用但尚未足以证明直接或间接产生了经济效益或社会效益的，我们都不称它们为商业数据挖掘应用。

我们的体会主要体现在如下几个方面。

1.4.1 项目关键点

数据挖掘应用项目的实施者，通常是一个包含应用业务人员和数据挖掘技术人员在内的完整的团队。我们说解决数据挖掘应用项目的关键，是在深入了解业务或专业知识的背景下，应用适当的数据挖掘技术解决相关专业问题，满足业务提出的要求。其中的难点包括：

- 能够应用足够的、系统的、实践成熟的数据挖掘技术。
- 对行业背景、业务规则、专业知识以及项目需求有深入的了解。
- 将以上两点真正结合起来，应用适宜的数据挖掘技术解决相关业务问题。

我们先谈谈第一条，数据挖掘项目中数据挖掘技术的重要性。首先，足够的数据挖掘技术是指拥有更多、更丰富的方法和手段，包括数据挖掘工程中的各个环节，如数据分析阶段、数据整合阶段、算法分析阶段和结果评估阶段等，各阶段所需的相关技术和方法。其次，系统性的数据挖掘技术是指全面地、完整地施加在数据挖掘项目的各个环节上的不同技术，而不是不少人认为重要的技术就是算法，甚至是唯算法论。系统性的技术应包括数据描绘技术（Data Profiling）、数据质量控制技术（Data Quality）、数据建模技术（Data Modeling）、数据整合技术（ETL）、算法技术（Mining Algorithm）和评估技术（Mining Evaluation）等。最后，实践成熟的技术是指大

多经过实践（商业应用项目）检验的数据挖掘技术，而不是停留在实验室中的技术。这样，项目成功的可靠性会更大一些。当然数据挖掘本身就是一个不断创新的学科，我们会尝试更多的新技术和新方法，但部署在商业项目之前，应尽量多做一些试验和验证工作。

接下来我们再谈一谈第二条，数据挖掘项目中对业务背景和知识了解的重要性。因为数据挖掘是揭示人文事物或自然事物发生、发展所蕴含的规律，因此对事务本身的深入了解是必需的，否则肯定不成功。对业务的深入了解，就需要实施团队中有专家级的业务人员，同时技术专家对业务知识了解越深入越好。我们在加拿大某大电信公司经历了一个项目。项目完成时，经测试数据挖掘模型运行良好，结果满足业务要求。但在试运行一段时间后，发现模型结果不稳定，造成业务部门不能可靠地依赖挖掘模型提供的数据。后来经全面分析，还是发现在项目的业务分析环节有不准确的地方！后来，团队特意从欧洲请来了两位对电信业务了解深厚的数据挖掘专家，重新分析了业务，并对后续的环节加以改进，最终取得了圆满的成功。这个例子说明了数据挖掘项目中对业务背景和知识的了解有多么的重要！

最后我们再谈谈第三条，将数据挖掘技术与业务的需求结合起来是最困难的一点，也是最重要的一点。做好商业项目，首先技术专家和业务专家需要相互学习一些对方的专业知识，以便业务专家和技术专家从各自的角度，能够知道现有技术在这个业务中能做些什么。这样才可能最大化数据挖掘在项目中的效益。然后，双方才能共同确定数据挖掘项目的目标。一个科学的挖掘目标是商业应用项目成功的关键。

键。最后，在项目的整个实施过程，必然需要技术专家与业务专家的通力配合，尤其是在挖掘结果的解释、评估和应用环节更是如此。

1.4.2 技术与应用创新

数据挖掘是一项具有挑战性的、探索性的和需要不断创新的学科与技术。基于我们的经验和认识，我们认为数据挖掘思想不能被限制和约束，不断探索与创新是在数据不断增长、需求不断增长、挑战不断增长的环境下的数据挖掘的本职任务。我们不要在以下两方面受到传统思想的束缚，而是要勇敢不断地探索、创新和突破。

1.现有理论和技术方法

我们知道，经典的有关数据挖掘的原理包括人工智能、机器学习、模式识别、模糊数学和数理统计等，但可能我们还可借鉴更多的学科知识应用于数据挖掘。吸纳现代科学的新技术，借用古代科学的老技术，开创数据挖掘新的原理和新的理论。另一方面，也可以通过不断增长的各种业务需求来寻找和探索新理论和新方法。发现新想法的最佳来源，并非仅仅是那些名字里有“数据库”、“算法”的理论领域，更多的来自于其他领域，来自于实践的需要。与其到数据挖掘内部的理论领域寻找创意，还不如在数据挖掘外部的专业应用领域中寻找创意。

2.现有的应用经验

目前，数据挖掘的理论和技術较数据挖掘的应用先行一步。从一个较高的视角来看，放眼国际，虽然近年来人们在数据挖掘应用方面

取得一些不错的成果，但相对于所谓大数据爆炸式的增长和各行各业包括信息技术在内的各项技术的快速发展，以及由业务产生的源源不断的新需求，数据挖掘现有的应用也仅仅是爬到了数据挖掘应用高峰的山脚，我们的实践经验还很有限，远远不能满足社会发展所提出的需求。因此，我们不能只是应用现有的经验去解决不断地涌现的业务问题，需要加大投入，加快步伐，不断地摸索数据挖掘在新领域的应用，不断积累数据挖掘在各行各业上的新的应用经验。

1.4.3 经验积累与应用

数据挖掘需要大量的理论知识与技术方法，还需要很多技巧与智慧，它既是勤奋者的沃土，更是天才们的宝藏。我们将知识、勤奋与智慧结合起来，在实践中不断探索，经历成功和失败，并不断加以总结、积累和升华，就形成了大量商业应用的宝贵经验，而这些经验就是我们打开奥秘大门的利器，是我们走出新迷宫的导航仪。木桶原理是由美国管理学家彼得提出的，如图1-37所示。说的是由多块木板构成的水桶，其价值在于其盛水量的多少，但决定水桶盛水量多少的关键因素不是其最长的板块，而是其最短的板块。我们可以把知识、勤奋、智慧、经验看作是构成数据挖掘应用成果水桶的四块木板！



图1-37 木桶原理

数据挖掘可能是十分复杂的一项工作，但有时也有很多技巧，就像一层窗户纸一样一捅就破。在我们经历的项目中，很多是借鉴过去相似的项目或跨行业的并不相似的项目的成功经验，有时需要积淀，有时需要灵感。我们很多数据挖掘项目的初级探索者往往收获的失败更多，其中一个重要原因是缺乏足够的应用经验，而经验有时又像一道不可逾越的鸿沟，成为人们前进的拦路虎。在此，我们也愿意把更多的经验，也包括失败的经验，写出来与大家分享，希望至少让大家少走弯路，这就是这本书的立意所在。

1.5 无限三维嵌套空间假说

作为准备数据挖掘的预备知识之一，本节将介绍一些有关高维空间的知识，以帮助读者更好地理解后续章节的一些内容。数据挖掘中，有很多需要处理高维数据的情况。人们在理解和解决高维度问题时经常遇到一些困扰和瓶颈，尤其是在利用一些数学算法处理高维空间下的问题时。为了帮助读者理解高维空间下数据挖掘方法，我们这里用一些简单、形象的可视化图形方法来描述高维空间。

1.5.1 一维空间

一维空间就是一条由无数个点组成的无限长的直线 X ，如果以 0 点作为原点，则直线空间有两个方向，如图1-38所示，它的空间坐标为图中的 X 轴：

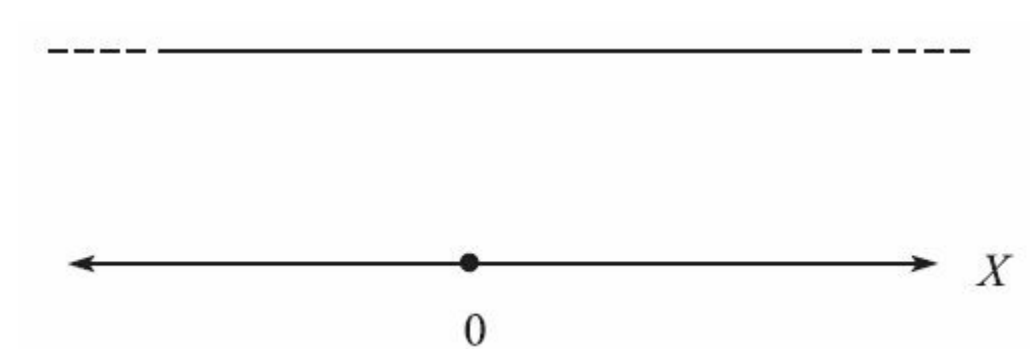


图 1-38

1.5.2 二维空间

二维空间是由无数个一维空间直线组成的。二维空间就是一个无限伸展的平面S，如图1-39所示。如果以 $(0, 0)$ 点为原点的话，则平面空间有四个方向，它的空间坐标图为1-40所示。

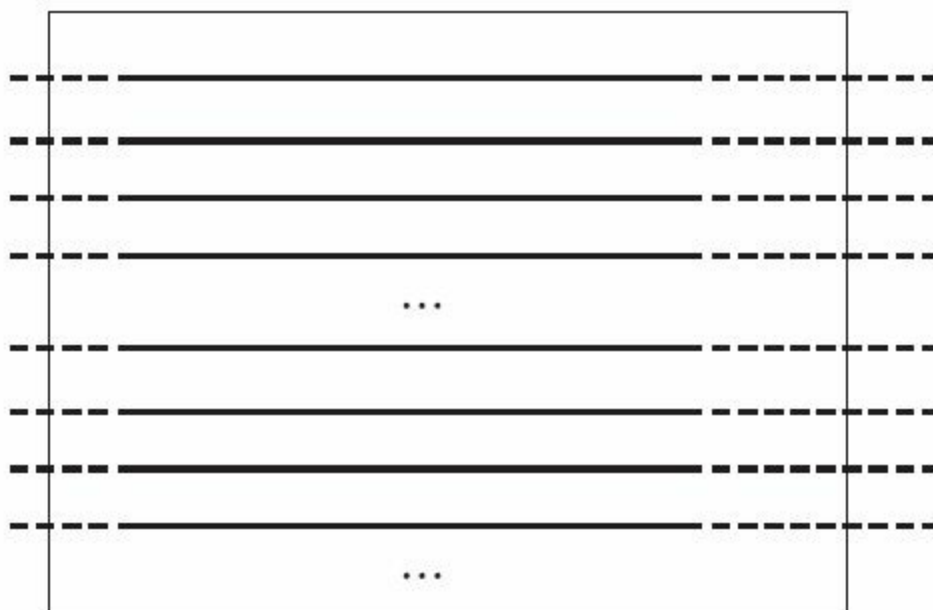


图 1-39

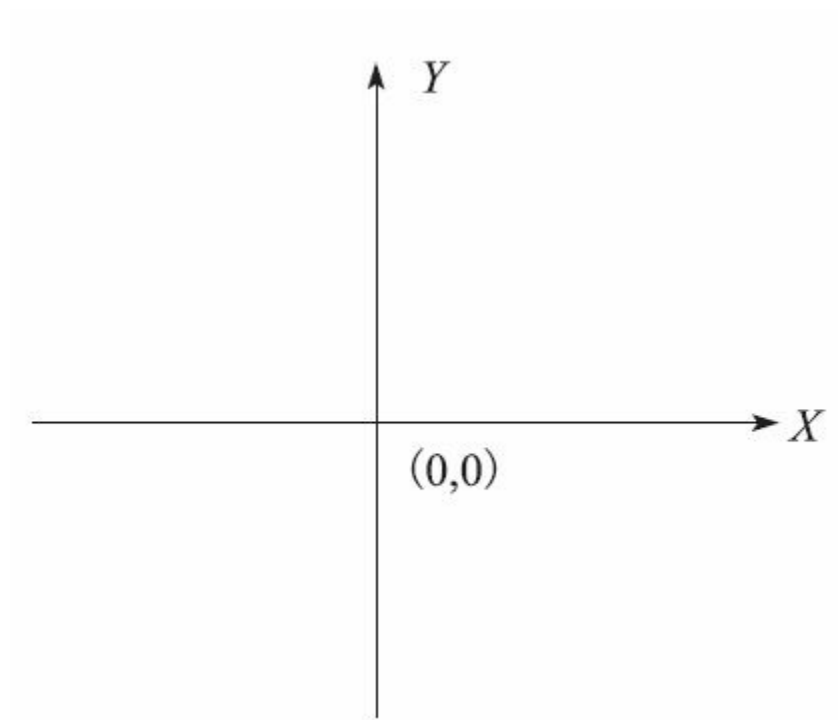


图 1-40

1.5.3 三维空间

三维空间是由无数个二维空间平面组成的。三维空间是一个无限扩展的立方体C，如图1-41所示。如果以 $(0, 0, 0)$ 为原点的话，则立方空间有六个方向，它的空间坐标为图1-42所示。

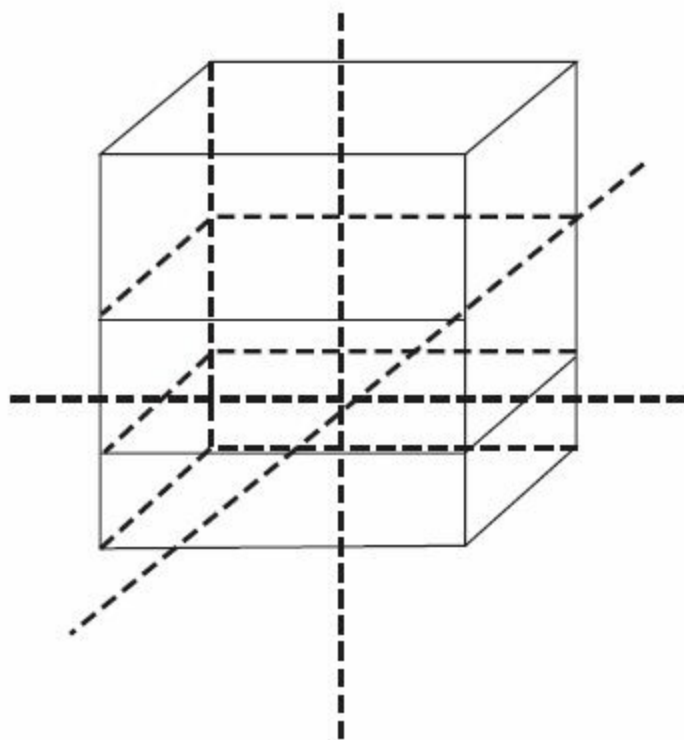


图 1-41

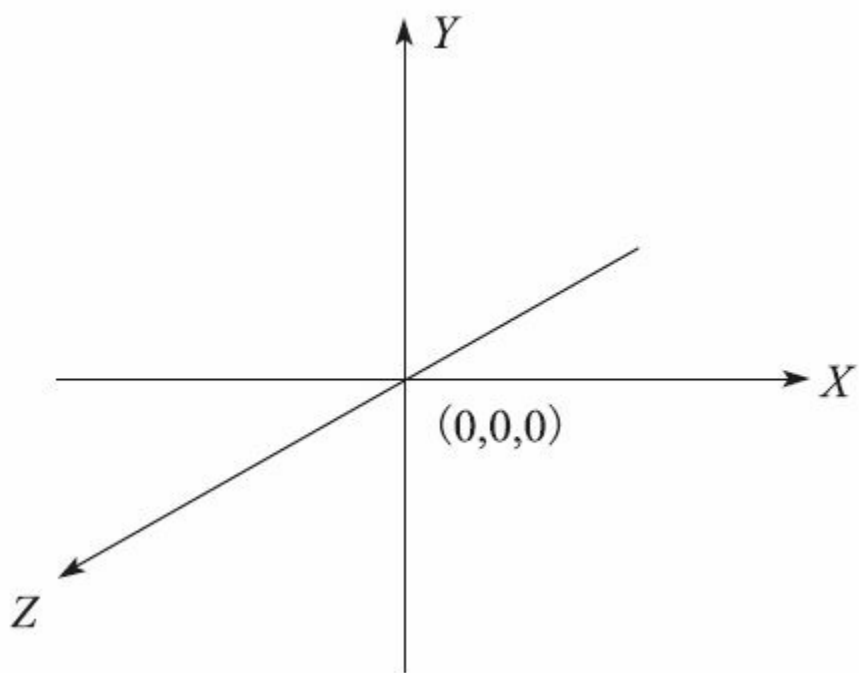


图 1-42

1.5.4 突破三维空间

什么是四维空间呢？爱因斯坦的四维空间是三维空间加时间，这是一个物理世界的四维空间。在数据挖掘中，更多用到的是数学世界的抽象空间。我们看到，由无数条直线（一维空间）可以组成一个平面（二维空间），由无数个平面（二维空间）可以组成一个立方体（三维空间）。那么，由无数个立方体如何形成一个四维空间呢？我们可以先把图1-41压缩成较薄的一个立方体，但仍保留着原先的三维坐标（图1-42），然后将无数个压薄的立方体倾斜着叠加起来，就生成了一个四维空间，如图1-43所示（注意，这只是个示意图），它的四维坐标为图1-44。

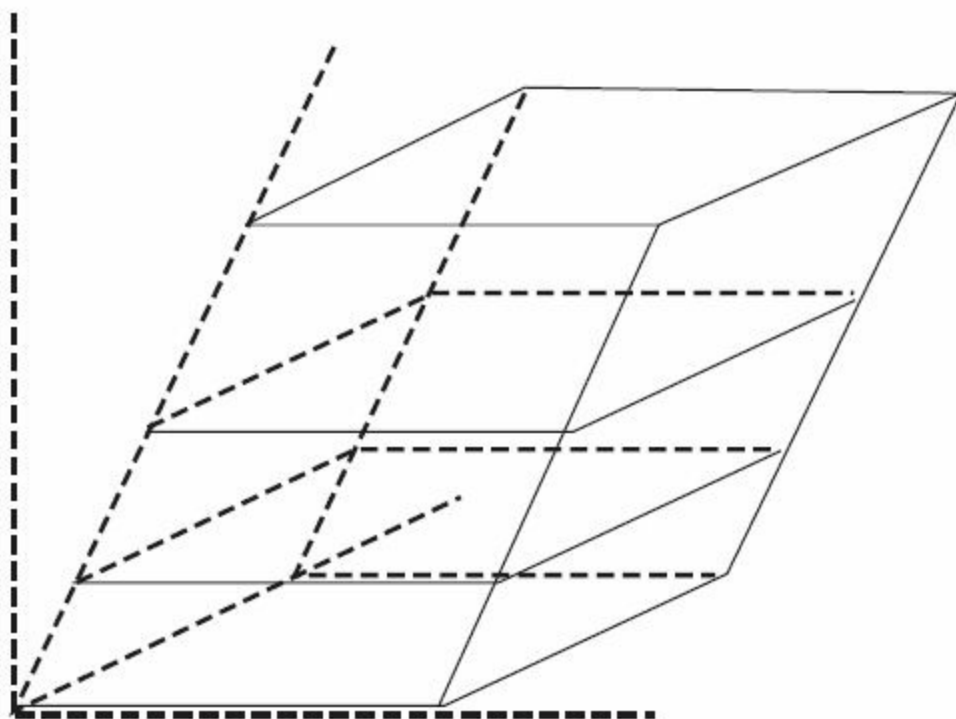


图 1-43

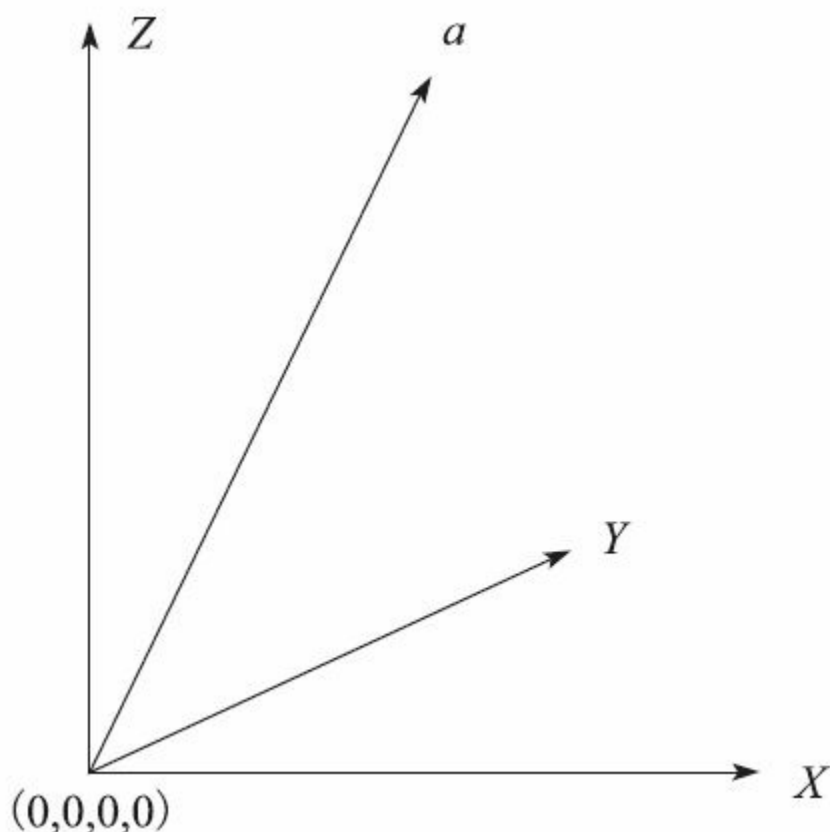


图 1-44

如果需要在四维空间中确立一个点，除了需要确定 x 、 y 、 z 三个维度的值外，还需要确定第四维度 a 上的值。我们可以把四维空间想象成‘串’在 a 轴上的无数个三维空间，其中位于 $a=0$ 的点上就有一个三维空间。我们生活中有一个实例可以更形象的表达这个四维空间。我们吃的一种由山楂制作的小吃叫糖葫芦，糖葫芦上的竹签子就像四维空间的 a 轴，每一个山楂就像一个独立的三维空间，有它自己的坐标系，四维空间就是由一条无限长的竹签子串起来的无数个山楂体，如图1-45所示。

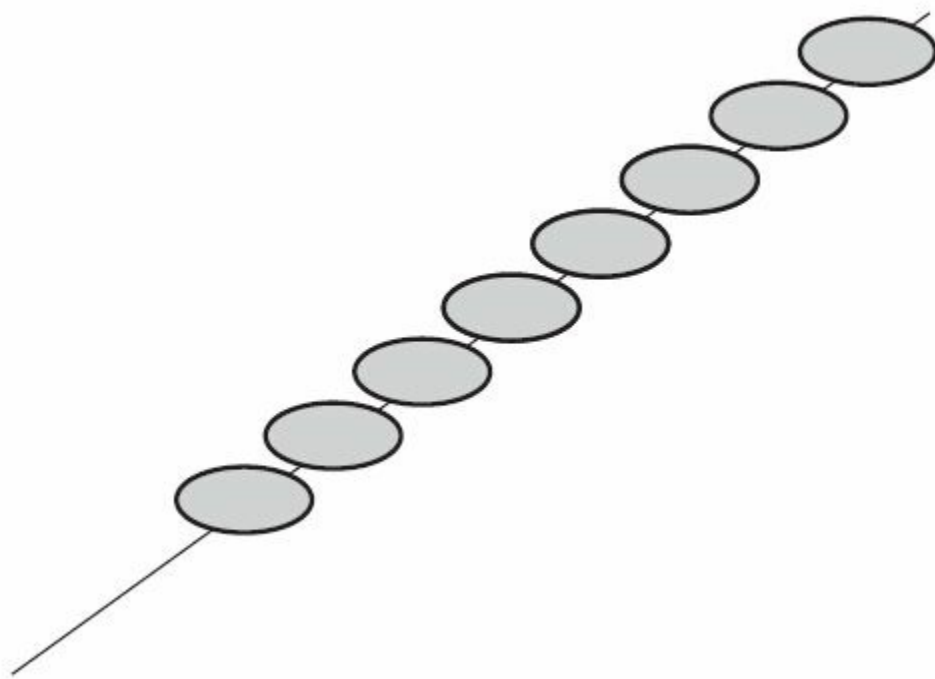


图 1-45

对于人们来说，突破三维空间的想象的确有些困难。我们可以这样想象，有一个四维空间直线“穿透”整个无限三维空间，该直线与无限三维空间当且仅当有一个交点，但四维空间直线除了交点外并不存在于这个无限三维空间中，我们看不见它是因为它存在于我们能见三维空间之外。假设在我们可见的三维空间中有一个点，如果它属于一个四维空间的点，则该三维空间点还要受我们看不见的四维空间点的约束，如果该三维空间点的四维空间取值不同，则该三维空间点可能对应着许多不同的四维空间点；反过来说，许多不同的四维空间点可能对应着（映射到、投影到）同一个我们可见的三维空间点。

1.5.5 五维空间

对于五维空间，同理，我们可以把很多糖葫芦串平摆在桌面上，就形成了一个五维空间，如图1-46所示。所以，五维空间就像是由无数个无限长的糖葫芦串（四维空间）构成的一个糖葫芦面空间。

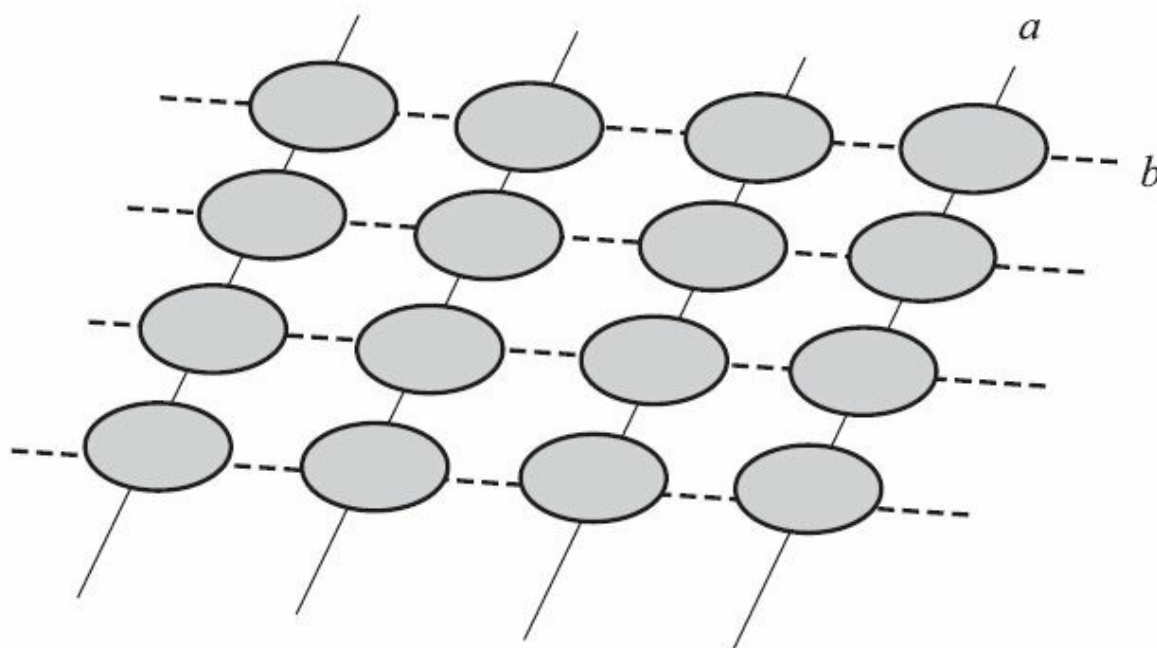


图 1-46

我们可以使用与1.5.4类似的方法来想象五维空间。我们可以这样想象，有一个由第四维和第五维空间组成的平面（我们称为超平面）“穿透”整个无限三维空间，但在这个无限三维空间中并不存在着这个超平面。如果我们需要将五维空间的一个点映射到我们可见的三维空间中，则我们需要投影两次，第一次将该点从五维空间投影到四维空间，第二次再将该点从四维空间投影到三维空间，我们就可见了。超平面的概念和应用在数据挖掘中普遍存在，例如，我们后面讲

到的一个分类算法——支持向量机的原理就利用超平面去划分两类数据，又例如，我们刚讲到的那个由第四维和第五维空间组成的超平面可以将一个六维空间数据集（拥有六个属性的数据）划分成两类。

1.5.6 六维空间

六维空间，以此类推，我们把许多糖葫芦面叠加在一起就形成了一个糖葫芦立方，如图1-47所示。我们可以发现六维空间由六个维度组成 a 、 b 、 c 、 x 、 y 、 z 。在图1-47中，每一个糖葫芦就好像是 a 、 b 、 c 三维空间中的一个点，而这个点本身又是一个由 x 、 y 、 z 三个维度组成的空间，我们把这种结构形象地称为三维嵌套空间。

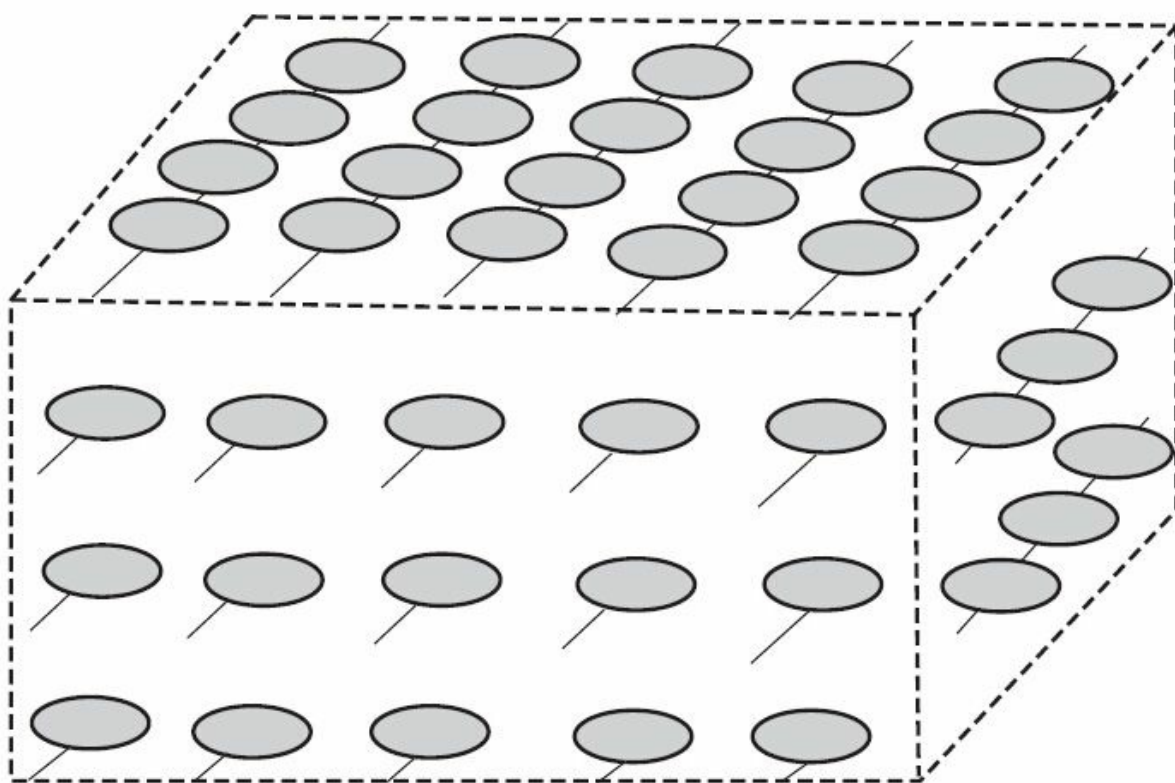


图 1-47

如果将 x 、 y 、 z 三个维度组成的空间作为原点空间层（每三维空间作为一个空间层），则由 x 、 y 、 z 三个维度组成的空间称为原点空间层的一层嵌套空间。对于六维以上的空间，我们仍可应用前面类似

的方法，构成六维以上的任意维度空间，形成**超高维空间**。但在这个超高维空间中，其结构均是每三维空间嵌套在更高维空间中。如果是无限维空间，我们可以想象为无限三维嵌套空间，同时也就存在着原点空间层的二层嵌套空间、三层嵌套空间.....无限层嵌套空间。另外，如果设定 x 、 y 、 z 三个维度组成的空间为原点空间层，则空间嵌套的方向是双向的，即原点空间层外嵌套着无限个更高维度的空间，同时，原点空间层内又包含着（反向嵌套）着无限个更低维度的空间。这时因为，在极点空间层中的任意一维空间的无限长直线是由无数个连续的点组成的，我们把其中的任意一个点称为零维空间，而零维空间内包含着（反向嵌套）着无限个更低维度的空间，如负一维空间、负二维空间、负三维空间等，而且其结构也是无限三维嵌套空间。我们把以上的高维空间描述称为“无限三维嵌套空间假说”，或通俗地称为“糖葫芦高维空间假说”，有待科学家去证明。但在本书中，可以暂且作为帮助读者理解数据挖掘的一个方法，总比没有办法想象出高维空间的情况要好吧？

以上的维度描述方法可以帮助我们想象和理解对高维空间数据的处理过程。但读者也可以充分发挥自己的想象力，以便能够找到更好的可视化高维空间的方法，并与大家共享。

1.6 本章小结

本章讲述了数据挖掘的一些基础方面的内容。从日常工作中，我们体会到数据挖掘的博大精深，正如挖掘宝藏一样，挖掘得越深，发现的现象就越多，未知的东西也越多，但是，我们的求知欲也就越强，也更加着迷于这项极富挑战 and 诱惑的工作。数据挖掘就是一项探索性质的工作，它不仅是一种探险任务，更需要的是智慧、经验和坚持。我们希望读者通过本章的阅读，对数据挖掘有个更贴近、更鲜活的认识，也希望引起更多人的兴趣与关注，并共同投入到数据挖掘的实践探索中。众人拾柴火焰高，我们希望越来越多的智囊和精英加入到数据挖掘的事业中来，也希望数据挖掘的理念和用途能为更多人了解，即使这些人不直接从事数据挖掘，但是可以引入数据挖掘并享受它所带来的收益。这样，走进生活并深入生活的数据挖掘会迎来一个百花争鸣、繁荣发展的时期。

第2章 数据探索与准备

本章节主要谈一下数据挖掘应用活动的前期工作，即数据探索的工作和数据准备工作。数据挖掘应用是一个系统工程，它包括算法应用在内的很多环节和步骤，仅仅会正确使用算法还是不足够的，在进入算法层面之前，首先需要处理好数据层面的工作，我们有大量的重要的探索性工作要做。数据层面的准备工作是算法层工作和应用层工作的基础，我们应该扎实做好数据层工作，不要使算法层、应用层如同建立在沙滩上的大厦，随时有倾覆的风险。

本章重点涉及三个方面：了解数据、选取数据、处理数据。

1) 了解数据。

- 了解数据的产生，是建立在对所分析业务的了解之上的，也就是对数据来源的分析和了解，及对先验知识的收集、学习、整理。这一步工作处理良好，就为后面分析解读数据打下了基础。

- 了解数据的关系和质量，是从数据技术的角度去了解的，包括数据的整体模型、数据之间的关系、数据的质量等。能否用数据的语言来准确地表达业务的关系和知识，也是数据挖掘项目成败的一个关键因素。

- 了解数据的特征，是应用一些数学工具对数据的基本特征有深入的了解。为了给数据挖掘进一步做好技术性的准备工作，**我们需要**

了解数据的分布特征以及更深入的数据质量问题，以便采用更恰当的方法，按照数据挖掘的目标，对数据进行处理和准备。

2) 选取数据。

- 数据挖掘是基于对大量数据的分析的，但商业项目既考虑数据挖掘的结果，又要考虑数据挖掘的成本。很多时候，需要从超大的数据集中选取有代表性的数据进行分析。因此，选择具有代表性的数据是一门技术。

- 数据挖掘算法中包含很多统计学的方法，这些统计学方法往往有一定的适用条件，如对数据量的要求。因此，选取适当的数据量以满足相应统计学算法的要求也是数据准备工作的一个重要环节。

3) 处理数据。

- 通过解析、理解业务系统的数据模型，深入分析数据关系模式，并重构分析型数据模型，应用数据整合技术构建一个完整的、一致的、有效的、用于分析的数据中心。

- 经选择的数据，通常并不能直接用于数据挖掘算法，而是需要进行很多处理，如数据标准化、数据离散化、数据质量的进一步处理等，以满足相应数据挖掘算法的要求。

另外，在第2.6节，我们还将介绍一些数据挖掘初探性应用，即利用某些统计学方法和商务智能（BI）工具对数据进行探索的应用，

以帮助用户快速了解数据的主要特征和规律，为今后的数据挖掘分析做好准备工作。

2.1 数据关系探索

要了解数据，需要从数据关系探索开始，主要包括两个方面，一个是从业务上了解，一个是从技术上了解，二者是密不可分的。数据关系探索通常从元数据分析着手。元数据从技术上讲，是描述数据的数据，就是类似我们通俗讲的数据结构的概念，所以元数据分析我们也称为数据结构分析。数据结构是描述数据源的信息，随着数据结构的分析，我们可以制订提高数据质量和数据整合最有效的策略和方案。通过元数据的分析，我们可以通过整合企业所有的数据源，组织有效地用于分析的数据逻辑模型；通过精确的分组相关数据和排除无关的数据，可以简化数据分析项目。元数据分析可以让我们确切地理解所有的企业数据源中的信息性质，并成功地组织它们，以确保所有的数据质量及项目的顺利进行。通过元数据的分析，我们的最终目的是要保证数据质量、数据整合和数据分析各个环节的顺利执行。

本节将先后介绍业务发现、关系发现、数据质量探索，以及数据整合等方面的内容。

2.1.1 业务发现

有关业务数据分析，主要考察两个方面：完整性和可信性。数据完整性是指需要研究的事物各维度的涵盖内容是不是足够反映事物的主要特点。例如，我们对医院的数据完整性进行分析。在实践中，完整性分析是从业务链开始的。图2-1是用药的业务流程，是医院业务链的一部分。

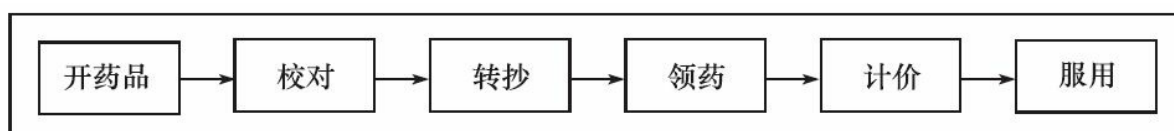


图 2-1

业务链的每个环节都是整个业务系统的重要组成部分，也是业务数据产生的源泉。业务链产生的数据进入数据库，就形成了数据链，如图2-2所示。

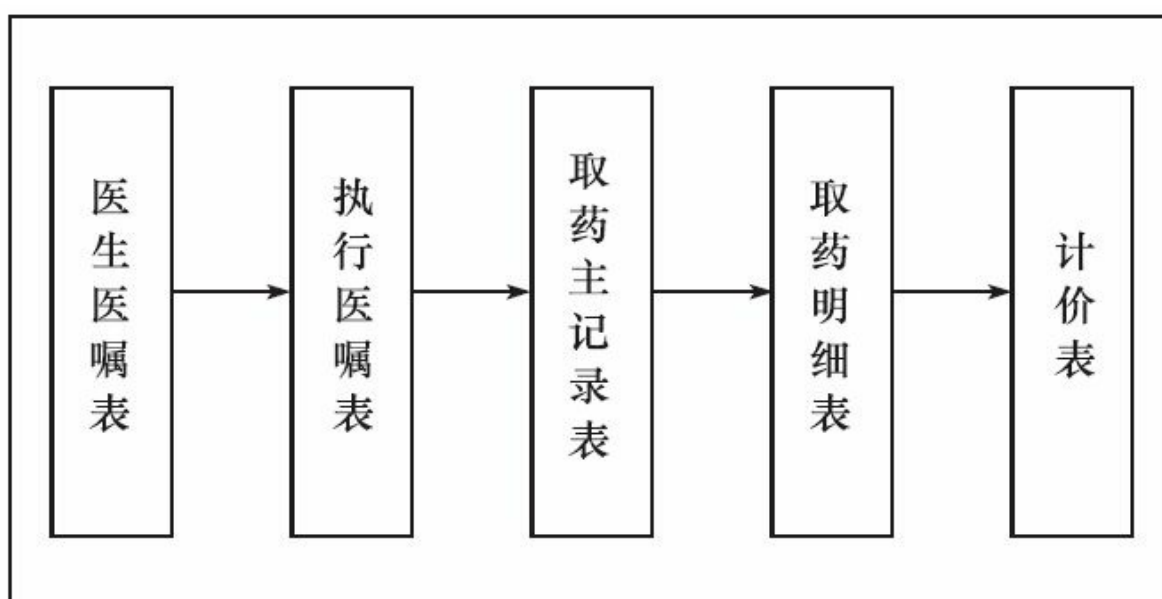


图 2-2

业务链的诊疗活动需要记录下来。这种记录活动，首先是医疗机构的业务需要，而不是首先为了满足数据挖掘的需要。在进行数据挖掘时，需要对照业务链和数据链，以判断数据的完整性，如图2-3所示。

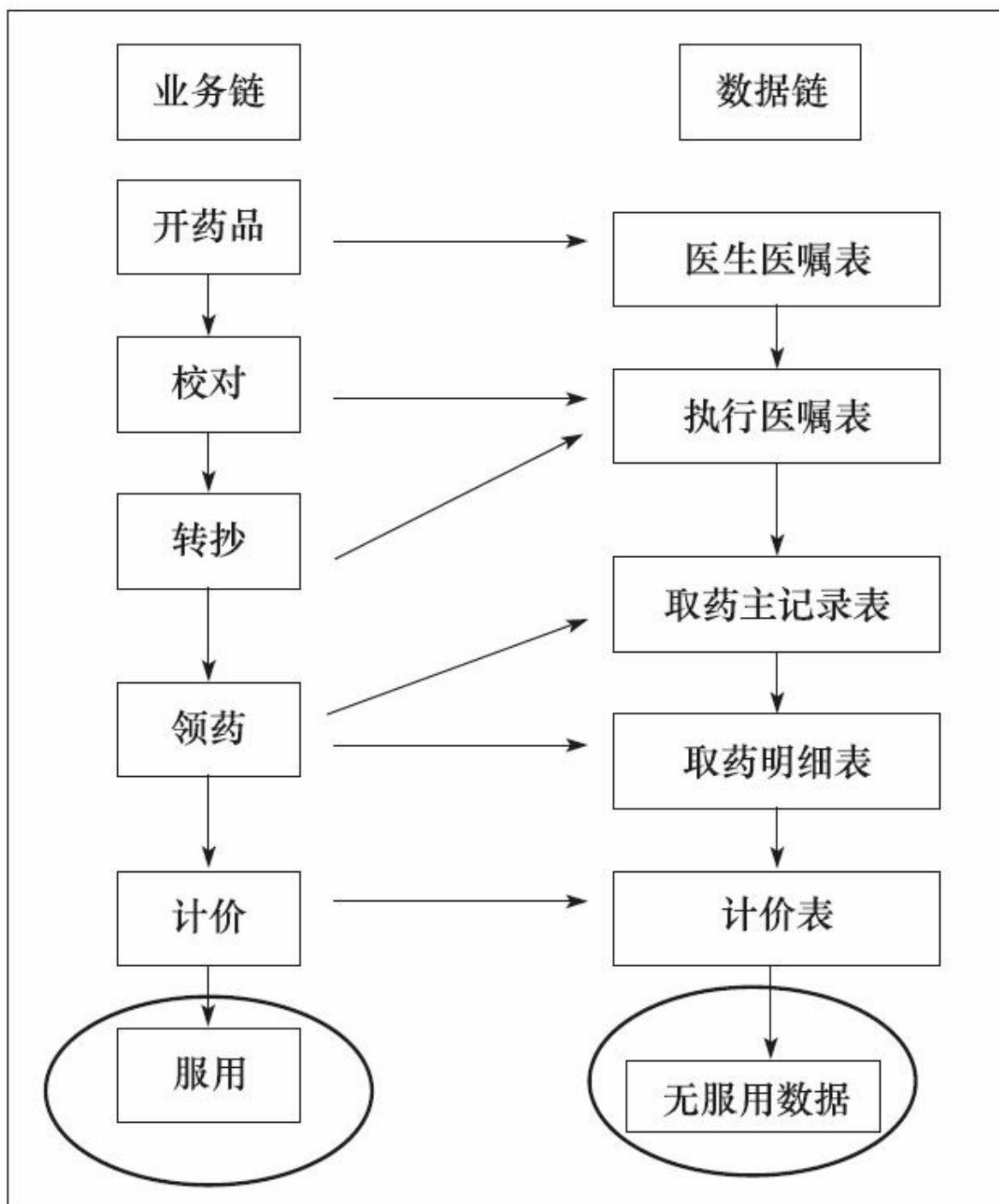


图 2-3

以用药来说，业务链的大部分被数据链所涵盖。而患者服用药物并没有数据链的相应记录，也就是没有一个药品服用结果表。在医疗行业，一般来说，药品的服用或使用由护士执行，只有药品记账后，

护士才可以从药房领取药品，然后监督病人正确服下（或给病人注射、输液）。也就是说计费的药品可以视为已被病人使用的，这个就视之为相关知识。

数据链是业务数据的可接触和采集部分，数据链的内容未必包括了流程的所有内容。我们说的完整性，指的是数据链的完整性是否包含了我们需要的所有数据。数据链有多种形式，常见的形式有表链。表链中需要表内具有完整性也需要表间具有完整性。表内的完整性，是指表内的数据没有缺失和错误，而对于涉及专业方面的逻辑错误，需要行业专家协助处理。**表间完整性**，主要考察主键与外键的关系，它们反映了表间关系是一对多的关系，还是多对多的关系，是依赖型的关系还是非依赖型的关系，我们会在后面的小节中谈及。

数据的可信性包括内部真实性和外部真实性。数据的可信性，一个指的是数据是否真实，比如数据是否虚假。建立在虚假数据基础上的知识和规律也不是真实可靠的；二是指数据是否可靠，比如病人就医的入院诊断，由于没有来得及进行各种检查，其可靠性就低于经历了各种必要检查、甚至手术和病理检查后的出院诊断。而可信性差的数据（低信度数据）是否无用呢？是否就成了垃圾数据了呢？比如上述的入院诊断信息的可信度低于出院诊断信息的可信度。入院诊断虽然可信度较低，但是我们可以分析在哪些病种上可信度低，为什么低，从而帮助提高入院诊断的正确率。所以说，**可信度低的数据也是有用的**。另一方面，有时所谓垃圾数据还可能没有用对地方的数据，不一定真是垃圾数据。下层出现的问题往往会影影响上层。很多数

据挖掘工程结果不理想，至少部分原因在于数据层没有处理好（可参照自发性回应导致的偏倚的例子，见本章第2.3节）。另外，不同行业领域的的数据源，必然受到行业特点的影响，只有了解行业才可以更好地处理数据。而了解行业也就是为数据挖掘的应用积累先验知识。

2.1.2 关系发现

商用数据挖掘通常是基于由业务信息系统产生的大量生产数据，而这些数据大多都存储在大型关系数据库中。所以在数据挖掘工作开始以前，首要的是分析业务数据库中的数据业务关系、数据技术关系和数据描述等。数据技术关系探索，主要分析诸如关系数据库的表间关系、外键和主键的关系、表间一对一关系、一对多关系、多对多关系、数据的依赖关系与非依赖关系，以及表内字段之间的关系、列冗余等。

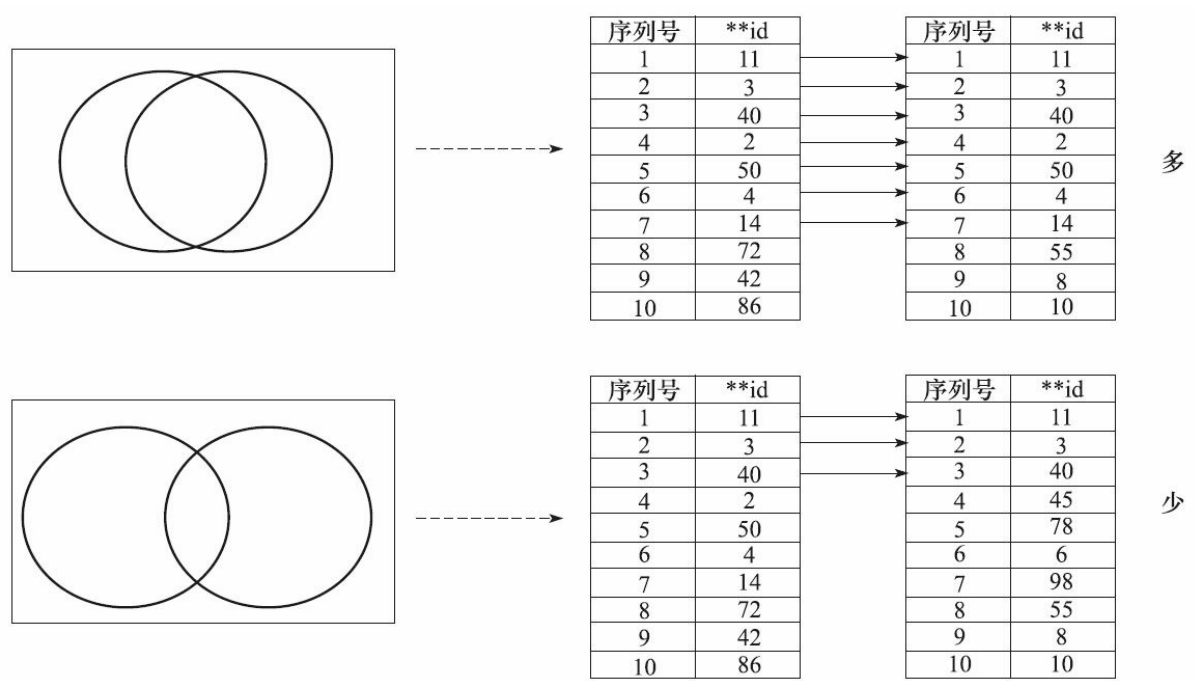
2.1.2.1 表间关系发现

关系发现揭示表间或数据库之间的关系，主要探索关系数据库的表间关系，如两个数据表的外键和主键的关系、两个数据表间的数量对应关系，以及两个表间的依赖关系等，可以使用关系匹配和发现工具来探索，也可以人工分析。很多企业保有大量的数据，如客户数据、供应商数据、产品数据、业务操作数据、业务分析数据、财务数据以及特定行业的特殊数据等。通常，这些数据都会横跨许多数据源。关系匹配和发现工具，可以通过不同的应用程序帮助数据分析师理解和管理所有这些数据，以及数据的来源和关系等。

这些数据中相关的记录可能存在于同一数据文件、不同数据文件或不同数据库中，可能有多条记录。通过关系发现，我们可以分析业务数据来回答下列问题：

- 关键表之间有什么潜在关系？
- 如果有一个主键和外键关系，是否可执行表间操作？
- 如果有一个明确的或推断的主键、外键关系，是否有任何孤立的数据？
- 有重复的记录吗？

图2-4显示了两个主键和外键的关系及两个数据表间的数量对应关系。图中两个左边的数据表是父表，两个右边的数据表是子表，两个父表的**id字段为主键，两个子表的**id字段为外键。图中上边的两个数据表的数据交集较大，下边的两个数据表的数据交集较小，正如图中上下两个示意图所示的那样。通过对元数据关系的理解，可以合并或者匹配不同表间的数据，并以此获得一个一致的、唯一的记录。



2.1.2.2 表内关系发现

关系发现揭示表内字段之间的关系，主要探索列属性、列逻辑相关等。对数据表内元数据模型分析，主要是从每个数据源中提取元数据，包括字段名称、字段数据类型、字段数据长度；比较字段的名称、类型、格式，使用模糊匹配算法来识别潜在的冗余和关系；列出冗余数据并在元数据仓库内编辑冗余目录。一旦数据关系被识别，可以通过识别冗余的可能性来验证和测试这些关系，通过数据源来验证关系，使用多个数据源识别分析来确定指定域的列。

2.1.3 数据质量探索

任何数据质量项目都需要调查和分析元数据，以便全面了解企业信息。我们可以扫描所有类型的数据来确定其关联的元数据，元数据表明当前数据内的特征，比如用数据类型、字段长度来判定数据的唯一性和字段是否可以缺失或为空值等。如图2-5所示，我们可以看到某个医学数据表EXAM_SUB_CLASS中字段的值。表2-1说明这个字段的相关信息。

~~通常，如果数据和元数据不一致，会给我们的数据质量控制和数据整合工作造成深远的影响。~~例如，考虑一个1000万行和长度为255个字符的字段。如果在数据中最长的数据元素是200个字符，设计的字段长度比要求的长，则会浪费550MB的磁盘空间。字段中的缺失值不应该有可能导致连接失败并报告错误结果的数据，表2-1显示了一个含有典型元数据的信息格式。

DEPT_ID	TEST_ITEM_NAME	REPORT_ITEM_NAME
289	血常规五分类	嗜碱细胞绝对值
289	血常规五分类	红细胞分布宽度CV
289	血常规五分类	大血小板比率
289	血常规五分类	白细胞数*
289	血常规五分类	红细胞数*
289	血常规五分类	血红蛋白*
289	血常规五分类	红细胞压积*
289	血常规五分类	红细胞平均体积*
289	血常规五分类	平均血红蛋白含量*
289	血常规五分类	平均血红蛋白浓度
289	血常规五分类	血小板数*
289	肝、肾、电、前...	前白蛋白
289	肝、肾、电、前...	A/G
289	肝、肾、电、前...	门冬氨酸氨基转移酶*
289	肝、肾、电、前...	总蛋白*
289	肝、肾、电、前...	白蛋白*
289	肝、肾、电、前...	总胆红素*
289	肝、肾、电、前...	直接胆红素*
289	肝、肾、电、前...	碱性磷酸酶*
289	肝、肾、电、前...	谷氨酰转肽酶*
289	肝、肾、电、前...	尿素*

图 2-5

表2-1 EXAM_SUB_CLASS字段信息

元数据名	元数据值
Data Type	VARCHAR
Primary Key Candidate	No
Unique Count	420
Uniqueness%	0.0969%
Minimum Value	1, 3-β-D 葡聚糖
Maximum Value	麝香草酚浊度试验
Minimum Length	6
Maximum Length	17
Null Count	0
Count	433634

2.1.3.1 模式分析

模式分析用于验证数据是否采用标准化的模式，是一个通常用于确定字段中的数据是否符合格式要求的技术。例如，电话号码或者产品标识符等字段往往都会具有特定格式。模式分析快速验证字段中的数据与数据源是否相同，以满足数据分析的需要。我们可以分析底层数据，然后开始构建校正和验证规则。对于任何数据，我们都可以把模式分析当成一个可执行的数据转换业务流程。

表2-2 电话号码表

模式	记录数	百分比
999-999-9999	3166	96.73
(999)999-9999	42	1.28
999 999 9999	34	1.04
999-999-AAAA	20	0.61
9-999-999-9999	5	0.15
others	6	0.18

表2-2为一个加拿大的电话号码格式的报。有许多有效的电话号码格式，但所有有效的格式由三组数字组成，其中三个编号为区域代码，三个编号为交换代码，四个编号为节点，这组数字可能是、也可能不是用空格或特殊字符分隔。有效的模式例如：

- 9999999999
- (999) 999-9999
- 999-999-9999

在这些例子中，“9”代表任何0至9的数字，从而构成了上述模式报告的一个电话号码字段。

下面为一个中国的电话号码格式的报。也有许多有效的电话号码格式，但所有有效的格式由三组数字组成，其中三到四个编号为区

域代码，三到四个编号为交换代码，四个编号为节点，这组数字可能是、也可能不是用空格或特殊字符分隔。有效的模式可能包括以下几种。

北京、天津：

- 999999999999
- 999-999999999
- 999-9999-9999
- (999) 9999-9999
- (999) 999999999

郑州、石家庄：

- 99999999999999
- 9999-9999999999
- 9999-9999-9999
- (9999) 9999-9999
- (9999) 999999999

太原、秦皇岛：

- 99999999999
- 99999-999999
- 9999-999-9999
- (9999) 999-9999
- (9999) 9999999

如果加上国际区号，就有诸如下面的号码：

- 1-9999999999 (多伦多)
- 86-999-99999999 (北京)
- 86-9999-99999999 (郑州)
- 86-9999-99999999 (太原)

2.1.3.2 统计分析

使用数据特征的统计结果可以发现趋势特点以及企业信息的共性，甚至非常规问题。仅通过检查一些基本的统计数据，我们就可以了解很多数据的情况。我们可以使用各种统计信息，包括最小值、最大值、平均值、中位数、方差和标准差等，来帮助我们评估数据的有效性。

表2-3为个人的财政组织住房贷款记录的统计数字，个人住房贷款额一般是20000（下限）到1000000（上限）美元不等。在贷款数据库中存储的不正确贷款记录可以导致很多问题，如无法有效分析结果、不正当贷款交易等。让我们看一下从贷款记账数据库中找到的一些不正常的贷款数据：

- 最小贷款值为负值
- 最高贷款值为9999999美元（超出上限）
- 贷款记录有缺失（即空值，但有贷款行为）
- 均值和标准差水平是一个出乎意料之外的很大值

这些数据全都表明了一个关于个人住房贷款记账的隐蔽性问题，这就需要构建一个拥有更完整细节的、正确的日常操作数据记录。另外，伴随着新数据的录入，跟踪基本统计信息，还可以令我们更加注意到新录入系统中的那些新数据的特征。这可以提醒我们一些前后不一致的信息，并且帮助我们避免向用于分析的数据资源中加入有问题的数据。

表2-3 个人财政组织住房贷款记录

元数据名	元数据值
Data Type	Double
Primary Key Candidate	No
Unique Count	1140
Uniqueness	70. 11
Pattern Count	(not applicable)
Minimum Value	- 223000 ←
Maximum Value	9999999 ←
Minimum Length	(not applicable)
Maximum Length	(not applicable)
Null Count	2 ←
Blank Count	(not applicable)
Actual Type	Double
Count	1628
Data Length	53 bit
Mean	114348. 170972 ←
Median	4888499. 5
Mode	0
Non-dull Count	1626
Nullable	YES
Ordinal position	7
Decimal places	0
Standard deviation	429438. 361236 ←
Standard Error	10649. 778281

应用统计学能帮助我们在深入挖掘和分析数据以前，做大量的重要的探索工作，往往对数据挖掘的应用结果产生巨大的影响，我们将在稍后的章节中重点讨论。

2.1.4 数据整合

对于一个拥有大量的客户、产品和供应商数据资源的企业，需要构建一个完整的、一致的、有效的、用于分析的数据中心，这就需要数据整合了。若要实现一个成功的分析性数据中心的商务计划，我们面临着两个最基本的挑战：了解数据的真实质量水平，并且将企业数据整合到一个独立可靠的平台。一旦这些基础数据有缺陷，那么不完善的数据整合会影响到企业资源计划、客户关系管理、产品和供应等各个环节。

2.1.4.1 数据链接和匹配

数据整合的核心是横跨多个数据源去理解数据、整合数据、清理重复信息等。这个技术可以使我们减少由于一个客户有多条记录所造成的混乱。我们还可以使用数据链接和数据匹配技术，从不同的数据库或应用程序来理解业务数据的特征。

1. 规范重复的客户记录

我们会发现在我们的数据存储中还有许多的隐藏的关系，通过实时方式或批处理方式匹配客户信息，可以完善这些发散的客户信息，使我们能够对客户基础构建一个单一的、统一的视图。有时，通过识别一些重复的记录、将顾客信息与数据仓库链接后，我们会发现一些欺诈行为。例如，要建立更有效的以客户为中心的经营策略，那么就

必须在企业中的每个信息源都建立更完善的客户信息。这项技术的构建能够协调不同的数据模式和格式，以及语音和拼写错误。为了便于理解，举一个简单的例子，如图2-6所示。

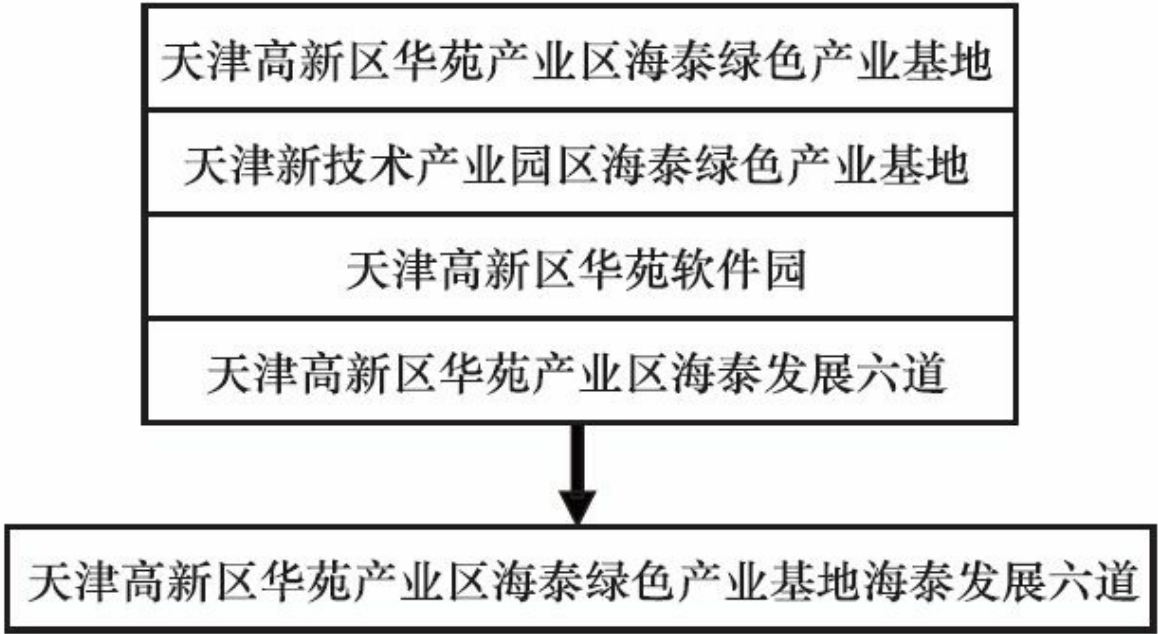


图 2-6

- “天津高新区华苑产业区海泰绿色产业基地”
- “天津新技术产业园区海泰绿色产业基地”
- “天津高新区华苑软件园”
- “天津高新区华苑产业区海泰发展六道”
- “天津高新区华苑产业区海泰绿色产业基地海泰发展六道”

以上都表示同一个地址，我们需要将前四个地址映射到第五个地

址上，同时保留存储前四个地址，由此建立起精确、高效的数据规范和数据备份。

2.匹配和同步复杂模式数据

在实际案例中，我们需要理解不同产品或项目的数据模型的复杂性和格式差异，通常是非名称和非地址类的数据。一旦了解这种知识，我们就可以对产品数据、库存数据或资产数据进行管理。通过识别数据格式、详细规范、度量数据，以及产品的缩写、数字代码和任何其他类型的业务数据，我们能够实时地或批处理地完成数据信息的匹配和同步。市场上有各种各样的具有自动匹配复杂模式的软件，这些软件可以帮助我们理解业务数据中不同元素的重要性，常用的功能包括：

- 标准化格式差异
- 多值字段中筛选重要的信息
- 翻译缩写或数字代码
- 执行模糊匹配规则

2.1.4.2 数据统一整合

如2.1.3.1的例子，在了解了不同数据的规则模式后，我们需要一个完整、统一的数据结构设计，如表2-4所示。

表2-4 统一的数据结构例子

	国际区号 1-2 位数	地区号 (省市) 3-4 位数	交换号 3-4 位数	节点号 4 位数
定义	VARCHAR(2)	VARCHAR(4)	VARCHAR(4)	VARCHAR(4)
多伦多	1	999	999	9999
北京	86	999	9999	9999
郑州	86	9999	9999	9999
太原	86	9999	999	9999

虽然常见的电话号码都为数字，但我们仍然定义它们为字符类型，因为有些号码的首位为0，尤其是区号，如北京的区号为010，郑州的区号为0371等。当然，这里只是为了展示有关数据探索内容的完整性，才提到一点有关数据整合的内容，其实数据整合是一个非常复杂又丰富的技术，深入涉及这方面的内容会超出本书的核心内容。

2.2 数据特征探索

通过对数据关系的探索分析，我们可以了解源数据的逻辑关系和关系模型，并通过对数据质量的控制和处理，进一步提高数据整合的效果。接下来的工作是要对整合后的数据进行较深入的特征探索和分析，主要包括对数据进行统计学描述分析、统计学图表绘制、常见概率分布分析、列（属性、变量）相关分析等。本节将重点介绍数据的统计学特征和统计学特征应用两个方面的内容。

2.2.1 数据的统计学特征

数据的统计学特征分析主要是从数据的变量类型、描述统计分析、概率分布分析等方面加以考察。

2.2.1.1 变量类型

1.连续变量

如图2-7所示，如果对于随机变量 x 的分布函数 $F(x)$ ，存在非负可积函数 $f(x)$ ，使对于任意实数 x 有

$$F(x) = \int_{-\infty}^x f(t) d(t) \quad (2.1)$$

则称 x 为连续型随机变量。

2.离散变量

如图2-8所示，一个随机变量，如果它的可能取值是有限个或可列无限多个，这种随机变量称为离散型随机变量。

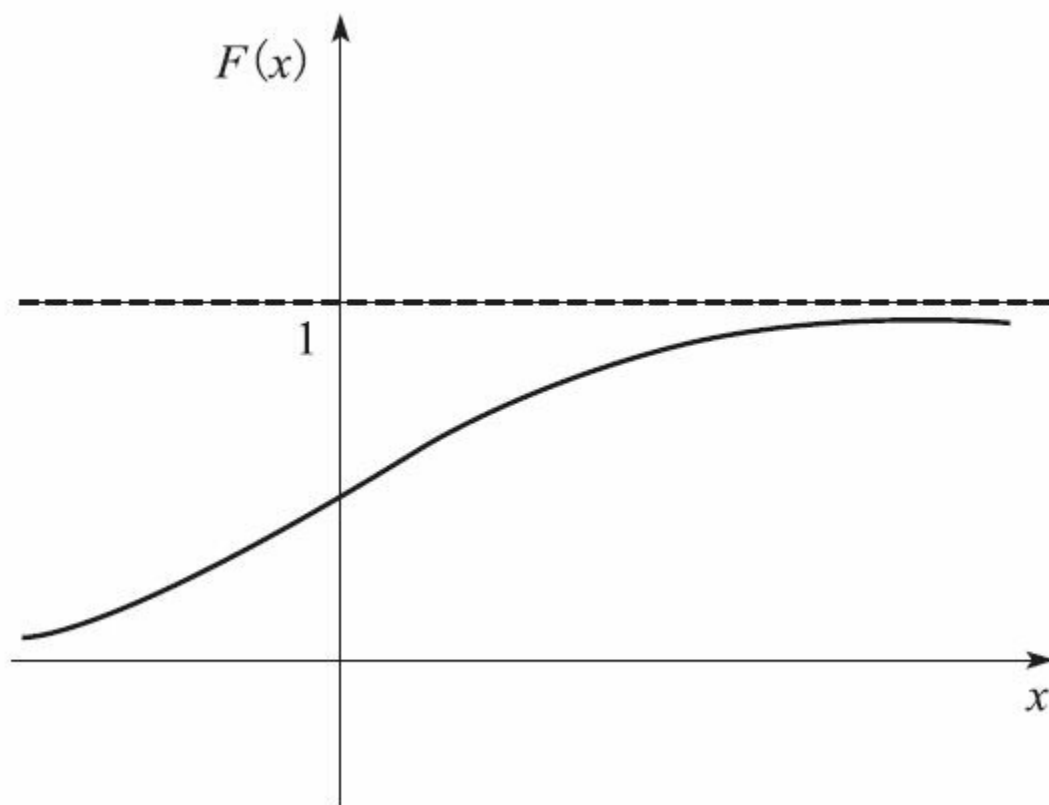
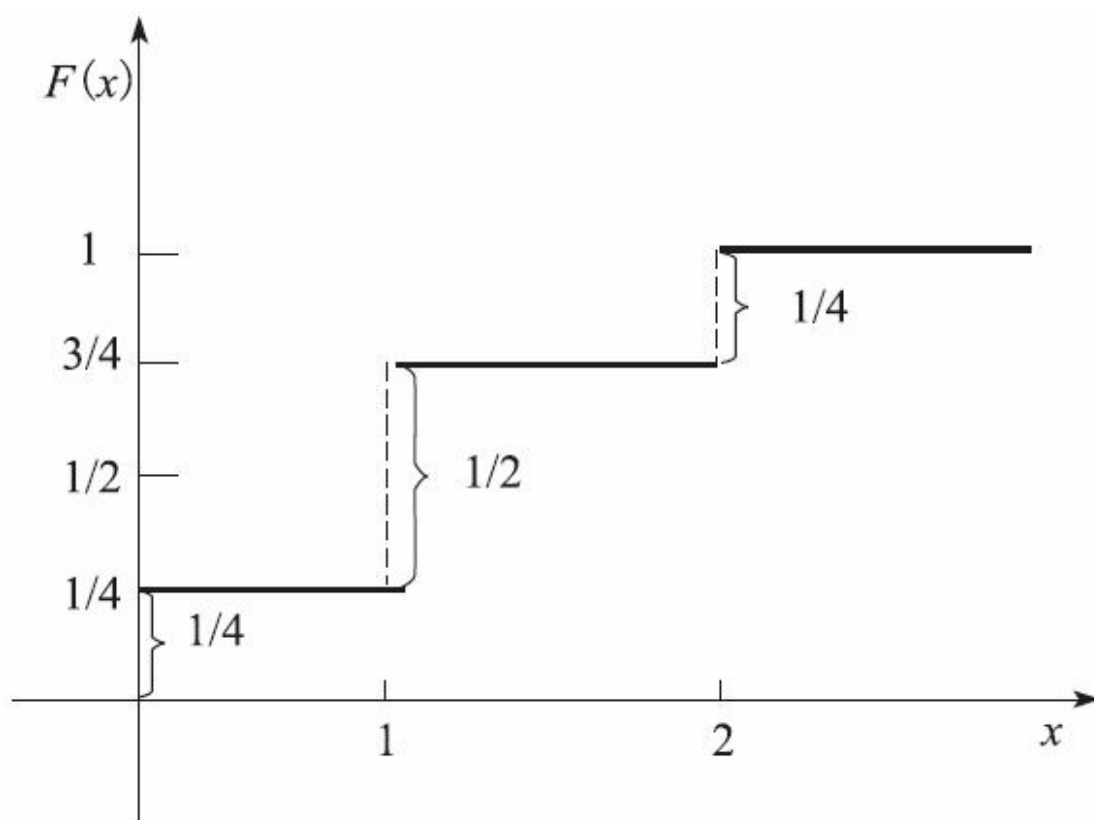


图 2-7



2.2.1.2 描述统计分析

描述性统计包括：

1.中心趋势分析

- 均值和加权算术均值
- 中位数
- 众数
- 中列数

2.离散趋势分析

- 极差
- 四分位数
- 方差和标准差

3.描述性图表

1) 散点图

如图2-9所示，散点图是用两组数据构成多个坐标点，考察坐标点的分布，判断两组变量是否存在某种关联或总结坐标点的分布模式。

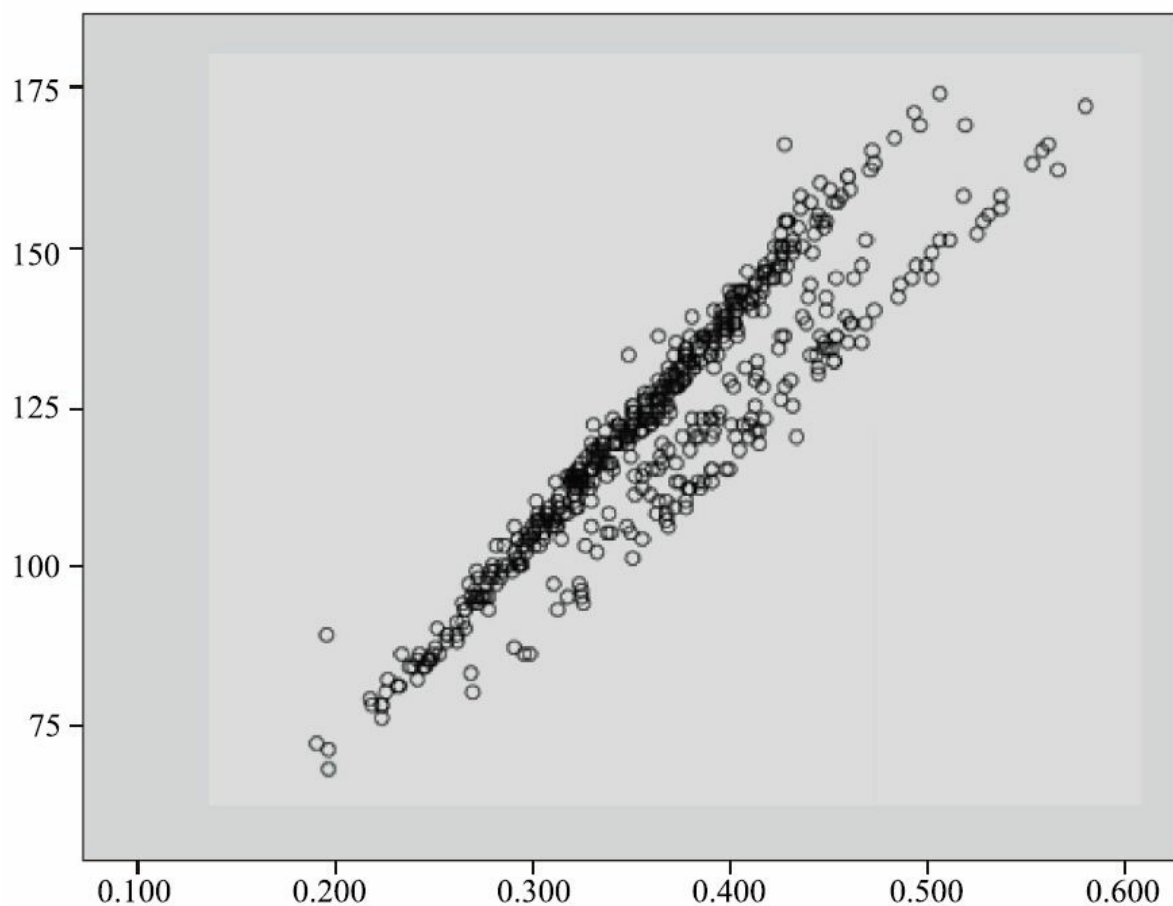


图 2-9

2) 直方图

如图2-10所示，直方图是由一系列高度不同的纵向条纹或线段表示的数据分布情况。

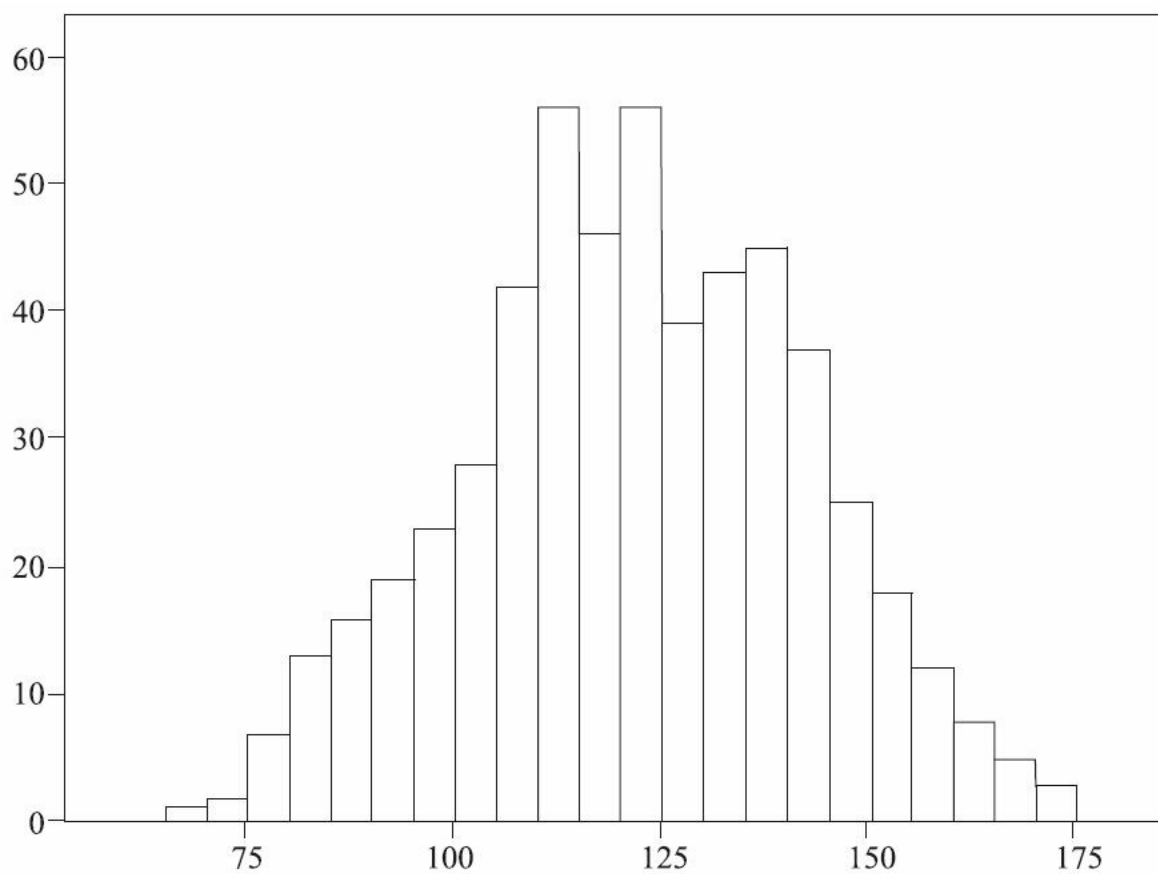


图 2-10

3) 线图

如图2-11所示，线图是用线条的升降表示事物的发展变化趋势。

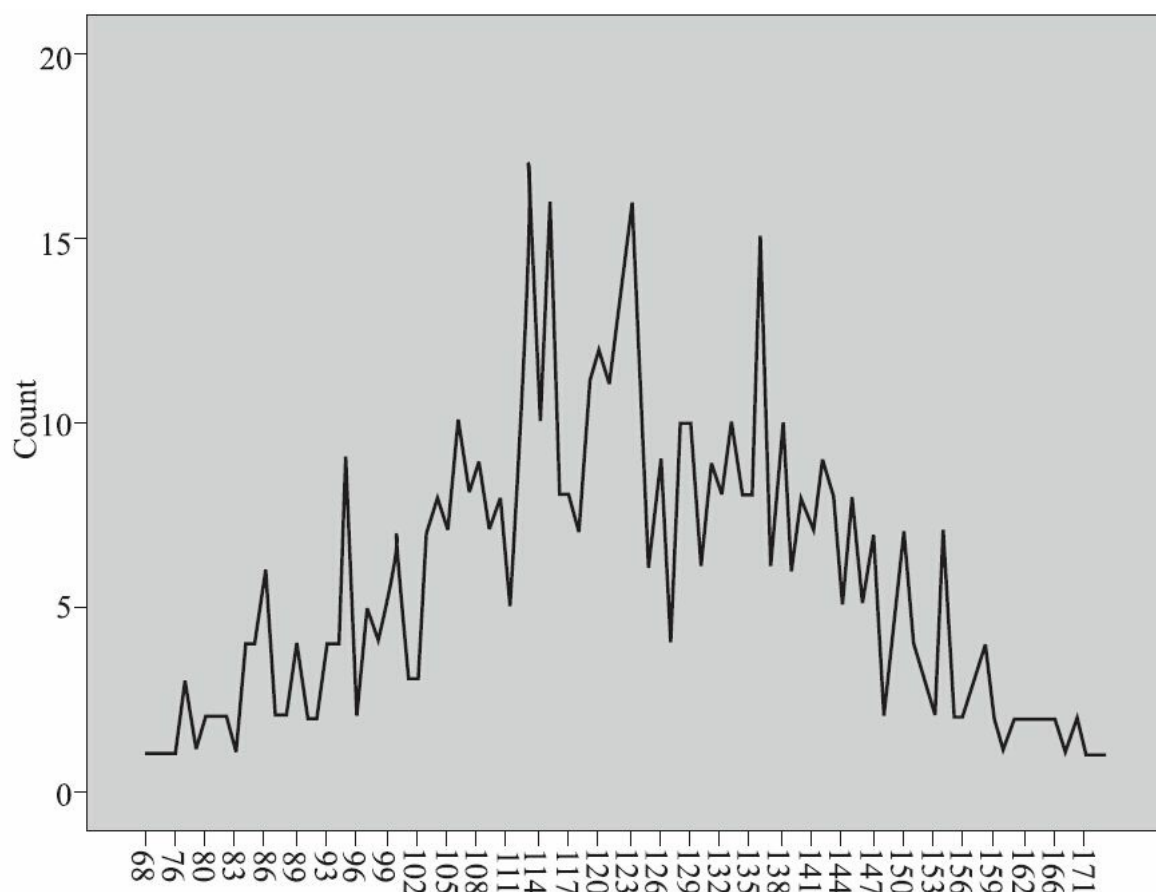


图 2-11

4) 饼图

如图2-12所示，饼图是描述构成比的大小，用颜色将不同比例表达出来。

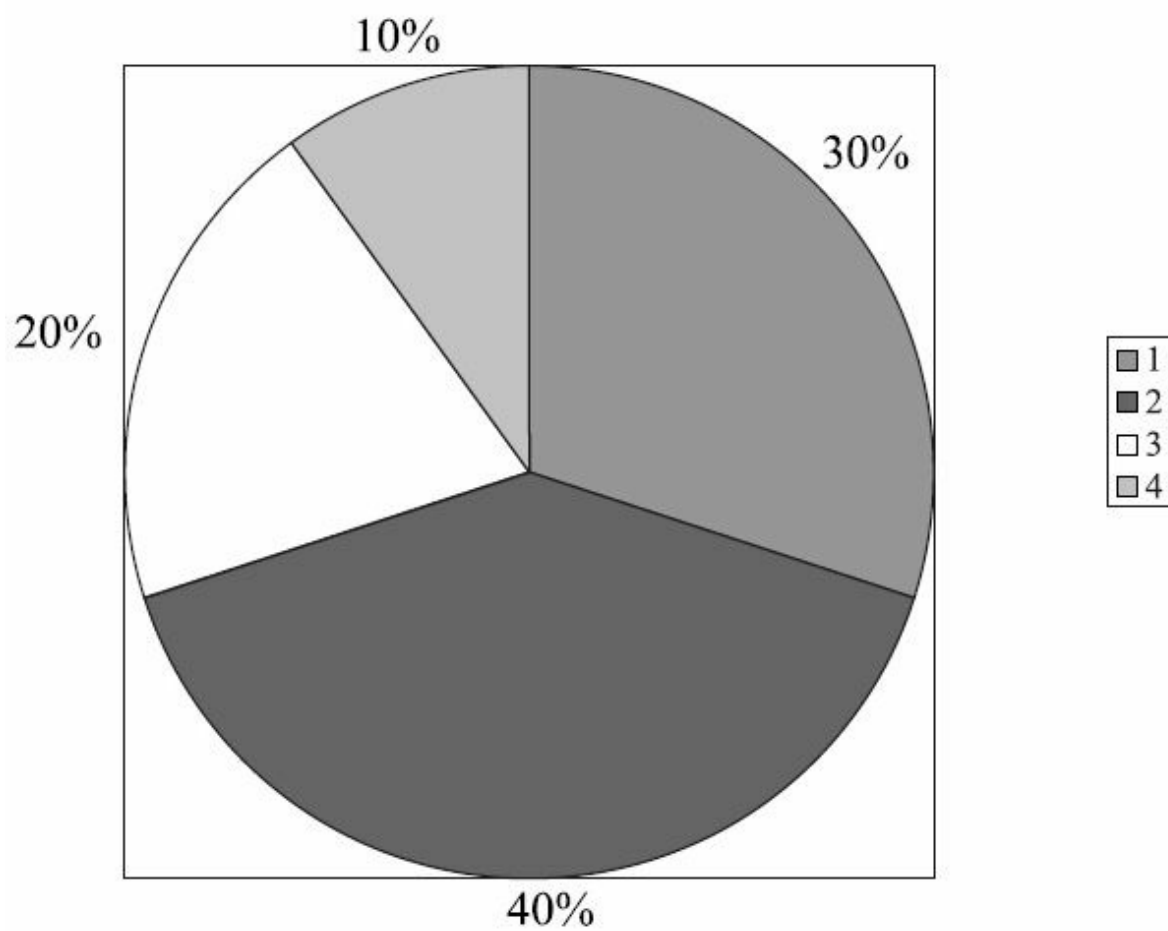


图 2-12

5) 直条图

如图2-13所示，表示独立指标在不同阶段的情况，用相同宽度长条的不同长度来表示。

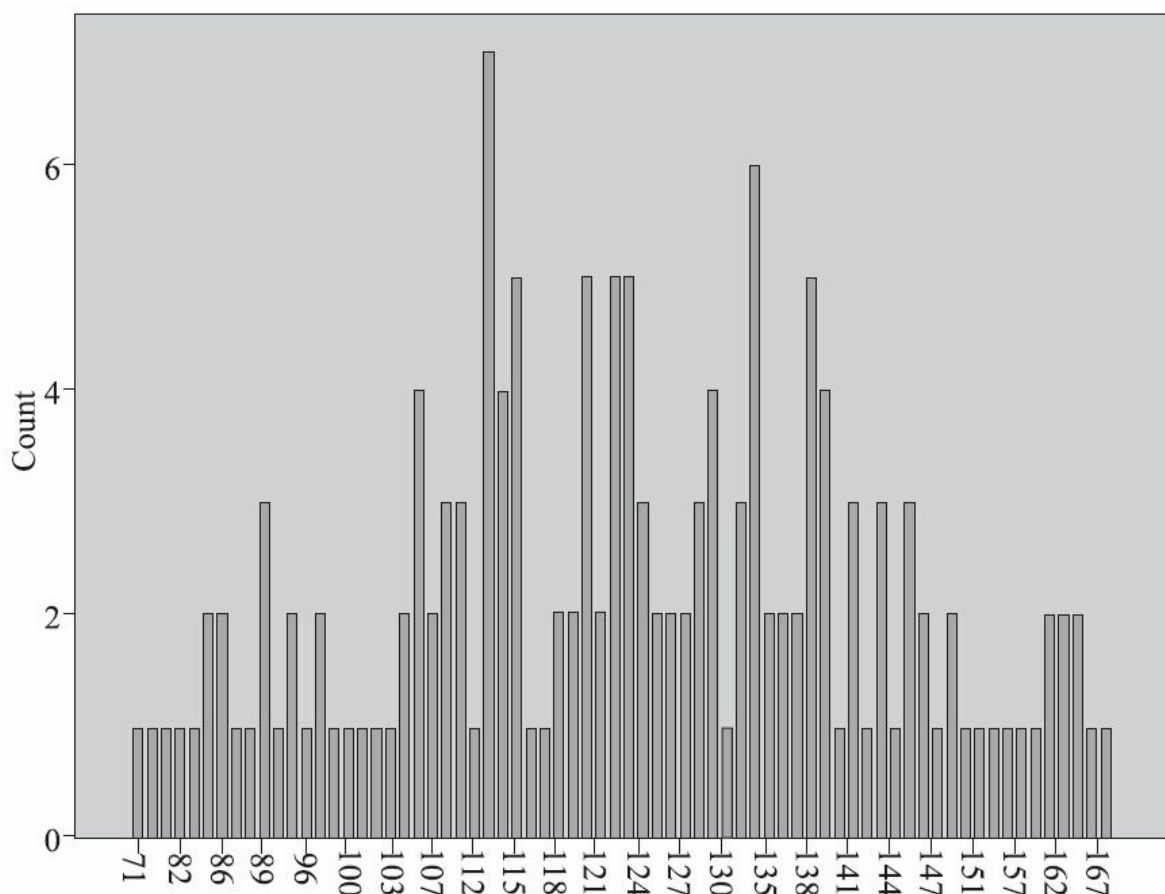


图 2-13

2.2.1.3 概率分布分析

数据概率分布分析主要是了解数据的自然属性特点。不同的维度、变量（如年龄、性别等）下的数据分布可能不同。不同种类的数据有不同的分布，例如体检人群和患病人群在年龄、性别上的分布可能不同。

常见的连续数据分布分析包括：

- 1) 高斯分布：若连续型随机变量 x 的概率密度为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty \quad (2.2)$$

其中 μ , σ ($\sigma>0$) 为常数, 则称 x 服从参数为 μ , σ 的高斯分布。如图2-14所示, 该曲线关于 $x=\mu$ 对称, 而且 x 离 μ 越远, $f(x)$ 的值越小。这说明对于同样长度的区间, 当区间离 μ 越远, x 落在这个区间上的概率越小。该分布曲线的高峰位于正中央, 即均值所在的位置, 具有集中性, 同时从均值开始均匀变动, 取值范围大致为 $(\mu-3\sigma, \mu+3\sigma)$ 。

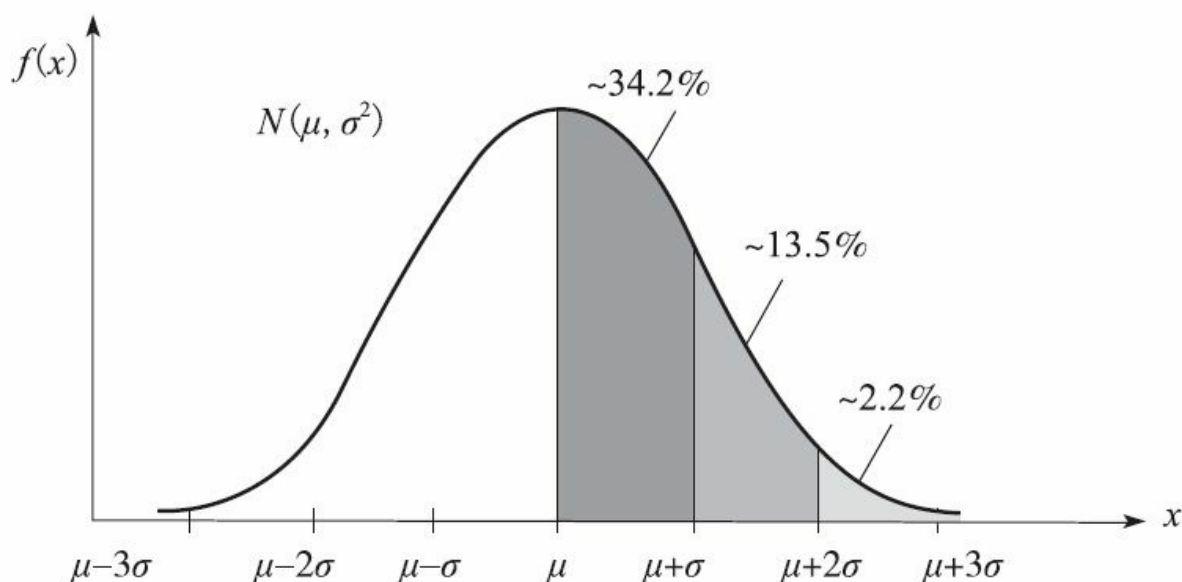


图 2-14

2) T分布: 假设 X 服从标准正态分布 $N(0, 1)$, Y 服从 $\chi^2(n)$ 分布, 且两者相互独立, 则

$$Z = \frac{X}{\sqrt{Y/n}} \quad (2.3)$$

的分布称为自由度为 n 的T分布。如图2-15所示, 该曲线关于 $t=0$ 对称, 当 n 足够大时, 该分布就类似于标准正态分布, 而对于较小的 n ,

T分布与标准正态分布相差较大，曲线较为低平。

3) F分布：假设X服从 $\chi^2(n_1)$ ，Y服从 $\chi^2(n_2)$ ，且两者相互独立，则称随机变量

$$F = \frac{X/n_1}{Y/n_2} \quad (2.4)$$

服从自由度为 (n_1, n_2) 的F分布。如图2-16所示，分布曲线是非对称分布，并且有两个自由度， n_1 通常称为分子自由度， n_2 称为分母自由度。从该分布的公式可以看出，该分布具有倒数性质，即 $\frac{1}{F} \sim F(n_2, n_1)$ 。

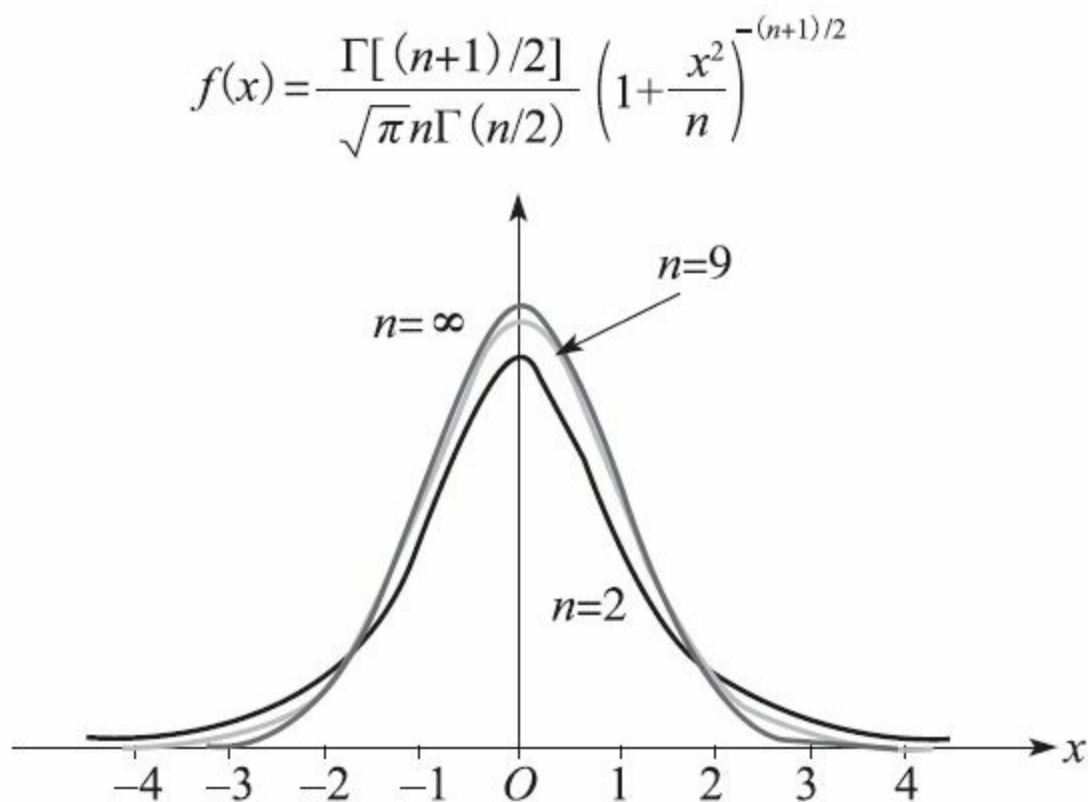


图 2-15

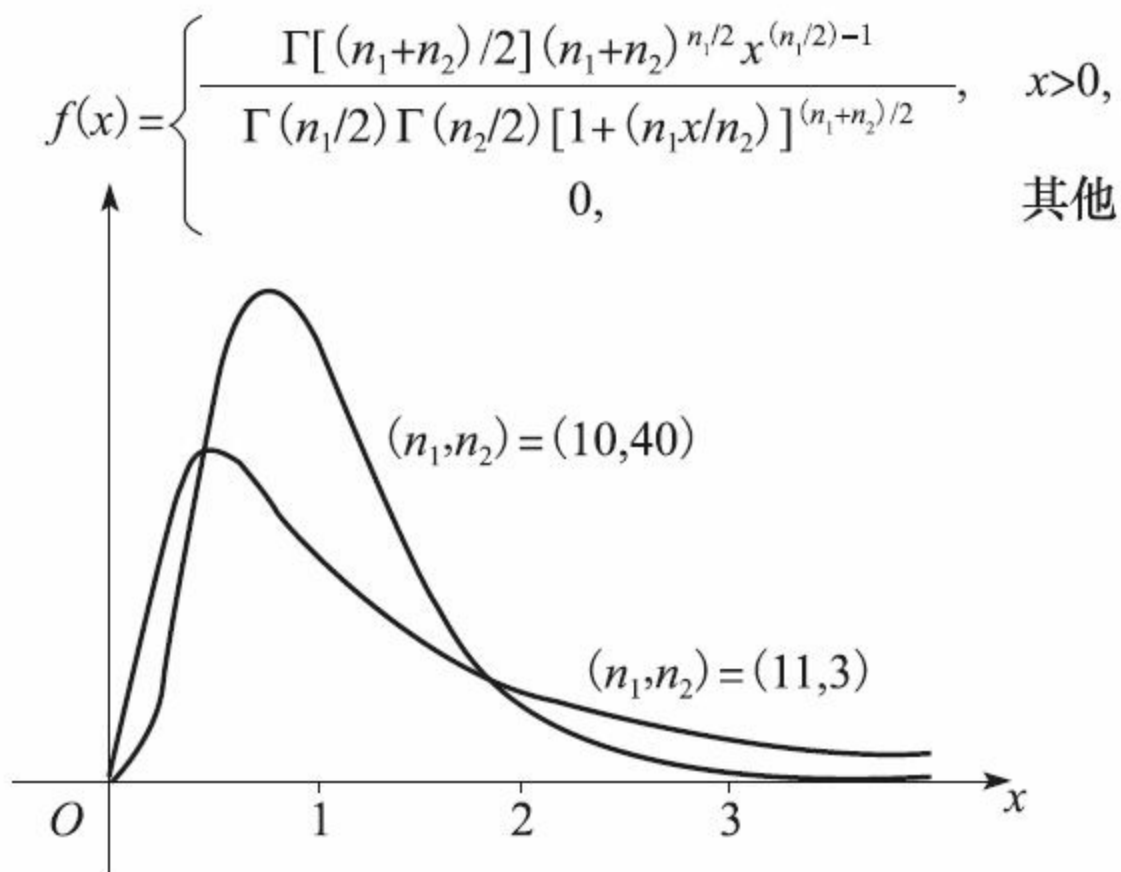


图 2-16

4) 卡方分布: 若 n 个相互独立的随机变量 $\xi_1, \xi_2, \dots, \xi_n$ 服从标准正态分布, 则这 n 个服从标准正态分布的随机变量的平方和构成一个新的随机变量, 其分布规律称为卡方分布。如图2-17所示, 卡方分布在第一象限内, 呈正偏态, 随着 n 的增大, 卡方分布趋近于正态分布, 而且向正无穷方向延伸, 分布线也越来越低阔。从图中还可以看出卡方值都是正值, 而且卡方分布曲线下的面积都是1。

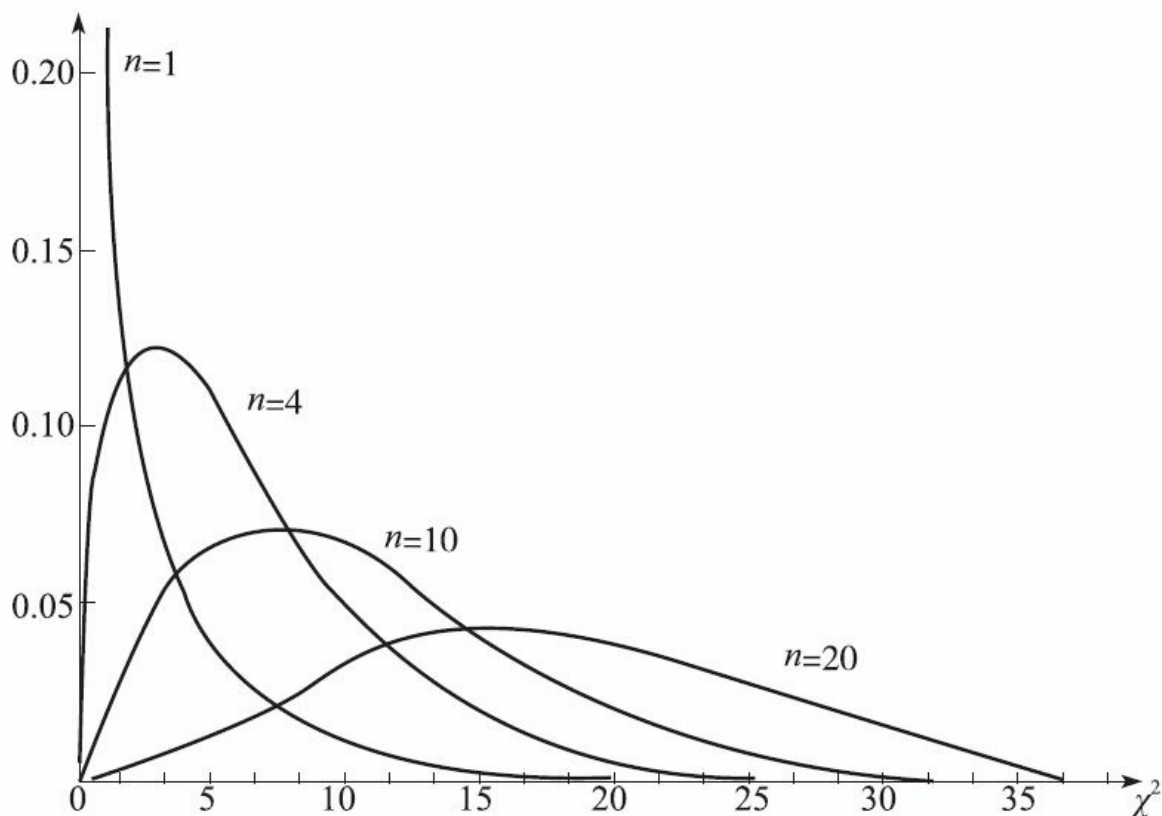


图 2-17

常见离散数据分布分析包括：

1) 二项式分布：在 n 次Bernoulli实验中，“结果A出现 k 次”这一随机事件的取值及其概率。如图2-18所示，二项式分布的形状取决于 π 和 n ，其中 π 是每次Bernoulli实验中结果A出现的概率， $f(x)$ 为结果A出现 k 次的概率，该图形的高峰在 $m=n\pi$ 处。当 $\pi=0.5$ 时，图形是对称的。当 n 增大到无穷大时，只要 π 不靠近0和1，二项式分布近似于正态分布。

2) 几何分布：在 n 次Bernoulli实验中，实验 k 次才得到第一次成功的概率。如图2-19所示。

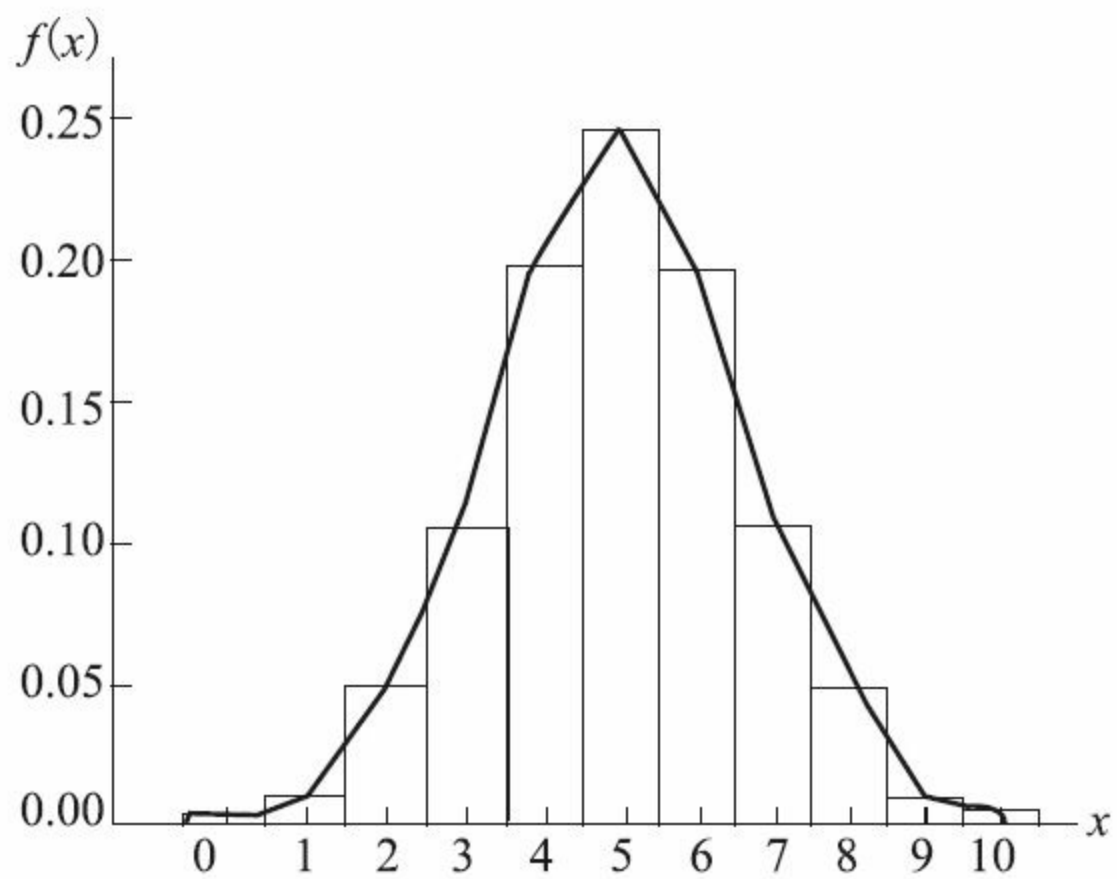


图 2-18

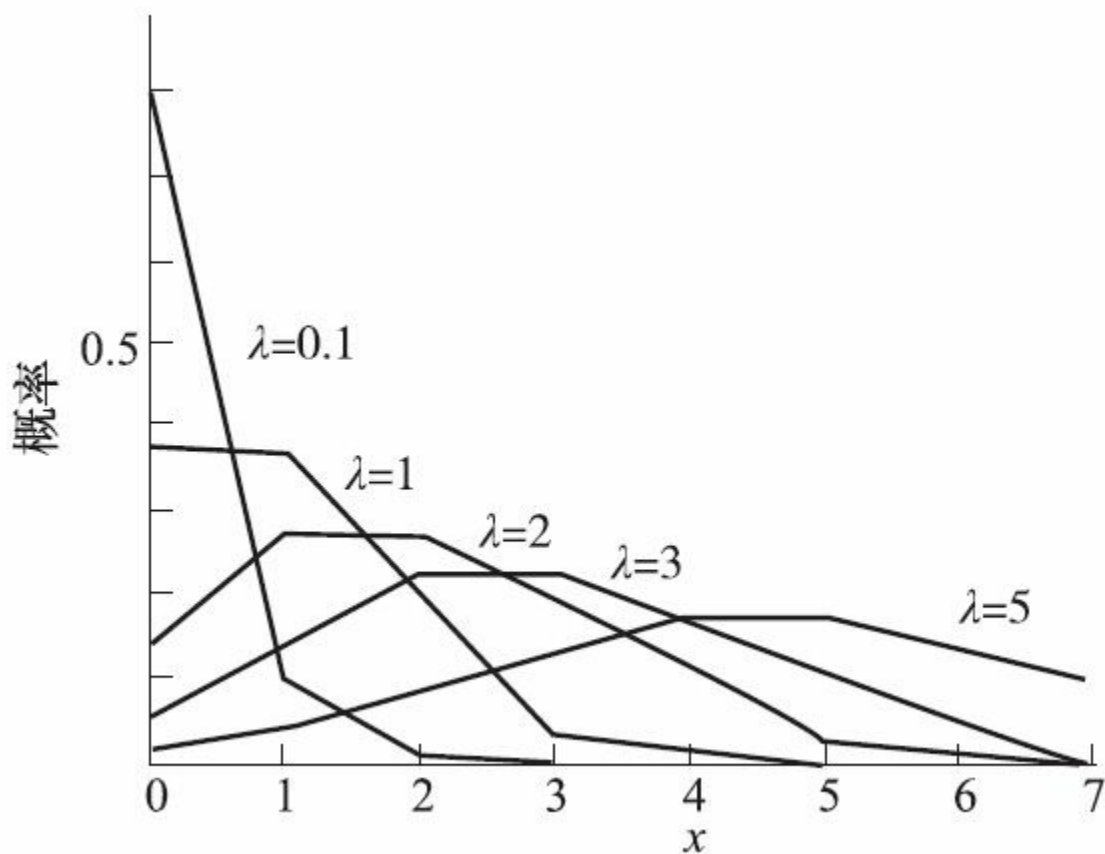


图 2-19

3) 超几何分布：统计学上一种离散概率分布，描述了由有限个物件中抽取 n 个物件，成功抽出指定种类的物件的次数（不归还）。如图2-20所示，当 N 趋于无穷时，就可以用二项式分布近似描述制定种类物品的个数。

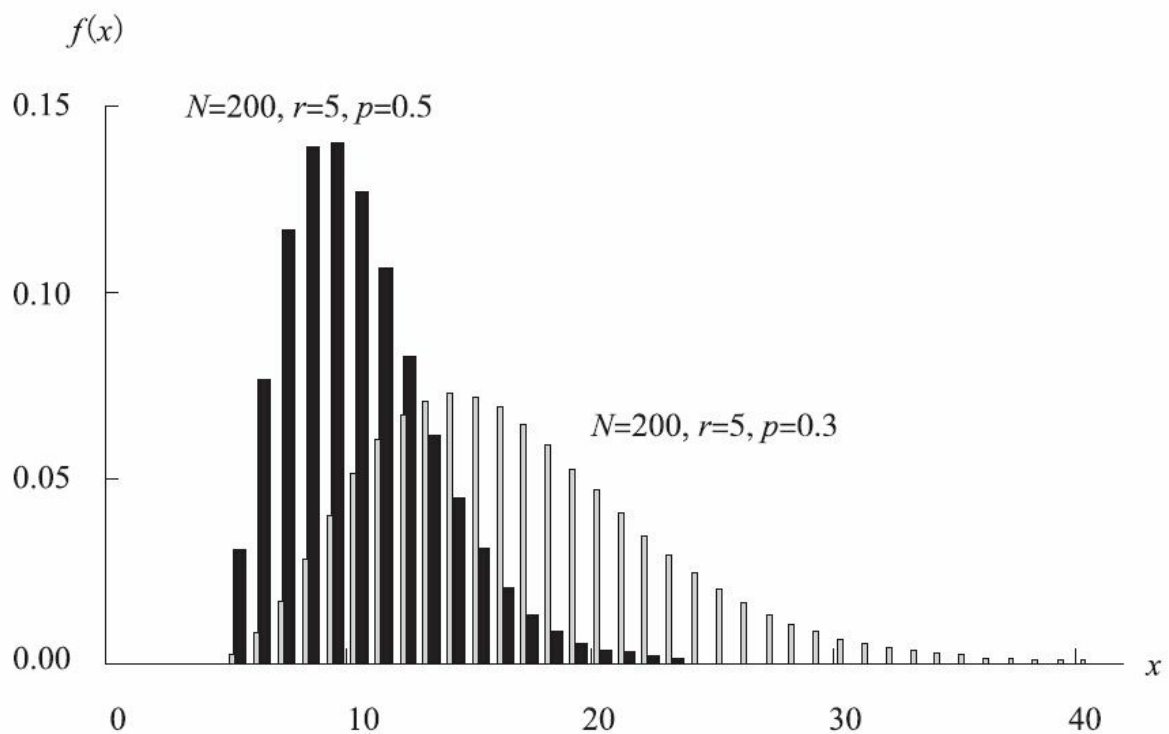


图 2-20

4) 泊松分布：随机变量 X 所有可能的取值为 $0, 1, 2, \dots$ ，而取各个值的概率为

$$P\{X = k\} = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots \quad (2.5)$$

其中 $\lambda > 0$ 是常数，则称 X 服从参数 λ 为的泊松分布。如图2-21所示，当 n 很大 p 很小时，该分布可用于二项式分布的近似计算。该分布适合于描述单位时间内随机事件发生的次数。

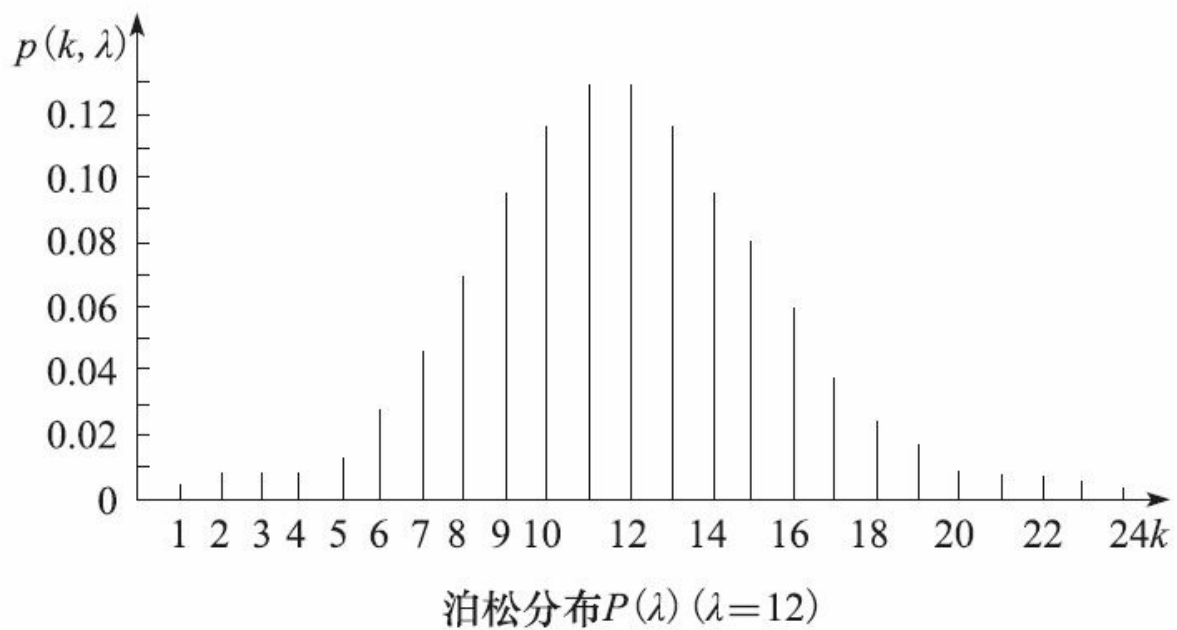


图 2-21

有关数据表列（属性、变量）的相关分析，参见第3.2节。

2.2.2 统计学特征应用

在长年的数据挖掘应用实践中，我们经历了许多行业的各种不同的应用，其中在医学科研领域中，我们开发过完整的数据挖掘应用系统，即智能医学科研系统，简称为IMRS（Intelligent Medical Research System）。它包含了从业务数据的探索性分析到样本数据的专业性选取，再到数据挖掘的深入分析，以及大量应用各种统计学方法对数据挖掘结果进行验证等各个环节，是一个比较典型的行业数据挖掘应用案例（详见第6章）。本书将从本节开始，使用该系统的许多功能模块作为各章的讲解案例，其中第7章更有对该系统比较集中和详细的介绍和讨论。默认的情况下，我们在很多地方以IMRS中的应用案例为背景。

统计分析分为统计描述和统计推断两部分。统计描述是统计推断的基础，是通过绘制统计图表和计算数据分布特征的基本统计量来了解样本观察值的分析特征，为进一步的统计推断打下基础。统计描述主要包括绘制样本数据的频数分布表和频数分布直方图，计算样本数据的最大值、最小值、均值、标准方差、中位数、变异系数和百分位数等。

可以通过下面的例子来认识以上统计描述有关统计量的概念。通过IMRS“统计子系统”的“描述统计”模块得到样本数据，如图2-22所示，疾病人群选择动脉硬化性心脏病病人，检验项为这些病人入院后首次的甘油三酯的化验值，并且是入院后5天内首次的化验值。得到数

据列表中“检验结果”列，即“检验结果”变量就是本次要统计描述的样本。

描述统计数据选择

上一步

下一步

观察人群：

动脉硬化性心脏病

(必选)

检验条件：

入院首次化验

检验项：

甘油三酯

(必选)

检验时间跨度：

5

(天)

生成数据

入院首次化验结果

病人Id	住院次数	检验结果
191179	2	6.06
191179	4	6.04
191179	6	5.66
206185	2	4.06
206185	2	4.06
206316	2	5.03
206316	2	5.03
207642	1	8.43
207642	2	7
207642	3	5.74
207642	4	7.69
214029	4	7.68
214918	12	10.28
222287	2	9.86

图 2-22

然后在IMRS中对图2-22中的“甘油三酯的检验结果”变量进行统计描述。点击图2-22中的“下一步”按钮，然后点击“入院首次数据分

析”按钮，得到统计描述结果，包括统计量值的列表（图2-23）、频数分布表（图2-24）、频数分布直方图（图2-25）。

图2-23列表内是各个基本统计量的值，包括集中趋势的描述和离散程度的描述。

集中趋势的主要统计指标有均值和中位数，这些指标也称为位置度量指标。

- “中位数”：连续型样本数据按数值大小排列，位于位次中点的数据即中位数。

- “均值”，即“算术均值”：等于所有样本值之和除以样本个数，用于反映样本数据的平均水平。

入院首次化验数据描述统计结果

中文名 ▲	英文名	统计量值
百分之二十五分位数	Quartile_25	.96
百分之九十五分位数	Quartile_95	3.19
百分之七十五分位数	Quartile_75	1.9
百分之五分位数	Quartile_05	.68
变异系数	Cv	53.69%
标准差	Stddev	.8269
方差	Variance	.6837
极差	Range	4.26
均值	Mean	1.5401
中位数	Median	1.345
最大值	Max	4.81
最小值	Min	.55

图 2-23

频数分布表4

组段	频数	频率	累计频数	累计频率
.45~	18	.26	18	.26
.98~	24	.35	42	.62
1.5~	11	.16	53	.78
2.03~	7	.1	60	.88
2.55~	3	.04	63	.93
3.08~	3	.04	66	.97
3.61~	1	.01	67	.99
4.13~	0	0	67	.99
4.66~	1	.01	68	1
5.18~	0	0	68	1

图 2-24

均值和中位数的关系：如果数据呈对称分布，那么数据的均值和中位数接近；如果数据呈左偏分布，则均值小于中位数，若呈右偏分布，则均值大于中位数。所以，可以比较两者大小来初步分析数据分布的对称性。

离散程度的主要统计指标有极差、方差、标准差、变异系数等。

- “最大值”和“最小值”：连续型样本数据按数值大小排列后的最大值和最小值。

- “极差”，也称为“全距”：是连续型样本数据的最大值与最

小值之差，反映数据变动的范围，绘制频数分布表时，由全距来决定组段数和组距。

- “方差”：描述一个变量所有的观察值与均值的平均离散程度的指标。方差越大，数据的离散程度越大。

- “标准差”：描述一个变量所有的观察值与均值的平均离散程度的指标。标准差是方差的算术平方根，对于计量单位相同的变量，标准差越大，数据的离散程度越大。

- “变异系数”：一个度量相对离散程度的指标，没有量纲。变异系数亦称为离散系数，是各个变异指标与其算术平均值的比值。比如：极差与其平均值的比称为极差变异系数，标准差与均值的比称为标准差系数。所以变异系数可以用来比较几个量纲不同的变量之间的离散程度的差异，也可以用来比较量纲相同但平均值相差悬殊的几个变量之间的离散程度的差异。变异系数值越大，离散程度越大。

- “分位数”：介于最大值和最小值之间的一个数值，使得变量的一部分观察值小于或等于这个数，另一部分观察值大于或等于这个数。两个分位数之差可以反映观察值的离散程度。最常用的分位数是百分位数。比如，“百分之五分位数”是0.125，则变量的5%的观察值小于或等于0.125，变量的95%的观察值大于0.125。最常用的百分位数是百分之五分位数、百分之二十五分位数、百分之五十分位数、百分之七十五分位数和百分之九十五分位数。

· “频数”指观察值出现的次数。为了解样本中观察值的分布规律和便于统计量计算，可以绘制频数分布表。频数分布表分为离散型数值变量资料频数分布表和连续型数值变量资料频数分布表。

离散型数值变量频数分布表绘制方法：以变量的不同取值进行分组，并分别列出各组所对应的频数；各组的频率是各组的频数除以总频数；各组的累计频数和累计频率均是该组的向下累计。

连续型数值变量频数分布表绘制方法：

1) 求全距：即极差，最大值与最小值之差。

2) 确定组段和组距：一般设8至13个组段，各组段的起点和终点分别称为下限和上限；相邻两个组段的下限之差称为组距，常用全距的1/10取整后作为组距；某组段的组中值是（下限+上限）/2。注意：第一组段应包括最小值，最末组段应包括最大值。

3) 列出频数表：将原始数据汇总，得出各组段的频数，即大于等于各组段的下限并且小于各组段的上限的观察值的例数；得出各组段的累计频数、频率和累计频率，如图2-24列表。

4) 绘制频数分布直方图：就是以分组变量为横轴，以频数为纵轴，以直方形的高度来表示次数分布的图形，如图2-25。

频数分布直方图4

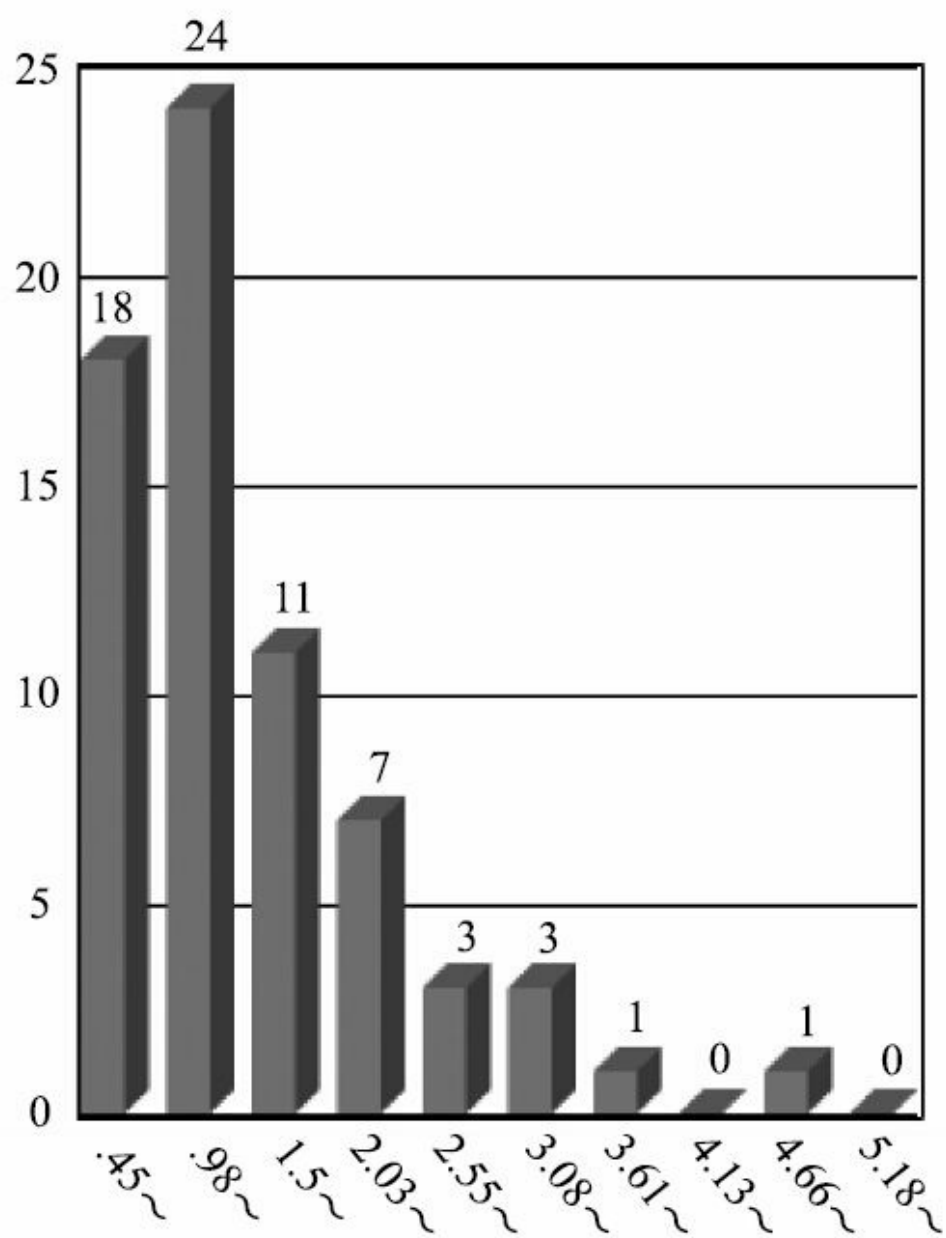


图 2-25

2.3 数据选择

上节我们介绍了如何从数据的业务关系、数据的技术表达和数据的数学特征等几个方面来了解数据。在对数据有了较全面而深入的了解之后，我们需要以数据挖掘的需求为目标选取适当的数据了。如何从数据挖掘角度思考数据的选取呢？我们应当考虑适当的数据规模和数据的代表性，然后再科学地进行数据的选取。

2.3.1 适当的数据规模

数据用于分析的规模是什么？是对总体进行分析还是对样本进行分析？我们先看一下总体和样本的定义。

总体：获取信息的全体对象，也可以说是要研究的是对象的全体。

样本：总体的一个子集，而且尽量是一个具有代表性的子集。从样本收集信息，以便对整个总体做某些结论。

统计学的一个核心思想是通常通过少量样本推测总体的某一个或多个维度或属性的特征。样本的目的一般是用来推测总体的，一锅汤是不是咸了，并不需要整个喝下去才知道，只要尝一勺就可以了。而数据挖掘的一个核心思想是通过尽量大的数据集甚至是总体本身来描绘总体，并可能根据描绘做出对未来的预测，或施加影响（人工干预）来达到希望的结果，这是数据挖掘工程更期望施展的方向。理论上的数据挖掘是描绘和预测，现实商业中的项目往往是“描绘-预测-影响”的循环。

但是，大千世界，各种各样的变化因素影响着我们要分析的数据：潜在变量的影响、变异无所不在、个体差异的影响、市场的正常波动、气温的时间变化、激素的脉冲影响等。我们会用基于数学的数据挖掘工具来处理这些动态变化中的大量数据。事实上，若想要加强我们对工作、生活中所遇到的大数据或大信息做出合理描绘或预测的

能力，仅仅通过提高和改进数据挖掘的算法是不够的。在数据挖掘的实际商业应用中，有诸多情况下需要对大数据进行科学地筛选和抽样，而不是对大数据本身直接应用数据挖掘算法，这里有很多原因，例如：

- 对大数据全集进行数据挖掘时间长、成本高。
- 大数据全集的数据质量不够高，混杂成分多，直接影响挖掘效果。
- 大数据全集的主题特征不易清晰地辨别，需要提炼后挖掘。
- 商业应用上需要及时（实时）得到挖掘结果的反馈和指导，需要提高数据挖掘的性能和效率。

因此，选取能代表大数据全集典型特征的样本数据成为某些数据挖掘商业应用的先决任务。我们数据选取的原则是，在选取的数据能够足以表达大数据集的主要特征的情况下，尽可能满足商业需求对数据挖掘性能的要求。

2.3.2 数据的代表性

首先，我们要考虑的是，不管样本是大的还是小的，无论是什么分析手段，要注意的是用于分析的样本是否能够代表总体。比如我们现在得到某种较少的疾病的若干案例，研究者使用这些案例来代表所有的患有该疾病的人群，作为我们研究该疾病的样本数据，实际上它往往是有偏倚的，而并不是精确意义上的样本，是从大量该疾病的案例中随机选取一些具有代表性的案例。

2.3.2.1 数据有偏

数据的代表性是数据可信性的一个方面，没有代表性的数据当然也是不可信的。我们常见的影响可信性的一个原因是抽取有偏。常见的抽取有偏的样本，一种是自发性回应，如例1；另一种是方便抽样，如例2。

例1：网络购物的差评！（自发性回应）。

我们注意到评论数据是来源于买家的评论，根据实际数据挖掘的行为实验，对购物满意度不高的买家更容易发布评论。发布评论是需要消耗时间和精力去完成的，如果不受到任何收益的诱导，差评数目应该大于好评。所以，对于评价商品而言该数据会有偏。

例2：摆在箱子上面的都是好梨子（方便抽样），如图2-26所示。再如，一盘菜上面都是牡蛎，下面看不见的都是生菜。

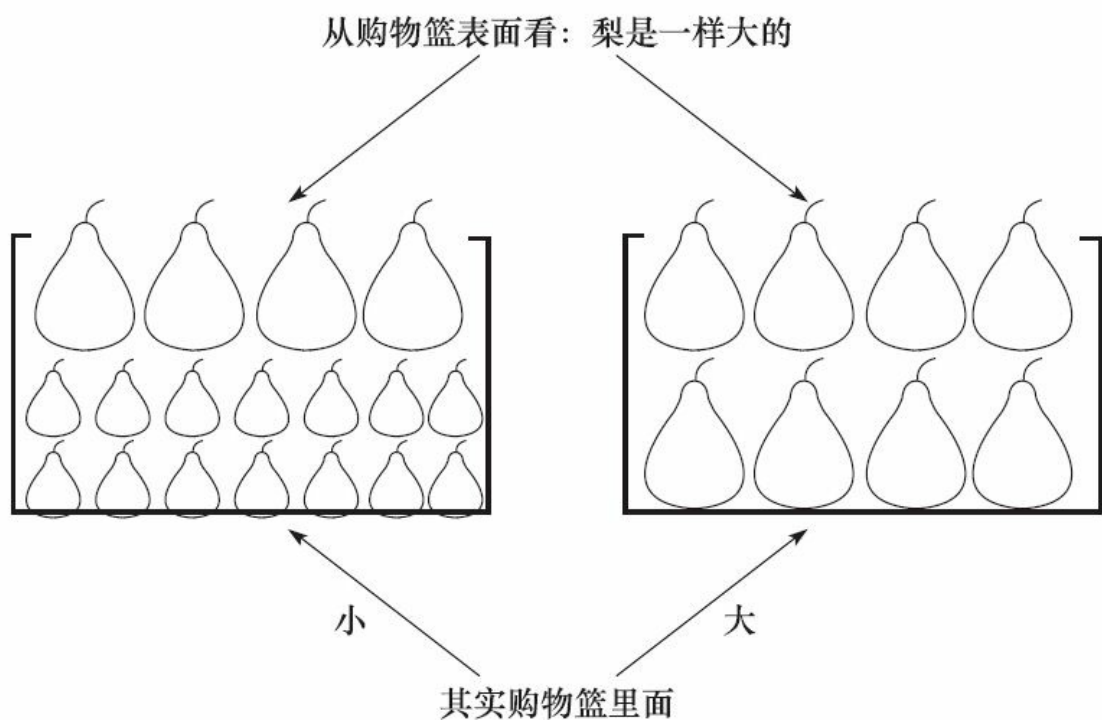


图 2-26

自发回应：由对某个诉求的自然回应形成。有强烈负面感觉的，比较会不嫌麻烦去回应，例如，去投诉的会比去表扬的多得多。

方便抽样：样本限于总体中易于抽到的一部分。比如从笼子里捉白鼠，被捉到的是反应速度慢的白鼠。

它的缺点是数据有偏，不具代表性。如买苹果，通常是看最上面的苹果来判断整箱苹果的好坏。

2.3.2.2 数据干扰

我们常见的影​​响可信性的原因，还有干扰。还需要注意的是收集数据的过程是否产生了对观测目标的干扰。例如，病人的心理预期对

疗效的干扰，心理预期差的病人比心理预期好的病人有较差的疗效。再如，购物场所调查中，调查对象的外表对调查目标的干扰，拥有比较和善或漂亮的外表的顾客更容易被调查到。

2.3.2.3 数据偏倚

影响可信性的常见原因还有偏倚。例如选择偏倚（selection bias），由于选入的研究对象与未选入的研究对象在某些特征上存在差异而引起的误差。这种偏倚常发生于数据分析的设计阶段，是指不同类型的研究对象，数据入选的机会不同。例如，健康工人效应，它指职业病研究中观察到工人总死亡率较一般人群为低的现象。这是由于有严重疾病或缺陷的人不能从事某些职业所致，如不考虑这一效应，而将工人死亡率与一般人群死亡率相比是不合适的，是选择偏倚的一种。

另外一种偏倚是信息偏倚，又称观察偏倚，指在收集整理资料阶段由于观察和测量方法上有缺陷产生的数据偏倚。例如，流行病学科研中的调查偏倚可能来自于调查对象及调查者双方。病例组和对照组的调查环境与条件不同，或者调查技术与质量不高，以及仪器设备的问题等均可产生调查偏倚。

2.3.3 数据的选取

怎样选取样本数据以便能最好地代表要分析的总体？一种方法是随机抽样，另一种方法是抽取总体本身。

2.3.3.1 随机抽样

我们不必吃完整头牛，才知道肉是老的！抽样的精髓就是从一部分来得知全体。人群之中研究调查对象的一部分人，选择这部分人并不是因为对他们特别感兴趣，而是因为他们具有代表性。

什么样的样本可以代表总体？首先，我们看一下图2-27，图中包括了数据提供者和数据抽样者。把信息想象成梨，数据提供者就是商人，数据抽样者是顾客。商人会尽量给小的，顾客会尽量挑大的。购物篮中的梨无法代表整箱里的梨，如图2-26所示，两个购物篮分别是大梨和小梨，均不能代表整箱里的梨。公平的方法是，在箱子里随机挑选，其结果很可能是有大有小。这里的随机不是随便，随机的意思是整个箱子（也就是梨的总体）里每一个梨都有可能被选中，也就是机会均等。随机抽样不许挑肥拣瘦，它让每个样本均有选中的机会，而且机会相等。例如随机数字和随机数字表，不同位置具有独立性。也就是说，知道了表的某一部分是什么数字，不会提供任何关于其他部分是什么数字的信息。

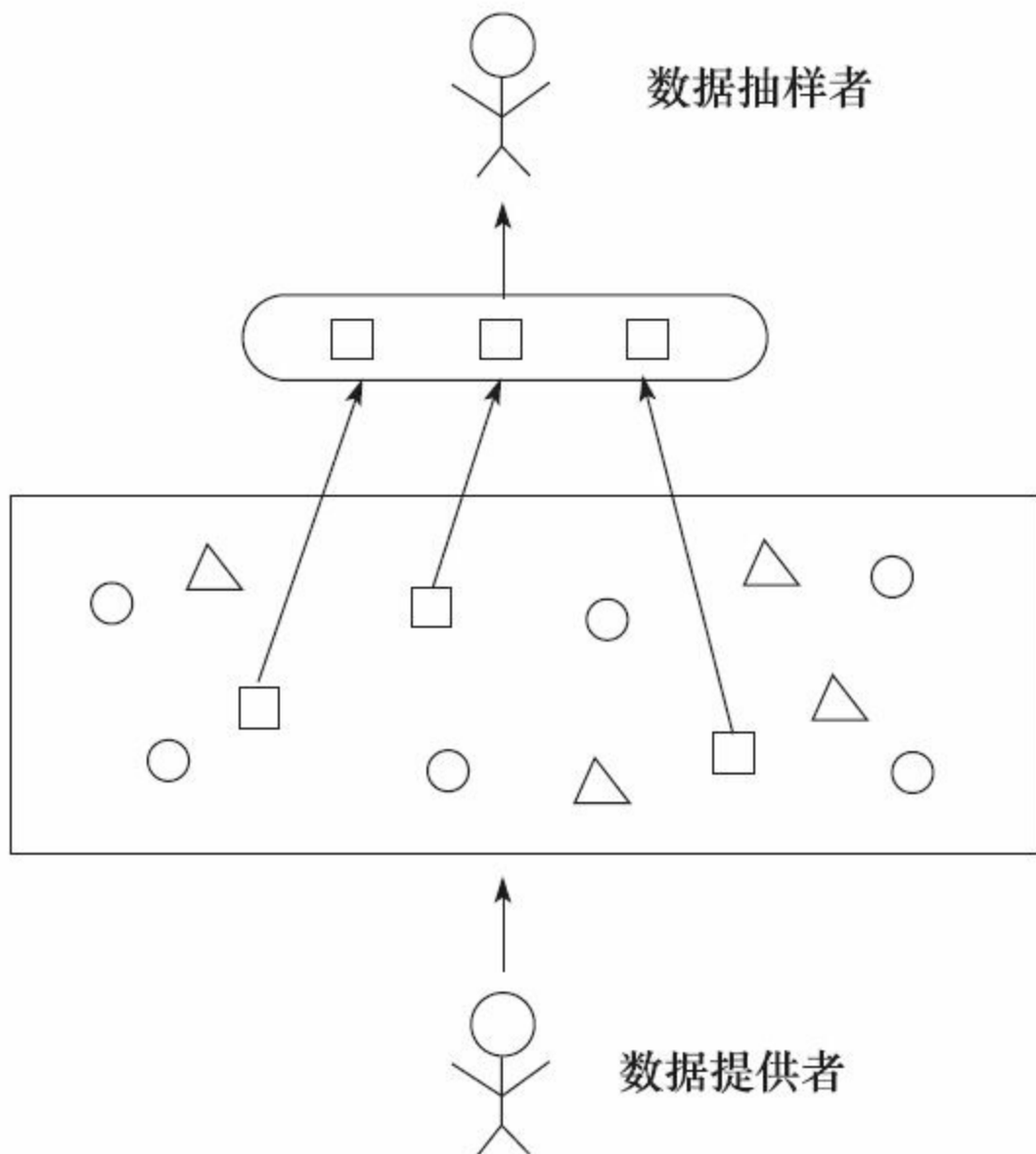


图 2-27

随机原则是利用不牵涉人为选择的机遇来选取样本，给每个个体同样的入选机会来消除偏差，如图2-28所示，随机样本避免的是人为选择带来的偏差。

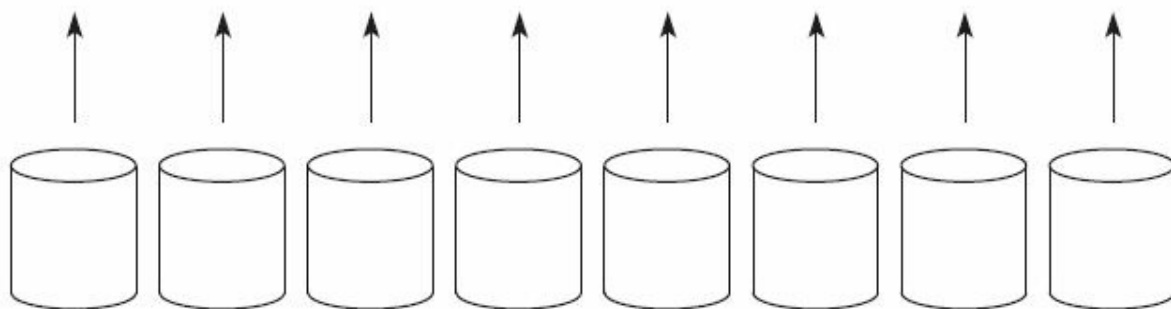


图 2-28

随机样本的两个优点：一是消除“偏心”。同一总体重复抽取许多大小一样的随机样本，所有样本的变异状况就会遵循某种可以预测的形态，如图2-29所示。

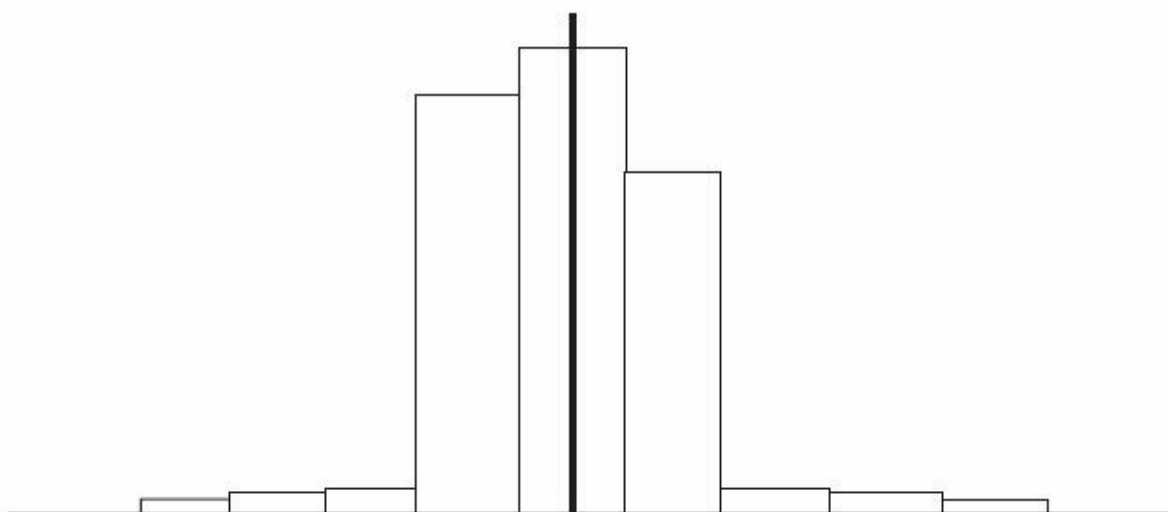


图 2-29

2.3.3.2 偏倚和变异

用样本估计总体，必然会遇到偏倚和变异。偏倚（**bias**）是指样本对总体真值的偏离，包括测量仪器的不准、样本过小、试验设计不合理、分配或分组不均衡、抽样未随机、测量者有主观倾向等。变异

（variability），是指在一个总体中，每个个体在相关属性上表现出的差异。从一个总体抽取的样本，即使是随机样本也会有偏倚和变异。从本质上理解，偏倚是后天造成的，而变异是先天形成的，但二者均会对估计总体特征的统计分析造成影响。

抽样误差是造成偏倚的最著名的一种，误差界限就是它的度量，我们可以选择随机抽样来控制抽样误差。抽样误差是抽样这个行动导致的误差，使用了糟糕的抽样方法，比如自发回应；涵盖不全，也就是漏掉了某些个体，如电话访问就会漏掉没安电话的人，再如人口普查会漏掉常常外出的人。除了抽样误差外，造成数据偏倚的还有随机误差和非抽样误差。所谓随机误差是选取样本时因为机会造成的样本统计量和总体参数的差距。而非抽样误差是指和抽样无关的误差，如：测量偏倚、诱导偏倚、记忆错误、回应误差，等等。

我们再来看数据的变异性。变异是统计的前提条件，有变异才有统计，没有变异也就没有统计的必要了。抽样与变异有着密切关系。例如，实际抽样往往只抽取一次，所以从大的随机样本算出的估计值，几乎会靠近真正值。从一个随机样本所得到的统计量的变异性，主要受随机样本量的影响，并不受总体大小影响（只要总体至少比样本大出足够的倍数）。因为，如果随机抽取的样本量小，则样本个体间的差异性显得更加明显，反之随着随机抽取的样本量的不断增加，数据就也趋向于正态分布，大多数样本间的差异性显得越来越小。显然，一个好的样本应该是尽量小的偏倚和尽量低的变异性。相应的处理方法是，利用随机抽样减低偏倚，抽取大样本来减低变异。

2.3.3.3 推测总体

样本没有办法告诉我们有关总体的准确信息。我们一般做的是把样本中推断的信息转换为总体的估计值。参数（Parameter）是描述总体的数字，我们实际无法知道参数的值。统计量（Statistic）是描述样本的数字。如图2-30所示，一旦选取了样本，统计量也就知道了。

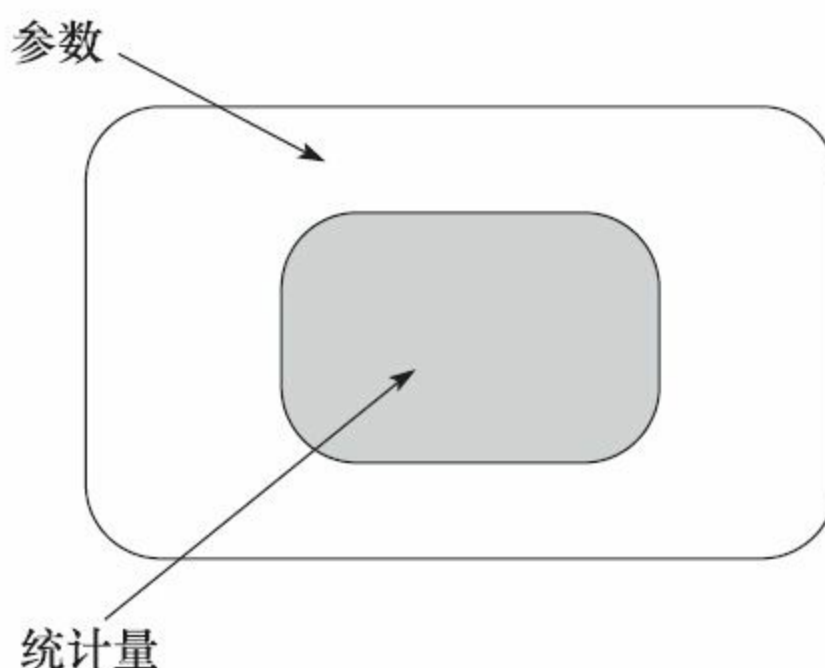


图 2-30

2.3.3.4 总体抽取

抽样的反面是普查，举个人口普查和总体的例子。人口普查时，把整个人口总体作为分析目标，期望把总体作为样本进行分析。但在实际人口普查时，总会漏掉一些人。当然在大数据时代以前，从经济角度看，抽样更合算一些。但是，生活在大数据时代，总体并不是可

望而不可及的了。普查是企图把整个总体纳入的调查，但直接研究总体有诸多缺点：

- 总体过大，花费高。这很容易理解，在此无需举例了。
- 容易有遗漏，同样也可能会把不是该总体的样本包括其中，即误纳。
- 在开始研究的时候总体还不存在。比如使用新药，如果把所有使用该新药的人都算总体的话，在开始的时候总体还不存在，而且总体会随着时间而越来越大，我们很难收集齐他们。
- 无法搞破坏性调查。比如测鞭炮是不是响，只有点燃了它才知道；保险丝会不会断，只有通上电后观察它是否断了等等。
- 获得结果的时间比较长。这很容易理解，在此也无需举例了。

2.4 数据处理

商业数据挖掘在应用算法以前，需要对数据进行大量处理工作，以便为相应的挖掘算法的成功实施提供数据准备。这一步对数据挖掘的结果有着非常重要的影响。常用的数据挖掘预处理方法主要包括：

- 数据生成：用单列或者多列根据某些规则生成新数据。
- 数据拆分：将待挖掘的数据分为两组或多组，一组用于训练模型，另外一组用于测试数据。
- 数据清理：将数据进行质量控制，即去除一些所谓的脏数据。
- 数据汇总：通常计算数据的总和、平均值、计数等信息。
- 数据规范化：同比例地更改数据的值以便满足要求，如要求在范围（-1.0~+1.0）之间、最小-最大规范化、z-score规范化、小数定标规范化等。
- 缺失值处理：删除、忽略有缺失值的数据，应用常量、均值来替换缺失值，手动添加不同缺失值，应用预测值来替换缺失值等。
- 连续变量离散化：如分箱（Binning）、等频划分、等宽划分、直方图分析、基于熵的离散化等。
- 噪声数据或离群点处理：应用一些简单的数学方法或使用异常侦测、聚类分析等算法处理异常数据。

