基于可靠性增强的社交网络链接预测算法

何玉林[1,2]，孙洪涛[1,2]，秦红莲[1]，黄舒影[3,4]，崔来中[2]，黄哲学[1,2]
1. 人工智能与数字经济广东省实验室（深圳），广东 深圳 518107；
2. 深圳大学计算机与软件学院，广东 深圳 518060；
3. 腾讯科技（深圳）有限公司，广东 深圳 518057；
4. 卡耐基梅隆大学电气与计算机工程学院，宾夕法尼亚州 匹兹堡 15213

摘要

社交网络链接预测（link prediction，LP）是复杂网络挖掘领域的重要研究方向之一，旨在检测网络节点之间的潜在联系。现有的社交网络链接预测算法包括基于静态相似性的指标、基于动态学习的预测器和基于内容的方法。尽管已有的实验结果表明，这些方法在特定的应用场景中可以表现出良好的性能，但它们仍然存在预测不可靠和探测能力弱的本质缺陷，且目前并没有得到有效解决。为此，提出了一种基于可靠性增强的链接预测（reliability enhancement-based link prediction，RE-LP）算法来弥补现有链接预测算法的上述缺陷。RE-LP算法有3个主要组成部分，即不存在链接的划分、可靠预测器的构建和连接概率的计算。该算法将可观察的不存在链接划分为高可靠的不存在（highly-reliable non-existing，HRNE）链接和可能被观察到的存在（possibly-observed existing，POE）链接，并以迭代的方式不断从POE链接中识别出HRNE链接，进而使用性能良好的贝叶斯链接预测器计算POE链接的连接概率，以达到准确可靠预测未知链接的目的。实验验证了RE-LP算法的可行性、合理性和有效性。结果表明，在选用的数据集上，RE-LP算法获得了高出其他5种先进LP算法21.7%~36.1%的预测精度，能够以较高的可信度探测出社交网络中的潜在链接。

关键词

社交网络；链接预测；贝叶斯分类器；可靠性增强；复杂网络分析

中图分类号：TP391　　文献标志码：A　　doi：10.11959/j.issn.2096-0271.2026023

Reliability enhancement-based link prediction algorithm for social network

He Yulin[1,2], Sun Hongtao[1,2], Qin Honglian[1], Huang Shuying[3,4], Cui Laizhong[2], Huang Zhexue[1,2]
1. Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen 518107, China
2. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
3. Tencent Technology (Shenzhen) Co. , Ltd. , Shenzhen 518057, China
4. Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh 15213, USA

Abstract

The link prediction (LP) for social network is an important research direction of complex network mining, which tries to detect the potential relationship between network nodes. The popular LP algorithms for social networks include the

static similarity-based indicators, dynamic learning-based predictors and content-based methods. Although the experimental results have reported the good performances in specific application scenarios of these algorithms, some essential defects such as prediction unreliability and detection incapability still exist and cannot be effectively solved up to now. Therefore, a reliability enhancement-based link prediction (RE-LP) algorithm is proposed to make up the above-mentioned shortcomings of existing LP algorithms. There are three main components in RE-LP algorithm, i.e., non-existing link partition, reliable predictor construction, and connection probability calculation. In RE-LP algorithm, the observable non-existing links are sophisticatedly partitioned into highly-reliable non-existing (HRNE) links and possibly-observed existing (POE) links. It then continuously identifies HRNE links from the POE links in an iterative manner, and uses a well-performed Bayesian classifier-based link predictor to calculate the connection probabilities of the POE links, in order to achieve the goal of accurately and reliably predicting unknown links. Through a series of verification experiments, the advantages of RE-LP algorithm in terms of feasibility, rationality and effectiveness are fully demonstrated. The experimental results demonstrate that RE-LP algorithm can obtain the 21.7%~36.1% higher prediction AUC than 5 advanced LP algorithms and meanwhile is able to detect the potential links with high credibility.

Key words
social network, link prediction, Bayesian classifier, reliability enhancement, complex network analytics

0 引言

社交网络是一种模拟社区中特定成员之间互动关系的通用模型。一般来说，社交网络中的一个节点代表一个成员，而一条边代表相应的成员之间某种形式的联系。由于社会关系的不断变化，社交网络中的节点之间的链接会随着时间的推移而动态变化，新增链接通常标志着现实社会结构中出现了新的互动关系。提前探测节点之间的链接关系对了解社交网络的进化机制至关重要。

链接预测（link prediction,LP）是社交网络领域的重要研究内容，目标是根据网络布局和节点相关统计信息，计算节点间建立链接的概率。它既包括对当前未知链接的预测，也包括对未来可能出现的链接的预测。随着社交网络研究的快速发展，链接预测因其重要的理论探索意义和实际应用价值而备受学术界和工业界关注，例如犯罪网络建设、传染病传播预测、蛋白质检测和个性化推荐。根据采用的技术方法不同，现有的链接预测方法可以分为以下3类。①基于相似性的方法，这些方法试图通过计算节点之间的相似性来确定链接存在的可能性，其核心思想是相似度高的节点对之间存在链接的可能性更大[1]。②基于学习的方法，这些方法主要包括基于分类的方法和最大似然方法。前者将链接预测视为一个两分类问题，其核心任务是根据基于节点的网络布局结构和社会理论提取一组合适的特征，并进一步构建有效的分类器；后者旨在识别最有可能的网络生成模型，然后根据识别出的模型估计未知链接的可能性。③基于内容的方法，这些方法通常不涉及对网络布局的分析，而是专注于检查用户的属性和行为，收集良好的非拓扑特征可以极大地提高链接预测算法的性能。

尽管现有研究中提出的预测方法在选用的测试数据集上获得了良好的实验结果，但它们仍然存在以下两个弱点。第一，预测可靠性较差。由于对网络结构的过度关注，对于不同类型的社交网络，相似性指

标的效果差异很大；另外，由于社交网络中存在边和缺失边的不均衡性，基于学习的链接预测方法的性能在很大程度上受限于数据平衡化的质量；基于内容的方法主要关注的是社交网络的非拓扑结构信息，这就使得预测性能更依赖非结构化的个体主观信息，这些都将导致现有链接预测方法的预测不稳定性。第二，探测能力较弱。现有的链接预测方法通常对不存在边的定义过于简单，事实上这些不存在边包括3类情况：确定不存在的边、实际存在但未被探测到的边、未来可能被观察到的边。这种简单的不存在边定义方式相当于在链接预测算法训练过程中引入了噪声数据，从而弱化链接预测算法对未知边的探测能力。

为了有效地应对上述边预测研究面临的挑战，进一步提升链接预测算法的稳定性和探测能力，本文提出了一种基于可靠性增强的链接预测（reliability enhancement-based link prediction，RE-LP）算法，该算法可以针对不同类型的社交网络进行更稳定、更准确的链接预测。RE-LP算法通过使用不同类型的负样本迭代训练增强型贝叶斯分类器来计算未知链接存在的可能性，进而处理链接预测任务，其主要的组成部分概括如下：①根据链接分布一致性将未知链接分为高可靠的不存在（highly-reliable non-existing，HRNE）链接和可能被观察到的存在（possibly-observed existing，POE）链接；②以迭代的方式训练基于已存在链接和HRNE链接组合的增强型贝叶斯分类器；③用稳定的贝叶斯分类器计算POE链接的连接概率，区分POE链接中的HRNE链接。

本文在8个代表性社交网络数据集上进行了一系列实验验证，其结果证实了RE-LP算法的可行性、合理性和有效性。实验结果表明：①RE-LP算法可以通过从POE链接中连续识别HRNE链接来达到收敛状态；②RE-LP算法能够检测未来可能出现的链接，具有较强的可靠性和稳定性；③与3种经典的基于静态相似性的方法和两种具有代表性的基于动态学习的方法相比，RE-LP算法具有更高的链接预测精度。

本文结构如下：第1节回顾了相关工作；第2节介绍了社交网络的相关概念、数据集划分方法以及链接预测问题的评估指标；第3节展示了RE-LP算法的详细构建过程；第4节给出了实验结果和分析；第5节总结了本文的主要研究和未来工作展望。

1 相关工作

本节将简要介绍3类链接预测方法，即基于相似性的方法、基于学习的方法和基于内容的方法，以及对应的代表性工作。

1.1 基于相似性的方法

基于相似性的方法凭借其在计算复杂度方面的优势，成为预测链接存在性最简单、最直接的方法。根据相似性计算中考虑的网络结构信息的不同，相似性指标主要包含局部、全局和准局部3种类型。

局部相似性指标仅考虑与相邻节点相关的信息。此类指标计算简单，适用于大规模网络。共同邻居（common neighbor，CN）指数[2]是最具代表性的指标之一，它将两个节点的相似度定义为它们共同邻居的数量。Jaccard系数[3]假设拥有更高共同邻居比例的节点对更相似，可被看作对CN指数的标准化。其他类似的指标包括Salton指数[4]、Sørensen指数[5]、枢

纽提升（hub promoted，HP）指数[6]、枢纽抑制（hub depressed，HD）指数[6]和LHN指数[7]等。AA指数[8]和资源分配（resource allocation，RA）指数[9]的计算式相似，都抑制了高度共同邻居的贡献，它们之间的主要区别在于RA指数的惩罚更重。此外，与其他指标不同，这两个指标不仅考虑了直接邻居，还考虑了邻居的邻居。

