黄德伟，王友国，侯浩杰. 基于传播模式聚类与 XGBoost 的社交网络流行度预测算法[J]. 智能计算机与应用，2025，15(5)：21-27. DOI：10.20169/j.issn.2095-2163.250503

基于传播模式聚类与 XGBoost 的社交网络流行度预测算法

黄德伟，王友国，侯浩杰
（南京邮电大学 理学院，南京 210023）

摘 要：对以微博为代表的社交媒体进行流行度预测具有重要价值，受到了广泛的关注。本文考虑信息传播模式之间的异质性，提出了基于传播模式聚类与 XGBoost 的微博流行度预测算法。在数据预处理阶段，使用 K-means++聚类算法，依据早期转发窗口内等时间间隔的转发增量序列，对微博传播模式进行聚类，得到了各个传播模式下的训练子集。在特征提取阶段，提取了相邻转发之间的平均时间间隔、微博首发到第一次转发的时间间隔、等时间间隔的流行度累计序列和流行度增量序列，作为微博数据的时序特征；提取了首发用户的一阶邻居节点数、微博转发的叶子节点数、观察窗口内的流行度和转发路径的平均深度，作为微博数据的结构特征。将这 2 类特征串联融合得到样本的特征。在离线训练阶段，采用 XGBoost 集成学习的框架进行回归学习，在不同子集上得到微博流行度预测模型。最后在新浪微博转发数据集上进行实验，验证了本文算法在 *MSLE* 和 *mSLE* 指标上的有效性。

关键词：流行度预测；信息传播模式；K-means++聚类；特征提取；XGBoost 算法
中图分类号：TP391 **文献标志码**：A **文章编号**：2095-2163(2025)05-0021-07

Social network popularity prediction algorithm based on propagation model clustering and XGBoost

HUANG Dewei, WANG Youguo, HOU Haojie
(School of Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Predicting the popularity of social media, with Weibo as a representative, holds significant value and has garnered widespread attention. This paper addresses the heterogeneity among information propagation patterns and proposes a Weibo popularity prediction algorithm based on propagation model clustering and XGBoost. In the data preprocessing stage, the K-means++ clustering algorithm is employed to cluster Weibo propagation patterns based on the increment sequence of retweets with equal time intervals within the early retweet window. This results in training subsets for different propagation patterns. In the feature extraction stage, temporal features for Weibo data are extracted, including the average time interval between adjacent retweets, the time interval from the initial post to the first retweet, popularity accumulation sequences at equal time intervals, and popularity increment sequences. Additionally, structural features for Weibo data are extracted, such as the first-order neighbor count of the initial user, the leaf node count of Weibo retweets, the popularity within the observation window, and the average depth of retweet paths. These two categories of features are concatenated to create the sample's feature set. During the offline training phase, regression learning is carried out using the XGBoost ensemble learning framework to obtain Weibo popularity prediction models on different subsets. Finally, experiments are conducted on a Sina Weibo retweet dataset to validate the effectiveness of this algorithm in terms of the *MSLE* and *mSLE* metrics.

Key words: popularity prediction; information propagation patterns; K-means++ clustering; feature extraction; XGBoost algorithm

0 引 言

随着互联网普及率的逐年提升，社交网络(Social Network)日益发达，信息可以迅速地在社交网络上得到广泛传播。社交网络让全球用户沟通互联，深刻地改变了信息传播方式并加速了信息的生

基金项目：国家自然科学基金(62071248)。

作者简介：黄德伟(1999—)，男，硕士研究生，主要研究方向：社交网络，机器学习。
通信作者：王友国(1968—)，男，博士，教授，博士生导师，主要研究方向：信号与信息处理，随机共振理论与研究，信息理论及应用，在线社交网络等。Email：wangyg@njupt.edu.cn。

收稿日期：2023-10-31

成与交互,产生了巨大的社会影响。当前,全球较为流行的社交网络平台包括 Facebook、Twitter 以及新浪微博等。新浪微博是国内一个影响力较广的主流社交媒体,微博 2022 年 9 月的日均活跃用户数为 2.53 亿。对微博的流行度预测问题进行研究,有助于计算信息未来的热度、发现热点话题和探究信息传播的规律,进而广泛应用于信息检索、舆情研判和企业营销等领域[1]。

