中 国 数 据 分 析 行 业 核 心 刊 物

数据分析

CHINA DATA ANALYSIS 用数据说话 · 做理性决策

中国商业联合会数据分析专业委员会 主办

《中国数据分析》会员特刊
2016年第01期 总第25期（季刊）
咨询热线：010-59000991 / 59000339
http://www.chinacpda.org
投稿请至 xiehui@chinacpda.org

China Data Analysis

中国数据分析
CHINA DATA ANALYSIS 用数据说话 · 做理性决策

中国数据分析 会员特刊 2016年第一期

本期目录CONTENTS

主办

中国商业联合会数据分析专业委员会

编委

冯伟 / 周子赫 / 宫辰 / 孙雪 / 杜长海(特邀)

出版时间

2016年第一期 04月出版

美工 / 设计

崔峻珩

联系我们

中国商业联合会数据分析专业委员会

地址: 北京市朝阳区朝外soho C座9层

电话: +86–10–59000991 / 59000339

传真: +86–10–59000991转 607

投稿

欢迎广大读者踊跃投稿，内容包括学术观点、教学体验、教学活动、学习感悟、实战经验、随笔文章等。 稿件附图格式为JPG或TIFF格式，大于1M，分辨率在300dpi以上。

感谢您对《中国数据分析》的支持!

投稿邮箱：xiehui@chinacpda.org

- 智能化，一键匹配
- 常用成熟的算法集成
- 操作简单，界面简洁
- 结果输出丰富，可视化度高
- 权威专家实时指导

数据分析 de 魔术师

Datahoop数据分析智能平台
助力您的事业腾飞

CREATIVE

登录 www.datahoop.cn了解更多信息

Datahoop
想你所想 做你所需

以最方便最快捷的方式解决问题
让您集中精力去关注更重要的业务

时间 Time
利用您的闲暇时间，实现化整为零，高效管理。

效率 Efficiency
最高效，最快捷的通过平台解决您所面对的问题，让您解放精力！

精准 Accuracy
海量的数据分析模型，与高效的算法，带给您无比精准的结论。

Datahoop
为数据而生，化整为零
随时随地，随你所用

Datahoop为无处不在的数据而生

- 一键智能匹配
- 操作简单 界面简约
- 成速算法集成
- 输出丰富 可视化度高

/ 从现在开始拥抱数据 /

近期，数据分析师培训系列丛书《CDA数据分析考试大纲》《CDA数据分析——零基础入门》《CDA数据分析实务》即将出版。有幸邀请到中国商业联合会数据分析专业委员会会长邹东生先生为本系列丛书作序，现将序文作为本期会刊的卷首语，与大家分享。

详文如下：

全球大数据时代到来了，不仅因为IT技术的变革使大数据得以产生和膨胀，不仅因为大数据分析使原来不可能的精确决策成为可能；中国大数据时代到来了，不仅因为国家将大数据提升为国家战略，不仅因为中国专业的数据分析师、数据分析师事务所正在崛起；大数据时代来到来更深的原因在于：人们探究科学、探究真理的努力从未停息。大数据将改变我们所有人、所有企业的行为轨迹和思维方式。

IDC咨询对2020年大数据市场的一组预测显示：到2020年，65%的大型企业会将自己武装成数据化公司；大数据分析市场将以23%的年复合增长率高速发展；利用大数据分析技术，各行业将在生产力方面节省超过4000亿美元；未来的全球2000强公司中将有85%的企业是借助数据化转型之机树立行业领先地位……

大数据时代，无论大家是否介意大数据可能产生的问题(如：个人隐私数据的泄露、人工智能带来的风险等)，大数据对人们的改变都无可避免，未来的世界是数据化世界，未来的企业一定是数据化企业，业务依靠数据而延展，人们依靠数据的分析能力来证明自己的实力。美国大数据人才需求分析报告显示：到2018年，美国数据分析师的人才需求将达150万人左右。我国相关部门统计：预计3–5年内，我国数据分析专业人才的需求将达100万人以上。

到目前为止，在我国所有的高校中，还没有真正意义上的数据分析专业，因为数据分析不仅仅需要数学、统计分析、信息管理等知识，更重要的是将数据决策的思维与企业的运维相结合，“做分析”比“做数据”更有意义、更重要。数据分析可以为企业带来的不是“概念”和“故事”，而是真正通过精确决策帮助企业实实在在的省钱和赚钱，这样的数据分析人才，在今天、在未来的商业价值都无可估量。

当然要成为真正的数据分析人才，掌握多元化的知识体系是基础，更要将知识与实践结合。现在获取数据的方式和途径越来越丰富，比如政府或行业的公开数据、互联网爬虫收集等，运作数据分析的方法和工具，对数据进行处理，然后分析、分析、再分析，尽力探寻数据背后的规律是每一个从事数据分析工作的个人进入大数据殿堂的必由之路。

中国商业联合会数据分析专业委员会作为我国大数据分析领域的行业带头人，深知数据分析人才的价值和重要，我们积极推动数据分析员（CDA）、数据分析师（CPDA）的培养体系的建立，希望通过本书帮助高校的大学生们提高对数据的兴趣、培养数据分析的素质，也希望由此能促进我国高校数据人才培养工作的开展。

面对越来越激烈的就业市场，选择一个朝阳产业、选择一个金领职业，是每个将要进入社会的大学生们需要谨慎思考的事情。如果将优秀职业的属性数据进行分析，你会发现数据分析的标签精彩而生动：企业对人才的需求将越来越旺盛、数据分析可以游刃有余的适应各个行业、薪酬诱人、数据分析经验日久弥新……

所以，如果你想成为我们中的一员，那就从现在开始拥抱数据吧。

中国商业联合会数据分析专业委员会

/ 北京 SALON /

文图 / 协会培训处 张楠 编辑 / 协会会员处 冯伟 插图 / 崔峻珩 日期 / 2016-03

3月12日，由中国商业联合数据分析专业委员会主办的2016年首次“用户画像”数据分析公益沙龙活动在北京朝阳冈措咖啡厅完美落下帷幕。

用户画像，又称作用户文档或用户模型，是对用户特征进行客观准确的描述，以此满足用户的个性化需求。抽象的定义可能大家都理解不了，换一种说法，用户画像就是贴标签！

孙雪老师通过形象的图片、生动的例子将用户画像的含义完美的诠释了出来。

那么用户画像应该如何设立、建模，它其中的重难点、应用须知都是什么？孙雪老师为大家耐心的解释。看看，大家的动作都好统一，拍，拍，拍！精彩的东西，不能错过！

孙雪老师为我们讲述了从含义到建立到采集的一系列知识，还给我们介绍了一个神奇的平台：Datahoop数据分析平台，有了它，我们不用再为技术苦恼，我们需要的只是数据精准分析以后的解读工作，能大大减少我们的工作量，提高工作效率，真是太棒了！

而接下来出场的李军老师为我们带来的则是用户画像的实际应用。

通过宏观、中观、微观不同层面的逐步深入，面面俱到的分析用户画像在企业应用中的实例，由此过度出用户画像的实际应用。

李军老师不但给我们悉心讲解了用户画像在企业中的实际案例和应用，还明确的指出了我们所面临的挑战，并结合他的专业角度为我们提出了不少中肯的意见以及建议。

用户画像是大数据时代衍生出来的产品，它很神奇，可以帮助我们完成个性化的定制，目标人群的细分，简直是无所不能的存在。但是通过沙龙知识的学习，我们一层一层的拨开迷雾，了解到用户画像的真谛。大数据，它的计算建立在理性的设计的基础上，它很实际，是用户最真实的反应，所以它才能如此精准定位。

这就是大数据，人人都可以了解的大数据。

fin

/ 河北 SALON /

文图 / 河北授权管理中心　编辑 / 协会会员处 周子赫　日期 / 2016-03

3月20日，由数据分析师(CPDA)河北授权管理中心主办的2016年首次“用户画像”数据分析公益沙龙活动在石家庄市汉庭酒店会议室完美落下帷幕。各大媒体对此次活动都进行了争相报道！

2015年8月国务院总理李克强先生召开国务院会议宣布大力推进数据分析行业，今年习主席又强调种种全民创新和大数据的重点发展方向，在大数据这个关键词高频环绕在我们的周围的同时，也反应出了当今民众对于数据分析的渴求程度。本次数据分析沙龙活动原计划与会人数50人，当日到场80余人，远远超过原定计划。

参加沙龙的人员分别来自房地产、金融、保险、通信、互联网、IT、软件公司、在校大学生等等各个行业，由此也可以看出各行业对数据分析都十分关注，本次分享的主题只是数据分析应用中的一个方向，而实际上，每个人的工作和生活已经逐渐被数据化，大家每时每刻都被数据围绕，数据分析也已经成为大家生活工作中的一种工具一种技能，大家在有意无意中就在用数据分析的思路去考虑和解决问题了。

用户画像，又称作用文档或用户模型，是对用户特征进行客观精确的描述，以此满足用户的个性化需求。抽象的定义可能大家都理解不了，换一种说法，用户画像就是贴标签！

本次沙龙活动邀请中国商业联合会数据分析专业委员会(中数委)的数据中心主任孙雪老师到场讲解，通过形象的图片、生动的例子完美诠释了用户画像的含义，以及用户画像应该如何设立、建模，它其中的重难点、应用须知都是什么等一系列知识，还给我们介绍了一个神奇的平台：Datahoop数据分析平台，有了它，我们不用再为技术苦恼，我们需要的只是数据精准分析以后的解读工作，能大大减少我们的工作量，提高工作效率！

精准分析客户需求 全民大数据 同享“用户画像”沙龙在石举办

16-03-21 08:41:29 来源：河北新闻网 责任编辑：杨林

全民大数据同享“用户画像”公益沙龙在石上演

2016-03-21 10:43:06 来源：中国日报网河北频道 打印文章 发送给我好友 分享 0

全民大数据 同享“用户画像”公益沙龙在石上演

来源：中国新闻网 作者：佚名 2016-03-21 10:40:00

用户画像是大数据时代衍生出来的产品，它很神奇，可以帮助我们完成个性化定制，目标人群的细分，简直是无所不能的存在，而通过沙龙知识的学习，我们一层一层的拨开迷雾，了解到用户画像的真谛。大数据，它的计算建立在理性的设计基础上，它很实际，是用户真实的反应，所以它才能如此精准定位。

本次沙龙活动，数据分析师(CPDA)河北授权管理中心不但请了非常专业的老师来给大家做专业讲解，同时还邀请了CPDA往期学员为大家分享在自己工作中使用课程中学习的数据分析方法，并深刻感受到在知识体系搭建和开拓思路方向收益匪浅，还跟大家交流了一些相关经验。

数据分析师(CPDA)河北授权管理中心在沙龙活动过程中还搭建了一个“人才需求、人才缺口”面对面互通的平台，本次授权中心邀请了在河北省内发展名列前茅的两家互联网行业及软件开发行业的企业，请他们现场做人才需求讲解，并表示了对数据分析师(CPDA)课程的高度认可，对于持有数据分析师(CPDA)从业证书的学员优先录取，现场在座的学员和即将参加学习的同学都跃跃欲试，反响热烈，对于未来数据分析领域发展非常看好。

通过本次主题沙龙活动，不单从感官上还是从数据上，都让数据分析门槛外的人员，深刻感受到了什么是大数据，什么是数据分析。而通过Datahoop这个平台，又可以让人人都可以了解的大数据，并从事数据分析行业！

fin

/ 陕西 SALON /

文图 / 陕西授权管理中心　编辑 / 协会会员处 周子赫　日期 / 2016-03

3月13日，2016年陕西地区第一场数据分析师沙龙活动在西安会展大厦举办。本次活动以“数聚”作为主题，希望通过本次活动不仅能够为广大数据分析师提供交流的平台，更希望能够将数据分析技术运用到实际生活，为大家的工作和学习提供帮助。

惊蛰经过，天上的春雷惊醒蛰居的动物，蛰虫惊醒，天气转暖，渐有春雷，古都西安进入了温和的春天。在周末的午后举行一场小小的下午茶沙龙聚会，邀约三五知己。相聚在熟悉的“老地方”，相信你的观点和他的看法定会产生一个美妙的邂逅!

