情报学报　2024 年 9 月　第 43 卷　第 9 期
Journal of the China Society for Scientific and Technical Information, Sept. 2024, 43(9): 1105-1115

DOI: 10.3772/j.issn.1000-0135.2024.09.009

基于网络表示学习的短视频流行度预测研究

朱恒民[1,2]，徐　凝[1]，魏　静[1]，沈　超[1]

（1. 南京邮电大学管理学院，南京　210003；
2. 江苏高校哲学社会科学重点研究基地—信息产业融合创新与应急管理研究中心，南京　210003）

摘　要　预测短视频的流行度不仅有助于短视频平台高效地管理信息，还可以对社会舆情进行监控。针对已有工作仅考虑短视频多模态内容特征构建流行度预测模型这一现实情况，本文基于网络表示学习，提出融合短视频内容和网络结构特征的流行度预测模型。首先，基于爬取的抖音数据构建包含短视频、发布者和评论者节点，以及发布和评论关系的异质信息网络，将其映射为短视频和发布者两个同质网络，选择node2vec算法表征网络结构，作为网络模态；其次，提取短视频的多模态内容特征，采用低秩多视图子空间学习方法融合短视频内容和结构特征，作为流行度预测模型的输入；最后，构建短视频流行度预测的多层感知机回归模型，并进行对比和消融实验。结果表明，融合网络结构能够降低短视频流行度预测的误差，各模态对短视频流行度预测的影响程度依次为文本、网络、社交、音频和视觉模态。本文融合了短视频内容和网络结构特征，为基于特征工程的短视频流行度预测提供了新思路。

关键词　短视频；网络表示学习；流行度预测；多层感知机

Study of Short Video Popularity Prediction Based on Network Representation Learning

Zhu Hengmin[1,2], Xu Ning[1], Wei Jing[1] and Shen Chao[1]

(1. *School of Management, Nanjing University of Posts and Telecommunications, Nanjing*　210003;
2. *Jiangsu University Philosophy and Social Science Key Research Base—Information Industry Integration Innovation and Emergency Management Research Center, Nanjing*　210003)

Abstract: Predicting the popularity of short videos not only helps short-video platforms with efficient information management but also plays an important role in monitoring public opinion. Unlike existing studies that focus only on multimodal content features of short videos, to construct a popularity prediction model, we propose a popularity prediction model based on network representation learning, fusing content and network structural features. First, based on the dataset crawled in Douyin, a heterogeneous information network consisting of nodes was constructed, including short videos, publishers, commenters, and edges. After mapping into two different homogeneous networks, namely, short-video and publisher networks, node2vec was selected to represent the network structure in the embedding space as a network modality. Second, the multimodal content features of short videos were extracted and fused using low-rank multiview embedding learning. Finally, a multilayer perceptron machine regression model was proposed for short-video popularity prediction. Comparisons and ablation experiments were further conducted. The results show that fusing network structure features can reduce the error of short-video popularity prediction. The degree of influence of the various modalities on short-video popu-

收稿日期：2024-01-30；**修回日期**：2024-04-28

基金项目：国家自然科学基金项目“基于网络表示学习的短视频舆情传播机理与演化研究”（72374111）；江苏省研究生科研与实践创新计划项目“异构视角下基于网络表示学习的短视频个性化推荐”（KYCX23_0933）。

作者简介：朱恒民，男，1974年生，博士，教授，博士生导师，研究方向为数据挖掘、舆情管理；徐凝，通信作者，女，1999年生，硕士研究生，研究方向为数据挖掘、舆情管理，E-mail：ningxu1022@163.com；魏静，女，1982年生，博士，教授，硕士生导师，研究方向为复杂网络、舆情传播；沈超，男，1983年生，博士，副教授，硕士生导师，研究方向为电子政务与网络舆情。

larity prediction consisted of the textual, network, social, acoustic, and visual modalities, in decreasing order. Our method, which combines short-video content and network structure features, provides new ideas for short-video popularity prediction based on feature engineering.

Keywords: short video; network representation learning; popularity prediction; multilayer perceptron

0 引 言

随着移动互联网的发展和智能手机的普及，短视频因其制作简单、信息量大、代入感强，已成为当前最具热度和代表性的传播形态，备受广大网民青睐。对短视频的流行度进行预测，不仅可以帮助短视频平台高效地管理信息、优化内容推荐，还可以帮助政府相关机构监测社会舆论的动向，预警可能涉及的敏感话题和事件。

相较于传统微博、论坛中发布的文字和图片数据，短视频承载的信息量更大，包含文本、音频和视觉等多模态数据，特征维度高，语义挖掘困难。基于深度学习技术充分挖掘短视频多模态数据特征，构建流行度预测模型，是目前短视频流行度预测的主要方法。已有研究验证了短视频的文本、音频、视觉及其发布者属性特征对其流行度预测的影响。部分学者将社交网络结构特征引入文本或图片等社交媒体的流行度预测，但主要基于传统社交网络分析方法描述同质网络的拓扑性质，不能充分挖掘稀疏性强、具有异质性的短视频平台社交网络。

本文提出基于网络表示学习的短视频流行度预测模型。从短视频平台中构建“用户-视频”异质信息网络，基于网络表示学习方法表征网络结构特征，将节点表示成统一、低维的稠密向量，综合短视频内容特征，作为后续短视频流行度预测模型的输入，有效解决了短视频流行度预测中的数据量大、维度高、稀疏性强等挑战性问题。研究工作证实了“用户-视频”异质信息网络的结构特征在流行度预测中的有效性，为短视频流行度预测研究提供了新的思路。

1 研究现状

随着深度学习方法数据处理效率和特征表示能力不断提高，基于特征工程的短视频流行度预测方法成为主流，即提取可能影响流行度的相关特征后，设计机器学习模型来预测短视频流行度。其中，部分学者探究了短视频视觉、文本、音频和发布者属性等内容特征对短视频流行度的影响机理。Trzciński 等[1]仅从短视频关键帧序列的视觉线索出发，构建了基于长期循环卷积网络（long-term recurrent convolution network，LRCN）的流行度预测方法；武维等[2]引入神经网络因子分解机（neural factorization machines，NFM）处理短视频种类、视频描述等文本内容特征，融合内容和时序信息进行流行度预测，发现内容特征在很大程度上影响了流行度预测模型的效果；Zohourian 等[3]收集了短视频的标题、音频、封面等 25 个内容特征来表征短视频，并对比了多个短视频流行度预测的回归和分类模型，发现局部多项式回归和决策树算法性能优于其他模型；Xie 等[4]和井佩光等[5]不仅考虑了短视频的文本、音频和视觉模态，还提取了短视频发布者的社交模态，并利用不同的深度学习方法融合短视频的多模态特征，验证了多模态特征对流行度预测的重要影响。

