专业英语

一、人工智能相关

1.人工智能系统安全与隐私风险。

Security and Privacy Risks in Artificial Intelligence Systems

摘要： 人类正在经历着由深度学习技术推动的人工智能浪潮，它为人类生产和生活带来了巨大的技术革新.在某些特定领域中，人工智能已经表现出达到甚至超越人类的工作能力.然而，以往的机器学习理论大多没有考虑开放甚至对抗的系统运行环境，人工智能系统的安全和隐私问题正逐渐暴露出来.通过回顾人工智能系统安全方面的相关研究工作，揭示人工智能系统中潜藏的安全与隐私风险.首先介绍了包含攻击面、攻击能力和攻击目标的安全威胁模型.从人工智能系统的4个关键环节——数据输入(传感器)、数据预处理、机器学习模型和输出，分析了相应的安全隐私风险及对策.讨论了未来在人工智能系统安全研究方面的发展趋势.

关键词: **智能系统安全, 系统安全, 数据处理, 人工智能, 深度学习**

Abstract: Human society is witnessing a wave of **artificial intelligence (AI)** 【人工智能】driven by **deep learning techniques**【深度学习】, bringing a technological revolution for human production and life. In some specific fields, AI has achieved or even surpassed human-level performance. However, most previous **machine learning theories**【机器学习理论】 have not considered the open and even adversarial environments, and the security and privacy issues are gradually rising. Besides of insecure code implementations, biased models, adversarial examples, sensor spoofing can also lead to security risks which are hard to be discovered by traditional security analysis tools. This paper reviews previous works on AI system security and privacy, revealing potential security and privacy risks. Firstly, we introduce a threat model of AI systems, including attack surfaces, attack capabilities and attack goals. Secondly, we analyze security risks and counter measures in terms of four critical components in AI systems: data input (sensor), data preprocessing, machine learning model and output. Finally, we discuss future research trends on the security of AI systems. The aim of this paper is to arise the attention of the computer security society and the AI society on security and privacy of AI systems, and so that they can work together to unlock AI's potential to build a bright future.

Key words: **intelligent system security, system security, data processing, artificial intelligence (AI), deep learning**

2.智慧教育研究现状与发展趋势

The State of the Art and Future Tendency of Smart Education

摘要： 当前，以大数据分析、人工智能等信息技术为支撑的智慧教育模式已成教育信息化发展的趋势，也成为学术界热点的研究方向.首先，对教学行为、海量知识资源2类教育大数据的挖掘技术进行调研分析；其次，重点论述了导学、推荐、答疑、评价等教学环节中的4项关键技术，包括学习路径生成与导航、学习者画像与个性化推荐、智能在线答疑以及精细化评测，进而对比分析了国内外主流的智慧教育平台;最后，探讨了当前智慧教育研究的局限性，总结出在线智能学习助手、学习者智能评估、网络化群体认知、因果关系发现等智慧教育的研究发展方向.

关键词: 智慧教育, 教育大数据, 大数据分析, 人工智能, 知识图谱

Abstract: At present the **smart education【智能教育】** pattern supported by information technology such as **big data analytics【大数据分析】** and artificial intelligence has become the trend of the development of education informatization, and also has become a popular research direction in academic hotspots. Firstly, we investigate and analyze the **data mining【数据挖掘】** technologies of two kinds of educational big data including teaching behavior and massive knowledge resources. Secondly, we focus on four vital technologies in teaching process such as learning guidance（导学）, recommendation（推荐）, Q&A （答疑）and evaluation（评估）, including **learning path generation and navigation【学习路径的生成与导航】**, **learner profiling【学习者画像】** and **personalized recommendations【个性化推荐】**, **online smart Q&A【在线智能答疑】** and **precise evaluation【精细化测评】**. Then we compare and analyze the mainstream（主流） smart education platforms（平台） at home and abroad. Finally, we discuss the limitations of current smart education research and summarize（总结） the research and development directions of **online smart learning assistants【在线智能学习助手】**, **learner smart assessment【学习者智能评估】**, **networked group cognition【网络化群体认知】**, **causality discovery【因果关系发现】** and other smart education aspects（方向）.