首先为我们做分享的是毕业于德国基尔大学的杨晓东老师，杨老师从企业转型进入大学从教，不但具备扎实的理论基础，同时还具备了丰富的实战经验，他通过实际案例，依据产品质量、价格、服务、交货期四个方面的因素来选择供货商。从而为大家深入浅出的剖析了层次分析法的原理和应用。

中场讨论环节我们围绕“什么才是科学的决策”，展开了热烈的讨论，我们从大数据聊到给孩子买奶粉、从苹果公司聊到全球信任危机，从天南聊到海北。最后陕西授权中心负责人刘敏老师亲自上场,为大家带来“大数据应用中决策分析的价值”主题演讲。从Kmart超市关注顾客的数据信息到美国大数据的发展。为我们诠释了如何运用大数据去做科学的决策。

活动在大家的热烈讨论中圆满结束，最后我们将一句话分享给大家，那就是“技术不是拿来学的，而是拿来用的”，期待数据分析技术不断发展，不断融入并丰富我们的生活。

fin

/ 2016年会员年检工作顺利结束 /

文 / 协会会员处 周子赫　日期 / 2016-03

按照《会员管理办法》的相关规定，2016年数据分析行业一年一度的会员资质年度审查工作已经结束。年检通过的会员属于行业正规从业单位或从业个人，在审核有效期内可以进行与数据分析相关的工作，而根据相关评审标准，年检未通过的团体单位或个人，协会已取消其从业资质。

随着数据分析行业快速发展，全国各地的团体会员和个人会员迅速增加，同时在从业过程中出现的问题也逐步凸显。协会2016年的年检工作汲取了以往的工作经验，结合行业实际情况，优化并进一步完善了年检内容，将年检工作标准化、制度化，真正发挥协会的职能作用、履行监管职责。

协会利用年检这一重要监督检查手段，了解会员真实的从业情况及存在的问题，从而帮助会员找出解决问题的方法，纠正和查处在备案管理、经营发展中存在的违规行为等。我会希望通过加强数据分析行业监督管理，从而保证行业更好的规范化发展。

fin

/ www.chinacpda.org 重装上阵 /

文 / 协会市场处 宫辰 编辑 / 协会会员处 冯伟 图 / 苏然 日期 / 2016-02

随着中国大数据行业的快速发展，每天所产生的信息量不断增长。作为中国唯一的数据分析行业网站，原有网站无论在栏目设计还是在功能实现上，都已经无法满足行业发展需求。为了更好的发挥协会网站的行业宣传和服务作用，增进社会各界对数据分析行业的认知度，协会结合各方意见和建议，对网站进行全新改版，值2016年新春佳节之际，正式上线。

此次新版网站对栏目设置、网页布局、网站功能进行了全新规划，丰富了首页内容和服务项目，增强了栏目和信息的条理性、全面性和规范性。改版后的网站无论在整体风格、内容展现，还是栏目设置等方面都进行了升级优化，力求贴近行业需求，为广大的学员、会员、企业以及所有数据分析从业人员提供更为优质的服务。

/ 2016年数据分析行业发展战略会议精彩回顾 /

文 / 协会会员处 周子赫　编辑 / 协会市场处 宫辰　图 / 赵金元　日期 / 2016-01

在过去的一年中，国务院印发了《促进大数据发展行动纲要》，与此同时工信部开始制定《大数据产业“十三五”发展规划》，“大数据”三个字反复在政府各大会议中出现，国家对大数据的关注可谓空前绝后，大数据万亿风口正在形成。

大数据发展正式升级为国家战略，2016年中国大数据产业必将迎来重大变革与机遇!

2016年1月9日-10日，在北京前门建国饭店群英厅，举办了“2016年数据分析行业发展战略会议”。作为每年一次的数据分析行业会议，组委会从众多报名者中，选拔出上百名与会者，分别来自全国数据分析师事务所负责人，数据分析师培训中心负责人以及企业和CPDA学员代表等。

本次会议主要讨论了2016年中国数据分析行业的发展机遇及战略布局，通过剖析目前事务所和企业所面临的困惑，借鉴大数据应用的成功案例以及大数据智能分析平台的技术导入，从而引出“技术+咨询”的行业发展战略。

本次会议我们有幸邀请到中国商业联合会数据分析专业委员会会长邹东生先生、中国邮政集团公司数据中心领导陈燕女士、数据堂联合创始人肖永红先生、香港绿洲游戏网络科技有限公司高级数据经理张炳出先生、美巢集团统计分析部主管沈惠娟女士、中颢润（北京）数据分析师事务所朱传东副总经理、河南明豫数据分析师事务所孙春光总经理以及协会专家团队等嘉宾，并分别进行了精彩的演讲。

在本次会议的开始，协会邹东生会长首先进行《大数据时代的机遇》的主题演讲，对大数据时代的机遇与挑战做出了阐述。针对当前大数据市场风口的现状，剖析了大数据时代面临的机遇，说明当前我们不仅拥有机遇，同时还面临着挑战。

邹会长还表示，这次参会的企业家都是从多年前，就开始真正进入了数据分析这个行业，在行业中摸爬滚打了很多年，积攒了很多宝贵的经验，具有非比寻常的远见卓识。但是机遇永远是摆在每一位从业者面前的，如何抓住机遇，将机遇转化为实力，如何快速完成数据分析行业的“弯道超车”，这就是我们本次会议想要和大家一起深度探讨的。

邹会长:《大数据时代的机遇》

大家对中国数据分析行业的发展寄予厚望，但首先还要先梳理一下行业的现状和问题，数据分析师事务所作为专业的数据分析咨询服务机构，具有一定的行业代表性，所以我们邀请到河南明豫数据分析师事务所孙春光总经理作为整个从业机构的代表，和大家一起探讨事务所目前所面临的一些问题与挑战。孙总表示：当前一些事务所缺乏数据来源，缺乏基于大数据创新设计的商业模式基本思路，没有认清信息化和数据化对企业发展的深刻意义，导致事务所缺乏大数据基因。

孙春光总经理：《数据分析师事务所面临的机遇与挑战》

事务所想要为企业提供更为优质的数据化服务，那我们就应该先了解一些企业在数据化过程中所面临的问题和需求。

沈惠娟女士：《大数据时代，我们可以做什么》

因此，我们邀请到美巢集团统计分析部经理沈惠娟女士为大家带来了《大数据时代，我们可以做什么》的主题演讲。沈经理在演讲中指出传统型企业在对接大数据时，所面临的一些问题，同时也提出了自己的一些观点和建议：要实现生产型企业的数据化建设，可以通过搭建行业数据共享平台，在生产及运营等各个环节中应用更多有价值的数据分析技术方法，从而提升企业在大数据时代的竞争优势。

事务所和企业站在不同的角度，提出了他们的困惑和建议。但是有一些问题是大家共同存在的，比如：数据来源问题，仅仅依靠企业内部数据已经无法满足大数据时代发展的需求。数据堂（北京）科技股份有限公司联合创始人肖永红先生，为大家带来了《数据银行》的主题演讲，希望能够帮助与会者解决一部分数据来源的问题。

肖永红先生：《数据银行》

肖永红先生提出了大数据时代发展的新思维，第一是通过交易凸显数据的交易属性，第二是数据的融汇，即数据之和的价值大于数据价值之和，第三是数据需要流通，数据的最大价值不应该由数据原始拥有者决定。肖永红先生还表示众包将成为大数据产业核心模式之一，贯穿在数据采集、开发、服务各个环节。

数据来源扩充了，我们又面临着下一问题，如何利用这些数据进行有效分析，为企业提供决策依据？在不同的行业领域，有很多大数据应用的成功经验。这些都是我们可以借鉴和引入的，所以，我们邀请到香港绿洲游戏网络科技有限公司高级数据经理张炳出老师，他为大家带来了《浅谈游戏行业中的数据分析》的主题演讲，为大家展现了在手游领域的大数据应用案例。

张炳出老师凭借多年的从业经验，还为希望从事游戏行业的数据分析师提出了几点建议：对业务有充分和正确的理解、能够明确地提出数据需求、能够给出可以落地的建议、极强的逻辑分析和归纳总结能力、善于用图形或BI系统进行可交互化的展示、高效灵活的数据处理能力。

张炳出老师：《浅谈游戏行业中的数据分析》

传统行业作为中国经济的支柱，借助数据分析技术提高自身竞争力显得尤为重要，看清我们的问题和需求才能更好的去迎合大数据时代的发展,来自中国邮政集团公司数据中心的陈燕老师为大家带来了《大数据时代下邮政大数据工作的思考》主题演讲。

陈燕老师表示：整合中国邮政自有核心数据资产，适度引入外部数据，按照大数据基础平台层、分析应用层、商业应用层等三个层级总体架构邮政大数据平台，基于“三个面向”（即：领导决策服务、业务板块服务、网络优化服务），实现“四个价值”（即：盈利价值、优化价值、服务改善价值、社会价值），实现“一个管控”（即：风险管控），支撑“一体两翼”战略的实施，从而助力邮政企业转型发展。

陈燕老师：《大数据时代下邮政大数据工作的思考》

通过半天的会议，我们认识到中国数据分析行业的发展不是一蹴而就的，我们要将这些想法与大数据时代相融合，认清自身的问题和不足，吸纳各行业各领域的成功经验，并通过梳理和调整制定出一系列的战略布局来实现行业

的快速发展。

经过短暂的午休，在9日下午的会议中，首先为大家揭晓了2015年优秀事务所获奖名单。

分别是：

北京中颢润数据分析师事务所

湖南翰林数据分析师事务所

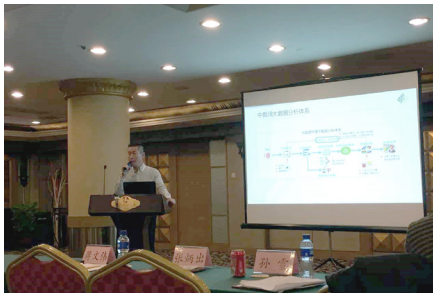
重庆传晟项目数据分析师事务所

上海天元项目数据分析师事务所

并由协会邹会长为优秀事务所颁奖，以此来表彰各优秀事务所在2015年所取得的成绩。同时协会鼓励更多的事务所能够快速转型，鼓励更多的数据分析师选择创业，共同抓住大数据时代的机遇。

邹会长为2015年度优秀事务所颁奖

在接下来得议程中，我们邀请到北京中颢润数据分析师事务所朱传东副总经理来为大家进行了《中颢润大数据案例分享》的主题演讲。朱总将中颢润的发展历史和业务开展情况和所有与会者进行了经验分享，多年来事务所坚持专业的数据分析咨询服务，在成为协会第一个战略合作事务所后，凭借技术实现能力的补充，使事务所的业务承接能力得到整体提升，从而获得了更多的优质业务和发展机遇。



朱传东副总经理：《中颢润大数据案例分享》

中颢润作为第一个与协会签订战略合作协议的事务所，近年来，坚持技术加咨询的业务开展模式，取得了可喜的成果，事实证明这一模式是可行的，是所有事务所都可以借鉴的。

如何才能将这种模式延续下去并进行升华，最终提升为中国数据分析行业2016年行业发展战略。接下来的环节是本次会议的重点，邹会长为大家阐述了2016年数据分析行业的发展战略。

邹会长：《2016年数据分析行业发展战略》

2016年数据分析行业将沿袭并深化“技术+咨询”的战略模式，借助智能大数据分析平台的完善应用，将为企业提供优质的数据化服务，打破大数据的应用瓶颈，让更多的企业有机会与大数据进行快速对接。

同时，事务所作为数据分析服务专业机构，近年来一直以量化咨询作为主体业务并日趋成熟，平台的引入无疑使事务所的整体实力得到显著提升，弥补了业务上的短板。通过战略调整，无论是协会自身还是从业机构，破局已经完成，行业变革即将开启。

在后续的议程中，我们就战略中提到的咨询进行展开讨论，由协会数据中心团队成员从不同的方面给大家展现了量化咨询的真正价值。

首先是祝捷博士为大家进行演讲，他演讲的主题是《决策咨询–大数据应用的精髓》， 展示了量化咨询的重要性，在数据分析领域，咨询的价值要高于技术实现，我们得知大数据分析来帮助决策价值巨大，玩转大数据不在于技术而在于咨询。

祝捷博士：《决策咨询–大数据应用的精髓》

咨询固然重要，但是我们如何才能利用量化咨询为企业提供决策，将量化咨询引用到各个领域之中。接下来是协会资深数据分析师卞静女士为大家进行的《大数据视野–让大数据成为生产力》的主题演讲，通过大数据在各个行业的应用解析，细致地对大数据进行解读，使我们了解到大数据之所以价值很大，是因为可以打破数据孤岛，体现数据价值，生动地为大家展现出了大数据宽阔的视野。

数据分析师卞静女士：《大数据视野–让大数据成为生产力》

会议首日的最后一位演讲嘉宾是协会数据中心主任孙雪女士，她为大家带来《数据分析让世界没有秘密》的演讲。帮助我们揭开了数据分析的神秘面纱，通过模型算法在企业应用中的案例解析，并引用餐饮行业的实际案例，让大家了解到数据分析建模的全过程是如何实现的。企业数据定制化服务是未来的行业发展趋势。同时，孙主任也希望协会数据中心在未来的合作中，能够成为事务所和广大分析师的咨询技术后盾，互相学习，共同发展。

孙雪主任：《数据分析让世界没有秘密》

1.10日会议议程进入第二天，在之前的会议中，邹会长提出了“技术+咨询”的行业发展战略，咨询是数据分析应用的精髓所在，但同时也需要技术来实现。

协会自主研发的Datahoop大数据分析平台为广大分析师及事务所提供了技术实现的基础。所以，在第二天的会议中，我们首先邀请到协会技术中心的Hadoop高级工程师张耀友先生，从专业的技术角度为大家展现了Datahoop大数据分析平台的技术构建。

张耀友工程师：《Datahoop平台介绍》

平台的技术构建固然重要，但对于广大分析师及事务所来说，平台的使用更加重要，协会数据中心主任陆开一先生和数据分析师张明珠女士为大家共同展示了Datahoop平台的功能特点及使用方法。

陆老师提出了平台的设计理念及初衷：我们自主研发设计的datahoop数据分析平台，采用智能化设计理念，从用户的角度出发，可以针对输入数据的类型，自动匹配算法模型，输出分析结果，避免了前期复杂专业知识的学习，设计遵循“小白原则”，操作更加人性化，采用一键式算法自匹配功能，为用户节省了操作时间，也省去了前期大量先验知识的积累。

陆开一主任：《Datahoop功能展示》

通过会议Datahoop平台的的如何才能利用这个平台与企业进行对接，在帮助企业进行数据化建设的同时，事务所又应该如何运用好这个平台，去承接更多的企业数据化业务？在接下来的会议议程中，协会副秘书长蒋文伟先生与大家共同探讨了大数据平台的商业合作运维。

蒋文伟先生：《基于平台的商业运维模式》

协会为所有从业者提供了坚实的咨询后盾以及技术支持，行业战略的实现是需要整体配合的，事务所需要内外兼修才能充分利用好周围的资源，在如今的大数据时代，品牌宣传必不可少。

崔旭主任：《奔跑吧，市场》•

在2016年协会基于平台将进行大量宣传和品牌推广工作，并且这些工作都和事务所有着密切的关系，在会议的最后一个议程， 协会市场部主任崔旭先生进行了《奔跑吧，市场》主题演讲。为大家阐述了在大数据时代，作为数据分析从业机构，应该具备的市场敏感度。