- 数据泛化：用高水平数据取代低水平数据，例如用“城市”值取代地址列的“街道”值等。

下面，我们将重点介绍数据标准化和数据离散化方面的内容。

2.4.1 数据标准化

由于不同变量常常具有不同的单位和不同的变异程度。不同的单位常使系数的实践解释发生困难。例如：第一个变量的单位是kg，第二个变量的单位是cm，那么在计算绝对距离时将出现将两个事例中第一个变量观察值之差的绝对值（单位是kg）与第二个变量观察值之差的绝对值（单位是cm）相加的情况。使用者会说5kg的差异怎么可以与3cm的差异相加？不同变量自身具有相差较大的变异时，会使在计算出的关系系数中不同变量所占的比重相差很大。例如，如果第一个变量（两水稻品种米粒中的脂肪含量）的数值在2%到4%之间，而第二个变量（两水稻品种的亩产量）的数值范围都在1000与5000之间。为了消除量纲影响和变量自身变异大小对数值大小的影响，故需要将数据标准化。

2.4.1.1 变量离差标准化

离差标准化是某变量中的观察值减去该变量的最小值，然后除以该变量的极差，即：

$$x_{ik} = \frac{x_{ik} - \min(x_k)}{R_k} \quad (2.6)$$

经过离差标准化后，各种变量的观察值的数值范围都将在〔0，1〕之间，并且经标准化的数据都是没有单位的纯数量。离差标准化是消除量纲（单位）影响和变异大小因素的影响的最简单的方法。有

一些关系系数（例如绝对值指数尺度）在定义时就已经要求对数据进行离差标准化，但有些关系系数的计算公式却没有这样要求，当选用这类关系系数前，不妨先对数据进行标准化，看看分析的结果是否为有意义的变化。

2.4.1.2 变量标准差标准化

标准差标准化是将某变量中的观察值减去该变量的平均值，然后除以该变量的标准差，即：

$$x_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k} \quad (2.7)$$

经过标准差标准化后，各变量将有约一半观察值的数值小于0，另一半观察值的数值大于0，变量的平均值为0，标准差为1。经标准化的数据都是没有单位的纯数量。对变量进行的标准差标准化可以消除量纲（单位）影响和变量自身变异的影响。但有人认为经过这种标准化后，原来数值较大的观察值对分类结果的影响仍然占明显的优势，应该进一步消除大小因子的影响。尽管如此，它还是当前用得最多的数据标准化方法。

2.4.2 数据离散化

数据离散化，即是将充满实数轴的数据划分为区间，它在数据挖掘的过程中具有重要的作用：

- 离散化结果将会减少给定连续变量值的个数，减少和简化原来的数据。
- 离散变量相对于连续变量来说更接近于知识层面的表示。
- 一些算法只能应用于离散型数据，使得离散化处理成为必要，而离散化又使很多算法的应用范围扩展了。

2.4.2.1 离散化概念

连续变量的离散化，就是具体性的问题抽象为概括性的问题，即是将它取值的连续区间划分为小的区间，再将每个小区间重新定义为一个唯一的取值。例如，学生考试结束后有了成绩，我们可以根据成绩给出划分两个区间， $[0, 60]$ 为不及格， $[60, 100]$ 为及格。60 是两个区间的分界点，称为断点。断点就是小区间的划分点，区间的一部分数据小于断点值，另一部分数据则大于等于断点值。选取断点的方法不同，从而产生了不同的离散化方法。

2.4.2.2 离散化处理过程

对连续变量进行离散化处理，一般经过以下步骤：

- 1) 对此变量进行排序。
- 2) 选择某个点作为候选断点，根据给定的要求，判断此断点是否满足要求。
- 3) 若候选断点满足离散化的要求，则对数据集进行分裂或合并，再选择下一个候选断点。
- 4) 重复步骤2) 和3)，如果满足停止准则，则不再进行离散化过程，从而得到最终的离散结果。

2.4.2.3 离散化常用方法

1. 无监督方法（分箱法）

分箱法包括等宽分箱法和等频分箱法，它们基本的离散化算法。分箱的方法是基于箱的指定个数自顶向下的分裂技术，在离散化的过程中不使用类信息，属于无监督的离散化方法。等宽分箱法，就是根据箱的个数得出固定的宽度，使得分到每个箱中的数据的宽度是相等的。等频分箱法，是使得分到每个箱中的数据的个数是相同的。在等宽或等频划分后，可用每个箱中的中位数或者平均值替换箱中的所有值，实现特征的离散化。

这两种方法简单，易于操作，但也有缺点：

- 需要人为给定区间个数；
- 等宽法分箱的时候容易出现不均匀的现象，它强调的是区间的长度，没有考虑数据点的个数，对异常点比较敏感；
- 等频法分箱的时候，可能导致把同类别的点分到不同的区间，它强调的是数据点的个数；

考虑到以上问题，在使用等宽法之前，首先进行异常侦测。对于等频法可能将具有相同类标签的实例分入不同箱中的问题，可用的解决方法是，先用等频法将特征值进行分箱，然后对各个相邻分箱的边界值进行调整，使得相同的值可以被分入同一个箱中。

2.有监督方法

1) 基于熵的离散化方法

熵是最常用的离散化度量之一，基于熵的离散化方法，使用类分布信息计算和确定分裂点，是一种有监督的、自顶向下的分裂技术。方法描述如下：

步骤1：定义区间的熵；

步骤2：把每个值看成分割点，将数据分成两部分，在多种可能的分法中寻找一种产生最小熵的分法；

步骤3：在分成的两个区间中，找较大熵的区间，继续步骤1；

步骤4: 满足用户指定个数时, 结束过程。

2) 基于卡方的离散化方法

基于卡方的离散化方法, 是采用自底向上的策略。首先将数据取值范围内的所有数据值列为一个单独的区间, 再递归地找出最佳邻近可合并的区间, 然后合并它们, 进而形成较大的区间。

常用的基于卡方的离散化方法是ChiMerge方法, 其过程简要如下:

步骤1: 首先将变量的每个不同值看作一个区间, 对相邻区间计算卡方统计量, 根据给定的置信水平的阈值, 判定其独立性, 将非独立的区间合并为一个区间;

步骤2: 递归步骤1, 直到根据给定的阈值找不到相邻可以合并的区间为止。

2.5 统计学算法的数量条件

正如前面一节谈到的，在实际商业数据挖掘应用中，有诸多情况需要对大数据进行科学地样本选取和抽样，选取能代表大数据全集的典型特征的样本数据成为某些商业应用数据挖掘的首要任务。而在数据挖掘算法的应用中，我们在有些情况下会用到统计学的一些方法，例如前面一节提到的基于卡方的离散化方法就是实际应用中很常见的一个数据准备算法。在商业数据挖掘应用中和实验性的数据挖掘工作中，常用的统计学算法包括检验样本均值的各种T检验、F检验、 χ^2 检验、线性回归、线性相关、对照分析等。通常情况下，我们挖掘的数据量一般比较大，因此可以应用很多统计学方法。但是，我们不能忽略了很多统计学方法对分析的数据量有一定的要求，而且很多人似乎无条件地使用这些算法，这对于一些实验性的数据挖掘任务似乎短时间内看不到负面影响，但对于时刻依赖数据挖掘结果的商业应用来说，有时在极端情况下无疑是一种灾难。所以我们这里有必要讨论一些在数据挖掘中常用的统计方法对最小数据量的要求，主要内容涉及：样本量估计概念、单样本总体均值比较的样本量估计、两样本总体均值比较的样本量估计、多样本总体均值比较的样本量估计、区组设计多样本总体均值比较的样本量估计直线回归与相关的样本量估计和对照分析的样本量估计等方面。

2.5.1 样本量估计概念

统计学方法传统上是用于科学实验研究或调查分析任务的，因此，我们在此沿用了这方面的相关术语，如实验研究、样本量、样本量估计、假设检验、检验效能，等等。而在数据挖掘应用中，有关术语的含义略有不同，对以上列举的术语，相对应地可以理解为挖掘应用、待挖掘的数据集、统计学算法所需的最小数据量、统计学验证、算法性能等。

- 样本量：指实验研究和调查研究中样本的观察单位数，又称样本大小。

- 样本量估计：指应用一定的统计方法在保证研究结论具有一定可靠性（精度与检验效能）的前提下所确定的最小样本例数。

- 样本量大小的效果：样本量过小，指标不稳定、检验功效低；样本量过大，增加研究困难、难于控制条件、浪费人力物力。

样本量估计的影响因素包含下列五个：

- 第一类错误概率 α 的大小。
- 第二类错误概率 β 或检验的效能 $(1-\beta)$ 的大小。
- 容许误差 δ 。
- 总体标准差 σ 或总体率 π 。

· 单双侧检验与设计类型。

α 的大小：本次研究允许的第一类错误概率； α 越小，所需样本量越大；一般取0.05或0.01； α 取值有单双侧之分，双侧检验比单侧检验所需样本量更多。

第一类错误：假设检验中，原假设为真而被拒绝的错误，也称假阳性错误或弃真错误。

β 的大小：检验效能（ $1-\beta$ ）越大，所需样本量越大；检验效能一般取0.8以上； β 是第二类错误概率。

第二类错误：假设检验中，原假设为不真而被接受的错误，也称假阴性错误或存伪错误。

容许误差 δ ：研究者要求的或客观存在的样本统计量与总体参数间的差值；容许误差值越小，所需样本量越大。

总体标准差 σ 或总体率 π ：标准差 σ 反映了数据的变异度，值越大，所需样本量越大； π 越接近0.5，所需样本量越大。

2.5.2 单样本总体均值比较的样本量估计 (T-Test)

单样本总体均值比较的样本量估计，用于单样本总体均值与某个特定值的比较，和配对设计中两均值的比较，适用于单样本均值的T检验（one sample t-test）和配对样本均值的T检验（paired t-test）。假设检验：单因素设计中与总体均值的比较、配对设计中两均值的比较。单侧假设检验的公式如下：

$$n = \frac{(t_{\alpha} + t_{\beta})^2 \sigma^2}{\delta^2} \quad (2.8)$$

例如IMRS中统计分析子系统的单因素单水平模块，对2型糖尿病患者的血清葡萄糖与正常值上限7.8比较，要确定最小样本量。首先选择相关数据，如图2-31所示，选择疾病人群为2型糖尿病，检验条件为入院首次化验，检验项为血清葡萄糖，选择时间跨度为3天。

描述统计数据选择

观察人群：2型糖尿病(非肥胖)(肥胖) (必选)

检验条件：入院首次化验 (必选)

检验项：★血清葡萄糖 (必选)

检验时间跨度：3 (天)

生成数据

系统说明

在“描述统计数据选择”功能区中选择目标疾病和目标检验项，也可以点击“重置”按钮来重新选择。然后点击“生成数据”按钮生成分析数据。点击“下一步”按钮，查看不同数据的描述统计结果、频数分布表和直方图。

图 2-31

点击图中的“生成数据”按钮，系统从数据库中自动选取所需数

据。其次确定所需的最小样本数，根据最小样本量的计算公式

$$n = \frac{(t_{\alpha} + t_{\beta})^2 \sigma^2}{\delta^2} \tag{2.9}$$

需要知道样本的标准差和总体误差，才能算出最小样本量。运行系统的描述性统计功能，即可得到如图2-32所示的描述统计结果。

入院首次化验数据描述统计结果

中文名 ▲	英文名	统计量值
百分之二十五分位数	Quartile_25	7.71
百分之九十五分位数	Quartile_95	23.010000000000002
百分之七十五分位数	Quartile_75	15.699999999999999
百分之五分位数	Quartile_05	5.0499999999999998
变异系数	Cv	45.83%
标准差	Stddev	5.5547
方差	Variance	30.855
极差	Range	29.459999999999999
均值	Mean	12.12
中位数	Median	10.9
最大值	Max	31.379999999999999
最小值	Min	1.9199999999999999

1 - 12

下载列表

图 2-32

图2-33中所示的差值标准差为σ，即图2-32所计算出的标准差5.5547，差异值为δ，即图2-32所计算出的均值12.12减去血清葡萄糖与正常值上限7.8，为4.32。选择显著水平α和检验能效β，计算出最小样本量为20例。如图2-33所示。

功能项

* 请选择功能项 5-单因素单水平

重置

单因素单水平

☐ $H_0: u = u_0 \text{ 和 } H_1: u > (<) u_0$
☒ $H_0: u = u_0 \text{ 和 } H_1: u \neq u_0$

* 请选择显著水平 α 0.05

* 请选择检验效能 $(1-\beta)$ 0.9

* 请输入差值标准差 5.5547

* 请输入差异值 4.32

样本量估计

样本均数与血糖正常值的
上限7.8的差值

单因素单水平样本量估计结果

双侧检验时，样本量=20

图 2-33

2.5.3 两样本总体均值比较的样本量估计 (T-Test)

两样本总体均值比较的样本量估计，用于两独立样本均值的比较，适用于两独立样本均值的T检验和两总体方法不等的近似T检验。假设检验：完全随机设计的两总体均值的比较。计算公式：

$$n = \frac{2\sigma^2(t_{\alpha} + t_{\beta})^2}{(u_1 - u_2)^2} \quad (2.10)$$

其中当 σ 未知时，通常用样本标准差估计，取合并方差的平方根或两个样本标准差中大的一个。

例如，分析冠状动脉粥样硬化性心脏病和脑梗死患者的超敏C反应蛋白的差异是否有统计学意义，经预实验得到以下数据，若要得出是否有差异的结论，需要计算最小的样本量。首先根据图2-34选取数据，病例组选择为冠状动脉粥样硬化性心脏病，对照组为脑梗死，观察条件为入院首次化验，观察项为超敏C反应蛋白，化验时间跨度为3天，本次筛选集名称为1009023，总计筛选了184条数据。

说明
样本分布
样本量估计
样本筛选
样本导出
用药信息
检验信息
体

素双水平

单因素双水平

- * 病例组： (必选)
- * 包含并发症： (可选) 当选择多个疾病时，
- * 不包含并发症： 用'%'分隔如（高血压%心脏病）
- * 对照组： (必选)
- * 包含并症： (可选) 当选择多个疾病时，
- * 不包含并症： 用'%'分隔如（高血压%心脏病）
- * 请选择性别： (必选)
- * 请选择年龄点： (可选) (年龄点之间用逗号隔开,如果为空默认全部年龄,例如 10,20,30)
- * 选择的年龄段为： 全部年龄段
- * 观察条件： (必选)
- * 观察项： (必选)
- * 化验时间跨度：
- * 请输入本次筛选名称： (必选)
- * 本次筛选数据量为： 184条

图 2-34

再如图2-35所示，首先计算两种病患的超敏C反应蛋白的标准差和均值。

再根据公式

$$n = \frac{2\sigma^2(t_\alpha + t_\beta)^2}{(u_1 - u_2)^2} \quad (2.11)$$

算出最小样本量，图2-36中的第一水平标准差和第二水平标准差为两种疾病超敏C反应蛋白的标准差，用于估算 σ 。差异值为两种疾病的

均值之差，即 $\mu_1 - \mu_2$ 。选择显著水平 α 的值和检验能效 β 的值，计算出最小样本量为783例。

2.5.4 多样本总体均值比较的样本量估计 (F-Test)

多样本总体均值比较的样本量估计，用于多个独立样本均值的比较，适用于多个独立样本均值的F检验（f-test, analysis of variance, one-way ANOVA）。假设检验：完全随机设计的多个总体均值相等。计算公式如下：

$$n = \frac{\psi^2 (\sum S_i^2 / k)}{\sum (\bar{X}_i - \bar{X}) / (k - 1)} \quad (2.12)$$

其中， $\bar{X} = \sum \bar{X}_i / k$ 且 k 是组数， ψ 是界值。

单因素双水平

- 病例组: 冠状动脉粥样硬化性心脏病 (必选)
- 包含并发症: (可选) 当选择多个疾病时, 用%分隔如(高血压%心脏病)
- 不包含并发症: (可选) 当选择多个疾病时, 用%分隔如(高血压%心脏病)
- 对照组: 脑梗死 (必选)
- 包含并发症: (可选) 当选择多个疾病时, 用%分隔如(高血压%心脏病)
- 不包含并发症: (可选) 当选择多个疾病时, 用%分隔如(高血压%心脏病)
- 请选择性别: 男,女 (必选)
- 请选择年龄点: (可选) (年龄点之间用逗号隔开, 如果为空默认全部年龄 例如 10,20,30)
- 选择的年龄段为: 全部年龄段
- 观察条件: 入院首次化验 (必选)
- 观察项: 超敏C反应蛋白 (必选)
- 化验时间跨度: 3
- 请输入本次筛选名称: 1009023 (必选)
- 本次筛选数据量为: 184条

描述统计数据选择

请选择筛选数据集。

请选择筛选数据集名: 1009023 (必选)

描述统计

描述统计分析结果, 本系统给出几个基础统计量, 频数分布表和频数分布图。

请选择疾病: 冠状动脉粥样硬化性心脏病

化验值描述统计结果

中文名称	英文名称	统计值
百分之二十五分位数	Quantile_25	.815999999999999995
百分之七十五分位数	Quantile_75	2
百分之二十五分位数	Quantile_25	6.190000000000000004
百分之七十五分位数	Quantile_75	292999999999999998
变异系数	Cv	296.37%
标准差	Stddev	30.8026
方差	Variance	948.7978
极差	Range	256.846
均值	Mean	10.3932
中位数	Median	1.94

描述统计

描述统计分析结果, 本系统给出几个基础统计量, 频数分布表和频数分布图。

请选择疾病: 脑梗死

化验值描述统计结果

中文名称	英文名称	统计值
百分之二十五分位数	Quantile_25	1.189999999999999999
百分之七十五分位数	Quantile_75	93.900000000000000006
百分之二十五分位数	Quantile_25	22.699999999999999999
百分之七十五分位数	Quantile_75	3260000000000000001
变异系数	Cv	159.84%
标准差	Stddev	26.2197
方差	Variance	687.4722
极差	Range	101.692
均值	Mean	15.4375
中位数	Median	4.465
最大值	Max	102
最小值	Min	.308

图 2-35

功能项

* 请选择功能项 **6-单因素双水平** 重置

单因素双水平

☐ $H_0: u = u_0 \text{ 和 } H_1: u > (\text{或 } <) u_0$ ☒ $H_0: u = u_0 \text{ 和 } H_1: u \neq u_0$

* 请选择显著水平 α **0.05**

* 请选择检验效能 $(1-\beta)$ **0.9**

* 请输入第一水平的标准差 **30.8026**

* 请输入第二水平的标准差 **26.2197**

* 请输入差异值 **5.0443**

样本量估计

单因素双水平样本量估计结果

双侧检验时, 每组所需样本量为**783**

图 2-36

例如，S病患者分别使用A药、B药和C药，分析三种药的治疗效果的差别是否具有统计学差异，经预实验得到以下三个组的均值和标准差，若要得出是否有差别的结论，则需要计算最小样本量。

计算出三种药的标准值和均值如图2-37所示，再根据公式

$$n = \frac{\psi^2 (\sum S_i^2 / k)}{\sum (\bar{X}_i - \bar{X}) / (k - 1)} \quad (2.13)$$



图 2-37

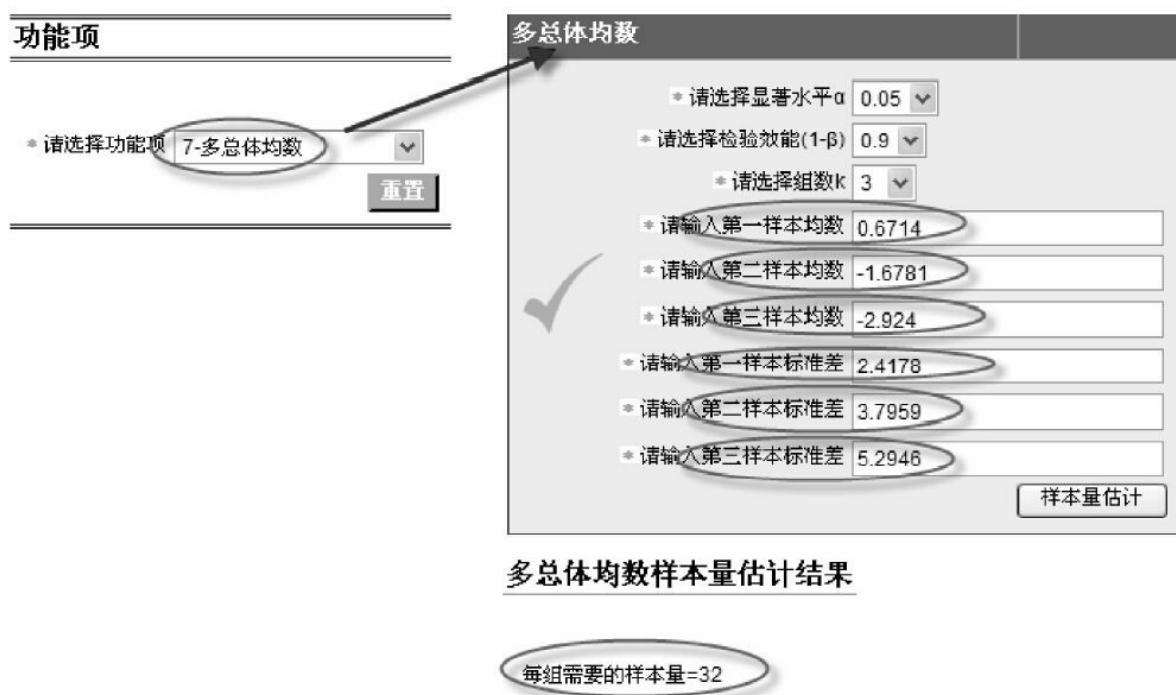


图 2-38

计算最小样本量。如图2-38所示，依次输入三种药的均值和标准差，选择显著水平 α 的值和检验能效1- β 的值，系统会依据公式计算出所需的最小样本量，计算结果为32例。

2.5.5 区组设计多样本总体均值比较的样本量估计 (F-Test)

区组设计多样本总体均值比较的样本量估计，用于基于区组设计的多个独立样本均值的比较，适用于基于区组设计的多个独立样本均值的F检验（f-test, analysis of variance, two-way ANOVA）。假设检验：计量资料的随机区组设计的多个总体均值相等。计算公式：

$$n = \frac{2 \times MS_e \times (Q + u_\beta)^2}{D^2} \quad (2.14)$$

其中， MS_e 是误差均方， D 为处理组间的差值（取差值最小值）。

例如，S病患者分别使用A药、B药和C药，对年龄分组，分析三种药的治疗效果有无差别，年龄对治疗有无影响，预实验得到图2-39左图的数据，方差分析结果如图2-39右图，每组需多少样本量？

数据列表

年龄组	第一种用药	第二种用药	第三种用药
17 至 70岁	-1.6	-1.375	-6.65
70 至 85岁	3.15	-0.4778	-2.1
85 至 99岁	0	-4.8429	-2.4429

1-3

下载列表

方差分析报表

离差来源	平方和	自由度	均方	F值	P值
处理因素(用药)	27.85	2	13.93	3.46	>0.05
区组因素(年龄)	19.02	2	9.51	2.36	>0.05
误差	16.13	4	4.03	0	0

1-3

下载列表

图 2-39

三组用药分别取均值 $(-1.6+3.15)/3$ ， $(-1.375-0.4778-4.8429)/3$ ， $(-6.65-2.1-2.4429)/3$ ，分别求差值，取最小差值为

1.499。

最小样本量的计算公式为

$$n = \frac{2 \times MS_e \times (Q + u_\beta)^2}{D^2} \tag{2.15}$$

由上述可知D为1.499， MS_e 为4.03，组数为3，如图2-40中输入相对应的数字，选择显著水平 α 的值和检验能效 $1-\beta$ 的值，计算出最小样本量为79。

功能项

* 请选择功能项

6-单因素双水平

重置

单因素双水平

☐ $H_0: u = u_0$ 和 $H_1: u > (<) u_0$

☒ $H_0: u = u_0$ 和 $H_1: u \neq u_0$

* 请选择显著水平 α

0.05

* 请选择检验效能 $(1-\beta)$

0.9

* 请输入第一水平的标准差

30.8026

* 请输入第二水平的标准差

26.2197

* 请输入差异值

5.0443

样本量估计

单因素双水平样本量估计结果

双侧检验时，每组所需样本量为783

图 2-40

2.5.6 直线回归与相关的样本量估计

直线回归与相关的样本量估计，用于线性回归（linear regression）和线性相关（linear correlation）分析。假设检验：线性回归系数与线性相关系数的假设检验。计算公式：

$$n = 4 \left[(u_{\alpha/2} + u_{\beta}) / \ln \left(\frac{1+r}{1-r} \right) \right]^2 + 3 \quad (2.16)$$

例如，分析G病患者的A检验项和B检验项是否有线性关系，首先要知道所需的最小样本量。如图2-41所示，经预实验得到相关系数 r 为0.924924，选择显著水平 α 的值和检验能效 $1-\beta$ 的值，再根据最小样本量估计公式，算出最小样本量为7。

The screenshot shows a software interface for sample size estimation. At the top, there is a '相关分析' (Correlation Analysis) section with a 'SPEARMEN秩相关分析' button. Below it, a search bar shows 'Spearman秩相关分析' and a value '924924' is entered. The main section is titled '直线回归与相关' (Linear Regression and Correlation). It contains several input fields: '请选择显著水平 α ' (Select significance level α) set to 0.05, '请选择检验效能 $1-\beta$ ' (Select test power $1-\beta$) set to 0.9, and '请输入相关系数' (Enter correlation coefficient) set to 0.924924. There are also radio buttons for hypothesis testing: 'H0:u=u0和H1:u>(或<)u0' and 'H0:u=u0和H1:u≠u0'. A '样本量估计' (Sample size estimation) button is at the bottom right. Below the main form, the text '直线回归与相关样本量估计结果' (Linear regression and correlation sample size estimation results) is displayed, followed by a circled result: '双侧检验时，每组所需样本量为7' (For two-sided test, the required sample size for each group is 7).

图 2-41

2.5.7 对照分析的样本量估计

对照分析的样本量估计用于对照分析中，适用于基于对照分析的优势比卡方 χ^2 检验（OR χ^2 -test, odds ratio chi-square test）。

估计总体的比数比时的样本量估计的计算公式如下：

$$n = \frac{u_{\alpha/2}^2 \{ 1/[\pi_1(1 - \pi_1)] + 1/[\pi_2(1 - \pi_2)] \}}{[\ln(1 - \varepsilon)]^2} \quad (2.17)$$

对比数比进行假设检验时，样本量估计的计算公式如下：

$$n = \frac{[u_{\alpha/2} \sqrt{2p_2(1 - p_2)} + u_{\beta} \sqrt{p_1(1 - p_1) + p_2(1 - p_2)}]^2}{(p_1 - p_2)^2} \quad (2.18)$$

例如，冠状动脉粥样硬化性心脏病和体检人群，暴露因素是发病年龄，从70岁到99岁，比较其暴露率，由统计分析子系统的病例对照模块得到的数据如图2-42所示，假设检验OR为1，则需多大的样本量？

显示行列表1

组别	暴露人数	非暴露人数	合计人数
冠状动脉粥样硬化性心脏病	247	93	340
体检人群	1821	9620	11441
合计	2068	9713	11781

1 - 3

图 2-42

计算出病例组的暴露率为 $247/340=0.726$ ，对照组的暴露率为 $1821/11441=0.159$ ，得到了 p_1 和 p_2 。再根据比例样本量估计公式计算最小样本量，如图2-43所示，选择显著水平 α 的值和检验能效 β 的值，输入对照组暴露率和病例组暴露率，即 p_1 和 p_2 ，系统就会计算出最小样本量为14。

图 2-43

上述例子说明了，对于不同的实验目的，首先要选择正确的统计方法，其次再根据统计方法去进行实验设计。而实验成功的基本要求就是要根据实验设计，计算出所需的最小样本量，只有这样，才能得出正确的结论。

在数据挖掘应用中，我们有时会用到一些统计学的方法进行数据处理或分析，如对两个数据的分布差异性进行比较，对数据中两个变量的相关性进行分析等。对一些比较特殊甚至极端的情况下，所分析的数据量很小，这时我们需要对所应用的统计学算法对样本量的前提

要求进行验证，才能保证数据挖掘相应环节的正确和可靠，并保证整个数据挖掘流程和项目的分析质量。

2.6 数据探索应用

上面的章节，介绍了在应用数据挖掘相关算法深入分析数据之前的一些准备工作，包括对数据的了解、对数据的选取和对数据的处理等。这些工作作为数据挖掘工程的一部分，通常情况下都是在后台完成的。本节我们将介绍运用一些统计学方法和商务智能工具对数据进行探索的应用，这些应用是在前台展现给用户的，目的也是帮助用户快速了解数据的主要特征和规律，为深入的数据挖掘分析做好铺垫工作，我们把这项工作也称为数据挖掘初探。

数据探索是个快速的概览过程，迅速地获取简略的特征信息。相对于传统意义的数据挖掘，快速和概括是数据挖掘初探最大的特点。一方面，数据探索的结果决定了数据挖掘的方法；另一方面，数据探索为数据挖掘算法提供了运算方向的参考。我们以IMRS的数据探索子系统（详见第5章）为例，说明实践中数据探索是如何发挥作用的。本节将介绍检验项的疾病分布、疾病中检验项的分布、成对检验项的相关分析、两种药物的应用分析等几个数据探索的实践应用。

2.6.1 检验项的疾病分布

通过对各种实验室指标在所有疾病中的分布分析，可以揭示实验室指标与疾病之间存在的特定关系。通过实验室指标来展示不同疾病的特征，并发现日常工作中不易察觉的规律和特征，从而引发科研的新线索和新思路。如图2-44所示，探索各种疾病中对特定的生理指标的影响，按照相关化验指标的高（H）、正常（N）、低（L）把疾病分类，并自动给出均值和标准差。

检验值：丙氨酸氨基转移酶, 比例：.32%

疾病名	例数	平均值	标准差	检验值区间
体检人群	216	3.2917	.9012	低

检验值：丙氨酸氨基转移酶, 比例：11.39%

疾病名	例数	平均值	标准差	检验值区间
不稳定性心绞痛，Brawnwald II级	2	130	21.2132	高
下肢动脉粥样硬化闭塞症	4	48.5	8.3865	高
体检人群	7552	65.13	38.2039	高

检验值：丙氨酸氨基转移酶, 比例：87.66%

疾病名	例数	平均值	标准差	检验值区间
颅内静脉窦血栓形成	3	16.6667	2.3094	正常
体检人群	58197	18.9515	7.9528	正常
输卵管妊娠流产伴腹腔内出血	2	8.5	3.5355	正常
恶性肿瘤化疗后的随诊检查	2	12.5	2.1213	正常
胆囊息肉	2	34	8.4853	正常
白血病化疗	2	9.5	4.9497	正常
丘脑出血	2	28.5	12.0208	正常
老年性白内障未熟期	4	16.25	8.3815	正常
下肢动脉粥样硬化闭塞症	4	18.5	7.5939	正常
2型糖尿病	2	27.5	2.1213	正常
输卵管妊娠破裂伴腹腔内出血	2	18.5	2.1213	正常
直肠癌术后化疗	2	26	15.5563	正常
乳腺癌术后化疗	3	17.3333	5.5076	正常
胆源性胰腺炎	2	30.5	6.364	正常

图 2-44

图2-44是运用BI技术开发的智能报表。我们截取了智能报表的一部分。报表显示，实验室指标丙氨酸氨基转移酶在所有疾病中的分布情况。其中，丙氨酸氨基转移酶出现低异常的体检人群有216例，其均值为3.29，标准差为0.90；丙氨酸氨基转移酶出现高异常的体检人

群有7552例，其均值为65.13，标准差为38.20；丙氨酸氨基转移酶出现高异常的下肢动脉粥样硬化闭塞症有4例，其均值为48.50，标准差为8.39；丙氨酸氨基转移酶正常的体检人群有58197例，其均值为18.95，标准差为7.95；丙氨酸氨基转移酶正常的下肢动脉粥样硬化闭塞症有4例，其均值为18.50，标准差为7.59。

我们进行横向比较发现，在丙氨酸氨基转移酶高异常的情况下，体检人群的平均值和标准差均高于下肢动脉粥样硬化闭塞症的平均值和标准差（当然，由于本例数据采集的有偏，体检人群和下肢动脉粥样硬化闭塞症的样本量相差悬殊，这里仅做示意性比较，真实医学比较见于医院现场数据）；同时，在丙氨酸氨基转移酶正常的情况下，体检人群的均值和标准差和下肢动脉粥样硬化闭塞症的均值和标准差非常接近。我们再进行纵向比较发现，对于体检人群的丙氨酸氨基转移酶来说，低异常的均值和标准差与正常的均值和标准差之比分别约为1：5.76和1：21.06；结果正常的均值和标准差与高异常的均值和标准差之比分别约为1：3.44和1：4.81。

当然，我们还可以类似地分析更多疾病之内的生理指标的分布特点，以及差异更多的疾病之间的生理指标的分布差异，并结合医学专家的兴趣开展更深入的分析，并提出新的科学假说，论证新的科学假说。

2.6.2 疾病中检验项的分布

与第2.6.1节检验项的疾病分布分析相反，本小节疾病中检验项的分布分析是对数据进行的反向探索。通过对各种疾病所做的全部实验室指标检查值的分布分析，我们可以揭示实验室指标与疾病之间存在的特定关系，通过实验室指标来展示某疾病的主要医学特征，也可以运用多次分析来分析不同疾病的特征差异。

如图2-45所示，以动脉硬化性心脏病为例，疾病中检验项的分布探索显示，白蛋白检查有55例，其中低的有12例占21.8%，正常的有43例占78.2%，异常与正常之比约为1：3.6。WBC（白细胞）检查有70例，其中低的有3例占4.3%，高的有11例占15.7%，正常的有56例占80%，异常与正常之比为1：4，异常中低异常与高异常之比约为1：3.7。以此类推，我们可以查看所有感兴趣的生理指标的检查特征，并建立指标特征之间的普遍联系，从中发现系统性的问题和潜在的规律。

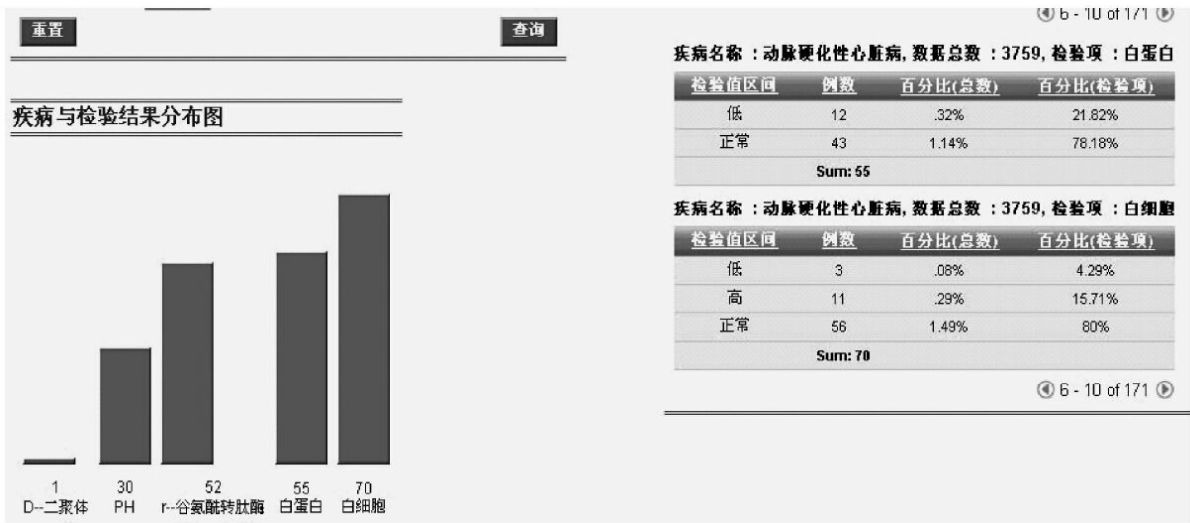


图 2-45

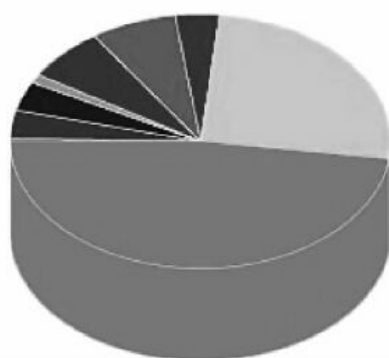
2.6.3 成对检验项的相关分析

通过对两维度指标在所有疾病中检查的结果，分析两个指标的相关性分布，快速概览两种特定指标之间及指标与疾病之间存在的特定关系。我们以血钾和血钠作为两维度指标的相关性分布分析的例子。在IMRS的数据探索子系统相关模块中，定义了两个维度生理指标，并给出上限和下限（可以使用系统自动给出的参考值，也可以自定义），系统会查询出与选定实验室指标相关的所有疾病，并按照指标值的高、中、低分类，如图2-46所示。

请选择第一个检验项目：	<input type="text" value="钾"/>	参考值：3.5-5 mmol/L
指定下限：	<input type="text" value="3.5"/>	指定上限： <input type="text" value="5"/>
请选择第二个检验项目：	<input type="text" value="钠"/>	参考值：135-145 mmol/L
指定下限：	<input type="text" value="135"/>	指定上限： <input type="text" value="145"/>
请选择开始日期：	<input type="text" value="01-11-2000"/>	
请输入结束日期：	<input type="text" value="11-11-2009"/>	
请选择检验数据：	<input type="text" value="入院首次检验数据"/>	
请选择检验跨度：	<input type="text" value="3天"/>	
<input type="button" value="重置"/>		<input type="button" value="查询"/>

实验组合疾病例数分布图

□□□□□□□□□□□□



■	□-H □-H 2	: 4%
■	□-H □-L 20	: 3.6%
■	□-H □-M 21	: 3.8%
■	□-L □-H 6	: 1.1%
■	□-L □-L 38	: 6.8%
■	□-L □-M 41	: 7.3%
■	□-M □-H 21	: 3.7%
■	□-M □-L 144	: 25.6%
■	□-M □-M 269	: 47.8%

图 2-46

我们以电解质指标钾和钠为例，上下限取标准值，分析结果为：钾高同时钠高（H，H）的病例仅占2例，约0.4%，而两者均正常的（M，M）有269例，约占47.8%。反过来的含义，就是入院病人中电解质不正常的占了一半以上。其他还有七种钾和钠的不同高低组合。通过本例，我们可以知道这个数据探索分析工具主要用于快速概览事物（这里是不同种类的疾病）两个特征维度的划分情况。

2.6.4 两种药物的应用分析

通过对两种药物在有关疾病中应用的效果分析，可以揭示两种特定药物之间，及药物与疾病之间存在的特定关系，还可以比较不同药物组合对疾病疗效的差异。在IMRS的数据探索子系统相关模块中，选择需要探索分析的两种药物：A药和B药。

如图2-47所示，左侧分析图显示数据仓库中所有的病种以及其例数，右侧用BI技术开发的智能报表显示了用药的情况，包括A药、B药单独以及联合使用（交叉使用）的例数。



图 2-47

数据探索分析通常可以同时正向和逆向使用，以期获得更全面的

信息，例如检验项疾病分布分析和疾病中检验项分布。另外，不同的数据探索分析偏重的方向也不同，它们提供了各个方面的快速概览功能。上述的数据探索分析模块都可以与数据挖掘算法引擎（如IMRS中的数据挖掘子系统）相互配合，探索疾病在医学指标中所反映的分布和联系。数据探索分析模块也可以与统计学工具（如IMRS中的智能统计子系统）配合应用，开展有关研究。数据探索更偏重于对知识的辅助发现。

数据探索可以直接针对粗数据，也就是未经过精确、严格筛选的数据，进行分析。数据探索的优势在于其高速性和概括性，而作为数据挖掘的初探，它的精准性则退而求其次。面对越来越巨大的数据，我们没有精力也没有时间来面面俱到地深入分析数据以确定每一个挖掘任务的方向。数据探索的高速概括有助于我们最大限度地了解数据和数据代表的事物。以数据探索高速地概括和总结配合数据挖掘算法的精准突击，互相协作，点面结合，可以更好地完成获取知识的任务。

2.7 本章小结

数据探索是一个复杂繁琐而又十分重要的工作，它是数据挖掘应用系统工程中初始的一个关键环节，它直接影响着数据挖掘应用的最终成败。有很多数据挖掘的应用项目，人们更多地关注数据挖掘算法的设计、开发和应用，在数据探索方面的工作投入得不多，也缺乏有效的数据探索技术和方法，这也是数据挖掘应用效果不佳的一个重要原因。数据挖掘是一广纳各个学科知识精华的、理论和技术永远不断发展的新兴学科和技术，其中的数据探索方法也是一样，需要不断地借鉴各个领域的知识、技术和方法，以便更好地为挖掘“真金”做好开路先锋，为数据挖掘应用系统工程的后续阶段打好坚实的基础。

第3章 数据挖掘应用算法

数据挖掘的算法是数据挖掘技术应用的核心内容之一。一个好的算法当然在数据挖掘应用项目中起着至关重要的作用，但是任何算法都不是万能的，都有着一定的适用条件。在不同的业务、不同特点的数据、不同的挖掘目标中，同一个算法可能表现出非常不同的性能。况且，数据挖掘的商业应用项目通常是一个复杂的系统工程，工程中的很多步骤对项目的成败都有决定性的作用。所以，我们在应用实践中，不能只重视算法而忽视其他环节，更不能唯算法论，否则就会像很多项目那样，有头无尾、无疾而终。对于数据挖掘的算法应用，关键的不是说一定要创造出多么高深的新算法，而是能最大限度地使用现有的算法。

作为数据挖掘应用实践类的一本书，本书用大量篇幅讲述一些数据挖掘应用实践相关的内容，自然也会涉及到很多数据挖掘算法的内容。本章将介绍一些数据挖掘应用项目中常用的算法，其中包括一些比较经典的算法，也有一些我们为应用编写的新算法。另外，在本章我们会提到一些算法流程，它们实际是应用于一些商业项目中的方法和技术，一般情况下是以某个算法为核心、以其他算法或处理步骤为前后过程的一整套技术方案，用于解决商业项目中的某一类问题。

本章将先后介绍聚类分析、特性选择、特征抽取、关联规则、分类和预测、以及时间序列等方面的算法，它们涵盖了数据挖掘商业项目中最常用的一些方法。

3.1 聚类分析

物以类聚，人以群分。聚类（Clustering）是将数据划分成群组（簇）的过程，根据数据本身的自然分布性质，数据变量之间存在的程度不同的相似性（亲疏关系），按照一定的准则将最相似的数据聚集成簇（Cluster）。很多数据挖掘算法可以分为有监督学习方法和无监督学习方法。有监督学习方法往往是由已知的结果信息来从数据中推导和验证数据中蕴含的某个能够导出结果的模型或规律，而无监督学习方法则通常是从结果信息未知的数据中来推导数据中蕴含的主要特征或规律。聚类与另一种类似的数据挖掘方法分类

（Classification，将在稍后的章节中介绍）不同，聚类的类别取决于数据本身自然特性，而分类的类别规则是预先定义好的。所以常说聚类属于无监督学习方法，而具有预先设定准则的分类则属于有监督学习方法。进行聚类工作，需要我们研究数据变量之间存在着程度不同的相似性（亲疏关系）的度量，也就是相异度的计算。根据不同的变量类型，不同的相异度计算方法，也就有了不同的聚类算法。我们依次介绍实践中常用的几个聚类算法，主要包括划分聚类算法、层次聚类算法、密度聚类算法，并尽可能举一些通俗易懂的简单实例，以有助于理解。

3.1.1 划分聚类算法（K均值）

传统上，聚类算法从技术上可以分为几类，如：基于划分的聚类、基于层次的聚类、基于密度的聚类，等等。每类聚类技术中又有很多不同的算法，它们都有各自的优缺点，都有一定的适用场合。下面，我们首先介绍一下基于划分的聚类。在划分方法中，最经典的就是K均值算法（K-Means）算法和K中心（K-Medoids）算法，很多算法都是由这两个算法改进而来的。

3.1.1.1 K-Means算法概念

给定一个含有N个数据点的数据集，以及要生成的簇的数目K。每一个分组就代表一个聚类， $K \leq N$ 。而N的数量不能过少，否则聚类的效果会大打折扣，一般实践中N最好大于 $K \times 10$ 。这K个分组满足下列条件：

- 每一个分组至少包含一个数据记录。
- 每一个数据记录属于且仅属于一个分组（注意：这个要求在某些模糊聚类算法中可以放宽）。

对于给定的K，算法首先的任务就是将数据构建成K个划分，以后通过反复迭代以改变分组的重定位技术，使得每一次改进之后的分组方案都较前一次好。将数据点在不同的划分间移动，直至满足一定的准则。一个好的划分的一般准则是：在同一个簇中的数据点尽可

能“相似”，不同簇中的数据点则尽可能“相异”。

所以聚类的过程，不仅仅是K-Means算法，是一个子类（簇）同质性上升的过程，也是子类之间异质性上升的过程。所以，聚类算法不但使用其划分功能，其同质性上升的功能在实践中也颇有用武之地（参见第5章）。

3.1.1.2 K-Means算法流程

1.用密度分类的方法将数据大致分类，确定k值

划分聚类在实际的使用当中，首先一个问题是确定K值，也就是聚类生成的簇数。如果认为人为确定簇数欠妥当，可以用密度分类的方法进行大致的分类，确定k值。

可以通过数据分布情况观察数据有几个密集区，例如单维的变量可以通过其分布峰的个数来确定k值。二维的可以通过其二维散点图来判定k值。例如，单维的数据可以通过直方图来判定，如图3-1所示可以确定 $k=2$ 。

二维的数据可以通过散点图来判定，如图3-2所示可以确定 $k=3$ 。

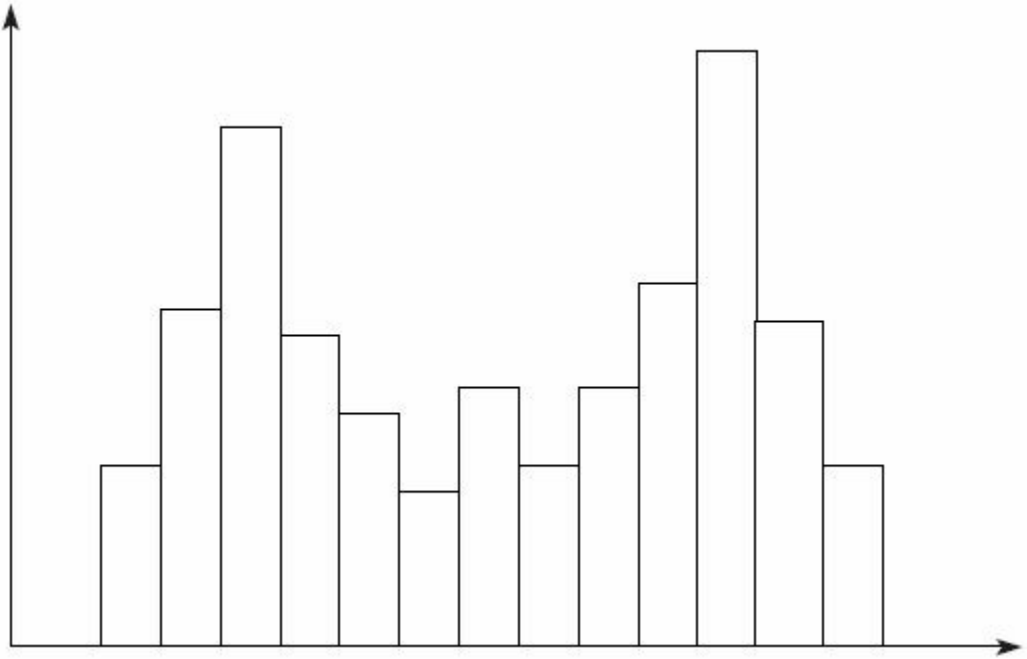


图 3-1

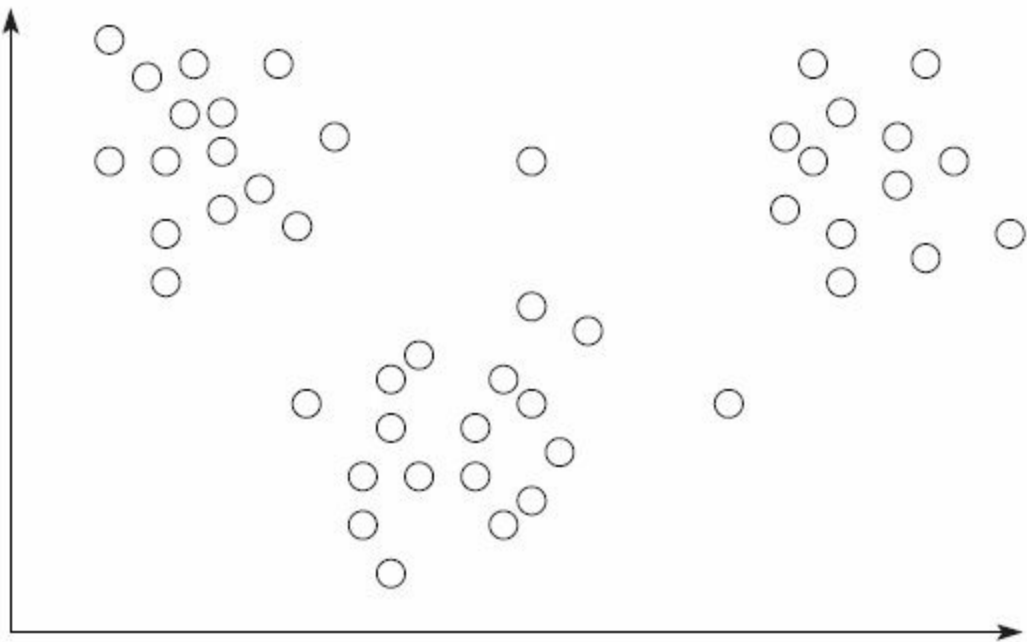


图 3-2

对于要求分类精度高的项目，我们甚至可以单独采用基于密度的算法来处理，如DBSCAN算法（参见3.1.3.2小节），以便获得比较可靠的k值。

2.用异常侦测等方法去离群点

通常我们会采用基于概率分布模型的方法，例如，采用一元正态分布分析或多元正态分布分析方法。如果数据点不能很好地拟合一个概率分布模型，或者说它不服从该分布，则将该数据点标志为离群点。如果上一步，我们采用了类似DBSCAN算法的基于密度的划分算法，由于基于密度的划分算法本身会具有良好的抗噪能力，我们可以将低密度区域的数据点标志为离群点。

其实，离群点也可能包含很重要的信息，毕竟一只黑天鹅就足以推翻天鹅都是白色的这一论断，如图3-3所示。观察离群点也有一套相应的算法，请大家参照其他书籍。

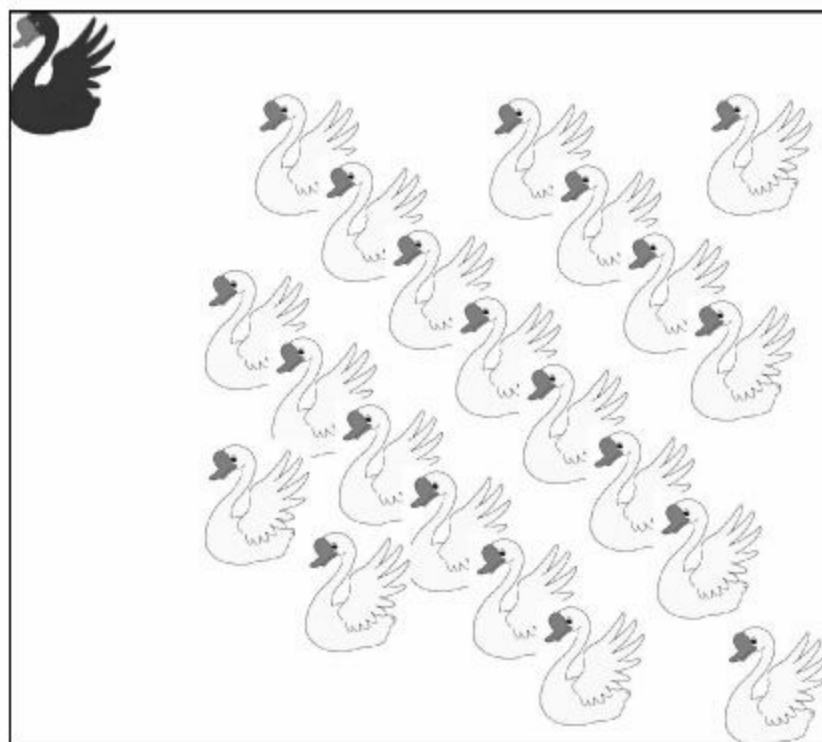


图 3-3

3.应用K-Means方法聚类

K-Means方法适合定量资料分析，而且在聚类前需要对数据进行标准化。K-Means算法接受输入参数 k ，然后将 n 个数据点划分为 k 个聚类以便使得所获得的聚类得到满足：同一聚类中的数据点相似度较高，而不同聚类中的数据点相似度较小。聚类相似度，是利用各聚类中数据点的均值所获得一个“中心数据点”（引力中心）来进行计算的。

K-Means算法的工作步骤如下：

- 1) 首先从 n 个数据数据点任意选择 k 个数据点作为初始聚类中心。
- 2) 对于剩下其他数据点，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的聚类中心所代表的聚类。
- 3) 然后再计算每个新聚类的聚类中心（该聚类中所有数据点的均值）。
- 4) 不断重复这一过程，直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数，即准则函数。

k 个聚类具有以下特点：各聚类本身尽可能紧凑，而各聚类之间尽可能地分开。样本点分类和聚类中心的调整是迭代交替进行的两个过程。

K-Means逻辑算法:

输入: 聚类个数k, 以及包含n个数据点的数据库

输出: 满足方差最小标准的k个聚类

处理流程:

- 1) 从n个数据数据点任意选择k个数据点作为初始聚类中心。
- 2) 根据簇中数据点的均值, 将每个数据点重新赋给最类似的簇。
- 3) 更新簇的均值, 即计算每个簇中数据点的均值。
- 4) 循环步骤2) 到步骤3), 直到每个聚类不再发生变化为止。

相异度计算方法, 两个数据点间的距离计算方法如下。

· 明氏距离 (Minkowski Distance)

$$d(x_i, x_j) = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^q \right)^{1/q} \quad (3.1)$$

这里的 $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ 和 $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$ 是两个p维的数据点并且 $i \neq j$ 。

· 欧氏距离 (Euclidean Distance)

当明氏距离中 $q=2$ 时, 欧氏距离如下:

$$d(x_i, x_j) = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^2 \right)^{1/2} \quad (3.2)$$

4.应用轮廓系数方法对K-Means聚类结果评估

可以通过考察簇的分离情况和簇的紧凑情况来评估聚类，利用数据集的数据点之间相似性度量。

轮廓系数的思想：

- 1) 从数据中随机取出第 i 个数据点，并计算该点到 S 簇中其他所有数据点的平均距离，记为 $d(i)$ 。
- 2) 计算该点到所有其他簇中所有数据点的平均距离，并找到最小平均距离，记为 $Md(i)$ 。
- 3) 计算轮廓系数 $SC(i) = (Md(i) - d(i)) / \max(d(i), Md(i))$ ，轮廓系数越接近1，说明该点的划分越好。将簇中所有点的轮廓系数取平均值，综合所有簇的平均轮廓系数，平均轮廓系数越高的分类方案越优。

具体操作步骤如下：

对于 n 个数据点的数据集 D ，假设 D 被划分成 k 个簇 S_1, \dots, S_k 。对于每一个数据点 $t_i \in D$ ，假设 t_i 所在的簇为 S_i ，计算 t_i 与 S_i 中其他数据点之间的平均距离 $d(i)$ ，计算 t_i 到不属于簇 S_i 的数据点的最小平均距离 $Md(i)$ 。

则

$$d(i) = \frac{\sum_{x \in S_i, t_i \neq x} dist(t_i, x)}{|S_i| - 1} \quad (3.3)$$

$$Md(i) = \min_{S_j: 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{x \in S_j} dist(t_i, x)}{|S_j|} \right\} \quad (3.4)$$

数据点 t_i 的轮廓系数定义为：

$$SC(i) = \frac{Md(i) - d(i)}{\max\{Md(i), d(i)\}} \quad (3.5)$$

轮廓系数的值在-1和1之间。 $d(i)$ 反映 t_i 所在的簇 S_i 的紧凑性，该值越小簇越紧凑， $Md(i)$ 的值捕获 t_i 与其他簇的分离程度， $Md(i)$ 值越大 t_i 与其他簇越分离。因此当 t_i 的轮廓系数值接近1时，包含 t_i 的簇是紧凑的，并且 t_i 远离其他簇，这是一种可取的情况。然而，当轮廓系数的值为负时（即 $Md(i) < d(i)$ ），表明聚类不理想。

观察下面一组单维数据，并应用K-Means聚类算法进行分簇。

$$X_1=-4, X_2=-3, X_3=-2, X_4=1, X_5=2, X_6=3$$

共六个点。应用K-Means聚类算法，按照如下步骤进行分簇：

1) 根据分布情况设定初始参数：

选择簇数 $k=2$ ，

选择 X_3 、 X_4 为初始簇中心，

$$M_1=X_3=-2,$$

$$M_2=X_4=1,$$

2) 计算其他各点到各个簇中心的距离，将它赋给最近的簇：

$$X_1: d(X_1, M_1) = |-4 - (-2)| = 2,$$

$$d(X_1, M_2) = 3,$$

$$d(X_1, M_1) < d(X_1, M_2),$$

故将 X_1 分配给 C_1 .

$$X_2: d(X_2, M_1) = 1,$$

$$d(X_2, M_2) = 4,$$

$$d(X_2, M_1) < d(X_2, M_2),$$

故将 X_2 分配给 C_1 .

$$X_5: d(X_5, M_1) = 4,$$

$$d(X_5, M_2) = 1,$$

$$d(X_5, M_1) > d(X_5, M_2),$$

故将 X_5 分配给 C_2 .

$$X_6: d(X_6, M_1) = 5,$$

$$d(X_6, M_2) = 2,$$

$$d(X_6, M_1) > d(X_6, M_2),$$

故将 X_2 分配给 C_2 .

得到簇集:

$$C_1 = \{X_1, X_2, X_3\}$$

$$C_2 = \{X_4, X_5, X_6\}$$

3) 计算簇的中心:

$$M_1 = 1/3 [(-4) + (-3) + (-2)] = -3;$$

$$M_2 = 1/3 [1+2+3] = 2$$

4) 重复过程2), 得到:

$$C_1 = \{X_1, X_2, X_3\}$$

$$C_2 = \{X_4, X_5, X_6\}$$

再计算簇的中心

$$M_1 = 1/3 [(-4) + (-3) + (-2)] = -3;$$

$$M_2 = 1/3 [1+2+3] = 2$$

两次迭代簇中心没变化, 所以停止迭代, 算法停止。

在K-Means算法中， k 需要事先给定，这也是K-Means算法的一个特点。这个特点可以用于很多特定方面。比如，我们需要把运动员分成一级、二级、三级三个档次，自然 $k=3$ ；同样，也可以把机构（比如医院）分成一级、二级、三级，这时 $k=3$ ；把俱乐部分级，分析某俱乐部属于几（ n ）流，这时 $k=n$ 。

3.1.2 层次聚类算法（组平均）

基于层次的聚类方法，是对给定的数据进行层次的分解，直到某种条件满足为止。首先将数据点组成一棵聚类树，然后根据层次，自底向上或是自顶向下分解。层次的方法可以分为凝聚的方法和分裂的方法。

凝聚的方法，也称为自底向上的方法，初始时每个数据点都被看成是单独的一个簇，然后通过逐步合并相近的数据点或簇，形成越来越大的簇，直到所有的数据点都在一个簇中，或者达到某个终止条件为止。

3.1.2.1 凝聚的方法

1) 用异常侦测等方法去离群点。

为了提高凝聚方法的分类效率和优化分类效果，首先应用异常侦测等方法去离群点。我们通常采用基于概率分布模型的方法，一元或多元正态分布分析方法，如果数据点不能很好地拟合一个概率分布模型，或者说它不服从该分布，则将该数据点标识为离群点。

2) 用密度分类的方法将数据分成许多小类。

采用了诸如DBSCAN的基于密度的划分算法，将数据划分足够小的类，以便减少凝聚算法的高昂成本，并且该步骤可逆，即如果后面

的凝聚算法得不到一个全局优解，可以重复本步。

3) 自底向上凝聚算法 (AGNES-AGglomerative NESting)。

自底向上凝聚的逻辑算法如下：

输入：包含 n 个数据点的数据库，终止条件簇的数目 k 。

输出： k 个簇，达到终止条件规定簇数目。

将每个数据点当成一个初始簇；

重复以下步骤：

- 1) 根据两个簇中最近的数据点找到最近的两个簇。
- 2) 合并两个簇，生成新的簇的集合。
- 3) 达到定义的簇的数目时终止。

凝聚法并不难理解，我们生活中处处可以发现类似的现象。如图3-4所示，从“人→家庭→社区→城市→大陆→星球”的凝聚过程，就是一个凝聚法的实例。

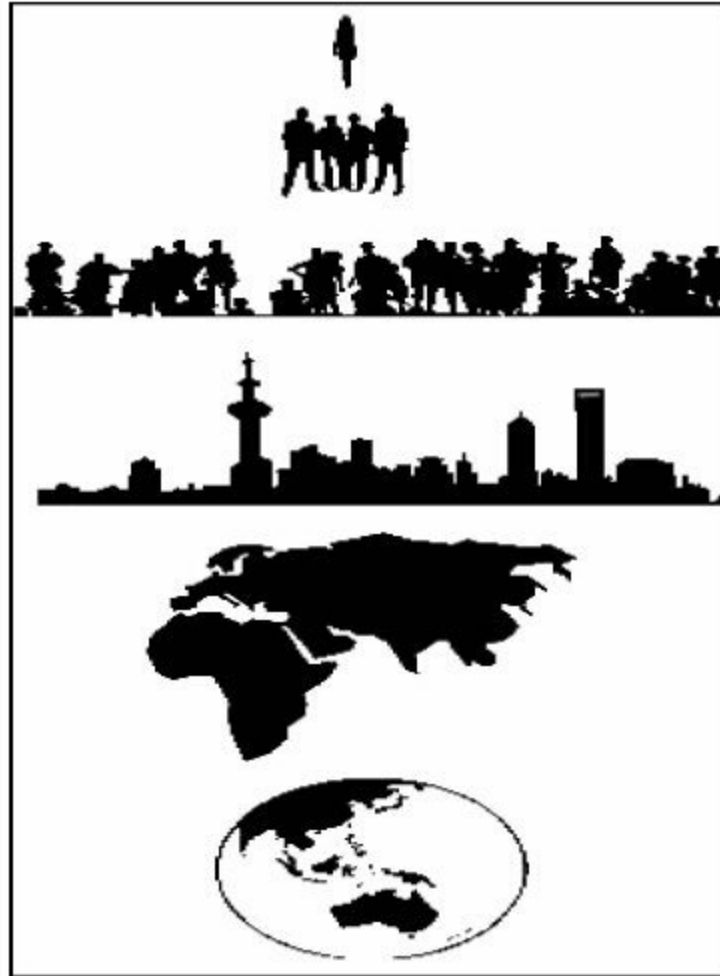


图 3-4

本书也是个例子，书中把各个概念、算法、应用，等等按照相似程度聚集成章节。再例如，有一组数据，共8个元素：A，B，C，D，E，F，G，H。

应用凝聚聚类方法进行分簇，步骤如下：

1) 初始簇： $\{A\} \{B\} \{C\} \{D\} \{E\} \{F\} \{G\} \{H\}$ ，

找到最近距离的 $\{A\} \{B\}$ ，合并成 $\{A, B\}$ ，

于是有新簇： $\{A, B\} \{C\} \{D\} \{E\} \{F\} \{G\} \{H\}$ 。

2) 重复1) 得到： $\{A, B\} \{C, D\} \{E, F\} \{G, H\}$ 。

3) 找到最近距离的 $\{A, B\} \{C, D\}$ ，合并成 $\{A, B, C, D\}$ ，

于是有新簇集： $\{A, B, C, D\} \{E, F\} \{G, H\}$ ，

继续得到： $\{A, B, C, D\}$ ， $\{E, F, G, H\}$ 。

4) 分簇完成。

3.1.2.2 分裂的方法

分裂的方法，也称为自顶向下的方法。它与凝聚层次聚类恰好相反，初始时将所有数据点置于一个簇中，然后逐渐细分为更小的簇，直到最终每个数据点都在单独的一个簇中，或者达到某个终止条件为止。

1) 用异常侦测等方法去离群点。

同3.1.2.1节的第一步。

2) DIANA (DIvisive ANAlysis自顶向下分裂算法) 逻辑算法。

自顶向下分裂的算法步骤如下：

输入：包含 n 个数据点的数据库，终止条件簇的数目 k 。

输出：k个簇，达到终止条件规定簇数目。

- 1) 将所有数据点设为初始簇S。
- 2) 找出S中与其他点平均相异度最大的点 X_1 ，将 X_1 放入新簇 S_1 。
- 3) 根据到最近点的距离将S中的剩余点重新分簇，于是产生两个新簇 S_1 和 S_2 。
- 4) 将 S_1 和 S_2 中直径大的簇，重复2)、3)步。
- 5) 生成k个新簇时终止。

3) 应用轮廓系数方法对分裂方法聚类结果评估。

同3.1.1.2的内容4。

例如有一组数据，共8个元素：A, B, C, D, E, F, G, H。应用分裂聚类方法进行分簇，步骤如下：

输入k=4，执行以下步骤：

- 1) S: {A, B, C, D, E, F, G, H}。
- 2) S_1 : {A, B, C, D},
 S_2 : {E, F, G, H}。
- 3) S (1, 1) : {A, B},

$S(1, 2) : \{C, D\}$,

$S(2, 1) : \{E, F\}$,

$S(2, 2) \{G, H\}$ 。

4) 分簇完成。

分裂法在生活中也是经常可以看到的。例如，下火车或下公交车的人群按照目的地分裂：先按照大方向分裂，再按小方向分裂，最后再到各自的目的地，如图3-5所示。

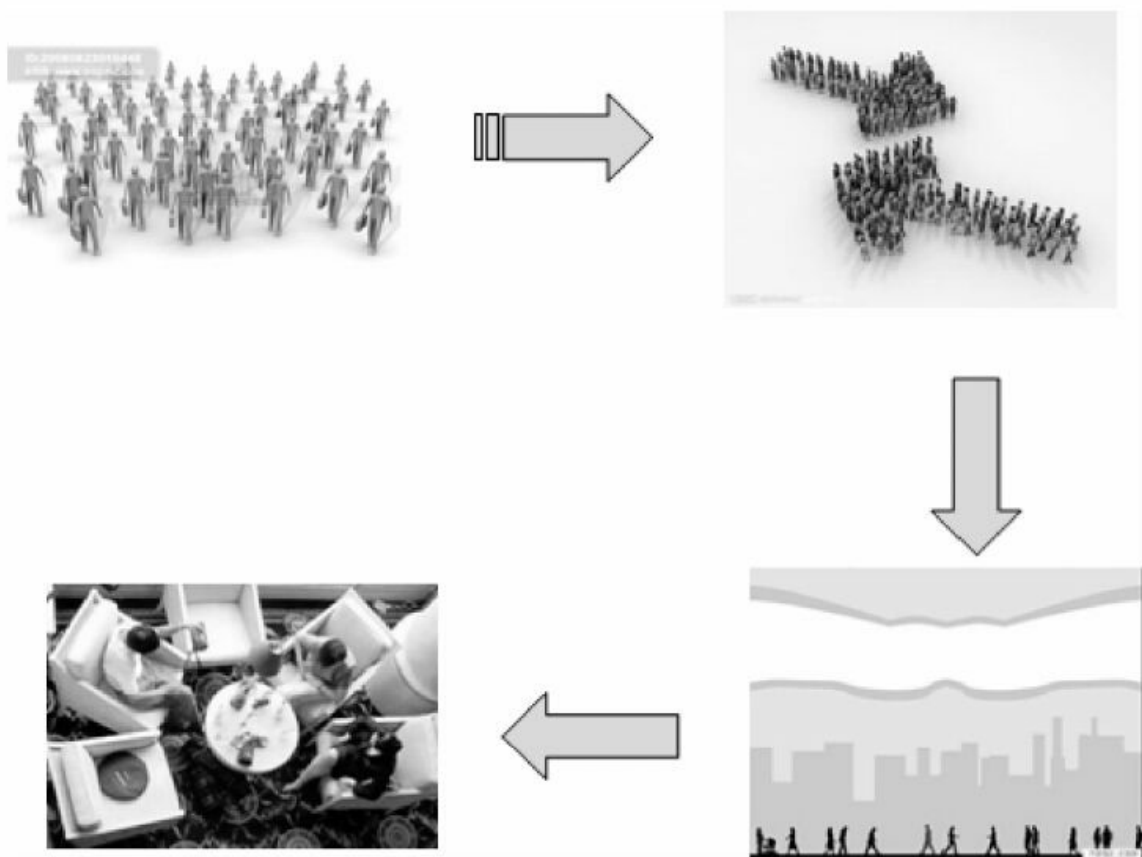


图3-5 自然分裂图（从人群到各个目的地）

3.1.2.3 综合层次聚类方法

1. BIRCH算法

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) 算法是一个综合性的层次聚类方法，它利用层次方法的平衡迭代进行归约和聚类。其核心是用一个聚类特征 (CF) 三元组表示一个簇的有关信息，从而使簇中的点可用对应的聚类特征表示。它通过构造满足分支因子和簇直径限制的聚类特征树来求聚类。该算法通过聚类特征可以方便地进行中心、半径、直径及类内、类间距离的运算。

聚类特征是一个三维向量， $CF = (n, LS, SS)$ ，其中 n 是数据点的个数， LS 是 n 个点的线性和， SS 是 n 个点的平方和。通过 CF 可以方便地求中心、半径、直径及类内、类间距离。 CF 树有两个参数：

- 分支因子 β ，通过它定义每个非叶子结点的子女的最大数目。
- 阈值 T ，存储在叶结点中的子簇的最大直径。

算法实现的流程如下：

1) 用异常侦测等方法去离群点。

同3.1.2.1节的第一步。

2) 扫描数据库，建立一个初始的 CF 树。它可以被看作一个数据

的多层压缩，试图保留数据内在的聚类结构。当一个数据点被插入到最近的叶节点（子聚类）中时，随着数据点的插入，CF树被动态地构造，不要求所有的数据读入内存，而可在外存上逐个读入数据项。因此，BIRTH方法对增量或动态聚类也非常有效。

3) 采用某个聚类算法对CF树的叶节点进行聚类。在这个阶段可以执行任何聚类算法，例如典型的划分方法。BIRCH算法试图利用可用的资源来生成最好的聚类结果。通过一次扫描就可以进行较好地聚类，故该算法的计算复杂度是 $O(n)$ ， n 是数据点的数目。

4) 应用轮廓系数方法对分裂方法聚类结果评估。

同3.1.1.2的内容4。

该算法的优点是具有数据点数目的线性可伸缩性及良好的聚类质量，一次扫描就可以进行较好的聚类。缺点是BIRCH算法只适用于类的分布呈凸形及球形的情况，对不可视的高维数据则不可行。

BIRCH逻辑算法：

输入：包括 n 个数据点的数据集合 $D = [x_1, x_2, \dots, x_n]$ 及阈值 T 。

输出：簇集合。

过程：

1) 扫描数据集，建立初始的CF树。

- 2) 对每个数据点 x_i 执行如下操作。
- 3) 为 x_i 找一个正确的待插入的叶子节点。
- 4) 如果阈值条件不被破坏, 则将 x_i 插入到叶子节点中, 并从被插入的叶子节点到根节点依次更新CF三元组。
- 5) 如果有叶子节点空间, 则将 x_i 作为单独的簇插入到树中, 并更新CF三元组。
- 6) 否则分类叶子节点, 重新安排CF三元组。

2.CURE算法

CURE (Clustering Using Representatives) 算法中既有层次部分, 也有划分部分, 所以CURE也是一个综合性的聚类算法。

前面讲到的聚类算法倾向于处理簇为球形、且簇的大小相似的聚类问题, 并且对孤立点较为敏感。CURE采用了多个点代表一个簇的方法, 可以较好地处理以上问题。并且在处理大数据量的时候采用了随机抽样、分区的方法, 来提高其效率, 使得其可以高效地处理大量数据。

算法实现的流程如下:

- 1) 用异常侦测等方法去离群点。

同3.1.2.1节的第一步。

2) 对原始数据进行随机抽样, 将样本进行等量划分, 划分后便形成了以这些样本点为代表的分区。

3) 对于每一个分区, 再使用层次聚类算法中的凝聚算法。在凝聚算法中的每一步, 距离最近的代表性点所对应的簇将被合并。它们之间的距离被定义为两个簇中代表性点之间距离的最小值。

4) 应用轮廓系数方法对分裂方法聚类结果评估。

同3.1.1.2的内容4。

CURE算法的具体步骤如下:

1) 从原数据集中抽取一个随机样本 S 。

2) 为了加速聚类, 把样本划分成 p 份, 每份大小相等。

3) 对每个划分进行局部地聚类。

4) 根据局部聚类结果, 通过随机抽样剔除孤立点。主要有两种措施: 如果一个簇增长得太慢, 就去掉它; 在聚类结束的时候, 非常小的类被剔除。

5) 对上一步中产生的局部的簇进一步聚类。从每个簇中选择 c (常数) 个点, 然后通过应用收缩因子 a , 将这些分散的点向簇的质心方向收缩。当 a 为1的时候, 所有点都收缩成一点, 即质心。通过多个有代表性的点, 簇的形状可以更好地被表示出来。

6) 用相应的簇标签来标记数据。

CURE算法的优点是它回避用所有点或单个质心来表示一个簇的传统方法，而是将一个簇用多个具有代表性的点来表示，使CURE可以适应非球形的几何形状。

另外，收缩因子降底了噪声对聚类的影响，从而使CURE对孤立点的处理更加健壮，而且能识别非球形和大小变化比较大的簇，对于大型数据库具有良好的可伸缩性。缺点是参数设置对聚类结果有很大的影响，不能处理分类属性。CURE的最小时间复杂度是 $O(n^2)$ ，其中 n 是数据点的数目。

3.1.3 密度聚类算法

基于密度的聚类算法由于能够发现任意形状的聚类，识别数据集中的噪声点，可伸缩性好等特点，在许多领域有着重要的应用。

3.1.3.1 密度算法概念

1) 如果一个数据点周围足够稠密，也就是以这个点为中心，给定半径 (ϵ) 的邻域内的数据点足够多，密度大于密度阈值（用户指定的参数MinPts），则称这个数据点为核心数据点，否则为边界数据点。

2) 研究数据点之间的关系，如果 p 是一个核心数据点， q 属于 p 的 ϵ 邻域，则称 q 是从 p 直接密度可达的。

3) 如果存在一条链 $\{p_1, p_2, \dots, p_i\}$ ，满足 $p_1=p$ ， $p_i=q$ ， p_i 直接密度可达 p_{i+1} ，则称 p 密度可达 q 。

4) 如果存在 r ， r 密度可达 q 和 p ，则称 p 和 q 是密度相连的。

5) 由一个核心数据点和其密度可达的所有数据点构成一个聚类。

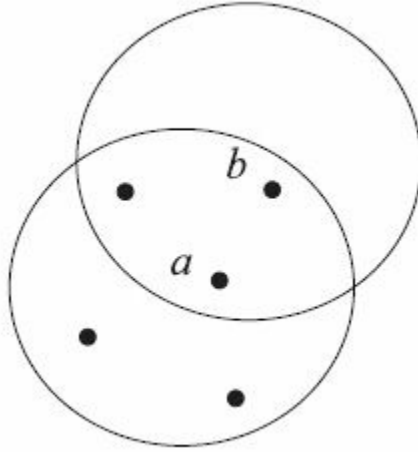


图 3-6

如图3-6所示，*a*为核心数据点，*b*为边界数据点，且*a*直接密度可达*b*，但*b*不直接密度可达*a*，因为*b*不是一个核心数据点。

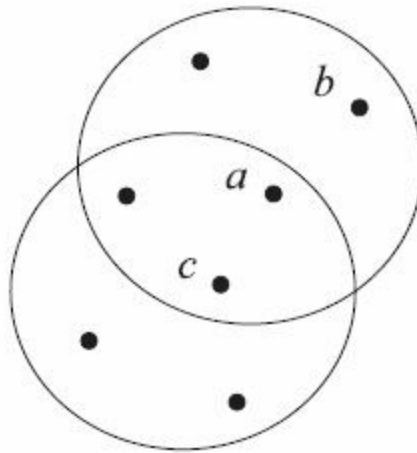


图 3-7

再如图3-7所示，*c*直接密度可达*a*，*a*直接密度可达*b*，所以*c*密度可达*b*，同理*b*不密度可达*c*，但*b*和*c*密度相连。

3.1.3.2 DBSCAN算法

DBSCAN算法思想：寻找数据集 D 的子集 S ， $S \subseteq D$ ， S 是密度相连的闭集， S 满足 S 中任意两点是密度相连的，并且 S 中任意点不能和 S 外的点是密度相连的。DBSCAN从任一数据点 p 开始，根据参数 ϵ 和 MinPts ，提取所有从 p 密度可达数据点，得到一个聚类。

DBSCAN算法的步骤如下：

- 1) 从任一数据点 p 开始，对 p 点根据 ϵ 和 MinPts 进行判定。如果 p 是核心数据点，则建立新簇 S ，并将 p 邻域内的所有点归入 S ；否则将 p 点标记为边界点或噪声点。
- 2) 对于 S 中除 p 点以外的点继续实施过程1)，继续扩充 S ，直到所有的点都被判定处理。

DBSCAN算法的优点：不易受噪声影响，可以发现任意形状的簇。DBSCAN算法的缺点：受设置参数的影响，判定的标准比较固定，较稀的聚类会被划分为多个类，或密度较大且离得较近的一类会被合并成一个聚类。

3.2 特性选择

特性选择是指为特定的应用在不失去数据原有价值的基础上选择最小的属性子集，去除不相关和冗余的属性。特性选择用于在建立分类模型前，或者预测模型之前，对原始数据库进行预处理。本节将介绍特性选择的概念、特性选择的三个常用算法：线性相关算法、相关因子SRCF算法，以及最小描述长度MDL算法。

3.2.1 特性选择概念

我们把被预测的那个属性称为目标变量（Target Variable），把用来预测目标变量值的属性称为预测变量（Predictor Variable）。特性选择，就是要考查数据库中的预测变量与目标变量的关联程度，或者说考察每个预测变量对最终预测结果的重要程度。这个重要程度用一个指数（importance）来表示，特性选择算法要计算每个预测变量的 importance 指数。

3.2.1.1 特性选择步骤

特性选择算法包含以下三个步骤：

- 筛选（screening）：除去不重要或者有问题的预测变量和数据，例如预测变量含有过多缺失值，或者预测变量的变化太大或太小而变得无用。
- 评分（ranking）：对所有预测变量根据其重要程度进行评价、打分、排序。
- 选择（selecting）：生成在后续模型中使用的功能子集，例如，仅保留最重要的预测变量，过滤或排除所有其他预测变量。

3.2.1.2 特性选择方法

在筛选阶段，要删除那些不适合参与建模的预测变量，以及一些有缺陷的样本数据。例如表3-1所示，我们发现B列的数据几乎是一个常数，C列几乎都是空值，因此我们会把这两列删去。我们还可以观察到第9行和第16行的多个字段为空值，因此我们会把这两行也删去。

表3-1 样本数据

ID	A	B	C	D	E	F	G
1	129	62.1		154	143	147	135
2	114	62.1		115	103	93	157
3	123	62.1		130	151	125	121
4	178	62.1		143	142	128	151
5	195			91	135	137	144
6	184	62.1	2	126	98	109	158
7	129	62.1		140	98	156	158
8	170	62.1		134	108	137	131
9		62.1		128		130	111
10	120	62.1		105	105	103	154
11	182	62.1		159	130	123	125
12	178	62.1		158	147	109	156
13	166	62.1		161	127	101	130
14	186	62.1		138	122	127	159
15	131	62.1		102	123	102	138
16	173	62.1					
17	198	62.1		125	110	154	139
18	144	62.1	3	116	117	136	116
19	178	62.1		124	146	146	93
20	131	62.1		121	140	122	107

进一步地，我们将处理有缺陷的样本数据归结为以下规则。

满足下列任一条件的预测变量将被删除：

- 变量值几乎全部缺失（如C列）。

- 变量全部取值都几乎是固定不变的值，即全部相同（如B列）。

- 多个变量同时代表了样本的标示号（如类似ID列的字段），只保留其中的一个变量。

满足下列任一条件的样本将被删除：

- 目标字段的值缺失的数据（如第9行）。

- 所有预测字段的值都缺失的数据（如第16行）。

根据用户的自行设置值（如 m_1 、 m_2 、 m_3 、 m_4 、 m_5 ）来删除预测变量：

- 有多于 $m_1\%$ 的值缺失，则删除该变量，因为缺失值百分比很大的字段几乎不提供任何预测信息。

- 对于离散型的预测变量，如果有超过 $m_2\%$ 的数据取同一个值，则删除该变量。

- 对于连续型的预测变量，如果其标准差小于 $m_3\%$ ，则删除该变量。

- 对于连续型的预测变量，如果其变异系数 $|CV| < m_4\%$ ， $CV = \text{标准差} / \text{均值}$ 。因为如果变异系数接近0，则变量值的变异性就不高。

- 对于离散型的预测变量，变量取值的个数大于总样本量的

$m_5\%$ ，则删除该变量。这一标准用于筛选掉相对于记录总数而言有过多类别取值的字段。如果很高百分比类别只含有一个观测值，则该字段用处有限。

在评分阶段，也就是特性选择的评分阶段，通常使用一些特征评价函数来对所有的属性相对于目标属性的预测重要性进行评价。常用的特征评估函数可分为五类：相关性、距离、信息增益、一致性和分类错误率。具体的评价方法有：欧式距离、非线性测量、Minkowski距离、信息增益、最小描述长度（MDL）、依赖性度量、相关性度量、一致性度量、分类错误率（正确率）等。其中最小描述长度（简称MDL算法）是一种应用普遍的主要评价方法，下面我们对MDL算法做一简单介绍。

MDL算法源自最小指数描述长度原则，由Rissanen基于数据压缩和通讯原理首先提出。其基本思想是：对于一个给定的数据集 D （Data Set），为了最大限度地节省数据集的存储空间，我们试图从 n 个可能的模型（或程序、或算法）中寻找一个模型 M_i （ $1 \leq i \leq n$ ）， M_i 能最大限度地提取数据集 D 中所有的规律，并将其进行数据压缩，然后将模型 M_i 本身连带压缩后的数据集 C_i 一同存储起来，其总存储大小为 S_i （Size）。由于不同的模型对 D 的压缩效率不同，一般来说 D 的压缩率越高模型的复杂度就越高。因此我们从可行的诸多压缩方案中，将最小的 S_i 称为 D 的最小描述长度。最小描述长度的原理就是要选择总描述长度最小的模型 M_i 。如图3-8所示。

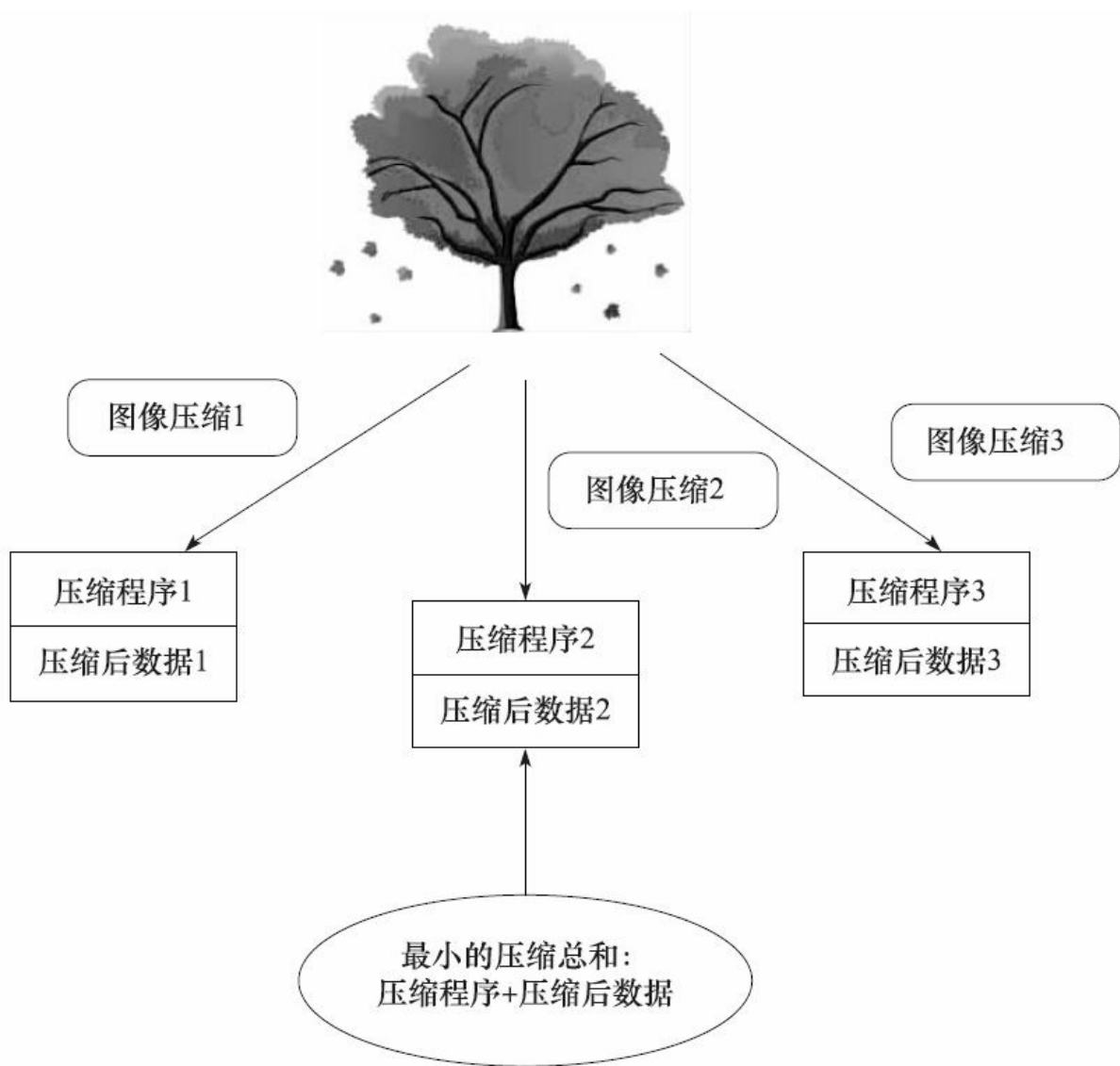


图 3-8

将MDL原理应用于特性选择的方法是，MDL算法将D中的每一个属性都看作是目标属性（如分类属性）的一个简单预测模型。使用其相应的MDL度量对这些单一预测模型进行比较和评分。使用MDL算法，模型选择问题变成了数据通信问题。属性评分使用两部分代码来传送数据。前一部分传送模型，模型参数是与每个预测值相关联的目标概率。后一部分传送的是使用该模型后预测错误的原始数据。如下公式：

$$S_i(\text{MODEL}_i, D) = S(\text{MODEL}_i) + S(C_i) \quad (3.6)$$

其中：

$S_i(\text{MODEL}_i, D)$ 为数据集D上应用第i个属性对目标属性建立简单预测模型后得到的总大小， $S(\text{MODEL}_i)$ 为应用第i个属性对目标属性建立简单预测模型（ MODEL_i ）的大小， $S(C_i)$ 是对第i个属性数据应用 MODEL_i 后，所有预测错误的原数据大小总和。

最小描述长度的英文表达为Minimum Description Length (MDL)。这里的Length（长度）是指压缩后的全部数据（程序+剩余数据）占存储空间（存储字节）的总大小，我们也可用L表示。由于通常情况下，某个预测属性对目标属性的预测正确率（压缩率）很难达到100%，因此不同的预测属性对目标属性的预测正确率表现为不同的概率，如下所示：

$$\text{建立预测模型} \left\{ \begin{array}{l} x_1 \rightarrow Y \quad \text{描述 } Y \text{ 的 } 80\%, \text{ 剩余 } 20\% \text{ 的数据, 总长度为 } L_1 \\ x_2 \rightarrow Y \quad \text{描述 } Y \text{ 的 } 75\%, \text{ 剩余 } 25\% \text{ 的数据, 总长度为 } L_2 \\ \dots \\ x_n \rightarrow Y \quad \text{描述 } Y \text{ 的 } 25\%, \text{ 剩余 } 75\% \text{ 的数据, 总长度为 } L_n \end{array} \right.$$

我们分别应用n个预测属性 X_1, X_2, \dots, X_n 对目标属性Y建立预测模型，其中 X_1 的预测正确率（压缩率）为80%，也就是说 X_1 可以正确地描述Y数据的80%，剩余的20%Y数据（压缩后数据）没能由 X_1 正确地描述，因此其总长度 L_1 为 $\text{Length}(\text{Model}(X_1, Y)) + \text{Length}(Y) * 20\%$ 。其余的 X_2, \dots, X_n 的预测模型的总长度为 L_2, \dots, L_n 。最小描述长度模型就是要从 L_1, L_2, \dots, L_n 中找到最小的一

个所对应的分类属性。

接着上面的数据集 D ，我们取最小的 $S_i(\text{MODEL}_i, D)$ ，即 $\text{Minimum}(S_i(\text{MODEL}_i, D))$ ，这样我们应用MDL算法将得到一个最佳预测属性，该属性可以理解为拥有最多的描述目标属性规律的属性，也就是该属性含有与目标属性相关的最多的信息，因此我们将其称为目标属性的一个相对最优特性。同理，我们按照MDL评分的排序（MDL值最小的排在最前面），依次得到不同属性相对于目标属性的特征评分。

在选择阶段，需要从前一阶段中评分较高的属性中选择一些属性，作为下一步数据挖掘分析（如分类或预测）的输入信息。这时，输入属性集的选择主要取决于下一步数据挖掘分析的目的（如分类或预测），和相应的备选算法（如贝叶斯算法、决策树算法、支持向量机算法等）。不同的备选算法选取的最优输入属性集可能是不同的。在评分阶段，特性选择的方法也称为过滤方法，即使用某种独立于备选算法的方法，在下一步数据挖掘分析前进行特性选择。而在选择阶段，特性选择的方法也称为包装方法，即将下一步数据挖掘分析算法作为黑盒子，使用在评分阶段得到的排名靠前的一个特性组合进行分类或预测试验，从中得到一个最佳特性组合，但通常不枚举所有可能的子集来找出最佳属性子集。过滤方法和包装方法的唯一不同是，它们使用了不同的特征子集评估方法。对于包装方法，子集评估使用目标数据挖掘算法；对于过滤方法，子集评估技术不同于目标数据挖掘算法。

另外，在选择阶段我们需要考虑一个问题。在评分阶段得到的一组特性是按照一定的评分算法，单独对每个属性与目标属性的预测重要性（也称相关性）进行评价，但并没有考虑这些属性之间的关系、以及这些属性共同参与对目标属性的预测时，对预测结果的影响。在实践中，我们经常会发现这样一个现象，对于目标属性Y有三个特征属性A、B、C，其中A和B一起预测Y的准确率要高于A和B任意一个单独预测Y的准确率，但是，A和C一起预测Y的准确率要不高于A和C任意一个单独预测Y的准确率。这里，我们把Y的三个特征属性A、B、C称为Y的矢量相关特性。其中，我们把A和B称为对于Y的矢量实相关，或A与B是同向与Y相关的，即A和B单独预测Y是同向的，A和B一起预测Y是互相增益的；把A和C称为对于Y的矢量虚相关，或A与C是不同向与Y相关的，即A和C单独预测Y是不同向的，A和C一起预测Y是互相干扰的。如图3-9所示。

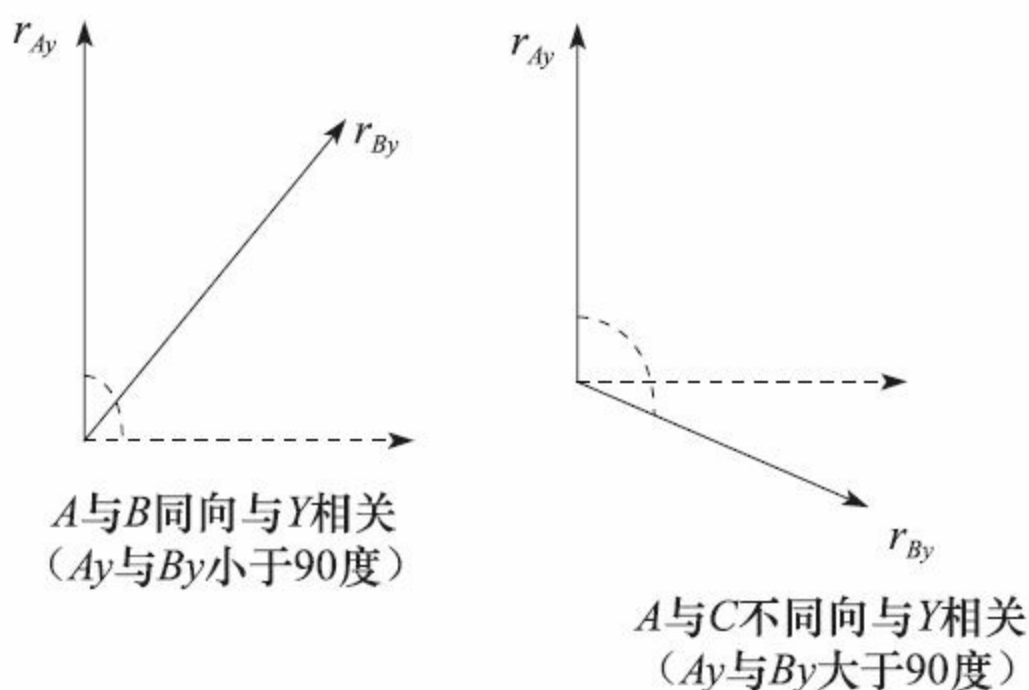


图 3-9

因此，在选择Y的预测属性时，应尽量避免矢量虚相关的属性对。

3.2.2 线性相关算法

对于探索与连续型的目标变量具有相关性的连续型变量，也就是分析两个连续型变量是否存在相关性，一种比较简单的方法是线性相关算法。直线关系是函数关系中最简单的一种，直线相关分析是判断两变量间是否有直线关系，给出相关方向和相关密切程度的度量，度量的指标为相关系数。应用于双变量正态分布资料，两变量是平等关系。

3.2.2.1 Pearson相关

相关系数亦称Pearson积差相关系数（coefficient of product-moment correlation），用 r 表示样本相关系数， ρ 表示总体相关系数，总体相关系数是未知的，通常用样本相关系数 r 进行估计。

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{l_{XY}}{\sqrt{l_{XX}l_{YY}}} \quad (3.7)$$

其中 X 和 Y 是两个连续变量， l_{XY} 是 X 和 Y 的离均差交叉乘积和， l_{XX} ， l_{YY} 分别是 X 和 Y 的离均差平方和。

相关系数没有单位，其值 $-1 \leq r \leq 1$ 。当两变量呈同向变化时， $0 < r < 1$ ，为正相关；两变量呈反向变化时， $-1 < r < 0$ ，为负相关； $r = 0$ 为零相关，表示无直线相关关系； $r = 1$ 或者 $r = -1$ 时为完全直线相关。直线相关的性质可由图3-10直观地说明。

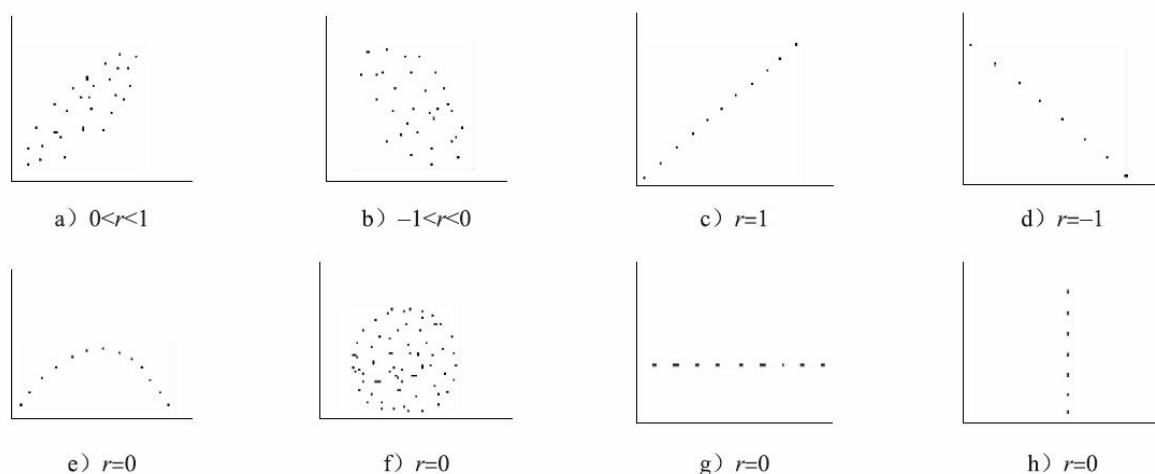


图3-10 相关系数示意图

相关系数的假设检验。相关系数 r 是样本相关系数，由于存在抽样误差， $\rho=0$ 但 r 不一定等于0，计算出 r 值后，接着应做 $\rho=0$ 的假设检验，以判断两变量的总体是否有直线相关关系。常用的有 t 检验，检验统计量 t 值的计算公式如下：

$$t = \frac{r - 0}{S_r} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}, \quad \nu = n - 2 \quad (3.8)$$

其中 n 是样本量。

3.2.2.2 Spearman秩相关

Spearman秩相关（rank correlation）是一种非参数统计方法，适用于资料不是双变量正态分布或总体分布未知；数据一端或两端有不确定值的资料或等级资料。用等级相关系数 r_s 来说明两个变量间相关关系的密切程度与相关方向，将两个变量分别排序求秩次，然后对秩次进行计算，分别用 P 、 Q 表示 X 和 Y 求完秩后的秩值，对于 P 和 Q 使用

公式（3.7），则：

$$r_s = \frac{\sum (P - \bar{P})(Q - \bar{Q})}{\sqrt{\sum (P - \bar{P})^2 \sum (Q - \bar{Q})^2}} = \frac{l_{PQ}}{\sqrt{l_{PP}l_{QQ}}} \quad (3.9)$$

等级相关系数 r_s 的假设检验如下， r_s 是总体等级相关系数 ρ_s 的估计值，存在着抽样误差，故计算出 r_s 后，需作 $\rho_s=0$ 的假设检验。当 $n \leq 50$ 时，可通过查 r_s 界值表实现 $\rho_s=0$ 的假设检验；当 $n > 50$ 时，

$$t = \frac{r_s - 0}{S_{r_s}} = \frac{r_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}}, \quad \nu = n - 2 \quad (3.10)$$

线性相关分析要求两个变量服从双变量正态分布，所以进行线性相关分析之前需要进行双变量正态分布验证，如果数据不服从双变量正态分布，则使用Spearman秩相关方法分析。

3.2.3 相关因子SRCF算法

对于探索与离散型的目标变量具有相关性的变量，也就是分析至少一个离散型变量的两变量是否存在相关性，方法相对复杂一些，我们这里应用相关因子SRCF算法来处理，并做详细介绍。

3.2.3.1 SRCF算法概念

所谓SRCF（Chi-Square Rank Correlation Factorization），是基于卡方-秩和检验的相关因子算法。它由两个子算法组成，一个是SRCF1，一个是SRCF2。其中SRCF1又分成两个小算法，分别为SRCF1.1和SRCF1.2；SRCF2又分成三个小算法，分别为SRCF2.1、SRCF2.2和SRCF2.3。我们先看一下它的逻辑算法，然后再逐步分解：

```
DECLARE
i NUMBER;
v_rank(i) ARRAY;
BEGIN
IF 数据类型 (Y) <> '离散'
THEN EXIT;
ELSE
  FOR i IN 1..n
  IF 数据类型 (Xi) = '连续'
  THEN CALL(SRCF1); --调用算法SRCF1
  IF COUNT(DISTINCT Y) = 2
  THEN
    v_re(i) := entropy (Xi,Y); --计算Xi与Y的相对熵
  ELSE
    v_rfsc(i) := rfsc (Xi,Y); ---计算Xi与Y的特征记分
  KW_compare(Xi,Y); -- 多组KW秩和比较及不同组量的计算
  END IF ;
  ELSE
  CASE WHEN 等级变量 (Xi) = 'NO' AND 等级变量 (Y) = 'NO'
  THEN CALL(SRCF2.1); --调用算法SRCF2.1;
  WHEN 等级变量 (Xi) = 'YES' AND 等级变量 (Y) = 'YES'
  THEN CALL(SRCF2.3); --调用算法SRCF2.3;
  ELSE CALL(SRCF2.2); --调用算法SRCF2.2;
  END CASE;
END;
```

```

IF (所有Xi与Y的组合中都至少有一个等级变量)=YES
THEN
    FM_compare(Xi,Y);--多组Friedmen秩和比较及不同组量的计算
ELSE
    Pearson_r(Xi,Y);-- 计算Pearson 列联系数
END IF;
END IF;
END FOR;
RANK(v_re(i),Xi); -对Xi及其相对熵进行排序并输出
RANK(v_rfsc(i),Xi); -对Xi及其特征记分进行排序并输出
RANK(KW_compare(Xi,Y)); -对KW秩和不同组量进行排序并输出
RANK(FM_compare(Xi,Y)); -对FM秩和不同组量进行排序并输出
RANK(Pearson_r(Xi,Y)); -对Pearson 列联系数排序并输出
END IF;
END;

```

3.2.3.2 算法一：SRCF1

SRCF1是相关因子算法SRCF的一个子算法，是用于探索与离散型的目标变量（分类变量）具有相关性的连续型变量，我们将这些具有相关性的连续型变量称为连续型相关因子，它也是特性选择的一种算法。SRCF1采用的是基于秩和检验的非参数统计方法，同时使用相对熵（Relative Entropy，以下用Entropy代替）、特征记分准则（RFSC）和多组K-W秩和不同组量方法进行相关因子排序。

对于目标变量，当分类取值不同时，我们采用不同的统计方法加以处理，这里分为两种情况：

- 目标变量是两分类时，采用两独立样本的秩和检验，也称Wilcoxon检验（Wilcoxon rank test for two independent samples）。

- 目标变量大于两分类时，采用多个独立样本的秩和检验，也称K-W检验（Kruskal-Wallis test for independent samples）或H检验。

1.目标变量是两分类（SRCF1.1）

例如，对于一个给定数据集： Y 为二分类目标变量（如 $Y=0$ 或 $Y=1$ ）， X 为连续型观察变量， $X=(X_1, X_2, \dots, X_m)$ 。我们的目的是从 X_1, X_2, \dots, X_m 连续变量中寻找 Y 的相关因子。如图3-11所示。

X_1	X_2	X_m	Y
⋮	⋮	⋮	⋮	0
				1
				0
				1
				1
				0

图 3-11

对于 X_1, X_2, \dots, X_m ，我们按照 Y 中值的分类，将某个连续变量 X_i 分为两组（ $Y=0$ 组和 $Y=1$ 组），然后对这两组数据的分布进行统计检验。对于两组连续变量数据的差异检验方法，通常采用的是参数统计方法正态性检验和T检验。但是，实际应用数据中，数据满足参数统计方法条件（数据正态分布）的情况并不多，所以我们采用非参数统计方法，即两独立样本的秩和检验，该方法虽然统计检验效能较参数统计方法的差一些，但通用性较好。

SRCF1算法实现如下所示：

(1) 首先是编秩，就是按照数值的大小进行排序。将两组数据由小到大统一编秩，可先将两组数据分别由小到大排序，然后再将两组数据合在一起由小到大统一排序。编秩时有数值相同的数据时，取它们的平均秩次。编秩完成后，将两组秩次（排名）分别相加，求秩和（排名总和）。

(2) 然后是确定统计量。无论两组数据的记录数是否相等，均选取记录数较小者对应的秩和为统计量。

(3) 最后确定P值。P值的确定可采用查表法和正态近似法。

1) 查表法：查统计量的界值表。当 $n_1 \leq 10$ ， $n_2 - n_1 \leq 10$ ，采用查表法，将检验统计量T值与T临界值相比，若T值在界值范围内，其P值大于相应的概率；否则小于相应的概率。

2) 正态近似法：组1的记录数 n_1 ，组2的记录数为 n_2 ，并且 $n_1 < n_2$ ，如果 n_1 或 $n_2 - n_1$ 超出界值范围，用正态近似检验：

$$u = \frac{|T - n_1(n_1 + n_2 + 1)/2| - 0.5}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \quad (3.11)$$

若两组有相同秩次时，公式（3.10）需要校正：

$u_c = \frac{u}{\sqrt{c}}$ 其中 $c = 1 - \sum (t_j^3 - t_j) / (n^3 - n)$ ， t_j 为第j个相同秩次的个数，

$$n = n_1 + n_2$$

(4) 当执行完上述的秩和检验后，就可以做出推断结论了。如果检验的结果为阳性，即 $p < 0.05$ ，则说明组1和组2的数据分布有显著差异。进一步，我们可以推断 X_i 变量（ $i \in 1 \dots m$ ）与目标分类变量Y具有相关性，也就是说 X_i 是Y的相关因子。还可以说， X_i 是从m个变量当中选择出来的Y的一个特性。

相关因子的相关度如下分析：对于目标变量是两分类的相关因子相关度的评估，可采用相对熵、特征记分准则方法。

相对熵方法，计算两组连续型数据的相对熵值。对于Y为两分类时，可以直接使用下面的公式计算两组之间的相对熵：

$$D(P \parallel Q) = \int_{x \in S} P(x) \log \frac{P(x)}{Q(x)} dx \quad (3.12)$$

其中，P和Q为两个概率密度函数。

相对熵的值越大， X_i 与Y的相关度就越高。

特征记分准则方法，计算多组连续型数据的特征记分。对于Y为多分类时，可以直接使用下面的公式计算多组之间的特征记分：

$$RFSC(g_i) = \frac{1}{2} \left| \frac{\mu_i^+ - \mu_i^-}{\sigma_i^+ + \sigma_i^-} \right| + \frac{1}{2} \ln \frac{(\sigma_i^+)^2 + (\sigma_i^-)^2}{2\sigma_i^+ \sigma_i^-} \quad (3.13)$$

特征记分的值越大， X_i 与Y的相关度就越高。

2. 目标变量是多分类（SRCF1.2）

例如，对于一个给定数据集：Y为多分类目标变量（如Y=0、Y=1、Y=2、Y=3），从 X_1 、 X_2 、...、 X_m 变量中寻找Y的相关因子。如图3-12所示。

X_1	X_2	X_m	Y
⋮	⋮	⋮	⋮	2
				0
				1
				3
				2
				1
			

图 3-12

对于 X_1 、 X_2 、...、 X_m ，我们按照Y中值的分类，将某个连续变量 X_1 分为四组（Y=0、Y=1、Y=2、Y=3），然后对这四组数据的分布进行统计检验。对于两组以上连续变量数据的差异检验方法，通常采用的参数统计方法为方差分析。方差分析需要首先检验多组数据的正态性，然后还要检验多组数据间的方差齐性。但是，实际应用数据中，数据满足参数统计方法的条件（即各组数据正态分布，同时多组数据方差齐性）的情况并不多。所以我们采用非参数统计方法，即多个独立样本的秩和检验，其目的是推断多个样本的中位数分别所代表的总体中位数是否相等。该方法虽然统计检验效能较方差分析方法的差一

些，但通用性比方差分析方法好。本方法由Kruskal和Wallis在Wilcoxon秩和检验的基础上扩展而来，所以称为K-W检验（Kruskal-Wallis test for independent samples）。

算法实现如下所示：

（1）K-W检验，其编制和求秩和的方法与Wilcoxon秩和检验的方法都是一样的，在此不赘述。

（2）统计量的计算，公式为

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1) \quad (3.14)$$

式中 R_i 为各组的秩和， n_i 为各组对应的例数， $n=\sum n_i$ 。

（3）P值的确定分两种情况。

1) 当组数 $g=3$ ，每组例数 $n_i \leq 5$ ，可查H界值表得到P值。

2) 当组数 $g>3$ ，或 $g=3$ 且最小样本例数 $n_i \leq 5$ 时，H近似地服从自由度为 $v=g-1$ 的 χ^2 分布，可查 χ^2 界值表得到P值。

若有相同秩次时，应计算公式（3.13）的校正值 H_c ，

$$H_c = \frac{H}{c} \quad (3.15)$$

其中 $c=1-\sum (t_j^3-t_j)/(n^3-n)$ ， t_j 为第j个相同秩次的个数。

（4）当执行完上述的秩和检验后，可以做出推断结论。如果检验的

结果为阳性，即 $p < 0.05$ ，则说明组1、组2、组3、组4的数据分布不完全相同。进一步我们可以推断 X_i 变量（ $i \in 1 \dots m$ ）与目标分类变量 Y 具有相关性，也就是说， X_i 是 Y 的相关因子。还可以说， X_i 是从 m 个变量当中选择出来的 Y 的一个特性。

相关因子的相关度分析：对于目标变量是多分类的相关因子相关度的评估，可采用特征记分准则和多组K-W秩和不同组量方法。特征记分准则方法同上一小节。多组K-W秩和不同组量方法如下：

使用SRCF1.2的Kruskal-Wallis秩和检验算法，当推断结论为拒绝原假设时，与方差分析类似，只能得出各总体分布不全相同的结论，但是不能说明任意两个总体分布不同。若要能做出任意两个总体分布有无不同的推断，需要作组间的多重比较。步骤如下：

首先是计算检验统计量。

设 R_i 和 R_j 分别为比较的第 i 组和第 j 组数据的秩和，其平均秩次分别为 \bar{R}_i 和 \bar{R}_j ，应用Wilcoxon秩和检验，方法同前。当数据量较大时，可以采用正态近似法，公式为：

$$u_{ij} = \frac{\bar{R}_i - \bar{R}_j}{\sigma_{\bar{R}_i - \bar{R}_j}} = \frac{\bar{R}_i - \bar{R}_j}{\sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad (3.16)$$

其中 $n = \sum_{i=1}^g n_i$ 为 g 个数据的总含量， n_i 、 n_j 分别为第 i 组和第 j 组的数据量。

当有相同秩次时，用校正值计算：

$$u_{je} = \frac{u_{ij}}{\sqrt{c}} \quad (3.17)$$

其中 $c = 1 - \frac{\sum (t_j^3 - t_j)}{n^3 - n}$, t_l 为第 l 个相同秩次的个数。

完成检验统计量的计算后，我们就可以利用标准正态分布表确定 P 值了。然后，我们需要将某两组比较所得 P 值与调整以后的检验水准 α' 比较，若 $P < \alpha'$ ，则拒绝 H_0 。进行多重比较时，不能用原来 $\alpha = 0.05$ 的检验水准作为每一个两两比较的检验水准，因为重复多次的假设将增大犯第一类错误的概率，必须对检验水准进行调整，常用的调整方法为 Bonferroni 法。该法经常用于多个数据的两两比较，先对需要比较的数据进行 χ^2 分割，形成多个四格表，再对每个四格表进行 χ^2 检验，采用 $\alpha' = \frac{\alpha}{\text{比较次数}}$ 计算调整的水准，其中 α 通常为 0.05。最终以调整的 α' 作为检验水准，推出有无统计学意义的结论。一般是 k 组数据间任两组都进行比较的情况较多，比较次数为 $k(k-1)/2$ ，检验水准 α' 为 $\alpha' = \frac{\alpha}{k(k-1)/2}$ ，最后输出所有 k 组间两两不同的组数，不同的组数越多， X_i 与 Y 的相关度就越高。

3.2.3.3 算法二：SRCF2

SRCF2 是相关因子算法 SRCF 的另一个子算法，是用于探索与离散型的目标变量（分类变量）具有相关性的离散型变量，如图 3-13 所示，我们将这些具有相关性的离散型变量称为离散型相关因子，它也是特性选择的一种算法。SRCF2 采用的是基于秩和检验、 χ^2 检验和 Gamma 系数的统计方法，同时使用 Friedman 秩和不同组量方法和

Pearson列联系数方法进行相关因子排序。

X_1	X_2	X_3	X_4	X_i	Y
...	a	...	1
				b		2
				c		3
				a		2
				b		3
				c		1
				a		1
				b		3
				c		2

图 3-13

在统计方法的选择上，主要是根据离散变量的等级类型来确定，下面我们分情况讨论。

假定Y与 X_i 均为离散变量：

1.SRCF2.1

如果Y与 X_i 都为非等级变量，则应用列联表 χ^2 检验方法来处理。

算法思路如下：首先将图3-13数据整理成列联表，如图3-14所示。一般情况下，列联表数据的行数和列数均大于2，所以需要运用行×列表 χ^2 数据检验的专用公式来进行列联表 χ^2 检验， $\chi^2 = n \left(\sum \frac{A^2}{n_R n_C} - 1 \right)$ ， $\nu = (R - 1)(C - 1)$ ，式中n为总例数，A为每个格子的实际频数， n_R 和 n_C 分别为与A所对应的行合计与列合计。由此来确定列联系数，根据列联系数选择相关因子。

y	a	b	c
1	A_{1a}	A_{1b}	A_{1c}
2	A_{2a}	A_{2b}	A_{2c}
3	A_{3a}	A_{3b}	A_{3c}

图 3-14

然后应用行列表 χ^2 检验，验证 X_i 变量的a、b、c各组对观察变量Y的影响是否相同（ $p < 0.05$ ），若 $p > 0.05$ 则说明 X_i 变量的a、b、c各组对观察变量Y的影响相同，也可以说 X_i 变量与观察变量Y不相关；若 $p < 0.05$ 则 X_i 变量的a、b、c各组对观察变量Y的影响不同，也可以说 X_i 变量与观察变量Y相关。

相关因子的相关度如下：SRCF2.1采用的是Pearson χ^2 检验计算关联系数，输出相关因子的相关度强弱。

我们采用Pearson列联系数公式：

$$r = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad (3.18)$$

该系数是用来度量变量之间的相关程度大小，列联系数越大， X_i 与Y的

相关度就越高。

2.SRCF2.2

如果Y与X_i有一个为等级变量，则使用Friedmen秩和检验。根据其检验统计量的符号也称M检验，目的是推断各处理组数据分别代表的总体分布是否都相同。

Friedmen秩和检验的算法思路如下：首先是编秩，将各区组内数据由小到大编秩，遇到相同数值时取平均秩次。再将各处理组的秩次相加，得到各组的秩和R_j。然后计算统计量M值。公式为

$$M = \sum (R_j - \bar{R})^2 \quad (3.19)$$

公式中 $\bar{R} = \sum R_j / g$ ，g为处理组数。

最后确定P值。在这里，P值的确定采用两种方法：

· 查表法：当n≤15，g≤15时，可以通过查M界值表确定P值。

· χ^2 近似法：当处理组数g或区组数n超出M界值表的范围时，可以采用服从自由度为(g-1)的 χ^2 分布。公式为

$$\chi_r^2 = \frac{12}{ng(g+1)} \sum_{j=1}^g R_j^2 - 3n(g+1) \quad (3.20)$$

当有相同秩次时，需要对此公式进行校正，公式为 $\chi_c^2 = \frac{\chi_r^2}{c}$ ，其中 $c = 1 - \sum (t_j^3 - t_j) / ng(g^2 - 1)$ ，t_j为第j (j=1, 2, ...) 个相同秩次的个数。这种校正方式通常用于，当数据个数在各区组中所占比重较大时，或当P值在检验水准附近时。

执行以上步骤后，我们就可以验证 X_i 变量（若 X_i 为等级变量）的a、b、c各组对观察变量Y的影响值是否相同（ $p>0.05$ ）。若 $p>0.05$ ，则输出 X_i 与Y无关，若 $p<0.05$ 则输出相关。

相关因子的相关度如下，Friedmen秩和不同组量方法：

- 对于Y为两分类时，直接输出不同组量为1；
- 对于Y为两分类以上时，使用如下步骤：

使用SRCF2.2的方法检验拒绝原假设后，需要对任意两组间进行多重比较。建立检验假设后，计算统计量并确定P值，设 R_i 和 R_j 分别为比较的第i组和第j组数据的秩和，其平均秩次分别为 \bar{R}_i 和 \bar{R}_j 。

当数据量较小时，应采用配对设计的符号秩和检验方法，求得统计量的数值后确定P值。方法如下：

将第i组和第j组的数据进行两两配对，假设共有m对数据，第l（ $l=1, \dots, m$ ）对具有观察值（ X_l, Y_l ），差值 $d_l=X_l-Y_l$ ， M_d 表示差值d的中位数。

1) 求差值：求各对数据（ X_l, Y_l ）的差值 $d_l=X_l-Y_l$ 。

2) 编秩：按差值的绝对值由小到大编秩，并保留差值的正负号。编秩时若差值为0，舍去不计；若差值的绝对值相等，取平均秩次。

3) 求秩和：将排出的秩次按正、负两组分别求秩次之和，分别以 T_+ 和 T_- 表示。

4) 确定统计量：任取正、负两组秩和之一为统计量 T 。记差值的总个数为 n （即 n 为差值不等于0的对子数），则 T_+ 和 T_- 之和为 $n(n+1)/2$ 。

5) 确定P值：

当 $5 < n \leq 50$ 时采用查表法，查配对设计用的 T 界值表。若检验统计量 T 值在上、下界值范围内，其 P 值大于相应的概率水平；若 T 值等于上、下界值或在界值范围之外，则 P 值小于相应的概率水平。

当 $n > 50$ 时采用正态近似法，可利用秩和分布的正态近似法检验，公式为

$$u = \frac{|T - n(n+1)/4| - 0.5}{\sqrt{n(n+1)(2n+1)/24}} \quad (3.21)$$

若有相同秩次，需要计算校正的统计量 u_c ：

$$u_c = \frac{|T - n(n+1)/4| - 0.5}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum(t_j^3 - t_j)}{48}}} \quad (3.22)$$

式中 t_j 为第 j （ $j=1, 2, \dots$ ）个相同秩次的个数。

这样我们就完成了配对设计的符号秩和检验。

当数据量较大时，可采用正态近似法，检验统计量为：

$$Z_{ij} = \frac{\bar{R}_i - \bar{R}_j}{\sigma_{\bar{R}_i - \bar{R}_j}} = \frac{\bar{R}_i - \bar{R}_j}{\sqrt{\frac{k(k+1)}{6n}}} \quad (3.23)$$

然后利用标准正态分布表确定 P 值。

得到P值后，将某两组比较所得P值与调整以后的检验水准 α' 比较，若 $P < \alpha'$ ，则拒绝原假设。与SRCF1.2类似，为了保证第一类错误的概率总共不超过 α ，我们还是按照Bonferroni法调整每次比较的错误概率 α' ，最后输出所有组间两两不同的组数。按两两不同的组数来判断相关度的强弱，即两两不同的组数越多表明Y与 X_i 的相关性越高，反之则越低。

3.SRCF2.3

如果Y与 X_i 都为等级变量，则计算Gamma系数（见图3-15）。

等级资料表				
X_i	Y			
	1	2	3	4
1	n_{11}	n_{12}	n_{13}	n_{14}
2	n_{21}	n_{22}	n_{23}	n_{24}
3	n_{31}	n_{32}	n_{33}	n_{34}
4	n_{41}	n_{42}	n_{43}	n_{44}

图 3-15

1) 共变（concordance）：变量 X_i 和Y具有相同变化趋势，即两者共同变化。

2) 异变（discordance）：变量 X_i 和Y具有不同变化趋势，即两者反向变化。

3) 共变频数 (N_s) : 变量 X_i 和 Y 具有相同变化趋势的对子数, 其计算公式为:

$$N_s = \sum_{x,y} n_{xy} A_{xy}$$

4) 异变频数 (N_d) : 变量 X_i 和 Y 具有不同变化趋势的对子数, 其计算公式为:

$$N_d = \sum_{x,y} n_{xy} D_{xy}$$

其中 n_{xy} 为具有 $X_i=x$ 和 $Y=y$ 特征的个体数:

$$A_{ij} = \sum_{x>i, y>j} n_{xy} + \sum_{x<i, y<j} n_{xy}$$

$$D_{ij} = \sum_{x>i, y<j} n_{xy} + \sum_{x<i, y>j} n_{xy}$$

以 n_{22} 为例对应的 $A_{22}=n_{11}+n_{33}+n_{34}+n_{43}+n_{44}$, 即为 i, j 从2, 2出发都变大或者都变小的趋势。

Gamma系数的计算原理是: 如果两变量相互独立, 那么两变量共变概率与异变概率之差应为零, 否则可认为两变量存在相关, 总体值用 γ 表示, 数据值用 G 表示: $G = \frac{N_s - N_d}{N_s + N_d}$, 取值范围 $[-1, 1]$

可以用Z检验来完成 $\gamma=0$ 的统计推断

$$Z = G \times \sqrt{\frac{N_s + N_d}{N(1 - G^2)}} \quad (3.24)$$

Y与 X_i 的相关性推断：若 $G=0$ 表示两变量不相关，若 $G>0$ 表示正相关，若 $G<0$ 表示负相关。

相关因子的相关度：当G的绝对值越接近1，表示两变量的关联程度越大，当G的绝对值越接近0表示关联程度越小。输出计算后的Gamma系数，以确定相关因子的相关度强弱。

3.2.3.4 SRCF相关因子评估

根据以上SRCF1.1、SRCF1.2、SRCF2.1、SRCF2.2、SRCF2.3的算法，我们可以将X中的所有观察变量逐个进行分析，以确定它们是否是Y的相关因子。下一步，我们需要对Y的所有相关因子的相关度进行比较评估。这里我们分两个类型分别评估。

（1）对所有连续变量的相关因子进行评估

如前面算法所示：当Y为两分类时，我们计算所有 X_i 与Y的相对熵，然后按相对熵大小降序排名。当Y为多分类时，我们计算所有 X_i 与Y的特征记分，然后按特征记分大小降序排名；或者，我们也可计算所有 X_i 与Y的多组K-W秩和不同组量，然后按不同组量的多少降序排名。

（2）对所有离散变量的相关因子进行评估

如前面算法所示：当所有的 X_i 与Y的组合中至少有一个等级变量时，我们计算所有 X_i 与Y的多组Friedmen秩和不同组量，然后按不同

组量的多少降序排名。当至少有一组 X_i 与 Y 的组合都是非等级变量时，我们计算所有 X_i 与 Y 的Pearson列联系数，然后按列联系数的大小降序排名。

另外，也可以使用计算离散变量的相对熵来对所有离散型的相关因子进行评估。计算两组离散型数据的相对熵的公式为：

$$D(P \parallel Q) = \sum_{i \in S} P(i) \log \frac{P(i)}{Q(i)} \quad (3.25)$$

其中， P 和 Q 为两组离散型数据的概率分布， i 为两组离散型数据的任意一个取值。无论是离散型数据之间，还是连续型数据之间，其相对熵都具有非对称的性质，即 $D(P \parallel Q) \neq D(Q \parallel P)$ 。另外相对熵还具有非负的性质，即 $D(P \parallel Q) > 0$ 。

扩展的应用是，对于一个连续型变量和另一个离散型变量的相关分析，可以将连续型变量进行离散化后，与另外一个离散型变量应用离散变量相对熵进行评估。我们会在后面章节中大量应用SRFC算法开展各类数据挖掘项目（见第4章、第6章）。

3.3 特征抽取

特征抽取是数据挖掘中另一种十分有用的方法，它与特性选择不同（在本书中为了更好地进行区分，我们约定俗成地称作特征抽取和特性选择，虽然特征和特性的字面含义接近）。特征抽取是对数据的特征进行概括和总结，而特性选择则主要是对数据中的不同特征进行比较和选取。

特征抽取，是数据挖掘技术中的一个常用方法，它是一个属性降维的过程，特征抽取实际上是变换属性。经变换了的属性，或者特性，是原来属性集的线性合并。特征抽取会导致更小更精的一组属性。用特征抽取建立的模型可能是质量更好的，因为数据被更少的和更有意义的属性所描述。特征抽取将高维数据集映射到一个低维的数据集中。特征抽取对数据可视化是很有用的，当一个复杂的数据集被降维到两维或三维空间时，数据可有效地被观测。

特征抽取的算法有很多，这里重点介绍主成分分析法、因子分析法、非负矩阵因子分解NMF法等算法。

3.3.1 主成分分析算法

假设待分析数据为 N 个 K 维数据，主要成分分析（PCA，又称Karhunen-Loeve或K-L方法）即是搜索 q 个最能代表数据的 k 维正交向量，这里 $q \leq k$ 。这样将 K 维数据空间压缩为 q 维数据空间。PCA不同于前面讲到的相关因子计算法（选择一个较小的属性子集），PCA创建新的、数量较小的属性集合，这些新的属性是原有属性的线性组合。

3.3.1.1 主成分分析算法思想

为了客观、全面地研究问题，需要记录多个观察指标 X_1 、 X_2 、 X_3 、... X_n 。虽然这些观察指标可以提供丰富的信息，但同时也使得数据的分析工作更趋复杂化。例如，在儿童生长发育的评价中，收集到的数据包括每一个儿童的身高、体重、胸围、头围、坐高、肺活量等十多个指标。事实上，在实际工作中，所涉及的众多指标之间经常是有相互联系和影响的。通过对原始指标相互关系的研究，找出少数几个综合指标，这些综合指标是原始指标的线性组合，它既保留了原始指标的主要信息，又互不相关，从而达到寻找少数综合指标以概括原始指标的目的。

3.3.1.2 主成分分析算法实现

- 1) 将 $X_1, X_2, X_3, \dots, X_n$ 构成的矩阵 M 标准化，标准化后记为

M_b ，计算 M_b 的相关系数矩阵，记为 R 。

2) 对于相关系数矩阵 R ，采用雅克比方法求解特征方程 $|R - \lambda I| = 0$ 的 P 个非负的特征根 $\lambda_1 > \lambda_2 > \dots > \lambda_p \geq 0$ ， λ_i 对应的特征向量记为 $v_i = (v_{i1}, v_{i2}, \dots, v_{in})$ ， $i=1, 2, \dots, P$ 并且 n 满足 $v_i v_j^T = \sum_{k=1}^n v_{ik} v_{jk} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$

3) 选择 ω 个主成分，使得前面 ω 个主分量的方差和占全部总方差的比例 $\eta = (\sum_{i=1}^{\omega} \lambda_i) / (\sum_{i=1}^P \lambda_i)$ 接近于1，并使所选的这 ω 个主成分尽可能多地保留原来 P 个特征的信息，从而达到特征抽取或降维的目的。

主成分分析算法的基本过程如下：

- 1) 对数据进行标准化。
- 2) PCA计算 q 个标准正交基。这 q 个标准正交基被称为主要成分，所有数据是主要成分的线性组合。
- 3) 根据主要成分对于变异的贡献量，即变异百分比，从大到小排序。
- 4) 当排序在前的几个因子的累积变异百分比达到一定值（比如75%）后，就可以将后面的因子去掉，达到压缩数据的目的。

PCA计算花费低，可以用于有序和无序的属性，并且可以处理稀疏和倾斜数据。例如，表3-2是一个应用PCA算法得到的因子分析表。表中列出了共12个因子特征根、变异百分比和累积变异百分比。

表3-2 因子分析表

12 个因子				12 个因子			
因子	特征根	变异百分比	累积百分比	因子	特征根	变异百分比	累积百分比
1	4	40%	40%	7	0.3	3%	92%
2	2	20%	60%	8	0.2	2%	94%
3	1	10%	70%	9	0.2	2%	96%
4	0.8	8%	78%	10	0.2	2%	98%
5	0.7	7%	85%	11	0.1	1%	99%
6	0.4	4%	89%	12	0.1	1%	100%

通过上面的表3-3我们可以观察到，前四个因子累积贡献率已经超过75%，可以将12个因子压缩为前四个因子。PCA算法的核心，实际上是维度旋转。通过维度旋转来重新定义维度（变量），维度旋转的原则是尽可能与主要数据的特征保持同向，以达到减少维度（变量）的目的。

3.3.2 因子分析算法

因子分析是一种变量简化技术，从分析多个原始变量的相关关系入手，主要是研究多个变量的相关矩阵，找出支配相关关系的有限个潜在变量，达到用少数变量来解释复杂问题的目的。因子分析跟主成分分析有联系也有区别，从方法学原理上看，因子分析可以看成是主成分分析的推广，因子分析主要是研究解释原始变量之间的关系，而主成分分析重点在综合原始变量的信息。

3.3.2.1 因子分析算法思想

因子分析算法是，目标变量为 Y ，预测变量为 X_1 、 X_2 、 $X_3 \dots X_n$ ，对 X_1 到 X_n 这 n 个特性进行重新整合，找出控制所有特性的少数公共特性 Z_1 、 $Z_2 \dots Z_m$ ，用少数特性来解释复杂问题的一种方法。

例如公司老板对应聘者进行面试，在15个方面打分，分别为申请书的形式、外貌、专业能力、讨人喜欢、自信心、精明、诚实、推销能力、经验、积极性、抱负、理解能力、潜力、交际能力和适应性。我们可以对这15个方面进行重新整合，归结为应聘者的外露能力、讨人喜欢的程度、经验、专业能力和外貌5个方面。

3.3.2.2 因子分析算法实现

因子分析算法的基本步骤如下：

- 1) 输入原始数据 $X_{n \times p}$ ，计算均值和方差，进行标准化计算；
- 2) 求样本相关系数矩阵 $R = (r_{ij})$ ；
- 3) 求相关系数矩阵的特征根 λ_i ($\lambda_i \geq 0$, $i=1, 2, \dots, p$) 和相对应的标准正交的特征向量 L_i ；
- 4) 确定公共因子数；
- 5) 计算公共因子的共性方差 h_i^2 ；
- 6) 对载荷矩阵进行旋转，以求更好地解释公共因子；
- 7) 对公共因子作出专业性的解释。

因子分析算法的实质是寻找数据中的潜在因子，并应用潜在因子来解释由原变量显示的现象。

3.3.3 非负矩阵因子分解NMF算法

非负矩阵因子分解NMF算法（Non-negative Matrix Factorization, NMF），是一种特征抽取算法，它常常用于属性很多、且属性模糊或者有很弱的可预测性的数据集。通过合并属性，NMF算法可以产生有意义的模式或主题。NMF算法通过创建用户定义的特征数来压缩多元变量数据，每一个特征是原来属性集的一个线性合并，这些线性合并的系数是非负的。NMF算法可以简单描述为，对于任意给定的一个非负矩阵 V ，NMF算法能够寻找到一个非负矩阵 W 和一个非负矩阵 H ，使得满足 $V=W \times H$ ，从而将一个非负的矩阵分解为两个非负矩阵的乘积。

3.3.3.1 NMF算法思想

NMF算法压缩一个数据矩阵 V 为两个低阶矩阵 W 和 H 的乘积，以近似地等于 $W \times H$ 。NMF算法使用交互式过程来修改矩阵 W 和 H 的初始值，以便这个乘积接近矩阵 V 。这个交互式过程当近似值差聚集或者达到一个给定的迭代次数时终止。NMF算法提供了基于简单迭代的求解 W 、 H 的方法，求解方法具有收敛速度快、左右非负矩阵存储空间小的特点，它可将高维的数据矩阵降维处理，适合处理大规模数据。利用NMF进行文本、图像大规模数据的分析方法，较传统的处理算法速度更快、更便捷。

特征抽取的本质，是对反映实际问题的复杂矩阵进行分解，即将

复杂矩阵 V 近似分解成两个具有较低秩的 W 和 H 的乘积： $V=W \times H$ ，前面讲过的主成分分析和因子分析也是这个思路。数学意义上的矩阵分解，因子 W 和 H 中的元素可为正或负，但是在解决文字聚类、人脸识别、计算机图像工程、基因及细胞分析等方面，负值便没有实际意义，非负矩阵因子分解NMF算法很好地解决了这些问题。

3.3.3.2 NMF算法实现

对一个 M 维的随机向量 v 进行了 N 次的观测，记这些观测为 v_j ， $j=1, 2, \dots, N$ ，取 $V= [V_{\cdot 1}, V_{\cdot 2}, \dots, V_{\cdot N}]$ ，其中， $V_{\cdot j}=v_j$ ， $j=1, 2, \dots, N$ ，

NMF要求发现 $M \times L$ 的基矩阵 $W= [W_{\cdot 1}, W_{\cdot 2}, \dots, W_{\cdot N}]$ ，

和 $L \times N$ 的系数矩阵 $H= [H_{\cdot 1}, H_{\cdot 2}, \dots, H_{\cdot N}]$ ，

使得 $V \approx WH$ ，直观地表示为 $V_{\cdot j} = \sum_{i=1}^L W_{\cdot i} H_{i \cdot j}$

由于通常设定 $L \ll \min(M, N)$ ，即只用很少的基去描述大量的数据，所以只有在 W 包含了随机变量的本质特征时，才可能使 $V \approx WH$ 。

实现NMF的过程是一个优化求解的过程，Donoho等从理论上分析了NMF存在唯一解的条件，这个条件的苛刻性告诉我们：合理地构造一个目标函数，以此交替地优化 W 和 H ，从而得到NMF的一个局部最优解才是进行NMF的可行方法，这也是目前NMF算法构造的基本

思想。主要步骤如下：

1) 目标函数可以设为：

$$\|V - WH\|^2 = \sum_{ij} (V_{ij} - (WH)_{ij})^2$$

当且仅当 $V=WH$ 时目标函数为0，得到近似分解的最优解。

2) 用K-L相对熵来衡量分解的逼近程度：

$$D(V \| WH) = \sum_{ij} (V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}) \quad (3.26)$$

当且仅当 $V=WH$ 时目标函数为0，得到近似分解的最优解。

设立目标函数后通过迭代来寻找 W 和 H 。

3.4 关联规则

世间万物普遍存在着联系，有些联系是我们知道的，比如说有些疾病有遗传问题、肺癌跟吸烟习惯有关联等。更多的联系是我们现在还是未知的，需要我们去探索。数据挖掘的关联规则算法，可以发现大量数据中项集之间有趣的未知联系。

通过关联规则可以发现未知的、想象不到的联系。例如，药品与疗效之间的联系。某医药公司研究了一种治疗心脏病的药物，通过临床数据检验发现对于心脏病疗效不明显，但通过关联规则挖掘后发现对性功能疗效明显，服用此药物的病例中有大量数据表明性功能改善，因此此药物最后用于治疗性功能障碍。本节将先后介绍关联规则的概念、关联规则的两个常用算法Apriori算法和FP树频集算法，以及关联规则的评价指标Lift的有关内容。

3.4.1 关联规则概念

关联规则是指大量数据中项集之间的有趣关联或相关关系。关联规则是以事务为单位的，而事物（Event）是由项（Item）组成（Item Set，即项的集合），它最终寻求的是项与项之间的关系。我们用 i_i 表示项； T 表示事务， T 是一些项的集合； D 表示所有事务的集合也就是数据库。 $I=\{i_1, i_2, \dots, i_m\}$ 是所有项的集合，每个事务 T 是一些项的集合， T 包含在 I 中，每个事务可以用唯一的标识符TID来标识。设 X 为某些项的集合，如果 X 包含在 T 中，则称事务 T 包含 X 。一个关联规则是形如 $X \Rightarrow Y$ 的蕴涵式，这里 $X \subset I$ ， $Y \subset I$ 并且 $X \cap Y = \Phi$ 。规则 XY 在事务数据库 D 中的支持度（Support）是事务集中包含 X 和 Y 的事务数与所有事务数之比，记为 $\text{Support}(X \Rightarrow Y)$ ，即 $\text{Support}(X \Rightarrow Y) = P(X \cup Y)$ 。规则 $X \Rightarrow Y$ 在事务集中的置信度（Confidence）是指包含 X 和 Y 的事务数与包含 X 的交易数之比，记为 $\text{Confidence}(X \Rightarrow Y)$ ，即 $\text{Confidence}(X \Rightarrow Y) = P(Y|X)$ 。给定一个事务集 D ，挖掘关联规则问题就是寻找支持度和置信度分别大于用户给定的最小支持度（minsupp）和最小置信度（minconf）的关联规则。

关联规则挖掘问题可分解为以下两个子问题：

- 找出事务数据库 D 中所有大于等于用户指定最小支持度的项集（itemset）。具有最小支持度的项集称为最大项集。项集的支持度指包含该项集的数目。

· 利用最大项集生成所需要的关联规则对每一最大项集A，找到A的所有非空子集a，如果比率： $\text{Support}(A) / \text{Support}(a) \geq \text{minconf}$ ，就生成关联规则 $a \Rightarrow (A-a)$ 。
 $\text{Support}(A) / \text{Support}(a)$ ，即规则 $a \Rightarrow (A-a)$ 的置信度。

3.4.2 Apriori算法

Apriori算法是关联规则的一个传统且比较有效的方法。基于经典的Apriori算法，后来又产生了很多不同的改进版本，但它们的基本原理都是类似的。这里，我们将经典的Apriori算法进行展示和回顾，因为很多商业项目中仍然使用它，并且它的基本思想在数据挖掘应用的很多地方还具有借鉴作用。

3.4.2.1 Apriori算法思想

Apriori算法基于先验原理，它反映了子集与超集之间的关系：即频繁项集的所有非空子集都必须是频繁的，非频繁项集的所有超集必是非频繁的。如果项集 I 不满足最小支持度阈值 s ，则 I 不是频繁的，即 $P(I) < s$ 。如果项 A 添加到 I ，则结果项集 $(I \cup A)$ 不可能比 I 更频繁出现。因此， (I, A) 也不是频繁的，即 $P(I \cup A) < s$ 。因此，Apriori算法的性质主要是用于搜索频繁项集的时候对候选式的筛选过程。Apriori算法中利用Apriori性质，能够比较好地避免盲目地搜索，提高频繁项集的查找效率。

3.4.2.2 Apriori算法实现

Apriori算法是使用逐层迭代找出频繁项集的。算法需要对数据集进行多步处理。

Apriori算法的基本过程如下：

输入：事务数据库D；最小支持度阈值。

1) 简单统计所有含一个元素项集出现的频率，并找出那些不小于最小支持度的项集，即一维最大项集。

2) 开始循环处理，直到再没有最大项集生成。

3) 循环过程是：第k步中，根据第k-1步生成的(k-1)维最大项集产生k维侯选项集，然后对数据库进行搜索，得到侯选项集的项集支持度，与最小支持度比较，从而找到k维最大项集。

输出：D中的频繁项集L。

算法缺点如下：

- 在每一步产生侯选项集时循环产生的组合过多，没有排除不应该参与组合的元素；

- 每次计算项集的支持度时，都对数据库D中的全部记录进行了一遍扫描比较。如果是一个大型的数据库的话，这种扫描比较会大大增加计算机系统的I/O开销，而这种代价是随着数据库的记录的增加呈现出几何级数的增加。因此人们开始寻求一种能减少这种系统I/O开销的更为快捷的算法。

3.4.3 FP树频集算法

上一小节，我们简要介绍了Apriori算法，本小节我们将介绍另外一种算法，即FP树频集算法。FP树频集算法思想完全不同于Apriori算法，它是一种从FP树紧凑数据结构中抽取频繁项集的方法。

3.4.3.1 FP树频集算法思想

在经过第一遍扫描之后，把数据库中的频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树分化成一些条件库，每个库和一个长度为1的频集相关，然后再对这些条件库分别进行挖掘。当原始数据量很大的时候，也可以结合划分的方法，使得一个FP树可以放入主存中。

3.4.3.2 FP树频集算法实现

基于FP树的最大频繁模式挖掘算法，主要包括构造频繁模式树FP树和利用FP树挖掘最大频繁模式两大步骤。逻辑算法如下：

· 频繁模式树FP树的构造：

输入：事务数据库D；最小支持度Support

输出：事务数据库D的频繁模式树，FP树

· 利用FP树挖掘最大频繁模式:

输入: D的频繁模式树FP树; 频繁项最小支持度Support

输出: D中的最大频繁模式集

3.4.4 提升Lift

无论是Apriori算法，还是FP树算法，在某些情况下，即使支持度和置信度两个指标都比较高，但产生的规则依然可能是没有用的。

Lift给出了评估规则质量的一个新指标。Lift指示给定的前件和后件随机同时出现的规则的强度，它提供了一个改进的信息，以提高给定前件下后件出现的概率。Lift的定义如下：

$$\frac{(\text{Rule Support})}{(\text{Support (Antecedent)} \times \text{Support (Consequent)})}$$

其中：

Rule Support: 是关联规则的支持度，

Support (Antecedent)：是关联规则前件的支持度，

Support (Consequent)：是关联规则后件的支持度。

任何Lift<1的规则都不能显示是一个真正的内在并发现象，无论它的支持度和置信度有多高，因为它不能比随机现象更能预测前件和后件的并发出现。

3.5 分类和预测

分类问题是通俗易懂的问题，分类技术是应用广泛的方法和手段。我们把分类和预测统称为推测。分类就是应用已知的一些属性数据去推测一个未知的离散型的属性数据，而这个被推测的属性数据的可取值是预先定义的。要很好地实现这种推测，就需要事先在已知的一些属性和未知的离散型属性之间建立一个有效的模型，即分类模型。可用于分类的算法有决策树、朴素贝叶斯分类、神经网络、logistic回归、支持向量机，等等。

大家都关心房子的价格，有时我们可以根据一套房子的地点，房间数、房屋面积及其他指标如月份、市场数据、经济数据等，来预测房子的价值。这就属于预测问题。这里的“预测”是数据挖掘中狭义的定义，它也被称为回归，或统称为回归预测，就是应用已知的一些属性数据去推测一个未知的连续型属性数据。同样，要很好地实现这种推测，也需要事先在已知的一些属性和未知的连续型属性之间建立一个有效的模型，即预测模型。用于预测的算法也很多，如神经网络、支持向量机、广义非线性模型等。以上提到的几个算法各有优缺点，但在商业应用中支持向量机方法综合性能更好一些，应用也得较多一些。下面我们就重点介绍一下支持向量机的算法。

3.5.1 支持向量机

支持向量机（Support Vector Machine，简称SVM）是由Cortes&Vapnik首先提出来的，是多年来引人注目的分类技术，它可很好地解决小样本、非线性及高维数据识别分类问题。支持向量机也称为最大边缘区分类器，可以同时最小化经验误差与最大化几何边缘区。支持向量机算法作为数据挖掘的一种算法，在实践应用中与其他算法比较时，总能表现出更好的性能和效果，是因为SVM在分类原理和方法上是一个根本性的解决方案，它给出的是全局最优解，而诸如神经网络和决策树分类等算法则通常给出的是一个局部最优解。

3.5.1.1 最大边缘超平面

图3-16是一个肿瘤病人数据集，三角形代表的是良性肿瘤实例，圆圈代表的是恶性肿瘤实例，良性肿瘤的数据要更多一些。我们试图寻找一条线，能将两类数据完全划分开来。

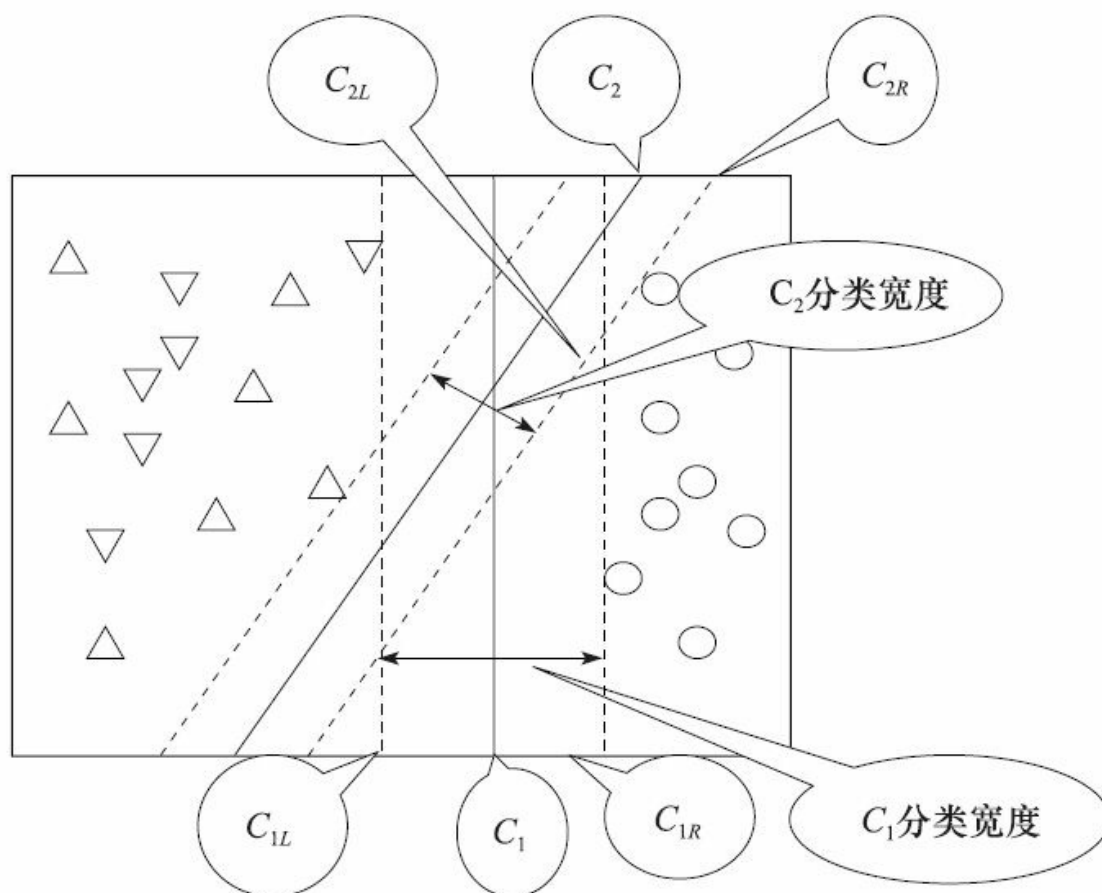


图 3-16

SVM的思想其实很简单，就是要我们找到类似图3-16中的这样一条线，我们称之为一个超平面，即在高维的空间中将数据分类切割的平面，“切入”数据集中能清楚地将数据分为两类，并且该超平面具有最大的分类边缘（或称为分类宽度），该超平面称为最大边缘超平面。我们来看图3-16，从图中我们可能找到好多个超平面，都能将两类数据完全分开。考察其中的两个，一个是超平面 C_1 ，一个是超平面 C_2 。其中 C_{1L} 是 C_1 向左平移到三角形数据最右边一个点的位置， C_{1R} 是 C_1 向右平移到圆图形数据集最右边一个点的位置， $C_{1R}-C_{1L}$ 称为 C_1 的分类宽度。同理，我们也会得到 C_2 分类宽度 $C_{2R}-C_{2L}$ 。显然， $(C_{1R}-C_{1L}) > (C_{2R}-C_{2L})$ 。假设 $C_{1R}-C_{1L}$ 是本图中能找到的所有超平面的

