**※2022年度广西自然科学奖公示材料**

1. **成果名称：**数据预处理的若干模型（自治区科技成果登记号：201712188）
2. **拟提名者：**广西壮族自治区教育厅
3. **成果简介**

信息技术与经济社会的交汇融合引发了数据迅猛增长，数据已在国务院《促进大数据发展行动纲要》（2015年8月）被列为国家基础性战略资源。作为数据分析的第一步，数据预处理通过对数据进行清理、集成、降维和压缩等操作输出高质量数据，满足现代数据方法进行数据分析的要求。图灵奖得主Michael Stonebraker教授认为数据分析需从干净、准确、统一的数据开始。计算机科学领域的先驱Charles Babbage指出没有高质量的数据，就没有高质量的数据分析结果。“数据分析方法通常建立在高质量数据基础”论断不但成为计算机科学与信息通讯技术的经典习语并对计算机科学的发展起到了积极推动作业，也推动了许多世界大公司的重大变革，如Google, 通用和丰田汽车等公司。但是，在真实世界中，数据通常是不完整的（缺少某些感兴趣的属性值）、冗余的（包含冗余的样本和属性）、极易受到噪声（错误或异常值）的侵扰的。数据的不准确、不完整影响了公司盈利，还因此造成资源浪费、生产力损失和交流上的额外开支。因此，现实世界的数据质量和现有数据分析方法存在着一条不可逾越的鸿沟，严重地限制了数据分析的发展。本项目开展对数据预处理的研究取得了以下创新成果：充分挖掘数据间隐含的信息去揭示数据间存在的复杂关系，进而发现属性重要性和稀疏理论的关联性，为此提出了基于稀疏理论的数据预处理方法突破性的解决了数据预处理若干问题，并构建了基于医学影像数据的计算机辅助诊断系统，为数据分析理论和应用的发展开辟了新途径。重要科学点如下：

1. 提出多种稀疏编码方法有效解决当前高维数据高存储成本和知识发现低效等问题，系统的解决了高维数据特征选择难题，并创建了一套神经影像分析的计算机辅助诊断系统。
2. 提出线性复杂度的单模态哈希和跨模态哈希方法，创新性地解决了无法有效在现代计算机内存进行高维大数据检索的难题，显著提高了大数据近似检索的性能。
3. 在不完备数据填补研究领域提出混合核填补方法和有序代价敏感填充方法，被国际同行应用到医学图像分析等领域。

上述理论和应用成果得到近十位ACM和IEEE双Fellow（如Philip S. Yu、Qiang Yang和周志华等）、20多位IEEE Fellow （如Sabine Van Huffel、Andrzej Cichocki、 Bastiaan Kleijn和Edwin R. Hancock等）、中科院院士郑南宁、多位欧洲科学院院士（如Xuemin Lin、李学龙和Heng Tao Shen等），俄罗斯工程院外籍院士吴信东，新加坡科学院院士Beng Chin Ooi和澳大利亚科学院和欧洲科学院外籍双院院士Dacheng Tao等国内外同行高度认可。

本项目的8篇代表性论文发表在权威期刊和会议，获谷歌他引1388次和SCI他引879次，含4篇ESI热点论文和5篇ESI高被引论文。第一完成人是国家千人计划创新人才获得者， Elsevier中国高被引学者（2020-2021）和科睿唯安全球高被引科学家（2019-2021）。

1. **代表性论文（专著）目录（不超过8篇）**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 排序 | 类型 | 论文专著名称 | 年卷页（版号） | 发表日期 | 作者 | 署名单位 | 刊名 | 通讯作者 | 他引次数 | 检索数据库 | 广西单位是否署名 | 附件编号 |
| 1 |  | Robust joint graph sparse coding for unsupervised spectral feature selection | 26 (6), 1263 – 1275(2016) | 2016-6 | Xiaofeng Zhu; Xuelong Li; Shichao Zhang; Chunhua Ju; Xindong Wu | 广西师范大学 | IEEE Transactions on Neural Networks and Learning Systems | Shichao Zhang  Xindong Wu | 267 | SCI | 是 |  |
| 2 |  | Block-row sparse multiview multilabel learning for image classification | 46 (2), 450-461(2015) | 2015-2 | Xiaofeng Zhu; Xuelong Li; Shichao Zhang | 广西师范大学 | IEEE Transactions on Cybernetics | Shichao Zhang | 244 | SCI | 是 |  |
| 3 |  | Linear cross-modal hashing for efficient multimedia search | 143-152(2013) | 2013-10 | Xiaofeng Zhu; Zi Huang; Heng Tao Shen; Xin Zhao | 广西师范大学 | ACM international conference on Multimedia | Xiaofeng Zhu | 216 | EI | 是 |  |
| 4 |  | Graph PCA Hashing for Similarity Search | 19(9): 2033-2044(2017) | 2017-9 | Xiaofeng Zhu; Xuelong Li; Shichao Zhang; Zongben Xu; Litao Yu | 广西师范大学 | IEEE Transactions on Multimedia | Shichao Zhang  Zongben Xu | 104 | SCI | 是 |  |
| 5 |  | Missing value estimation for mixed-attribute data sets | 23(1): 110-121(2011) | 2011-1 | Xiaofeng Zhu; Shichao Zhang; Zhi Jin; Zili Zhang; Zhuoming Xu | 昆士兰大学 | IEEE Transactions on Knowledge and Data Engineering | Shichao Zhang | 274 | SCI | 否 |  |
| 6 |  | Unsupervised feature selection by self-paced learning regularization | 132: 4-11(2020) | 2020-4 | Wei Zheng; Xiaofeng Zhua; Guoqiu Wen; Yonghua Zhu; Hao Yu; Jiangzhang Gan | 广西师范大学 | Pattern Recognition Letters | Xiaofeng Zhu | 85 | SCI | 是 |  |
| 7 |  | A novel relational regularization feature selection method for joint regression and classification in AD diagnosis | 38: 205-214(2017) | 2017-5 | Xiaofeng Zhu, Heung-Il Suk, Li Wang, Seong-Whan Lee, Dinggang Shen, Alzheimer’s Disease Neuroimaging Initiative | 北卡罗莱纳大学 | IEEE transactions on neural networks and learning systems | Seong-Whan Lee, Dinggang Shen | 156 | SCI | 否 |  |
| 8 |  | Efficient utilization of missing data in cost-sensitive learning | *33*(6), 2425-2436(2019) | 2019-11 | Xiaofeng Zhu; Jianye Yang; Chengyuan Zhang; Shichao Zhang | 电子科技大学 | IEEE Transactions on Knowledge and Data Engineering | Jianye Yang; Chengyuan Zhang; | 42 | SCI | 否 |  |
| 合计 | | | | | | | | | 1388 | / | / | / |

**五、 候选人姓名**

朱晓峰，文国秋，郑威，甘江璋

**六、候选单位名称**

广西师范大学