第二天会议下午，协会邹会长和各处室负责人与本次会议的参会代表进行了圆桌会议，大家各抒己见，分别就2016年行业发展战略提出了自己的想法和建议，同时邹会长也为大家更加深入的解析了战略的精髓，无论是事务所还是授权中心，我们都应该是行业发展的主角，为数据分析行业的发展贡献应有的力量。

协会与参会代表进行圆桌会议

两天的会议充实紧张，最后为了庆祝此次会议的圆满成功，演讲嘉宾和与会者合影留念。大家期待在2016年能够跟随协会的脚步，所有数据分析从业者能够共同努力，紧密联合在一起，同发展，共创造，并且期待下次的行业聚会。

fin

/ “2015—2020”百家企业经营模式创新与战略扶持 /

文 / 协会会员处 周子赫 编辑 / 市场处 宫辰 日期 / 2016-01

2015年6月1日，针对我国中小型企业的大数据扶持计划——“百家企业经营模式创新与战略扶持合作方案”拉开帷幕，一经发布吸引了大批中小型企业和媒体的关注与报道。通过半年多的推广与摸索，我们扶持一部分企业完成了数据化建设，同时更加深入了解了中小型企业的数据化程度，对于企业的数据化需求也更加清晰。根据企业的不同基础和需求，我们将从企业数据化建设、数据分析人才培养、运营成本控制、客户行为分析、营销方案预测以及大数据创业等多个方面，提供技术+咨询的全面扶持与帮助，让更多的企业感受到大数据的魅力。

一、方案背景

大数据概念在中国火速发展，2015年3月两会上“互联网+”工作计划的提出，更加速了企业与云计算、大数据、物联网的结合，然而对于如何有效存储数据、利用数据指导经营决策，从而实现企业的盈利，是当前企业遇到的最大问题。

二、方案目的

目前，很多有远见的企业已经高度重视数据化服务的建设，部分企业已经构架自己的“大数据”分析平台，但由于数据分析平台搭建具有“技术难度大”、“费用昂贵”、“技术服务公司不懂数据算法和分析”等问题，使企业在大数据时代，心有所向，力所不及！

中商联数据委是我国大数据领域的行业组织，监管和服务全国数据分析师事务所，是全国性非盈利行业管理协会。数据委具有先进的技术开发能力及数据分析深度研究能力，拥有自主知识产权的Datahoop数据分析平台。

战略扶持计划开展的目的有三：

其一、响应主管机构（中国商业联合会）的号召，更好的服务于会员单位，使中国企业更好的适应大数据时代的变化，取得发展先机；

其二、通过免费为企业搭建数据分析平台，提供技术支持，降低数据应用门槛，使企业真正见识大数据的魅力，帮助企业构建内部数据的完整性以及与外界数据的有效交互；

其三、通过成功企业的推广，引导更多的企业重视数据分析，重视数据化建设工作，进而壮大数据分析市场，使行业内事务所或企业有更好的发展契机和空间。

三、方案周期

2015年6月1日–2020年5月31日（5年计划），在此期间，正式签约企业，我会提供免费扶持期6个月或1个完整项目周期。

四、具体内容

（一）企业标准

1、数据化程度不高，但企业有传统方式的基础数据保存（如ERP或CRM系统等），具备一定的数据基础；

2、企业对数据及数据分析工作高度重视，有意愿迅速将企业数据进行科学的存贮及应用。

（二）我会提供扶持项目

1、对企业数据程度进行评估，为企业数据保全提供方案；

2、根据企业需求及硬件情况，为企业引入Datahoop平台接入服务，使企业数据得以有效存贮、根据企业数据情况，对数据进行基础数据处理及数据展现分析，帮助企业利用先进的大数据平台迅速提高企业决策分析能力；

3、协助企业进行业务数据梳理工作，根据企业决策的需求急迫性，逐步引入经营数据建模分析及运营方案，其中可能包括：

– 企业数据化建设 – 经营战略分析
– 运营成本控制 – 品牌舆情分析
– 企业人才培养 – 销售数据分析
– 客户行为分析 – 产品定价策略
– 精准营销 – 用户流失分析
– 优化客户体验 –大数据创业

（三）合作企业需知

1、企业数据需对协会研究团队公开（双方签订数据安全协议，有效保证企业自有数据安全）

2、如数据服务过程中要构架硬件或互联网接入等工作，会产生相应费用，企业需自行承担；

3、在数据分析服务过程中，我会数据中心提供免费的远程支持和服务，如企业需要到场支持，则需承担相应交通及接待费用；

4、作为战略合作伙伴出现在企业的对外宣传上。

五、方案流程：

填写“百家企业扶持报名表”

发送邮件至xiehui@chinacpda.org

选拔出符合标准的企业

15个工作日内通知获选企业

洽谈扶持方案，签订合约

完成扶持计划，后续跟踪服务

咨询方式：

协会会员处：周老师 冯老师

010-59000991-651/652

/ 关于组织实施促进大数据发展重大工程的通知 /

文 / 协会会员处 周子赫　编辑 / 协会会员处 冯伟　插图 / 苏然　日期 / 2016-01

近日，国家发展改革委办公厅印发了关于组织实施促进大数据发展重大工程的通知。从国家发改委这次下发的文件中我们可以看到国家对于大数据发展的重视以及强力支持，数据开放已经是大势所趋。在发达国家推动数据开放和流通已成为共识，美欧多国通过国家战略为数据开放背书。自从“互联网+”上升为我国战略后，中央不断加大力度推动数据开放。

要实现国家的大数据战略目标，大量专业性的人才无疑是重中之重。但我国目前数据分析人才缺口严重，因为大数据需要的是复合型人才，即能够对数学、统计学、数据分析、项目决策和自然语言处理等多方面知识综合掌握的人才。

由中国商业联合会数据分析专业委员会及工业与信息化部教育与考试中心统一主办的CPDA数据分析师培训考试是中国目前数据分析行业最权威的培训考试体系，能够从数据分析实践操作能力、战略分析能力、营销分析能力、运营分析能力和投资分析能力五方面使学员深入的理解数据分析行业，使学员能够对行业特征具有感性的理解，能够对行业未来技术加咨询的发展方向有较为前瞻的认知，获得对目前大数据背景下的数据产业链较为纵深的洞察，并深刻理解数据分析在其中所起到的灵魂作用。帮助学员形成清晰的从业规划，发现个人发展的最佳路径，与中国大数据行业共同发展成长。


中华人民共和国国家发展和改革委员会

发改办高技〔2016〕42号

国家发展改革委办公厅关于
组织实施促进大数据发展重大工程的通知

中央和国家机关有关部门、直属机构办公厅（室），各省、自治区、直辖市及计划单列市、新疆生产建设兵团发展改革委，有关中央管理企业：

为贯彻落实《国务院关于印发促进大数据发展行动纲要的通知》（国发〔2015〕50号，以下简称《大数据纲要》），加快推动数据资源开放共享流通，强化数据资源在各领域应用，促进产业转型升级，培育发展新业态，我委拟依托国家重大建设项目库组织开展促进大数据发展重大工程，现将有关事项通知如下：

一、总体思路

加快落实《大数据纲要》，从破解制约大数据创新发展的突出矛盾和问题出发，重点推进数据资源开放共享，推动大数据基础设施统筹，打破数据资源壁垒，深化数据资源应用，积极培育新兴繁荣的产业发展新业态。各地方、各部门及中央企业要加强统筹研究，提出具有全国性示范效应、产业带动性强的促进大数据发展重

特此通知。

联系人：国家发展改革委高技术产业司

电　话：68502544　68501272　68501530（传真）

国家发展改革委办公厅

2016年1月7日

/ 中国数据分析行业的未来之路 /

编辑 / 协会会员处 冯伟　插图 / 崔峻珩　日期 / 2016-03

2015年10月，国家正式把大数据作为国家战略提出来，而且还要大力推动整个大数据的发展。这一战略的提出让所有人都看到了大数据的价值，越来越多的政策、资金开始向大数据行业倾斜，大数据行业也迎来了发展的春天。

在这个千载难逢的机遇下，大数据企业如何能够抓住这一发展契机，搭上大数据发展的快车呢？为此，本期会客厅专门请到了中国商业联合会数据分析专业委员会会长邹东生先生，就怎样抓住大数据行业的新机遇，开拓大数据行业的新模式，做一次深入的访谈。

本刊记者：

邹会长，您好，当前大数据行业已经成为万众焦点，您认为行业的发展契机有哪些？

邹会长：

首先是政策层的支持，2015年5月8日李克强总理签批，国务院发布《中国制造2025》规划，明确提出应用工业大数据打造中国制造2025；2015年7月1日 国务院办公厅发布《关于运用大数据加强对市场主体服务和监管的若干意见》，提出明确的时间表和任务表，推动政府在大数据上的工作进展；2015年9月5日李克强总理签批，国务院发布《促进大数据发展行动纲要》，系统部署大数据发展工作，大数据上升到中国的国家战略的高度。2015年10月，十三五规划正式将大数据发展提升为国家战略，是中国政府在经济下行的情况下重点布局的新的经济增长点。

其次是国内企业已完成大数据概念普及，国内企业数据化需求即将引爆。根据中数委近期的调研报告显示，国内企业数据化程度普遍较低，有超过50%的单位没有或计划成立数据分析部门。目前，随着大数据概念的普及和商业生态的改变，越来越多的企业谋求转变或模式创新，他们拥抱互联网+，转型O2O，关注大数据，这些新商业模式依赖于数据，新模式的运转依赖于数据服务，使得越来越多的企业谋求数据化，企业数据化的需求即将爆发，这些都为大数据行业的腾飞契机提供了有利的条件。

本刊记者：

您认为在如此有利的情形下大数据行业的未来发展趋势是什么？

邹会长：

总的来说我认为未来大数据的发展会有三个阶段。第一阶段是私有云服务即企业数据化：针对数据化程度很低的企业，有针对性地帮助企业部署信息系统，在企业内部部署大数据系统，使得企业具备大数据能力的软硬件基础，重点在于要以咨询带动技术，以性价比最高的模式切入市场、迅速占有中国优质企业，从而拉低技术门槛，提高咨询竞争力。第二阶段是混合云服务：结合企业的业务特点，采取有针对性的混合云模式提供服务，重点在于技术带服务、技术换数据，数据融合产品。第三阶段是公有云服务：针对企业的特点和需求，接入公有云服务，重点在于与数据分析师事务所业务相融合。

本刊记者：

您认为中国的大数据企业为此做好准备了吗？

邹会长：

放眼市场，国内很多大数据企业是没有做好准备的，很多大数据企业是从传统软件企业借助大数据技术转型而来的，本身大数据基因就不强。而且很多大数据企业的商业模式、技术服务模式都有很大的问题，我主要列举其中的四个方面：第一、中国主板中大数据板块的企业多数是从软件、信息化建设公司转入，无论是从专业技术、研究能力，还是从对大数据的理解都远远达不到有效服务企业的目的；第二、数据缺乏有效融合：现阶段市场上很多大数据系统和大数据方案，只是简单的把数据做了汇总存储，数据之间缺乏关联，特别是非结构数据没有有效的利用和融合；第三、缺乏价值发现能力：市场上大多数大数据系统还是按照传统的数据库系统组织和利用数据，属于一种后验模式，而大数据最大的魅力在于价值发现，数据驱动业务，数据中挖掘价值，是一种先验模式；第四、缺乏落地应用：市场上大多数软件停留于数据浅层次的展示层面，缺少有效的大数据应用模型，因此没有充分发挥大数据的价值。

本刊记者：

数据分析师事务所作为专业咨询机构，您有哪些建议？

邹会长：

现在中国的大数据行业发展还处在我所说的第一阶段即企业数据化阶段。在这一阶段中，数据分析师事务所应着重于技术+咨询的新商业模式，主要为企业提供三方面的服务：第一、企业数据化服务，帮助企业将各类业务信息进行数据化，引入大数据智能服务平台，达到业务即数据，数据即业务，使得企业具备数据化运营的基础；第二、企业数据融合服务，实现企业各部门间的工作联动及数据共享，同时使得企业内部数据与外部数据融合，使得企业初步具备大数据的能力；第三、企业数据咨询服务，让数据驱动业务，从数据中根据分析算法找寻企业存在的问题，改善经营，发掘机会，减小试错成本，提升企业应变能力。

总结：通过本次访谈，邹会长为大家展现了大数据行业的未来发展趋势。随着中国经济的发展，大数据技术的引入将成为未来企业的必争之地，从业机构需要理清思路，转化经营理念，才能把握机遇。毫无疑问，大数据必将改变中国未来！

/ 基于MATLAB模糊聚类的交通状态辨识研究 /

文 编辑 / 杜长海 博士　插图 / 崔峻珩　日期 / 2016-03

一、引言

随着国民经济的高速发展，城市道路交通状况正变得日益严峻，交通拥堵已经成为中国各大城市首要解决的顽疾。城市道路交通状态辨识是实际交通管理中评价交通拥挤状态、解决交通拥挤的基础，为交通管理者和交通出行者提供动态的决策依据，从而有利于快速地疏散拥挤的交通流，具有重要的学术价值和现实意义。

交通状态是一种随时间和空间而不断变化的动态过程，难以用确切的数字或数字界限来说明、判断交通系统处于诸如“通畅”、“拥挤”之类的状态，这充分表明交通状态具有概念的模糊性。模糊聚类是建立在样本类属的不确定性描述下，可以很好的表达和处理对象的这种不分明的类属性质，更客观的反映交通状态。

被誉为第四代计算机语言的MATLAB是以矩阵运算为基础的一种面向科学与工程计算的高级数学分析与运算软件，它在矩阵处理和图形处理上有着得天独厚的优势，因而使用MATLAB软件编制计算程序可以使工作大大简化，计算精度更高。因此，通过利用MATLAB语言中的模糊C均值聚类函数——FCM（ ）对交通流状态进行识别，以期为交通状态实时辨识提供一个新的研究思路。