在预测文本或图片等社交媒体流行度时，部分学者引入了信息在社交平台传播形成的社交网络结构特征，为流行度预测提供了新思路。例如，Hong 等[6]和 Ma 等[7]分别验证了推特社交媒体中图结构属性和情境特征对推文或标签流行度的影响；Meghawat 等[8]综合考虑视觉特征、文本特征和社交网络特征（平均视图和组计数等）来预测社交媒体中图片的流行度；Shulman 等[9]发现，出入度、连通性、网络密度等图结构特征对不同社交媒体流行度预测影响不同，例如，在图片社交平台中，网络密度越高，流行度越高，但在音乐社交平台中，网络密度越高，流行度越低。在视频领域，少数学者提出了基于拓扑建模的流行度预测方法。例如，Li 等[10]基于早期视频传播过程构建未来的视频传播网络，揭示出底层在线社交网络拓扑结构在视频传播和流行度演变中发挥着重要作用；Zhang 等[11]将短视频按地理区域聚类，构建短视频区域相关性网络，并通过改进的图卷积神经网络进一步模拟区域之间的影响，以捕捉短视频未来的流行度。上述研究方法对网络结构特征的提取大多是基于传统社交网络分析方法描述同质网络的拓扑性质，难以适用于数据量大、维度高、稀疏性强的短视频网络，且没有区分对象及其间关系的异质性，易造成不可逆的信息

损失。

异质网络的表示学习不仅能够把不同类型的异质信息融合为整体，有效缓解网络数据的稀疏性难题，而且能高效地实现语义相关性计算，显著提升节点表示质量和下游任务的计算效率[12]。Wei 等[13]基于图卷积网络，利用相邻节点的拓扑结构特征来丰富特定模态的“用户-短视频”二元网络中节点的表示，从而捕获用户的偏好；Guo 等[14]利用 DeepWalk 对“用户-短视频”异质信息社交网络结构进行表征，将其作为短视频推荐模型的输入之一；Sang 等[15]基于元路径网络表示学习方法提取短视频多模态异质信息网络结构，用于视频点击率的预测。上述研究将短视频网络的结构特征应用于短视频推荐等任务中，但是，鲜有研究涉及基于网络表示学习如何将结构特征应用于短视频流行度预测。

2　研究思路和方法

本文提出了基于异质信息网络表示学习的短视频流行度预测方法，具体研究思路如图 1 所示。首先，构建“用户-视频”异质信息网络，基于定义的关系将异质网络映射成短视频和发布者同质信息网络，利用 node2vec 学习网络结构特征，获得网络模态；其次，提取短视频的文本、音频和视觉模态，以及短视频发布者的社交模态，并基于低秩多视图嵌入学习，进行多模态融合，获得模型的输入；最后，构建短视频流行度预测的多层感知机模型，设计对比实验验证模型的有效性，并通过消融实验对短视频流行度影响因素进行评估和讨论。

