2018年度自治区科技奖励提名项目公示信息
（自然奖）

一、 项目名称

高维复杂数据的特征分析及高效学习方法与应用

二、 提名单位

中国科学院新疆分院

三、 提名单位（专家）意见

项目针对高维复杂数据的海量、高冗余、高噪声、多类别、多标记等特性，提出了一套完整表征复杂数据的特征建模、特征提取及维数约简的新方法；探索了生物序列与自然语言中词的等价类、语法和语义规则，建立了生物语言学模型，定义了“生物语言处理”，提出了若干基于统计语言学和计算语言学技术和理论的生物序列结构和功能识别新方法。建立了基于智能计算的复杂数据高效学习方法。

项目研发难度大、成果丰富，发表代表性SCI学术论文58篇、申请发明专利8件、软件著作权登记2项，培养硕博士研究生20人，项目研究成果得到国内外同行认可，其中SCI他人引用1555次，项目在基础研究上有重大创新，应用研究上有重大突破。

提名该项目为自治区自然科学奖一等奖。

四、项目简介

随着信息技术的飞速发展，国民经济各行业所获得的数据越来越多地呈现出海量、高维、异构、非线性、不完全、不精确与动态时变等高复杂性特征。传统的数据处理方法在面对这些高复杂性数据时，往往收效甚微，使得蕴含在这些

数据中的信息或规律无法被探索和理解，导致“数据资源”变成“数据灾难”。同时，信息技术和国民经济的发展却迫切需要去探索和揭示隐藏在这些数据中的规律和奥秘。因此，如何有效地从复杂数据中获取信息或规律已成为当今信息科学技术领域所面临的基本科学问题之一。本项目所开展的是针对高复杂性数据的智能信息处理及其应用研究，无论是从信息处理基本的理论研究，还是对国民经济的发展来说，都具有极其重要的意义。

1、项目研究对高复杂性异构数据建立相关的分析系统和处理模型，提出了高复杂性异构数据处理的相关理论，并能大力推动智能技术和方法在各领域的广泛应用和发展。项目的研究发展并丰富了信息技术处理的方法，因而为了解更多的未知世界提供更多可以选择的工具和手段。

2、信息技术的不断发展对现有的机器学习理论不断提出挑战，因此，现有的机器学习理论或算法在应用到复杂数据上还存在很多未解决的问题。针对现有机器学习方法在处理复杂数据上的不足，项目提出和发展了针对复杂数据处理和实际应用的高效学习理论和算法，使之能在当前实际应用研究中发挥更大作用。

3、近年来，在生命科学研究中涌现出了大量复杂数据，使得数据获取的能力已远远超出数据处理技术的发展。生物信息学是一个融合多门学科的领域，其主要目标是从异常复杂、数目繁多的实验数据及相关数据库中提取有用的生物信息。机器学习具有从数据和经验中获取知识的学习能力，因此成为生物信息学中数据分析的重要手段。项目研究还能够帮助了解蕴涵于高复杂性生物信息数据中的一些规律，从而

更加有利于人们利用这些规律为人类的生活和生产服务。

五、代表性论文专著目录：

(1) You Z. H., Li X., Chan KCC. An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers[J]. NEUROCOMPUTING, 2017,228(SI):277-282. (WOS:000393017900030)

(2) Zhu H. J., Jiang T. H., Ma B., You Z. H., Shi W. L., Cheng L., Chen X.. HEMD: A highly efficient Random Forests based malware detection framework for Android[J]. NEURAL COMPUTING AND APPLICATIONS, 2017, 30(11): 3353-3361. (WOS: 000451178200007)

(3) Wang Y. B., You Z. H., Li X., Jiang T. H., Chen X., Zhou X., Wang L.. Predicting protein-protein interactions from protein sequences by stacked sparse auto-encoder deep neural network[J]. MOLECULAR BIOSYSTEMS, 2016, 12(11): 3702-3710.

(4) Huang Y. A., You Z. H., Li X., Chen X., Hu P. W., Li S., Luo X. Construction of reliable protein-protein interaction networks using weighted sparse representation based classifier with pseudo substitution matrix representation features[J]. NEUROCOMPUTING, 2016,218:131-138.(WOS:000388053700014)

(5) Luo X., Zhou M. C., Li S., You Z. H., Xia Y. N., Zhu Q. S. A Nonnegative Latent Factor Model for Large-Scale Sparse Matrices in Recommender Systems via Alternating Direction Method[J]. IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, 2016,27(3):579-592. (WOS:000372022900007)

(6) Huang Y. A., You Z. H., Chen X., Yan G. Y. Improved protein-protein interactions prediction via weighted sparse representation model combining continuous wavelet descriptor and PseAA composition[J]. BMC SYSTEMS BIOLOGY, 2016,104(120).(WOS:000392598000010)

六、主要完成人情况

公示姓名	排名	行政职务	技术职称	工作单位	完成单位	对本项目技术创造性贡献
尤著宏	1	无	研究员	中国科学院新疆理化技术研究所	中国科学院新疆理化技术研究所	提出了一套完整表征复杂数据的特征建模、特征提取及维数约简的新方法。建立了基于智能计算的复杂数据高效学习方法。对主要科技创新点(1)、(2)、(3)做出了创造性贡献。

程力	2	无	研究员	中国科学院新疆理化技术研究所	中国科学院新疆理化技术研究所	建立了大规模疾病-微生物关联关系预测计算模型，为生物信息学技术应用于复杂疾病-微生物关联关系的临床研究奠定了一定的基础。对主要科技创新点(2)、(3)做出了重要贡献。
周喜	3	研究室主任	研究员	中国科学院新疆理化技术研究所	中国科学院新疆理化技术研究所	探索了生物序列与自然语言中词的等价类、语法和语义规则，建立了生物语言学模型，定义了“生物语言处理”对主要科技创新点(2)、(3)做出了重要贡献。
李晓	4	无	研究员	中国科学院新疆理化技术研究所	中国科学院新疆理化技术研究所	提出了若干基于统计语言学和计算语言学技术和理论的生物序列结构和功能识别新方法。对主要科技创新点(2)、(3)做出了创造性贡献。
王轶	5	无	副研究员	中国科学院新疆理化技术研究所	中国科学院新疆理化技术研究所	在智能计算模型训练、数据预处理及模型验证方面取得了多项进展。对主要科技创新点(1)做出了重要贡献。

七、完成人合作关系说明

第一完成人尤著宏研究员与其他四位完成人为同事关系。其中，尤著宏研究员是代表作 1、5、7 的第一作者，代表做 2-4、6 的通讯作者。第二完成人程力研究员，作为共同作者，与尤著宏研究员共同发表了代表作 2。第三完成人周喜研究员，作为共同作者与尤著宏研究员共同发表了代表作 3。第四完成人李晓研究员，作为共同作者与尤著宏研究员共同发表了代表作 1、3、4。

八、知情同意证明

知情同意证明

本人知晓中科院新疆理化技术研究所尤著宏研究员使用我们的合作论文申报 2018 年新疆自治区自然科学奖，并同意不作为完成人推荐奖励。

特此证明。

代表作中有署名的作者	本人知情同意签名	代表作中有署名的作者	本人知情同意签名
李政伟		罗辛	
马博		桂杰	
陈兴		雷迎科	
施炜雷		周小波	
王延斌		严桂英	
黄裕安		黄志安	
胡鹏伟		蒋同海	
Chan KCC			

尤著宏

项目总负责人

2019 年 8 月 19 日