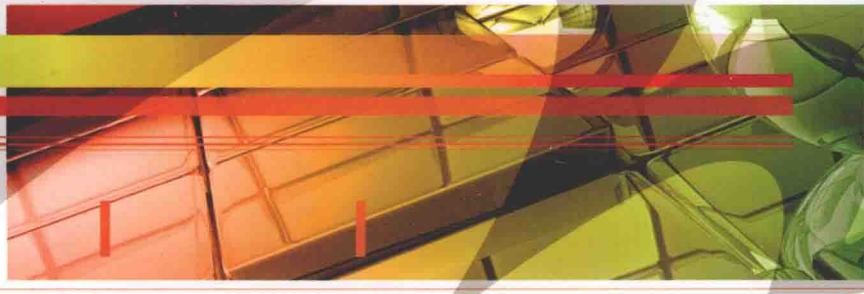
数据挖掘理论、方法与应用

◆ 罗泽举　著



電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http: //www.phei.com.cn

相关参考图书

数据挖掘理论、方法与应用

罗泽举 著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书共分 10 章。第 1 章介绍了数据挖掘方法的历史，主要是人工神经网络与支持向量机的产生背景，另外讨论了统计学习的一般模型，这是通过数据挖掘建立数学模型的基础，本章还介绍了常用的数据预处理变换方法。第 2、3、4 章介绍了 3 种数据挖掘模型：人工神经网络、支持向量机和隐马尔可夫模型。站在独特的角度，用形象生动和朴实易懂的语言分析了 3 种数据挖掘方法的思想、原理、理论。第 5 章介绍了一种新型支持向量诱导回归模型，第 6 章介绍了一种基于快速训练算法的 HMM/SVM 混合系统，第 7 章介绍了分解向前算法及 PCA/ICA 降维 SVM 模型，第 8 章介绍了不对称支持向量机改进算法，第 9 章介绍了一种基于隐马尔可夫模型的多重序列分析方法，第 10 章介绍了一类基于 SVM/RBF 的气象模型预测系统。

本书内容丰富，可供理工科中应用数学、计算机科学，计算生物学，统计学等相关专业具有一定数学背景并对数据挖掘方法有兴趣的高校教师、研究生使用，也可供从事机器学习与模式识别的相关领域研究的科研人员和数据挖掘工作者参考。

图书在版编目（CIP）数据

数据挖掘理论、方法与应用/罗泽举著. —北京：电子工业出版社，2014.12

ISBN 978-7-121-24633-3

Ⅰ.①数… Ⅱ.①罗… Ⅲ.①数据采集 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2014）第 246864 号

策划编辑：任欢欢
责任编辑：郝黎明
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：720×1 000　1/16　印张：11.5　字数：294.4 千字
版　　次：2014 年 12 月第 1 版
印　　次：2014 年 12 月第 1 次印刷
定　　价：35.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

数据挖掘领域目前越来越引起人们的重视，在过去的几十年里，数据挖掘技术已经从人们的零散的方法发展成为一系列系统的知识理论。随着大数据时代的到来，人们在这方面投入了更大的研究兴趣，数据挖掘的内容越来越丰富，数据挖掘的应用范围也越来越广泛，从经济到管理、从生物到医学、从农业到商业，到处都呈现出数据挖掘学科发展的欣欣向荣的局面。进行数据挖掘的动力在于挖掘的信息能对我们的经济生活提供重要的帮助。商业数据挖掘可以为商家提供重要的商机，气象数据挖掘可以帮助我们防止自然灾害和提高农作物产量，医学数据挖掘可以帮助我们检测疾病并进行及早治疗，管理数据挖掘可以帮助管理者提高决策效率等。