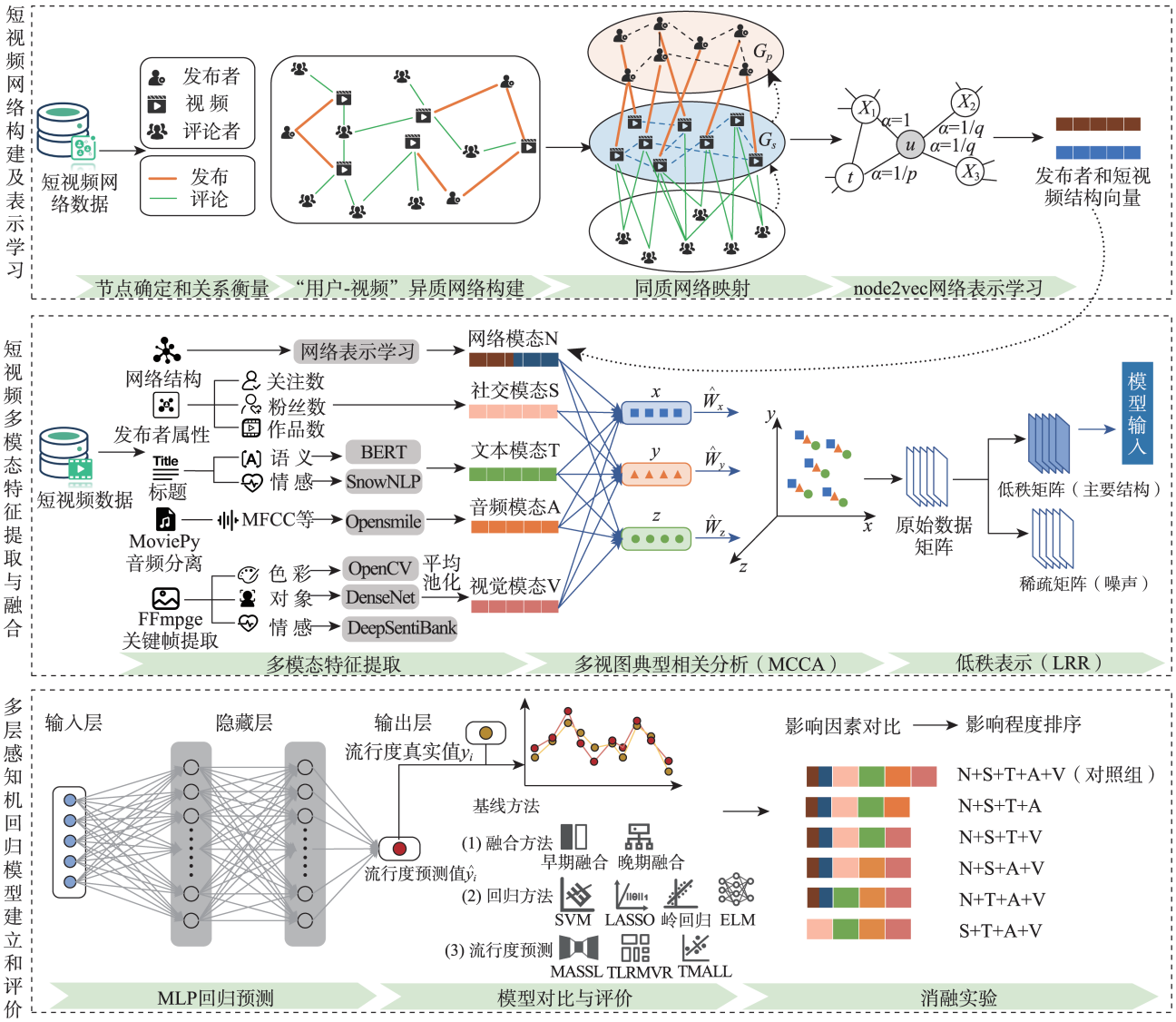

图 1　研究思路

2.1 “用户-视频”异质信息网络的构建及网络表示学习方法

（1）“用户-视频”异质信息网络构建。短视频社交平台具有庞大的用户群体、视频作品，以及错综复杂的信息扩散路径。若将用户和视频作品看作网络节点，将用户通过发布、评论和分享等信息传播行为形成的节点之间联系看作网络连边，则短视频社交平台本质上就是一个异质的复杂信息网络。挖掘“用户-视频”异质信息网络的结构特征，获得质量更高的节点向量表示，可以提高流行度预测的准确性。

定义.“用户-视频”异质信息网络为 $G_{\mathrm{HIN}}=\{V,E\}$。其中，V 代表节点集合，且 $V=\{P,S,C\}$，P、S 和 C 分别表示发布者（publisher）、短视频（short video）和评论者（commenter）；E 代表边的集合，且 $E=\{E_{ps},E_{cs}\}$；E_{ps}、E_{cs} 分别表示发布者和短视频之间的发布关系、评论者和短视频之间的评论关系。

（2）“用户-视频”异质信息网络的表示学习方法。网络的结构蕴含着节点间丰富的语义关系，两个具有相似受众群体的视频，其内容更有可能相近；发布了内容相似短视频的不同发布者，更有可能具有相近的创作风格和目标受众。为了充分利用“用户-视频”异质信息网络的结构特征，首先，依据共同评论和视频发布关系，将异质信息网络映射成短视频和发布者两个同质信息网络；其次，基于 node2vec 分别对短视频和发布者网络进行表示学习，生成短视频和发布者的结构特征向量，以期为短视频流行度预测提供更全面的上下文信息。

（3）短视频网络。在真实的短视频平台中，视频与视频之间的直接关系难以获取。考虑到具有共同评论的短视频在内容上往往具有相似性这一现象，将共同评论关系定义为视频之间的连边，即当两条短视频被一定数量的共同评论者评论时，两者之间存在一条连边。因此，短视频网络可定义为 $G_s=\{S,E_s\}$，其中，S 为短视频集合，E_s 为短视频之间的共同评论关系集合，即 $E_s=\{(S_m,S_n):\exists C_i\in C,(C_i,S_m)\in E_{cs}\wedge(C_i,S_n)\in E_{cs}\}$，$C$ 为评论者集合。

（4）发布者网络。将两个发布者之间的关系定义如下：如果发布的短视频被一定数量的共同评论者评论，那么这两个发布者之间也存在一条连边。这可解释为他们发布了内容相似的短视频。因此，发布者网络可定义为 $G_p=\{P,E_p\}$，其中，P 为发布者集合，E_p 为发布者之间的关系集合，即 $E_p=\{(P_m,P_n):\exists(S_i,S_j)\in E_s,(P_m,S_i)\in E_{ps}\wedge(P_n,S_j)\in E_{ps}\}$。

（5）同质网络表示学习。短视频网络和发布者网络均属于同质网络。在同质网络中，相互连接的节点间往往存在着相似的模式，因此，可以从邻居节点入手，通过随机游走对目标节点进行表示学习。DeepWalk 是一个经典的、基于随机游走的网络表示学习算法，其先利用随机游走策略在图中生成节点序列，再利用自然语言处理领域中基于 word2vec 框架的方法来学习节点的表示，从而捕获图的结构特征。node2vec 算法是在 DeepWalk 的基础上，通过参数 p 和 q 控制随机游走过程中的广度优先搜索（breadth first search，BFS）邻域和深度优先搜索（deep first search，DFS）邻域采样策略。Grover 等[16]指出，这种采样的灵活性可以使 node2vec 找到最佳的平衡点，充分捕捉网络中节点的同质性和结构相似性，从而综合考虑网络中的结构信息。因此，本文利用 node2vec 方法，对 G_s 和 G_p 进行网络表示学习，分别获得短视频和发布者的结构特征，定义为网络模态（network modality，NM）。

2.2 短视频多模态特征的提取

与纯文本和图片信息不同，短视频同时包含文本、音频和视觉等多模态特征，这使得短视频在传达信息时具有更丰富和多样化的表现形式。在短视频流行度预测模型中，本文综合考虑短视频的文本、音频和视觉模态以及发布者的社交模态，各模态特征的提取方法阐述如下。

2.2.1 社交模态

本文爬取发布者的粉丝数、作品数和关注数 3 种相关属性作为短视频的社交模态。粉丝数反映了发布者的受众规模和用户忠诚度，作品数反映了发布者的创作活跃度和内容生产能力，关注数反映了发布者的社交互动程度，这些属性综合衡量了发布者的影响力。受抖音推荐机制的作用，影响力越高的发布者，其发布的视频往往能被更多的用户看到，也更容易成为热门视频。为减少数值范围差异对流行度预测性能的影响，先将粉丝数、作品数和关注数进行对数函数归一化处理，再拼接成 3 维社交模态（social modality，SM）。