二、模糊C均值聚类

模糊C均值聚类用于将多维数据空间分布的数据点分成特定数目的类，每一个数据点以某种程度属于某一类，用隶属度来表示每个数据点属于某个聚类的程度，使得非相似性指标的目标函数最小。

数据集合为$X=\{x_1,x_2,\cdots,x_n\}$，$x_k=(x_{k1},x_{k2},\cdots,x_{km})^T$；c是类别数，且$2\le c\le n$；$p_i=(x_{i1},x_{i2},\cdots,x_{im})^T$为第$i$类的聚类原型，则$P=(p_1,p_2,\cdots,p_c)(\in R^{m\times c})$构成聚类原型矩阵；$U=(\mu_{ik})(\in R^{c\times n})$是隶属度矩阵，$\mu_{ik}$表示$x_k$对于$p_i$的隶属度，且$\mu_{ik}\in[0,1]$。则模糊C均值聚类模型表示如下：

$$\begin{cases}\min\ J_b(U,P)=\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^b d_{ik}^2\\ s.t.\ \sum_{i=1}^{c}\mu_{ik}=1,\ k=1,2,\cdots,n\end{cases}\quad(1)$$

式中：$b\in[1,\infty)$为模糊指数；d_{ik}为x_k与p_i的相异度，可取为欧式距离。

FCM算法就是搜索最优的U和P，使得式（1）中的

$J_b(U,P)$达到最小，具体迭代步骤如下：

（1）设置参数：b、c、终止阈值ε、迭代计数器$t=1$、最大迭代次数T；初始化聚类原型$P^{(t)}$；

（2）计算划分矩阵$U^{(t)}$：$\forall i,k$，如果$\exists d_{ik}^{(t)}>0$，则有

$$\mu_{ik}^{(t)}=\left\{\sum_{j=1}^{c}\left(\frac{d_{ik}^{(t)}}{d_{jk}^{(t)}}\right)^{\frac{2}{b-1}}\right\}^{-1} \qquad (2)$$

如果$\exists i,r$，使得$d_{ir}^{(t)}=0$，则有$\mu_{ir}^{(t)}=1$，且对$j\neq r$，$\mu_{ij}^{(t)}=0$。

（3）更新聚类原型矩阵$P^{(t+1)}$：

$$p_i^{(t+1)}=\frac{\sum_{k=1}^{n}\left(\mu_{ik}^{(t)}\right)^b\cdot x_k}{\sum_{k=1}^{n}\left(\mu_{ik}^{(t)}\right)^b} \qquad (3)$$

（4）如果$\left\|P^{(t)}-P^{(t+1)}\right\|<\varepsilon$或$t=T$，则算法停止，并输出$U$和$P$，否则令$t=t+1$，转向步骤（2）。其中，$\|\cdot\|$为某种合适的矩阵范数，可取为矩阵F–范数。

三、基于MATLAB的模糊C均值聚类——FCM（）介绍

语法格式：

[center, U, obj_fcn] = FCM(data, cluster_n, options)

用法：

1. [center,U,obj_fcn] = FCM(Data,N_cluster,options);

2. [center,U,obj_fcn] = FCM(Data,N_cluster);

输入变量：

data —— n*m矩阵,表示n个样本,每个样本具有m维特征值

cluster_n —— 标量,表示聚合中心数目,即类别数

options —— 4*1列向量，其中

options(1): 隶属度矩阵U的指数，>1(缺省值: 2.0)

options(2): 最大迭代次数(缺省值: 100)

options(3): 隶属度最小变化量,迭代终止条件(缺省值: 1e–5)

options(4): 每次迭代是否输出信息标志(缺省值: 0)

输出变量：

center —— 聚类中心

U —— 隶属度矩阵

obj_fcn —— 目标函数值

四、实例分析

4.1 数据预处理

本文所用的交通流数据来源于文献——“基于梯度校正法的交通数据融合和行程时间预测研究[D]. 北京: 北京交通大学, 2008”，特征指标是：流量、速度、占有率，数据周期为2分钟，共计720组，如图1所示。

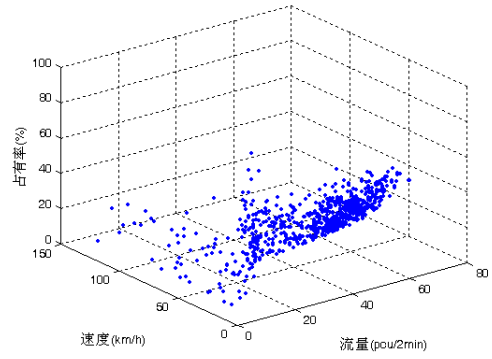


图1交通流数据示意图

原始交通流数据量比较大，而且往往分布不均，因此，有必要对这些交通流特征数据进行合理地采样，以提高数据利用的准确性和算法使用的高效性。提取原则要求采样后交通流特征数据尽可能覆盖交通流数据变化的整个范围，而且采样后的交通流特征数据的分布密度要均衡。

本文采用以占有率为划分标准，采样范围为占有率分别在[0,20)、[20,30)、[30,40)、[40,50)、[50,60)、[60,100]共计6个区间的交通流特征数据，每个区间的样本数量均为25，如表1和图2所示。

表1数据采样表

区间	原始数据（组）	提取数据（组）
[0, 20)	27	25
[20, 30)	71	25
[30, 40)	235	25
[40, 50)	263	25
[50, 60)	98	25
[60, 100]	26	25
共计	720	150

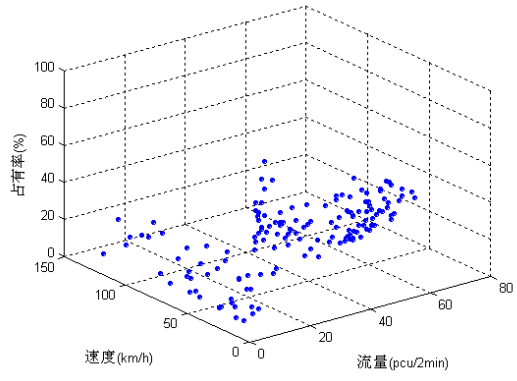


图2数据采样示意图

显然，从原始数据720组中采样了150组，由图2可见，采样的数据分布比较均匀，基本涵盖了交通流状态的所有特征。

4.2参数设置

参考我国公安部2002年颁布的《城市交通管理评价指标体系》与交通运输部2012 年颁布的《公路网运行监测与服务暂行技术要求》对拥挤程度的分类，将道路交通状态划分为4个等级：畅通、平稳、拥挤、堵塞，因而规定聚类原型数目 c 取为4。由4.1节知交通流特征维数 $m=3$。最大迭代次数为50，$b=2$，$\varepsilon=0.00001$。

4.3结果分析

根据FCM的计算结果，得到代表4 种交通状态的聚类中心矩阵 （ 式4 ）：

$$P=\begin{pmatrix} 10.2797 & 19.0433 & 11.4511 & 44.0008 \\ 76.9546 & 45.6642 & 2.6173 & 11.7194 \\ 12.7663 & 21.8564 & 56.7537 & 45.8890 \end{pmatrix}$$

式（ 4 ）中第1个列向量p_1=(10.2797,76.9546, 12.7663)T，即流量为10.2797pcu/2min，速度为76.9546km/h，占有率为12.7663%，流量较小，速度较高，占有率很低，表示畅通状态；同理，p_2流量偏中，速度较高，占有率偏中，表示稳定状态；p_3流量偏少，速度较小，占有率较高，表示堵塞状态；p_4流量较大，速度偏中，占有率较高，表示拥挤状态。

选取8组测试样本，分别计算其对式（ 4 ）中4种交通状态的隶属度，根据最大隶属度原则，辨识样本所处的交通状态，结果见表2。

表2 样本归属状态

序号	测试样本	隶属度（ 畅通，平稳，堵塞，拥挤 ）	归属状态
1	(8, 92, 5)	(0.8528, 0.0975, 0.0233, 0.0264)	畅通
2	(33, 48, 28)	(0.1106, 0.7372, 0.0524, 0.0998)	平稳
3	(5, 1, 43)	(0.0275, 0.0699, 0.7904, 0.1122)	堵塞
4	(56, 12, 54)	(0.0221, 0.0501, 0.0850, 0.8428)	拥挤
5	(10, 59, 14)	(0.4659, 0.4697, 0.0301, 0.0342)	平稳
6	(5, 5, 14)	(0.1174, 0.3197, 0.3261, 0.2367)	堵塞
7	(38, 68, 27)	(0.3756, 0.4462, 0.0673,0.1109)	平稳
8	(30, 6, 54)	(0.0211, 0.0551, 0.4137, 0.5101)	拥挤

通过“最大隶属度”原则进行去模糊化后，样本仅归属到一种交通状态，如：样本4(56, 12, 54)对于4种交通状态(畅通，平稳，堵塞，拥挤)的隶属度分别为(0.0221, 0.0501, 0.0850, 0.8428)，其中对于“拥挤”状态的隶属程度最为显著，为0.8428，那么把样本4归为“拥挤”状态是比较合理的。但是，这样处理往往不能有效体现模糊聚类的优势，如：样本5对“平稳”、“堵塞”两种状态的隶属度分别为0.3197, 0.3261，差别不大，若按最大隶属度原则，则严格归为“堵塞”状态，并没有体现出交通状态的模糊性特点，以及交通状态之间的过渡性。

为此，可以给定阈值 $\lambda \in [0,1]$，只要满足 $u_{ik} \leqslant \lambda$ 的类别均符合要求。例如：阈值 λ 依次取0.45、0.40和0.30，并按隶属度降序列出，得到归属情况如表3所示。

表3样本归属状态

序号	测试样本	归属状态 λ =0.45	归属状态 λ =0.40	归属状态 λ =0.30
5	(10, 59, 14)	畅通/平稳	畅通/平稳	畅通/平稳
6	(5, 5, 14)	/	/	平稳/堵塞
7	(38, 68, 27)	/	平稳	畅通/平稳
8	(30, 6, 54)	拥挤	堵塞/拥挤	堵塞/拥挤

显然，这种归类方法充分体现样本“亦此亦彼”的性质，在不同阈值水平下，将样本依次归到主要的若干个类，能够更客观地反映出道路实际情况，所得交通状态的量化结论更为合理，并具有较强的实际指导意义。

五、总结与展望

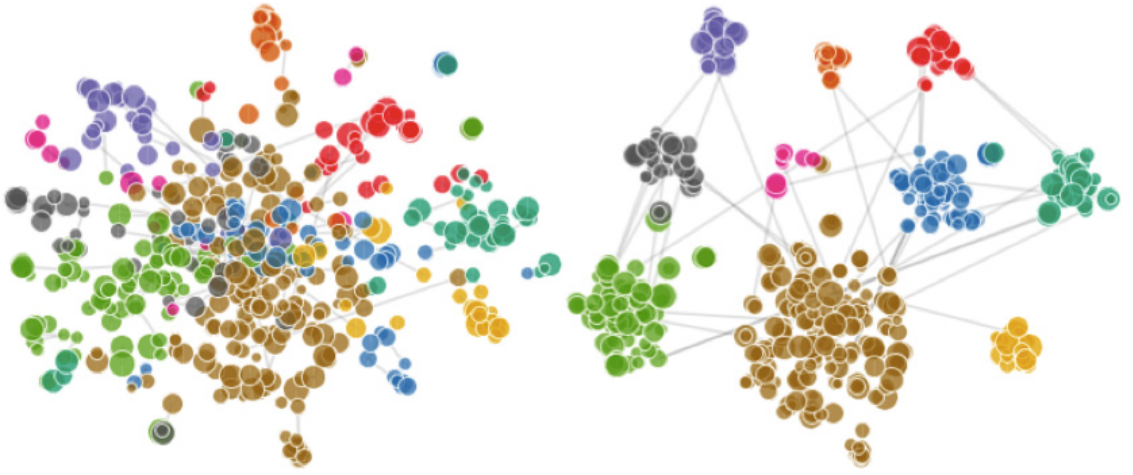
及时而有效的交通状态辨识是进行动态交通诱导、主动避免交通拥堵、保持道路畅通的前提和依据。本文将MATLAB语言中的FCM函数应用于城市道路交通状态辨识研究。

实验结果表明，该方法对解决城市道路交通状态划分等数据聚类问题是可行、有效的，为以后解决交通拥堵问题提供思路，有一定的参考和实用价值。同时，FCM算法如何选取最优化参数组合以进一步提高运算精度，及如何广泛应用于各种工程实际问题，还有待于继续深入研究。

fin

/ 矩阵分解在推荐系统中的应用:NMF和经典SVD实战 /

供稿 / 乐天笔记 编辑 / 协会会员处 冯伟 插图 / 崔峻珩 日期 / 2016-02



本文以NMF和经典SVD为例，讲一讲矩阵分解在推荐系统中的应用。

Item\user	Ben	Tom	John	Fred
Item2	5	0	3	4
Item3	3	4	0	3
Item4	0	0	5	3
Item5	5	4	4	5
Item6	5	4	5	6

User \ Item	Item1	Item2	Item3	Item4
Ben	5	5	3	0
Tom	5	0	4	0
John	0	3	0	5
Fred	5	4	3	3

NMF
用户和物品的主题分布

```
#!/usr/bin/python2.7
# coding: UTF-8
import numpy as np
from sklearn.decomposition import NMF
import matplotlib.pyplot as plt

RATE_MATRIX = np.array(
    [[5, 5, 3, 0, 5, 5],
     [5, 0, 4, 0, 4, 4],
     [0, 3, 0, 5, 4, 5],
     [5, 4, 3, 3, 5, 5]]
)

nmf = NMF(n_components=2)
user_distribution = nmf.fit_transform(RATE_MATRIX)
item_distribution = nmf.components_

print '用户的主题分布：'
print user_distribution
print '物品的主题分布：'
print item_distribution
```

