第八届国际数据挖掘与应用统计研究会学术简述

李　勇[1]，张　敏[2]，朱建平[3]

（1.重庆工商大学；2.云南财经大学；3.厦门大学）

第八届国际数据挖掘与应用统计研究会于2016年7月23－26日在大庆市隆重召开。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和重庆允升科技大数据研究中心联合承办。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100所高校、政府和企事业单位200多位专家学者莅临参加。大会入选论文52篇，分为大数据分析及应用、统计理论及方法应用等专题。

一、大数据发展面临问题与未来趋势研究

台北医学大学谢邦昌教授在《大数据发展现况与未来发展趋势》报告中，首先阐述了BIG DATA，并指出BIG DATA基本方法主要有数据可视化分析、Data Mining算法、预测性分析、语义引擎和数据质量与数据管理等五方面。最后阐述了BIG DATA的未来及其瓶颈：①数据收集和提取合法性，数据隐私保护和数据隐私应用间的权衡。指出目前全世界对于用户隐私应如何保护、商业规则应如何制定、触犯用户隐私权应如何惩治、法律规范应如何制定等系列管理问题大大滞后于大数据发展速度，数据源头采集受限将大大限制大数据的商业应用。②大数据发挥协同效应需要产业链各个环节的企业达成竞争与合作的平衡。共享信息的企业间如何权衡利弊、平衡利益，而权威第三方中立机构的缺乏将制约大数据发挥其最大潜力。③大数据结论的解读和应用。大数据可从数据层面揭示各变量可能的关联，但如何具像到行业实践中？如何制定可执行方案？这需要从技术层面、行业背景和管理能力三者融汇的执行者，而深谙技术又具有系统思维的卓越管理人才的稀缺将制约大数据的发展。

国家统计局潘璠博士在报告《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，列举了中国大数据应用中存在五个非议：若大数据不分析，就成为一堆“大垃圾”；全国性普查数据够不上大数据；“全样本是永远不可能的”；大数据只能说明总体现在，若要说明未来，还不如小数据；要对所谓“伪大数据说不”。针对这些非议，他阐述了自己的观点，指出尽管大数据在发展中确实出现了失密、造假等严重问题，但这正需要正视和应对，必须完善法律法规、明确牵头单位统筹和规范标准等。东北石油大学辛华博士在《基于密度分布的聚类算法研究》中提出，基于先验分布的引力聚类法选择最优类簇中心。一些学者对大数据进行了应用研究。

二、统计模型的理论与应用研究

北京大学房祥忠教授在报告《EM算法及其在置信推断中的作用》中指出，面对小样本问题的精确置信推断很重要，但处理多维参数或不完全观测数据会涉及计算困难。首次将用于求解最大似然估计和后验分布最大值的EM算法应用于精确置信推断。给出了Buehler置信限新型式，利用EM算法计算其概率最大值，并用指数单元串联系统、两个二项分布比例差和Gamma随机变量均值的精确置信限三个实例进行说明。

台湾淡江大学蔡宗儒教授在报告《Accelerated Degradation Tests》中，探讨如何针对加速退化数据进行统计推论，评价其可靠度，以及考虑成本下的实验设计，以利后续测试参考。指出寿命测试传统设限方法和近代加速寿命测试法，面对高可靠度数据仍会失效。为此提出加速退化测试方法，观测与产品寿命高度相关的代理变数退化的数据。以伽马随机过程代替布朗运动与对数布朗运动，对产品进行可靠度推断构建和最佳实验设计。

重庆工商大学李勇在《层次贝叶斯统计方法及应用研究》报告中，针对层次贝叶斯基本方法进行了系统研究。构建了基于不同层次来源的嵌套结构复杂数据多层次模型；将贝叶斯分析应用于层次模型，利用先验分布优先等对参数进行推断；构建了基于贝叶斯后验分布的层次模型参数假设检验和基于贝叶斯因子与预测分布的层次评价模型；构建了层次贝叶斯雾霾监测评估模型。天津财经大学杨贵军等在《三种线性回归多重插补法的模拟研究对比分析》中，采用普通线性回归多重插补法、贝叶斯线性回归多重插补法、贝叶斯自助线性回归多重插补法对无回答进行插补研究。云南财经大学张敏博士在报告《基于高层次结构的多水平发展模型的统计建模及应用》中，构建了农户收入二、三、四水平发展模型，通过比较体现多水平模型拟合高层次嵌套数据的优势。一些学者对统计方法进行了应用研究。

（责任编辑：杜一哲）