全局相似性指标利用网络的完整拓扑信息来计算相似性，因此具备较高的预测精度。但其计算复杂度偏高，难以扩展到较大规模网络的链接预测中。著名的Katz指数[10]将两个节点之间的所有路径相加，并对较短的路径赋予较高的权重。SimRank指数[11]假设，节点的相似性可由其连接节点的相似程度决定。重启随机游走（restart random walk，RRW）指数[12]通过随机移动迭代探索网络的整体结构。与其他随机游走指数不同，RRW指数在每次游走后都有一定的概率返回起点。

准局部相似性指标考虑了更多的拓扑信息，复杂度低于全局相似性指标，是一种兼顾预测准确性和计算复杂度的指标。局部路径（local path，LoPa）指数[13-14]是最广泛使用的准局部相似性指标之一，它同时考虑了两个节点间长度为2和3的路径信息。局部随机游走（local random walk，LRW）指数[15]和叠加随机游走（superposed random walk，SRW）指数[15]是基于随机游走策略设计的。其他准局部相似性指标还包括命中时间（hitting time，HT）指数[16]和交换时间（commuting time，CT）指数[16]。

1.2 基于学习的方法

基于学习的方法一般包括基于机器学习的分类方法和基于统计推断的最大似然方法两类。

基于机器学习的分类方法已成为当前链接预测研究的一个重要部分。从算法实现角度，链接预测问题可被形式化为一个典型的二分类学习任务。网络中连通节点对是正样例，非连通节点对是负样例。对于机器学习算法来说，选择合适的特征集进行分类器训练至关重要。Hasan等[17]从3个维度（包括接近度、聚合度和拓扑结构）选择了对共同作者社交网络中的链接预测性能至关重要的特征来训练支持向量机、决策树、多层感知机等。De等[18]专注于加权网络，并使用加权结构相似性指数来构建训练特征。Bütün等[19]考虑了链接方向对预测结果的影响，进而提出一种基于近邻的拓展链接预测算法。除了常用的监督学习方法[20]外，半监督学习[21]、深度学习[22]、集成学习[23]等方法逐渐成为链接预测研究的重点。在集成学习方面，罗凯靖等[24]提出了基于Bootstrap样本划分的大数据模型与分布式集成学习方法，为大规模社交网络链接预测提供了新途径。

基于统计推断的最大似然方法假设网络结构具有一定的组织原则，某些链接能够反映网络内在的层次结构，从而推断未知链接存在的可能性。Clauset等[25]提出了一种从网络数据中推断层次结构的通用技术，并证明了其在重现网络拓扑特性和预测缺失链接方面的有效性。Guimerà等[26]使用随机块模型进行网络重构，识别缺失和虚假交互。实验结果表明，基于统计推断的最大似然方法对具有明显层次结构的网络（如攻击网络和草原食物链网络）具有良好的预测效果。

1.3 基于内容的方法

基于内容的方法专注于检查用户的属

性和行为。

在用户属性方面，相关研究主要考虑用户兴趣、地理信息和文本分析等因素。Pizzato等[27]的研究表明，相似的爱好可以提高推荐的成功率。Shi等[28]提出用户情绪及其社会关系对链接预测有正向影响，并设计了一种结合节点结构和用户情绪属性的算法，有效地提高了预测性能。

在用户行为方面，研究主要集中在两个方面：个体用户行为和社会行为。Shahmohammadi等[29]将不同的用户活动（包括点赞、评论、发布和分享）加权到现有网络中，并提出了3种基于协同策略的链接预测算法，即协同随机游走算法、多层协同随机游走算法、协同关联规则算法。然而，这类研究偏重微观互动，缺乏宏观传播建模。王续澎等[30]提出的5W传播模型框架从主体、内容、渠道、受众和效果5个维度解析信息扩散过程，提供了新的链接预测特征建模思路。

除了上述方法外，基于网络嵌入的方法，尤其是基于神经网络的方法近年来越来越受到人们的关注，相关研究可以参阅近年的综述[31]。

2 预备知识

2.1 问题陈述

社交网络是由不同实体及其关系组成的结构，它通常可以被视为一个无向图$G=(V,E)$，其中V和E分别是节点集和链接集。给定某一时刻的社交网络，链接预测问题的预测对象可被划分为当前网络中缺失的链接和未来可能出现的新链接。其中，缺失的链接往往是由信息收集不完整导致的，而未来的新链接将会伴随着社交网络的演化而出现。

图G中链接数量最多为$\frac{|V|\cdot(|V|-1)}{2}$，其中$|V|$是节点集中包含的元素个数。将完整的可能链接集表示为U，那么当前不存在的链接集可以表示为$N=U-E$。如前文所述，链接预测问题的任务是在集合N中找出当前缺失的链接或未来可能出现在集合中的链接。

设计可用的链接预测算法来确定两个节点x和y的连接概率。为每对节点$(x,y)\in N$分配一个概率值S_{xy}，然后降序排列这些节点对的概率值，具有最高值的一对节点最有可能存在或建立链接。

2.2 数据划分

为了验证链接预测算法的性能，将所有样本按比例8∶1∶1随机划分为训练集、验证集和测试集。此划分在算法开始前一次性完成，并且对测试集不做任何处理，以确保评估指标的无偏性。

将所有已存在的链接定义为正样本集E，将所有不存在的链接定义为负样本集N。首先分别从正样本和负样本中各随机选择出80%的链接，共同构成训练集U^{P}，其中包含正样本训练子集E^{P}和负样本训练子集N^{P}，即$U^{\mathrm{P}}=E^{\mathrm{P}}\cup N^{\mathrm{P}}$。然后从剩余20%的正样本与负样本中各随机选出50%（即总样本数的10%）的链接共同组成验证集U^{V}，余下未被选择的正样本与负样本将组成测试集U^{T}。链接预测算法将根据来自训练集U^{P}的已知信息预测未知链接的可能性，并使用测试集U^{T}来验证链接预测算法的有效性。

图1展示了上述社交网络链接样本的划分过程，其中网络包含5个节点、7条现有链接和3条不存在链接(1,3)、(1,5)和(3,4)。为了衡量链接预测算法的性能，随机选择现有链接(1,4)和(2,3)作为测试集，

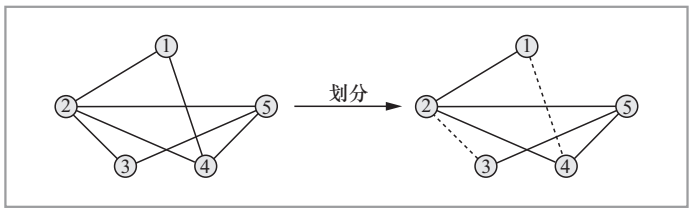


图1 社交网络数据划分示意图

如**图1**中虚线所示；其他现有链接集和不存在的链接集组成训练集，并分别标记为正样本和负样本。常见的数据集划分方法包括随机抽样法、K折叠交叉验证法、随机游走抽样法等。本文采用随机游走抽样法划分数据集。

2.3 评估指标

在解决链接预测问题时，经常使用的评估指标是曲线下面积（area under curve，AUC）。在计算AUC时，分别从测试集的正样本和负样本中各随机选择一条链接。然后比较所选链接的预测概率值。如果前一条链接的值更高，则将在AUC分子上加1分；如果值相等，则加0.5分；否则不加任何分，其计算式如式（1）所示。

$$\mathrm{AUC}=\frac{n_1+0.5n_2}{n} \tag{1}$$

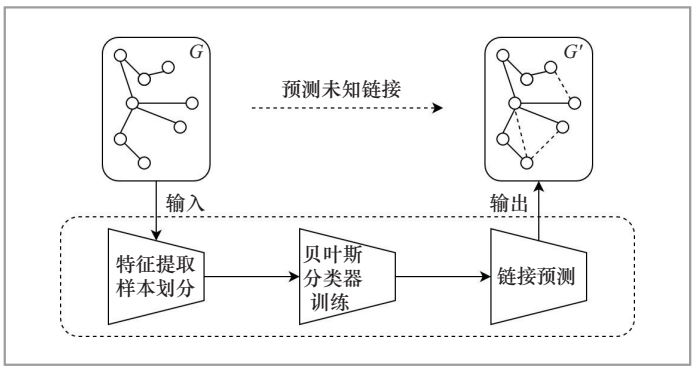


图2 RE-LP算法总体架构

其中，n是比较的总数，n_1是测试集中缺失的链接具有较高分数的次数，n_2是所选两条链接的概率值相等的次数。AUC越接近1，表明链接预测算法的排序性能越优；若AUC为0.5，则说明算法无判别能力。