运行后输出：

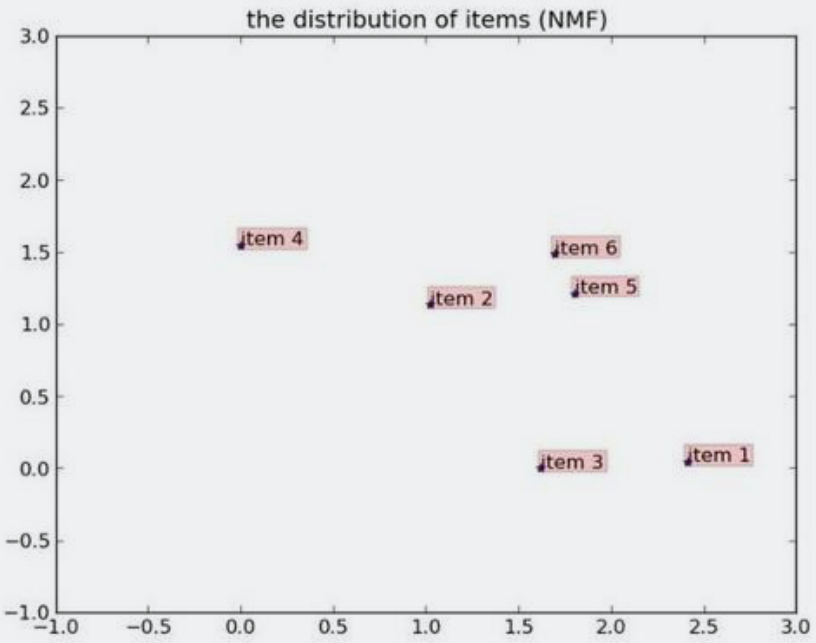
```
用户的主题分布：
[[ 2.20884275  0.84137492]
 [ 2.08253282 -0.        ]
 [-0.          3.18154406]
 [ 1.84992603  1.60839505]]
物品的主题分布：
[[ 2.4129931   1.02524235  1.62258152  0.          1.80111078  1.69591943]
 [ 0.0435741   1.13506094  0.          1.54526337  1.21253494  1.48756118]]
```

可视化物品的主题分布：

```
#!/usr/bin/python2.7
# coding: UTF-8
import numpy as np
from sklearn.decomposition import NMF
import matplotlib.pyplot as plt

RATE_MATRIX = np.array(
    [[5, 5, 3, 0, 5, 5],
     [5, 0, 4, 0, 4, 4],
```

```
    [0, 3, 0, 5, 4, 5],
    [5, 4, 3, 3, 5, 5]]
)

nmf = NMF(n_components=2)
user_distribution = nmf.fit_transform(RATE_MATRIX)
item_distribution = nmf.components_

item_distribution = item_distribution.T
plt.plot(item_distribution[:, 0], item_distribution[:, 1], "b*")
plt.xlim((−1, 3))
plt.ylim((−1, 3))

plt.title(u'the distribution of items (NMF)')
count = 1
for item in item_distribution:
    plt.text(item[0], item[1], 'item '+str(count), bbox=dict(facecolor='red', alpha=0.2),)
    count += 1

plt.show()
```

结果：



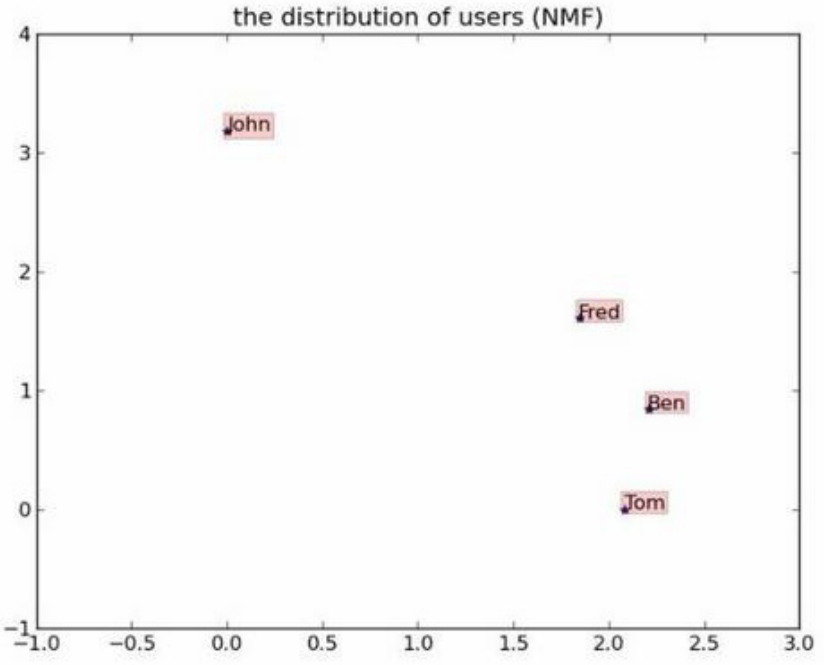
从距离的角度来看，item 5和item 6比较类似；从余弦相似度角度看，item 2、5、6 比较相似，item 1、3比较相似。

可视化用户的主题分布：

```
#!/usr/bin/python2.7
# coding: UTF−8
```

```
import numpy as np
from sklearn.decomposition import NMF
import matplotlib.pyplot as plt

RATE_MATRIX = np.array(
    [[5, 5, 3, 0, 5, 5],
     [5, 0, 4, 0, 4, 4],
     [0, 3, 0, 5, 4, 5],
     [5, 4, 3, 3, 5, 5]]
)

nmf = NMF(n_components=2)
user_distribution = nmf.fit_transform(RATE_MATRIX)
item_distribution = nmf.components_
users = ['Ben', 'Tom', 'John', 'Fred']
zip_data = zip(users, user_distribution)

plt.title(u'the distribution of users (NMF)')
plt.xlim((-1, 3))
plt.ylim((-1, 4))
for item in zip_data:
    user_name = item[0]
    data = item[1]
    plt.plot(data[0], data[1], "b*")
    plt.text(data[0], data[1], user_name, bbox=dict(facecolor='red', alpha=0.2),)

plt.show()
```

结果:



从距离的角度来看，Fred、Ben、Tom的口味差不多；从余弦相似度角度看，Fred、Ben、Tom的口味还是差不多。

如何推荐?
现在对于用户A，如何向其推荐物品呢?
方法1：找出与用户A最相似的用户B，将B评分过的、评分较高、A没评分过的的若干物品推荐给A。
方法2：找出用户A评分较高的若干物品，找出与这些物品相似的、且A没评分的若干物品推荐给A。
方法3：找出用户A最感兴趣的k个主题，找出最符合这k个主题的、且A没评分的若干物品推荐给A。
方法4：由NMF的评分结果，重建评分矩阵。
例如：

```
#!/usr/bin/python2.7
# coding: UTF-8
import numpy as np
from sklearn.decomposition import NMF
import matplotlib.pyplot as plt

RATE_MATRIX = np.array(
    [[5, 5, 3, 0, 5, 5],
     [5, 0, 4, 0, 4, 4],
     [0, 3, 0, 5, 4, 5],
     [5, 4, 3, 3, 5, 5]]
)

RATE_MATRIX[1, 2] = 0  # 对评分矩阵略做修改
print '新评分矩阵：'
print RATE_MATRIX

nmf = NMF(n_components=2)
user_distribution = nmf.fit_transform(RATE_MATRIX)
item_distribution = nmf.components_
reconstruct_matrix = np.dot(user_distribution, item_distribution)
filter_matrix = RATE_MATRIX < 1e-6  # 小于0
print '重建矩阵，并过滤掉已经评分的物品：'
print reconstruct_matrix*filter_matrix
```

运行结果：

```
新评分矩阵：
[[5 5 3 0 5 5]
 [5 0 0 0 4 4]
 [0 3 0 5 4 5]
 [5 4 3 3 5 5]]

重建矩阵，并过滤掉已经评分的物品：
[[ 0.          0.          0.          0.80443133  0.          0.        ]
 [ 0.          2.19148602  1.73560797  0.          0.          0.        ]
```

```
[ 0.02543568  0.          0.48692891  0.          0.          0.        ]
[ 0.          0.          0.          0.          0.          0.        ]]
```

对于Tom（评分矩阵的第2行），其未评分过的物品是item 2、item 3、item 4。item 2的推荐值是 2.19148602 ，item 3的推荐值是 1.73560797 ，item 4的推荐值是 0，若要推荐一个物品，推荐item 2。

如何处理有评分记录的新用户

NMF是将非负矩阵V分解为两个非负矩阵W和H：

$$V = W \times H$$

在本文上面的实现中，V对应评分矩阵，W是用户的主题分布，H是物品的主题分布。

对于有评分记录的新用户，如何得到其主题分布?

方法1：有评分记录的新用户的评分数据放入评分矩阵中，使用NMF处理新的评分矩阵。

方法2：物品的主题分布矩阵H保持不变，将W更换为新用户的评分组成的行向量，求W即可。

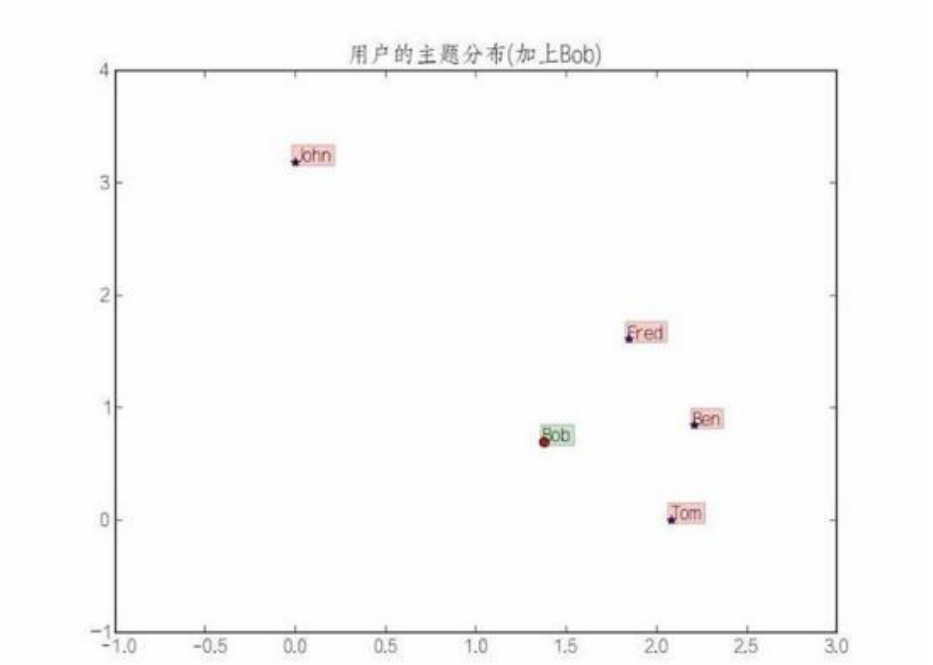
下面尝试一下方法2。

设新用户Bob的评分记录为：

```
[5,5,0,0,0,5]
#!/usr/bin/python2.7
# coding: UTF-8

import numpy as np
from sklearn.decomposition import NMF
import matplotlib.pyplot as plt

RATE_MATRIX = np.array(
    [[5, 5, 3, 0, 5, 5],
     [5, 0, 4, 0, 4, 4],
     [0, 3, 0, 5, 4, 5],
     [5, 4, 3, 3, 5, 5]]
)

nmf = NMF(n_components=2)
user_distribution = nmf.fit_transform(RATE_MATRIX)
item_distribution = nmf.components_

bob = [5, 5, 0, 0, 0, 5]
print 'Bob的主题分布：'
print nmf.transform(bob)
```

运行结果是：

```
Bob的主题分布：

[[ 1.37800534  0.69236738]]
```



经典SVD

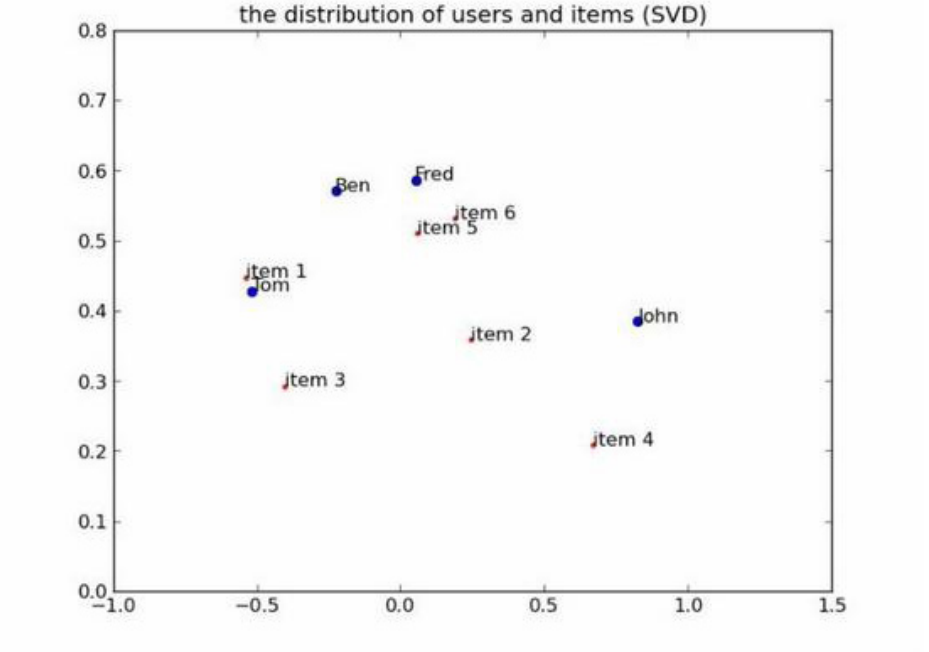
```
#!/usr/bin/python2.7
# coding: UTF-8
import numpy as np
from scipy.sparse.linalg import svds
from scipy import sparse
import matplotlib.pyplot as plt

def vector_to_diagonal(vector):
    """
    将向量放在对角矩阵的对角线上
    :param vector:
    :return:
    """
    if (isinstance(vector, np.ndarray) and vector.ndim == 1) or \
            isinstance(vector, list):
        length = len(vector)
        diag_matrix = np.zeros((length, length))
        np.fill_diagonal(diag_matrix, vector)
        return diag_matrix
    return None

RATE_MATRIX = np.array(
    [[5, 5, 3, 0, 5, 5],
     [5, 0, 4, 0, 4, 4],
     [0, 3, 0, 5, 4, 5],
     [5, 4, 3, 3, 5, 5]]
)
RATE_MATRIX = RATE_MATRIX.astype('float')
```

