# 数据挖掘与应用统计现状及趋势研究

## ——第八届国际数据挖掘与应用统计研究会年会学术综述

李 勇[1] 张 敏[2] 刘 浩[1] 李禹锋[1] 朱建平[3]

1.重庆工商大学 2.云南财经大学 3.厦门大学

【摘 要】国际数据挖掘与应用统计研究会是我国从事数据挖掘领域研究最早的学术研究团体之一。从2006年以来,为政府、高校、研究机构以及企业界的数据挖掘专家和学者提供了一个学术交流的高端平台。2016年第八届年会的主题是“卓越数据共享统计的理论及应用研究”,此次会议会聚了国内外众多专家学者,共同聚焦数据挖掘和统计应用的发展趋势。

【关键词】共享数据时代; 数据挖掘; 应用统计

【中图分类号】C81 【文献标识码】A 【文章编号】1004-5937(2016)22-0024-02

第八届国际数据挖掘与应用统计研究会年会于2016年7月23—26日在油城大庆隆重召开。本届会议由国际数据挖掘与应用统计研究会主办,东北石油大学、厦门大学数据挖掘研究中心、台北医学大学大数据研究中心、重庆允升科技大数据研究中心和重庆誉锋宸数据信息技术有限公司联合承办。会议主题为“卓越数据共享统计的理论及应用研究”。来自国内外近百所高校、政府和企事业单位的200多位专家学者参会。

会议开幕式由东北石油大学数学与统计学院院长王玉学教授主持。东北石油大学副校长吕延防教授介绍了大庆市貌、学校环境和铁人精神等,对本次会议的作用和意义进行了高度评价。教育部统计学类专业教学指导委员会主任、厦门大学曾五一教授从统计学科如何适应大数据时代的发展角度,对会议的召开提出了进一步的期望。台北医学大学谢邦昌教授结合大庆石油,畅谈了大数据的应用前景。厦门大学朱建平教授从学会的起源到现状,对学会未来的发展前景作了展望。

本届大会除特邀报告外,入选论文52篇。按照论文所涉及的理论领域和方法应用,将入选论文分为数据挖掘与大数据应用、统计理论、统计方法应用及实证分析等专题进行了分组交流讨论。主要学术观点综述如下:

### 一、数据挖掘与大数据研究现状及未来趋势研究

谢邦昌教授在《大数据发展现况与未来发展趋势》中首先阐述了何谓BIG DATA。当你连上脸书按赞打卡、上传照片到网络相簿与朋友分享、上班收发e-mail、用悠游卡买杯咖啡、通过ATM领钱、走进大卖场刷卡购物甚至是进家门开灯,都正在源源不断地创造“海量数据”。这正是云端时代的新金脉。其次是BIG DATA的理论及其应用。最重要的是如何对大数据进行分析,其基本方面如下:⑴数据可视化分析。决策者需要的不是数据本身及分析后的数值,而是庞大数据经分析之后的结果、趋势或现象,利用可视化效果易于被接受。⑵Data Mining算法。这是大数据分析的理论核心,而深入挖掘和快速处理是两大重要课题。⑶预测性分析。如何找出特性、科学建模、预测未来。⑷语义引擎。非结构化数据的多元化给数据分析带来新的挑战,要提高语义引擎设计的智能化水平。⑸数据质量和数据管理。高质量的数据和有效的数据管理可保证分析结果的真实和有价值。最后,真正制约或者成为大数据发展和应用的三个瓶颈:数据收集的合法性、产业链各个环节企业的均衡、大数据有效解读。

国家统计局潘璠博士在《我看当前对大数据的一些非议——兼议大数据应用面临的问题》中指出近几年中国的大数据应用取得了一定的进展,但面临的诸多障碍依然存在,且不断出现一些对大数据的非议之声。这些非议有的有一定道理,有的则失之偏颇。潘璠博士针对这些非议指出大数据是科学技术及社会生产力发展到特定阶段的必然。尽管其发展进程中确实出现了失密、造假等严重问题,但这正说明必须正视大数据的扑面而来,并尽快制定各种应对措施,抓住机遇,保存价值,着力解决出现的各种问题。最后,提出完善法律法规、明确牵头单位、统筹各部门和规范标准等措施。