3 RE-LP算法

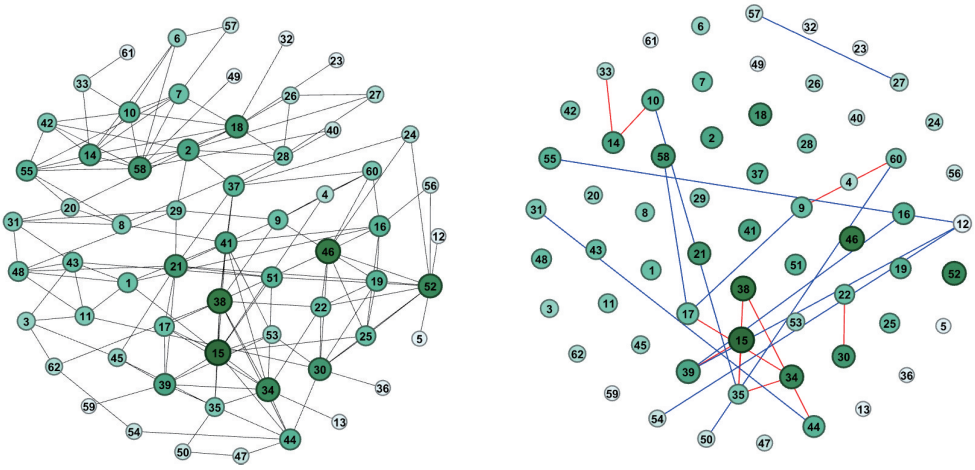
本节给出了RE-LP算法，详细介绍了其设计过程及具体执行步骤。**图2**展示了RE-LP算法的总体架构，总体功能为预测给定图中的未知链接，具体功能模块包括对输入图的样本特征提取和样本划分、训练贝叶斯分类器以及使用训练好的贝叶斯分类器预测链接并输出。

3.1 当前链接预测算法的缺陷

如前所述，当前的链接预测算法仍然存在预测不可靠、探测能力不强等本质缺陷。有效的LP算法应该正确预测链接的存在状态，并为现有链接赋予更高的概率值。

• 对于预测的不可靠性，从dolphins数据集中随机选择了10条存在链接和10条不存在链接，如**图3（a）**所示。**表1**显示了AA指数和DeepWalk[32]算法分配的预测概率值，其中概率值较高的10条链接被算法输出为存在，较低的10条链接被输出为不存在。显然，可以发现二者错误地将较高的概率值赋予了一些标记为红色的不存在的链接。

• 至于探测能力，随机丢弃了aves-sparrow-social数据集上的8条现有链接，如**图4（b）**黄色线所示，并基于剩余数据测试了3种LP算法（RE-LP、node2vec[33]和DeepWalk）。从**表2**可以看出，node2vec



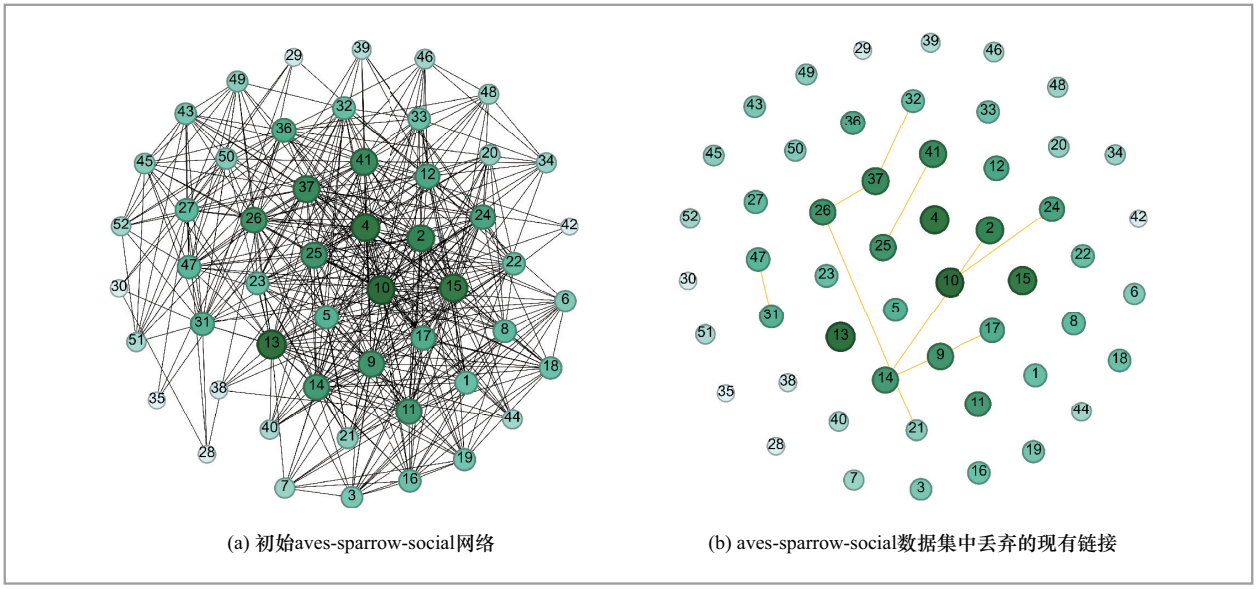
(a) 初始dolphins社交网络　　(b) dolphins数据集中的选定链接

（红色：现有链接；蓝色：不存在链接）

图3　经典LP算法的预测不可靠性示意图

表1　AA指数与DeepWalk算法的链接预测概率值

链接信息			AA指数		DeepWalk算法	
编号	链接	链接标签	预测概率值	预测结果	预测概率值	预测结果
1	(10，14)	1	1.000000	Right	0.938398	Right
2	(15，39)	1	0.598626	Right	0.187481	Right
3	(17，34)	1	0.487537	Right	0.387452	Right
4	(22，30)	1	0.366871	Right	0.629521	Right
5	(9，60)	1	0.356672	Right	0.621107	Right
6	(9，17)	0	**0.297487**	**Wrong**	**0.162218**	**Wrong**
7	(15，35)	1	0.228773	Right	0.119183	Right
8	(35，38)	1	0.224849	Right	0.007599	Right
9	(38，44)	1	0.224849	Right	0.496571	Right
10	(34，35)	1	0.220210	Right	0.009210	Right
11	(14，33)	1	0.138098	Right	0.533920	Right
12	(10，35)	0	0.000000	Right	0.000817	Right
13	(12，39)	1	**0.000000**	**Wrong**	0.000458	Right
14	(12，54)	0	0.000000	Right	0.000362	Right
15	(12，55)	0	0.000000	Right	0.000664	Right
16	(16，39)	0	0.000000	Right	**0.017564**	**Wrong**
17	(17，58)	0	0.000000	Right	0.005446	Right
18	(27，57)	0	0.000000	Right	0.008427	Right
19	(31，44)	0	0.000000	Right	0.003888	Right
20	(50，60)	0	0.000000	Right	0.000174	Right



(a) 初始aves-sparrow-social网络　　(b) aves-sparrow-social数据集中丢弃的现有链接

图4　经典LP算法的探测低效性示意图

表2　3种LP算法的链接预测结果

丢弃链接	预测标签		
	RE-LP算法	node2vec算法	DeepWalk算法
(26, 37)	1	1	0
(25, 41)	1	0	1
(2, 14)	1	1	0
(10, 24)	1	1	0
(14, 17)	1	1	0
(31, 47)	1	0	1
(32, 37)	1	0	1
(22, 26)	1	0	1

和DeepWalk算法受干扰影响很大，其预测准确率远低于本文的RE-LP算法（后续将详细介绍其实现过程）。

因此，为了解决现有的LP算法问题，本文提出了RE-LP算法，旨在探索高可靠的不存在（HRNE）链接，并有效地找出当前缺失的链接和未来可能出现的链接，即可能被观察到的存在（POE）链接。

3.2 样本特征提取

RE-LP算法使用基于分类的方法来解决链接预测问题。以网络中的节点对为样本，每个节点对包含8个特征属性和1个类别标签。给定社交网络$G=(V,E)$，其中V和E分别是节点和链接（节点对）的集合。$\{u, v\}\subseteq V$代表网络中任意两个不同节点，$(u,v)\in E$代表节点u和v之间的链接，则未链接的节点对集合$N=\{\{u,v\}\subseteq V|u\neq v, (u,v)\notin E\}$。以节点对为样本，记$\boldsymbol{x}^{(u,v)}=(x_1, x_2, \cdots, x_8)$代表样本$(u,v)$的八维特征向量，$\boldsymbol{x}_i$代表第$i$个特征的值。

$l^{(u,v)}$是样本标签，它由节点对之间是否存在链接来确定。对于任意节点对(u,v)，如果它们之间存在链接，则将其视为正样本，即$l^{(u,v)}=1$；否则，(u,v)被视为负样本，即$l^{(u,v)}=0$。此外，对于无向网络中的任意节点u，定义其邻居节点集合为$\Gamma(u)$。