分类宽度最大的一个，那 C_1 就是我们要找的SVM分类器，也成为该数据集的分类决策边界。

3.5.1.2 线性完全可分支持向量机

我们在前面看到的具有最大的分类宽度的超平面 C_1 ，能够将恶性肿瘤数据和良性肿瘤数据完全正确地分开，我们将 C_1 称为是一个线性完全可分类的分类器。线性完全可分情形，即是可以用一个最大分类宽度超平面将两类具有多个维度的点完全分开，分类超平面方程是 $f(x) = w^T x + b = 0$ 。其中， w 和 b 是多元线性方程的系数矩阵和常数，也正是我们要求得的参数。我们将判别函数进行归一化，使两类所有样本数据都满足 $|f(x)| \geq 1$ ，此时离分类面最近的样本的 $|f(x)| = 1$ ，而要求分类面对所有样本都能正确分类，就是要求它满足：

$$y_i (w^T x_i + b) - 1 \geq 0, i=1, 2, \dots, n. \quad (3.27)$$

公式（3.27）中使等号成立的那些样本即为支持向量（Support Vectors）。两类样本的分类空隙（Margin）的宽度：

$$\text{Margin} = 2/\|w\| \quad (3.28)$$

因此，最优分类面问题可以表示成如下的约束优化问题，即在条件（3.27）的约束下，求函数：

$$\phi(w) = \frac{1}{2} \|w\|^2 = \frac{1}{2} (w^T w) \quad (3.29)$$

的最小值。为此，可以定义如下的拉格朗日（Lagrange）函数：

$$L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^n \alpha_i [y_i (w^T x_i + b) - 1] \quad (3.30)$$

其中， $\alpha_i \geq 0$ 为拉格朗日（Lagrange）系数，我们的问题是对 w 和 b 求拉格朗日函数的最小值。把公式（3.30）分别对 w 、 b 、 α_i 求偏微分并令它们等于0，得：

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^n \alpha_i y_i x_i \quad (3.31)$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \alpha_i y_i = 0 \quad (3.32)$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \Rightarrow \alpha_i [y_i (w^T x_i + b) - 1] = 0 \quad (3.33)$$

以上三式加上原约束条件，可以把原问题转化为如下凸二次规划的对偶问题：

$$\begin{cases} \max \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i^T x_j) \\ s. t. \quad a_i \geq 0, i = 1, \dots, n \\ \sum_{i=1}^n a_i y_i = 0 \end{cases} \quad (3.34)$$

这是一个不等式约束下二次函数机制问题，存在唯一最优解。若 α_i^* 为最优解，则

$$w^* = \sum_{i=1}^n a_i^* y_i x_i \quad (3.35)$$

α_i^* 不为零的样本即为支持向量，因此，最优分类面的权系数向量是支持向量的线性组合。

b^* 可由约束条件 $\alpha_i [y_i (w^T x_i + b) - 1] = 0$ 求解，由此求得的最优

分类函数是：

$$g(x) = \text{sgn}((w^*)^T x + b^*) = \text{sgn}\left(\sum_{i=1}^n a_i^* y_i x_i^* x + b^*\right) \quad (3.36)$$

$\text{sgn}()$ 为符号函数。

线性完全可分SVM是一种理想的分类器，它完全正确地将数据分为两类。但在实践中，更多的情况是，我们找不到一个超平面能完全正确地分开两类数据，只能找到相对分类错误率较小的超平面。还有一种情况，即使能找到一个线性完全可分的超平面，但它的分类宽度可能很小，这种分类器在对将来未知的数据进行分类时，决策边界任何轻微的干扰都会对分类效果产生很大的影响，我们称之为过度拟合。因此，我们可能更倾向于选择一个虽然划分有一定误差，但具有更大分类宽度、抗干扰能力更强的超平面分类器，我们将这类分类器称为线性不完全可分SVM。

这里，我们再简单解释一下过度拟合现象，应该说任何有监督学习算法都会可能产生过度拟合的情况。举个简单的例子，有一些灵长类动物（如猩猩、猴子、狒狒等）在体貌、习性等方面有着一定的相似性，我们构建了一个灵长类动物自动分类模型，根据灵长类动物的体貌特征，我们力争能将猩猩、猴子、狒狒等比较准确地区分开来。该模型以猩猩、猴子、狒狒各十只作为训练数据，但这三十只动物中有几个异常点，就是有几只猩猩、猴子、狒狒长得较特殊，与它们本类动物的体貌特征有点差异。但灵长动物自动分类模型过分训练了这三十只动物的所有个体的体貌特征，使得该模型对于这三十个样本数

据拟合得非常好，也就是它过度描绘了三十只动物的综合“画像”。当从其他大量的灵长动物中另外随机抽出三十一只猩猩、猴子、狒狒对灵长动物自动分类模型进行测试时，该模型的分类正确率大幅下降，它错误地区分了不少动物，这就是分类模型的过度拟合。

再简单来说，如果找来一只猩猩，按照它的相貌画出一只猩猩的主要相貌特征，如果我们过分地描绘了该猩猩体貌特征的每个细节，然后按照这幅画像作为标准来衡量另外一只猩猩是否也属于猩猩类动物，就有可能认为它不符合猩猩的“一般”特征，从而认为这只猩猩不是猩猩了。产生过度拟合的原因，是我们人为地将模型的训练精度设置得过高，导致模型失去普适性。过度拟合会造成模型对当前数据训练、拟合得过强，精度过高，但是模型对于其他新来数据的区分能力下降了很多。我们在后面的其他分类和预测模型的讨论中，都会涉及过度拟合的问题。实践中，我们需要根据实际情况，在模型对训练数据的学习度和对测试数据的验证力两方面找到一个最佳的平衡点，使得模型既能较好地拟合现有数据又能较好地拟合测试数据以及能够应用于新来的数据。

3.5.1.3 线性不完全可分支持向量机

我们看图3-17，有两个超平面 C_1 和 C_2 ，它们具有不同的分类误差和分类宽度。在很多情况下，数据集就像图3-17中显示的那样，数据分布呈现两类数据在中间地带有不少交叉的情况。我们就是要寻找一个分类宽度和分类误差均衡的分类器，它既具有较好的分类准确率，

同时又具有较强的抗噪声能力，避免过度拟合现象。这时，我们把这个分类宽度和分类误差均衡的超平面称为线性不完全可分支持向量机，因为它的分类正确率小于100%。从图中我们可看到， C_2 比 C_1 具有较小的分类误差和分类宽度，我们会根据分类的不同目的，根据分类目标所要求的不同分类准确率和成本来决定是选用 C_2 还是 C_1 。

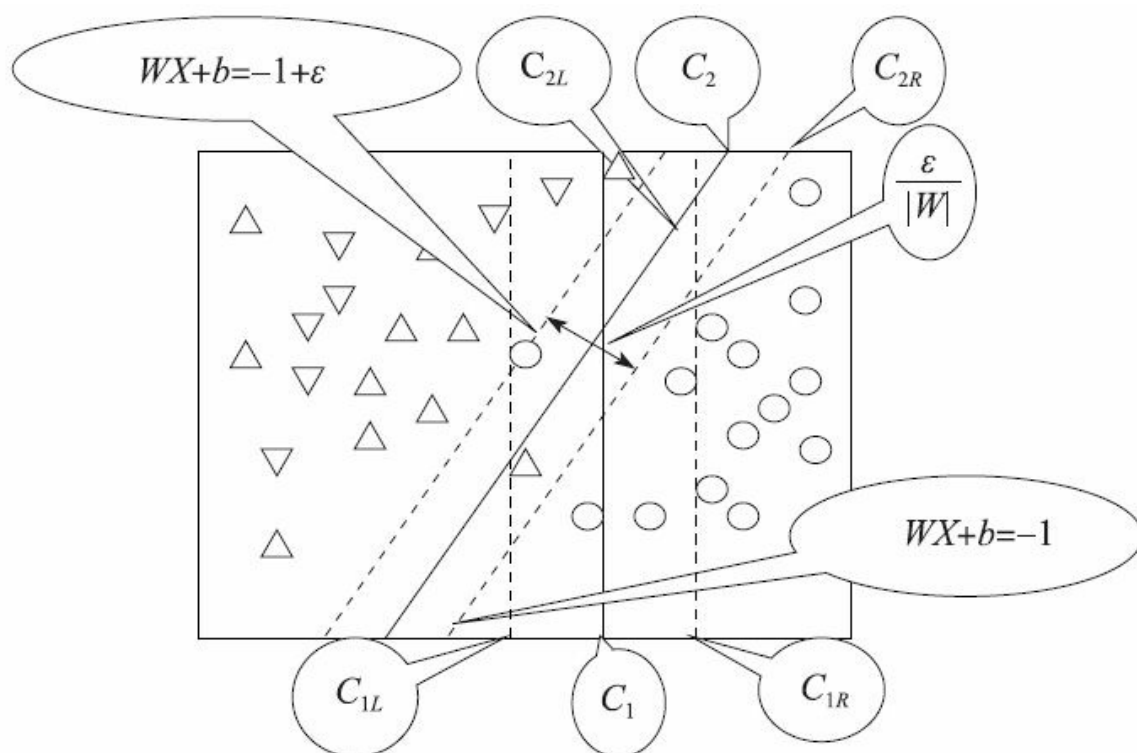


图 3-17

线性不完全可分支持向量机的算法实现如下：当用一个超平面不能把两类点完全分开时（有少数点被错分），可以引入松弛变量 ξ_i ($\xi_i \geq 0$, $i=1, \dots, n$)，使超平面 $w^T x + b = 0$ 满足：

$$y_i (w^T x_i + b) \geq 1 - \xi_i \quad (3.37)$$

当 $0 < \xi_i < 1$ 时样本点 x_i 仍旧被正确分类，而当 $\xi_i \geq 1$ 时样本点 x_i 被错分。为

此，引入以下目标函数：

$$\psi(w, \xi) = \frac{1}{2}w^T w + C \sum_{i=1}^n \xi_i \quad (3.38)$$

其中C是一个正常数，称为惩罚因子，此时SVM可以通过二次规划（对偶规划）来实现：

$$\begin{cases} \max \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i^T x_j) \\ s. t \quad 0 \leq a_i \leq C, i = 1, \dots, n \\ \sum_{i=1}^n a_i y_i = 0 \end{cases} \quad (3.39)$$

3.5.1.4 非线性支持向量机

我们前面介绍了SVM的两种情况，线性完全可分和线性不完全可分支持向量机，它们都是基于线性空间进行分类的。但实践中，很多情况是需要非线性空间下进行数据分类的。目前尚没有一个能够直接处理非线性数据分类的SVM算法，我们需要将原来的坐标空间进行变换，应用映射函数将其变换为新的坐标空间，以便可以使用线性支持向量机方法进行分类。要将非线性空间映射为线性空间，从技术上就需要利用核函数的支持。

核函数算法实现如下：若在原始空间中的简单超平面不能得到满意的分类效果，利用核函数将输入样本空间映射到高维特征空间，令：

$$K(x_i, x_j) = \langle \Phi(x_i) \cdot \Phi(x_j) \rangle \quad (3.40)$$

用核函数 $K(x_i, x_j)$ 代替最优分类平面中的点积 $x_i^T x_j$ ，就相当于把原特征空间变换到了某一新的特征空间，此时优化函数变为：

$$Q(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (3.41)$$

而相应的判别函数式则为：

$$g(x) = \text{sgn}[(w^*)^T \phi(x) + b^*] = \text{sgn}\left(\sum_{i=1}^n \alpha_i^* y_i K(x_i, x) + b^*\right) \quad (3.42)$$

应用核函数将数据的非线性维度变换为线性维度后，再使用前面的线性支持向量机方法对数据进行分类。

其实，SVM的核心思想就是利用核函数将输入样本空间映射到高维特征空间，在这个空间中求一个最优分类面 $f(x) = w^T \cdot x + b = 0$ ，根据 $f(x)$ 构造新的符号函数 $g(x)$ ，根据 $g(x)$ 的取值将数据点分类。我们会在后面章节中，大量应用SVM算法开展各类数据挖掘项目（见第4章、第5章、第6章、第7章、第8章）。

3.5.2 Logistic回归算法

Logistic回归算法（Logistic Regression）是统计学习中经典的分类算法，属于对数线性模型，传统上它在流行病学研究中应用较多。例如，探索代谢综合征的危险因素，根据危险因素预测代谢综合征发生的概率等（见第5章）。作为基于概率的一种回归算法，Logistic回归是数据挖掘中一种全局最优解的分类方法，而且该模型一旦验证成立并准确，它不仅为我们带来了一种可靠的分类方案，而且还带来了极具可解释性的一个显式模型。

3.5.2.1 Logistic回归算法思想

我们观察一个数据集（LOGISIC数据）的样本数据，如表3-3所示。数据集有 X_1 、 X_2 、 X_3 、 X_4 四个自变量（三个连续型变量和一个离散型变量），应变变量 Y 为二元离散变量。我们试图 $Y=f(X_1, X_2, X_3, X_4)$ 的Logistic回归模型，以期望根据 X_1 、 X_2 、 X_3 、 X_4 的取值来确定 Y 的分类取值。计划应用数据集中70%的数据建立Logistic回归模型，然后将数据集中其余30%的数据用Logistic回归模型进行测试，如果测试的结果令人满意，则可以应用该模型对与数据集（LOGISIC数据）同质的新数据进行预测分类。

表3-3 LOGISIC数据

X_1	X_2	X_3	X_4	Y
3	22	NO	300	0
5	3	NO	350	0
25	344	YES	446	1
14	28	YES	121	0
2	77	NO	45	0
16	34	NO	233	1
28	66	YES	654	1
19	100	YES	277	0
13	87	YES	321	1
33	45	YES	335	1

下面，具体讨论一下Logistic回归分类模型的建立方法。

设 X_1, X_2, \dots, X_n 为一组自变量， Y 为因变量。当 Y 是阳性反应时，记为 $Y=1$ ；当 Y 是阴性反应时，记为 $Y=0$ 。用 P 表示发生阳性反应的概率；用 Q 表示发生阴性反应的概率，显然 $P+Q=1$ 。Logistic回归目的是从特征学习出一个0/1分类模型，而这个模型是将特性的线性组合作为自变量，由于自变量的取值范围是负无穷到正无穷。因此，使用Logistic函数（或称作sigmoid函数）将自变量映射到 $(0, 1)$ 上，映射后的值被认为是属于 $y=1$ 的概率。

假设函数

$$f_{\beta}(x) = g(\beta^T x) = \frac{1}{1 + e^{-\beta^T x}} \quad (3.43)$$

其中 x 是 n 维特征向量，函数 g 就是Logistic函数 $g(z) = \frac{1}{1+e^{-z}}$ 的图像，如图3-18所示。

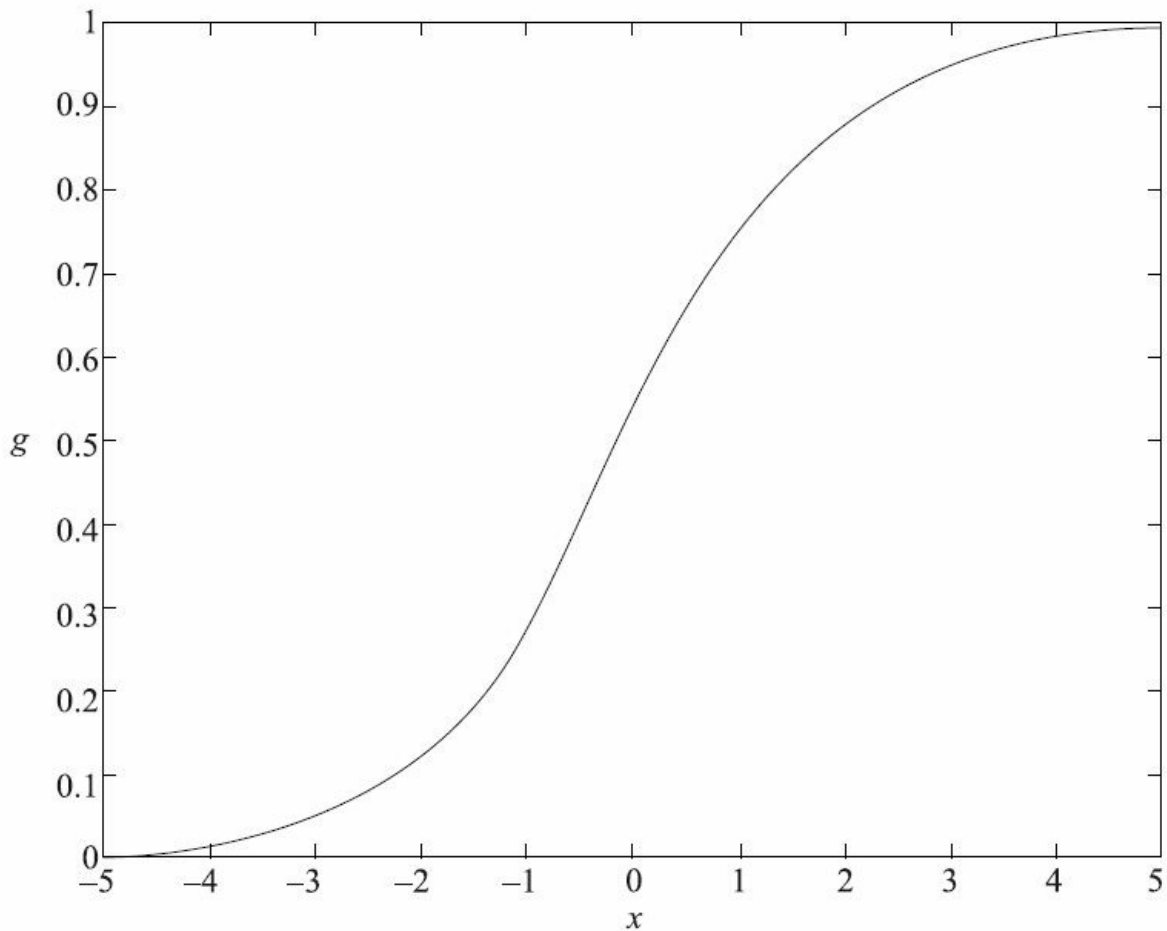


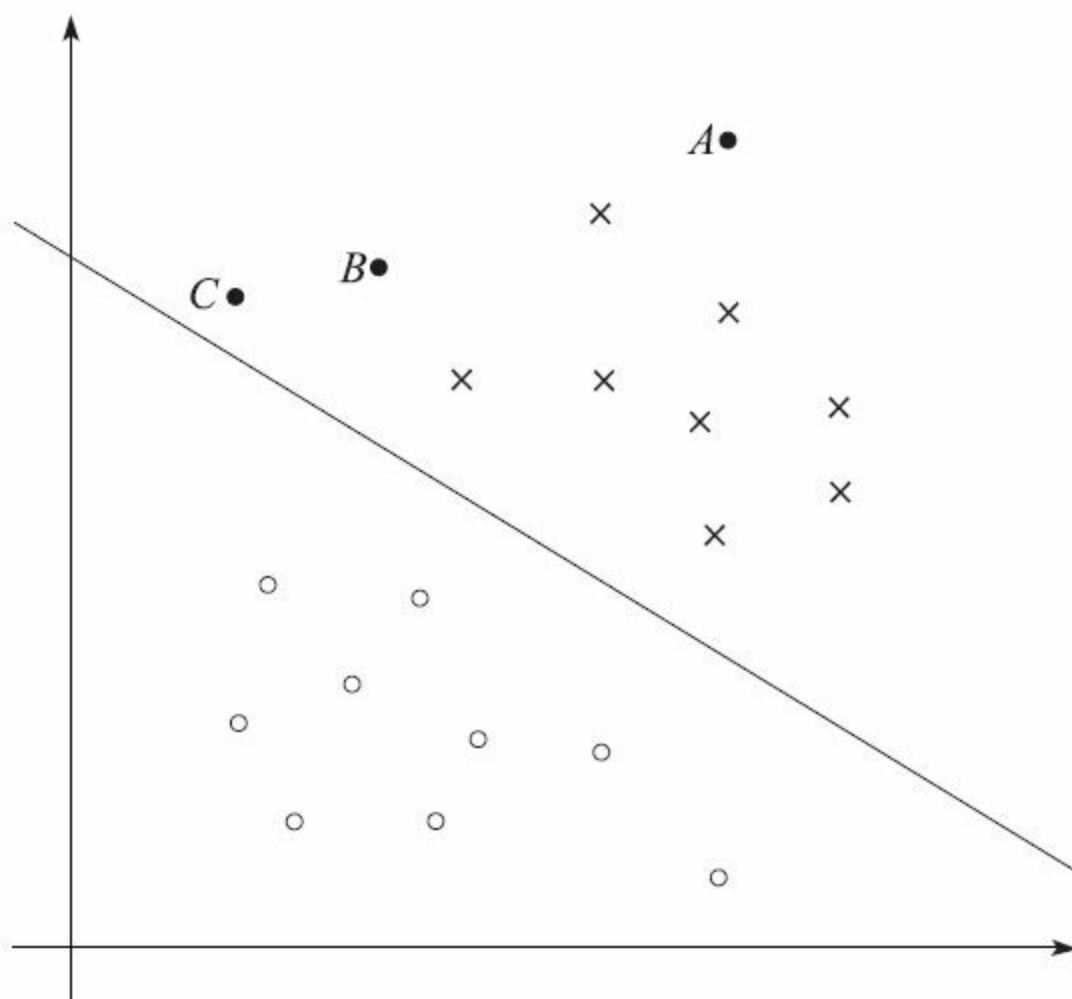
图 3-18

从图中可以看到，将无穷映射到了 $(0, 1)$ ，而假设函数就是特征属于 $y=1$ 的概率。

$y=1$ 的条件概率为： $p(y=1|x; \beta) = f_{\beta}(x)$ ； $y=0$ 的条件概率为： $p(y=0|x; \beta) = 1 - f_{\beta}(x)$ 。当我们要判别一个新来的特征属于哪个类时，只需求 $f_{\beta}(x)$ ，若大于0.5就是 $y=1$ 的类，反之属于 $y=0$ 类。

再审视一下 $f_{\beta}(x)$ ，发现 $f_{\beta}(x)$ 只和 $\beta^T x$ 有关， $\beta^T x > 0$ ，那么 $f_{\beta}(x) > 0.5$ ， $g(z)$ 只不过是用来映射，真实的类别决定权还在 $\beta^T x$ 。还有当 $\beta^T x \rightarrow +\infty$ 时， $f_{\beta}(x) = 1$ ，反之 $f_{\beta}(x) = 0$ 。如果我们只从 $\beta^T x$ 出发，希望模型达到的目标无非就是让训练数据中 $y=1$ 的特征 $\beta^T x$ 远大于0，而 $y=0$ 的特征 $\beta^T x$ 远小于0。Logistic回归就是要学习得到 β ，使得正例的特征远大于0，负例的特征远小于0，强调在全部训练实例上达到这个目标。

分类思想的图形化表示如图3-19所示。



中间那条线是 $\beta^T x=0$ ，Logistic回归强调所有点尽可能地远离中间那条线。学习出的结果也就中间那条线，这点有些像支持向量机（SVM）算法。Logistic回归与SVM都是要在数据中找出一条线（超平面），能准确地将两类数据分开，并且两类数据尽可能地远离这条线（超平面）；不同的是Logistic回归是基于概率的非线性回归模型，而SVM是基于多元线性方程的模型。

Logistic 回归模型为：

$$P = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n}} \quad (3.44)$$

同时可以写成：

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}} \quad (3.45)$$

式中 β_0 是常数项； β_j （ $j=1, 2, \dots, k$ ）是与研究因素 X_j 有关的参数，称为偏回归系数。事件发生的概率 P 与 $\beta^T X$ 之间呈曲线关系，当 $\beta^T X$ 在 $(-\infty, +\infty)$ 之间变化时， P 在 $(0, 1)$ 之间变化。

得到 β_j （ $j=0, 1, 2, \dots, n$ ）的方法如下。 $P(y_i=1)=p_i$ ， $P(y_i=0)=1-p_i$ ，所以 y_i 得概率函数为：

$$P(y_i) = p_i^{y_i} (1-p_i)^{1-y_i}, \quad y_i=0, 1 \quad i=1, 2, \dots, n$$

于是 y_1, y_2, \dots, y_n 的似然函数为：

$$L = \prod_{i=1}^n P(y_i) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} \quad (3.46)$$

将 $p_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in}}}$ 代入并取对数得到:

$$\ln L = \sum_{i=1}^n [y_i(\beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in}) - \ln(1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_n x_{in}})] \quad (3.47)$$

通过 Newton-Raphson 迭代法计算得到 $\beta_j (j=0, 1, 2, \dots, k)$ 的估计值。

3.5.2.2 Logistic回归算法流程

设 X_1, X_2, \dots, X_n 为一组自变量, Y 为应变量。Logistic回归算法主要步骤如下:

- 1) 筛选自变量, 可以通过特征选择算法选择与因变量 Y 相关性强的自变量。
- 2) 对选择后的变量检验其相关性, 剔除相关性强的冗余变量。
- 3) 对于剩余的自变量和 Y , 通过实验数据对未知参数 $\beta_j (j=0, 1, 2, \dots, k)$ 进行极大似然估计, 形成 Logistic 回归模型。
- 4) 用似然比、计分检验和 Wald 检验等方法对 $\beta_j (j=0, 1, 2, \dots, k)$ 进行检验, 剔除没有显著性的变量。
- 5) 对模型结果进行拟合优度和预测准确度检验, 选择最优模型。

3.5.2.3 Logistic回归算法的用途

Logistic回归除了可以作为二元分类算法外，还有诸如流行病学等其他典型用法，如探索参数 β_j 的意义。

通过 X_j 的系数 β_j ($j=0, 1, 2, \dots, n$) 研究 X_j 对调整优势比的影响，由 $P = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}}$ 得到

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (3.48)$$

1) β_0 的意义：当相关因素不起作用时， β_j , $j=1, 2, \dots, n$ 都为0, $\ln\left(\frac{P}{1-P}\right) = \beta_0$, 所以 β_0 得意义为当各相关因素不存在作用时，阴性与阳性的概率之比的自然对数。

2) $\beta_j, j=1, 2, \dots, n$ 的意义，设 X_j 的两个不同取值为 $X_j=b$ 和 $X_j=a$, 假设其他值保持不变，假设两值分别对应的P为 P_1 和 P_0 ,

于是有：

$$\ln OR_j = \ln\left[\frac{P_1/(1-P_1)}{P_0/(1-P_0)}\right] = \beta_j(b-a) \quad (3.49)$$

$$OR_j = e^{\beta_j(b-a)} \quad (3.50)$$

OR_j 称为调整优势比，表示扣除其他自变量影响后自变量 X_j 的作用。

当 X_j 仅取二值， $b=1$, $a=0$,

$$X_j = \begin{cases} 1 & \text{1 为暴露} \\ 0 & \text{0 为非暴露时即为流行病学上的优势比} \end{cases}$$

$\beta_j=0$, $OR_j=1$, 说明 X_j 对是否出现阳性结果不存在影响；

$\beta_j > 0$, $OR_j > 1$, 说明 X_j 是危险因子;

$\beta_j < 0$, $OR_j < 1$, 说明 X_j 是保护因子。

样本量较大时, 可以利用正态近似法得到 β_j 和 OR_j 的置信区间。

3.5.3 朴素贝叶斯分类算法

朴素贝叶斯分类算法是一个经典的、但在很多场景下都比较有效的分类方法。我们先看一下朴素贝叶斯分类算法的原理。朴素贝叶斯分类的算法基础是贝叶斯公式。

3.5.3.1 贝叶斯公式

1. 贝叶斯公式原理

贝叶斯决策理论方法是统计模型决策中的一个基本方法，其基本思想是：

- 已知类条件概率密度参数表达式和先验概率。
- 利用贝叶斯公式转换成后验概率。
- 根据后验概率大小进行决策分类。

贝叶斯公式，是在已知 $P(A|B)$ 的前提下求 $P(B|A)$ ：

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (3.51)$$

观察如下推导。有A、B两个独立事件，如图3-20所示，绿色代表事件A，红色代表事件B。描述两个条件概率之间的关系，在事件B发生的情况下事件A也同时发生的概率为 $P(A|B)$ ，在事件A发生的情况

下事件B也同时发生的概率为 $P(B|A)$ 。

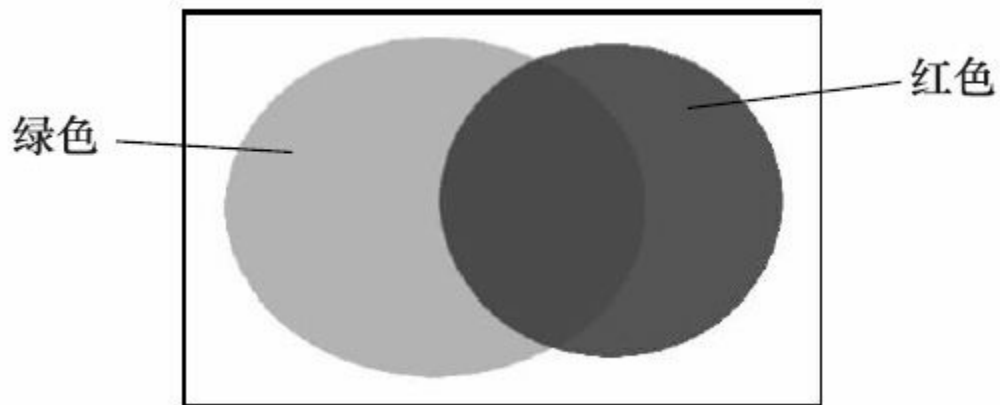


图 3-20

按照乘法法则： $P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$ ，公式也可变形为：

$P(B|A) = P(A|B) \times P(B) / P(A)$ 。对于两个以上变量的情况，贝叶斯定理亦成立：

$P(A|B, C) = P(A) \times P(B|A) \times P(C|A, B) / (P(B) \times P(C|B))$ ，也就是：

$(P(B) \times P(C|B) \times P(A|B, C) = P(A) \times P(B|A) \times P(C|A, B))$ ，如图3-21所示。

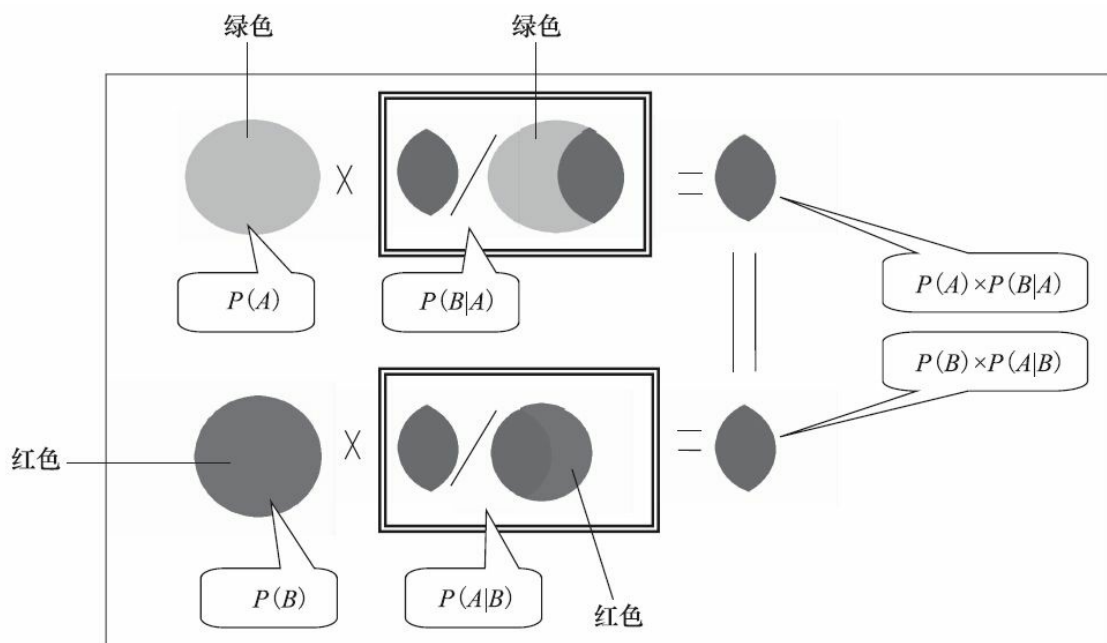


图 3-21

2. 贝叶斯公式的应用举例

例如，现分别有M、N两个容器，在容器M里分别有七个红球和三个白球，在容器N里有一个红球和九个白球。抽出了一个红球，且已知球来自M、N两个容器里任意一个。这个红球是来自容器M的概率是多少？

首先，“红球来自M容器”可以表达为：

- 事件B：抽出的是红球，
- 事件A：来自M容器，
- 事件A发生时，事件B也发生。

按照公式 $P(B|A) = P(A|B) \times P(B) / P(A)$ ，要推测“红球来

自M容器”，需要知道三个信息：

· 从两个容器中抽出红球的概率 $P(B) : (7+1) / (7+3+1+9)$
 $=8/20$

· 从两个容器中抽出来自M容器的概率 $P(A) : 1/2$

· 单从M容器抽出红球的概率 $P(AB) : 7 / (7+3) =7/10$

则： $P(BA) = (7/10) \times (8/20) / (1/2) =0.56$

答：这个红球是来自容器M的概率是56%。

3. 贝叶斯公式的应用实践

我们再看一下专业领域的贝叶斯公式应用。假设患者症状为咳嗽，那么既可能是某型流感患者，也可能仅仅是普通感冒患者。患有某型流感的患者除了咳嗽还有7/10的患者伴有高热，而患有普通感冒的患者只有1/10伴有高热。假如一位患者同时有咳嗽和高热，那么该患者患有流感的概率有多高？

首先，我们把医学问题转换成为数学问题。假设上面例子M、N容器都是咳嗽患者。来自M容器是流感患者、N容器是普通感冒患者。再以红色球表示高热，白球表示非高热。那么，这个流感诊断问题被转换成上述“抽取红球来自M容器的概率”的问题了。

1) “普通感冒患者和流感患者高热”的概率相当于“抽出红球”的

概率：

$$P(B) : (7+1) / (7+3+1+9) = 8/20$$

2) “患有流感（或普通感冒）”的概率相当于“来自M(N) 容器”的概率：

$$P(A) : \text{为} 1/2$$

3) “患有某型流感的患者除了咳嗽还有7/10的患者伴有高热”相当于“从M容器抽出红球”的概率：

$$P(AB) : 7 / (7+3) = 7/10$$

则“一位患者同时有咳嗽和高热，那么该患者患有流感”的概率相当于“这个红球是来自容器M的概率：

$$P(BA) = (7/10) \times (8/20) / (1/2) = 0.56$$

也就是说，高热且咳嗽的患者有56%的可能性患有流感。

3.5.3.2 贝叶斯公式扩展应用

读者请参见本章3.4节“关联规则”。数据挖掘中，关联规则中的支持度（Support）是事件A和事件B同时发生的比例，即概率；置信度（Confidence）是事件A发生的条件下事件B同时发生的比例，即条件概率。

假设：前件：事件A

后件：事件B

则：

支持度= $P(A \cap B)$ ，

置信度= $P(B|A) = P(A \cap B) / P(A)$ 。

再假设：前件：事件B

后件：事件A

则：

支持度= $P(A \cap B)$

置信度= $P(A|B) = P(A \cap B) / P(B)$

通过贝叶斯公式可以很方便地从一个关联方向推测它的反向规则：

$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$ ，即：

支持度 ($A|B$) = 支持度 ($B|A$)

= 概率 (A) \times 置信度 ($B|A$)

= 概率 (B) \times 置信度 ($A|B$)，

所以有：概率（A）/概率（B）=置信度（A|B）/置信度（B|A）。

因为置信度最大为1，所以如果经典关联规则置信度取>0.6，那么当前件和后件中一个发生的概率是另一个发生的概率两倍（至少1.67倍）时，必然仅仅有一个方向的关联规则可能入选。关联规则的方向是事物的重要信息，单向比双向更可能有因果关系。

3.5.3.3 朴素贝叶斯分类算法

1.朴素贝叶斯分类算法原理

朴素贝叶斯分类算法思想是，对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个概率最大就认为此待分类项属于哪个类别。表述如下：

- 确定X: $\{x_1, \dots, x_n\}$ 项元素集合, Y: $\{y_1, \dots, y_m\}$ 为类别集合。
- 计算 $P(y_1|x_i)$; $P(y_2|x_i)$; ...; $P(y_k|x_i)$ 。
- 如果 $P(y_k|x_i) = \max\{P(y_1|x_i), P(y_2|x_i), \dots, P(y_m|x_i)\}$ 则 $x_i \in y_k$;

算法需要解决的是怎样来计算 $P(y_1|x_i)$; $P(y_2|x_i)$; ...; $P(y_k|x_i)$ 。可以找到一个已知分类的待分类项集合，称这个集合为训

训练样本集。统计各个类别下的条件概率估计，即通过训练样本集得到：

$$P(x_i|y_1); P(x_i|y_2); \dots; P(x_i|y_m),$$

再利用贝叶斯定理得到：

$$P(y_1|x_i); P(y_2|x_i); \dots; P(y_k|x_i)。$$

2.朴素贝叶斯分类算法实现

逻辑代码如下：

```
DECLARE
i NUMBER;
j NUMBER;
num NUMBER;
num1 NUMBER;
BEGIN
删除数据表中空值行;
SELECT * FROM 目标表 SAMPLE(数量);
-----从数据表中随机选取两组数据，一个定义为训练表，一个为测试表。
SELECT COUNT(*) FROM 训练表 GROUP BY '目标分类变量';
-----每组目标分类变量对应的个数
求P(每种分类出现)值;
-----求出每组目标分类变量的概率
FOR i IN 1..num -----num为测试表中分类因子的个数
    IF 数据类型='离散' THEN
        FOR j IN m -----m为测试表中每种分类因子的分组数
            计算P(每种分类因子与每种分类共同出现)值;
        END
    ELSE(连续型)
        计算P(每种连续型分类因子与每种分类共同出现)值;-----应用连续变量概率公式
    END IF;
END;
P(每种分类出现)* P(每种分类因子与每种分类共同出现);-----求出先验概率，得到了一个朴素贝叶斯分类模型。
FOR i IN 1... num1 -----为测试表中的数据总数
    FOR j IN 1...s -----s为测试表中分类因子个数
        IF 数据类型='离散' THEN
            计算P(每个分类因子与每种分类共同出现)值;
        ELSE (连续型)
            计算P(每个连续型分类因子与每种分类共同出现)值;-----应用连续变量概率公式
        END IF ;
    END;
    计算每条数据的对所有分类的后验概率，并取后验概率最大的分类作为该数据的预测分类;
    将得到的预测分类与数据的原类别进行比较，如果相同输出预测正确，否则输出预测错误;
END;
统计所有测试数据预测结果，输出模型的预测正确率;
```

3.朴素贝叶斯分类算法应用

朴素贝叶斯分类法可以输出最大的后验假定。下面我们就举个简单的例子，了解一下朴素贝叶斯分类算法的计算过程。

一个关于心脏病患者的临床数据集，有六个分类变量（分类因子），包括性别、年龄、KILLP评分、饮酒、吸烟、住院天数；一个目标分类变量疾病，如表3-4所示。

表3-4 朴素贝叶斯分类法举例

行数	性别	年龄	KILLP	饮酒	吸烟	住院天数	疾病
1	男	>80	1	是	是	7-14	心梗病
2	女	70-80	2	否	是	<7	心梗病
3	女	70-80	1	否	否	<7	不稳定性心绞痛
4	女	<70	1	否	是	>14	心梗病
5	男	70-80	2	是	是	7-14	心梗病
6	女	>80	2	否	否	7-14	心梗病
7	男	70-80	1	否	否	7-14	心梗病
8	女	70-80	2	否	否	7-14	心梗病
9	女	70-80	1	否	否	<7	心梗病
10	男	<70	1	否	否	7-14	心梗病
11	女	>80	3	否	是	<7	心梗病
12	女	70-80	1	否	是	7-14	心梗病
13	女	>80	3	否	是	7-14	不稳定性心绞痛
14	男	70-80	3	是	是	>14	不稳定性心绞痛
15	女	<70	3	否	否	<7	心梗病
16	男	70-80	1	否	否	>14	心梗病
17	男	<70	1	是	否	7-14	心梗病
18	女	70-80	1	否	是	>14	心梗病
19	男	70-80	2	否	否	7-14	心梗病
20	女	<70	3	否	否	<7	不稳定性心绞痛

根据该表，我们要计算一组分类变量X（性别、年龄、KILLP评分、饮酒、吸烟、住院天数）中的一个实例（性别=‘男’，年龄<70，KILLP=‘1’，饮酒=‘是’，吸烟=‘是’，住院天数<7）最可能导致目标分

类变量Y（疾病）取哪个取值，（心脏病，还是不稳定性心绞痛）。

如表3-4所示，我们可以先计算出全部目标分类变量的概率（ C_1 表示心脏病， C_2 表示不稳定性心绞痛）：

$$P(C_1) = 16/20 = 0.8$$

$$P(C_2) = 4/20 = 0.2$$

每一个分类因子对应每种目标分类变量的概率：

$$P(\text{性别}=\text{'男'}|C_1) = 7/16 = 0.4375$$

$$P(\text{性别}=\text{'男'}|C_2) = 1/4 = 0.25$$

$$P(\text{年龄}<70|C_1) = 4/16 = 0.25$$

$$P(\text{年龄}<70|C_2) = 1/4 = 0.25$$

$$P(\text{KILLP}=\text{'1'}|C_1) = 9/16 = 0.5625$$

$$P(\text{KILLP}=\text{'1'}|C_2) = 1/4 = 0.25$$

$$P(\text{饮酒}=\text{'是'}|C_1) = 3/16 = 0.1875$$

$$P(\text{饮酒}=\text{'是'}|C_2) = 1/4 = 0.25$$

$$P(\text{吸烟}=\text{'是'}|C_1) = 7/16 = 0.4375$$

$$P(\text{吸烟}=\text{'是'}|C_2) = 2/4 = 0.5$$

$$P(\text{住院天数} < 7 | C_1) = 4/16 = 0.25$$

$$P(\text{住院天数} < 7 | C_2) = 2/4 = 0.5$$

我们可以计算出，全部该组分类变量组合下患有心脏病或不稳定性心绞痛的P值：

$$P(X|C_1) = 0.4375 \times 0.25 \times 0.5625 \times 0.1875 \times 0.4375 \times 0.25 = 0.0012$$

$$P(X|C_2) = 0.25 \times 0.25 \times 0.25 \times 0.25 \times 0.5 \times 0.5 = 0.0009$$

该组分类变量组合的每个分类与该目标分类的组合概率：

$$P(C_1) \times P(X|C_1) = 0.00096$$

$$P(C_2) \times P(X|C_2) = 0.00018$$

以上算法过程只是针对于X中的变量为离散型变量，对于连续性变量，则给出了另一种算法：将 $P(X|C_i)$ 用 $g(\chi, \upsilon, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\chi-\upsilon)^2}{2\sigma^2}}$ 替代，只要求出X中的变量均值 υ 和标准差 σ ，带入上面公式即可算出在 C_i 发生时X中的变量发生的概率。

3.5.4 决策树

决策树（Decision Tree）是数据挖掘领域应用比较广泛的一种分类模型，它主要是以离散型的属性（连续型的属性需要做离散化）作为分类因子，将数据按照目标属性进行分类。决策树模型是一个类似流程图的倒置树状结构，树中所有非叶节点均是一个某分类属性的条件判定枢纽，也就是一个决策节点，根据数据的相关属性值对每条数据进行层层判定划分，直至将该数据目标属性的分类结果输出到叶子节点。决策树模型是属于有监督的学习算法。一棵简单的决策树如下图3-22所示。

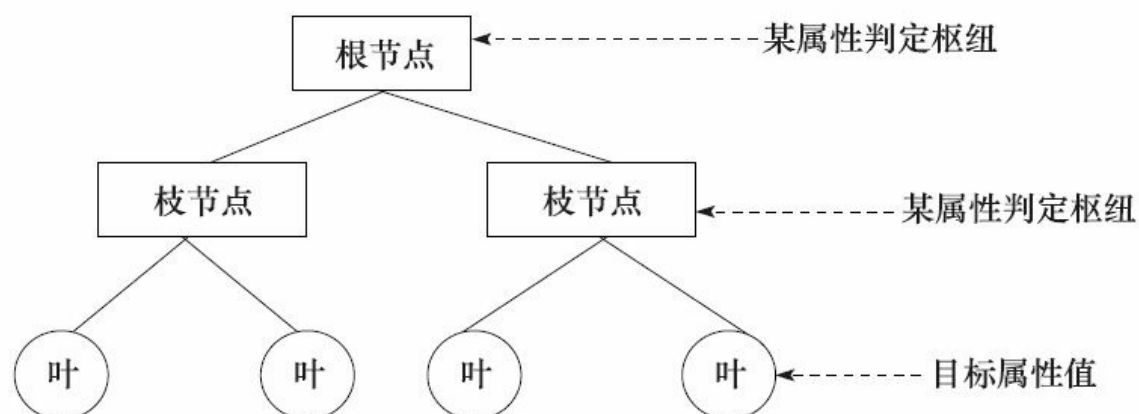


图 3-22

根节点向下生长成两个或两个以上的枝节点，每个枝节点再向下再生长成两个或两个以上的枝节点或叶节点。

3.5.4.1 决策树算法原理

决策树的算法主要包括决策树的生成和决策树的剪枝两个部分，下面简单总结算法的主要步骤。

1.决策树的生成

主要步骤如下：

- 1) 确定决策树的目标分类属性和分类因子属性集；
- 2) 在当前分类因子属性集中选择对于目标分类属性最具有分类标示能力的属性；
- 3) 根据当前决策节点属性取值的不同，将训练数据进行分组，生成几个数据子集；
- 4) 针对每个子集重复2)、3)操作。

当满足下述三个条件之一时，不再划分：

- 数据子集中所有的记录都属于一类；
- 没有可以继续分解的分类因子属性了；
- 数据子集中的所有剩余的分类因子属性取值完全相同。

2.决策树的剪枝

在创建决策树时，由于数据中存在噪声和离群点，许多决策分枝

反映的是训练数据中的异常点，这种情况称为决策树模型的过度拟合。我们可以通过剪枝来处理这种过度拟合问题。常用的剪枝方法有两种：先剪枝和后剪枝。

先剪枝是通过提前停止树的构造来对树进行剪枝，一旦停止，该节点便成为叶节点。**后剪枝**是决策树经过训练，生成全部树形模型之后，将测试结果误差大于设定误差范围的枝叶剪掉。删除节点的分支用树叶来替换，并用该节点最频繁的目标分类来标示该树叶。

决策树的特点如下：

- 决策树模型是一种构建分类模型的非参数方法；
- 决策树模型常采用贪心算法和自顶向下进行递归划分的策略；
- 决策树模型建立模型较快，分类也较快，相对容易理解；
- 决策树模型是对离散值进行分类的代表算法；
- 决策树模型对噪声的干扰具有一定的健壮性。

3.5.4.2 决策属性选择

我们从决策树模型的算法原理中不难看出，决策树分类的一个核心步骤就是在当前分类因子属性集中选择对于目标分类属性最具有分类标示能力的属性，而且这一步骤需要反复运行。只有在确立了决策属性后，才能根据当前决策节点属性取值的不同，将训练数据进行分

组，并生成若干个数据子集。因此，决策属性的选择几乎决定决策树模型的优劣。决策树很多不同的算法也是很大一部分区别于决策属性的选择方法的不同。

所谓决策属性的选择方法就是数据挖掘中的特性选择方法，只不过，通常的情况是只针对离散型的属性进行特性选择分析，连续型的属性需要首先离散化，再进行特性选择分析。在典型的决策树算法中，常见的决策属性的选择方法包括：信息增益、信息增益率、Gini指数、基于卡方的统计检验和基于G统计量等方法。我们在本章中的特性选择算法中也介绍了一些方法，包括线性相关算法、最小描述长度算法和SRCF相关因子算法。其中最小描述长度算法和SRCF相关因子算法都适用于决策树模型决策属性的选择。

典型的决策树算法包括：

- ID3算法：使用信息增益作为决策属性选择的策略，它倾向于选择具有大量值的属性。

- C4.5算法：试图克服ID3的这种决策属性选择的偏倚，它使用增益率概念，将信息增益规范化，来确定决策属性。

- C5.0算法：是C4.5算法的升级版，它以信息熵的下降速度作为确定最佳分支变量和分割阈值的依据，通常不需要很长的训练次数进行估计，具有一定的稳定性。

- CHAID算法：使用卡方检验作为树分类的基本方法，只适用

于离散型分类属性。

· QUEST算法：是对CHAID改进后的算法，将变量选择和分叉点选择分开进行，可适用于任何类型的分类属性。克服了CHAID某些缺点，在变量选择上基本无偏。

3.5.4.3 决策树应用举例（ID3算法）

以上算法在很多书籍中都有描述，大家可以参考。这里，仅以最简单的ID3算法为例，通过一个实例来演示一下决策树模型的工作过程。ID3算法是基于信息论的基本原理来构造决策树的，它在实际应用中容易实现，算法比较简单、直观，准确率较高。ID3算法是基于概率的一种方法，主要用到信息熵、条件熵和信息增益等几个信息：

信息熵的公式为：

$$H(U) = - \sum_i p(u_i) \log_2 p(u_i) \quad (3.52)$$

条件熵的公式为：

$$H(U/V) = \sum_j p(v_j) \sum_i p(u_i | v_j) \log_2 \frac{1}{p(u_i | v_j)} \quad (3.53)$$

其中， $p(v_j|u_i)$ 表示收到信息 v_j 后判定输入为 u_i 的概率。

信息增益的公式为：

$$I(U, V) = H(U) - H(U/V) \quad (3.54)$$

我们下面来看一下ID3决策树算法的一个应用。选取某医院不稳定性心绞痛和心梗病例30个，病人年龄在66-88岁之间，对年龄属性进行

离散化处理，分为66~70、71~75、76~80、81~88四个年龄阶段，分别用1、2、3、4来表示，具体数据如表3-5所示（表中数据按是年龄进行排序）。

表 3-5

序号	属性						类别
	性别	年龄	饮酒	吸烟	高血压病史	糖尿病史	
1	男	1	有	有	有	有	不稳定性心绞痛
2	男	1	无	有	无	有	不稳定性心绞痛
3	男	1	无	无	有	无	心梗
4	男	1	无	有	有	无	心梗
5	男	1	无	有	有	有	心梗
6	男	1	无	无	有	有	不稳定性心绞痛
7	男	1	无	无	有	有	不稳定性心绞痛
8	女	2	有	有	有	无	心梗
9	女	2	无	无	有	无	心梗
10	女	2	无	有	有	有	心梗
11	女	2	无	无	无	无	不稳定性心绞痛
12	女	2	无	无	无	有	不稳定性心绞痛
13	男	2	无	无	无	无	不稳定性心绞痛
14	男	3	无	有	有	无	不稳定性心绞痛
15	男	3	有	有	有	无	心梗
16	女	3	无	无	有	无	不稳定性心绞痛
17	女	3	无	无	有	无	不稳定性心绞痛
18	男	3	有	有	有	无	不稳定性心绞痛
19	男	3	无	有	无	有	不稳定性心绞痛
20	男	3	无	无	有	无	心梗
21	男	3	无	无	无	无	心梗
22	女	3	无	有	无	有	心梗
23	男	4	无	有	有	无	心梗
24	男	4	有	有	有	无	不稳定性心绞痛
25	女	4	无	无	无	无	不稳定性心绞痛
26	女	4	无	无	有	无	心梗
27	男	4	有	有	有	无	心梗
28	女	4	无	无	无	无	心梗
29	男	4	无	有	无	有	不稳定性心绞痛
30	男	4	有	有	无	无	心梗

1) 按照上表中属性顺序，首先计算“性别”属性的信息增益。

计算信息熵：

$$H(U) = - \sum_i p(u_i) \log_2 p(u_i) = - \frac{11}{30} \log_2 \frac{11}{30} - \frac{19}{30} \log_2 \frac{19}{30} = 0.9481$$

计算条件熵：

$$\begin{aligned} H(U/V) &= \sum_j p(v_j) \sum_i p(u_i | v_j) \log_2 \frac{1}{p(u_i | v_j)} \\ &= \frac{15}{30} \left(\frac{10}{15} \log_2 \frac{1}{10/15} + \frac{5}{15} \log_2 \frac{1}{5/15} \right) + \frac{15}{30} \left(\frac{9}{15} \log_2 \frac{1}{9/15} + \frac{6}{15} \log_2 \frac{1}{6/15} \right) = 0.9446 \end{aligned}$$

计算“性别”的信息增益：

$$I(\text{性别}) = H(U) - H(U/V) = 0.9481 - 0.9446 = 0.0035$$

2) 根据 $I(\text{性别})$ 计算步骤，计算出其他属性的信息增益。结果为：

$$I(\text{年龄}) = 0.0183,$$

$$I(\text{饮酒}) = 0.0045,$$

$$I(\text{吸烟}) = 0.0129,$$

$$I(\text{高血压病史}) = 0.0314,$$

$$I(\text{糖尿病史}) = 0.0589。$$

3) 比较 $I(\text{性别})$ 、 $I(\text{年龄})$ 、 $I(\text{饮酒})$ 、 $I(\text{吸烟})$ 、 $I(\text{高血压病史})$ 、和 $I(\text{糖尿病史})$ 6项的大小，选择值最大的属性 $I(\text{糖尿病史})$ 为第一层决策属性。然后根据当前决策节点年龄属性取值的不

同，将训练数据进行分组，生成几个数据子集。

4) 对于上步生成的每一个数据子集，对当前存在的所有剩余属性再进行信息增益的计算，重复上述1~3步，直到训练数据不能再进行分组。

5) 利用测试数据组对整个决策树模型进行测试，剪去正确率较低的枝叶。

ID3算法不能直接处理属性值为连续型的数据。对于连续型的属性，在分类前需要对数据进行离散化处理。对于属性值有缺失的数据，ID3算法处理不了，需要先进行缺失处理。

3.5.5 人工神经网络

人工神经网络是一个局部最优解的分类和预测算法，它得到的是一个隐式模型，我们通常称为“黑盒子”模型，其中存放着一个复杂的网状感知结构，如图3-23所示，网状的中间节点是一些动态的权重参数，“黑盒子”的输入端是一些分类因子或预测因子，“黑盒子”的输出端是一个目标分类变量或目标预测变量。

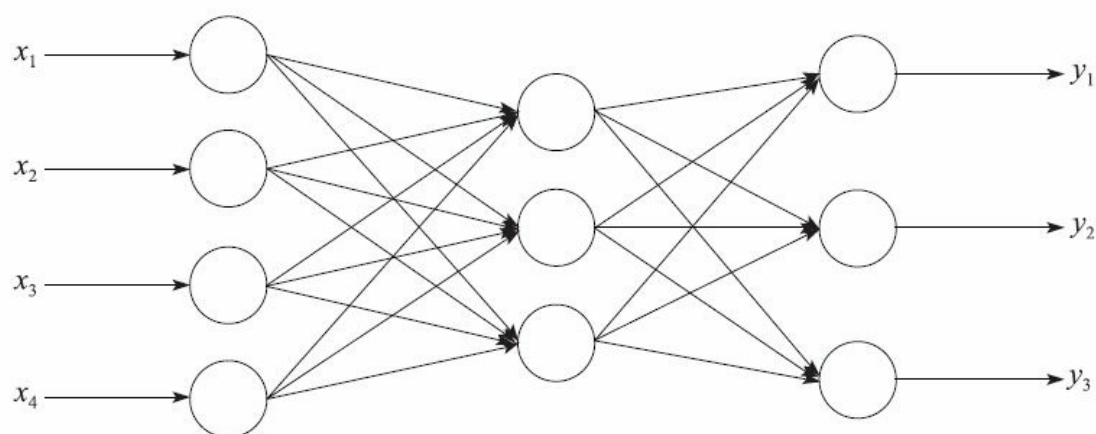


图 3-23

人工神经网络这个结构有点像生物神经系统，其实它的原理就是由生物神经系统的功能特点启发而来的，是一个人造的类生物神经系统的数学模型。

我们简单观察一下这个模拟生物神经系统的原理。如图3-24所示，生物的神经元单元由树突、细胞体和轴突构成（我们在此把问题简单化了）。生物体外界的刺激（输入信息）由树突（接收信息）传输到细胞体，然后经过细胞体的感知（处理信息）后再将处理后刺激

信号传递到轴突（输出信息），由轴突将刺激信号再传递给下一层神经元单元（接收信息、处理信息、输出信息），最后形成最终信号（输出结果）传送给大脑。

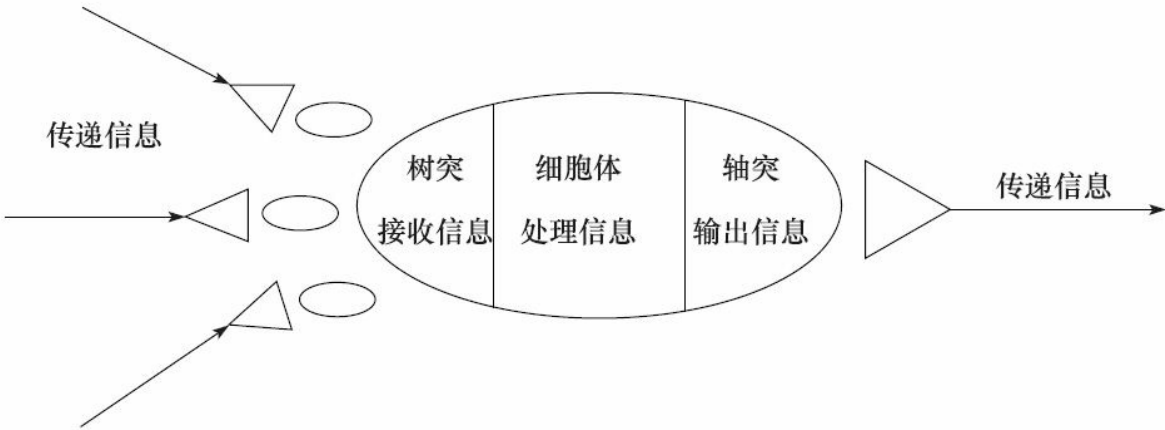


图3-24 神经元处理信息模型

3.5.5.1 信息感知器原理

我们观察这样一个数据集，它由三个输入变量和一个输出变量组成，如表3-6所示。

这里我们暂且假定三个输入变量与这个输出变量的相关度是相当的，也就是说，如果我们将这三个输入变量对输出变量的影响度进行评价的话（特性选择挖掘），我们得到三个类似重要的分类因子或预测因子。因此，这里我们给予它们（图中的三条“网线”）相同的权重值，如0.3。这样，我们构建了一个信息感知器，如图3-25所示。

表 3-6

X_1	X_2	X_3	y
2	1	1	0
2	1	2	1
2	2	1	1
2	2	1	1
1	1	2	0
1	2	1	0
1	2	2	1
1	1	1	0

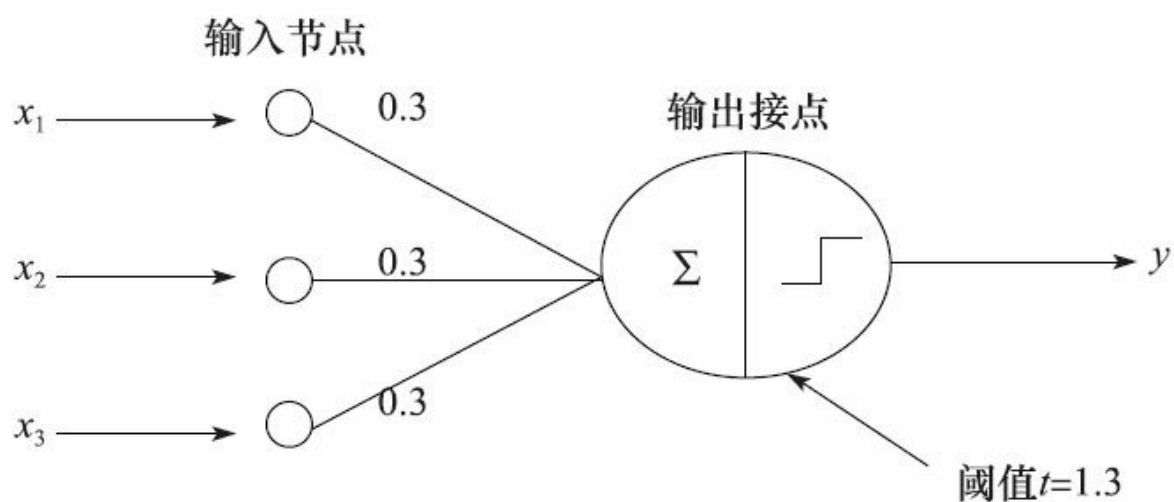


图 3-25

它将三个输入变量与一个输出变量构成一个关系函数（如式 3.55），也称为激活函数，在这里是一个符号函数，以反映输入变量与输出变量的数据规律。

$$\hat{y} = \begin{cases} 1, & \text{IF: } 0.3 \times x_1 + 0.3 \times x_2 + 0.3 \times x_3 - 1.3 > 0 \\ 0, & \text{IF: } 0.3 \times x_1 + 0.3 \times x_2 + 0.3 \times x_3 - 1.3 < 0 \end{cases} \quad (3.55)$$

这是一个点积公式，即权重向量×输入向量。

由式（3.55），我们可以将信息感知器的阈值设置为1.3，如果三个输入变量的和大于1.3，信息感知器则输出结果1，如果三个输入变量的和小于1.3，信息感知器则输出结果0。

我们可以将公式写成如下一个通用公式：

$$\hat{y} = \text{Sign}(w \cdot x) = \text{Sign}(w_d \cdot x_d + w_{d-1} \cdot x_{d-1} + \dots + w_1 \cdot x_1 + w_0 \cdot x_0 - t)$$

(3.56)

通过以上实例，我们就模拟构建了一个类生物神经元的信息感知器。类似地，我们可以构建很多个功能类似但激活函数和阈值都不同的信息感知器，并构成像图3-22那样的各种网状结构，我们把它称之为信息感知网络。这个信息感知网络是经过很多数据不断训练、不断调整“网线”权重、最后达到一个满意的分类和预测精度而得到的。接下来，我们再观察这个信息感知网络是如何工作的。

3.5.5.2 BP人工神经网络原理

为了简化说明，我们将信息感知网络简化描述成如图3-26那样一个三层结构，即一个输入层、一个中间层（也称隐含层）和一个输出层。原始数据通过输入层进入信息感知网络，经过隐含层的处理，将结果信息传输到输出层。BP（Back-Propagation，后传播）神经网络的工作原理是，首先我们人为设定了隐含层的层数和节点个数，这里

我们设置了一个隐含层和两个隐含节点（节点4和节点5）。然后再人为设定节点4、节点5和节点6的激活函数和阈值，以及所有“网线”的初始权重值，包括 w_{14} 、 w_{15} 、 w_{24} 、 w_{25} 、 w_{34} 、 w_{35} 、 w_{46} 和 w_{56} 共8个权重值。

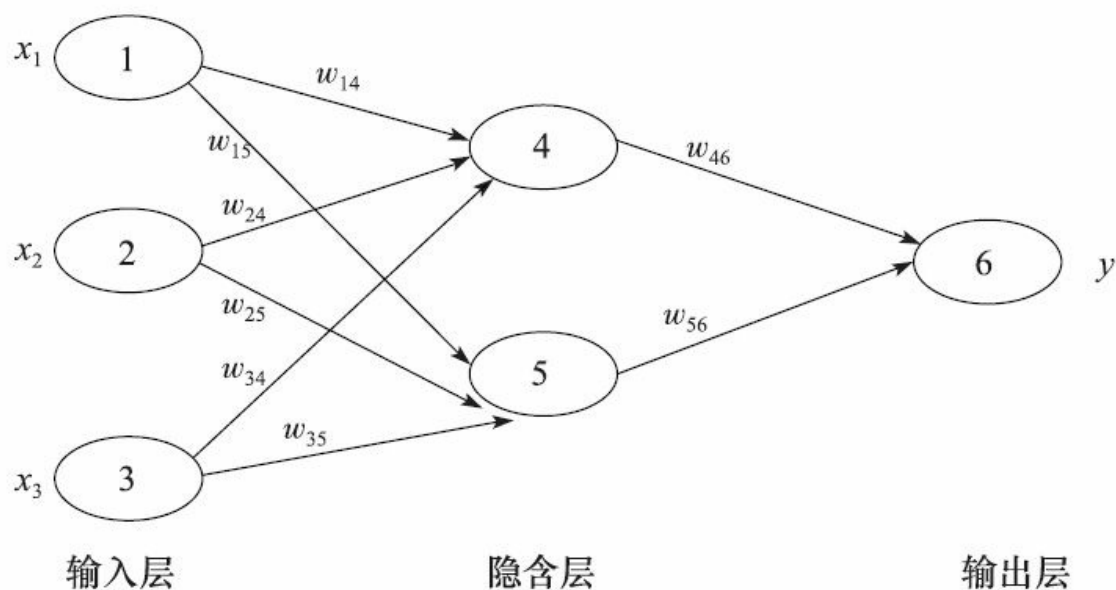


图 3-26

接下来，我们开始训练这个网络。首先输入一组原始数据（ x_1 ， x_2 ， x_3 ），经过节点4和节点5两个信息感知器的处理，得到了两个输出值，记为（ z_1 ， z_2 ）。问题是，我们不知道这两个完全人造的（ z_1 ， z_2 ）的结果是否合理，因为我们没有这样一个中间的真实数据，所以无法对它们的准确性进行评估，也就是无法评价节点4和节点5所产生的误差，也就无法调整 w_{14} 、 w_{15} 、 w_{24} 、 w_{25} 、 w_{34} 、 w_{35} 这6个权重。这时，我们只得暂时利用这两个数据（ z_1 ， z_2 ）继续向前计算（“传播”）。经过 w_{46} 和 w_{56} 以及输出层的最后一个信息感知器节点6的处理，我们得到一个最终结果数据，记为 y_1 。因为作为训

训练数据，原始的真实目标变量数据 y 是知道的，因此我们将可以评估 y_1 与 y 之间的拟合情况了。经过计算，我们可以得到节点6的误差，然后我们利用节点6的误差结果向后继续计算（“传播”）节点4和节点5所产生的误差值。最后，我们得到了所有节点的误差值后，再从输出层到隐含层逐层重新计算所有“网线”的权重值，首先调整 w_{46} 和 w_{56} 的权重，然后调整 w_{14} 、 w_{15} 、 w_{24} 、 w_{25} 、 w_{14} 、 w_{34} 、 w_{35} 的权重，方向也是向后“传播”，这就是BP神经网络。图3-27总结了上述计算的迭代过程。其中，假设该神经网络共有 m 层（ $m-1$ 个隐含层+1个输出层）， $1 \leq k \leq m$ 。

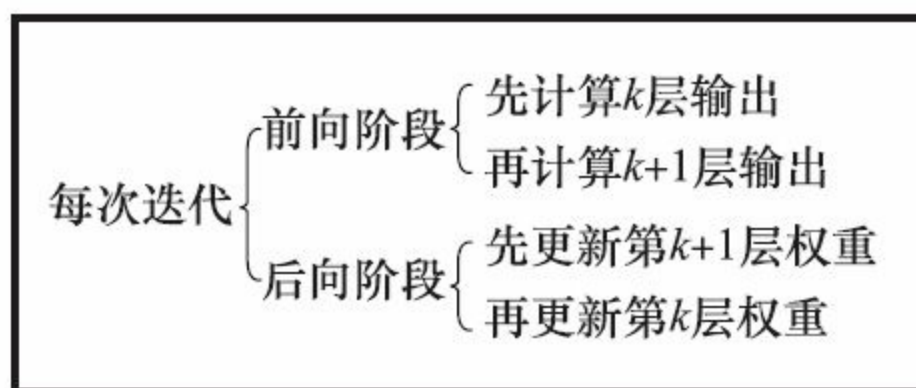


图 3-27

以上是一条原始数据的训练过程，我们再用其他原始数据不断训练这个信息感知网络，使之不断地改进，并能适应更多数据的拟合，最终得到一个比较满意的人工BP神经网络。

人工BP神经网络的计算方法中的关键是权重更新的计算，它计算公式为：

$$w_j^{(k+1)} = w_j^{(k)} + \lambda (y_i - \hat{y}_i^{(k)}) x_{ij} \quad (3.57)$$

其中, $w^{(k)}_j$ 为第 k (层) 次循环后第 j 条输入“网线”的权重; λ 为学习率; x_{ij} 是训练数据第 i 条记录 x_i 的第 j 个属性值; $y_i - y_i^{(k)}$ 是第 k 层的预测误差;

因此有, 如果 $y=1$, $\hat{y}=0$, 则 $(y - \hat{y}) = 1$, 那么需要提高正输入“网线”的权重值, 同时降低负输入“网线”的权重值;

相反, 如果有 $y=0$, $\hat{y}=1$, 则 $(y - \hat{y}) = -1$,

那么需要降低正输入“网线”的权重值, 同时提高负输入“网线”的权重值。

影响BP神经网络的因素主要有:

- 权重和阈值的初始化会对神经网络的性能产生重要的影响。
- 学习率 λ 会对神经网络产生一定影响, 其取值 $0 \sim 1$ 之间, 如果 λ 越趋近于 0 , 则新旧权重值的变化不大, 如果 λ 越趋近于 1 , 则新权重值的调整会越大。
- 隐含层的层数和节点数也会对神经网络产生影响。过多的隐含层数和节点数会导致神经网络对训练数据出现过度拟合的现象。

BP神经网络具有如下的特点:

- 运算量大, 收敛速度较慢。
- 可能会造成对训练数据过度拟合的现象。

- 其得到的权重和阈值为局部最优解，非全局最优解。
- 容易受到噪声数据的影响。
- 数据属性间的冗余性（即非独立属性）对神经网络影响不大。
- 神经网络的可解释性较差。

3.5.6 分类与聚类的关系

最后再简要谈谈有监督学习与无监督学习之间的关系。我们以聚类和分类为例。通常认为聚类是无监督学习算法，而分类是有监督学习算法。然而无论如何，聚类和分类都是遵守着一定的规则。有无监督并不是有无规则，只不过分类是显规则，在分类之前就存在着预先定义。而聚类是潜规则，在分簇之前没有预设，而是在数据中隐含存在。如果我们能将潜规则显示出来的话，则我们就可将之用于分类。如果聚类分簇良好，那么对于同样的数据，分类也很有可能良好。观察一个简单的例子，如图3-28所示，假设一组数据，进行聚类分析划分两个簇，簇A小于S域、簇B大于S域，划分良好。有新点T，需要知道属于簇A还是簇B，则根据T小于S域，认为T属于簇A。

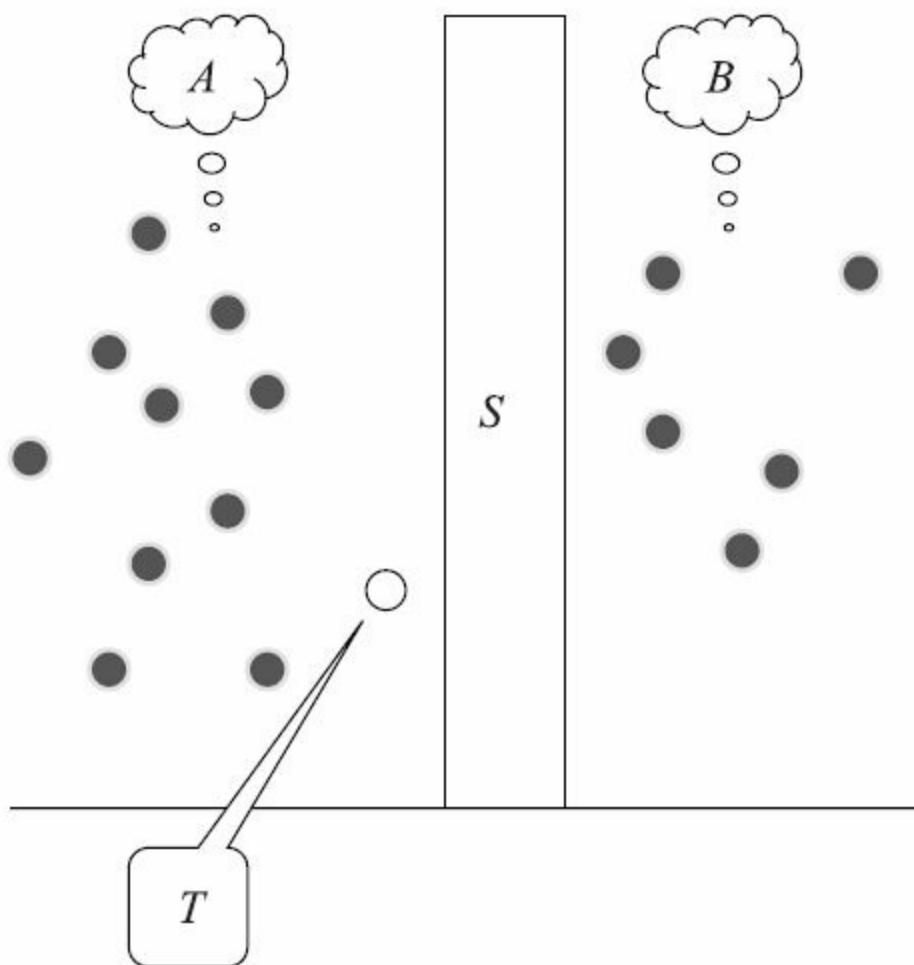


图 3-28

如果分类良好，说明我们已经得到了一个分类模型或规则，那么说这个分类模型或规则可用于对此数据进行聚类的“算法”，该“算法”得到的聚类分簇效果一定不会差。反过来，如果数据难以进行分类，那么聚类分簇也很困难，因为数据中没有隐含的分类模型或规则。就像硬币的两个面一样，聚类和分类就是规则的两个面。

3.6 时间序列

数据挖掘的一个重要方向是序列模式探索，典型的序列模式探索包括时间序列分析和非时间序列分析。其中，大量的时间序列问题广泛分布在现实生活的许多领域中，对时间序列的分析我们也称之为趋势预测探索。常见的时间序列的分析方法有如简单的移动平均模型、较复杂的平稳时间序列模型、更复杂的非平稳时间序列模型等。数据挖掘应用中，比较常用的趋势预测探索方法有自回归平均移动模型（ARIMA），灰色系统预测模型（GM）等。其中，ARIMA模型多用于时间序列数据量较多的趋势预测应用中，通常的数据量级为数十、数百、至数千；而GM模型多用于时间序列数据量较少的趋势预测应用中，通常的数据量级为8-20、20-100。本节将重点介绍灰色系统预测模型，并对ARIMA模型做一简单介绍。

3.6.1 灰色系统预测模型

我们在第1章“数据的形状”小节中曾谈到过小数据的概念，我们需要一些技术从规模较小的、有限的数据中探索其中的规律和知识，而灰色系统预测模型正是这样一个典型的专门处理小数据知识挖掘的方法。

3.6.1.1 灰色系统理论

灰色系统是指已知部分信息的样本数据能反映的不确定性系统。不完全的信息包括：

- 系统因素不完全明确；
- 因素关系不完全明确；
- 系统结构不完全明确；
- 系统作用原理不完全明确；

相应地，我们将信息完全的（包括内部特征）的系统称为白色系统，将信息完全未知（只有该系统与外界的联系信息）的系统称为黑色系统。如图3-29所示。

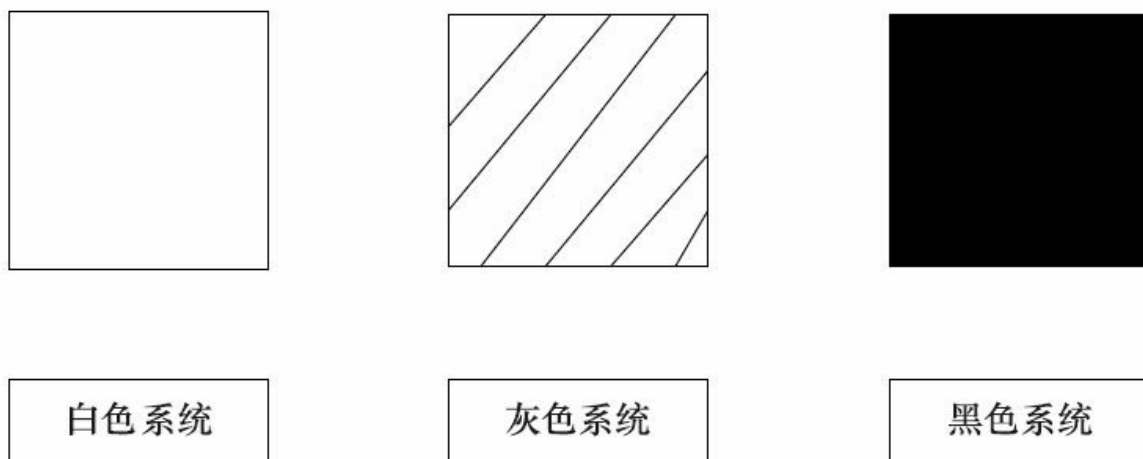


图 3-29

3.6.1.2 灰色系统预测

灰色系统预测是通过分辨系统因素之间的发展趋势的相似度或相异度（关联分析），经过对原始数据的生成处理来探索系统变化的规律，它通过生成数据序列较强的规律性来建立相应的微分方程模型，从而预测事物的未来发展趋势。灰色系统预测可用于以下几方面：

- 时间序列预测：利用等时距观测到的反映预测目标特征的一系列数值构造灰色系统模型，预测未来某时刻的特征量，或达到某特征量的时间。
- 灾变预测：通过模型预测异常值出现的时刻，预测异常值何时出现在特定时区内。数据挖掘中我们将这种用法称为异常侦测。
- 拓扑预测：通过灰色模型预测事物未来变化的趋势和轨迹。
- 系统预测：对系统行为指标建立一组相互关联的灰色系统预测

模型，预测系统各个环节的变化。

1.生成数的概念

生成数分为累加生成数（AGO）和累减生成数（IAGO）。

累加生成数（AGO）： 是一次累加生成的数列：

有一原始数列：

$$X^{(0)} = \{x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)}\} \quad (3.58)$$

则一次累加生成序列为：

$$X^{(1)} = \{x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}\} \quad (3.59)$$

其中，

$$x_k^{(1)} = \sum_{i=1}^k x_i^{(0)} = x_{k-1}^{(1)} + x_k^{(0)} \quad (3.60)$$

累减生成数（IAGO）： 是一次累减生成的数列：

有一原始数列：

$$X^{(1)} = \{x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}\} \quad (3.61)$$

则一次累减生成序列为：

$$X^{(0)} = \{x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)}\} \quad (3.62)$$

其中，

$$x_k^{(0)} = x_k^{(1)} - x_{k-1}^{(1)} \quad (3.63)$$

累加生成与累减生成互为逆过程。

均值生成数 (MEAC)

设原始数列：

$$X^{(1)} = \{x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}\} \quad (3.64)$$

α 为常数： $\alpha \in [0, 1]$ ，

则

$$Z_k^{(0)} = \alpha x_k^{(0)} + (1 - \alpha) x_{k-1}^{(0)} \quad (3.65)$$

数列 $X^{(0)}$ 在生成系数 α 下的邻值生成数，特别的，当 $\alpha = 0.5$ 时，则称

$$Z_k^{(0)} = 0.5x_k^{(0)} + 0.5x_{k-1}^{(0)} \quad (3.66)$$

为邻均值生成数。

2. 关联度分析

与数据挖掘中的变量相关分析和影响度分析（特性选择模型）相类似，灰色系统预测中的关联度分析也是探索事物动态关联的特征与程度，它是根据因素（变量）之间发展趋势的相似度和相异度来评价因素（变量）间的关联程度，通常这些因素（变量）是时间的一个函数，这是它与其他特性选择模型所分析的非时间性变量相关性的区别。

有4个时间性的变量A、B、C、D，它们随时间变化的趋势图如图3-30所示。我将A和B、C、D的关联度记为 r_{AB} 、 r_{AC} 、 r_{AD} 。显然，A和B的运行轨迹更加相似，A和C的运行轨迹部分相似，而A与D的运行模式则完全相反。因此我们得到了一组关联度大小不同的相关评价数据，按其数值大小排序得到 $\{r_{AB}, r_{AC}, r_{AD}\}$ ，我们称之为关联序

列。关联度分析就是一种曲线间几何形状的分析比较，几何形状越接近，则发展变化的趋势就越接近，关联程度越大，反之也是一样。

关联度分析是探索多时间变量间的相互关系，它的计算方法如下：

(1) 关联系数计算

设原始数列：

$$X^{(0)} = \{x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)}\} \quad (3.67)$$

比较数列为：

$$X^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\} \quad (3.68)$$

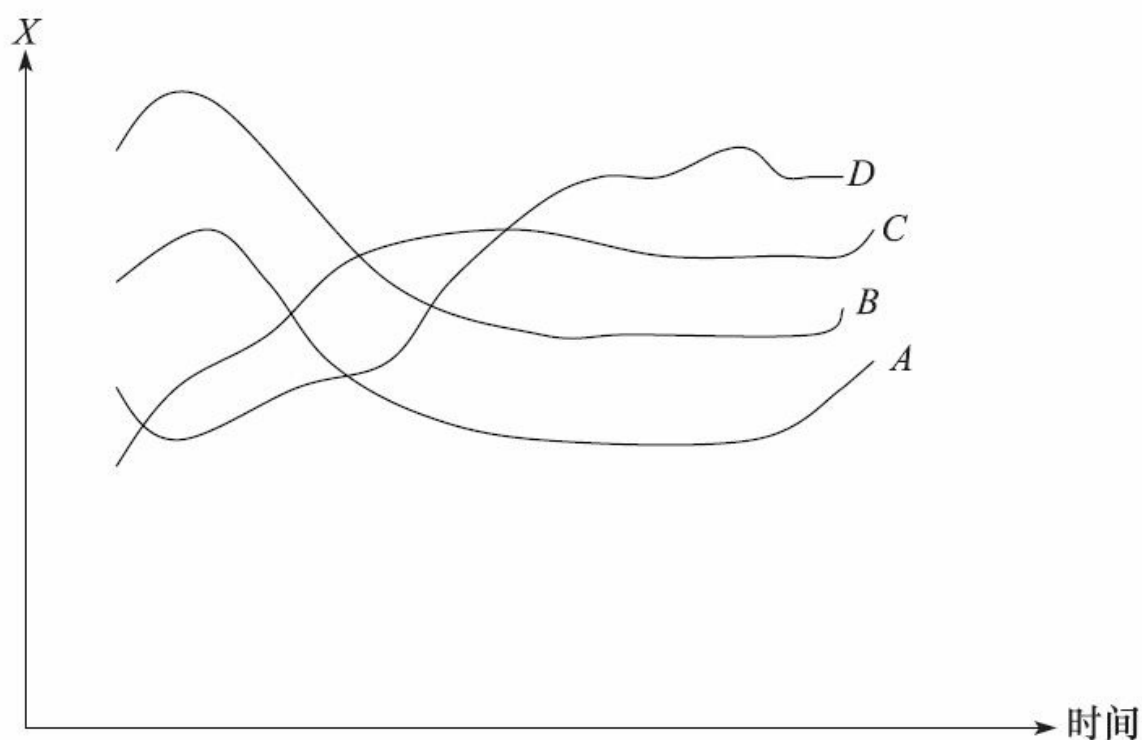


图 3-30

并联系数为：

$$\eta_k^{(i)} = \frac{\min_j \min_l |x_l^{(0)} - x_l^{(j)}| + \rho \max_j \max_l |x_l^{(0)} - x_l^{(j)}|}{|x_k^{(0)} - x_k^{(i)}| + \rho \max_j \max_l |x_l^{(0)} - x_l^{(j)}|} \quad (3.69)$$

其中：

$|x_k^{(0)} - x_k^{(i)}|$ 为第k点 $x^{(0)}$ 和 $x^{(i)}$ 的绝对误差； $\min_j \min_l |x_l^{(0)} - x_l^{(j)}|$ 为两级最小差； $\min_l |x_l^{(0)} - x_l^{(i)}|$ 是第一级最小差，即在 $x^{(i)}$ 序列中找出各差与 $x^{(0)}$ 的最小差； $\min_j \min_l |x_l^{(0)} - x_l^{(j)}|$ 为第二级最小差，即在各序列中找出最小差，再寻找所有序列的最小差； $\min_j \max_l |x_l^{(0)} - x_l^{(j)}|$ 为两级最大差，含义与最小差相似； ρ 为分辨系数， $\rho \in [0, 1]$ ，一般采用 $\rho=0.5$ 。

对于量纲不一致的情况，即初值单位不同的序列，应先进行数值的初值化，即将序列的所有数据分别除以第一个数据（量纲），然后再计算关联系数。

（2）关联度的计算

上面计算得到的关联系数值表示了各时刻参考序列与比较序列间的关联程度。为了从总体考察序列间的关联程度，需计算它们的时间平均值，即关联度，计算公式为：

$$\gamma^i = \frac{1}{n} \sum_{k=1}^n \eta_k^{(i)} \quad (3.70)$$

3.6.1.3 灰色系统预测模型原理

灰色系统预测模型是一种时间数列预测模型，它有很多不同类型的模型。灰色系统理论的微分方程称为GM模型（Gray Model），下

面我们简要介绍一下其中最基本一种数列预测模型GM（1，1）。

数列预测模型GM（1，1）是最简单的、较常用的一种灰色系统预测模型，它是一个单变量的一阶灰微分方程模型，而GM（1，N）是多变量的一阶灰微分方程模型。GM（1，1）模型适用于与具有较强指数规律的序列，通常描述单调的变化过程。

1.GM（1，1）模型的建立方法

有一原始非负数列 $X^{(0)}$ ：

$$X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), \dots, x^{(0)}(n)\} \quad (3.71)$$

其中， $x^{(0)}(k) \geq 0, k=1, 2, \dots, n$

首先，为了确保建模方法的可行性，要对该数列进行必要的检验处理，计算数列的等级比：

$$\lambda(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, k = 2, 3, \dots, n \quad (3.72)$$

如果所有的级比 $\lambda(k)$ 都落在可容覆盖 $(e^{-\frac{2}{n+1}}, e^{\frac{2}{n+2}})$ 内，则数列 $X^{(0)}$ 可以作为模型GM（1，1）的数据进行灰色预测。否则，需要对数列 $X^{(0)}$ 做必要的变换处理，使其落入可容覆盖内，即取适当的常数 c ，作平移变换：

$$y^{(0)}(k) = x^{(0)}(k) + c, k=1, 2, \dots, n, \quad (3.73)$$

则使数列 $y^{(0)}(k) = (y^{(0)}(1), y^{(0)}(2), \dots, y^{(0)}(n))$ 的等级比:

$$\lambda(k) = \frac{y^{(0)}(k-1)}{y^{(0)}(k)} \in X, k = 2, 3, \dots, n \quad (3.74)$$

然后, 设该原始非负数列的生成数列为 $X^{(1)}$:

$$X^{(1)} = \{x^{(1)}(1), x^{(1)}(2), x^{(1)}(3), \dots, x^{(1)}(n)\} \quad (3.75)$$

其中, $x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i), k = 1, 2, \dots, n$

该生成数列 $X^{(1)}$ 的紧邻均值生成数列为 $Z^{(1)}$:

$$Z^{(1)} = \{z^{(1)}(1), z^{(1)}(2), \dots, z^{(1)}(n)\} \quad (3.77)$$

其中, $Z^{(1)}(k) = 0.5x^{(1)}(k) + 0.5x^{(1)}(k-1), k=1, 2, \dots, n$ (3.78)

我们将 $x^{(0)}(k) + az^{(1)}(k) = b$ (3.79)

称为GM(1, 1)模型, 其中a, b是模型求解的参数。

若 $a = (a, b)^T$ 为参数列, 且

$$Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}, \quad B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ -z^{(1)}(4) & 1 \\ -z^{(1)}(5) & 1 \end{bmatrix} \quad (3.80)$$

则求微分方程 $x^{(0)}(k) + az^{(1)}(k) = b$ 的最小二乘估计系数列,

$$\text{满足 } \hat{a} = (B^T B)^{-1} B^T Y \text{ 时, 称 } \frac{dx^{(1)}}{dt} + ax^{(1)} = b \quad (3.81)$$

为灰微分方程 $x^{(0)}(k) + az^{(1)}(k) = b$ 的白化方程 (影子方程),

其中, 白化方程 $\frac{dx^{(1)}}{dt} + ax^{(1)} = b$ 的解是一个时间响应函数, 记为:

其中, 白化方程 $\frac{dx^{(1)}}{dt} + ax^{(1)} = b$ 的解是一个时间响应函数, 记为:

$$\hat{x}^{(1)}(t) = \left(x^{(1)}(0) - \frac{b}{a} \right) e^{-at} + \frac{b}{a} \quad (3.82)$$

GM(1, 1) 灰微分方程 $x^{(0)}(k) + az^{(1)}(k) = b$ 的时间响应数列为:

$$\hat{x}^{(1)}(k+1) = \left(x^{(1)}(0) - \frac{b}{a} \right) e^{-ak} + \frac{b}{a}, k = 1, 2, \dots, n \quad (3.83)$$

我们取 $x^{(1)}(0) = x^{(0)}(1)$, 则

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a} \right) e^{-ak} + \frac{b}{a}, k = 1, 2, \dots, n \quad (3.84)$$

还原到原始数据为:

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k), k = 1, 2, \dots, n \quad (3.85)$$

2. GM(1, 1) 模型的检验

建立GM(1, 1)模型完成后, 需要对该GM(1, 1)模型进行检验。模型检验包括分为三个方面: 残差检验、关联度检验和后验差检验。

残差检验。即残差大小检验, 是对模型值和实际值的残差进行逐点检验。

设一原始数列：

$$X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)\}$$

其相应的模拟数列为：

$$\hat{X}^{(0)} = \{\hat{x}^{(0)}(1), \hat{x}^{(0)}(2), \dots, \hat{x}^{(0)}(n)\} \quad (3.86)$$

则绝对残差数列为：

$$\begin{aligned} \varepsilon^{(0)} &= \{\varepsilon(1), \varepsilon(2), \dots, \varepsilon(n)\} \\ &= \{x^{(0)}(1) - \hat{x}^{(0)}(1), x^{(0)}(2) - \hat{x}^{(0)}(2), \dots, x^{(0)}(n) - \hat{x}^{(0)}(n)\} \end{aligned} \quad (3.87)$$

其相对误差数列为：

$$\begin{aligned} \Delta &= \left\{ \left| \frac{\varepsilon(1)}{x^{(0)}(1)} \right|, \left| \frac{\varepsilon(2)}{x^{(0)}(2)} \right|, \dots, \left| \frac{\varepsilon(n)}{x^{(0)}(n)} \right| \right\} \\ &= \{\Delta_k\}_1^n \end{aligned} \quad (3.88)$$

$$\text{其中, 对于 } k < n, \text{ 称 } \Delta_k = \left| \frac{\varepsilon(k)}{x^{(0)}(k)} \right| \quad (3.89)$$

$$\text{为 } k \text{ 点模拟相对误差, } \Delta_n = \left| \frac{\varepsilon(n)}{x^{(0)}(n)} \right| \quad (3.90)$$

$$\text{为滤波相对误差, } \bar{\Delta} = \frac{1}{n} \sum_{k=1}^n \Delta_k \quad (3.91)$$

为平均模拟相对误差。进一步地，我们将 $1 - \bar{\Delta}$ 称为平均相对精度， $1 - \Delta_n$ 为滤波精度。对于一个给定的 α ，当 $\bar{\Delta} < \alpha$ ，且 $\Delta_n < \alpha$ 成立时，称模型为残差合格模型。

关联度检验。即通过考察模型值曲线和建模序列曲线的相似程度进行检验。

设 $X^{(0)}$ 为原始数列， $\hat{X}^{(0)}$ 为相应的模拟数列，计算两个数列的关联系数，然后计算其关联度，方法见3.6.1.2的“关联度分析”。一般来说，当 $\rho=0.5$ 时，关联度大于0.6是比较满意的模型。

后验差检验。设 $X^{(0)}$ 为原始数列， $\hat{X}^{(0)}$ 为相应的模拟误差数列，

$\varepsilon^{(0)}$ 为残差数列，则，

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x^{(0)}(k) \tag{3.92}$$

为 $X^{(0)}$ 的均值, $s_1^2 = \frac{1}{n} \sum_{k=1}^n (x^{(0)}(k) - \bar{x})^2$ (3.93)

为 $X^{(0)}$ 的方差, $\bar{\varepsilon} = \frac{1}{n} \sum_{k=1}^n \varepsilon(k)$ (3.94)

为残差均值, $s_2^2 = \frac{1}{n} \sum_{k=1}^n (\varepsilon(k) - \bar{\varepsilon})^2$ (3.95)

为残差方差，我们将 $c = \frac{s_2}{s_1}$ (3.96)

称为均方差比值。对于给定的 $c_0>0$ ，当 $c<c_0$ 时，称模型为均方差比合格模型。另外，将

$$p=P\{|\varepsilon(k) - \bar{\varepsilon}|<0.6745S_1\} \tag{3.97}$$

称为小误差概率。指标 c 和 p 是后验差检验的两个重要指标，可以根据这两个指标对模型的精度进行评级，精度由后验差和小误差概率共同说明，可参照表3-7。

表 3-7

精度等级	均方差比值 c	小误差概率 p
1 级（好）	$c \leq 0.35$	$0.95 \leq p$
2 级（合格）	$0.35 < c \leq 0.5$	$0.8 \leq p < 0.95$
3 级（勉强）	$0.5 < c \leq 0.65$	$0.7 \leq p < 0.8$
4 级（不合格）	$0.65 < c$	$p < 0.7$

3.6.2 ARIMA模型预测

ARIMA模型全称是自回归移动平均模型（Autoregressive Integrated Moving Average Model），它是处理带有趋势、季节因素平稳随机项数据的模型。ARIMA的一般模型为ARIMA（ p, d, q ），称为差分自回归移动平均模型，AR是自回归，P为自回归项，MA为移动平均，q为移动平均项数，d为时间序列成为平稳时所做的差分次数。

3.6.2.1 ARIMA模型的核心思想

ARIMA模型的核心思想是将预测对象随时间推移而形成的数据序列视为一个随机序列，以时间序列的自相关分析为基础，用一定的数学模型来近似描述这个序列。这个模型一旦被识别后，就可以从时间序列的过去值及现在值来预测未来值。ARIMA模型考虑了数据在时间序列上的依存性，又考虑了数据随机波动的干扰性，对于短期趋势的预测准确率较高。ARIMA模型主要分为以下三种：

- 自回归模型（简称AR模型）；
- 滑动平均模型（简称MA模型）；
- 求和自回归滑动平均混合模型（简称ARIMA模型）。

3.6.2.2 ARIMA模型预测的流程

1) 根据时间序列的散点图、自相关函数和偏自相关函数图，以ADF单位根检验其方差、趋势及其季节性变化规律，对序列的平稳性进行识别。

2) 对非平稳序列进行平稳化处理。如果数据序列是非平稳的，并存在一定的增长或下降趋势，则需要对数据进行差分处理。如果数据存在异方差，则需对数据进行技术处理，直到处理后的数据的自相关函数值和偏相关函数值与零无显著差异。

3) 根据时间序列模型的识别规则，建立相应的模型。若平稳序列的偏相关函数是截尾的，而自相关函数是拖尾的，可判定序列适合AR模型；若平稳序列的偏相关函数是拖尾的，而自相关函数是截尾的，则可判定序列适合MA模型；若平稳序列的偏相关函数和自相关函数均是拖尾的，则序列适合ARMA模型。

4) 进行参数估计，检验是否具有统计意义。

5) 进行假设检验，诊断残差序列是否为白噪声。

6) 利用已通过检验的模型进行预测分析。

ARIMA模型对非平稳序列需要平稳化处理，这是一个数据变换的过程。我们知道，任何数据变换都会造成不同程度的信息损失，基于数据变换的模型是一种近似求解方案，必然会不同程度地影响模型的准确度和精度。ARIMA模型也是这样，非平稳序列进行平稳化处理时，需要对数据进行 n 次（1~3次较多）差分处理及还原处理，每进

行一次差分处理会损失一部分信息。如果能够找到一种直接挖掘非平稳时间序列的方法，将会进一步提高模型的预测准确度。

3.7 本章小结

本章比较简明扼要地介绍了数据挖掘应用中常用的一些算法。与很多专门论述数据挖掘算法的书籍相比较，本章节可谓短小精悍，其主要目的一是试图能较全面地介绍一下数据挖掘商业项目中经常用到的算法，二是使读者能对数据挖掘常用算法有个快速入门，并把算法融入到整个系统性工程的数据挖掘实践中去，把更多的精力均匀分配到数据挖掘应用的各个重要环节中，而不至于总是纠结于算法。另外一个感受是，虽然目前很多人投入到数据挖掘算法的研究与开发工作中，也的确在产生了一些实验证明是有效的新算法，但面对需求越来越多、越来越高的各类商业项目以及越来越复杂、越来越庞大的大数据，真正能够在应用中胜任的算法其实少得可怜！这也正是我们今后不断努力的方向！

第4章 数据挖掘应用案例

上一章我们简要介绍了数据挖掘应用中常用的一些算法，本章我们将介绍应用上一章中提到的一些算法开展商业应用的案例。这些案例不仅体现了算法的实践应用，也展现了数据挖掘应用各个环节的一些工作内容。本章将主要介绍特性选择模型的应用、分类模型的应用和算法的综合应用，应用中包括临床科研、网络异常侦测、肿瘤标志物的研究等方面的案例，并简要介绍了一些数据挖掘在银行和电信领域中的应用。

4.1 特性选择的应用

本节选取了一家综合性三甲医院心脏病患者的临床数据，对该数据的有关变量指标应用特性选择算法，将相关系数排名靠前的因子利用不同的分类算法进行分类，对得出的结果进行了分析。本节将介绍该案例的数据整合、数据描绘、数据标准化、特性选择探索等各阶段内容，这些工作实际上是该案例数据挖掘流程的一个体现。

4.1.1 数据整合

首先我们进行数据整合。原始的数据是从该院提供的Excel文件中导入到Oracle数据库中的，数据的主要信息包括病人的基本信息、既往史、既往病史、检验信息、用药信息和其他的一些信息，共有917例数据，其中心梗病例作为观察组，心绞痛病例作为对照组。临床数据涵盖如下几个方面：

- 病人的基本信息，主要有病例号、性别、年龄、入院时间、主诉等信息。
- 既往史信息，主要有吸烟、饮酒等信息。
- 既往病史信息，主要有高血压、心梗病史和糖尿病史等信息。
- 检查信息，主要有心电图、冠脉造影等信息。
- 用药信息，主要有抗生素、降压用药、调脂用药、抗凝用药、强心药、利尿剂等信息。
- 其他信息，主要有住院天数、转归^[1]和死亡时间等信息。

导入数据的工具使用的是Oracle SQL Developer，导入的时候增加了字段Case_id作为数据记录的唯一标识键。

[1] 转归是指病情的转移和发展的意思，比如病情的恶化或好转，以及扩散或减轻。

4.1.2 数据描绘

首先计算数据字段中所有连续型变量的平均值、最大值、最小值、标准差和中位数，然后统计出连续型变量字段的空值记录数，如表4-1所示。

表4-1 连续型变量字段分析

字段	类型	AVG	MAX	MIN	标准差	中位数	NULL
AGE(年龄)	NUMBER(20, 10)	75. 501	97	63	6. 556	75	0
KILLP(评分)	NUMBER(20, 10)	1. 903	4	1	0. 959	2	0
IN_DAYS(住院天数)	NUMBER(20, 10)	9. 251	83	0	6. 715	8	0

下一步，计算数据字段中所有离散型变量的不同取值个数（Distinct Values Count），然后统计出所有离散型变量字段的空值记录数及空值百分比，如表4-2所示。

表4-2 离散型变量字段分析

字段	类型	不同取值个数	空值个数	空值百分比
SEX(性别)	VARCHAR2(1000 BYTE)	2	0	0
DRINK(饮酒)	VARCHAR2(1000 BYTE)	2	0	0
SMOKE(吸烟)	VARCHAR2(1000 BYTE)	2	0	0
HIGH_PRESURE(高血压)	VARCHAR2(1000 BYTE)	2	0	0
MCD_INFC(心梗)	VARCHAR2(1000 BYTE)	2	0	0
CORONARY(冠心病)	VARCHAR2(1000 BYTE)	2	0	0
DIABETES(糖尿病)	VARCHAR2(1000 BYTE)	2	0	0
ST_UP(ST 抬高)	VARCHAR2(1000 BYTE)	2	0	0
STATUS(预后)	VARCHAR2(1000 BYTE)	2	0	0

接下来，要对数据库表中所有字段的空值进行处理。如果某个字段缺失值的百分比不大，则对于连续变量中的空值数据将赋予本字段的平均数，对于离散变量中的空值数据将赋予一个不同于本字段所有

取值的特殊值，如‘未知’、‘Unknown’等。如果某个字段缺失值的百分比很大，一般情况下该字段对于数据分析工作来说没有太大的意义，我们可以选择在分析时排除该字段。但在某些特定情况下，这样的字段本身却提供了很多信息，比如该字段代表的是一个非常规操作或非常规现象。例如，住院病人“血常规”是个常规检查，代表它的字段值缺失很少，而“肥达式反应”不是一个常规检查，相关字段值缺失就多，我们通过这些可以分析哪些是常规项目，哪些是非常规项目。

4.1.3 数据标准化

将原来Excel表中的中文字段在Oracle数据库表中统一用英文来表示。然后按照专业含义对数据进行整理，分别对观察组数据和对照组数据进行标签分类，1表示观察组，0表示对照组。

4.1.4 特性选择探索

我们将应用第3章中介绍的MDL算法和SRCF算法中的相对熵Entropy和特征准则计分Rfsc的方法开展特性选择的探索。首先应用SRCF算法计算目标属性中性别SEX的相关因子，结果如表4-3-A和表4-4所示，它们给出的结果是使用SRCF算法计算出相对于目标属性SEX的各个相关因子的Entropy和Rfsc值。表4-3-A中的相关因子是按照Entropy降序排列的，表4-4中的相关因子是按照因子的Rfsc值降序排列的。

这里的属性集中，既有AGE、KILLP、IN_DAYS这样的连续型数据（其中，KILLP是数值为1-4的等级数据，也可看做是有序离散数据），也有像SEX、DRINK、SMOKE等这样的离散型数据。对于这样的混合型的属性数据，我们需要运用统一的一种属性相关评价方法，以便使混合型的属性数据具有可比性。有两种策略，一种是将连续型数据离散化，另一种是离散数据连续化。这里，我们采用了后一种策略，将离散数据连续化就是将离散型的符号数据（如0和1）转化成数值计算，同时对不同的连续型数据进行了去量纲化处理，然后应用连续型相对熵算法对所有属性因子进行评分。同时，为了验证这一方法的有效性，我们还应用了离散型相对熵算法对所有离散型的属性因子进行评分，并与连续型相对熵算法对所有离散型属性因子的评分进行比较。结果发现，连续型相对熵算法与离散型相对熵算法对不同的离散型属性因子的评价是基本一致的。参见表4-3-A和表4-3-B，以

及表4-5-A和表4-5-B。当然，我们还可将连续型数据离散化后，应用离散型相对熵算法对原连续型的属性因子进行评分。

表4-3-A 用SRCF计算SEX的Entropy和Rfsc（连续型相对熵算法）
（按Entropy排序）

列名	Entropy	Rfsc	列名	Entropy	Rfsc
IN_DAYS	32.72848	0.949859	STATUS	1.287258	-0.16023
AGE	32.23134	0.455337	DIABETES	0.347341	0.144639
DRINK	1.784574	0.318073	SMOKE	0.086668	0.073434
KILLP	1.326928	0.170407	ST_UP	0.039373	-0.04758
MCD_INFC	1.321028	0.275301	HIGH_PRESURE	0.026271	-0.03919

表4-3-B 用SRCF计算SEX的Entropy（离散型相对熵算法）

列名	相对熵
DRINK	0.562
MCD_INFC	0.460
STATUS	0.374
DIABETES	0.158
SMOKE	0.043
ST_UP	0.018
HIGH_PRESURE	0.012
CORONARY	0.003

表4-4 用SRCF计算SEX的Entropy和Rfsc（按Rfsc排序）

列名	Entropy	Rfsc
IN_DAYS	32.728 48	0.949 859
AGE	32.231 34	0.455 337
DRINK	1.784 574	0.318 073
MCD_INFC	1.321 028	0.275 301
KILLP	1.326 928	0.170 407
DIABETES	0.347 341	0.144 639
SMOKE	0.086 668	0.073 434
CORONARY	0.006 565	0.020 21
HIGH_PRESURE	0.026 271	-0.039 19
ST_UP	0.039 373	-0.047 58
STATUS	1.287 258	-0.160 23

从两表中可以看出，住院时间IN_DAYS、年龄AGE两个因子的Entropy和Rfsc的值都是最大的，这说明这两个因子与性别SEX目标属性相关性较大，同时也是因为IN_DAYS和AGE是连续型的属性因子，它们与其他离散型的属性因子的Entropy和Rfsc的值有较大的距离，因此，最好将连续型的属性因子与离散型的属性因子分别进行比较。而后面离散型的因子中的饮酒DRINK，评分KILLP，和心梗MCD_INFC相对值较高，说明不同的性别组在这三个方面可能有较明显的差别。

Entropy与Rfsc值的关系图如图4-1所示。我们看到对于Entropy和Rfsc值来说，它们的数值大小所反映的相关系数的方向是一致的。以Rfsc值分布来说，大致可分为三段。第一段 $Rfsc < 0$ ，此时Rfsc与Entropy大致呈现线性下降的趋势；第二段 $Rfsc > 0$ 且 $Rfsc < 0.4$ （离散型因子），此时Rfsc与Entropy大致呈现缓慢线性增长的趋势；第三段 $Rfsc > 0.4$ （连续型因子），此时Rfsc与Entropy大致呈现更缓慢的线性

增长的趋势。其实，第二段与第三段的关系特征是相似的，只是由于两段中不同的数据类型（离散和连续）在量级上体现的差别。

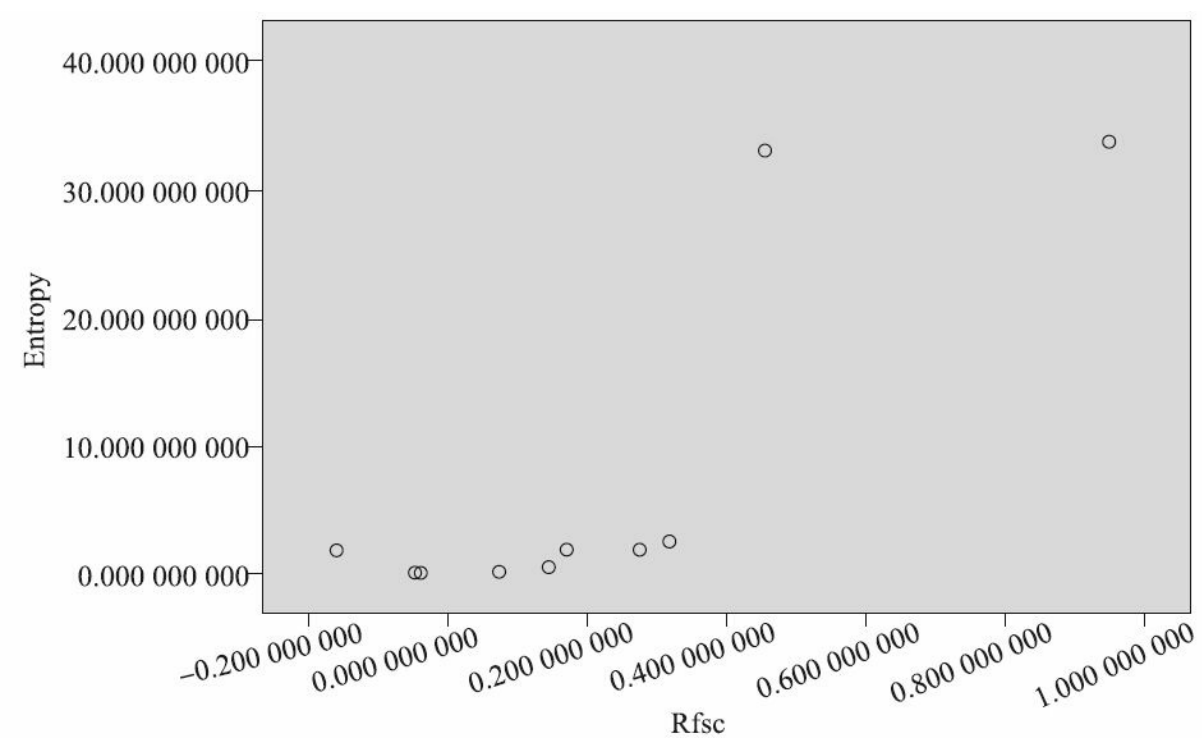


图 4-1

相似地，我们应用SRCF算法计算预后^[1]STATUS字段的相关因子的Entropy和Rfsc值，如表4-5-A所示。

表4-5-A 用SRCF计算STATUS的Entropy和Rfsc（连续型相对熵算法）
（按Entropy排序）

列名	Entropy	Rfsc
IN_DAYS	55.800 44	1.347 158
AGE	11.2501	0.589 335
DRINK	6.108 681	0.617 894
MCD_INFC	5.247 643	0.572 087
DIABETES	3.083 999	0.431 768
KILLP	2.497 269	0.390 784
SMOKE	2.122 455	0.351 175
CORONARY	1.500 87	0.287 504
SEX	1.287 258	0.262 394
HIGH_PRESURE	0.906 889	0.211 835

表4-5-B 用SRCF计算STATUS的Entropy（离散型相对熵算法）

列名	相对熵
DRINK	1.616
MCD_INFC	1.431
DIABETES	0.828
SMOKE	0.522
CORONARY	0.336
SEX	0.278
HIGH_PRESURE	0.180
ST_UP	0.162

我们再看一下由表4-5-A绘制的Entropy与Rfsc值之间的关系图，

如图4-2所示。

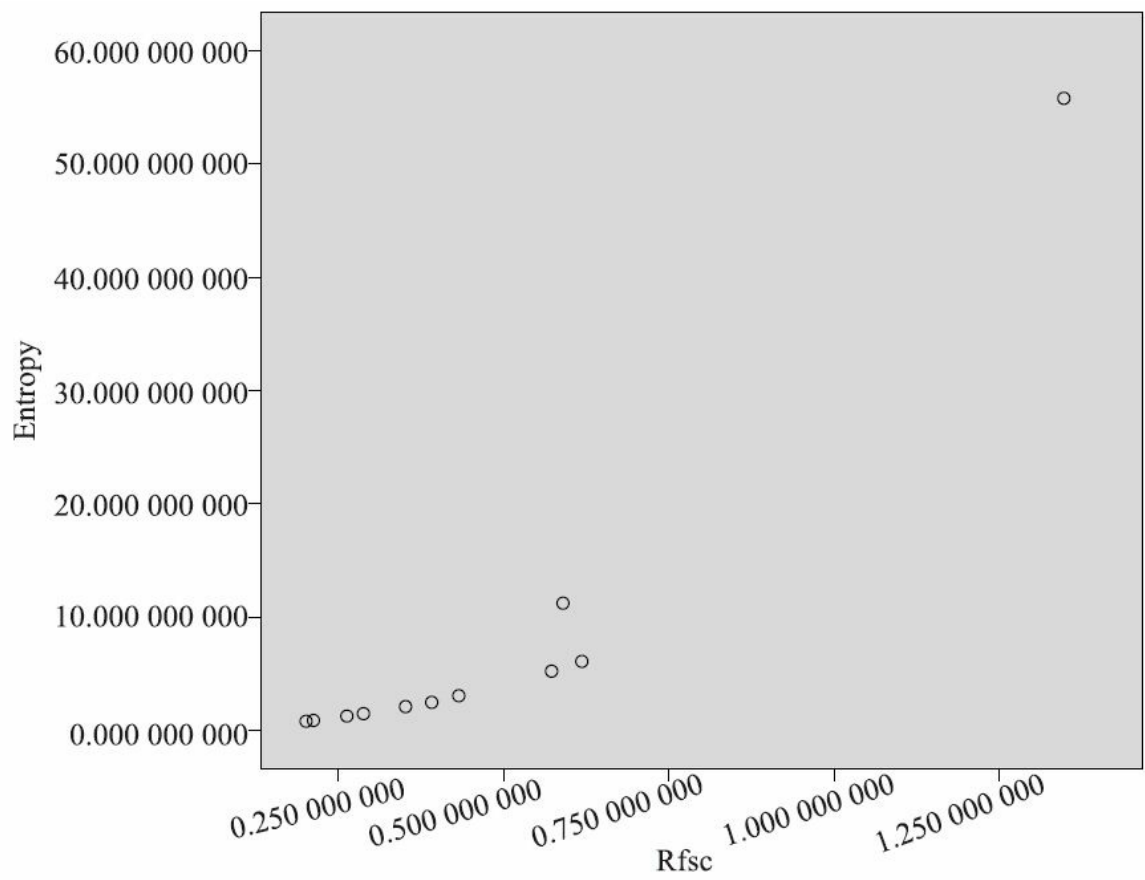


图 4-2

从图中我们可以看出，预后STATUS字段的相关因子的Entropy和Rfsc值呈现一种近似增长曲线的关系。

接下来，我们对SRCF算法和MDL算法进行比较。SRCF算法和MDL算法对性别SEX和预后STATUS的相关因子的比较，如表4-6和表4-7所示。

表4-6 用SRCF和MDL计算SEX的Entropy、Rfsc和Importance（按Rfsc排序）

列名	SRCF 算法		MDL 算法	
	Entropy	Rfsc	Rank	Importance
IN_DAYS	32.728 48	0.949 859	10	-0.012 41
AGE	32.231 34	0.455 337	6	-0.002 69
KILLP	1.326 928	0.170 407	11	-0.012 43
TANGNIAO	0.144 639	0.144 639	9	-0.006 97
SMOKE	0.086 668	0.073 434	2	0.026 341
GUANXIN	0.006 565	0.020 21	5	-0.0022

表4-7 用SRCF和MDL计算STATUS的Entropy、Rfsc和Importance（按Rfsc排序）

列名	SRCF 算法		MDL 算法	
	Entropy	Rfsc	Rank	Importance
IN_DAYS	55.800 44	1.347 158	1	0.083 72
AGE	11.2501	0.589 335	4	-0.003 86
KILLP	2.497 269	0.390 784	2	0.039 816
SMOKE	2.122 455	0.351 175	5	-0.004 16
SEX	1.287 258	0.262 394	3	-0.002 15

表4-6是SRCF算法和MDL算法对于目标属性SEX得到的分类因子，表4-7是SRCF算法和MDL算法对于目标属性STATUS得到的分类因子，分类因子均按照Rfsc值降序排列。从表中可以看出，使用SRCF算法得到的Entropy和Rfsc值高的因子，在使用MDL算法时不一定能够得到较高的权重和排序。这两种分类算法得到的因子排序不同，这说明SRCF算法和MDL算法具有不同的特性选择方法和体系。

观察SRCF算法和MDL算法评分散点图。首先观察关于性别SEX目标属性，由SRCF算法得到的相对熵（Entropy）与MDL算法得到的显著性评分（Importance）之间的关系图，如图4-3所示。

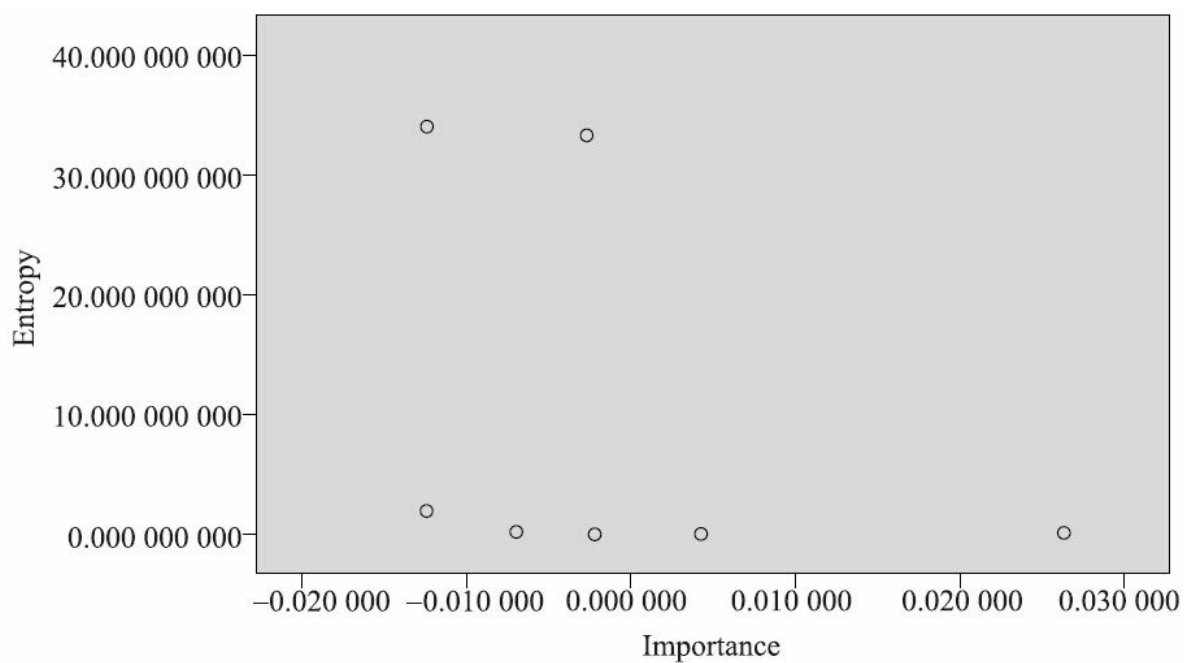


图 4-3

然后观察关于性别SEX，由SRCF算法得到的特征计分（Rfsc）与MDL算法得到的显著性评分（Importance）之间的关系图，如图4-4所示。

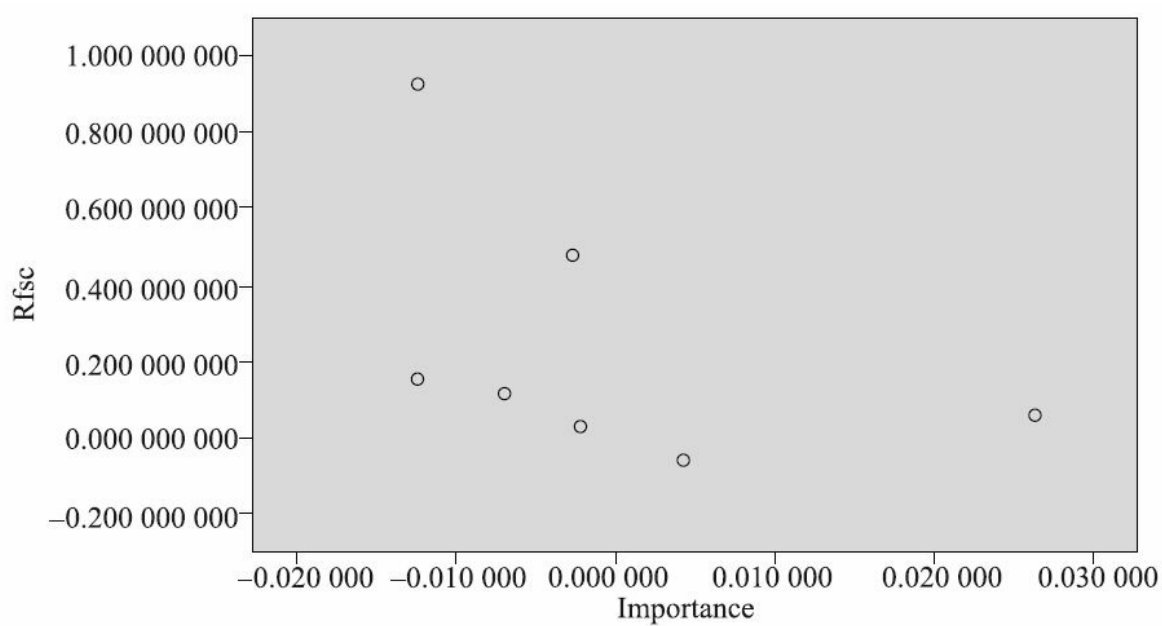


图 4-4

从图4-3和图4-4中，我们可以看出SEX的Entropy与Importance关系和其Rfsc与Importance关系有一定的相似性。连续型因子的IN_DAYS和AGE之间具有相对可比性，其他离散型因子之间具有相对可比性。在离散型因子部分，Importance与Entropy显示出十分低缓的线性下降趋势，Importance与Rfsc在主流上显示出比较明显的线性下降趋势。

然后，我们再观察一下预后STATUS的Entropy与Importance关系图，如图4-5所示。

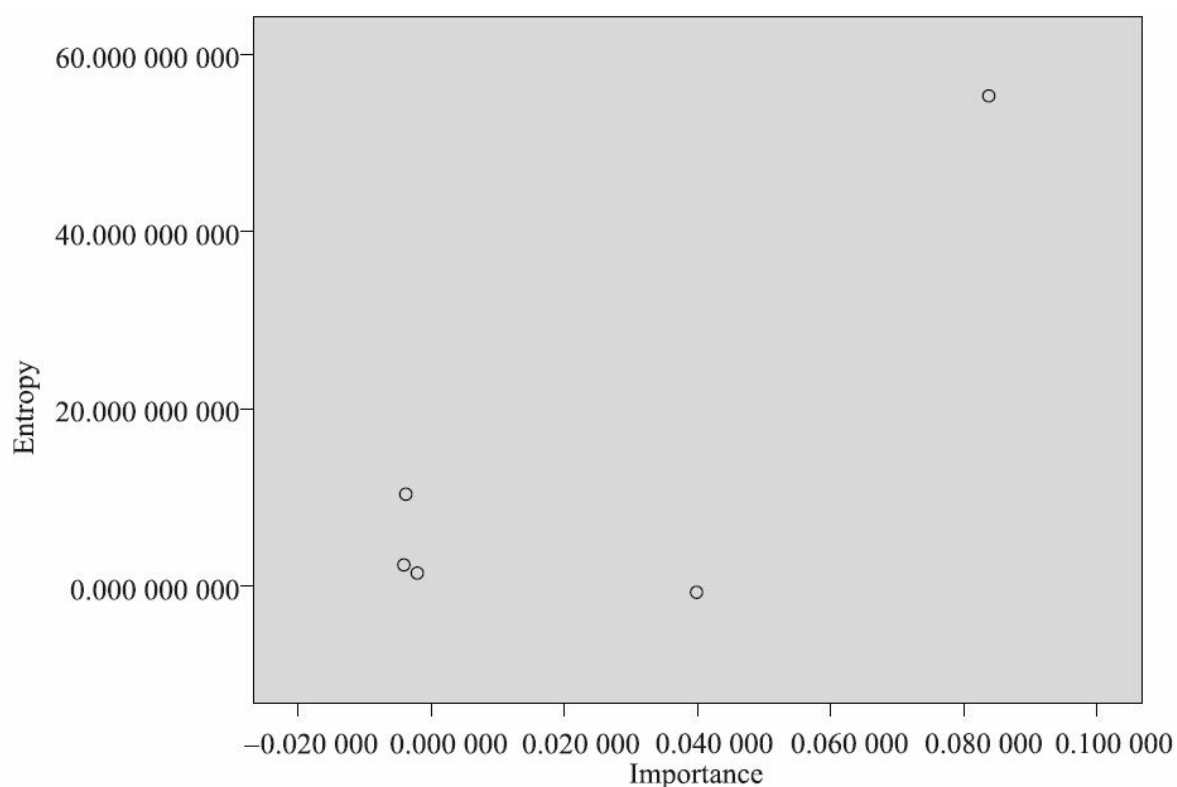


图 4-5

预后STATUS的Rfsc与Importance关系图，如图4-6所示。

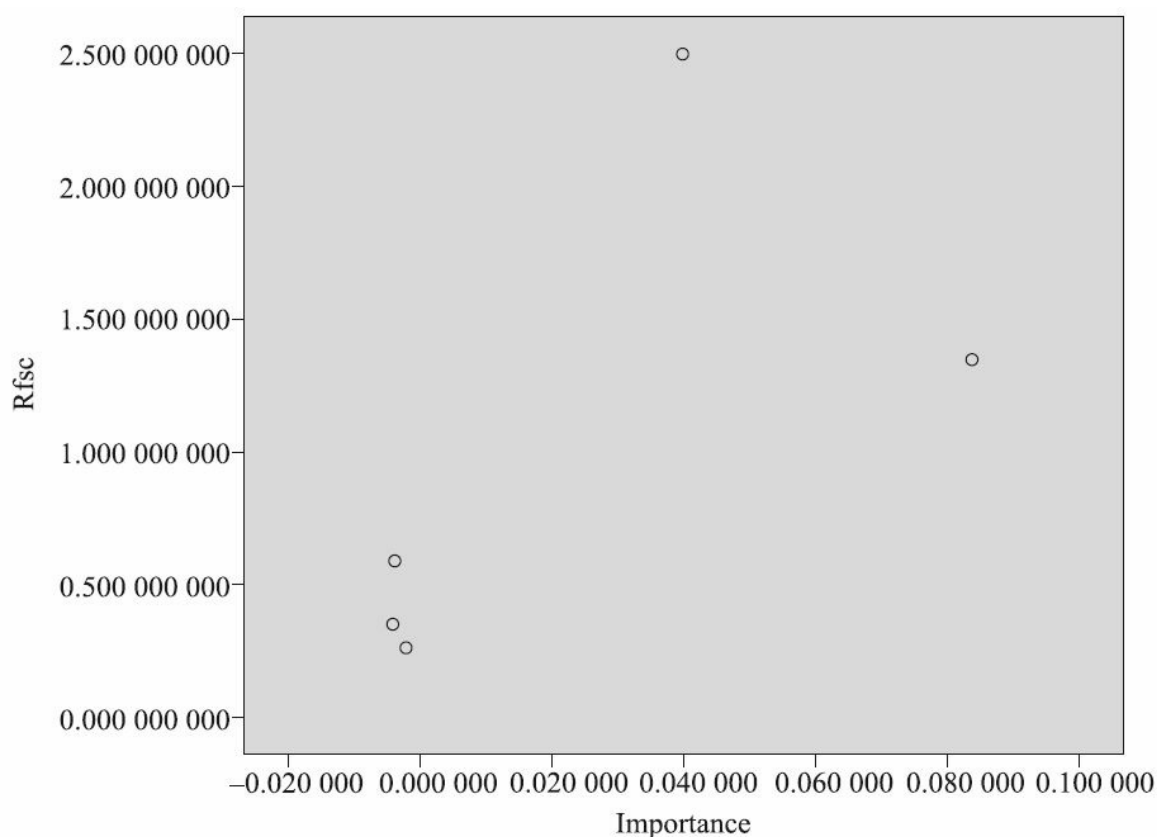


图 4-6

从图4-5和图4-6中，我们可以看出STATUS的Entropy与Importance关系和它的Rfsc与Importance关系也有相似性，但呈现的主流趋势不太清晰。这说明，一方面Entropy与Rfsc所表达的信息量有着类似方式，另一方面基于MDL算法的Importance系数是从不同于前两者的角度来表达数据间的关系的。在应用中，我们会根据案例的实际情况选用不同的特性选择方法，下一节我们会比较它们在案例中的应用效果。

[1] 预后是指预测疾病的可能病程和结局。

4.2 分类模型的应用——算法比较

本节选取了某大型三甲医院的肿瘤病人的临床数据，将与肿瘤疾病相关的特定生化指标应用特性选择算法进行指标影响度分析，对排名靠前的因子（生化指标）再利用不同的分类算法进行分析，得出分类结果。本节将介绍该案例的数据整合、数据描绘、数据标准化、特性选择探索和分类模型等各阶段内容，这些环节实际上反映了该案例的数据挖掘流程。首先我们要进行的准备工作是数据的整合。

4.2.1 数据整合

原始的数据是从该医院提供的Excel文件中导入到大型数据库系统Oracle。原始数据包括观察组（恶性肿瘤病人组）的数据和对照组的数据（非恶性肿瘤病人组）。观察组的数据可分为几大类：病人的基本信息、病人的既往史、病人的生化检查信息（肿瘤标志物）、病人的病理数据、专业的检测信息、病人的其他些信息，总共有570例数据。临床数据涵盖如下几个方面：

- 病人的基本信息，主要有病人编号、性别、年龄、入院时间、确诊时间等。
- 病人的既往史信息，主要有呼吸病史、肿瘤家族史、吸烟等。
- 病人的生化检测信息，主要包括癌胚抗原、神经烯醇化酶NSE、细胞角蛋白、胃泌素前体、鳞癌抗原等。
- 病人的病理数据，主要是病理类型、组织学分级、治疗方法等数据。
- 专业的检测信息，主要包括F检测、L检测、C检测等专业化的检测数据。
- 其他的信息，主要有进展时间、死亡时间、总生存期等。

对照组的数据又细分成两组，分别是健康组（健康人群）和良性

组（带良性肿瘤的病人），总共有233例数据。对照组中的数据基本上是专业的检测信息，例如，F检测、L检测、C检测等，此外还有一些病人的其他检测信息，如P04、P08、P09等信息。

我们对观察组和对照组进行了分析，此外也对对照组中的健康组和良性组进行了分析。导入数据的工具使用的是Oracle SQL Developer。将Excel文件中的数据导入到Oracle数据库后，再将观察组和对照组的数据整合到一起，存储在同一张数据表中，并增加字段Case_id作为每条数据的唯一标识键。

4.2.2 数据描绘

首先，计算数据字段中所有连续型变量的平均值、最大值、最小值、标准差和中位数，然后统计出连续型变量字段的空值记录数，如表4-8所示。

表4-8 连续型变量字段分析

字段	类型	AVG	MAX	MIN	标准差	中位数	NULL
P4	FLOAT	0.155	3.102	0	0.189	0.145	77
P8	FLOAT	0.281	3.268	0.005	0.196	0.238	22
P9	FLOAT	0.489	215.937	0.009	7.623	0.192	22
P15	FLOAT	0.256	2.7265	0.044	0.174	0.256	240
P25	FLOAT	0.236	1.479	0.058	0.116	0.236	240
P27	FLOAT	0.200	2.241	0.037	0.148	0.174	22
P29	FLOAT	0.481	1.6565	0.037	0.216	0.481	242
F	FLOAT	6.095	127.333	0	11.542	3.835	121
L	FLOAT	1192.626	7703.786	0	1061.322	922.905	133
C	FLOAT	5.282	15.0459	1.310	1.734	5.282	0
P	FLOAT	616.529	3284.169	-3.562	494.468	531.707	129
AGE	FLOAT	55.69	87	0	13.512	57	2
H3_TIME	FLOAT	20.109	65	0	18.715	20	252

(续)

字段	类型	AVG	MAX	MIN	标准差	中位数	NULL
H3_AMOUNT	FLOAT	24.825	70	1	12.949	20	469
N	FLOAT	1.314	3	0	1.177	2	257
M	FLOAT	0.303	1	0	0.460	0	289
PFS	FLOAT	7.723	65	0	7.740	6	689
ST	FLOAT	19.892	106	0	13.654	18	292

下一步，计算数据字段中所有离散型变量的不同取值个数，然后统计出所有离散型变量字段的空值记录数及空值百分比，如表4-9所示。

表4-9 离散型变量字段分析一

字段	类型	不同取值个数	空值个数	空值百分比%
NO	VARCHAR2(50 BYTE)	801	0	0
SEX	VARCHAR2(50 BYTE)	2	2	0.25
H1	VARCHAR2(50 BYTE)	2	244	30.46
H2	VARCHAR2(50 BYTE)	2	244	30.46
H3	VARCHAR2(50 BYTE)	2	245	30.59
CEA	VARCHAR2(50 BYTE)	3	254	31.71
NSE	VARCHAR2(50 BYTE)	3	254	31.71
LAB1	VARCHAR2(50 BYTE)	3	254	31.71
LAB2	VARCHAR2(50 BYTE)	3	254	31.71
LAB3	VARCHAR2(50 BYTE)	3	255	31.84
LAB4	VARCHAR2(50 BYTE)	3	277	34.58
LAB5	VARCHAR2(50 BYTE)	8	242	30.21
LAB6	VARCHAR2(50 BYTE)	2	285	35.58
T	VARCHAR2(50 BYTE)	5	515	64.29
STAGE	VARCHAR2(50 BYTE)	8	236	29.46
TYPE	VARCHAR2(50 BYTE)	9	232	28.96
GRADE	VARCHAR2(50 BYTE)	6	252	31.46
TREAT	VARCHAR2(50 BYTE)	9	233	29.09
STATUS	VARCHAR2(50 BYTE)	5	250	31.217

从表4-9中可以看出，与表4-2不同，数据中许多字段的空值记录数及比例都不小。进一步，需要计算所有离散型变量不同取值的记录个数，以便进一步了解数据分布情况，当然也包括离散型变量字段的空值记录数，如表4-10所示。

表4-10 离散型变量字段分析二

字段	字段取值 1 (记录数)	字段取值 2 (记录数)	字段取值 3 (记录数)	字段取值 4 (记录数)	字段取值 5 (记录数)	字段取值 6 (记录数)	字段取值 7 (记录数)	字段取值 8 (记录数)	字段取值 9 (记录数)	空值 个数
SEX	SEX =0; 266	SEX =1; 533								2
H1	H1 =0; 514	H1 =1; 43								244
H2	H2 =0; 515	H2 =1; 42								244
H3	H3 =0; 213	H2 =1; 343								245
CEA	CEA =1; 101	CEA =2; 276	CEA =3; 170							254
NSE	NSE =1; 101	NSE =2; 276	NSE =3; 170							254
LAB1	LAB1 =1; 39	LAB1 =2; 289	LAB1 =3; 219							254
LAB2	LAB2 =1; 4	LAB2 =2; 323	LAB2 =3; 220							254
LAB3	LAB3 =1; 19	LAB3 =2; 309	LAB3 =3; 218							255
LAB4	LAB4 =1; 224	LAB4 =2; 298	LAB4 =3; 2							277
LAB5	LAB5 =1; 1	LAB5 =2; 211	LAB5 =3; 291	LAB5 =4; 1	LAB5 =5; 52	LAB5 =6; 2	LAB5 =7; 1	LAB5 =8; 0		242
LAB6	LAB4 =0; 400	LAB4 =1; 116								285
T	T =1; 63	T =2; 125	T =3; 27	T =4; 58	T =5; 13					515
STAGE	STAGE =1; 4	STAGE =2; 45	STAGE =3; 202	STAGE =4; 67	STAGE =5; 56	STAGE =6; 43	STAGE =7; 124	STAGE =8; 24		236
TYPE	TYPE =1; 3	TYPE =2; 1	TYPE =3; 237	TYPE =4; 1	TYPE =5; 1	TYPE =6; 10	TYPE =7; 6	TYPE =8; 307	TYPE =9; 3	232
GRADE	GRADE =1; 77	GRADE =2; 63	GRADE =3; 185	GRADE =4; 7	GRADE =5; 22	GRADE =6; 195				252
TREAT	TREAT =1; 27	TREAT =2; 9	TREAT =3; 29	TREAT =4; 138	TREAT =5; 22	TREAT =6; 4	TREAT =7; 335	TREAT =8; 3	TREAT =9; 1	233
STATUS	STATUS =2; 31	STATUS =3; 217	STATUS =4; 263	STATUS =5; 40						250

接下来，要对数据库表中所有字段的空值进行处理。对于连续变量中的空值数据将赋予本字段的平均数，对于离散变量中的空值数将赋予一个不同于本字段所有取值的特殊值，如‘未知’、‘Unknown’等。

4.2.3 数据标准化

将原来Excel表中的中文字段在Oracle数据库表中统一用英文来表示。然后按照专业含义对数据进行整理，分别对观察组数据和对照组数据进行标签分类。在数据表中对每一行数据增加一个Diag_New字段作为目标分类变量，以区分对照组病例和观察组病例，如用9表示观察组，1表示良性组，0表示健康组。

4.2.4 特性选择探索

在应用分类模型之前，需要以分类变量为目标，与所有其他变量做相关性分析，即采用特性选择算法对目标分类变量的相关因子进行影响度显著性评价，以得到影响度显著性高的相关因子作为目标分类变量的分类因子。

4.2.4.1 应用SRCF算法

首先，应用我们创造的SRCF算法对目标分类变量Diag_New进行分类因子探索和选择，具体方法与4.1.4节同。

1.使用SRCF算法计算Diag_New的相关因子。

应用SRCF算法计算字段Diag_New的相关因子，结果如表4-11所示，表中的相关因子是按照Entropy降序排列的。

表4-11 用SRCF计算Diag_New的Entropy和Rfsc（按Entropy排序）

列名	Entropy	Rfsc	列名	Entropy	Rfsc
P9	114.9888	1.227 118	AGE	6.189 588	-0.042 27
P4	34.588 28	0.898 817	P29	5.121 004	0.316 604
P15	29.096 41	0.833 069	SEX	0.669 048	-0.203 57
P25	17.2346	0.668 753			

表4-11中给出的是使用SRCF算法分析的结果，计算相对于目标属性Diag_New的各个相关因子的Entropy和Rfsc值。表4-12中的相关因子是使用同一SRCF算法分析的结果按照每个因子的Rfsc值降序排列

的。

表4-12 用SRCF计算Diag_New的Entropy和Rfsc（按Rfsc排序）

列名	Entropy	Rfsc	列名	Entropy	Rfsc
P9	114.9888	1.227 118	P29	5.121 004	0.316 604
P4	34.588 28	0.898 817	AGE	6.189 588	-0.042 27
P15	29.096 41	0.833 069	SEX	0.669 048	-0.203 57
P25	17.2346	0.668 753			

从表4-11和表4-12中，我们能够看出P9、P4两个相关因子的Entropy和Rfsc值都是最高的，并且两个表中大部分因子的Entropy值和Rfsc值的排序基本是相同的。这说明对于Diag_New字段而言，P9、P4这两个因子有很好的相关性，也就是说P9、P4与Diag_New的不同取值（即对照组病例和观察组病例）有很强的相关关系。反过来可以说，对照组病例和观察组病例的差异性指标可能是P9、P4。

2.使用SRCF算法计算SEX的相关因子。

同理，我们可以计算出性别属性SEX的相关因子，如表4-13和表4-14所示。

表4-13 用SRCF计算SEX的Entropy和Rfsc（按Entropy排序）

列名	Entropy	Rfsc	列名	Entropy	Rfsc
P27	63.495 71	1.183 315	LAB6	1.029 358	0.252 279
ST	21.001 81	0.856 437	M	0.916 578	0.232 468
P29	10.101 29	0.646 571	N	0.672 193	0.179 364
AGE	3.691 425	0.282 237	TYPE	0.647 154	0.083 611
H3_TIME	2.500 036	0.360 365	GRADE	0.382 197	0.067 975
TREAT	1.763 848	0.187 375	H3	0.012 947	0.026 431
T	1.542 025	0.244 438			

表4-13中的相关因子是按照Entropy的大小降序排列的，表4-14中

的相关因子是按照Rfsc值降序排列的。

表4-14 用SRCF计算SEX的Entropy和Rfsc（按Rfsc排序）

列名	Entropy	Rfsc	列名	Entropy	Rfsc
P27	63.495 71	1.183 315	M	0.916 578	0.232 468
ST	21.001 81	0.856 437	TREAT	1.763 848	0.187 375
P29	10.101 29	0.646 571	N	0.672 193	0.179 364
H3_TIME	2.500 036	0.360 365	TYPE	0.647 154	0.083 611
AGE	3.691 425	0.282 237	GRADE	0.382 197	0.067 975
LAB6	1.029 358	0.252 279	H3	0.012 947	0.026 431
T	1.542 025	0.244 438			

比较表4-13和表4-14可以看出，按照Entropy排序和Rfsc值排序，相关因子的顺序不同但波动较大，其中排名靠前的P27、ST、和P29与性别的相关性可能较强，而在表4-13中排在H3_TIME之后的属性因子则可能与性别的相关性较弱了。

4.2.4.2 SRCF算法和MDL算法比较

我们使用了另外一种常用的特性选择算法MDL算法对以上数据进行了分析运算，结果在此不再赘述。下面我们把两种算法分析的结果进行比较，并讨论一下两种算法的异同。对Diag_New相关因子的比较，如表4-15所示。

表4-15 用SRCF和MDL计算Diag_New的Entropy、Rfsc和Importance（按Entropy排序）

列名	SRCF 算法		MDL 算法	
	Entropy	Rfsc	Rank	Importance
P9	114.9888	1.227 118	4	0.02
P4	34.588 28	0.898 817	2	0.04
P15	29.096 41	0.833 069	6	-0.001
P25	17.2346	0.668 753	7	-0.002
AGE	6.189 588	-0.042 27	1	0.16
P29	5.121 004	0.316 604	3	0.02
SEX	0.669 048	-0.203 57	5	-1.69

表4-15是SRCF算法和MDL算法计算出的对于分类属性Diag_New的各个因子的部分相关系数值。SRCF算法给出的参考值是相对熵Entropy和特征计分值Rfsc，MDL算法给出的是影响显著性评分Importance。表中的相关因子是按照Entropy降序排列的。从表中可以看出，Importance和Entropy对各个因子的相关性的评价不太相同，其中二者对P4和SEX的评分比较相近，而对其他的属性因子评分差异较大。我们用SRCF算法和MDL算法对数据中其他所有属性因子相对于分类属性Diag_New都做了属性相关性评价，并应用到下面的分类模型中。

4.2.5 分类模型

分类模型是用SRCF算法和MDL算法得到不同的分类因子，然后将这些分类因子应用到不同的分类算法中，最后对分类结果进行分析比较。本次分析采用了四种不同的分类算法，分别是：支持向量机（SVM）、贝叶斯网络（ABN）、决策树（DT）、贝叶斯算法（NB）。其中，贝叶斯算法作为验证算法。

表4-16中的数据是以MDL和SRCF因子选择算法得到的对于目标属性Diag_New不同的相关因子，利用四个分类算法进行分类后在预测准确率、代价成本和置信度（Confidence）上得到不同的结果。表中分类目标值栏里的1表示的是对照组，9表示的是观察组。

表4-16 对照组与观察组比较

	分类目标值 Target	实际总数 Total Actual	正确预测率 Correctly Predicted%	成本 Cost	成本率 Cost %	置信度 Confidence
基于 MDL 因子选择的 贝叶斯网络 abn_md1	1	90	80.00	62.15	47.39	57.10%
	9	214	77.10	68.98	52.61	
基于 SRCF 因子选择的 贝叶斯网络 abn_srcf	1	90	80.00	62.15	47.39	57.00%
	9	214	77.10	68.98	52.61	
基于 MDL 因子选择的 决策树 dt_md1	1	90	71.11	89.77	85.29	65.97%
	9	214	94.86	15.49	14.71	

(续)

	分类目标值 Target	实际总数 Total Actual	正确预测率 Correctly Predicted%	成本 Cost	成本率 Cost %	置信度 Confidence
基于 SRCF 因子选择的决策树 dt_srcf	1	90	70.00	93.22	85.75	64.86%
	9	214	94.86	15.49	14.25	
基于 MDL 因子选择的支持向量机 svm_md1	1	90	92.22	7.00	14.29	72.60%
	9	214	80.37	42.00	85.71	
基于 SRCF 因子选择的支持向量机 svm_srcf	1	90	86.67	12.00	19.05	62.83%
	9	214	76.17	51.00	80.95	
基于 MDL 因子选择的贝叶斯 nb_md1	1	90	80.00	62.15	100.00	80.00%
	9	214	100.00	0.00	0.00	
基于 SRCF 因子选择的贝叶斯 nb_srcf	1	90	76.67	72.50	47.03	49.56%
	9	214	72.90	81.65	52.97	

首先是四个算法应用不同相关因子之间的区别。从表4-16中我们能够看出，对于用MDL和SRCF得到的分类因子，使用贝叶斯网络算法得到的置信度和预测准确率都是非常相近的，说明这两组分类因子对于贝叶斯网络算法来说，对于分类目标Diag_New的分类效果相差不多。决策树算法使用这两组分类因子得到的置信度和预测准确率也很近似，并且总体效果比贝叶斯网络算法要好，这说明决策树算法的性能在这个数据中要优于贝叶斯网络算法。支持向量机的算法在这4类算法中，置信度和预测准确率都是比较适中的，使用MDL算法产生的相关因子在置信度和预测准确率上要比SRCF算法的好些。贝叶斯算法（NB）对于不同组的分类因子，置信度变化十分明显，说明贝叶斯算法对Diag_New的分类效果，用SRCF算法得到的分类因子不如使用MDL得到的分类因子好。此外，还可看出使用SRCF得到的分类因子在每个算法上的分类置信度都低于MDL算法得到的分类置信度。

其次是同一组分类因子对不同算法的影响。由SRCF算法得到的分类因子是AGE、C、P27和P4，在这组分类因子中，决策树算法的

分类置信度最高并且在观察组中的预测准确率最高、代价开销最小；贝叶斯算法在四个算法中的置信度最低且预测准确率也是最低的。由MDL算法得到的分类因子是AGE、H1、H2和P4，在这组分类因子中，贝叶斯网络算法的分类置信度是最低的，但是与SRCF算法得到的分类因子的置信度接近，这说明这两组因子对贝叶斯网络算法的影响不大；分类置信度最高的是贝叶斯算法，而且预测的准确率也是最高的，这与使用SRCF算法得到的分类因子形成鲜明的对比，说明这两组分类因子对贝叶斯算法的影响是很大的。进一步说，由于贝叶斯算法是基于条件概率原理的，它要求分类因子之间是独立的、不相关的，因此，由SRCF算法得到的这组因子间的独立性没有MDL算法得到的因子间的独立性要好。

4.3 分类模型的应用——网络异常侦测

上一节，我们讲述了不同分类算法的分类模型应用和比较。本节我们将介绍应用分类模型开展网络异常侦测。这里提到的异常侦测是业务上的，而不是技术上的，不要与数据挖掘的异常侦测技术相混淆。我们将从计算机网络异常行为的分析开始，重点介绍网络异常数据模型及分类模型算法的应用。