Key words: smart education智慧教育, educational big data教育大数据, big data analytics大数据分析, artificial intelligence人工智能, knowledge graph知识图谱

3.智能芯片的评述和展望。

A Survey of Artificial Intelligence Chip

摘要： 近年来，人工智能技术在许多商业领域获得了广泛应用，并且随着世界各地的科研人员和科研公司的重视和投入，人工智能技术在传统语音识别、图像识别、搜索/推荐引擎等领域证明了其不可取代的价值.但与此同时，人工智能技术的运算量也急剧扩增，给硬件设备的算力提出了巨大的挑战.从人工智能的基础算法以及其应用算法着手，描述了其运算方式及其运算特性.然后，介绍了近期人工智能芯片的发展方向，对目前智能芯片的主要架构进行了介绍和分析.而后，着重介绍了DianNao系列处理器的研究成果.该系列的处理器为智能芯片领域最新最先进的研究成果，其结构和设计分别面向不同的技术特征而提出，包括深度学习算法、大规模的深度学习算法、机器学习算法、用于处理二维图像的深度学习算法以及稀疏深度学习算法等.此外，还提出并设计了完备且高效的Cambricon指令集结构.最后，对人工神经网络技术的发展方向从多个角度进行了分析，包括网络结构、运算特性和硬件器件等，并基于此对未来工作可能的发展方向进行了预估和展望.

关键词: 人工智能, 加速器, FPGA, ASIC, 权重量化, 稀疏剪枝

Abstract: In recent years, artificial intelligence （AI）technologies have been widely used in many commercial fields. With the attention and investment of scientific researchers and research companies around the world, AI technologies have been proved their irreplaceable（不可替代的） value in **traditional speech recognition【传统语音识别】**, **image recognition【图像识别】**, **search/recommendation engine【搜索/推荐引擎】** and other fields. However, at the same time, the amount of computation(运算量） of AI technologies increases dramatically, which poses a huge challenge to the computing power of hardware equipments. At first, we describe the basic algorithms（算法） of AI technologies and their application algorithms in this paper, including their **operation modes【运算方式】** and **operation characteristics【运算特性】**. Then, we introduce the development directions of AI chips（芯片） in recent years, and analyze the main architectures of AI chips. Furthermore, we emphatically（着重） introduce the researches of DianNao series **processors（处理器）**. This series of processors are the latest and most advanced researches in the field of AI chips. Their architectures and designs are proposed（提议） for different technical features（技术特征）, including deep learning algorithms, **large-scale deep learning algorithms【大规模深度学习算法】**, machine learning algorithms, **deep learning algorithms for processing two-dimensional images【处理二维图像的深度学习算法】** and **sparse deep learning algorithms【稀疏深度学习算法】**. In addition, a complete and efficient instruction architecture(ISA) for deep learning algorithms, Cambricon, is proposed. Finally, we

analyze the development directions of **artificial neural network technologies【人工神经网络技术】** from various angles, including **network structures【网络结构】**, **operation characteristics【运算特性】** and hardware devices. Based on the above, we predict and prospect the possible development directions of future work.

Key words: artificial intelligence, accelerators【加速器】, FPGA, ASIC, weight quantization【权重量化】, sparse pruning【稀疏剪枝】