本文综合考虑局部和准局部相似度，选择了8个广泛使用的基于相似度的指标来构建样本特征向量。为观察样本分布特

征，**图5**展示了两个具有代表性的相似度指标（SRW指数与LHN指数）在不同数据集上的正负样本分布情况，可知这些指数在正样本和负样本之间呈现出较大的分布差异。出于对时间复杂度的考虑，没有使用任何全局指数，而是使用准局部相似性指数作为研究社交网络拓扑特征的替代方案。选定指标的计算式见**表3**，除了这8个指数之外，本文还给出了最经典的相似性指数之一——CN指数的计算式，用于后续算法效率的比较。

$$\mathrm{CN}(u,v)=\left|\Gamma(u)\cap\Gamma(v)\right| \tag{2}$$

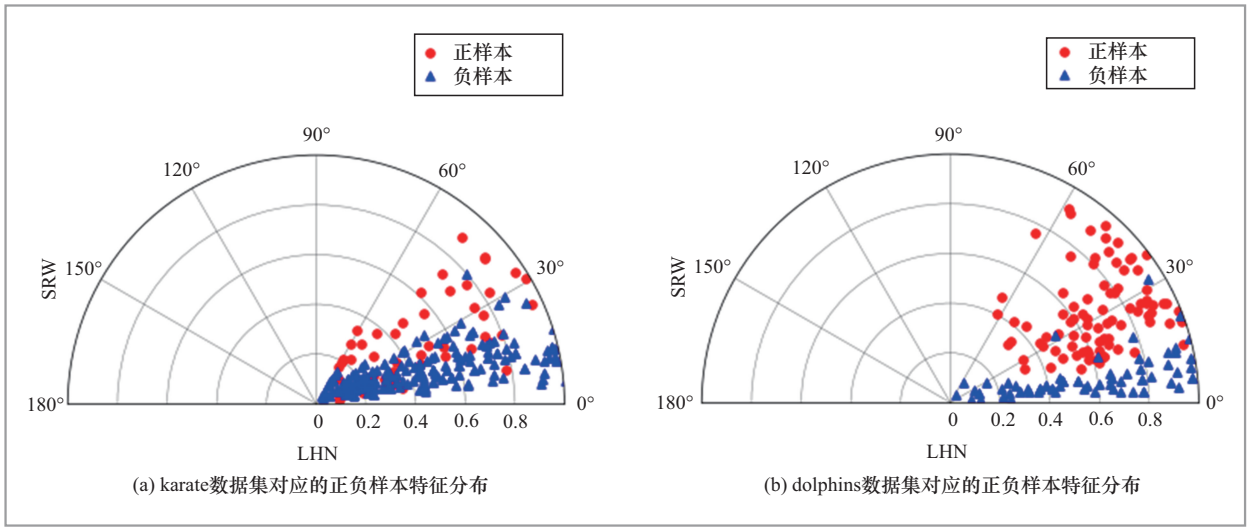
表3中，LoPa指数中的$\boldsymbol{A}^n$表示任意两个节点u和v之间距离为n的路径数量，系数α用于调节3跳路径的拓扑影响力。

对于SRW索引，随机步行器最初从节点u开始测量节点u和v之间的相似性。步行器将以概率c迭代地移动到随机邻居，并以概率$1-c$返回节点u。在游走t步之后步行器随机位于节点v的稳态概率为$\boldsymbol{\pi}_{\mathrm{uv}}(t)$，计算式如式（3）所示：

$$\boldsymbol{\pi}_{\mathrm{uv}}(t)=(1-c)\sum_{\tau=0}^{t}c^{\tau}(\boldsymbol{P}^{\tau})_{\mathrm{uv}} \tag{3}$$

其中，$\boldsymbol{P}$是随机游走的转移概率矩阵，维度为$|N|\times|N|$，如果节点对(u,v)之间存在链接，则$\boldsymbol{P}_{uv}=\dfrac{1}{\deg(u)}$（分母为节点$u$的度），否则$\boldsymbol{P}_{uv}=0$。从节点$u$开始的随机步行器的初始转移概率向量表示为$\boldsymbol{\pi}_u^{(0)}=\boldsymbol{e}_u$，其中$\boldsymbol{e}_u$是第$u$个分量为1且其余分量为0的单位向量。该概率向量的演变方式为：$\boldsymbol{\pi}_u^{(t+1)}=c\boldsymbol{\pi}_u^{(t)}\boldsymbol{P}+(1-c)\boldsymbol{e}_u$。SRW指数通过对有限步长内的随机游走到达概率进行叠加，并同时考虑从节点u和节点v出发的双向随机游走，从而刻画节点对之间的结构相似性。

本文进一步分析了所选用的8种局部与准局部相似性指数，发现它们大多依赖于共同邻居的数量、权重或相似的局部拓扑结构，因此在特征层面存在较强的相关性。若将这些高度相关和冗余的特征直接输入贝叶斯分类器训练，则会违背其特征独立性假设这一前提，从而削弱RE-LP算法的分类性能与泛化能力。为克服这一问题，引入了主成分分析（principal component analysis，PCA）对特征进行去相关与降维。PCA通过正交变换将原始的、相关的特征映射为一组新的、不相关



(a) karate数据集对应的正负样本特征分布　(b) dolphins数据集对应的正负样本特征分布

图5　正样本和负样本之间的相似性指标分布差异示例

表3 相似性指数的计算式

相似性指数	计算式
Jaccard	$\mathrm{JC}(u,v)=\frac{\Gamma(u)\cap\Gamma(v)}{\sqrt{\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert}}$
Salton	$\mathrm{SI}(u,v)=\frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\sqrt{\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert}}$
Hub Promoted	$\mathrm{HPI}(u,v)=\frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\min(\lvert\Gamma(u)\rvert,\lvert\Gamma(v)\rvert)}$
Hub Depressed	$\mathrm{HDI}(u,v)=\frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\max(\lvert\Gamma(u)\rvert,\lvert\Gamma(v)\rvert)}$
LHN	$\mathrm{LHN}(u,v)=\frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert}$
AA	$\mathrm{AA}(u,v)=\sum_{z\in\Gamma(u)\cap\Gamma(v)}\frac{1}{\log\Gamma(z)}$
Local Path	$\mathrm{LoPa}=\boldsymbol{A}^2+\alpha\boldsymbol{A}^3$
Superposed Random Walk	$\mathrm{SRW}_{\mathrm{uv}}(t)=q_{\mathrm{u}}\sum_{\tau=1}^{t}\pi_{\mathrm{uv}}(\tau)+q_{\mathrm{v}}\sum_{\tau=1}^{t}\pi_{\mathrm{vu}}(\tau)$

的主成分，并按方差贡献率进行排序。选择前 k 个主成分，将每个样本特征向量 $\boldsymbol{x}^{(u,v)}\in\mathbb{R}^8$ 转化为降维后的向量 $\boldsymbol{x}'^{(u,v)}\in\mathbb{R}^k$，使其既满足特征独立性假设，又有效减少了冗余维度。

3.3 HRNE/POE链接的划分

为解决传统链接预测算法对非现有链接误判率高的问题，RE-LP算法提出HRNE链接的概念，其核心思想是：部分非现有链接因拓扑特征与真实链接差异显著，可靠性相对于其余非现有链接较高。这一步的核心在于对训练集中的负样本进行可靠性划分，旨在为下游分类器的训练筛选出可靠的负样本，以优化其链接预测性能。

如**图5**所示，根据可视化的特征分布结果，容易观察到部分负样本与正样本之间存在很高的分布不一致性。以正样本的特征分布为参照，标记与正样本分布差异显著的负样本为HRNE链接，而与正样本分布较接近的负样本被标记为POE链接。具体划分过程中需要使用到 K-means（K均值聚类）算法，同时为负样本特征向量增加一个新的维度，以衡量与正样本的分布差异。步骤如下。①将训练正样本 E^{P} 中所有降维后的特征向量输入 K-means（设 K=1）算法，得到正样本簇中心点 C_{E}。②对于训练负样本 N^{P} 中的每个负样本，计算其降维后的特征向量 $\boldsymbol{x}'_i$ 到 C_{E} 的欧氏距离 d_i：

$$d_i=\|\boldsymbol{x}'_i-C_{\mathrm{E}}\| \tag{4}$$

然后，将 $\boldsymbol{x}'_i$ 与 d_i 拼接，形成一个新的聚类特征向量 $\boldsymbol{z}_i=[\boldsymbol{x}'_i,d_i]$。③将所有训练负样本的聚类特征向量 $\boldsymbol{z}_i$ 输入 K-means（设 K=2）算法得到两个簇，将簇中心在距离维度上较小（即与正样本分布较相似，可靠性较低）的簇标记为POE，其样本集记为 POE_0；将簇中心在距离维度上较大的簇标记为HRNE，其样本集记为 HRNE_0。

3.4 RE-LP算法训练流程

贝叶斯分类器作为RE-LP算法的核心模块，本节将详细介绍它的训练过程，包括样本迭代细节与训练终止条件。

将不存在的链接 N^{P} 划分为HRNE和POE链接集后，基于样本集 E^{P} 和 HRNE_0 训练一个初始贝叶斯分类器 $\mathcal{B}_0$，之后使用POE链接集不断对贝叶斯分类器进行迭代优化。每轮迭代训练贝叶斯分类器均需样本集 E^{P} 及上一轮得到的HRNE与POE链接集参与，并输出用于下一轮迭代的HRNE与POE链接集，如**图6**上半部分所示。

迭代训练和链接集演变细节如**图6**下半部分与算法1所示，设当前迭代次数为 t，在第 t 个迭代单元内使用由本次输入样本集训练好的贝叶斯分类器 $\mathcal{B}_{\mathrm{t}}$（算法1第

