面向大数据的挖掘方法研究

MIANXIANG DASHUJU DE WAJUEFANGFA YANJIU

赵 妍 著

电子科技大学出版社

面向大数据的挖掘方法研究

赵妍 著

电子科技大学出版社

图书在版编目(CIP)数据

面向大数据的挖掘方法研究 / 赵妍著. —
成都:电子科技大学出版社,2015.6
ISBN 978-7-5647-3015-4

Ⅰ.①面… Ⅱ.①赵… Ⅲ.①数据处理
Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2015)第 112223 号

面向大数据的挖掘方法研究

赵 妍 著

出 版: 电子科技大学出版社(成都市一环路东一段 159 号电子信息产业大厦 邮编:610051)
策划编辑: 谭炜麟
责任编辑: 谭炜麟
主 页: www.uestcp.com.cn
电子邮箱: uestcp@uestcp.com.cn
发 行: 新华书店经销
印 刷: 郑州宏达印务有限公司
成品尺寸: 145mm×210mm 印张 11 字数 203 千字
版 次: 2015 年 6 月第一版
印 次: 2015 年 6 月第一次印刷
书 号: ISBN 978-7-5647-3015-4
定 价: 26.00 元

前　言

大数据的爆炸式增长在大容量、多样性和高增速方面，全面考验着现代企业的数据处理和分析能力；同时，也为企业带来了获取更丰富、更深入和更准确地洞察市场行为的大量机会。对企业而言，能够从大数据中获得全新价值的消息是令人振奋的。然而，如何从大数据中发掘“真金白银”则是一个现实的挑战。大数据是现在非常热门的一个话题。从工程或者技术的角度来看，大数据的核心是如何存储、分析、挖掘海量的数据解决实际的问题。那么对于一个工程师或者分析师来说，如何查询和分析 TB/PB 级别的数据是在大数据时代不可回避的问题。

数据挖掘是帮助人们在海量数据中发现信息和知识的工具。近年来，数据挖掘技术成了商业智能的核心技术，被广泛应用到了诸多领域，引起了学术界极大的关注。数据挖掘是一个决策支持过程，技术基础是人工智能。目前数据挖掘主要利用人工智能中的一些算法和技术，包括统计分析、决策树、人工神经网络等来进行预测、模式识别、分类和聚类分析。

BP(Back Propagation)算法，即误差反传训练算法，以其良好的非线性映射逼近能力和泛化能力以及易实现性成为人工神经网络应用最广泛的训练算法。但是BP算法也有其明显的缺陷，即训练速度慢、容易陷入局部极值等。遗传算法(Gentic Algorithm)能在复杂而庞大的搜索空间中进行自适应地搜索，并且寻找出最优或准最优解，它具有算法简单、适用性和鲁棒性强等优点。将GA与BP网络结合，可以使BP网络系统扩大搜索空间、提高计算效率以及增强BP网络建模的自动化程度。也可以用GA来解决BP网络中一些难题，如对输入样本的高品质要求，解释神经网络的黑箱行为，减轻手工调整网络的负担等。GA通过对BP网络进行预处理，把关于解空间的知识内嵌到BP网络初始化状态中，使BP网络计算工作量大大减少。综合两者优点，用GA优化BP神经网络初始权值来建立预测模型，并应用于商业行为预测。支持向量机作为一种新的机器学习方法，是近40年统计学习理论研究成果的结晶，具有训练样本少、推广能力好、全局最优以及广泛的适用性等优点，已被应用于模式识别、信号处理、控制系统等多个领域，并取得极好的效果。在信息安全领域，攻击者往往是安全漏洞的发现者和使用者，因为要对一个系统进行攻击，如果不能发现和使用系统中存在的安全漏洞是不可能成功的。通过对网络数据包的截获，从这些数据报文中进行黑客攻击特征的提取，提取的报文信息作为SVM算法的输入特征向量，从得出的分类结果输出判断是否为黑客进行攻击。

目前，许多领域和行业需要越来越多的大数据挖掘和分析人员，高校研究生也是参与科研和开发的重要力量。因此，本书读者类型多样，适合在校大中专学生、研究生、科研人员和系统开发人员。本书是作者多年的研究成果的积累，其中第一章、第二章、第六章是苏玉召老师撰写的，同时也感谢王瑞静同学帮助收集资料。