二、机器学习相关

1.基于机器学习的智能路由算法综述

A Survey on Machine Learning Based Routing Algorithms

摘要： 互联网的飞速发展催生了很多新型网络应用，其中包括实时多媒体流服务、远程云服务等.现有尽力而为的路由转发算法难以满足这些应用所带来的多样化的网络服务质量需求.随着近些年将机器学习方法应用于游戏、计算机视觉、自然语言处理获得了巨大的成功，很多人尝试基于机器学习方法去设计智能路由算法.相比于传统数学模型驱动的分布式路由算法而言，基于机器学习的路由算法通常是数据驱动的，这使得其能够适应动态变化的网络环境以及多样的性能评价指标优化需求.基于机器学习的数据驱动智能路由算法目前已经展示出了巨大的潜力，未来很有希望成为下一代互联网的重要组成部分.然而现有对于智能路由的研究仍然处于初步阶段.首先介绍了现有数据驱动智能路由算法的相关研究，展现了这些方法的核心思想和应用场景并分析了这些工作的优势与不足.分析表明，现有基于机器学习的智能路由算法研究主要针对算法原理,这些路由算法距离真实环境下部署仍然很遥远.因此接下来分析了不同的真实场景智能路由算法训练和部署方案并提出了2种合理的训练部署框架以使得智能路由算法能够低成本、高可靠性地在真实场景被部署.最后分析了基于机器学习的智能路由算法未来发展中所面临的机遇与挑战并给出了未来的研究方向.

关键词: 机器学习, 数据驱动路由算法, 深度学习, 强化学习, 服务质量

Abstract: The rapid development of the Internet accesses many new applications including **real time multi-media service【实时多媒体流服务】**, **remote cloud service【远程云服务】**, etc. These applications require various types of service quality, which is a significant challenge towards current best effort **routing algorithms【路由算法】**. Since the recent huge success in applying machine learning in game, **computer vision【计算机视觉】** and **natural language processing【自然语言处理】**, many people tries to design “smart” routing algorithms based on machine learning methods. In contrary with（相比于） traditional model-based, decentralized（分布式） routing algorithms (e.g.OSPF), machine learning based routing algorithms are usually **data-driven【数据驱动】**, which can adapt to dynamically（动态） changing network environments and accommodate（适应） different service quality requirements. Data-driven routing algorithms based on machine learning approach have shown great potential in becoming an important part of the next generation network. However, researches on artificial intelligent routing are still on a very beginning stage. In this paper we firstly introduce current researches on data-driven routing algorithms based on machine learning approach, showing the main ideas, application scenarios（场景） and pros and cons（正反两方面/优点和缺点） of these different works. Our analysis shows that current researches are mainly for the principle of machine learning based routing algorithms but still far from deployment in real scenarios. So we then analyze different training and deploying（部署） methods for machine learning based routing algorithms in real scenarios and propose two reasonable approaches to train and deploy such routing algorithms with low overhead（低成本） and high reliability. Finally, we discuss the opportunities and challenges and show several potential research directions for machine learning based routing algorithms in the future.

Key words: machine learning, data driven routing algorithm【数据驱动路由算法】, deep learning, reinforcement learning【强化学习】, quality of service (QoS)【服务质量】

2.编码技术改进大规模分布式机器学习性能综述

Coding-Based Performance Improvement of Distributed Machine Learning in Large-Scale Clusters

摘要： 由于分布式计算系统能为大数据分析提供大规模的计算能力，近年来受到了人们的广泛关注.在分布式计算系统中，存在某些计算节点由于各种因素的影响，计算速度会以某种随机的方式变慢，从而使运行在集群上的机器学习算法执行时间增加，这种节点叫作掉队节点(straggler).介绍了基于编码技术解决这些问题和改进大规模机器学习集群性能的研究进展.首先介绍编码技术和大规模机器学习集群的相关背景;其次将相关研究按照应用场景分成了应用于矩阵乘法、梯度计算、数据洗牌和一些其他应用，并分别进行了介绍分析;最后总结讨论了相关编码技术存在的困难并对未来的研究趋势进行了展望.