11步）预测POE链接，按预测概率高低将POE链接集划分为$\mathbf{POE}_{pos}$（高概率）和$\mathbf{POE}_{neg}$（低概率）两个集合（算法1第12步），然后把$\mathbf{POE}_{pos}$作为新的POE链接集$\mathbf{POE}_{t+1}$，而$\mathbf{POE}_{neg}$中的链接则全部加入HRNE链接集以得到$\mathbf{HRNE}_{t+1}$（算法1第13步），再将更新后的HRNE链接集与E^{P}重新组合以训练下一轮的贝叶斯分类器，直到分类器无法从POE链接集中筛选出新的HRNE链接，此时贝叶斯分类器不能学习到更多的可靠负样本，训练终止。同时在每个迭代周期结束时，计算当前分类器在验证集U^{V}上的AUC，并记录表现最优的分类器$\mathcal{B}_{best}$，以防止分类器在训练集上出现过拟合。这种迭代改进过程通过逐步澄清模糊的链接状态来持续增强可靠性，从而有效减少错误预测，因此可认为POE链接集中的其余链接在未来出现的可能性很大。

每次迭代都会细化HRNE和POE链接之间的边界，可以通过不断使用分类置信度更高的负样本来更新训练数据，从而提升分类器的决策能力。贝叶斯分类器是该迭代方法的核心，它利用概率决策来动态提高预测准确率。因此，该算法降低了错误分类的可能性，并提高了POE集中链接的预测置信度。最后，使用测试集U^{T}对训练得到的贝叶斯分类器$\mathcal{B}_{best}$进行性能评估。计算贝叶斯分类器在U^{T}上的预测结果与真实标签的AUC值，从而量化分类器对不存在的链接与真实链接的区分能力。

算法1：RE-LP算法的训练过程

输入：训练样本U^{P}，样本标签$l^{(u,v)}$，样本特征$\boldsymbol{x}^{(u,v)}$

输出：训练完成的贝叶斯分类器$B_{best}(\cdot)$

1. 样本特征向量降维$\boldsymbol{x}'^{(u,v)}=PCA(\boldsymbol{x}^{(u,v)})$

2. 计算正样本簇中心$C_{E}=$

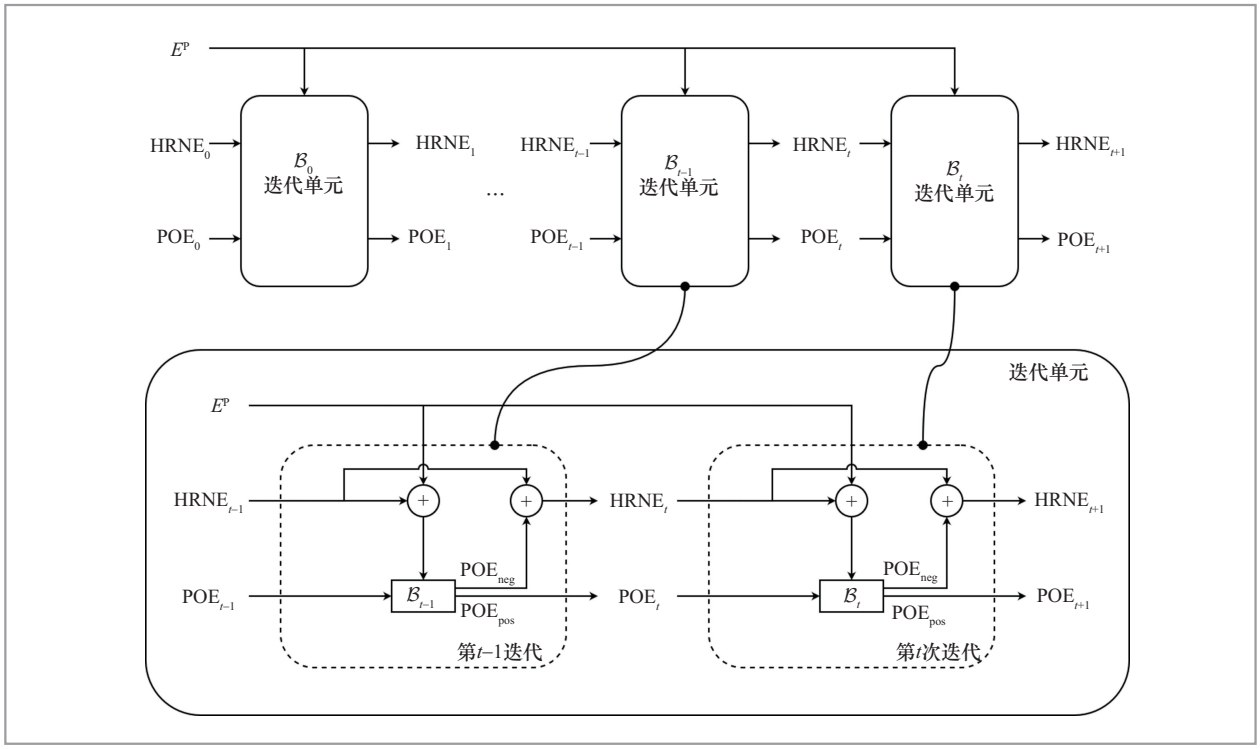


图6 贝叶斯分类器的迭代式训练过程

get_cluster_center(E^{P})

3.计算每个负样本到C_E的欧氏距离$d_i=||\boldsymbol{x}'_i-C_E||$

4.构建负样本聚类特征向量$\boldsymbol{z}_i=[\boldsymbol{x}'_i,d_i]$

5.划分负样本得到两个簇中心点 2_means(N^{P})$\rightarrow C_1+C_2$

6.**If** $||C_1-C_E||>||C_2-C_E||$

7.$HRNE_0\leftarrow C_1$ **and** $POE_0\leftarrow C_2$

8.**Else**

9.$HRNE_0\leftarrow C_2$ **and** $POE_0\leftarrow C_1$

10.**Repeat**：

11.训练贝叶斯分类器 $B_t(\cdot)\leftarrow E^{P}+HRNE_t$

12.划分 POE 链接集 $B_t(POE_t)=POE_{pos}+POE_{neg}$

13.更新样本集 $HRNE_{t+1}\leftarrow POE_{neg}+HRNE_t$ **and** $POE_{t+1}\leftarrow POE_{pos}$

14.**If** $AUC_{best}^{val}<AUC_t^{val}$

15.记录验证集上表现最优分类器 $B_{best}(\cdot)=B_t(\cdot)$

16.**Until** $POE_{t+1}=POE_t$

17.**Return** $B_{best}(\cdot)$

3.5 算法时间复杂度分析

下面分析RE-LP算法的时间复杂度，假设网络G中节点数量为n，节点对数量为$\frac{n(n-1)}{2}$，则时间复杂度主要来源于3个方面。

（1）相似性指标计算：若计算每种相似性指标时间为O(1)，则总的复杂度为$O(8\cdot n^2)$。

（2）2-means聚类：每次迭代需计算所有负样本到两个簇中心点的距离，假设迭代次数为t_c，则复杂度为$O(t_c\cdot n^2)$。

（3）迭代贝叶斯分类器：每次迭代都需要训练1个贝叶斯分类器，样本数为$|E^{P}|+|HRNE|$，因此训练复杂度为$O(|E^{P}|+|HRNE|)$；从POE集合中预测新的HRNE链接的复杂度为O(|POE|)；假设迭代次数为t_i，则迭代过程复杂度为$O(t_i\cdot(|E^{P}|+|HRNE|+|POE|))$。由于$|E^{P}|+|HRNE|+|POE|$的数量上限为$\frac{n(n-1)}{2}$，且最大迭代次数$t_i$取决于POE集的收敛速度，因此总的迭代复杂度为$O(t_i\cdot n^2)$。

在实际中，t_c和t_i通常远小于n，所以RE-LP算法总体时间复杂度仍为$O(n^2)$。

4 实验与分析

本节通过详细的实验验证所提出的RE-LP算法的可行性、合理性和有效性。在从Network repository和Ucinet数据库中获得的8个社交网络数据集上对RE-LP算法的性能进行测试，数据集的基本信息已经总结在表4中，详细信息如下。

（1）aves-sparrow-social

该数据集记录了麻雀之间的社交互动，属于动物社交网络范畴。

（2）aves-weaver-social-08

该数据集记录了28只织巢鸟在2008年的社交网络数据，展示了鸟类之间的互动关系。

（3）ENZYMES_g1

该数据集包含了600个蛋白质的3级结构图，每张图代表1种酶，分为6个类别，本文选取其中1张结构图。

（4）Les Misérables

该数据集是基于维克多·雨果小说《悲惨世界》中的人物共现关系构建的社交网络。如果两个角色在同一章节中出现，则两者之间存在1条边。

（5）karate

该数据集记录了20世纪70年代美国某大学空手道俱乐部成员之间的社交关系。

（6）dolphins

该数据集记录的是从1994年至2001年间观察得到的新西兰Doubtful Sound地区62只瓶鼻海豚之间的社交关系。

（7）ia-southernwomen

该数据集记录的是20世纪30年代收集的美国18位南方女性参加14次社交活动的数据，构成一个二部图。

（8）brunson_club-membership

该数据集记录了25名企业高管在各种社会组织（如俱乐部和董事会）中的会员关系。

每轮训练-验证-测试中，存在的链接和不存在的链接分别被标记为正样本和负样本，按8:1:1的比例划分为训练集、验证集和测试集，其中负样本的HRNE/POE链接划分仅在训练集上进行。

