不只介绍R语言

更深入数据挖掘的本质：探寻数据背后的逻辑，挖掘人们的欲望、需求及态度

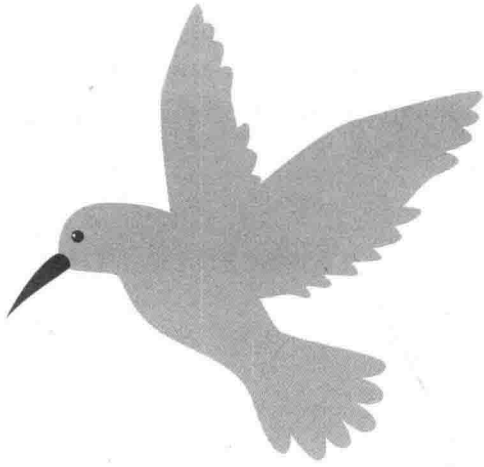
Broadview®
www.broadview.com.cn

# 探寻数据背后的逻辑

## R语言数据挖掘之道

宋云生 张坚洪 黎新年◎著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn



# 探寻数据背后的逻辑

## R语言数据挖掘之道



宋云生 张坚洪 黎新年◎著

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

## 内 容 简 介

数据分析、数据挖掘的本质是探寻数据背后的逻辑，挖掘人们的欲望、需求、态度等。本书不仅仅教会读者如何掌握数据挖掘相关技能，更教会读者如何从数据挖掘结果中分析出更深层次的逻辑。

本书主要介绍使用 R 语言进行数据挖掘的过程。具体内容包括 R 软件的安装及 R 语言基础知识、数据探索、数据可视化、回归预测分析、时间序列分析、算法选择流程及十大算法介绍、数据抓取、社交网络关系分析、情感分析、话题模型、推荐系统，以及数据挖掘在生物信息学中的应用。另外，本书还介绍了 R 脚本优化相关内容，使读者的数据挖掘技能更上一层楼。

本书适合从事数据挖掘、数据分析、市场研究的工作者及学生群体，以及对数据挖掘和数据分析感兴趣的初级读者。



**图书在版编目（CIP）数据**

探寻数据背后的逻辑：R 语言数据挖掘之道 / 宋云生，张坚洪，黎新年著. —北京：电子工业出版社，2018.8
ISBN 978-7-121-33861-8

Ⅰ. ①探… Ⅱ. ①宋… ②张… ③黎… Ⅲ. ①数据采集－研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2018)第 049541 号

策划编辑：王　静
责任编辑：石　倩
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
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：787×1092　1/16　印张：27　字数：742.4 千字
版　　次：2018 年 8 月第 1 版
印　　次：2018 年 8 月第 1 次印刷
定　　价：89.00 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）88254888，88258888。

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。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

# 序言

# 提纲挈领式心诀：一名数据挖掘工程师的成长之路

## 我的学习之路

不知不觉毕业两年多了，有一些大音如霜工作室的读者总想了解一下我是怎么学习数据挖掘、数据分析的，下面就综合大家常见的问题分享一下自己的经历、经验。

我不是学数学的，也不是学计算机的，研究生的专业是植物学，而且方向是植物分类，可以说很难和数据挖掘、市场研究等领域扯上关系。唯一能扯上关系的也就是我的舍友是做生物信息学研究的。

说一句丢人的话，在读本科时上的 SPSS 课我都不知道在讲什么。那时没考过计算机等级考试，原因是我每分钟打汉字的速度都不过关，讲这么多，只是为了告诉读者，我的基础并不扎实。

需要说明的是，我的英语还不错，在大一和大二分别通过了英语四、六级考试（而其他科目则学得比较一般，因为我每学期只有一两个主要学习目标），在大四我读了很多英文文献。因为在读研究生期间需要查阅大量文献，我需要给这些文章建立一个数据库，于是年少无知的我就选择了 Access。选择 Access 的原因并不是我比较熟悉它，而是我的老师用它，我至今也不会太多的操作。这应该算是我开始接触数据分析了。

使用高级语言时，记不住函数不要紧，但是你要有很强的搜索能力。

之所以讲这一段经历，不是为了说明我起步晚，而是为了说明建立 Access 文献数据库锻炼了我的英文搜索能力。我一碰到问题，就在 Google 里搜索，很快就能找到答案。于是 Google 几乎成了我的眼睛，真正做到了用 Google 搜索、发邮件、社交、阅读和写作。在公司里曾经传说，如果是连我都搜不到的内容，那么别人更不可能搜到。有些年轻人就怕英文，我并不是崇洋媚外，客观地想一想，现在的很多知识都是从欧美起源的，如果你连这门世界语言都不掌握，那么你获得的资料永