4.3.1 计算机网络异常行为

常见的网络恶意行为是指网络系统的硬件、软件及其系统中的数据受到恶意代码攻击而遭到破坏、更改、泄露，致使系统不能连续、可靠、正常地运行，并可能造成网络服务中断的行为。随着信息化的普及、网络新应用的大量出现，网络恶意代码所表现出的行为也层出不穷。目前最流行的网络恶意行为是网页挂马、盗取网游帐号、端口扫描、漏洞扫描、ARP欺骗、IP劫持、DDOS攻击、溢出攻击、木马攻击等。针对这些大量的网络恶意行为分析，我们大致可以把网络恶意行为分为四类：

- 信息探测行为（如端口和漏洞扫描）
- 信息欺骗行为（如ARP欺骗和IP劫持）
- 信息淹没行为（如SYN Flood和DDOS攻击）
- 信息伪传递行为（溢出攻击、ICMP木马、HTTP隧道木马）

从上述的四类行为可知，网络恶意行为总体上可以从两个方面来描述，一个方面是基于协议栈的行为，针对这种行为可以从数据包中给出的信息来描绘；另一方面是基于主机操作系统的行为，针对这种行为可以从进程的行为信息来描绘。

互联网工作的架构如图4-7所示。

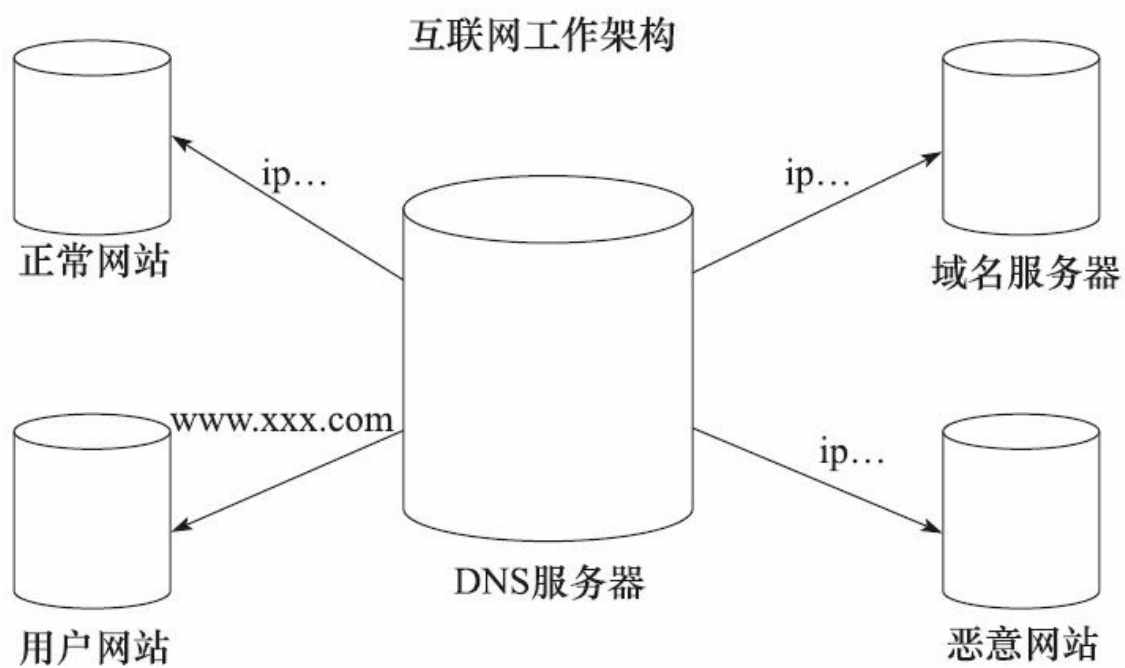


图 4-7

有关互联网工作的一些技术细节，请参见有关参考书，在此不宜详细讨论。

4.3.2 网络异常数据模型

从互联网服务商的网络服务器中产生的大量实时网络访问日志数据中，我们获得了有关每一个网络访问活动的信息，如图4-8所示。

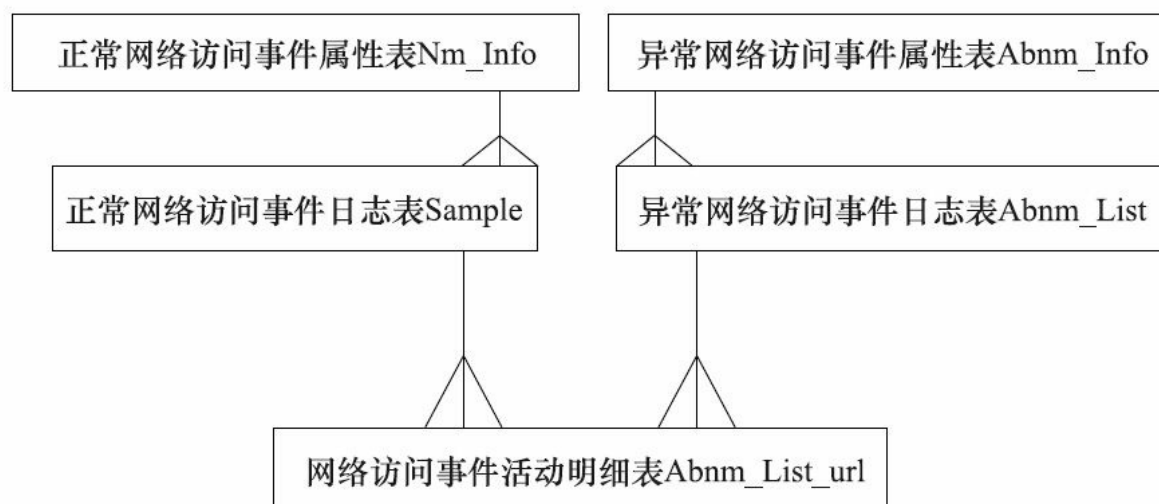


图 4-8

数据模型中，网络访问事件日志表是反映业务数据的基本单位，包括正常网络访问事件日志表（Sample）和异常网络访问事件日志表（Abnm_List），它们有相应的属性父级表，即正常网络访问事件属性表（Nm_Info）和异常网络访问事件属性表（Abnm_Info）。其实，Nm_Info表和Abnm_Info具有基本相同的数据结构，只是随着新的异常访问网络事件不断被侦测出来，其相应的属性信息从Nm_Info表中不断地复制到Abnm_Info，列为了黑名单。网络访问事件日志表的子级表Abnm_List_url则记录了网络访问事件的详细活动细节。

4.3.2.1 数据表Abnm_Info

数据表Abnm_Info是异常网络访问事件的属性表部分数据，数据结构如表4-17所示。

表4-17 Abnm_Info的数据结构

字段	含义	类型	空值
DOMAINID	域 ID	NUMBER	NO
URL_DOMAINSTR	URL 域结构	VARCHAR2	YES
DNPARTNUM	域名分段号	NUMBER	YES
DNPART1	域名分段 1	VARCHAR2	YES
DNPART2	域名分段 2	VARCHAR2	YES
DNPART3	域名分段 3	VARCHAR2	YES
DNPART4	域名分段 4	VARCHAR2	YES
URL_DOMAIN_TYPE	URL 域类型	VARCHAR2	YES
URL_IP	URLIP	VARCHAR2	YES
URL_LOCATION	URL 位置	VARCHAR2	YES

部分字段样本数据如图4-9所示。

4.3.2.2 数据表Sample

数据表Sample是网络访问事件日志表，它反映了互联网上大量的实时网络站点间的访问活动，互联网服务商的网络服务器每小时捕获大约200多万条网络访问事件记录，数据结构如表4-18所示。

1	2	3	4	5	6	7	8	9	10
URL_DOMAINSTR	DNPARTNUM	DNPART1	DNPART2	DNPART3	DNPART4	URL_DOMAIN_TYPE	URL_IP	URL_LOCATION	
22747474.com	2	22747474	com	(null)	(null)	.com	218.25.11.6	辽宁省沈阳市 联通IDC机房	
234uuu.com	2	234uuu	com	(null)	(null)	.com	67.159.20.238	美国 CZ88.NET	
3579.cc	2	3579	cc	(null)	(null)	其它	110.45.196.152	韩国 CZ88.NET	
3d-line.nl	2	3d-line	nl	(null)	(null)	其它	83.137.194.110	美国 CZ88.NET	
500h.info	2	500h	info	(null)	(null)	.info	173.208.228.146	北美地区 CZ88.NET	
51cai8.com	2	51cai8	com	(null)	(null)	.com	58.247.128.22	上海市 联通	
7j5.bd12.in	3	7j5	bd12	in	(null)	其它	61.159.225.18	云南省昆明市 电信ADSL	
97.41bo.com	3	97	41bo	com	(null)	.com	67.159.20.235	美国 CZ88.NET	
a.web9700.com	3	a	web9700	com	(null)	.com	174.139.50.42	美国 theplanet	
arpocc.3322.org	3	arpocc	3322	org	(null)	.org	218.145.29.237	韩国 CZ88.NET	
bar.netkuu.com	3	bar	netkuu	com	(null)	.net	218.25.58.49	辽宁省沈阳市东陵区 联通	
bdown.dynssl.com	3	bdown	dynssl	com	(null)	.com	203.251.224.71	韩国 CZ88.NET	
cxry.gsau.edu.cn	4	cxry	gsau	edu	cn	.edu.cn	61.178.69.74	甘肃省兰州市 电信	
f47.bd12.in	3	f47	bd12	in	(null)	其它	61.159.225.18	云南省昆明市 电信ADSL	
ftpl101212.24315...	4	ftpl101212	243155	ftpdo	com	.com	204.45.72.147	澳大利亚 TELSTAR公司	
gvr.ns02.us	3	gvr	ns02	us	(null)	其它	121.22.6.39	河北省秦皇岛市 联通	
gwd.ns02.us	3	gwd	ns02	us	(null)	其它	121.22.6.39	河北省秦皇岛市 联通	
gro.ns02.us	3	gro	ns02	us	(null)	其它	221.10.97.90	四川省绵阳市 联通	
grs.ns02.us	3	grs	ns02	us	(null)	其它	221.10.97.90	四川省绵阳市 联通	
grw.ns02.us	3	grw	ns02	us	(null)	其它	221.10.97.90	四川省绵阳市 联通	
gzy.ns02.us	3	gzy	ns02	us	(null)	其它	221.10.97.90	四川省绵阳市 联通	
gzz.ns02.us	3	gzz	ns02	us	(null)	其它	221.10.97.90	四川省绵阳市 联通	

图 4-9

表4-18 Sample数据结构

字段	含义	类型	空值
ID	事件 ID	NUMBER	NO
RTIME	事件发生时间	DATE	NO
MSIP	访问发出者 IP 地址	VARCHAR2	NO
SPORT	访问发出者端口号	NUMBER	NO
MDIP	访问目标 IP 地址	VARCHAR2	NO
DPORT	访问目标端口号	NUMBER	NO
DNAME	访问域名	VARCHAR2	NO
RRIP	RR IP 地址	VARCHAR2	NO
FLAG	事件异常标志	VARCHAR2	NO

部分字段样本数据如图4-10所示。

由于数据的敏感性，这里我们将访问发出者的IP地址和访问目标的IP地址两个字段的数据进行加密。

4.3.2.3 数据表Abnm_List

数据表Abnm_List与表Sample类似，也是来自网络访问事件日志

表，但它是侦测人员从互联网上大量的实时网络站点间的访问活动中，采用各种人工方法和半自动方法（非数据挖掘方法）发现的网络异常事件，数据结构与Sample表相同。

ID	RTIME	MSIP	SPORT	MDIP	DPORT	DNAME	RRIP	FLAG
27640066	2012-03-08 13:22:21	e7e4c3d769e227...	1028	84934dc9da3adea...	53	cloud-u.duba.net	0.0.0.0	0
27940069	2012-03-08 13:22:21	e7e4c3d769e227...	1028	f995d9d1483973a...	53	videodown.baofeng.com	0.0.0.0	0
28040070	2012-03-08 13:22:21	4d10d967d3188e...	65355	2b1a1872b7c4a43...	53	afs.opap.co.kr.cdnge.net	0.0.0.0	0
28140071	2012-03-08 13:22:21	ed9d2d87fd3efc...	17582	5774fea9e8ee985...	53	33.pcpop.chinacache.net	0.0.0.0	0
28240072	2012-03-08 13:22:21	ffe9d49a9f7454...	30864	603f92d66f44c44...	53	dcm-zupl.com	0.0.0.0	0
28340073	2012-03-08 13:22:21	4d10d967d3188e...	37166	528400277dc458e...	53	8car.sxcartg.com	0.0.0.0	0
28440074	2012-03-08 13:22:21	e6b76c031b86a1...	48361	9a319ad33b023e3...	53	ww4.sinaimg.cn	0.0.0.0	0
28540075	2012-03-08 13:22:21	5157674a7ec158...	57429	2c3a3d9f1df0ec6...	53	www.update.microsoft.com	0.0.0.0	0
28640076	2012-03-08 13:22:21	f6b65e8621115a...	18431	2c3a3d9f1df0ec6...	53	imgcache.qq.com	0.0.0.0	0
28740077	2012-03-08 13:22:21	eedbe76e633120...	28374	2ac0fcc8d184026...	53	7v7kqld.mnshow99.com	0.0.0.0	0
28840078	2012-03-08 13:22:21	ac3a2835b462bf...	35166	178c5811c3fe0ee...	53	bn.eeaagllleerz.info	0.0.0.0	0
28940079	2012-03-08 13:22:21	65175922d54b6e...	52094	c37013d9ec4bc1e...	53	ns6.360safe.com	0.0.0.0	0
29040080	2012-03-08 13:22:21	e6b76c031b86a1...	48361	9a319ad33b023e3...	53	ssl1.sinaimg.cn	0.0.0.0	0
29140081	2012-03-08 13:22:21	3e1c416999d0d4...	50929	f45d6aa7551bb27...	53	grx.ns02.us	0.0.0.0	1
29240082	2012-03-08 13:22:21	953fd06c84eb24...	63566	8565183881d0502...	53	www.aviliv.com	0.0.0.0	0
29340083	2012-03-08 13:22:21	727f5502569aac...	30320	e9d2374f5b5f19...	53	www.hg00882.com	0.0.0.0	0
29440084	2012-03-08 13:22:21	6264d857a4df2e...	8956	4bc229e6c81da38...	53	bsgkabrndw.router	0.0.0.0	0
29540085	2012-03-08 13:22:21	6264d857a4df2e...	45055	9f4a12a65b66156...	53	s49.app28982.qqopenap...	0.0.0.0	0
29640086	2012-03-08 13:22:21	aa22bc05996a4f...	64226	6849b3597f82551...	53	dyn.www.rayfile.com	0.0.0.0	0
29740087	2012-03-08 13:22:21	953fd06c84eb24...	20438	f4199b0f732ed37...	53	www.01rph.com	0.0.0.0	0
29840088	2012-03-08 13:22:21	7e0fc3eb67c549...	59962	06df3d3b328857b7...	53	4567.info	0.0.0.0	0
29940089	2012-03-08 13:22:21	b4536a7c2f763d...	51070	eb7aa7341911467...	53	mx1.hanmail.net	0.0.0.0	0
30040090	2012-03-08 13:22:21	6264d857a4df2e...	36769	fla27a0bba039bb...	53	jump.szvr.com	0.0.0.0	0
30140091	2012-03-08 13:22:21	65175922d54b6e...	56905	22815a5253ff5ab...	53	al.zassets.com	0.0.0.0	0

图 4-10

4.3.2.4 数据表Abnm_List_url

数据表Abnm_List_url是网络访问事件的详细活动信息，其部分数据结构如表4-19所示。

表4-19 Abnm_List_url数据结构

字段	含义	类型	空值
URL_ID	URL ID	NUMBER	NO
URL_MD5	URL MD	VARCHAR2	YES
URLSTR	URL 结构	VARCHAR2	YES
URL_DOMAINSTR	URL 域结构	VARCHAR2	YES
URL_LAYER	URL 层次	NUMBER	YES
URL_FILENAME	URL 文件名	VARCHAR2	YES
URL_FILEEXT	URL 文件名扩展	VARCHAR2	YES

部分字段样本数据如图4-11所示。

4.3.2.5 表结构关系

网络访问事件属性表作为网络访问事件日志表的父级表，网络访问事件属性表与网络访问事件日志表是一对多的关系，而且二者是依赖关系，即对于网络访问事件日志表中的每一条记录，必有一条有关它的属性记录存在于网络访问事件属性表中。同样，网络访问事件日志表作为网络访问事件活动明细表的父级表，网络访问事件日志表与网络访问事件活动明细表是一对多的关系，而且二者是依赖关系，即对于网络访问事件活动明细表中的每一条记录，必有一条有关它的属性记录存在于网络访问事件日志表中。

URL_ID	URL_MD5	URLSTR	URL_DOMA...	URL_LAYER	URL_FILENAME	URL_FILEEXT	REPORTDATE
167147719	c859f84c3c2ec96e...	http://22747474.com/car.asp?bdc...	22747474.com	2	car	asp	.. 20120222
267148969	3fecc958cc073b5d...	http://22747474.com/car.asp	22747474.com	2	car	asp	.. 20120222
366013998	b46805d8ceefbef2...	http://234uuu.com/html/9611.html	234uuu.com	3	9611	html	.. 20120206
466014029	e1af52edb541e96e...	http://234uuu.com/list/class_70...	234uuu.com	3	class_70_14	html	.. 20120206
566014034	9a550fc78f3162b0...	http://234uuu.com/html/14712.html	234uuu.com	3	14712	html	.. 20120206
666014052	590110f9793c20aa...	http://234uuu.com/list/class_65...	234uuu.com	3	class_65_1	html	.. 20120206
766014086	2dbdd0141a1e03...	http://234uuu.com/html/13953.html	234uuu.com	3	13953	html	.. 20120206
866014112	579276f19779d3c6...	http://234uuu.com/html/21787.html	234uuu.com	3	21787	html	.. 20120206
966014122	5a55bef195d9caf3...	http://234uuu.com/html/21785.html	234uuu.com	3	21785	html	.. 20120206
1066014130	59e54aa6860cd227...	http://234uuu.com/list/class_70...	234uuu.com	3	class_70_8	html	.. 20120206
1166014138	c481f9fdb084af33...	http://234uuu.com/html/8777.html	234uuu.com	3	8777	html	.. 20120206
1266014146	282ad0397b85c6a2...	http://234uuu.com/html/8785.html	234uuu.com	3	8785	html	.. 20120206
1366014172	bcef72cc6a8aca79...	http://234uuu.com/html/8041.html	234uuu.com	3	8041	html	.. 20120206
1466014191	bfa6b743dc59a6e...	http://234uuu.com/html/7210.html	234uuu.com	3	7210	html	.. 20120206
1566014209	a1fc6b33de17df1e...	http://234uuu.com/html/22181.html	234uuu.com	3	22181	html	.. 20120206
1666014210	313f8dcfd4681f7e...	http://234uuu.com/html/8744.html	234uuu.com	3	8744	html	.. 20120206
1766014222	f35564a42023d382...	http://234uuu.com/html/2912.html	234uuu.com	3	2912	html	.. 20120206
1866014258	cad9ebee67e8b2cf...	http://234uuu.com/list/class_52...	234uuu.com	3	class_52_33	html	.. 20120206
1966014262	eca99685d36b09c0...	http://234uuu.com/html/6362.html	234uuu.com	3	6362	html	.. 20120206
2066014270	0e58637521121363...	http://234uuu.com/html/13939.html	234uuu.com	3	13939	html	.. 20120206
2166014484	8bc18cbca92620cf...	http://234uuu.com/html/6332.html	234uuu.com	3	6332	html	.. 20120206
2266014526	c30fe2f8f2ad7c7c...	http://234uuu.com/html/6372.html	234uuu.com	3	6372	html	.. 20120206

图 4-11

4.3.2.6 数据准备

这是一个不多见的、数据准备非常简单的、几乎可以直接应用数据挖掘算法的案例。由于数据采自互联网服务商的网络服务器所产生的大量网络实时访问日志数据，它的数据特点是数据关系模型简单，数据质量比较好，但单位时间产生的数据量非常大。因此，不同于业务关系复杂、数据质量参差不齐的医学数据，我们为这个信息安全的数据挖掘项目所做的数据准备阶段的任务相对比较简单。

我们得到的实验数据来自Sample表的某一小时的数据，以及由此表中人工侦测到的异常事件数据，并将其存储在Abnm_List表中。其中Sample表的数据有2042379条记录，Abnm_List表有77条数据，合计有2042456条记录。我们将两个表的数据合在一起，并增加了一个字段Flag，表示每条记录是正常还是异常，我们用0表示正常，用1表示

异常。由于表中的ID字段为不重复的键值，可以直接用作该表的唯一键。在数据采样过程中，我们所使用的算法是按照随机选取ID字段的值来取相应的记录行的，而不是按照数据库表中的记录行存放的物理顺序，如ROWNUM或ROWID来选取数据，并且Sample表的数据和Abnm_List表的数据的ID值并不是顺序存储的，所以不影响包括正常和异常数据的采样的均衡性和随机性。

4.3.3 分类模型算法应用

虽然我们的业务目的是从大量的数据中侦测出少数的异常行为，但从数据挖掘技术上讲，这实际是一个典型的分类问题。虽然看上去似乎也可以应用异常侦测的方法来处理，但是，由于异常侦测算法的实质是一种分类法，即将大多数数据划分为正常，而少数数据划为异常，它的特点是划分正常和异常的界限不是很清晰，往往比较依赖于算法中人工设定的离群率等参数。在本案例中，虽然我们实验的数据的异常比例是已知的，但由于网络活动的动态变化很大，网络事件的异常率变化也很大，我们很难预先判断异常率。另外，从业务经验和业务知识中，我们知道异常网络事件应该有比较明显的行为特征，所以我们应当采用基于分类算法的两分类方法，它较少受到动态异常率的影响，对数据中隐藏的行为差别比较敏感。

4.3.3.1 分类算法的选择

我们从由Sample表和Abnm_List表的2042456条记录组成的数据中，随机抽取了816050条数据，其中正常网络活动数据816021条、异常网络活动数据29条。我们尝试了近十种不同的分类模型算法，如朴素贝叶斯、决策树、多元Logistic回归、支持向量机、神经网络模型等。经实际应用与对比，我们发现在网络访问数据环境中，支持向量机和朴素贝叶斯模型的效果不错。具体结果比较如下：

将应用支持向量机算法得到的分类模型称为模型A，图4-12为模

型A分别预测正常网络活动和异常网络活动的准确度。

应用朴素贝叶斯算法得到的分类模型称为模型B，图4-13为模型B分别预测正常网络活动和异常网络活动的准确度。

异常标志	记录数	预测准确度
0	816 021	99.98%
1	29	58.62%

图 4-12

异常标志	记录数	预测准确度
0	816 021	93.83%
1	29	82.76%

图 4-13

如图4-13显示，模型A预测0（即正常网络访问事件）的正确率为99.98%，预测1（即异常网络访问事件）的正确率为58.62%；模型B预测0（即正常网络访问事件）的正确率为93.83%，预测1（即异常网络访问事件）的正确率为82.76%。

模型A从29个实际异常网络访问事件（阳性事件）中检测出了17个，漏检出12个；同时模型A将816021个实际正常网络访问事件（阴性事件）中的815858个事件标识为正常事件（检验通过），将其中的

151个事件错检为异常事件，如图4-14所示。

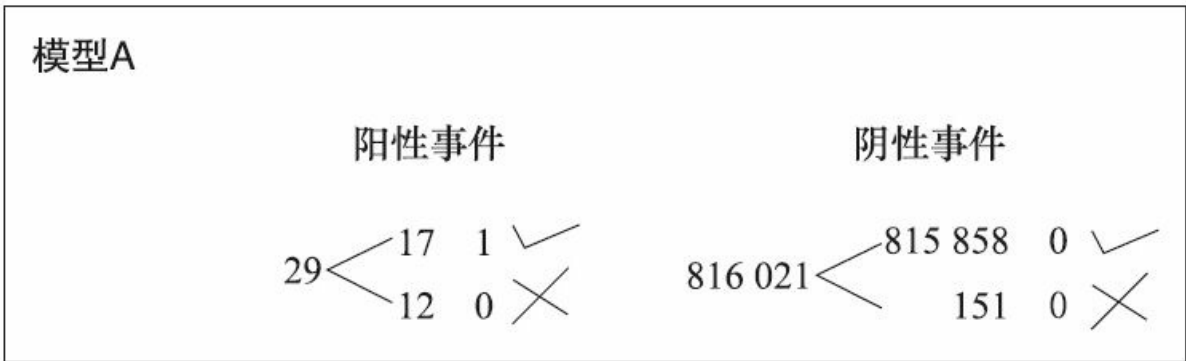


图 4-14

因此，模型A的阳性检出率为 $17/29=58.62\%$ ，

阴性检出率为 $815858/816021=99.98\%$ ，

模型总成本为 $(12+151) / (17+815858) \times 100\%=0.02\%$ 。

模型B从29个实际异常网络访问事件（阳性事件）中检测出了24个，漏检出5个；同时模型B将816021个实际正常网络访问事件（阴性事件）中765673个事件标识为正常事件（检验通过），将其中的50348个事件错检为异常事件，如图4-15所示。

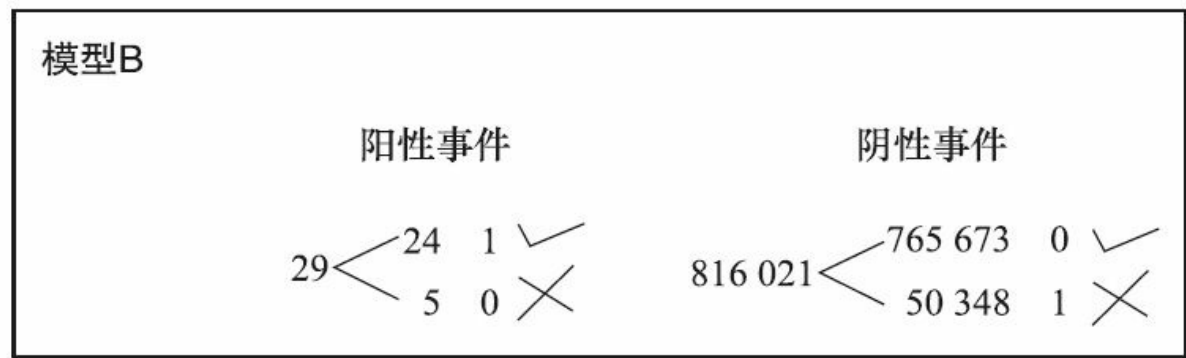


图 4-15

因此，模型B的阳性检出率为 $24/29=82.76\%$ ，

阴性检出率为 $765673/816021=93.83\%$ ，

模型总成本为 $(5+50348)/(24+765673)\times 100\%=6.58\%$ 。

两个模型比较起来，我们会发现模型B较模型A有更高的“纠错”能力，阳性检出率高出 $82.76\%-58.62\%=24.14\%$ ；但是模型B较模型A有更高的“冤案”率，阴性检出率低出 $93.83\%-99.98\%=6.15\%$ 。在不同的案例中，业务的需求不太一样。有时，我们希望“宁可错杀一万，也不能漏网一个”；有时我们希望“宁可漏网一万，也不能错杀一个”。当然，理想的模型应是既具有较高的识真能力又有较强的辨伪能力。如果二者不能兼顾的话，我们就需要根据业务的特点来取舍了。

在本案例中，我们当然希望模型具有较高“纠错”能力，但是由于异常数据与正常数据的比例差距过大，如果识真率较低，则会带来大量的人工工作量。因为，B模型较高的“纠错”能力带来了高昂的预测成本，它的模型总成本为6.58%，也就说它给信息安全人员提交 $50348+24=50388$ 个“异常数据”，需要人工从中找到24个真异常数据。相反，A较模型较低的“纠错”能力带来了很低的预测成本，它的模型总成本只有0.02%，也就说它给信息安全人员仅提交了 $151+17=168$ 个“异常数据”，只需要人工从中找到17个真异常数据，大大减轻了人工工作量。

因此，通过两个模型的比较，在本案例中应选用模型A。

4.3.3.2 分类算法的组合应用

通过上面的比较分析，我们发现模型A和模型B各有自己的优缺点，模型A更适用于本案例的业务需求。我们能否在上述方案基础上加以改进呢？为了弥补单一分类模型各自的不足，我们采用了组合分类模型算法，即将支持向量机算法和朴素贝叶斯算法组合一起使用，应用模型A+模型B的方案，这样做到了多模型的优势互补。方法如下：

1) 首先利用模型A的低成本优势，将原数据816050条记录应用模型A，得到168个（151+17）异常标识数据，其中的10%为真异常，数据异常正确检出率为58.6%。应用模型A后得到分类结果，其中标识为1的共有168条，分布如图4-16所示。

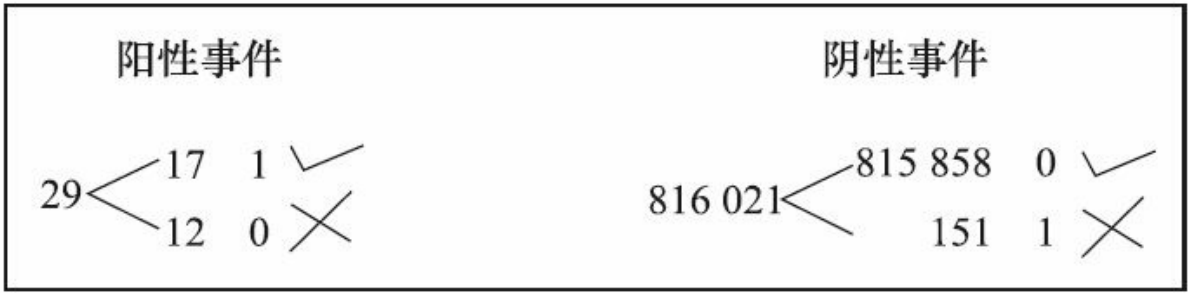


图 4-16

2) 然后利用模型B的高“纠错”能力，对上一步得到的168条数据应用模型B，再次得到23个（14+9）异常标识数据，如图4-17所示。

异常标志	记录数	预测准确度
0	151	93.83%
1	17	82.76%

图 4-17

应用模型B后得到分类结果，其中标识为1的有23条，分布如图4-18所示。

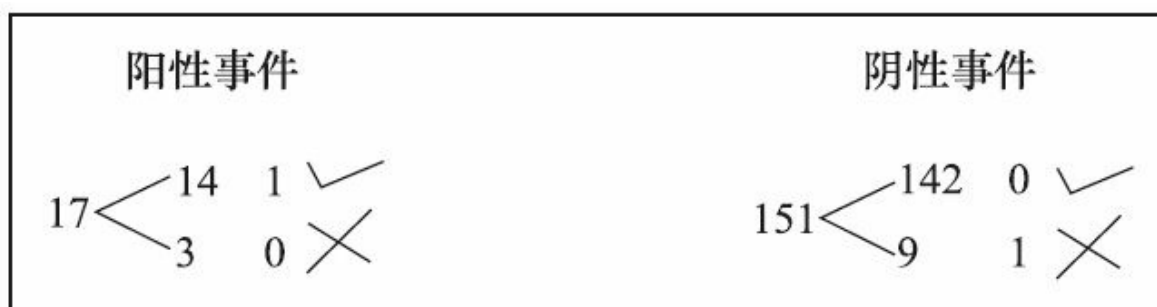


图 4-18

3) 最后由人工从23个数据中找出14个真异常数据，其中60.9%为真异常，数据异常正确检出率是82.4%。

通过对23个标识异常数据进行人工筛选得到14个真异常，数据异常侦测总正确率是 $14/29=48.3\%$ ，人工总成本：

$23/816050=0.0000282$ ，为十万分之2.8。按200万条数据/小时计算，每小时只需人工筛选56条数据，其中60.9%为真异常的数据，即34条数据为真异常。按每人每小时查50条计算，约需一个人，平均每人检出30条。

因此，在本案例中，我们采用了支持向量机和朴素贝叶斯的组合模型作为最终的分类模型算法。

4.4 算法的综合应用——肿瘤标志物的研究

这个案例是一个关于某肿瘤标志物的研究，这里我们对肿瘤标志物癌胚抗原进行研究。癌胚抗原（英文简称CEA）是我们常用的肿瘤标志性辅助指标，在临床试验、临床诊断方面广泛应用，但是在临床当中我们发现CEA与临床免疫学方面的因子存在一定联系，但目前尚未证实CEA在临床免疫学方面的确切意义。我们还发现CEA的异常与糖尿病有相关关系，并发现该肿瘤标志物的异常等级往往与一些临床症状存在相关的现象，另外，CEA是否还与其他实验室指标有相关性，也值得我们深入探讨。

在这个案例当中，我们应用了很多数据挖掘算法。包括在研究的样本设计与选取、临床特征主题分析、临床特征关联分析、相关因子分析、不同等级差异分析等各个环节，都应用了数据挖掘算法和统计学算法，并给我们带来以下好处：

- 观察组与对照组的设计与选择非常严谨和均衡，为科学的对照研究奠定了坚实基础。
- 主动式的探索发现和论证了科学规律，帮助科研者提出科研假说，并论证了假说。
- 帮助科研者大范围地研究多变量的复杂关系，多角度、多维度地探寻、论证变量之间的普遍联系和特异关系。

癌胚抗原作为重要的肿瘤标志性辅助指标，在临床试验和诊断方面已被广泛应用。近年，有研究表明癌胚抗原还与临床免疫学方面的因子存在一定联系。本研究旨在癌胚抗原与临床免疫学等关系方面做较深入的研究，以期揭示两者之间的关系，并探讨其免疫学的影响意义。

4.4.1 样本选取

科研样本的设计与选取在整个科研活动中至关重要。传统的样本选取方法和技术比较简陋，往往造成观察组与对照组的设计不严谨、不均衡，直接影响了科研的质量。本案例中，首次采用数据挖掘的一些方法辅助科研样本的设计与选取，大大提高了观察组与对照组的质量，使得科研结论的可信度大大提高。

4.4.1.1 观察组样本的选取

我们收集了2008年1月到2010年12月某大型三甲医院体检中心的体检数据。选取男性，且终检包含“癌胚抗原（CEA）增高”，同时将血脂、肌酐和尿酸化验的体检数据作为观察组（记为G1），共1391例。数据集采集项目包括所有体检的生化指标以及终检结果描述。

4.4.1.2 观察组样本的特征分析

年龄和性别是观察组临床特征的重要因素，它们往往与许多临床指标存在密切的联系。因此，对观察组样本的年龄和性别的分布特征分析具有重要意义。为了更科学地反映观察组样本的年龄和性别的客观分布，本研究采用数据挖掘的聚类算法（见第3章）加以分析。

本研究采用数据挖掘的聚类模型，就是按照观察组数据在年龄方面的相似性来找出数据的不同分簇，簇内的成员比簇间成员更具有相

似性，聚类分析的目的就是要使得簇内有足够的相似性和簇间具有足够的差异性。

由于观察组所有样本均为男性，我们采用Oracle公司的Oracle Data Mining (ODM) 工具（见第8章）提供的改进K-Means算法，只对观察组1391例样本的年龄变量数据进行运算，得到了一组高置信度的分类规则，本例得到了13个分组。为了更好地反映观察组样本在年龄上的主要分布特征，我们将13个分组合并为四个大组，具体如下：

- 24岁到41岁置信度为0.9462，支持度 (Support) 为88例。
- 41岁（不包括41）到58岁置信度为0.856，支持度为496例。
- 58岁（不包括58）到75岁置信度为1，支持度为243例。
- 75岁（不包括75）到86岁置信度为1，支持度为192例。

用统计学方法做描述统计，结果如下：

- 24岁到41岁有135例，均值和标准差为 35 ± 4.99 。
- 41岁（不包括41）到58岁有753例，均值和标准差为 51 ± 4.51 。
- 58岁（不包括58）到75岁有327例，均值和标准差为 67 ± 5.12 。
- 75岁（不包括75）到86岁有160例，均值和标准差为 80 ± 3.05 。

分析G1组中各个年龄组例数比例，如表4-20所示。

表4-20 各个年龄组例数比例

组别	例数	比例
24 岁到 41 岁	135	0. 097
41 岁（不包括 41）到 58 岁	753	0. 541
58 岁（不包括 58）到 75 岁	327	0. 235
75 岁（不包括 75）到 86 岁	160	0. 115

4.4.1.3 对照组样本的选取

我们使用聚类分析方法找到观察组G1中的年龄特征后，然后以相同特征（表4-19分组规则）选取对照组，记为G2。具体方法为，从与G1同一数据源的体检数据中选取数据，条件为男性，终检不包含“癌胚抗原（CEA）增高”，同时做了血脂、肌酐和尿酸化验，得到一个大样本集，记为G，约22000条数据。我们编写一个随机选取算法，设计从G组中随机选取样本例数与观察组相当的样本，约为2000左右的样本，记为G2组。具体做法是，找出观察组G1的每个年龄段中每个年龄所占比例，分别与对照组中每个年龄段例数相乘，得到对照组中每个年龄段中每个年龄的例数，然后从G中随机取出相同例数的记录，最后得到表4-21中第四列的例数。

表4-21 对照组中每个年龄段例数

组别	观察组例数	比例	对照组例数
24 岁到 41 岁	135	0. 097	194
41 岁（不包括 41）到 58 岁	753	0. 541	1082
58 岁（不包括 58）到 75 岁	327	0. 235	470
75 岁（不包括 75）到 86 岁	160	0. 115	230

G2组的选取方法如下：

1) 观察组G1的24岁到41岁年龄层设为A1，其中每个年龄段为 $a(i)$ ，例数为 $L(i)$ ，占比为 $q(i)$ 。例如：

$$a(1)=25, L(1)=3, q(1)=0.022;$$

$$a(2)=26, L(2)=6, q(2)=0.044;$$

.....，以此类推，直到最后一段，

$$a(17)=41, L(17)=23, q(17)=0.17。$$

2) 对照组G2的24岁到41岁年龄层需要194例，那么 $q(i) \times 194$ 可以求出对照组G2的24岁到41岁年龄层中每个年龄段的例数 $Q(i)$ 。
例如：

$$Q(1)=q(1) \times 194=0.022 \times 194=4,$$

$$Q(2)=q(2) \times 194=0.044 \times 194=9,$$

.....，以此类推，直到最后一段，

$$Q(17)=q(17) \times 194=0.17 \times 194=33。$$

从G中的年龄为25的体检者中，用随机选取算法随机取出4个，年龄为26的体检者中随机取出9个，以此类推，可以取出对照组G2的24岁到41岁年龄组B1，共195例。

3) A1和B1做均值比较。

对A1和B1做两独立样本秩和检验， H_0 ：两总体分布相同； H_1 ：两总体分布不相同，检验水准为0.05，结果为 $P=1$ ，大于0.05，所以在0.05水准下，A1和B1两样本的总体分布的差异无统计学意义。

4) 重复以上1) 至3) 步，得到另外三层年龄组的对照样本，如下：

- 对照组G2的41岁（不包括41）到58岁年龄组，共1083例， $P=0.961>0.05$ 。

- 对照组G2的58岁（不包括58）到75岁年龄组，共468例， $P=0.992>0.05$ 。

- 对照组G2的75岁（不包括75）到86岁年龄组，共230例， $P=0.929>0.05$ 。

加上G2组24岁到41岁层样本，实际总共得到对照组G2共1976例。其均值和标准差如下：

- 24岁到41岁有195例，均值和标准差为 35 ± 4.97 。

- 41岁（不包括41）到58岁有1083例，均值和标准差为 51 ± 4.5 。

- 58岁（不包括58）到75岁有468例，均值和标准差为 67 ± 5.13 。

- 75岁（不包括75）到86岁有230例，均值和标准差为 80 ± 3.03 。

4.4.1.4 对照组样本的特征验证

我们对观察组样本进行描述统计验证。对观察组G1的年龄做描述统计，结果如图4-19所示。

图4-19显示，观察组平均年龄为56岁，其中48到55岁例数最多，共330例，40岁到62岁之间有862例，占总例数的62%。观察组的频数分布直方图如图4-20所示。

然后，我们对对照组G2的年龄做描述统计，结果如图4-21所示。

中文名 ▲	英文名	统计量值
中位数	Median	54
变异系数	Cv	24.11%
均值	Mean	56.4076
方差	Variance	184.9654
最大值	Max	91
最小值	Min	18
极差	Range	73
标准差	Stddev	13.6002
百分之七十五分位数	Quartile_75	65
百分之九十五分位数	Quartile_95	81
百分之二十五分位数	Quartile_25	47
百分之五分位数	Quartile_05	36
下载列表		
1 - 12		

频数分布表5				
组段	频数	频率	累计频数	累计频率
18~	8	.01	8	.01
25~	42	.03	50	.04
33~	67	.05	117	.08
40~	259	.19	376	.27
48~	330	.24	706	.51
55~	273	.2	979	.7
62~	127	.09	1106	.8
70~	157	.11	1263	.91
77~	100	.07	1363	.98
85~	28	.02	1391	1
下载列表				
1 - 10				

图 4-19

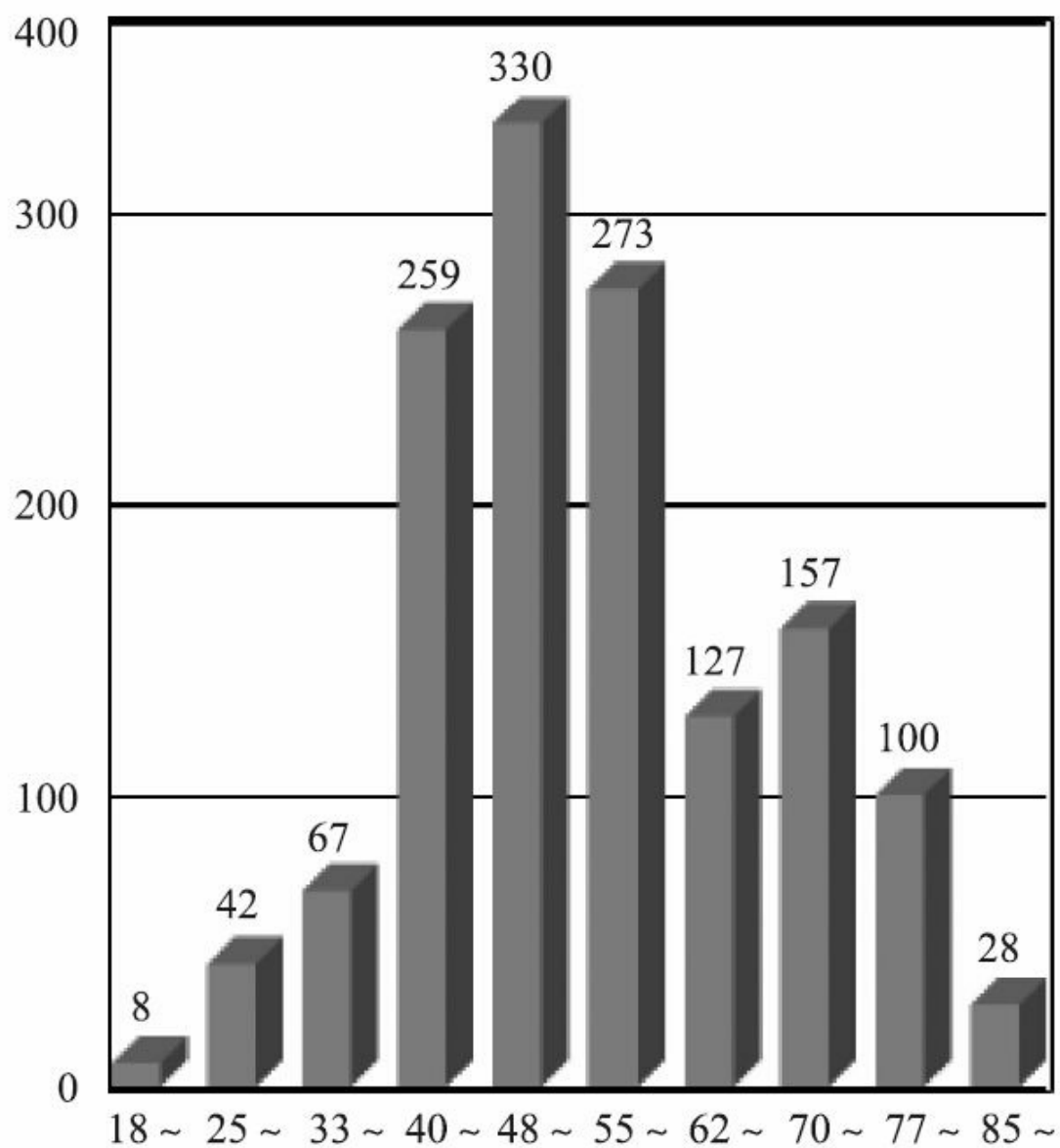


图 4-20

中文名 ▲	英文名	统计量值	组段	频数	频率	累计频数	累计频率
中位数	Median	54	25~	56	.03	56	.03
变异系数	Cv	23.44%	31~	61	.03	117	.06
均值	Mean	56.2454	37~	134	.07	251	.13
方差	Variance	173.8866	44~	414	.21	665	.34
最大值	Max	86	50~	413	.21	1078	.55
最小值	Min	25	56~	323	.16	1401	.71
极差	Range	61	62~	153	.08	1554	.79
标准差	Stddev	13.1866	68~	163	.08	1717	.87
百分之七十五分位数	Quartile_75	65	75~	167	.08	1884	.95
百分之九十五分位数	Quartile_95	80	81~	92	.05	1976	1
百分之二十五分位数	Quartile_25	47					
百分之五分位数	Quartile_05	36					

图 4-21

图4-21显示，对照组平均年龄为56岁，其中44岁到50岁例数最多，共414例；44岁到62岁之间有1150例，占总例数的58%。对照组的频数分布直方图如图4-22所示。

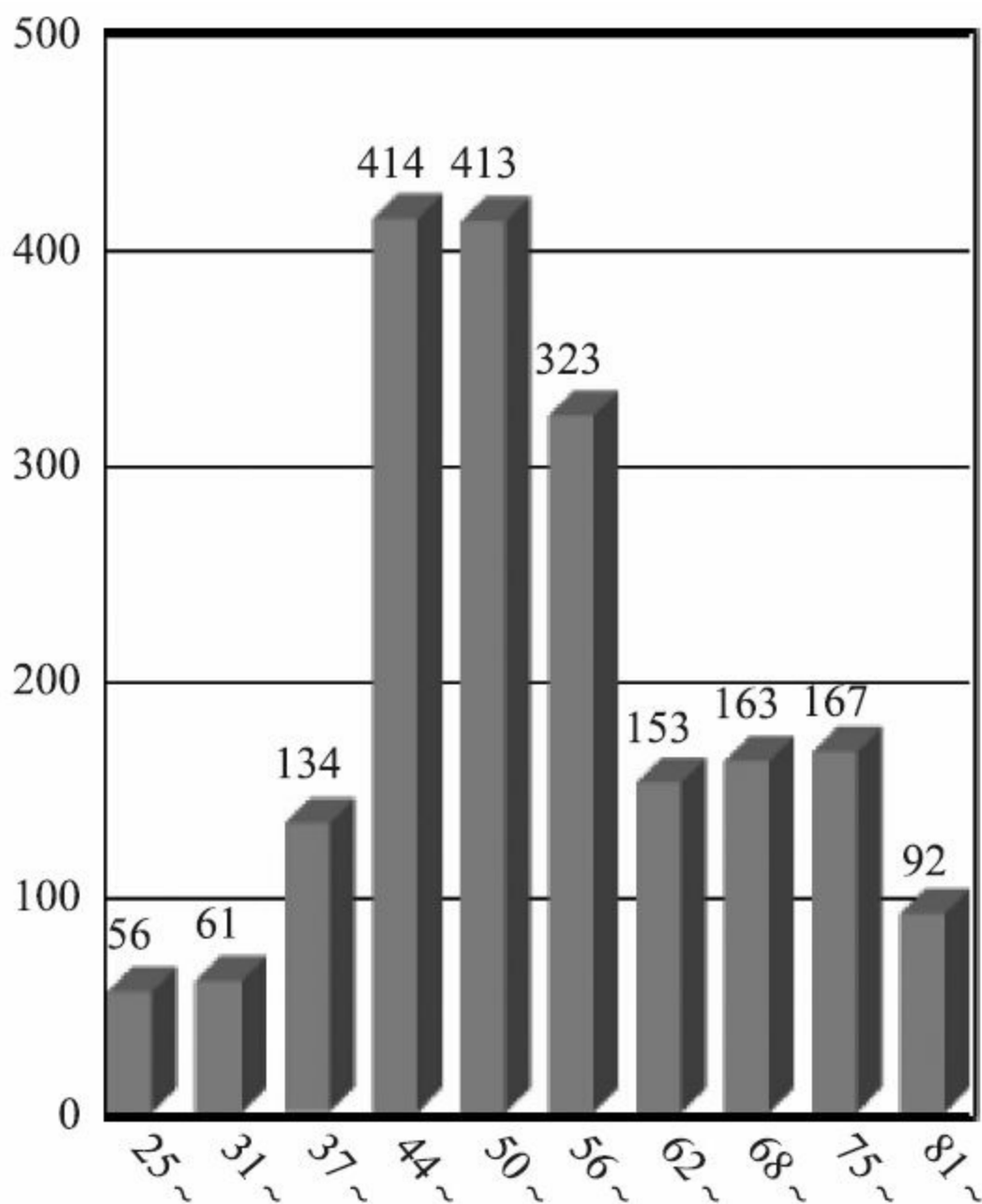


图 4-22

从图4-19至图4-22可以看出，两组的平均年龄都是56岁；观察组G1的标准差为13.6，对照组G2的标准差为13.2；变异系数几乎相同；40到62岁之间的例数一个为62%，一个为58%，很接近。可见两组分布几乎相同。

最后，我们对观察组G1和对照组G2做两独立样本秩和检验，结果 $P=0.774$ ，大于0.05，接受 H_0 ，即在0.05水准下，观察组G1和对照组G2的总体分布相同。

4.4.2 癌胚抗原临床特征主题分析

上一节，我们采用数据挖掘的聚类技术辅助开展科研样本的设计与选取，大大提高了观察组与对照组之间的可比性。本节中，我们将运用特征抽取技术开展癌胚抗原临床特征的主题分析。

4.4.2.1 应用特征抽取模型

特征抽取（Feature Extraction）模型（见第3章），常应用于潜在语义分析、数据压缩、数据分解、映射和模式识别等。这里，我们应用特征抽取模型分析本研究中观察组G1和对照组G2的终检结果的主题特征，或者也可以说应用降维的方法将终检描述变量进行维度压缩，以获取临床症状合并模式。

4.4.2.2 数据准备

体检终检结果数据表的结构如表4-22所示。

表4-22 体检终检结果表结构

Customer_id	Check_id	Gender	Age	Check_date	Final_Result
体检者 ID	本次体检 ID	性别	年龄	体检日期	终检结果

例如这样一条数据，如表4-23所示。

表4-23 体检终检结果表

Customer_id	Check_id	Gender	Age	Check_date	Final_Result
12345	0001	Female	47	2001-12-12	“高血压病 球蛋白增高 右侧卵巢囊肿 血流变异常 子宫肌瘤 骨质疏松 腰臀比值异常 超重 丙氨酸氨基转移酶 ALT 增高 血脂异常”

我们将以上数据变换为交易型的多记录数据结构，如表4-24所示。

表4-24 交易型多记录数据结构表

Row_id	Customer__check_id	Final_Result
1	123450001	高血压病
2	123450001	球蛋白增高
3	123450001	右侧卵巢囊肿
4	123450001	血流变异常
5	123450001	子宫肌瘤
6	123450001	骨质疏松
7	123450001	骨质疏松
8	123450001	腰臀比值异常
9	123450001	超重
10	123450001	丙氨酸氨基转移酶 ALT 增高
11	123450001	血脂异常

更详细的技术细节，请参见第6章。

4.4.2.3 特征抽取算法

我们采用ODM工具提供的非负矩阵因子分解（Non-negative Matrix Factorization）NMF算法（见第3章）对观察组1391例样本进行运算。

非负矩阵因子分解NMF算法是一种特征抽取算法，它常常用于属性很多而且属性模糊或者有很弱的可预测性的数据集。通过合并属

性，NMF算法可以产生有意义的模式或主题。NMF算法通过创建用户定义的特征数来压缩多元变量数据，每一个特征是原来属性集的一个线性合并，这些线性合并的系数是非负的。NMF算法压缩一个数据矩阵V为两个低阶矩阵W和H的乘积，以近似地等于 $W \times H$ 。NMF算法使用交互式过程来修改矩阵W和H的初始值，以便这个乘积接近矩阵V。这个交互式过程当近似值差聚集或者达到一个给定的迭代次数时终止。

在NMF模型应用的时候，该模型将原始数据映射到一个由模型发现的新的属性集上。在本研究中，我们将每一个单独终检描述项视为一个特殊变量，它的取值为Exist或Null。我们将原始数据变换为表4-24后，基于这个变换表创建了一个NMF模型，将大量的终检描述变量进行合并压缩，并映射到新的特征集中，从而找到观察组G1和对照组G2的终检结果的主题特征。

4.4.2.4 观察组模型结果

用特征抽取方法对观察组G1（癌胚抗原增高）进行分析，得到排序在前7位的特征，其特性总参数和特性序号如表4-25所示。

表4-25 特征总参数和特征序号

特征 ID	特征系数	特征排序
7	.076417447455388954868312	1
31	.074994735623321114921267	2
21	.07341831945569003497124	3
4	.073405033057366836630199	4
11	.07118816913693513109347	5
13	.068658949765274956256723	6
28	.067246622504113313470008	7

特征7中包括8个并发症：白细胞分类异常、血脂异常、腰臀比值异常、冠心病、空腹血糖高、高血压病、体重指数偏低和双眼视网膜动脉硬化Ⅱ。其中白细胞分类（免疫系统指标）异常占主要因素，说明癌胚抗原增高与免疫系统异常同时出现的概率很高，如图4-23所示。



图 4-23

表4-26中，该特征提示出非常强的主题，即CEA增高组有明显的免疫方面的反应。经对CEA增高组的生化指标进行全面分析，发现淋巴细胞分布异常，95%的案例都有淋巴细胞增高现象，说明CEA与免疫系统因子有较密切联系。

表4-26 特征属性

特征 ID	属性 - 分变量	属性系数
7	白细胞分类异常	.0719578859343747
7	血脂异常	.00162243349064348
7	腰臀比值异常	.00158936332587202
7	冠心病	.00105056285944897

我们再来观察特征抽取模型得到的另外一个特征20，如图4-24所示。

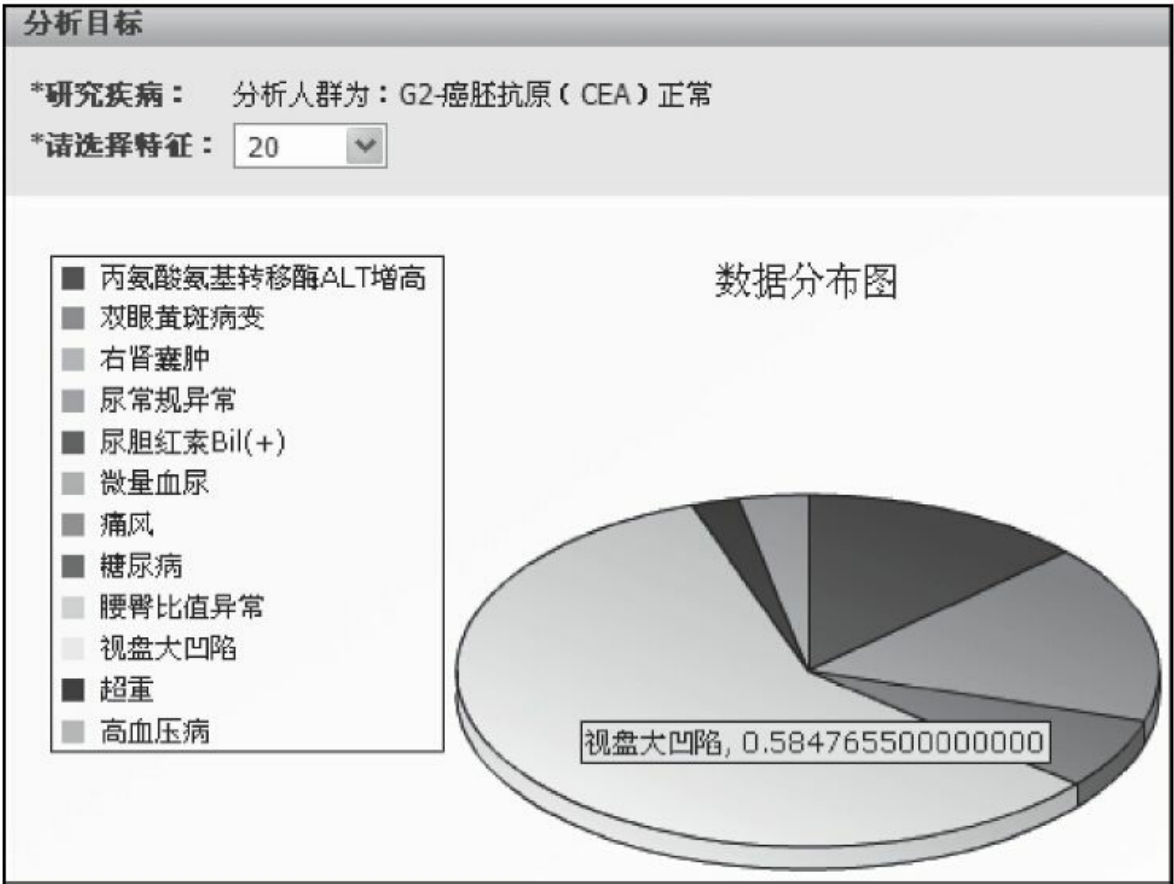


图 4-24

特征20包括六个并发症：视盘大凹陷、尿常规异常、丙氨酸氨基转移酶ALT增高、糖尿病、高血压病、超重。其中视盘大凹陷占主要因素，其次是尿常规异常、ALT增高和糖尿病，这些症状都反映了与糖尿病有关的现象，说明癌胚抗原增高与糖尿病有一定的关联，如表4-27所示。

表4-27 特征属性

特征 ID	属性 - 分变量	属性系数
20	视盘大凹陷	.0436013879598403
20	尿常规异常	.0122705874336597
20	丙氨酸氨基转移酶 ALT 增高	.00977715137348934
20	糖尿病	.00488969149263277
20	高血压病	.00236796108889889
20	超重	.00163286137986745

以上两个有关CEA增高组的临床主题抽取特征，反映了两个完全不同的临床特征，一个是与免疫学有关，一个是与糖尿病有关，从两个不同的系统揭示了CEA增高组的临床特征。当然，有研究表明，糖尿病与免疫学有关也有一定的联系，这样也间接印证了特征抽取模型的特征提取效果。

4.4.3 癌胚抗原临床特征规则分析

上一节，我们运用特征抽取技术开展癌胚抗原临床特征的主题分析。本节，我们将应用数据挖掘的另外一个重要技术——关联规则模型，进一步探索癌胚抗原临床特征规则。

4.4.3.1 应用关联规则模型

关联规则（Association）模型（见第3章），用于发现数据集中同时发生事件的概率，同时发生事件之间的关系可以表达为关联规则。关联规则的结果就是产生一些规则用来辨别数据中的关联模式。关联模型的应用也称作货篮分析，它在市场分析、产品促销和发现商业趋势等方面是很有价值的。这里，我们应用关联规则模型分析本研究中观察组G1和对照组G2的终检结果伴随特征。

1. 终检描述项集

关联规则分析的第一步是找出项集。一个项集就是在一个交易中找出两个或两个以上的项的组合，最大的项数是由用户指定的。本研究中，我们在一个人次的终检结果描述中，找出两个或两个以上的终检描述项的组合，并且给定最大的项数为三，即一个终检描述项集是两个终检描述项或三个终检描述项的组合。

2. 终检描述频繁集

关联规则是由项集推导出来的。如果从所有可能的项集中产生规则，可能会得到非常多的规则，而且很多规则未必有意义，建立模型也需要花很长的时间。通常我们希望仅得到一些在数据中具有代表性的规则。频繁集就是由用户指定最小频度的项集。最小频度项集支持度是一个用户定制的百分比，它限制了关联规则项集的数目。一个项集必须出现所有交易中指定百分比及以上的才被采用为关联规则。本研究中观察组G1和对照组G2的给定置信度为 $\geq 70\%$ ，支持度为 $\geq 10\%$ 。

数据准备同第4.4.2.2节。

4.4.3.2 关联规则算法

在关联规则算法中，最小支持度和置信度是影响关联规则建模的两个重要指标，支持度和置信度也是评估模型规则质量的。支持度是指示在一个规则中项集共同出现的频度的指标，支持度是一个在前件和后件中包含的所有项集的交易数与所有交易数的比率。置信度是指示在一个规则中项集共同出现的概率的指标，置信度是在给定前件的条件下同时出现后件的条件概率。

支持度和置信度用于确定规则是否有效。然而，在某些情况，即使支持度和置信度两个指标都比较高，但产生的规则依然可能是没有用的。Lift指标给出了评估规则质量的一个新指标。Lift指示给定的前件和后件随机同时出现的规则的强度，它提供了一个改进的信息，以

提高给定前件下后件出现的概率。任何Lift<1的规则都不能显示是一个真正的内在并发现象，无论它的支持度和置信度有多高，因为它不能比随机现象更能预测前件和后件的并发出现。

4.4.3.3 观察组模型结果

我们应用ODM提供的关联规则算法分析本研究中观察组G1和对照组G2的终检结果伴随特征。首先，我们应用关联规则算法分析了观察组G1的终检结果数据，然后将得到的关联规则模型按照后件、支持度、置信度优先顺序进行降序排列，得到如表4-28所示的关联规则结果集。

表4-28 规则表

规则序号	前件	后件	置信度	支持度
55	超重	腰臀比值异常	0. 7013	0. 3047
61	高血压病	腰臀比值异常	0. 7407	0. 2456
122	超重和血脂异常	腰臀比值异常	0. 7833	0. 2055
131	高血压病和血脂异常	腰臀比值异常	0. 824	0. 1603
46	肥胖	腰臀比值异常	0. 8601	0. 1523
57	轻度脂肪肝	腰臀比值异常	0. 804	0. 1166
140	高血压病和超重	腰臀比值异常	0. 8051	0. 1144
88	血脂异常和空腹血糖高	腰臀比值异常	0. 7525	0. 1108
33	糖尿病	腰臀比值异常	0. 7053	0. 1064
112	血脂异常和肥胖	腰臀比值异常	0. 8512	0. 1042
128	高尿酸血症和血脂异常	腰臀比值异常	0. 7114	0. 1042
27	空腹血糖高	血脂异常	0. 7138	0. 1472
87	腰臀比值异常和空腹血糖高	血脂异常	0. 7755	0. 1108
127	高尿酸血症和腰臀比值异常	血脂异常	0. 7079	0. 1042
65	轻度脂肪肝	血脂异常	0. 7035	0. 102

如表4-28所示，在观察组中如果有超重的现象，则同时出现腰臀比值异常的概率比较大，达到了70.13%，并且在表中所有后件为腰臀

比值异常的规则中，本规则出现的频率最高，达到30.47%。同时，在观察组中如果有血脂异常的现象，那同时出现空腹血糖高的概率比较大，达到了71.38%，并且在表中所有后件为血脂异常的规则中，本规则出现的频率最高，达到14.72%。

我们再将关联规则模型的结果重新排序，按照后件、置信度、支持度的优先顺序进行降序排列，得到如表4-29所示的结果。

表4-29 规则表

规则序号	前件	后件	置信度	支持度
46	肥胖	腰臀比值异常	0. 8601	0. 1523
112	血脂异常和肥胖	腰臀比值异常	0. 8512	0. 1042
131	高血压病和血脂异常	腰臀比值异常	0. 824	0. 1603
140	高血压病和超重	腰臀比值异常	0. 8051	0. 1144
57	轻度脂肪肝	腰臀比值异常	0. 804	0. 1166
122	超重和血脂异常	腰臀比值异常	0. 7833	0. 2055
88	血脂异常和空腹血糖高	腰臀比值异常	0. 7525	0. 1108

(续)

规则序号	前件	后件	置信度	支持度
61	高血压病	腰臀比值异常	0. 7407	0. 2456
128	高尿酸血症和血脂异常	腰臀比值异常	0. 7114	0. 1042
33	糖尿病	腰臀比值异常	0. 7053	0. 1064
55	超重	腰臀比值异常	0. 7013	0. 3047
87	腰臀比值异常和空腹血糖高	血脂异常	0. 7755	0. 1108
27	空腹血糖高	血脂异常	0. 7138	0. 1472
127	高尿酸血症和腰臀比值异常	血脂异常	0. 7079	0. 1042
65	轻度脂肪肝	血脂异常	0. 7035	0. 102

如表4-29所示，在观察组中如果有肥胖的现象，则同时出现腰臀比值异常的概率在表中所有后件为腰臀比值异常的规则中最高，达到80.4%，换句话说只要肥胖，那么腰臀比值就基本异常。同时，在观察组中如果有腰臀比值异常和空腹血糖高的现象，则同时出现血脂异

常的概率比较大，达到了77.55%，在表中所有后件为血脂异常的规则中最高。

4.4.3.4 观察组模型的Lift指标

我们再将Lift指标添加到关联规则模型的结果中，并按照后件、Lift的优先顺序进行降序排列，得到如表4-30所示的关联规则表。

表4-30 规则表

规则序号	前件	后件	置信度	支持度	LIFT
46	肥胖	腰臀比值异常	0.8601	0.1523	1.466365993
112	血脂异常和肥胖	腰臀比值异常	0.8512	0.1042	1.451132897
131	高血压病和血脂异常	腰臀比值异常	0.824	0.1603	1.40465934
140	高血压病和超重	腰臀比值异常	0.8051	0.1144	1.372587892
57	轻度脂肪肝	腰臀比值异常	0.804	0.1166	1.370863535
122	超重和血脂异常	腰臀比值异常	0.7833	0.2055	1.335543616
88	血脂异常和空腹血糖高	腰臀比值异常	0.7525	0.1108	1.283326438
61	高血压病	腰臀比值异常	0.7407	0.2456	1.262890558
128	高尿酸血症和血脂异常	腰臀比值异常	0.7114	0.1042	1.212887198
33	糖尿病	腰臀比值异常	0.7053	0.1064	1.20259677
55	超重	腰臀比值异常	0.7013	0.3047	1.196120531
87	腰臀比值异常和空腹血糖高	血脂异常	0.7755	0.1108	1.368961296
27	空腹血糖高	血脂异常	0.7138	0.1472	1.259588988
127	高尿酸血症和腰臀比值异常	血脂异常	0.7079	0.1042	1.249176571
65	轻度脂肪肝	血脂异常	0.7035	0.102	1.241236609

如表4-30所示，在观察组中按Lift指标由高到底排序，排在前三的规则显示，如果有肥胖，或血脂异常和肥胖，或高血压病和血脂异常，则同时出现腰臀比值异常的概率比较大，达82%至86%，比率在10%至16%，而且是比较强的规则，它们的Lift值都超过了1.4。

4.4.3.5 对照组模型结果

同理，我们应用关联规则算法分析了观察组G2的终检结果数据，然后将得到的关联规则模型按照后件、支持度、置信度优先顺序进行降序排列，得到如表4-31所示的规则表。

表4-31 规则表

前件	后件	置信度	支持度
超重	腰臀比值异常	0.7234	0.3163
高血压病	腰臀比值异常	0.7654	0.251
超重和血脂异常	腰臀比值异常	0.7673	0.1969
肥胖	腰臀比值异常	0.8895	0.167
高血压病和血脂异常	腰臀比值异常	0.8234	0.1604
轻度脂肪肝	腰臀比值异常	0.7883	0.1301
空腹血糖高	腰臀比值异常	0.7159	0.125
高血压病和超重	腰臀比值异常	0.7682	0.1174
血脂异常和肥胖	腰臀比值异常	0.892	0.1129
高尿酸血症和血脂异常	腰臀比值异常	0.7251	0.1068
高尿酸血症和腰臀比值异常	血脂异常	0.7276	0.1068

如表4-31所示，在对照组中如果有超重的现象，则同时出现腰臀比值异常的概率比较大，达到了72.34%，并且在表中所有后件为腰臀比值异常的规则中，本规则出现的频率最高，达到31.63%。同时，在观察组中如果有血脂异常的现象，那同时出现高尿酸血症和腰臀比值异常的概率比较大，达到了72.76%，本规则出现的频率达到10.68%。

我们再将关联规则模型的结果重新排序，按照后件、置信度、支持度的优先顺序进行降序排列，得到如表4-32所示的规则表。

表4-32 规则表

前件	后件	置信度	支持度
血脂异常和肥胖	腰臀比值异常	0.892	0.1129
肥胖	腰臀比值异常	0.8895	0.167
高血压病和血脂异常	腰臀比值异常	0.8234	0.1604
轻度脂肪肝	腰臀比值异常	0.7883	0.1301
高血压病和超重	腰臀比值异常	0.7682	0.1174
超重和血脂异常	腰臀比值异常	0.7673	0.1969
高血压病	腰臀比值异常	0.7654	0.251
高尿酸血症和血脂异常	腰臀比值异常	0.7251	0.1068
超重	腰臀比值异常	0.7234	0.3163
空腹血糖高	腰臀比值异常	0.7159	0.125
高尿酸血症和腰臀比值异常	血脂异常	0.7276	0.1068

如表4-32所示，在对照组中如果有血脂异常和肥胖的现象，则同时出现腰臀比值异常的概率在表中所有后件为腰臀比值异常的规则中最高，达到89.2%，换句话说只要肥胖，那么腰臀比值就基本异常。

我们再将关联规则模型的结果重新排序，按照后件、Lift的优先顺序进行降序排列，得到如表4-33所示的规则表。

表4-33 规则表

前件	后件	置信度	支持度	LIFT
血脂异常和肥胖	腰臀比值异常	0.892	0.1129	1.435 917 363
肥胖	腰臀比值异常	0.8895	0.167	1.431 258 488
高血压病和血脂异常	腰臀比值异常	0.8234	0.1604	1.324 704 916
轻度脂肪肝	腰臀比值异常	0.7883	0.1301	1.268 922 827
高血压病和超重	腰臀比值异常	0.7682	0.1174	1.236 051 789
超重和血脂异常	腰臀比值异常	0.7673	0.1969	1.234 847 34
高血压病	腰臀比值异常	0.7654	0.251	1.231 612 901
高尿酸血症和血脂异常	腰臀比值异常	0.7251	0.1068	1.166 954 109
超重	腰臀比值异常	0.7234	0.3163	1.164 021 396
空腹血糖高	腰臀比值异常	0.7159	0.125	1.152 037 011
高尿酸血症和腰臀比值异常	血脂异常	0.7276	0.1068	1.293 130 479

如表4-33所示，在观察组中按Lift由高到底排序，排在前两个的

规则提示，如果有血脂异常和肥胖，或单纯肥胖，则同时出现腰臀比值异常的概率比较大，达89%，比率在11%至16%，而且是比较强的规则，它们的Lift值都超过了1.4。

4.4.4 癌胚抗原临床特征规则的比较分析

通过上节的案例分析，从观察组和对照组的关联规则模型中，我们发现临床特征规则总体上比较接近，在给定置信度为 $\geq 70\%$ ，支持度为 $\geq 10\%$ 的条件下，我们得到了有关观察组和对照组的很多关联提示。

观察组和对照组腰臀比值异常的强规则基本相似，主要包括：单纯肥胖，或血脂异常和肥胖，或高血压病和血脂异常等。两组支持度和Lift也相当，但置信度相差3%至7%，如表4-34所示。说明观察组中CEA增高可能引起一部分人的临床体征的特殊变化，有待进一步深入研究。

表4-34 关联规则模型分析

	观察组	对照组
置信度	82% ~ 86%	89%
支持度	10% ~ 16%	11% ~ 16%
LIFT	1.4 +	1.4 +

通过对上面关联规则模型的分析，我们发现观察组和对照组的临床特征规则总体上比较接近。我们按照置信度下限70%和支持度下限10%的条件筛选关联规则，发现：

- 1) 观察组的腰臀比值异常规则，主要前提要素包括：超重、高

血压、血管异常、肥胖、轻度脂肪肝、空腹血糖高、糖尿病、高尿酸血症等八个要素及其组合。

2) 观察组的血脂异常规则，主要前提要素包括：空腹血糖高、腰臀比值异常、高尿酸血症、轻度脂肪肝等四个要素及其组合。通过分析，我们发现腰臀比值异常与血脂异常互为前后件，说明二者密切相关。

3) 对照组的腰臀比值异常规则，主要前提要素包括：血脂异常、肥胖、高血压病、轻度脂肪肝、超重、高尿酸血症、空腹血糖高等七个要素及其组合。

4) 对照组的血脂异常规则，主要前提要素包括：高尿酸血症，腰臀比异常两个要素及其组合。

我们在腰臀比值异常规则中发现，观察组比对照组多了糖尿病，在血脂异常规则中，观察组比对照组多了空腹血糖高和轻度脂肪肝。也就是说，两组的差异主要是糖尿病及疑似糖尿病。

我们通过分析，进一步发现了下面的规律。观察组和对照组的腰臀比值异常的强规则基本一致，主要包括：单纯肥胖，或血脂异常和肥胖，或高血压病和血脂异常等。两组的支持度和Lift也相当，但是置信度相差3%~7%，说明观察组中的CEA增高可能引起一部分人的临床体征特殊的变化。这样，我们可以初步得出一个重要结论：CEA增高可能与糖尿病有关。

4.4.5 癌胚抗原相关因子分析

上一节，我们运用关联规则技术探索了癌胚抗原的临床特征规则。本节，我们将应用数据挖掘的又一个重要技术——特性选择模型，深入探索癌胚抗原的临床相关因子。

4.4.5.1 样本选取

首先是观察组样本的选取。我们收集了某大型三甲医院体检中心的2008年1月到2010年12月的体检数据，共47664例。数据集采集的内容包括所有体检的生化指标以及终检结果描述。

4.4.5.2 应用特性选择模型

下一步来看癌胚抗原相关因子的研究分析。在这里，我们用的数据挖掘方法是特性选择（Feature Selection）模型（见第3章）。特性选择也被称为因子权重模型，它是一个有监督学习的模型，给预测目标的每个属性因子的显著性打分。发现最有价值的预测因子，是很多数据挖掘项目的目标，权重评估和打分用来选取目标属性的预测属性。在这里，我们的目标属性是癌胚抗原。

我们应用ODM提供的特性选择算法MDL算法，以癌胚抗原作为目标属性，对所有与癌胚抗原同时检测的生化指标进行相关因子权重分析，得到图4-25所示的结果。

图4-25显示，与癌胚抗原相关性最高的是天冬氨酸氨基转移酶，其次是高密度脂蛋白、低密度脂蛋白。肌酐和尿酸对癌胚抗原相关度明显小于前3个指标，但又明显大于后面的所有指标。以上5个生理指标对CEA的影响显著性系数均大于0，说明它们与目标属性CEA成“实相关”的关系（见第7章）。以上特征的医学意义，我们不在此讨论了。下面，我们将应用统计学工具对MDL算法挖掘的结果进行评估和验证。

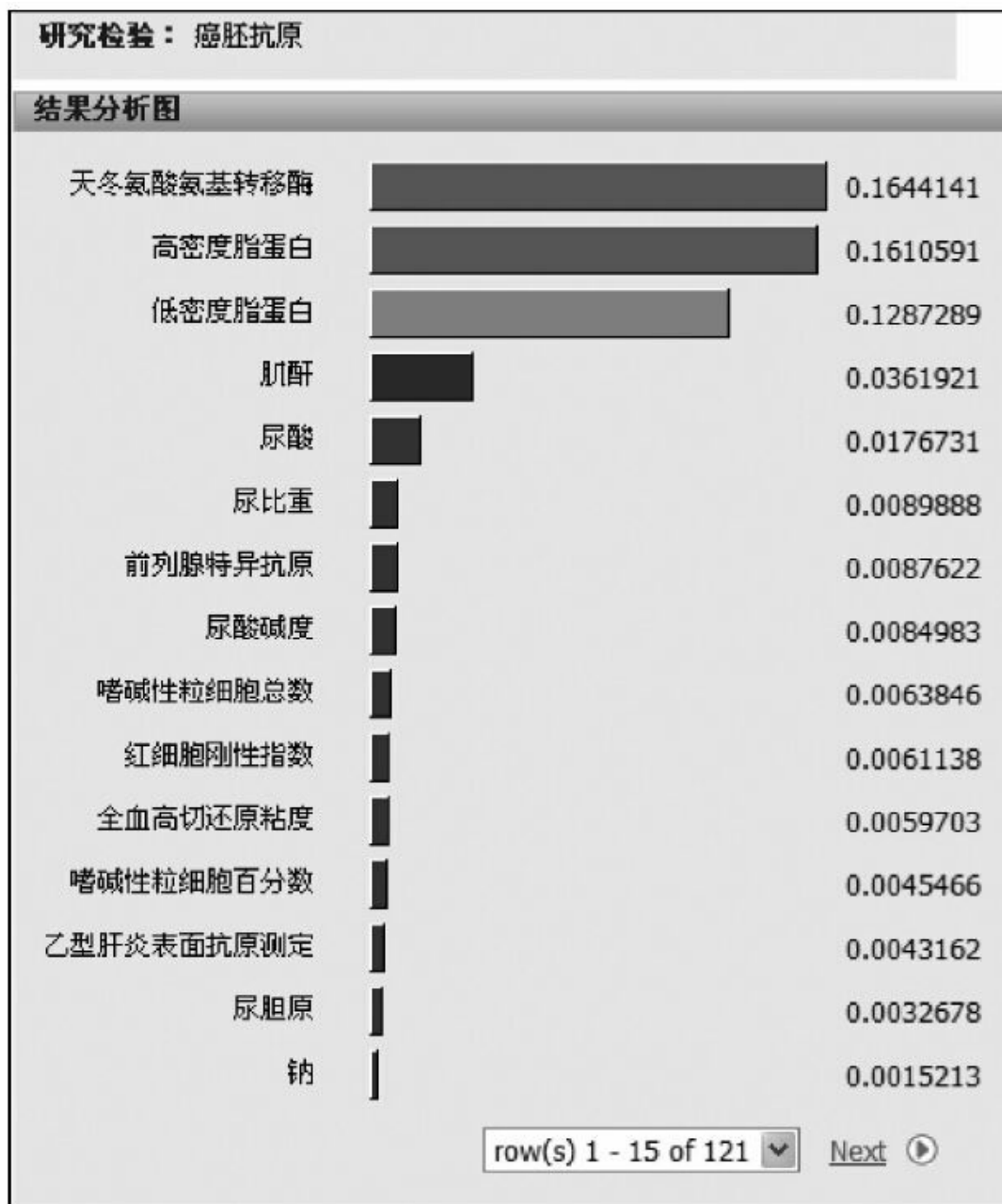


图 4-25

4.4.5.3 统计学验证分析

我们将应用统计学的相关性检验，来进一步验证CEA与通过特性

选择模型得到的具有“实相关”关系的生理指标的关系。

1.癌胚抗原与天冬氨酸氨基转移酶的相关性

首先，我们对天冬氨酸氨基转移酶与CEA进行二元正态分布验证。验证的结果显示天冬氨酸氨基转移酶与CEA不服从二元正态分布，因此，我们采用基于秩和检验的Spearman秩相关分析，结果如图4-26所示。

通过对天冬氨酸氨基转移酶与CEA的相关性秩和检验，得出的结论是Spearman秩相关系数：0.137，双侧概率 $p < 0.05$ ，具有统计学意义，说明天冬氨酸氨基转移酶与CEA成弱相关的关系。

2.癌胚抗原与高密度脂蛋白等的相关性

同理，我们对高密度脂蛋白与CEA进行二元正态分布验证。验证的结果显示高密度脂蛋白与CEA不服从二元正态分布，因此，我们采用基于秩和检验的Spearman秩相关分析，结果如图4-27所示。

*评估目标：癌胚抗原

*评估目标：天冬氨酸氨基转移酶

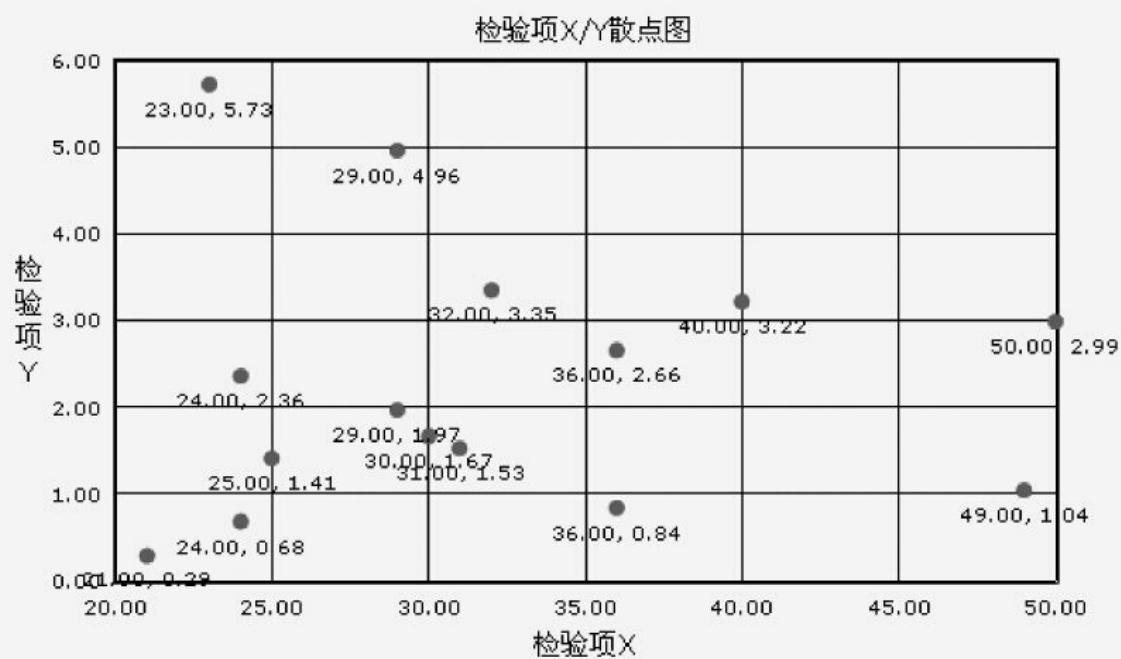
正态检验

正态验证检验结果(X检验) 0



正态验证检验结果(Y检验标准残差e) 0

样本数量 35168



Spearman秩相关分析: .1370099

双侧: $p < 0.05$

图 4-26

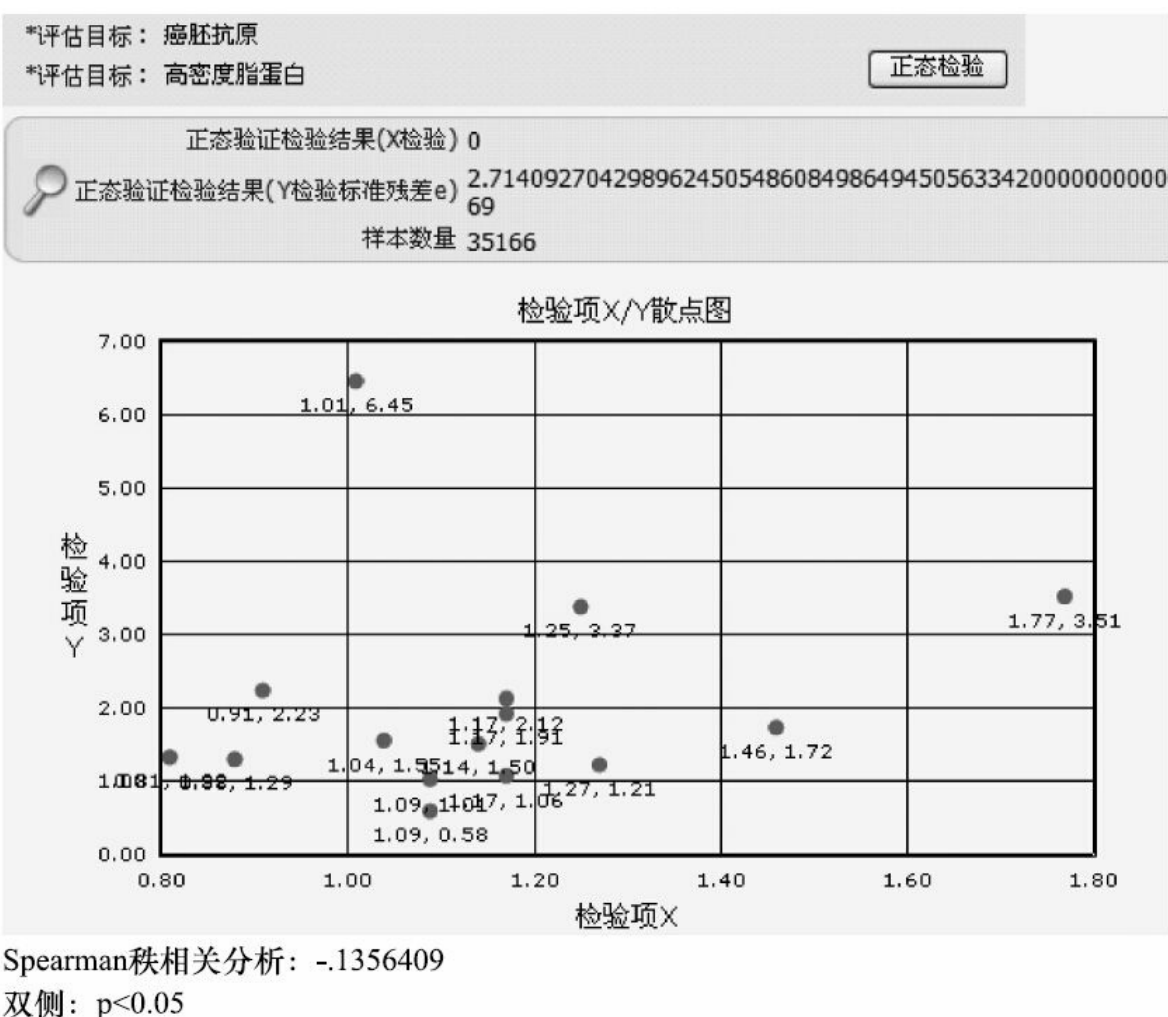


图 4-27

我们可以看到，癌胚抗原与高密度脂蛋白得到的是一个负相关的结果，系数为-0.135，双尾 $P < 0.05$ ，也就是说如果CEA增高的话，高密度脂蛋白就降低，反之，如果CEA降低的话，高密度脂蛋白就增高，但相关系数较低，成弱相关关系。

同理，我们可以计算出CEA和低密度脂蛋白的相关系数为0.0956， $P < 0.05$ ，CEA与肌酐的相关系数0.2573， $p < 0.05$ ，CEA与尿酸的相关系数0.2175， $P < 0.05$ 。这说明CEA与肌酐和尿酸有一定程度的相关性，且较CEA与其他指标的相关性要更强一些，而CEA与低密度

脂蛋白的相关性最弱。

3.癌胚抗原与肌酐、尿酸和天冬氨酸氨基转移酶的复相关性

我们还可以利用统计学中的复相关分析方法来验证CEA与肌酐、尿酸和天冬氨酸基转移酶多种指标组合的相关性，如图4-28所示。

The screenshot shows a software window for statistical analysis. It contains the following elements:

- *请选择分析类型:** A dropdown menu set to "复相关分析" (Multiple Correlation Analysis) with "(必选)" (Required) next to it.
- *观察项X1:** A dropdown menu set to "癌胚抗原" (Carcinoembryonic Antigen) with "(必选)" (Required) next to it.
- A search icon (magnifying glass) and two buttons labeled "Add" and "Remove" are positioned to the right of the "观察项X1" dropdown.
- *请选择控制检验项:** A list box containing three items: "尿酸" (Uric Acid), "肌酐" (Creatinine), and "天冬氨酸氨基转移酶" (Aspartate Aminotransferase). Below this list box is the label "(必选)" (Required).

图 4-28

复相关分析的结果：相关系数 $r=0.1302$ ，统计检验结果为： $p<0.05$ 。这样，我们可以看出CEA与肌酐、尿酸和天冬氨酸基转移酶多种指标组合呈现一种较弱的相关性。由于特性选择模型只是考察目标属性CEA与每个其他生理指标单独的影响关系，因此复相关分析可以补充目标属性CEA与某一组生理指标的组合关系。

4.4.6 不同等级癌胚抗原组差异分析

不同等级癌胚抗原是观察组临床特征的重要因素，对观察组样本的癌胚抗原的分布特征分析具有重要意义。为了更科学地反映观察组样本的癌胚抗原的客观分布，本研究采用数据挖掘的聚类算法加以分析。

4.4.6.1 应用聚类模型

本研究采用数据挖掘的聚类模型就是按照观察组数据在年龄方面的相似性来找出数据的不同分簇，簇内的成员比簇间成员更具有相似性，聚类分析的目的就是要使得簇间有足够的相似性和簇间有足够的差异性。

我们采用改进K-Means算法，对观察组47664例样本的癌胚抗原变量数据进行运算，得到了一组高置信度的分类规则，并得到13个分组。为了更好地反映观察组样本在癌胚抗原的主要分布特征，我们将13个分组间的相似度特点合并为3个大组，具体如下：

第1组：CEA<5.1

第2组：5.1≤CEA<8.1

第3组：8.1≤CEA

然后我们进一步考察这3个大组间的生理指标的特征差异，如组

间天冬氨酸氨基转移酶差异和组间密度脂蛋白差异。

4.4.6.2 天冬氨酸氨基转移酶差异分析

我们首先进行的是不同等级癌胚抗原组的天冬氨酸氨基转移酶差异分析。

1.天冬氨酸氨基转移酶的描述性统计结果

运用描述性统计分析方法，对全部3个大组的天冬氨酸氨基转移酶数据进行分析，结果如图4-29所示。

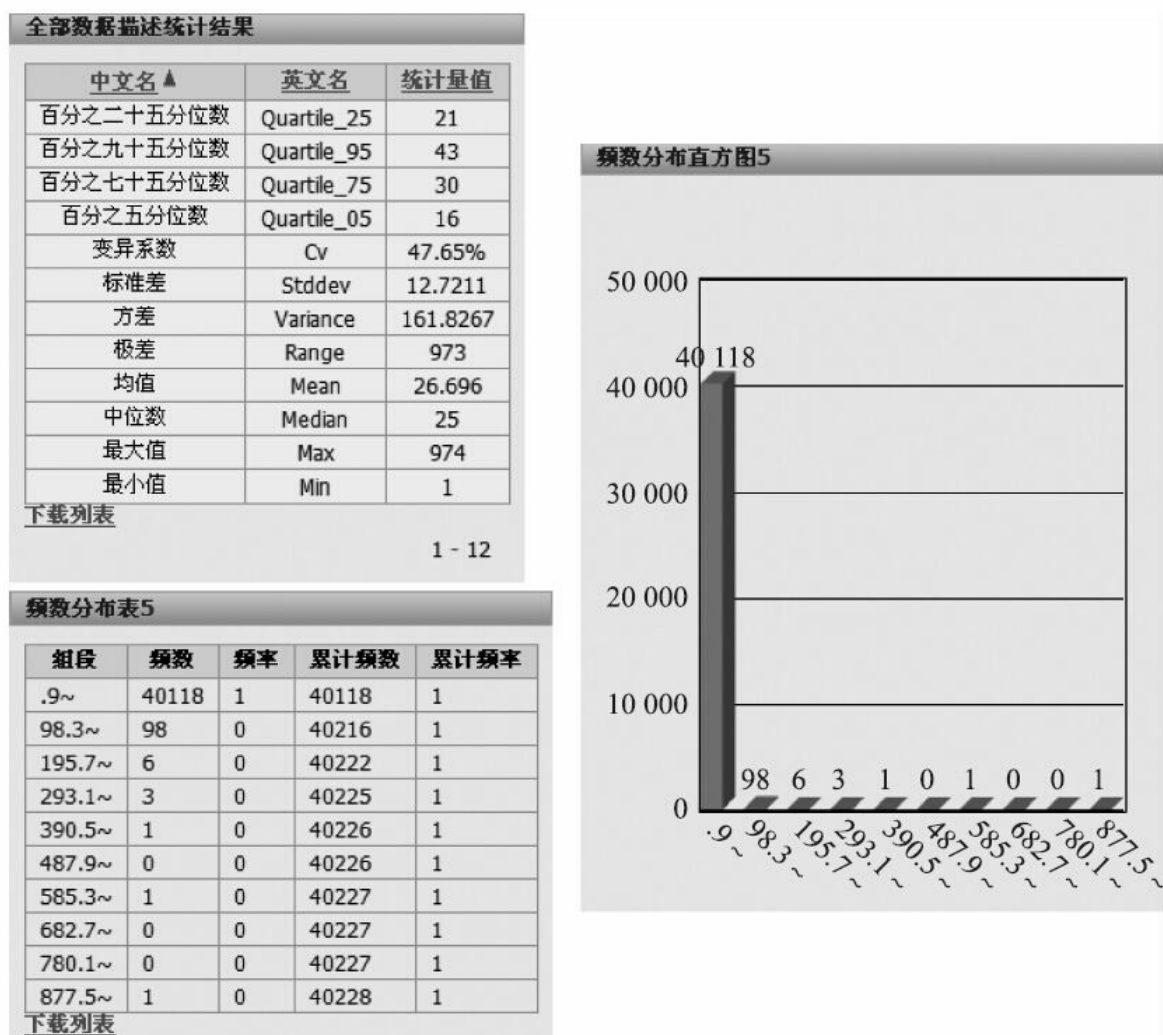


图 4-29

图4-29显示，天冬氨酸氨基转移酶数据存在明显的离群点，分布在几个不同的区段内，范围包括195.7以上到877.5以上之间，总共有12个离群点，而40118个点集中在0.9到98.3之间，98个点分布在98.3至195.7之间，我们把这98个点成为亚离群点。我们给出这样一个图，可能看不到绝大多数点的实际分布，当然我们可以再给出一个只针对40118个点的描述性统计图表。但这里，是因为医学家可能对离群点或亚离群点更感兴趣，他们的兴趣点是天冬氨酸氨基转移酶的离群点是怎样分布的，它们代表了什么临床现象，原因是什么等问题。

2.三组天冬氨酸氨基转移酶的秩和分析

接下来，我们要对按照癌胚抗原划分的3个大组间天冬氨酸氨基转移酶进行差异性分析。由于3组数据的正态分布验证和方差齐性验证结果未通过检验，因此，我们直接应用基于非参数检验方法的多个独立样本的秩和分析。分析结果如表4-35所示。

表4-35 天冬氨酸氨基转移酶的秩和分析

	第 1 组与第 2 组	第 1 组与第 3 组	第 2 组与第 3 组
天冬氨酸氨基转移酶	uc = 7. 7663, p < 0. 01	uc = 3. 8657, p < 0. 01	uc = . 0871, P > 0. 05

我们看到，第1组分别与第2组和第3组天冬氨酸氨基转移酶都有统计学差异，但第2组与第3组天冬氨酸氨基转移酶尚不能认为有统计学差异。说明癌胚抗原以5.1为界，天冬氨酸氨基转移酶有明显的不同，换句话说癌胚抗原与天冬氨酸氨基转移酶有影响关系是体现在癌胚抗原以5.1为界的现象下。进一步地，第2组与第3组天冬氨酸氨基转移酶尚不能认为有统计学差异，并不代表癌胚抗原5.1以上的天冬氨酸氨基转移酶没有分层，只是尚不能用癌胚抗原以8.1为界来分层，我们可以进一步对该段进行分层分析，并找到癌胚抗原相应的分层特点。此处，不再做分析了。

4.4.6.3 高密度脂蛋白差异分析

接下来，我们进行的是不同等级癌胚抗原组的高密度脂蛋白差异分析。

1. 高密度脂蛋白的描述性统计分析

运用描述性统计分析方法，对全部3个大组的高密度脂蛋白数据进行分析，结果如图4-30所示。

从图4-30中我们可以看到，高密度脂蛋白粗略地呈现出一种偏态分布的趋势，但需要进一步验证。与天冬氨酸氨基转移酶相比，高密度脂蛋白的变异系数要小得多，为23.89%。这些数据能反映什么问题，医学专家可能会比我们更清楚。

2. 三组高密度脂蛋白的秩和分析

接下来，我们同样要对按照癌胚抗原划分的3个大组间高密度脂蛋白进行差异性分析。基于与天冬氨酸氨基转移酶同样的原因，我们直接应用基于非参数检验方法的多个独立样本的秩和分析，分析结果如表4-36所示。

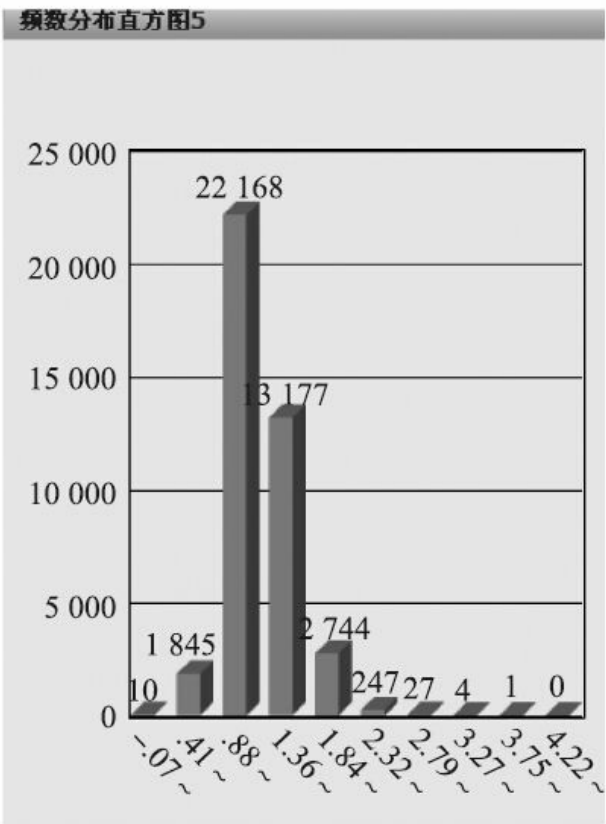


图 4-30

表4-36 高密度脂蛋白的秩和分析

	第 1 组与第 2 组	第 1 组与第 3 组	第 2 组与第 3 组
高密度脂蛋白	uc = 6. 0597, p < 0. 01	uc = 5. 348, p < 0. 01	uc = 2. 1068, p < 0. 05

我们看到，与天冬氨酸氨基转移酶不同，三组高密度脂蛋白两两间都有统计学差异，说明癌胚抗原的三个分组对高密度脂蛋白很敏感，换句话说癌胚抗原与高密度脂蛋白有明显的影响关系，并从这样的分层角度上说，这种影响关系比癌胚抗原与天冬氨酸氨基转移酶的

影响关系要强。当然，究竟哪两对之间的影响关系更强，需要对全体数据的相关性进行分析，而不是这样的三段数据分析。

通过以上的过程分析，我们看到应用多种分析方法可以互相提示、相互印证，可以多角度、深层次地分析大数据中的隐含规律。当然，我们会发现，随着运用的方法不断丰富和深入，我们的视野也在不断扩大。在发现越来越多的现象和规律的同时，我们也发现了越来越多的问题，也就是说，已知越多，未知也就越多，但我们可能就越接近真理，也有可能为人类做出越来越多的贡献。

4.5 数据挖掘在其他领域中的应用

数据挖掘大规模应用于商业项目大约有20余年。从应用范围来讲，已经深入到各行各业中。例如，数据挖掘应用于金融行业可以智能辅助银行的风险控制工作，如卡类风险控制系统、支票风险控制系统、信贷风险控制系统等。其中信用卡、借记卡防欺诈系统就是卡类风险控制系统的一种。随着银行借记卡、信用卡业务的推广，银行卡防欺诈的工作被提到管理层的议事日程。通过建立有效的防信用欺诈系统，例如，使用潜在边界分配模型描述欺诈客户特性、采用自举模型和逻辑回归模型建立防欺诈模型、设计贝叶斯模型辨别欺骗使用信用额度、利用连锁分析建立商业欺诈侦测模型等，从而及时侦测到欺诈行为，及时联系顾客，维护顾客的权益，保证客户、银行的资金安全，提高顾客对银行的信任度，增加银行卡类业务的收入。如图4-31所示。

信用卡借记卡防欺诈系统系统架构

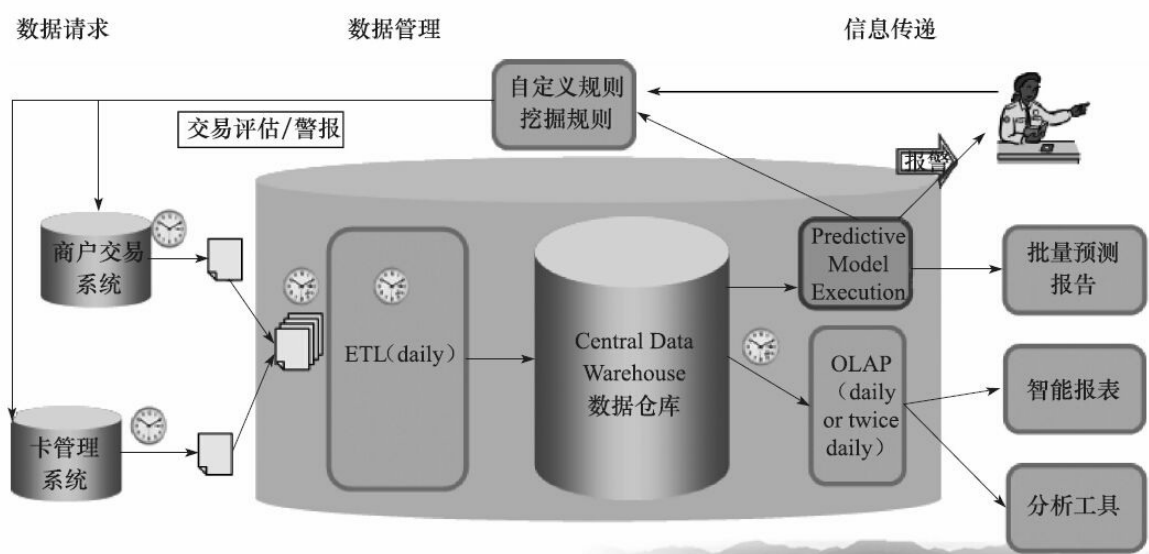


图 4-31

再例如，支票风险控制系统是数据挖掘在金融领域的又一应用。Check Kiting（挪用支票周转，用空头支票骗取钱财），即利用支票进行恶意透支是西方国家常见的一种金融诈骗行为，银行每年因此损失巨大。现在，各大银行除了加强制度管理外，还纷纷采用数据挖掘的方法加大风险控制力度，从而收到了显著的效果，现在预测模型已经成为北美各银行不可缺少的支票风险控制手段。在中国，个人支票的广泛使用是必然的趋势，相应的风险控制必然也随之而来，预测模型必将成为我们的一个有力工具。如图4-32所示。

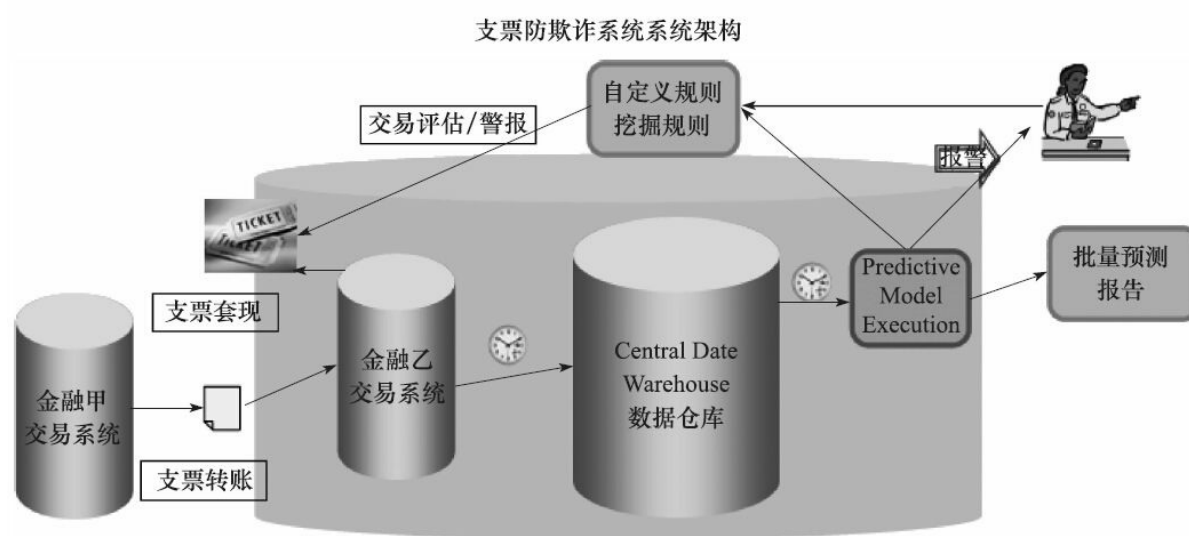


图 4-32

信贷防欺诈系统是银行业采用的另一个风险控制工具。随着市场经济的进一步发展和完善，信贷市场必将有一个更大的飞跃。中国信贷市场已初具规模，诸如房屋贷款、汽车贷款已经越来越广泛，随之而来的各种信贷欺诈也不乏少见。与发达国家相比，中国信贷管理方

法和使用的管理系统还相对滞后。建立信贷防欺诈系统，利用世界最先进的管理方法和技术，为防止信贷防欺诈提供了新的手段，从而降低银行信贷风险。如图4-33所示。

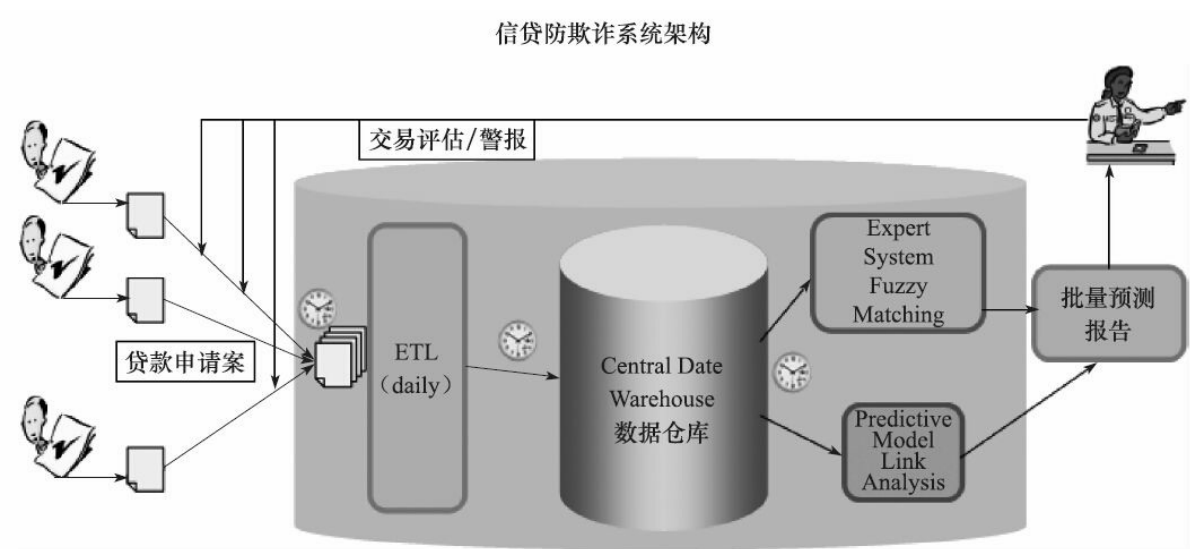


图 4-33

房屋信贷欺诈侦测就是信贷防欺诈系统中的一个具体应用。在西方国家，房屋贷款欺诈涉及很广，涉案者往往伪造申请者信息或房产信息，包括故意抬高他们的收入和借贷，以便总能获得贷款；或者隐瞒购买意图，如将转卖改为自住、改变贷款条件以获得更好的贷款利率，更有甚者，通过倒卖房产洗钱。房屋贷款欺诈的目标就是找到申请案的不一致点，并在看似无关的申请贷款案中分析寻找潜在的相关性，以找出贷款欺诈案。表4-37是房屋信贷欺诈侦测模型所提示的专家系统规则。

表4-37 房屋信贷欺诈侦测模型所提示的专家系统规则

序号	业务规则：满足以下条件标识为异常	序号	业务规则：满足以下条件标识为异常
1	申请人已经有过不诚信记录	4	超过 3 人购买同一房产
2	房产报价高于参考估价的 50% 以上	5	申请人的工资信息与实际不符
3	居住房产距离工作地点超过 100 英里		

数据挖掘还广泛应用于电信行业。如电信网络智能管理系统，包括通信网络性能评估、网络通信故障预测、网络通信故障预警、部分网络通信故障自动预处理等功能。

通信网络管理预测主要应用了数据挖掘的特性选择、回归预测和分类模型。预测网络管理是有效的网络管理策略的重要组成部分，预测模型方法是实现这一目标的主要工具。使用这些预测模型，可以用预测因子建立对某些特定事件发生的风险级别。例如：

- 每日信息包准确率模型
- 每日信息传输延迟模型
- 每月信息传输延迟模型

这些模型在实际应用中取得了很好的模型收益，例如：

- 降低了违反质量标准协议的风险。
- 提高了网络稳定性。
- 可更好地侦测网络异常。
- 可更好地侦测网络性能下降。
- 降低了网络崩溃的概率。

通过数据挖掘模型提供的预警机制，工作人员可以及时地采取故障纠正措施。例如：

- 检查网络设备软硬件配置，纠正或替换有故障的软硬件。
- 检查全部网络设计，帮助网络工程团队查找问题。
- 验证第三方供应商的网络设备，并纠正或替换不兼容的软硬件。
- 查验不充足的客户资源分配，包括软件、硬件和服务，以便对这些资源进行优化、升级。

这样，从整体质量上大大提高了通信网络的可靠性，从而进一步提高了通信服务质量。

数据挖掘在其他行业的应用也很多，包括传统行业和互联网行业，市场上也出了不少的相关书籍，读者可以参考。

4.6 本章小结

随着大数据的不断发展，各行各业商业数据挖掘的需求不断涌现，数据挖掘技术将越来越多地应用于商业项目中。这需要我们一方面不断地进行数据挖掘应用的各种创新，另一方面应不断总结和积累成功的数据挖掘商业应用经验。实践是检验真理的唯一标准，应用是检验数据挖掘技术的唯一标准。只有不断地将数据挖掘技术应用于实践中，不断地验证、改进、创新和再应用，数据挖掘才能真正地不断发展，并成为一种不断成熟的、广为认可的应用技术。

第5章 数据挖掘行业应用原理

在我们数据挖掘应用的经历中，做过银行业、电信业、制造业、高教业、网游业和医疗业等行业项目。在本书中，我们在大部分章节引用的案例都是与医疗行业和医学领域相关的。这是因为，一方面，数据挖掘在商用中的应用题材广泛，博大精深，如果我们从各行业中均等选取案例，很难做到对数据挖掘在商业中应用的各个环节做一个系统性的、承前启后的、技术和业务都很深入的剖析，会使读者对内容有蜻蜓点水的感觉；另一方面，医学领域的数据挖掘案例具有数据挖掘技术与传统行业结合的代表性，因为医院工作流程是一个联系比较紧密、关系比较复杂的业务，医学的临床工作、临床科研和基础科研是一个内容复杂、涉及学科广泛和理论技术要求高的领域，数据挖掘在其中有很大的发挥空间，而且很多应用方法对其他行业的数据挖掘应用也有很好的借鉴作用。所以，在本章中我们主要论述了数据挖掘在医学方面的应用解决方法，偏重于科研方向，其中包括应用需求的产生、解决思路、系统设计、应用实现、效果的评价与总结等各个环节，涵盖应用各个阶段的完整过程。正如我们在第2章中提到的，我们研发了完整的医学领域的数据挖掘应用系统，即智能医学科研系统，我们简称为IMRS（Intelligent Medical Research System），如图5-1所示，它包含了从业务数据的探索性分析到样本数据的专业性选取，再到数据挖掘的深入分析，以及大量应用各种统计学方法对数据挖掘结果进行验证、评估等各个环节，是一个比较典型的行业数据挖掘应用案例。

我们在本章中以数据挖掘在医学（科研）中的应用为例，阐述数据挖掘在一个行业中的应用方法和原理，主要内容包括：首先考察传统医学科研方法的现状，然后分析医学科研对数据挖掘的需求，再提出智能医学科研系统的设计思想，并逐步阐述了智能医学科研系统的核心技术方法、智能医学科研系统的科研数据仓库建设、智能医学科研系统的核心功能设计、智能医学科研系统的整体功能设计，最后总结了智能医学科研系统的应用价值。



图 5-1

5.1 传统医学科研方法的现状

医学科研主要包括医学临床科研和基础科研，在本书的大多数场合中的科研主要指的是临床科研。如果我们谈论如何应用数据挖掘技术为医学科研服务，首先应当从医学科研应用需求的产生开始，这样，我们必然需要对传统医学科研方法的现状做个了解和分析，看看传统医学科研存在哪些问题和瓶颈，我们将如何应用数据挖掘去解决这些问题和瓶颈。本节先后分析了传统医学科研的命题与假说、传统医学科研的数据应用、传统医学科研的统计学应用、传统医学科研的流程等方面。

5.1.1 传统医学科研的命题与假说

科研命题和假说是科研工作的源泉。一个好的、高水平的科研，命题和假说是它的核心和生命。因此有一个好的命题和假说是科研成败的首要条件。在医学科研上，传统的医学科研命题和假说的来源主要包括医学实践、科学文献、实际问题等几个方面。

1.从医学实践中来

医学实践的概念比较宽泛，我们最常见的医学实践主要是医学临床工作。医院每天接待、诊断、治疗大量病人，从这些频繁的医疗活动中，医学工作者有所感悟，有所发现，经过许多年的总结、推理、思考，经验和推测能升华成经过验证的知识。这些经过实践初步验证的知识会成为一个科研课题的命题和假说，需要进行进一步深入研究和科研验证。这种方法是广大医学工作者最常用的，也是最熟悉和自然的科研方法。

2.从科学文献中来

他山之石可以攻玉。当今的科研，包括医学科研，随着经济和科技的发展，呈现加速发展的态势，全球的医药学术成果和学术论文有逐年增长的趋势。因此，大量阅读相关医学文献，从中得到新的启示，从而进一步提炼成新的科研命题与假设，是另外一条重要的科研渠道。

3.从实际问题中来

有些科研思想的产生来源广泛，在大量的临床实践中产生了很多疑问和问题。另一方面，大量的临床科研和基础医学科研中尚有许多未解决的问题，还有社会生活中的诸多方面触及了许多有关医学理论与先验知识解释不了的现象。这些问题，以及问题的演绎问题，也促使新的科研思路的产生，并酝酿新的医学命题与假说。

5.1.2 传统医学科研的数据应用

传统医学科研的数据通常称为科研资料，是从大量临床工作中收集、整理而来的，它们是医学科研的重要依据和资源，是决定科研成败的关键因素之一。因此，医学科研数据的应用至关重要。

传统医学科研的数据来源包括：

- 有特定目的性的数据，由科研设计主动采集，多为前瞻性研究的资料，包括临床实验数据、调查问卷、基础实验数据等。
- 非特定目的性的数据，是从大量的临床数据中筛选符合特定科研设计的数据，多为回顾性研究的资料。

传统医学科研的数据形式，最早主要是记录于纸质介质上的资料，随着电子信息技术的发展和普及，越来越多的电子化数据被用于医学资料中。现在，我们把传统医学科研的数据形式大致分为三类：

- 非电子化的数据。无论是前瞻性研究，还是回顾性研究，所有的实验数据均记录在纸质的介质中，尚未电子化。
- 不完全电子化的数据。有些数据经业务系统或专用科研系统录入为电子化数据，但不能覆盖医学科研所需要的全部数据，还有一部分资料记录在纸质上。
- 电子化的数据。由专业设计的科研软件或医学应用软件录入了

某些科研所需要的全部数据。这部分数据往往是为某些或某类科研定制的，有些能做到具有较广泛的通用性。

传统医学科研的数据采集方法可大致分为如下两类：

· **人工录入采集**。将数据采集到纸质记录上，然后电子化到Excel等初级的电子化文档中。例如，将科研的调查问卷印发几百份至几千份分发到调研对象手中，然后采集填好的问卷，再将有关数据录入到Excel或一些简单的管理软件中。再如，医生从纸质病案抄写大量的科研用资料，然后输入到Excel等电子文件中，等等。

· **数据库化采集**。随着近十几年医院信息化的不断发展，大量的临床数据被记录到电子化的数据库中。因此，越来越多的科研者开始从这些电子化的临床数据库中调取一些研究所需的资料，但方法大多是请求信息部门的帮助，人工从数据库中导出或复制出某些数据，而且这些数据都还是业务数据格式，还很难保证数据的完整性和可靠性。还有，科研人员和信息人员双方需要对医学目的有非常良好的沟通，使信息人员采集的数据能精确定位于科研人员所需的样本数据。

传统医学科研数据的样本量一般都不会太大，常见的量级有几十例、几百例，上千例以上就称为大样本了。传统的医学科研数据的使用多为一次性或少数多次性的，大部分数据的收集常常是为某一科研专题服务的，一旦研究结束，这些数据也就束之高阁了。

5.1.3 传统的医学科研的统计学应用

本书中提到的医学，大多数情况下指的主要是西医。现代医学主要来源于西医，当然中医也是现代医学的一个重要组成部分，并发挥着独特的优势。现代医学科研主要是基于现代医学技术的科学论证，很多科学论证都采用统计学等数学工具。因此，统计学被广泛应用于医学科研项目中。

传统医学科研的统计学大多应用于实验设计，如定量分析、定性分析及描述性统计分析等。统计学的一个核心是基于概率的、由小样本推测总体规律的方法，实践中通常会有诸如误差、偏倚及数学应用条件限制等问题。统计学的医学应用一是观察有无差异（大部分是低维度情况下），二是观察相关关系（多数为线性），三是获得预测模型（显式模型，线性模型较非线性模型的多）。

传统的医学科研的统计学应用大多采用通用统计学工具软件（以前甚至使用数学公式人工计算）。但大多统计学软件并不是专为医学设计的，更不是为专业的医学科研定制的。应用这些软件进行统计学处理需要数据具有一定的前提条件，所以要求使用者应有足够的统计学知识，同时也比较熟悉医学专业知识和科研的设计方法，否则可能使科研成果的可信度大打折扣，甚至得出相反的结论。

传统的医学科研的统计学应用，一般限于验证某科研假说，如果得不到期望的设计结果，往往需要重新设计实验并采集实验数据。统

计学的应用一般尚不能用于帮助科研者开展科研创新，使得应用非常局限。

5.1.4 传统医学科研的流程

一般情况下，传统医学科研的流程可以简单地描述为以下过程：科研工作者从医学实践中发现、总结某些现象，或查阅大量的医学文献，从中发现某些新的线索，然后提出一个新的科研假说，初步计划一个科研设计，然后搜集有关数据，初步做一些统计学分析和专业分析，得到初步性结论后正式立题、写标书、申请有关基金，中标后再正式开始研究，正式研究的步骤与前期预研究的过程相似。

5.2 智能医学科研系统的需求

从前面介绍的有关传统医学科研方法的现状来看，传统医学科研方法在医学科研的命题与假说方面、医学科研的数据应用方面、医学科研的统计学应用方面，以及医学科研的流程方面，都存在着一定的问题和瓶颈，这些问题和瓶颈大大阻碍着医学科研的有效发展。因此，有必要从医学科研的方法学上加以改进和提升。运用以智能计算技术为核心手段的智能医学科研工具就是一个很好的医学科研的改良方法，它将显著地改进传统医学科研方法的各个环节，加速提升医学科研的生产力。我们将在本小节讨论一下临床医学科研的需求和临床医学科研的解决思路。

5.2.1 临床医学科研的问题

下面，我们将从医学科研的有关业务环节具体分析一下现存的一些问题和瓶颈，并有针对性提出一些解决方案，满足提升临床医学科研方法的需求。

1.费时费力

医生既要担负临床的重任，又要进行科学研究，显然时间上和精力上都很捉襟见肘。由于传统科研流程，从科研创意与选题到科研预实验，再从科研设计到数据收集以及统计分析，最后总结论文，周期相当长，而且如果其中某个环节有误，返工的工作量较大，无疑更加大了科研周期，而基于长周期的科研活动，也必然带来人力、财力成本的增加，并且难以对科研新热点进行快速反应。

2.思路枯竭

由于当前的临床工作和传统的科研活动方法很大程度上束缚了医学信息的获取和学术接触面，科研者在科研思路上和选题途径上存在瓶颈。另外，受科研经验、临床经验的局限，尤其是中初级科研者还囿于学术素养的限制，导致科研假设和科研设计上的困难。

3.科研设计和统计分析质量不高

在统计学方面，尤其是复杂严谨的数理统计设计，也是很多没有

专业统计学背景的科研者感到棘手的问题。如造成统计设计不严谨甚至错误，会导致整个科研成果的可信度下降，从而影响了成果发表的档次与水平。

4.海量电子化临床数据利用低

对于现有的电子化临床数据，由于缺乏专业的数据整合、数据挖掘和数据分析的知识和技术，并由传统的科研思维所局限，很多科研者尚没有意识到这些海量的临床数据是宝贵的科研素材，很多医学规律，甚至重要规律可直接从中获取，科研者也缺乏对这些数据进行整合、挖掘、分析的智能工具。反过来，有时数据资源的不足也是一个重要问题。不同的医疗机构数据的数量，病种的分布，数据链的完整性，数据的可信度都存在巨大差异。

5.2.2 临床医学科研的解决思路

我们把智能医学科研方法解决临床医学科研的问题总结为以下几个要素：

1) 从海量数据中寻找科研思路。主要是积极吸取国外有关先进的技术和经验，结合医院现有的资源和实际工作情况，应用智能计算技术搭建一个完整的智能化科研平台，即智能医学科研系统，从日积月累的大量临床数据中精心提炼所需的科研资料，全方位地提供智能化的科研工具，提高临床信息的科研使用效果。

2) 从智能化科研平台中进行虚拟科研（见图5-2），如科研预实验。一是快速预观察，提供数据信息和维度特点，配合医务人员的科学素养提供知识发现的辅助，二是使用数据挖掘引擎进行的虚拟研究，为科研提供全面支持。

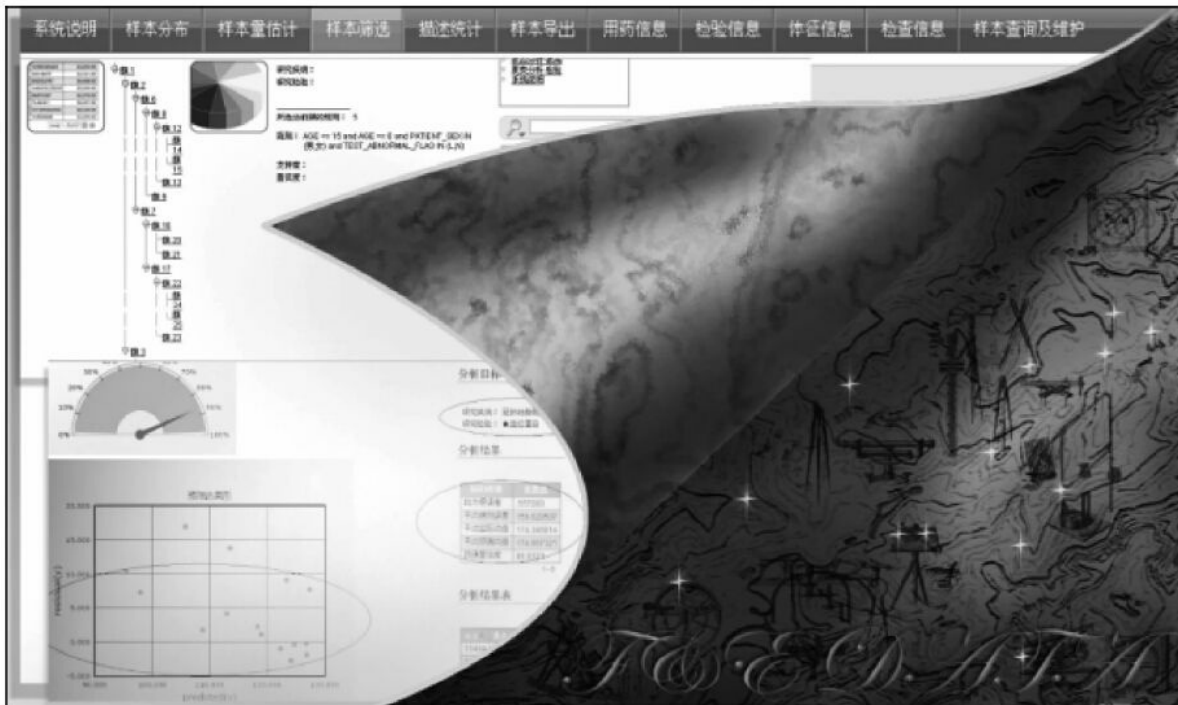


图 5-2

3) 从海量临床数据中严格筛选符合科研设计的样本数据，医务人员可以独立使用数据筛选工具，如图5-3所示，灵活与方便并重。

4) 统计学分析在智能式科研统计流程中进行，如图5-4所示，最大化避免了统计学不适当的使用。

5) 从海量临床数据挖掘中直接发现医学知识并进行统计学验证，从现象中探索变量的相关性，再从相关性中进一步探索出因果关系，最后再用统计学工具进行验证。

根据这些思路，我们就可以创建并部署智能化的科研平台，如图5-5所示，首先从医院的业务系统——医院信息平台中采集临床数据，然后将数据集中存储到临床数据中心CDR中，接着，再将CDR中的数据采集到科研数据仓库中，并按照科研分析的需求对临床数据进行

行整合、变换，最后将可分析的数据输送到智能医学科研系统中，提供给各个科研组或科研人员研究应用。

观察条件：用药后首次化验 (必选)

包含用药：即时注射液 (必选)

不包含用药： (可选)

用药时间跨度：3 (天)

用药方法：每次用量：10ug用法：静脉滴注频次：每日一次 (可选)

药剂总量范围：下限10 - 上限200

开始日期：01-06-2000 (必选)

结束日期：01-06-2010 (必选)

观察项：白蛋白 (必选)

化验时间跨度：3 天(必选)

输入本次筛选名称：test_lab (必选)

生成数据

图 5-3

智能统计流程

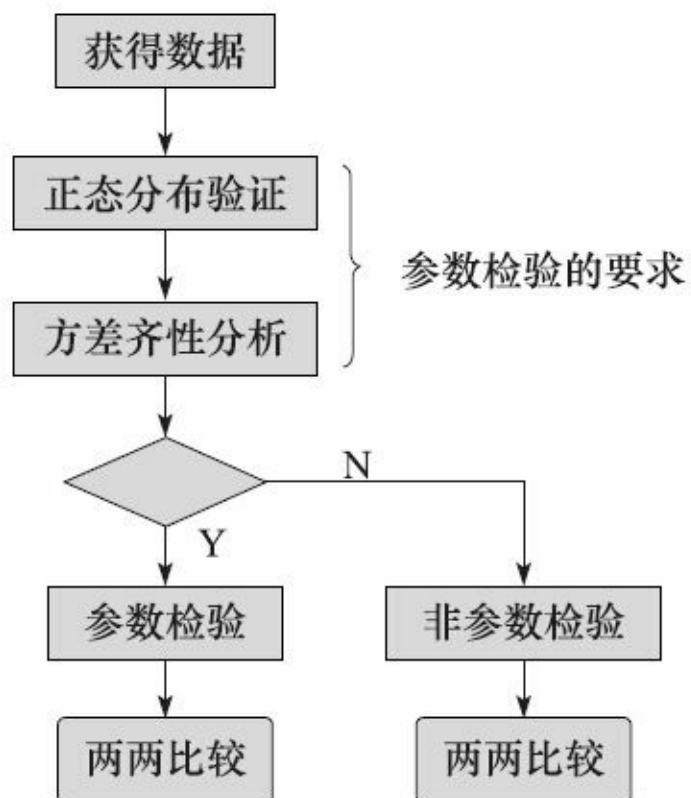


图 5-4

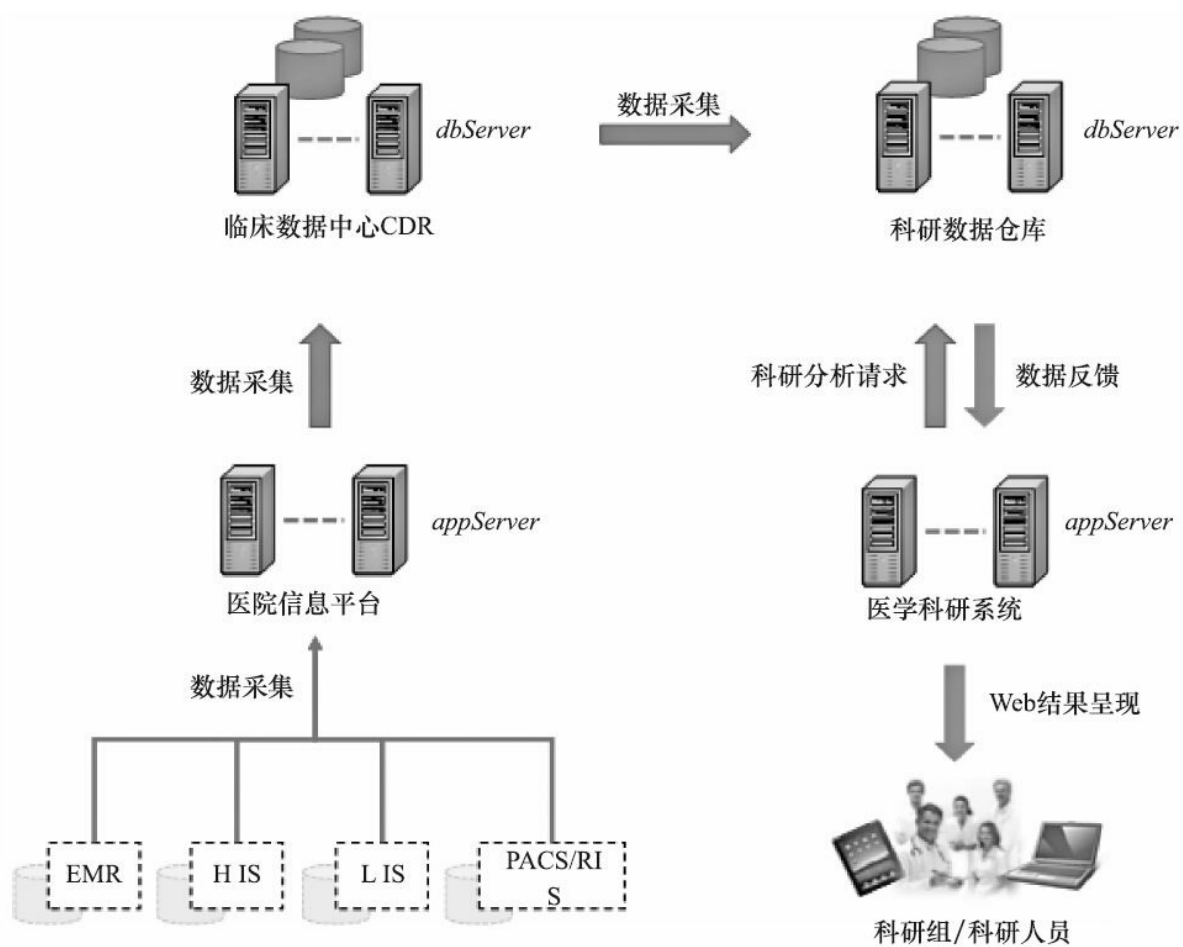


图 5-5

5.3 智能医学科研系统的设计思想

根据前面智能医学科研方法解决临床医学科研问题的思路，我们在如下几个方面总结了智能医学科研系统的设计思想：科研立项、科研设计与统计分析，以及样本数据收集与分析等。

5.3.1 科研立项

选好科研课题的关键在于思路，临床医学科研工作者面临的主要问题是，由于受学术素养和临床经验的积累速度和医学信息量的局限，在发现医学理论与实践的矛盾、医学理论体系内部的矛盾并由此提出科研假说，以及发挥科研创造性思维等方面显得力不从心。在选题策略上，医研人员主要是从临床实践问题中选题和从学科交叉发展中选题，以及从已有课题的延伸中选题。这些方法在没有思维辅助工具帮助的情况下，都是繁重的脑力劳动。所以在命题来源上，由于受医学实践经验和传统选题资源和方法的局限，医研人员往往很难确立较好的命题。

专门设计的智能医学科研系统，如前面章节提到的IMRS，其中的科研预分析子系统，针对如上科研思路与选题方面的问题，设计了帮助医研人员开拓科研思路、做好课题立项的数据探索工具。系统基于大量的医学临床数据，从疾病到实验室指标，再从药物与用药前后的实验室指标变化的关系中找出大量的关联现象和作用，如图5-6所示。如不同疾病的实验室指标之间、同一疾病的不同实验室指标之间，药物与实验室指标之间、以及不同药物之间的特有关系，从而为科研人员挖掘了大量的传统临床和科研不易得到的线索和规律，获得很多新的医学特征，从而为科研人员打开了广阔的思路，为选好课题提供了正确的方向。基于科研预分析子系统的数据探索过程，实际上也是一个很好的科研预实验的一部分，课题的成功率也会大大提高。

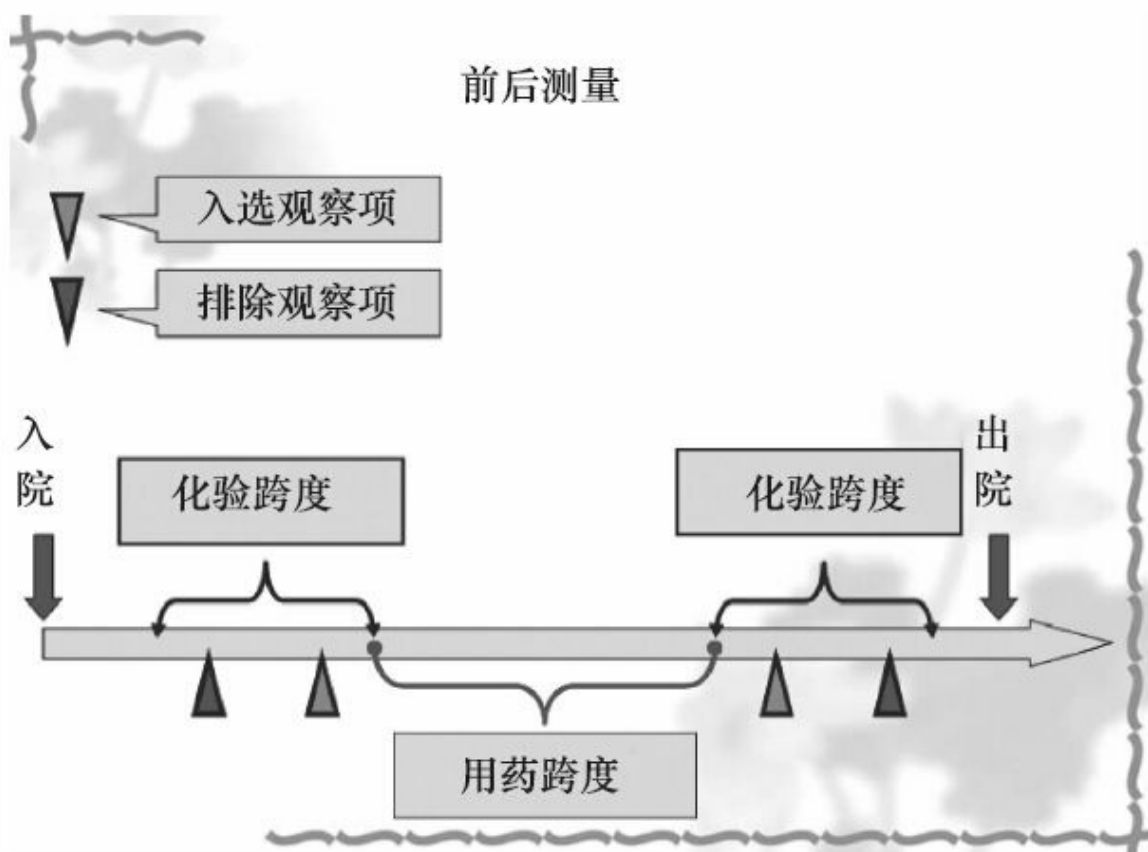


图 5-6

5.3.2 科研设计与统计分析

医学研究从立项、设计、实施到评价，均需用到统计学的思维和方法。由于缺乏足够的医学科研专业教育与培训以及丰富的科研设计经验，不少科研人员往往感到医学实验设计比较困难，并常常出现各种设计缺陷和漏洞，例如在实验设计的“三要素”、“四原则”和“设计类型”上把握不当。在另一个科研关键因素——医学统计设计上，由于很多医学人员没有专业统计学背景，在统计分析设计方面感到困难。而现有的一些通用统计软件并没有专门的医学科研统计设计，不能从医学科研设计上把好关，也不能实现与在线网络数据的实时的无缝衔接。

IMRS系统是根据国际经典的医学科研设计理论和方法，精心设计研发的智能临床医学科研平台。它是一个专业的、可靠的科研分析工具。这个全在线式（院内局域网）的科研平台提供科研全过程的一揽子工具，包括智能式的科研统计设计流程，省去了繁琐复杂的人工数据处理。系统采用基于世界最大的数据计算厂商Oracle（甲骨文）公司最新提供的与SAS、SPASS等国际主流统计软件同步验证的统计算法和业界公认的数据挖掘算法，如图5-7所示。系统中运用多种数据挖掘技术探索数据规律，直接挖掘医学新知识，不仅可以大大简化科研工作，提高效率，为科研设计提供科学依据，保证了科研的成功率，还大大增加了重大科研发现的概率。

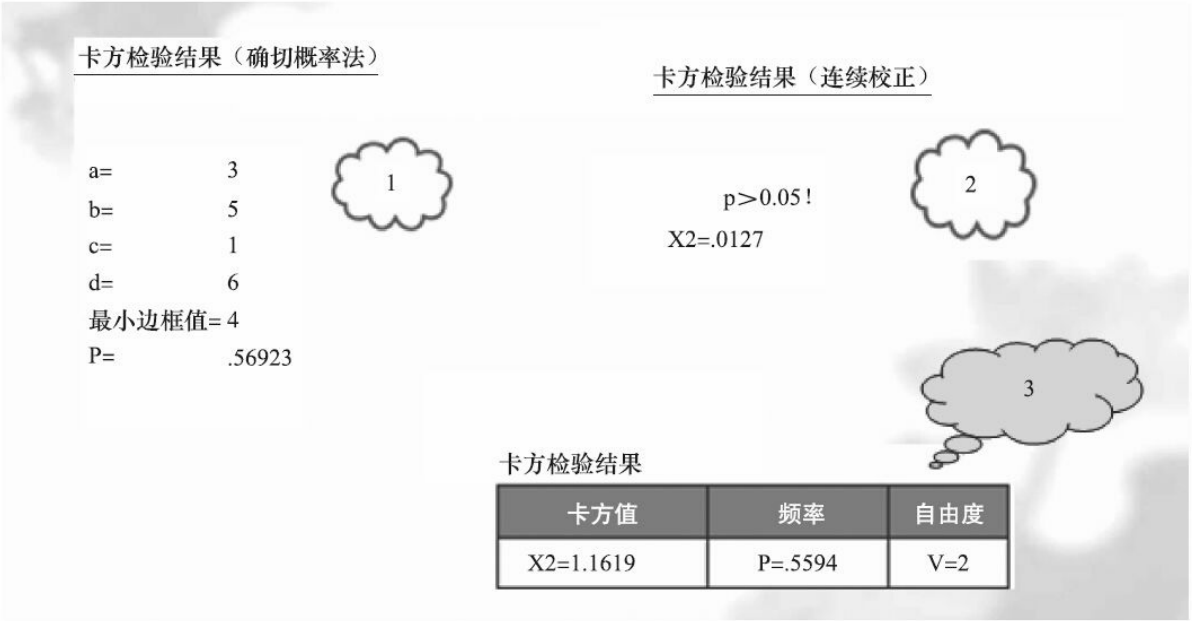


图 5-7

5.3.3 样本数据收集与分析

传统的样本数据收集方法往往存在周期长、效率低、随机误差大和人为主观造成的偏倚大等缺点。数据收集不仅耗费大量的人力物力，而且由于数据未电子化，造成数据分析与共享困难，如要将数据电子化又要花费可观的人力和费用，总的结果是样本数据成本高、利用率低和产出小。由于传统的科研预实验方法的限制，造成样本数据量大小的估计不精确，导致样本量估计不够而得不出可信的结论，或者样本量过大造成成本浪费，如图5-8所示。

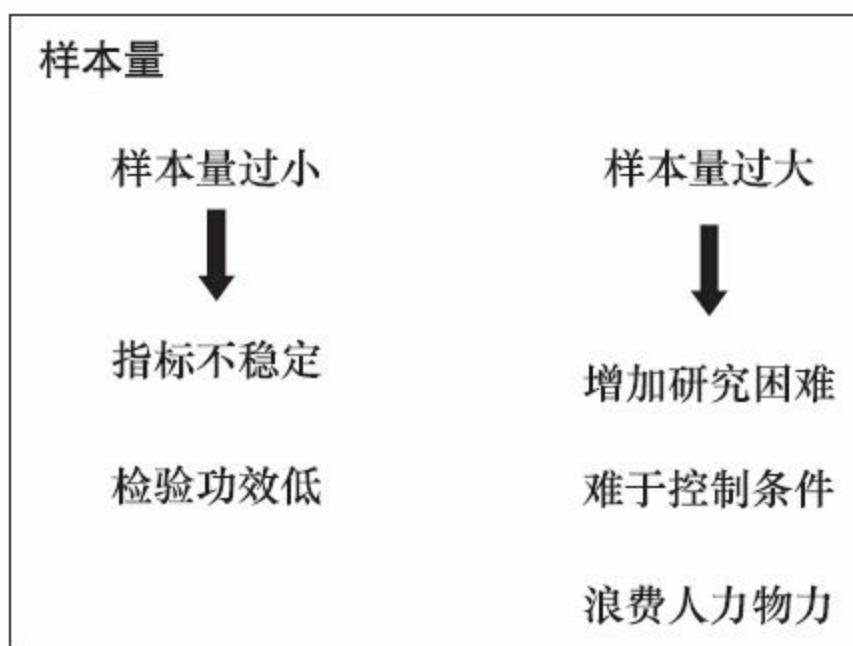


图 5-8

我们在IMRS中，设计了数据筛选子系统解决如上问题。该子系统直接利用医院积累的HIS/LIS/PACS/病案系统等海量医学数据建立专用医学数据仓库，从中收集医学科研样本数据。样本分布分析模块

可使科研者快速了解全院系统内各病种分类样本数目；样本量估计模块使科研者很容易地得到相应的科研设计所需的精准的最小样本数；满足各类科研设计的样本分类筛选器使科研人员快速、准确、严谨地收集到复杂条件下的样本数据。直接多课题交叉重复利用科研人员积累的医学数据，使科研成本大大降低。强大易用的样本筛选器，使科研人员科研数据的收集高效、准确，获得大样本数据易如反掌，图形化的数据分析器更是科研人员进行科研准备工作的有力工具。

5.4 智能医学科研系统的核心技术方法

实现智能医学科研系统，需要运用智能技术、数据整合技术、IT展现技术等多种技术和方法。我们将智能医学科研系统的核心技术方法总结为以下三个方面：以数据仓库为核心的数据整合技术、以数据挖掘为核心的知识发现技术、以数理统计和科研流程为核心的统计分析技术。

1.以数据仓库为核心的数据整合技术

以数据仓库技术为核心的医学数据整合系统，独立于已有的医疗机构业务系统，以全新的设计将分散的业务系统产生的不一致的数据进行整理、变换、集成，整合得到全面、高效、一致的信息。数据仓库技术还能对历史的全部海量数据进行在线的、实时的、深入的分析，并使其变得很轻松。直接利用积累的现有医学数据，使科研成本大大降低，能利用节省下的科研经费取得更多科研成果。应用数据仓库技术的整合技术，使获得大数据科研样本数据易如反掌。

2.以数据挖掘为核心的知识发现技术

数据挖掘，又称数据库中知识发现，正如我们在第1章提出的，是有目的地探索数据中隐含的规律和知识的过程。数据挖掘所探寻的模式是一种客观存在的、但隐藏在数据中未被发现的知识。例如，数据挖掘可直接挖掘疾病高发人群、疾病及症状间的未知联系、化验指

标间的影响关系及化验指标与疾病间的潜在影响、对未知的检验项值进行预测等。通过可观测指标推断不可观测指标，或通过简单易行的观测指标推断昂贵的或临床有创的指标。由简而知繁，由易而知难。再如，在科研设计中利用聚类分析、因子权重分析，我们可以对数据进行分布更加适合的分组、考察多因素的不同权重、帮助确定析因分析或嵌套分析等不同的科研设计。数据挖掘在医学中应用非常广泛，为医学研究提供传统方法不能企及的技术手段，例如：聚类分析、关联规则分析、特性选取分析、回归预测分析、异常侦测分析和特性挖掘分析，等等。

3.以数理统计和科研流程为核心的统计分析技术

将医学统计方法以流程控制的方式软件化，使每一位医学专业人士不再将数学和统计学视为科研难点和烦恼。以数理统计为核心的统计分析技术涵盖了很多的医学定量分析和定性分析统计算法，统计工具涵盖了所有常用的正态检验、各种t检验、方差齐性检验、方差分析、秩和检验、卡方检验、回归分析和相关分析，等等，如图5-9所示，满足了最常用的科研课题的统计学要求。