本书向读者介绍了几种重要的数据挖掘概念与算法，包括神经网络、支持向量机、隐马尔可夫理论，本书由浅入深，循序渐进，采用生动朴实的语言向读者介绍了这些理论的来龙去脉，使读者对这些理论与方法有更深入的理解。神经网络部分，从人的大脑结构开始，从生物的刺激反应到建立数学模型，遵循了知识的发现过程。支持向量机部分，由最大间隔分类超平面和支持向量介绍入手，从一维到多维，从独特的角度介绍了这一理论的深刻内涵。隐马可夫理论部分，从马尔可夫链的介绍开始，通过引入双重随机过程使这一理论得到自然展现。

本书分为三个主要部分。绪论部分、主要理论部分及应用部分。绪论部分介绍了神经网络和支持向量机理论的产生过程，同时介绍了数据挖掘的基本数学问题，对众多算法概念进行了剖析，并讨论了统计学习的一般模型和常用数据变换方法。主要理论部分由神经网络挖掘理论、基于支持向量的挖掘理论及隐马尔可夫挖掘理论 3 章组成，神经网络挖掘理论部分包括神经智能、生物神经元和人工神经元、LMS 和 SDA 方法及后向传播算法；基于支持向量的挖掘理论部分由支持向量与分类超平面、风险控制策略、样本被错分的讨论、最优化策略、分类与回归、几种经典算法描述组成；隐马尔可夫挖掘理论部分由马尔可夫思想、隐马尔可夫链、隐马尔可夫模型组成。应用部分介绍了这三种主要的数据挖掘技术在经济和计算生物学中的应用，在这些章节，改进了传统算法，提出了多个数据挖掘模型，包括第 5 章的新型支持向量诱导回归模型，第 6 章的基于快速训练算法的 HMM/SVM 混合系统，第 7 章的分解向前算法及 PCA/ICA 降维 SVM 模型，第 8 章的不对称支持向量机改进算法模型，第 9 章的基于隐马尔可夫模型的多重序列分析模型，第 10 章的一类基于

SVM/RBF 的气象模型预测系统。这些应用章节的内容也较为丰富，包含了许多有关数据挖掘的知识内容，如降维方法、数据编码、计算复杂度分析等。

本书内容丰富，可供理工科中应用数学、计算机科学、计算生物学、统计学等相关专业具有一定数学背景并对数据挖掘方法有兴趣的高校教师、研究生使用，也可供从事机器学习与模式识别的相关领域研究的科研人员和数据挖掘工作者参考。

感谢在本书编写过程中给予过我帮助、扶持和关心的朋友，同时也感谢我的妻子，每天承担了许多家务，以便我有时间来静心工作，也感谢我的女儿，容忍了我由于睡眠不足而导致的喜怒无常。

在本书撰写时，参考了国内外学者所著的有关文献，受益匪浅，特此向原作者致谢。由于作者水平有限，书中肯定存在不少疏漏与不足，恳请读者批评指正。

罗泽举

2014 年 8 月于丽水学院

目　录

第1章 绪　论

生活是一曲美丽而曲折的探索赞歌。

“滚滚长江东逝水，浪花淘尽英雄，是非成败转头空，江山依旧在，几度夕阳红。”当人们都在为这脍炙人口的佳作感叹称道的同时，从另一个层面，我们可曾想到，自从人类这个文明诞生以来，多少朝代更迭，多少王国新生又湮灭，多少英雄洒下一个又一个传奇，而探索不尽的大自然，还有人与自然组成的社会，大的生态系统，它却静静地依然流淌在那里，任由时间演变，人们就在大自然的怀抱上成长。而征服自然、改造自然成了人类的重要使命，从自然身上挖掘那取之不尽的真理与规律，成为多少仁人志士的终生追求与奋斗目标，真所谓“江山如此多娇，引无数英雄竞折腰”。阿基米德从澡盆洗澡时水往外溢而发现了浮力，进而推动了静态力学和流体静力学的大发展，我国蔡伦“闭门绝宾，暴体田野”，历尽千辛万苦终于琢磨出了一整套完善的造纸术，研制出了第一批用废麻和树皮做原料的植物纤维纸，推动了全球文明的发展，牛顿通过苹果落地而发现了万有引力定律，并推动了现代科学的巨大发展，瓦特通过观察蒸汽把水壶盖顶开而引发了他对蒸汽的兴趣，从而有了蒸汽机的发明，并推动了世界第一次工业技术革命的兴起。可见，人类生存的历史，本质上就是一部真理探索史，其实就是一部数据挖掘史。