由于作者水平和知识有限，对于书中不足和错误之处，恳请读者批评指正！

赵　妍

2015 年 3 月

目　　录

第一章 概 述

1.1 “网络”带来了什么

20世纪50年代初期，美国地面防空系统用通信线路将远程雷达与测量控制设备连接到同一台中央计算机上，这就是当今世界首次尝试将计算机技术与通信技术结合到一起。这为计算机网络的出现做好了技术准备，奠定了良好的理论基础。70年代中期，微型计算机的出现和微处理技术的发展，使得它对计算机在短距离间的通信要求也越发明显，局域网便应运而生了。计算机网络要求开始正式步入标准化的网络体系结构时代，美国IBM公司率先公布了SNA标准(System Network Architecture，系统网络结构)，在这之后，DEC公司也提出了DNA标准(Digital Network Architecture，数字网络体系结构)。在此期间，国际标准化组织(ISO)也成立了一个委员会，专门服务于开放系统互连分技术，并陆续制定了一系列的国际标准，用以促进计算机网络的发展方向越来越趋向于标准化。

80年代人工智能(Artificial Intelligence)被引入市场，它是研究、开发用于模拟、延伸和扩展人的智能的理论、方

法、技术及应用系统的一门新的技术科学，并显示出实用价值。信息化的发展使得人工智能相关技术更大地促使新的进步不断出现。人工智能已经并且将继续不可避免地改变我们的生活，网络发展的方向之一就是网络的进一步融合，所以加快信息网络建设在信息化进程中具有非常重要的意义。

近年来，信息技术的快速发展，特别是信息获取技术、物理信息系统、互联网、物联网、社交网络等技术的突飞猛进，引发了数据规模的爆炸式增长，大数据已经普遍存在，能源、制造业、交通运输业、服务业、科教文化、医疗卫生等领域都积累了 TB 级、PB 级乃至 EB 级的大数据，这些数据已经开始造福于人类，成为信息社会的重要财富。例如，著名的全球连锁超市沃尔玛每小时需要处理 100 余万条的用户请求，维护着一个超过 2.5PB 的数据库；在高能物理实验中，2008 年开始投入使用的大型强子对撞机每年产生超过 25PB 的数据；社交网络 Facebook 现已存储超过 500 亿张照片。大数据蕴含着巨大的价值，对社会、经济、科学研究等各个方面都具有重要的战略意义，为人们更深入地感知、认识和控制物理世界提供了前所未有的丰富信息。例如，著名国际咨询机构 Gartner 在 2012 年预测全球大数据相关产业的规模将达到 2320 亿美元；2010 年时代杂志刊载的医学界年度十大突破中，医疗科技公司 CardioDX 通过对 1 亿个基因样本的分析，最终识别出能够预测冠心病的 23 个主要基因；2009 年 Google 的研究人员通

过对每日超过30亿次搜索请求和网页数据的挖掘分析，在H1N1流感爆发几周就预测出流感传播；通过对微博等网络大数据的挖掘分析能够发现社会动态，预警重大和突发性事件。

大数据的迅速涌现及其巨大价值，已经引起国内外学术界、工业界和政府部门的广泛关注。美国等世界发达国家都制定和启动了大数据研究计划，投入大量资金支持大数据研究。我国对建设大数据管理基础设施的需求已经提出了指导性的方针。《国家中长期科技发展规划纲要(2006—2020)》指出："信息领域要重点研究开发……海量存储和安全存储等关键技术。"《国民经济和社会发展第十二个五年规划纲要》提出："重点研究……海量信息处理及知识挖掘的理论与方法……"

虽然目前大数据研究已经蓬勃兴起，但是工作主要集中在大数据的存储、管理、挖掘分析等方面，数据可用性问题没有得到足够重视。随着大数据的爆炸性增长，劣质数据也随之而来，导致数据质量低劣，极大地降低了数据的可用性。事实表明，大数据在可用性方面存在严重问题(以下简称数据可用性问题)。国外权威机构的统计表明，美国企业信息系统中1%～30%的数据存在各种错误和误差，美国医疗信息系统中13.6%～81%的关键数据不完整或陈旧。国际著名科技咨询机构Gartner的调查显示，全球财富1000强企业中超过25%的企业信息系统中的数据不正确或不准确。随着大数据的不断增长，数据可用性问