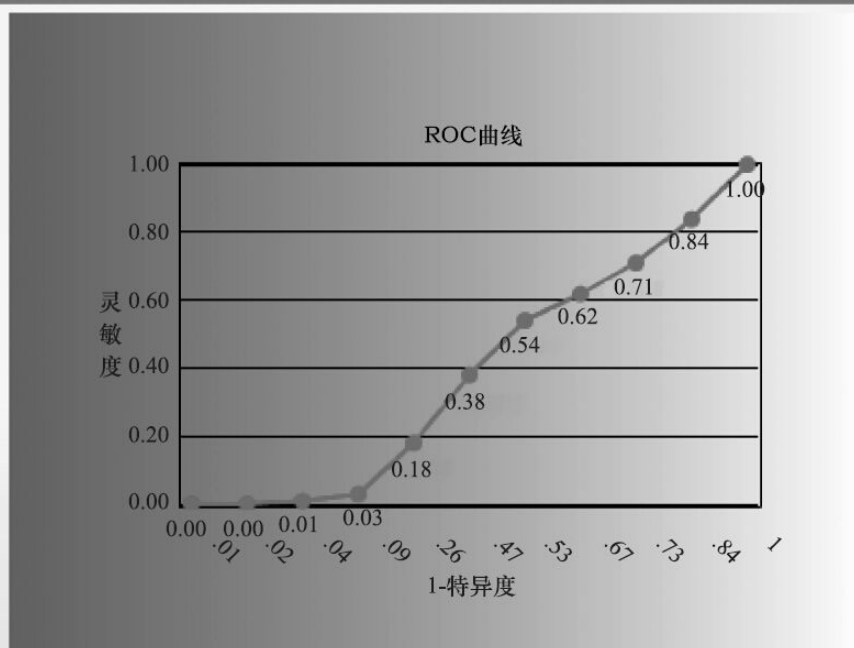


图 5-9

5.5 智能医学科研系统的科研数据仓库建设

在日常科研活动中，病历资料是我们极其珍贵的科研素材，无论是在论文撰写还是在临床科研项目申报过程中均起到主导地位。此外，临床教学医院目前承担的很多科研课题，其数据提取、统计、挖掘和分析能力也亟待提高。因此，作为IMRS智能医学科研系统的基础功能，通过将医院分散的临床数据整合起来，建立一个独立的、完整的医学科研数据仓库可以更好地为科研服务。下面我们将从医学科研数据仓库建设的技術方法、医学科研数据仓库的建设过程、科研数据仓库的数据安全等几个方面谈一谈科研数据仓库的建设。

5.5.1 医学科研数据仓库建设的技術方法

现在，数据仓库技术已经广泛应用于各行各业中。但由于每个行业都有其业务自身特有的一些需求和特点，因此数据仓库的开发、建设应具有行业的针对性。我们在分析、总结了医学科研的业务特点和特殊需求后，提出了一套建设医学科研数据仓库的技术和方法。

1.医学数据仓库概念

医学数据仓库是将医院的事务性数据经过提炼、加工集成为对医院有用的信息，是面向主题的、稳定的、随时间变化的、用于决策支持和数据分析的数据库系统。数据仓库是定期更新的，存放大量的有用数据，应用智能报表工具、数据挖掘工具，结合专业人员的专业知识，从数据中获得有用的信息，为医疗、科研、管理和决策提供宝贵的信息和辅助手段，从而大幅度提升医院在医疗、科研和经营管理方面的水平。

2.面向主题的数据仓库技术

数据仓库按规模和用途可划分为企业级数据仓库和部门级数据仓库。其中企业级医院数据仓库能为整个医院各个部门的运行提供决策支持手段；而部门级医院数据仓库，也称数据集市（Data Mart），是部门级的或面向医院中某一主题的数据仓库，一般只能为某个局部范围内的人员服务。数据集市是将从业务数据库或者其他地方获得的数

据进行储存的数据仓库，它被设计用来服务于某一部门或者某一主题，因此数据集市技术也被称为面向主题的数据仓库技术。建立医学科研数据仓库，就是在医院层面上建立面向医学科研主题的数据仓库或数据集市。

3.数据仓库技术在医学科研中的应用

我们将数据仓库技术应用于医院的实例项目上，就是将医院现有的电子化业务数据中有关临床医学的海量数据进行整合，以集成的电子化数据形成医学科研的专用数据仓库。以数据仓库技术为核心的医学数据整合系统，独立于已有的医院业务系统，以全新的设计将分散的业务系统产生的不一致的数据进行整理、变换、集成和整合得到全面、高效、一致的信息，如图5-10所示。数据仓库技术还能对历史的全部海量数据进行在线的、实时的、深入的分析，并使其变得更方便。

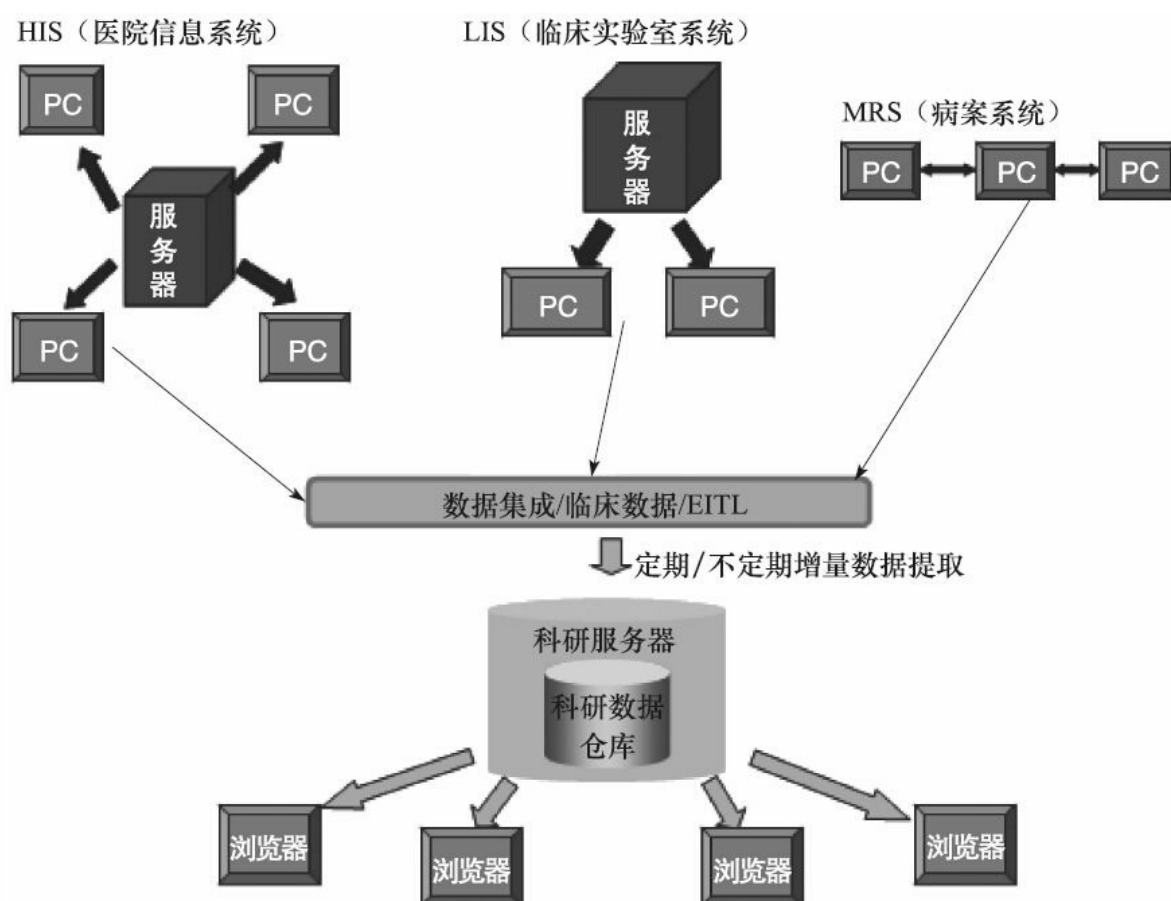


图 5-10

4.医学科研数据仓库的设计方法

数据仓库的用途决定了其设计和实施都与传统的交易型数据库不同。一个成功的数据仓库的创建和实施，不仅需要良好的数据模型和优化的实现技术，而且需要把众多相关的因素很好地进行集成和组织，形成一套有效的设计方法和实施规则。医学科研数据仓库的设计通常会分多个阶段来实施，又称分阶段项目技术（Phase to Phase Project），其中的每一个阶段（Phase）的开发通常又采用瀑布（Waterfall）模式。数据仓库的设计通常都采用范式技术（Normalization）设计数据的关系层次（Hierarchy），运用逆范式技

术（De-Normalization）和星型（Star Schema）模型设计方法设计多维（Multi-Dimensional）的数据库。其中有效地进行数据的采集、传送、装载（E、T、L）在开发过程中非常关键。应用商务智能（BI）工具、数据挖掘工具分析数据仓库中的信息，为用户提供分析结果和依据，体现数据仓库建设的价值。最后，高效的数据仓库还少不了数据优化，如利用数据分区（Partitioning Data）、数据索引（Indexing Data）等优化查询速度。医学科研数据仓库的建设是一个不断完善的过程，但目的只有一个，就是从数据中获得有用的信息，挖掘和共享临床科研数据，帮助提高科研者的科研能力和效率，多、快、好省地获取科研成果。

5.5.2 医学科研数据仓库的建设过程

现在我们以某大型三甲医院为例，简要说明医学科研数据仓库的建设过程，更详尽、更技术性的介绍参见第6章。

5.5.2.1 分散的业务数据源分析

医院的临床业务系统主要包括住院信息系统（HIS）、实验室信息系统（LIS）、病案信息系统、影像报告系统（RIS）以及体检信息系统等。这些信息系统涵盖了医院大部分临床业务，分散在多个不同的数据库中，数据库环境为SQL Server、DB2、My SQL，以及桌面级文件。这几个数据库的历史数据不尽相同，除了体检信息系统（数据独立于其他业务系统），最长的积累约八年的数据，最短的积累约五年的数据。按照数据集成的完整性，约有五年的完整临床数据，包含了临床业务的各主要环节，可为临床科研提供完整的研究资料。我们需要做的是，将这些异种数据库的临床数据集成在一起，并存储到Oracle数据仓库中，如图5-11所示。

5.5.2.2 医学科研数据仓库的技术设计

数据仓库系统往往基于超大数据集，而且需要配备专用的数据分析工具，因此我们选用了Oracle 10g数据库系统作为我们的科研数据库平台。各业务数据库与科研数据仓库的数据通信，采用异种数据库

连接的架构设计。由于SQL Server是基于Windows环境的，所以我们的Oracle 10g也选用基于Windows环境的。因此，异种数据库连接采用基于Windows的开放数据库互联ODBC。

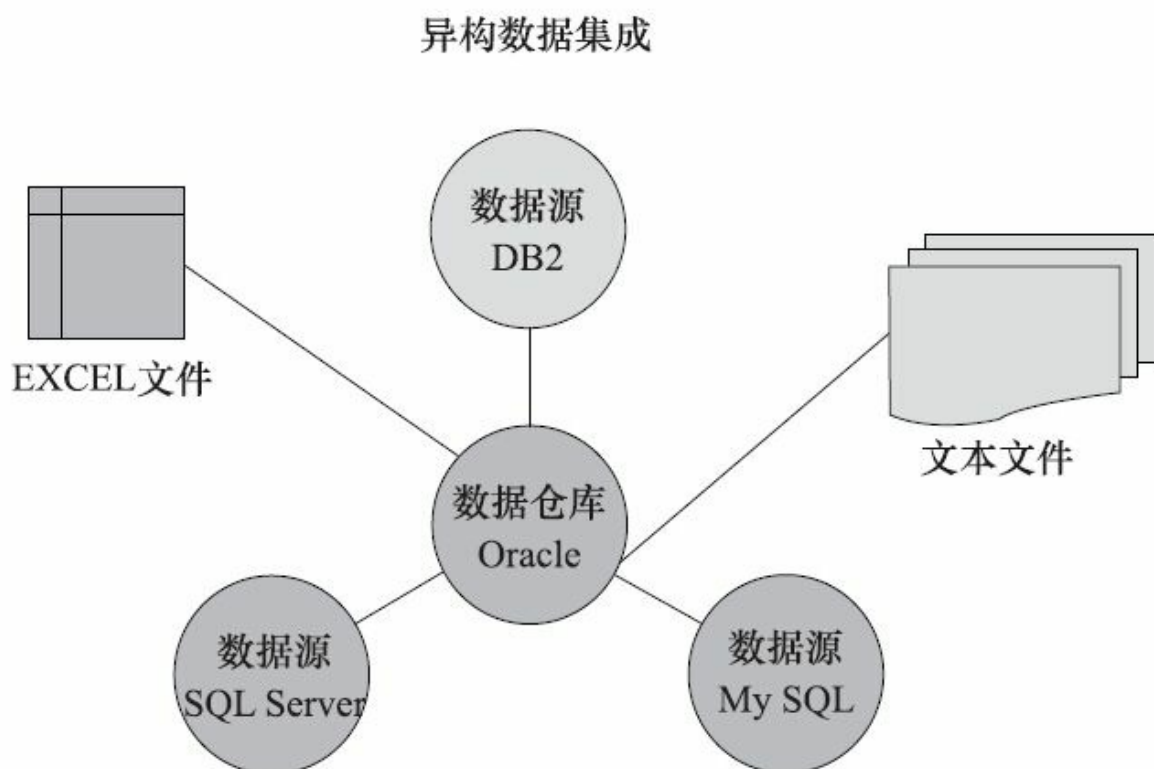


图 5-11

5.5.2.3 医学科研数据仓库的数据整合

数据整合技术称为ETL（Exact抽取、Transformation变换、Loading装载）技术。实际上，抽取、变换、装载只是概括性地代表了数据仓库整合的主要过程，在实际项目中应用的具体数据整合技术和方法会很多。在项目中，我们应用了很多的整合方法，如源数据导出、数据导入、数据清洗、数据约束性清洗（除去重复记录）、数据

类型转换、数据格式映射、数据追加、数据更新等。我们将ETL的处理流程分为四个阶段，第一个是将数据从业务系统中抽取到数据仓库中的登录区（Landing Area），然后再将数据送到变换区（Staging Area）进行变换，接着再将变换后的数据传输到数据中心（Data Center Area），最后将可分析的数据调到数据集市（Data Mart Area），提供给应用程序，如图5-12所示（详见第6章）。

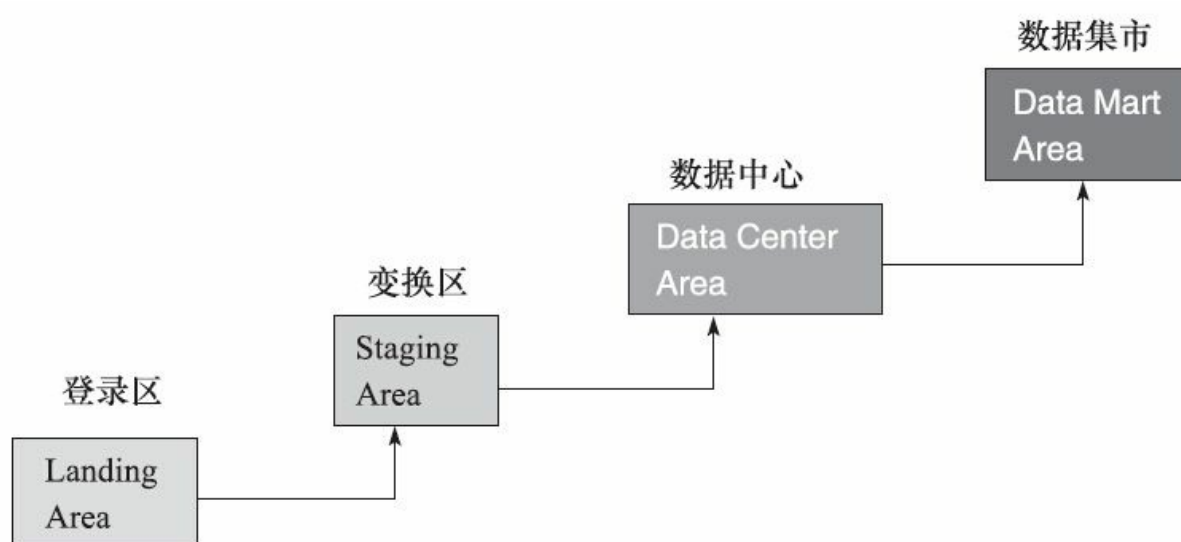


图 5-12

由于数据整合不是本书的核心内容，因篇幅所限，在此只给出数据整合的一个实例。项目中需要整合来自两个异种数据源的数据：病案数据库和住院病人数据库。在异种数据源的匹配分析中发现，直接将住院病人数据库中住院病人表的主键INPATIENT_NO与病案数据库中病人和病案表的主键BAID进行连接时，它们的交集非常少，按照业务数据设计的规则，因此需要建立一个数据变换的映射。经过数据变换的映射后，得到了较满意的整合结果，如表5-1所示。

表5-1 映射后的整合结果

	住院病人表	病人和病案表
两表匹配度： 114700/143357 = 80%	匹配字段：NAME, SEX, BIRTH_DATE, IN_DATE, OUT_DATE, INPATIENT_NO (经处理)	匹配字段：xm, xb, csrq, ryrq, cyrq, baid

为了进一步增加住院病人表与病人和病案表的交集，我们按照姓名、性别、出生日期、入院日期和出院日期进行匹配连接，结果使得两表的匹配度进一步提高：在住院病人表中，按照姓名、性别、出生日期、入院日期、出院日期比较，有124374条记录与病案表匹配，占出院病人143357的86.76%；有18983条记录与病案表不匹配，其中11046条记录是产科出院婴儿；2254条记录是短时间多次出入院，并已合并到病案表中；5683条属于其他特殊情况。如表5-2所示。

表5-2 匹配链接结果

	数据表：住院病人	数据表：病案和病人
两表匹配度： 124374/143357 = 86.76%	匹配字段：NAME, SEX, BIRTH_DATE, IN_DATE, OUT_DATE	匹配字段：name, xb, csrq, ryrq, cyrq
其余 18 983 条不匹配记录分析与处理	11 046 条记录是出院产科婴儿。 2880 记录是短时间多次出入院，并已合并到病案中。 5058 条属于其他特殊情况。	5058 条中有部分可能是病案丢失的数据

5.5.3 科研数据仓库的数据安全

临床科研项目的目标始终是研究群体的医学规律，从不涉及个案及医、患双方的隐私信息。因此，智能医学科研系统中没有与病人隐私有关的任何信息，如姓名、身份ID、住院ID、住址、单位、电话等，更没有与临床工作人员和科室有关的任何信息。此外，根据医院对业务系统和临床数据安全的要求，我们在医学科研仓库系统中增加了诸多安全技术措施，如：

- 将医学科研仓库建在医院外网，与内网业务系统物理隔离。
- 对所有涉及病人ID的信息均予以剔除，在应用系统中看不到任何病人和工作人员的个人信息。
- 封闭了所有源数据的下载功能，普通用户只能导出科研结果数据。
- 对应用系统的用户进行分级授权管理，不同的用户拥有不同的数据访问权限等。

5.6 智能医学科研系统的核心功能设计

按照前面对科研业务需求的分析，我们设计了智能医学科研系统IMRS的核心系统——临床科研数据挖掘系统的架构功能。按照临床科研的普遍需求，我们将临床科研数据挖掘系统的功能划分为六个方面：

- **分布探索**。所谓分布探索就是探索临床数据的客观分布。数据挖掘的主要技术是聚类分析。智能医学科研系统中具体设计如下两个模块：聚类疾病和聚类检验。

- **关系探索**。关系探索是探索事物与事物间的伴随关系，以及变量与变量之间的影响关系。数据挖掘的主要技术是关联规则和特性选择。智能医学科研系统中具体设计如下3个模块：合并症关联规则、疾病因子权重和检验因子权重。

- **特征探索**。特征探索是探索由高维变量表达的事物的主题特征，数据挖掘的主要技术是特征抽取。智能医学科研系统中具体设计如下模块：合并症特征抽取。

- **异常探索**。异常探索是探索由高维变量表达的事物的离群个案，数据挖掘的主要技术是异常侦测。智能医学科研系统中具体设计如下两个模块：多维指标负向异常点侦测和多维指标正向异常点侦测。

· **推测探索**。所谓推测探索就是根据有限变量的已知数据推测目标变量的未知值，通常有两种类型，一种是目标变量为离散型，我们将这种推测探索称之为分类模型；另一种是目标变量为连续型，我们将这种推测探索称之为预测模型，也称回归预测。我们把用于推测目标变量的有限已知变量称之为分类因子（用于分类模型）或预测因子（用于回归预测）。推测探索是建立在分类因子或预测因子与目标变量之间存在着一定的关系模型的基础上，而这种关系模型可能是显式模型也可能是隐式模型。所谓显式模型就是可以用某种数学公式或逻辑规则或其他显式方法明确表达出来的多变量关系规律；而隐式模型就是尚不能用某种数学公式或逻辑规则或其他显式方法明确表达出来的多变量关系规律，我们通常称之为黑盒子，我们可以证明它存在，并可以应用它，只是暂时不能表达它。智能医学科研系统中具体设计如下两个模块：疾病推测和指标预测。

· **趋势探索**。所谓趋势探索是按照某种次序或序列，通常是时间次序，考察并推测事物发生的变化，数据挖掘的主要技术是时间序列。智能医学科研系统中具体设计了生命指征监测模块。

智能医学科研系统中的临床科研数据挖掘系统功能架构图，如图5-13所示。

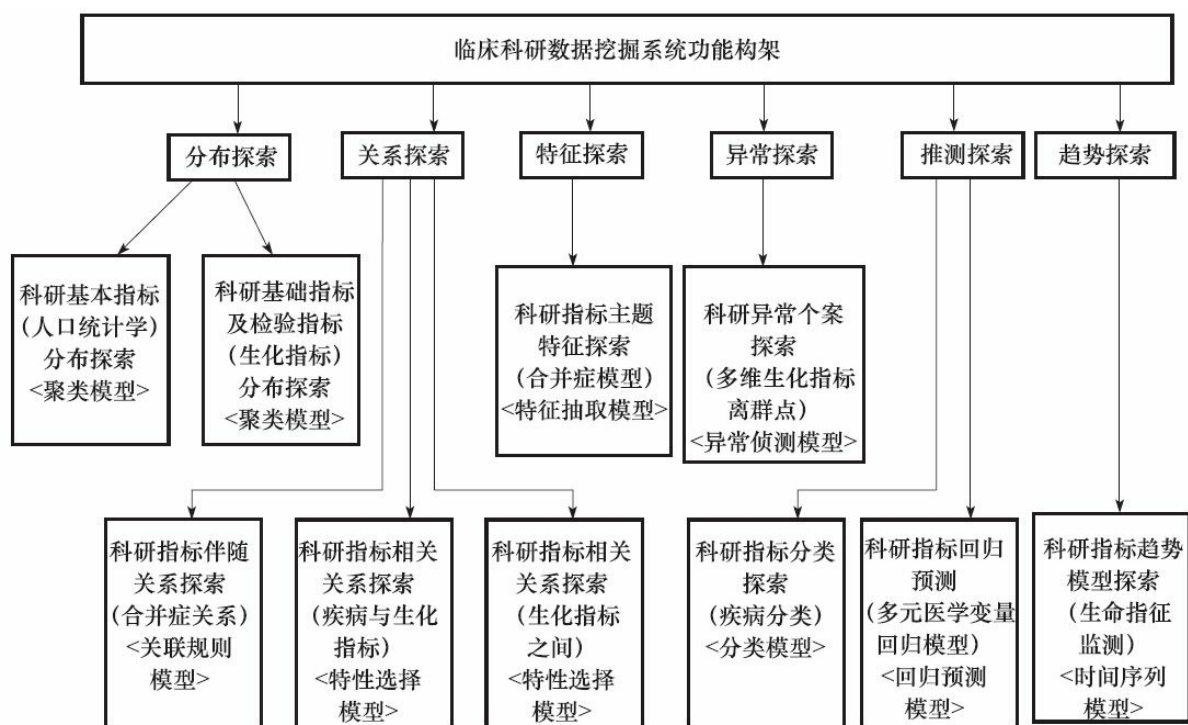


图 5-13

5.7 智能医学科研系统的整体功能设计

前面我们设计了智能医学科研系统IMRS的核心系统——临床科研数据挖掘系统的架构功能。除了临床科研数据挖掘子系统外，IMRS还有一些其他重要的子系统，与临床科研数据挖掘子系统相互配合，共同为临床科研工作服务。本节我们将探讨智能医学科研系统主要功能、智能医学科研系统的模块设计与应用实现、智能医学科研系统的评估方法等内容。

5.7.1 智能医学科研系统主要功能

通过对临床科研的业务分析，以及对亟待解决的科研瓶颈的分析，我们设计了智能医学科研系统的主要功能，如图5-14所示，包括：

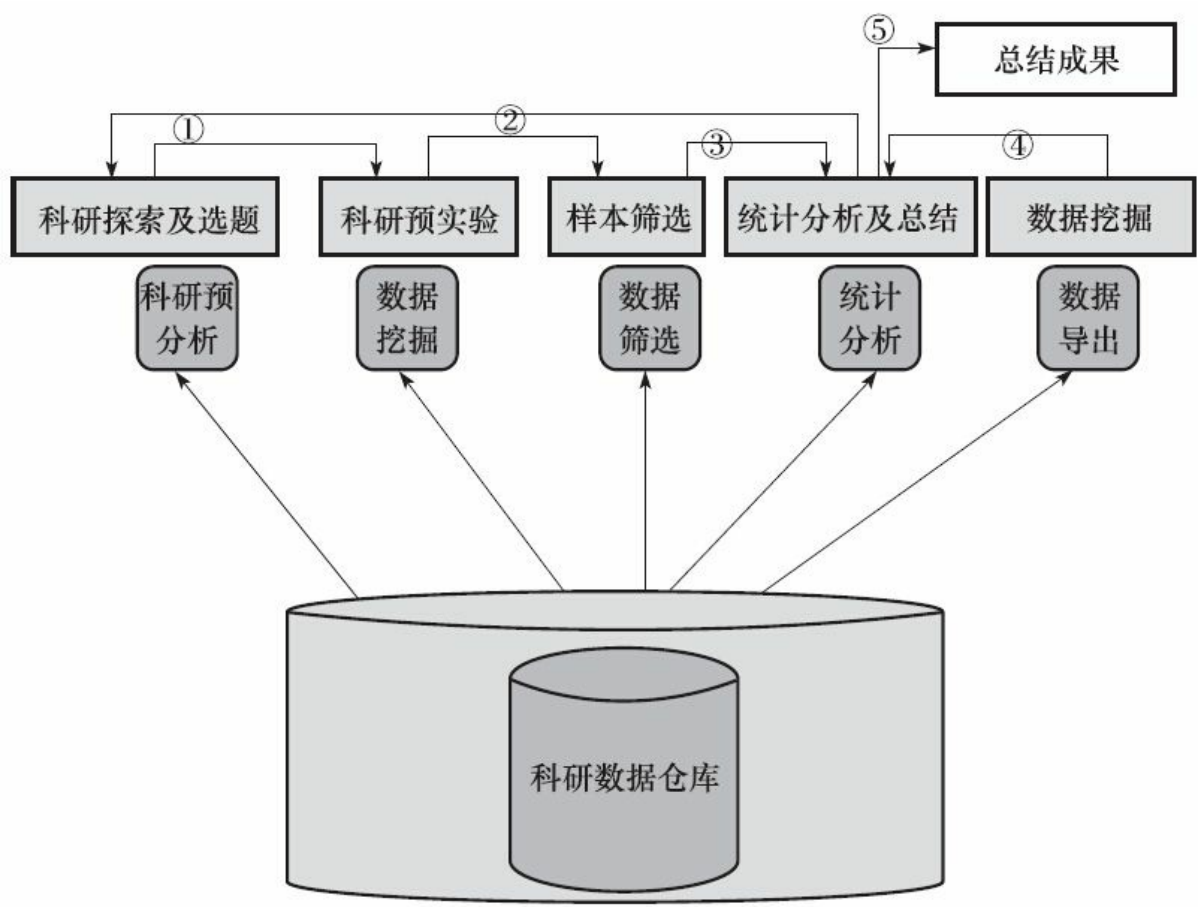


图 5-14

- 数据挖掘子系统
- 数据筛选子系统
- 统计分析子系统

这几个子系统互相配合，构成一个完整的有机体。首先，可以应用科研预分析（也称为数据探索）子系统从科研数据仓库中寻找、探索新的科研线索和方向，然后应用数据挖掘子系统按照这些新发现的科研线索和方向，对海量数据进行深入地探索和挖掘，并找到数据中蕴含的医学规律和医学靶向，然后再利用数据筛选子系统严格选取由数据挖掘新发现所确定的科研样本数据，最后再运用智能统计子系统对筛选的数据进行统计学验证。以上步骤可以循环往复，直至获得满意的科研结果为止，并对科研结果进行总结和论述，形成学术论文和科研成果。以上过程体现了一个完整的智能化科研理念和体系方法。

这里我们重点介绍智能数据挖掘子系统的设计方法。我们选取了智能数据挖掘子系统具有代表性的几个主要功能，有如下几个方面：

1. 疾病聚类分析

对各种疾病在年龄和性别特征上的聚类分析，发现各种疾病不同的年龄和性别特征；对各种疾病在其相关的各种生理指标上进行聚类分析，研究各种疾病相关生理指标的分布特征。

2. 疾病关联分析

对各种疾病在各种合并症上进行关联规则分析，分析发生某一合并症的前提下发生另一种合并症的概率，探索合并症的关联关系及因

果关系。

3.疾病特征抽取分析

对各种疾病最可能发生的多种合并症模型进行特征提取分析，并研究模型中各合并症分量之间的权重和关系，抽取各种疾病合并症的主题特征。

4.疾病相关因子分析

挖掘疾病与生理指标的相关关系，探索不同的疾病在生理指标上的差异性，研究特定疾病的不同生理指标的相关关系。

5.疾病回归预测分析

挖掘疾病中多变量的高维数学模型，探索疾病的显式或非显式回归模型，并对模型进行验证和评估，从而揭示疾病的内在规律。

6.疾病异常侦测分析

对各种疾病在高维变量中的离群点的挖掘，探索疾病特殊病例的特征和规律等。

5.7.2 智能医学科研系统的模块设计和应用实现

临床科研的基本方法是遵循所谓的“三要素”、“四原则”的基本方法。“三要素”是指研究人群、处理因素和观察结果，“四原则”就是要求科研设计要满足随机、对照、重复和均衡的原则。在科研数据上，可以理解为数据分布的要求和数据对比的要求。临床科研的很多目标是要揭示潜在的医学现象、医学指标间的关系，及复杂的医学规律等。智能化科研探索的核心就是应用数据挖掘的技术从海量的临床数据中，主动地探索、发现蕴含的潜在医学规律，这样就大大提高了科研探索的效率，大幅拓宽了科研的题材面，并可深度探究和分析非常复杂的医学问题。

我们从大量的各类临床科研活动中，归纳、总结了带有普遍性的科研分析环节，以下面几种模型为例：聚类模型、关联规则、特征抽取、特性选择、分类模型、异常侦测、回归预测。下面分别介绍。

5.7.2.1 聚类模型

临床科研通常是以医院的病人为研究对象，并对病人的诸多基本变量和临床指标进行分析、总结，其中首要的、最基本的病人指标就是年龄和性别。通常不同年龄或（和）不同性别的病人组具有不同的医学特征，因此常常需要将病人样本按年龄和性别进行分组，也就是科研的区组设计。但区组设计的方法大多采用人工划分，如将年龄按十年分为一组。如下例，选取了340个患有冠状动脉粥样硬化性心脏

病的病人，以年龄为基准，采用等距法将340个病人分为七组数据，如表5-3所示。

表5-3 等距分组表

组别	年龄段	人数（人）	性别分布
1	30 岁~40 岁	2	男性 2 人
2	40 岁~50 岁	20	男性 17 人，女性 3 人
3	50 岁~60 岁	38	男性 28 人，女性 10 人
4	60 岁~70 岁	33	男性 19 人，女性 14 人
5	70 岁~80 岁	132	男性 54 人，女性 78 人
6	80 岁~90 岁	102	男性 62 人，女性 40 人
7	90 岁以上	13	男性 4 人，女性 9 人

这样划分的数据不能科学地反映研究人群的客观的年龄分布规律。因此，基于这样的分组对研究目标指标进行统计学分析，往往得不到组间的差异性，最终得到的科研结论的可信性大打折扣。

针对这个普遍存在的问题，我们采用数据挖掘中分析数据客观分布规律的聚类技术，将科研人员任意选定的某一研究人群的性别和年龄，进行科学划分。聚类技术的特点是探索未知数据的分布规律，最终找到一组最佳分组，使得组间具有足够的相异性，组内具有足够的相似性。聚类的算法主要采用划分算法和层次算法。例如，将上例中的数据进行聚类分析，得到了不同的区组划分结果。运用聚类方法将340位病人划分为八组，如表5-4所示。

表5-4 聚类分组表

组别	年龄段	性别分布	支持度	置信度
1	37 岁 ~41 岁	男性	2	1
2	41 岁 ~50 岁	男性	19	0. 8636
3	50 岁 ~55 岁	男性	12	0. 8
4	55 岁 ~64 岁	男性, 女性	28	1
5	64 岁 ~74 岁	男性, 女性	68	0. 8395
6	74 岁 ~78 岁	男性, 女性	61	0. 7922
7	78 岁 ~88 岁	男性, 女性	88	0. 9072
8	88 岁 ~97 岁	男性, 女性	14	1

近一步，科研中还需要对病人的诸多临床指标进行深入分析和总结，例如不同年龄和性别分组下的某生化指标的特征是什么。与上一点类似，我们也将采用聚类技术进行分析，但不同的是聚类分析的变量由两个变为了三个。这时，从技术上要做些改进与优化，因为三个维度与两个维度的划分是不同的，多一个维度的引入对划分的影响很大。

我们选取了该疾病患者的白细胞检验项进行检验，总计3115人次，聚类结果如表5-5所示。

表5-5 聚类结果表

组别	年龄分布	性别分布	检验分布	支持度	置信度
1	39 岁 ~56 岁	男性, 女性	N	29	0. 7838
2	56 岁 ~60 岁	男性, 女性	H, N	16	1
3	60 岁 ~68 岁	男性, 女性	H, N, L	14	1
4	68 岁 ~81 岁	男性, 女性	N	111	0. 7872
5	81 岁 ~89 岁	男性, 女性	N	71	0. 8452
6	89 岁 ~94 岁	女性	N	6	0. 8571

5.7.2.2 关联规则

科研中很多情况要做有关病因学方面的探讨，如某种合并症是否

是导致另一种合并症的诱因。我们这里采用数据挖掘的关联规则技术进行分析。关联规则技术的特点就是从大量的随机发生的并发事件中，找到强关联的现象，使得在某事件发生的前提下，另一事件的发生具有很高的概率，并且是一种具有业务意义的强规则。关联规则的一个可贵之处，就是它的单向性，更容易观察谁导致了谁，谁可能是因，谁可能是果。更为有趣的是，如果前后件互推的置信度都很高，那么它们就是双向的，也就是说二者互为因果，或者说是同生共寄。

例如，有一组心绞痛病人大约2000例，他们的合并症有上千种，每个人的合并症可能不一样，有的合并症多，有的合并症少。但医学家感兴趣的是，在这些大量的合并症案例中是否蕴含着一定的规律，即如有某合并症发生，很可能伴随着另一种合并症的发生，而且发生的概率非常高，同时不同的支持度也具有不同的医学含义。运用关联规则模型，我们选取了1908例患有不稳定心绞痛的案例，采用关联规则的挖掘方法，判定1908例病患不同合并症之间的联系。经过计算的结果如表5-6所示（选取了Lift值大于1的13种规则）。

表5-6 关联分析结果

规则编号	前件	后件	支持度	置信度	LIFT
8	非胰岛素依赖型糖尿病	冠状动脉粥样硬化性心脏病	0.1022	0.9112	1.768 644
25	高血压Ⅲ和非胰岛素依赖型糖尿病	冠状动脉粥样硬化性心脏病	0.0597	0.8906	1.727 299
10	高血压Ⅱ	冠状动脉粥样硬化性心脏病	0.0849	0.6255	1.062 232
1	冠心病	高血压Ⅲ	0.0519	0.6149	1.123 006
21	非胰岛素依赖型糖尿病	高血压Ⅲ	0.0671	0.5981	1.092 318
23	非胰岛素依赖型糖尿病和冠状动脉粥样硬化性心脏病	高血压Ⅲ	0.0597	0.5846	1.066 547
6	心功能Ⅲ级	冠状动脉粥样硬化性心脏病	0.0624	0.5561	1.079 876
24	高血压Ⅲ和冠状动脉粥样硬化性心脏病	非胰岛素依赖型糖尿病	0.0597	0.2184	1.938 144
7	冠状动脉粥样硬化性心脏病	非胰岛素依赖型糖尿病	0.1022	0.1984	1.768 644
9	冠状动脉粥样硬化性心脏病	高血压Ⅱ	0.0849	0.1648	1.062 232
22	高血压Ⅲ	非胰岛素依赖型糖尿病	0.0671	0.1468	1.092 318
5	冠状动脉粥样硬化性心脏病	心功能Ⅲ级	0.0624	0.1211	1.079 876
2	高血压Ⅲ	冠心病	0.0519	0.1135	1.123 006

编号8的规则前件与后件说明，患有不稳定心绞痛的病人，在出现非胰岛素依赖型糖尿病的情况下发生冠状动脉粥样硬化性心脏病的概率最高。

5.7.2.3 特征抽取

对于合并症的研究，除了了解某两个合并症之间的关联关系以外，科研者往往还需要做更深入的研究。如需要了解多个合并症的组合模型，在实际临床案例中很可能有多个不同的疾病亚型，不同的疾病亚型在合并症上表现不同，所以经由合并症研究疾病亚型是一个方法。这时，我们采用特征抽取技术来解决这一类问题。特征抽取技术的特点是探寻数据所表达的通常由高维变量组成的事物的主要特征。特征抽取技术也是一种降维的技术，就是将高维变量进行线性合并或

非线性合并，使得高维变量合并到低维变量，从而更有效地、更清晰地表达事物的主要特征。合并后的变量是新产生的变量，它不同于原有的变量，往往更具有代表性。如上例子，一组大约2000例的心绞痛病人，他们的合并症有上千种，但每一个病人的合并症只可能是其中的几种。类似交易型数据，我们把某一病人的几种合并症看作是一组货篮数据，它们具有并发性。而下一步，又将这总共上千种合并症看作是上千个特殊的变量，变量的取值是有或无。然后，我们将这1000多个变量进行特征抽取，看看能压缩成多少个新的组合变量，而压缩后的每一个变量就是一个具有医学意义的合并症模型。它给人们提示了这一人群在某一方面的医学特征，如免疫学方面、内分泌代谢方面、血管系统方面等，而不再是更细微的医学特征。这里的特征抽取模型采用的算法是非负矩阵因子分解NMF算法。挖掘的结果如下：选取了2169例患有不稳定心绞痛的案例，运用特征抽取技术对其进行分析，得到排序在前七个的特征，结果如表5-7所示。

表5-7 特征抽取

特征	特征总参数	特征排序
5	.053816345033859554662758	1
26	.051795403879748618766556	2
49	.05170465393359742492492	3
44	.051342110751736022035855	4
27	.046175818817979017548347	5
32	.045977571577092711106727	6
23	.044168557525068571388803	7

其中，特征5中包含14种合并症分量：脑血管病后遗症、2型糖尿病性肾病、高血压Ⅱ级高危等，合并症分布如表5-8所示。

表5-8 合并症分布表

特征	疾病名称	特征系数
5	脑血管病后遗症	0.020 251
5	2 型糖尿病性肾病	0.016 189
5	高血压Ⅱ级高危	0.008 811
5	2 型糖尿病	0.006 999
5	慢性肾功能不全，氮质血症期	0.001 565
5	陈旧性广泛前壁心肌梗塞	1.09E-07
5	腰椎间盘突出	7.24E-08
5	陈旧性肺结核	2.73E-08
5	慢性气肿型支气管炎（慢支伴肺气肿）	2.56E-08
5	2 型糖尿病性周围神经病变	2.26E-08
5	陈旧非 Q 波心梗	3.65E-09
5	高血压Ⅲ级高危	8.1E-10
5	短暂性脑缺血（TIA）	3.34E-10
5	老年性白内障未熟期	1.12E-10

其中脑血管病后遗症分量的相对权重为0.376，2型糖尿病性肾病分量的相对权重为0.300，高血压Ⅱ级高危分量的相对权重为0.164等。从该特征的合并症情况可以看出，不稳定心绞痛与脑血管病后遗症同时出现的概率最高，其次是2型糖尿病性肾病、高血压Ⅱ级高危等疾病。

5.7.2.4 特性选择

医学研究中另外一个普遍的需求，就是了解哪些医学变量指标对某些疾病具有很强的相关关系。我们采用数据挖掘的方法，可以很好地解决医学科研中类似这样的需求。具体的技术是特性选择模型，更形象和更便于理解的叫法是因子权重算法，也就是在大量的医学指标变量中，探索哪些变量因子对特定的疾病具有较强的影响权重，哪些

变量因子对这些疾病具有较弱的影响权重，哪些变量因子对这些疾病的影响权重属于混淆因子。因子权重算法的特点正是发现不同的变量因子与目标变量之间的相关关系，它属于有监督的学习模型。例如，从诊断上有两种极易混淆的疾病X病和Y病，它们在很多症状、临床指标甚至临床表现中具有相似的特征，现有的医学诊断方法较难分清二者，我们希望能从很多客观的医学指标，如生化指标中发现两者的差别。这时，我们需要采用因子权重算法加以探索。

例如：患有冠状动脉粥样硬化性心脏病与患有脑梗死的患者共计424例，对所有化验项经因子权重模型分析，结果如表5-9所示（由于化验项目过多，故只选取了其中的五个实相关的因子、两个虚相关因子和两个零相关因子）。

表5-9 因子权重：疾病结果

检验项	权重	排名	检验项	权重	排名
腺苷脱氨酶	0.002 546	2	细胞总数	0	39
血氧饱和度	0.0003	11	脂肪球	0	39
嗜中性粒细胞	0.000 206	14	血小板	-0.0217	198
肌钙蛋白	0.000 186	15	淋巴细胞百分比	-0.02258	199
血红蛋白总量	0.000 185	16			

从表中结果可以看出，冠状动脉粥样硬化性心脏病与脑梗死的最大差异是腺苷脱氨酶，其次是血氧饱和度和嗜中性粒细胞等指标。冠状动脉粥样硬化性心脏病与脑梗死的细胞总数和脂肪球没有显著性差异，而血小板与淋巴细胞百分比属于混淆因素，无法作为相关指标。

在临床科研中，类似的需求还有探索不同的医学指标之间的相关

关系。也就是在大量的医学指标变量中，探索哪些变量因子对特定的医学指标（目标变量）具有较强的影响权重，哪些变量因子具有很弱的影响权重。例如：检查冠状动脉粥样硬化性心脏病病人的甘油三酯，总计1347例，与病人所做的227项生化检查指标做因子权重分析，结果如表5-10所示（仅选取了甘油三酯的七个实相关因子、两个零相关因子和两个虚相关因子）。

表5-10 因子权重：检验

序号	检验项	权重	排名
1	高密度脂蛋白胆固醇	0.393 08	1
2	总胆固醇	0.379 806	2
3	低密度脂蛋白胆固醇	0.359 77	3
4	A/G	0.165 74	4
5	总酸性磷酸酶	0.145 812	5
6	血清铁	0.097 677	6
7	血清镁	0.096 177	7
22	脂肪球	0	15
23	尿细胞管型	0	15
228	红细胞	-0.773 69	180
229	红细胞压积	-0.797 86	181

从表5-10中可以看出，冠状动脉粥样硬化性心脏病病人的甘油三酯指标与高密度脂蛋白胆固醇的相关性最强，而红细胞压积、红细胞对甘油三酯而言，属于混淆因素。

5.7.2.5 分类模型

在第5.7.2.4节的例子中，我们找到了辨别两种极易混淆的疾病X病和Y病的差异性生化指标。进一步地，我们希望通过差异性生化指标将这两种疾病分辨出来，这里的差异性生化指标就是用于分类的预

测因子。这时，就用到了数据挖掘中的分类模型技术。分类模型的特点是根据若干个预测因子，将数据类型为离散变量的目标变量进行分类。分类模型是有监督的学习模型。常用的分类模型算法主要有决策树算法、朴素贝叶斯算法、支持向量机算法、多元Logistic回归算法等。需要特别注意的是，在很多商业项目中，我们发现了不少目标变量的预测因子，它们对目标变量都具有很强的影响力，但预测因子之间的影响矢量却往往是不同的。对于影响矢量相似的预测因子，我们称之为同向因子，对于影响矢量不相似的预测因子，我们称之为异向因子。一组同向预测因子对分类模型的结果起着影响叠加的作用，一组异向预测因子对分类模型的结果起着影响抵消的作用。因此，在预测因子的组合选择上，我们还需要一个因子同向选择算法，来选定一组最佳的预测因子的组合。然后我们就可以应用这个预测因子组合，采用适当的分类算法进行分类运算了。当然，我们要通过模型的评估来评价这个预测因子组合的性能优劣。另一方面，对于不同的分类算法，可能有不同的预测因子组合，更能发挥分类模型的分类效果。

再例如一个代谢综合症研究的案例，研究者从医学专业角度认为十个医学指标与该病有关，并建立一个多元Logistic回归模型，试图证明其中哪些指标是代谢综合症的保护因子，哪些指标是代谢综合症的危险因子。于是，研究者运用统计学软件，将所收集到的代谢综合症的数据进行Logistic回归建模分析，结果发现统计学假设检验始终不能通过验证。后来他们借助IMRS中的疾病因子权重模块，对代谢综合症所有采集到的生化指标进行相关性分析。结果发现，他们原先

通过医学知识选定的十个医学指标中，有两个指标是该疾病的混淆因子，它们干扰了多元Logistic回归模型的建立。而在这十个预先选定的医学指标之外，还有一些指标与该病有较强的相关关系。于是，研究者对代谢综合症的影响变量做了调整，即去掉混淆因子，替换为疾病因子权重模块提示的相关性较强的因子，重新建立多元Logistic回归模型，结果一次性就通过了建模的所有过程和统计学检验。

5.7.2.6 异常侦测

在临床科研中，很多研究是探索一些特殊临床案例的共同特征。怎样高效地发现这些特殊病例，并且准确地将其进行归类，是科研中很重要的一件事。我们怎样用数据挖掘帮助科研者解决这个问题呢？数据挖掘中的异常侦测技术就是一个很好的方法。异常侦测是以寻找和划分数据中的离群点为目标，通常是一类划分。在商业项目应用中，异常侦测可以是无监督学习，也可以是有监督学习。有时我们可以将异常侦测看作是一个聚类分析的特殊模型，有时也可以将异常侦测看作是一个分类预测的特殊模型，这时分类只有两类，但两类的实例却相差悬殊，实例很少的一类就是异常侦测的离群点。需要注意的是，商业项目中的离群点，从业务角度或者专业意义上往往不是异常的，但它们的发生和出现却是异常的。在医学科研中，我们需要将给定维度下的大量数据，寻找出离群点来，而这些离群点不一定都具有相似的特征，往往有几个不同的特征，我们称之为异常亚型。例如，有296例肺炎患者的胆红素代谢（TBIL，DBIL）检验项数据，其中包

括间接胆红素、直接胆红素和总胆红素，进行异常侦测分析，得出38例异常数据，结果如表5-11所示。

表5-11 异常侦测结果

年龄	性别	序号	异常	置信度
91. 42	男	128 815	0	0. 952 54
12. 97	女	529 837	0	0. 952 54
89. 01	男	27 768	0	0. 952 54
94. 97	男	529 599	0	0. 845 789
65. 38	男	144 196	0	0. 845 788
84. 24	男	121 751	0	0. 845 779
94. 97	男	529 598	0	0. 842 746
90. 85	女	3505	0	0. 832 401
84. 24	男	121 750	0	0. 818 846
79. 36	男	29 487	0	0. 78 975
93. 03	女	133 153	0	0. 737 883
84. 24	男	121 749	0	0. 684 431
46. 65	男	72 444	0	0. 661 265
86. 04	女	56 083	0	0. 651 86
48. 73	男	197 764	0	0. 637 514

(续)

年龄	性别	序号	异常	置信度
76.71	女	52 572	0	0.634 099
10.53	男	131 610	0	0.598 502
4.77	男	102 275	0	0.598 15
91.42	男	128 816	0	0.596 355
60.73	男	18 307	0	0.587 099
91.42	男	128 817	0	0.576 175
79.36	男	29 489	0	0.575 014
79.36	男	29 486	0	0.572 927
60.73	男	18 309	0	0.554 92
69.41	男	78 903	0	0.553 721
89.01	男	27 769	0	0.553 065
94.97	男	529 597	0	0.533 629
79.36	男	32 169	0	0.527 522
70.52	男	45 037	0	0.521 346
84.24	男	121 748	0	0.520 045
79.36	男	29 488	0	0.518 316
84.24	男	121 744	0	0.515 918
85.42	男	89 133	0	0.512 208
86.47	男	24 244	0	0.510 408
85.67	女	65 850	0	0.507 974
92.33	男	35 434	0	0.504 561
82.02	女	140 865	0	0.502 158
93.27	男	514 935	0	0.500 08

从38例病人的生化指标可以看出，这些异常点的某些生化指标要明显高于正常值。例如序号为128815的病人的异常置信度达95%，观察病人的有关生化指标，发现总胆红素值为70.2，是296例中最高的。

5.7.2.7 回归预测

在临床科研中，还有很多研究试图发现由多变量指标组成的多元关系模型，尤其在那些复杂的慢性病中，有很多医学指标互动的系统性疾病往往可能蕴含着一个多元关系模型。对于这种情况，我们需要

应用数据挖掘中的回归预测技术。回归预测是利用一组证明是目标变量的相关因子，对连续型目标变量进行拟合预测的方法。回归预测常用的算法包括多元线性模型、多元非线性模型、神经网络模型、支持向量机模型，甚至是灰色理论模型等。需要注意的是，在很多商业项目中，我们虽然采集了足够大的数据，但不一定能得到满意的预测模型。一种可能是我们现有的技术尚不足以发现数据中蕴藏的规律，另一种可能是数据中根本没有蕴藏的规律。以当前的技术来讲，我们能发现的规律还很有限，而且这些规律还不都是显式模型，很多都是隐式模型。所谓显式模型，就是可以用现有的知识和技术表达出来的模型，如显式模型可以表达为一个数学公式，也可以表达为一个规则集等。所谓隐式模型，就是尚不能用现有的技术表达出来的模型，我们通常将它们称之为黑盒子，我们仅能证明它是存在的，甚至可以应用它，但它是什么我们尚不清楚。

临床科研中的例子，如选取患有冠状动脉粥样硬化性心脏病的病人，化验项为血红蛋白，共计1347例数据。首先以血红蛋白作为目标变量，与229项生化指标进行相关因子分析，结果如表5-12所示。

表5-12 相关因子

序号	检验项	权重	排名
1	平均血红蛋白含量	0.765 196 297	1
2	红细胞压积	0.716 408 37	2
3	单核细胞	0.685 863 033	3
4	红细胞	0.683 357 132	4
5	单核细胞百分比	0.657 451 722	5
6	嗜碱细胞百分比	0.597 471 56	6
7	淋巴细胞	0.591 979 371	7
8	平均血小板体积	0.498 335 253	8
9	血小板分布宽度	0.473 294 191	9

表5-13 回归预测表

回归参数	参数值
均方根误差	0.548 238
平均实际均值	159.2593
平均预测均值	159.3933
预测置信度	83.8880

血红蛋白决定了平均血红蛋白含量的值，两者的相关性虽然较强，但平均血红蛋白含量不能算作与血红蛋白独立的变量，因此不选择平均血红蛋白含量作为预测因子。我们选取了独立变量红细胞压积、红细胞、嗜碱细胞百分比、平均血小板体积和血小板分布宽度五种化验项作为血红蛋白的预测因子，建立了一个回归预测模型，得到的模型预测精度较高，如表5-13所示。

从表中可以看出，预测均值与实际均值两者相差不大，且预测精度为83.89%，说明该模型的预测准确度比朴素预测模型的预测准确度要好83.89%，还说明模型拟合得较好，预测精度较高。通过模型得到的预测值与真实值如表5-14所示。

表5-14 预测值与真实值

序号	血红蛋白	预测值	序号	血红蛋白	预测值
4262	177.8909	178.5581	33982	136.7636	135.8737
4263	171.5636	172.4484	35564	113.0364	113.4039
5079	151	151.7942	38913	136.7636	136.0745
6742	165.2364	165.1523	41808	171.5636	171.9861
12597	173.1455	173.6943	189122	192.1273	192.9924

表中的结果说明了预测值与真实值差异很小，模型预测精度较高。

5.7.3 智能医学科科研系统的评估方法

对数据挖掘应用系统的评估是很重要的，这决定着系统的可用性。数据挖掘应用的评估技术有很多，其中运用统计学进行验证是一个实践证明很有效的方法。这里有必要说明一下统计学与数据挖掘之间的关系。有些人常常把统计学和数据挖掘混为一谈，认为统计学就是数据挖掘、数据挖掘也是统计学。也有些人对这个问题纠结不清。我们认为数据挖掘与统计学是交集的关系，它们之间有很强的关系，但不是一个涵盖另一个。从本质上讲，统计学是从小样本数据分析，从而推测总体的特征；而数据挖掘则正好相反，是从尽可能大的数据集中直接寻找特征。另外一个区别是，应用方法上，数据挖掘多用于主动式地发现，而统计学多用于被动式地验证。从大量的商业项目实践中，我们体会数据挖掘探索与统计学分析很多时候结果是一致的，但数据挖掘的结果往往更优于统计学分析的结果，原因可能是，有时数据挖掘分析的数据比统计学分析的数据更大，因此数据挖掘得到的模型可能更精确些。因此，我们常常应用统计学对数据挖掘的结果进行评估和验证。

5.7.3.1 聚类分析的评估方法

聚类技术的目标是探索未知数据的分布规律，最终找到一组最佳分组，使得组间具有足够的相异性，组内具有足够的相似性。因此，我们可以运用统计学中的各组间差异分析方法，对数据挖掘聚类的分

组进行验证。统计学中也有聚类分析，而数据挖掘聚类与统计学聚类有很多相似之处。但二者的区别是，有些数据挖掘聚类可以处理很大的数据集，但统计学聚类通常只能处理较小的样本数据。统计学中常用的组间差异分析方法，包括T检验、方差分析、秩和检验等。我们首先运用这些统计学工具对数据挖掘聚类结果进行验证，看看分组结果是否好，也就是组间是否具有足够的相异性、组内是否具有足够的相似性。评估的第二步，是对几个不同的聚类模型进行比较，也就是综合评判它们的统计学验证效果，从中找到一个最优的聚类模型，它的综合统计学验证置信度最高。

例如，我们对先前的生化指标聚类模型做统计验证，利用聚类各组的欧式距离来判别各组之间的差距。选择病人的年龄和生化指标作为评估变量，如选取四组与五组有关变量的距离做秩和检验，计算出正态近似检验统计量 $Z=12.7618$ ，P值小于0.001，说明在95%显著水平下，两组距离存在差异。

5.7.3.2 关联规则的评估方法

由关联规则的定义中，我们可以看出它是从大量的、随机发生的并发事件中找到强关联的现象，使得在某事件发生的前提下，另一事件的发生具有很高的概率。因此，在关联规则算法中已经整合了统计学的验证。

如关联规则算法中的支持度，是指在一个规则中项集共同出现的

频度的指标，是一个在前件和后件中包含的所有项集的交易数与所有交易数的比率，用概率公式表达如下：

$$\text{Support (A implies B)} = P(A, B)$$

关联规则算法的置信度是指示在一个规则中项集共同出现的概率的指标，是在给定前件的条件下同时出现后件的条件概率，用概率公式表达如下：

$$\text{Confidence (A implies B)} = P(B/A) = P(A, B) / P(A)$$

关联规则算法的Lift是指给定的前件和后件随机同时出现的规则的强度，它提供了一个改进的信息，以提高给定前件下后件出现的概率。Lift的公式表达如下：

$$\frac{(\text{Rule Support})}{(\text{Support (Antecedent)}) \times \text{Support (Consequent)}}$$

详见第3章3.4节。

在实际商业项目中的关联规则模型中，我们常常对如下的统计指标感兴趣：

- 具有高置信度、高支持度、高Lift值的规则：

高置信度是规则的可信度高，高支持度是发生这种现象的实例数量多，说明关联规则的前后件的相关现象很普遍，且具有很强的必然

性，前后件的关系很可能像宿主与寄宿的关系，高Lift值是大量事件随机产生的强规则。

- 具有高置信度、低支持度、高Lift值的规则：

高置信度是规则的可信度高，低支持度是发生这种现象的实例数量少，说明关联规则的前后件的相关现象不普遍，但具有很强的必然性。这样的低支持度的规则往往更具有商业价值，高Lift值是大量事件随机产生的强规则。

- 具有中度置信度、高支持度、高Lift值的规则：

中度置信度是规则具有一定的可信度，高支持度是发生这种现象的实例数量多，说明关联规则的前后件的相关现象很普遍，但必然性稍差，前后件的关系非强耦合的关系，高Lift值是大量事件随机产生的强规则。

例如，从表5-6中可以看出规则编号8的置信度为0.91，说明这种规则的可信度高。支持度为0.10，说明在出现非胰岛素依赖型糖尿病的情况下发生冠状动脉粥样硬化性心脏病的案例较为普遍。Lift值为1.77，说明此规则具有较好的实际意义。

5.7.3.3 特征抽取的评估方法

特征抽取技术的原理，是将高维变量进行线性合并或非线性合并，使得高维变量合并到低维变量，从而更有效地、更清晰地表达事

物的主要特征。合并后产生了新的变量，它带有合并变量的不同程度的特征。我们应用统计学中的各种相关分析方法对特征抽取的模型进行验证，如双变量相关分析Person相关分析、Spearman相关分析、Kendall相关分析以及多变量复相关分析等。评估的第一步是应用适当的变量相关分析对抽取后的每个变量与目标变量做相关分析。第二步是应用多变量复相关分析对抽取后的所有变量作为一个变量组合与目标变量做相关分析。应用这种方法对不同的特征抽取模型进行评估，最后选出综合相关系数最高的模型作为最终模型。

5.7.3.4 特性选择的评估方法

特性选择（因子权重）算法的原理，是发现不同的观察变量与目标变量之间的相关关系，这里我们考察的是某一单个观察变量与目标变量之间的相关关系。因此与特征抽取的评估方法相似，我们采用统计学中的双变量相关分析方法进行验证。在商业项目应用中，特性选择有两种用途，一种是只需找出哪些变量与目标变量最相关，然后分别应用这些变量处理与目标变量的关系；一种是找出一组变量与目标变量最相关，然后用这一组变量作为目标变量的预测因子或分类因子去拟合或分类目标变量。对于第一种情况，验证比较简单，但对于第二种情况，验证就比较复杂了。首先将所有待测变量与目标变量进行相关性统计学验证，然后再对某一组合变量与目标变量进行多元复相关性统计学验证，然后对所有特性选择模型的评估结果进行比较，从而选出最优算法。

例如，IMRS中“因子权重-疾病”模块中，选取疾病变量（目标变量）的两个取值X病和Y病，将X病转换为1，Y病转换为0，计算疾病变量与腺苷脱氨酶、血氧饱和度、嗜中性粒细胞和肌钙蛋白的Spearman相关系数，运用IMRS统计学子系统计算，结果如表5-15所示。

表5-15 相关系数

系数	腺苷脱氨酶	血氧饱和度	嗜中性粒细胞	肌钙蛋白
Spearman	-0.171	0.213	0.367	-0.223
P 值	0.010	0.029	0.030	0.000

Spearman系数的P值均小于0.05，说明在95%的显著水平下，X疾病和Y疾病与腺苷脱氨酶、血氧饱和度、嗜中性粒细胞和肌钙蛋白的相关系数具有显著性。

再如，IMRS中“因子权重-检验”模块中，计算甘油三酯（目标变量）与高密度脂蛋白胆固醇、总胆固醇、低密度脂蛋白胆固醇和A/G的Spearman相关系数，运用IMRS统计学子系统计算，结果如表5-16所示。

表5-16 Pearson、Spearman系数

系数	高密度脂蛋白胆固醇	总胆固醇	低密度脂蛋白胆固醇	A/G
Spearman	-0.267 84	0.275 21	0.138 37	0.125 79
P 值	<0.0001	<0.0001	0.0060	0.0181

Spearman系数的P值均小于0.05，说明在95%显著性水平下，甘油三酯与高密度脂蛋白胆固醇、总胆固醇、低密度脂蛋白胆固醇和A/G

的相关系数具有显著性。

5.7.3.5 异常侦测的评估方法

对异常侦测的评估方法与聚类评估的方法类似，如使用统计学中组间差异分析方法T检验、方差分析、秩和检验等。但与聚类评估的方法不同的是，由于异常侦测模型得到的正常组与异常组数量差异巨大，为了统计分析的均衡，我们从正常组中随机选取少量具有代表性的样本数据，与异常组进行统计学分析，样本数据量选取原则是在数量级上与异常组相当的条件下尽可能的大。考虑更复杂的情况，有的时候异常组内的数据并不集中或收敛，而是分为几个异常亚型，如果把它们作为整体与正常组进行统计学比较往往存在偏倚。因此，有必要将几个异常亚型组分别与同一正常组样本数据进行统计学比较。最后将几个不同的异常侦测模型进行比较，综合评判它们的统计学验证效果，从中找到一个最优的异常侦测模型，它的综合统计学置信度最高。

例如，IMRS中“异常侦测”模块，通过计算异常组与正常组的欧式距离，来判断两者之间是否具有差异。选取与异常组等数量、置信度较高的正常组，计算结果如表5-17所示。

表5-17 欧式距离

异常	欧式距离	异常	欧式距离	异常	欧式距离	异常	欧式距离
0	110.6392	1	12.023 31	0	20.733 31	1	22.461 97
0	43.327 01	1	10.273 27	0	47.504 21	1	21.402 34
0	97.049 73	1	10.5366	0	102.3922	1	24.458 54
0	241.42	1	8.718 945	0	57.091 59	1	16.417 06
0	147.0093	1	7.754 998	0	15.135 72	1	15.434 38
0	142.2283	1	7.349 83	0	38.326 36	1	13.997 14
0	112.9973	1	19.973 48	0	48.512 47	1	18.170 86
0	40.512 84	1	19.257 73	0	26.269 56	1	10.322 79
0	93.866 39	1	15.839 19	0	11.967 04	1	18.3483
0	76.769 07	1	16.370 09	0	8.512 344	1	11.037 21
0	23.922 79	1	7.348 469	0	43.911 84	1	16.476 04
0	92.868 62	1	19.719 53	0	98.8958	1	13.490 74
0	17.130 38	1	20.376 95	0	12.5996	1	14.277 25
0	21.901 14	1	10.5366	0	56.100 71	1	16.305 21
0	10.907 34	1	24.879 31	0	24.7548	1	10.051 87
0	13.161 69	1	20.645 58	0	20.201 24	1	17.488 28
0	11.990 83	1	6.871 681	0	6.685 058	1	8.114 185
0	22.7785	1	18.553 17	0	8.318 654	1	12.4996
0	50.033 69	1	16.797 02	0	8.998 889	1	16.499 09

再进行秩和检验，判别两组数据是否来自同一分布，计算出正态近似检验统计量 $Z=4.2230$ ， P 值小于 0.0001 ，拒绝原假设，说明两组距离存在差异。

5.7.3.6 回归预测的评估方法

回归预测的评估方法与以上数据挖掘技术评估方法有所不同。虽然我们也可以运用统计学中的多元线性模型或多元非线性模型对数据进行拟合验证，但往往很有限，这是因为数据挖掘的回归预测模型大多应用了非统计学的方法，而挖掘到的预测模型不一定能用统计学中的多元线性模型或多元非线性模型来表达，甚至模型是一个黑盒子。因此，商业项目应用中，我们常常采用金标准验证的方法进行评估。

也就是采用测试数据评估训练好的不同模型，最终评优出最佳预测模型。下面，我们给出一个运用统计学中的多元线性模型，对应用支持向量机算法得到的回归预测模型进行评估的例子。

例如，对第5.7.2.8节中的回归预测模型进行统计验证。根据前面所选的目标变量和预测因子，利用统计软件做多元回归，得到模型结果为：

$$y=20.4078+190.09591x_1+10.10438x_2-3.52811x_3-1.94848x_4+0.73625x_5$$

1) 回归预测模型的决定系数 $R^2=0.8847$ 。通常情况下， R^2 大于0.7就认为模型是可取的，说明该回归模型拟合程度较好。

2) 模型整体显著性检验

原假设为系数全为零，备择假设为系数不全为零，

$F=825.64$ ，拒绝原假设，认为系数不全为零，模型具有统计学意义。

3) 系数显著性检验

对5个变量进行T检验，取 $\alpha=0.05$ ，计算结果如表5-18所示。

表5-18 系数显著性检验

变量	常数项	x_1	x_2	x_3	x_4	x_5
T 值	6.12	14.90	8.56	-4.13	-5.21	4.47
P 值	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

从计算结果可以看出，在95%的显著水平下，常数项和5个自变量的系数显著不为零。

5.7.3.7 分类模型的评估方法

对分类模型的评估方法，我们应用基于统计学的朴素贝叶斯算法验证非朴素贝叶斯的分类算法，例如决策树算法、支持向量机算法、多元Logistic回归算法等。但基于统计学的朴素贝叶斯算法有它的局限性，它的前提条件是所有的分类因子必需是独立的，分类因子之间不具有相关性，否则验证的可信度会大打折扣。因此，我们在用朴素贝叶斯算法进行验证之前，首先要对分类因子做两两相关性验证，正如第5.7.3.4节特性选择的评估方法那样。如果分类因子的独立性比较好，则我们应用朴素贝叶斯算法进行验证，如果不好，我们不得不采用类似第5.7.3.6节中的回归预测评估方法，用金标准验证的方法进行评估，即运用已知的测试数据评估训练好的不同模型，最终评优出最佳预测模型。

5.8 智能医学科研系统的应用价值

以上章节比较系统和完整地阐述了数据挖掘医学应用的原理。基于这些原理和方法，我们最终成功完成了数据挖掘医学应用的一个实现，即开发、建设并应用智能医学科研系统。我们对该系统总结了如下几个特点：

- 主动式医学规律发现，帮助临床科研不断创新。
- 探索超大样本蕴含的医学知识，自动揭示科研靶向。
- 协助高水平复杂医学科研，攻克重大科研难题。

如图5-15所示，是对智能医学科研系统的一个形象化的概括和总结。



图 5-15

通过对智能医学科研系统在实际应用中的观察和总结，我们得出了如下一些结果。首先，将医院的医院信息系统、实验室信息系统、影像信息系统、病案系统，以及体检系统等信息系统中有关医学临床的海量数据进行整合，以集成的电子化数据形成医学科研和临床分析的专用数据仓库。然后，基于该数据仓库及配套的智能化科研分析工具，可以达成以下目标：

- 完成某些医学科研的全部过程。
- 完成某些医学科研的主要过程。
- 替代、简化、提升某些定制化科研的部分过程。
- 加快科研进程，提高科研质量。
- 完成用于指导临床的医学分析任务。

最后，我们将用户应用智能医学科研系统的经验和体会总结为多、快、好、省四个字。

1.“多”，指科研工作收益多

· 科研线索多，科研思路多，科研立题广泛、新颖。智能医学科研系统提供的数据探索工具可以通过知识发现技术，和大量的智能分析报表来揭示各种疾病的特征，以及疾病、检验项、用药之间的相互关系，从而发现日常工作中不易发觉的、新的、有趣的现象，为新的科研立项拓展思路，并为相应的课题立项做好准备工作，使临床科研

有更多的创新性。

- 科研样本种类多、样本量大。由于智能医学科研系统建立了真正的临床科研数据仓库，使科研者可以轻易地获取基于十几年积累的大量临床样本数据，利用智能医学科研系统的智能样本筛选工具，科研者可以广泛获取各类科研设计所需的样本数据，基于大样本的高水平科研更加容易了。

- 科研分析手段多。智能医学科研系统提供了所有常用的数理统计工具，还具有医学统计智能化，并提供了一般统计软件不具备的智能医学数据挖掘工具，因此，科研者可运用更强大的分析手段，使深层科研分析能够广为应用。

- 课题多，成果多。由于以上几个优点，医院还可以申请到更多的国家级课题和省市级课题，再加上智能医学科研系统提供了全在线共享式（院内局域网）的智能科研平台，可以支持多课题的并发研究，因此，医院课题可能成倍地增长，最终获得的成果和论文也会成倍地增长。

2.“快”，科研工作进程快、获取各类成果和论文速度快

- 预试验快。智能医学科研系统提供了试验数据与统计分析工具的高度集成，因此，各种科研预试验可以随时地、大量地、反复地进行，各类科研预试验工作可以更快、更好地完成。

- 科研设计快。智能医学科研系统为科研者事先定制了大量常用

的临床医学定量分析和定性分析，科研者只需了解并适当地选用所需的科研设计即可，从而加快了科研设计阶段的工作。

- 样本数据收集快。智能医学科研系统为科研者提供各类专业级的定量、定性分析样本筛选工具，向导式样本筛选工具，以及智能化自定义样本筛选工具，因此，样本数据收集速度大大加快，从通常的几周至几个月的周期加快为几分钟、几小时。

- 统计试验分析快。智能医学科研系统提供的统计工具是按照经典医学定量、定性分析设计的智能式统计流程，而且可以直接分析样本筛选工具得到的样本数据，无需向其他统计软件那样将数据反复整理、导出导入，因此，统计试验分析阶段的工作的速度提高了几倍至几十倍。

- 科研成功率高、获得成果快。智能医学科研系统不仅提供了可以提高科研成功率的数据探索工具、样本筛选工具和智能统计工具，更重要的是，提供了智能医学数据挖掘工具，这一基于大量数据的主动式分析工具可以直接有效地获取科研靶向甚至科研成果，可以使基于数据挖掘成果的科研设计和统计试验获得更高的成功率，从而极大地缩短了获得科研成果的周期。

3.“好”，科研质量好，科研水平高

- 科研立意高新。智能医学科研系统提供的数据探索工具从大量的数据中总结出题材广泛的临床医学特征，解决了传统科研立题瓶颈

问题，使立题更具新颖性，为高质量、高水平的科研奠定了基础。

- 科研方法专业。智能医学科研系统提供的样本筛选工具是严格按照科研“三要素”、“四原则”设计的专业医学科研设计和样本选取工具，从而保证了科研的质量。

- 统计设计工具先进。智能医学科研系统提供的智能统计工具严格符合经典定量分析和定性分析的完整统计方案，并充分考虑了数学在医学科研上的适用性，使用户不必因统计设计失误或误用统计方法而使科研功亏一篑。

- 科研水平提升。智能医学科研系统提供的专用智能医学数据挖掘工具，突破了传统医学科研手段，使高层次的临床医学科研更易于普及，也大大增加了科研者重大科研发现的可能性。

4.“省”，医院总体科研成本省

如果建设了完整的智能医学科研平台，医院每年用于支持院级的各类科研经费可以相应大大减少，从而节省了医院总体的科研成本，这样长远节省下来的科研经费是相当可观的。

传统的科研是从科研设计开始，经过预实验、实验、收集数据、分析、得出结果，有花费时间长且成本高、收集数据量小且不准确、结果难以预料等多种缺点。医学是一门实践的科学，很多的知识、定理、规律都是来自于临床、应用于临床。我们相信，医学上还有太多未经发现的规律，很可能就隐藏在海量的医学数据中，传统的手段是

力不从心的。应用智能医学科研系统，可以大大简化科研工作，提高效率，使科研者快速获取高置信的科研成果，甚至重大科研发现。

一项新技术的采用，往往意味着全新的方向。如同伦琴射线在医学上的应用，开创了全新的医学视角一样，随后的CT、MRI、B-US、PETS等新技术的采用一次次地推动了医学的发展，扩展了医生的视野。如今，影像学已经是医学不可缺少的组成部分，当年伦琴射线引入医学的时候，一定不会想到今日的局面。而以数据挖掘、数理统计、数据仓库等技术为核心的智能计算技术引入医学领域后，就像当年伦琴射线引入医学，它将透视医学潜在的知识与规律，加速医学科研过程，开创了全新的医学科研视角。

5.9 本章小结

目前，数据挖掘技术现已应用于许多行业中，不同行业中的应用方法有相似之处，也有很多不同的地方，这些主要是由各行业固有的行业规律和特色所造成的。如果想把数据挖掘技术成功地应用到某个行业中，就必需深入剖析和了解该行业的业务规则和专业知
识，并将数据挖掘的知识和技术与之有机结合起来，摸索出具有该行业特色的数据挖掘应用规律和体系，并可能创新出许多该行业特有的应用技术和方法。只有将数据挖掘的知识和技术与该行业的结合上升到更高的层面，甚至上升到数据挖掘在行业应用的理论高度，才会在该行业中更好地发挥数据挖掘的作用，达到令人满意的、可产生经济效益的、更好的应用效果。

第6章 数据挖掘应用系统的开发

上一章我们介绍了医学方面、偏重科研方向的一个数据挖掘应用解决方案，包括从应用需求的产生、解决思路、系统设计、应用实现、效果的评价与总结等涉及所有环节的完整过程，我们称之为智能医学科研系统（IMRS）。该系统包含了从业务数据的探索性分析到数据的专业性选取，再到数据挖掘的深入分析，以及大量应用各种统计学方法对数据挖掘结果进行验证等各个环节，是一个比较典型的行业数据挖掘应用案例。这一章，我们将重点剖析一下，IMRS数据挖掘应用系统的主要功能开发要点，包括技术架构、数据变换、模型创建、功能实现等。我们先简要讨论一下数据挖掘应用系统的意义，然后详细讨论IMRS系统设计，接下来我们再具体剖析IMRS的几个重要模块的开发方法，包括异常侦测模型的开发、特征抽取模型的开发、智能统计模型的开发，最后介绍一下IMRS的算法开发。

6.1 数据挖掘应用系统的意义

在介绍数据挖掘应用系统的开发之前，我们首先探讨一下开发、建设数据挖掘应用系统的意义。

数据挖掘常用的应用方式为工程性的应用或定制化的应用，其特点是应用需求十分明确和具体，数据挖掘的应用围绕着具体业务需求和业务问题来开展，因此，这类应用都是针对业务特定需求的，一般通用性较差。在这里，我们要讨论的是为了解决某一类的业务需求，而开发的类似平台性质的数据挖掘应用系统，它们具备相当的通用性，而且又不同于工具类的数据挖掘系统。数据挖掘工具类软件提供的往往是比较纯技术的功能，比如数据的整理、各种数据挖掘算法，以及数据挖掘流程等，一般没有行业的针对性，更不提供专业的知识。但是数据挖掘应用领域中，最大的瓶颈正是数据挖掘技术与行业知识和专业需求很难有机地结合起来。数据挖掘在商业应用中效果不理想的原因，一是从事数据挖掘技术的人员不能深入理解行业应用的专业知识，二是行业领域从事业务的专业人员不太理解和掌握数据挖掘技术。另外，数据挖掘工作本身也极具挑战性。因此，在现实项目中，我们看到虽然数据挖掘技术人员和业务专业人员的热情都很高，对数据挖掘应用结果的期望也很高，但最后，在业务工作中真正依赖于数据挖掘应用提供的结果做业务指导的案例并不多见。

数据挖掘应用系统试图解决应用效果欠佳的瓶颈问题，即将数据挖掘技术与行业知识和专业需求有机地结合起来，使得应用人员在

必深入了解数据挖掘具体技术的情况下，就能很好地开展专业数据探索和挖掘工作，而数据挖掘应用系统一定是很“专业化”的。从功能上，数据挖掘应用系统既要提供适当的数据挖掘算法、有针对性的数据处理过程，还要提供一定的专业应用知识和经验。这样，就使得数据挖掘应用系统开发的要求和条件相当高，相应的开发团队既需要对相关行业和专业知识的深入理解，并有相当专业挖掘成功经验的技术专家，也仍然需要在该行业领域中有很深造诣的、对数据挖掘技术的应用有很好理解的专家。两方面的专家密切配合，互相学习和借鉴，共同研究和开发，才能搞好数据挖掘应用系统的开发工作。

行业应用系统最大的作用之一是使得知识的获取变得方便。当用户有了什么想法，立刻就可以去验证这个想法，而不是费事费力地获取数据、整理数据、亲自写作算法或使用工具包处理数据，最后还要借助其他技术工具来验证评估结果。这一系列环节消耗人们的时间、精力，足以使一些有价值的商业想法还未来得及挖掘出来，就由于时间的延误而失去了它应有的价值，并可能错失了商机。

行业应用系统把商业数据挖掘活动的这些环节集成在了一起。数据是收集好了的，放在一个不断自动增长的数据仓库里面，通过面向行业设计的数据筛选器可以很方便地取出来。数据在数据仓库ETL阶段时已经经过了初步的技术性整理和清洗（详见第2章）；在使用专用数据筛选器的时候，又再次从行业角度进行了专业清洗、整理；最后在数据展现给最终用户时，使用者还可以进行一遍人工选择。专门设计的这个人工环节，令用户可以完全控制让什么数据进入分析引

擎。用户在将数据送入分析引擎之前，可以把精心筛选的数据在应用系统中保留一份副本，并按照数据管理的严格程度，导出应用系统之外或只能存在服务器上面。

数据经过专业筛选之后，用户可以使用集成在应用系统之上的专门为行业数据进行优化的分析引擎进行分析，用户不必亲自写算法或再利用各种工具软件，也不必再花费大量时间深入学习数据挖掘的理论知识、开发技术，也不用在每次搞数据挖掘应用时，都需要数据挖掘专业技术人员的支持和定制化服务。所有的数据挖掘技术对于用户而言都是透明的，运行在后台的。当然应用系统开发人员需要对用户进行相应的培训，使用户了解这些数据挖掘的引擎都可以干什么，基本原理是什么，能解决哪些商业问题。用户只需把他们既有限又宝贵的时间和精力集中到解决业务问题上，以便及时把握商机。

6.2 IMRS系统设计

智能医学科研系统（IMRS）是一个专门为临床科研工作设计的智能化软件工具，它的主要功能涵盖了临床科研工作的主要环节，包括科研数据的探索分析、样本数据的设计与选取、科研数据的预实验分析、科研统计方法的设计与验证、科研规律与数学模型的探索等。由于本书的主题是数据挖掘的应用，所以我们将重点介绍与数据挖掘紧密相关的一些内容。我们首先介绍一下智能医学科研系统与数据挖掘相关的部分模块，如图6-1所示。在这里，“数据挖掘”子系统是用于临床医学科研发现新规律、新靶向的科研创新工具；“统计分析”子系统一方面可以作为临床科研的统计实验设计工具，另一方面可以作为数据挖掘子系统的部分验证工具；“其他分析”子系统是用于从临床业务系统中抽取数据挖掘所需要的数据、数据挖掘模型的维护、数据挖掘算法的定制化，以及数据挖掘用户的管理等。

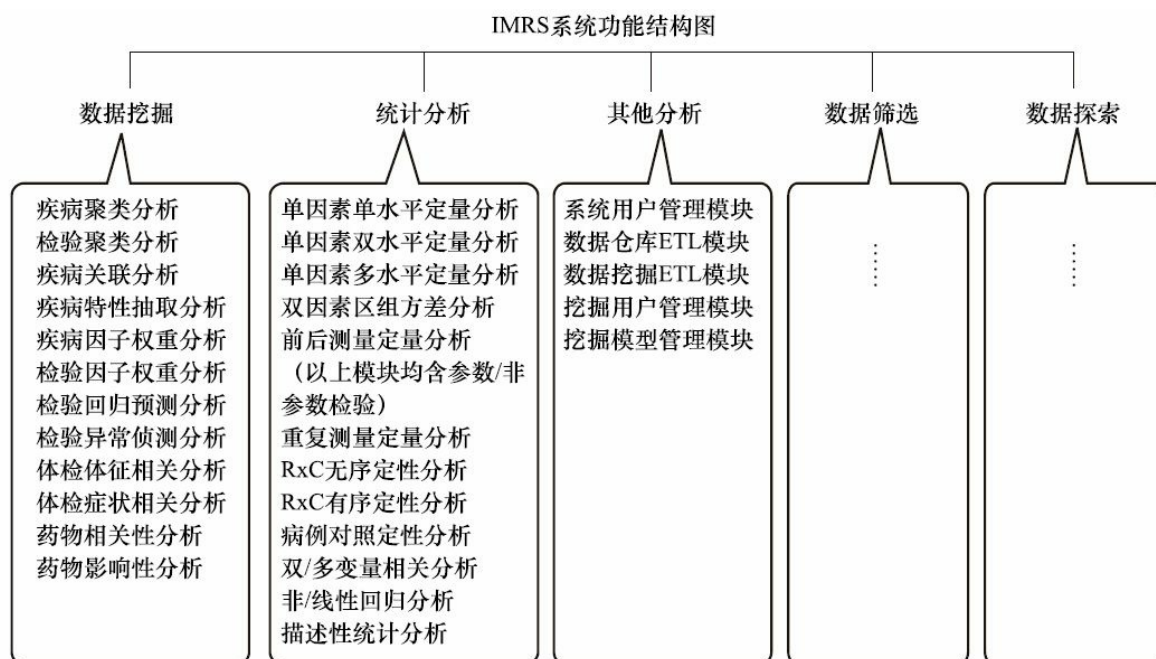


图 6-1

数据挖掘的目的，是通过与各种专业领域的结合，并通过提供新方法、新工具改进原有的流程和方法，帮助业务的发展。数据挖掘的实践是离不开行业的，而每个行业都有特殊之处，数据挖掘需要针对这些特性而定制。数据挖掘是数据与挖掘并重的一门应用科学。在设计挖掘之前，首先要了解数据。因此，如果对医学科研数据挖掘应用系统进行设计，首先需要了解系统所需的数据源。

6.2.1 对数据源的分析

我们以一家典型的综合性三甲医院的主要业务系统作为案例，分析一下临床数据的主要构成。

6.2.1.1 业务系统的数据源构成

医院的临床业务系统主要包括：医院信息系统（Hospital Information System, HIS）、实验室信息系统（Laboratory Information System, LIS）、图像存储与传输系统（Picture Archiving and Communications System, PACS）、病案管理系统（Medical Records System, MRS）、电子病历系统（Electronic Medical Record, EMR）、体检信息系统（Physical Examination Information System, PEIS）。

智能医学科研系统（IMRS）所采集的数据主要来源于这些业务系统：

- HIS系统中的住院系统。HIS是指利用计算机软硬件技术、网络通信技术等现代化手段，对医院及其所属各部门的人流、物流、财流进行综合管理，对在医疗活动各阶段产生的数据进行采集、储存、处理、提取、传输、汇总、加工生成各种信息，从而为医院的整体运行提供全面的、自动化的管理及各种服务的信息系统。智能医学科研系统主要采集HIS中住院病人的临床数据，用于临床分析和医学科

研。因为住院病人的临床数据比较完整，它包括病人在院期间的用药、治疗、手术、护理、费用等各方面临床信息，反映了住院病人医疗的整个过程。

· LIS系统。LIS是一类用来处理实验室过程信息的软件系统。实验室信息是比较客观的又是结构化的临床数据，包括病人在院期间的各种化验的结果，如血液学、化学、免疫学、外科病理学、解剖病理学和微生物学等方面的检查结果。智能医学科研系统主要应用大量的LIS数据进行医学规律和靶向的探索、挖掘。

· PACS系统。PACS即图像存储与传输系统，是用于管理医院中医疗设备如CT、MR等产生的医学图像的信息系统。它把医学图像从采集、显示、储存、交换和输出进行数字化处理，最后实现图像的储存和传送。医学图像信息是多样化的，如B超扫描图像、彩色多普勒超声图像、核磁共振（MRI）图像、X-CT图像、X线透视图像，各种电子内窥镜图像，显微镜下病理切片图像等。智能医学科研系统主要采集住院病人的PACS检查报告，用于帮助临床科研者深入了解、研究病人的医学状态，以便筛选符合科研设计要求的临床数据。

· MRS系统。MRS是集病人病例录入、查询、管理等一体的医院病历管理软件系统。智能医学科研系统采集已出院病人的病案信息作为主索引数据，如住院病案首页信息，包括姓名、性别、出生日期、入院日期、出院日期、入院科室、出院主诊断、各种合并症诊断等信息。

· EMR系统。EMR是医院通过电子病历以电子化方式记录患者就诊的信息，我们称作电子病历系统，包括：病历首页、病程记录、检查检验结果、医嘱、手术记录、护理记录等，其中既有结构化信息，也有非结构化的自由文本，还有图形图像信息。智能医学科研系统主要采集电子病历系统中结构化的数据，用于帮助临床科研人员全面了解研究对象的病程病史，以便筛选、分析符合科研设计要求的临床数据。

· PEIS系统。PEIS通过软件实现提取相关检测仪器的数据，并将体检结果存储到计算机系统中，我们称之为体检信息系统，其业务主要包括健康体检、从业人员体检、职业病体检等，它通常是独立于医院其他临床信息系统的。智能医学科研系统采集体检信息系统的数据，一方面用于临床科研中的对照组，另一方面也用于对健康人群和亚健康人群开展医学数据挖掘和科研分析。

以上这些信息系统的数据存储环境及数据来源如表6-1所示。

表6-1 数据源构成

数据源/数据仓库	数据库环境	涉及科室
数据源：住院系统	SQL Server 2005 Windows Server 2003 独立数据库	内、外科临床等
数据源：实验室系统	SQL Server 2005 Windows Server 2003 与体检系统共享数据库	化验室
数据源：影像系统	SQL Server 2005 Windows Server 2003 独立数据库	放射科
数据源：病案管理系统	SQL Server 2000 Windows Server 2000 独立数据库	病案室
数据源：电子病历系统	XML 文件	门诊科室
数据源：体检系统	SQL Server 2005 Windows Server 2003 与实验室系统共享数据库	体检中心
合计	5 个数据库分布在 4 个服务器上	医院大部分临床科室

6.2.1.2 业务系统的数据模型分析

医院的主要临床业务系统的数据模型，通常属于OLTP类型的临床数据模型，它的特点是业务关系复杂，数据的实时性强、并发量大，主要用于代替手工实现业务流程。OLTP类型的临床数据系统主要用于日常的业务操作，因此也称为业务操作系统（Operational System），这种系统不优先支持数据查询、数据分析工作。现在，我们简要分析一下临床数据模型的主要数据关系，如图6-2所示。

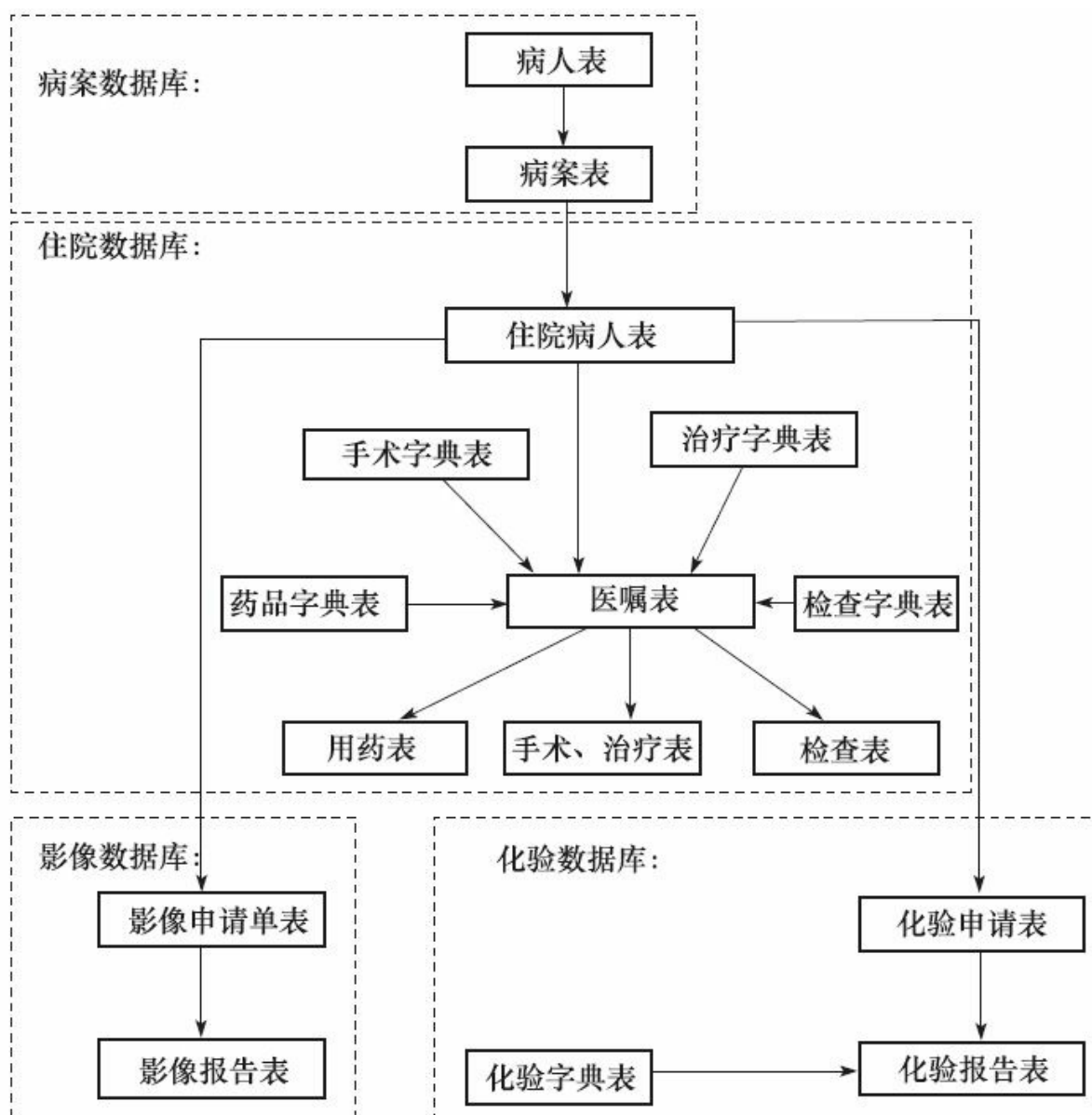


图 6-2

病案系统MRS的数据位于该数据模型的顶层，主要包括病人信息表和病案信息表。病人表作为病案表的父级表，病人表与病案表之间可能是一对一（一个病人一次住院）、一对多（一个病人多次住院）的关系。而且二者是依赖关系，即对于病案表中的每一条记录必有一条有关它的属性记录存在于病人表中。

病案系统MRS的数据与下一级的住院系统HIS衔接。病案表只包括已出院病人的信息，它作为住院系统的住院病人表的父级表，与住院病人表是一对一的关系，而且二者是非依赖关系，即对于住院病人表中的每一条记录并不一定有一条有关它的属性记录存在于病案表中，例如现住院病人就没有一个有关他的病案信息存储于病案表中，因为病人只有出院以后才会有病案人员将住院期间的临床信息整理成病案信息。

HIS的住院系统以住院病人表为父级表，在住院系统内与病人医嘱表连接，在住院系统外与实验室系统LIS的化验申请表、影像系统PACS的影像申请表连接。它们表间的关系大多为一对多的关系，反映了某个住院病人在住院期间的多次各种医疗活动，如病人做各种物理检查（核磁、CT、X透视、B超、脑电图、心电图等），病人做生化检查（验血、验尿、免疫学检查、病理学检查、分子生物学检查等），病人接受各种药物治疗，病人接受各种物理治疗，病人接受各种手术、治疗等。

6.2.2 数据挖掘应用系统IMRS的总体设计

通过以上对医学临床业务系统的分析，我们了解了主要数据的来源、功能，以及它们之间的关系。下面，我们需要将这些数据源的数据进行整合，对于这些独立的临床业务系统，以全新的设计将分散的业务系统产生的不一致的数据进行整理、变换、集成，整合得到全面、高效、一致的信息。通过建设医学科研专用的数据仓库，可以形成专业的、完整的、不断增长的、可重复利用的、超大规模的临床科研数据仓库，为更多的临床科研和更高水平的临床科研提供重要的数据资源。下面，我们将从IMRS的数据源开始，遵循着数据的“流动”方向，经过整合、变换、准备、应用等各个阶段，来详细介绍一下IMRS各个阶段的设计，包括：

- 异种数据连接的架构设计
- 数据仓库的ETL设计
- 数据仓库的数据模型设计
- 数据挖掘的API工具选择
- 系统架构设计

6.2.2.1 异种数据连接的架构设计

相对于基于前台业务的OLTP在线事物处理系统，智能医学科研

系统IMRS是一个后台分析型的应用系统，它的数据模型属于用于分析的数据仓库类型。IMRS的数据运行平台可以有多种选择，如操作系统可以是基于Unix或Linux的，也可以是基于Windows的；数据库系统可以选择Oracle、DB2、SQL Server或其他大型数据库系统等，各自有不同的优缺点。

在我们这个案例里，所有的现有业务系统都是采用基于Windows的SQL Server（以前用过DB2和My SQL），有32位版的，也有64位版的，电子病历系统的数据主要存储在XML文件中。因此，为了客户信息化管理和系统集成的方便，我们采用了基于Windows的Oracle数据库系统。

由于智能医学科研系统IMRS是基于数据仓库的分析型应用系统，它对系统资源的要求较高，因此我们选择64位的环境平台，即64位Windows Server操作系统和64位版的Oracle 10g/11g。业务数据从32位版或64位版Windows的SQL Server到数据仓库的64位版的Oracle的异种数据库的连接，我们选择基于Windows的ODBC和Oracle Client。

对电子病历系统的XML数据，我们采用Oracle的XDB将XML数据抽取、转换到Oracle的关系型数据库表中。数据从业务系统经过ETL（概念见第5章）数据变换后，再传送到数据挖掘应用服务器中。

在数据挖掘应用服务器中安装了64位版的Oracle数据库用于建立医学科研专用数据仓库（也叫数据集市Data Mart），同时还安装了应

用服务器软件和Web服务器软件，如图6-3所示。

方案特点：整体架构为64位Windows平台。采用一台高档服务器，安装64位Oracle for Windows作为连接异种数据库的数据仓库，与基于32位Windows的SQL Server数据源直接互连，再通过Oracle Database Link传送到64位的数据挖掘数据库中。基于64位Windows上的Oracle数据库最大有效可用内存可突破32位操作系统3.25GB的限制，基本配置16GB即可，系统性能好，尤其随着数据量的不断增加，系统有很好的可扩展性。

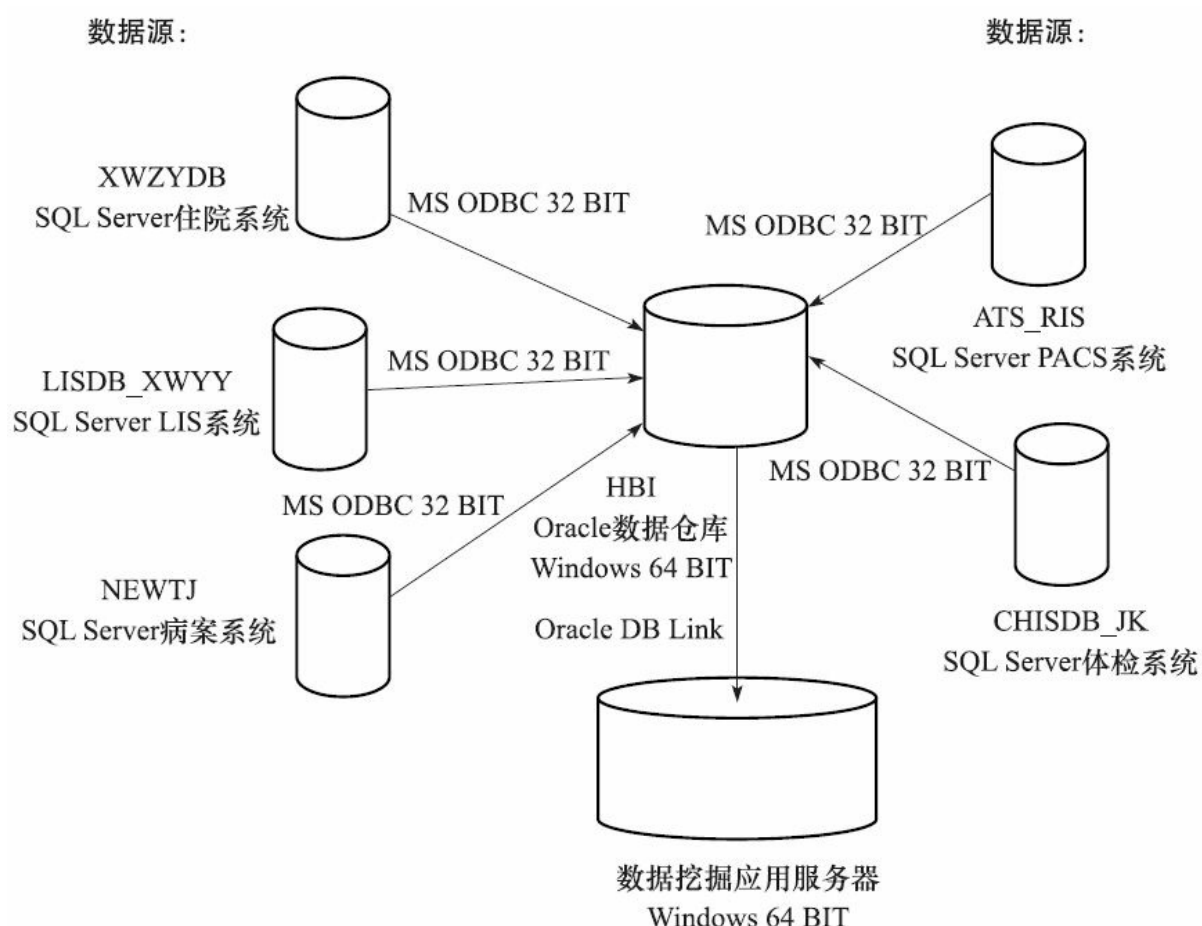


图 6-3

6.2.2.2 数据仓库的ETL设计

做好数据挖掘应用，我们不仅需要对所有的数据源、业务数据关系有深入的了解，而且应当对数据从业务系统到数据仓库的抽取和变换的过程有全面了解。因此，我们有必要简要叙述一下数据仓库的ETL设计，如图6-4所示。

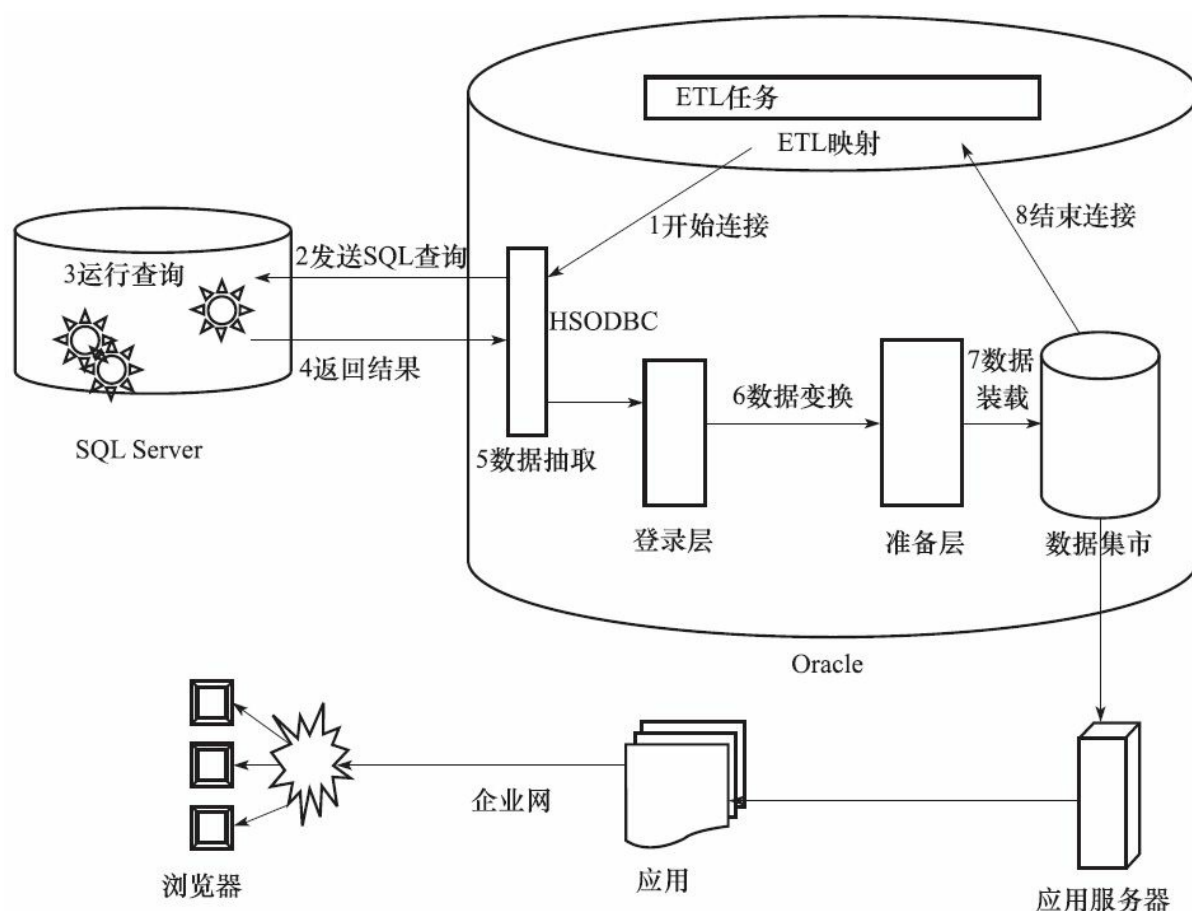


图 6-4

1) 首先，在作为数据仓库的Oracle数据库中配置异种数据库互连网关，如基于MS ODBC建立Oracle HS ODBC，并与业务系统SQL Server数据库（通常情况下是备份数据库）建立远程数据库链接（DB

Link)。

2) 然后从数据仓库中发送SQL查询命令到业务数据库中，业务数据库将查询到的数据经远程数据库链接传送到数据仓库的登录层（Landing Layer），即用于存放没有经过任何处理的业务数据。这一步就是ETL中的数据抽取（E-Extract）。

3) 将登录层（Landing Layer）没有经过任何处理的业务数据经过各种变换，加工处理成数据仓库需要的数据形式，然后传送、存放于数据仓库的准备层（Staging Layer），即用于清理、变换原始数据，并存储变换后的数据。这一步就是ETL中的数据变换（T-Transformation）。

4) 再将准备层（Staging Layer）的数据装载到数据仓库的数据集市层（Data Mart Layer），即用于存储为分析应用提供可分析的数据。这一步就是ETL中的数据装载（L-Loading）。

5) 以上2)至4)的步骤，通过在数据仓库内建立ETL Mappings，即数据变换程序，并由设定的ETL Jobs（数据变换程序）自动运行和控制任务，来定时地自动驱动执行或不定时地人工驱动执行。

6) 存储在数据集市层（Data Mart Layer）上的数据，经过应用程序的调取展现在应用服务器（Application Server）上。

7) 最后，展现在应用服务器上的数据分析结果，经过企业内网

络（Intranet）传送到终端用户的浏览器中。

6.2.2.3 数据仓库的数据模型设计

前面分析了基于OLTP的临床数据模型，它主要是反映复杂的业务关系，主要性能需要满足大量的并发性操作，在数据库中主要为DML操作，包括数据插入、数据增加和数据删除。而数据挖掘应用需要的是用于分析的数据类型，一般情况不能直接利用OLTP的业务数据，而是需要将它们转换成用于分析的数据形式，即数据仓库类型的数据模型或数据挖掘特需的数据模型，这种数据模型多表现为结构简单的、利于大量查询任务的多维数据模型，例如RLOAP或MLOAP。典型的数据关系结构如星形结构（Star Schema）或雪花形结构（Snow Flake Schema）。

我们设计的临床数据仓库（Data Warehouse）的数据模型就是一个雪花形结构模型，如图6-5所示。

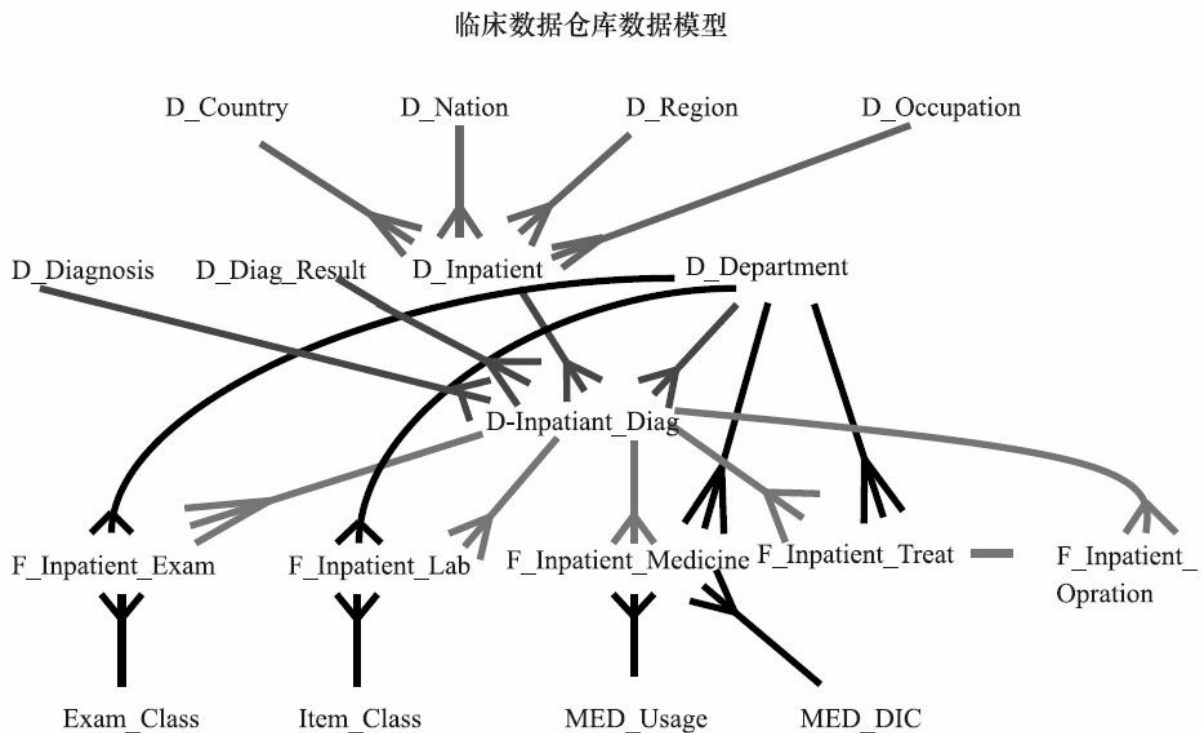


图 6-5

从图6-5中，我们可以看出，数据模型的最上层是一些在数据仓库里我们称作维度表（Dimensional Tables）的数据表，如D_Country（国家维度表）、D_Nation（民族维度表）、D_Region（地区维度表）、D_Occupation（职业维度表）等，用于描述D_Inpatient（病人信息）的各种属性，它们与D_Inpatient（病人维度表）之间的关系是一对多的父子表关系。

数据模型的第二层是另外一些维度表，如D_Inpatient（病人维度表）、D_Diagnosis（疾病分类维度表）、D_Diag_Result（诊断结果维度表）、D_Department（住院科室维度表），用于描述病案信息的各种属性，它们与D_Inpatient_Diag（病案维度表）之间的关系也是一对多的父子表关系。

接下来的一层中有一个表，就是D_Inpatient_Diag（病案维度表），它是描述病人某次住院治疗的结果信息，该表中的每条数据反映了一个病人的诊疗案例，是临床业务的基本单位。

在病案维度表下面的一层数据表中，包含了几个由基本单位病案信息所产生的各种流水性的数据，在数据仓库里我们称作Fact Tables（事实表），如F_Inpatient_Exam（物理检查事实表）、F_Inpatient_Lab（生化检查事实表）、F_Inpatient_Medicine（医嘱用药事实表）、F_Inpatient_Treat（医嘱治疗事实表）、F_Inpatient_Operation（医嘱手术事实表）等，这些表反映了一个住院病人在院期间的各种临床活动。病案维度表与这些事实表间的关系是一对多的父子表关系。

在数据模型的最下一层也是一些类似于维度表的属性字典表，如Exam_Class（检查大类表）、Item_Class（化验大类表）、MED_Usage（药品用法表）、MED_DIC（药品字典表）等，用于描述相关事实表的一些属性，它们与相关事实表的关系是一对多的父子表关系。从数据模型技术上，我们完全可以把它们视为维度表，但由于我们为突出病人—病案—临床活动这个主题线（临床主数据模型），不把它们划入临床主数据模型的维度表了。

6.2.2.4 数据挖掘的API工具选择

数据挖掘应用的开发方法有很多，可以用很多不同的软件工具来

实现。例如，可以用编程能力很强的C++语言，移植性很强的Java语言，等等。数据挖掘的算法可以完全由开发人员编写，也可以通过调用第三方提供的API，如IBM Data Modeler Server等。在IMRS的开发中，我们选择了采用基于Oracle数据库环境的数据挖掘API（Oracle Date Mining API）开发应用。

Oracle Data Mining API主要优点如下：

- 可直接处理存储在大型数据库中的、数据关系复杂的、超大规模的数据。
- 采用如PL/SQL的数据库编程语言可实现数据挖掘应用的可编程任务。
- 可充分利用集成于Oracle数据库中的数据挖掘算法API和统计学函数API。
- 通过数据库的互联网关，如开放数据库互联工具MS ODBC或Unix ODBC，第三方专用数据网关，如MS Transparent Gateway等，访问、挖掘更多的来自不同数据结构的数据。

Oracle Data Mining API主要是通过Oracle数据挖掘程序包DBMS_DATA_MINING Package的调用来提供数据挖掘的主要功能，如：

- 创建、删除、修改数据挖掘模型：CREATE_MODEL、

DROP_MODEL、RENAME_MODEL。

- 应用数据挖掘模型：APPLY。

- 评估数据挖掘模型应用结果：RANK_APPLY。

- 描述数据挖掘模型：GET_MODEL_DETAILS、
GET_MODEL_SETTINGS、GET_DEFAULT_SETTINGS、
GET_MODEL_SIGNATURE。

- 计算数据挖掘模型测试评估图表：
COMPUTE_CONFUSION_MATRIX、COMPUTE_LIFT、
COMPUTE_ROC。

- 输出和导入数据挖掘模型：EXPORT_MODEL、
IMPORT_MODEL。

数据挖掘模型的实施流程如下：

1) 创建数据挖掘模型。CREATE_MODEL过程用于创建一个数据挖掘模型。通过应用GET_MODEL_DETAILS函数，我们可以查看各种算法运算的模型结果，GET_MODEL_SIGNATURE和GET_MODEL_SETTINGS函数则可以得到数据挖掘模型的详细描述信息。

2) 应用数据挖掘模型。APPLY过程可以创建和操作一个预定义的数据表。这个数据表是基于一个特定的数据挖掘函数、算法、目标

属性类型（数值型或类别型）。RANK_APPLY过程对结果表的数据进行排序后，将前几名的结果输出到另外一个结果表中。分类模型可以基于成本对结果进行排名。结果表的字段结构根据特定的数据挖掘函数、算法、目标属性类型（数值型或类别型）的不同而不同。

3) 数据挖掘模型测试。分类模型的结果测试，由DBMS_DATA_MINING包提供的COMPUTE程序广泛应用于分类模型中评估量表METRICS的计算，它并不依从于特定的模型。只要输入的数据表的字段结构符合应用结果表和目标表的规范，它可以计算任何有意义的输入数据的评估量表。回归模型的结果测试，最常用的评估量表是平方根误差和平均绝对误差。

6.2.2.5 系统架构设计

IMRS的系统架构设计，采用数据层嵌入数据挖掘引擎的三层架构，如图6-6所示。

第一层为数据层（数据挖掘引擎），采用Oracle 10g/11g 64位企业级数据库系统，内嵌Oracle Data Mining（ODM）数据挖掘引擎，作为IMRS的数据中心，并提供一揽子数据处理和高级分析功能，包括：

- 数据准备（Data Preparation）：主要是一些存储在数据库中的存储过程，用于准备最终用户业务分析时需要的特定数据。

- 数据变换 (Data Transformation) : 也是一些存储在数据库中的存储过程, 用于为数据挖掘模型计算提供的数据库前期处理过程。

- ODM API (Oracle Data Mining Engine) : 是一些存储在Oracle数据库中的封装程序包, 由Oracle数据库提供的数据挖掘算法API。

- 定制化数据挖掘算法 (Customized Algorithm for Data Mining, CADM) : 是一些存储在Oracle数据库中的程序包, 是我们为IMRS定制开发的专用数据挖掘算法。

- Oracle统计函数 (Oracle Statistics Functions) : 是一些存储在Oracle数据库中的程序包和函数, 是由Oracle数据库提供的统计学程序和函数。

- 定制化统计函数 (Customized Statistics Functions) : 是一些存储在Oracle数据库中的程序包和函数, 是我们为IMRS定制开发的专用统计学程序和函数。

第二层为应用服务器层, 采用Oracle Application Express和带有Mod PL/SQL数据网关的Oracle HTTP Server, 作为IMRS的应用服务器和Web服务器, 为最终用户提供基于Web的应用服务, 包括:

- Oracle Application Express (APEX) : 是Oracle的一套基于Web架构的快速应用开发平台, 应用开发的语言为SQL和PL/SQL, 它可以集成于Oracle HTTP Server。

· Oracle HTTP Server：是Oracle的一套提供Web服务的中间件，与Oracle Application Express集成，为最终用户提供界面服务。

· Mod PL/SQL：是集成于Oracle HTTP Server的数据网关，它负责Oracle Application Express与Oracle数据库之间的数据通信。

第三层为最终用户层：是用户界面层，通过安装特定Flash的Internet浏览器，例如IE、Firefox等调用应用服务器层的应用，与数据挖掘引擎嵌入的数据层进行交互。

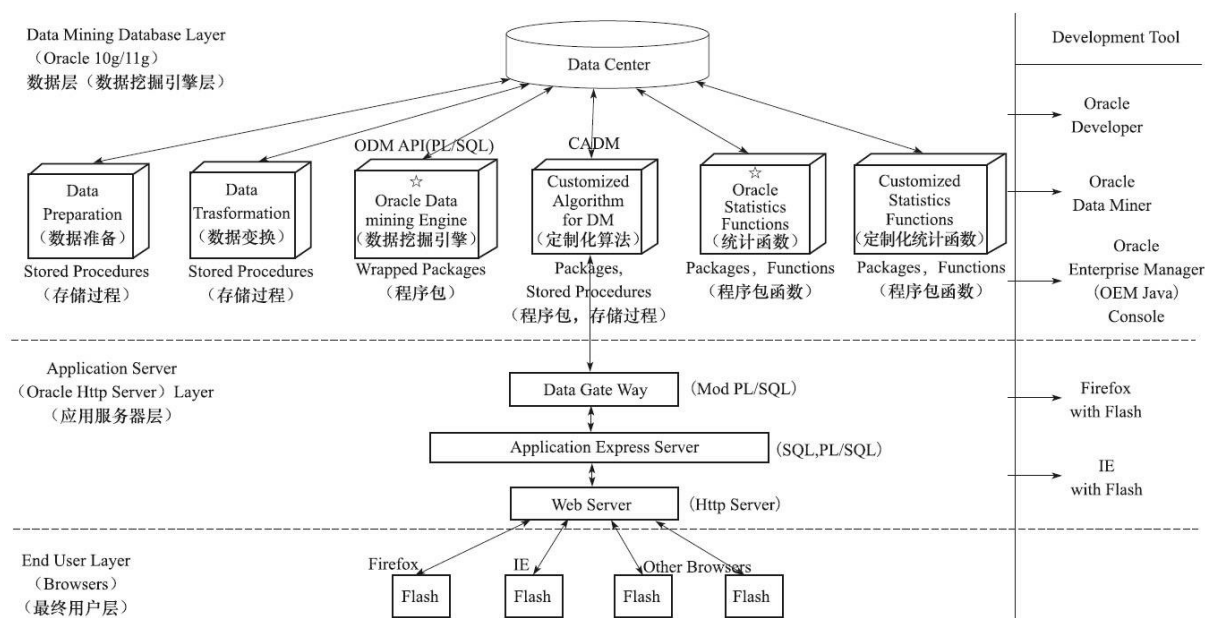


图 6-6

系统的开发工具也是采用Oracle提供的客户端开发软件，包括：

· Oracle Developer：是一个功能强大的SQL和PL/SQL客户端程序开发软件，提供所有的存储在数据库层的后台程序的开发和调试，也可用于存储在应用服务器APEX（最终代码也是存储在数据库中）的

前台程序的调试。

- Oracle Data Miner：是一个功能全面的数据挖掘应用客户端开发软件，提供数据处理、调用Oracle数据挖掘引擎Oracle Data Mining等功能，并可由向导式界面操作完成的数据挖掘功能生成PL/SQL代码。

- Oracle Enterprise Manager (OEM)：是Oracle数据库的管理工具，为数据挖掘应用开发项目提供数据层的管理、开发、安全和优化服务。

- Oracle Application Express：通过IE或Firefox等浏览器展现的APEX是基于Web的应用开发工具，这种快速开发工具为最终用户提供完善的应用展现和操作功能。

该架构设计方案具有如下特点：

- 数据库系统、数据挖掘引擎、应用服务器软件、Web服务器软件全部采用Oracle产品，系统整合性好、稳定、高效。

- 数据挖掘引擎和数理统计程序嵌入数据库，方便大规模、复杂数据的运算，并便于性能调优。

- 基于SQL和PL/SQL的应用服务器软件与数据库之间的交互更方便、更灵活，基于Web的数据挖掘应用界面更友好。

6.3 IMRS异常侦测模型的开发

如上节介绍的，IMRS的核心功能是数据挖掘子系统，它是一个用于临床医学科研发现新规律、新靶向的科研创新工具，其中包含很多不同的数据挖掘应用模块，如：疾病聚类分析、检验聚类分析、疾病关联分析、疾病特征抽取分析、疾病因子权重分析、检验因子权重分析、检验回归预测分析、检验异常侦测分析等。我们这里介绍其中的两个模块，首先介绍的是异常侦测分析模块的开发，主要讨论异常侦测模型的功能和相关数据挖掘技术开发要点。

6.3.1 异常侦测模型的功能展示

IMRS的异常侦测分析模块用于探索高维医学变量下的某生理指标的异常分布。虽然从形式上与聚类模型的分布分析类似，它也是展示数据的一种分布状态，一种特殊的分布状态，但异常侦测模型的核心是揭示数据中的异常，因此，我们把这一类模型和算法称为异常探索。下面，我们先展示一下异常侦测分析模块设计后的功能效果，然后再详细剖析相关的开发与实现。

6.3.1.1 数据准备

在本例中，研究人员选取了“体检人群”作为目标研究人群。主要筛选条件为做了“甲功五项”检验的男性人群和女性人群，如图6-7所示。

基础数据准备成功!

系统操作起始页

返回主页 下一步>

此模块为整个数据挖掘程序的一部分。用于疾病的聚类分析。你需要按照步骤操作此程序：“准备数据”->“清除数据”->“运行模型”->“取模型结果”->“查看结果”。如果开始数据分析，请点击下面的“开始分析”

体检人群-000.000

请选择疾病

（必选）

请选择化验项大类：甲功五项

（必选）

请选择性别：男/女

（可选）

入院日期：23-11-2000

（必选）

出院日期：30-11-2013

（必选）

图 6-7

经选择该模块的各下拉单式的参数选项后，我们点击“准备分析”按钮，该模块会自动从科研数据仓库中筛选满足以上要求的数据。系统显示本次筛选了1955条数据。

6.3.1.2 模型运行

数据准备完成后，我们需要适当的数据挖掘模型对该数据进行分析。我们点击该模块“运行模型”按钮，该模块会自动运行事先设计好

的数据挖掘引擎，运用适当的算法进行分析，如图6-8所示，此时系统界面自动蒙屏，将屏幕锁定防止用户再进行其他操作。



图 6-8

该模块在后台运行一段时间后，系统会显示“模型运行成功”提示，此时系统界面自动解屏，用户可以继续进行其他操作。我们点击“取出模型”按钮，稍后，该模块会将本次筛选的1955条数据的挖掘结果详细地显示出来，如图6-9所示。



图 6-9

6.3.1.3 模型分析

图6-10便是上节的运行结果部分数据，从中我们可以看出，“甲功五项”包括“血清游离T3”、“血清游离T4”、“血清总T3”、“血清总T4”和“TSH”（促甲状腺激素）。这五个指标都有相应的正常值范围，如果其中任一指标超出了正常值范围，我们都会认为该指标可能是异常的。即使同时有几个指标超出正常值范围，我们也是容易理解的，此时并不意味着几个指标“综合”异常，只要各指标之间遵从一定的相互影响关系，仍然属于单指标异常。以上出现的异常情况，我们

都称之为单变量异常。而医学的临床和科研中还会关心一个以上的指标（变量）的“综合”异常案例及它们的医学特征，这里的“综合”异常是指几个变量指标违反了它们之间的固有关系模型。例如本案例，研究者会关心包括性别和年龄在内的“甲功五项”综合异常的情况，这就需要我们应用异常侦测的算法来分析包括性别和年龄在内的“甲功五项”的共七个维度指标的异常数据。

经过模型的运算，我们把七个变量的异常数据标志为0，例如ID号为197457的女性体检人和ID号为192285的男性体检人。另外，算法给数据打出的正常和异常推断是建立在一定的概率基础上的，如ID号为197457的女性体检人有66.4%的概率是异常，表的第一行ID号为155669的女性体检人有68.3%的概率是正常的等。图6-11所示的条形图，给出了各个案例数据的正常和异常概率示意。

年龄	性别	序号	异常	置信度	血清游离T3	血清游离T4	血清总T4	血清总T3	TSH
28.45	女	155669	1	.68265759842223983	3.05	1.15	8.7	1.66	2.51
28.25	女	155683	1	.66429453785591164	2.84	1.2	8.6	1.37	1.22
46.59	女	155154	1	.53825009180789296	2.52	.91	10.3	2.02	12.6
26.45	女	155690	1	.59249253520423606	3.15	1.29	6.8	1.5	2.47
30.11	男	155128	1	.67952754784868119	3.19	1.21	8.8	1.66	3.82
24.99	男	155133	1	.64148868292621775	3.54	1.36	8.4	1.77	1.13
27.39	女	155147	1	.60911055470757847	2.81	1.3	6.8	1.31	2.27
56.74	女	162798	1	.61329206818376603	3.15	1.33	10.9	1.48	1.78
56.42	女	180297	1	.5780389459741947	3.01	1.67	10.2	1.18	1.61
31.03	女	188392	1	.61494288480643289	2.94	1.2	7.6	1.89	1.97
37.22	男	192285	0	.50298334226136587	2.56	1.09	8.7	1.84	4.66
30.33	男	204400	1	.6201927005915544	3.07	1.04	7.5	1.74	1.02
47.17	男	192534	1	.64375612357546219	3.01	1.3	9.1	2.04	1.41
27.12	男	190342	1	.61473385458016871	3.44	1.48	7	1.52	1.26
47.51	男	192318	1	.65006502426106028	2.92	1.31	8.6	1.89	1.68
39.32	男	198032	1	.63317029868003216	3.1	1.1	7.9	1.44	8.71
34.1	女	197457	0	.66371500212859935	3.81	.95	10.6	2.72	0
52.98	女	202884	1	.57119680581990862	2.82	1.39	11.5	2.1	1.33
53.42	女	199349	1	.63919833238574675	2.75	.9	7.8	1.17	1.26
48.78	男	198277	1	.65481454027023578	3.31	1.02	8.3	1.71	1.83
24.48	女	211807	1	.62543737167383939	3.17	1.13	11.3	1.81	5.27
33.96	女	211593	1	.60637694133543196	3.04	1.22	12	1.58	2.51
40.26	男	205016	1	.63128920547794021	3.58	1.4	8.4	2.08	1.28
46.4	男	216303	1	.50052499393518024	3.95	1.79	11.4	1.34	1.64

图 6-10

我们可以从图6-11中看出所有的数据被划分为两类，一类是“甲功五项”综合正常，一类是“甲功五项”综合异常。大多数数据被划归为正常（线条向右的点），小部分数据被划分为异常（线条向左的点）。我们看到正常点中，ID号为11454的点的正常概率为71.8%，说明该点的综合正常概率较高；ID号为105713的点的正常概率为50%，说明该点的综合正常概率较低，处于正常与异常的临界线处。而异常点中，ID号为94799的点的异常概率高达90%，说明该点的综合异常的概率非常高；ID号为92725的点的异常概率为57.6%，说明该点的综

合异常的概率较低，处于正常与异常的临界线处。

研究者对于三类数据点比较感兴趣，一类是正常概率比较高的数据，一般的正常概率达80%以上，属于比较典型的“甲功五项”综合正常案例；一类是异常概率比较高的数据，一般的异常概率达80%以上，属于比较典型的“甲功五项”综合异常案例；还有一类是正常概率或异常概率比较低的数据，一般的正常或异常概率小于60%，属于比较典型的“甲功五项”正常与异常的临界案例。在医学研究中，我们可能用“甲功五项”综合正常案例作为对照组，用“甲功五项”综合异常案例作为观察组，用“甲功五项”综合正常与异常的临界案例用作观察组或对照组，开展有关“甲功五项”综合特征的研究。

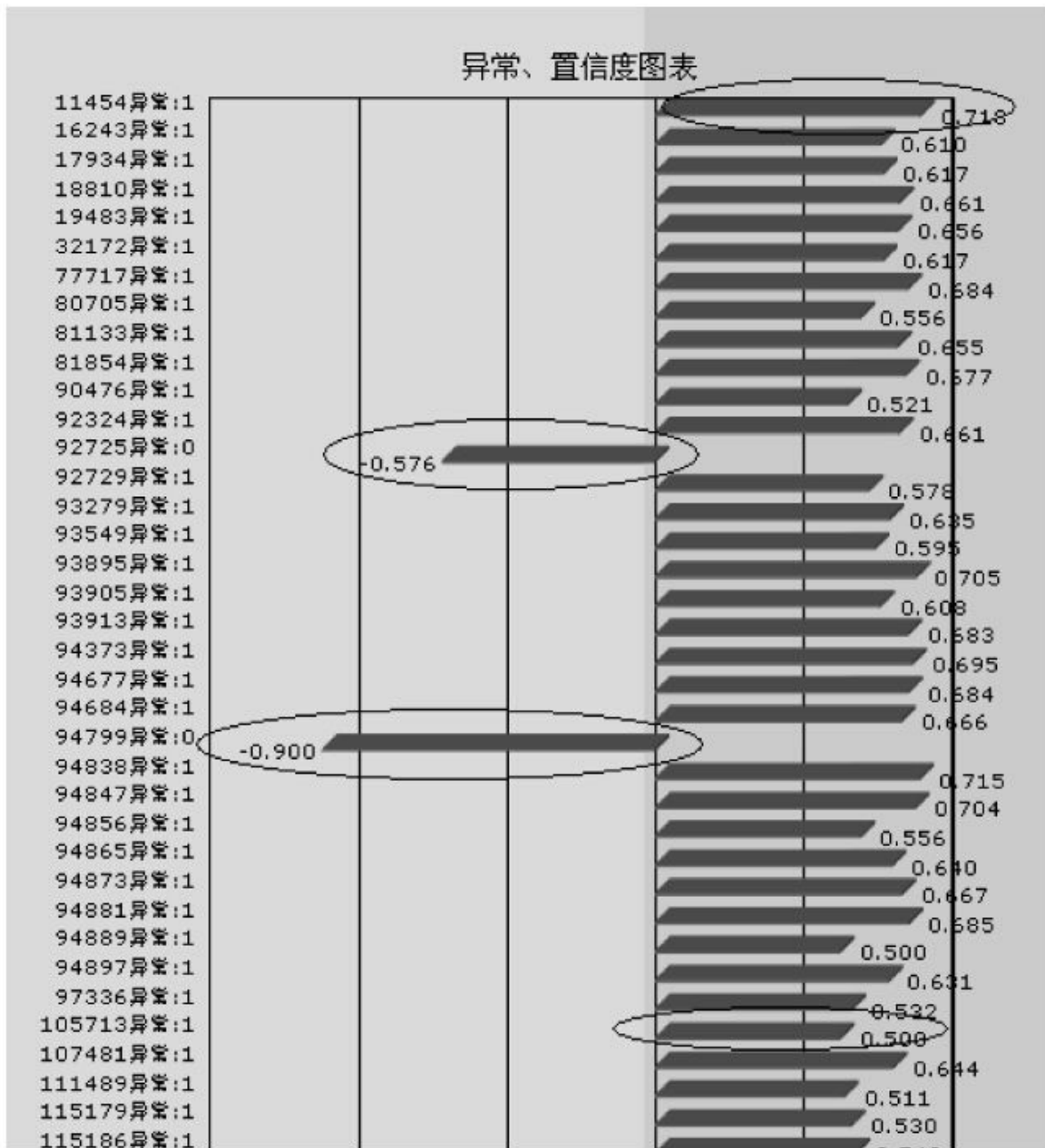


图 6-11

6.3.2 数据挖掘技术开发要点

我们总结了一些异常侦测模型应用开发的技术要点，主要涉及数据的变换与准备、异常侦测模型的建立和异常侦测模型的应用展现等方面。

6.3.2.1 数据变换与准备

例如选取“甲功五项”建立五维度异常侦测模型，数据准备的任务是自动从体检数据中选取做了“甲功五项”的体检人案例，并将“甲功五项”所包含的“血清游离T3”、“血清游离T4”、“血清总T3”、“血清总T4”和“TSH”五个指标项自动找出，列出体检人所做的五项化验结果。如表6-2所示。

表6-2 DM_INPATIENT_LAB_EXT6

TRANS_ID	PATIENT_AGE	PATIENT_SEX	OUT_DIAG...	COL400	COL397	COL399	COL401	COL398
32172	35.3	男	000.000	1.6	3.17	6.7	2.03	1.28
77717	37.7	女	000.000	1.35	3.1	7.9	1.16	1.26
81133	47.42	女	000.000	1.05	2.52	8.2	2.98	1.42
80705	52.11	男	000.000	1.31	3.07	10.9	1.25	1.68
81854	37.38	女	000.000	1.43	3.02	7.3	1.96	1.24
92729	28.81	男	000.000	1.54	4.13	8	5.68	1.31
92324	37.99	女	000.000	1.79	2.39	9.3	3.99	1.19
90476	48.02	男	000.000	0.95	2.82	7	1.86	1.29
93549	33.31	女	000.000	1.24	2.8	6.3	1.12	1.07
94677	37.43	男	000.000	1.86	3.3	10.3	1.91	1.6
94684	38.7	男	000.000	1.97	3.53	8.8	1.48	1.29
94799	34.06	男	000.000	2.02	3.27	8.9	2.67	8.9
93895	44.58	女	000.000	1.93	3.3	9.8	3.46	1.47
93905	34.87	男	000.000	1.99	3.81	11	2.5	1.69
94838	32.83	女	000.000	1.54	3.09	9.5	4.62	1.33
94847	48.92	女	000.000	1.48	2.91	9.4	0.99	1.38
94873	36.53	男	000.000	1.67	3.82	8.8	1.05	1.38
94881	37.94	男	000.000	1.66	3.01	9.4	1.66	1.19
94897	39.5	女	000.000	1.93	3.17	7.5	3.1	1.19
97336	38.04	男	000.000	1.09	2.92	6.2	0.93	1.22
105713	32.23	女	000.000	1.79	3.19	13.2	2.98	0.99

其中COL400、COL397、COL399、COL401、COL398五个字段分别代表“血清游离T3”、“血清游离T4”、“血清总T3”、“血清总T4”和“TSH”五个指标，OUT_DIAG为‘000.000’代表体检人群。由于该数据的提取技术与数据挖掘没有特异性关系，在此我们不做详细介绍。

计算目标表DM_INPATIENT_LAB_EXT6中的数据量，如果数据量低于数据挖掘引擎的最低要求，即如果总例数不大于100，系统将显示报错并提示用户重新选择数据。

部分PL/SQL代码如下：

```
DECLARE
    str varchar2(1000);
BEGIN
    str:='SELECT COUNT(*) FROM dm_inpatient_lab_ext6';
    EXECUTE IMMEDIATE str INTO :p30_datacount;
    IF :p30_datacount>100 THEN
        RETURN TRUE;
    ELSE
        RETURN FALSE;
    END IF;
END;
```

6.3.2.2 异常侦测模型建立

1.算法数据准备

计算COL400、COL397、COL399、COL401、COL398和AGE（年龄）的最大值、最小值、平均值、标准差，并生成一个数据分布视图‘DM_INPATIENT_AD_SV’。

PL/SQL代码如下：

```
CREATE OR REPLACE FORCE VIEW dm_inpatient_ad_sv AS
SELECT
AVG(col400) AVG_COL400,MAX(col400) MAX_COL400, MIN(col400)
MIN_COL400,STDDEV(col400) STD_COL400, AVG(col397) AVG_COL397,MAX(col397)
MAX_COL397, MIN(col397) MIN_COL397,STDDEV(col397) STD_COL397, AVG(col399)
AVG_COL399,
MAX(col399) MAX_COL399, MIN(col399) MIN_COL399,STDDEV(col399) STD_COL399,
AVG(col401)
AVG_COL401,MAX(col401) MAX_COL401, MIN(col401) MIN_COL401,STDDEV(col401)
STD_COL401,
AVG(col398) AVG_COL398,MAX(col398) MAX_COL398, MIN(col398)
MIN_COL398,STDDEV(col398)
STD_COL398, AVG(patient_age) AVG_AGE,MAX(patient_age) MAX_AGE,MIN(patient_age)
MIN_AGE ,STDDEV(patient_age) ,STD_AGE
FROM dm_inpatient_lab_ext6;
```

2.游标的定义

生化实验大类细项表为DM_LAB_ITEM_CLASS，它的部分数据如表6-3所示。

表6-3 DM_LAB_ITEM_CLASS

TEST_ITEM_NAME	REPORT_ITEM_NAME	LAB_COL	ENAME
生化六项	葡萄糖	COL302	GLU
生化二十项	丙氨酸氨基转移酶	COL301	ALT
白蛋白	雌二醇	COL496	E2
生化十项	天冬氨酸氨基转移酶	COL309	AST
生化全项	白蛋白	COL339	ALB
生化全项	肌酸激酶	COL348	CK
血常规	平均红细胞血红蛋白浓度	COL316	MCHC
血常规	嗜中性粒细胞百分数	COL328	NEUT%
生化全项	乳酸脱氢酶	COL350	LDH
肝功全项	白蛋白	COL339	ALB

该表包括生化实验大类TEST_ITEM_NAME、生化实验细项REPORT_ITEM_NAME、生化实验细项代码LAB_COL、生化实验细项英文简称ENAME的四个字段。（按照生化实验大类）取生化实验

细项代码的游标的定义如下：

```
DECLARE
TYPE T2TYPE IS REF CURSOR;
s_cur T2TYPE ;
CURSOR C2 IS SELECT lab_col FROM dm_lab_item_class;
s_rec C2%rowtype;
```

3.算法数据准备：缺失值的处理

对于数据中缺失值的处理，代码如下（由于代码中都有较详细的注释，故不再重复叙述）：

```
----- Start: Missing Values Transformation -----
--对于生化指标，使用其平均值来填补缺失值：
ls_append(v_tmp_lstmt, 'DECODE(''||s_rec.lab_col||'' , NULL,'||to_char(v_avg)||',
''||s_rec.lab_col||'' ) ''||s_rec.lab_col||'' ,');
--对于年龄指标，使用其平均值来填补缺失值。由于表中的性别字段没有缺失值，因此无需处理。
ls_append(v_tmp_lstmt,'DECODE("PATIENT_AGE",NULL,'||TO_CHAR(v_avg)||',
"PATIENT_AGE")"PATIENT_AGE","DMRMYMCASE_ID" FROM');
----- End: Missing Values Transformation -----
```

4.算法数据准备：数据规范化

对于数据中数值的规范化处理，代码如下：

```
----- Start: Normalize Transformation -----
--将生化指标规范到0和1之间的一个数
ls_append(v_tmp_lstmt, 'LEAST(1, GREATEST(0, (ROUND(('||s_rec.lab_col||'' -
' ||TO_CHAR(v_min)||') / DECODE(' ||TO_CHAR(v_max- v_min)||',0,0.01, '
||TO_CHAR(v_max- v_min)||'),15) * (1.0 - 0.0) + 0.0))''||s_rec.lab_col||''
,');
----- End: Normalize Transformation -----
```

5.模型参数设定

这里应用Oracle Data Mining提供的One-class SVM一次划分支持向

量机算法，算法的参数设置代码如下：

```
v_build_setting := ADD_TEMP_TABLE(v_tempTables,
create_new_temp_table_name('DM$T'));
EXECUTE IMMEDIATE 'CREATE TABLE ' || v_build_setting || ' (setting_name
VARCHAR2(30), setting_value VARCHAR2(128))';
EXECUTE IMMEDIATE 'INSERT INTO ' || v_build_setting || ' VALUES
('SVMS_ACTIVE_LEARNING', 'SVMS_AL_ENABLE')';--算法为主动式学习
EXECUTE IMMEDIATE 'INSERT INTO ' || v_build_setting || ' VALUES ('ALGO_NAME',
'ALGO_SUPPORT_VECTOR_MACHINES')'; --异常侦测算法为一次划分支持向量机:
EXECUTE IMMEDIATE 'INSERT INTO ' || v_build_setting || ' VALUES
('SVMS_OUTLIER_RATE', '0.1')';-- 算法的离群率为0.1
EXECUTE IMMEDIATE 'INSERT INTO ' || v_build_setting || ' VALUES
('SVMS_CONV_TOLERANCE', '0.0010')';-- 默认的最小收敛公差为0.01
```

6.建立模型，算法引擎API调用

应用ODM的One-class SVM算法创建一个异常侦测模型，代码如下：

```
-- BUILD MODEL
DBMS_DATA_MINING.CREATE_MODEL(
  model_name => model_name,
  mining_function => dbms_data_mining.classification,
  data_table_name => v_build_data,
  case_id_column_name => v_case_id,
  target_column_name => NULL,
  settings_table_name => v_build_setting);
-- End: BUILD MODEL
```

7.应用模型

异常侦测模型建立后，对同质的新数据应用这个模型，代码如下：

```
DECLARE
  tab varchar2(500);
  modell varchar2(300);
  result_tb varchar2(200);
  str varchar2(3000);
BEGIN
  DELETE from dm_ad_lab_venousbld_rst; --将模型结果表清空
  tab:='dm_inpatient_lab_ext6'; --选择需要异常侦测分析的数据表
```

```

modell:='dm_inpatient_ad_sv'; --给异常侦测模型一个特定名称
result_tb:='dm_ad_lab_venousbld_rst';--将清空的表作为模型结果表
"pke_ad_lab_venousbld_apl"."dm_inpatient_lasv"(tab, NULL,modell, NULL,result_tb,
FALSE, TRUE,:p30_itemclass);
--调用异常侦测模型应用程序包，对给定数据表（tab）的给定生化检查组合（:p30_itemclass，应用程序
变量传入的生化检查组合名，通常由多个生化指标构成），建立多维度的异常侦测挖掘模型（modell），运
行模型后将结果存储到结果表（result_tb）中。
End;

```

在这个异常侦测模型的数据挖掘过程里，对模型影响较大的主要参数是离群率SVMS_OUTLIER_RATE。我们比较一下设置不同离群率SVMS_OUTLIER_RATE得到的不同的结果，这里1表示预测为正常的点，0表示预测为异常的点。

1) 设置离群率为0.05，代码如下：

```

svms_outlier_rate = 0.05
SELECT prediction,MAX(probability), MIN(probability),AVG(probability),COUNT(*)
FROM dm_inpatient_lab783501790_a
--设置为svms_outlier_rate = 0.05的模型结果表
GROUP BY prediction;

```

得到的异常侦测模型结果，如表6-4所示。

表6-4 结果1

预测值	最大的预测概率	最小的预测概率	平均预测概率	记录数
1	.72925367021769583	.50000742452986424	0.61093	233
0	.80869050474153825	.50002741989533583	0.58394	15

2) 设置离群率为0.10，代码如下：

```

svms_outlier_rate = 0.10
SELECT prediction,MAX(probability), MIN(probability),AVG(probability),COUNT(*)
FROM dm_inpatient_lab187220622_a
--设置为svms_outlier_rate = 0.10的模型结果表
GROUP BY prediction;

```

得到的异常侦测模型结果，如表6-5所示。

表6-5 结果2

预测值	最大的预测概率	最小的预测概率	平均预测概率	记录数
1	.71774584904434746	.50010998624299896	0.60575	218
0	.90191470168067456	.50000174148176968	0.61748	30

3) 设置离群率为0.20, 代码如下:

```
svms_outlier_rate = 0.20
SELECT prediction,MAX(probability), MIN(probability),AVG(probability),COUNT(*)
FROM dm_inpatient_lab128303631_a
--设置为svms_outlier_rate = 0.20的模型结果表
GROUP BY prediction;
```

得到的异常侦测模型结果, 如表6-6所示。

表6-6 结果3

预测值	最大的预测概率	最小的预测概率	平均预测概率	记录数
1	.6707084952113409	.50003139552928322	0.57435	193
0	.95155302711156164	.5026131151659986	0.62679	55

从以上结果数据中我们可以看出, 随着离群率 SVMS_OUTLIER_RATE 的不断增加 0.05->0.10->0.20, 预测异常 0 的记录数、最大的预测概率、平均预测概率都在不断增加, 同时, 预测正常 1 的记录数、最大的预测概率、平均预测概率都在不断减小。说明随着离群率的不断设高, 支持向量机的分类曲面在不断地向正常点区域推进, 使得正常点区域在不断减小, 同时区域内点的可信度在不断降低; 异常点区域在不断增大, 同时区域内点的可信度在不断增强。

那么, 什么样的一个离群率的设置是适当的呢? 在商业项目中, 可以通过检验先验离群率来估计一个当前比较恰当的离群率。如果商

业业务中没有可以检验的先验离群率，我们可以用金标准进行抽样试验。例如：

- 1) 从大量数据中随机抽取一小部分数据。
- 2) 用不同的离群率的模型进行挖掘。
- 3) 然后将较小数量的预测结果进行金标准的验证，比较哪一个设定离群率比较接近实际的离群率。
- 4) 然后再随机抽取另外一小部分数据重复以上2)至3)的试验。
- 5) 综合评价多次对随机数据试验的结果，以确定一个当前比较适当的离群率。

6.3.2.3 异常侦测模型的展现

经过上面建立和应用异常侦测模型的一系列过程之后，我们可以将异常侦测模型应用的结果通过APEX界面展现出来，代码如下：

```
DECLARE
TYPE array_char IS table of dm_lab_item_class.test_item_name%TYPE;
array1 array_char;
i number;
con number;
str2 varchar2(2000);
str1 varchar2(3000);
str varchar2(1000);
BEGIN
SELECT
CASE
WHEN ename IS NULL THEN
lab_item
ELSE
```

```

        ename
    END bulk collect INTO array1 FROM dm_lab_item_detail
    WHERE lab_item IN (SELECT SUBSTR(report_item_name,1,14) AS item_name FROM
    dm_lab_item_class WHERE test_item_name=:p30_itemclass )
ORDER BY lab_col;
--将用户选定的生化实验组和名 (:p30_itemclass) 所包含的多个生化指标筛选出来，并存储到数组
(array1) 中
    FOR i IN 1..array1.COUNT LOOP
        str:=str||','||'COL'||i||' AS '||array1(i)||'';
    END LOOP;
    --数组 (array1) 中的生化指标名连接成一个用逗号分开的字串
    str1:=' CREATE OR REPLACE FORCE VIEW dm_ad_lab_vldrst
    AS
        SELECT "patient_age1" AS "年龄","patient_sex1" AS "性
别","trans_id" AS "序号","prediction" AS "异常标志","probability" AS "置信
度"||str||'
        FROM dm_ad_lab_venousbld_rst';
    --用连接由数组 (array1) 组成的字串，创建一个视图 (dm_ad_lab_vldrst)，从模型结果表
(dm_ad_lab_venousbld_rst) 中取出异常侦测结果
EXECUTE IMMEDIATE str1;
COMMIT;
END;

```

6.4 IMRS特征抽取模型的开发

上面我们介绍了异常侦测分析模块的开发。IMRS的数据挖掘子系统中另外一个非常重要的分析模块是特征抽取模块，下面我们将对特征抽取模块的设计和开发进行详细分解。与上一节一样，我们将主要讨论特征抽取模型的功能和相关数据挖掘技术开发要点。

6.4.1 特征抽取模型的功能展示

由于医学临床科研的需要，我们需要开发一个研究特定病人人群合并症特征的数据挖掘分析模块，以下简称该模块。首先，我们先看一下，业务需求希望得到什么样的功能和应用效果。在本例中，研究者选取了“不稳定性心绞痛”作为目标疾病人群，该人群患有多种不同的合并症，医学研究需要从大量的各种合并症案例中探索、挖掘出主要的合并模式，以及各个合并症分量的不同组合及其权重。详细请参见第7章的7.3节特征探索的应用之一。

6.4.2 数据挖掘技术开发要点

6.4.2.1 数据变换与准备

1.将单行的长字符串数据变换为交易型多记录数据结构。

有关“不稳定心绞痛”病人的合并症信息存储在视图‘V_DM_INPATIENT_DIAG_INST’中的“OUT_DIAGNOSIS_DESC”字段中，如图6-12所示。

PATIENT_INSTANCE_ID	OUT_DIAGNOSIS	OUT_DIAGNOSIS_NAME	OUT_DIAGNOSIS_DESC
100002518300016	I20.001	不稳定心绞痛	冠状动脉粥样硬化性心脏病 完全性右束枝传导阻滞（CRBBB） 心功能II级...
200002518300018	I20.001	不稳定心绞痛	冠状动脉粥样硬化性心脏病 完全性右束枝传导阻滞（CRBBB） 心功能II级...
30000342623002	I20.001	不稳定心绞痛	心房纤维性颤动（心房纤颤） 高血压III级，很高危
400002735780017	I20.001	不稳定心绞痛	冠状动脉粥样硬化性心脏病 心功能不全（I级） 高血压III
50000281750003	I20.001	不稳定心绞痛	陈旧性前壁心肌梗死 心功能III级 高血压III 钙化性主动脉瓣狭窄 左心扩大
60000317548003	I20.006	不稳定心绞痛，Braunwald III级	高血压III级，很高危
70000319317003	I20.001	不稳定心绞痛	陈旧性前间壁心肌梗死 急性左心衰竭 心功能IV级 急性支气管炎 高血压...
80000274780004	I20.001	不稳定心绞痛	冠状动脉粥样硬化性心脏病 心房扩大 心功能不全（I级） 高血压III
90000243885005	I20.001	不稳定心绞痛	陈旧性前间壁心肌梗死 心功能III级 胸腔积液 高血压III级，很高危 肺栓塞
100000253680004	I20.006	不稳定心绞痛，Braunwald III级	持续性房颤 心功能IV级 急性左心衰竭 高血压III级，很高危 高血压心脏...
110000247470007	I20.001	不稳定心绞痛	冠状动脉粥样硬化性心脏病 非胰岛素依赖型糖尿病（II型糖尿病） 继发...
1200002448570012	I20.006	不稳定心绞痛，Braunwald III级	高血压II级，很高危 阵发性心房纤颤（PAF） 下肢深静脉血栓形成
1300002448570013	I20.006	不稳定心绞痛，Braunwald III级	病态窦房结综合征 阵发性心房纤颤（PAF） 高血压II级，很高危
140000288158002	I20.006	不稳定心绞痛，Braunwald III级	陈旧性下壁后壁心肌梗死 心功能III级 一度房室传导阻滞 慢性阻塞性肺疾...
150000258538007	I20.001	不稳定心绞痛	心功能II级 骨质疏松
160000258538008	I20.001	不稳定心绞痛	心功能II级 骨质疏松
170000331549002	I20.006	不稳定心绞痛，Braunwald III级	高血压II级，很高危
180000247456009	I20.006	不稳定心绞痛，Braunwald III级	高血压II级，很高危 支气管扩张
190000243646002	I20.001	不稳定心绞痛	冠状动脉粥样硬化性心脏病 永久房颤 心功能不全（I级） 高血压III 非胰...
200000243646007	I20.001	不稳定心绞痛	心房纤维性颤动（心房纤颤） 高血压II 2型糖尿病 甲状腺机能减退
210000295612007	I20.001	不稳定心绞痛	阵发性心房纤颤（PAF） 急性左心衰竭 心功能IV级 局限性肺炎 2型糖尿...
220000248594007	I20.001	不稳定心绞痛	冠状动脉粥样硬化性心脏病 高血压II

图 6-12

我们的任务是，类似第4章中的数据变换，将表4-23的数据自动转换成表4-24的数据形式。数据转换前的形式如表6-7所示。

表6-7 转换前的表

PATIENT_INSTANCE_ID	OUT_DIAGNOSIS	OUT_DIAGNOSIS_NAME	OUT_DIAGNOSIS_DESC
0000243885005	I20. 001	不稳定性心绞痛	陈旧性前间壁心肌梗死心功能Ⅲ级胸腔积液高血压Ⅲ级，很高危肺栓塞

将以上数据变换为交易型的多记录数据结构，如表6-8所示。

表6-8 交易型多记录数据结构

Row_id	PATIENT_INSTANCE_ID	PATIENT_OUT_DIAGNOSIS	PATIENT_SYMPTOM
1	0000243885005	I20. 001	陈旧性前间壁心肌梗死
2	0000243885005	I20. 001	心功能Ⅲ级
3	0000243885005	I20. 001	胸腔积液
4	0000243885005	I20. 001	高血压Ⅲ级，很高危
5	0000243885005	I20. 001	肺栓塞

我们的方法是：

1) 将视图‘V_DM_INPATIENT_DIAG_INST’作为动态游标读入到服务器内存中。

2) 截断（TRUNCATE TABLE）一个表DM_INPATIENT_SYMPTOM，准备用于存储变换后的数据。

3) 使用两层循环将动态游标中的“OUT_DIAGNOSIS_DESC”字段中长字符串数据以空格作为分隔符，逐个将每个“词”（合并症）用INSTR()函数和SUBSTR()截取子串函数分别截取出来，并存储到表DM_INPATIENT_SYMPTOM中的PATIENT_SYMPTOM字段中，部分PL/SQL代码如下：

```

sql1 := 'SELECT patient_instance_id, out_diagnosis, out_diagnosis_desc FROM
v_dm_inpatient_diag_inst' ;--定义动态游标
EXECUTE IMMEDIATE 'TRUNCATE TABLE dm_inpatient_symptom';-- 截断目标表
OPEN lab_cur FOR sql1;-- 打开动态游标
<<bigloop>> --开始大循环

```

```

LOOP
  FETCH lab_cur INTO lab_rec;
  EXIT bigloop WHEN lab_cur%notfound;
  i:=1;
  k:=INSTR(lab_rec.out_diagnosis_desc,' '); --查找字段空格位置
  x:=LENGTH(lab_rec.out_diagnosis_desc); --计算字段长度
  <<smallloop>>--开始小循环
  LOOP
    y:= INSTR(lab_rec.out_diagnosis_desc,' ',1,i);
    IF i-1=0 THEN
      j:=0;
    ELSE
      j:= INSTR (lab_rec.out_diagnosis_desc,' ',1,i-1);
    END IF;--将截取后的字串插入到目标表中
    IF y=0 THEN
      EXECUTE IMMEDIATE 'INSERT INO dm_inpatient_symptom
        (patient_instance_id,patient_out_diagnosis,patient_symptom)
        VALUES('''||lab_rec.patient_instance_id||'', ''||
        lab_rec.out_diagnosis||'',SUBSTR('''||lab_rec.out_diagnosis_desc||'',
        '||TO_CHAR(j+1)||',''||TO_CHAR(x-j)||'))';
    END IF;
    EXIT smallloop WHEN y=0; --将截取后的字串插入到目标表中
    EXECUTE IMMEDIATE 'INSERT INTO dm_inpatient_symptom
      (patient_instance_id,patient_out_diagnosis,patient_symptom)
      VALUES('''||lab_rec. patient_instance_id ||'', ''||lab_rec.
out_diagnosis ||'',
      SUBSTR ('''||lab_rec. out_diagnosis_desc ||'', '||TO_CHAR(j+1)||',
      '||TO_CHAR(Y-1-j)||'))';
    i:=i+1;
  END LOOP; --结束小循环
END LOOP; --结束大循环
COMMIT;
CLOSE lab_cur; ;--关闭动态游标

```

2.数据规范化

字段值出现很多不规范的专业词汇，例如“高血压Ⅱ级”可能出现“高血压Ⅱ级，很高危”、“高血压（Ⅱ级）”、“高血压二级”等多种表达方式，我们会将它们映射成一个规范的表达“高血压Ⅱ级”。这样去伪存真，使得数据更能反映实际临床的有关信息，如表6-9所示。

表6-9 词汇

	不规范词汇	规范词汇
1	“心功能不全（Ⅰ级）”、“心功能不全Ⅰ级”	“心功能不全Ⅰ级”
2	“高血压Ⅱ级，很高危”、“高血压（Ⅱ级）”、“高血压二级”、“高血压Ⅱ级”	“高血压Ⅱ级”
...		
n	“心功能Ⅲ级，很高危”、“心功能（Ⅲ级）”、“心功能Ⅲ级”	“心功能Ⅲ级”

采用的相应的字符串处理技术进行数据清洗：

1) 去掉字段中的左右括号：

```
UPDATE dm_inpatient_symptom SET patient_out_diagnosis = REPLACE(REPLACE
(patient_out_diagnosis,'(',')',' ');
```

2) 截取字段中逗号前面的字符串：

```
UPDATE dm_inpatient_symptom SET patient_out_diagnosis =
SUBSTR(patient_out_diagnosis,1, INSTR(patient_out_diagnosis,' , ')-1);
```

当然，还有很多种不同的数据不规范的情况，我们都要采用相应的数据处理技术加以处理、规范，在此不再赘述。

3. 数据分析

举例，我们选取了某疾病的合并症数据，经以上处理后做如下分析：

1) 计算交易型数据表中的总记录数：

```
SELECT COUNT(*) FROM dm_inpatient_symptom;
```

得到结果13553例。

2) 计算交易型数据表中的不同的合并症数：

```
SELECT COUNT (DISTINCT patient_symptom) FROM dm_inpatient_symptom;
```

得到结果875种。

3) 计算交易型数据表中的病人个案数:

```
SELECT COUNT (DISTINCT patient_instance_id ) FROM dm_inpatient_symptom;
```

得到结果2617例。

综合以上分析结果，我们得知，本次数据中有2617例病人总共患有875种不同的合并症，平均每人患有 $13553/2617=5$ 种不同的合并症。

4.计算目标表DM_INPATIENT_SYMPTOM中的数据量

如果数据量低于数据挖掘引擎的最低要求，即如果病人总例数小于500或者平均每个病人患有两种以下的合并症，系统将显示报错，并提示用户重新选择数据。部分PL/SQL代码如下：

```
DECLARE
    str varchar2(1000);
    str1 varchar2(1000);
BEGIN
    str:='SELECT COUNT(*) FROM dm_inpatient_symptom';
    EXECUTE IMMEDIATE str INTO :p40_datacount;
    str1:='SELECT COUNT (DISTINCT patient_instance_id) FROM dm_inpatient_symptom';
END;
EXECUTE IMMEDIATE str1 INTO :p41_datacount;
IF :p41_datacount>500 AND (:p40_datacount/ :p41_datacount) > 2 THEN
    RETURN TRUE;
ELSE
    RETURN FALSE (数据量太小，不能进行数据挖掘，请重新选择数据! );
END IF;
```

5.将交易型的数据结构转换成正常的二维数据表格式

应用数据库交集合并处理技术将上表交易型的数据结构转换成正常的二维数据表格式，生成一个超高维度矩阵表，如表6-10所示。

表6-10 超高维度矩阵表

PATIENT_INSTA NCE_ID	陈旧性 前间壁 心肌梗死	心功能Ⅲ级	胸腔积液	高血压 Ⅲ级， 很高危	肺栓塞	...	冠状动脉 粥样硬化性 心脏病	...
0000243885005	1	1	1	1	1	0	0	0
00002518300016	0	0	0	0	0	1	1	1
...								
0000281750003	1	1	0	1	0	0	0	1
...								