远都是二手资料。另外，无论你是找函数还是找包、模块，抑或是为问题寻找答案，使用 Google 进行英文搜索会为你节省很多时间。掌握这门语言并不需要你听、说、读、写样样精通，而是将其作为一种工具，应用起来比较方便就可以了。

要善用英文搜索，原因很简单，你所用的编程语言或软件大多是外国人构建的，并且在国外已经普及，相关的问答社区早已完善，你碰到的问题可能早就有人解决了。

在搜索文献的过程中，我喜欢上了《经济学人》的 *Graphic Details* 栏目，发现其绘制的图表非常漂亮、专业，于是我就开始学习 Excel，尽自己所能将 Excel 图表做得更漂亮、更专业，这些经历为我日后做数据可视化打下了坚实的基础：我知道了商务色彩搭配及图表要简洁、易读等原则，我知道怎么使自己的图表特色鲜明。后来看了大前研一先生的著作，了解了专业精神，我曾经写下这样一句话，以勉励自己：

所谓专业，即每一个细节都经得起推敲。

有一天，舍友看到我用 Excel 作图，嘲笑我孤陋寡闻，推荐我学习 R 语言，然后我就开始搜寻一些 R 语言入门读物进行阅读，慢慢地知道了关于这门语言的粗浅知识。

这个时候已经到研二下半学期了，我需要为自己未来的工作做打算了：是步入园林行业还是就此转行？必须做一个决断。我发现自己真的对植物分类不感兴趣，而我做家教的学生的妈妈是星空传媒的一个经理，平时待我很好，她说毕业可以介绍我去做市场研究。我了解了一下市场研究，发现其中涉及一些数据分析的内容（现在看起来很简单），于是，我从此决定踏上数据分析这条“不归路”。

为了快速上手、熟悉统计学知识，我并没有马上深入地学习 R 语言，而是像以往一样懒懒散散地学习（后悔当时没有实战学习）。我通过搜索发现，市场研究的岗位大多将熟练使用 SPSS 作为硬性要求，偶尔也会要求熟悉 R 语言，但 SPSS 对我来说更容易上手，于是就开始学习 SPSS。SPSS 帮助我巩固了统计学知识，当学习完简单的统计学知识后，我发现 SPSS 不够灵活，很多功能不够用，做出的图表很难看（这对于我来说是无法忍受的），因此，网络上有一些人鄙视 SPSS，但很推崇 R 语言。于是我决定要深入地学习 R 语言。我先将 SPSS 的功能在 R 中做了一遍，有了一些自己的理解后，我开始在自己的论文里做一些数据分析的内容。

现在想来，如果我直接在实战中学习可能会节省更多的时间。

实战更能锻炼技能水平。

研二快结束了，开始找工作了。我找工作的目的很明确，如果工作不是做数据分析、数据研究的，那么我宁愿放弃这个工作的机会。非数据研究的岗位我也不去面试，这样又省下了大量的时间学习。

## 在工作中学习

2013 年毕业后，我去了一家医药市场研究公司，当时的工作并不太忙，我有大量的时间学习。但这时也暴露了我的弱点，公司的数据并不是很规整（raw data）的，往往需要标准化等，而且数据规模也不再是之前练习时那么小，在面对这些脏数据、大一点的数据时，我的数据清洗水平显得捉襟见肘。周围的人都是 Excel 高手，如果跟着他们学，估计也能成为高手，但是我一定要在 R 中做数据清洗整理，反正公司的工作不是很多，我就一点点地学习和积累，这样我的数据处理能力就逐渐扎实起来了。其间我用两天读完了《异类》这本书，感触很深，阅读经历已经写成一篇文章在我们的公众号里分享了。

> 任何一个工具在刚开始学习时都会觉得它很糟糕，其实这并不是工具的问题，而是自己的知识体系跟不上节奏，或者是它的很多方法与自己原有的认知相反，这时不要急于否定它，而是要深入地学习它。知识体系是一个积累过程，为自己准备一万个小时计划吧。

