TURING 图灵程序设计丛书

ELSEVIER

Predictive Analytics and Data Mining
Concepts and Practice with RapidMiner

预测分析与数据挖掘
RapidMiner实现

【美】Vijay Kotu Bala Deshpande 著
严云 译

- 以易于理解的方式梳理数据挖掘背后的基础知识
- 全面展示预测分析领域广泛的实践案例和方法
- 无需编写代码，即可解决数据分析问题

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Predictive Analytics and Data Mining
Concepts and Practice with RapidMiner

预测分析与数据挖掘
RapidMiner实现

【美】Vijay Kotu Bala Deshpande 著
严云 译

人民邮电出版社
北 京

图书在版编目(CIP)数据

预测分析与数据挖掘：RapidMiner实现/（美）瓦杰·考图（Vijay Kotu），（美）巴拉·达什潘德（Bala Deshpande）著；严云　译. —北京：人民邮电出版社, 2018.1
（图灵程序设计丛书）
ISBN 978-7-115-47366-0

Ⅰ. ①预…　Ⅱ. ①瓦… ②巴… ③严…　Ⅲ. ①数据处理软件　Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2017）第295481号

内 容 提 要

本书旨在帮助读者理解数据挖掘方法的基础知识，并实现无需编写代码就能在自己的工作中实践这些方法。书中围绕分类、回归、关联分析、聚类、异常检测、文本挖掘、时间序列预测、特征分析等数据挖掘问题，着重介绍了决策树、k近邻、人工神经网络、线性回归、k均值聚类等当今广泛使用的二十多种算法，针对每一种算法都先以通俗的语言解释其原理，再使用开源数据分析工具RapidMiner加以实现。

本书适合在日常工作中大量接触数据的分析师、金融专家、市场营销人员、商务专业人士等阅读。

◆ 著　〔美〕Vijay Kotu　Bala Deshpande
◆ 译　严　云
责任编辑　朱　巍
责任印制　彭志环
◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路11号
邮编 100164　电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京圣夫亚美印刷有限公司印刷
◆ 开本：800×1000　1/16
印张：21.25
字数：476千字　2018年1月第1版
印数：1–3 000册　2018年1月北京第1次印刷
著作权合同登记号　图字：01-2015-3675号

定价：99.00元

读者服务热线：(010)51095186转600　印装质量热线：(010) 81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字20170147号

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

版 权 声 明

Elsevier (Singapore) Pte Ltd.
3 Killiney Road, #08-01 Winsland House I, Singapore 239519
Tel: (65) 6349-0200; Fax: (65) 6733-1817

This translation of *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner,1st* by Vijay Kotu and Bala Deshpande was undertaken by POSTS & TELECOM PRESS and is published by arrangement with Elsevier (Singapore) Pte Ltd.
Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner,1st by Vijay Kotu and Bala Deshpande 由人民邮电出版社进行翻译，并根据人民邮电出版社与爱思唯尔（新加坡）私人有限公司的协议约定出版。

《预测分析与数据挖掘：RapidMiner 实现》（严云 译）
ISBN: 9787115473660

献　　词

献给为开源软件的发展做出贡献的人们。

谨以此书献给世界上所有才华横溢又慷慨大方的开发者，他们仍在继续为开源软件工具增加巨大的价值。没有他们，这本书不可能面世。

序

每个人都能成为数据科学家，也应该是数据科学家。通过阅读本书，你会了解到为什么每个人都应该是数据科学家以及如何成为一名数据科学家。当今世界，如果不首先理解可用的数据就作出任何复杂决策，应该会很尴尬。成为“数据驱动型组织”是发展趋势，而且往往是大幅改进业务成果的最佳办法。因此，帮助我们快速成功的“利器”正迅速地推陈出新。建立数据仓库并基于数据仓库搭建数据分析及汇报的平台，在许多大企业里已经成为常态，而这种巨变仅花了几年的时间。科技的发展日新月异，更加速了这个进程。事实上，许多数据发现工具令业务用户不再需要 IT 部门的指导，就能自己独立搭建数据仓库以及数据分析平台。尽管我们已经能够基于过去的数据有效地解答问题，但整个时代的车轮正向新的方向驶去：如果我们还能预测未来，岂不美哉？这正是高级分析和数据科学的初衷，也就是把注意力从过去转移到未来，并且积极地调整业务运作以提高收益。