人类探索真理的目的就是为了获得相对的自由，发现了真理就不会为那羁绊所跌倒，有了真理就不会为那无知而受苦难。人类掌握的真理越多，所受自然的奴役就越少；人类挖掘到的自然规律和社会规律越多，人类所受的苦难就越少。

人类是很有挖掘和学习能力的。从生物的刺激反应学会了感知，从水的承载力发明了船，从鸟的飞行发明了飞机，从无穷小的探索中发现了微积分，从数据计算中发明了计算机，从互联网的交往中又发明了电子商务、移动终端、支付宝等。人类总是不断创造一个又一个的奇迹，制造一个又一个的传说。

1.1 研究背景

人类能从实践中不断学习和改进，通过人类自身能挖掘设计出能模仿人类思维的智能挖掘机器吗？

计算机当初是作为数据统计、数据分析工具的。机器是否也具有学习和进行数据挖掘的能力？能否构造从经验中学习的机器一直是哲学界和科学界研究的伟大目标之一。例如，人类能够利用大脑神经进行思考和学习，那么，这些神经学习模式能够被提炼和升华吗？1943 年，心理学家 W.McCulloch 和数理逻辑学家 W.Pitts 在合作的《A logical calculus of the ideas immanent in nervous activity》论文中提出并给出了人工神经网络的概念及人工神经元的数学模型，从而开创了人工神经网络研究的时代。1949 年，心理学家唐纳德·赫布在《The Organization of Behavior》论文中描述了神经元学习法则。终于在 20 世纪 50 年代末，在 Cornell 航空实验室中，美国神经学家 F.Rosenblatt 设计出一种从样本中学习的感知器，成功在 IBM 704 机上完成了感知机的仿真，并于 20 世纪 60 年代早期建立了关于感知器算法的基本理论，如 1958 年发表的论文《The Perceptron：A Probabilistic Model for Information Storage and Organization in the Brain》和 1962 年出版的著作《Principles of Neurodynamics：Perceptrons and the theory of brain mechanisms》，这些论著实际上预示着数据挖掘中机器学习理论研究的开始。F.Rosenblatt 于 1962 年证明了对于一个无限的样本序列，感知器经过有限多次训练后，就可以构造出将这一无限序列进行正确分类的超平面。这其实是历史上第一个被挖掘出的具有学习型的神经网络，尽管比较简单，却具有神经网络的基本性质，如分布存储、并行处理、可学习性、连续计算等。1962 年，B.Widrow 和 M.Hoff 提出了一种连续取值的线性加权求和阈值网络，即自适应线性元件网络，可以看成是感知器的发展，它成功地应用于自适应信号处理和雷达天线控制等连续可调过程。他们在人工神经网络理论上创造了一种被人们熟知的 Widrow-Hoff 学习训练规则，即最小均方（Least Mean Square，LMS）算法，并将人工神经网络用硬件电路实现，为今天用大规模集成电路实现神经网络奠定了重要研

究基础。20 世纪 60 年代中期，M.A.Aizerman 等提出了位势法，利用符号函数从样本数据中估计函数的依赖关系，从中找出分类规则。1970 年，他们已经构造出势函数的一致性理论。1969 年，人工智能创造人之一 M.Minsky 和 S.Papert 出版一本名为《感知器》的专著，书中指出线性感知器功能是有限的，简单的神经网络只能进行线性分类和求解一阶谓词问题，而不能进行非线性分类和解决比较复杂的高阶谓词问题。1972 年，芬兰的 T.Kohonen 教授提出了自组织映射理论。1982 年美国加州理工学院物理学家 J.Hopfield 教授提出一种递归网络——Hopfield 网络，利用 Hopfield 网络的神经计算能力来解决约束优化问题，从而开辟了神经网络用于联想记忆和优化计算的新途径。20 世纪 80 年代中期，美国认知心理学家 D.E.Rumelhart 和 J.L.Mccelland 为首的 PDP 研究小组发表了多层神经网络学习的误差反向传播算法（Backpropagation Algorithm，BP 算法），用损失函数的梯度来迭代修正神经网络的权系数，使神经网络理论和应用走向热潮。