我们公司当时在做 BI（商业智能），于是我接触了市面上常见的 BI 工具，包括 Tableau、QV 等，我熟悉它们的优、劣势，也熟悉它们的数据可视化效果。因为需要将 R 语言的页面融入 BI 中，所以我熟悉了 shiny 包，做了一些页面，但我渐渐看到 R 语言在做这些通用语言的工作时所暴露的缺点，于是开始接触 Python。

后来，我们的合作公司的总经理听说我比较熟悉 R 语言，就向我请教，我们一起讨论了 R 语言和数据挖掘。得知他们在做文本挖掘，于是在我闲暇时间开始学习中文文本挖掘的内容。没有成型的数据和书，我就看帖子，去一个个地实现，然后积累经验，这时我对 R 的操作算得上非常熟练了，从实现到速度优化（并行计算等）等也已经非常熟练，积累的代码也非常多了。

后来，那个经理找我做医院处方数据挖掘工作，之后，他请我去负责法院文本数据挖掘，我没去，但成了他们的外援，仍然没收过钱，他们搭建的一台服务器也帮助我了解了不少 Linux 的知识。

> 刚开始，锻炼自己的机会远远比钱重要，反正自己闲着也是闲着，但是这种情况只适用于刚开始。

后来，我们公司推出了微信公众号平台，我开始给公司的公众号写文章。其间我为公司的公众号写了多篇关于综合排名的文章，阅读量最高达到 4 万多人次，当时公众号的粉丝才 2 000 人左右。后来我又制作了评价医院市场趋势的综合指标体系，现在公司也一直在沿用这套指标，这些工作中的小点子都是我在公交车上想出来的。

> 除要把工作当成一种谋生手段外，还必须有极大的兴趣，要么不做，要么做好。

另外，我在公交车上读完了 *Data Mining with R learning by case studies*、*Machine Learning for Hackers*、*R Graphics Cookbook* 等书籍，之所以提这三本书，是因为我不止读过一遍，这三本书很有特色，第一本帮助我学习了各种算法，第二本帮助我接触了实际应用中的知识，第三本帮我熟练了 ggplot 的函数及图表元素结构。我开始学会利用零散的时间，坚持积累，也开始学习高度自律。

古之成大事者，不唯有超世之才，亦必有坚韧不拔之志。

——苏轼

其实，我一直幻想着有一个自己想写什么就写什么的平台，于是，我和小伙伴们开通了微信公众号，直到现在，我们更注重文章的可读性、趣味性，而不仅仅是技术，但是每一篇文章都可以作为一个小项目让希望学习数据分析的读者能锻炼一下自己的技能。

经常有读者问学数据分析就一定要学编程吗？以及为什么要看英文资料？针对这两个问题，我写下了这样一段对话，希望你能在对话中找到答案。

为什么学习数据分析？

赚钱！

什么样的人容易赚钱？

技能比别人高的！

英语是不是一般人的难关？是不是大家都想学习傻瓜式操作软件？

是！

那么如果大家都这么想，你应该怎么做？

很明显，你要做其他人不愿意做的事情，才能赚到别人不能赚的钱！

作为数据分析师，一定要将自己和技术区分开，分析数据、挖掘数据本质上是探寻数据背后的人心，挖掘人们的欲望、需求、态度等，所以数据分析师还要尽量拓宽自己的视野和知识结构，尽自己所能博览群书。

我的经历大体如此，中间会有各种迷茫、各种苍白无力，但是如果你缺少什么，就去搜集资料，做出判断，努力去争取，这一点总不会错，千万不要一味地否定你不了解的东西，这也是我对待未知领域的态度。