重庆工商大学李勇在《网络舆情数据挖掘方法及其在意识形态传播新特点中的应用研究》中系统研究了当前网

【作者简介】李勇(1970—),男,重庆人,重庆工商大学统计系教授,硕士生导师,研究方向:大数据分析与贝叶斯统计学习;张敏(1989—),女,重庆人,云南财经大学博士研究生,研究方向:多水平贝叶斯统计;刘浩(1993—),女,重庆人,重庆工商大学硕士研究生,研究方向:经济统计;李禹锋(1995—),男,重庆人,重庆工商大学硕士研究生,研究方向:大数据分析技术;朱建平(1962—),男,山西太原人,厦门大学管理学院教授,博士生导师,研究方向:数据挖掘

络舆情数据挖掘的主要方法，并将这些方法应用于网上意识形态传播新特点的研究中。对互联网出现前后意识形态传播呈现的不同特点进行了对比分析，提炼出意识形态传播在当前 DT 时代的本质特征，结合主流意识形态提出相应的有效传播方式和防范措施。

东北石油大学辛华博士在《基于密度分布的聚类算法研究》中通过密度聚类方法 DBSCAN 二次聚类提高了聚类精度。湖北经济学院陈战波、陶前功、黄小舟和王磊的《基于阿里云音乐平台大数据的歌手流行趋势预测及推荐研究》，山西财经大学舒居安、赵丽琴、刘逸萌的《基于网络舆情的居民购买力倾向指数构造研究》和重庆工商大学李禹锋的《基于网络团购的重庆火锅消费行为分析》等进行了大数据的应用研究。光环国际杨恩博的《大数据人才发展与培养》、广州泰迪智能科技赵云龙的《大数据形势下数据科学人才培养初探》和刘彬的《大数据双创实践探索与服务体系》，从业界不同角度探索了大数据人才培养。

二、统计基本理论及应用研究

台湾淡江大学蔡宗儒教授在《Accelerated Degradation Tests》中，回顾了可靠度分析近期的发展，指出随着制造技术的进步，产品可靠度大幅提升，进而提升了对产品可靠度分析的难度。而传统设限方法和近代加速寿命测试法具有一定局限性，通过研究加速退化测试方法，指出如何针对加速退化数据进行统计推断、评价其可靠度，如何在成本的考察下对加速退化测试实验进行设计，以利后续的测试实验参考。

北京大学房祥忠教授在《EM 算法及其在置信推断中的作用》中指出医学或产品试验费用昂贵等小样本情况，其精确置信推断尤为重要；Buehler 置信限在多维参数或删失数据时，难以计算，并将 EM 算法用于求精确置信限，给出了可靠性领域中的实证。

重庆工商大学李勇在《灰色统计基本理论及其应用》中系统研究了灰数的统计学基本理论和方法。他从随机样本产生灰色估计量和直接从灰色数据开始，构建了一套从数理统计逐步过渡到主要以灰色系统为研究对象的灰色统计方法，如灰数的区间估计、灰数的假设检验、灰数的相关分析和回归分析等，并进行了实例分析。

哈尔滨工业大学张孟琦、田波平在《空间模型参数拟极大似然估计量的渐近性和实证》中提出了双权重矩阵空间回归模型参数的极大似然估计量，包括对数似然函数、集中似然函数和参数估计；证明了相合性和渐进分布性质，并实例进行了空间自相关检验和空间计量模型分析。