关键词: 编码技术, 机器学习, 分布式计算, 掉队节点容忍, 性能优化

Abstract: With the growth of models and **data sets【数据集】**, running large-scale machine learning algorithms in **distributed clusters【分布式集群】** has become a common method. This method divides the whole machine learning algorithm and training data into several tasks and each task runs on different worker nodes. Then, the results of all tasks are combined by master node to get the results of the whole algorithm. When there are a large number of nodes in distributed cluster, some worker nodes, called **straggler【掉队节点】**, will inevitably（不可避免地） slow down than other nodes due to resource competition and other reasons, which makes the task time of running on this node significantly higher than that of other nodes. Compared with running **replica task【复制任务】** on multiple nodes, coded computing shows an impact of efficient utilization of computation and **storage redundancy【存储冗余】** to alleviate（减轻） the effect of stragglers and communication **bottlenecks【瓶颈】** in large-scale machine learning cluster.This paper introduces the research progress of solving the straggler issues and improving the performance of large-scale machine learning cluster based on coding technology. Firstly, we introduce the background of coding technology and large-scale **machine learning cluster【机器学习集群】**. Secondly, we divide the related research into several categories according to application scenarios: **matrix multiplication【矩阵乘法】**, **gradient computing【梯度算法】**, **data shuffling【数据洗牌】** and some other applications. Finally, we summarize the difficulties of applying coding technology in large-scale machine learning cluster and discuss the future research trends about it.

3.贝叶斯机器学习前沿进展综述

Recent Advances in Bayesian Machine Learning

摘要： 随着大数据的快速发展，以概率统计为基础的机器学习在近年来受到工业界和学术界的极大关注，并在视觉、语音、自然语言、生物等领域获得很多重要的成功应用，其中贝叶斯方法在过去20多年也得到了快速发展,成为非常重要的一类机器学习方法.总结了贝叶斯方法在机器学习中的最新进展，具体内容包括贝叶斯机器学习的基础理论与方法、非参数贝叶斯方法及常用的推理方法、正则化贝叶斯方法等. 最后，还针对大规模贝叶斯学习问题进行了简要的介绍和展望，对其发展趋势作了总结和展望.

关键词: 贝叶斯机器学习, 非参数方法, 正则化方法, 大数据学习, 大数据贝叶斯学习

Abstract: With the fast growth of big data, statistical machine learning has attracted tremendous attention from both industry and academia（学术界）, with many successful applications in vision, speech, natural language, and biology. In particular, the last decades have seen the fast development of **Bayesian【贝叶斯】** machine learning, which is now representing a very important class of techniques. In this article, we provide an overview of the recent advances in Bayesian machine learning, including the basics of Bayesian machine learning theory and methods, **nonparametric Bayesian methods【非百叶窗学方法】** and **inference algorithms【推理算法】**, and regularized Bayesian inference. Finally, we also highlight the challenges and recent progress on large-scale Bayesian learning for big data, and discuss on some future directions.

Key words: Bayesian machine learning【贝叶斯机器学习】, nonparametric methods【非参数方法】, regularized methods【正则化方法】, learning with big data, big Bayesian learning【大数据贝叶斯学习】

三、大数据相关

1.基于高性能密码实现的大数据安全方案

A Big Data Security Scheme Based on High-Performance Cryptography Implementation

摘要： 目前信息技术发展的趋势是以大数据计算为基础的人工智能技术.云计算、雾计算、边缘计算等计算模式下的大数据处理技术，在给经济发展带来巨大推动力的同时，也面临着巨大的安全风险.密码技术是解决大数据安全的核心技术.大数据的机密性、认证性及隐私保护问题需要解决海量数据的高速加解密问题；高并发的大规模用户认证问题；大数据的隐私保护及密态计算问题等，这些问题的解决，需要底层密码算法的快速实现.针对大数据安全应用的逻辑架构，对底层的国产密码标准算法SM4-XTS，SM2以及大整数模幂运算，分别给出快速计算的算法，并在基于Xilinx公司的KC705开发板上进行了验证，并给出实验数据.实验表明：该工作具有一定的先进性：1)SM4-XTS模式的实现填补了国内该方向的空白；2)SM2签名具有较高性能，领先于国内同类产品；3)大整数的模幂运算应用于同态密码的产品化，填补了国内该产品的空白.