题将日趋严重，也必将导致源于数据的知识和决策的严重错误。数据可用性问题及其所导致的知识和决策错误已经在全球范围内造成了恶劣后果，严重困扰着信息社会。在美国，由于数据错误而引发的医疗事故，每年导致约98000名患者死亡，约占全部医疗事故致死人数的50%；由于数据错误和陈旧而引起的生产事故和决策失误，每年给美国工业企业造成约6110亿美元的损失，约占美国GDP的6%；美国零售业每年因标价数据错误而导致25亿美元的损失；在美国银行业，由于数据不一致问题而失察的信用卡欺诈在2006年就造成48亿美元的损失。据有关专家推算，在数据仓库项目的开发过程中，清理不洁数据通常需要花费30%～80%的开发时间和开发预算；数据可用性问题平均给每个企业增加的成本是该企业产值的10%～20%。此外，由于网络的普及，很多应用可以从不同的数据源抽取和集成信息，致使劣质信息产生和传播的风险达到了空前的水平。事实上，数据可用性问题是信息化社会中固有的问题。它们不仅在西方发达国家存在，而且在任何一个信息化社会都普遍存在。尽管我国尚未公布相关统计信息，我们没有理由相信我国不存在类似的问题。例如，我们通过对某国有大型企业信息中心的TB级数据的抽样检验，发现10%的信息存在各种类型的错误。综上所述，确保数据可用性是关系到大数据时代的国计民生、社会和谐等方面的一项重大战略任务，是圆满完成大数据管理基础设施建设、有效发挥大数据作用的重要前提。因

此，深入开展数据可用性基础理论和关键技术的研究具有重要战略意义。

1.2 “啤酒与尿布”的故事

在一家超市中，人们发现了一个特别有趣的现象：尿布与啤酒这两种风马牛不相及的商品居然摆在一起，而且这一举措居然使尿布和啤酒的销量大幅增加了。这并不是一个玩笑，这个著名的“啤酒与尿布”的故事产生于20世纪90年代的美国沃尔玛超市，沃尔玛的超市管理人员分析销售数据时发现了一个令人难于理解的现象：在某些特定的情况下，“啤酒”与“尿布”两件看上去毫无关系的商品经常出现在同一个购物篮中，这种独特的销售现象引起了管理人员的注意，经过后续调查发现，这种现象出现在年轻的父亲身上。

在美国有婴儿的家庭中，一般是母亲在家中照看婴儿，年轻的父亲去超市购买尿布。父亲在购买尿布的同时，往往会顺便为自己购买啤酒，这样就会出现啤酒与尿布这两件看上去不相干的商品经常会出现在同一个购物篮的现象。如果这个年轻的父亲在卖场只能买到两件商品之一，则他很有可能会放弃购物而到另一家商店，直到可以一次同时买到啤酒与尿布为止。沃尔玛发现了这一独特的现象，开始在卖场尝试将啤酒与尿布摆放在相同的区域，让年轻的父亲可以同时找到这两件商品，并很快地

完成购物；而沃尔玛超市也可以让这些客户一次购买两件商品，而不是一件，从而获得了很好的商品销售收入，这就是“啤酒与尿布”故事的由来。

当然“啤酒与尿布”的故事必须具有技术方面的支持。1993 年美国学者 Agrawal 提出通过分析购物篮中的商品集合，从而找出商品之间关联关系的关联算法，并根据商品之间的关系，找出客户的购买行为。艾格拉沃从数学及计算机算法角度提出了商品关联关系的计算方法——Aprior 算法。沃尔玛从 20 世纪 90 年代尝试将 Aprior 算法引入到 POS 机数据分析中，并获得了成功，于是产生了“啤酒与尿布”的故事。

“啤酒与尿布”的故事其实就是数据挖掘技术的应用。随着数据库技术的不断发展及数据库管理系统的广泛应用，大型数据库系统已经在各行各业普及，数据库中存储的数据量急剧增大。在大量的数据背后隐藏着许多重要信息，而这些重要信息可以很好地支持人们的决策。可是目前用于对这些数据进行分析处理的工具却很少。目前人们用到的主要是数据库的存储功能，而隐藏在这些数据之后的更重要的信息则没有充分利用。这些信息是关于数据的整体特征的描述及对发展趋势的预测，在决策生成的过程中具有重要的参考价值。数据库技术的日益成熟和数据仓库的发展为数据挖掘提供了发挥的平台。