天津财经大学杨贵军、于洋、孟杰的《基于 AIC 的粗糙集择优方法》和杨贵军、孙玲莉、董世杰的《三种线性回归多重插补法的模拟研究对比分析》分别从粗糙集择优和回归插补进行了研究。云南财经大学张敏博士在《基于高层次结构的多水平发展模型的统计建模及应用》中研究了拟合高层次嵌套数据的多水平发展建模问题。集美大学纪崑的《模糊数据 Jonckheere- Terpstra 检验法及应用》探讨了模糊数据检验。广东财经大学的刘照德、林海明在《因子分析五个争议的解答》中定量分析了因子分析的争议问题。湖南大学周四军、王佳星、罗丹在《基于门限面板模型的我国能源利用效率研究》中，基于柯布—道格拉斯生产函数理论构建了我国能源利用效率门限面板模型，并进行了实证分析。

三、统计方法及实证研究

天津财经大学杨贵军、孟杰、邹文慧在《基于模型平均的中国总和生育率估计》中指出目前国内学者对中国总和生育率的估计尚未形成一致性的结论，缺少高质量的数据源以及不完善的估计方法是影响总和生育率估计的主要问题；提出使用社会和经济等“人口系统”外部数据，引入当前统计学和计量经济学前沿的模型平均方法对中国总和生育率进行估计。

华侨大学项后军和浙江财经大学何康在《自贸区的影响与资本流动——以上海为例的“自然实验”估计》中，从自然实验角度考察了样本期内上海自贸区的设立对上海地区资本流动的影响。得出：基于双重差分模型估计的自贸区对上海资本流动的影响显著；基于改进后合成控制法得到的“合成上海”对上海设立自贸区之前的模拟程度更高；基于安慰剂检验，证实了自贸区政策的有效性。

湖南大学晏艳阳、邓嘉宜、文丹艳在《邻里效应与居民政治信任——基于中国家庭追踪调查（CFPS）的证据》中，指出近年来居民对政府的信任危机频发，矛盾不断出现，严重制约着政府的行政效率；基于中国家庭追踪调查（CFPS）截面数据，建立回归模型进行实证分析，证实了其他信息获取渠道与社会互动之间具有相互替代的关系，有效解决了关联效应和反射性问题对邻里效应估计带来的影响。

中国南方电网科学研究院冷媛、傅蔷、陈政和厦门大学范新妍在《基于 MCP，Group MPC 的先行、一致、滞后指标筛选》中，提出了基于 MCP 惩罚法的单一指标先行、一致、滞后性的判定方法和基于 Group MCP 的多指标系统下各个指标的先行、一致、滞后性的判定方法。冷媛、傅蔷和厦门大学孙俊歌、梁振杰在《经济景气指数研究比较及思考》中梳理了国内外景气指数的研究状况。辽宁大学马树才、宋琪在《中国人口年龄结构变动对资本投入及经济增长影响研究》中通过构建数理模型，就人口年龄结构对资本投入及经济增长的影响进行研究，得出充足的劳动供给会提高教育人力资本和物质资本的使用效率，促进经济增长，政府公共教育支出增加会提高教育人力资本对经济增长的贡献；并对面板数据进行实证分析。厦门大学刘云霞在《我国高技术产业创新绩效影响因素动态比较研究——基于状态空间和门槛模型相结合的研究》中确定了

# 共享数据聚变时代的经济统计趋势研究

## ——第十届全国企业经济统计学年会学术综述

张 敏[1] 刘 浩[2] 周世铭[2] 李禹锋[2] 李 勇[2]

1.云南财经大学 2.重庆工商大学

【摘 要】全国企业经济统计学年会是我国经济统计学界一年一度的盛会。前后近 30 年,为我国经济统计的研究、教学和服务构建了一个高学术、高规格的产—学—研一体化平台。2016 年第十届年会的主题是“共享数据聚变时代下的经济统计理论及应用研究”。汇聚了国内众多专家学者,共同探讨大数据时代下我国经济统计的未来发展趋势。