下面用几个例子来解释我刚才提到的巨大转变。

- 传统商务智能系统回答的问题：*去年流失了多少老客户*？尽管这个问题还是有意义的，但知道答案的时候未免为时已晚，再想挽回客户已经无能为力了。然而，预测分析却能告诉你*未来10天内我们可能流失多少客户*，并且可以指导你*如何提早应对以留住客户*。
- 传统商务智能系统回答的问题：*过去最成功的营销策略是怎样的*？虽然问得也有意义，但这个问题的答案对你为即将上市的产品所设计的策略提供的帮助非常有限。相比之下，预测分析能够明确建议*如何针对每一位潜在客户定制吸引他们的营销策略*。
- 传统商务智能系统回答的问题：*过去生产停滞问题出现的频率和原因*。虽然这也实用，但由于这对于现在并不是最优的，肯定无法改变收益下降的现状。但是，预测分析则能够精准地告诉我们*什么时候以及为什么流水线机器会出故障，什么时候应更换零件，而不是在出问题后无奈地积压订单*。

这些都是非常有价值的问题，而提前获知这些问题的答案将很可能为你的业务流程带来积极的影响。而且这并非天方夜谭，如今基于过去的数据和数据内的固有模式，我们能够在一定程度上预测未来。不过，为什么并不是每个公司都能充分利用这种方法呢？原因就是缺乏数据科学技能。

执行高级分析（比如预测分析、数据挖掘、文本分析、必要的数据预处理）需要高级技能。如果一个人不仅拥有统计学博士学位、优秀的编程能力，而且还精通商业领域的每个实际业务问题，那么他将被视为数据科学家。显然这样集通才与精才于一身的人物实在凤毛麟

角。事实上，麦肯锡预测，截至 2018 年，仅在美国，数据科学家的缺口就将达到 180 万。这真的是很尴尬的境地：一方面，我们明知道预测未来能力的经济价值，也拥有现成的数据科学方法；然而另一方面，我们并没有这样的人才，又何谈预测未来。目前来看，唯一的出路就是*对高级分析方法进行解耦*。我们应该让更多的人具备构建预测模型的能力，这些人包括业务分析师、精通 Excel 的用户、玩转数据的业务经理等。显然我们没有魔法把他们变成数据科学家，但是我们可以设计工具，并告诉他们如何使用，让他们像*真正的数据科学家一样工作*。这本书的要义也在于此。

我们处在采用现代分析方法的时代，“大数据”使得人们对答案的需求越发强烈。我们应该意识到一点，大数据不仅指数据的量大，也指复杂程度高。数据越多意味着数据的组织形式越来越新、也越来越复杂。未结构化的数据需要新的存储与检索手段。另外，有时候由于数据产生的速度非常快，我们真正需要存储的并不是原数据，而是直接在原数据上分析所得到的结果。实时数据分析、数据流挖掘、物流网等已经成为了现实。与此同时，我们还应该清楚的一点是，数据本身并没有价值，有巨大价值的是隐藏于数据背后的模式。而发掘这些宝藏的能力绝不能只属于数据专家，每一位分析师、业务经理都应该具备。如果我们能解耦高级分析方法，就能移除数据科学应用的瓶颈，立刻获得巨大的商业价值。

这种转变对于本来就是数据科学家的人而言也是好消息。如果业务分析师、高级 Excel 用户、精通数据的业务经理都能够独立解决各自 95% 的高级分析问题，那么这对于稀缺的数据科学家来说是一种解脱。过去在数据科学家眼里不过是玩转数据表的事情，如今则下放给业务分析师。这样一来，一方面整体业务运作会更快，另一方面数据科学家们终于能专注于更富挑战性的问题（比如开发新的算法），而不是为了业务枯燥地重复造轮子。

我们开发 RapidMiner 也正是出于这一目的：让非专业人士获得与数据科学家一样的分析能力。我们要帮助用户高效地分析数据，真正让分析和部署在弹指一挥间就能完成。RapidMiner 也能让业务分析师以及数据科学家们更快地探索隐藏的模式，实现新的商业价值。它打开了隐藏在整个商业市场里的巨大藏宝盒。我衷心希望 Vijay 与 Bala 的这本书也能促成分析模式的转变，打破你使用数据科学时的瓶颈。最后希望通过这本书，你能够发现一片新的天地，走向成功。好好享受这段奇妙的发现之旅吧。