关键词: SM4-XTS, SM2, 大整数模幂, 密码算法快速实现, 大数据

Abstract: At present, the trend of information technology development is the artificial intelligence technology based on big data computing. Although it has made enormous（极大的） contribution in the economic development, big data processing technology which includes cloud computing, **fog computing【雾计算】**, **edge computing【边缘计算】** and other computing modes also brings a great risk of data security. **Cryptographic【加密】** technology is the kernel（核心） of the big data security. Confidentiality, authentication and privacy protection of big data need to solve the following three security problems: firstly, high-speed encryption and decryption of massive data; secondly, the authentication problem of high concurrency（高并发性） and large scale user; thirdly, privacy protection in data mining. The solution of these problems requires the fast implementation of the underlying cryptographic algorithm. Aiming at the logic architecture of big data security application, this paper gives a fast calculation algorithm for the cryptographic standard algorithm SM4-XTS, SM2 and **modular exponentiation【模幂运算】** of large integers（大数）. It is verified（验证） on the KC705 development board based on Xilinx company, the results of experiment show that our work has certain advancement: 1) The implementation of SM4-XTS fills the blank of this direction in China. 2) SM2 signature has high performance, leading domestic similar products. 3) Modular exponentiation is applied to the productization of **homomorphism cryptography【同态加密】**, and its performance is ahead of other similar products.

Key words: SM4-XTS, SM2, modular exponentiation【大整数模幂】, high-speed implementation of cryptographic algorithm【密码算法快速实现】, big data

2.知识图谱研究综述及其在医疗领域的应用。

Research Review of Knowledge Graph and Its Application in Medical Domain

摘要： 随着医疗大数据时代的到来，知识互联受到了广泛的关注.如何从海量的数据中提取有用的医学知识，是医疗大数据分析的关键.知识图谱技术提供了一种从海量文本和图像中抽取结构化知识的手段，知识图谱与大数据技术、深度学习技术相结合，正在成为推动人工智能发展的核心驱动力.知识图谱技术在医疗领域拥有广阔的应用前景，该技术在医疗领域的应用研究将会在解决优质医疗资源供给不足和医疗服务需求持续增加的矛盾中产生重要的作用.目前，针对医学知识图谱的研究还处于探索阶段，现有知识图谱技术在医疗领域普遍存在效率低、限制多、拓展性差等问题.首先针对医疗领域大数据专业性强、结构复杂等特点，对医学知识图谱架构和构建技术进行了全面剖析;其次，分别针对医学知识图谱中知识

表示、知识抽取、知识融合和知识推理这4个模块的关键技术和研究进展进行综述，并对这些技术进行实验分析与比较.此外，介绍了医学知识图谱在临床决策支持、医疗智能语义检索、医疗问答等医疗服务中的应用现状.最后对当前研究存在的问题与挑战进行了讨论和分析，并对其发展前景进行了展望.

关键词: 知识图谱, 智慧医疗, 大数据, 知识融合, 自然语言处理

Abstract: With the advent of the **medical big data era【医疗大数据时代】**, **knowledge interconnection【知识互联】** has received extensive（广泛的） attention. How to extract useful medical knowledge from massive data is the key for medical big data analysis. **Knowledge graph technology【知识图谱技术】** provides a means to extract structured knowledge from massive texts and images.The combination of knowledge graph, big data technology and deep learning technology is becoming the core driving force for the development of artificial intelligence. The knowledge graph technology has a broad application prospect in the medical domain. The application of knowledge graph technology in the medical domain will play an important role in solving the contradiction（矛盾） between the supply of high-quality medical resources and the continuous increase of demand for medical services.At present, the research on medical knowledge graph is still in the exploratory stage（探索阶段）. The existing knowledge graph technology generally has several problems such as low efficiency, multiple restrictions（多限制） and poor expansion（拓展性差） in the medical domain. This paper firstly analyzes the medical **knowledge graph architecture【知识图谱架构】** and **construction technology【构建技术】** for the strong professionalism and complex structure of big data in the medical domain. Secondly, the key technologies and research progress of the three modules of **knowledge extraction【知识提取】**, **knowledge expression【知识融合】**, **knowledge fusion【知识抽取】** and **knowledge reasoning【知识推理】** in medical knowledge map are summarized. In addition, the application status of medical knowledge maps in **clinical decision support【临床决策支持】**, **medical intelligence semantic retrieval【医学智能语义检索】**, medical question answering system and other medical services are introduced. Finally, the existing problems and challenges of current research are discussed and analyzed, and its development is prospected.