2.2.2 文本模态

短视频往往都会带有一条简短的标题，不仅简

明概括了短视频要传递的内容和情感信息，还可以吸引用户观看并进行互动。提取标题的语义特征，有助于更好地理解短视频的核心内容；提取标题的情感特征，有助于了解短视频传递的情感色彩，情感对于用户与短视频的互动至关重要，能够影响用户在社交媒体上对短视频的分享、点赞和评论行为。BERT（bidirectional encoder representations from transformers）[17]是目前常用的预训练语言模型，Chinese-BERT-wwm[18]则是一个基于全词掩码（whole word masking，wwm）技术和BERT模型的中文自然语言处理预训练模型。与传统BERT模型采用的随机掩码不同，wwm策略是在预训练任务中，选择整个词而不是词的一部分进行掩码，这样的操作更加适用于中文语言结构，可以帮助模型更好地理解和捕捉语境和语义信息。SnowNLP（simplified Chinese text processing）是一个用于识别中文文本情感的Python库，可以根据文本内容，返回一个0~1的情感值，表示文本的情感倾向。越接近1，情感越积极；越接近0，情感越消极。因此，本文分别采用Chinese-BERT-wwm模型和SnowNLP库提取短视频标题的768维语义和1维情感特征，拼接形成769维文本模态（textual modality，TM）。

2.2.3　音频模态

与传统的图文信息表现方式不同，音频是短视频不可或缺的一部分。音频的录制和制作质量直接影响用户的听觉体验，高质量的音频能够提高视频的整体质感，增加用户的满意度，有助于视频的流行度提升。此外，音频还可以传递情感信息，音频中的音调、音量、节奏等元素能够表达视频中的情感色彩，对于用户产生共鸣和情感连接至关重要。不同于Zohourian等[3]在分析短视频音频特征时，简单地将其划分为全局音乐、局部音乐和不含音乐3类，本文着重分析短视频音频的声学和情感特征，以期深入理解音频模态。首先，采用Python中的MoviePy库分离视频的音频文件，保存为“.wav”无损音轨格式；其次，选择Opensmile[19]开源工具和eGeMAPS[20]（extended Geneva minimalistic acoustic parameter set）特征集提取音频特征，不仅包含基本频谱、梅尔频率倒谱系数（Mel-frequency cepstral coefficients，MFCCs）、声音强度和能量、语音质量和速率在内的声学特征，还包括情感强度、愉悦度、激动度等语音情感特征。基于上述方法，生成88维音频模态（acoustic modality，AM）。

2.2.4　视觉模态

画面是呈现短视频内容的主要载体，直接关系到用户的第一印象。吸引人的画面、丰富的色彩和有趣的对象特征可以引起用户的兴趣，引发用户的观看和互动行为。本文将短视频的画面特征定义为视觉模态，主要包括色彩、对象和情感特征。色彩是衡量视觉吸引力的重要标准，良好的色彩设计有助于提高用户的观看体验，更容易吸引用户的注意力；提取画面的对象特征有助于更精准地理解视频内容，增强对用户兴趣的把握；视频中的情感特征不仅能表达作者的情感，也能影响观众对短视频的情感态度，如果短视频中的情感能够引起观众的共鸣和情感共振，那么观众更有可能积极地与视频互动，从而提高短视频的流行度。

（1）色彩特征提取。首先，利用OpenCV（open source computer vision library）开源计算机视觉库，先生成画面的RGB（red, green, blue）颜色直方图，再在HSV（hue, saturation, value，色调、饱和度、亮度）颜色空间上计算HSV直方图，综合RGB直方图和HSV直方图，进行多维直方图展平、归一化、标准化和降维运算，形成画面在RGB-HSV通道上的色彩特征向量空间。与传统的图片色彩特征提取方式不同，本文不仅考虑了RGB直方图，还进一步考虑了更符合人眼对颜色感知模式的HSV直方图特征，从而全面地描述图像中不同颜色的分布情况，更好地反映人眼对图像色彩的主观感受。

（2）对象特征提取。卷积神经网络（convolutional neural network，CNN）广泛应用于图像的特征提取。DenseNet（densely connected convolutional networks）是CNN的一种变体，Huang等[21]指出，与普通的CNN模型相比，其密集连接的结构使得信息的传递更加充分，有效缓解了梯度消失问题，有助于捕捉多层次、高级别的特征，提高网络的表示能力。DenseNet-121是一种基于DenseNet架构的预训练模型，在ImageNet数据集上取得很高的准确率，其通过密集连接结构提取具有丰富语义信息图片对象特征。本文采用DenseNet-121提取画面的对象特征。

（3）情感特征提取。DeepSentiBank[22]是一个基于CNN的图像情感语义分类器。给定一张图片，DeepSentiBank能够生成由231个形容词和424个名词组成的2089个情感概念（如“young adult”和“safe driver”等）及其对应的概率分布。杨瀚森等[23]指出，这些基于语义的情感概念能够准确地描

述图像内容，同时具有明显的情感倾向，可以很好地体现画面的情感特征。本文采用 DeepSentiBank[22] 模型提取画面的情感特征。

短视频本质上是在时间序列上不断变化的动态图像。本文采用 FFmpeg 工具提取每个视频的关键帧，针对若干关键帧图片，分别采用 OpenCV、DenseNet-121 和 DeepSentiBank 提取 50 维色彩、1024 维对象和 2048 维情感特征，平均池化后，再进行拼接操作，最终形成每个视频的 3122 维视觉模态（visual modality，VM）。

2.3 多模态融合

提取后的多模态特征具有维度高、冗余性和数据不平衡等问题，且初始的特征向量都是在独立的空间中提取的，不仅造成了数据的异构性，还忽略了多模态特征之间的语义相关性，进而影响后续流行度预测模型的稳定性。解决这些问题的一种常用方法是将异构特征投影到一个公共子空间中，保证具有相似语义的多模态数据具有相关的向量表示。首先，对多模态数据进行均值中心化操作，消除不同特征之间的尺度差异，提高模型训练的效果。其次，考虑到提取的网络、社交、文本、音频和视觉模态之间可能存在丰富的信息关联，如短视频的音频通常与画面内容相辅相成，音频情感色彩也与视频画面的情感特征有关联，参考 Jing 等[24]提出的低秩多视图嵌入学习，先使用 MCCA[25]（multi-view canonical correlation analysis）将网络、社会、文本、音频和视觉 5 种模态映射到一个视图间相关性最大的共享子空间之中；再采用低秩学习方法 LRR[26]（low-rank representation）进行矩阵分解，将数据矩阵分解为稠密低秩矩阵和稀疏误差矩阵，从而减少数据矩阵中的冗余信息，提高流行度预测模型的泛化和鲁棒性。采用低秩多视图子空间学习方法融合短视频的 NM、SM、TM、AM 和 VM 这 5 个模态特征，最终将短视频表示为一个 5 维向量。