【关键词】共享数据; 聚变时代; 经济统计

【中图分类号】C81 【文献标识码】A 【文章编号】1004- 5937(2016)22- 0026- 03

第十届全国企业经济统计学年会于 2016 年 7 月 16—18 日在兰州隆重召开。会议由全国企业经济统计学会主办,兰州财经大学统计学院、重庆允升科技大数据研究中心和重庆誉锋宸数据信息技术有限公司联合承办。会议的主题是:“共享数据聚变时代下的经济统计理论及应用研究”。全国近百所高校、政府和企事业单位的 200 位专家学者参会。

国家统计局副局长许宪春博士针对我国当前经济发展态势作了《2016 年上半年经济形势分析》报告,北京师范大学邱东教授针对空间经济比较中由购买力平价推断存在的宾大效应等问题作了《BHPPP 中的纯价比假设与宾大效应的弱存在》报告,厦门大学杨灿教授基于投入产出分析的扩展框架作了《产业关联测度与关键产业甄别》报告,暨南大学刘建平教授针对我国政府统计调查体系在新时代面临的问题作了《深化我国政府统计调查体系改革的思考与建议》报告,浙江财经大学李金昌教授针对大数据

【作者简介】张敏(1989— ),女,重庆人,云南财经大学博士研究生,研究方向:多水平贝叶斯统计;刘浩(1993— ),女,重庆人,重庆工商大学硕士,研究方向:经济统计;周世铭(1994— ),女,重庆人,重庆工商大学硕士,研究方向:数据挖掘;李禹锋(1995— ),男,重庆人,重庆工商大学硕士,研究方向:大数据分析技术;李勇(1970— ),男,重庆人,重庆工商大学统计系教授,硕士生导师,研究方向:大数据分析与贝叶斯统计学习

反映创新绩效的指标以及影响创新绩效的因素,再将状态空间模型和门口模型进行有机结合,找出了各影响因素对创新绩效的动态影响轨迹以及轨迹改变的关键点,并提出对策建议。

南京财经大学李昌峰、何红、李珂在《产业结构对跨越“中等收入陷阱”的影响研究》中,构建了基于中等收入陷阱的固定效应变截距定量分析模型并进行实证研究。吉林财经大学燕苗霞在《中国各地区城市基础设施水平综合评价研究》中,利用因子分析和聚类分析法构建我国城市基础设施水平综合评价模型。红豆集团魏昊和集美大学庄赟在《交通运输设备制造业服务化及影响因素研究》中,构建了我国交通运输设备制造业服务化影响因素指标体系,并基于 Cobb- Douglas 生产函数模型,对服务化影响因素发挥效用的方式及强度进行了实证检验。对外经济贸易大学凌志明、王景乐在《基于 Copula 模型变点检测的投资者情绪传染分析》中,建立了主成分分析的综合投资者情绪指标,构建了基于非参数的最佳 Copula 函数模型并进行了实证分析。福建农林大学阙翔、吴冲龙,刘金福在《基于地质统计学法的岩石含 Pb 空间分布特征预测分析》中,基于区域化变量理论和变异函数,探讨地质统计学的估值、局部不确定性预测、随机模拟和多点地质统计等方法。山西财经大学刘逸萌、赵丽琴、舒居安在《BP 神经网络在太原市 PM2.5 浓度预测中的应用》中,构建了预测太原市 PM2.5 浓度的 BP 神经网络模型。河北经贸大学刘金玲在《宏观政策对房地产价格的动态研究——基于 VAR 模型的实证》中,构建了 VAR 模型分析土地交易价格、货币供应量与利率变化冲击对中国房地产价格的动态影响。

综上所述,每年一届的“国际数据挖掘与应用统计研究会”现已成为我国数据挖掘与应用统计领域高水平的学术会议之一。本届会议较全面地总结和交流了我国数据挖掘与应用统计领域的最新研究进展和成果,对进一步促进相关领域的发展起到了积极作用。●