由于数据存储技术的日渐成熟，数据库和联机事务处理(OLTP)已经被广泛应用于金融、证券、保险、销售以及

天气预报、工业生产、分子生物学、基因工程研究等各行各业，因而积累了大量数据。人们已经不满足于简单的统计分析，而需要发现更深层次的规律，提供更有效的决策支持。而传统专家系统靠人工获取知识这一“瓶颈”在日益膨胀的“数据山”面前显得更加无力。数据挖掘的对象是某一专业领域中积累的数据；挖掘过程是一个人机交互、多次反复的过程，挖掘的结果要应用于该专业。因此数据挖掘的整个过程都离不开应用领域的专业知识。目前数据挖掘技术在货篮数据(Basket data)分析、金融风险预测、产品产量、质量分析、分子生物学、基因工程研究、Internet站点访问模式发现以及信息搜索和分类等许多领域得到了成功的应用。一套金融风险预测系统一年可以挽回数千万美元的损失；如果你通过 Internet 访问著名的亚马逊网上书店，会发现当你选中一本书后，会出现“购买该书者中有百分之 XX 同时购买了 XX”的推荐。可见，数据挖掘技术已经步入人们日常生活。

因此，数据挖掘是应用需求推动下跨学科发展的产物，而且在近几年里迅速发展起来。这个领域的实质是智能技术与数据库技术的结合，不但为决策者提供知识和策略，而且为投资者带来经济效益。这就是数据挖掘技术产生的背景。

数据挖掘是一个决策支持过程，技术基础是人工智能。人工智能是通过模拟人类宏观外观的思维行为，从而高效率地解决现实世界的问题。目前数据挖掘主要利用

人工智能中的一些算法和技术，包括人工神经网络技术等来进行预测、模式识别、分类和聚类分析。人工神经网络有表示任意非线形关系和学习等能力，给解决这类问题提供了新思想和新方法。

1.3 大数据挖掘

2013 年 8 月 11～14 日，第 19 届知识发现与数据挖掘大会（ACM Conference on Knowledge Discovery and Data Mining，SIGKDD（2013））在美国芝加哥召开。大会吸引了来自全球 50 多个国家的 1200 多人参加，打破了历届大会的参会人数纪录。

SIGKDD 是数据挖掘领域的顶级国际会议，由 ACM 数据挖掘及知识发现专委会负责协调筹办。会议内容涵盖数据挖掘的基础理论、算法和实际应用。SIGKDD 的发展历史可以追溯到 1989 年开始组织的一系列关于知识发现及数据挖掘的研讨会 KDD（Knowledge Discovery and Data Mining）。自 1995 年以来，KDD 以大会的形式连续举办了 18 届。由于 KDD 的学科交叉性和广泛应用性，大会吸引了来自统计、社会网络分析、机器学习、大数据挖掘、数据库、万维网、生物信息学、多媒体、自然语言处理、人机交互及高性能计算等众多领域的专家、学者，影响力越来越大。SIGKDD 2013 是自 SIGKDD 2005 后，第二次来到芝加哥。大会从 2012 年起增加了暑期学校，2013 年

的暑期学校名为“大数据训练营”(Big Data Camp)。

2013年的大会主席是前通用汽车高级研究经理拉马萨米·尤瑟鲁萨米(Ramasamy Uthurusamy)博士和芝加哥大学的罗伯特·格洛斯曼(Robert L. Grossman)教授,程序委员会主席由来自德克萨斯奥斯丁大学的因德里特·迪伦(Inderjit S. Dhillon)教授和谷歌公司的耶和达·科伦(Yehuda Koren)博士担任,另外有50名高级程序委员会委员和300名程序委员负责论文评审。主会期间,除了学术研究论文,SIGKDD还设有面向工业和政府应用的专题研讨会以及工业应用博览的邀请报告会。此次大会的主题是“大数据挖掘”,邀请了相关领域的知名专家作大会主旨报告,包括微软的技术院士拉胡·罗摩克里希纳(Raghu Ramakrishnan)、在线教育系统Coursera的创始人、斯坦福大学的吴恩达教授、威斯康辛大学的史蒂芬·赖特(Stephen J. Wright)教授以及谷歌公司首席经济学家、加州伯克利大学的哈尔·瓦里安(Hal Varian)教授。

在研究热点方面,以社交网络和信息网络为中心的大数据分析成为热点,图挖掘、推荐系统、用户行为分析也吸引了很多科研人员。此外,值得关注的是,SIGKDD还吸引了工业界的广泛关注,参会单位不仅涵盖了几乎所有的大型IT公司,还包括很多传统行业的企业:谷歌、脸谱、雅虎、微软、IBM、推特、甲骨文、易趣、通用电器、迪斯尼研究中心、福特汽车、美国军事学院等企业和机构均在研讨会上发布了报告。毫无疑问,大数据挖掘和社交网络分析已