6.4.2.2 特征抽取模型建立

由D.D.Lee和H.S.Seung率先提出的非负矩阵因子分解方法，为我们提供了一种大规模数据降维技术。该方法中非负矩阵分解的乘性迭代公式有效地深化了数据的非负性。算法详见第3章。

非负矩阵分解算法主要用于大规模非稀疏数据中图像的模式识别，和基于较稀疏型的文本数据中的文本挖掘。由于在病人的合并症数据中，病人个体患有的合并症是很有限的几种，但对于全体病人来说（总体）合并症的种类是非常多的，常见的就有成千上万种。

因此由病人合并症数据组成的矩阵是一个典型的稀疏矩阵。在这个矩阵中，行代表症状，列代表病人个案。在非负矩阵因子分解算法中： $V=W \times H$ ，将大矩阵（高维）分解为两个小矩阵（低维），W为基矩阵，H为系数矩阵。采用约束非负矩阵的稀疏度，可以约束低维

矩阵 W ，也可以约束低维矩阵 H 。在本案例中，我们应该约束 H 系数矩阵，使得较少的合并症组合更能清晰地代表某类疾病的主题特征。适当的稀疏性可以保留数据的主要特征，同时还可减少存储空间，提高运算效率。

非负矩阵因子分解是一种低秩逼近算法。由于矩阵分解前后仅仅包含非负元素，原数据矩阵中的列向量可以表达为对左矩阵（也就是基矩阵）中所有列向量的加权和，而权重系数形成右矩阵（也就是系数矩阵）对应列中的元素。这里， V 的每一列为观测数据（病人个案），每一行为合并症特征向量元素， W 的列远远小于 V 的列数，也就是从 V 中抽取的合并症特征。分解后的 W 的每一列元素之和为1，分解后的 H 的每一行元素之和也为1。

1. 算法数据准备

第一步，离群点数据处理

超高维度矩阵表中，去掉单维度（变量）合并症数据，也就是只有一个合并症的数据。例如，0000250811004只有“陈旧性前间壁心肌梗死”合并症，0000310432003只有“冠状动脉粥样硬化性心脏病”合并症，如表6-11所示。

表6-11 超高维度矩阵表

PATIENT_INSTA NCE_ID	陈旧性 前间壁 心肌梗死	心功能Ⅲ级	胸腔积液	高血压 Ⅲ级， 很高危	肺栓塞	...	冠状动脉 粥样硬化性 心脏病	...
0000250811004	1	0	0	0	0	0	0	0
0000257457002	0	0	0	0	0	0	0	0
...								
0000310432003	0	0	0	0	0	0	1	0
...								

超高维度矩阵表中，去掉超高维度（合并症变量）数据，也就是有超过二十个合并症（由医学专业来界定）的数据。例如，0000283287004有“陈旧性前间壁心肌梗死”、“心功能Ⅲ级”、“胸腔积液”、“高血压Ⅲ级”、“肺栓塞”、“冠状动脉粥样硬化性心脏病”等超过20个的合并症，这种情况一种可能是极罕见的病例，另一种可能是病案录入员的误操作，如表6-12所示。

表6-12 去掉超高维度的矩阵表

PATIENT_INSTA NCE_ID	陈旧性 前间壁 心肌梗死	心功能Ⅲ级	胸腔积液	高血压 Ⅲ级， 很高危	肺栓塞	...	冠状动脉 粥样硬化性 心脏病	...
0000283287004	1	1	1	1	1	1	1	1
...								
0000253332006	1	1	0	1	1	1	1	1
...								

第二步，准备挖掘矩阵

应用数据库的行列转换技术，将数据准备成非负矩阵因子分解算法所需矩阵V的形式，得到超高维度矩阵转置表，如表6-13所示。

表6-13 超高维度矩阵转置表

PATIENT_INSTANCE_ID	0000243885005	00002518300016	...	0000281750003	...
陈旧性前间壁心肌梗死	1	0		1	
心功能Ⅲ级	1	0		1	
胸腔积液	1	0		0	
高血压Ⅲ级，很高危	1	0		1	
肺栓塞	1	0		0	
...	0	1		0	
冠状动脉粥样硬化性心脏病	0	1		0	
...	0	1		1	

本例的数据可生成875×2617的矩阵。

2.算法参数定义

部分PL/SQL代码如下：

```
v_build_setting := ADD_TEMP_TABLE(v_tempTables, create_new_temp_table_name('DM
$T'));
EXECUTE IMMEDIATE 'CREATE TABLE ' || v_build_setting || ' (setting_name
VARCHAR2(30),
setting_value VARCHAR2(128))'; --设置模型名字串长度为128个字符内
EXECUTE IMMEDIATE 'INSERT INTO ' || v_build_setting || ' VALUES
(''NMFS_NUM_ITERATIONS'', ''50'')'; --设置模型的迭代次数为50
EXECUTE IMMEDIATE 'INSERT INTO ' || v_build_setting || ' VALUES (''ALGO_NAME'',
''ALGO_NONNEGATIVE_MATRIX_FACTOR'')'; --设置模型的算法为非负矩阵因子分解
EXECUTE IMMEDIATE 'INSERT INTO ' || v_build_setting || ' VALUES
(''NMFS_CONV_TOLERANCE'', ''0.05'')'; --设置模型的收敛公差为0.05
COMMIT;
```

如上代码所示，特征抽取算法为非负矩阵因子分解，参数设置为：

```
ALGO_NAME=ALGO_NONNEGATIVE_MATRIX_FACTOR
```

默认的迭代次数为50，参数设置为：

```
NMFS_NUM_ITERATIONS=50
```

默认的最小收敛公差为0.05，参数设置为：

```
NMFS_CONV_TOLERANCE=0.05
```

3.算法引擎API的调用

部分PL/SQL代码如下：

```
-- BUILD MODEL
DBMS_DATA_MINING.CREATE_MODEL(
  model_name          => model_name,
  mining_function     => dbms_data_mining.feature_extraction,
  data_table_name     => v_build_data,
  case_id_column_name => v_case_id,
  settings_table_name => v_build_setting);
```

4.模型运行结果

按照前面的步骤准备数据、设置算法参数后，可以运行模型。最终我们得到了一个特征抽取列表，如表6-14所示。

表6-14 特征抽取列表

合并症	特征 1	特征 2	特征 3	...	特征 67
高血压Ⅲ	0.000292617890	0.000000000000	0.000000000000		0.000388375565
冠状动脉粥样硬化性心脏病	0.000063371881	0.0261112660712	0.021393524778		0.005708851551
脑梗塞后遗症	0.000000000000	0.001346647787	0.000000000000		0.000000000000
高血压Ⅲ级，很高危	0.000000000000	0.000000000000	0.000000000000		0.007023409411
...					
2 型糖尿病	0.000000000000	0.000000000000	0.000005759340		0.000000000000
...					

表6-14中，我们一共得到了67个合并症新特征。这些合并症新特征是由多个不同的合并症组合而成的，它们概括了全部数据的主要特

征，可以解释为主要特征反映在67个方面。从非负矩阵因子分解的技术上看，这个特征矩阵是从875×2617的高维矩阵中分解而来的，但从医学专业角度上看，这个特征矩阵是从875×2617的高维矩阵中抽取（总结）出来的。

在非负矩阵因子模型的参数设置里，对模型影响较大的主要是最小收敛公差MIN_CONV_TOLERANCE。下面我们比较、分析一下设置不同最小收敛公差MIN_CONV_TOLERANCE得到的不同的结果。

1) 设置NMFS_CONV_TOLERANCE=0.05（默认设置）

从模型运行的结果表中选取数据：

```
SELECT *  
FROM dm_fe_symptom_rslt;
```

列表如图6-13所示。

	FEATURE_ID	ATTRIBUTE_VALUE	COEFFICIENT	USERNAME
1	1	陈旧性前壁心肌梗死	0.0000235329316188246	mngl
2	1	陈旧性下壁心肌梗死	0.0352070350748534	mngl
3	1	高血压 I 级，很高危	0.00000000024875858528263	mngl
4	1	高血压 II 级，很高危	0.0000500121872365273	mngl
5	1	高血压 III	0.000292617890372326	mngl
6	1	冠状动脉粥样硬化性心脏病	0.000063371880766761	mngl
7	1	心功能 IV 级	0.000352355164299055	mngl
8	2	胆囊结石	0.00000484780179847179	mngl
9	2	冠状动脉粥样硬化性心脏病	0.026111266071243	mngl
10	2	类风湿性关节炎	0.00000000010040786356215	mngl
11	2	慢性浅表性胃炎	0.00000044052893277118	mngl
12	2	脑梗塞后遗症	0.00134664778704138	mngl
13	2	完全性右束支传导阻滞（CRBBB）	0.0000000624343155757379	mngl
14	2	胃溃疡	0.00000001816059958829	mngl
15	2	阵发性心房纤颤（PAF）	0.000229040888607468	mngl
16	2	脂肪肝	0.00000113552424642434	mngl
17	3	2 型糖尿病	0.0000057593408353523	mngl
18	3	陈旧性肺结核	0.000000000939954367633192	mngl
19	3	二型糖尿病性肾病	0.0000000018907348248517	mngl
20	3	冠状动脉粥样硬化性心脏病	0.0213935247787791	mngl
21	3	2 型糖尿病	0.0000000640360333575308	mngl

图 6-13

计算模型运行的结果表的记录总数：

```
SELECT COUNT(*)
FROM dm_fe_symptom_rslt;
```

得到5881条数据。

计算模型运行的结果表的特征总数：

```
SELECT COUNT(DISTINCT feature_id)
FROM dm_fe_symptom_rslt;
```

得到67个特征，由5881个特征分量组成。

计算模型运行的结果表的特征分量元素种类：

```
SELECT COUNT(DISTINCT attribute_value)
FROM dm_fe_symptom_rslt;
```

得到670个特征分量元素。

计算模型运行的结果表的特征系数总和以及特征分量元素数，并按照特征系数总和的大小降序排列：

```
SELECT feature_id, SUM(coefficient), COUNT(*)
FROM dm_fe_symptom_rslt
GROUP BY feature_id
ORDER BY SUM(coefficient) DESC;
```

例如特征24的特征系数总和为0.046，特征分量元素数为222个，如图6-14所示。

	FEATURE_ID	SUM(COEFFICIENT)	COUNT(*)
1	24	0.046614609769046833486809	222
2	57	0.041778657121877364936648	224
3	37	0.039458175737412039782053	188
4	61	0.037724040424351510669837	173
5	33	0.036322897556709665607641	183
6	9	0.035378983790307758589521	199
7	53	0.034272035786128468829016	171
8	42	0.034205761885205408443576	160
9	25	0.028905416368398304072555	165
10	62	0.026078020976362408865126	108
11	4	0.022727222247207248072525	162
12	49	0.020886170811601266298338	92
13	20	0.019767411088335765541779	146
14	1	0.019000948736273751723141	176
15	46	0.018929518912139172060439	85
16	2	0.018779984605280549666949	61
17	10	0.018586034630515701453747	69
18	29	0.01737536642023197629482	81
19	13	0.017320673600607406454081	81

图 6-14

对于所有的特征分量元素，计算每个特征分量元素的系数总和以及该特征分量元素存在于特征的个数，并按照特征系数总和的大小降序排列：

```
SELECT attribute_value, SUM(coefficient), COUNT(*)
FROM dm_fe_symptom_rslt
GROUP BY attribute_value
ORDER BY SUM(coefficient) DESC;
```

例如特征分量元素“高脂血症”的特征系数总和为0.32，包含该特

征分量元素的特征有64个，如图6-15所示。

2) 设置NMFS_CONV_TOLERANCE=0.10（默认设置的2倍）

同理，我们应用上面的方法步骤得到相应的模型结果。

3) 设置NMFS_CONV_TOLERANCE=0.025（默认设置的1/2）

同理，我们应用上面的方法步骤得到相应的模型结果。

	R2	ATTRIBUTE_VALUE	R2	SUM (COEFFICIENT)	R2	COUNT (*)
1		高脂血症		0.320187395351199333567707		64
2		高血压III		0.241705367820809856001006		60
3		2型糖尿病 (非肥胖) (肥胖)		0.12830930240697457990779		59
4		高血压 II		0.042397627988615044320836		47
5		脑血管病后遗症		0.038423862764010765216071		54
6		心功能III级		0.038204033191164834235973		51
7		颈动脉动脉硬化		0.032277253162191627868302		45
8		肢体动脉硬化		0.03045624250337833287847		43
9		心功能 II 级		0.027669185764582925135148		48
10		慢性支气管炎		0.024461807598845552431797		48
11		高血压 I		0.024267234869622017716195		45
12		颈椎病		0.02293520509079530623072		29
13		心房颤动 (心房纤颤)		0.019929255993348849699818		39
14		心律失常		0.019482296733710432730338		48
15		前列腺增生		0.018719416934269616275762		49
16		动脉硬化		0.015865346000031488870961		43
17		泌尿道感染		0.01411074359231268753516		43
18		慢性胃炎		0.005501499572585545053754		38
19		心功能I级		0.00457391889630362024312		43
20		颈椎关节强硬		0.004276029653372481854135		37
21		脂肪肝		0.003773086926603101538137		32
22		腔隙性脑梗死		0.002917143931858267733275		40
23		慢性肾功能不全		0.002113559733560122272169		39
24		甲状腺功能减退		0.001800037604588590455385		37

图 6-15

然后，我们将以上三种设置不同最小收敛公差得到的模型结果加以汇总、计算，得到以下结果，如表6-15所示。

表6-15 汇总结果表

特征抽取结果表	NMFS_CONV_ TOLERANCE =0. 100	NMFS_CONV_ TOLERANCE =0. 050	NMFS_CONV_ TOLERANCE =0. 025
记录总数 COUNT(*)	5881	533	377
特征总数 COUNT(DISTINCT feature_id)	67	67	67
特征分量元素种类 COUNT(DISTINCT attribute_value)	670	216	158
特征系数总和 SUM(coefficient)	1. 098	1. 358	1. 378
平均特征分量元素数/每特征 COUNT(*) / COUNT (DISTINCT feature_id)	87. 8	8	5. 6

(续)

特征抽取结果表	NMFS_CONV_ TOLERANCE =0. 100	NMFS_CONV_ TOLERANCE =0. 050	NMFS_CONV_ TOLERANCE =0. 025
平均特征系数总和/每特征 SUM(coefficient) / COUNT (DISTINCT feature_id)	0. 016	0. 020	0. 020
平均特征数/每特征分量元素 *	8. 8	2. 5	2. 4
平均特征系数总和/每特征分量元素 SUM(coefficient) / COUNT (DISTINCT attribute_value)	0. 0016	0. 0063	0. 0087

```
* SELECT SUM(cnt) / COUNT(DISTINCT attribute_value)
FROM (
SELECT attribute_value,SUM(coefficient), COUNT(*) cnt FROM dm_fe_symptom_rslt
GROUP BY attribute_value);
```

比较表6-15的结果，我们可以看到随着最小收敛公差 NMFS_CONV_TOLERANCE 的值越小 0.100->0.050->0.025，也就是特征矩阵分解得越收敛，特征抽取结果表的记录（特征分量元素）总

数、特征分量元素种类总体减少；相应地，每特征平均特征分量元素数、每特征分量元素平均特征数也总体减少，其中从0.100到0.050段降幅很大，从0.050到0.025段降幅较小。另一方面，随着最小收敛公差NMFS_CONV_TOLERANCE的值越小0.100->0.050->0.025，特征系数总和、每特征平均特征系数总和、每特征分量元素平均特征系数总和总体增长，其中从0.100到0.050段增幅很大，从0.050到0.025段增幅较小。

从以上的数据中我们可以看出，随着特征矩阵分解得越收敛，模型表现为：

- 原数据的主题特征被抽取得越概括、越抽象。
- 特征分量元素种类越集中在较少的范围里。
- 特征系数总和也越大，总的主题特征越突出。
- 每特征平均特征分量元素数在减少，每特征被抽取得越概括、越抽象。
- 每特征平均特征系数总和在增大，每特征的主题特征越突出。
- 每特征分量元素平均出现的特征数在减少，每特征分量元素表达的特征范围越集中。
- 每特征分量元素平均特征系数总和在增大，每特征分量元素表达的特征权重越高。

不难想象，随着特征矩阵分解得越发散，模型表现应与上面的方向相反。那么，在实际应用中，我们应该怎样把握这个“收敛度”呢？我们说，不同的商业项目的需求是不一样的，这主要根据业务的要求来决定。如果我们的模型提取的特征比业务需求更概括、更抽象，则在应用中就会漏掉了很多有用的特征和信息；同理，如果我们的模型提取的特征比业务需求更具体、更细化，使得太多的信息混杂在特征里，会使我们在应用中很难抓住事物的主要特征和关键信息。因此，特征抽取的“收敛度”要较好地适应商业项目的实际需求。

6.4.2.3 特征抽取模型的展现

1.从Oracle Data Mining API中取模型结果

```
DECLARE
    model varchar2(200);
BEGIN
    model:='DM_INPATIENT_FE_NMF';----赋值特征抽取模型名
    DELETE dm_fe_symptom_rslt;----将特征抽取模型结果表清空
    EXECUTE IMMEDIATE 'INSERT INTO dm_fe_symptom_rslt SELECT t.feature_id,
    s.attribute_value, s.coefficient FROM
    table(dbms_data_mining.get_model_details_nmf(''||model||'')) t,
    table (t.attribute_set) s
    WHERE s.coefficient > 0.0000000001';-
    --将名为'DM_INPATIENT_FE_NMF'的特征抽取模型的结果取出，存到特征抽取模型
    结果表'DM_FE_SYMPTOM_RSLT'中
    COMMIT;
END;
```

2.应用交互式报表（Interactive Report）展示模型结果

```
SELECT feature_id, sum_coeff,rownum
FROM
(SELECT feature_id,SUM(coefficient) sum_coeff FROM dm_fe_symptom_rslt
GROUP BY feature_id
ORDER BY SUM(coefficient) DESC );
--按照特征ID: feature_id 进行分组，累加特征系数总和: SUM(coefficient) sum
```

3.应用统计图表（Chart Report）展示模型结果

```
DECLARE
    sum1 number;
    str varchar2(3000);
    str1 varchar2(1000);
BEGIN
SELECT SUM(coefficient) INTO sum1 FROM "dm_fe_symptom_rslt"
WHERE feature_id=p40_feature;--计算选定特征ID的特征系数总和
    str:='select null link, attribute_value label, coefficient value1
FROM (
SELECT attribute_value,coefficient/'||sum1||' coefficient
FROM "dm_fe_symptom_rslt"
WHERE feature_id='''||:p40_feature||''')';
--计算选定特征ID的每个特征分量在该特征中的相对权重
RETURN str;
END;
```

由于本书之重点在于讨论数据挖掘的思想和技术要点，因此对于应用PL/SQL语言解决一些软件功能不做过于详细的讲解。

6.5 IMRS智能统计模型的开发

我们知道，数据挖掘与统计学是一个交集的关系。数据挖掘的很多方法和技术都含有统计学的思想和方法，因此说二者是密不可分的。不但从技术说，两者是紧密相关的，而且在行业应用中也是这样。例如，在医学科研方面，我们一方面需要应用数据挖掘技术来探索新规律、新靶向，另一方面我们也需要应用统计学方法来验证我们的规律和模型。因此，在IMRS设计中，就将医学统计方法以流程控制的方式软件化，我们称为智能统计分析系统，使每个医学专业人士不再将数学和统计学视为科研的瓶颈。IMRS系统包含了大多数最常用的医学定量分析和定性分析，统计工具涵盖了大量的医学统计方法，满足了科研者常用的科研课题的统计学设计及实验的需求。智能式统计流程工具可以直接分析数据筛选工具得到的数据，无需向其他统计软件那样将数据反复整理、导出导入，因此，统计实验分析阶段的工作的效率得到了大大的提高。智能统计分析系统主要有如下几个特点：

- 为医学专门定制，具有严谨的数学适用性。
- 统计分析以医学科研方案为单位，专业方便。
- 统计学实验与数据集成，可随时、大量、反复分析。
- 验证数据挖掘的探索，提高科研的成功率。

智能统计分析系统涵盖的功能很多，我们这里以医学领域里的回归预测模型和线性相关模型应用在数据库环境下的开发实现为例，做一简要介绍。

6.5.1 回归模型的开发实现

回归模型可分为线性回归模型、非线性回归模型，还可分为一元回归模型和多元回归模型。限于篇幅，这里仅介绍一元线性和非线性回归模型的功能实现，主要内容将涉及回归模型的开发实现、计算线性回归方程、计算线性回归模型的五个统计图、统计检验、计算置信带可信区间、回归预测，以及非线性回归模型的建立。

在IMRS系统中，选择“回归分析”模块。首先进入简单线性回归界面，从事先建好的科研数据仓库中选择数据。这里选择临床科研课题的研究对象，本例中选择“恶性肿瘤术后化疗”的病人组为科研课题病例组。然后依次选择病例组数据的其他筛选条件，如入组性别、医学观察指标（如“红细胞压积”、“血红蛋白”等）、入院年月等，如图6-16所示。

病例组数据的筛选条件选好后，点击“下一步”按钮，系统将自动生成所需的数据，如图6-17所示。

选择数据-开始页

返回主页

下一步

在两个变量的资料中，如果一个变量Y随另一个变量X呈现出线性变化规律，则变量Y与X间便构成一种线性依存关系，而这种线性依存关系的有效统计学方法就是线性回归。

请选择疾病:

恶性肿瘤术后化疗

(必选)

* 观察条件:

全部数据分析

(必选)

* 请选择性别:

男女

(可选)

* 观察变量X:

★红细胞压积

(必选)

* 观察变量Y:

★血红蛋白

(必选)

* 入院日期:

(可选)

* 出院日期:

(可选)

图 6-16

显示列表

检验项X	检验项X值	检验项Y	检验项Y值
★红细胞压积	.335000000000000002	★血红蛋白	106
★红细胞压积	.364999999999999999	★血红蛋白	110
★红细胞压积	.302999999999999999	★血红蛋白	88
★红细胞压积	.406999999999999997	★血红蛋白	122
★红细胞压积	.339000000000000002	★血红蛋白	102
★红细胞压积	.317	★血红蛋白	111

图 6-17

这里，我们将“红细胞压积”作为回归模型的自变量X，“血红蛋白”作为回归模型的自变量Y，试图建立一个回归模型，并去验证它，如果回归模型成立，我们会应用这个回归模型作为由“红细胞压积”预测“血红蛋白”的一个预测模型。以下是这个预测模型创建、验证和应用的简要技术实现。

6.5.1.1 计算线性回归方程

线性回归方程的计算结果如下：

线性回归方程： $Y=17.255+282.981X$

$SS_{\text{总}}=SS_{\text{回}}+SS_{\text{剩}}$

$SS_{\text{剩}}=343868.46-267461.34=76407.12$

线性回归方程的计算代码如下：

```
SELECT ROUND( REGR_INTERCEPT( tab1.tr1 (检验项X) , tab2.tr1 (检验项Y) ), 3) ,  
        ROUND( REGR_SLOPE( tab1.tr1, tab2.tr1 ), 3) INTO aa, bb  
FROM 表1;  
--求解一元线性回归方程的系数。
```

代码中用到了Oracle计算线性回归方程的聚合函数，
REGR_INTERCEPT()返回Y的截距，REGR_SLOPE()返回Y的斜率。

6.5.1.2 计算线性回归模型的五个统计图

第一个是线性回归模型的散点图，如图6-18所示。

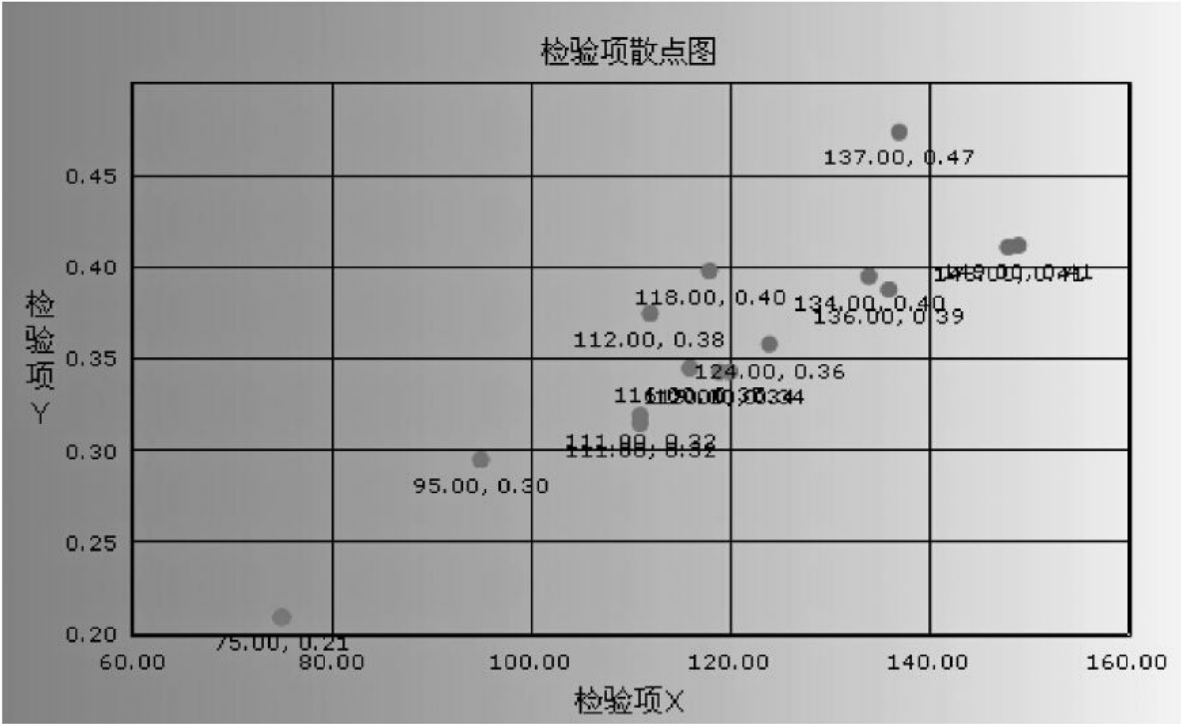


图 6-18

线性回归模型的散点图代码如下：

```
SELECT test_result 1 value1, test_result 2 value2
FROM 指定表;
```

第二个是线性回归模型LOG变换的散点图，如图6-19所示。

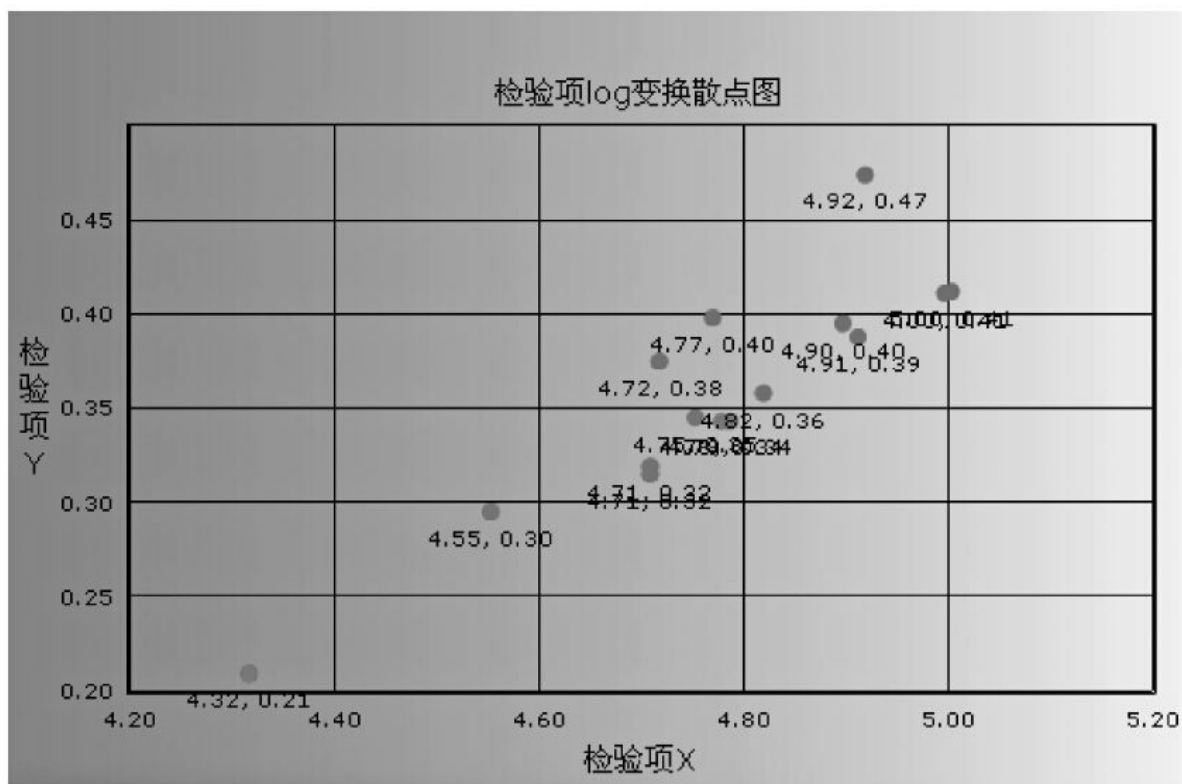


图 6-19

线性回归模型LOG变换的散点图代码如下：

```
SELECT IN (test_result 1) value1, IN (test_result 2) value2
FROM 指定表;
```

第三个是线性回归模型的残差直方图，如图6-20所示。

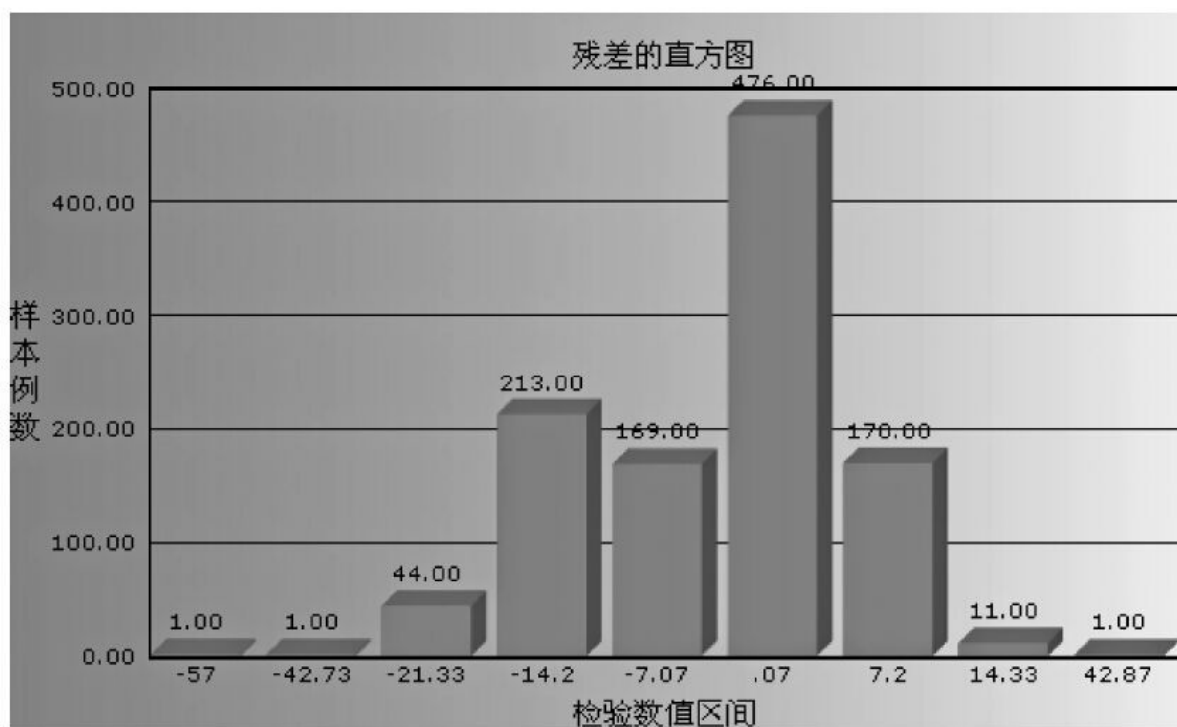


图 6-20

线性回归模型的残差直方图代码如下：

```

SELECT  [y-(aa+bb*x)] e label, COUNT (*) value
FROM  指定表
WHERE  条件
GROUP BY e;
--aa为截距, bb为斜率, e为残差。

```

计算出 e_i 后，将结果插入新表中，然后绘制柱状图。

第四个是线性回归模型的正态概率图，如图6-21所示。

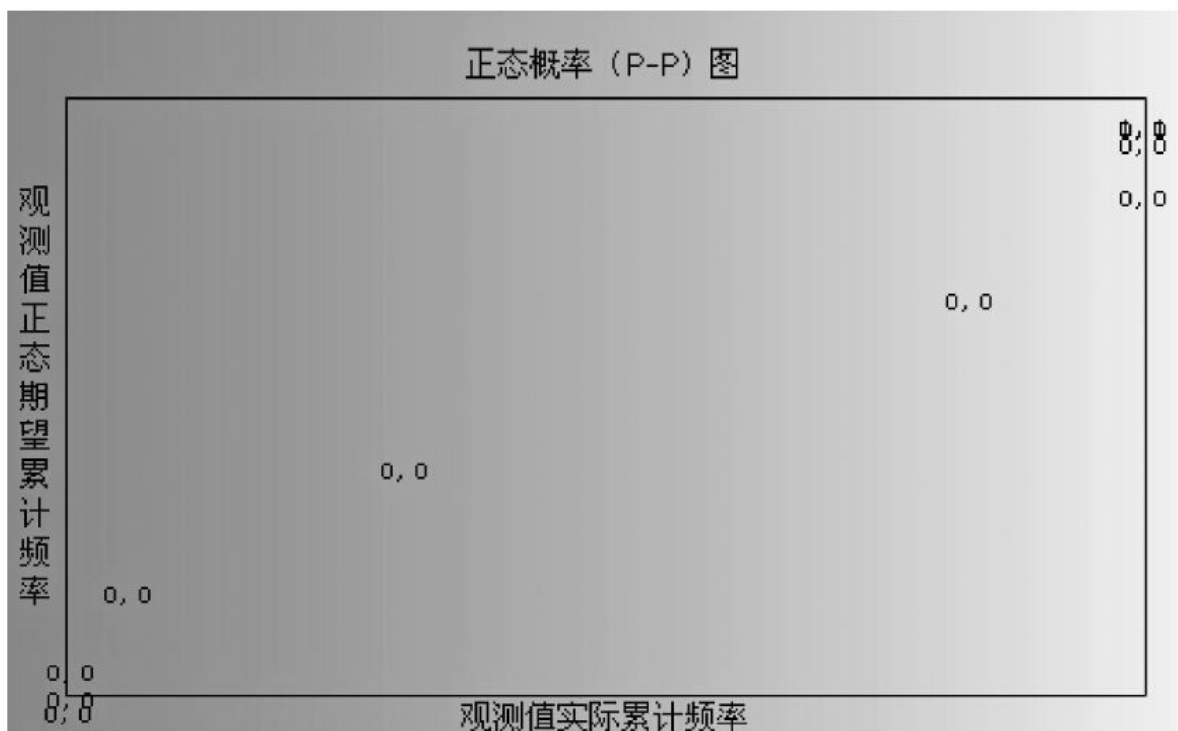


图 6-21

线性回归模型的正态概率图代码如下：

```
SELECT test_result label, test_result value1, test_result _y value2
FROM dm_regress_result;
--test_result为X的累积频数，test_result_y为Y的累积频率。
```

第五个是线性回归模型的标准化残差散点图，如图6-22所示。

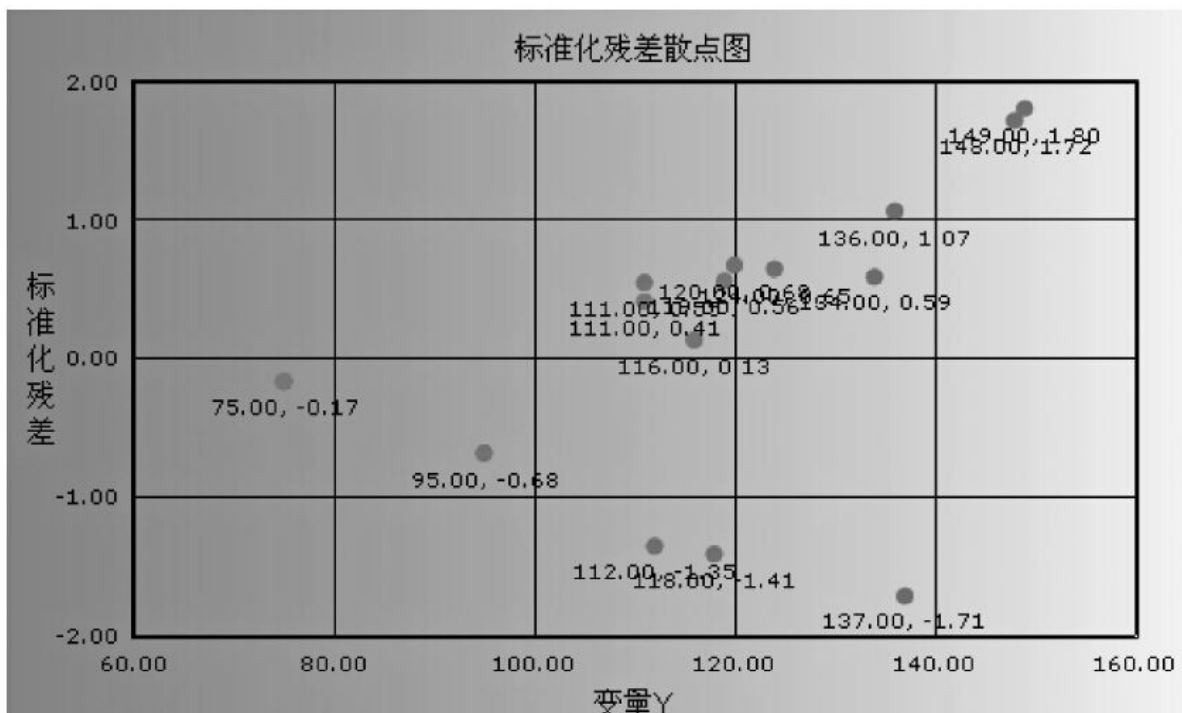


图 6-22

线性回归模型的标准化残差散点图代码如下：

```
SELECT test_result_y label, test_result_y value1, ((test_result - ('|| mean||')) / ('||
stddevy||')) value2
FROM v_regress_residual;
--mean代表残差的均值，stddevy代表的是残差的方差。
```

本例中，从以上线性回归模型的五个相关图中，我们发现基本符合线性回归模型的特征。

6.5.1.3 统计检验

线性回归模型的五个相关图绘制完成后，进入下一步统计检验，得到检验结果如下：

F检验：F=3787.47762

求得：P<0.05

T检验：T=61.54249

求得：P<0.05

计算得回归系数（1— ）的95%可信区间：（273.59893，
292.36307）

$$R^2 = \frac{SS_{\text{回}}}{SS_{\text{总}}}$$
$$R^2 = 0.7775$$

表示可解释发生变异性77.75%，另外22.25%为其他因素解释。

1) 计算方差检验F值，步骤如下：

首先计算方差F公式中的有关参数，该处用到了Oracle的四个聚合函数：

```
SELECT  
ROUND(REGR_SLOPE(tab1.tr1,tab2.tr1),7),ROUND(REGR_SYY(tab1.tr1,tab2.tr1),7),  
ROUND(REGR_SXX(tab1.tr1,tab2.tr1),7),REGR_COUNT(tab1.tr1,tab2.tr1) INTO  
bb,ss,xx,  
count1  
FROM 指定表;
```

聚合函数REGR_SLOPE()返回斜率，REGR_COUNT返回n，
REGR_SYY返回 l_{YY} ，REGR_SXX返回 l_{XX} 。

其中, l_{yy} 的公式为:

$$\sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} \quad (6.1)$$

l_{xx} 的公式为:

$$\sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n} \quad (6.2)$$

然后将计算出的有关参数带入 F 公式中, 代码如下:

```
F: = (xx*bb*bb) / ((ss - ROUND(xx*bb*bb, 7)) / (count1 - 2));
```

即公式:

$$F = \frac{b^2 l_{xx}}{(l_{yy} - b^2 l_{xx}) / (n - 2)} \quad (6.3)$$

P值通过查表可得。

本例中 $P < 0.05$, 则说明线性回归模型可以成立。

2) 计算斜率的T检验值, 步骤如下:

首先计算T值公式中的有关参数:

```
SELECT ROUND(REGR_SLOPE(tab1.tr1,tab2.tr1),7),
ROUND(REGR_SYY(tab1.tr1,tab2.tr1),7),ROUND(REGR_SXX(tab1.tr1,tab2.tr1),7),
REGR_COUNT(tab1.tr1,tab2.tr1) INTO bb,ss,xx,count1
FROM指定表;
```

然后将计算出的有关参数带入T公式中, 代码如下:

```
T: = (bb*(SQRT(xx)/(SQRT((ss-ROUND(xx*bb*bb,7))/(count1-2))));
```

即公式:

$$T = \frac{b \sqrt{l_{xx}}}{\sqrt{(l_{yy} - b^2 l_{xx}) / (n - 2)}} \quad (6.4)$$

P值通过查表可得。

本例中 $P < 0.05$ ，则进一步说明线性回归模型可以成立。

3) 计算置信区间和决定系数。

首先计算与置信区间有关的参数，如下代码：

```
SELECT
ROUND( REGR_SLOPE (tab1.tr1,tab2.tr1),3),ROUND( REGR_SYY (tab1.tr1,tab2.tr1),2),
ROUND( REGR_SXX (tab1.tr1,tab2.tr1),2),REGR_COUNT (tab1.tr1,tab2.tr1),
REGR_COUNT (tab1.tr1,tab2.tr1) INTO bb,ss,xx,count1,count2
FROM 指定表;
```

然后将计算出的参数代入置信区间的公式，代码如下：

```
ss_b:= SQRT((ss-ROUND(xx*bb*bb,2))/(count1-2))/SQRT(xx);
```

即公式：

$$S_b = \frac{\sqrt{\frac{l_{YY} - b^2 l_{XX}}{n - 2}}}{\sqrt{l_{XX}}} \quad (6.5)$$

由此求出回归系数置信区间（interval1，interval2），代码如下：

```
interval1 := bb-t2*ss_b;
interval2 := bb+t2*ss_b;
```

即公式： $b \pm t_{\alpha/2, \nu} S_b$ 。

本例中，回归系数置信区间为（274，293）。

4) 计算决定系数 R^2 公式。

代码如下：

```
SELECT ROUND (REGR_R2 (tab1.tr1,tab2.tr1),4) INTO r2
FROM 表;
```

Oracle的聚合函数regr_r2是返回 R^2 ,

即公式：

$$R^2 = \frac{l_{XY}^2}{l_{XX}l_{YY}} \quad (6.6)$$

本例中， R^2 的值为0.7775，说明“血红蛋白”变化的77.75%可由“红细胞压积”的变化来解释。

6.5.1.4 计算置信带可信区间

建立和验证线性回归模型后，可以计算应变变量Y的置信带可信区间，代码如下：

```
S:= SQRT ((POWER ((x0-x1),2)/xx)+(1/count1))* SQRT ((ss- ROUND (xx*bb*bb,2))/
(count1-2));
```

其中 x_0 为输入的 X 值, bb 为斜率, $count1$ 为 n , $x1$ 为 \bar{X} , xx 为 l_{XX} , ss 为 l_{YY} 。
公式为:

$$S_{\hat{Y}_0} = S_{Y \cdot X} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X - \bar{X})^2}} \quad (6.7)$$

$$S_{Y \cdot X} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - 2}} \quad (6.8)$$

计算出 $S_{\hat{Y}_0}$ 后, 再根据 $\hat{Y}_0 \pm t_{\alpha/2, n-2} S_{\hat{Y}_0}$ 计算置信区间。

计算 Y 的容许区间的部分代码:

```
s1:=SQRT((POWER((x0-x1),2)/xx)+(1/count1)+1)*SQRT((ss-ROUND(xx*bb*bb,2))/(count1-2));
```

即公式:

$$S_{Y_0} = S_{Y \cdot X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X - \bar{X})^2}} \quad (6.9)$$

然后再根据 $\hat{Y}_0 \pm t_{\alpha/2, n-2} S_{Y_0}$ 计算 Y 的容许区间。

6.5.1.5 回归预测

最后, 我们可以应用这个线性回归模型进行预测。例如可以输入一个“红细胞压积” X 值, 进行“血红蛋白” Y 值的区间预测, 结果如下:

预测值 Y_0 的总体均值的95%双侧可信区间: (103.701, 121.717),

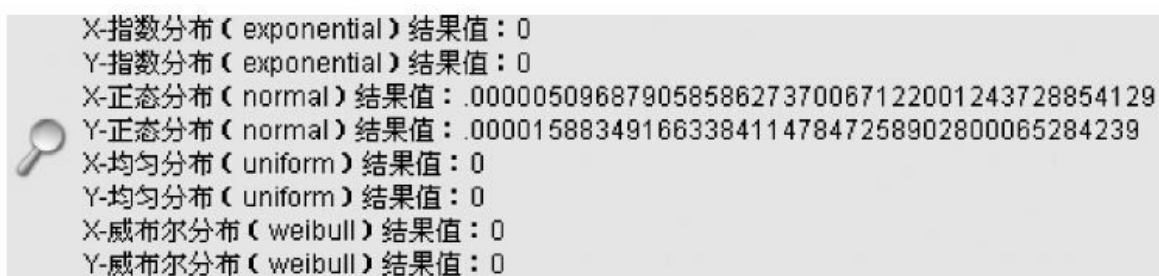
个体 Y 值的95%容许区间: (77.092, 148.326)。

实际应用中, 不是所有的实验室指标都可以得到线性回归模型。当上述各种检验的某一项未通过, 说明该模型有可能不适合做线性回归。如果数据不适合线性回归模型, 我们可以尝试一些非线性回归模型。

6.5.1.6 建立非线性回归模型

1.进行四种分布验证

首先对实验室指标做四种分布验证，即指数分布、正态分布、均匀分布和威布尔分布，分析实验室指标服从哪种分布。四种分布的检验结果如图6-23所示。



X-指数分布 (exponential) 结果值: 0
Y-指数分布 (exponential) 结果值: 0
X-正态分布 (normal) 结果值: .00000509687905858627370067122001243728854129
Y-正态分布 (normal) 结果值: .0000158834916633841147847258902800065284239
X-均匀分布 (uniform) 结果值: 0
Y-均匀分布 (uniform) 结果值: 0
X-威布尔分布 (weibull) 结果值: 0
Y-威布尔分布 (weibull) 结果值: 0

图 6-23

图6-23中的四种分布检验都是用Oracle的DBMS_STAT_FUNCS函数计算的。这里我们说明一下四种分布检验的使用方法。

第一种，指数分布验证，代码如下：

```
DBMS_STAT_FUNCS.EXPONENTIAL_DIST_FIT (  
  ownname    IN    VARCHAR2,--数据库用户名  
  tablename  IN    VARCHAR2,--数据表名  
  columnname IN    VARCHAR2,--进行检验的列名  
  test_type  IN    VARCHAR2 DEFAULT 'KOLMOGOROV_SMIRNOV',  
  --统计检验方法，默认为KOLMOGOROV_SMIRNOV  
  lambda     IN    NUMBER,--比例参数  
  mu         IN    NUMBER,--位置参数  
  sig        OUT   NUMBER);--输出结果P值，越接近1越服从该分布
```

第二种，正态分布验证，代码如下：

```
DBMS_STAT_FUNCS.NORMAL_DIST_FIT (--正态分布验证
```

ownername	IN	VARCHAR2, -数据库用户名
tablename	IN	VARCHAR2, -数据表名
columnname	IN	VARCHAR2, -进行检验的列名
test_type	IN	VARCHAR2 DEFAULT 'SHAPIRO_WILKS', -检验所用的方法, 默认为SHAPIRO_WILKS
mean	IN	NUMBER, -检验列的均值
stdev	IN	NUMBER, -检验列的标准差
sig	OUT	NUMBER); -输出结果, 越接近1, 越服从该分布

第三种, 均匀分布验证, 代码如下:

DBMS_STAT_FUNCS.UNIFORM_DIST_FIT (-均匀分布验证		
ownername	IN	VARCHAR2, -数据库用户名
tablename	IN	VARCHAR2, -数据表名
columnname	IN	VARCHAR2, -进行检验的列名
var_type	IN	VARCHAR2 DEFAULT 'CONTINUOUS', -方差检验的方法, 默认为CONTINUOUS
test_type	IN	VARCHAR2 DEFAULT 'KOLMOGOROV_SMIRNOV', -分布检验的方法, 默认为KOLMOGOROV_SMIRNOV
paramA	IN	NUMBER, -a的样本估计量
paramB	IN	NUMBER, -b的样本估计量
sig	OUT	NUMBER); -输出结果, 越接近1, 越服从该分布

第四种, 威布尔分布检验, 代码如下:

DBMS_STAT_FUNCS.WEIBULL_DIST_FIT (-威布尔分布检验		
ownername	IN	VARCHAR2, -数据库用户名
tablename	IN	VARCHAR2, -数据表名
columnname	IN	VARCHAR2, -进行检验的列名
test_type	IN	VARCHAR2 DEFAULT 'KOLMOGOROV_SMIRNOV', -分布检验的方法, 默认为KOLMOGOROV_SMIRNOV
alpha	IN	NUMBER, -尺度参数
mu	IN	NUMBER, -位置参数
beta	IN	NUMBER, -形状参数
sig	OUT	NUMBER); -输出结果, 越接近1, 越服从该分布

2.非线性模型的建立

由于无法预知非线性模型的类型, 故我们首先尝试五种非线性模型的建立, 分别计算方程、相关系数, 并绘制其散点图。我们分别介绍每一个模型。这里用到了Oracle的三个聚合函数, 包括 REGR_INTERCEPT、REGR_SLOPE和CORR。

1) 第一种非线性方程为非线性幂函数方程。

非线性幂函数方程： $Y=5.665X^{0.867}$

相关系数 $r=0.89256$

非线性幂函数方程代码如下：

```
SELECT ROUND( REGR_INTERCEPT (LN (y) , LN (x) ) , 3) , ROUND( REGR_SLOPE (LN (y) , LN (x) ) , 3) ,  
        ROUND (CORR (LN (y) , LN (x) ) , 5)  
FROM 数据表  
WHERE 条件;
```

—REGR_INTERCEPT (LN (y)、LN (x)) 是返回LN (y)、LN (x) 的截距，REGR_SLOPE (LN (y)、LN (x))
是返回LN (y)、LN (x) 的斜率，CORR (LN (y)、LN (x)) 是返回LN (y)、LN (x) 的相关系数。

非线性幂函数方程的散点，如图6-24所示。

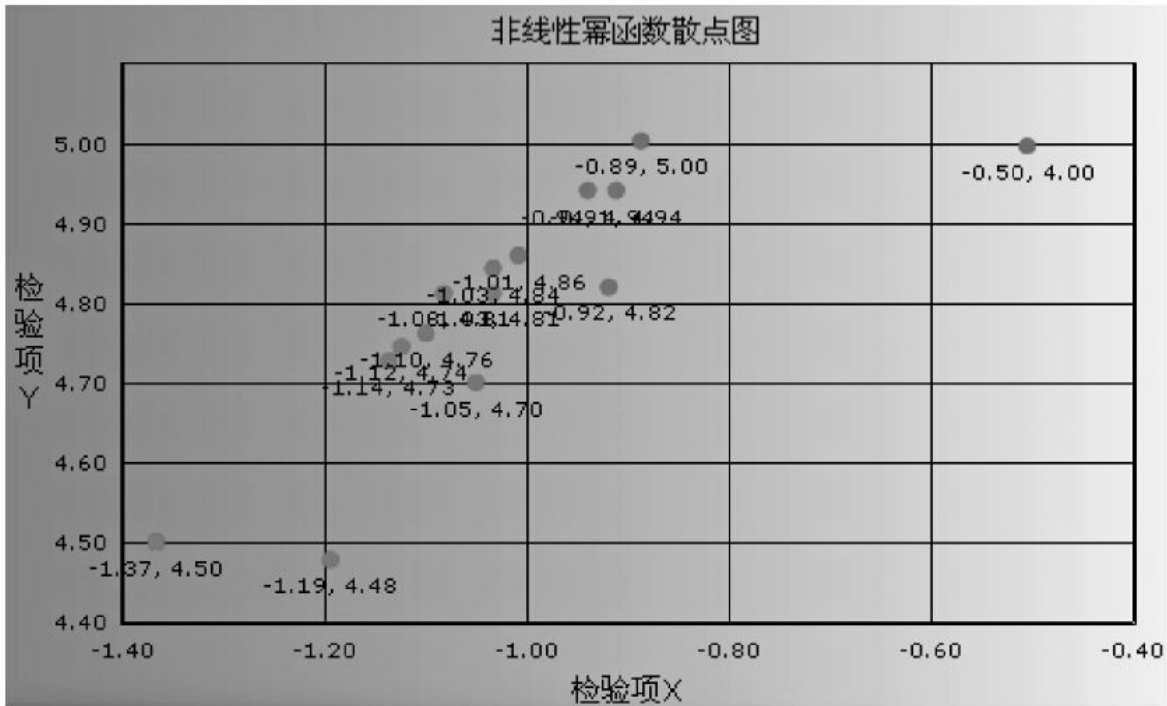


图 6-24

非线性幂函数方程的散点图代码：

```
SELECT LN(x) value1, LN(y) value2
FROM数据表;
```

从图6-24中，我们可以看出非线性幂函数模型的相关系数为0.89，这是一个很强的系数，说明该两组数据相关性很高。类此地，我们可以尝试其他非线性模型。

2) 第二种非线性方程为非线性指数函数方程。

非线性指数函数方程： $Y=3.833 \times e^{2.603X}$

相关系数 $r=0.87682$

实现代码为：

```
SELECT
ROUND (REGR_INTERCEPT (LN (y) , x) , 3) , ROUND (REGR_SLOPE (LN (y) , x) , 3) , ROUND (CORR (LN (y) , x) , 5)
FROM 数据表
WHERE 条件;
```

非线性指数函数方程的散点，如图6-25所示。

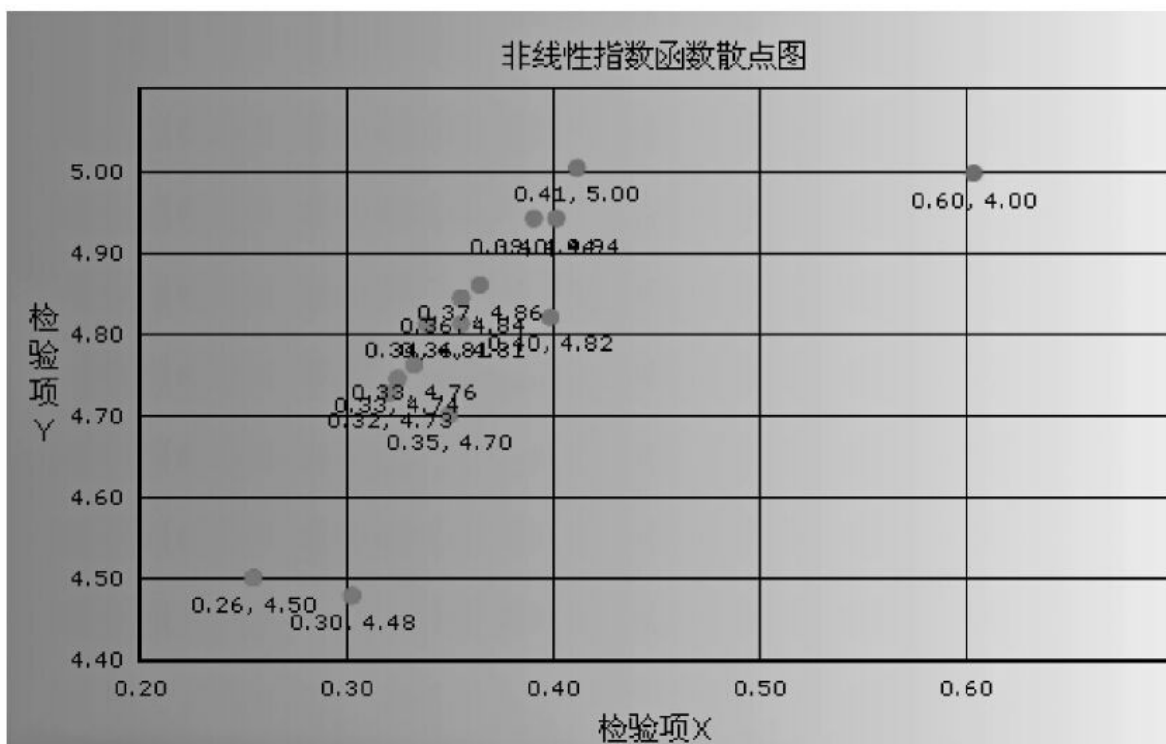


图 6-25

实现代码为:

```
SELECT NULL LINK, patient_id|| patient_adms_id label, X value1, LN(Y) value2
FROM 表;
```

3) 第三种非线性方程为非线性对数函数方程。

非线性对数方程: $Y=214.96+92.827\ln X$

相关系数 $r=0.88439$

实现代码为:

```
SELECT ROUND( REGR_INTERCEPT(y, LN(x)), 3), ROUND( REGR_SLOPE(y, LN(x)), 3),
        ROUND(CORR(y, LN(x)), 5)
FROM 数据表
WHERE 条件;
```

非线性对数函数方程的散点，如图6-26所示。

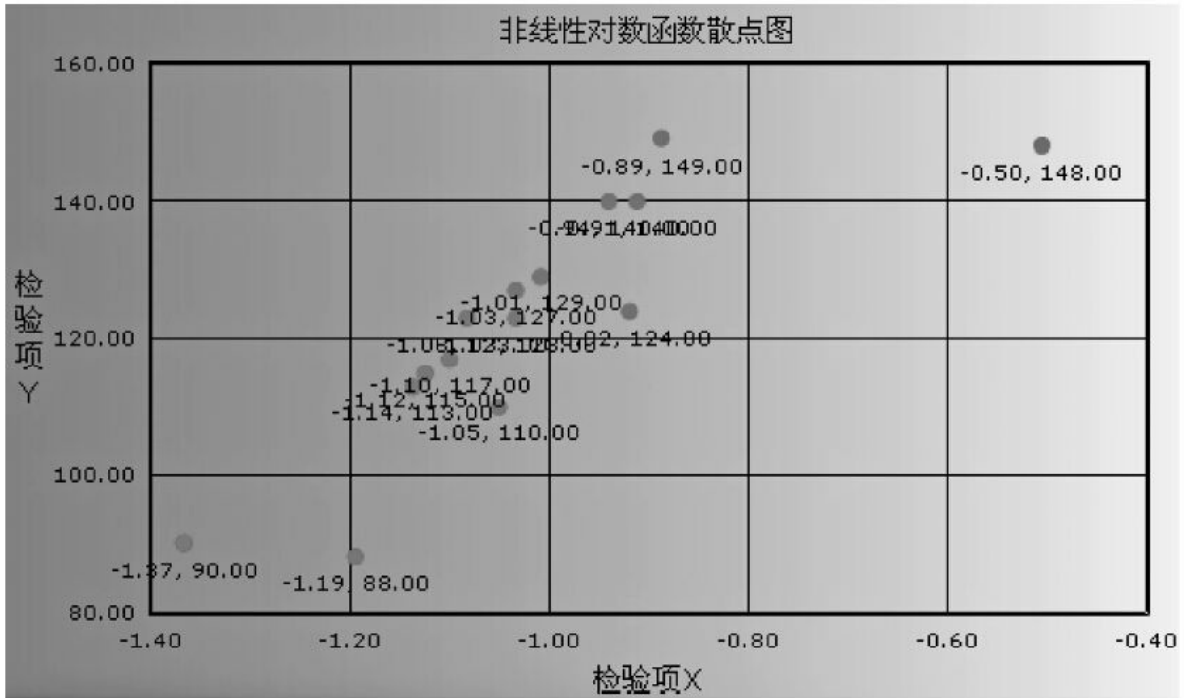


图 6-26

实现代码为：

```
SELECT NULL LINK, patient_id|| patient_adms_id label, LN(x) value1, y value2
FROM 表;
```

4) 第四种非线性方程为非线性双曲线函数方程。

非线性双曲线方程： $y = \frac{x}{0.001x + 0.003}$

相关系数r=0.89599

实现代码为：

```
SELECT ROUND(REGR_INTERCEPT(1/y,1/x),3),ROUND(REGR_SLOPE(1/y,1/x),3),
ROUND(CORR(1/y,1/x),5)
FROM 数据表
```

WHERE 条件;

非线性双曲线函数方程的散点，如图6-27所示。

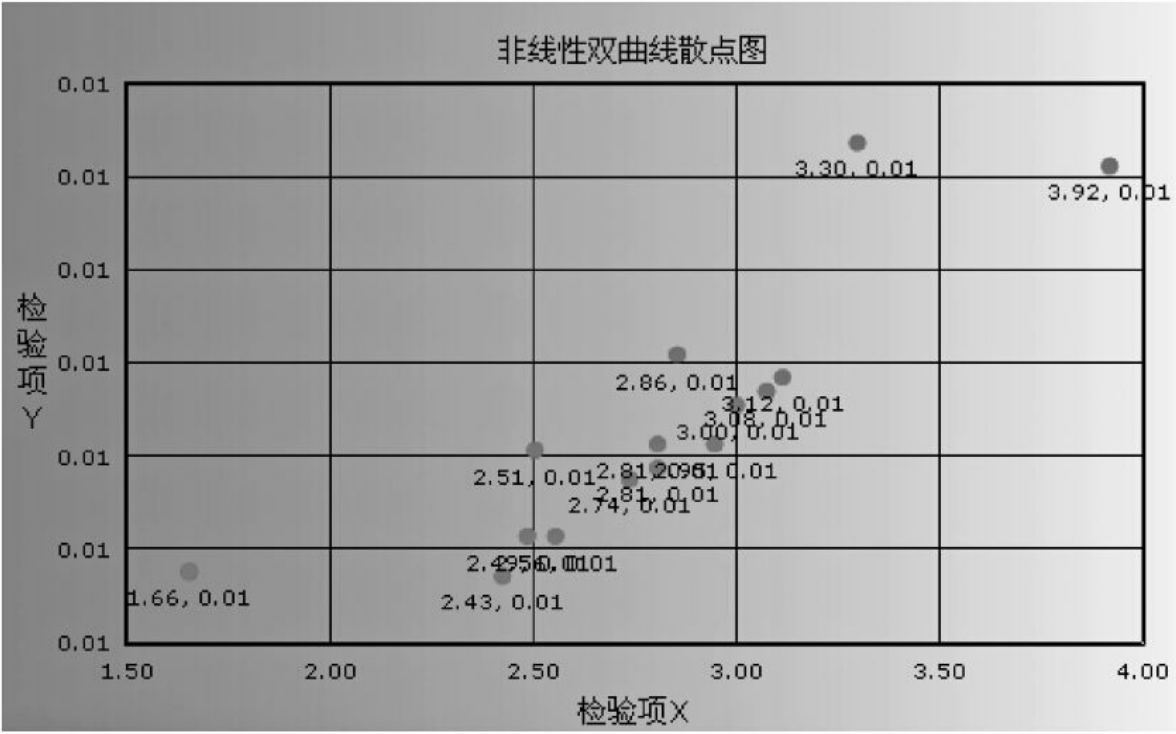


图 6-27

实现代码为:

```
SELECT NULL LINK, patient_id|| patient_adms_id label, 1/x value1, 1/y value2
FROM 表;
```

5) 第五种非线性方程为非线性S型函数方程。

非线性S型方程: $y = \frac{1}{-0.016 + 0.035 \times e^{-x}}$

相关系数r=0.86574

实现代码为:

```
SELECT  
ROUND (REGR_INTERCEPT (1/y, EXP (-x) ) , 3) , ROUND (REGR_SLOPE (1/y, EXP (-x) ) , 3) ,  
    ROUND (CORR (1/y, EXP (-x) ) , 5)  
FROM 数据表  
WHERE 条件;
```

非线性S型函数方程的散点，如图6-28所示。

实现代码为：

```
SELECT NULL LINK, patient_id|| patient_adms_id label, EXP(-x) value1, 1/y value2  
FROM 表;
```

6.5.2 线性相关模型的开发实现

在第3章中，我们详细介绍了线性相关的有关概念和原理。这里，我们介绍一下线性相关模型应用的开发和实现，主要包括二元正态分布验证、计算线性相关模型散点图、计算Spearman秩相关系数，以及计算Pearson相关系数。

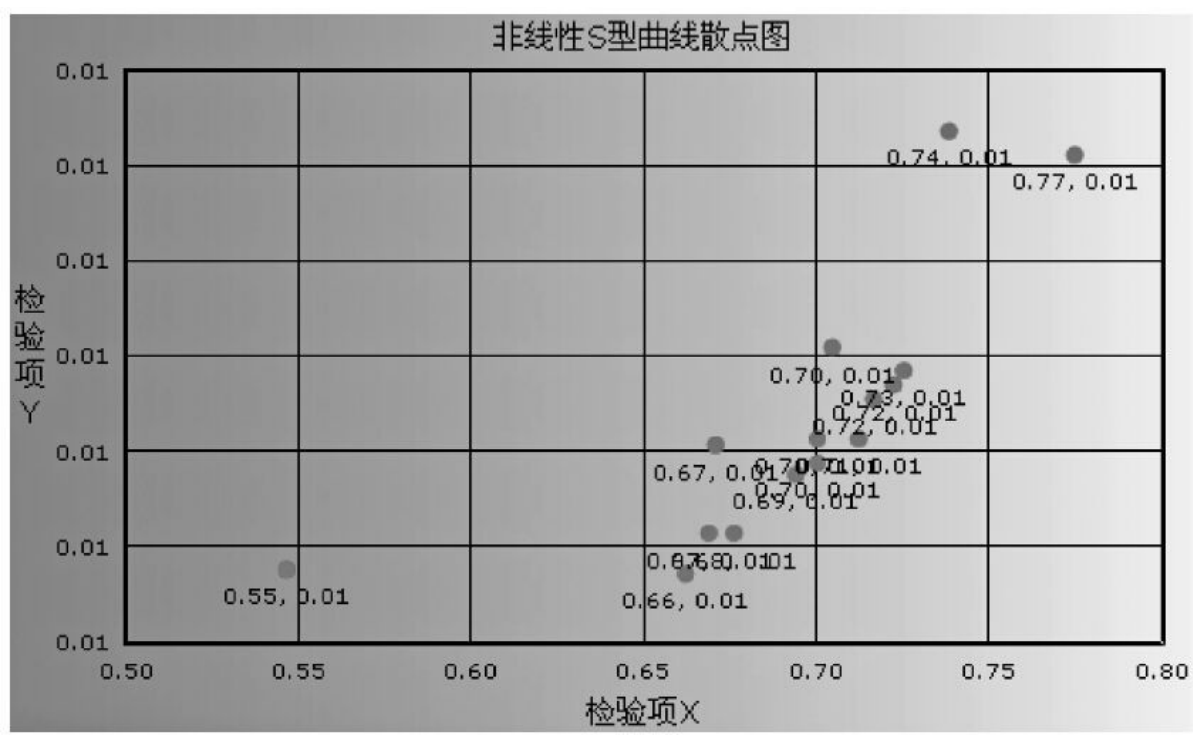


图 6-28

在IRMS系统中，选择“相关分析”模块，其中可以应用三种相关分析：双变量相关分析、复相关分析和偏相关分析。由于复相关分析和偏相关分析比较复杂，占用篇幅较大，这里仅对双变量相关分析的实现做一简单介绍。

6.5.2.1 二元正态分布验证

在本例中，选择“恶性肿瘤术后化疗”的病人组为科研课题病例组，“甘油三脂”和“白细胞”作为线性相关模型分析的两个连续型变量。如图6-29所示，选择好各项参数后，系统自动准备所需数据。

* 观察人群：	恶性肿瘤术后化疗	▼
	(必选)	
* 观察条件：	全部数据分析	▼ (必选)
* 请选择性别：	男/女	▼ (可选)
* 请选择分析类型	双变量分析	▼
* 观察项X1：	★甘油三酯	▼ (必选)
* 观察项X2：	★白细胞	▼ (必选)
* 入院日期：	<input type="text"/>	
	(可选)	
* 结束日期：	<input type="text"/>	
	(可选)	
<div>生成列表</div>		

图 6-29

数据筛选后，首先要做双变量正态验证，判断两种化验指标是否服从二元正态分布。本模块使用了两种正态分布检验方法，一种是基于KOLMOGOROV_SMIRNOV的检验频数分布的正态性检验方法，另一种是基于SHAPIRO_WILKS的检验顺序统计值W的正态性检验方法。结果如图6-30所示。

[illegible]

在此处调用了Oracle的DBMS_STATS_FUNCS中的
NORMAL_DIST_FIT过程，其使用方法如下：

```

DBMS_STAT_FUNCS.NORMAL_DIST_FIT (
  ownername   IN   VARCHAR2,--数据所在的库
  tablename   IN   VARCHAR2,--存储数据的表名
  columnname  IN   VARCHAR2,--进行检验的列名
  test_type   IN   VARCHAR2 DEFAULT 'SHAPIRO_WILKS',
--检验所用的方法，默认为SHAPIRO WILKS
  mean        IN   NUMBER,--检验列的均值
  stdev       IN   NUMBER,--检验列的标准差
  sig         OUT  NUMBER--输出结果
);

```

其中，test_type为检验方法选项，包括四种用法：

- CHI_SQUARED为卡方检验，用于非参数检验。
- KOLMOGOROV_SMIRNOV为检验频数分布的正态性检验。
- ANDERSON_DARLING是基于经验累计分布函数的检验。
- SHAPIRO_WILKS是用顺序统计值W来检验正态性。

6.5.2.2 计算线性相关模型散点图

图6-31为该线性相关模型双变量X和Y的散点图，上一节中已介绍了绘制方法，故此处不再叙述。

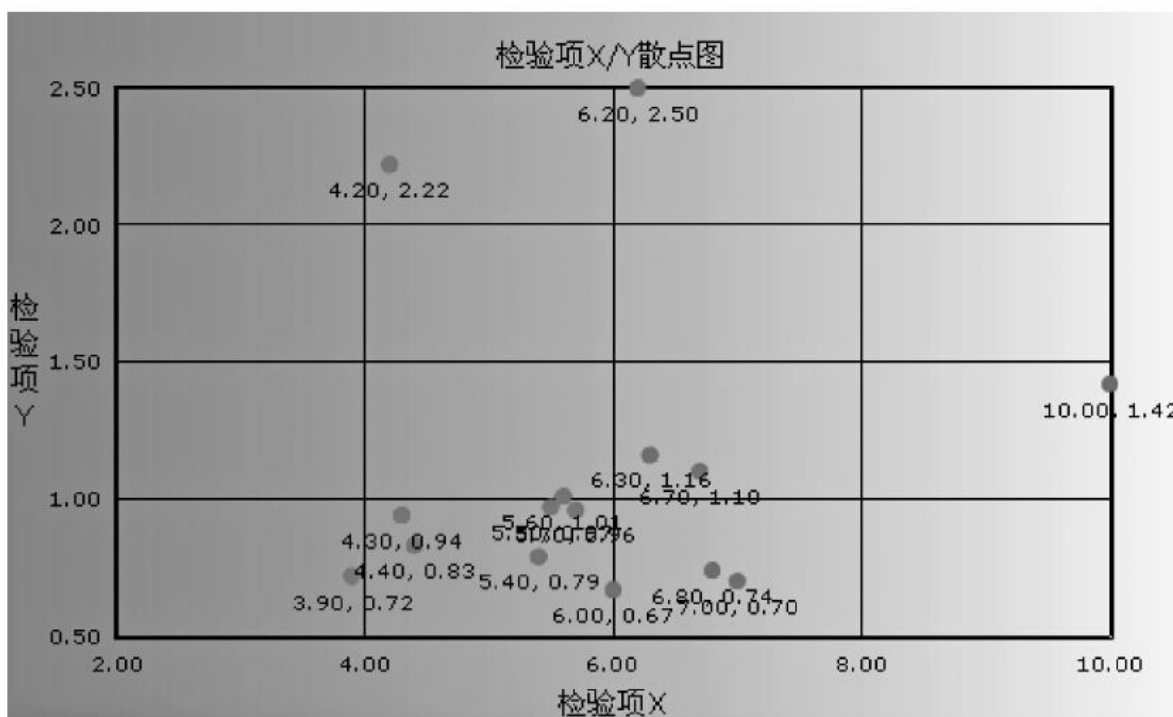


图 6-31

若二元正态分布检验结果说明数据服从正态分布，则执行Pearson相关系数算法，反之则执行Spearman秩相关算法（有关原理和公式参见第3章）。

6.5.2.3 计算Spearman秩相关系数

本例中数据并不服从二元正态分布，故使用Spearman秩相关算法计算相关系数，结果如下：

Spearman秩相关分析：0.0762694

双侧：P>0.05

其代码如下：

```
SELECT SUM((rx-x)*(ry-y)), SQRT(SUM(POWER(rx-x))*SUM(POWER(ry-y))) INTO r1, r2
FROM 数据表;
--其中rx是检验项1的秩次, ry是检验项2的秩次, x是检验项1的平均秩次, y是检验项2的平均秩次
rs:= r1/r2;
```

当数据量小于50时, 通过查表法得到P值, 若数据量大于50, 计算检验统计量Ts, 代码如下:

```
Ts:=rs/(SQRT(1-POWER(rs,2))/SQRT(nn-2));
```

再查表得到P值, 依据P值大小判定两检验项是否相关。

6.5.2.4 计算Pearson相关系数

如果数据服从二元正态分布, 则可以应用Pearson相关系数算法, 代码如下:

```
SELECT CORR(test_result_x1, test_result_x2) INTO st
FROM 数据表;
--CORR() 是返回检验项1和检验项2的相关系数
```

然后计算统计量Tr, 代码如下:

```
Tr:=st/(SQRT(1-POWER(st,2))/SQRT(nn-2));
```

查表得到P值, 依据P值大小判定两检验项是否相关。

6.6 IMRS的算法开发

本节我们将探讨一下IMRS中应用的算法开发。限于篇幅，仅讨论一个我们自己创新的算法，相关因子算法SRCF的实现，以及一个经典的朴素贝叶斯分类算法的实现，并且两个算法均使用数据库环境中的数据库语言实现。

6.6.1 相关因子算法SRCF的实现

SRCF是基于秩和检验的相关因子算法，它由两个子算法组成，一个是SRCF1，一个是SRCF2。本节是有关SRCF1的第一部分的算法实现，用PL/SQL代码完成，算法参见第3章。不同于第3章的SRCF的大算法的一部分，这里将SRCF1.1设计成一个相对独立的小算法，以体现SRCF的大算法的整体算法思路。所以，这里有必要对SRCF1.1的逻辑设计再简要叙述一下。

6.6.1.1 算法思想

有一个数据集table1，Y为二元离散变量， X_i 为连续变量，如表6-16所示。

表6-16 table1

X_1	...	X_i	...	X_n	Y
⋮	⋮	⋮	⋮	⋮	0
					1
					0
					1
					1
					0

1) 对于一个给定数据集：Y为二元观察变量，从 X_1 、...、 X_i 、...、 X_n 变量中寻找Y的相关因子。选取Y=0组和Y=1组，对连续变量 X_i 做Wilcoxon两独立数据秩和检验，将 $p<0.05$ 的变量选为相关因子；

- 2) 当步骤1) 中 $p < 0.05$ 时, 计算 $Y=0$ 组和 $Y=1$ 组的 X_i 的相对熵, 同时也计算两组的特征记分;
- 3) 对 X_1, \dots, X_n 中其他所有连续变量重复步骤1) 和步骤2);
- 4) 将步骤3) 中计算的所有相对熵以及对应的连续变量, 按照相对熵的大小降序排列, 并输出。

6.6.1.2 逻辑算法

```

DECLARE
i NUMBER;
v_rank(i) ARRAY;
BEGIN
IF 数据类型 (Y) <> '离散' AND
(SELECT COUNT(DISTINCT Y) FROM table1 ) <> 2
THEN EXIT; --如果观察变量Y不是二元离散变量, 则退出程序
ELSE
FOR i IN 1..n
IF 数据类型 (Xi) <> '连续'
THEN EXIT; --如果变量Xi不是连续变量, 则退出程序
ELSE
wilcoxon(Xi,Y); --调用两独立数据秩和检验
IF wilcoxon(Xi,Y).p_value < 0.05
THEN --当秩和检验 $p < 0.05$  时
v_re(i):=entropy (Xi,Y);--计算Xi与Y的相对熵
v_rfsc(i):= rfsc (Xi,Y);---计算Xi与Y的特征记分
END IF;
END IF;
END FOR;
END IF;
RANK(v_re(i),Xi); --对Xi及相对熵其进行排序并输出
END;

```

6.6.1.3 算法实现

1.数据准备。

连接数据库, 从数据库中列出所有数据表供用户选择分析, 代码

如下：

```
SELECT table_name
FROM all_tables
WHERE owner='用户名称 ';
```



图 6-32

如图显示该库中所有的表，用户通过程序界面选择其中某个表，例如图6-32中的表“BJ_XK_ALL”。

然后再选择性别SEX作为观察变量Y，它有两个取值1和0，代表男性和女性，最后选择一个连续型变量，如“T”（某实验室生化指标），如图6-33所示。

图 6-33

代码如下：

```
SELECT T CASE WHEN SEX=1 THEN T AS T_m
            WHEN SEX=0 THEN T AS T_f
            ELSE NULL;
END ;
FROM BJ_XK_ALL;
```

2.两独立数据秩和检验。

1) 选取自变量、目标变量并存入数据库中：

```
INSERT INTO 临时表1(a, b, row_num ) SELECT '||变量Xi||','||目标变量Y||',
ROWNUM FROM 所选表 ORDER BY a;
-----在临时表1中按自变量a的升序排列，b为目标变量
```

2) 进行秩和检验：

首先进行编秩，求出自变量对应的秩次，代码如下：

```
INSERT INTO 临时表2 (a, n, rowid) (SELECT a, COUNT(a) ROWNUM FROM临时表1 GROUP BY a
ORDER BY a);
-----在临时表2中对自变量a进行分组，排序及计数
FOR i IN 1...num
-----num为临时表2中的数量
SELECT n FROM 临时表2 WHERE ROWNUM=i INTO m;
SELECT SUM(row_num) FROM 临时表1 WHERE 临时表1.a=(SELECT a FROM 临时表2
WHERE rownum=i) INTO n;
UPDATE 临时表1 SET rank1=n/m WHERE row_num=i;
END;
```

-----求出自变量a对应的所有秩次

然后求秩和，并确定统计量T。按照目标变量的分组，求出分组秩次、秩和及统计量，代码如下：

```
FOR j IN 1..num
-----num为临时表1中目标变量b的分组个数
    UPDATE 临时表3 SET z(j)=' 目标变量b的每组秩次', h(j)=' 每组秩和',
        t(j)=' 每组个数统计量';
    END;
-----将分组秩次、秩和及每组个数统计量放入到临时表3中
```

然后确定P值，代码如下：

```
IF n1<=10 AND n2-n1<=10 AND n1>=2 THEN
-----n1,n2为每组个数统计量，
用查表法确定p值;
ELSE
    EXECUTE IMMEDIATE 'SELECT (ABS('|| t|| '-' || n1|| '*( '|| n1|| '+' ||
n2|| '+1)/2) -
        0.5)/SQRT('|| n1|| '* '|| n2|| '*('|| n1|| '+' || n2|| '+1)/12) FROM dual'
    INTO u;
    -----用正态近似法检验
    IF EXIST(相同秩次) ----如果出现相同秩次
    THEN Y:=0;
        FOR o IN 1..tj
            EXECUTE IMMEDIATE 'SELECT o FROM table_name WHERE row_num='|| o|| ''
        INTO x;
            j:=x*x*x-x;
            Y:=Y+j;
            EXECUTE IMMEDIATE 'SELECT COUNT(*) FROM table_name ' INTO n;
            cc:=1-Y/(n*n*n-n);
            uc:=u/SQRT(cc);
        END IF ;
        查表确定p值;
    END IF;
```

3.对于其他所有自变量，重复2的操作，进行两数据秩和检验。

4.所有通过秩和检验的自变量为目标变量Y的相关因子。

最后，再使用相对熵对所有相关因子进行相关强度分析。

```
EXECUTE IMMEDIATE 'SELECT u FROM 结果表' INTO str7;
IF str7='两组数据有差异'
```

```

THEN r:=(((d*d)/(f*f)+(f*f)/(d*d)-2)+(a-b)*(1/(d*d)+1/(f*f)))/2;
-----计算相对熵的数值
rf:=(((a-b)/(d+f))+(((d*d)+(f*f))/(2*d*f)))/2;
-----计算特征记分的数值
END IF;

```

6.6.1.4 界面功能

下面我们将上述代码在界面应用上的实现做一简述。我们以表BJ_XK_ALL_1为例，选取性别SEX为观察变量，变量分组 $Y_1=0$ ， $Y_2=1$ ，选取T（实验室生化指标）为自变量 X_i ，如图6-34所示。

然后运行程序，点击“单个X变量相对熵”按钮进行相关性分析，运行结果如图6-35所示。我们得到T的相对熵值为0.3982，同时变量分组 Y_1 和 Y_2 组的秩和检验P值 $P<0.05$ ，通过了假设检验，说明T是性别SEX的一个相关因子。

选择分析表

请选择分析表: BJ_XK_ALL_1

请选择观察Y变量: SEX (必选)

请选择分组变量Y1: 0 (必选)

请选择分组变量Y2: 1 (必选)

请选择X变量: T (必选)

按钮: 打开分析表, 进行分组, 单个X变量相对熵, 全部X变量相对熵

图 6-34

列名	相对熵	操作日期	P值	Rfsc
T	.3982	02-DEC-13	p≤0.05	.6091

1 - 1

图 6-35

另外，程序也计算出了T的两组数据的特征计分RFSC，RFSC的值为0.6091。

这里，T的相对熵值和特征计分RFSC的值都比较大，说明T是性别SEX的一个较强的相关因子。

如果用户需要对目标变量与所有自变量进行相关因子分析，可以利用“全部X变量相对熵”按钮的功能，便可得到一个包括所有相关因子分析的交互式报表。如本例中，点击“全部X变量相对熵”按钮，程序自动计算该表中所有的X_i变量与性别SEX的相关性，将所有秩和检验 $P < 0.05$ 的变量筛选出来，运行结果如图6-36所示。



列名	相对熵	操作日期	P值	Rfsc
AGE	.0501	16-JUN-14	$p \leq 0.05$.4399
T	.3982	16-JUN-14	$p \leq 0.05$.6091
TYPE	.6066	16-JUN-14	$p \leq 0.05$.7474
GRADE	.2295	16-JUN-14	$p \leq 0.05$.6051
N	.231	16-JUN-14	$p \leq 0.05$.5536
ST	.153	16-JUN-14	$p \leq 0.05$.4654

图 6-36

从图中，我们可以看出数据表BJ_XK_ALL_1中的“TYPE”、“T”、“N”、“GRADE”几个字段的相对熵值较大，说明它们与性别有较强的相关性；而“H3”、“P27”、“P29”、“TREAT”几个字段的相对熵值较小，说明它们与性别的相关性较弱。另外，我们也得到了这些字段的RFSC特征计分值。与相对熵的值相比较，RFSC特征计分值与相对熵有相似的度量方向，但二者并不完全相同，详细讨论可参见第4章。这里是采用相对熵的值选择相关因子。

6.6.2 朴素贝叶斯分类算法的实现

朴素贝叶斯分类算法是个比较经典的自动分类方法，它具有分类准确率较高（分类变量的独立性较好的条件下）、算法简单、分类效率较高的特点，在很多实践应用中被广泛采用。

虽然Oracle数据库中提供了朴素贝叶斯分类算法API，但我们还是自己编写了一个运行在数据库环境下的朴素贝叶斯分类算法程序，一方面它可以被移植、部署到没有朴素贝叶斯分类算法API的任何数据库中，另一方面我们还可以根据实际应用需要对算法进行改进和优化。下面就简要叙述一下朴素贝叶斯分类算法数据库的实现方法。

6.6.2.1 数据准备

连接数据库，从数据库任意选择一个需要分析的数据表，代码如下：

```
SELECT table_name
FROM all_tables
WHERE owner='用户名称' ;
```

图6-37显示该库中所有的表。



图 6-37

将该用户的所有数据表名在应用程序中列出，并由人工选择某一数据表名（table_name），用于进行分类分析。

清空临时表，逻辑代码如下：

```
EXECUTE IMMEDIATE 'TRUNCATE TABLE "临时表1"';
```

存放准备数据到临时表中：

```
EXECUTE IMMEDIATE 'INSERT INTO "临时表1" SELECT * FROM table_name';
```

临时表中null值的处理:

```
DELETE FROM table_name WHERE '任意字段' IS NULL ;
```

6.6.2.2 算法过程

1) 将该数据按一定比例（例如70%、30%）分成两组，前一组数据用于训练朴素贝叶斯分类模型，后一组数据用于对分类模型进行测试和评估。

这里涉及两组数据如何选取的问题，我们使用数据随机选取函数SAMPLE()，按照事先设定的数据比例，从数据表中随机选取两组数据，代码如下：

```
SELECT * FROM table_name SAMPLE(行数);
```

2) 假定该数据表中有s个分类因子，目标分类变量有m个取值，即m个分类，数据总数有n条，计算目标分类变量的先验概率：

首先计算目标分类变量的总数，代码如下：

```
SELECT COUNT(*) INTO ALL_CNT FROM table_name;
```

然后计算目标分类变量的每个类别Ci的个数，代码如下：

```
FOR i IN 1..m-----m是类别个数  
  SELECT '目标分类变量', COUNT(*) INTO NUMi  
  FROM table_name
```

```
GROUP BY '目标分类变量'-----NUMi 为一数组变量;
END ;
```

其中每个分类值的概率 $P(C_i)$ 的计算公式为 NUM_i/ALL_CNT 。

3) 接下来计算每个分类因子的每种取值与每种目标分类同时出现的概率 $P(X_i|C_j)$ ，代码如下：

```
FOR i IN 1..num
-----num是目标类别个数
    SELECT DATA_TYPE FROM user_tab_columns WHERE user_tab_columns.Table_Name =
        table_name AND column_name = '分类因子';
-----读取分类因子的数据类型
IF
    DATA_TYPE =离散型
THEN
    P(Xi| Cj)= 第i个分类因子取值与第j个目标分类同时出现的个数/第j个目标分类出现的总个数;
ELSE- -DATA_TYPE =连续型
    P(Xi| Cj):=(1/sqrt(2*pi*STDDEV('分类因子')))*EXP(-(Xi-avg('分类因子'))*(
        Xi-avg('分类因子'))/(2*STDDEV('分类因子')*STDDEV('分类因子')));
    --计算公式: P(Xi| Cj)=g(x,u,σ)=1/√2πσe-(x-u)2/2σ2 (u为平均值, σ为标准差)
END IF;
END ;
```

将以上代码计算得到的所有 $P(X_i|C_j)$ 的结果存放于'临时表2'，代码如下：

```
FOR i IN 1..num-----num是类别个数
    INSERT INTO '临时表2'(P_字段) VALUE(P(Xi| Cj));
COMMIT;
END;
```

最后计算每个分类因子组合与每个目标分类同时出现在整个先验数据中的概率： $P_{ij} = P(C_j) * P(X_i|C_j)$ ；

3) 将计算得到的所有先验概率 P_{ij} 存放到临时表组中，我们得到了一个朴素贝叶斯分类模型。

6.6.2.3 模型验证

我们将对计算出来的朴素贝叶斯分类模型进行验证。正如前面提到的，我们将该数据按一定比例（例如70%、30%）分成两组，前一组数据用于建立朴素贝叶斯分类模型，后一组数据用于对分类模型的测试。测试的方法是，将测试数据中的每一条数据的分类因子组合（用 Y_i 表示），应用训练得到的朴素贝叶斯分类模型，计算出最有可能的目标分类：

```
FOR j IN 1..m -----m是类别个数
  Pij:=P(Cj)* P(Yi| Cj);
END;
```

我们比较所有 m 个概率值，取值最大的一个分类作为该条数据的目标分类，然后用这个目标分类值与数据原目标值进行比对，如果两值相等则推断预测正确，反之推断预测错误。

数据测试的计算方法与建立模型的过程类似，逻辑代码如下：

```
BEGIN
-----P(X| C) 值计算:
FOR i IN 1.. n*30% -----n*30%为测试数据量
  FOR g IN 1..s -----s是分类因子个数
    IF Xg的取值没有在先验数据中出现过 -----Xg为第g个分类因子
    THEN EXIT; -----那么要把该数据排除在测试数据之外
    ELSE
      select DATA_TYPE from user_tab_columns WHERE user_tab_columns.Table_Name =
        测试数据表 AND column_name = 'Xg';
      -----读取分类因子的数据类型
      IF
        DATA_TYPE =离散型
      THEN
        FOR j IN 1..m -----m是类别个数
          从先验模型中取i条数据的第g个分类因子与第j个目标分类同时出现的概率P(Xg| Cj);
          END;--结束m次循环
        ELSE -----连续型概率计算
          FOR j IN 1..m -----m是类别个数
            P(Xg| Cj):=(1/sqrt(2*pi*STDDEV('Xg')))*EXP(-(Xg-avg('Xg'))*
              (Xg-avg('Xg'))/(2*STDDEV('Xg')*STDDEV('Xg')));
```

```

END;-----结束m次循环
END IF ;
END IF;
END; ;-----结束s次循环
FOR j IN 1..m (m是类别个数)
P(Y| Cj): = P(X1| Cj)* P(X2| Cj)*...* P(Xs| Cj); ----- [Y=(X1,X2,...,Xs)]
P(Y)j:=P(Y| Cj)* P(Cj);
END; -----结束m次循环
IF max(P(Y)j)对应的Cj=Ci
-----对于所有测试数据，取每条数据的组合因子与目标分类变量同时出现的概率值最大的
分类作为预测分类
THEN
flag=1
ELSE
flag=0
END IF;
INSERT INTO '临时表3' SELECT flag WHERE rownum=i;
END; -----结束n*30%次循环
SELECT flag, COUNT(*) AS cnt_num, (COUNT(*)/(n*30%))*100% AS cnt_percent
FROM '临时表3'
GROUP BY flag
ORDER BY flag;
-----计算测试数据分类模型预测的正确率
END;

```

6.6.2.4 界面应用

下面我们将说明在实际项目中是如何运用该算法的。分类因子以五个为例，如图6-38所示。

图 6-38

首先选取目标表（以表BJ_XK_ALL为例），然后选取该表中的字段作为目标分类变量和分类因子。这里，我们选取性别“SEX”字段作为目标分类变量，实验室生化指标“LAB1”、“LAB2”、“LAB3”、“LAB4”、“LAB5”作为性别SEX字段的分类因子。然后选取分类模型测试数据与训练数据的比例，这里我们输入0.3，即数据表中30%的数据用作测试数据，数据表中70%的数据用作训练数据。

所有参数设定后，点击“模型运算”按钮，我们事先写好的朴素贝叶斯分类算法在数据库后台进行自动计算、训练，最后得到一个分类模型，如图6-39所示。

分类因子1	分类因子2	分类因子3	分类因子4	分类因子5	分类变量	概率
2	2	2	2	3	1	.009944
2	2	2	1	3	1	.007896
2	2	2	2	2	1	.007543
3	2	2	2	3	1	.007373
2	2	3	2	3	1	.006749
2	3	2	2	3	1	.006629
2	2	2	1	2	1	.00599
3	2	2	1	3	1	.005855
3	2	2	2	2	1	.005593
2	2	3	1	3	1	.005359
2	3	2	1	3	1	.005264
2	2	3	2	2	1	.00512
2	3	2	2	2	1	.005029
3	2	3	2	3	1	.005004
3	3	2	2	3	1	.004915

图 6-39

图中详细列出了实验室生化指标“LAB1”、“LAB2”、“LAB3”、“LAB4”、“LAB5”在各种不同的取值时与目标分类变量性别“SEX”不同取值共同出现的概率。例如第一行显示，当：“LAB1”=2，且“LAB2”=2，且“LAB3”=2，且“LAB4”=2，且“LAB5”=3，且“SEX”=1时的概率是0.0099，也就是接近0.01。

分类模型训练后，对该分类模型进行评估。点击“模型测试”按钮，系统自动运行模型的评估算法。在本例中，运行结果如下：输出结果：

输出结果：
正确率=0.7375
错误率=0.2625

错误率=0.2625我们看到，这个朴素贝叶斯分类模型的预测正确率为73.8%，说明“LAB1”、“LAB2”、“LAB3”、“LAB4”、“LAB5”对“SEX”的区分能力较强。换句话说，男性组与女性组在综合“LAB1”、“LAB2”、“LAB3”、“LAB4”、“LAB5”五个实验室生化指标方面有较明显的差异。

当然，以上只是朴素贝叶斯分类算法在数据库上的一个简单实现，还有很多功能可以改进和完善，在此不再赘述。

6.7 本章小结

数据挖掘的应用，现在更多的是作为一个定制化方案服务于商业应用中，而作为面向行业业务的数据挖掘工具的应用还不是太多，但随着数据挖掘技术的不断发展、数据挖掘应用的不断普及，以及数据挖掘在各行各业应用的不断成熟和改进，专业性的数据挖掘工具的应用将越来越先进、越来越普及，也必将在各种业务中发挥越来越大的作用。正像近三十年来IT技术和各种各样的信息化系统的快速发展，并已经渗透到人类社会发展的各个角落那样，数据挖掘技术和各种各样的数据挖掘应用系统也必将融入到大数据时代的各个领域。

第7章 数据挖掘应用系统的应用

上一章从技术角度详细介绍了数据挖掘应用系统的开发，重点剖析了智能医学科研系统IMRS的技术架构、数据变换、模型创建、功能实现等方面内容。既然数据挖掘应用系统已经建好了，那么本章将着重探讨一下数据挖掘应用系统的应用了。这听起来有点绕口，但的确是本书的一个重点章节。因为任何技术，即使再高明也是为业务服务的，帮助业务产生新的价值、新的收益，这才能体现技术的价值。数据挖掘技术更是如此，我们不希望它成为追求时髦的概念炒作、故弄玄虚的IT技术、高深莫测的技术手段，甚至沽名钓誉的“忽悠”工具，而是真正为实践服务、为业务服务的好方法。任何不能为业务提供有效帮助的数据挖掘项目，都是毫无意义的、毫无价值的。当然，另一方面，失败的数据挖掘项目是托起成功者的巨人，也是通向成功目标的必经之路。

临床医学是一门知识体系庞大、复杂的学科，人类对医学科学的探索有着巨大的需求和推力。临床医学的日常工作产生了海量的临床医学数据，而这些临床医学大数据中蕴含着很多尚未发现的医学规律。临床医学数据挖掘系统正是一套试图从临床医学大数据中挖掘、探索未知的医学特征、医学关系的有力工具，它是一套通用于医学各学科的临床医学科研平台。应用系统针对临床医学数据的特点，首先建立了专用的医学科研数据仓库（参见第5章有关章节），然后转换成适于数据挖掘算法建模的可挖掘数据。应用系统中应用了多种类型

的算法，而且很多算法的应用方法也不同于一般的常用方式，根据医学科研的目的和需求具有一定医学应用特色，我们将在下面的章节中详细叙述。

数据挖掘的行业应用系统往往与数据仓库或规模小一些的数据集市结合在一起，临床医学数据挖掘系统也是这样，它还拥有快速探索、专业数据筛选、专业统计学等模块，与数据挖掘引擎构成一个完整的流程环。这些模块建立在共同的数据仓库的基础上，提供给使用者从数据获取，到数据预分析，再到数据筛选，数据挖掘，最后到结果评估等一系列功能的一站式服务。数据仓库作为应用系统各个功能模块的基础，不但使数据挖掘系统成为建立在海量数据上的有根之木、有源之水，也为各个功能模块的协同和数据挖掘引擎各个算法的平滑协同提供了基础。临床医学数据挖掘系统所有的这些功能，都应该体现到这样一句话：方便发现隐蔽的“真相”，助于挖掘深藏的“真金”。

下面，我们就详细介绍临床医学数据挖掘系统的一个实现实例：智能医学科研系统IMRS的应用。正如我们第5.6节所归纳的，按照临床科研的普遍需求，我们将临床科研数据挖掘系统的功能划分为六个方向：分布探索、关系探索、特征探索、异常探索、推测探索和趋势探索。本章将在IMRS中实现的分布探索、关系探索、特征探索、异常探索、推测探索五个方面的应用进行逐一介绍。

7.1 分布探索

事物的分布是事物重要的特征之一，也是最容易被观察和理解的。我们生活中接触到以分布为特色区分事物的例子比比皆是。比如按照克拉数来衡量钻石的价值，按照斤两多少来定价大闸蟹等。这里克拉数就是反映钻石的一个重要特征的分布度量；斤两数也是反映大闸蟹的一个重要特征的分布度量，位于不同斤两数区间的大闸蟹其定价不同，这些都属于数据分布的概念。数据挖掘对数据分布的探索工具很多，我们现在谈一下IMRS系统应用中常用的聚类模型的应用，包括两维度聚类应用和高维度聚类应用。

数据挖掘的聚类分析，是一种用途广泛的无监督学习分类算法，旨在揭示数据的客观分布规律。在行业应用系统中，在满足行业需求的同时，聚类的维度尽量选择数据容易获取、可解释性强而且可信度高的。在IMRS系统中，年龄和性别是两个最基本和最主要的临床指标之一。年龄和性别的数据获取容易，一般不需要特殊的设备即可获得，数据采集的可靠性较高，因为人们对性别和年龄的记忆相当准确。年龄和性别的分析结果可解释性强，也具有较广泛的通用性。

7.1.1 两维度聚类模型应用

使用聚类分析是探索某一疾病在年龄（连续型变量，范围通常为1至100）、性别（离散型变量，取值为‘男’，‘女’）两变量上的分布规律，为科研者和临床医师研究和分析特定疾病的性别特征和年龄规律提供重要的依据。科研者可根据聚类分析的结果进一步深入探讨疾病发病的影响因子和规律，并可能提出新的医学假说。聚类分析还可作为智能分析工具，帮助科研者快速发现高发年龄段和高发性别组等医学规律。通过不同疾病年龄、性别的分布比较，还可研究不同疾病发病规律的异同。该工具还可应用于将挖掘到的聚类组（如年龄聚类组）用于科研的区组设计上，从而为科研区组设计提供更科学、更严谨的依据和方法。

1. 操作流程

我们在这里给出一个IMRS系统应用的具体步骤，以便使读者对系统应用有更直观和更具体的了解。临床医学数据挖掘系统（IMRS系统的核心子系统）是一个全向导式的智能分析工具，用户只需按照向导的指示工作即可。基于数据仓库，系统的数据获取非常容易，如图7-1所示。



图 7-1

选择“聚类分析-疾病”模块。选择需要分析的目标疾病人群，如“冠状动脉硬化性心脏病”，然后点击“准备分析”按钮进入下一页面，如图7-2所示。



图 7-2

这时，系统会为选定的“冠状动脉粥样硬化性心脏病”的病例组准

备聚类分析所需要的数据。点击“准备数据”按钮，系统会自动准备数据挖掘所要求的带有性别和年龄变量的数据。

数据准备成功后进入下一界面，如图7-3所示。点击“运行模型”按钮，系统自动运行配置好的数据挖掘引擎算法，并自动产生最佳聚类模型。

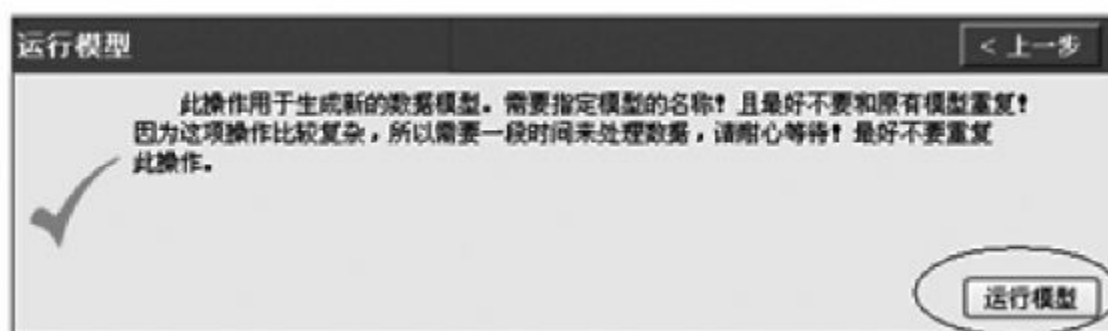


图 7-3

模型运行成功后进入下一界面，如图7-4所示。此时聚类分析的结果已经存储在数据库中了。所有的数据挖掘的步骤都是在数据库后台，通过自动生成的代码执行的。用户点击“取模型结果”按钮，系统从成功运行的模型中取出结果。



图 7-4

模型结果取出后，系统进入下一界面，如图7-5和图7-6所示。我们得到该疾病“冠状动脉粥样硬化性心脏病”按年龄和性别聚类的分布结果，包括左边的树形分布规则集，中间的规则明细，和“叶子”（最底层）规则表。其中“叶子”规则表可以根据需要，输出到Excel文档中，用于生成论文和科研报告。

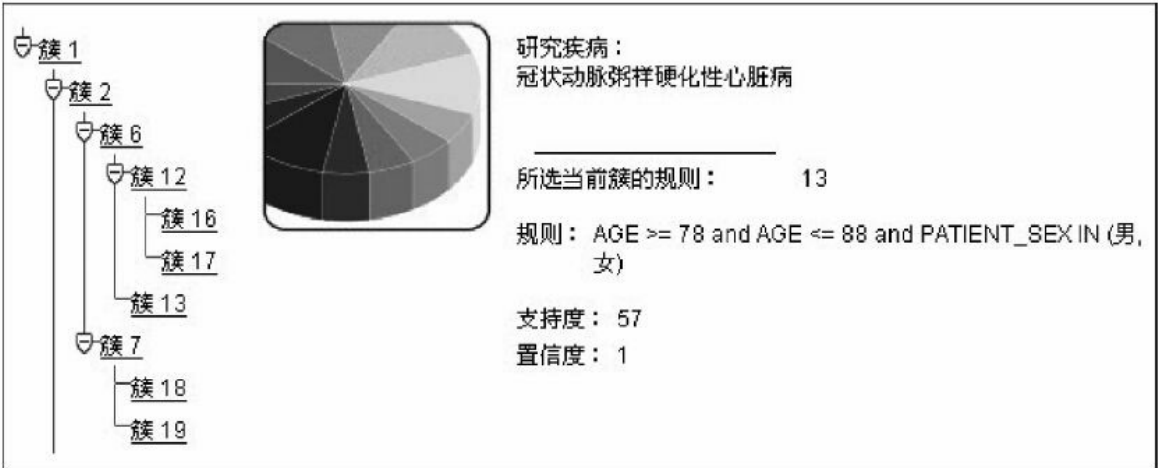


图 7-5

规则序号	规则描述	支持度	置信度
8	PATIENT_SEX IN 男 and AGE <= 55,AGE >= 50	12	.800000000000000004
13	PATIENT_SEX IN 男,PATIENT_SEX IN 女 and AGE <= 88,AGE >= 78	57	1
14	PATIENT_SEX IN 男 and AGE <= 50,AGE >= 41	10	.76923076923076905
16	PATIENT_SEX IN 男,PATIENT_SEX IN 女 and AGE <= 97,AGE >= 88	14	1
17	PATIENT_SEX IN 男,PATIENT_SEX IN 女 and AGE <= 88,AGE >= 83	37	.80434782608695699

图 7-6

2.应用结果

在这里我们会遇到数据挖掘行业应用系统应用的一个难点，即输

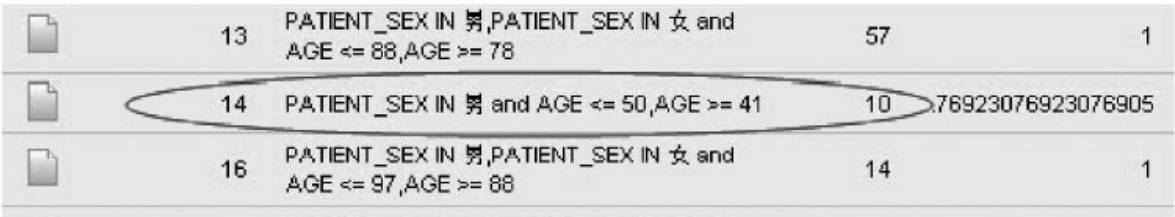
出结果的专业性分析。这个环节需要数据挖掘人员与行业专业人员紧密配合，取长补短，共同讨论分析，也可由掌握数据挖掘的一定原理并具备相当的本领域数据挖掘应用经验的行业专业人员独立进行研究分析。

下面我们就从三个方面对此模型进行研究分析。

(1) 性别倾向

这里，经医学专业与技术专家的合议，对上面这个数据挖掘的输出结果，我们建议从以下几个方面来分析和理解：

1) 存在单一性别分支。如图7-7圈中所示。






	13	PATIENT_SEX IN 男, PATIENT_SEX IN 女 and AGE <= 88, AGE >= 78	57	1
	14	PATIENT_SEX IN 男 and AGE <= 50, AGE >= 41	10	.76923076923076905
	16	PATIENT_SEX IN 男, PATIENT_SEX IN 女 and AGE <= 97, AGE >= 88	14	1

图 7-7

模型提示该病种在性别上具有发病的偏向性，这个例子是偏向于男性。

2) 单性别分支的置信度较高。

性别维度对于一个单一个体而言是个二元类型变量，而对于群体来说则要复杂得多，存在从单纯男性组到单纯女性组之间有各种不同性别组合分布，如图7-8所示。

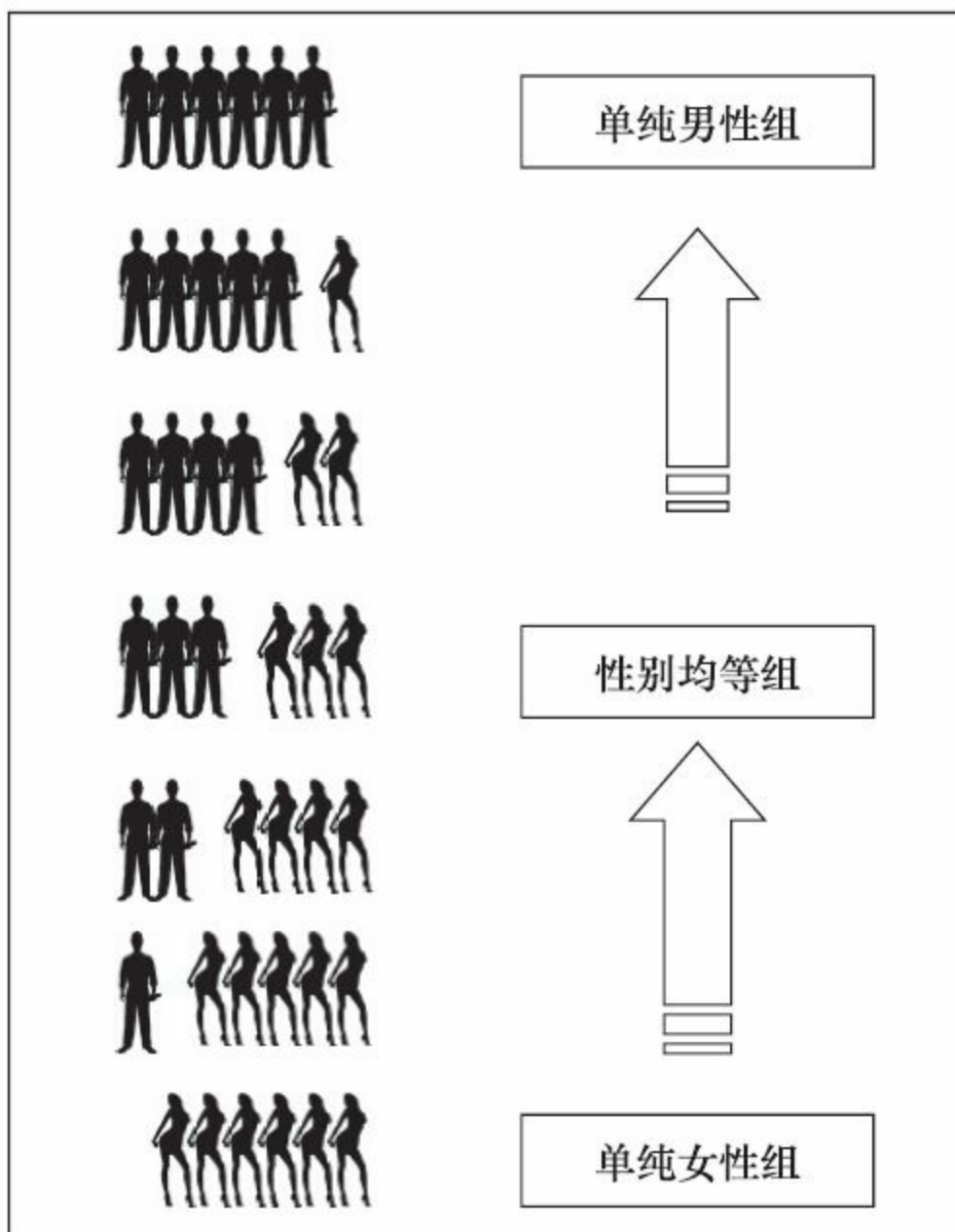


图 7-8

分支的置信度较高，提示这个分支对性别的偏向程度高。

3) 划分的次序。

本聚类算法是按照多维度变量权重来划分的，因此聚类分析的输出结果会在规则中使用不同的次序来体现维度的不同权重。观察图7-

5可以发现在年龄和性别聚类的第一个划分是年龄。对比一下图7-9乳腺疾病的双维度聚类，乳腺疾病的第一个划分是性别。

而本例第一个划分的是年龄，可以推断冠心病的年龄维度区别更加明显，年龄分簇先于性别分簇。而乳腺疾病的性别维度区别更加明显，说明乳腺疾病的性别分簇先于年龄分簇。分簇次序的不同，提示疾病的年龄分布和性别分布哪个更强一些。乳腺疾病具有更强的性别倾向性，性别的倾向性大于年龄；而冠心病的不同年龄段分布更强一些，所以第一次划分是年龄维度划分。这是疾病本身的属性特点反映到了数据挖掘分簇现象的结果上。

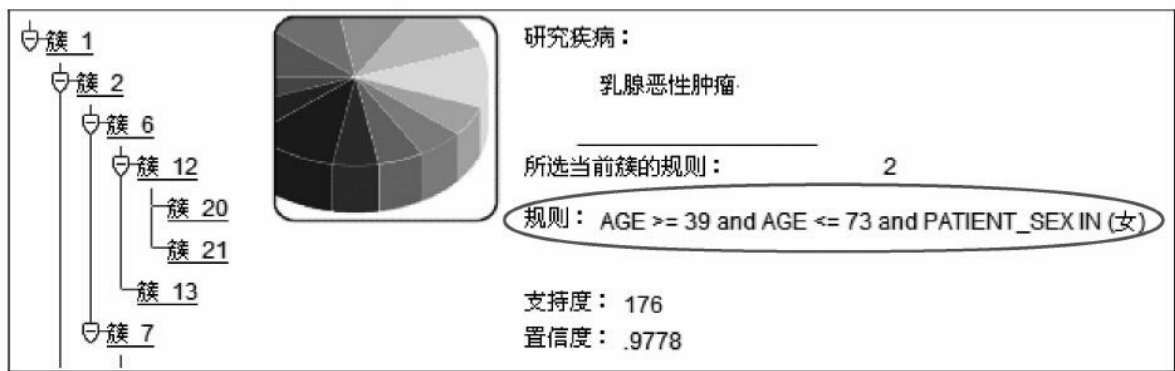


图 7-9

（2）高发人群

图7-10中的“叶子规则表”是一功能强大的交互式智能报表。点击其右上角的“支持度”按钮，将弹出一对话框，点击“降序排列↓”按钮，将得到支持度由高至低排列的规则集，如图7-10所示，从中我们可以得到高支持度的规则。如示例中的规则13：年龄在46.69与61.92之间，性别不限，共有951个案例，居于所有规则的首位，我们可据

此理解为冠状动脉粥样硬化性心脏病的高发年龄在46.69与61.92之间的可能性很大。

用类似的方法，可以得到高置信度的规则集。同样，点击其右上角的“置信度”按钮，将弹出一对话框，点击“降序排列↓”按钮，将得到置信度由高至低排列的规则集，从中我们可以轻易地得到高置信度的规则。一般地说，70%以上的置信度可视为高置信度规则，当然最好还要结合专业领域的实际情况评价规则置信度的高低。



规则序号	规则描述	支持度
11	AGE >= 1 AGE <= 8.615 PATIENT_SEX IN (男)	29
13	AGE >= 46.692 AGE <= 61.923 PATIENT_SEX IN (男) (女)	951
17	AGE >= 77.154 AGE <= 84.769 PATIENT_SEX IN (男) (女)	336
19	AGE >= 31.462 AGE <= 39.077 PATIENT_SEX IN (男)	267
24	AGE >= 54.308 AGE <= 69.538 PATIENT_SEX IN (男) (女)	464

图 7-10

(3) 不同疾病年龄性别分布比较

通过开展多次疾病年龄性别分布分析，我们还可以进行不同疾病在性别和年龄上发病分布的比较研究。如我们在前面已经分析了冠状动脉粥样硬化性心脏病在年龄、性别上的发病分布，也得到了如图7-10所示的高发年龄分布表，我们用类似的方法还可以研究另外一个心脑血管疾病脑梗死在年龄、性别上发病分布。分别如图7-11和图7-12

所示，脑梗死病人的年龄在61.54与69.23之间，性别不限，共有590个案例，居于所有年龄组的首位，我们可据此理解为脑梗死的高发年龄在61.54与69.23之间的可能性很大。比较冠状动脉粥样硬化性心脏病的高发年龄在46.69与61.92之间，可以看出脑梗死的高发年龄的平均值65.39较冠状动脉粥样硬化性心脏病高发年龄的平均值54.31大出11.08岁，但是冠状动脉粥样硬化性心脏病高发年龄的跨度15.23为脑梗死高发年龄的跨度7.76的将近两倍。

系统操作起始页

此模块为整个数据挖掘程序的一部分。用于疾病的聚类分析。你需要按照步骤操作此程序：“准备数据” -> “清除数据” -> “运行模型” -> “取模型结果” -> “查看结果”。如果开始数据分析，请点击下面的“开始分析”

☒ 请选择疾病

脑梗死-163.902

☒ 入院日期：

(可选)

☒ 出院日期：

(可选)

图 7-11

	20	AGE >= 61.538 AGE <= 69.231 PATIENT_SEX IN (男)(女)	590
	22	AGE >= 53.846 AGE <= 61.538 PATIENT_SEX IN (男)(女)	110
	9	AGE >= 84.615 AGE <= 92.308 PATIENT_SEX IN (男)(女)	270
	13	AGE >= 76.923 AGE <= 84.615 PATIENT_SEX IN (男)(女)	298

图 7-12

通过对两种心脑血管疾病脑梗死和动脉硬化心脏病的比较研究，我们发现，与脑梗死疾病相比，冠状动脉粥样硬化性心脏病患病年龄更小且年龄范围更广泛。年龄维度的划分是一个具体数字，而在实践中更可能是一个范围，一个可信区间。因为数据不是完全覆盖所研究的总体，即使是完全覆盖总体，也需要考虑还有后续病例的情况。所以，分布的探索结果都是建立在一定置信度区间上的，而且随着时间的变化，这个置信度区间也可能会变化，因此需要我们持续地探索、不断地校正并用之于实践。因此，数据挖掘的模型是需要长期维护的。

7.1.2 高维度聚类模型应用

检验聚类分析，是揭示某一疾病在年龄和性别及某一生理指标，三个变量上发病的分布规律。在实践中我们很少使用高维度的聚类，一是可能会出现维度灾难，二是输出的结果难以理解和应用。我们比较容易解释和分析两维度和三维度聚类，但会很难理解更高维度聚类的输出结果。如果我们在分析、解释模型结果时产生困难，又何谈应用数据挖掘的结果呢？下面，我们看一个三个维度的例子，检验聚类分析。

检验聚类分析作为智能分析工具，可挖掘疾病年龄、性别、生理指标（离散型变量，取值为‘H’高、‘N’中、‘L’低）分布，分析疾病与高频生理指标的聚类分布特征，分析疾病与高频生理指标的聚类关系，和比较不同疾病生理指标的分布等。该工具还可用于将挖掘到的聚类组（如年龄和生理指标聚类组）用于科研的区组设计上，从而为科研区组设计建立在一个更科学的方法上，而不再是依靠人为划分。

选择“聚类分析-检验”模块。这里我们只看输出结果，操作流程与前面例子大同小异，不再赘述。我们得到该动脉硬化心脏病按年龄和性别及甘油三酯的分布结果，包括左边的树形分布规则集，中间的规则明细，和右边的“叶子”规则表，如图7-13所示。



图 7-13

如图7-14所示，从中我们可以选取高支持度的规则。如示例中的规则10：年龄在64至73之间，性别为男性，甘油三酯的指标为中、高（即高于正常范围的上限，是一种异常），居于所有规则的首位，我们可据此理解为动脉硬化心脏病病人中甘油三酯偏高的多发于64至73岁的男性病人。

规则序号	规则描述	支持度	置信度
10	PATIENT_SEX IN 男 and TEST_ABNORMAL_FLAG IN N,TEST_ABNORMAL_FLAG IN H and AGE <= 73,AGE >= 64	127	1
13	PATIENT_SEX IN 女,PATIENT_SEX IN 男 and TEST_ABNORMAL_FLAG IN N,TEST_ABNORMAL_FLAG IN H and AGE <= 85,AGE >= 77	65	.97014925373134298
14	PATIENT_SEX IN 女,PATIENT_SEX IN 男 and TEST_ABNORMAL_FLAG IN N,TEST_ABNORMAL_FLAG IN H and AGE <= 51,AGE >= 43	10	.83333333333333304

图 7-14

通过开展多次疾病年龄、性别及生理指标的分布分析，我们还可以进行不同疾病发病分布的比较研究。用类似的方法还可研究脑梗死的高甘油三酯、高发性别和年龄分布。如图7-15所示，年龄在51至60之间，性别男性，甘油三酯为中、高，我们可据此理解为脑梗死病人

中甘油三酯偏高的多发于51至60岁的男性组病人。

18	PATIENT_SEX IN 男 and TEST_ABNORMAL_FLAG IN N,TEST_ABNORMAL_FLAG IN H and AGE <= 60,AGE >= 51	113	.866666666666666703
19	PATIENT_SEX IN 女,PATIENT_SEX IN 男 and TEST_ABNORMAL_FLAG IN N,TEST_ABNORMAL_FLAG IN H and AGE <= 77,AGE >= 68	60	.967741935483871
20	PATIENT_SEX IN 女,PATIENT_SEX IN 男 and TEST_ABNORMAL_FLAG IN N,TEST_ABNORMAL_FLAG IN H and AGE <= 68,AGE >= 60	13	.92857142857142905

图 7-15

我们分析的数据是临床数据，也就是患病人群的数据。通过对这些临床数据的分析，我们发现甘油三酯对两种疾病发病的影响中，脑梗死的高甘油三酯年龄段比动脉硬化心脏病高甘油三酯年龄段要小得多。

7.2 关系探索

关系探索是探索事物与事物间的伴随关系，以及变量与变量之间的影响关系，数据挖掘也拥有很多工具处理这方面的需要，例如关联规则和特性选择就是关系探索的两个重要技术，关联规则用于揭示事物之间发生的伴随关系，而特性选择是探索变量之间的相关关系。本节将介绍关联规则模型的应用和特性选择模型的应用，其中后者可分为连续型特性选择的应用和离散型特性选择的应用。

7.2.1 关联规则的应用

关联规则分析是一种从事物数据集中挖掘频繁项集的工具，旨在揭示事物之间存在的关联关系。在IMRS系统中，关联规则的应用有几种，我们看几个例子。

我们先来看合并症的关联分析。疾病间的关联分析用于分析合并症之间、症状之间或各量化指标之间存在的相关关系和关联规律，为研究疾病的合并症之间的相关关系或症状之间的相关关系，及合并症的病因有很大帮助。

选择“关联分析-疾病”模块，首先选择目标疾病，例如选择“动脉硬化心脏病”。动脉硬化心脏病是一种典型的心血管方面的慢性病，患者通常同时患有许多其他合并症，如高血压、高脂血等症。这里，关联规则模型的目的是试图揭示动脉硬化心脏病人群的诸多合并症之间存在着哪些内在的关联关系，如图7-16所示。



图 7-16

点击“准备分析”按钮，系统输出结果如图7-17所示。

分析目标

研究疾病：冠状动脉粥样硬化性心脏病

分析结果

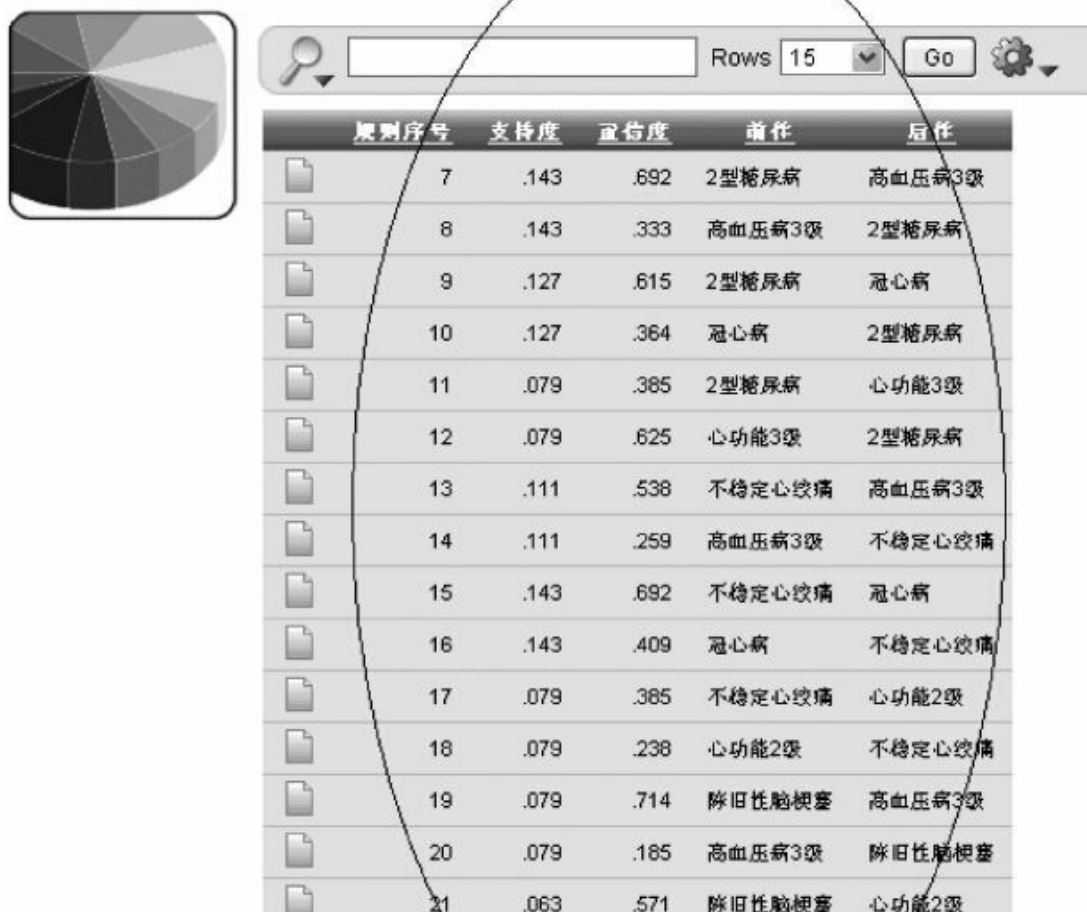


图 7-17

我们得到一个动脉硬化心脏病人群合并症的关联规则集。关联规则模型的读法为：由前件病症“导致”后件病症的情况，这里有几个指标来度量规则，如置信度、支持度、提升（Lift）等。关联规则的语义就是：在前件（事件）发生的情况（条件）下后件（事件）发生的概率（置信度）和比率（支持度）。多数情况下，我们关心的是那些在前件事件发生的情况下后件事件发生的置信度和支持度都较高的规

则。除此之外，还有一个Lift指标用来衡量某条规则是否是存在于大量事件中的、随机发生的、内在的强规则。在这里，我们关心的关联规则模型是那些在前件合并症发生的情况下发生后件合并症的高概率规则，包括有较高支持度的规则和较低支持度的规则，并且是在动脉硬化心脏病人群中发生的大量合并症现象中随机发生的内在的强规则，即Lift指标高的（至少要大于1）规则。

那么，怎样理解“大量事件中的、随机发生的、内在的强规则”呢？举一个例子，比如一个烟酒专卖店，该店以销售香烟和酒类商品为主，同时兼营一些食品等小商品。我们从该店大量的销售数据中发现一个现象，即顾客购买香烟同时购买各种酒的交易事件比较多，反过来，顾客买酒时也买香烟的交易事件也不少，而且发生这两种情况的置信度和支持度都挺高。那么，它们是否是反映大量商品销售中随机产生的有商业价值的规则吗？不是，因为烟酒专卖店的主营商品就是香烟和酒类，该店所有商品的销售由商店的经营方向决定，这两条规则反映的是商品销售事务以外的因素影响的情况，而不是客观的由市场需求决定的销售情况，因此它们不属于有商业价值的强规则，它们的Lift指标通常会小于1（参见第5章）。

现在，我们再来看本案例。我们可以按置信度进行排序，以得到高置信度的规则。例如示例中所示规则50，提示如果动脉硬化心脏病患者患有2型糖尿病和心功能三级合并症的话，很有可能也会患有高血压三级的合并症。

我们还可以在保持高置信度的情况下对支持度进行排序，以得到高置信度和高支持度的规则。例如示例中所示规则7，如果动脉硬化心脏病人患有2型糖尿病的话很有可能也患有高血压三级，规则的置信度为69%，支持度为0.14，如图7-18所示。

规则序号	支持度	置信度	前件	后件
50	.079	1	2型糖尿病,心功能3级	高血压病3级
56	.063	1	陈旧性脑梗塞,心功能2级	高血压病3级
41	.111	.875	冠心病-不稳定心绞痛	心功能2级
52	.095	.857	不稳定心绞痛,高血压病3级	冠心病
60	.095	.857	冠心病,心功能2级	高血压病3级
25	.079	.833	肺炎	高血压病3级
51	.079	.833	高血压病3级,心功能3级	2型糖尿病
23	.063	.8	充血性心衰	冠心病
40	.063	.8	心律失常	冠心病
44	.063	.8	冠心病,	高血压病3级
55	.063	.8	陈旧性脑梗塞,高血压病3级	心功能2级
61	.063	.8	高血压病3级,冠心病-不稳定心绞痛	心功能2级
34	.095	.75	心功能3级	高血压病3级
19	.079	.714	陈旧性脑梗塞	高血压病3级
7	.143	.692	2型糖尿病	高血压病3级

图 7-18

关联规则反映的是处在某个环境下的事物的独特属性，主要体现事物的伴随关系属性，包括它的前件、后件、支持度、置信度和Lift值等描述和度量。在本例中，每种合并疾病都会有独特的前件、后件和不同的支持度、置信度和Lift值。不同的合并疾病前、后件规则会

反映不同的合并疾病伴随关系。例如：

- 前件合并症A “导致”了后件合并症B的置信度非常高，同时前件合并症B “导致”了后件合并症A的置信度也非常高，则可能说明合并症A和合并症B的并发关系非常强，二者可能有着平等的强相关关系。

- 前件合并症A “导致”了后件合并症B的置信度非常高，但是前件合并症B “导致”了后件合并症A的置信度非常低，则可能说明合并症A是合并症B的诱发原因，二者可能有着类似因果的强相关关系。

通过前件事物和后件事物的伴随出现关系，我们可以观察疾病合并症之间的关系类型，并据此推测病因和制订相应的治疗方法。

7.2.2 特性选择的应用

特性选择分析，有时我们也称为因子权重分析，是一种针对某一目标变量考察其他多个变量（因子）对该目标变量产生影响的重要性的技术，以揭示不同变量（因子）对目标变量的相关程度，它是一种用途广泛的数据挖掘工具。在IMRS中，我们称作“因子权重”的分析模块，实际上就是一种特性选择的分析方法。在应用中，因子权重一般分为两类，一类是连续型因子权重分析，另一种是离散型因子权重分析。

7.2.2.1 连续型特性选择的应用

所谓连续型因子权重分析是指目标变量是连续型的，在IMRS中，目标变量通常是连续型的生理指标。“因子权重-检验”模块用于生理指标间影响度的分析，如揭示不同的实验室指标之间的相关关系，探索生理指标之间的互相影响，观察各个因素（因子）对某一个特定指标的影响程度（权重），其影响程度越高，相关性就越高。而且不同于传统的相关性分析，因子权重对于非线性相关的探索有着天然的优势。例如，我们仍然以动脉硬化心脏病的人群作为研究对象。在“因子权重-检验”模块，选择“动脉硬化心脏病”后进入下一页面，从“检验项目”下拉列表中选择“甘油三酯”作为目标变量，然后点击“运行模型”按钮，系统将自动运算出的结果显示出来，如图7-19所示。



图 7-19

我们得到对于动脉硬化心脏病人群的一个生化实验观测目标“甘油三酯”的一组生理指标的权重分析结果，如图7-20所示，包括左边的条形权重系数图和右边的权重系数表。其中右边的权重系数表基于交互型报表技术，它有很多查询功能和图形化工具，使得数据挖掘结果的展现更丰富。

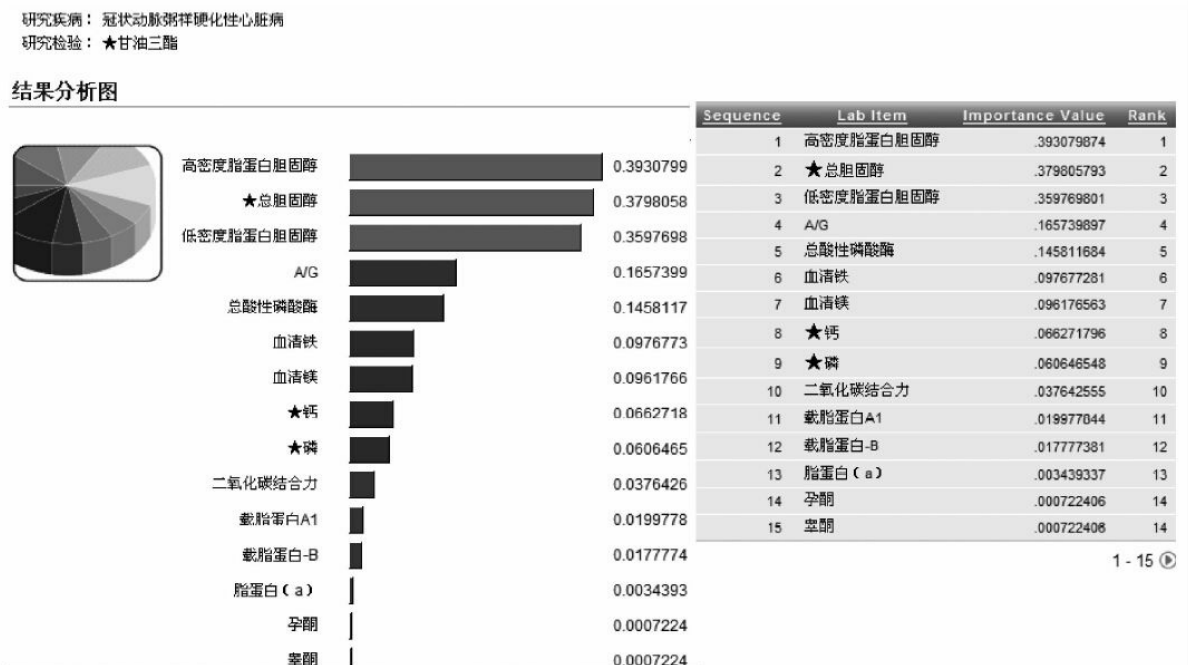


图 7-20

按照示例中的分析结果，我们可以看出：对于动脉硬化心脏病的

观测目标项“甘油三酯”的影响度，从高到低依次为：高密度脂蛋白胆固醇、总胆固醇、低密度脂蛋白胆固醇，其他生理指标对甘油三酯的影响度逐步减小直至接近0。

我们再选择“脑梗死”人群作为研究对象，同样观察“甘油三酯”的相关因子，如图7-21所示。

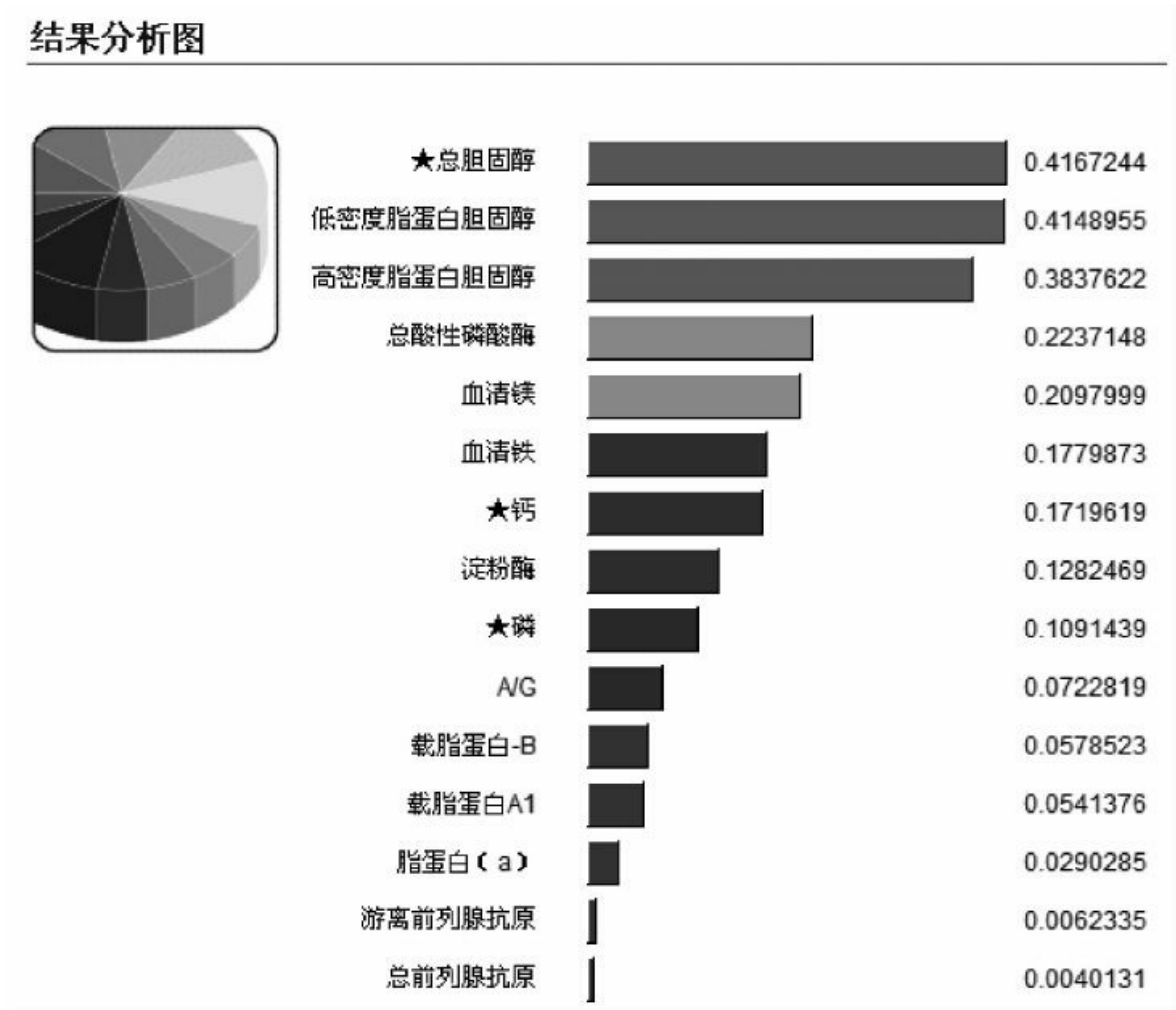


图 7-21

按照示例中的分析结果，我们可以看出：对于脑梗死的观测目标项“甘油三酯”的影响度为：总胆固醇最高，其次是低密度脂蛋白胆固醇

醇，高密度脂蛋白胆固醇，其他生理指标对甘油三酯影响度逐步减小直至接近0。通过以上示例中的分析研究，我们发现，对于动脉硬化心脏病和脑梗死的共同观测目标项“甘油三酯”来说，有着相似的生理指标影响因子，主要为：总胆固醇、低密度脂蛋白胆固醇、高密度脂蛋白胆固醇，这三者为前三位，而且权重都较高，不同的是，对于脑梗死的“甘油三酯”还有一定影响关系的是“淀粉酶”。当然我们还需要应用其他临床实验的方法进一步验证它与脑梗死的关系。

7.2.2.2 离散型特性选择的应用

所谓离散型因子权重分析是指目标变量是离散型的疾病变量。“因子权重-疾病”模块用于分析病人与健康人之间、两种疾病之间或多种疾病之间潜在的生理指标的差异，为研究病人与健康人之间的医学差异、不同疾病之间的医学差异以及生理指标与疾病之间的内在联系提供了很好的挖掘工具。

例1：两种相近疾病生理指标权重分析。

选择两种相近的疾病，往往属于某一大类疾病，应用“因子权重-疾病”探索对这一类疾病影响较大的医学指标因素，同时也可研究这两种相近疾病的典型医学差异指标。当然，我们也可以选择两种以上的疾病作为研究对象，但对于挖掘出的差异性变量指标的解释比较困难，因为得出的差异是指所有三个及三个以上的疾病之间的差异，但我们并不清楚每两个疾病之间的差异，这还需要做所有的两两差异性