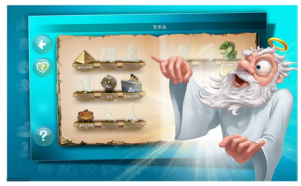
```
U, S, VT = svds(sparse.csr_matrix(RATE_MATRIX), k=2, maxiter=200)
S = vector_to_diagonal(S)
print '用户的主题分布：'
print U
print '奇异值：'
print S
print '物品的主题分布：'
print VT
print '重建评分矩阵，并过滤掉已经评分的物品：'
print np.dot(np.dot(U, S), VT) * (RATE_MATRIX < 1e-6)
```

运行结果：

```
用户的主题分布：
[[-0.22279713  0.57098887]
 [-0.51723555  0.4274751 ]
 [ 0.82462029  0.38459931]
 [ 0.05319973  0.58593526]]
奇异值：
[[ 6.39167145  0.        ]
 [ 0.         17.71392084]]
物品的主题分布：
[[-0.53728743  0.24605053 -0.40329582  0.67004393  0.05969518  0.18870999]
 [ 0.44721867  0.35861531  0.29246336  0.20779151  0.50993331  0.53164501]]
重建评分矩阵，并过滤掉已经评分的物品：
[[ 0.          0.          0.          1.14752376  0.          0.        ]
 [ 0.          1.90208543  0.         -0.64171368  0.          0.        ]
 [ 0.21491237  0.         -0.13316888  0.          0.          0.        ]
 [ 0.          0.          0.          0.          0.          0.        ]]
```

可视化一下：

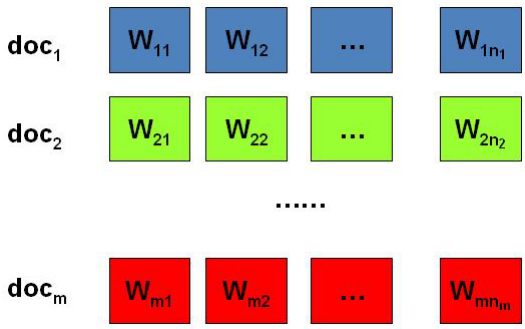




/ 统计文本建模——一个猜测上帝的游戏 /

文 / 统计之都　编辑 / 协会数据中心 孙雪　插图 / 崔峻珩　日期 / 2016-03

我们日常生活中总是产生大量的文本，如果每一个文本存储为一篇文档，那每篇文档从人的观察来说就是有序的词的序列 d=(w1,w2,.....,wn)。



包含M 篇文档的语料库

统计文本建模的目的就是追问这些观察到语料库中的的词序列是如何生成的。统计学被人们描述为猜测上帝的游戏，人类产生的所有的语料文本我们都可以看成是一个伟大的上帝在天堂中抛掷骰子生成的，我们观察到的只是上帝玩这个游戏的结果 —— 词序列构成的语料，而上帝玩这个游戏的过程对我们是个黑盒子。所以在统计文本建模中，我们希望猜测出上帝是如何玩这个游戏的，具体一点，最核心的两个问题是

* 上帝都有什么样的骰子；

* 上帝是如何抛掷这些骰子的；

第一个问题就是表示模型中都有哪些参数，骰子的每一个面的概率都对应于模型中的参数；第二个问题就表示游戏规则是什么，上帝可能有各种不同类型的骰子，上帝可以按照一定的规则抛掷这些骰子从而产生词序列。

上帝掷骰子

1 Unigram Model

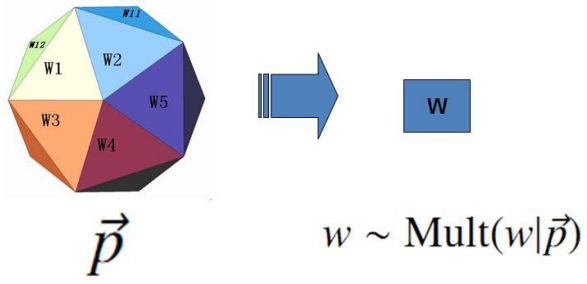
假设我们的词典中一共有 V 个词 v1,v2,....vV，那么最简单的 Unigram Model 就是认为上帝是按照如下的游戏规则产生文本的。

Game 1 Unigram Model

1: 上帝只有一个骰子，这个骰子有V 个面，每个面对应一个词，各个面的概率不一；

2: 每抛一次骰子，抛出的面就对应的产生一个词；如果一篇文档中有n 个词，上帝就是独立的抛n次骰子产生这n 个词；

上帝的这个唯一的骰子各个面的概率记为 $\vec{p} = (p1, p2, \ldots, pV)$，所以每次投掷骰子类似于一个抛硬币时候的贝努利实验， 记为 w~Mult(w|p→)。



上帝投掷V 个面的骰子

对于一篇文档d=w→=(w1,w2,...,wn)，该文档被生成的概率就是

p(w→)=p(w1,w2,...,wn)=p(w1)p(w2)...p(wn)

而文档和文档之间我们认为是独立的， 所以如果语料中有多篇文档 W=(w1→,w2→,⋯,wm→),则该语料的概率是

p(W)=p(w1→)p(w2→)...p(wm→)

在 Unigram Model 中， 我们假设了文档之间是独立可交换的，而文档中的词也是独立可交换的，所以一篇文档相当于一个袋子，里面装了一些词，而词的顺序信息就无关紧要了，这样的模型也称为词袋模型(Bag-of-words)。

假设语料中总的词频是N，在所有的 N 个词中,如果我们关注每个词 vi 的发生次数 ni，那么 n→=(n1,n2,...,nV) 正好是一个多项分布

p(n→)=Mult(n→|p→,N)=(Nn→)∏k=1Vpnkk

此时， 语料的概率是

p(W)=p(w1→)p(w2→)...p(wm→)=∏k=1Vpnkk

当然，我们很重要的一个任务就是估计模型中的参数p→，也就是问上帝拥有的这个骰子的各个面的概率是多大，按照统计学家中频率派的观点，使用最大似然估计最大化P(W)，于是参数pi的估计值就是

pi^=niN.

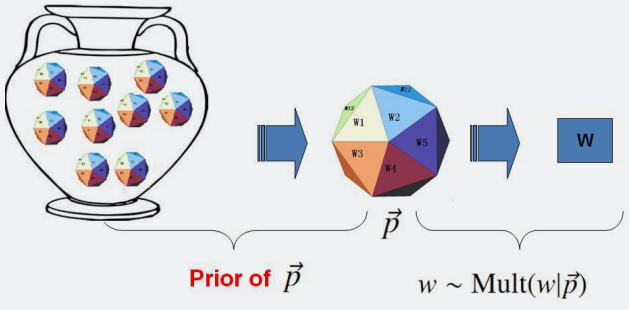
对于以上模型，贝叶斯统计学派的统计学家会有不同意见，他们会很挑剔的批评只假设上帝拥有唯一一个固定的骰子是不合理的。在贝叶斯学派看来，一切参数都是随机变量，以上模型中的骰子 p→不是唯一固定的，它也是一个随机变量。所以按照贝叶斯学派的观点，上帝是按照以下的过程在玩游戏的

Game 2 贝叶斯Unigram Model假设

1: 上帝有一个装有无穷多个骰子的坛子，里面有各式各样的骰子，每个骰子有V 个面；

2: 上帝从坛子里面抽了一个骰子出来，然后用这个骰子不断的抛，然后产生了语料中的所有的词；

上帝的这个坛子里面，骰子可以是无穷多个，有些类型的骰子数量多，有些类型的骰子少，所以从概率分布的角度看，坛子里面的骰子p→ 服从一个概率分布 p(p→)，这个分布称为参数p→ 的先验分布。



贝叶斯观点下的 Unigram Model

以上贝叶斯学派的游戏规则的假设之下，语料W产生的概率如何计算呢？由于我们并不知道上帝到底用了哪个骰子p→，所以每个骰子都是可能被使用的，只是使用的概率由先验分布p(p→)来决定。对每一个具体的骰子p→，由该骰子产生数据的概率是 p(W|p→)，所以最终数据产生的概率就是对每一个骰子p→上产生的数据概率进行积分累加求和

p(W)=∫p(W|p→)p(p→)dp→

在贝叶斯分析的框架下，此处先验分布p(p→) 就可以有很多种选择了，注意到

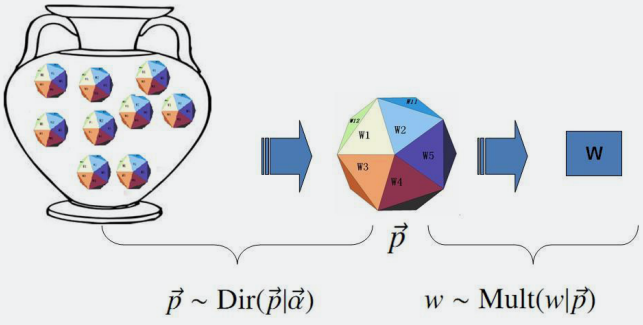
p(n→)=Mult(n→|p→,N)

实际上是在计算一个多项分布的概率，所以对先验分布的一个比较好的选择就是多项分布对应的共轭分布，即 Dirichlet 分布

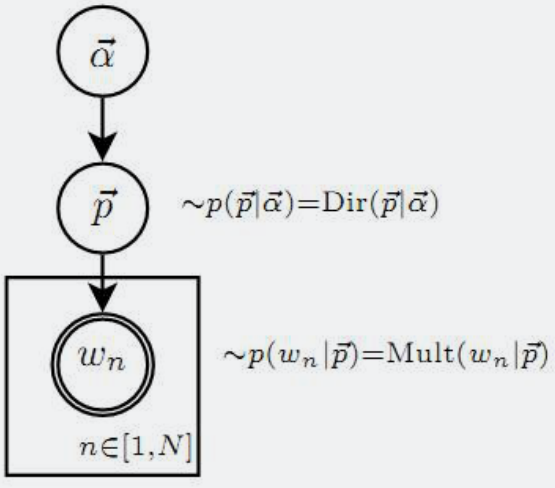
Dir(p→|α→)=1Δ(α→)∏k=1Vpαk−1k， α→=(α1, ..., αV)

此处，Δ(α→) 就是归一化因子Dir(α→)，即

Δ(α→)=∫∏k=1Vpαk−1kdp→.



Dirichlet 先验下的 Unigram Model



Unigram Model的概率图模型

回顾一下 Drichlet 分布的知识，其中很重要的一点就是

Dirichlet 先验 + 多项分布的数据 → 后验分布为 Dirichlet 分布

Dir(p→|α→)+MultCount(n→)=Dir(p→|α→+n→)

于是，在给定了参数 p→的先验分布 Dir(p→|α→) 的时候，各个词出现频次的数据 n→~Mult(n→|p→,N) 为多项分布，所以无需计算，我们就可以推出后验分布是

p(p→|W,α→)=Dir(p→|n→+α→)=1Δ(n→+α→)∏k=1Vpnk+αk−1kdp→(1)

在贝叶斯的框架下，参数p→如何估计呢？由于我们已经有了参数的后验分布，所以合理的方式是使用后验分布的极大值点，或者是参数在后验分布下的平均值。在该文档中，我们取平均值作为参数的估计值。使用上个小节中的结论，由于 p→ 的后验分布为 Dir(p→|n→+α→)，于是

E(p→)=(n1+α1∑Vi=1(ni+αi),n2+α2∑Vi=1(ni+αi), ...,nV+αV∑Vi=1(ni+αi))

也就是说对每一个 pi,

我们用下式做参数估计

$$\hat{p}_i = \frac{n_i + \alpha_i}{\sum_{i=1}^{V}(n_i + \alpha_i)} \quad (2)$$

考虑到 α_i 在 Dirichlet 分布中的物理意义是事件的先验的伪计数，这个估计式子的含义是很直观的：每个参数的估计值是其对应事件的先验的伪计数和数据中的计数的和在整体计数中的比例。

进一步，我们可以计算出文本语料的产生概率为

$$p(W|\vec{\alpha}) = \int p(W|\vec{p})p(\vec{p}|\vec{\alpha})d\vec{p} = \int \prod_{k=1}^{V} p_k^{n_k} Dir(\vec{p}|\vec{\alpha})d\vec{p} = \int \prod_{k=1}^{V} p_k^{n_k+\alpha_k-1} d\vec{p} = \frac{\Delta(\vec{\alpha})}{\Delta(\vec{n}+\vec{\alpha})} \quad (3)$$

2 Topic Model 和 PLSA

以上 Unigram Model 是一个很简单的模型，模型中的假设看起来过于简单，和人类写文章产生每一个词的过程差距比较大，有没有更好的模型呢?