Key words: knowledge graph【知识图谱】, medical wisdom【智慧医疗】, big data, knowledge fusion【知识融合】, natural language processing【自然语言处理】

3.基于MOOC数据的学习行为分析与预测

Learning Behavior Analysis and Prediction Based on MOOC Data

摘要： 随着近2年慕课(massive open online course, MOOC)的兴起，教育大数据分析正成为一个新兴的研究方向.2013年秋，北京大学在Coursera上开设了6门慕课.通过分析挖掘约8万多人次参与这6门课的海量学习行为数据，力图展现慕课学习活动多个侧面的风貌.同时，首次针对中文慕课中学习行为的特点，将学习者分类，以更加深入地考察学习行为与学习效果之间的关系.在此基础上，通过选择学习者的若干典型行为特征，对他们最后的学习成果进行预测的工作也尚属首次.数据表明：基于学习行为的特征分析能有效地判别一个学习者能否成功完成学习任务获得通过证书，并能找出潜在的认真学习者，这为今后更加精准的慕课教学测评提供了一种依据.

关键词: 慕课, 学习者类型, 学习行为, 数据分析, 成绩预测

Abstract: With the booming of MOOC (massive open online course) in the past two years, educational data analysis has become a promising research field where the quality of teaching and learning can be and is being quantified to improve the educational effectiveness and even to promote the modern higher education. In the autumn of 2013, Peking University released its first six courses on the Coursera platform. Through **mining and analyzing【挖掘分析】** the massive data of learning behavior of over 80000 participants from the courses, this paper endeavors（努力） to manifest more than one side of learning activity in MOOC. Meanwhile, according to the characteristic of learning behavior in Chinese MOOC, learners are classified into several groups and then the relationship between their learning behavior and performance is thoroughly

studied. Based on the above work, we find out that learners□ performance, regarding whether he□she could get certificated eventually, can be predicted by looking into several features of their learning behavior. Experiment results indicate that these features can be trained to effectively estimate whether a learner is probably to complete the course successfully. Besides, this method has the potential to partially evaluate the quality of both teaching and learning in practice.

Key words: massive open online course (MOOC), engagement style【学习者类型】, learning behavior, data analysis, performance prediction【成绩预测】

4.知识图谱构建技术综述

Knowledge Graph Construction Techniques

摘要： 谷歌知识图谱技术近年来引起了广泛关注，由于公开披露的技术资料较少，使人一时难以看清该技术的内涵和价值.从知识图谱的定义和技术架构出发，对构建知识图谱涉及的关键技术进行了自底向上的全面解析.1)对知识图谱的定义和内涵进行了说明，并给出了构建知识图谱的技术框架，按照输入的知识素材的抽象程度将其划分为3个层次：信息抽取层、知识融合层和知识加工层;2)分别对每个层次涉及的关键技术的研究现状进行分类说明，逐步揭示知识图谱技术的奥秘，及其与相关学科领域的关系;3)对知识图谱构建技术当前面临的重大挑战和关键问题进行了总结.