挖掘和分析。本例中，选择“因子权重-疾病”模块，选择“动脉硬化心脏病”和“脑梗死”作为分析对象，也就是说目标变量疾病有两个离散取值，一个是“动脉硬化心脏病”，一个是“脑梗死”。如图7-22所示。

使用IMRS的功能进行分析的过程与前例类似。最后我们得到如下结果，如图7-23、图7-24和图7-25所示。

在医学应用方面，使用疾病病种作为特性选择的离散型目标变量时，可以考察不同生理指标对病种的区分能力，也就是说，疾病因子权重分析用于分析两种疾病之间或多种疾病之间存在的生理指标的差异。特性选择是探索对目标变量变化有影响的观察变量（因子）有哪些，影响强度（权重）是怎么样的。对于离散型特性选择来说，目标变量的变化就是取不同的离散值，也就是不同的目标簇之间的变化被哪些因素所影响。反过来说，就是哪些因素影响了簇的划分，不同的目标簇之间存在哪些最明显的差异。

按照示例中的分析结果，我们可以看出：动脉硬化心脏病和脑梗死最大的差异在于生理指标“铁”，其次是“维生素B12”和“网织红细胞”，最后是“叶酸”、“D-二聚体”等其他生理指标。如图7-23所示。

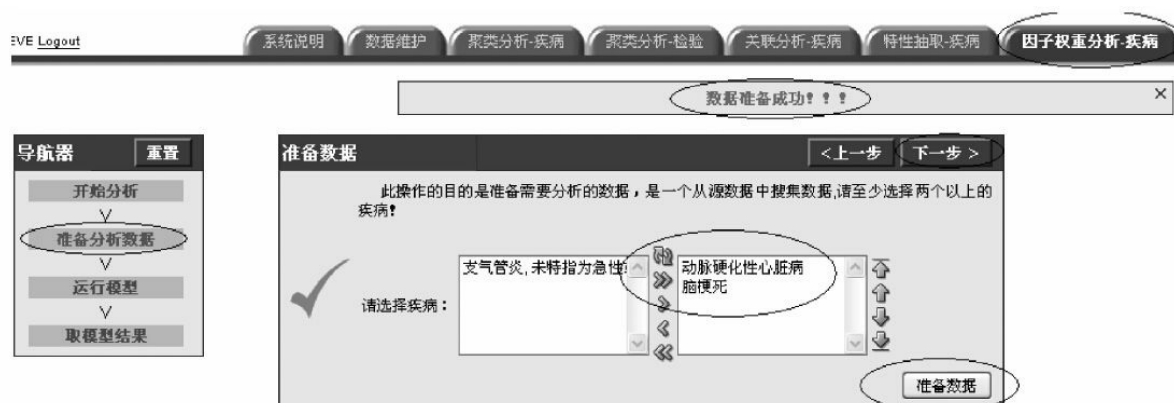


图 7-22

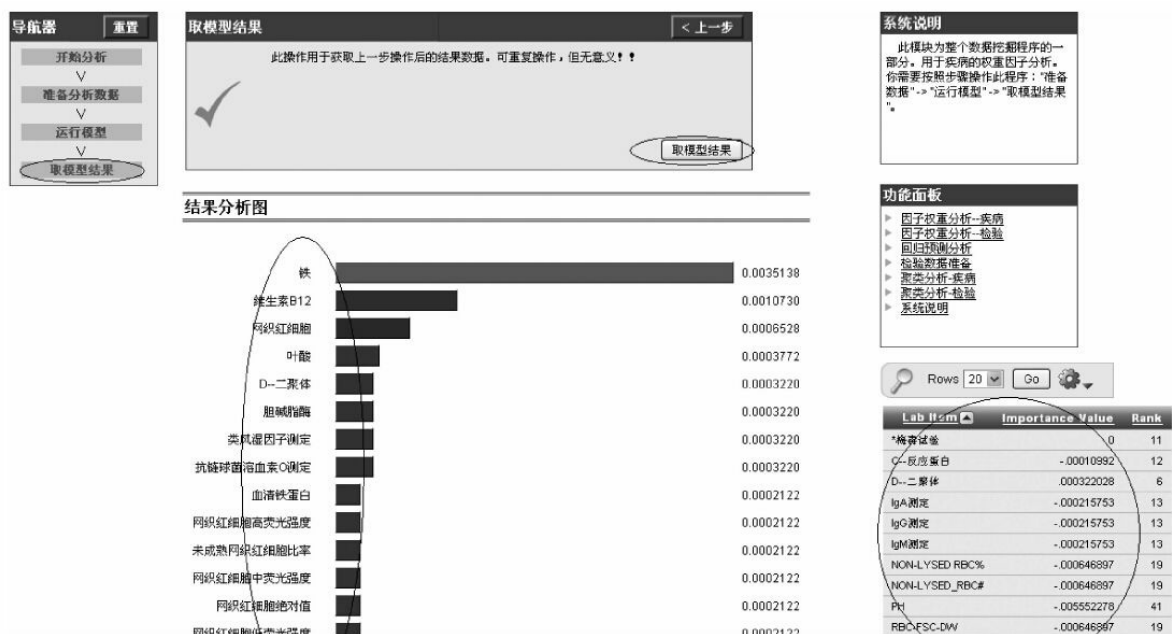


图 7-23

从图中我们可以看出，条形图都是从左至右展开的（疾病变量的相关因子的相关系数 >0 ），而且是从长到短由上至下依次排列的，条形的长短代表了相关因子与疾病变量间相关系数的大小。图中显示的这一组相关系数（影响系数） >0 的因子，我们将它们与疾病的相关关系称作“实相关”，也就是它们与目标变量疾病肯定存在着一定的实际关联，目标变量与相关因子是“互动”的关系。在此处，由于目标变量是二值的离散变量。因此，一方面我们可以认为这些生理指标与这两个疾病有一定的相关性，另一方面，我们也可以将这些生理指标看作是这两个疾病的差异性指标。

我们再看图7-24，这是一组与疾病变量成零相关的生理指标，所有的观察变量（因子）的影响系数都是零，它们对于疾病变量的影响

关系（权重）为零，说明这些因子与疾病变量不相关。

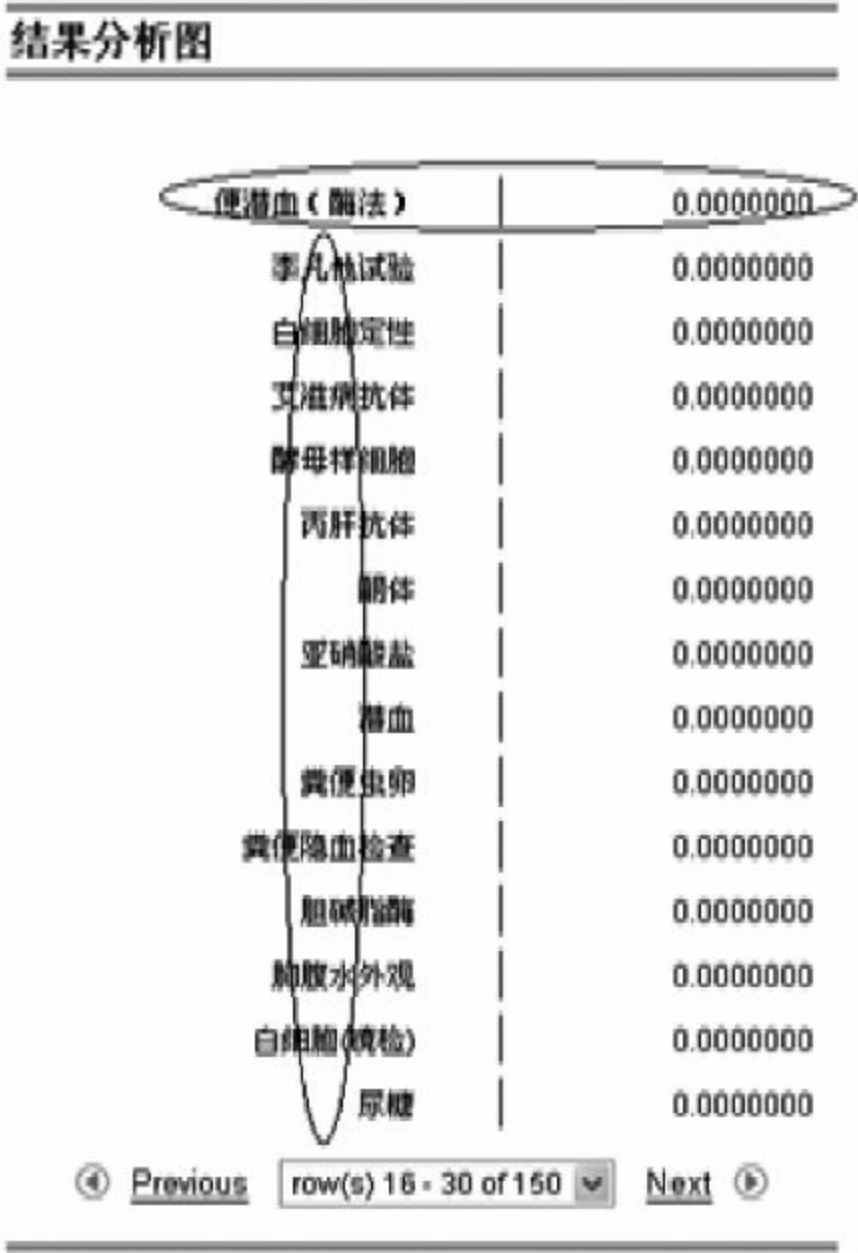


图 7-24

我们再看图7-25，可以看出，条形图都是从右至左展开的（疾病变量的相关因子的相关系数 <0 ），而且是从短到长由上至下依次排列的，条形的长短代表了观察变量（因子）与疾病变量间相关系数的大

小，但方向是相反的。图中显示的这一组相关系数 <0 的因子，我们将它们与疾病的相关关系称作“虚相关”，也就是它们是目标变量疾病的干扰因素，目标变量与观察变量是“互相混淆”的关系。在本例，目标变量是二值的离散变量，包括动脉硬化心脏病和脑梗死。因此，我们不建议用这些指标来判定或区别这两种疾病，否则会造成误诊、误判。

对于数据挖掘得到的结果，我们应当尽可能多地与有关领域的专家进行交流讨论与评估，听取他们的意见和建议，以便在专业领域得到最佳的解释与应用。数据不能完全代替专家，这点读者一定要牢记。

结果分析图

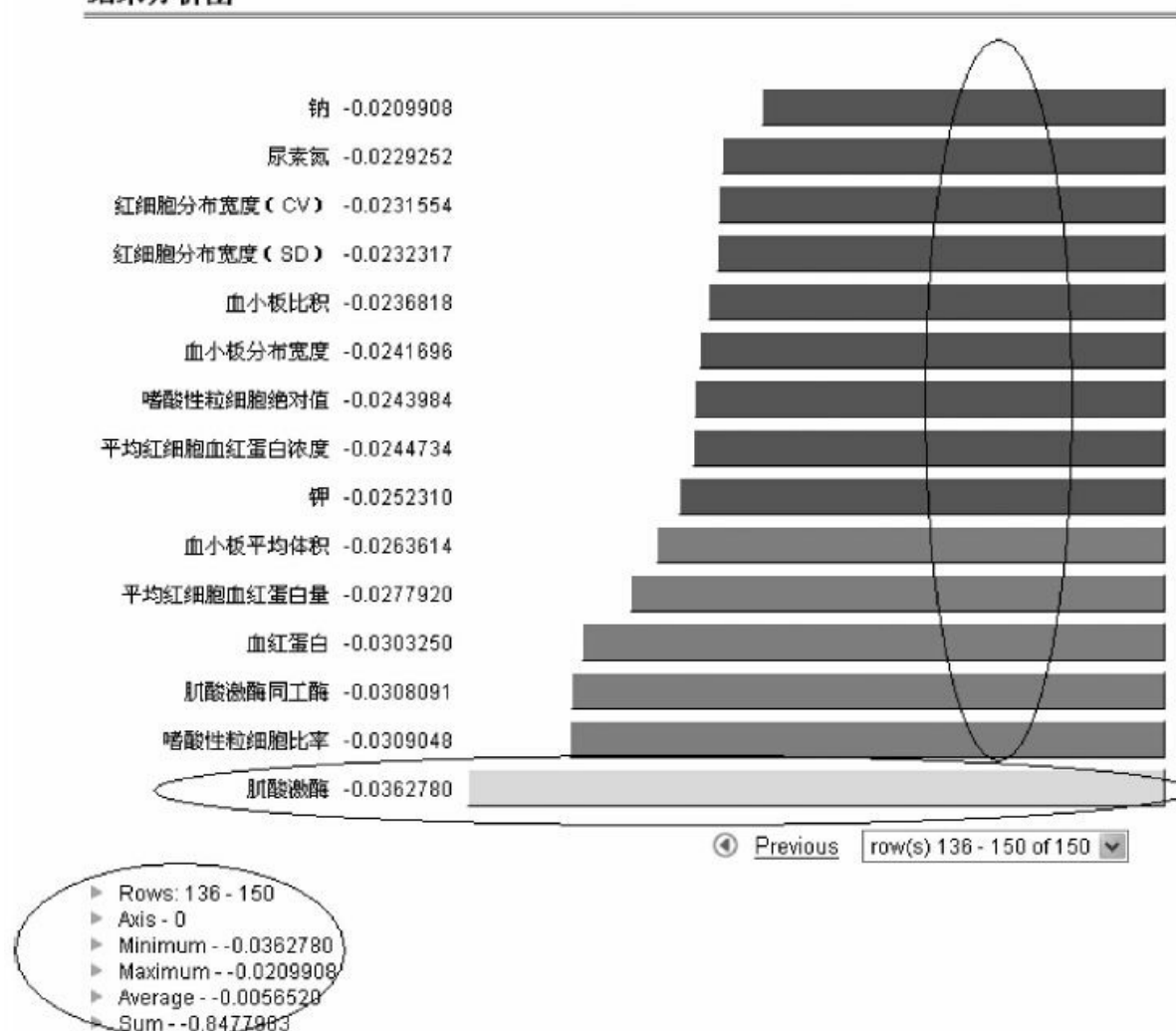


图 7-25

对特性选择的分析，不论是连续型还是离散型，都要首先注意特性选择影响系数的正负，还有注意选择出来的特性（因子）的相关度强弱。我们将特性选择系数为正值的因子与目标变量的关系称为“实相关”，这些“实相关”的因子可能成为目标变量的预测因子或分类因子，“实相关”的因子与它们的目标变量很可能构成一个多元关系模型：一个或多个因子的变化会引起目标变量的变化，这种变化存在一定的规律。如果目标变量是连续变量，这种“实相关”的相关因子与目

标变量可能成线性相关，也可能成非线性相关，可能完全相关，也可能不完全相关。

通过这两个模块的应用，我们可以发现生理指标之间的互相影响关系，不仅是对于病人人群，对健康人群也是这样。我们得到的三组因子（相关系数 >0 ，相关系数 $=0$ ，相关系数 <0 ）在医学专业上，都是有意义的，但不同于我们熟悉的统计学中相关系数为正值的表示线性正相关，相关系数为负值的表示线性负相关。这里的影响系数正值因子是“实相关”，它们对目标变量有积极影响力；影响系数零值因子是“不相关”，它们对目标变量既没有积极影响力也没有消极影响力；影响系数负值因子是“虚相关”，它们对目标变量有消极影响力。

在医学上，建议尽量避免采用含有“虚相关”的因子来诊断或鉴别疾病。不同的疾病之间，“不相关”因子也是不同的。有些在专业经验上看似“不相关”的变量，经数据挖掘后发现是“实相关”的。例如，在一个科研课题的数据挖掘探索中，我们发现刮风与某种心脏病的发病成“实相关”关系。当时很难理解，后来与医学专家深入讨论，专家认为刮风减低了空气中的微粒（可能是PM2.5），微粒可能会诱发心脏病的发病。因此，刮风可能导致减低心脏疾病的发病概率。

7.3 特征探索

作为特征探索的一个重要工具，特征抽取是一种通过降低维度的方法来抽取事物的主要特征的挖掘分析工具，旨在揭示事物内存在的主要特征。特征抽取用来处理关系型数据中的高维度问题，也可以处理从文本数据中提取主题的问题。在IMRS中，我们用特征抽取模型提取医学的临床主题特征。例如，疾病特征抽取用于分析合并症之间，或症状之间，或各量化指标之间存在的组合关系和关联规律，为研究疾病的合并症之间、症状之间相关关系及病因学研究有很大帮助。本节我们将从不稳定心绞痛的特征总结和动脉硬化心脏病的临床特征两个应用上来探讨特征抽取模型的应用。

7.3.1 不稳定心绞痛的特征总结

1) 数据准备

该模块可以按照ICD10的国际疾病分类标准，选取所要分析的对象。在本例中，研究人员选取了“不稳定性心绞痛”作为目标疾病人群。具体的纳入标准为近十年内40岁以上（含40岁）、80岁以下曾经住院治疗的以“不稳定性心绞痛”为主诊断的男性病人和女性病人，如图7-26所示。

系统操作起始页 [返回主页](#)

此模块为整个数据挖掘程序的一部分。用于疾病的聚类分析。你需要按照步骤操作此程序：准备数据 -> 清除数据 -> 运行模型 -> 取模型结果 -> 查看结果。如果开始数据分析，请点击下面的“开始分析”

请选择疾病

不稳定性心绞痛, BRAUNWALDIII级-I20.006
不稳定性心绞痛-I20.001

请选择性别：男女 (可选)

请选择年龄段：40 至 79 (可选)

入院日期：30-11-2004 (可选)

出院日期：30-11-2013 (可选)

[准备分析](#)

图 7-26

经选取该模块的各下拉单式的参数选项后，我们点击“准备分析”按钮，该模块会自动从科研数据仓库中筛选满足以上要求的数据。系统显示本次筛选了2469条数据，如图7-27所示，图中为灰朦状态，表示系统正在运行中，以下同。



图 7-27

2) 模型运行

数据准备完成后，我们需要适当的数据挖掘模型对该数据进行分析。点击该模块“运行模型”按钮，该模块会自动运行事先设计好的数据挖掘引擎，运用适当的算法进行分析，如图7-28所示。



图 7-28

该模块在后台运行一段时间后，系统会显示“模型运行成功”提示。点击“取模型结果”按钮，该模块会将本次筛选的2469条数据的挖掘结果详细地显示出来，如图7-29所示。



图 7-29

3) 模型分析

首先我们会得到一个抽取合并症主题特征的特征集，该模块会按照特征系数的强弱（特性总参数）进行降序排列，如图7-30所示。

特性排序		
特性	特性总参数	特性排序
43	.04866859630368802823298	1
49	.047818385142394319	2
16	.046619010104583001792066	3
7	.044710570935884265556134	4
15	.04320964934476545393	5
5	.042782528610180640030542	6
10	.042197210509050084	7

1 - 7 ➡

图 7-30

特征ID号为43的合并症主题特征排在最前面，依次是特征49、特征16等。

下一步，我们可以在下面的“分析目标”栏里，通过“选择特征”下拉选项来选取某个具体特征的详细结果信息。例如我们选择排在第三位的特征16，如图7-31所示。

分析目标

* 研究疾病： 不稳定性心绞痛，Brawnwald III级;不稳定性心绞痛

* 请选择特征： 16

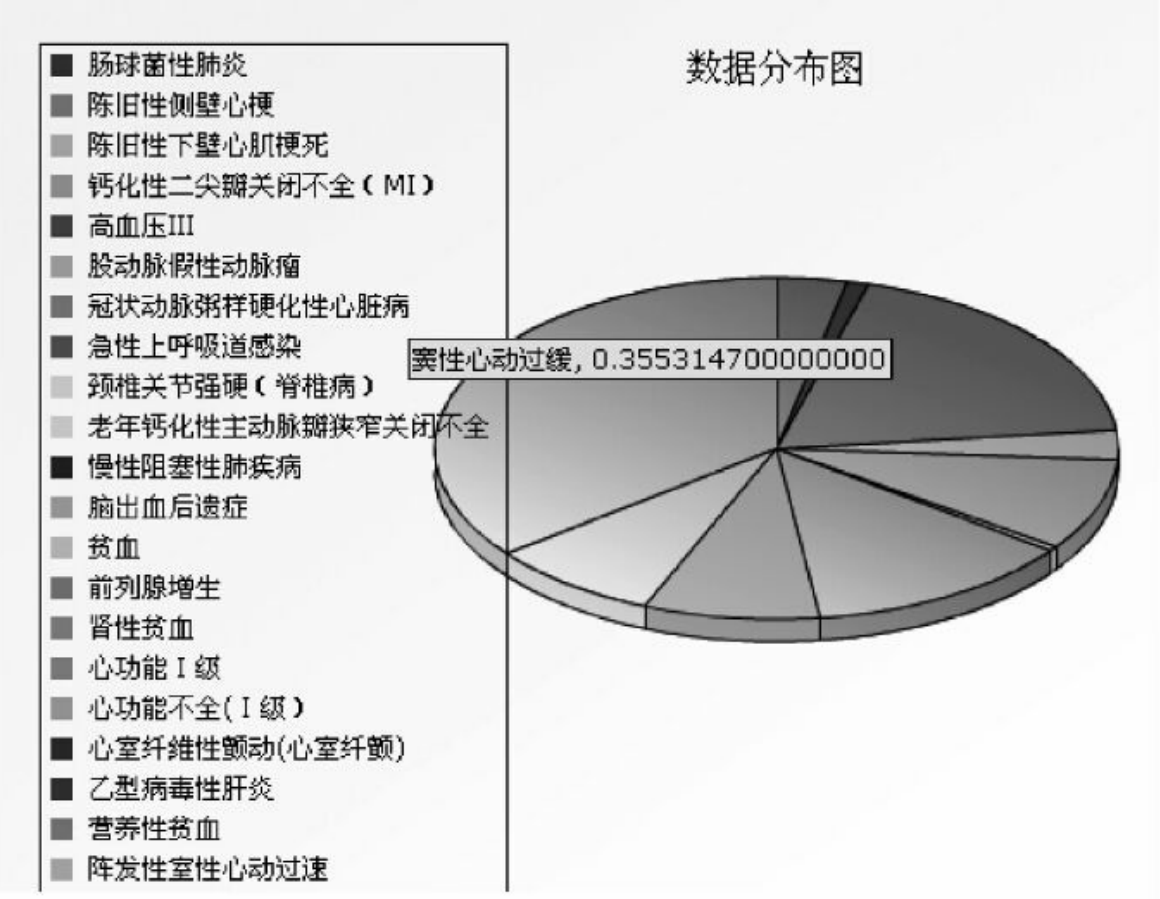


图 7-31

将鼠标滑动到饼图的浅蓝色区域，系统自动显示该区域所代表的合并症名称及在饼图中的相对比重，它表示该区域所代表的分量对整个饼图所代表的新特征的贡献度。

在交互式特征报表中，通过条件筛选，选择特征ID等于16的所有分变量信息，如图7-32所示。

按照分量的特征系数进行降序排列，我们得到了特征ID16的所有合并症组成成分。其中特征系数所表达的是每个组成成分的绝对值，而不是相对比重。我们发现特征系数相对较大的前六个分量是：“窦性心动过缓”、“冠状动脉粥样硬化性心脏病”、“前列腺增生症”、“脑出血后遗症”、“心功能不全”和“阵发性室性心动过速”。

我们再在下方的“分析目标”栏里，通过“选择特征”下拉选项来选取排在第一位的特征43，并将鼠标滑动到饼图的深蓝色区域，系统自动显示该区域所代表的合并症名称“陈旧性前壁心肌梗死”及在饼图中的相对比重0.821，如图7-33所示。

Rows 15

Go

特征 = 16

特征	疾病名称	特征系数
16	窦性心动过缓	.0165644180181709
16	冠状动脉粥样硬化性心脏病	.00882110316276906
16	前列腺增生	.00596432999073177
16	脑出血后遗症	.00404184543602703
16	心功能不全(I级)	.00388748675909512
16	阵发性室性心动过速	.00380151770475423
16	陈旧性侧壁心梗	.00155774585357284
16	颈椎关节强硬(脊椎病)	.00130700136127663
16	高血压III	.00046708496948358

图 7-32

分析目标

* 研究疾病： 不稳定性心绞痛，Braunwald III级;不稳定性心绞痛

* 请选择特征： 43

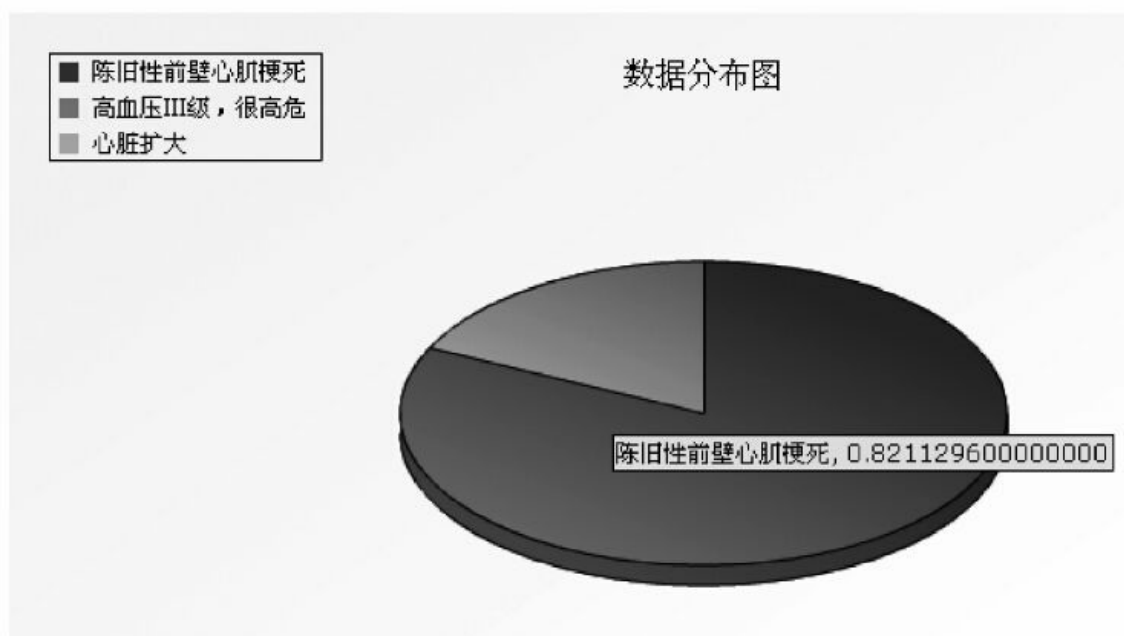


图 7-33

在交互式特征报表中，通过条件筛选，选择特征ID等于43的所有分变量信息，如图7-34所示。

按照分变量的特征系数进行降序排列，我们得到了特征ID43的所有合并症组成成分。我们发现特征系数相对较大的前两个分量是：“陈旧性前壁心肌梗死”和“高血压III级”。

综合分析特征ID16和特征ID43。我们发现如下特点：

- 特征ID16所组成的分量成分比较多且均匀，而且主要是各种不同类型的心脏疾病。“前列腺增生症”显示了该特征的男性成分，及前列腺增生症与心脏疾病具有一定的相关性，“脑出血后遗症”则提

示了心脑血管疾病的相关性。

· 特征ID43所组成的分量成分比较少且不均衡，主要突出表现为“陈旧性前壁心肌梗死”和“高血压III级”的特征，同时说明“陈旧性前壁心肌梗死”和“高血压III级”之间可能具有较强的相关性。



Rows 

 ☒ 

特征

43

陈旧性前壁心肌梗死

.0399632226351536

43

高血压III级，很高危

.00870537258937725

43

心脏扩大

.00000000107915717823298

1 - 3

图 7-34

7.3.2 动脉硬化心脏病的临床特征

我们再看下面一个合并症特征抽取分析的应用例子。在IMRS中，选取“特征抽取-疾病”模块，选取“动脉硬化心脏病”作为研究对象，对患有动脉硬化心脏病的人群发生的大量合并症进行临床特征抽取挖掘，如图7-35所示。



图 7-35

我们得到一组合并症特征集。我们从动脉硬化心脏病的合并症中提取了19个可能的特征组。其中，每一个挖掘后的动脉硬化心脏病的合并症特征，是经过特征抽取算法从高维度的、更具体的合并症变量

中压缩、降维、合并而形成的新的、更概括的合并症变量。如图7-36所示，我们得到了左边的特征分布饼图，和右边特征分布表。选择显示其中的一个特征，如特征2进行分析，我们得到一组合并症包括高血压三级、不稳定心绞痛、急性冠脉综合症等。该特征显示动脉硬化性心脏病的一个合并症组合模型，它反映了动脉硬化性心脏病的某个方面的合并症临床特征，这里主要表现与心脏病有关的疾病。

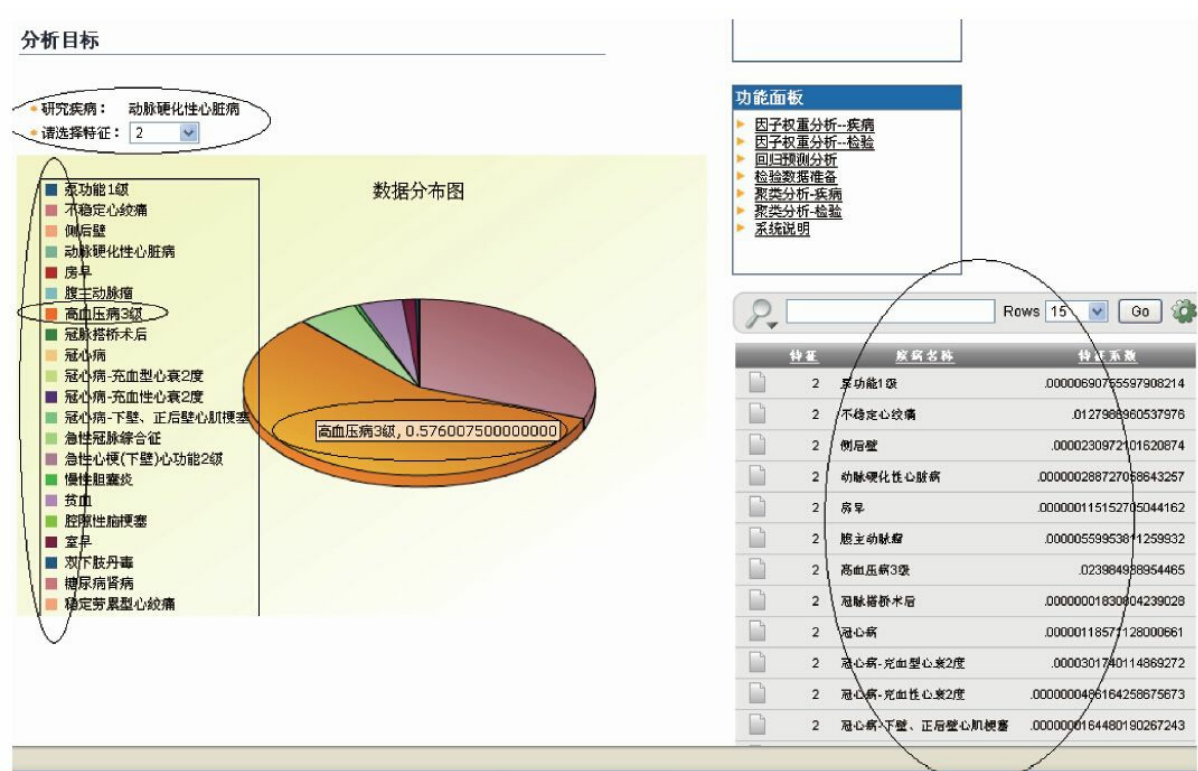


图 7-36

另外，在这个动脉硬化性心脏病的合并症组合模型中，有多个合并症分量元素，它们以不同的权重共同出现在这个合并症组合模型中。该特征提示：高血压三级分量是该特征的主要合并症，不稳定心绞痛分量其次，急性冠脉综合症分量再次，等等。我们用右边的特征分布表对合并症分量进行系数排序，可以得到如图7-37所示的特征规律。

<div> <input type="text"/> Rows 15 Go </div>		
特征	疾病名称	特征系数
2	高血压病3级	.028984938954465
2	不稳定心绞痛	.0127986960537976
2	急性冠脉综合征	.00227717627290806
2	贫血	.00168144263868304
2	室早	.000502763360030964
2	慢性胆囊炎	.000187755832541894
2	双下肢丹毒	.000116870896342599
2	冠心病-充血型心衰2度	.0000301740114869272
2	侧后壁	.0000230972101620874
2	泵功能1级	.00000690755597908214
2	乙状结肠溃疡	.00000662749263472002
2	消化性溃疡	.00000660601663496403
2	腹主动脉瘤	.00000559953811259932
2	腰椎管狭窄	.00000401927133840243
2	纵隔淋巴结肿大	.00000362389571628819

1 - 15

图 7-37

这是动脉硬化心脏病的合并症中的另外一个新合并症特征组合，编号为15，如图7-38所示。

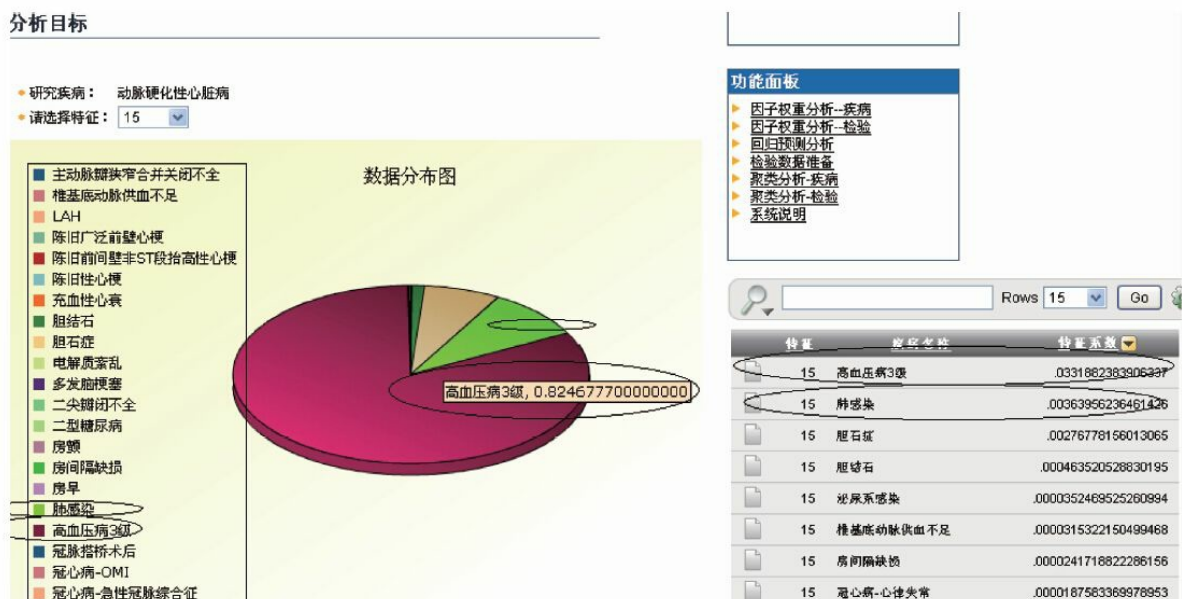


图 7-38

我们可以看出每个合并症新特征都有着自身的特点。从动脉硬化心脏病合并症的角度来总结，该疾病的临床主题特征主要表现在19个方面，每个方面反映了不同的动脉硬化心脏病合并症组合模式。当然我们也可以把合并症换成其他的临床指标，来观察疾病在选定临床指标上所表现的临床主题特征，从而为疾病的内部分型提供了依据和分析思路。

7.4 异常探索

异常探索作为数据挖掘技术的一个重要分支，在商业应用中发挥着独特的作用。例如，在医学上，多年的临床素养和天赋往往赋予医学专家们一种直觉，看到一组数据和症状本就能感觉异常。实际上，我们的异常侦测模型（异常探索类的一个常用工具）正是辅助医学专家增强这种直觉的一个工具。异常侦测模型可以作为临床分析工具单独使用，也可作为体检数据的分析工具和科研创新的探索工具等。我们现在讨论一下IMRS系统应用中两个异常侦测模型的应用：生理指标的异常侦测和异常侦测模型的比较。

7.4.1 生理指标的异常侦测

我们首先看一个高维度的生理指标分析的例子。在IMRS中，选择“异常侦测分析”模块，操作步骤同前章。依然选择“动脉硬化心脏病”作为研究对象，我们也把动脉硬化心脏病人群简称为一个观察群组，如图7-39所示。

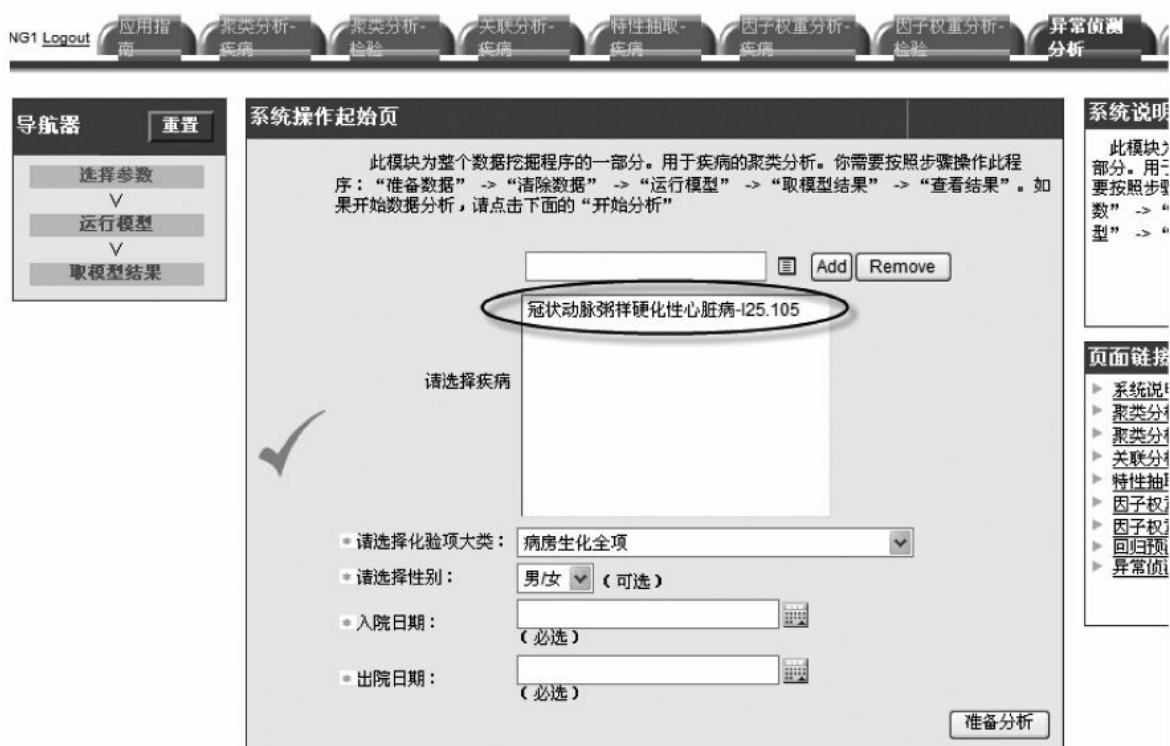


图 7-39

点击“取模型结果”按钮，系统自动从成功运行的模型中取出结果，然后进入下一界面，如图7-40所示。我们得到该种疾病的一组特定生理指标（多维变量）异常侦测的分析结果，包括图左边的侦测出的异常案例概率图，和图右边的侦测出的异常案例列表，它包含了每个案例所有特定生理指标（多维变量）的值。

分析这个输出结果，我们可以看到，如异常案例概率图所示，侦测异常的“树干”左右两侧的“树枝”数目比例一般足够小（比如1：50）。我们将“树干”右侧的居大多数的“树枝”视为观察群组里面一组特定生理指标下的正常点，将“树干”左侧的居少数的“树枝”视为观察群组里面一组特定生理指标下的异常点。这样，我们可以发现一个观察群组里面哪些是异常点或离群点。在医学专业上，少量的异常“树枝”可能意味着：

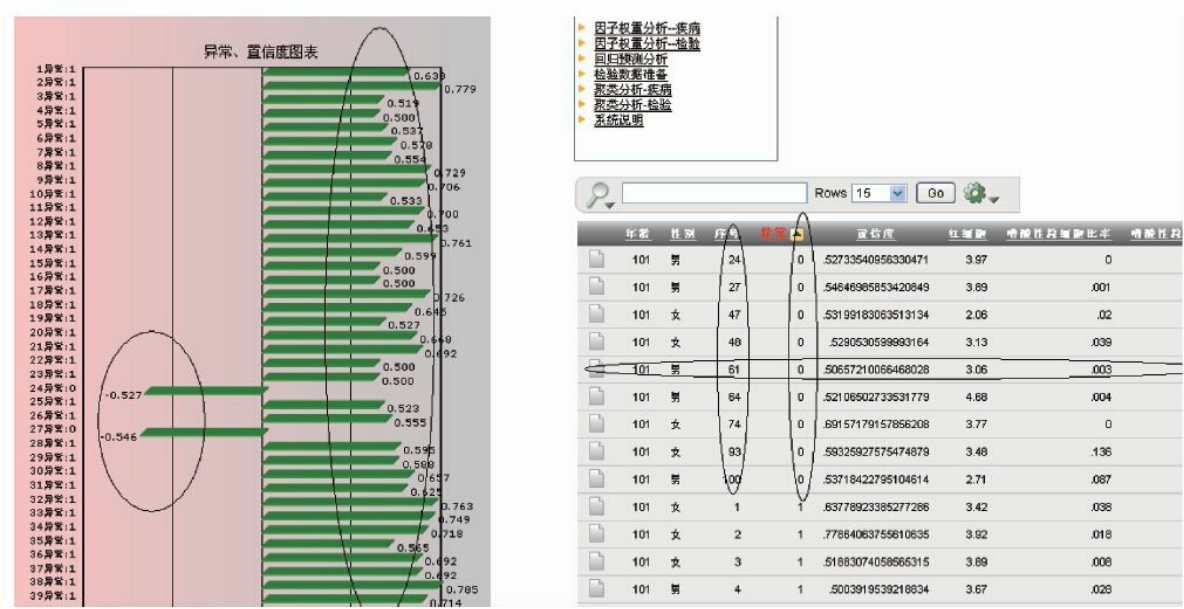


图 7-40

- 罕见“正常”类型，即符合医学逻辑（现代医学理论和知识的）的稀有类型。
- 罕见“异常”类型，即不符合医学逻辑（现代医学理论和知识的）的类型，新发现。
- 误诊，由于设备或人为的因素，错误地测量了或记录了数据。

作为异常侦测工具，还可以侦测正常人的体检数据，来区别出那些看上去每个化验数据个体都正常，而整体组合却不正常的人，进而提醒体检人注意，或做进一步检查。

当异常侦测算法不是基于设定离群率的时候，我们还可以应用异常侦测对高维生理指标组合进行分型分析。例如“树干”左右两侧的“树枝”数目比例足够大（比如1：4）时，与前面的情况有所不同，这可能反映一个大的疾病群组里面包含几种疾病亚型，或者是几个疾病群组混杂在一起。作为分型工具，在异常侦测过程中就验证了哪种化验或化验组合（比如血糖和免疫全项是否可以组合）可以作为最佳分型鉴别试验。如果“树枝”左右分布比例不大（比如1：5），而且置信度不高，也可能意味着选定的标准不能很好地完成异常侦测。如果分型的同时可以把不同类型的病人挑出来，研究诸如对不同类型疾病的病人病史、疗效等，则可能使疾病的分型和治疗更加准确、高效。

分型时还需要注意一点，就是异常侦测的分型与聚类的分型的不同，异常侦测划分的两组之中有一个组的同质性升高，而另外一组的同质性未必升高，如图7-41所示。

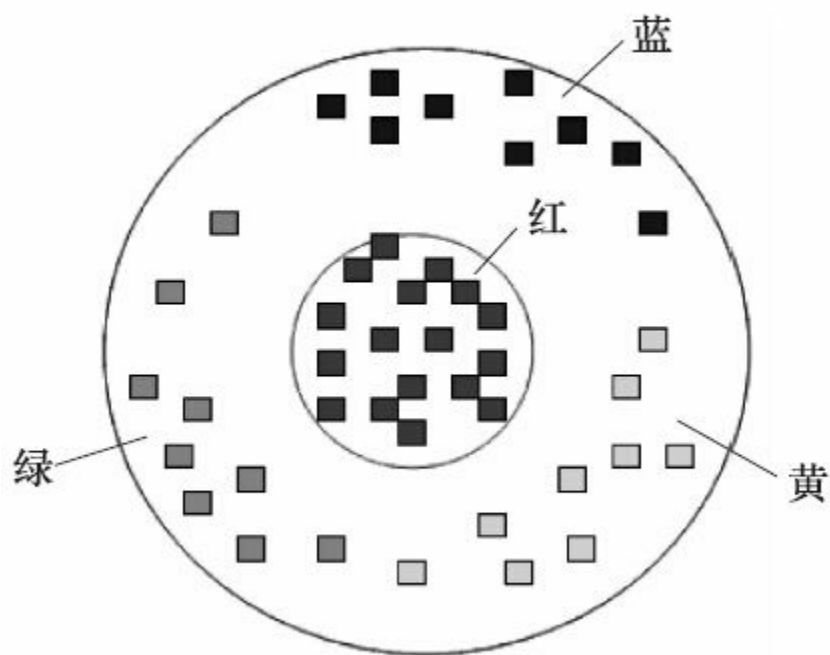


图 7-41

异常侦测是一个特殊的二分类引擎，划分情况可能如图所示。中心红色圈中的组同质性提高，可以作为一个分型使用；而中心红色圈外的组同质性实际上是下降了，是否可以直接将它们作为一个分型使用，还需要结合专业情况和使用目的。这点与聚类模型是不同的，聚类的每个簇的同质性均有提高，如图7-42所示。

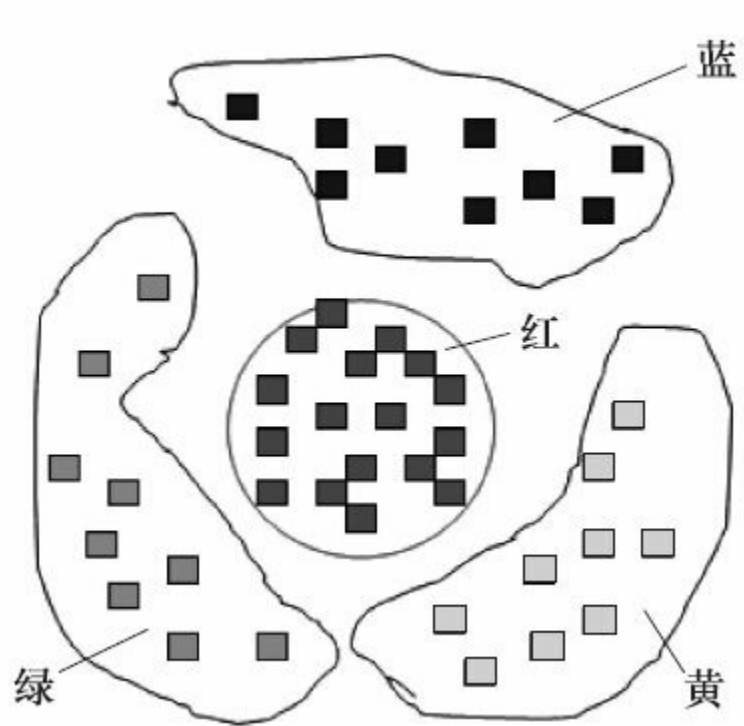


图 7-42

另外，我们还要关注置信度的问题。本例中，我们可以观察到有部分“树枝”的置信度不是很高，甚至为接近50%，它表示有接近50%的可能性是正常，也有接近50%的可能性是异常。这部分数据，我们称之为临界点数据。在医学研究中，异常侦测工具对于多元高维度资料的分析有着传统方法不可比拟的优势。

7.4.2 异常侦测模型的比较

选择脑梗死患病人群为研究对象，异常侦测分析的操作过程与上类似。我们同样得到相对于该种疾病的一组特定生理指标的异常侦测分析结果，包括图左边的侦测出的异常案例概率图，和图右边的侦测出的异常案例列表。但是与动脉硬化心脏病人群的数据相比较，脑梗死人群数据的异常侦测模型得到的数据分布结果大为不同。动脉硬化心脏病的血常规分布基本为一类，有少数几个离群点；但是，脑梗死的血常规分布可以基本视为两类，比率可能为3：7或2：8，一个局部异常案例概率图如图7-43所示，显示局部异常点多于正常点。

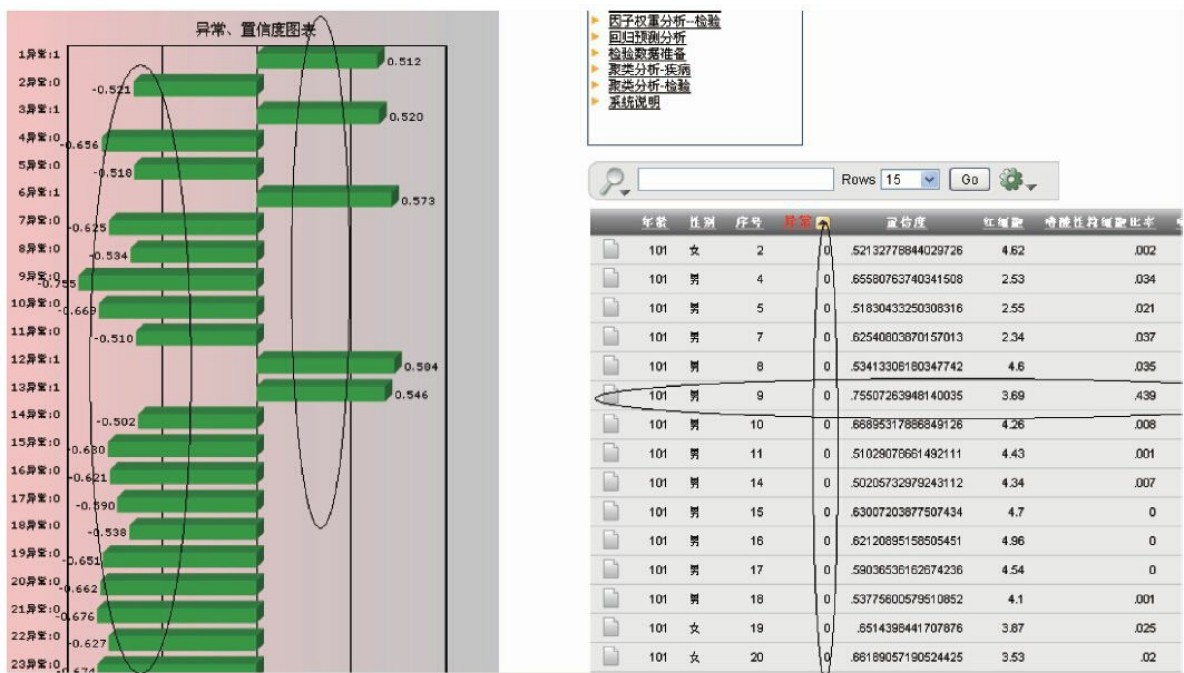


图 7-43

从以上分析中，我们可以得到如下临床指导与参考。我们对这两个病人群组数据应用的是参数设置完全相同的同一个异常侦测算法引

擎。这个结果提示，对于脑梗死病人的血液成分而言，至少存在两种不同的类别，这两种以上不同的类别是否完全可以作为亚型，尚需要进一步研究，但在对该病进行科研设计时，需要考虑在血液成分方面进行分层；而动脉硬化心脏病在血液成分方面的同质性较高，在研究这种疾病时，血液成分方面可以不考虑分层。

7.5 推测探索

数据挖掘的算法中包含着一类推测方法，即由若干个预测因子来推测一个目标变量的取值，我们称为推测探索。按照被推测的目标变量的类型，又可以分为两种方法，一种是推测离散类型目标变量的方法，我们称为分类模型，另一种是推测连续类型目标变量的方法，我们称为预测模型。本节将介绍预测模型，也称为回归模型或回归预测模型。

回归预测模型是应用多个自变量拟合、预测一个连续型的应变量的方法。在医学中，回归预测模型主要有两个方面的应用：一是通过建立和验证一个回归预测模型来证实所发现的一个医学规律，往往用于科研；另一个是通过应用一个已经得到验证的回归预测模型来推测一个临床中非常需要又不易直接得到的医学指标值（回归预测模型的应变量），这时多用于临床。下面，我们就观察一个在IMRS中的回归预测模型应用。

生理指标预测分析

我们使用“回归预测分析”模块为例子来说明。首先选择需要分析的目标人群，我们以“体检人群”为例子。然后在“选择检验项目”菜单选择一个生理指标作为预测的目标变量，我们以“血钠”为例子。系统自动准备数据后，显示所准备的数据总量，本例为11438条数据。然

后点击“因子分析”按钮，系统进入因子权重分析阶段，此时界面变灰，如图7-44所示。



图 7-44

系统会得到权重值较高的前n个生理指标作为目标变量“血钠”的待选相关因子，按照相关系数的高低降序排列，并列出所有相关因子的排名。本例n为79，如图7-45所示。

因子权重排名

SEQUENCE	ATTRIBUTE_NAME	EXPLANATORY_VALUE	RANK
1	★氯	.12961784	1
2	★钙	.12600387	2
3	血清铁	.1225183	3
4	★磷酸肌酸激酶	.120604867	4
5	血清镁	.120500164	5
6	★钾	.120212884	6
7	★磷	.119374627	7
8	白蛋白	.118242668	8
9	总蛋白	.118050886	9
10	A/G	.117933479	10
11	球蛋白	.117277997	11
12	总胆红素	.115480337	12
13	间接胆红素	.11380677	13
14	直接胆红素	.111874513	14
15	促甲状腺素	.034550411	15



row(s) 1 - 15 of 79  

图 7-45

预测准确度

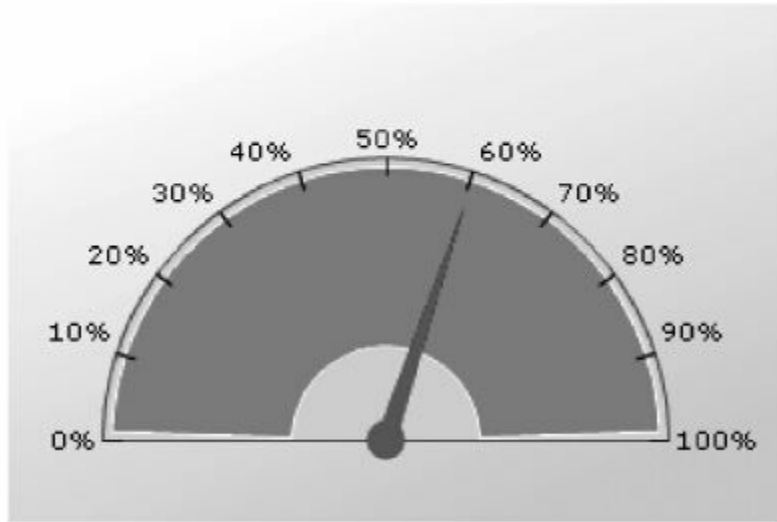


图 7-46

从相关因子排名靠前的生理指标中，根据专业意义和需求任意选择几个生理指标作为因变量“血钠”的自变量，也可以说是作为目标预测变量“血钠”的预测因子，包括：氯、钙、血清铁、磷酸肌酸激酶、血清镁、钾、磷等七个指标。

预测因子选择后，点击“回归预测”按钮，系统自动运行配置好的数据挖掘引擎算法，自动探索由预测变量“血钠”和选好的预测因子所组成的一个多元线性或非线性模型，并对这个模型进行金标准验证。

模型运行成功后进入下一界面，如图7-46所示。此时回归预测分析的结果已经存储在数据库中了。模型运行结束后，系统自动从数据库中取出模型结果。该模型显示预测置信度为60.3%，也就是说，它比朴素预测模型的准确率高出60.3%。如该模型的仪表盘指示图显

示，说明模型性能较好。

我们再看预测模型的残差散点图。图7-47显示的是预测残差示意图，Y坐标标值为0的横线代表的是实际的“血钠”值，零线上线下的散点是预测模型的预测值，我们看到绝大部分点都集中在零线上线下的附近的区域内（图中的数据只为示意作用），模型拟合较好。

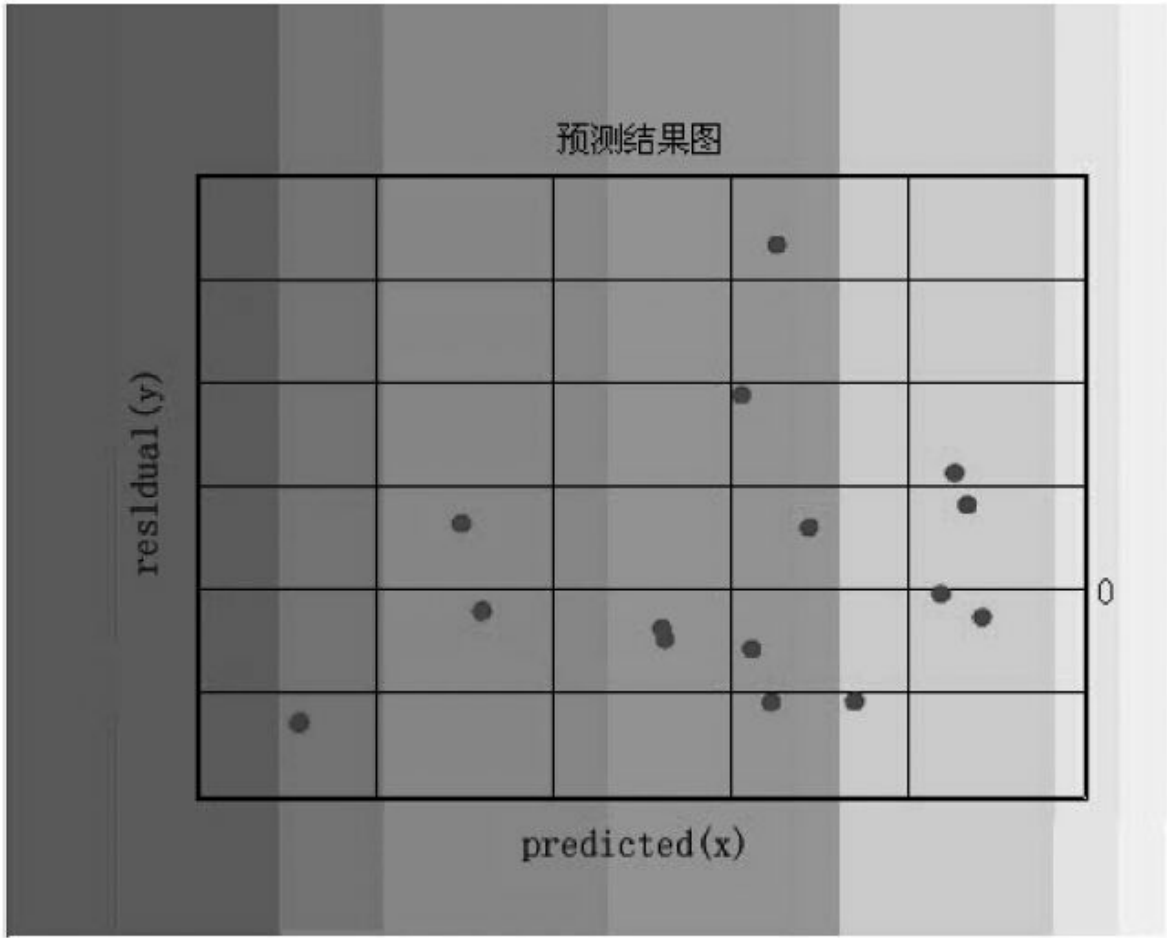


图 7-47

从上述应用过程中，我们得到了氯、钙、血清铁、磷酸肌酸激酶、血清镁、钾、磷等七个指标与“血钠”组成一个多元的回归关系模型，基于金标准的验证证实了该模型的拟合精度较好，初步证明了该

模型的存在。但一方面模型的预测准确度不是非常高，因此可能尚不能应用于医学临床实践；另一方面，由于基于SVM的模型算法虽然可以比较好地开展回归或分类分析，但并不能进一步显式地揭示模型规律，所以即使我们可以通过算法得到一个精度很好的模型，但若想得到可表达的显式模型，还需要其他的数学工具进一步分析。

通过以上应用，IMRS为医学家提供一个面向医学专业的全自动数据挖掘工具。医学家不必深入了解数据挖掘的理论和技術，就可以方便、快捷地应用数据挖掘工具开展他们感兴趣的专业研究，并使得他们把更多的精力集中于医学专业问题。

7.6 应用系统的高级应用

前面的章节介绍了数据挖掘在IMRS系统中五个方向的应用。本节将就其中的两个方向进行进一步深入探讨，一个方向是异常侦测模型的高级应用，另一个方向是关联规则的高级应用。其中在异常侦测模型的高级应用中，我们将讨论健康与非健康的应用和同质性与异质性的应用两个内容，而关联规则的高级应用将涉及四个内容：对不同并发模式的兴趣、负模式规则和零模式规则、关联规则的方向性，以及关联规则的时间性。

7.6.1 异常侦测的高级用法

异常侦测是一种处理高维度异常数据的算法，它是两分法的实践。对于高维度数据的分类，避免维度灾难的方法之一是两分法，根据维度把数据划分成两个部分，‘正常’和‘异常’，有时候叫A组和B组也许更合适一些。实践中，比如在医学中，两分（也就是二元变量）的情况很多见，如人群的健康和非健康、两种疾病的鉴别等。

7.6.1.1 健康与非健康的应用

健康和非健康的研究始终是临床医学科研的一个主题。首先看一下某个体检人群的异常侦测模型结果，如图7-48所示。

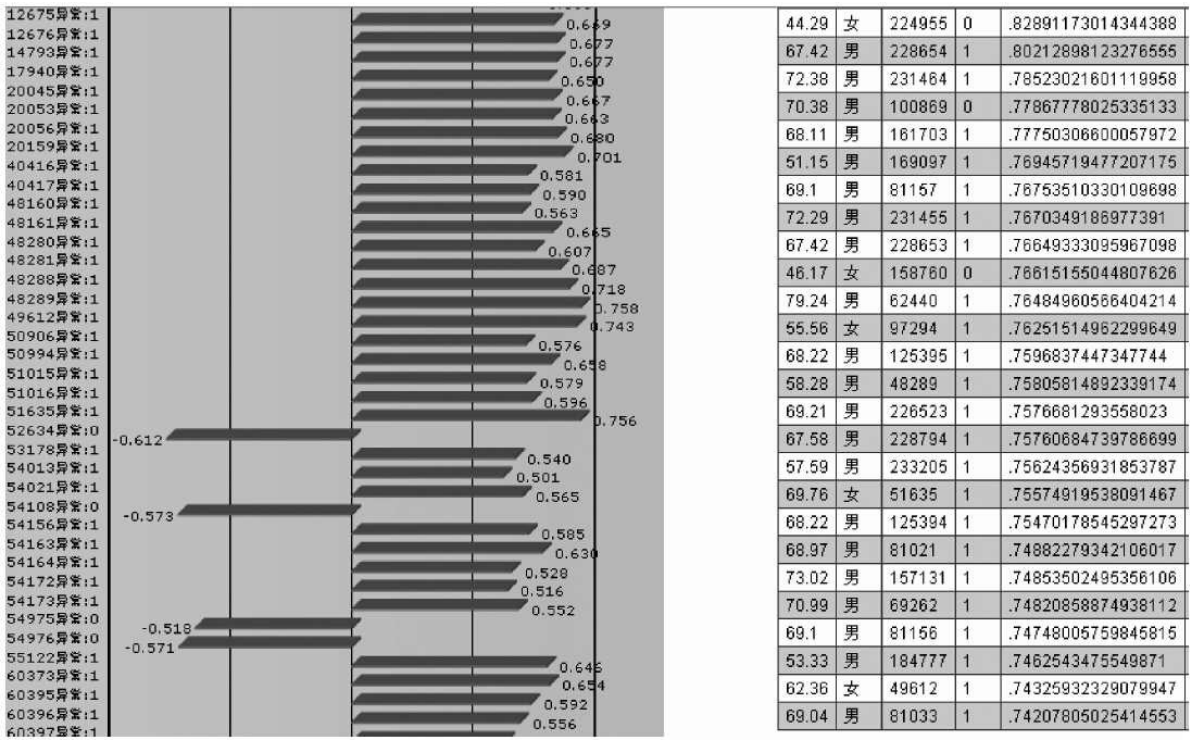


图 7-48

异常侦测模型是一个二分类的工具，在医学上可以充分利用这个二分类特点。健康人群和非健康人群（可能是亚健康或疾病）是两个人群的分类，而且还是互斥的分类，如图7-49所示。

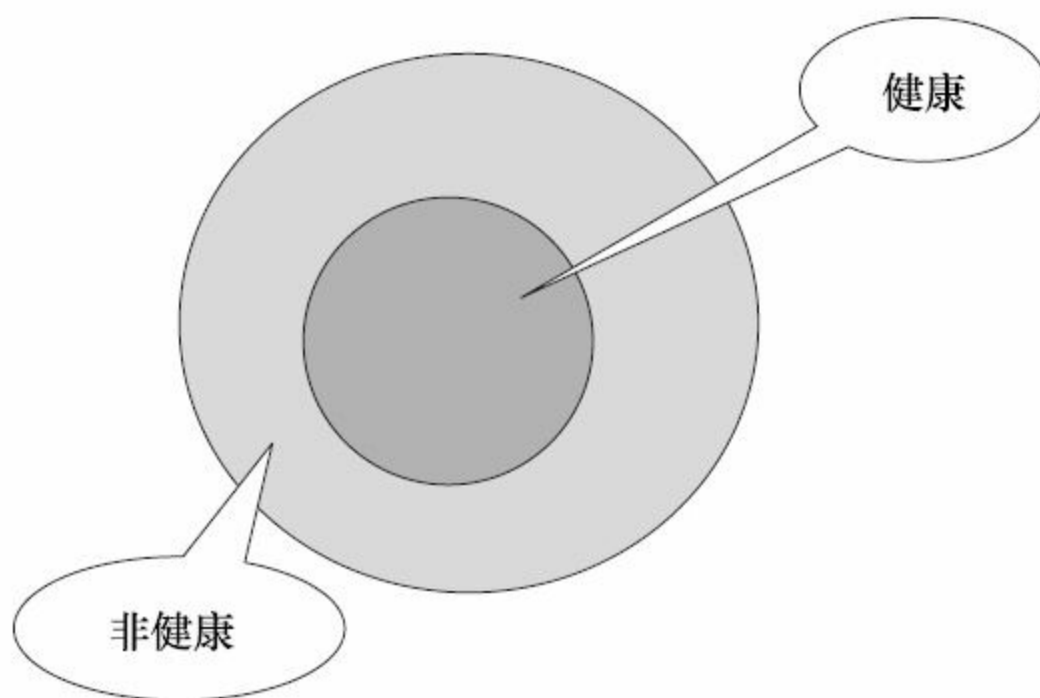


图7-49 人群中的健康和非健康

医学上的各种非健康，从数据角度看，都是一个或几个维度偏离了一个特定的范围。比如血压、血脂、血糖，及血常规的各项指标等，都是具有一个特定的范围，偏离了这个范围就算是非健康。如图7-50所示，血钾高了或低了都算是异常。

如图7-51所示，非健康人群作为一个群体，以各种维度为特征被划分为各种疾病，这样便于医疗活动的进行。医学数据的特点是，全体健康人群的各个维度数据具有同质性，而全体非健康人群的数据在各个维度上具有异质性增加的趋势。这种情况很像一句诗所说

的，“欢乐只有一种，而悲伤各不相同”。同样，也可以说健康只有一种，而非健康各不相同。

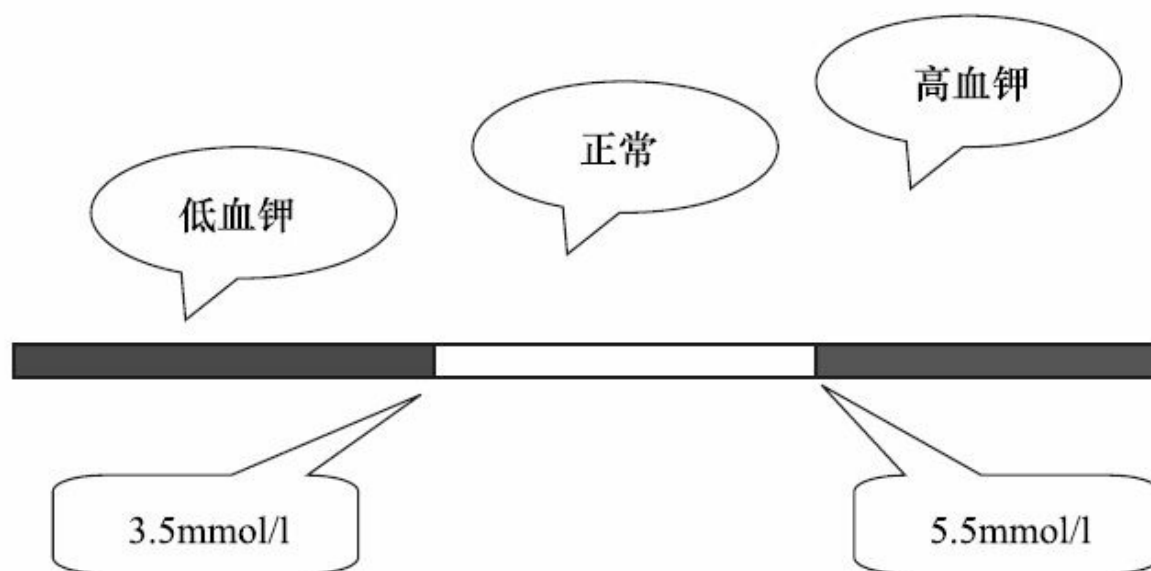


图 7-50

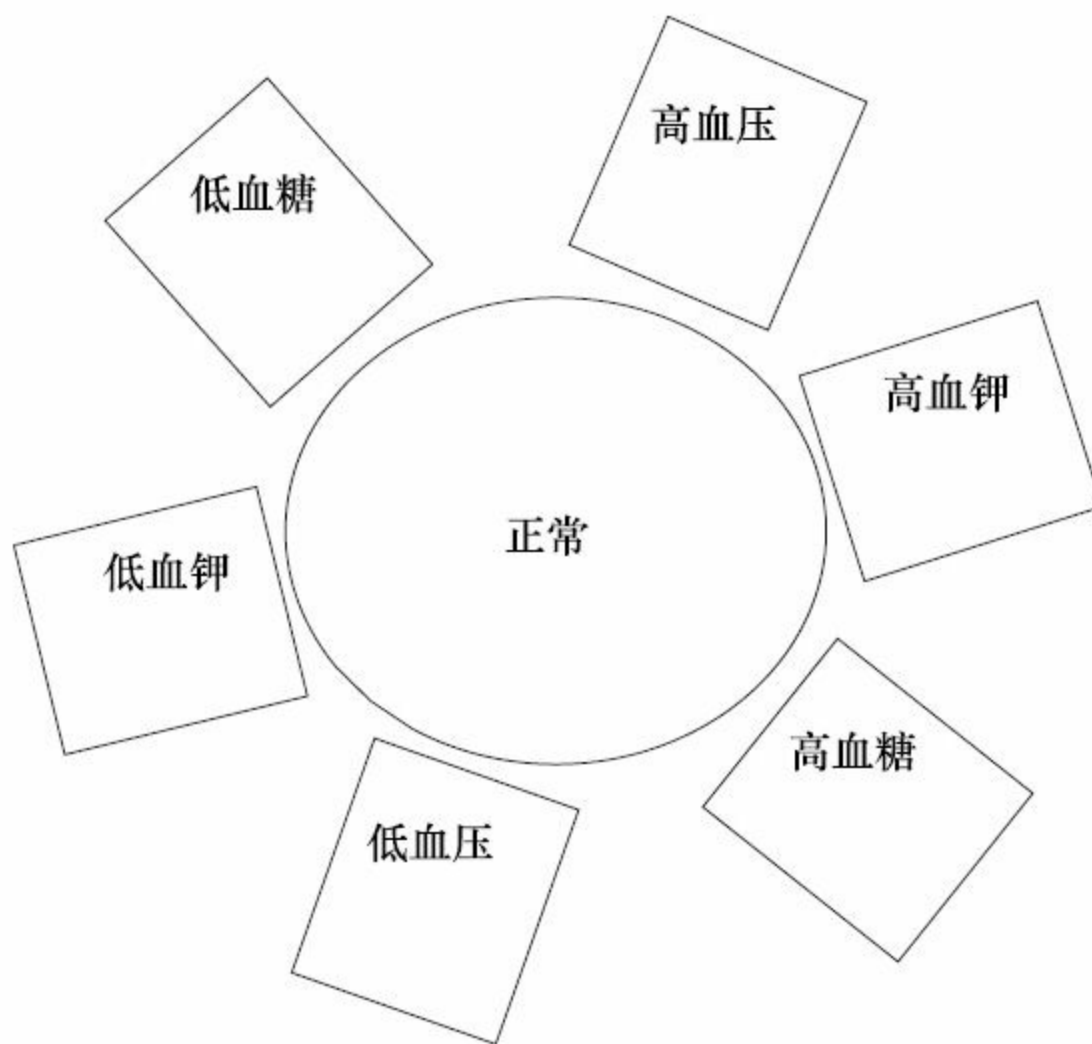


图 7-51

医学数据具有高维度的特点和二分类的需要。比如体检数据，各种检查项目几十个、上百个，维度很高，需要检出健康和非健康这两个分类。我们利用这个特点，对体检数据应用异常侦测引擎的处理。每使用异常侦测模型处理一次，均质性都有提高。这样可以把人群从疾病、亚临床.....亚健康、健康一层一层分开，而不是直接划分健康和疾病，如图7-52所示。

健康和亚健康的分级可以用于很多方面，比如体检，比如与健康

相关的医疗保险等。这样按照不同应用目的，根据不同的参数标准将健康人群和非健康人群一层一层地分开，从同质性最高的健康人群到各种各样的非健康人群，中间层层过渡，也是对健康程度的分级。将人群划分为健康和疾病两组，再将人群划分为健康、亚健康.....亚临床、疾病多个层面，更贴近人群健康水平的自然分布。

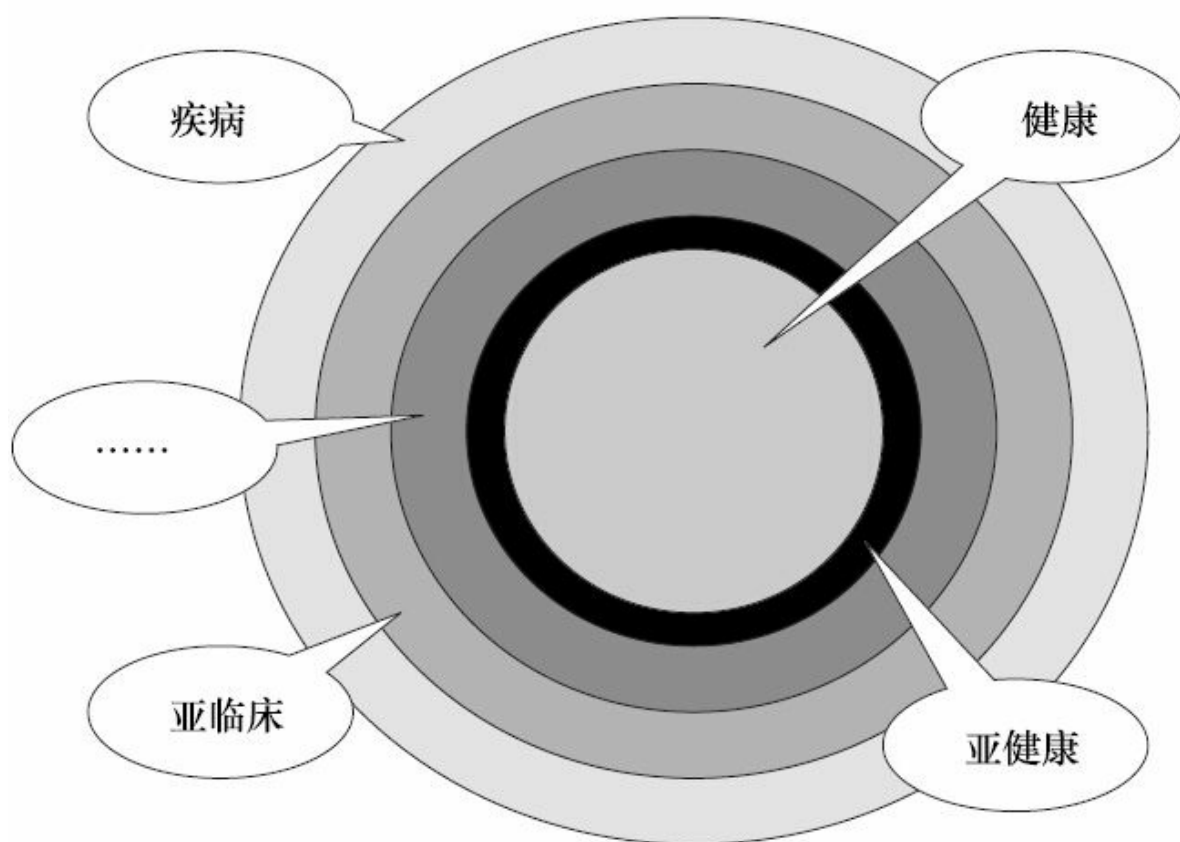


图 7-52

7.6.1.2 同质性和异质性的应用

1.构建同质性组

同样是发挥异常侦测模型的特点，异常侦测的两分特点在数据上会提高至少一个组的同质性（这点与聚类不同，聚类的最后结果是每个簇的同质性均会提高）。这个特点可以用来构建一个同质性高的组。例如在医学科研上（也包括其他领域科研），对于满足均衡性原则很有好处。所谓均衡性就是要求实验组（试验组）与对照组之间或各个实验组（试验组）之间，除了处理因素外其他一切条件尽可能相同。因为实验结果不仅取决于处理因素，也会受到非处理因素的影响。

图7-53显示了均衡性好的部分（两个红圈部分）。

64.45	女	4922	0	.5343965120269393	14.470000000000001	57	308	11	28	5.0800000000000004	1.05	4.
64.45	女	4923	1	.50032368380384074	5.9199999999999999	64	349	39	53	4.1500000000000004	1.79	6.
81.08	女	6282	0	.74124318047925275	24.649999999999999	228	678	36	36	3.5699999999999998	4.4299999999999997	22.
71.82	女	6546	1	.65617919440263492	5.9699999999999998	80	140	8	13	3.7400000000000002	.7299999999999998	4.
69.81	女	6899	1	.643715693309582	6.2199999999999998	78	299	11	15	4.3899999999999997	2.3100000000000001	5.
76.11	女	9053	1	.71703793118919124	8.2400000000000002	102	286	9	21	4.9100000000000001	2.4900000000000002	6.
77.15	女	9375	1	.73998917700782252	6.9900000000000002	102	376	10	22	4.0899999999999999	1.45	7.
87.73	男	21554	1	.75836364975327675	6.0700000000000003	111	464	34	26	3.9399999999999999	.9300000000000005	11.
84.5	男	21768	1	.56311242157075858	6.9500000000000002	72	344	14	16	6.29	.9899999999999999	4.
86.13	女	21947	1	.59036952032531628	5.2999999999999998	121	375	40	31	3.5800000000000004	.75	9.
75.04	女	23032	1	.73941510600901095	4.9299999999999997	164	352	8	15	4.6200000000000001	1.47	9.

图 7-53

在需要考虑的非实验因素比较多时，使用异常侦测模型比使用聚类模型更加适合，而聚类模型更适合在非实验因素较少的情况下分层使用。

异常侦测引擎对数据的作用是把离群点不断地划分出来。我们利用这个特点，把数据输入异常侦测引擎后，从获得的输出结果数据中选取其中的一个部分，然后把这部分继续输入异常侦测引擎，以此不断循环。异常侦测引擎会不断地划分出离群点，通常情况下剩下的数

数据集在均质性上会越来越提高，如图7-54至图7-56所示。

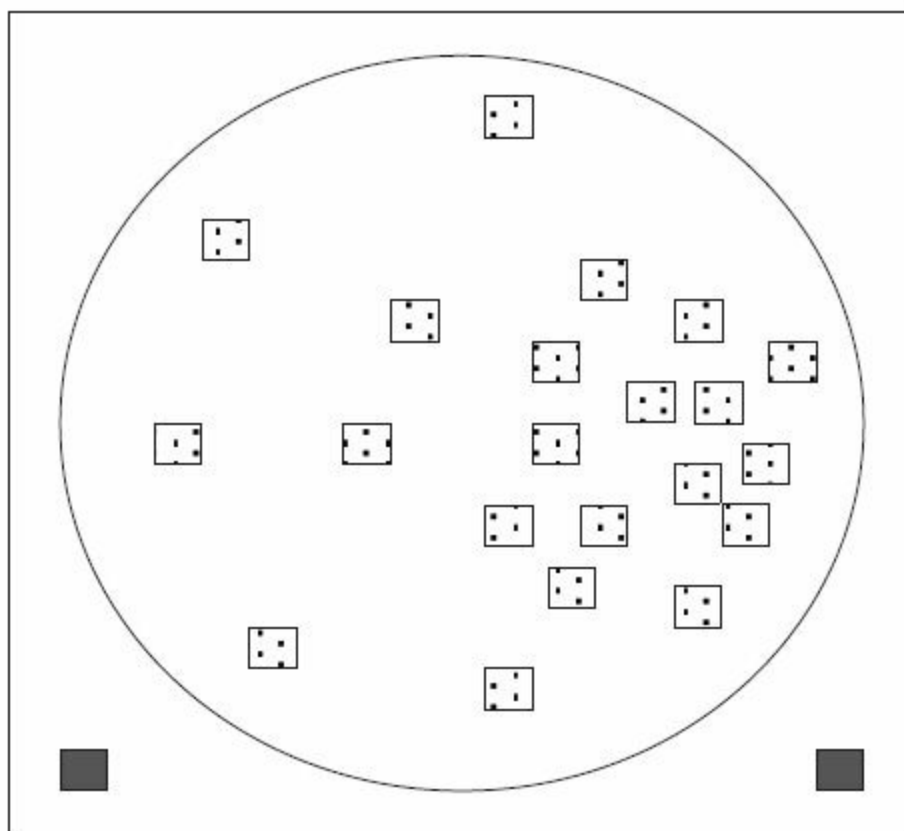


图 7-54

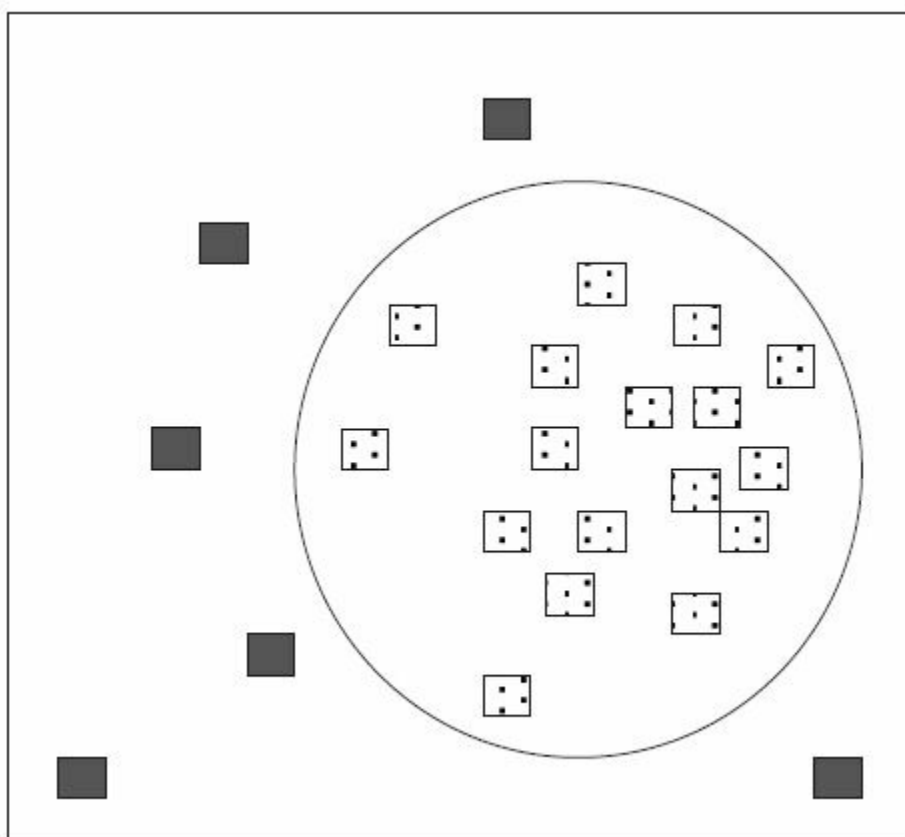


图 7-55

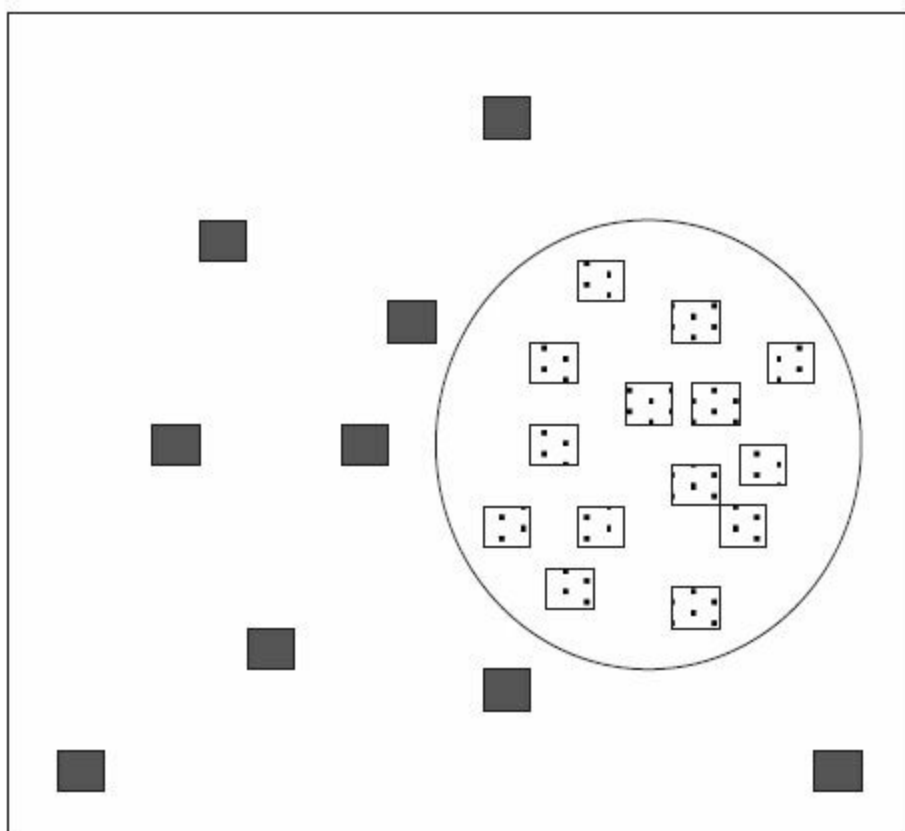


图 7-56

均质性高的数据其均衡性也不会差，在医学研究中受到混杂因素的影响会减小。

而反过来使用，我们获得的是离群点，在医学上也就是可以发现异常人群，并加以关注。

2. 异质性组现象的启发

在以医学数据为代表的一类数据中，异常侦测模型处理的结果是，在提高一组数据的同质性时另外一个组的异质性也获得了提高。这个异质组包括了很多离群点，也是特殊现象的集合。这些特殊现象也是有很多用处的，如可以从不同寻常的现象中启发思路，可以找到群组的特殊个体，如图7-57所示。

64.45	女	4922	5343965189369333	14.470000000000001	57	308	11	28	5.080000000000001	1.95	4
64.45	女	4923	50032368380384074	5.919999999999999	64	349	39	53	4.150000000000004	1.79	6
81.08	女	6282	74124318047925275	24.649999999999999	228	678	36	36	3.569999999999998	4.429999999999999	22
71.82	女	6546	65617919440363492	5.969999999999998	80	140	8	13	3.740000000000002	7.299999999999998	4
69.81	女	6899	643715693309582	6.219999999999998	78	299	11	15	4.389999999999997	2.310000000000001	5
76.11	女	9053	71703793118919124	8.240000000000002	102	286	9	21	4.910000000000001	2.490000000000002	6
77.15	女	9375	73998917700782252	6.990000000000002	102	376	10	22	4.089999999999999	1.45	7
87.73	男	21554	75836364975327675	6.070000000000003	111	464	34	26	3.939999999999999	9.300000000000005	10
84.5	男	21768	56311242157075858	6.950000000000002	72	344	14	16	6.29	.989999999999999	4
86.13	女	21947	59036952932531628	5.299999999999998	121	375	40	31	3.580000000000001	.75	9
75.04	女	23032	73941510600901095	4.929999999999997	104	352	9	15	4.620000000000001	1.47	9

图 7-57

离群点代表的很有可能是重大发现，如在医学上发现了一种新的疾病、一种新的微生物。在股票上可能是新的动向趋势，新闻分析中可能是新政策出台的标志等。

7.6.2 关联规则的高级应用

前面，我们介绍了关联规则的经典应用，下面我们讨论一下关联规则在医学上的扩展应用。我们观察一下事件A和事件B的并发组合，如表7-1所示。

表7-1 A、B事件的并发组合

并发 ID	前件 (位置 1)	后件 (位置 2)
001	A	B
002	A	- B
003	- A	B
004	- A	- B

注：-A表示非A，-B表示非B。

作为事件A和事件B的并发组合一共四种情况：

001：前件A->后件B

002：前件A->后件非B

003：前件非A->后件B

004：前件非A->后件非B

其中“非A”表示事件A没有出现，“非B”表示事件B没有出现。如果我们将以上每种情况的支持度和置信度简化为高低两个取值，则不同支持度和置信度下的事件A和事件B并发组合总共有16种，如表7-2

所示。

表7-2 A、B事件的并发组合情况

前件	后件	置信度	支持度
A	B	H	H
A	B	H	L
A	B	L	H
A	B	L	L
A	–B	H	H
A	–B	H	L
A	–B	L	H
A	–B	L	L
–A	B	H	H
–A	B	H	L
–A	B	L	H
–A	B	L	L
–A	–B	H	H
–A	–B	H	L
–A	–B	L	H
–A	–B	L	L

7.6.2.1 对不同并发模式的兴趣

不同支持度和置信度的并发组合模式分为如下几种：

HH模式：高支持度高置信度

HL模式：高支持度低置信度

LH模式：低支持度高置信度

LL模式：低支持度低置信度

第一种HH模式是很常见的应用，前面谈过很多了，这里我们不再讨论了。我们主要看一下后面三种模式在医学上的应用。

1) HL模式

在合并症研究中，关联规则模型的支持度高时，说明该合并症病例的基数很大，置信度虽然可能不高，但因为病例的基数大，事件A（合并症A）伴随事件B（合并症B）出现的情况也是一个常见的情况。比如某种代谢病的基数很大，伴随出现酮症酸中毒症虽然置信度不高，但是因为代谢病的基数大，酮症酸中毒症伴随代谢病（或者说代谢病合并酮症酸中毒症），也算是一个常见的情况，需要加以注意。

2) LH模式

这是一种低支持度和高置信度的模式。因为这种模式的置信度高，所以该伴随现象从前件推知后件的正确概率很高。这样的模式见于一些少见疾病的特殊伴随现象。这种特殊伴随现象作为一个特征，可以用来发现和提示少见疾病的特殊伴随现象的出现，比如伤寒和相对缓脉的同时出现。或者是常见疾病的少见伴随现象，通过前件来推

测后件的出现，如感冒引发病毒性心肌炎。

3) LL模式

低支持度和低置信度模式看似无用，但是在医学这种特殊行业（当然其他行业也可能有类似情况），会有一些很少见的伴随现象，却具有极高危险性，如某症状A引发急性心肌梗塞这样的情况。低的置信度是因为这种情况下，心肌梗塞不表现出典型的症状，而是以少见的症状A为表现。如果没有想到症状A后面的心梗，则会耽误救治的最佳时机，可能是一个致命的疏漏。LL模式主要应用于寻找这类拥有高破坏力后件的少见伴随现象。

7.6.2.2 负模式规则和零模式规则

如表7-1中所示，我们将001（前件A->后件B）这样的关联规则，称为关联规则的“正模式”；将002（前件A->后件非B）或003（前件非A->后件B）这样的关联规则，称为关联规则的“负模式”；将004（前件非A->后件非B）这样的关联规则，称为关联规则的“零模式”。

经典的关联规则发现的是表7-1的ID001的部分。当事件A出现时事件B也出现，且具有高支持度和高置信度，如某种疾病的高血压症状和高血胆固醇症状的伴随出现具有高置信度和支持度。这样的关联规则我们称之为“正模式”规则。ID002和ID003是另外一种关联形式，浅灰色的是事件A出现时事件B不出现，如买了A品牌啤酒的顾客不买

B品牌啤酒，再如有高血压症状的病人没有高血胆固醇症状的伴随出现。深灰色的是事件A不出现时事件B出现，如没有买A品牌啤酒的顾客购买B品牌啤酒，再如没有高血压症状的病人却有高血胆固醇症状的出现。这样的关联规则我们称之为“负模式”规则。ID004是又一种模式，没有出现事件A时事件B也没有出现，如没有购买A啤酒的顾客也不会买B啤酒，再如没有高血压症状的病人也没有高血胆固醇症状的出现。这样的关联规则我们称之为“零模式”规则。

医学数据的特点与购物数据不同。医学数据的非A（非B）不是购物篮中的没有购买，而是可能某医学指标处于正常范围、低于某值、高于某值或未出现某症状等这些非异常的状况。

1) 负模式规则应用

负模式规则是一种特殊的伴随现象，在医学上的用途之一是可以发现致病因素或疾病的保护因素。当非A事件与疾病伴随出现时，尤其是非A事件与疾病的伴随是LL模式时，提示事件A可能是个保护性因素。当然这只是个现象和可能，具体事件A是不是该病的保护性因素还需要大量的医学验证。而相反，如果事件B是对某疾病有益的因素，当事件A与非B事件伴随出现时，A可能作为某疾病的危险因素而出现。另外，当发现某种症状不出现时能确诊某疾病，这样的情况可以发现、排除诊断某种疾病的特征维度（症状）。

2) 零模式规则应用

零模式也是一种特殊的伴随现象，医学上的用途之一是可以发现某疾病的致病因素或保护因素。在合并症的研究中，当非A事件与不利的非B事件伴随出现时，非A事件可能是个保护因素。如事件B是“高密度脂蛋白在正常范围”，非A事件与非B事件“高密度脂蛋白不足”伴随出现，事件A有可能是个保护因素。如果非A事件与有利的非B事件伴随，则事件A可能是个致病因素或危险因素。

当然，我们提出的负模式规则应用和零模式规则应用绝不能像前面讨论的这样简单操作，实际情况远远比我们想象的要复杂，我们需要将所有的影响因素考虑进去。这里，我们给大家提供比经典关联规则应用更宽泛的应用思路。

7.6.2.3 关联规则的方向性

关联规则前件和后件的关系是有方向的。前件引发了后件的出现，后件伴随着前件的出现。伴随数量占总量的比是支持度，伴随数量占后件的比例是置信度。置信度的大小决定了关联规则的方向。

- 如果“前件A->后件B”的置信度和“前件B->后件A”的置信度都高，我们说事件A与事件B是双向的，即事件A前件引发了事件B后件出现的置信度与事件B前件引发了事件A后件的置信度相当且都较高，说明二者可能是“共生”的。

- 如果“前件A->后件B”的置信度高，但“前件B->后件A”的置信度低，我们说事件A与事件B是单向的，即事件A前件引发了事件

B后件出现的置信度远大于事件B前件引发了事件A后件的置信度，说明二者可能是“寄生”的。

· 如果“前件A->后件B”的置信度和“前件B->后件A”的置信度都低，我们说事件A与事件B是无向的，即事件A前件引发了事件B后件出现的置信度与事件B前件引发了事件A后件的置信度都较低，说明二者可能是“无关”的。

一般说来，单方向规则比双方向规则更有可能提示了前件导致了后件的出现、后件伴随着前件出现现象。我们再看，考虑了关联规则的方向性后，不同支持度（H、L）和置信度（H、L）的事件A和事件B并发组合总共有32种，如图7-58所示。

位置1	位置2	置信度	支持度
A	B	H	H
A	B	H	L
A	B	L	H
A	B	L	L
A	-B	H	H
A	-B	H	L
A	-B	L	H
A	-B	L	L
-A	B	H	H
-A	B	H	L
-A	B	L	H
-A	B	L	L
-A	-B	H	H
-A	-B	H	L
-A	-B	L	H
-A	-B	L	L

位置1	位置2	置信度	支持度
B	A	H	H
B	A	H	L
B	A	L	H
B	A	L	L
B	-A	H	H
B	-A	H	L
B	-A	L	H
B	-A	L	L
-B	A	H	H
-B	A	H	L
-B	A	L	H
-B	A	L	L
-B	-A	H	H
-B	-A	H	L
-B	-A	L	H
-B	-A	L	L

图 7-58

我们这里，将关联规则模型的支持度和置信度的定义仅划分为高、低（H、L），实际上，在很多应用里，需要将支持度和置信度定义成更多的不同等级，如分为高、中、低，或者定义在1%到100%之间的几个区段。

我们看一个例子，假定某地区所有急诊患者中，高热为事件A，咳嗽为事件B，同时假定该地区可能受到某种呼吸道烈性传染性疾病的侵袭，高热和咳嗽同时出现者有可能是该病患者，需要密切注意。将患者数据使用关联规则引擎处理，输出的结果中的（A->B）模式和（B->A）模式肯定是需要特别关注的，这里我们将（A->B）模式简称为AB模式，（B->A）模式简称为BA模式。我们看一下引擎需要提供什么基本功能。首先需要引擎定义：前件A=高热，后件B=咳嗽。因为我们对这个AB模式或BA模式最感兴趣，不论是AB模式还是BA模式，它们的支持度是相同的，不同的是置信度（参见第3章）。在一般情况下，支持度也可以理解为可能患病的人数，是在一个范围内波动的；同样，AB模式或BA模式的两个置信度也有其波动范围。在遭受疾病侵袭的情况下，AB模式和BA模式的支持度和置信度三个值会偏离正常范围。比如可以表现为潜在患者的大量增加，也可能是AB模式和BA模式的支持度上升，也可能表现为AB模式和BA模式的置信度的变化。

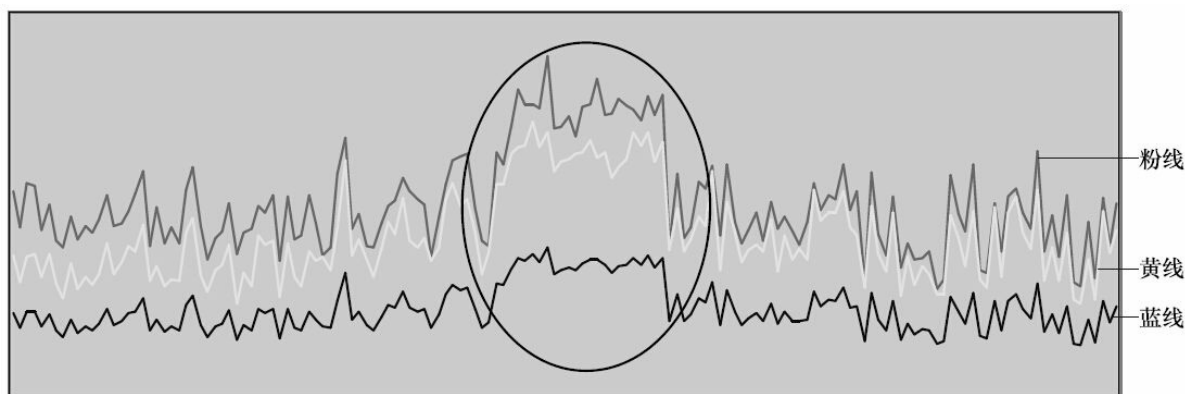


图 7-59

图7-59为AB模式的支持度（蓝，最低线）、BA模式的置信度（粉，最高线）、AB模式的置信度（黄，中间线）的时间变化图。这些变化也可能影响有关事件A和事件B的关联规则的方向性。

7.6.2.4 关联规则的时间性

关联规则反映的伴随关系是随着时间的变化而变化的，比如前面章节的取药高峰伴随出现于年终。在商业上，销售量与各个节日的伴随关系、与季节的伴随关系也都与时间有关。在考虑了时间之后，关联规则形成了这样的形态：每种不同的伴随关系有自己独特的变化形状，某一时段或某一时刻的伴随关系可能随着时间变化而完全不同，我们把这种关联规则称为时间关联规则，如图7-60所示。

时间关联规则形成了如图7-60这样类似两条（也可能多条）互相缠绕的水管形态，每个不同的截面所代表的时间伴随关系有自己独特的形状，随着时间的变化，某种伴随关系可能变化很小，或者变化很大，可演变成与前面截然不同的伴随关系。在实践中我们也需要考虑

这样的问题，如下面例子。

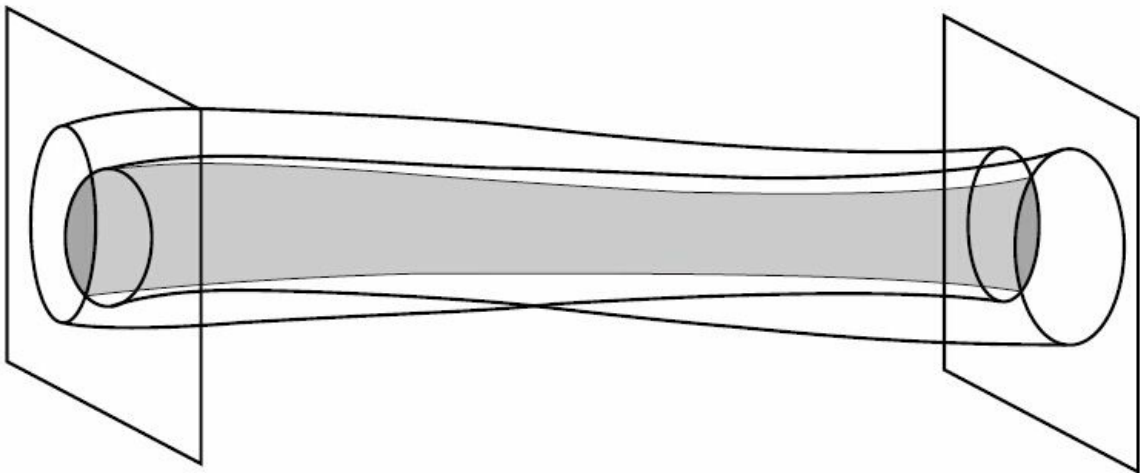


图7-60 关联和水管

图7-61和图7-62是我们发现的同一种疾病的两种亚型（图7-61为亚型1，图7-62为亚型2），在一个时间段内的两个主要生理指标（由粉线和蓝线代表）关联规则伴随关系的时间演变。诸如这样的医学上的精密关系，可以使用面向专业的数据挖掘系统方便、简单地发现。

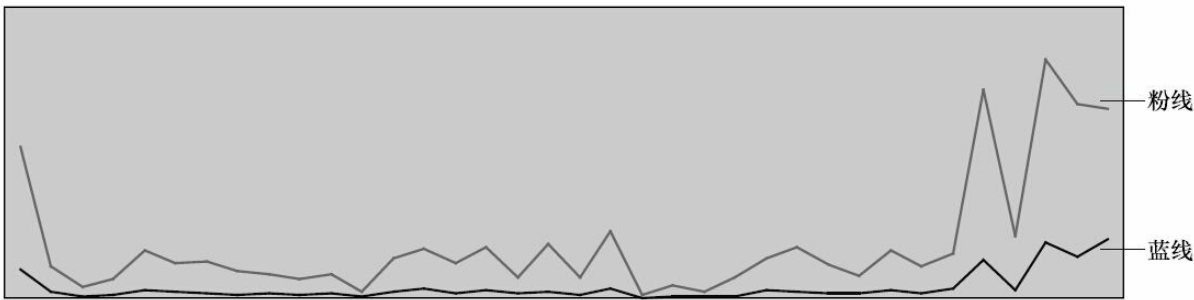


图 7-61

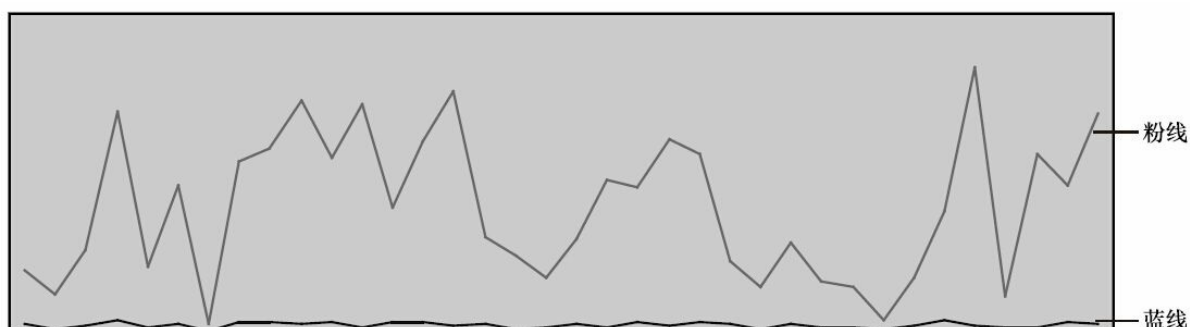


图 7-62

支持度和置信度对于一个特定关联规则来说是其属性之一。而随着时间的变化，某特定关联规则的支持度和置信度也在不断地变化。在很多实际应用中，随时间的变化，关联规则的属性变化也是有一定规律的，通常我们可以制订某特定关联规则的属性值的正常范围。通过建立大量的属性正常范围，对于观察和比较某特定关联规则的属性值的各种异常偏离来说，既省时又省力，可以使用正常范围来过滤出那些偏离的规则。

7.7 本章小结

我们试图较深入地展示一个数据挖掘的商业应用案例。我们认为，数据挖掘应用越是专业、越是深入，从某种意义上讲，也就越是通用、越是具有广泛借鉴意义，因为数据挖掘的实践应用是脱离不开某一具体行业和专业。从理论到实践需要一个过程，需要一个学习、借鉴、探索、实践、验证的过程；同样，从一个实践到另外一个实践，也需要一个学习、借鉴、探索、实践、验证的过程。

本章讲的是医学行业的应用，但是其中很多思想和方法也适用于其他行业。比如本章的第7.6.2.1节中所谈的几种并发组合，在商业方面，低支持度和高置信度的LH模式经常见于小规模忠实用户群体的购买行为模式；在金融市场、国际关系预测中也需要寻找低支持度和低置信度LL模式所代表的高破坏力小事件等。我们希望本书对其他行业的数据挖掘应用能起到一个抛砖引玉的作用。

第8章 数据挖掘工具的应用

本章将介绍使用数据挖掘通用工具开展的一些应用。虽然形式上，我们将利用一些数据挖掘工具软件，其实“醉翁之意不在酒”，本章关注的重点并不是工具软件本身，而依然是实际的项目应用。

随着数据挖掘技术和应用的不断发展，市场上产生了许多数据挖掘的通用工具软件。目前为大众所熟悉的数据挖掘软件主要有以下几种：

- 传统的数据挖掘软件，如IBM公司的SPSS Modeler和SAS公司的Enterprise Miner。
- 开源的数据挖掘软件，如RapidMiner、KNIME、Weka等。
- 自动化数据挖掘软件，如KXEM Analysis Framework。
- 专门化数据挖掘软件，如Viscovery SOMiner、Prudsys Discovery、Bissantz Delta Master等。
- BI产品内置的数据挖掘软件，如SAP Netwear的Data Mining Work bench、Oracle的Data Mining、Microsoft SQL Server的Analysis Services等。

利用工具软件搞数据挖掘项目有便利性和优越性。这里，我们将介绍应用两个数据挖掘通用工具软件的一些案例。这两个工具，一个

是Oracle的Data Mining，另一个是IBM的SPSS Modeler。这两个工具是比较具有代表性的商业数据挖掘应用软件，其中Oracle Data Mining是引擎基于数据库的可编程的数据挖掘工具的代表，IBM SPSS Modeler则是典型的独立引擎的客户端数据挖掘工具（它同时也提供服务端的批处理运算）。

8.1 应用Oracle Data Mining

Oracle的数据挖掘工具Oracle Data Mining（简称ODM）有着较长的历史，它是一个集成在Oracle数据库中的数据挖掘引擎，主要功能包括探索数据中的特征模式、识别关键属性、发现新的群组 and 关联模式，并揭示有价值的规律等。按照Oracle公司的提议，Oracle的数据挖掘工具主要应用于以下几个方向：

- 预测和防止客户流失，主要采用分类技术
- 获得新客户并识别最能带来收益的客户，主要采用特性选择技术
- 识别会成功的交叉销售机会，主要采用关联规则技术
- 洞察违规的和欺诈性的活动，主要采用异常侦测技术
- 发现新的集群或片断，主要采用聚类侦测技术
- 开发客户概要文件，主要采用特征抽取技术
- 识别药物发明中的可能目标和有希望的线索，主要采用回归预测技术
- 找出同时出现的项目或事件的关联关系，主要采用关联规则技术

- 挖掘非结构化数据，即文本挖掘，主要采用特征抽取技术

应用Oracle Data Mining，可以体现知识自动发现的思想，即数据挖掘通过运行模型自动来完成。知识自动发现是基于数据挖掘模型的运行，一个模型的运行是通过算法作用于数据集的过程实现的。本节我们将介绍有关ODM的数据挖掘流程、算法模型和算法应用等几个方面。

8.1.1 ODM数据挖掘流程

图8-1展示了一个ODM数据挖掘项目的各个阶段。流程图显示，当部署一个数据挖掘特定方案时，数据挖掘项目不会静止，而是按照流程图不断循环往复执行的过程。

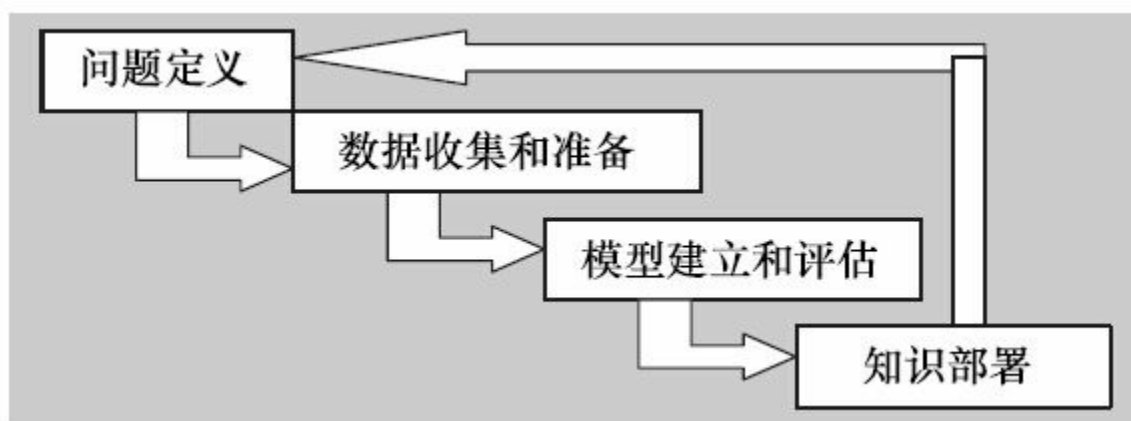


图 8-1

8.1.1.1 问题定义

数据挖掘目标的初始阶段集中于理解项目的目标和要求。一旦从业务角度确定了项目，我们就可以详细定义数据挖掘需要解决的问题，这里我们称之为“数据挖掘问题”（数据挖掘专业问题），并且开始实施计划。例如，我们的业务问题可能是：如何向客户销售更多的产品？我们可以把它转化为一个“数据挖掘问题”：哪些客户最可能购买我们的产品？基于过去已购买产品的客户的数据，我们可以建立一个预测购买产品客户的模型。在建立模型之前，我们假设，在已购买产品客户的数据和尚未购买产品客户的数据之间，一定存在某种联

系。客户的属性可能包含年龄、房主或租房者信息、居住时间，甚至孩子数目等。这是个从对问题的业务描述到对问题的数据挖掘描述的过程，从业务术语到数据挖掘术语的转换过程，从业务思维转换成数据挖掘思维的过程。

8.1.1.2 数据收集和准备

数据理解阶段还包括数据搜索和收集（Data Gathering and Preparing）。当我们深入分析数据时，我们会发现数据对业务问题的描述程度。我们可能决定删除或增加一些数据，这也是辨别数据质量和了解数据分布模式的阶段（详见第2章）。

8.1.1.3 模型建立和评估

在这个阶段，我们可以选择和应用各种建模技术，并调整最佳参数。如果某个算法需要数据处理，我们可以返回上一步。项目在这个阶段，需要评估模型在多大程度上能满足先前定义的业务目标。如果模型可用于预测最可能购买产品的客户，那么这个模型是否能充分地将两类客户区分开来？增加一些属性数据是否可以改进模型？造成假阳性和假阴性的成本，是否在模型中考虑到了？从工程角度说，获取数据的成本是否考虑到了？可能的收益是否可以抵消成本支出？

8.1.1.4 知识部署

知识部署（Knowledge Deployment）就是在目标环境下使用数据挖掘模型。在这个阶段，可从模型中导出数据内在的、有指导性的信息。知识部署包括：

- 模型应用分析，如将模型应用于新数据
- 抽取模型细节，如抽取决策树规则
- 在应用、数据仓库、报表工具中集成数据挖掘模型

数据挖掘的流程显示，问题定义阶段的业务问题数据挖掘化，与知识部署阶段的数据挖掘问题业务化互为逆过程。前者需要将业务问题的描述转换成数据挖掘的语言，而知识部署阶段需要反过来，将数据挖掘语言再次“翻译”回业务描述。

8.1.2 ODM算法模型

按照算法分类，Oracle Data Mining主要包括七类算法模型：

- 聚类模型
- 特性选择
- 关联规则
- 特征抽取
- 异常侦测
- 分类模型
- 回归预测

8.1.2.1 聚类

聚类（Clustering）有些类似分类，可以用来区分数据。与分类不同的是，聚类将数据分到没有预先定义类别的簇中，而分类则将数据划分到预先定义类别的簇中，它们需使用目标类值。聚类在探索数据的未知分布中是非常有用的。如果有大量个案，但没有明显的类别，聚类算法就可以用于发现它们的自然类别。聚类还可以用于建立用来识别分类的有监督模型的预处理过程。聚类还可以用来异常侦测，一旦数据被分到不同的簇中，就会发现有少数数据不适合于任何簇，这

些数据要么是异常点，要么是离群点。

聚类就是按照数据在某些方面的相似性，来找出数据的不同分簇，簇内的成员比簇间成员更具有相似性。聚类分析的目的，就是要使得簇内有足够的相似性和簇间足够有差异性。ODM聚类算法（Clustering Algorithms）主要包括：

- 增强K-Means聚类
- 层次划分O-Cluster聚类

8.1.2.2 特性选择

特性选择（Feature Selection），它是一个有监督学习的模型，它是给预测目标或分类目标的属性因子的显著性打分。发现最有价值的预测因子或分类因子是很多数据挖掘项目的目标。例如，一个模型试图寻找具有高信用风险客户的基本特征，这时就可以应用特性选择模型，来辨别有关客户的信用风险的属性特征。特性选择方法也常常用于分类模型的预处理过程，特别是在ODM朴素贝叶斯算法和支持向量机算法建模之前，而在ODM决策树算法中已经包括了特性选择的算法部分。

在商业数据挖掘应用中，有时太多的信息可能会降低数据挖掘的效率。有些属性在建立和测试模型时未必能提供有价值的信息，有些信息甚至会降低模型的质量和精度。例如，我们可能搜集到某一群人

的大量全面数据，我们期望从这些数据中能预测发生某一疾病的概率。这些数据中，可能有很多信息对于推测某疾病有很少的价值，或根本没有价值。例如，每家有几辆车的属性可能对于这个主题根本没有意义。无关的属性只是增加了数据的噪声，而且也影响了模型的准确度。噪声数据还增加了模型的建立和应用所需的时间和系统资源。ODM特性选择，就是将所有对预测目标或分类目标有影响的属性因子打分，以评估这些属性因子辨别预测目标或分类目标的能力，并将没有影响力的属性因子删除。ODM特性选择算法（Feature Selection Algorithms）主要包括：最小描述长度MDL算法（Minimum Description Length）。

8.1.2.3 关联规则

关联规则（Association）用于发现数据集中同时发生事件的概率，同时发生事件之间的关系可以表达为关联规则。关联规则的结果，就是产生一些规则用来辨别数据中的关联模式。关联规则常常用于分析销售交易。例如，实例显示在食品店购买奶酪的客户也常常同时购买牛奶。事实上，关联分析发现，有85%的交易中同时包括奶酪和牛奶。因此，我们得到这样一条规则：奶酪=>牛奶85%置信度。关联模型的应用也称作货篮分析，它在市场分析、产品促销和发现商业趋势等方面是很有价值的，关联模型在其它领域也有着重要的应用。ODM关联规则算法（Association Algorithms）主要包括：Apriori算法。

8.1.2.4 特征抽取

特征抽取（Feature Extraction）就是一个属性降维的过程。不同于特性选择（对于给定目标属性，将其他所有因子按照其对目标属性预测的影响度进行打分），特征抽取实际上是属性变换。经变换了的属性，或者特性，是原来属性集的线性合并。特征抽取会产生更小、更精的一组属性。用特征抽取建立的模型可能是质量更好的，因为数据被更少的和更有意义的属性所描述。特征抽取将高维数据集映射到一个低维的数据集中。特征抽取对数据可视化是很有用的，当一个复杂的数据集被降维到二维或三维空间时，数据可有效地被观测。

特征抽取可应用于潜在语义分析、数据压缩、数据分解和映射、模式识别等。特征抽取也可用于提高有监督学习的速度和效率。特征抽取还可以用于抽取文本数据中的主题，这些主题被描述为一系列关键词及它们的频度。每个主题特征用一些关键词的组合来表达，文本集可以表达为一些关键词推导出来的主题。

在项目应用中，很多数据具有很多的属性，许多属性之间还具有相关性。具有相关性的属性，实际可以被划为同一属性类别之内。这些相关属性的共同出现扭曲了算法逻辑，也影响了模型的精度。具有大量属性的数据给数据挖掘算法带来了运算上的挑战。模型属性就是算法处理的空间维度，处理的空间维度越高，算法运算的开销就越大。为了最小化噪声、相关性、高维度的影响，维度压缩是数据挖掘必须的处理过程，特征抽取正是一种很好的维度压缩方法。ODM特

征抽取算法（Feature Extraction Algorithms）主要包括：非负矩阵因子分解NNMF算法（Non-Negative Matrix Factorization）。

8.1.2.5 异常侦测

异常侦测（Anomaly Detection）的目的，就是要找出数据中看似异类的非正常案例。异常侦测是一种用于侦测欺诈、网络入侵，和其他一些难于发现却又有重大意义的罕见事件的重要工具。例如，执法机构希望从数据中甄别出违法活动，又不能误判合法活动，如何准确地辨别出可疑行为？再如，保险公司可以处理上百万件保险申报案，其中有一小部分是骗保案，如何识别出这些骗保案？这些都需要异常侦测的技术支持。

异常侦测也是一种分类，即单分类（One-Class Classification），它执行的是一类分类，因为训练数据集中只有一个类别。异常侦测模型预测一个数据点是否是一个典型的给定分布。一个异常数据点要么是一个噪声，要么就是一个未知类的个案或离群点。

通常，一个分类模型需要在包括正类和负类的数据中进行训练，以学会识别它们。例如，一个模型需要预测一种药物的负作用，就需要在包括该药广泛的药物反应数据中进行训练。异常侦测开发了一个框架，它能一般性地挑选训练集中的典型案例，偏离这个框架就视为异常点。单分类器有时候也可以用于正面探索模型，因为它要从大部分坏的行为中找出少数好的行为。

异常侦测用于发现离群点（Anomaly Detection for Finding Outliers）。离群点是一种异常，因为它们落入数据的正常分布之外。从与正常分布的中心点的距离，就可以看出一个点是否是一个正常分布点。每一个案例都可以按照它是正常或异常的概率进行评估。ODM异常侦测算法（Anomaly Detection Algorithms）主要包括：一类支持向量机One-Class SVM算法（One-Class Support Vector Machine）。

8.1.2.6 分类模型

分类模型（Classification），就是一种根据已知的分类标号将输入的数据集建立分类的数据挖掘方法。分类的目标就是将数据的每个个案都尽可能准确地预测到一个目标分类中。例如，分类模型可以将很多贷款申请区分为高风险、中风险和低风险三个类别中去。首先考察一段时间内贷款申请数据的基于历史信用等级，还要考虑贷款申请人的工作历史、本地居住年限、投资的类型和数量等数据。将信用等级作为目标属性，其他条件作为预测属性，每个贷款申请人作为一个案例数据，建立一个分类模型，从而通过有限的预测属性来比较准确地推测案例数据的目标属性。实际上，分类模型的应用十分广泛。

ODM分类算法（Classification Algorithms）主要包括：

- 支持向量机SVM算法（Support Vector Machine）
- 决策树算法（Decision Tree）

- 贝叶斯算法 (Naive Bayes)
- 自适应贝叶斯网络算法 (ODM10.2, Adaptive Bayes Network)
- Logistic回归算法 (ODM11.3, Logistic Regression)

8.1.2.7 回归预测

回归 (Regression/Prediction) 是一种预测数值变量的数据挖掘方法。利润、销售、贷款利率、温度、距离等都可以用回归进行预测。例如，一个回归模型可以用来根据一套房子的地点，房间数、房屋面积及其他指标来预测房子的价值。回归分析要从目标值已知的数据集开始。例如，建立预测房屋价值的模型，要收集、观察许多房屋在相当一段时间内的数据。除了收集房屋的价值外，还收集有关房龄、占地面积、房间数、税率、校区，甚至购物中心等属性数据。房价是目标属性，其他属性可能是预测因子，而每套房子的数据就是一个案例。

在模型建立、训练过程中，回归算法评估每个案例的因子导出的目标值。预测因子和目标属性之间的关系可以在模型中进行概括，然后将该模型应用于目标属性值未知的数据集中。回归模型可以通过计算各种统计结果来进行测试，这些统计结果度量了预测值与期望值之间的差别。用于回归模型的历史数据可典型地分为两组，一组用于建立模型，另一组用于测试模型。ODM的回归算法

(Regression/Prediction Algorithms) 主要包括：

- 支持向量机SVM算法 (Support Vector Machine)
- 广义线性回归GLM算法 (ODM11.3, Generalized Linear Models)

SVM是一个强大的线性和非线性回归算法。SVM回归支持两个内核：Gaussian非线性回归和Linear线性回归。SVM支持主动学习（无监督学习）。GLM是广泛使用的统计学技术，用于线性模型。GLM提供可信的参数设计和模型设计，也用于支持置信度判定。回归模型可应用于很多方面，如：趋势分析、业务规则、市场营销、金融预测、时间序列分析、生物医学和药物反应模型、环境分析模型等。

8.1.3 ODM算法应用

ODM客户端工具是Oracle Data Miner。我们将应用Oracle Data Miner展现ODM在特性选择、分类模型、聚类模型，和回归预测几种算法方面的应用，我们选择恶性肿瘤病人的数据和股票交易的数据作为数据挖掘应用案例。这些算法和应用案例都是比较典型的商业应用，具有较好的代表性。

另外，目前我们没有发现完全中文的Oracle Data Miner版本，在我们展现的版本里，界面里有一些英文的注释，我们尽可能把更多的信息留给读者，以便于读者更好地理解相关内容，而且我们认为阅读本书的读者应该具有读懂的能力，况且本书本身不是ODM的工具书，因此我们保留了原有界面的内容，没有给出中文说明，如果读者专门学习工具软件，请参考有关工具文献。

8.1.3.1 特性选择：区分恶性肿瘤病人的属性

肺癌是一种严重威胁人类健康的恶性疾病。在中国乃至世界范围内，肺癌发病率和死亡率均居恶性肿瘤之首。随着全球自然环境的恶化，肺癌发病率仍呈快速上升趋势。由于缺乏有效的技术手段，难以实现早期发现，80%以上的肺癌确诊患者均为中晚期肺癌，并且难以获得满意的疗效，五年存活率仅为10%。因此寻找能够早期发现肺癌的新的肺癌标记物和治疗靶点的研究具有重要的意义。本案例中采集了观察组恶性肿瘤病人组570例的数据，对照组（健康人群和良性肿

瘤病人组) 233例数据(详细见4.2节)。将观察组和对照组相关的基本信息、既往史、生化检测信息(肿瘤标志物)、病理数据、专业的检测信息等作为区分两组的特征属性加以考察,以期得到对区分两组贡献比较大的特征属性,从而为医学专家寻找新的肺癌标记物和治疗靶点提供线索和依据。

有关数据准备阶段的一些工作如数据描绘、数据标准化等按第4章的方法实施。下面我们将数据准备阶段完成后的数据表BJ_XK_ALL_2进行特性选择分析,我们采用的是ODM客户端工具Oracle Data Miner的特性选择Attribute Importance模块。由于此处是首次通过ODM来展示应用,我们会简要附带Oracle Data Miner的应用流程,以便读者在了解数据挖掘项目过程的同时,对Oracle Data Miner的应用也有一些了解。

首先打开Oracle Data Miner(这里的版本为10.2),在数据表和视图列表中点击目标分析数据表“BJ_XK_ALL_2”,如图8-2所示,我们可以从图中右侧的“Structure”栏中观察该表的数据结构,重点查看一下每个字段类型和大小的定义。

此时,我们可以点击鼠标的右键,系统会弹出多个功能菜单,如图8-3所示。我们可以看到一级功能菜单包括“Transform”(数据变换)、“Predict”(预测)、“Explain”(解释)等功能。其中,“Transform”的二级功能菜单包括“Aggregate”(数据聚合)、“Compute Field”(字段计算)等功能(详细功能介绍参见有关

工具手册)。如果需要, 我们可以利用这些功能对数据进行数据挖掘准备阶段的处理和即时分析(如简单预测和解释等)。

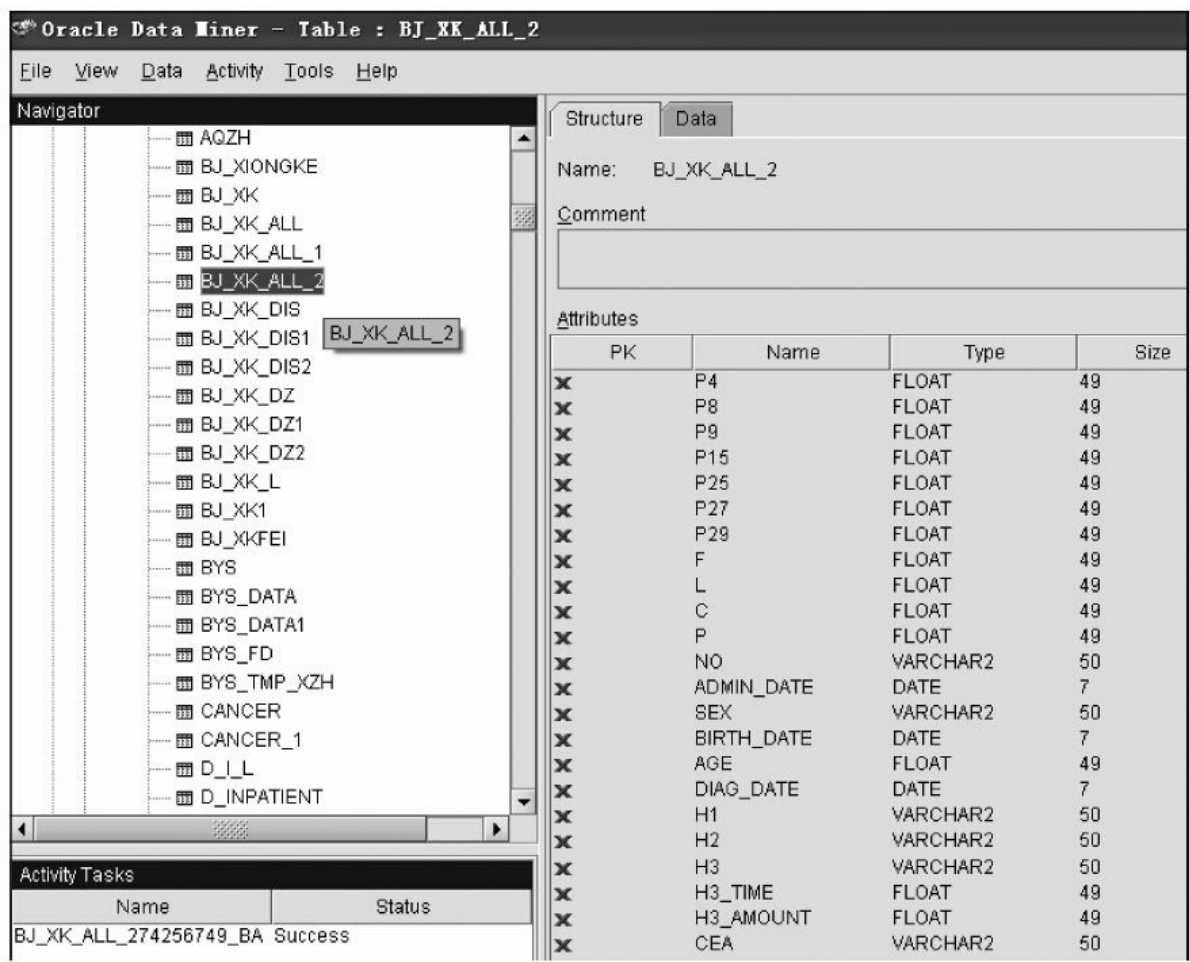


图 8-2

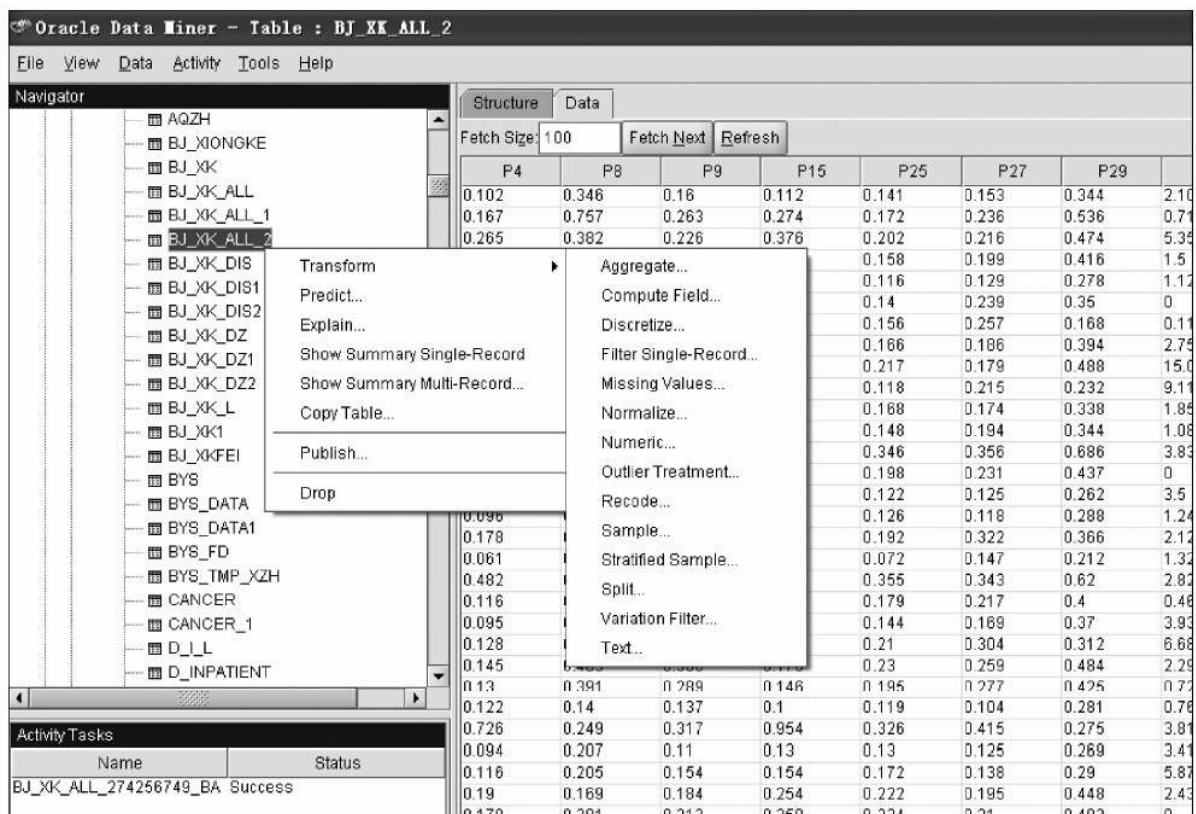


图 8-3

然后，我们点击Oracle Data Miner主菜单“Activity”下的“Build”选项，系统自动弹出一个向导式对话框，如图8-4所示。点击“下一步”按钮，进入下一个页面。



图 8-4

在“Function Type”下拉选框中，系统列出了有关数据挖掘算法类别，如用于异常侦测的“Anomaly Detection”、用于关联规则的“Association”、用于特征抽取的“Attribute Importance”、用于分类的“Classification”、用于聚类的“Clustering”、用于特征抽取的“Feature Extraction”、用于回归预测的“Regression”。这里，我们选择“Attribute Importance”模块，如图8-5所示。

Select Mining Activity Type

Choose a model function type and algorithm. Review the descriptions to be sure you have picked the most appropriate selections. Click the Help button for additional details.

Function Type:

Attribute Importance

Algorithm:

Anomaly Detection
Association Rules
Attribute Importance
Classification
Clustering
Feature Extraction
Regression
Anomaly Detection

Description:

Attribute Importance ranks the predictive attributes by eliminating redundant, irrelevant, or uninformative attributes and identifying those predictor attributes that may have the most influence in making predictions.

图 8-5

点中“Attribute Importance”选项后，系统自动在“Algorithm”栏中显示可以选择的具体算法，如图8-6所示。这里，Oracle Data Miner仅提供了一种算法，即“Minimum Description Length”最小描述长度MDL算法。

Function Type:

Attribute Importance

Algorithm:

Minimum Description Length

Description:

Attribute Importance function:
- Select most important attributes for classification

Minimum Description Length Algorithm:

Usage:

Attribute Importance ranks the predictive attributes by eliminating redundant, irrelevant, or uninformative attributes and identifying those predictor attributes that may have the most influence in making predictions.

图 8-6

点击“下一步”按钮，进入下一个页面，如图8-7所示。选择要分析的数据表的有关选项，如数据表的用户、表名、唯一键字段名、参加分析的字段名等。

Select the Case Table

Select the table containing the "cases" (individual records/rows) that will be input to your mining activity. You can unselect any table columns that you know should not be considered as mining attributes. You can also join additional data in with the case table by selecting the checkbox below.

Schema: HEALTH

Table/View: BJ_XK_ALL_2

☐ Join additional data with case table

Unique Identifier: ☒ Single Key: CASE_ID
☐ Compound, or None
NOTE: Compound (multiple supporting table. This

Select Columns:

Select	Name	
<input checked="" type="checkbox"/>	ADMIN_DATE	
<input checked="" type="checkbox"/>	AGE	
<input checked="" type="checkbox"/>	BIRTH_DATE	
<input checked="" type="checkbox"/>	C	
<input checked="" type="checkbox"/>	CASE_ID	NUMBER
<input checked="" type="checkbox"/>	CEA	VARCHAR2
<input checked="" type="checkbox"/>	DIAG	NUMRFR

Sampling Settings... Include All Exclude All

图 8-7

点击“下一步”按钮，进入下一个页面，如图8-8所示。选择特性选择模型的目标字段，这里是“DIAG_NEW”（诊断）字段，它包括“观察组”和“对照组”两个取值类别。然后选择特性选择模型目标字段的特征属性作为输入字段，这里包括“AGE”、“C”、“CEA”等多个

指标等。注意，系统会根据事先对每个字段的初始计算，给出一组默认的特征属性输入字段，其中自然不会包括数据表中的唯一键（这里是“CASE_ID”），我们根据业务需要和技术需要对输入字段进行取舍。

Review Data Usage Settings

Select the target column, and review the column settings. You can change the column settings to better match your understanding of the data. The default settings have been determined for each column based on the activity type and the characteristics of the data. The options of changing input and mining type vary based on the algorithm choosen. Click Help for more details.