2.4 短视频流行度预测模型的构建

（1）流行度计量

在抖音等平台上，用户的评论和分享行为促进了短视频的扩散，点赞和收藏行为关系到短视频能否获得平台的推荐。因此，本文采用评论数、分享数、点赞数和收藏数 4 个指标来衡量短视频的流行度。图 2 是 4 个指标与其平均值之间的相关性热力图。4 个指标两两之间的皮尔逊相关系数范围是 0.30~0.65，任意 2 个指标之间的相关性均不大，因此，需要综合考虑 4 个指标，以期较为全面地计量短视频的流行度。从图 2 可以看出，4 个指标与平均值之间的相关性较大，平均值能够相对全面地描述短视频的流行度。借鉴 Chen 等[27]、Jing 等[24]量化 Vine 平台上短视频流行度的方法，选取 4 个指标的均值作为本文中的抖音短视频流行度值，即

$$y_i = \frac{n\_\text{comment}_i + n\_\text{share}_i + n\_\text{like}_i + n\_\text{favorite}_i}{4} \tag{1}$$

其中，y_i 为数据集中第 i 个短视频的流行度值；n_comment、n_share、n_like 和 n_favorite 分别表示短视频的评论数、分享数、点赞数和收藏数。

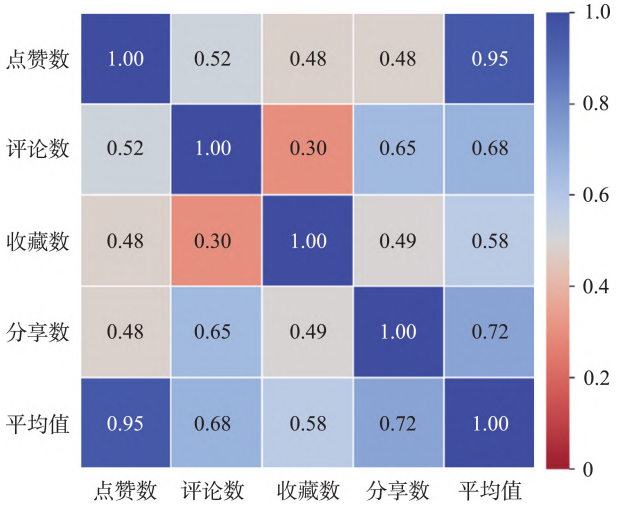

图 2 短视频流行度衡量指标与其平均值相关性热力图

（2）多层感知机回归模型建立

短视频流行度的值为一个连续变量，可将流行度预测任务看作一个回归问题，定义 $y=f(x)$，其中，y 为短视频流行度，x 为基于低秩多视图嵌入学习融合的短视频特征向量，f 为短视频流行度预测模型。

PyTorch 是一种用于深度学习任务的开源机器学习库，以其动态计算图、简洁易懂的 API（application programming interface）设计、丰富的文档和社区支持以及 GPU（graphics processing unit）加速等特点，成为深度学习领域的主流选择。考虑到人工神经网络能够利用深层次的结构处理非线性关系和高维特征等数据中的复杂模式，建立一个基于 PyTorch 的多层感知机（multilayer perceptron，MLP）人工神经网络用于回归任务学习。具体的神经网络结构包含：5 个神经元的输入层，分别对应短视频的 5 维向量；2 个隐藏层；只含 1 个神经元的

输出层，对应短视频流行度的预测值。采用均方误差作为模型的损失函数。为缓解梯度消失及加快收敛速度，选取ReLu作为激活函数。

3　实验和结果分析

3.1　数据采集和预处理

（1）数据采集。为验证模型的有效性，本文以抖音平台的“反诈”短视频为研究对象，创建了一个真实的短视频流行度数据集。作为影响力最大的平台之一，抖音短视频平台积累了大量的用户，创作了丰富多样的短视频内容。近年来，国内网络电信诈骗案件频发，打击电信诈骗成为社会关注的热点事件。本文以“反诈”为关键词进行搜索，针对官方认证的104个反诈抖音号进行数据爬取，数据采集的截止日期为2023年8月10日。考虑到短视频传播的时间效应，2023年7月下旬之后的视频各项数据尚处于波动期，未纳入研究范围。最终，选取2023年7月25日之前发布的历史数据作为数据集，具体信息包括原视频、发布者的相关属性（粉丝数、关注数和作品数）、短视频的流行度信息（点赞数、评论数、收藏数和分享数）、评论者ID（identity）和评论内容。删除缺失值和重复内容后，共计采集到短视频5436条，评论者34.2663万个，评论43.9147万条。

（2）数据预处理。视频与视频之间的关系通过共同评论来定义。删除只评论了一条视频的评论者数据，确保每个评论者都发起了共评关系。数据集中仅被一个用户共同评论的视频对有433063对，占总数的73.5%。考虑到用户评论视频存在偶然性，仅仅被一个用户共评的两个视频未必在内容上具有相关性。因此，去除共同评论次数仅为1的视频对，剩余视频对数量155849，共涉及46个发布者、2890条视频、17773个评论者、2890条发布关系和61996条评论关系。由此构建的部分“用户-视频”异质信息网络如图3所示，网络拓扑性质如表1所示。可以看出，“用户-视频”异质信息网络和发布者网络较为稀疏，但前者具有明显的社区特征；短视频网络平均度较大，稀疏性得到了缓解。