关键词: 知识图谱, 语义网, 信息检索, 语义搜索引擎, 自然语言处理

Abstract: Google's **knowledge graph technology【知识图谱技术】** has drawn a lot of research attentions in recent years. However, due to the limited public disclosure（披露） of technical details, people find it difficult to understand the connotation（内涵） and value of this technology. In this paper, we introduce the key techniques involved in the construction（构建） of knowledge graph in a bottom-up way（自下而上的）, starting from a clearly defined concept and a technical architecture of the knowledge graph. Firstly, we describe in detail the definition and connotation of the knowledge graph, and then we propose the technical framework for knowledge graph construction, in which the construction process is divided into three levels according to the abstract level of the input knowledge materials, including the **information extraction layer【信息提取层】**, **the knowledge integration layer【知识集成层】**, and **the knowledge processing layer【知识处理层】**, respectively. Secondly, the research status of the key technologies for each level are surveyed comprehensively and also investigated critically for the purposes of gradually revealing the mysteries（奥秘） of the knowledge graph technology, the state-of-the-art progress, and its relationship with related disciplines. Finally, five major research challenges in this area are summarized, and the corresponding key research issues are highlighted.

Key words: knowledge graph【知识图谱】, semantic Web【语义网】, information retrieval【信息检索】, semantic search engine【语义搜索引擎】, natural language processing【自然语言处理】

5.一种面向大规模序列数据的交互特征并行挖掘算法

A Parallel Algorithm for Mining Interactive Features from Large Scale Sequences

摘要： 序列是一种重要的数据类型，在诸多应用领域广泛存在.基于序列的特征选择具有广阔的现实应用场景.交互特征是指一组整体具有显著强于单独个体与目标相关性的特征集合.从大规模序列中挖掘交互特征面临着位点的"组合爆炸"问题，计算挑战性极大.针对该问题，以生物领域高通量测序数据为背景，提出了一种新的基于并行处理和演化计算的高阶交互特征挖掘算法.位点数是制约交互作用挖掘效率的根本因素.摈弃了现有方法基于序列分块的并行策略，采用基于位点分块的并行思想，具有天然的效率优势.进一步，提出了极大等位公共子序列(maximal allelic common subsequence, MACS)的概念并设计了基于MACS的特征区域划分策略.该策略能将交互特征的查找范围缩小至许多"碎片"空间，并保证不同"碎片"间不存在交互特征，避免计算耦合引起的高额通信代价.利用基于置换搜索的并行蚁群算法，执

行交互特征选择.大量真实数据集和合成数据集上的实验结果，证实提出的PACOIFS算法在有效性和效率上优于同类其他算法.

关键词: 交互特征, 数据挖掘, 大规模序列, 蚁群算法, 并行计算, 极大等位公共子序列

Abstract: Sequence(序列） is an important type of data which is widely existing in various domains, and thus **feature selection【特征选择】** from sequence data is of practical significance in extensive applications. **Interactive features【交互式功能】** refer to a set of features, each of which is weakly correlated（关联） with the target, but the whole of which is strongly correlated with the target. It is of great challenge to **mine【挖掘】** interactive features from large scale sequence data for the **combinatorial explosion problem of loci【面临着位点的"组合爆炸"问题】**. To address the problem,（解决这个问题） against the background of **high-throughput sequencing in biology【生物领域高通量测序数据】**, **a parallel evolutionary algorithm【并行进化算法】** for **high-order interactive features mining【高阶交互功能挖掘】** is proposed in this paper. Instead of **sequence-block based parallel strategy【基于序列块的并行策略】**, the work is inspired by loci-based idea since the number of loci is the fundamental factor that restricts the efficiency. Further, we propose the conception of **maximal allelic common subsequence (MACS)【极大等位公共子序列】** and **MACS based strategy for feature region partition【基于MACS的特征区域划分策略】**. According to the strategy, the search range of interactive features is narrowed to many **fragged spaces【碎片空间】** and interactions are guaranteed not to exist among different fragments（片段）. Finally, **a parallel ant algorithm based on substitution search【基于置换搜索的并行蚁群算法】** is developed to conduct interactive feature selection. Extensive experiments on real and **synthetic datasets【合成数据】** show that the efficiency and effectiveness（有效性和效率） of the proposed PACOIFS algorithm is superior to that of competitive algorithms（同类其他算法）.

Key words: interactive features【交互特征】, data mining, large scale sequence【大规模序列】, ant colony algorithm【蚁群算法】, parallel computation【并行计算】, maximal allelic common subsequence (MACS)【极大等位公共子序列】