我们可以看看日常生活中人是如何构思文章的。如果我们要写一篇文章，往往是先确定要写哪几个主题。譬如构思一篇自然语言处理相关的文章，可能 40\% 会谈论语言学、30\% 谈论概率统计、20\% 谈论计算机、还有10\%谈论其它的主题：

* 说到语言学，我们容易想到的词包括：语法、句子、乔姆斯基、句法分析、主语…；

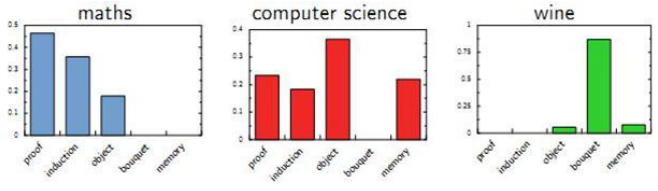
* 谈论概率统计，我们容易想到以下一些词: 概率、模型、均值、方差、证明、独立、马尔科夫链、…；

* 谈论计算机，我们容易想到的词是： 内存、硬盘、编程、二进制、对象、算法、复杂度…；

我们之所以能马上想到这些词，是因为这些词在对应的主题下出现的概率很高。我们可以很自然的看到，一篇文章通常是由多个主题构成的、而每一个主题大概可以用与该主题相关的频率最高的一些词来描述。

以上这种直观的想法由Hoffman 于 1999 年给出的PLSA(Probabilistic Latent Semantic Analysis) 模型中首先进行了明确的数学化。Hoffman 认为一篇文档(Document) 可以由多个主题(Topic) 混合而成， 而每个Topic 都是词汇上的概率分布，文章中的每个词都是由一个固定的 Topic 生成的。

下图是英语中几个Topic 的例子。



Topic 就是Vocab 上的概率分布

所有人类思考和写文章的行为都可以认为是上帝的行为，我们继续回到上帝的假设中，那么在 PLSA 模型中，Hoffman 认为上帝是按照如下的游戏规则来生成文本的。

Game 3 PLSA Topic Model 假设

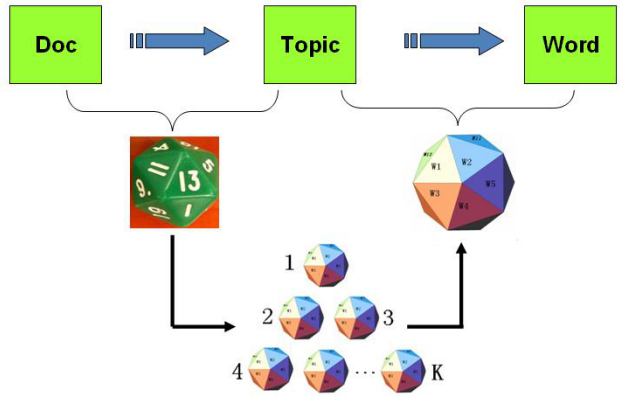
1: 上帝有两种类型的骰子，一类是doc-topic 骰子,每个doc-topic 骰子有K 个面，每个面是一个topic 的编号；一类是topic-word 骰子，每个topic-word 骰子有V 个面，每个面对应一个词；

2: 上帝一共有K 个topic-word 骰子, 每个骰子有一个编号，编号从1 到K；

3: 生成每篇文档之前，上帝都先为这篇文章制造一个特定的doc-topic 骰子，然后重复如下过程生成文档中的词

- 投掷这个doc-topic 骰子,得到一个topic 编号z
- 选择K 个topic-word 骰子中编号为z的那个，投掷这个骰子，于是得到一个词

以上PLSA 模型的文档生成的过程可以图形化的表示为



PLSA 模型的文档生成过程

我们可以发现在以上的游戏规则下，文档和文档之间是独立可交换的，同一个文档内的词也是独立可交换的，还是一个bag-of-words 模型。游戏中的K 个topic-word 骰子，我们可以记为 $\phi_{\rightarrow 1}, \ldots, \phi_{\rightarrow K}$, 对于包含M篇文档的语料 C=(d1,d2, ... ,dM) 中的每篇文档dm，都会有一个特定的doc-topic骰子 $\theta_{\rightarrow m}$，所有对应的骰子记为 $\theta_{\rightarrow 1}, \ldots, \theta_{\rightarrow M}$。为了方便，我们假设每个词w 都是一个编号，对应到topic-word 骰子的面。于是在 PLSA 这个模型中，第m篇文档 dm 中的每个词的生成概率为

$$p(w|dm) = \sum_{z=1}^{K} p(w|z)p(z|dm) = \sum_{z=1}^{K} \phi_{zw}\theta_{mz}$$

所以整篇文档的生成概率为

$$p(\vec{w}|dm) = \prod_{i=1}^{n}\sum_{z=1}^{K} p(w_i|z)p(z|dm) = \prod_{i=1}^{n}\sum_{z=1}^{K} \phi_{zwi}\theta_{dz}$$

由于文档之间相互独立，我们也容易写出整个语料的生成概率。求解PLSA 这个 Topic Model 的过程汇总，模型参数并容易求解，可以使用著名的 EM 算法进行求得局部最优解。

fin

/ 分类算法——决策树CART算法原理及实现 /

编辑 / 协会市场处 宫辰 插图 / 崔峻珩 日期 / 2016-03

1.算法介绍

分类回归树算法：CART(Classification And Regression Tree)算法采用一种二分递归分割的技术，将当前的样本集分为两个子样本集，使得生成的的每个非叶子节点都有两个分支。因此，CART算法生成的决策树是结构简洁的二叉树。

分类树两个基本思想：第一个是将训练样本进行递归地划分自变量空间进行建树的想法，第二个想法是用验证数据进行剪枝。

建树:在分类回归树中，我们把类别集Result表示因变量，选取的属性集attributelist表示自变量，通过递归的方式把attributelist把p维空间划分为不重叠的矩形，具体建树的基本步骤参见：http://baike.baidu.com/view/3075445.htm。

CART算法是怎样进行样本划分的呢？它检查每个变量和该变量所有可能的划分值来发现最好的划分，对离散值如{x,y,x}，则在该属性上的划分有三种情况({{x,y},{z}},{{x,z},y},{{y,z},x})，空集和全集的划分除外；对于连续值处理引进“分裂点”的思想，假设样本集中某个属性共n个连续值，则有n-1个分裂点，每个“分裂点”为相邻两个连续值的均值 (a[i] + a[i+1]) / 2。将每个属性的所有划分按照他们能减少的杂质（合成物中的异质，不同成分）量来进行排序，杂质的减少被定义为划分前的杂质减去划分之后每个节点的杂质量*划分所占样本比率之和，目前最流行的杂质度量方法是：GINI指标，如果我们用k，k=1,2,3……C表示类，其中C是类别集Result的因变量数目，一个节点A的GINI不纯度定义为：

$$\text{Gini}(A) = 1 - \sum_{k=1}^{C} p_k^2$$

其中，Pk表示观测点中属于k类得概率，当Gini(A)=0时所有样本属于同一类，当所有类在节点中以相同的概率出现时，Gini(A)最大化，此时值为(C-1)C/2。

对于分类回归树，A如果它不满足“T都属于同一类别or T中只剩下一个样本”，则此节点为非叶节点，所以尝试根据样本的

每一个属性及可能的属性值，对样本的进行二元划分，假设分类后A分为B和C，其中B占A中样本的比例为p，C为q(显然p + q = 1)。

则杂质改变量：Gini(A) –p*Gini(B)–q*Gini(C)，每次划分该值应为非负，只有这样划分才有意义，对每个属性值尝试划分的目的就是找到杂质该变量最大的一个划分，该属性值划分子树即为最优分支。

剪枝:在CART过程中第二个关键的思想是用独立的验证数据集对训练集生长的树进行剪枝。

分析分类回归树的递归建树过程，不难发现它实质上存在着一个数据过度拟合问题。在决策树构造时，由于训练数据中的噪音或孤立点，许多分枝反映的是训练数据中的异常，使用这样的判定树对类别未知的数据进行分类，分类的准确性不高。因此试图检测和减去这样的分支，检测和减去这些分支的过程被称为树剪枝。树剪枝方法用于处理过分适应数据问题。通常，这种方法使用统计度量，减去最不可靠的分支，这将导致较快的分类，提高树独立于训练数据正确分类的能力。

决策树常用的剪枝常用的简直方法有两种：事前剪枝和事后剪枝，CART算法经常采用事后剪枝方法：该方法是通过在完全生长的树上剪去分枝实现的，通过删除节点的分支来剪去树节点。最下面未被剪枝的节点成为树叶。

CART用的成本复杂性标准是分类树的简单误分(基于验证数据的)加上一个对树的大小的惩罚因素。惩罚因素是有参数的，我们用a表示，每个节点的惩罚。成本复杂性标准对于一个数来说是Err(T)+a|L(T)|,其中Err(T)是验证数据被树误分部分，L(T)是树T的叶节点树，a是每个节点的惩罚成本：一个从0向上变动的数字。当a=0对树有太多的节点没有惩罚，用的成本复杂性标准是完全生长的没有剪枝的树。在剪枝形成的一系列树中，从其中选择一个在验证数据集上具有最小误分的树是很自然的，我们把这个树成为最小误分树。

2.算法实现

本文根据一个样本集，进行了CART算法的简单实现。该样本集中每个样本有十六个特征属性和一个结果属性，为了降低划分的难度，每个特征属性取两个不同的离散值，结果属性有两个离散值：Yes和No。

数据结构定义：在该算法中定义了三种数据结构：存储样本属性名称及取值的Node属性，存储单个样本的EXampleSet属性，树的节点属性dataNode；存放在DataStructure.h中，代码如下：

```
<span style="font–size: 18px;">typedef struct tagNode
    {           //存储属性
                string name;        //属性的名称
                string value;       //属性取值
    }Node;
                typedef struct tagExampleSet
    {           //样本存储
                string example[16];         //样本的每个属性上的属性值
                string decision;    //样本的结果类
    }ExampleSet;
                typedef struct Data_Node{           //节点的数据结构，结果分为两类yes类和No类
                int Yesnum;         //类yes得样本数目
                int Nonum;          //类no得样本数
                vector<ExampleSet> myVector;      //存储样本
                Data_Node *LeftNode;      //左子树
                Data_Node *RightNode;     //右子树
                int Property;       //划分选取的属性
                string Proper_value;        //所选的属性的值
                int nodenum;        //标示节点
                bool leavenode;             //标示叶节点
    }dataNode;
</span>
```

```
<span style="font-size:18px;">typedef struct tagNode
    {           //存储属性
    string name;        //属性的名称
                string value;       //属性取值
    }Node;
                typedef struct tagExampleSet
    {           //样本存储
                string example[16];         //样本的每个属性上的属性值
                string decision;    //样本的结果类
    }ExampleSet;
    typedef struct Data_Node{           //节点的数据结构，结果分为两类yes类和No类
                int Yesnum;         //类yes得样本数目
                int Nonum;          //类no得样本数
                vector<ExampleSet> myVector;      //存储样本
                Data_Node *LeftNode;      //左子树
                Data_Node *RightNode;     //右子树
                int Property;       //划分选取的属性
                string Proper_value;        //所选的属性的值
                int nodenum;        //标示节点
                bool leavenode;             //标示叶节点
    }dataNode;
</span>
```

样本读取及处理：用两个文件分别存储样本的属性及所有样本。文件t存储样本的十六个自变量属性、类别属性的名称和离散值集合，文件t1是所有样本的集合，用ReadFile类读取文件，并把它们分别存储在两个向量中。建树的过程在MySufan类中，该类地方法列表如下：

```
<span style="font-size: 18px;">MySuanfa();
    ~MySuanfa();
    void Method();              //调用建树、剪枝方法
    void BuildTree(Data_Node*thisNode);         //建树方法，每次调用DeviceTree对非叶节点进行划分
    void DeviceTree(Data_Node*thisNode,int i);              //对非叶结点进行划分，分出左节点，有节点
    int Choose_Property(Data_Node* thisNode);               //返回选择的属性值
    double pure(int i1,int i2,int i3);                  //纯度计算函数，每次计算最优划分时用
    void Deal(Data_Node* d);            //剪枝函数，此函数对建好的树用测试样本进行剪枝
    void levelorder(Data_Node * p);     //层次遍历，此方法按曾给决策点分配序号，用于剪枝
    void inorder(Data_Node *p);                 //中序遍历，和建树的前序遍历用于确定树的结构
    void BuildTest(Data_Node *d,int t);   //此方法用于计算当取不同决策点时，建树样本的错误样本数，t为决策点数
目
    void CutTree(Data_Node *d,int k,int t);         //k为单个样本，t为决策点数，根据决策点对测试样本集进行测试
    void ClassOfNode(vector<ExampleSet>);     //本方法用于切割原始样本集，将样本分为测试样本和建树样本
</span>

<span style="font-size:18px;">MySuanfa();
    ~MySuanfa();
```

```
    void Method();    //调用建树、剪枝方法
    void BuildTree(Data_Node*thisNode);          //建树方法，每次调用DeviceTree对非叶节点进行划分
    void DeviceTree(Data_Node*thisNode,int i);          //对非叶结点进行划分，分出左节点，有节点
    int Choose_Property(Data_Node* thisNode);          //返回选择的属性值
    double pure(int i1,int i2,int i3);       //纯度计算函数，每次计算最优划分时用
    void Deal(Data_Node* d);          //剪枝函数，此函数对建好的树用测试样本进行剪枝
    void levelorder(Data_Node * p);       //层次遍历，此方法按曾给决策点分配序号，用于剪枝
    void inorder(Data_Node *p);          //中序遍历，和建树的前序遍历用于确定树的结构
    void BuildTest(Data_Node *d,int t); //此方法用于计算当取不同决策点时，建树样本的错误样本数，t为决策点数目
    void CutTree(Data_Node *d,int k,int t);          //k为单个样本，t为决策点数，根据决策点对测试样本集进行测试
    void ClassOfNode(vector<ExampleSet>);     //本方法用于切割原始样本集，将样本分为测试样本和建树样本
</span>
```