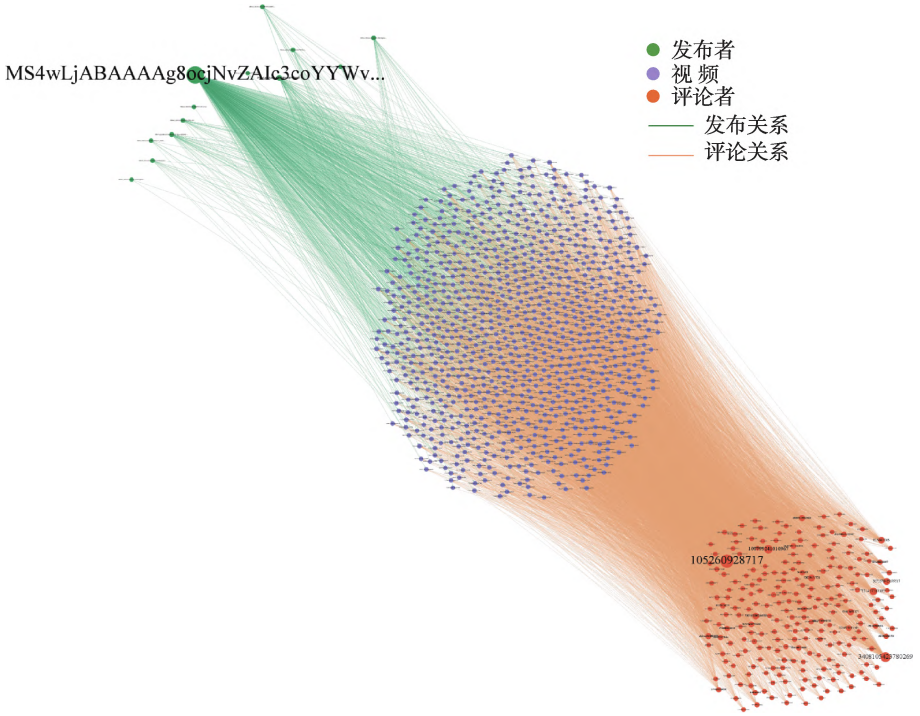

图3　“用户-视频”异质信息网络（部分）（彩图请见https://qbxb.istic.ac.cn）

3.2　实验设置与评价指标

在划分数据集进行实验时，随机选取整个数据集的80%作为训练集，剩余20%作为测试集。模型的训练和测试均在GPU上完成，显卡配置是NVIDIA GeForce RTX 3070，内存16 G，显存8 G。模型框架运行环境具体如下：Python=3.10，PyTorch=2.0.1，NumPy=1.25.2。

表 1 网络拓扑性质

网络名称	网络类型	节点数	连边数	平均度	模块度
“用户-视频”网络	异质有向	20709	64446	3.112	0.581
短视频网络	同质无向	2890	155849	107.854	0.364
发布者网络	同质无向	46	158	6.870	0.054

采用均方误差（mean squared error，MSE）来衡量预测值和真实值之间的差异，即

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{2}$$

MSE越小，模型的预测性能越好。其中，n是测试集短视频样本的个数；y_i是短视频流行度的真实值；$\hat{y}_i$是短视频流行度的预测值。为了提高实验结果的稳定性，重复实验50次，取MSE的平均值作为评价指标。

3.3 对比实验

3.3.1 基线方法

从融合方式、回归模型和现有流行度预测方法3个方面设计对比试验，验证本文方法的有效性。在融合方式上，将早期融合和晚期融合基线方法与低秩多视图子空间学习进行对比；在回归模型上，将支持向量回归、岭回归、LASSO（least absolute shrinkage and selection operator）回归和极限学习机基线方法与MLP回归进行对比；将现有的流行度预测模型TMALL（transductive multi-modal learning）、TLRMVR（transductive low-rank multi-view regression）和MASSL（multi-modal variation auto-encoder regression learning）基线方法与本文方法进行对比。各基线方法阐述如下。

（1）早期融合：水平拼接的方式通常用于简单融合不同向量或具有相同行数但不同列数的矩阵。本文将多模态数据水平拼接，作为MLP回归预测的输入。

（2）晚期融合：先用任务模型对不同模态分别进行训练，再融合模型输出的多个结果。融合方法主要有平均值、贝叶斯规则以及集成学习等。本文采用多元线性回归融合各模态输出结果。

（3）支持向量回归（support vector regression，SVR）[28]：SVR是一种用于回归分析的机器学习方法，通过选择适当的核函数处理复杂数据，以实现精确的回归预测。本文利用高斯径向基函数核对高维的多模态特征进行处理，用于训练一个非线性的SVR模型。

（4）岭回归（ridge regression，RR）[29]：RR是一种有偏估计的回归算法，本质上是对最小二乘估计法的改良。其通过在最小二乘法的基础上增加一个正则化项，有助于防止模型过度拟合，提高模型的稳定性和泛化能力，但本质上是一个线性回归模型。

（5）LASSO回归[30]：LASSO回归全称为最小绝对值收敛和选择算子回归，又称套索算法。LASSO回归兼顾了正则化和特征选择，将一些不重要的回归系数压缩到零，防止过拟合，并实现变量简化。

（6）极限学习机（extreme learning machine，ELM）[31]：ELM是一种单层前馈神经网络算法，其主要特点在于训练过程相对简单且高效，尤其适用于大规模数据集，具有更高的可扩展性和更低的计算复杂性。

（7）TMALL[27]：TMALL是一种基于多视图学习的短视频流行度预测模型，该模型将SM、TM、VM、AM特征映射到一个潜在的公共空间中，以解决短视频因时长短导致的模态信息不充分问题以及模态质量差问题。

（8）TLRMVR[24]：TLRMVR是一个基于低秩多视图子空间学习的短视频流行度回归预测模型，融合SM、TM、VM、AM特征，使用多图正则化最小二乘回归，并引入岭正则化，更好地平衡模型的拟合能力和泛化能力。

（9）MASSL[32]：MASSL是一个基于多模态特征提取和变分自编码器框架的短视频流行度预测回归模型，先采用MLP将SM、TM、VM、AM编码成一个向量表示，再组合卷积神经网络和LSTM（long short-term memory）网络构成解码器网络，输出预测的短视频流行度序列。