作为一名技术人员，要让自己的知识时刻在进步！这是一种宿命。

作　者

# 前言

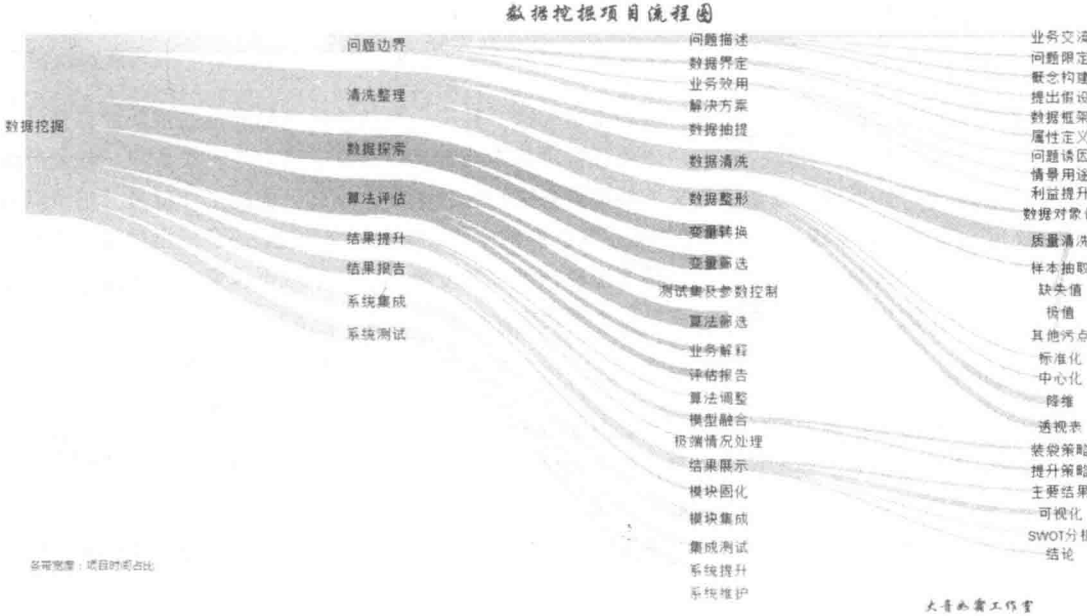
## 什么是规范化的数据挖掘流程

人总是被自己日常从事的工作所蒙蔽双眼，看不到事态发展的整体面貌，为了手里的工作而工作，这就是所谓的迷失吧。一个数据挖掘项目不仅仅是数据挖掘工程师手中的一部分工作（虽然它是工作的核心），作为一个力求向上的人，要跳出来看看项目的全貌。只有对项目全程有了足够的了解，才能更加有效地使用数据挖掘、机器学习、数据分析的工具。

数据挖掘项目一般可以分为问题边界、清洗整理、数据探索、算法评估、结果提升、结果报告、系统集成和系统测试 8 个主要的模块。一般而言，可以尽量将这些模块合并，但无论怎么合并，它们在项目中都是不可或缺的（当然，有些项目并没有其中某些模块的需求，比如市场研究项目，它们可能就不需要系统集成）。从右图中可以看到数据清洗、数据探索、算法评估占据了项目的大部分时间，这也说明它们是项目的核心内容，缺了这 3 项，就不能再称为是数据挖掘项目了。

| 项目 | 项目分支 | 时间占比% |
|---|---|---|
| 数据挖掘 | 问题边界 | 10 |
| 数据挖掘 | 数据清洗 | 25 |
| 数据挖掘 | 数据探索 | 15 |
| 数据挖掘 | 算法评估 | 20 |
| 数据挖掘 | 结果提升 | 5 |
| 数据挖掘 | 结果报告 | 8 |
| 数据挖掘 | 系统集成 | 8 |
| 数据挖掘 | 系统测试 | 9 |

数据挖掘项目流程如下图所示。



### 问题边界

问题边界一般是项目的开头部分，可以分为 4 点。首先要和业务部门细致沟通，从业务背景中提炼出对业务问题的描述，限定项目要解决的问题，便于组织力量集中对这些问题设计解决方案。然后根据解决构想将业务问题转换为数据语言，限定将要使用的数据界限，搞清楚要牵涉哪些数据。之后为了吸引业务部门必须整理出业务效用，告诉业务部门如果解决这些问题能够得到哪些改善，完成业务部门哪些具体目标。最后要将以上问题整理成一个可行的解决方案。很多人忽略了这个阶段，其后果就是业务部门觉得挖掘出来的结果不是他们想要的，或者节外生枝补充各种相关的或不相关的业务问题，最终扭曲了项目本身，使工作反复无常。所以，在项目实施之前，非常仔细地沟通并制定一个完善的问题边界非常重要。

### 清洗整理

清洗整理是数据挖掘工程师非常熟悉的工作，但是，很少有人认识到这是项目中花费时间最多的部分，很多人会以为算法评估部分才会花费最多的时间。其实不然，如果数据清洗进行得不顺利，则将直接影响后面的工作和模型的效果。首先要设计畅通、高效的数据抽取程序，将数据从各种数据平台抽取出来供数据挖掘工具使用，然后进行数据清洗，将数据转化为数据挖掘工具便于处理的对象类型（在 R 里指 list、data.frame、arrary 等），再进行质量清洗，包括处理缺失值、异常值、其他污点（在文本挖掘中多如牛毛）等。之后要对数据整形，包括一些统计变化，例如中心化、标准化、降维等，更重要的是数据形状的变化。