然而，伴随神经网络理论的发展，人们逐渐面临如下一些问题：启发式算法在训练过程中如何保证网络结构的最优化？对于多层网络，如何克服模型的复杂性？如何克服网络的过学习和过拟合问题？神经网络的优化结果是局部最优，怎样才能得到全局最优解？当输入向量的维数增加时，神经网络的计算量大增，如何克服算法的复杂性和解决输入空间的维数灾难？样本数据仅是输入空间中的稀疏分布，传统上仍以样本数目无穷多为假设来推导各种算法，如何处理小样本学习问题？如何刻画网络的收敛条件和收敛速度等。

要克服上述这些问题，人们试图通过用更加聪明的启发式算法来解决，例如，20 世纪 80 年代中期，再次出现了利用势函数（径向基函数）研究函数逼近的兴趣，花费很大的力气对径向基函数模型进行研究，采用经验风险最小化原则替代随机逼近推理来构造逼近函数，大家提出了许多非监督学习启发式算法，如 D.E.Rumelhart 和 J.L.Mccelland 的梯度下降算法，但这些方法最终都未能解决上述机器学习中遇到的若干问题。

统计学习理论的早期研究者们逐渐认识到，要克服传统经验学习和启发式学习的弊端，必须去寻求严格理论的支持，他们开始用统计学习理论（Statistical Learning

Theory，SLT）来系统地研究机器学习问题。这些新的思想火花是受最小化错误率的理论界而被激发的，与传统启发式构造学习算法的旧思想根本不同，新的算法不但要有良好的数学性质，如解的最优性、小样本的学习推广能力、处理大样本的简单方法，以及不依赖于输入空间的维数等，而且这些新的算法得到的解表现出比旧方法更优秀的性能，如全局最优的、一致收敛的、风险和错误率得到良好控制等。1992年，V.N.Vapnik 等人用自己早年提出的最优分类超平面思想，结合 M.A.Aizerman 等在对位势函数的收敛特性的分析中提出的将低维输入空间映射到高维特征空间的理论，V.N.Vapnik 等人发现，为了在特征空间中构造分类超平面，并不需要以显式的形式来表示特征空间，而只需要能够计算支持向量与特征空间中向量的内积——他们提出并创立了支持向量机（Supporting Vector Machines，SVM）理论。支持向量机是统计学习理论和机器学习的历史结晶。这一新理论一被提出，立即引起国际人工智能技术研究界的关注，并逐渐成为机器学习研究的新热点。1995 年，V.N.Vapnik 等在自己领导 AT&TBell 实验室，第一次对美国邮件服务数据库里的 9300 个邮政编码数字进行手写数字识别，使得支持向量机比五层神经网络高出 2.6 个百分点的识别结果。

支持向量机起初是作为模式识别而被提出来的，后来迅速发展应用于函数逼近，线性和非线性回归函数估计，近十几年来，有众多学者研究和应用了支持向量机理论，在理论方面，有以改进算法为方向的，也有以核函数研究为方向的；应用方面，已成功用于语音识别，图像识别，手写数字识别，文本识别，非线性系统识别，各种时间序列分析，包括混沌时间序列分析等；SVM 还成功用生命科学的前沿，如 DNA 建模与控制，基因序列分析，病理诊断等。

支持向量机成为机器学习新的热点领域，并成功地得到广泛的研究和应用，在于它成功地克服了传统神经网络学习的众多弊病以及自身的一系列新特点所至，归纳起来，主要有以下几个方面。