3.3.2 融合方式对比

对比实验的回归任务部分均采用MLP回归模型，依据不同方式融合NM、SM、TM、AM和VM这5个模态特征作为回归模型的输入，实验结果如表2所示。可以看出，采用低秩多视图子空间学习这一多模态融合方式的MLP回归模型在流行度预测上误差的均值（MSE）是0.0087，标准差为1.3729×10^{-4}，明显低于早期融合和晚期融合方式，这说明考虑多模态特征之间相关性的融合方式效果更佳。

表 2　不同融合方式对比结果

融合方式	MSE
早期融合	$0.0121\pm8.8457\times10^{-4}$
晚期融合	$0.0142\pm4.1339\times10^{-4}$
低秩多视图子空间学习	$\mathbf{0.0087\pm1.3729\times10^{-4}}$

注：粗体表示最优值。

3.3.3　回归模型对比

回归模型对比实验中，均采用低秩多视图子空间学习融合短视频 NM、SM、TM、AM 和 VM 这 5 个模态特征，将其作为各回归模型的输入，实验结果如表 3 所示。可以看出，MLP 回归模型在流行度预测上的误差最小，其次是 RR、LASSO 回归和 SVR，ELM 模型的预测误差最大，这说明 MLP 回归具有优越性。

表 3　不同回归模型对比结果

回归模型	MSE
SVR	$0.0112\pm8.9999\times10^{-4}$
RR	$0.0095\pm9.0706\times10^{-4}$
LASSO	$0.0098\pm8.9484\times10^{-4}$
ELM	$0.0321\pm7.2641\times10^{-4}$
MLP	$\mathbf{0.0087\pm1.3729\times10^{-4}}$

注：粗体表示最优值。

3.3.4　现有短视频流行度预测模型对比

TMALL、TLRMVR 和 MASSL 均是短视频流行度预测的经典模型，这些模型仅根据短视频的内容特征进行流行度预测，而没有研究短视频网络的结构特征对流行度预测的影响，对比实验结果如表 4 所示。可以看出，本文提出的基于网络表示学习的短视频流行度预测模型具有最小的预测误差，这说明综合考虑短视频网络结构和内容特征的流行度预测方法具有优越性。

表 4　不同流行度预测模型对比结果

预测模型	MSE
TMALL	$0.3107\pm2.8843\times10^{-1}$
TLRMVR	$0.0101\pm9.0320\times10^{-4}$
MASSL	$0.0131\pm9.2184\times10^{-4}$
本文模型	$\mathbf{0.0087\pm1.3729\times10^{-4}}$

注：粗体表示最优值。

3.4　消融实验与影响因素分析

为了进一步验证短视频网络模态的有效性以及各个模态对短视频流行度预测的影响程度，设计消融实验，进行细粒度的影响因素分析。为了简化表示，分别用 N（network）、S（social）、T（text）、A（audio）和 V（visual）表示网络、社交、文本、音频和视觉 5 种模态。表 5 展示了不同模态组合对短视频流行度预测性能的影响，以包含所有模态的特征组合（N+S+T+A+V）为参照组，对其他 5 组分别去除某一模态后的特征组合进行 MLP 回归预测实验。实验结果显示：①组合所有模态特征进行流行度预测的效果最好，说明每个模态特征都是不可或缺的一部分。②当去除网络模态时，S+T+A+V 特征组合让流行度预测的误差变大，说明短视频流行度预测任务中考虑网络模态是非常有必要的。③当去除文本模态时，N+S+A+V 特征组合的误差最大，说明与其他模态特征相比，文本模态对短视频流行度预测的影响程度最大。这是因为用户进入抖音短视频平台后，首先看到的就是视频的标题，其中使用的话题标签和表达方式不仅可以为短视频引流，还可以引发用户的互动行为，从而提高视频的流行度。对数据集内具有较高流行度的短视频进行浏览可以发现，其标题内容较为丰富，带有话题标签并采用了疑问、反问、感叹的句式进行表达，这也为此结论提供了事实依据。④各模态特征对短视频流行度预测的影响程度，由高到低依次为文本模态、网络模态、社交模态、音频模态和视觉模态。

表 5　不同模态组合对比

模态组合	MSE
N+S+T+A+V	$\mathbf{0.0087\pm1.3729\times10^{-4}}$
N+S+T+A	$0.0088\pm9.0937\times10^{-5}$
N+S+T+V	$0.0088\pm1.0678\times10^{-4}$
N+S+A+V	$0.0102\pm1.0440\times10^{-4}$
N+T+A+V	$0.0091\pm9.1055\times10^{-5}$
S+T+A+V	$0.0093\pm1.0689\times10^{-4}$

注：粗体表示最优值。

为了进一步探究短视频和发布者两个子网对流行度预测的贡献度，选取不同的网络作为网络模态，设计消融实验，结果如表 6 所示。①组合短视频和发布者两个子网的网络结构特征作为网络模态进行流行度预测的效果最好，说明两个子网的结构特征在短视频流行度预测任务中都发挥着重要作用；②相比于短视频网络，发布者网络结构特征对减少流行度预测误差的作用更大，说明发布者是影响短视频受欢迎程度的一个较为重要的因素。

表 6 网络模态各子网对比

网络	MSE
短视频网络	$0.0092\pm1.1425\times10^{-4}$
发布者网络	$0.0090\pm1.0093\times10^{-4}$
短视频网络+发布者网络	$0.0087\pm1.3729\times10^{-4}$

4 总结与展望

针对短视频流行度预测中的数据量大、维度高、稀疏性强等挑战性问题，本文提出了基于网络表示学习的短视频流行度预测方法。一方面，通过构建“用户-视频”网络，将不同类型的异质信息融合为整体，有效缓解网络数据的稀疏性问题；另一方面，基于网络表示学习方法表征“用户-视频”网络的结构特征，融合短视频的内容特征，将短视频表示成统一、低维、稠密的向量，提升流行度预测的准确性。与基线方法的对比实验表明，综合考虑短视频网络结构和内容特征的流行度预测方法具有优越性。消融实验进一步证实了短视频流行度预测中考虑网络模态的必要性。该方法不仅突破了当前工作主要基于特征工程预测短视频流行度的研究框架，而且可为短视频平台优化信息管理和内容推荐，相关机构有效监管网络舆论，以及短视频创作者制作高质量视频提供了参考。

本文尚存在一些局限性：短视频流行度预测模型考虑到的网络结构和内容特征都是静态的，没有考虑到影响流行度预测的动态特征，这可能会影响预测的精准性和全面性。未来研究将考虑采集在时间维度上短视频的点赞数、评论数、分享数和收藏数等时序变化数据构建短视频动态特征，研究特征的时间衰减效应，在此基础上，提升流行度预测模型的准确性和实用性。

参 考 文 献

[1] Trzciński T, Andruszkiewicz P, Bocheński T, et al. Recurrent neural networks for online video popularity prediction[C]// Proceedings of the 23rd International Symposium on Methodologies for Intelligent Systems. Cham: Springer, 2017: 146-153.

