HZ BOOKS
华章科技

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

大数据
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）
（中） 庄映辉 李堃 著

机械工业出版社
China Machine Press

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）
（中） 庄映辉 李堃 著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据挖掘技术与工程实践 /（加）洪松林等著 . —北京：机械工业出版社，2014.9
（大数据技术丛书）

ISBN 978-7-111-48076-1

I. 数… II. 洪… III. 数据采集 IV. TP274

中国版本图书馆 CIP 数据核字（2014）第 232037 号

本书系统讲解数据挖掘应用系统的实现方法，包含数据挖掘的基本概念与系统实现的全过程。本书作者根据自己20多年数据挖掘方面的工程经验，总结了数据挖掘的理论知识和实践经验，提供了大量一线资料。本书首先介绍数据挖掘的基本概念和误区，然后根据实际工作流程来讲解如何实现一个数据挖掘应用系统，最后总结了数据挖掘的常用工具。数据挖掘应用系统实现的流程包括数据的探索与准备、算法的应用、案例分析、行业应用特点、应用系统的开发、应用系统的充分使用等。书中介绍了大量数据挖掘的相关算法，包括：相关因子算法、聚类算法、分类算法、回归与测试算法等，不仅列举了详细示例，还介绍了算法在工程实践中的具体应用，特别是总结了自己独特的一些新算法，例如秩相关因子选择算法、矢量相关因子选择算法、密度分布聚类算法、概率特征模型算法等。还剖析了几个热门领域的实际应用，涉及医药学、信息安全等领域的应用。

本书可供数据挖掘、数据仓库、数据库等领域的技术人员参考，也可供想建立智能计算系统的企业信息系统管理人员参考。

数据挖掘技术与工程实践

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：吴 怡 责任校对：殷 虹
印 刷：藁城市京瑞印刷有限公司 版 次：2014 年 11 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：24
书 号：ISBN 978-7-111-48076-1 定 价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

献给我亲爱的 Eddie!

Preface 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。“特征抽取则”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很有效的。

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、肖芃、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。

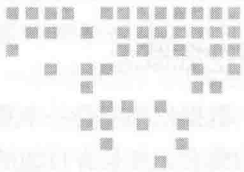
洪松林（Hong Song Lin）

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 人工神经网络原理 （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		

Contents 目　录



第 1 章 Chapter 1

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