（1）以严格的数学理论（统计学习理论）为基础，克服了传统神经网络学习中靠经验和启发的先验成分。

（2）结构风险最小化（Structural Risk Minimization，SRM）原则，克服了传统神

经网络只靠经验风险最小化（Empirical Risk Minimization，ERM）原则，提高了置信水平，克服了过学习问题，使学习机器有良好的泛化能力。

（3）通过解决凸二次规划问题，得到问题的全局最优解，而不是传统神经网络学习的局部最优解。

（4）用内积的回旋克服了特征空间的维数灾难问题，巧妙地构造核函数，通过非线性映射，只需计算支持向量与特征空间中向量的内积，不需要以显式形式表示特征空间。

（5）成功解决了小样本学习问题，克服了传统上以样本数目无穷多为假设来推导各种算法，得到了小样本条件下的全局最优解。

（6）通过引入 VC 维的概念，使网络的收敛速度、样本被错分的界和风险泛函得到了控制。

1.2　数据挖掘的基本数学问题

1.2.1　相关概念

1．输入空间（Input Space）

一般意义上，输入空间就是我们讨论的数据样本空间。输入空间 I 是 R^n 空间的子集，即 $I \subseteq R^n$，要求存在未知的概率分布函数 $F(x)$，使得输入空间 $I=\{x \mid x \in R^n, x \sim F(x)\}$。

2．输出空间（Output Space）

输出空间 O 是 R 空间的子集，即 $O \subseteq R$，$O=\{y \mid y \in R\}$。对于二元分类问题，一般取 $O=\{y \mid y=0$或$y=1\}$ 或 $O=\{y \mid y=1$或$y=-1\}$；对于多分类问题，通常取 $O=\{y \mid y \in \{0,1,\cdots, k-1\}\}$；对于函数回归或密度估计问题，$O$ 可以是 R 的任意子集。

3．训练机器（Training Machine）

训练机器 S 是根据某种训练规则确定的进行样本训练的机器。对每个输入向量

x 返回一个输出值 y，产生输出的数据应满足未知的条件分布函数 $F(y|x)$。

4．训练集（Training Set）

对于输入空间 $I=\{x|x\in R^n, x\sim F(x)\}$ 中的任一变量 x，存在未知的概率分布函数 $F(x)$ 和按训练规则 $F(y|x)$ 进行训练的训练机器 S，因而联合分布 $F(x,y)=F(x)F(y|x)$ 存在。训练集 T 是由 k 个独立同分布（Independent and Identically Distributed，i.i.d）观测数据 $(x_1,y_1),(x_2,y_2)\cdots,(x_k,y_k)$ 组成的，它们满足存在但未知的联合分布函数 $F(x,y)=F(x)\ F(y|x)$，即 $T=\{(x_i,y_i)|(x_i,y_i)\sim F(x,y), i\in\{1,2,\cdots,k\}\}$。

5．损失函数（Loss function）

对参数集（指标集）Λ 确定的函数集合 $\Omega=\{f(x,\alpha)|\alpha\in\Lambda\}$，设观测样本为 (x,y)，其中 x 为观测输入，y 为观测输出，若学习机器的输出为 $f(x,\alpha)$，损失函数 $L(y,f(x,\alpha))$ 定义为 y 与 $f(x,\alpha)$ 之间的某种差值。常见的损失函数有平方损失函数 $L(y,f(x,\alpha))=(y-f(x,\alpha))^2$ 和绝对值损失函数 $L(y,f(x,\alpha))=|y-f(x,\alpha)|$。