Ingo Mierswa

RapidMiner CEO 与联合创始人

前　言

大多数新兴技术都要经历被高德纳咨询公司称为“光环曲线”的过程。这个曲线是用来比较新兴技术“光环化”程度与其实际产生的效益的一种方法。光环曲线主要有三大时期：过高期望的峰值期，幻灭失望的低谷期，实质生产的高台期。第三个时期指的是技术成熟且带来价值的时期。预测分析（本书写作时）正处在其“光环曲线”的成熟期。

那么，这是不是意味着该领域已经停滞不前了，或者已经到达饱和点了呢？非也。相反，这个学科的应用已经超出了起初的市场营销范畴，开始进军科技、互联网、医疗健康、政府管理、金融和制造业等领域。所以，不同于许多已有的数据挖掘和预测分析方面的图书，它们可能或专攻理论知识，或着眼于与市场营销相关的应用，本书旨在展示出这个令人振奋的领域里更广泛的用例，同时介绍一系列不同的实际应用和实践方法。

对于如何形容数据的增长势头，我们真的已经穷尽辞藻了。简单而言，技术的革新促成了这样的需求：以有意义的方式去处理、存储、分析和理解海量、多样的数据。数据的规模以及多样性对组织迅速揭示隐藏趋势和模式提出了新的要求，而这正是数据挖掘方法变得至关重要的原因。它们逐渐介入到了诸多有关商务与政府职能的日常活动中，比如识别出哪些顾客更有可能转投他处，又如借助社交媒体信息定位流感。

数据挖掘是一大类扎根于应用统计和计算机科学的方法，其过程包括很多步骤：确定问题、理解数据、准备数据、应用正确的方法构建模型、解析结果和搭建流程以部署模型。本书旨在全面概述这些用于揭示模式、预测结果的数据挖掘方法。

所以，这本书究竟要讲些什么呢？非常广泛。预测分析是一门能够把未来的不确定性转化为有意义的概率的学科。本书涵盖了许多聚焦于该领域的重要方法，同时还包括了数据挖掘（一个有些烂大街的术语）里更宽广的领域。数据挖掘也包括被称为描述性分析的方法。本书约三分之一的内容主要介绍数据挖掘里描述性的一面，而余下的部分则专注于其预测性的一面。目前被广泛应用的最常规的数据挖掘任务都有详细的介绍：分类问题、回归分析、关联性问题、聚类分析，以及异常检测、文本挖掘、时间序列预测等有关技术。本书的目的在于引领感兴趣的读者一览这些令人振奋的领域，同时提供有足够深度的技术以帮助积极进取的读者在自己的工作中实践这些方法。

关于本书

本书的目标有两个：以易于理解的方式梳理数据挖掘方法背后的基础知识，同时帮助所有人基本掌握一些数学知识，做到无需编写一行代码就能在自己的工作中实践这些方法。尽

管可以实现算法和应用开发的商业软件有很多，但解决数据挖掘问题的方法都是类似的。我们想挑选一款功能齐备、开源、具有图形用户界面的数据挖掘工具，这样读者就可以一边轻松消化这些概念，一边亲自实践数据挖掘的算法。RapidMiner，一个数据挖掘与预测分析的先进平台，正好满足需要。如此一来，我们可以搭配着它去实践接下来每一章里介绍的数据挖掘算法。这款工具最棒的一点在于开源，这意味着除了花时间之外读者无需花一分钱就能使用它开始学习数据挖掘。

读者对象

本书讲解的内容和用例是为那些在日常工作中大量接触数据的商务及分析专业人士量身打造的。通读本书，读者将会对用于预测和模式识别的各种数据挖掘算法有全面的认知，针对给定数据问题能选出正确的解决方法，对于创建通用的分析过程可以十拿九稳。

我们尝试以清晰的逻辑讲解主要知识，着重介绍如今正在广泛使用的二十多种重要算法。算法的内容将会按照以下框架来呈现。

(1) 每一种算法的高级实际用例。

(2) 以通俗的语言解释每一种算法的原理。当然，许多算法都深深根植于统计学或计算机科学，不过我们的讲解极力实现一种平衡，既恪守学术的严谨，又让不具备数学背景的读者容易接受。