Data Summary

Name	Alias	Target	Input	Data Type	Mining Type	Sparsity	
HEALTH.BJ_XK_ALL_2							
ADMIN_DATE	ADMIN_DATE	<input type="radio"/>	<input type="checkbox"/>	DATE	numerical	<input type="checkbox"/>	
AGE	AGE	<input type="radio"/>	<input checked="" type="checkbox"/>	FLOAT	numerical	<input type="checkbox"/>	
BIRTH_DATE	BIRTH_DATE	<input type="radio"/>	<input type="checkbox"/>	DATE	numerical	<input type="checkbox"/>	
C	C	<input type="radio"/>	<input checked="" type="checkbox"/>	FLOAT	numerical	<input type="checkbox"/>	
CASE_ID	CASE_ID	<input type="radio"/>	<input type="checkbox"/>	NUMBER	numerical	<input type="checkbox"/>	
CEA	CEA	<input type="radio"/>	<input checked="" type="checkbox"/>	VARCHAR2	categorical	<input type="checkbox"/>	
DIAG	DIAG	<input type="radio"/>	<input type="checkbox"/>	NUMBER	categorical	<input type="checkbox"/>	
DIAG_DATE	DIAG_DATE	<input type="radio"/>	<input type="checkbox"/>	DATE	numerical	<input type="checkbox"/>	
DIAG_NEW	DIAG_NEW	<input checked="" type="radio"/>	<input type="checkbox"/>	NUMBER	categorical	<input type="checkbox"/>	
F	F	<input type="radio"/>	<input checked="" type="checkbox"/>	FLOAT	numerical	<input type="checkbox"/>	
GRADE	GRADE	<input type="radio"/>	<input checked="" type="checkbox"/>	VARCHAR2	categorical	<input type="checkbox"/>	
H1	H1	<input type="radio"/>	<input checked="" type="checkbox"/>	VARCHAR2	categorical	<input type="checkbox"/>	

Include All

Exclude All

图 8-8

点击“下一步”按钮，进入下一个页面，如图8-9所示。系统会显示一个对话框，提示用户输入本次挖掘活动的名称、注释等信息。



图 8-9

点击“下一步”按钮，进入下一个页面，如图8-10所示，系统会显示向导最后的对话框。这里，我们可以选择特性选择模型的一些高级选项。点击“Advanced Settings”按钮，系统弹出多页选项框，如在“Discretize”（离散化）选项中，我们可以选择连续变量离散化的方法，如“Quantile Binning”等分法、“Equal Width Binning”等距法。

点击“完成”按钮，系统将按照前面所选条件，执行特性选择模型的创建任务。任务完成后，系统显示有关结果，如图8-11所示。点击页面下红色的“Result”按钮，系统弹出一个新的页面，显示特性选择模型的图形结果及详细数据。

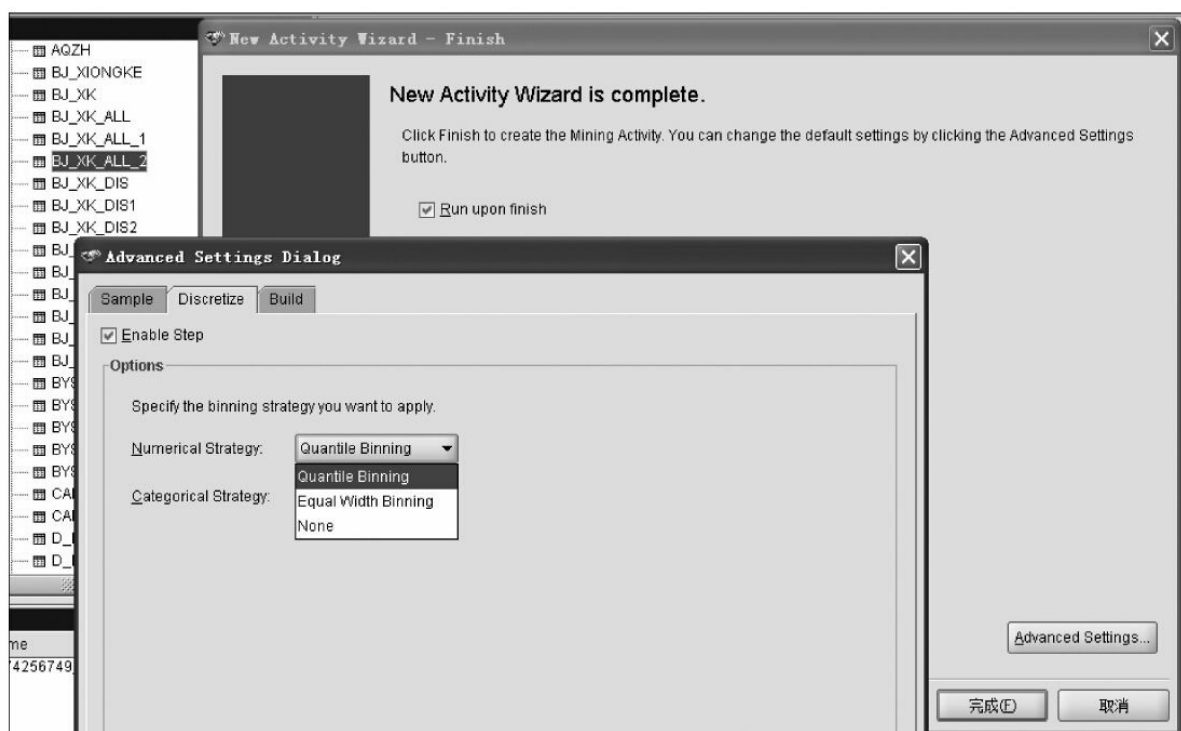


图 8-10

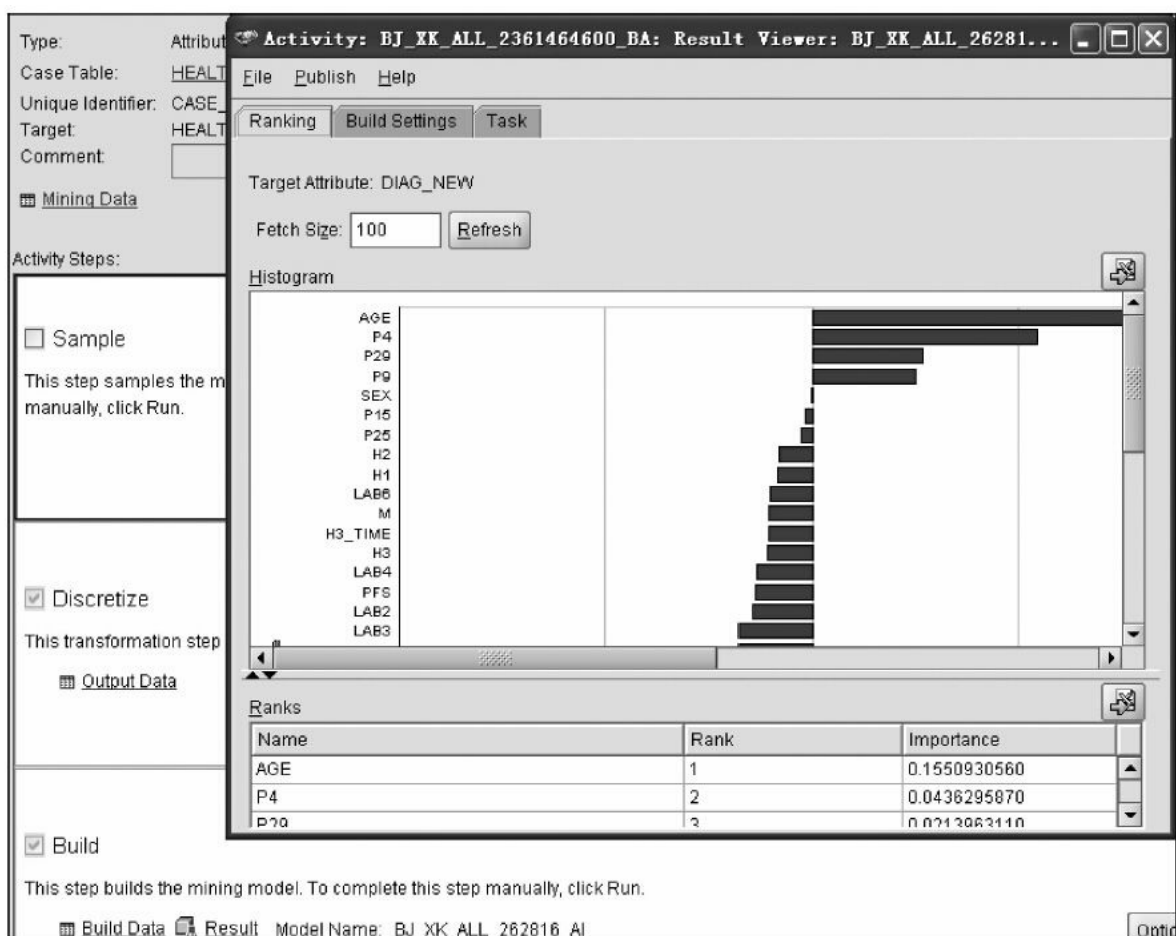


图 8-11

至此，我们完成了对有关恶性肿瘤“观察组”和“对照组”的目标字段“DIAG_NEW”的特性选择挖掘任务。接下来，我们需要对挖掘结果进行详细分析。我们将“Result”按钮弹出的特性选择模型的图形结果及详细数据展开，如图8-12和图8-13所示。

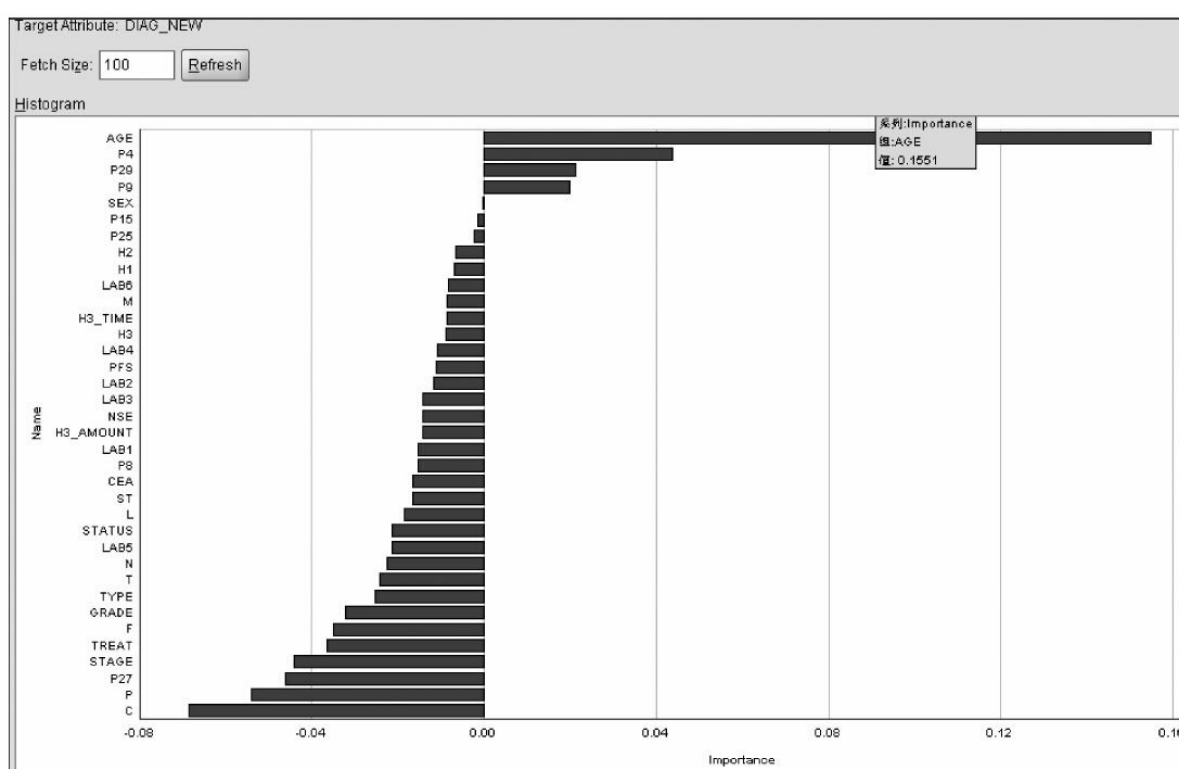


图 8-12

从图8-12中，我们可以看出，“Age”、“P4”、“P29”、“P9”几个属性的显著性为明显大于0的正值，说明这几个属性对“DIAG_NEW”的影响度较大，我们把这四个属性与“DIAG_NEW”称为“实相关”。其中，尤其是“Age”年龄与“DIAG_NEW”的相关性最大，远大于其他三个相关属性。

图中显示“Sex”、“P15”、“P25”等几个属性的显著性值接近零，说明这几个属性与“DIAG_NEW”“零相关”，即不相关。

图中显示“C”、“P”、“P27”等十几个属性的显著性为明显小于0的负值，说明这些属性对“DIAG_NEW”是混淆影响的关系，我们把它们与“DIAG_NEW”称为“虚相关”，将这些属性称为“DIAG_NEW”的混淆因子。

Ranks		
Name	Rank	Importance
AGE	1	0.1550930560
P4	2	0.0436295870
P29	3	0.0213963110
P9	4	0.0201160230
SEX	5	-0.0001693710
P15	6	-0.0012574140
P25	7	-0.0022917150
H2	8	-0.0066334210
H1	9	-0.0066713340
LAB6	10	-0.0081196030
M	11	-0.0084474440
H3_TIME	12	-0.0085460080
H3	13	-0.0087967510
LAB4	14	-0.0107198140
PFS	15	-0.0110292840
LAB2	16	-0.0116749670
LAB3	17	-0.0140792040
NSE	18	-0.0142290720

图 8-13

图8-13列出了目标字段“DIAG_NEW”的特性选择排名和具体的显著性系数。我们也可以将这个结果与其他特性选择算法得到的结果进

行比较，如第4章中的表4-16。最后再根据分类算法的选型，综合选用一些适当的特征属性作为目标分类字段的分类因子。

8.1.3.2 分类模型：恶性肿瘤病人的分类预测

在第8.1.3.1节里，我们完成了对有关恶性肿瘤“观察组”和“对照组”的目标字段“DIAG_NEW”的特性选择挖掘任务。接下来，我们将选用一些适当的特征属性作为目标分类字段的分类因子，并与目标分类字段建立分类模型。

我们首先对如上“Age”、“P4”、“P29”、“P9”等几个相关因子间再做相关性分析（此处略），以考察它们之间的独立性。结果我们发现，P4与P29有强相关关系，因此我们去掉显著性系数较小的“P29”，不将它作为分类因子。

然后我们对如上“Age”、“P4”、“P9”等几个相关因子与“C”、“P”、“P27”等十几个“虚相关”属性再做相关性分析，以考察它们与混淆因子的相关性。结果我们发现，P9与P27有强相关关系，因此我们去掉“P9”作为分类因子。

最后我们再根据医学专家从专业角度提出的建议，从绝对值较小的显著性为负数的属性中选择几个，测试它们与“Age”、“P4”的相关性。我们发现，“H1”、“H2”与“Age”、“P4”的独立性较好。最终，我们选择了“Age”、“P4”、“H1”、“H2”作为“DIAG_NEW”的分类因子，使用要求分类因子具有独立性的贝叶斯分类算法进行分析，如图8-14

所示。

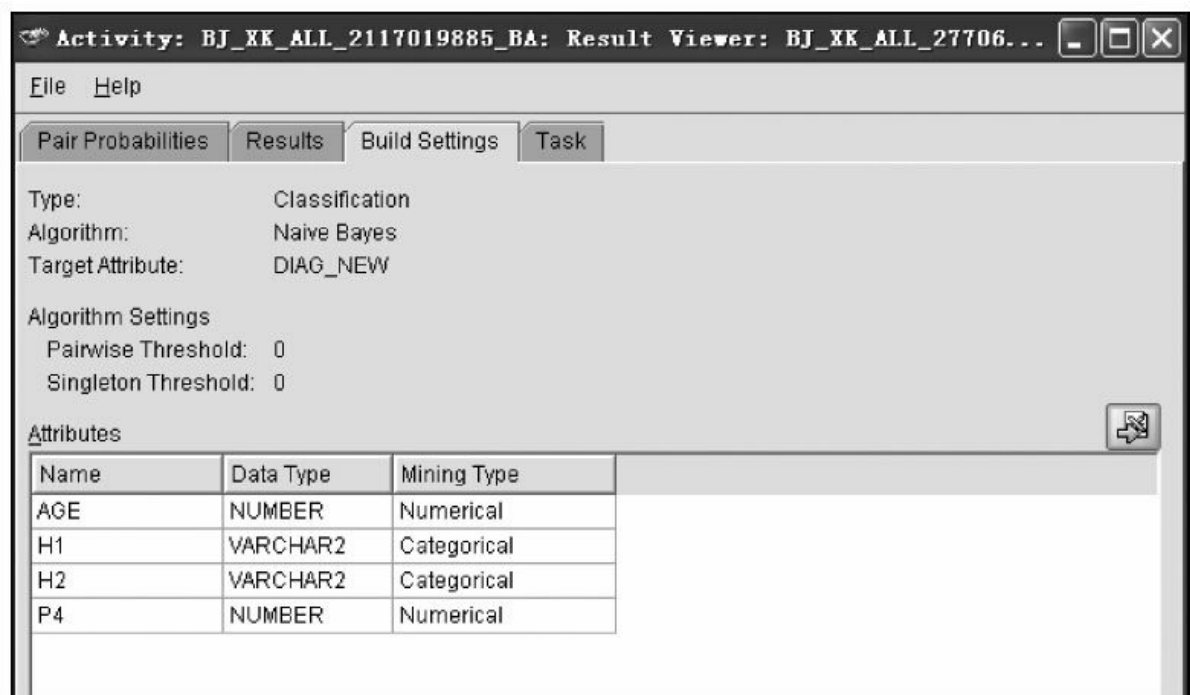


图 8-14

经过算法运行，我们得到了一个分类模型。该模型显示预测置信度为80%，也就是说，它比朴素预测模型的准确率高出80%。该模型的仪表盘指示图显示，模型位于绿色的最佳模型区域，如图8-15所示。

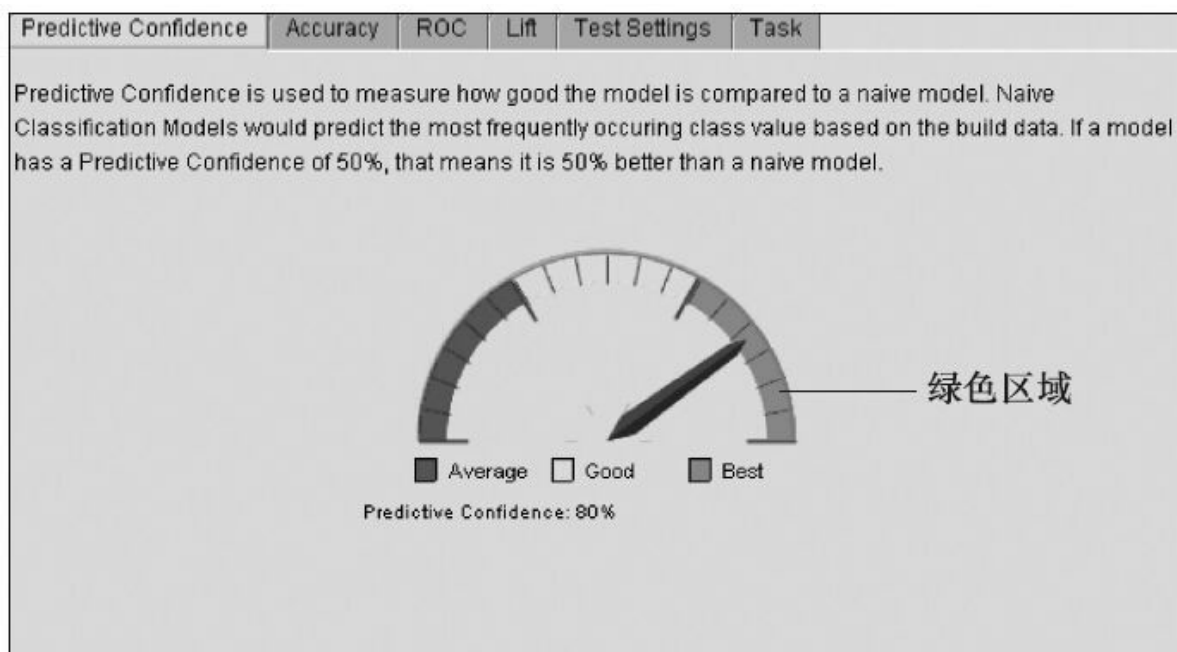


图 8-15

从图8-16模型性能的详细描述中，我们可以看出，该模型的平均预测准确率为90%，整体预测准确率为94.1%。例如，预测目标“1”（恶性肿瘤）正确的有72例，错误的有18例；预测目标“9”（非恶性肿瘤）正确的有214例，错误的有0例；因此，总预测准确率为： $(72+412) \div (90+214) = 94.1\%$ 。该模型对阴性（非恶性肿瘤）的预测性能非常好，达到了100%；对阳性（恶性肿瘤）的预测性能较好，达到了80%。

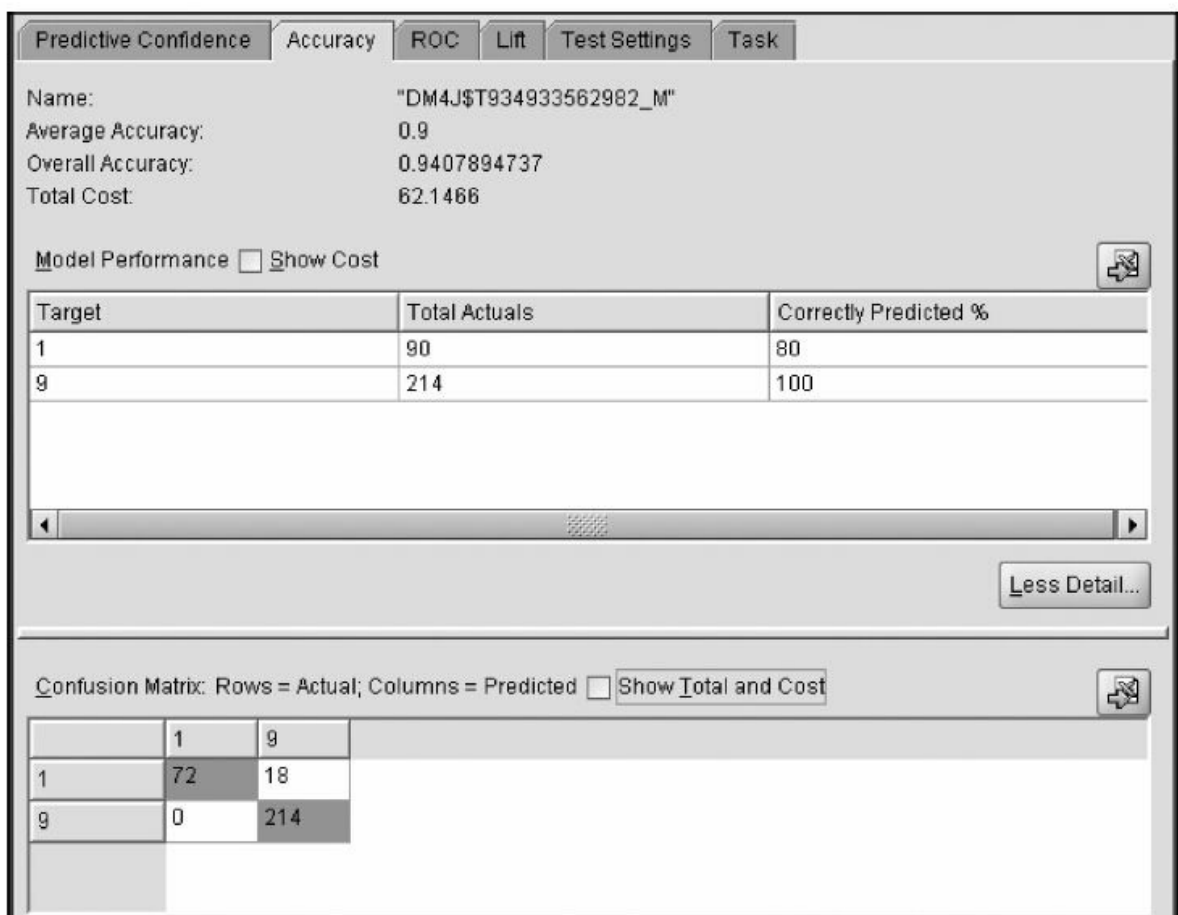


图 8-16

在医学专业领域，对于恶性肿瘤，因为早发现、早治疗的医疗效果更好，所以阳性预测的价值更高。因此我们根据专业需求，寻求该模型对阴性（非恶性肿瘤）和阳性（恶性肿瘤）的预测性能平衡，可以通过调整图8-17所示的ROC曲线图的性能阈值来实现。图中的红线可以左右移动，模型的综合预测性能也随之改变。我们调整一个接近最佳的模型性能阈值，图中显示该模型的平均预测准确率可达98.8%，整体预测准确率可达98.4%。例如，预测目标“1”（恶性肿瘤）正确的有90例，错误的仅有五例；预测目标“Others”（非恶性肿瘤）正确的有209例，错误的有0例；因此，总预测准确率为：

$(90+209) \div (95+209) = 98.4\%$ 。该模型对阴性（非恶性肿瘤）的预测性能非常好，达到了100%；同时对阳性（恶性肿瘤）的预测性能也非常好，达到了94.7%。

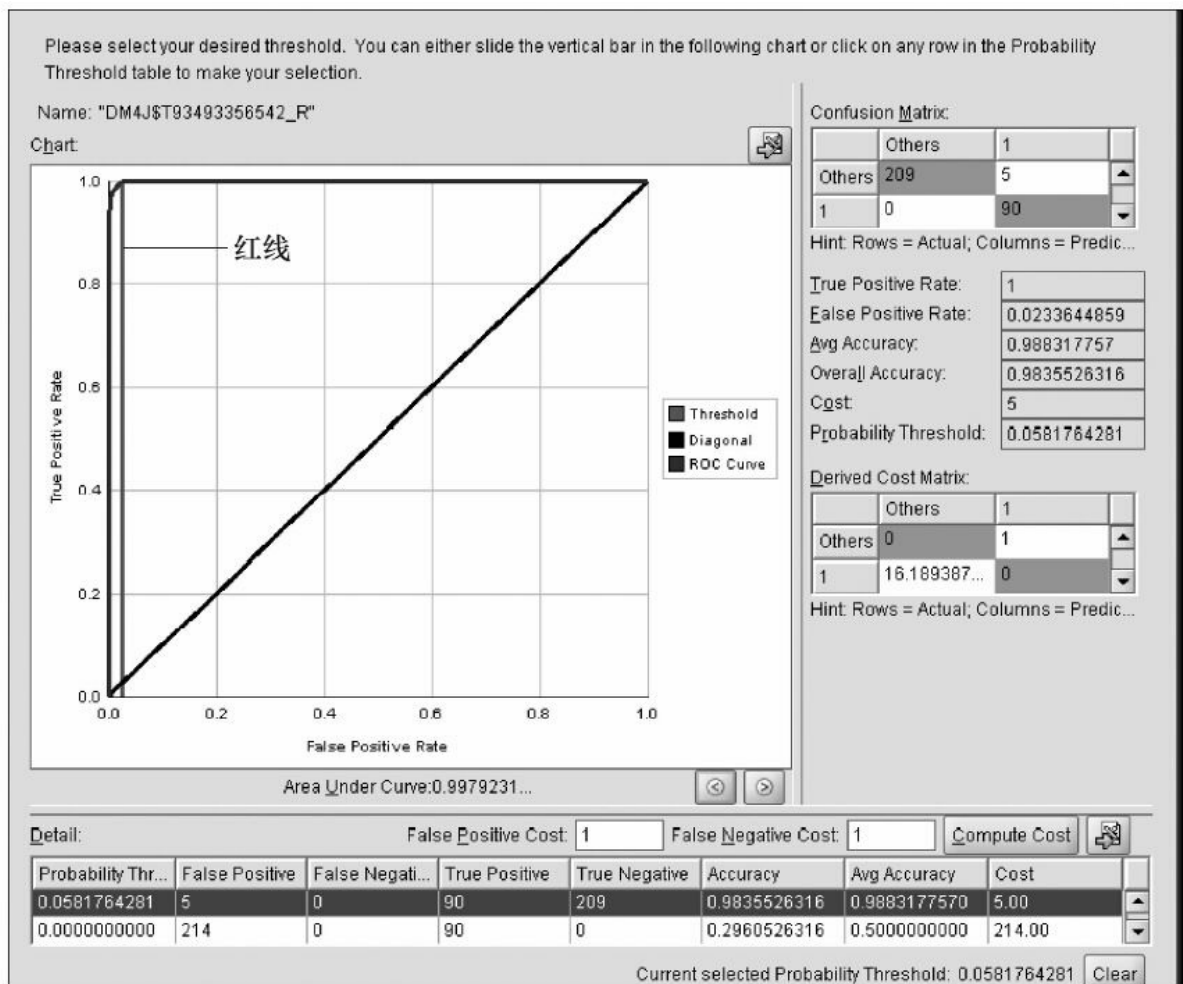


图 8-17

我们还可以查看计算不同Lift指标的解释图形。如选择“Cumulative Lift”单选按钮，我们看到该模型的Cumulative Lift值可以达到3.4，如图8-18所示，也就是说该模型的预测性能可比朴素预测模型（随机预测）好3.4倍，说明该预测模型可以达到很好的性能。

针对本案例的任务，我们对包括第8.1.3.1节和第8.1.3.2节的步骤

的整体方法和全部结果进行总结。

分析步骤如下：

- 1) 首先对原数据进行了数据清洗、整理、变换等处理，生成可分析数据。
- 2) 应用数据挖掘多种算法对数据进行分析、建模、评估。
- 3) 根据专业领域的特色需求来调整模型，并应用多种统计学工具对数据挖掘的结果进行验证。

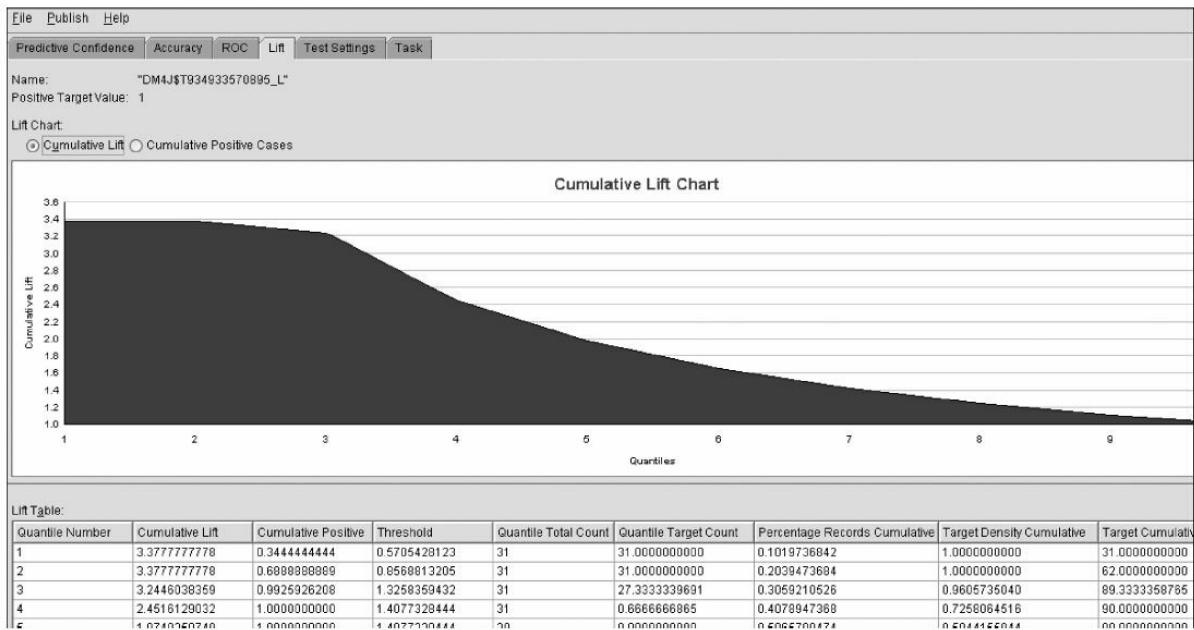


图 8-18

分析结果如下：

- 病例组和对照组合集中P04与P15、P25、P29有强相关关系，并可建关系模型。

- 病例组和对照组合集中P08与P09、P27有强相关关系。
- 病例组和对照组在年龄上有统计学差异。
- 病例组和对照组在P04上有统计学差异。
- 年龄、P4、H1和H4与病例组和对照组的类别有较强的相关关系，并可以建立分类预测模型，模型的预测置信度达80%，模型性能为优良。

此外，我们还用类似方法对对照组内的两个子集，健康组与良性组进行了分析。结果发现P15是两组最强的相关因子，性别与L检测有弱相关关系，并通过了统计学验证。

通过将观察组和对照组有关的医学信息作为区分两组的特征属性进行考察分析，本项目获得了对区分两组贡献比较大的特征属性，很好地为医学专家寻找新的肺癌标记物和治疗靶点提供线索和依据，达到了项目的预期目标。

8.1.3.3 聚类模型：肺癌病人的合并症与分型

上例中，医学专家试图通过数据挖掘技术寻找新的肺癌标记物和治疗靶点，并获得了令人满意的成果。在本案例中，医学专家在大量的临床工作中，进一步发现不少肺癌的患者同时合并患有糖尿病。依据长年的临床经验，并查证了有关国内外文献资料，研究者初步推断糖尿病与肺癌有着较密切的联系。如果能从大量的临床数据中找到二

者关联的证据，并加以科学地论证，将有助于为肺癌及糖尿病的诊断和治疗带来新的技术和方法。本例中，研究者搜集了700多例各种肺癌合并糖尿病的数据。采集了病人有关临床的全面数据多达100多项（临床指标）。

有关数据准备阶段的一些工作过程，如数据描绘、数据标准化等，在此不再叙述。我们采用的是ODM客户端工具Oracle Data Miner的聚类分析Clustering模块，算法采用基于层次聚类的O-Cluster。此处，我们省略了Oracle Data Miner的应用流程，直接将分析结果给出。如图8-19显示，我们只将最终的叶子（Leaf Clusters）规则筛选出来。

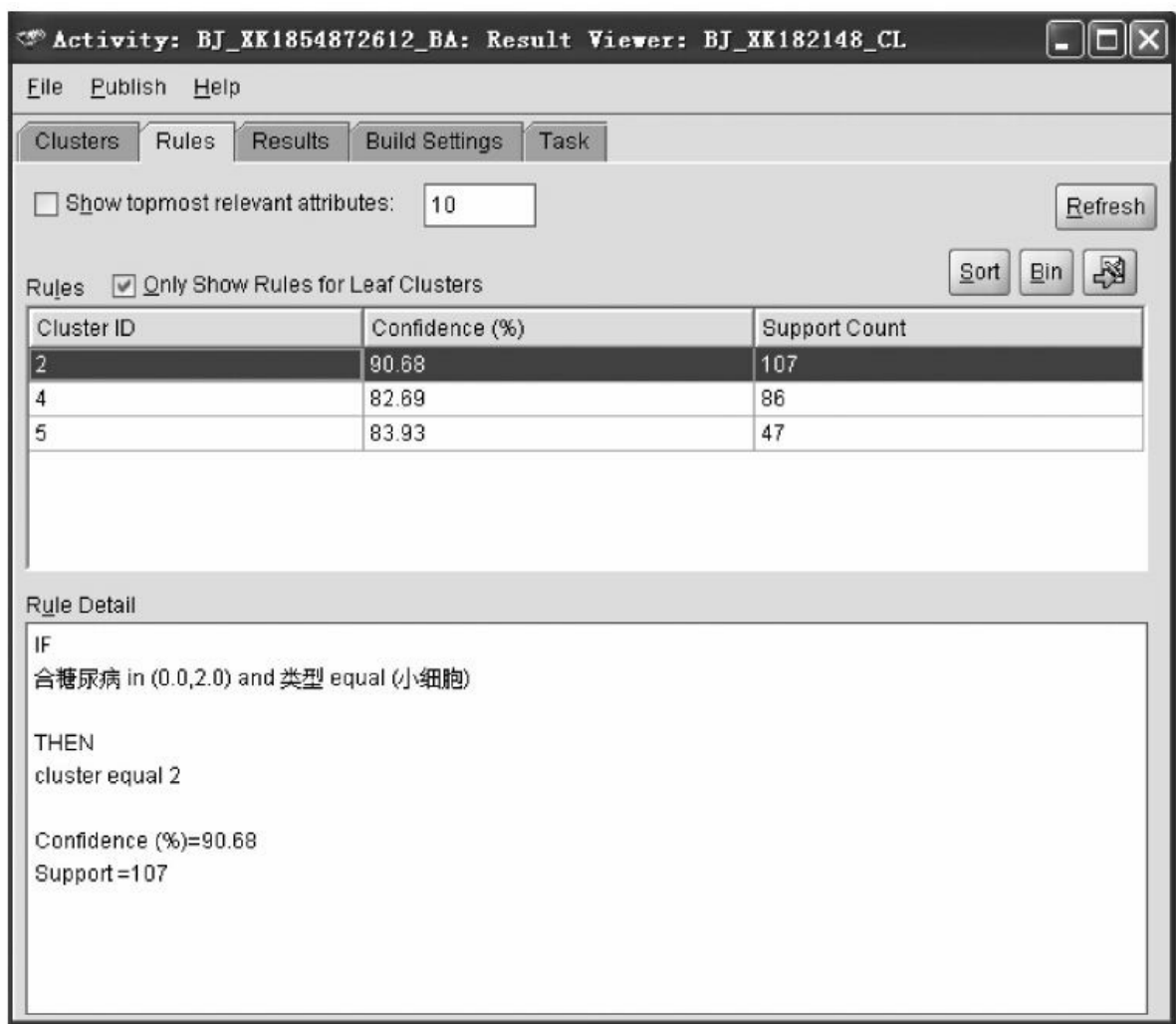


图 8-19

我们得到了三个高置信度的聚类分簇，如表8-1所示。所有分簇的置信度都超过了80%，而且支持度相对比较均衡，没有出现差距过于悬殊的情况。

表8-1 分簇的置信度

规则 ID	置信度 (%)	支持度
2	90.68	107
4	82.68	86
5	83.93	47

将三个聚类分簇逐个展开，我们可以查看每个分簇的详细规则。图8-19提示规则ID为2的一簇聚类，该簇中一些病人同时患有2型糖尿病和小细胞型肺癌，一些病人单纯患有小细胞型肺癌，总共有107例，占三个聚类分簇总和的44.6%，是三簇中最大的一簇，该簇的置信度高达90.68%，也是三簇中置信度最高的一簇。我们再来看另外一簇规则ID为4的聚类，如图8-20所示。

该簇中所有病人同时患有2型糖尿病和腺癌型肺癌，总共有86例，占三个聚类分簇总和的35.8%，是三簇中居第二位的一簇，该簇的置信度达82.69%，是三簇中置信度稍低的一簇，也是三簇中成分最简单的一簇。我们再来看最后一簇规则ID为5的聚类，如图8-21所示。

The screenshot shows a software interface with tabs for Clusters, Rules, Results, Build Settings, and Task. The 'Rules' tab is selected. At the top, there is a checkbox for 'Show topmost relevant attributes:' with a value of 10, and a 'Refresh' button. Below this, there is a checkbox for 'Only Show Rules for Leaf Clusters' which is checked. To the right of this checkbox are buttons for 'Sort', 'Bin', and a trash icon. The main part of the interface is a table with three columns: 'Cluster ID', 'Confidence (%)', and 'Support Count'. The table contains three rows: Cluster ID 2 with Confidence 90.68 and Support Count 107; Cluster ID 4 with Confidence 82.69 and Support Count 86; and Cluster ID 5 with Confidence 83.93 and Support Count 47. The row for Cluster ID 4 is highlighted. Below the table is a section titled 'Rule Detail' which shows the rule for Cluster 4: 'IF 合糖尿病 = 2.0 and 类型 equal (腺癌) THEN cluster equal 4'. It also displays 'Confidence (%)=82.69' and 'Support=86'.

Cluster ID	Confidence (%)	Support Count
2	90.68	107
4	82.69	86
5	83.93	47

Rule Detail

IF
合糖尿病 = 2.0 and 类型 equal (腺癌)

THEN
cluster equal 4

Confidence (%)=82.69
Support=86

图 8-20

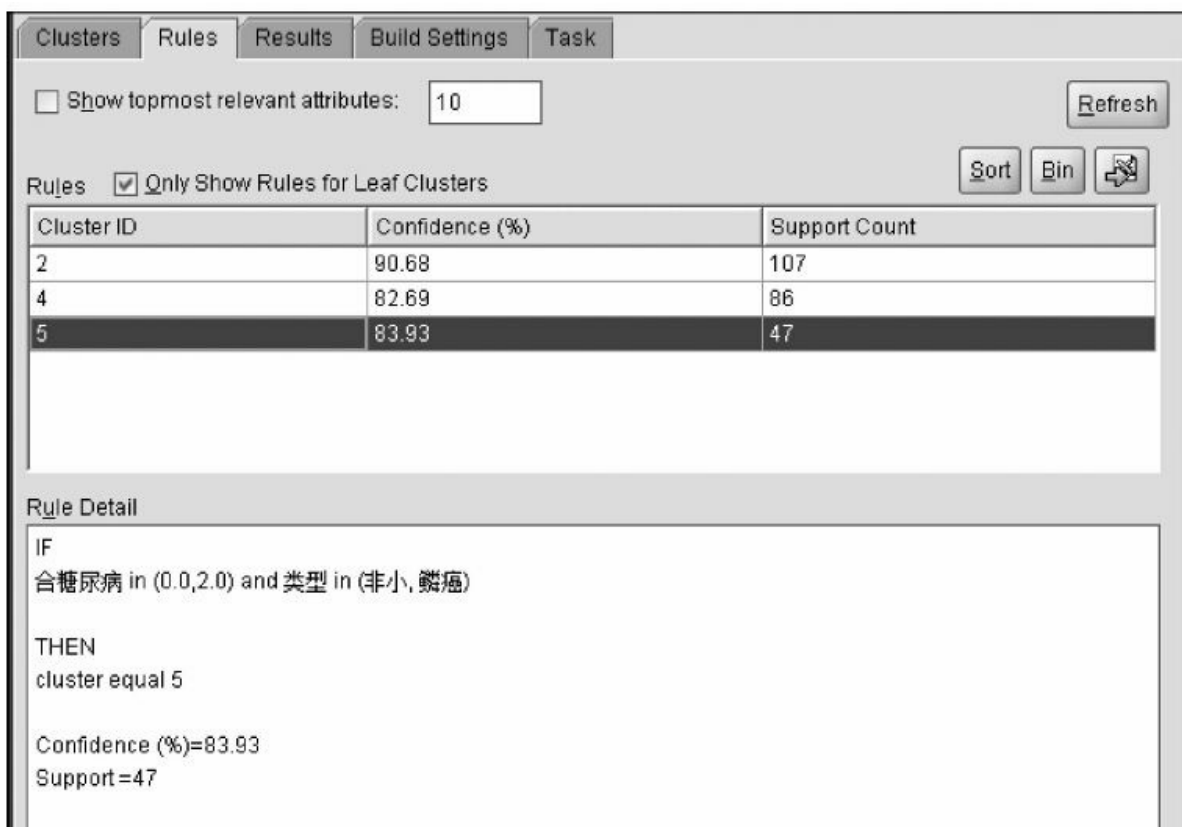


图 8-21

该簇中所有病人都是患有或未患2型糖尿病与非小细胞型肺癌或鳞癌的组合型，总共有47例，占三个聚类分簇总和的19.6%，是三簇中最小的一簇，该簇的置信度为83.93%，是三簇中置信度居中的一簇，也是三簇中成分最复杂的一簇。

从以上分析中，我们看到，ODM的聚类模型将原始数据做出了初步的、比较清晰的划分，应该说在某种程度上揭示了各种类型肺癌合并糖尿病的一些关系特征。但是，我们发现在本案例中的三个分簇中，第一个分簇和第三个分簇还有一些混合成分存在，使得我们对所有类型肺癌与是否合并合并糖尿病的关系还不完全清楚。因此，有必要对该数据再进行更详细地划分。但是，由于我们使用的ODM的工

具是算法封装的引擎，有限的参数调整不足以大幅度改变算法的性能，因此，我们没能用ODM达到令医学专家们满意的结果。然而，这并不说明是ODM本身的问题，而是通用工具的共同特点，一方面它们给使用者带来了很大的方便和高效率，但另一方面它们也不是能够解决任何问题的万能灵丹。因此，针对本案例，我们需要借助数据挖掘专家创作的定制化算法进一步深入分析。由于超出了本章涉及的内容，在此不再详细讨论。

8.1.3.4 回归预测：股票日预测和周预测

股票行情分析与股票预测是每个股民十分关系的事。市场变化莫测，股市也是瞬息万变。现实中，有太多的因素（变量）影响着股市行情。很多属于客观的因素，也有不少主观的人为因素，更有各种政治、经济、军事、科技、社会的突发事件，都左右着股价的变化。因此，很多人都认为股票是不可预测的。一方面，多数人（也许是绝大多数人）认为股票行情的影响因素（变量）太多了（成百上千），人（类）很难找到这些大变量间复杂的关系和规律；另一方面，多数人（也许是绝大多数人）认为有些股票有时候完全受人为操纵，没有规律而言。

真是这样的吗？有些数学家可能不这么认为。我们也认为，虽然影响股票行情的因素（变量）很多，但它们（股票的变化）还是有着自身的一定规律的。一方面，虽然股票行情的影响因素（变量）很多，凭借人（类）的大脑很难理解和分析成百上千的变量关系，但对

于高级数学模型和超高速计算机而言，这些还是有限变量而已，借助它们是有可能解决成百上千变量模型的问题；另一方面，即使有些股票有时候完全受人为操纵，它也是有一定规律的。因为所有股票操纵的目的都是一样的，那就是为了获利，它们所表达出来的数学模型也并不复杂。股票操纵的行为从时间的截面上看是很难把握规律的，但把它放入股票长期的历史中，很多股票操纵的行为是相似的、可分类的，既有时间上的规律也有行为特点的规律。我们把股票的所有影响因素叠加在一起，会得到一个根本上是为获利而驱动模型。

我们的观点是，既然大变量很复杂，很难理清这么多变量（因素），那我们干脆就不梳理它们，我们只关心这么多变量（因素）施加在一起后的数据规律和模型。从数据挖掘的角度上看，理论上，细到某个人的一举一动都是可预测的。这里，我们跟大家讨论这些，只是展示一种数据挖掘的思维方式。下面，我们尝试对一些股票数据进行预测分析。

1.股票A日预测

上面阐述了我们从数据挖掘的角度对股票行情分析的一些观点。实践中，我们发现不同的股票在行为表现上有着很大的不同。有的规律性较强，有的规律性较差。限于技术的投入与收益的产出权衡（这是数据挖掘实践项目首先需要考虑的），我们选择那些较为容易分析和预测的股票。经过对一些股票数据的初步挖掘分析，我们选择了一两只规律性较强的股票数据进行回归预测模型的建立。

通过数据分析和预挖掘评估，我们决定根据上一日交易的最高价和当日的开盘价来预测该股票当日交易的最高价，目的是为股民当日的交易提供预测和参考。

首先，对数据进行处理，代码如下：

```
SELECT a."case_id",a."datetime",a."open",a."high",a."low",
       a."close",a."quantity",a."amount",a."month",a."year",
       a."day_year",a."week_year",a."season_year",a."stock_id",
       a."stock_name",a."stock_block",
       c.high last_high, c.low last_low, c.open last_open, c.close
last_close,ln(a.open)
       ln_open,ln(c.open) ln_last_open,ln(c.high) ln_last_high, ln(c.low)
ln_last_low,
       ln(c.close) ln_last_close
FROM stock_day a, stock_day c
WHERE a.case_id = c.case_id + 1
AND a.stock_name = '股票A'
AND c.stock_name = '股票A';
```

这里，我们将股票A的当天交易与前一天交易的数据放在一起，包括当日开盘价、当日最高价、当日最低价、当日收盘价、前一日开盘价、前一日最高价、前一日最低价、前一日收盘价，前一日数据的LOG变换数据，以及一些时间维度数据（如年、月、日）。如图8-22和图8-23所示。

Structure Data View Lineage												
Fetch Size	100	Fetch Next	Refresh									
CASE_ID	OPEN	HIGH	LOW	CLOSE	QUANTITY	AMOUNT	MONTH	YEAR	DAY_YEAR	WEEK_YEAR	SEASON_YEAR	
1574	14.83	14.92	14.63	14.75	5818	8551000	6	2000	167	24	2	
1575	14.82	15.02	14.65	14.71	4865	7213000	6	2000	168	24	2	
1576	14.65	14.7	14.38	14.49	5689	8260000	6	2000	171	25	2	
1577	14.43	14.55	14.3	14.36	5865	8436000	6	2000	172	25	2	
1578	14.3	14.59	14.26	14.52	4261	6138000	6	2000	173	25	2	
1579	14.5	14.5	14.28	14.32	7255	10398000	6	2000	174	25	2	
1580	14.32	14.6	14.3	14.52	4515	6513000	6	2000	175	25	2	
1581	14.56	14.8	14.4	14.79	7617	11106000	6	2000	178	26	2	
1582	14.7	14.8	14.48	14.55	4685	6842000	6	2000	179	26	2	
1583	14.7	14.74	14.47	14.6	5969	8731000	6	2000	180	26	2	
1584	14.7	14.7	14.41	14.56	5550	8068000	6	2000	181	26	2	
1585	14.6	14.6	14.33	14.36	8628	9536000	6	2000	182	26	2	
1586	14.36	14.7	14.16	14.29	4801	6942000	7	2000	185	27	3	
1587	14.3	14.49	14.18	14.28	4232	6041000	7	2000	186	27	3	
1588	14.28	14.45	14.2	14.23	4026	5745000	7	2000	187	27	3	
3571	32.7	32.94	31.5	32.1	8024	25826644	12	2008	347	50	4	
3572	32.27	33.2	32.26	33	6342	20886440	12	2008	350	50	4	
3573	32.99	33.67	32.5	33.64	8320	27835680	12	2008	351	51	4	
3574	33.8	34.9	33.66	34.7	11386	39353764	12	2008	352	51	4	
3576	34.51	35.3	34.45	35.09	11928	41849932	12	2008	354	51	4	
3577	35	35.2	33.8	34.15	12429	42807840	12	2008	357	51	4	
3578	34.32	34.32	33.05	33.19	11822	39589668	12	2008	358	52	4	
3579	32.8	34.5	32.6	33.78	12478	42232304	12	2008	359	52	4	
3580	33.78	34.6	33.7	34	11743	40117444	12	2008	360	52	4	
3581	34.01	34.22	33.55	33.57	6463	21955060	12	2008	361	52	4	
3582	33.5	33.58	32.58	33.58	8912	29367486	12	2008	364	52	4	
3583	33.65	34.9	33	34.68	51844	174391040	12	2008	365	53	4	
3584	34.68	34.8	33.85	34.41	20716	71291680	12	2008	366	53	4	
3585	34.28	34.5	33	34.12	17702	59318628	1	2009	5	1	1	
3586	32.67	32.7	32.9	32.4	26924	102424442	1	2009	6	1	1	

图 8-22

我们采用的是ODM客户端工具Oracle Data Miner的回归预测Regression模块，算法采用基于支持向量基的SVM算法。值得提出的是，我们没有像很多数据挖掘应用采用的方法，为了得到更平滑的拟合，去掉了一些离群点（这也是很多预测软件不能预测股价离群点的原因），而是保留了所有的所谓的离群点，因为它们是很重要、很有价值的信息。

LAST_HIGH	LAST_LOW	LAST_OPEN	LAST_CLO...	LN_OPEN	LN_LAST_OP...	LN_LAST_HI...	LN_LAST_LOW	LN_LAST_C
15.2	14.66	14.95	14.83	2.697	2.705	2.721	2.685	2.697
14.92	14.63	14.83	14.75	2.696	2.697	2.703	2.683	2.691
15.02	14.65	14.82	14.71	2.684	2.696	2.709	2.684	2.689
14.7	14.38	14.65	14.49	2.669	2.684	2.688	2.666	2.673
14.55	14.3	14.43	14.36	2.66	2.669	2.678	2.66	2.664
14.59	14.26	14.3	14.52	2.674	2.66	2.68	2.657	2.676
14.5	14.28	14.5	14.32	2.662	2.674	2.674	2.659	2.662
14.6	14.3	14.32	14.52	2.678	2.662	2.681	2.66	2.676
14.8	14.4	14.56	14.79	2.688	2.678	2.695	2.667	2.694
14.8	14.48	14.7	14.55	2.688	2.688	2.695	2.673	2.678
14.74	14.47	14.7	14.6	2.688	2.688	2.691	2.672	2.681
14.7	14.41	14.7	14.56	2.681	2.688	2.688	2.668	2.678
14.6	14.33	14.6	14.36	2.664	2.681	2.681	2.662	2.664
14.7	14.16	14.36	14.29	2.66	2.664	2.688	2.65	2.66
14.49	14.18	14.3	14.28	2.659	2.66	2.673	2.652	2.659
33.2	32.51	32.9	32.65	3.487	3.493	3.503	3.482	3.486
32.94	31.5	32.7	32.1	3.474	3.487	3.495	3.45	3.469
33.2	32.26	32.27	33	3.496	3.474	3.503	3.474	3.497
33.67	32.5	32.99	33.64	3.52	3.496	3.517	3.481	3.516
35	33.82	34.65	34.48	3.541	3.545	3.555	3.521	3.54
35.3	34.45	34.51	35.09	3.555	3.541	3.564	3.54	3.558
35.2	33.8	35	34.15	3.536	3.555	3.561	3.52	3.531
34.32	33.05	34.32	33.19	3.49	3.536	3.536	3.498	3.502
34.5	32.6	32.8	33.78	3.52	3.49	3.541	3.484	3.52
34.6	33.7	33.78	34	3.527	3.52	3.544	3.517	3.526
34.22	33.55	34.01	33.57	3.512	3.527	3.533	3.513	3.514
33.58	32.58	33.5	33.58	3.516	3.512	3.514	3.484	3.514
34.9	33	33.65	34.68	3.546	3.516	3.552	3.497	3.546
34.8	33.85	34.68	34.41	3.535	3.546	3.55	3.522	3.538
34.5	33	34.28	34.12	3.515	3.535	3.541	3.497	3.53
32.7	32.8	32.62	32.4	3.509	3.515	3.517	3.49	3.509

图 8-23

此处，我们同样省略了Oracle Data Miner的应用流程，直接将分析结果给出。如图8-24所示，经过算法运行，我们得到了一个回归预测模型。该模型显示预测置信度为96.56%，也就是说，它比朴素预测模型的准确率高出96.56%。该模型的仪表盘指示图显示，模型位于绿色的最佳模型区域的最大值附近，从技术上说明模型性能非常优良。

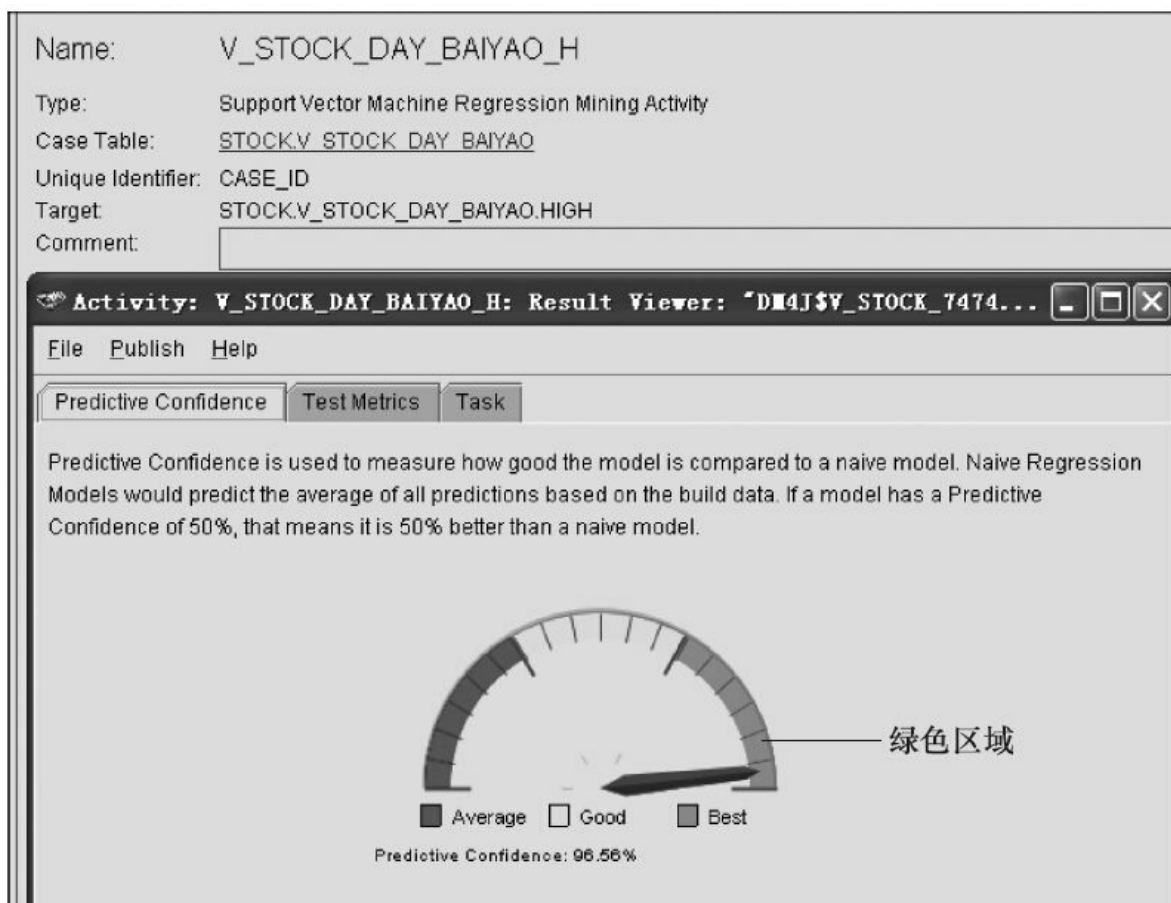


图 8-24

从图8-25中，我们可以看到模型的详细性能指标如下：

Activity: V_STOCK_DAY_BAIYAO_H:	
File Publish Help	
Predictive Confidence	Test Metrics Task
Mean Absolute Error	0.3301042872
Mean Actual Value	20.2676694915
Mean Predicted Value	20.312975571
Root Mean Square Error	0.5002458336

图 8-25

· 平均绝对误差：0.33

- 平均实际值：20.27
- 平均预测值：20.31
- 平均平方根误差：0.5

图8-26显示该模型的参数设置：我们使用了经过Log变换的上一日最高价和今日开盘价，以及日期维度变量作为今日最高价的预测因子。SVM算法的几个主要参数设置为，内核函数为“Gaussian”，标准离差为0.12，模型复杂因子为0.62，收敛度为0.001等。

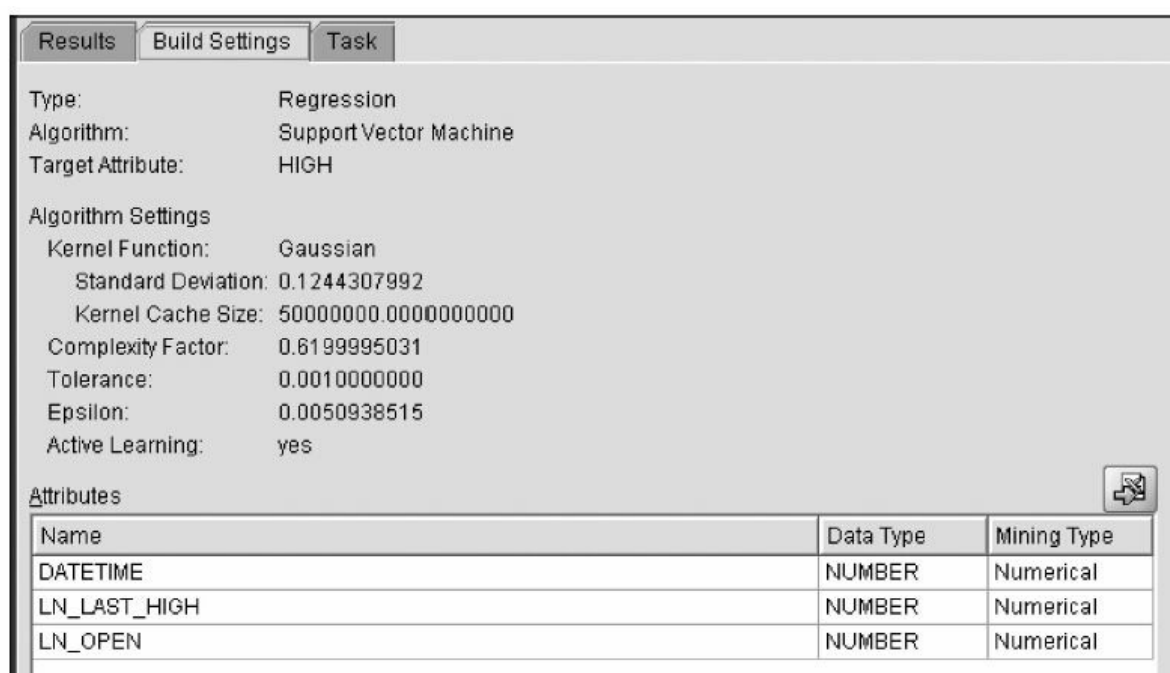


图 8-26

图8-27为回归预测模型的残差散点图，显示了该模型对股票A每日最高价历史数据的回归拟合情况。从图中，我们可以观察到该股票的历史发展可能大致分为四个阶段（这里，我们将残差值<-2和>1的

点视为离群点)：

1) 从横坐标0开始到20左右，图形比较收敛，散点比较密集，股票A交易数量较大，模型拟合比较好，但也发生过少量离群点。

2) 从横坐标20开始到40左右，图形比较发散，散点比较稀疏，股票A交易数量明显减少，模型拟合比较差，离群点发生的次数也有所增加。

3) 从横坐标40开始到55左右，图形比较发散而且非常稀疏，股票A交易数量进一步大幅减少，甚至似乎在个别时段出现了零交易的现象，模型拟合尚可，离群点发生的次数很少了。

4) 从横坐标55开始到70左右，图形依然比较发散，散点比较稀疏，但比上一阶段有所好转，股票A交易数量有了明显增加，模型拟合欠佳，离群点也时有发生。

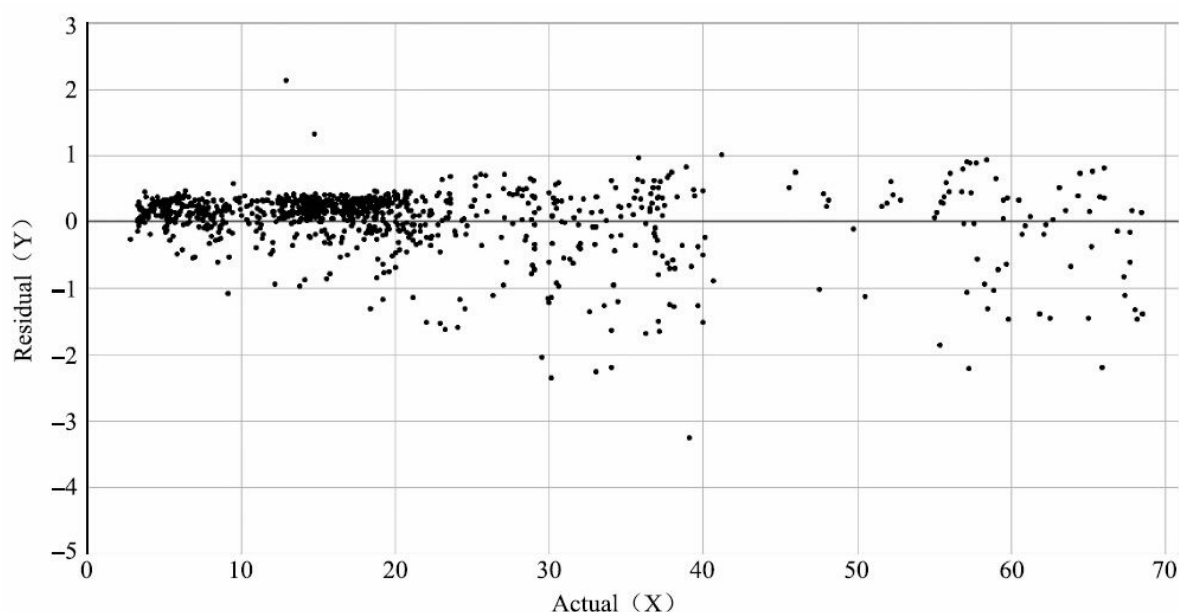


图8-27 日高点图

从以上分析中，不难看出，股票A历史交易的变化较大。我们的回归预测模型反映的是对股票A历史数据的拟合。模型结果显示，虽然整体性能还算不错，但一方面，由于股票A近期历史表现得更发散，所以该模型对股票A现阶段的走势拟合得不一定很好；另一方，拟合的精度的好坏，不能仅从技术上考虑，还要考虑业务实际操作的需求。因为每个模型的预测都有一定的误差，如果一个模型剔除了技术误差造成的影响，刨去了业务操作的成本之外，能给模型使用者带来一定利润的话，该模型才具有实用价值。

图8-28显示了该模型预测股票A的实例数据，“HIGH”栏是股票A当日最高价的实际值，“PREDICTION”是股票A当日最高价的预测值。

Residual Plot Task		
Apply Output Table:		
DMR\$CASE_ID	HIGH	PREDICTION
522	3.25	3.2551
523	3.25	3.2916
561	3.3	3.5993
521	3.3	3.2489
560	3.33	3.5912
526	3.33	3.3729
534	3.33	3.3872
528	3.35	3.3383
524	3.38	3.4323
553	3.39	3.5771
550	3.39	3.5823
530	3.44	3.4759
569	3.46	3.69
540	3.46	3.5488
559	3.49	3.6919
571	3.56	3.7235
136	3.65	3.7173
503	3.68	3.6711
510	3.68	3.6755

图 8-28

应用与前面类似的方法，我们得到了一个预测股票A当日最低价的回归预测模型。该模型显示预测置信度为96.92%。仪表盘指示图显示，该模型从技术上说也属于性能非常优良，如图8-29所示。图8-30则提示该模型的平均绝对误差为0.26，平均实际值为19.50，平均预测值为19.50，平均平方根误差为0.43。

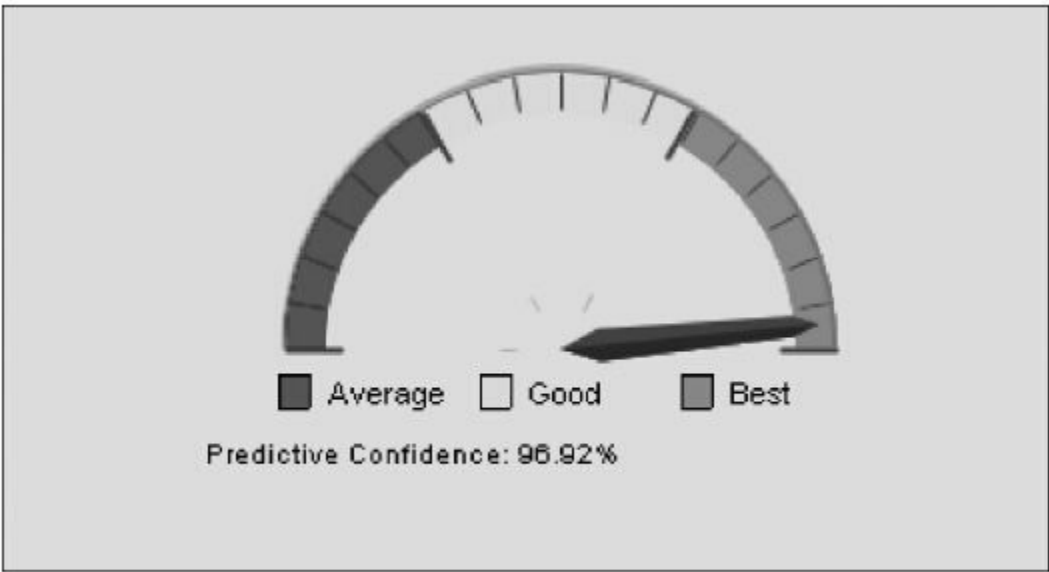


图 8-29

Predictive Confidence	Test Metrics	Task
Mean Absolute Error	0.263986936	
Mean Actual Value	19.5044552058	
Mean Predicted Value	19.5038845551	
Root Mean Square Error	0.4305254823	

图 8-30

图8-31为残差散点图，显示了该模型对股票A每日最低价历史数据的回归拟合。从图中，我们可以观察到该股票的历史发展大致与每日最高价残差散点图的规律类似，不同的是，在第四阶段，该图显示有“两个”非常大的最低价离群点，模型没能很好地拟合，似乎发生了什么突发事件？

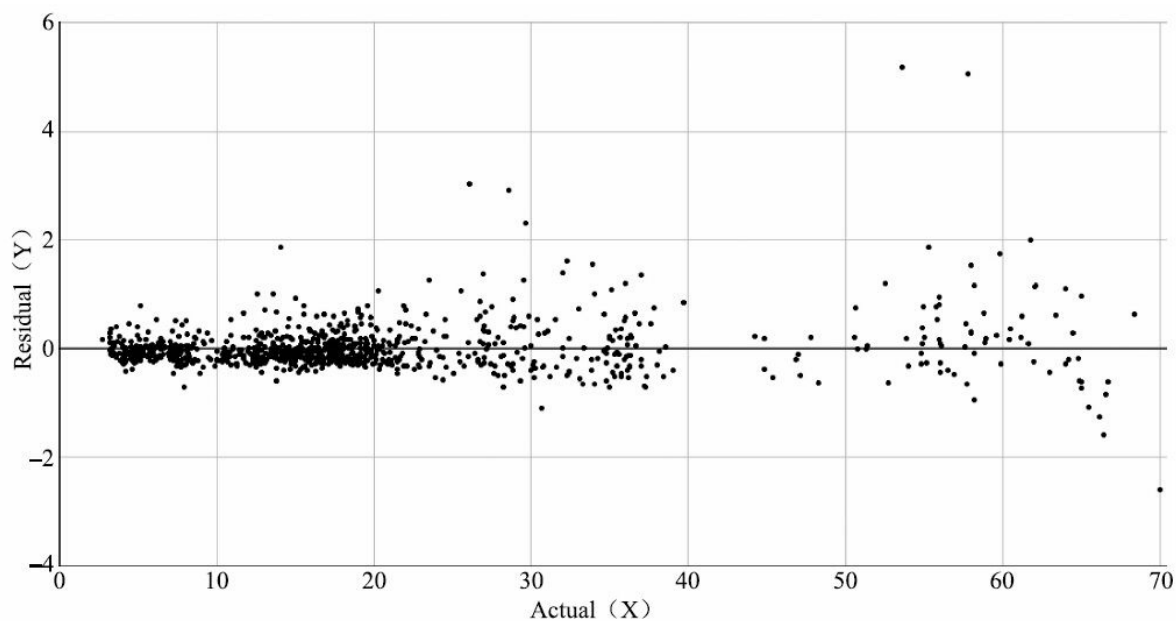


图 8-31

图8-32也给出了该模型对股票A每日收盘价历史数据的回归拟合残差散点图。情况与其他残差散点图区别不太大，留给读者自己观察、参考。

我们的许多读者可能都是很优秀的股票专家，以上仅做了一些简单分析，深入的探讨留给读者，大家一定自有见解。

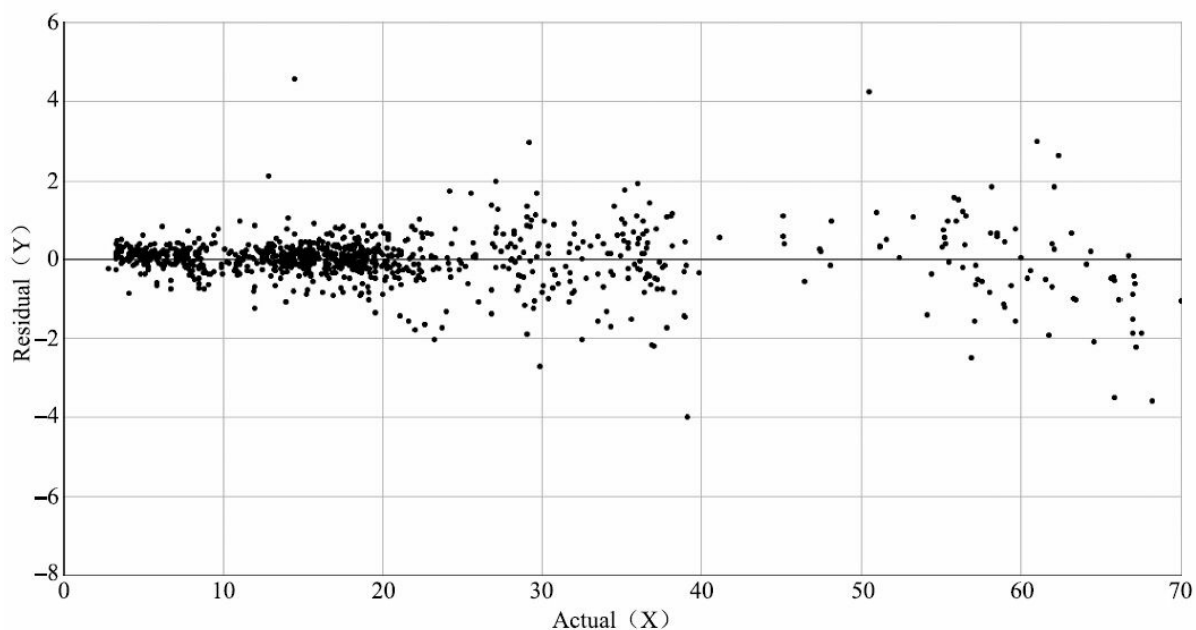


图 8-32

2.股票A周预测

与前面的思路类似，我们按照每周为单位，将根据上一周交易的最高价和本周的开盘价来预测该股票本周交易的最高价，目的是为股民本周的交易提供预测和参考，并在时间上为股民提供更好的可操作性。

同样，我们将股票A本周交易与上一周交易的数据合在一起，包括本周开盘价、本周最高价、本周最低价、本周收盘价、上一周开盘价、上一周最高价、上一周最低价、上一周收盘价，上一周数据的Log变换数据，以及一些时间维度数据（如年、月、日）。代码如下：

```
SELECT a.open_id as case_id, a.year, a.week_year,b.datetime, LN(b.open) open,
a.high, a.low,
LN(c.high) last_high, LN(c.low) last_low, LN(d.close) last_close, LN(d.open)
last_day_open,
```

```

LN(d.high) last_day_high, LN(d.low) last_day_low
FROM
(SELECT year, week_year , MIN(case_id) open_id, MAX(high) high, MIN(low) low
FROM "v_stock_day_faiyao" GROUP BY year, week_year ) a,
" v_stock_day_faiyao " b,
(SELECT year, week_year , MAX(case_id) close_id, MAX(high) high, MIN(low) low
FROM " v_stock_day_faiyao " GROUP BY year, week_year ) c,
" v_stock_day_faiyao " d
WHERE a.open_id = b.case_id
AND a.open_id = c.close_id+1
AND c.close_id = d.case_id;

```

相关数据如图8-33所示。

Structure	Data	View Lineage										
Fetch Size	100	Fetch Next	Refresh									
CASE_ID	YEAR	WEEK_YEAR	...	OPEN	HIGH	LOW	LAST_HIGH	LAST_LOW	LAST_CLO...	LAST_DAY...	LAST_DAY...	LAST_DAY...
1581	2000	26	...	2.678	14.8	14.33	2.688	2.657	2.676	2.662	2.681	2.66
3578	2008	52	...	3.536	34.6	32.58	3.564	3.481	3.531	3.555	3.561	3.52
3585	2009	1	...	3.535	34.5	32.8	3.552	3.497	3.538	3.546	3.55	3.522
3979	2010	34	...	4.045	57.8	56	4.08	4.005	4.049	4.052	4.059	4.033
1606	2000	31	...	2.717	15.72	14.82	2.727	2.646	2.711	2.695	2.727	2.695
1611	2000	32	...	2.708	15.75	14.72	2.755	2.696	2.707	2.724	2.733	2.701
1651	2000	41	...	2.759	16.65	15.61	2.824	2.722	2.762	2.734	2.763	2.731
3923	2010	22	...	4.197	69.01	65.8	4.202	4.103	4.188	4.182	4.19	4.171
3934	2010	25	...	4.135	66	61.9	4.234	4.14	4.142	4.215	4.217	4.14
3242	2007	30	...	3.526	36.8	33.98	3.543	3.497	3.526	3.523	3.535	3.503
3252	2007	32	...	3.615	39.04	36	3.624	3.55	3.598	3.574	3.61	3.57
3272	2007	36	...	3.663	40	36.5	3.664	3.598	3.653	3.626	3.664	3.608
3277	2007	37	...	3.589	37.1	35.21	3.689	3.597	3.599	3.626	3.642	3.597
2181	2002	53	...	2.845	17.38	17.18	2.86	2.824	2.846	2.851	2.852	2.836
3613	2009	8	...	3.643	39.44	37.02	3.733	3.592	3.653	3.679	3.733	3.639
3618	2009	9	...	3.628	38.43	35.6	3.675	3.611	3.627	3.638	3.644	3.621
2415	2003	52	...	2.713	15.48	15.01	2.719	2.695	2.713	2.709	2.719	2.701
1275	1999	11	...	2.438	12.58	11.35	2.454	2.407	2.441	2.408	2.447	2.408
4026	2010	45	...	4.186	71	63.39	4.186	4.143	4.186	4.156	4.186	4.149
1324	1999	21	...	2.317	10.95	9.8	2.322	2.169	2.321	2.282	2.322	2.251
1338	1999	24	...	2.434	12.8	11.1	2.46	2.389	2.434	2.451	2.451	2.418
1735	2001	8	...	2.818	16.87	15.8	2.834	2.773	2.809	2.821	2.821	2.785
4046	2010	49	...	4.156	63.97	61	4.22	4.134	4.145	4.151	4.167	4.139
4051	2010	50	...	4.111	63.66	61.03	4.158	4.111	4.111	4.13	4.131	4.111
4056	2010	51	...	4.127	63.02	58.88	4.154	4.111	4.131	4.128	4.146	4.122
2923	2006	7	...	3.05	23.76	21.11	3.054	2.99	3.05	3.016	3.051	3.016
2928	2006	8	...	3.139	23.89	22.39	3.168	3.05	3.118	3.129	3.14	3.11
2938	2006	10	...	3.207	25.17	23.4	3.257	3.116	3.209	3.205	3.229	3.193
3355	2008	2	...	3.611	38.05	35.68	3.642	3.55	3.616	3.622	3.624	3.59
2992	2006	28	...	3.434	31.18	17.7	3.463	3.334	3.43	3.45	3.463	3.428
3395	2008	11	...	3.611	38.45	31.8	3.731	3.618	3.623	3.675	3.675	3.618
2672	2005	3	...	2.773	16.9	16	2.82	2.715	2.794	2.82	2.82	2.779
2687	2005	7	...	2.867	17.66	17.18	2.867	2.79	2.867	2.833	2.867	2.833

图 8-33

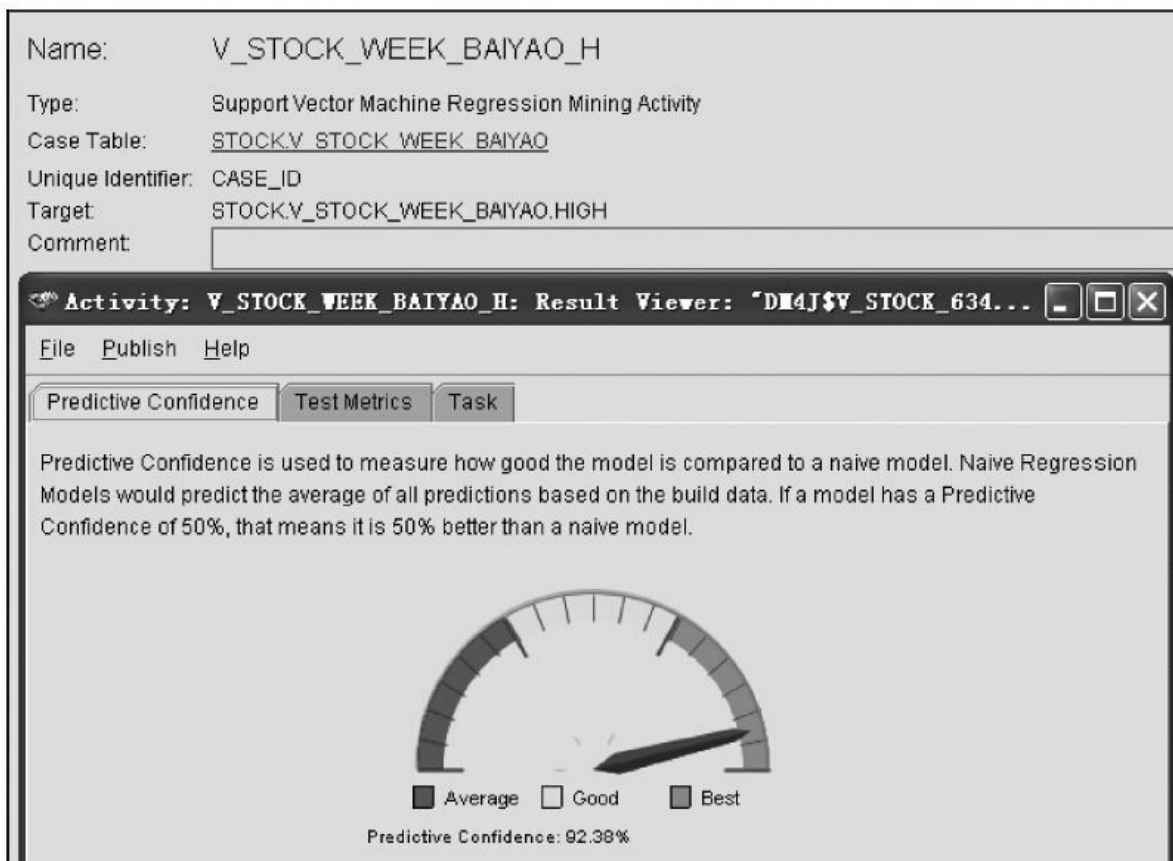


图 8-34

经过算法运行，我们得到了一个回归预测模型。该模型显示预测置信度为92.38%，如图8-34所示，模型性能尚属优良，但较每日最高价的回归模型低了几个百分点。不要小看这几个百分点，它在股票价格预测中的影响是很大的。在实际应用中，如果说日最高价模型尚有实际应用的可能的话，那么周最高价模型基本就不能应用于实际股票操作中去了，因为误差对于使用者来说是无法容忍的。这个场景与在医学科研上的情况大不一样。在医学应用中，我们得到的模型性能有可能达不到股票数据预测模型的水准，但在某些场合下，对于医学研究的需求来说可能足够了。

Activity: V_STOCK_WEEK_BAIYAO_H:	
File Publish Help	
Predictive Confidence	Test Metrics Task
Mean Absolute Error	0.7382178502
Mean Actual Value	20.8195211268
Mean Predicted Value	20.8019177876
Root Mean Square Error	1.1548883449

图8-35

是该预测模型的测试结果，图8-36是该模型的建模参数和预测因子，图8-37是该预测模型的预测残差图，留给读者自行参考。

Activity: V_STOCK_WEEK_BAIYAO_H: Result Viewer:

File Help

ResultsBuild SettingsTask

Type:Regression

Algorithm:Support Vector Machine

Target Attribute:HIGH

Algorithm Settings

Kernel Function:Gaussian

Standard Deviation:0.1215995530

Kernel Cache Size:50000000.0000000000

Complexity Factor:0.6347846035

Tolerance:0.0010000000

Epsilon:0.0080718976

Active Learning:yes

Attributes

Name	Data Type	Mining Type
DATETIME	NUMBER	Numerical
LAST_CLOSE	NUMBER	Numerical
LAST_HIGH	NUMBER	Numerical
OPEN	NUMBER	Numerical

图 8-36

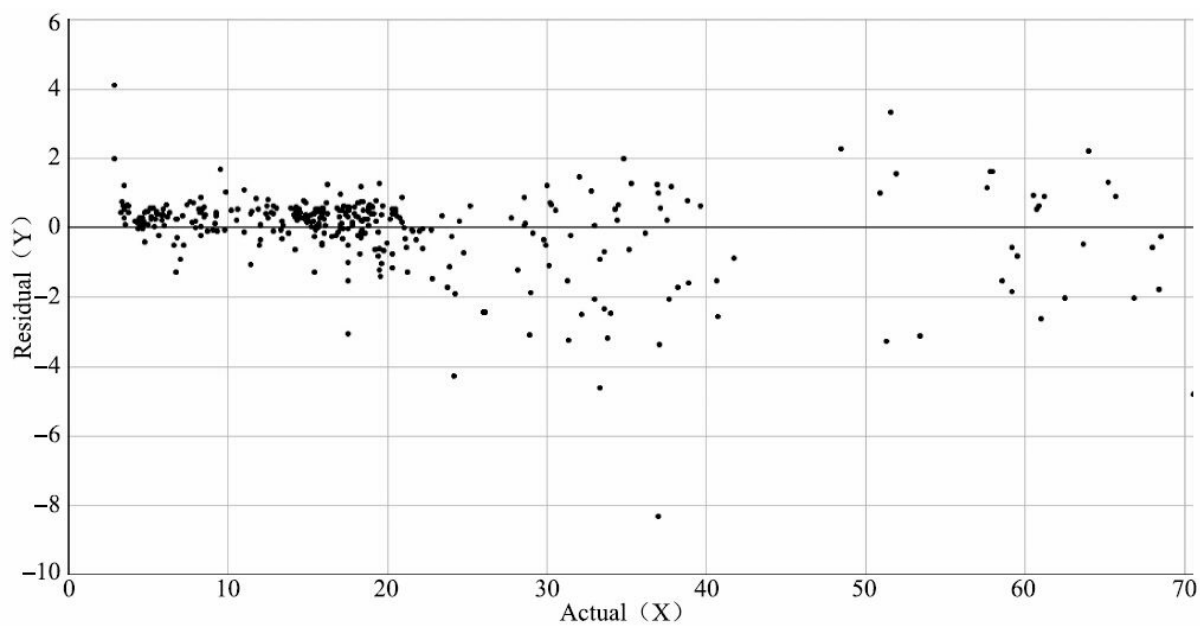


图 8-37

图8-38~图8-41是同一只股票每周最低价的回归预测模型，这里不再赘述，留给读者自行参考。

Residual Plot Task		
Apply Output Table:		
DMR\$CASE_ID	HIGH	PREDICTION
482	4.13	4.3281
338	4.34	4.4842
477	4.35	4.331
342	4.38	4.4534
123	4.4	4.6972
609	4.5	4.6471
614	4.6	4.5899
293	4.6	4.6816
273	4.62	4.9272
303	4.7	4.9076
308	4.75	5.2066
288	4.77	4.381
333	4.83	5.3129
268	4.95	5.4402
619	4.98	5.0959
445	5.1	5.6607
234	5.15	5.4649
440	5.18	5.4304
...

图 8-38



图 8-39

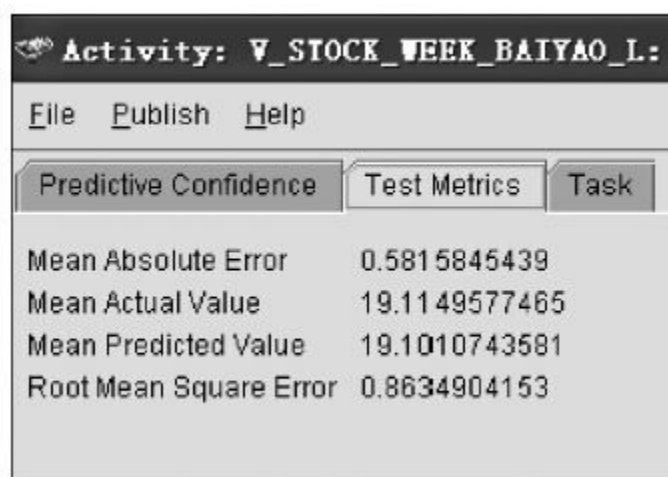


图 8-40

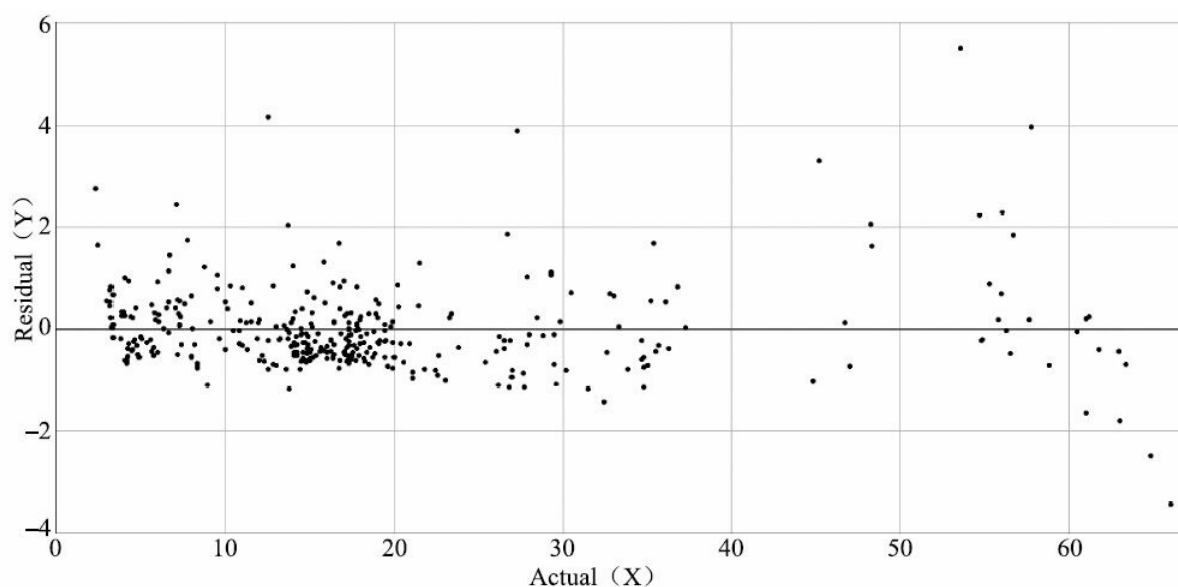


图 8-41

3.股票B周预测

我们再来观察另外一只股票的特征。我们得到了一个股票B周最高价回归预测模型，用于与股票A周最高价回归预测模型进行比较。该模型显示预测置信度为87.43%，模型性能尚属良好，但较股票A周最高价回归预测模型低了好几个百分点。如图8-42~图8-44所示。



图 8-42

Activity: V_STOCK_WEEK_PZ_H: Result	
File Publish Help	
Predictive Confidence	Test Metrics Task
Mean Absolute Error	0.3430351784
Mean Actual Value	9.3912323944
Mean Predicted Value	9.3618511872
Root Mean Square Error	0.4906346932

图 8-43

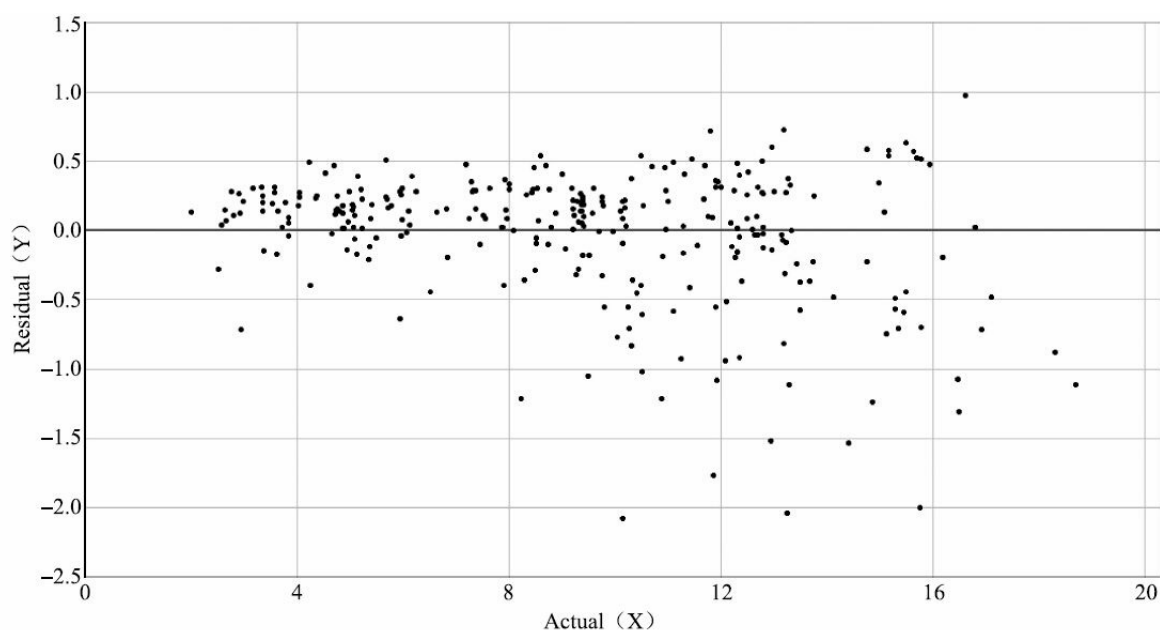


图8-44 周高点

从图中，我们可以看出该只股票周高点的表现大致分为两个阶段，一个是横坐标从0至8之间，股票B交易数量较稳定，而且模型拟合得比较好，几乎没有明显的离群点；一个是横坐标从8至20之间，股票B交易数量（散点数）较前一阶段变化不大，但图形明显发散，模型拟合较差，而且明显的离群点显著增加。

同时，我们也运行了股票B周最低价的回归预测模型，模型性能与股票B周最高价的模型相当，为了节省篇幅在此不再给出，仅给出了一个股票B周最低价的回归预测模型的残差散点图。

从图8-45中，我们可以看出该只股票周低点的表现与周高点相差较大，主要表现在：图形总体比较收敛，但有逐步发散的趋势，明显的离群点不太多，但也有增多的趋势。

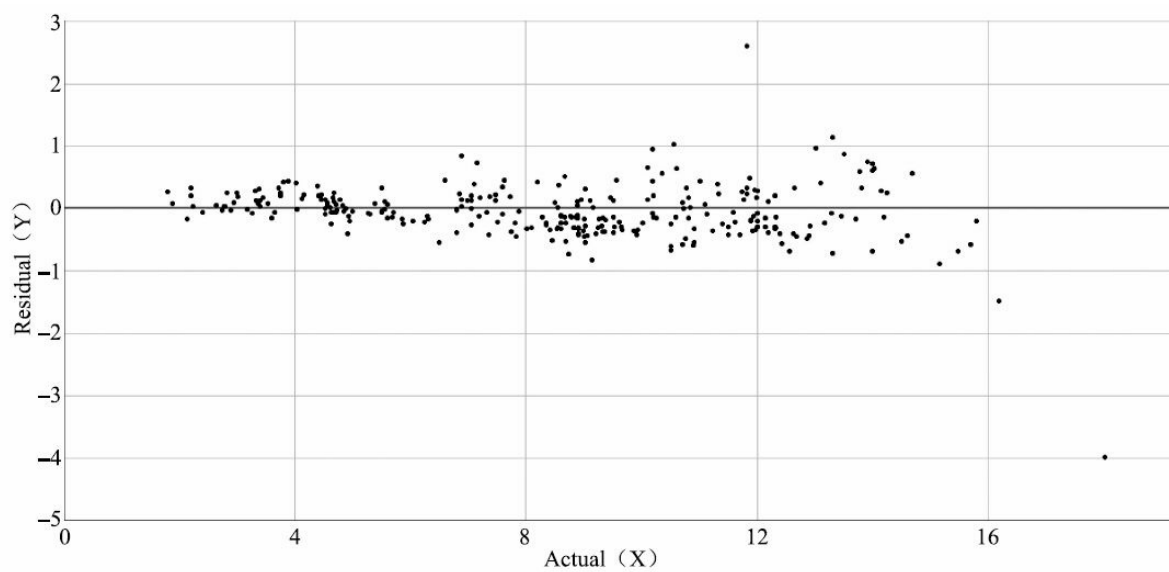


图8-45 周低点

读者还可以自行将该股票B的高价模型和低价模型与股票A的高价模型和低价模型进行比较，从中发现一些有意义的现象。

8.2 应用IBM SPSS Modeler

在这一节里，我们将介绍使用IBM公司的数据挖掘工具SPSS Modeler（以前称为SPSS Clementine）软件开展数据挖掘应用的案例。与数据挖掘引擎嵌入到数据库的Oracle Data Miner不同，SPSS Modeler是数据挖掘引擎独立于数据库之外的工具软件，它需要将要挖掘的数据读入到SPSS Modeler中，或者由SPSS Modeler连接到数据库，再将数据库中的数据读入到SPSS Modeler进行处理、运算，然后将结果输出到文件中。本节我们首先对SPSS Modeler做一些基本的介绍，然后展示SPSS Modeler作为独立工具的应用，主要讨论了线性模型和关联规则的应用，最后讲解了SPSS Modeler与应用系统的联合应用，主要探讨了分类模型和回归预测的应用。

8.2.1 IBM SPSS Modeler介绍

IBM SPSS Modeler原名SPSS Clementine（SPSS公司产品），2009年被IBM收购。SPSS Modeler是一组应用于商业数据挖掘活动的工具，通过这些工具用户可以快速、简便地建立商业预测性模型，并应用这些模型改进业务决策过程。按照IBM公司的提示，SPSS Modeler有以下几种优势：

- 支持不同层次用户

SPSS Modeler的每种算法和模型都有自动选择模型和专家模型两种工作模式，方便不同水平的用户使用。

- 挖掘流程易于管理

SPSS Modeler建立的所有数据流、模型等均可保存在数据挖掘项目中，保证了数据挖掘项目的可重用性和共享性。

- 节省时间

SPSS Modeler中有三种自动建模器，可以快速识别更好的模型，综合多种模型的不同性能得到最佳结果。

SPSS Modeler中包含了多种数据挖掘算法，主要分为以下几方面：

- 1) 维度简约与异常侦测

数据挖掘应用中常常包含上百个变量，许多应用通常是在高维度下研究和分析数据，这样就需要在建模的过程中花费大量的时间分析这些变量。“特性选择”模型能够排除对决策影响不大的变量；“主成分/因子分析”算法是将数据维度简化的另外一种方法。“异常侦测”算法可以在数据分析探索阶段侦测不寻常的事物。

2) 决策树模型

决策树模型可以基于一组决策规则来预测或分类未来的观测值，并构建分类规则。SPSS Modeler中提供了多种算法支持决策树分类：

- 决策列表模型：目的在于找出一组有独立行为模式的人群，例如高概率购买某种商品的人群；
- 神经网络模型：一般函数预测器，用于分类模型；
- 最近相邻元素模型：根据观测值与其他观测值的类似程度对观测值进行分类的方法，可用于离散变量或连续变量的分类。

3) 聚类模型

SPSS Modeler聚类模型主要将相似的数据划归为一组，并按照所属的组来标记数据。

4) 关联规则模型

SPSS Modeler可以通过Apriori、GRI及CARMA三种算法来发现数

据中的关联规则。

5) 时间序列模型

时间序列算法集成了指数平滑、单变量ARIMA和多变量ARIMA算法来预测时间序列数据。

6) 其他统计模型

SPSS Modeler还包括了线性回归、逻辑回归、广义线性模型、判别分析、Cox回归、SVM（Support Vector Machines）支持向量机、贝叶斯网络等多种算法。

8.2.2 SPSS Modeler独立应用

这里讲的SPSS Modeler独立应用是相对于ODM基于数据挖掘API开发的数据挖掘应用而言，实际上，它就是SPSS Modeler常用的一种使用方式，即用户直接使用SPSS Modeler中提供的各种算法对数据进行处理和挖掘，基本上不涉及算法或数据流程的二次开发。相对于这节的SPSS Modeler独立应用，后面的第8.2.3节中会涉及结合SPSS Modeler二次开发的数据挖掘应用。

8.2.2.1 线性模型：实验室指标的初探

病人到医院就诊时，通常需要化验一些生理指标（如实验室指标）来协助医生的诊断。这些通过化验得到的生理指标之间是否存在一定关系呢？医学家会研究实验室指标之间的关系，通过研究某种指标是否受其他因素（指标）的影响，及受到多大影响，来揭示生理指标之间的内在联系，从而发现疾病产生的根源，以便开展更广泛、深入的临床研究。解决这一类问题，数据挖掘常用的方法包括线性回归模型、非线性模型和相关分析等方法。

本例采集了545例某种疾病患者的11项化验指标，研究该疾病患者的血红蛋白是否受其他实验室指标的影响，如果存在影响，试图建立简单线性模型，对化验指标进行初步探索。（与本书中的大多数案例类似，本例的目的仅是为了展示数据挖掘应用的方法和技术，因此对指标的选择并不是基于医学专业角度考虑，并且也不做医学专业上

的讨论。)

与上一节类似，在首次应用SPSS Modeler时，我们会简要附带SPSS Modeler的应用流程，以便读者在了解数据挖掘案例的同时，对SPSS Modeler的功能及其应用也有一些了解。

首先启动SPSS Modeler（此处版本为15）。本例中的数据源为Excel文件，因此在SPSS Modeler的源中添加Excel文件节点，如图8-46所示，选择文件类型为Excel97-2003格式，单击导入文件旁的“...”按钮，可以在文件中选择数据格式为.xls的文件，以便读取数据源。

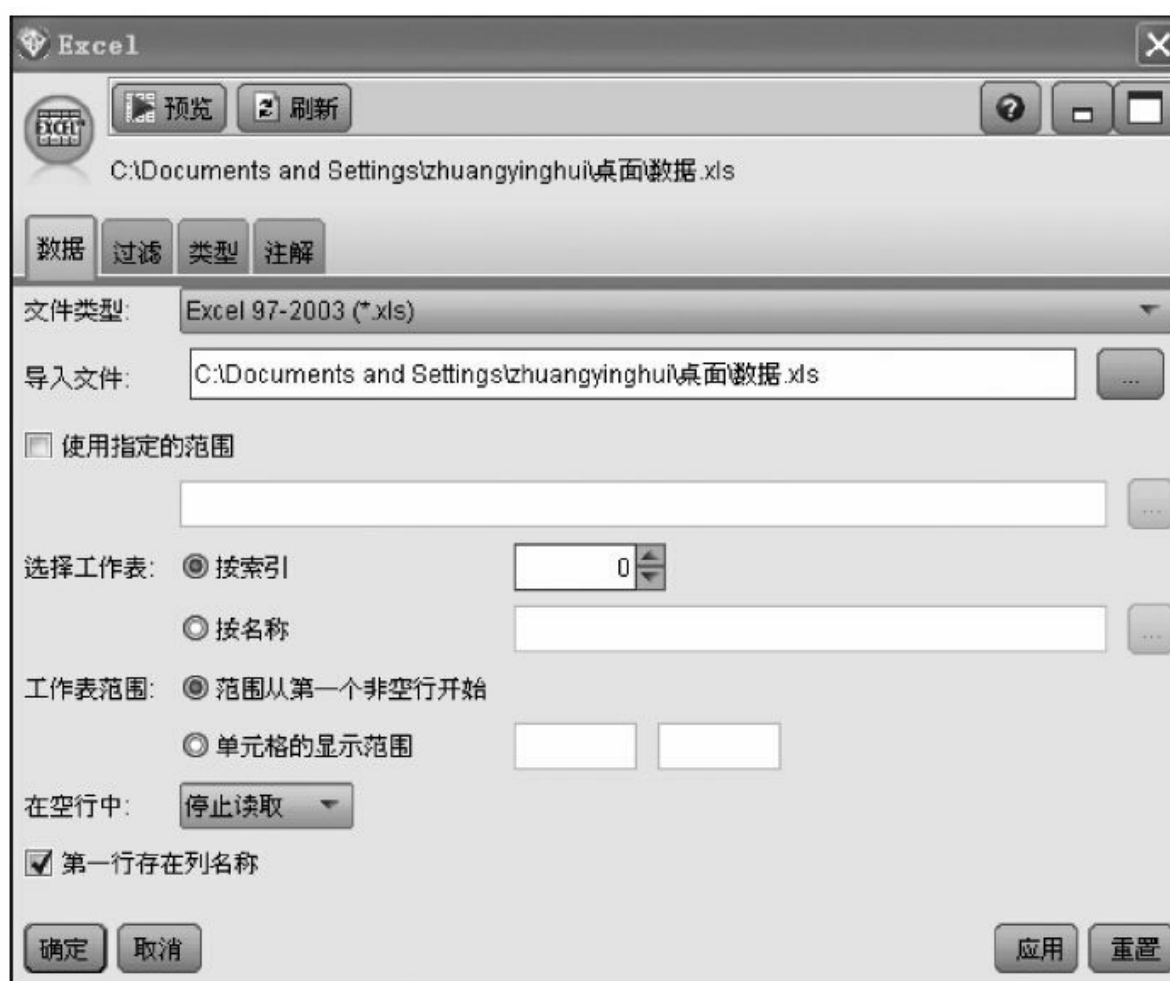


图8-46 载入Excel文件

该数据源中包括该种病人的血红蛋白、红细胞压积、单核细胞、白细胞、红细胞、单核细胞百分比、嗜碱细胞百分比、淋巴细胞、平均血小板体积、血小板分布宽度、平均血红蛋白浓度，总计11项化验数据。点击上图中的“类型”选项页，可以看到这11项指标的名称和类型，如图8-47所示。



图8-47 数据名称和类型

单击“读取值”按钮，就可以从图8-47中看到11项指标的类型、值、缺失情况和角色等详细情况。点击“预览”按钮，还可以看到原始

数据。在建立线性回归模型之前，要观察血红蛋白与其余十项指标的散点图，初步探索血红蛋白与其他指标在图形上是否呈现线性关系。以血红蛋白和淋巴细胞为例，绘制其散点图。点击图形节点中的散点图节点，将其添加到数据流中，与源节点连接。双击该节点，设定X字段为淋巴细胞，Y字段为血红蛋白，如图8-48所示。



图 8-48

点击“运行”按钮，得到血红蛋白和淋巴细胞的散点图，如图8-49所示。

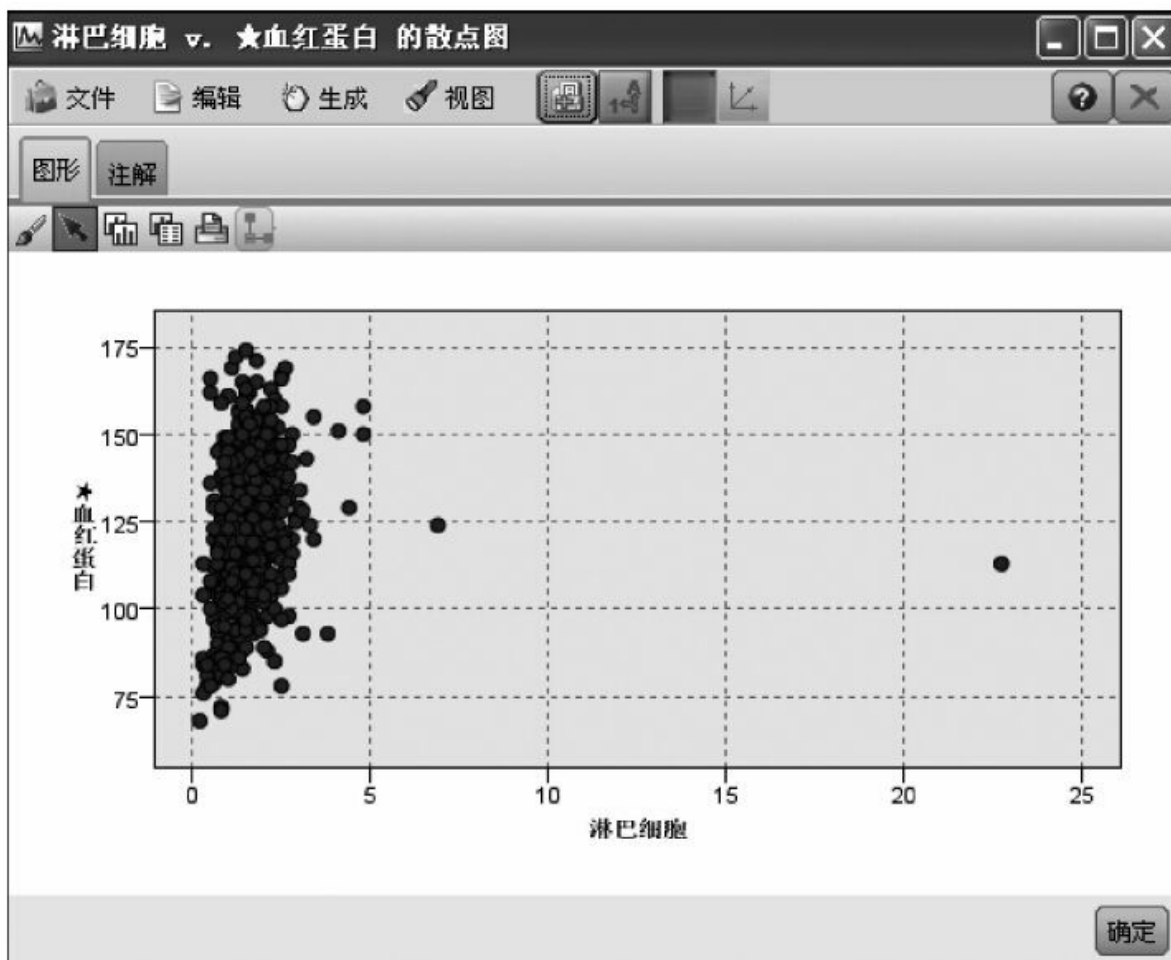


图 8-49

从散点图中可以看出，这两种化验指标之间不存在明显的线性关系，无法建立线性回归模型。在此，我们可以假设血红蛋白不是与这十项化验指标都存在明显的线性关系，由此在选择模型方法时，不选择“进入法”，而选择“逐步法”。将字段选项中的类型节点添加到数据流中，在其选项中选择输入和输出变量。如图8-50所示，选择血红蛋白为目标变量、其他变量为输入变量，表示将血红蛋白作为因变量，其余指标作为自变量。



图 8-50

选择建模中的回归模型节点，添加到流的末端。该模型包括四种方法：“进入法”、“逐步法”、“前进法”和“后退法”。上文中的散点图已说明血红蛋白不是和其余所有的化验指标呈明显线性相关，因此在方法上选择“逐步法”，如图8-51所示。



图 8-51

点击“运行”按钮，得到线性模型结果，如图8-52所示。

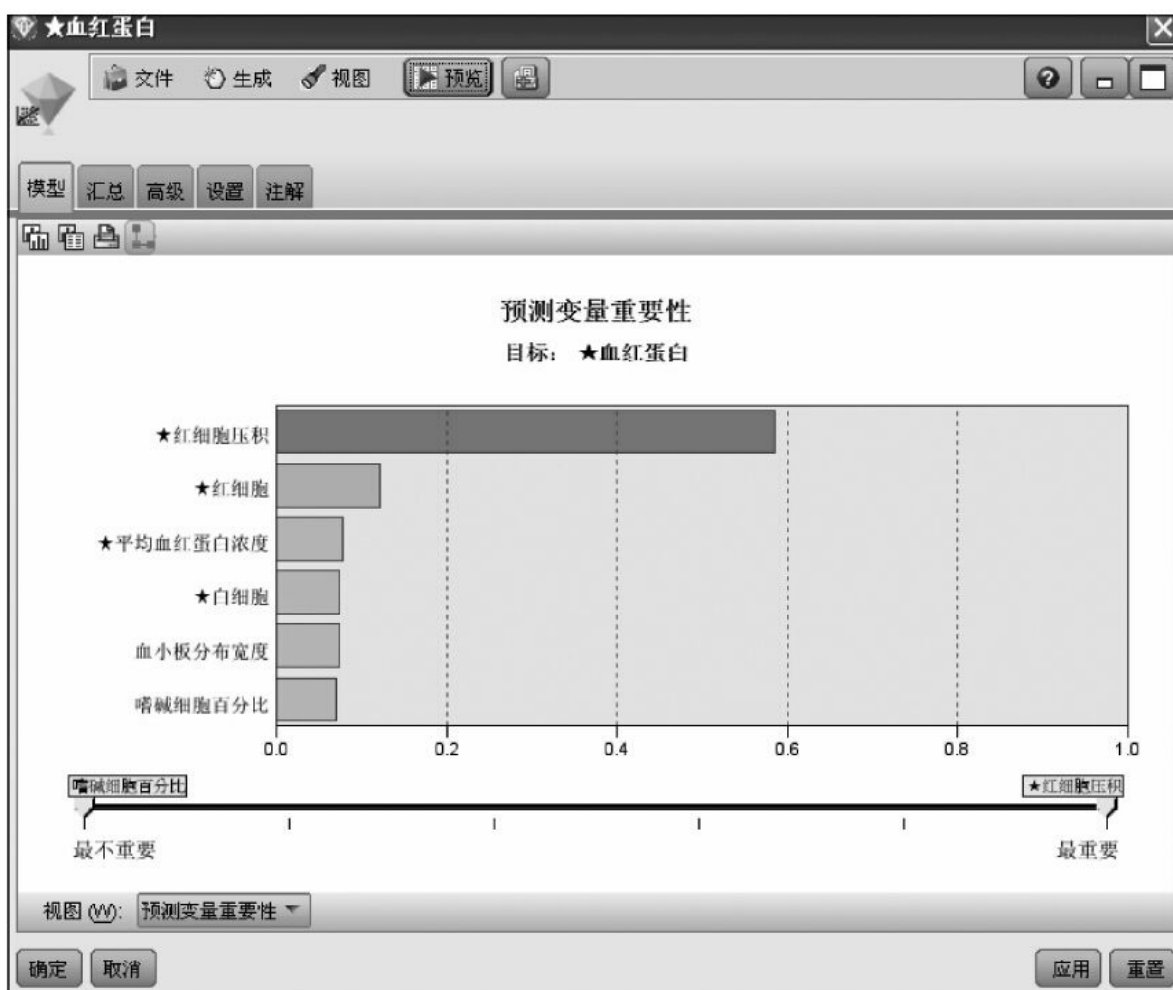


图 8-52

图8-52中的结果是一个评价自变量重要性的示意图，柱状图形越长说明该变量的重要性越高。本例中，红细胞积压的重要性最高，而嗜碱细胞百分比的重要性最低，说明该模型中红细胞积压对血红蛋白的影响最大，嗜碱细胞百分比对血红蛋白的影响最小。然后单击“汇总”按钮，我们就可以观察该回归模型的详细信息，如图8-53所示。



图 8-53

点击“展开全部”，其中的“分析”栏解释了回归模型的总方程，自变量后面是每项指标在回归模型方程中的系数，例如红细胞压积的系数为306，常数项为-117.4。“字段”栏解释了该回归模型的因变量是血红蛋白，自变量是红细胞积压、白细胞、红细胞、嗜碱细胞百分比、血小板分布宽度和平均血红蛋白浓度。通过建立回归模型，我们从十个实验室指标中筛选出六个与血红蛋白相关的指标。该回归方程中红细胞积压的系数说明了，在其他指标不变的情况下，每当红细胞积压平均变化306个单位，血红蛋白就平均变化一个单位，其他指标系数的解释方法与此类似。因为本例中采用的“逐步法”，所以图中显示的

回归模型方程系数均通过假设检验，证明此线性回归模型是成立的。

SPSS Modeler提供了一个称为“Stream”的流程，将数据挖掘活动的整个过程串联为一个整体，并自动按“Stream”的顺序执行。本例数据挖掘的流程如图8-54所示。

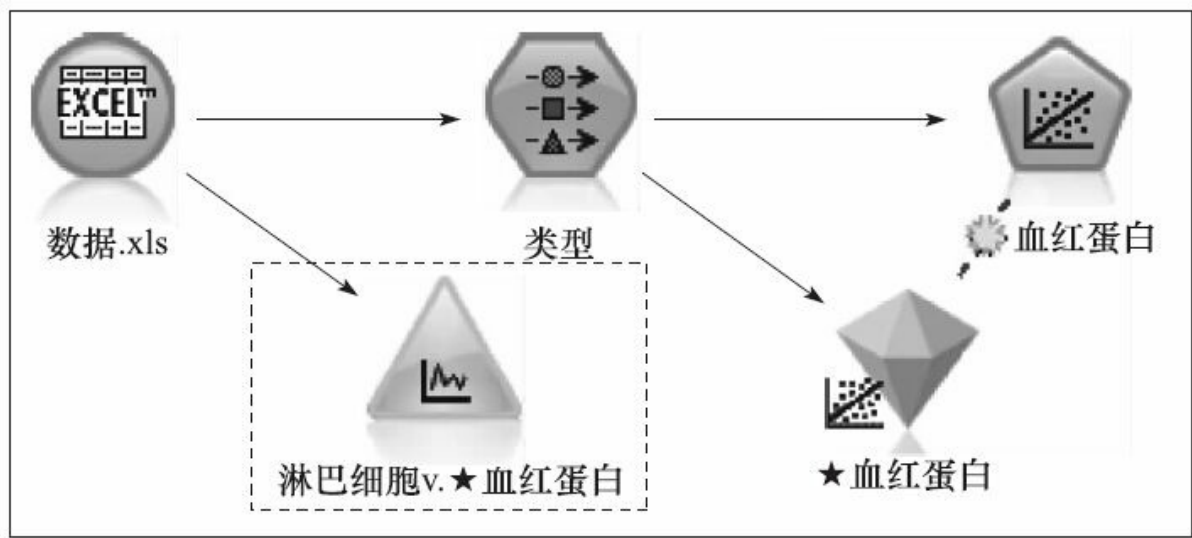


图8-54 流程图

该图显示了从数据源的读取，到输入变量和输出变量的确定，再到线性回归模型的运行，以及模型运行结果的展示等几个环节的全部过程。

综上对11项化验指标的线性回归分析，我们发现该疾病患者的血红蛋白确实受其他化验指标的影响，这种影响体现为一种线性回归模型的关系。

8.2.2.2 关联规则：多种并发症的关联探索

很多病人在疾病发作的时候，可能不只是由一种疾病引起的，而是两个或更多的疾病相互影响引起的症状，因为人体本身就是一个由许多子系统组成的一个生理系统。“牵一发而动全身”，所以医学临床研究中会涉及到很多与疾病的并发症相关的研究内容。我们应用数据挖掘中的关联分析方法，可以帮助医学家发现一些隐藏在大量数据中对疾病并发症有意义的联系，并试图用关联规则的形式将隐藏规律“显式”地表达出来。

在本例中，我们收集了包含某种疾病的上百种并发症的数据，并存储在数据库表中。我们用字段“PATIENT_INSTANCE_ID”作为数据的唯一键，用字段“PATIENT_SYMPTOM”存储病人的并发症信息。不同于上个案例的Excel数据源，本例的数据源为大型关系数据库Oracle。因此，我们需要为SPSS Modeler配置ODBC开放数据库互连网关。

数据连接配置好后，在SPSS Modeler中选择源节点中的数据库子节点，数据源为之前配置成功的数据库名称。将类型节点附加到源节点，由于关联规则算法没有目标变量，所以在类型设置中的“角色”设置两个字段均为输入。

上文中已经介绍了关联规则的三种算法，本例中应用的是其中的Apriori算法。设置模型中Apriori算法的有关参数，其中使用“定制设置”，格式为“事物处理格式”，“ID”为“PATIENT_INSTANCE_ID”字段，“内容”为“PATIENT_SYMPTOM”字段，如图8-55所示。



图 8-55

点击“运行”按钮后，得到如图8-56的结果。结果显示，置信度最高的规则为：前件为冠心病和不稳定心绞痛，后件为心功能2级，置信度达87.5%，其支持度为12.7%，也是规则中最高的支持度。此规则说明病人在患有冠心病和不稳定心绞痛的情况下，那么同时患有心功能2级的可能性也较大，而且这样的现象不在少数，因为12.7%的支持度在大量的疾病并发现象中比例已经不小了。

<div> <div>文件 生成 预览</div> <div>?</div> </div>			
<div> <div>模型 设置 汇总 注解</div> </div>			
<div> <div>按以下内容进行排序: 置信度 %</div> <div>3 属于 3</div> </div>			
后项	前项	支持度 %	置信度 %
心功能2级 = T	冠心病-不稳定心绞痛 = T	12.698	87.5
冠心病 = T	不稳定心绞痛 = T	11.111	85.714
高血压病3级 = T	心功能2级 = T	11.111	85.714
	冠心病 = T		

图 8-56

<div> <div>模型 设置 汇总 注解</div> </div>	
<div> <div>折叠全部 展开全部</div> </div>	
<div> <div>分析</div> <div> <div>规则数: 3</div> <div>有效事务数: 63</div> <div>最小支持度: 11.111%</div> <div>最大支持度: 12.698%</div> <div>最小置信度: 85.714%</div> <div>最大置信度: 87.5%</div> <div>最小提升: 2.0%</div> <div>最大提升: 2.625%</div> <div>最小部署能力: 1.587%</div> <div>最高部署能力: 1.587%</div> <div>最低规则支持度: 9.524%</div> <div>最高规则支持度: 11.111%</div> </div> </div>	
<div> <div>字段</div> <div> <div>ID</div> <div>PATIENT_INSTANCE_ID</div> </div> </div>	
<div> <div>内容</div> <div>PATIENT_SYMPTOM</div> </div>	

图 8-57

点击输出结果中的“汇总”选项页，我们可以勘察该模型的详细信息。图8-57显示，按照预先设定的规则阈值，该模型中包含三个有效的规则，规则的置信度在85.714%至87.5%之间，这说明规则的可信度是比较高的。规则的支持度在11.111%至12.698%之间。Lift提升指标为2.0%到2.625%之间，说明规则是随机的较强规则。

通过本例的数据分析，我们从大量数据中发现了对该种疾病并发症有意义的一些联系，并用关联规则的形式将这些联系显式地表达出来，为医学专家的相关研究提供了一些有价值的线索。

8.2.3 SPSS Modeler与应用系统的联合应用

作为一个独立的数据挖掘第三方通用工具，SPSS Modeler在大多数应用场景中是单独应用的，正如前面的案例展示的那样。但是，我们会发现这样的应用有时还存在着一些瓶颈。例如在数据准备方面，尤其是对于关系复杂的存储在大型数据库中的数据源，数据挖掘通用工具很难提供灵活的、强大的数据筛选功能，而且，有时我们希望通用工具与面向业务的应用软件能紧密地结合起来。因此，我们需要将二者进行某种程度的集成，并联合应用，以便互相取长补短，更加有效地服务于业务。

本例中，我们将IMRS系统与SPSS Modeler软件做了尝试性的集成应用。如图8-58所示，我们在IMRS系统中开发了“扩展挖掘”模块，为SPSS Modeler算法模型提供数据接口。在IMRS中，我们可以进行四种算法的应用：分类、预测、异常侦测和影响分析。选择要分析的病人人群、相关的化验指标项、病人年龄范围几个条件后，IMRS会在数据库中自动生成相应的数据视图，作为SPSS Modeler分析的数据源。

8.2.3.1 分类模型：疾病分类

根据实验室指标来判断患者是否患有某种疾病是临床常用的手段。分类算法是一种自主学习方法，它能够根据输入的数据，来确定对象属于哪个预定义的目标类。比如说，根据核磁共振扫描结果区分

肿瘤是恶性的还是良性的。本例中，将肺炎和恶性肿瘤术后化疗两种疾病作为研究对象。医学上，有人认为肿瘤也是一种炎症。我们通过本例中的探索性研究，来探索这两种疾病是否可以依据某些化验指标将其有效地区分开。

如上文所述，从IMRS系统中选择数据。我们采集了肺炎和恶性肿瘤术后化疗的生化全项（共34项）数据，总计1620条，存储在数据库中的v_dm_lab视图中。打开SPSS Modeler，选择源为数据库节点，选择已配置成功的数据库名称，查找视图v_dm_lab，并读取数据。

从我们得到的1620条数据的视图结构中看出，每一个病人的生化全项的多个化验细项数据，存储在有同一个TEST_ID的多条数据行中。我们需要依据病人的生化全项的有关指标变量对病人进行分类，按照SPSS Modeler提供的功能，我们需要对数据结构进行变换，即数据表的行列转换。在字段选项中选择重新结构化节点，“可用字段”为report_item_name（化验细项），将其选入到“创建重新构建的字段”框中，“值字段”为test_result（化验结果），这样我们就将所有的检验细项转化为新的字段（列）了。如图8-59所示。

分类

预测

异常侦测

影响分析

条件选择

重置

Add

Remove

* 选择疾病分类：

(必选)

Add

Remove

* 选择化验大类：

(必选)

* 选择年龄段：

下限

-上限

(可选)

生成

图 8-58



图 8-59

化验细项转化为列后，数据表的数量并没发生变化，只是增加了字段数，一个病人仍有34条数据，需将这些数据整合成一条数据。点击记录选项中的“汇总”，“关键字段”为TEST_ID，“汇总字段”为所有的化验细项，默认输出模式为最大值，即根据最大值进行汇总，保存设置。

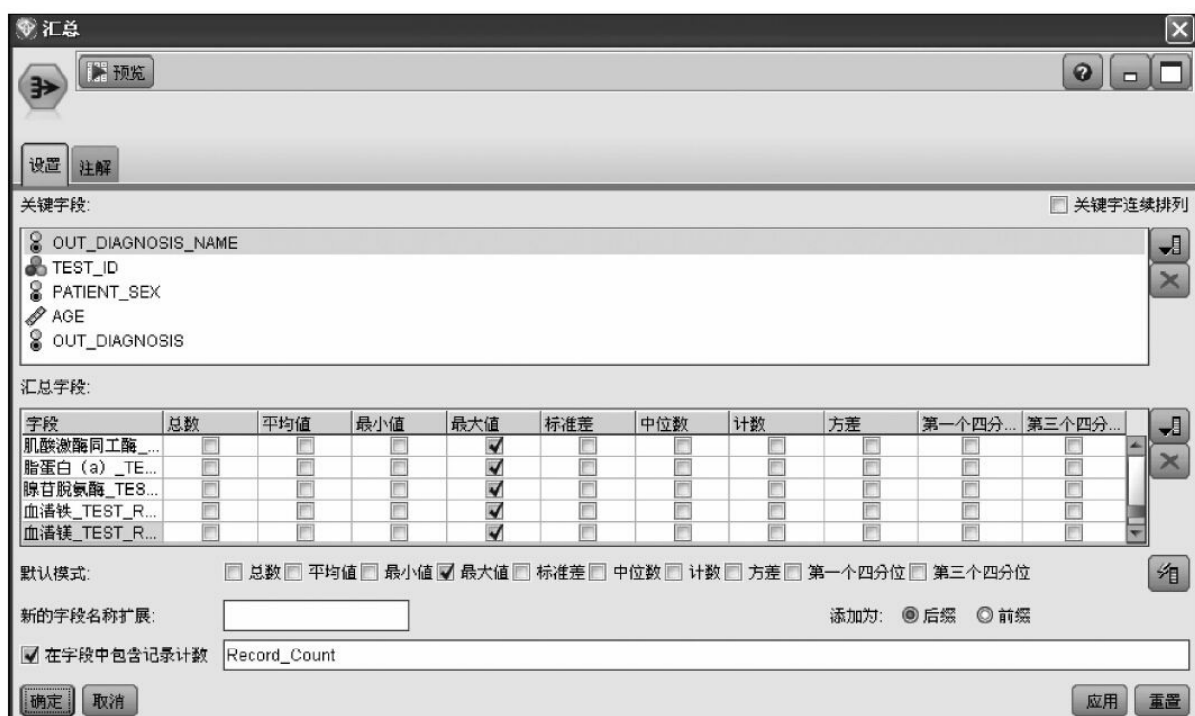


图 8-60

为了确保字段转化和汇总的正确性，在输出选项中选择表节点，将转换后的结果输出。同样也可在图8-60中点击“预览”按钮，达到同样的效果。

数据转换后，对字段做过滤处理。将字段选择中的过滤节点添加到流中，把TEST_ID、OUT_DIAGNOSIS和Record_Count过滤。再将类型节点添加到流中，选择疾病为输出变量，年龄、性别和生化全项的所有细项为输入变量。

选择模型中的自动分类器节点，添加到流的末端。自动分类器是SPSS Modeler算法集成器的一种，其中包含了C5、Logistics、决策列表、贝叶斯网络、判别式、C&R树、Quest、CHAID、神经网络、KNN和SVM，总计11种算法。其默认选项包括除去KNN和SVM的九

种算法，可以根据情况自主选择。本例中为默认选择。字段的设置选择中，可以选择模型的排序方式、是否计算预测变量重要性等，本例为默认选项。点击“运行”按钮，得到的结果如图8-61所示。

最终有三种算法可供选择，包括Quest、C&R树和CHAID。模型精确度最高的是CHAID算法，总体精确度为97.619%，其图形也说明了该模型将两种疾病划分正确的效果最好。双击该算法的“图形”，就可以看到详细图形，结果如图8-62所示。



图 8-61

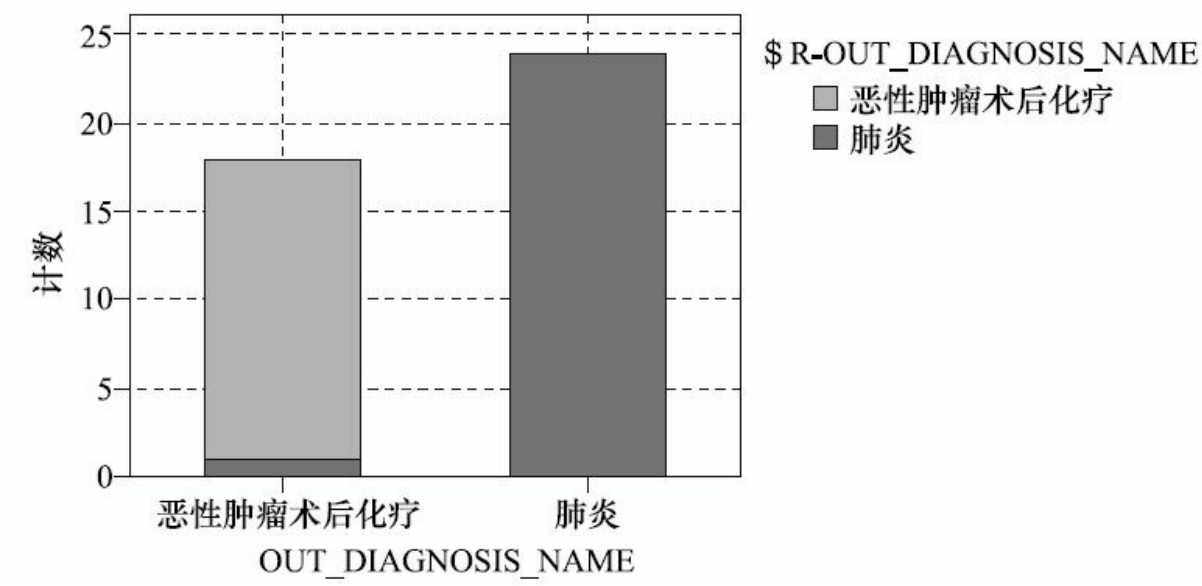


图 8-62

从图中可以看出，该算法的预测正确率很高，只有少数病例分错。双击“CHAID算法”，就可看到该算法的详细信息，如图8-63所示。

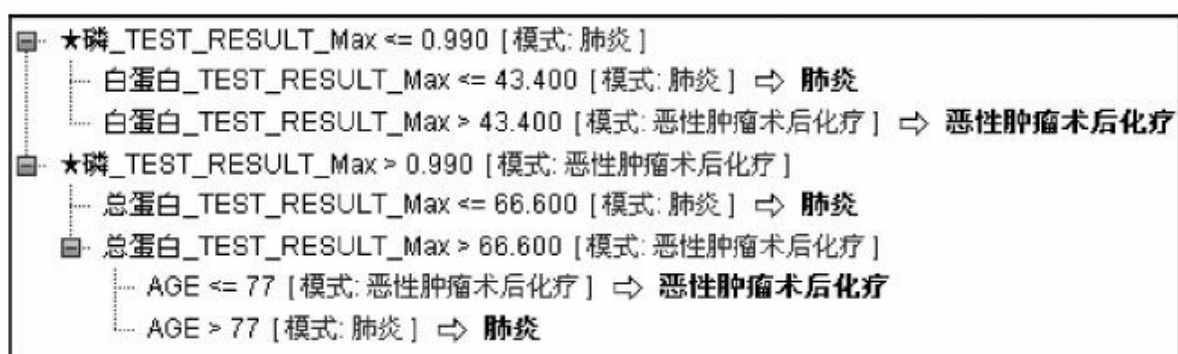


图 8-63

从图中可以看出其分类规则：

- 当患者的磷指标小于等于0.990时，再根据白蛋白的化验结果判断。当白蛋白小于等于43.4时，该病人很有可能是肺炎患者；当白蛋白的化验结果大于43.4时，该病人很有可能是恶性肿瘤术后化疗患者。

- 当患者的磷指标大于0.990时，则根据年龄和总蛋白进行判断。当总蛋白小于等于66.600时，该疾病人者很有可能是肺炎患者；而当总蛋白大于66.600时，又要根据年龄去判断。当年龄大于77岁，很可能是肺炎患者，而小于等于77，则很可能是恶性肿瘤术后化疗患者。

通过本例的探索分析，我们发现肺炎和恶性肿瘤术后化疗这两种

疾病是可以依据一些化验指标将其有效地区分开来。

8.2.3.2 回归预测：疾病的预测探索

未来是不可知的，但它有一定的可预测性。我们可以根据实际中的某些情况去预测未来。不是所有的疾病都会有显著的症状，有些疾病，当病人感觉到症状的时候，很有可能已经到了很严重的地步。根据一些相关的生理指标，可以预测与某种疾病相关的生理指标，就能进一步预测病人潜在患有该疾病的可能性。本例中，采集了恶性肿瘤术后化疗的生化全项数据，根据该病人的一些实验室指标数据，来预测该病人群的 α -羟丁酸脱氢酶指标，分析该指标是否与其他指标之间存在某种模型关系。

SPSS Modeler中还有一个算法集成器：自动数值器，其中包含了八种算法，即线性回归、广义线性、KNN、SVM、C&R树、CHAID、神经网络和自动线性。从IMRS系统中选择数据，选择疾病为恶性肿瘤术后化疗，检验项为生化全项，得到数据总计695条数据，数据挖掘过程与前例类似。在本例中，选择 α -羟丁酸脱氢酶为目标变量，性别、年龄和其余检验项为输入变量。将建模中的自动数值器添加到数据流的末端，其参数为默认设置，KNN和SVM不包含在默认算法中。如图8-64所示。

要执行的模型估计数: 6

字段 模型 专家 设置 注解

模型名称: ☐ 自动 ☐ 自定义

☒ 使用分区数据

☒ 为每个分割构建模型

模型排序依据: 相关

模型排序方式: ☐ 训练分区 ☒ 测试分区

要使用的模型数量:

☒ 计算预测变量重要性

在下列情况下不保存模型:

☐ 相关小于

☐ 字段数大于

☐ 相对误差大于

图 8-64

系统将会自动运行所有默认算法，所有可行算法模型运行成功后，输出结果如图8-65所示。

模型

图形

汇总

设置

注解

排序方式:

使用

升序

降序

刷新

删除未使用模型

视图:

训练集

是否使用?	图形	模型	构建时间 (分钟)	相关	使用的 字段编号	相对错误
<input checked="" type="checkbox"/>		 回归 1	< 1	1	14	0
<input checked="" type="checkbox"/>		 广义线性 1	< 1	1	40	0
<input checked="" type="checkbox"/>		 CHAID(C) 1	< 1	1	3	0

图 8-65

设定中输出前三个最好的模型，本例是回归、广义线性和CHAID这三种模型。三种模型的构建时间、相关性和相对错误指标显示这三个模型性能良好，其中，CHAID模型使用的预测变量数最少，可以更简洁、清晰地表达预测模型，故在本例中主要介绍CHAID模型。双击CHAID模型的“图形”，可以详细观察其图形，如图8-66所示。

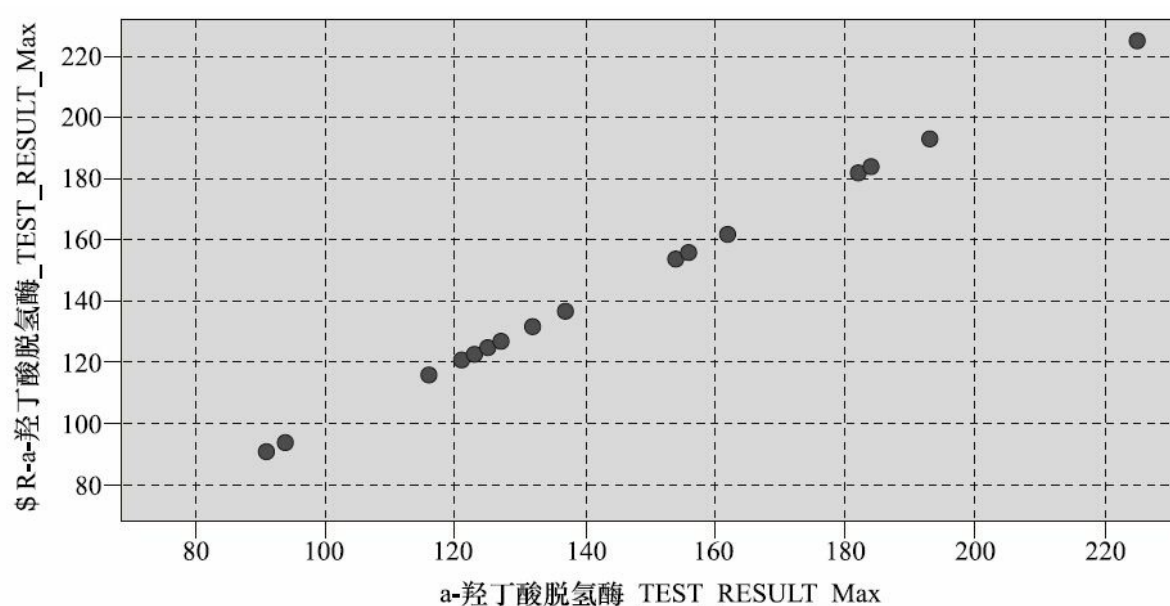


图 8-66

该图形说明模型拟合度较高，可以选用该模型为预测模型。双击“CHAID (C) 模型”，弹出该模型的详细结果，如图8-67所示。

从预测变量重要性的图8-67中可以看出，该算法的输入变量是乳酸脱氢酶、碱性磷酸酶和尿素酶这三种指标，最重要的预测变量是乳酸脱氢酶，占0.43；其次是碱性磷酸酶，占0.34；最后是尿素酶，占0.23。

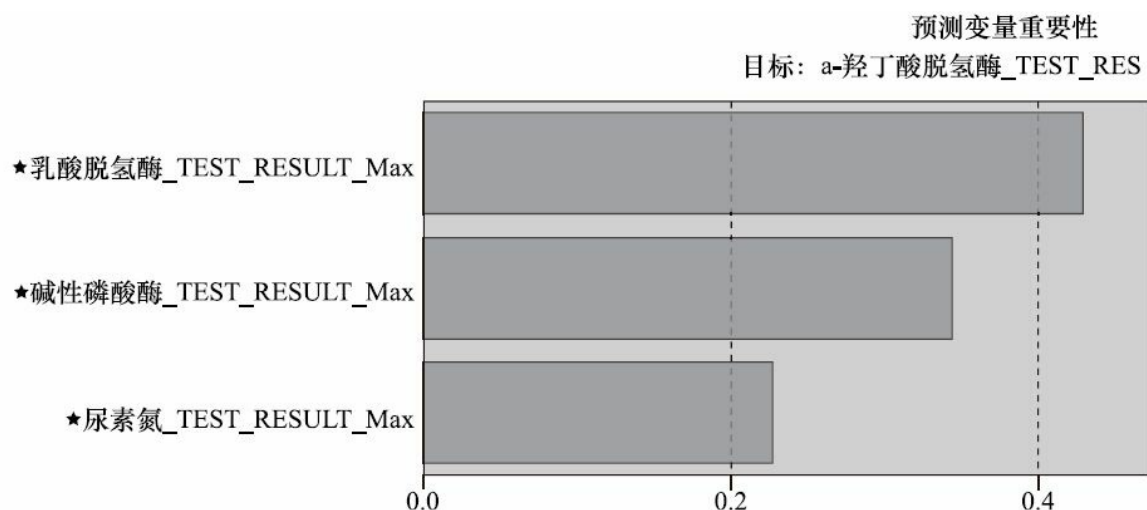


图8-67 预测变量重要性

第二个图，即图8-68，显示了决策树的每一个节点，该决策树总共包含了三个主节点，17个子节点。节点的末端就是落在该节点的a-羟丁酸脱氢酶的预测指标，该预测值基本反映了a-羟丁酸脱氢酶的分布情况。

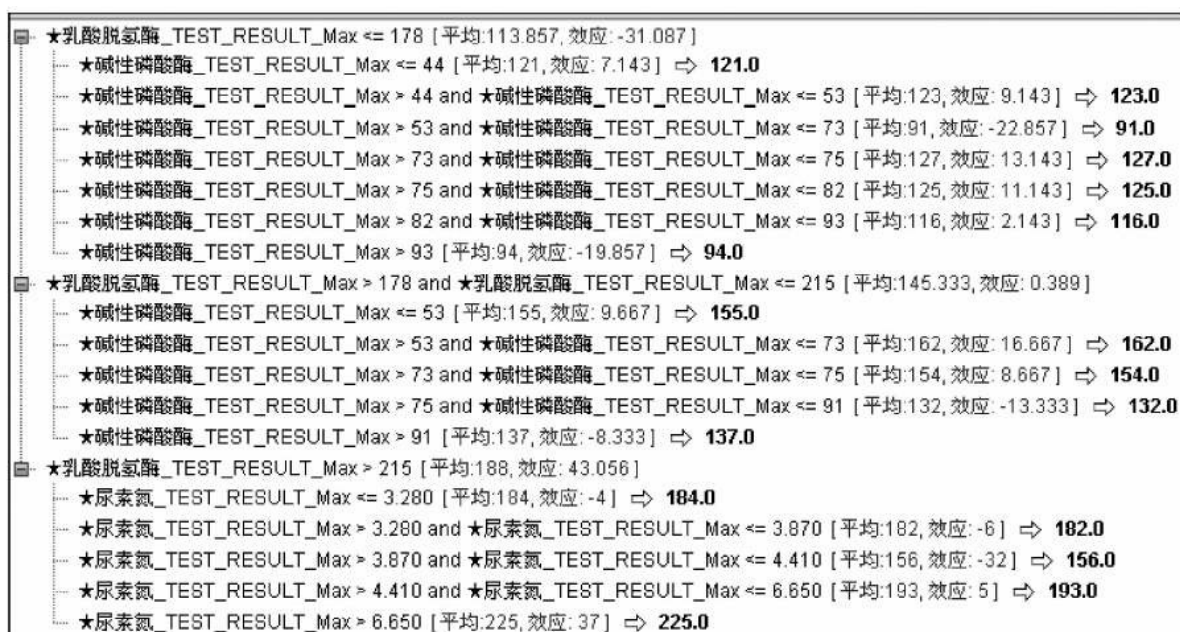


图 8-68

再单击输出结果的“查看器”按钮，如图8-69所示，是以树形方式显示的模型，树的节点都是一个决策点。查看树的上部分，第一个节点（节点0）为我们提供数据集中所有记录的摘要。数据集总共有18个数据，为总数据的100%，预测平均值为144.944。输出结果还包括了P值、F值和自由度，其统计检验结果说明该节点预测显著。再观察树的第二部分，节点1是当乳酸脱氢酶小于等于178时的决策点。该节点预测了七个值，占总数据的38.889%，预测的平均值为113.857。节点1又分为十个节点，分别在不同的碱性磷酸酶取值下进行的预测。例如节点5是当碱性磷酸酶小于等于44时，预测a-羟丁酸脱氢酶的值

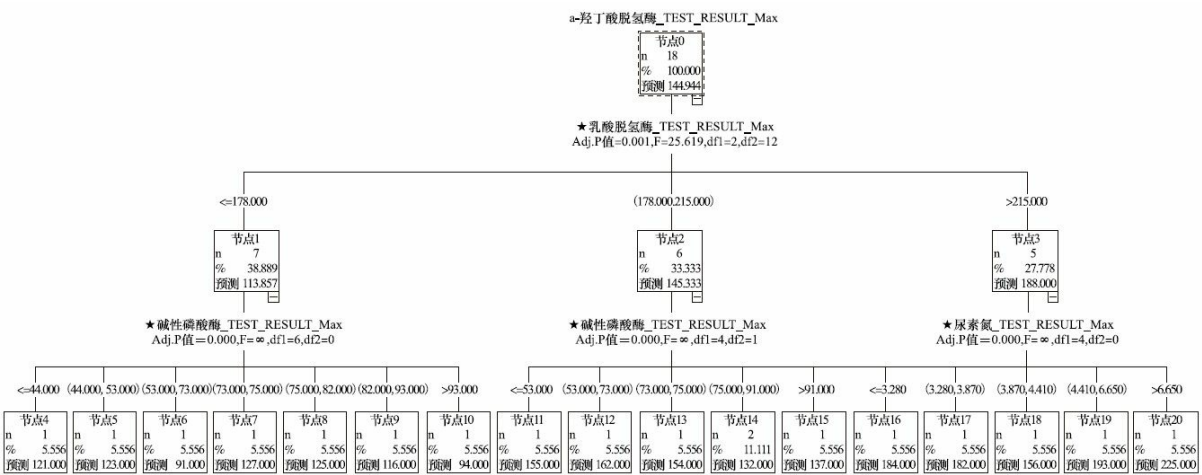


图 8-69

从本例的研究结果可得，恶性肿瘤术后化疗患者的a-羟丁酸脱氢酶指标与乳酸脱氢酶、碱性磷酸酶和尿素酶三个指标存在一定的相关关系，其中对a-羟丁酸脱氢酶影响最高的是乳酸脱氢酶，并且这三个指标作为预测因子与目标变量a-羟丁酸脱氢酶构成一个预测模型，结果显示预测模型性能良好。

8.3 本章小结

本章介绍了应用数据挖掘通用工具软件开展数据挖掘商业应用的案例。我们的体会是数据挖掘通用工具软件的确给我们提供了一些便利，对于数据挖掘应用的普及和推广起到了积极的作用。但是仅仅有了功能强大、使用方便数据挖掘软件工具，对于数据挖掘的成功应用来说还远远不够。很多没有数据挖掘应用经验的单位，人们的认识往往有种误区，认为我们拥有了很好的数据挖掘工具软件，就可以做好商业应用项目。其实这种认识错了，当他们将数据挖掘项目真正付诸实施的时候，才发现过程和结果远远不是当初所想象的！其实，数据挖掘应用的关键、挑战和瓶颈都是在于开展数据挖掘应用的人才，这种人才不仅仅要求能深入掌握数据挖掘技术，还需要对商业知识有深入的了解，并具有丰富的数据挖掘应用实战经验。当然，有一套好的数据挖掘软件工具对于数据挖掘应用来说也是很必要的，我们相信，更多更好的数据挖掘工具的出现会促进数据挖掘实践应用的发展和普及。

参考文献

- [1] Jiawei Han, Micheline Kamber.数据挖掘概念与技术
[M].范明, 孟小峰, 译.北京: 机械工业出版社, 2012.
- [2] 科弗 (Cover, T.M.), 托马斯 (Thomas, J.A.).信息论基础.阮吉寿, 张华, 译.北京: 机械工业出版社, 2007.
- [3] 威滕 (Witten, I.H.), 弗兰克 (Frank, E.), 霍尔 (Hall, M.A.).数据挖掘: 实用机器学习与技术.李川, 等译.北京: 机械工业出版社, 2014.
- [4] Pang-Ning Tan, Michael Steinbach, Vipin Kumar.数据挖掘导论 [M].范明, 等译.北京: 人民邮电出版社, 2011.
- [5] 胡良平.科研课题的研究设计与统计分析——错误案例辨析与释疑 (第一集) [M].北京: 军事医学科学出版社, 2008.
- [6] 颜宏等.医学统计学 [M].北京: 人民卫生出版社, 2005.
- [7] 黄德双, 等.生物信息学中的智能计算理论与方法研究 [M].合肥: 中国科技大学出版社, 2006.
- [8] 王学民.应用多元分析 [M].上海: 上海财经大学出版社, 2009.
- [9] 戴维S穆尔.统计学的世界 [M].郑惟厚, 译.北京: 中信出

版社，2003.

[10] Oracle官网: www.oracle.com.

[11] IBM官网: www.ibm.com.

Table of Contents

前言	13
第1章 数据挖掘应用绪论	22
1.1 认识数据挖掘	23
1.1.1 数据挖掘概念	24
1.1.2 数据挖掘与生活	30
1.1.3 数据挖掘与知识	35
1.2 数据挖掘应用基础	37
1.2.1 事物与维度	39
1.2.2 分布与关系	44
1.2.3 描绘与预测	51
1.2.4 现象和知识	54
1.2.5 规律与因果	56
1.3 数据挖掘应用系统工程	58
1.3.1 数据层	60
1.3.2 算法层	70
1.3.3 应用层	84
1.4 数据挖掘应用体会	90
1.4.1 项目关键点	91
1.4.2 技术与应用创新	94
1.4.3 经验积累与应用	96
1.5 无限三维嵌套空间假说	98
1.5.1 一维空间	99
1.5.2 二维空间	100
1.5.3 三维空间	102
1.5.4 突破三维空间	104
1.5.5 五维空间	107
1.5.6 六维空间	109
1.6 本章小结	111
第2章 数据探索与准备	112
2.1 数据关系探索	115
2.1.1 业务发现	116

2.1.2 关系发现	121
2.1.3 数据质量探索	124
2.1.4 数据整合	132
2.2 数据特征探索	136
2.2.1 数据的统计学特征	137
2.2.2 统计学特征应用	153
2.3 数据选择	161
2.3.1 适当的数据规模	162
2.3.2 数据的代表性	164
2.3.3 数据的选取	167
2.4 数据处理	173
2.4.1 数据标准化	175
2.4.2 数据离散化	177
2.5 统计学算法的数量条件	181
2.5.1 样本量估计概念	182
2.5.2 单样本总体均值比较的样本量估计 (T-Test)	184
2.5.3 两样本总体均值比较的样本量估计 (T-Test)	187
2.5.4 多样本总体均值比较的样本量估计 (F-Test)	190
2.5.5 区组设计多样本总体均值比较的样本量估计 (F-Test)	194
2.5.6 直线回归与相关的样本量估计	196
2.5.7 对照分析的样本量估计	197
2.6 数据探索应用	200
2.6.1 检验项的疾病分布	201
2.6.2 疾病中检验项的分布	204
2.6.3 成对检验项的相关分析	206
2.6.4 两种药物的应用分析	208
2.7 本章小结	210
第3章 数据挖掘应用算法	211
3.1 聚类分析	213
3.1.1 划分聚类算法 (K均值)	214
3.1.2 层次聚类算法 (组平均)	225
3.1.3 密度聚类算法	236
3.2 特性选择	239

3.2.1 特性选择概念	240
3.2.2 线性相关算法	249
3.2.3 相关因子SRCF算法	252
3.3 特征抽取	270
3.3.1 主成分分析算法	271
3.3.2 因子分析算法	274
3.3.3 非负矩阵因子分解NMF算法	276
3.4 关联规则	279
3.4.1 关联规则概念	280
3.4.2 Apriori算法	282
3.4.3 FP树频集算法	284
3.4.4 提升Lift	286
3.5 分类和预测	287
3.5.1 支持向量机	288
3.5.2 Logistic回归算法	297
3.5.3 朴素贝叶斯分类算法	305
3.5.4 决策树	316
3.5.5 人工神经网络	324
3.5.6 分类与聚类的关系	332
3.6 时间序列	334
3.6.1 灰色系统预测模型	335
3.6.2 ARIMA模型预测	346
3.7 本章小结	349
第4章 数据挖掘应用案例	350
4.1 特性选择的应用	351
4.1.1 数据整合	352
4.1.2 数据描绘	353
4.1.3 数据标准化	355
4.1.4 特性选择探索	356
4.2 分类模型的应用——算法比较	366
4.2.1 数据整合	367
4.2.2 数据描绘	369
4.2.3 数据标准化	372
4.2.4 特性选择探索	373

4.2.5 分类模型	377
4.3 分类模型的应用——网络异常侦测	380
4.3.1 计算机网络异常行为	381
4.3.2 网络异常数据模型	383
4.3.3 分类模型算法应用	390
4.4 算法的综合应用——肿瘤标志物的研究	397
4.4.1 样本选取	399
4.4.2 癌胚抗原临床特征主题分析	408
4.4.3 癌胚抗原临床特征规则分析	414
4.4.4 癌胚抗原临床特征规则的比较分析	422
4.4.5 癌胚抗原相关因子分析	424
4.4.6 不同等级癌胚抗原组差异分析	431
4.5 数据挖掘在其他领域中的应用	438
4.6 本章小结	443
第5章 数据挖掘行业应用原理	444
5.1 传统医学科研方法的现状	446
5.1.1 传统医学科研的命题与假说	447
5.1.2 传统医学科研的数据应用	449
5.1.3 传统的医学科研的统计学应用	451
5.1.4 传统医学科研的流程	453
5.2 智能医学科研系统的需求	454
5.2.1 临床医学科研的问题	455
5.2.2 临床医学科研的解决思路	457
5.3 智能医学科研系统的设计思想	462
5.3.1 科研立题	463
5.3.2 科研设计与统计分析	465
5.3.3 样本数据收集与分析	467
5.4 智能医学科研系统的核心技术方法	469
5.5 智能医学科研系统的科研数据仓库建设	472
5.5.1 医学科研数据仓库建设的技術方法	473
5.5.2 医学科研数据仓库的建设过程	477
5.5.3 科研数据仓库的数据安全	481
5.6 智能医学科研系统的核心功能设计	482
5.7 智能医学科研系统的整体功能设计	485

5.7.1 智能医学科研系统主要功能	486
5.7.2 智能医学科研系统的模块设计和应用实现	489
5.7.3 智能医学科研系统的评估方法	505
5.8 智能医学科研系统的应用价值	515
5.9 本章小结	521
第6章 数据挖掘应用系统的开发	522
6.1 数据挖掘应用系统的意义	523
6.2 IMRS系统设计	526
6.2.1 对数据源的分析	528
6.2.2 数据挖掘应用系统IMRS的总体设计	534
6.3 IMRS异常侦测模型的开发	548
6.3.1 异常侦测模型的功能展示	549
6.3.2 数据挖掘技术开发要点	557
6.4 IMRS特征抽取模型的开发	566
6.4.1 特征抽取模型的功能展示	567
6.4.2 数据挖掘技术开发要点	568
6.5 IMRS智能统计模型的开发	587
6.5.1 回归模型的开发实现	589
6.5.2 线性相关模型的开发实现	610
6.6 IMRS的算法开发	615
6.6.1 相关因子算法SRCF的实现	616
6.6.2 朴素贝叶斯分类算法的实现	624
6.7 本章小结	632
第7章 数据挖掘应用	633
7.1 分布探索	635
7.1.1 两维度聚类模型应用	636
7.1.2 高维度聚类模型应用	646
7.2 关系探索	649
7.2.1 关联规则的应用	650
7.2.2 特性选择的应用	656
7.3 特征探索	666
7.3.1 不稳定心绞痛的特征总结	667
7.3.2 动脉硬化心脏病的临床特征	676
7.4 异常探索	680

7.4.1 生理指标的异常侦测	681
7.4.2 异常侦测模型的比较	686
7.5 推测探索	688
7.6 应用系统的高级应用	694
7.6.1 异常侦测的高级用法	695
7.6.2 关联规则的高级应用	704
7.7 本章小结	715
第8章 数据挖掘工具的应用	716
8.1 应用Oracle Data Mining	718
8.1.1 ODM数据挖掘流程	720
8.1.2 ODM算法模型	723
8.1.3 ODM算法应用	731
8.2 应用IBM SPSS Modeler	776
8.2.1 IBM SPSS Modeler介绍	777
8.2.2 SPSS Modeler独立应用	780
8.2.3 SPSS Modeler与应用系统的联合应用	794
8.3 本章小结	807
参考文献	808