[2] 武维, 李泽平, 杨华蔚, 等. 融合内容特征和时序信息的深度注意力视频流行度预测模型[J]. 计算机应用, 2021, 41(7): 1878-1884.

[3] Zohourian A, Sajedi H, Yavary A. Popularity prediction of images and videos on Instagram[C]// Proceedings of the 4th International Conference on Web Research. Piscataway: IEEE, 2018: 111-117.

[4] Xie J Y, Zhu Y C, Zhang Z B, et al. A multimodal variational encoder-decoder framework for micro-video popularity prediction [C]// Proceedings of the Web Conference 2020. New York: ACM Press, 2020: 2542-2548.

[5] 井佩光, 叶徐清, 刘昱, 等. 基于双向深度编码网络的短视频流行度预测[J]. 激光与光电子学进展, 2022, 59(8): 300-308.

[6] Hong L J, Dan O, Davison B D. Predicting popular messages in Twitter[C]// Proceedings of the 20th International Conference Companion on World Wide Web. New York: ACM Press, 2011: 57-58.

[7] Ma Z Y, Sun A X, Cong G. On predicting the popularity of newly emerging hashtags in Twitter[J]. Journal of the American Society for Information Science and Technology, 2013, 64(7): 1399-1410.

[8] Meghawat M, Yadav S, Mahata D, et al. A multimodal approach to predict social media popularity[C]// Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval. Piscataway: IEEE, 2018: 190-195.

[9] Shulman B, Sharma A, Cosley D. Predictability of popularity: gaps between prediction and understanding[J]. Proceedings of the Tenth International AAAI Conference on Web and Social Media, 2016, 10(1): 348-357.

[10] Li H T, Ma X Q, Wang F, et al. On popularity prediction of videos shared in online social networks[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2013: 169-178.

[11] Zhang Y C, Li P M, Zhang Z L, et al. GraphInf: a GCN-based popularity prediction system for short video networks[C]// Proceedings of the 27th International Conference on Web Services. Cham: Springer, 2020: 61-76.

[12] 齐金山, 梁循, 李志宇, 等. 大规模复杂信息网络表示学习: 概念、方法与挑战[J]. 计算机学报, 2018, 41(10): 2394-2420.

[13] Wei Y W, Wang X, Nie L Q, et al. MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video[C]// Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM Press, 2019: 1437-1445.

[14] Guo D Y, Hong J S, Luo B L, et al. Multi-modal representation learning for short video understanding and recommendation[C]// Proceedings of the 2019 IEEE International Conference on Multimedia & Expo Workshops. Piscataway: IEEE, 2019: 687-690.

[15] Sang L, Xu M, Qian S S, et al. Context-dependent propagating-based video recommendation in multimodal heterogeneous information networks[J]. IEEE Transactions on Multimedia, 2021, 23: 2019-2032.

[16] 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.

[17] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep

bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[18] Cui Y M, Che W X, Liu T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[19] Eyben F, Wöllmer M, Schuller B. Opensmile: the Munich versatile and fast open-source audio feature extractor[C]// Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press, 2010: 1459-1462.

[20] Eyben F, Scherer K R, Schuller B W, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing[J]. IEEE Transactions on Affective Computing, 2016, 7(2): 190-202.

[21] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]// Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2261-2269.

[22] Chen T, Borth D, Darrell T, et al. DeepSentiBank: visual sentiment concept classification with deep convolutional neural networks[OL]. (2014-10-30) [2023-10-04]. https://arxiv.org/pdf/1410.8586.

[23] 杨瀚森, 樊养余, 吕国云, 等. 基于语义概念的图像情感分析[J]. 西北工业大学学报, 2023, 41(4): 784-793.

[24] Jing P G, Su Y T, Nie L Q, et al. Low-rank multi-view embedding learning for micro-video popularity prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(8): 1519-1532.

[25] Rupnik J, Shawe-Taylor J. Multi-view canonical correlation analysis[C]// Proceedings of the Conference on Data Mining and Data Warehouses. Piscataway: IEEE, 2010: 1-4.

[26] Liu G C, Lin Z C, Yu Y. Robust subspace segmentation by low-rank representation[C]// Proceedings of the 27th International Conference on Machine Learning. Madison: Omnipress, 2010: 663-670.

[27] Chen J Y, Song X M, Nie L Q, et al. Micro tells macro: predicting the popularity of micro-videos via a transductive model[C]// Proceedings of the 24th ACM International Conference on Multimedia. New York: ACM Press, 2016: 898-907.

[28] Smola A J, Schölkopf B. A tutorial on support vector regression [J]. Statistics and Computing, 2004, 14(3): 199-222.

[29] Hoerl A E, Kennard R W. Ridge regression: biased estimation for nonorthogonal problems[J]. Technometrics, 2000, 42(1): 80-86.

[30] Kukreja S L, Löfberg J, Brenner M J. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification[J]. IFAC Proceedings Volumes, 2006, 39(1): 814-819.

[31] Huang G B, Zhou H M, Ding X J, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE Transactions on Systems, Man, and Cybernetics Part B, Cybernetics. Piscataway: IEEE, 2012, 42(2): 513-529.

[32] Zhang Z R, Xu S B, Guo L, et al. Multi-modal variational auto-Encoder model for micro-video popularity prediction[C]// Proceedings of the 8th International Conference on Communication and Information Processing. New York: ACM Press, 2022: 9-16.

（责任编辑　王克平）