RE-LP算法基于Python语言实现。所有实验均在Windows环境（CPU: AMD Ryzen AI 9 HX 370 w/ Radeon 890M 2.00 GHz, GPU: NVIDIA GeForce RTX 4060 8 GB）下进行。

4.1 RE-LP算法的可行性验证

本部分实验将测试随着HRNE链接集合的改变，RE-LP算法在验证集上预测性能（即AUC值）的变化情况，选用的两个代表性数据集为karate和ia-southernwomen。

图7展示了在两个数据集上分别进行的两次独立实验的收敛过程，共计4组结果。每张图的柱状部分表示迭代训练过程中生成的HRNE链接数随迭代训练次数的变化情况，折线部分表示该次迭代训练下分类器在验证集上的AUC值。可以观察到，随着迭代训练次数的增加，HRNE链接数逐渐趋于稳定，而AUC值也表现出逐步上升并最终收敛的趋势。

在贝叶斯分类器的迭代训练过程中，每次训练都会不断地从POE链接集中选择新的HRNE链接，直到HRNE/POE链接集稳定收敛。**图8**使用数据集ia-southernwomen演示了HRNE和POE链接集的迭代训练变化过程。**图8（a）**显示了不存在的链接初始分类情况，其中绿、蓝色链接分别属于HRNE和POE链接集；在**图8（b）**到**图8（e）**显示的迭代训练过程中，贝叶斯分类器从POE集中选择要加入HRNE集中的链接，这些链接被标记为红色；**图8（f）**显示了迭代训练结束时的HRNE和POE链接集。

表4 数据集的基本信息

数据集	节点数/个	链接数/条	密度	最大度数
aves-sparrow-social	31	211	0.454	23
aves-weaver-social-08	28	145	0.384	22
ENZYMES_g1	37	168	0.252	7
Les Misérables	77	254	0.087	36
karate	34	78	0.139	17
dolphins	62	159	0.084	12
ia-southernwomen	18	78	0.490	16
brunson_club-membership	25	93	0.307	21

以上实验结果表明，RE-LP算法具有良好的收敛性能，具备处理社交网络链接预测问题的能力。

4.2 RE-LP算法的合理性验证

本部分实验将验证RE-LP算法预测结果在正样本和负样本之间的差异。使用dolphins数据集从现有和不存在的链接集中分别随机提取10条链接，然后通过RE-LP算法构建的贝叶斯分类器来预测所选链接的存在概率值。如果分类器能够为现有链接赋予更高的概率值，则意味着分类器能够对社交网络中的链接做出合理的预测。

图9显示了初始的dolphins社交网络，并显示了随机选择链接后的社交网络，所选的正例和负例分别用蓝色线和红色线标记。**表5**显示了RE-LP算法的预测结果，根据预测的概率按降序对链接进行排序。实验结果显示RE-LP算法能够正确预测现有和不存在的链接，这表明该算法具备正确预测社交网络链接的能力。

4.3 RE-LP算法的有效性验证

本部分实验在8个有代表性的社交网络上，将RE-LP算法的链接预测表现（即AUC值）与3种常用的相似性指数（CN、LoPa和AA）和两种基于学习的算法（node2vec和DeepWalk）进行了比较。实验结果见**表6**，其中最好的结果用粗体