还有一项就是数据抽样，面对大数据，在数据处理阶段就要进行抽样，不能因为要清洗一个点就清洗全量的数据，那样会花费大量的时间。不如抽取小样本进行测试，等进行完数据清洗程序后，再进行全量数据的整体清理，这样反而更加省时省事，这里的样本量需要尽量保证抽到足够多的问题数据，同时要让程序运行起来非常轻松、高效。

### 数据探索

数据探索要完成两个目标：变量转换和变量筛选。其中变量转换既包括变量的重新计算变形，也包括概念变量的构建，比如，在客户流失预警项目中要定义什么样的客户是流失客户，就会产生出一个新的变量。如果这个变量的定义不能用业务进行合理解释，那么下面的工作就是“瞎子点灯白费油”了。有些变量不仅不会对模型产生正向的影响，而且除了影响速度，还会降低模型的效果。显然进行变量筛选就非常重要了。谷歌预测流感模型筛选变量足以证明数据探索多么重要，而且在大数据环境下，数据探索已不再是一件轻而易举就能完成的事情了。

### 算法评估

算法评估是数据挖掘项目的灵魂。算法评估首先要求我们充分了解算法或模型的参数意义，然后需要预留测试数据集。模型评估不是仅仅比较模型结果的准确性是否存在差异（别忘了统计学教导我们比较差异时要判断差异的显著水平），所以，模型比较是对不同模型准确性均值的比较。算法筛选完成后，工作就告一段落了，这时要和业务部门一起对结果进行业务解释，不能进行业务解释

的数据挖掘结果就是为了数据挖掘而数据挖掘，这显然就是迷失在了项目中，遗忘了项目要解决的问题边界。最后要对结果进行完整的评估报告。评估报告是必需的，因为除了将它给领导看，更重要的是它能帮助你总结发现这个过程中可以改善的节点。

### 结果提升

首先要判断是否需要调整算法或模型，包括更换算法或调整参数。如果模型调整没有必要，那么就要考虑使用模型融合提高模型效率。模型融合的方法包括装袋（Bagging）和提升（Boosting）等，有些方式可能用业务解释起来比较困难，这也是数据挖掘工程师要考虑的问题之一，显然，有些问题选择可解释模型比较好。在项目中对一些极端情况最好另做处理。

### 结果报告

“丑媳妇也要见公婆，”分析结果报告最终要给业务部门的同事学习，教他们如何使用数据挖掘的结果进行业务分析和部署，其中主要成果要突出，吸引他们的眼球，一定要联系业务具体的困境或具体的业务情景，即所谓的对症下药。规律和结果必须通过易读的方式传达给受众，充满技巧的数据可视化是不二之选，将美妙的可视化图表嵌入具体的应用情景中进行宣讲，往往能达到事半功倍的效果，因此，在此处无论多么努力都不为过。SWOT 分析是业务部门最喜欢的分析方式，我们当然不能放过，以对方熟悉的方式表达自己的诉求，是交流的法宝。

模块固化这一步工作的快慢取决于之前的工作，如果之前已经考虑到后面要进行模块固化，那么就会将代码写得比较规范、注释良好，这种情况下就很容易将数据清洗、数据整形、变量转换、模型构建、结果输出等模块的内容固定下来，成为一个数据有进口及有出口的脚本文件。

### 系统集成

将固化下来的模块按照一定的秩序集成在一起，就成为一个分析的脚本体系。在这个体系中，有输入就有产出，中间不需要人工干预，是一个有序的自动化脚本体系。这一步考验数据挖掘工程师对每一步任务的理解。良好的模块集成可以提升整个系统的速度，减少后期维护的时间和次数。

模块集成后要与其他系统集成在一起，首先要和数据平台（数据库、Hive、Hadoop）对接，为分析模块提供数据来源和存储分析结果，同时要和前台展示对接，将结果可视化，让结果真正接触受众，即所谓的为决策者提供支持。

### 系统测试

这么一个“五脏俱全”的系统需要维护在所难免，总有一些极端情况会导致数据分析模块宕机，所以，代码一定要写得尽量规范，注释要尽量清晰，否则在维护时会有一种再造系统的感觉。关于规范请参看 *Google's R Style Guide*。

# 目录