递归建树：建树按照递归方式进行建树，采用全部样本的2/3进行建树，首先找到一个划分值，如果不存在返回–1,然后判断一个树是否为叶子节点，不为叶子节点按照划分值进行划分，关键代码如下：

```
<span style="font–size: 18px;">void MySuanfa:: BuildTree(Data_Node* thisNode)
    {
    if(thisNode!=NULL){          //节点不为空
                nodenum++;
                thisNode–>nodenum=nodenum;
                int getProperty=Choose_Property(thisNode);          //找到划分
                thisNode–>Property=getProperty;
                if((thisNode–>Yesnum*thisNode–> Nonum==0)||getProperty==–1)
                {          //如果划分为–1，则无法再次划分
                thisNode–>Property=–1;
                thisNode–>leavenode=true;
    } else {          //递归建树
    thisNode–>leavenode=false;
    DeviceTree(thisNode,getProperty);          //将父节点按照划分属性进行划分
    BuildTree(thisNode–>LeftNode);          //递归建立左子树
    BuildTree(thisNode–>RightNode);          //递归建立右子树
    }
    }
}</span>

<span style="font–size:18px;">void MySuanfa :  BuildTree ( Data_Node* thisNode )
    { if(thisNode!=NULL){          //节点不为空
    nodenum++;
    thisNode–>nodenum=nodenum;
    int getProperty=Choose_Property(thisNode);          //找到划分
    thisNode–>Property=getProperty;
    if((thisNode–>Yesnum*thisNode–> Nonum ==0)||getProperty==–1) {    //如果划分为–1，则无法再次划分
                thisNode–>Property=–1;
                thisNode–>leavenode=true;
```

```
    }
else {        //递归建树
    thisNode->leavenode=false;
    DeviceTree(thisNode,getProperty);          //将父节点按照划分属性进行划分
    BuildTree(thisNode->LeftNode);             //递归建立左子树
    BuildTree(thisNode->RightNode);            //递归建立右子树
    }
    }
}</span>
```

分析上面代码，Choose_Property(thisNode);函数的作用是将thisNode中的样本尝试进行最优划分，划分的依据就是杂质最大该变量，如果划分成功返回属性下标，否则返回-1，我们在样本中每个属性默认取两个离散值。注意到方法中对书中定义的leavenode和nodenum两个变量的操作，他们的用途是什么呢？nodenum的第一个作用是树的遍历，将每一个节点赋予一个唯一的值，建树的过程是前序建树，建树结束后根据树的中序遍历可以唯一确定树的结构，nodenum的第二个作用和leavenode的作用将会在剪枝过程中用到，后面将会提到。

当建树结束后，树的前序即为nodenum从小到大的排序，然后通过调用中序遍历函数输出树的中序序列，确定树的结构。该树含有17个决策点（非叶子节点），18个叶子节点。

树中决策点的划分代码对应的属性名称：

0 ———— handicapped-infants；
1 ———— water-project-cost-sharing；
2 ———— adoption-of-the-budget-resolution；
3 ———— physician-fee-freeze；
4 ———— el-salvador-aid；
5 ———— religious-groups-in-schools；
6 ———— anti-satellite-test-ban;
7 ———— aid-to-nicaraguan-contras；
8 ———— mx-missile；
9 ———— immigration；
10 ———— synfuels-corporation-cutback；
11 ———— education-spending；
12 ———— superfund-right-to-sue；
13 ———— crime；
14 ———— duty-free-exports；
15 ———— export-administration-act-south-africa；

按照递归分类的算法，最终生成的树的叶子节点中或者同属一类或者只有一个样本，分析树的结构我们可以发现，有两个叶子节点8和23不符合这种情况，却成了叶子节点。

这与所选样本有关，在这两个叶节点中两个样本的十六个特征属性值都相同，只有所属类别不同，所以无法根据递归算法进行分类。另当选取physician-fee-freeze 和adoption-of-the-budget-resolution两种属性进行决策时，样本所属的类别已经基本判定，造成这种情况我们可认为这两种属性在样本中所占的权重很大，只要确定这两种情况，树的大部分样本的分类就已确定。

剪枝：用训练样本建树结束后，就是进行树的剪枝阶段，本算法采用样本集的后1/3作为测试进行剪枝。

树的决策点：如果一个节点为非叶节点，则称该节点为一个树的决策点。树的剪枝就是减去过分拟合给树带来的的冗余，用尽可能少的决策点、尽可能低的树高获取尽可能大的正确率。

如何获取树的决策点？逐层确定树的决策点，并根据决策点数目进行剪枝是剪枝的关键。

根据二叉树的特性可知树的非叶节点=叶节点-1；所以可以从树的节点数中得知树种非叶结点的数量。

本程序根据这一特性将树的决策点逐层赋值，根节点赋值1，根节点的左节点赋值2……，这一过程通过层次遍历实现。并将

该值赋给nodenum，对于叶子节点nodenum为0，

关键代码如下：

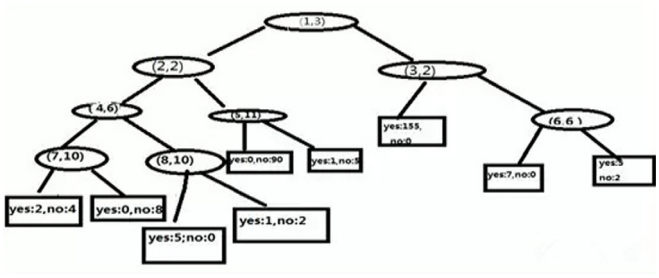
```
<span style="font-size: 18px;">void MySuanfa:: levelorder(Data_Node* p)
    {
            int node=1;
            list<Data_Node *>q;
            if(p)q.push_back(p);
            p->nodenum=node;
    while(!q.empty()) {
            p=q.front();
q.pop_front();
    if(p->LeftNode)  {
    if(p->LeftNode->leavenode) {                    //如果该节点的左节点是子节点，则将nodenum赋0
            15.p->LeftNode->nodenum=0;
    } else {            //否则将该节点赋一个node值，该值表示此决策点的顺序
            node++;
            p->LeftNode->nodenum=node;
            q.push_back(p->LeftNode);
    }
} if(p->RightNode) {
            if(p->RightNode->leavenode){      //如果该节点的右节点是子节点，则将nodenum赋0
    p->RightNode->nodenum=0;
} else {                    //否则将该节点赋一个node值，该值表示此决策点的顺序
    node++;
    p->RightNode->nodenum=node;
    q.push_back(p->RightNode);
    }
    }
    }
}
</span>

<span style="font-size:18px;">void MySuanfa:: levelorder (Data_Node* p)
    {
            int node=1;
            list<Data_Node *>q;
    if(p)q.push_back(p);
            p->nodenum=node;
    while(!q.empty()) {
            p=q.front();
            q.pop_front();
            if(p->LeftNode) {
            if(p->LeftNode->leavenode) {      //如果该节点的左节点是子节点，则将nodenum赋0
            p->LeftNode->nodenum=0;
```

```
    } else {   //否则将该节点赋一个node值，该值表示此决策点的顺序
               node++;
               p->LeftNode->nodenum=node;
               q.push_back(p->LeftNode);
    }
} if(p->RightNode) {
    if(p->RightNode->leavenode)
               {          //如果该节点的右节点是子节点，则将nodenum赋0
               p->RightNode->nodenum=0;
    } else {           //否则将该节点赋一个node值，该值表示此决策点的顺序
               node++;
               p->RightNode->nodenum=node;
               q.push_back(p->RightNode);
               }
    } }
}
</span>
```

遍历结束后，每一个决策点数目可以确定一个树，我们就可以根据树的决策点数对训练样本和测试样本的误差进行统计，怎样根据决策点数确定树的结构？可以将树的前序遍历进行改进，对于t个决策点，节点为0或大于t的都是叶子节点，一旦确定叶子节点，树的结构就清楚了，下图为重新赋值后的树，在该图中，如当有3个决策点时，2的子节点和3的子节点都是叶子节点，当用改进的前序遍历便立时会输出有3个决策点:(1,2,3);4个叶子节点(4,5,0,6)的子树。

不同决策点可对应不同子树，通过前序遍历可以将叶子节点中的错误样本统计出来计算该树情况下错误样本的个数，然后再用测试样本遍历树，统计测试样本再改树下错误样本个数。

通过比较可知当树有8和9个决策点时，测试误差最小，我们取8，因为此时树比9个决策点简单，我们取含有8个决策点为最小误分树。最小误分树结构如下：



最小误分树

上图中最小误分树非叶节点中的两个值，第一个表示决策点表示，第二个表示选择的属性的代码，叶子节点中两数表示每一类的数目。

我们定义最优剪枝的方法是在剪枝序列中含有误差在最小误差树的一个标准差之内的最小树，算出的最小误差率被砍做一个带有标准差等于

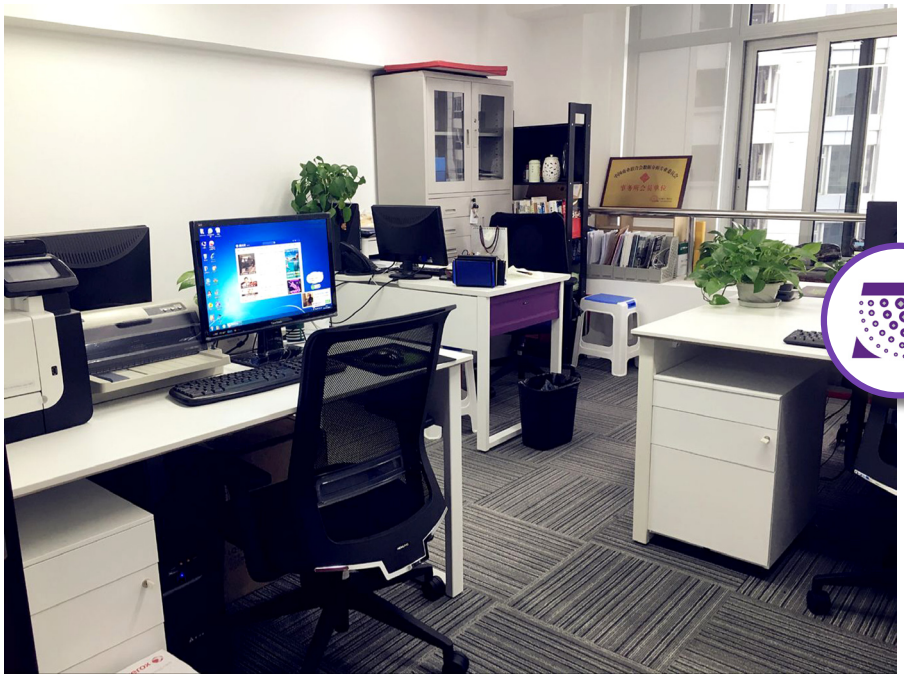
$$\sqrt{E\min(1-E\min)\ /\ Nval}$$

产生的的随机变量的观测值，其中Emin对最小误差树的错误率，Nval是验证集的个数:Emin=5.41%,Nval=148,所以到当树有4个决策点时，为最优剪枝。

fin

/ 上海天元项目数据分析师事务所 /

文图 / 上海天元项目数据分析师事务所　编辑 / 协会会员处 冯伟　插图 / 崔峻珩　日期 / 2016–03



上海天元项目数据分析师事务所（以下简称上海天元）于2012年由多位CPDA数据分析师发起成立。自成立以来，已连续三届获得优秀事务所的称号，得到行业内及行业外的一致认可。**并在2016年获得中国商业联合会数据分析专业委员会常务委员身份**。

从无到有，从有到优，上海天元专注以数据分析为基础，为创业者、企业以及行业研究机构提供智能数据洞察决策服务。

上海天元以决策数据分析为主，同时拓展经营类数据分析，熟练利用R语言、SPSS、MATLAB等数据分析软件，挖掘大量非结构化数据的隐藏信息，并总结其内在规律；以此不断为中小型企业、国内外银行、政府组织等机构提供最专业、最有价值的数据分析服务；帮助客户发现问题，认识问题，规避决策风险，提高企业核心竞争力,保持企业良性发展。

随着大数据分析不断在各行业领域获得广泛的认可，各行业领导、专家与上海天元一直保持着长期广泛的接触和沟通。上海天元充分了解市场所需，对大数据最核心的价值（海量数据的存储和分析）有了新的理解和认知；运用不同维度为企业进行数据分析服务，通过大数据全视角的洞察分析，深度了解企业合作伙伴的客户和战略层次需求；对相关数据进行有效梳理分析，发现问题→研究问题→提出解决方案。从而帮助客户构建适用于自身的数据解决方案平台，帮助客户配置最优质的长期技术支持及数据服务，真正为合作伙伴提升核心竞争力。

上海天元在做好现有数据分析服务业务的基础上，将突破数据分析行业局限，不断深入大数据的各个方面，站在新起点，创造新成绩。

fin

上海天元项目数据分析师事务所有限公司
ShangHai TianYuan Project Data Analyst Firms Co.,Ltd.

- 中商联数据分析专业委员会常务委员单位
- 连续三届获得优秀事务所称号单位

着眼客户核心利益

突破行业局限

专注深度数据分析

上海天元秉承诚信、客观、科学、实效、公正的经营理念；

帮助客户发现问题，认识问题，规避决策风险；

提高企业核心竞争力，保持企业良性发展；

努力成为客户发展历程中 *值得信赖的合作伙伴！*

上海天元项目数据分析师事务所有限公司
ShangHai TianYuan Project Data Analyst Firms Co.,Ltd.
上海市徐汇区天钥桥路329号B栋7层（200030）
Tel：021-24193019、13917778657
Fax：021-24193100
Email：tianyuanfx@126.com、2676759620@qq.com
Web：www.shtianyuan.com