(3) 通过讲解常用的设置选项，我们会仔细地介绍如何使用 RapidMiner 去实现算法。若是有必要，我们会把章节开头介绍的用例按照以下框架来拓展：描述问题，概述目标，应用章节里讲解的算法，解析结果，以及部署模型。最后，尽管我们讲解应用时的确采用了“攻略”的行文方式，但这本书绝对不是 RapidMiner 软件的用户使用文档，更不仅仅是一本手册。

今天或是不远的将来，分析师、金融专家、市场营销人员、商务专业人士或是任何分析数据的人，都极有可能在他们的工作中应用这些高级分析方法。对于不参与实际数据分析过程的业务主管而言，重要的是要知悉这些高级分析的长处和短处，这样他们才能提出正确的问题，设立合理的期望目标。基本的电子表格分析法以及借助标准商务智能软件进行的数据分解与切割，虽然仍将作为商业数据探索的基石而存在，尤其是针对过去的数据而言，但数据挖掘和预测分析对于商业数据分析的完整构架却是不可或缺的一环。数据挖掘和预测分析的商业软件支持图形用户界面，注重实践应用而非算法的内在原理，从而稳固了二者的地位。我们不仅仅勾勒出理论的框架，而且还会手把手指导如何实践重要的算法，以此拓宽预测分析和数据挖掘的受众面。这正是我们的核心愿景。希望本书能够有助于达成这个目标。

Vijay Kotu

Bala Deshpande

致谢

在一个人愿意竭力去做的事情里，最有趣也最有挑战的是写书。我们严重低估了它需耗费的精力以及它所带来的成就感。当我们花大量时间写这本书时，我们的家人给予了极大的包容。若是没有他们的支持，恐怕这本书不会问世。我们想要感谢 RapidMiner 团队。他们在很多事情上提供了巨大帮助，从技术支持到审阅章节内容，再到回答关于他们产品的问题。我们要特别感谢 Ingo Mierswa 作序，为本书开篇。特别鸣谢技术审稿人提出了全面而又深刻的意见：来自 Slalom Consulting 的 Doug Schrimager、来自L&L Products 的 Steven Reagan，以及来自 RapidMiner 的 Tobias Malbrecht。感谢来自英特尔的 Mike Skinner 就模型评估那一章提供的专业指导。我们获得了 Morgan Kaufmann 出版团队的极大支持和协助：Steve Elliort、Kaitlin Herbert 和 Punithavathy Govindaradjane。感谢同事和朋友就这本书展开的建设性讨论以及提出的建议。

目　录

第1章 引 言

近年来，预测分析领域愈发火热。然而，以预测分析为分支的数据挖掘，其流行程度已日趋稳定。这门学科近年的发展态势与流行程度固然有目共睹，但其实它背后的科学至少已有四五十年了。从第一个登月计划以来，工程师和科学家们就一直在使用预测模型。人类是一类始终“向前看”的物种，而预测性科学恰好折射出人类这种好奇的本性。

那么，今天究竟谁在使用预测分析和数据挖掘呢？谁又是最大的用户呢？其中有三分之一的应用集中在市场营销领域（Rexer, 2013），比如客户定位与划分、客户赢取、客户流失和客户生命周期价值管理。另有三分之一受银行业、金融服务、保险业（BFSI）主导，包括用于诈骗识别和风险评估等。最后的三分之一则广泛分布于各行各业，包括制造业、科技互联网业、医药业、政府、学术界。其应用从传统的销量预测到商品推荐再到选举人气建模等。

在科学与工程领域，预测建模是将物理或化学原理应用于模型开发。本书所介绍的预测建模与之不同，是基于经验知识，更具体地说，是历史数据。我们收集、存储、处理数据的能力按摩尔增长曲线增强，这意味着硬件计算能力每两年翻一倍，因此数据挖掘在各种领域里有了越来越多的应用。不过，大部分早期工作是市场营销领域的研究人员做的。Olivia Parr Rud 在其 *Data Mining Cookbook* 一书中写了一段有趣的轶事，讲述了她在 20 世纪 90 年代早期为了构建 Logistic 回归模型如何花费了 27 个小时。重点在于，因为模型构建工作中相当大的一部分是数据准备，所以预测分析的整个流程需要非常精心的安排。于是她用了整整一周的时间做数据准备，最后模型被提交到一台配有 600MB 硬盘的 PC 上，她还花了整个周末运行（同时祈祷不会出现死机）。不到 20 年的时间，科技已经取得了长足进步。面向成百上千条记录（即样本），构建一个具有上百个自变量的 Logistic 回归模型，如今用一台笔记本电脑在几分钟内就能完成。