6．期望风险泛函（Expectation Risk Functional，ERF）

设训练集满足未知联合分布 $F(x,y)=F(x)F(y|x)$，给定损失函数为 $L(y,f(x,\alpha))$，期望风险泛函 $R(\alpha)$ 就是在 (x,y) 的联合分布函数 $F(x,y)$ 下损失函数 $L(y,f(x,\alpha))$ 的数学期望，即

$$R(\alpha)=\int L(y,f(x,\alpha))\mathrm{d}F(x,y) \tag{1-1}$$

如果将损失函数 $L(y,f(x,\alpha))$ 记为：$L(y,f(x,\alpha))\triangleq Q(z,\alpha),\alpha\in\Lambda$，其中 $Q(z,\alpha)$ 的向量 z 作如下规定：z 包含 $n+1$ 个坐标，输入向量 x 的 n 个坐标再加上坐标输出 y，即设 $x=(x^1,x^2,\cdots,x^n)$，则 $z=(x,y)=(x^1,x^2,\cdots,x^n,y)$，于是训练集 $(x_1,y_1),(x_2,y_2),\cdots,(x_k,y_k)$ 就可以记为 $z_1,z_2,\cdots,z_k$，其中 $z_i=(x_i,y_i)=(x_i^1,x_i^2,\cdots,x_i^n,y_i),i=1,2,\cdots,k$，由于观测数据 $x_1,x_2,\cdots,x_k$ 独立同分布，则 $z_1,z_2,\cdots,z_k$ 也是独立同分布的，于是期望风险泛函 $R(\alpha)$ 可记为

$$R(\alpha)=\int Q(z,\alpha)\mathrm{d}F(z),\alpha\in\Lambda \tag{1-2}$$

7．经验风险泛函（Empirical Risk Functional，ERF）

$$R_{\text{emp}}(\alpha)=\frac{1}{k}\sum_{i=1}^{k}Q(z_i,\alpha)=\frac{1}{k}\sum_{i=1}^{k}L(y_i,f(x_i,\alpha)),\alpha\in\Lambda \tag{1-3}$$

称为经验风险泛函。可见，经验风险泛函是期望风险泛函的近似表示，在数据挖掘中知道的仅是有限个样本的信息，经典学习理论常常以之最小化为目标。

8．经验风险最小化原则（Empirical Risk Minimization Principle，ERM）

设 $Q_{\text{emp}}(z,\alpha_1)$ 表示使式（1-3）中的经验风险泛函 $R_{\text{emp}}(\alpha)$ 达到最小化的某个函数，$Q(z,\alpha_0)$ 表示使式（1-2）期望风险泛函 $R(\alpha)$ 达到最小化的某个函数，即

$$Q_{\text{emp}}(z,\alpha_1)\triangleq\arg\min_{\alpha\in\Lambda}R_{\text{emp}}(\alpha), \tag{1-4}$$

$$Q(z,\alpha_0)\triangleq\arg\min_{\alpha\in\Lambda}R(\alpha), \tag{1-5}$$

在概率分布 $F(z)$ 未知的条件下，用 $Q_{\text{emp}}(z,\alpha_1)$ 去逼近 $Q(z,\alpha_0)$，即用使经验风险最小的函数 $Q_{\text{emp}}(z,\alpha_1)$ 去逼近使期望风险达到最小的函数 $Q(z,\alpha_0)$，这一原则称为经验风险最小化原则（Empirical Risk Minimization，ERM）。传统回归估计中的最小二乘方法，概率密度估计中的最大似然方法（Maximum Likelihood，ML）都是 ERM 原则的具体体现。

9．监督学习与非监督学习（Supervised Learning and Unsupervised Learning）

监督学习是指在机器学习中，样本训练数据是由输入和输出对给出的，同时将相应的期望输出与网络输出相比较，得到误差信号，以此控制权值连接强度的调整，经多次训练后收敛到一个确定的权值。感知器神经网络、线性神经网络、BP 神经网络、支持向量机等属于监督学习。

非监督学习是指在机器学习中，样本训练数据只是给出输入数据，网络在训练过程中处理输入，并根据数据本身计算可能的输出，学习规律的变化服从连接权值的演变方程。Hebb 学习规则，自组织映射、适应谐振理论网络、主成分分析、隐马尔可夫方法等属于非监督学习。

10．欠学习与过学习（Underfitting Learning and Overfitting Learning）

欠学习就是给定一个函数集 $\Omega=\{f(x,\alpha)\,|\,\alpha\in\Lambda\}$，一个函数 $f_0\in\Omega$，使得不仅