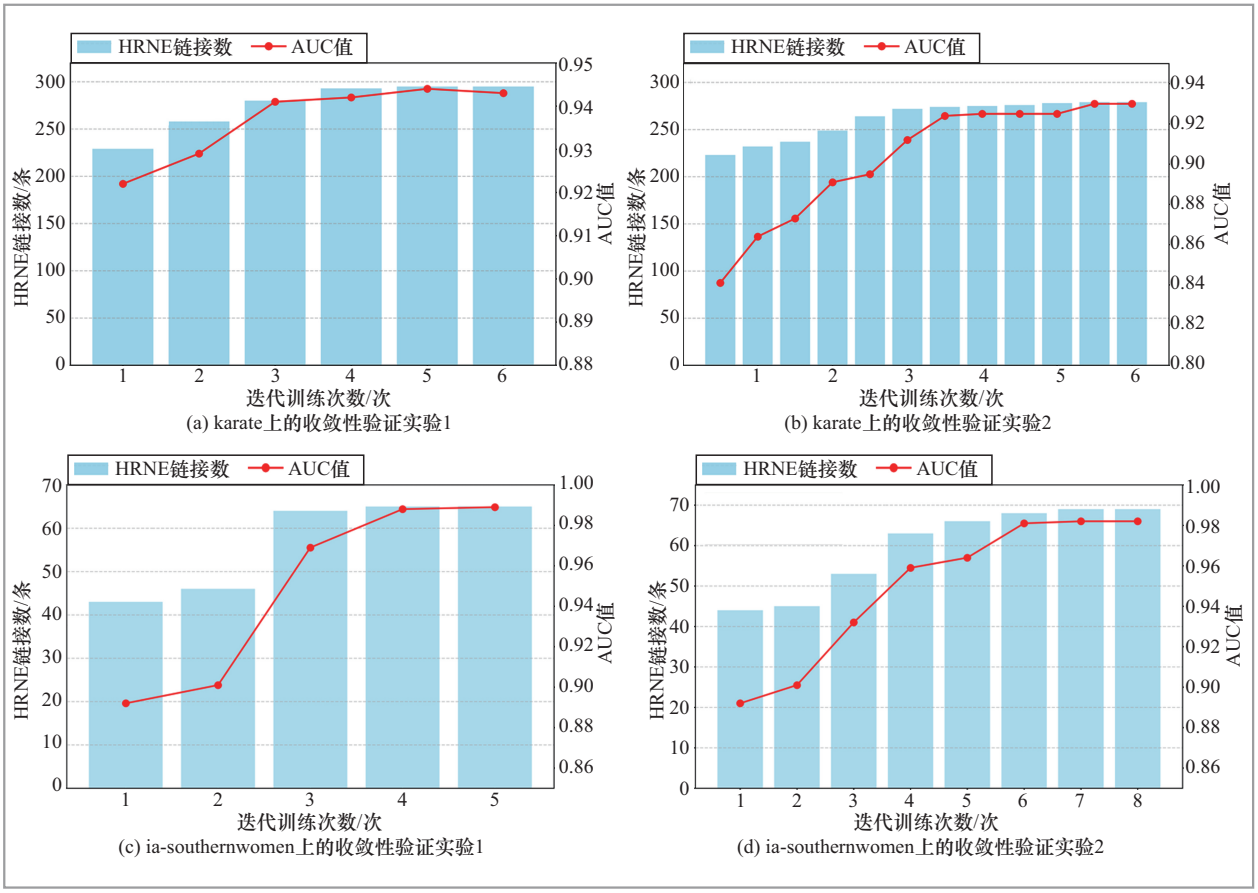


图7 RE-LP算法的收敛性验证

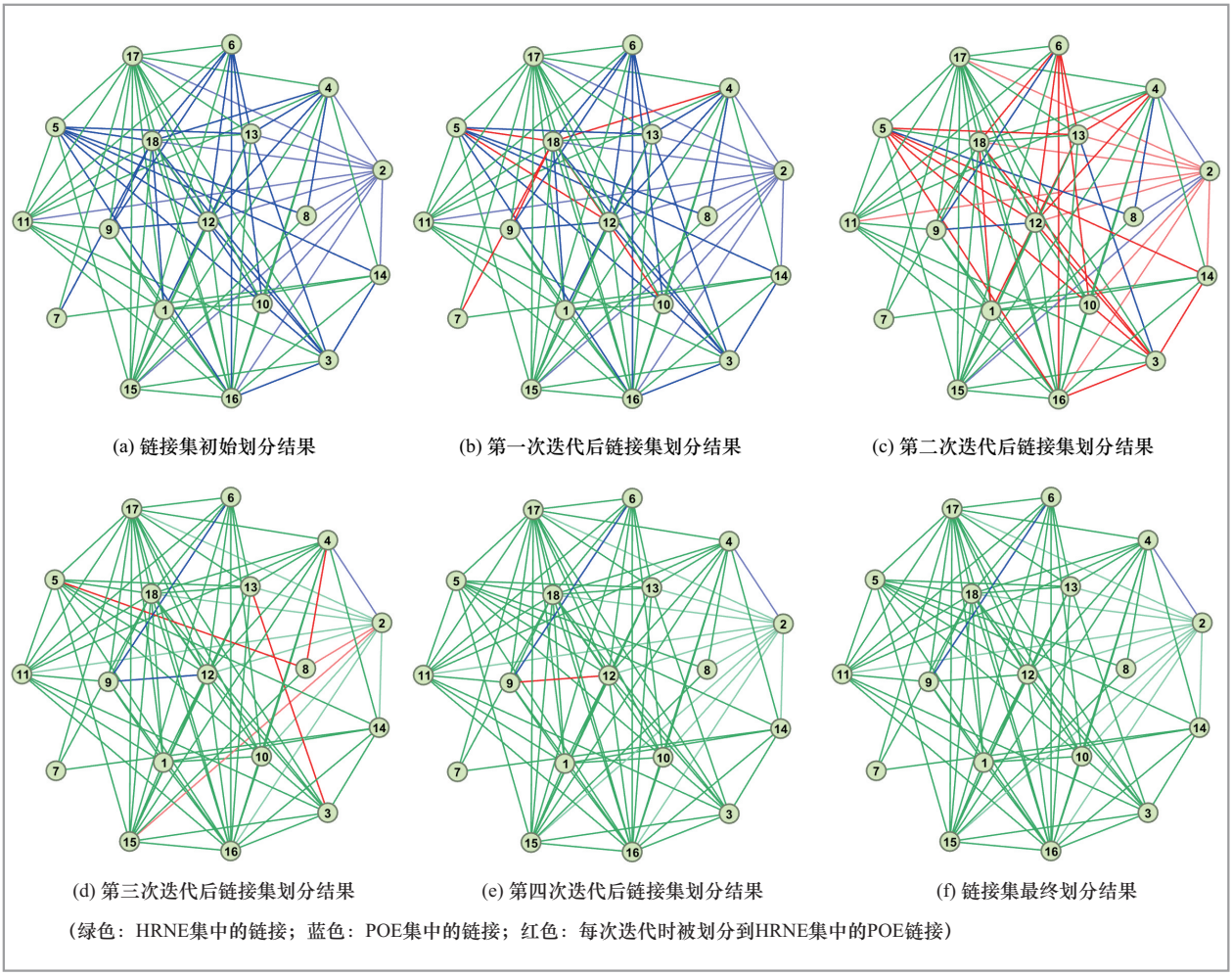
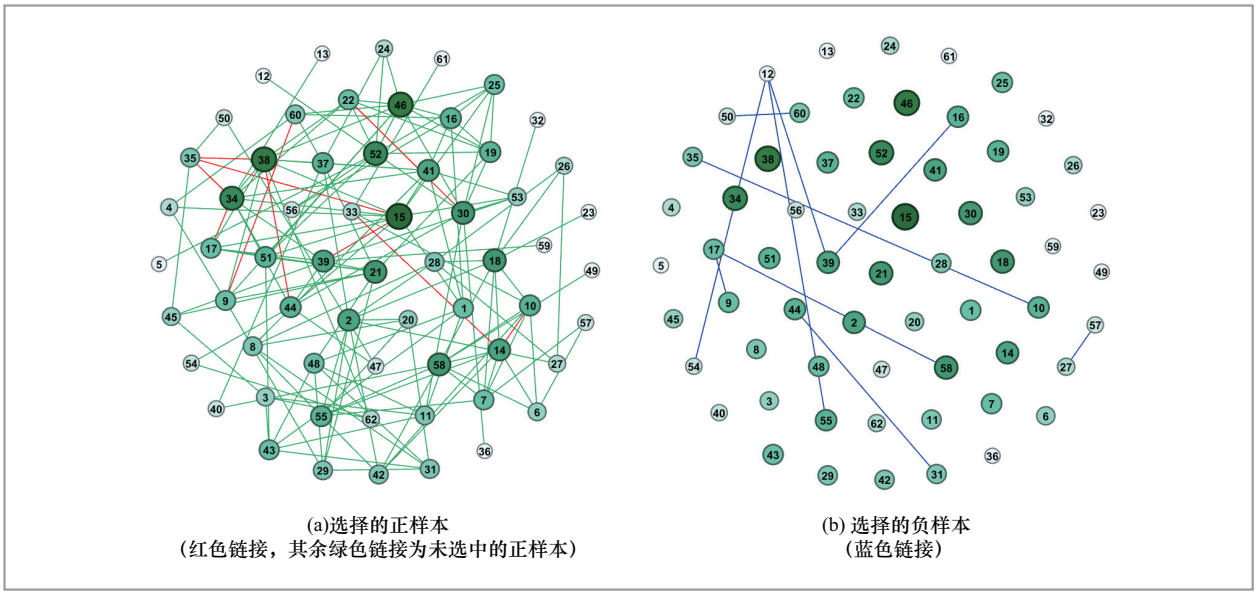


图8 HRNE和POE链接集随迭代次数增加的划分改变情况

标记。结果表明RE-LP算法获得了比其他链接预测算法更好的预测性能，即更高的AUC值。

在RE-LP算法的实际训练过程中，其迭代训练次数与最终表现均依赖HRNE/POE的初始划分情况，所以通过随机选择*K*-means的初始中心点独立执行10次RE-LP算法，分析了算法的鲁棒性。如**图10**所示，分别在两个数据集上展示了RE-LP算法（蓝色箱线图）和其余5种算法（折线图）在3轮独立训练-验证-测试划分下的AUC值分布情况。每轮测试均包含了10次不同HRNE/POE初始划分情况下的运行结果，箱体中线代表AUC值中位数，其余算法以折线图形式叠加。通过对比可见，箱体高度和紧凑的分布不仅表明了RE-LP算法具有较高的鲁棒性，同时平均性能也领先于其他算法。

由于假设属于POE链接集的不存在链接在未来出现的可能性很大，算法对它们的预测概率较高，这对算法整体表现可能存在负面影响。因此还测试了仅选取HRNE链接而不是任意不存在的链接作为测试负样本时RE-LP算法的性能。见**表7**，通过仅使用HRNE链接作为测试负样本计算的AUC值具有显著提高。这进一步验证了HRNE链接的高可靠性，并证明先前的假设合理。



(a)选择的正样本
（红色链接，其余绿色链接为未选中的正样本）

(b) 选择的负样本
（蓝色链接）

图9 RE-LP算法在dolphins数据上的合理性验证

表5 RE-LP算法预测结果

链接	链接标签	预测概率值
(10, 14)	1	1.000000
(15, 39)	1	1.000000
(9, 60)	1	1.000000
(17, 34)	1	1.000000
(22, 30)	1	1.000000
(34, 35)	1	1.000000
(35, 38)	1	1.000000
(15, 35)	1	1.000000
(14, 33)	1	1.000000
(38, 44)	1	0.999858
(9, 17)	0	0.999772
(10, 35)	0	0.000002
(16, 39)	0	0.000001
(50, 60)	0	0.000001
(31, 44)	0	0.000001
(12, 39)	0	0.000000
(27, 57)	0	0.000000
(12, 54)	0	0.000000
(17, 58)	0	0.000000
(12, 55)	0	0.000000

上述实验结果表明RE-LP算法可以通过分配适当概率值的链接，成功地确定现有链接的状态。此外，RE-LP算法还具有检测未来可能观察到的链接的能力。

5 结束语

本文提出了一种基于可靠性增强的链接预测算法来解决现有社交网络链接预测算法可靠性低和探测能力弱的问题。RE-LP算法首先根据选定的相似度指标为每个样本构建特征向量；其次，根据样本特征分布不一致性将不存在的链接划分为HRNE和POE链接集；最后，该算法基于划分的链接集迭代训练一个贝叶斯分类器并不断地重新划分POE链接集直至收敛。在一系列公开数据集上，本文对RE-LP算法的可行性、合理性和有效性进行了系统性的验证。实验结果表明：RE-LP算法的平均表现优于当前通用的5种链接预测算法，同时具有较高的鲁棒性；在实际

表6 在8个社交网络上6种链接预测算法的AUC值比较

数据集	链接预测算法					
	RE-LP	CN	LoPa	AA	node2vec	DeepWalk
aves-sparrow-social	**0.897**	0.774	0.747	0.794	0.812	0.853
aves-weaver-social-08	**0.882**	0.613	0.728	0.692	0.730	0.726
ENZYMES_g1	**0.807**	0.498	0.593	0.520	0.619	0.727
Les Misérables	**0.916**	0.826	0.832	0.831	0.876	0.876
karate	**0.938**	0.688	0.699	0.702	0.630	0.677
dolphins	**0.879**	0.684	0.730	0.681	0.689	0.688
ia-southernwomen	**0.989**	0.681	0.694	0.694	0.680	0.686
brunson_club-membership	**0.865**	0.504	0.618	0.533	0.669	0.661
均值	**0.897**	0.659	0.705	0.680	0.713	0.737