然而，数据挖掘流程从那时起就再没变过，并且在可预见的未来也不大可能发生太多改变。在算法能着手处理数据之前，我们仍然需要将大部分精力用于数据的准备、清理、清除或标准化，以便从数据中获取有意义的结果。但也许存在可变余地的是，在这流程中究竟能有多少环节可以实现自动化。虽然现在这个流程是重复的，并且要求分析人员知道最佳实践，但也许很快就会出现聪明的算法替我们完成这些工作。自动化将使得我们可以专注于预测分析里最重要的方面，即解释分析结果以作出决策。同时，这将使得更多的分析人员和业务用户能够从事数据挖掘。

那么数据挖掘由什么构成？是否存在一套必须掌握的核心步骤和原理？最后，这两个术语 —— 预测分析和数据挖掘，究竟有何不同？下一节中将给出更加正式的定义，但在此之前

不妨透过当前的调查（Rexer, 2013）来了解当今数据挖掘者的实战经验。结果是，现在绝大多数的数据挖掘者为了达成他们的目标，仅仅使用着少数几个强大的方法，如决策树（第 4 章）、回归模型（第 5 章）、聚类（第 7 章）。甚至 80/20 法则在这里仍然适用：多数数据挖掘任务可以通过少数方法实现。然而，正如 80/20 法则所述，位于分布中长尾部分的大量方法，虽不常用，却是真正价值所在。针对你的需要，最佳解决方案可能是某种相对少见的方法，或者是几个不常用方法的某种组合。所以系统地学习数据挖掘和预测分析将有所裨益，而这也正是本书将会帮助你实现的。

1.1 什么是数据挖掘

简而言之，数据挖掘就是寻找数据中的有用模式。作为一个流行术语，数据挖掘有各式各样的定义和标准。数据挖掘也被称为知识发现、机器学习、预测分析。但是，在不同的背景之下，每个术语有着不同的含义。在本章中，我们会对数据挖掘进行总体概述，并指出它的一些重要特性、用途、分类以及常用方法。

数据挖掘从*数据*出发，而数据的形式可以简单如一个数组，由几个数值型观察值组成；也可以复杂如一个矩阵，由百万条具有上千个变量的观测值组成。为了挖掘数据中有意义且有用的结构，我们进行数据挖掘时会使用一些专业的计算*方法*。它们源于统计学、机器学习、人工智能等领域。数据挖掘这个学科与以下众多有关领域并存且紧密相连：数据库系统、数据清理、可视化、探索性数据分析、性能评估。通过探讨数据挖掘的一些关键特性和动机，可以进一步定义数据挖掘。

1.1.1 有意义模式的提取

数据库中的知识发现是一个重要的过程，它可以识别出数据中可用、新颖、潜在实用和本质上易于理解的模式或者关系，我们以此作出重要决策（Fayyad et al., 1996）。“重要”一词道出了数据挖掘与直接的统计计算（如均值或标准差的计算）之间的不同。数据挖掘涉及推断以及反复提出多种不同的猜想。数据挖掘的一个重要方面是对数据集模式的*泛化*。这个泛化过程必须是可靠可信的，无论是对于用来观察模式的数据集，还是新的未知数据集而言都是如此。数据挖掘同时也是一个步骤明确的流程，每一个步骤都有一系列的任务。“新颖”一词则强调，我们借助数据挖掘方法所挖出的数据模式通常都是未曾知晓的。数据挖掘的最终目标是得出潜在有用的结论，作为分析人员的决策准则。

1.1.2 代表性模型的构建

统计学中，模型是一种表现形式，代表了数据中各个变量之间的关系。它描述了数据中一个或多个变量是如何与其他变量相关的。而建模是一个过程，是对观测数据集进行有代表性的概括。举个例子，我们基于信用评分、收入水平、申请贷款金额建立一个模型，确定贷款