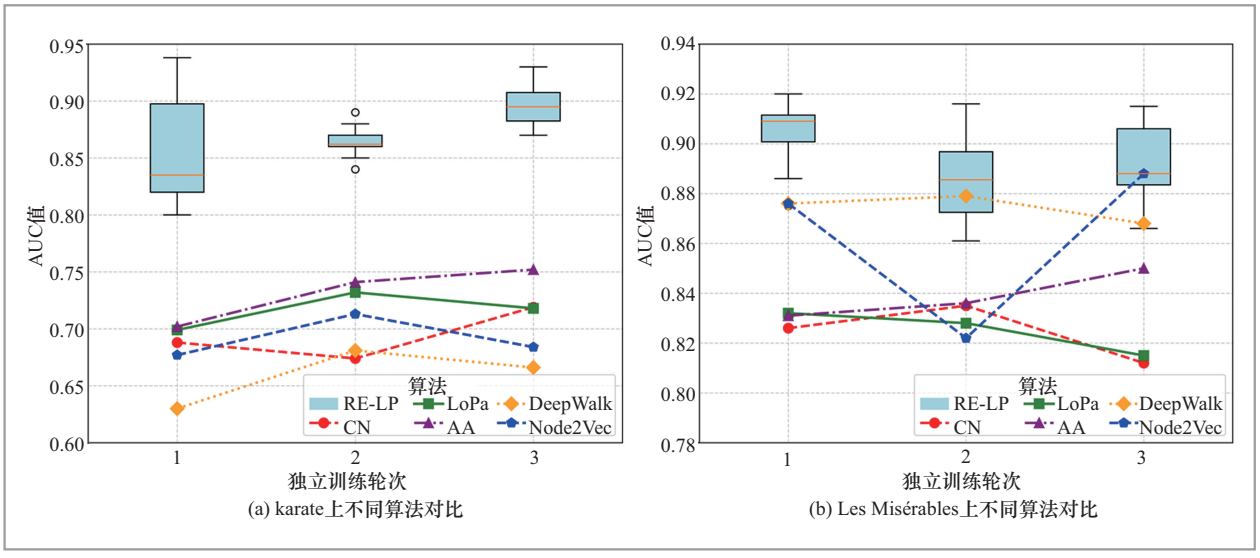


图10 RE-LP算法与主流算法的预测表现对比及鲁棒性分析

执行方面，RE-LP算法不仅可以识别当前缺失的链接，还可以检测未来可能建立的链接。

未来研究将着重围绕以下3个方面开展：①使用RE-LP算法处理分布式环境下大规模社交网络的链接预测问题；②利用深度学习技术生成更有代表性的特征来描述链接信息；③寻找RE-LP算法的目标应用场景，例如精准内容推送和服务行为监控[34]。

表7 RE-LP算法对应不同测试负样本的性能

数据集	测试负样本	
	不存在的链接集	HRNE链接集
aves-sparrow-social	0.897	0.947
aves-weaver-social-08	0.882	0.891
ENZYMES_g1	0.807	0.824
Les Misérables	0.916	0.935
karate	0.938	0.962
dolphins	0.879	0.910
ia-southernwomen	0.989	1.000
brunson_club-membership	0.865	0.903
均值	0.897	0.921

参考文献：

[1] Lyu L Y, Zhou T. Link prediction in complex networks: a survey[J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170.

[2] Lorrain F, White H C. Structural equivalence of individuals in social networks[J]. The Journal of Mathematical Sociology, 1971, 1(1): 49-80.

[3] Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et des Jura[J]. Bulletin De La Societe Vaudoise des Sciences Naturelles, 1901, 37(142): 547-579.

[4] Salton G, Mcgill M J. Introduction to modern information retrieval[M]. New York: McGraw-Hill, 1983.

[5] Carass A, Roy S, Gherman A, et al. Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis[J]. Scientific Reports, 2020, 10(1): 8242.

[6] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(5586): 1551-1555.

[7] Leicht E A, Holme P, Newman M E J. Vertex similarity in networks[J]. Physical Review E, 2006, 73(2): 026120.

[8] Adamic L A, Adar E. Friends and neighbors on the web[J]. Social Networks, 2003, 25(3): 211-230.

[9] Zhou T, Lu L Y, Zhang Y C. Predicting missing links via local information[PP]. arXiv preprint, 2009, arXiv: 0901.0553.

[10] Katz L. A new status index derived from sociometric analysis[J]. Psychometrika, 1953, 18(1): 39-43.

[11] Jeh G, Widom J. SimRank: a measure of structural-context similarity[C]//Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 538-543.

[12] Tong H H, Faloutsos C, PanAN J Y. Fast random walk with restart and its applications[C]//Proceedings of the Sixth International Conference on Data Mining (ICDM'06). Piscataway: IEEE Press, 2006: 613-622.

[13] Pastor-Sattorras R, Vespignani A. Epidemics and immunization in scale-free networks[M]. Berlin: Wiley-VCH, 2003.

[14] Lyu L Y, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 2009, 80(4): 046122.

[15] Liu W, Lyu L. Link prediction based on local random walk[J]. Europhysics Letters, 2010, 89(5): 58007.

[16] Fouss F, Pirotte A, Renders J M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 355-369.

[17] Hasan M A, Chaoji V, Salem S, et al. Link prediction using supervised learning[C]//SDM06: Workshop on Link Analysis, Counter-Terrorism and Security. Philadelphia: Society for Industrial and Applied Mathematics, 2006: 798-805.

[18] De Sá H R, Prudêncio R B C. Supervised link prediction in weighted networks[C]//Proceedings of the 2011 In-

ternational Joint Conference on Neural Networks. Piscataway: IEEE Press, 2011: 2281-2288.

[19] Bütün E, Kaya M, Alhajj R. Extension of neighbor-based link prediction methods for directed, weighted and temporal social networks[J]. Information Sciences, 2018, 463/464: 152-165.

[20] Lu Z D, Savas B, Tang W, et al. Supervised link prediction using multiple sources[C]//Proceedings of the 2010 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2010: 923-928.

[21] Kashima H, Kato T, Yamanishi Y, et al. Link propagation: a fast semi-supervised learning algorithm for link prediction[C]//Proceedings of the 2009 SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics, 2009: 1100-1111.

[22] Li X Y, Du N, Li H, et al. A deep learning approach to link prediction in dynamic networks[C]//Proceedings of the 2014 SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics, 2014: 289-297.

[23] Li K Y, Tu L L, Chai L. Ensemble-model-based link prediction of complex networks[J]. Computer Networks, 2020, 166: 106978.

[24] 罗凯靖, 张育铭, 何玉林, 等. Bootstrap样本大数据模型和分布式集成学习方法[J]. 大数据, 2024, 10(3): 93-108.
Luo K J, Zhang Y M, He Y L, et al. Bootstrap sample partition data model and distributed ensemble learning[J]. Big Data Research, 2024, 10(3): 93-108.

[25] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature, 2008, 453(7191): 98-101.

[26] Guimerà R, Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(52): 22073-22078.

[27] Pizzato L, Rej T, Akehurst J, et al. Recommending people to people: the nature of reciprocal recommenders with a case study in online dating[J]. User Modeling and User-Adapted Interaction, 2013, 23(5): 447-488.

[28] Shi S L, Li Y P, Wen Y M, et al. Adding the sentiment attribute of nodes to improve link prediction in social network[C]//Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Piscataway: IEEE Press, 2015: 1263-1269.

[29] Shahmohammadi A, Khadangi E, Bagheri A. Presenting new collaborative link prediction methods for activity recommendation in Facebook[J]. Neurocomputing, 2016, 210: 217-226.

[30] 王续澎, 何洪波, 王闰强. 基于5W传播模型的技术体系: 计算传播技术综述[J]. 大数据, 2025, 11(3): 139-166.
Wang X P, He H B, Wang R Q. Technology system based on the 5W communication model: a review of computational communication technology[J]. Big Data Research, 2025, 11(3): 139-166.

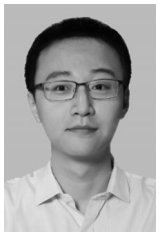
[31] Wu H X, Song C Y, Ge Y, et al. Link prediction on complex networks: an experimental survey[J]. Data Science and Engineering, 2022, 7(3): 253-278.

[32] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2014: 701-710.

[33] Grover A, Leskovec J. node2vec: scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 855-864.

[34] 马文聪, 谭毓安, 冯硕, 等. 基于Android无障碍服务的行为监控[J]. 电子学报, 2023, 51(12): 3572-3581.
Ma W C, Tan Y A, Feng S, et al. Behavior monitoring based on Android accessibility service[J]. Acta Electronica Sinica, 2023, 51(12): 3572-3581.

作者简介



何玉林（1982– ），男，博士，人工智能与数字经济广东省实验室（深圳）研究员、高级工程师，主要研究方向为数据挖掘与机器学习、大数据处理与分析、大数据近似计算技术、多样本统计分析理论与方法等。

孙洪涛（2000– ），男，深圳大学计算机与软件学院硕士生，主要研究方向包括社交网络的链接预测、人工智能算法设计与应用。

秦红莲（1995– ），女，人工智能与数字经济广东省实验室（深圳）算法工程师，主要研究方向包括大数据处理与分析技术、机器学习和数据挖掘算法。

黄舒影（2001– ），女，卡耐基梅隆大学硕士生，主要研究方向包括机器学习算法和分布式数据处理系统。

崔来中（1984– ），男，博士，深圳大学计算机与软件学院特聘教授，主要研究方向包括下一代互联网体系结构、软件定义网络、边缘计算、大数据分析、机器学习和智能计算。

黄哲学（1959– ），男，博士，深圳大学计算机与软件学院特聘教授，主要研究方向为数据挖掘、机器学习、大数据处理与分析、大数据系统计算技术等。

收稿日期: 2025-10-17

通信作者: 何玉林，yulinhe@gml.ac.cn

基金项目: 广东省自然科学基金项目(No. 2023A1515011667)；深圳市科技重大专项项目(No. KJZD20230923114809020)

Foundation Items: The Natural Science Foundation of Guangdong Province (No. 2023A1515011667), Science and Technology Major Project of Shenzhen (No. KJZD20230923114809020)