现有的社交网络上的信息流行度预测方法主要包括:基于生成式点过程的方法、基于特征提取的方法和基于深度学习的方法等。

基于生成式点过程的预测方法是一种从微观角度对信息传播进行建模的方法,建立在随机过程的基础上,通常使用泊松过程或霍克斯过程对信息的传播过程进行建模,着力于描述信息传播的 3 个关键因素:自身影响力、时间衰减效应、富者愈富机制。Shen 等学者[2]提出使用基于增强泊松过程(Reinforced Poisson Process, RPP)的概率模型,来预测信息未来的累积转发数。Bao 等学者[3]提出基于自激霍克斯过程(Self-Excited Hawkes Process, SEHP)的概率模型,显式地对每次转发的贡献进行了建模。高金华等学者[4]提出一种叠加的基于增强泊松过程的方法,刻画了信息“去中心化”的特点。

基于特征提取的预测方法是一种从宏观角度对信息传播过程进行建模的方法。方法遵循特征工程的基本流程,通常需要领域专家针对特定平台进行考察,结合相关领域的专业知识,人工构造和提取特征,然后将流行度预测的问题归结为一个回归或者分类任务,使用提取到的特征对问题进行建模,从而预测未来流行度。已有的研究发现,信息的未来流行度与社交网络的用户、传播的结构、内容以及时序特征息息相关。Szabo 等学者[5]对 Digg 和 YouTube 平台上在线内容进行研究,发现其最终流行度与早期流行度呈现对数线性相关性,使用线性回归方法构建 SH 模型预测在线内容的流行度。Bakshy 等学者[6]对 Twitter 上的信息传播进行分析,发现用户的粉丝数是影响流行度的重要因素。Bao 等学者[7]从转发网络中提取出链路密度和扩散深度两种结构特征,建立了线性回归模型来预测信息流行度。Cheng 等学者[8]发现,仅使用时序特征的预测模型与在增加内容特征、用户特征及结构特征后的预测模型预测性能几乎相当。任敏捷等学者[9]将流行度预测问题转换为互动值档位分类问题,提取出博文特征、话题特征和用户特征,使用 XGBoost 算法对微博流行度情况进行预测。基于特征提取的方法具有较好的可解释性,其预测性能主要取决于特征,这些特征大多是人工定义或构造的,需要投入领域专家大量的时间精力。经过实证的特征可以成为研究者们进行信息传播规律探究与流行度预测的重要参考。

随着深度学习在图像、语音等各领域的成功应用,研究者们开始考虑将深度学习算法应用于流行度预测[10]。Li 等学者[11]将深度学习引入到流行度预测领域,提出了 DeepCas 模型。利用循环神经网络对每条游走序列分别建模,通过注意力机制对不同游走序列赋以不同的权重进行加权聚合,得到参与用户子图的表示。Cao 等学者[12]将霍克斯过程与深度学习结合,提出 DeepHawkes 模型,融入用户影响力,自激机制和时间衰减效应三种可解释因素进行信息级联预测,在保持高解释性的同时提高了预测的准确性。

本文利用基于特征提取的方法进行研究。微博的传播模式具有异质性,针对不同模式下的样本进行训练,可以提高子集内样本的相似性,从而提高预测性能。在现有的特征提取工作中,大多考虑单独的特征,这样只能从某个单独的方面对样本进行描述。从多角度提取训练样本特征,并对特征进行融合,能够更加充分地描述训练样本的信息,提高算法性能。此外,已有算法大多使用单独的算法,可能会导致离线训练精度不高,而使用集成学习算法能够克服单个学习方式的不足。

本文提出了基于传播模式聚类与 XGBoost[13-15]的流行度预测算法。本文工作如下:

首先,针对微博数据集,提出了使用 Kmeans++ 算法对训练样本进行预处理,使用微博早期转发时间窗口内等时间间隔的转发增量序列作为聚类的特征。通过数据预处理得到不同模式下的训练样本子集。其次,在特征提取方面,本文从考察微博传播的时间演化动态和社交网络结构的角度出发,分别提取了时序特征和结构特征。接着,为了对特征进行充分的利用,将 2 类特征进行串联融合作为训练的特征。最后,在充分提取特征的基础上,为了提高离线学习的系统性能,本文使用了 XGBoost 这一集成学习框架,分别对不同训练样本进行训练。在新浪微博转发数据集上的实验表明,本文所提出的方法在性能上较优。

1 算法框架

将微博作为考察对象,微博未来流行度增量预

测问题定义为：记原始微博 i 发布时刻为 t_0，通过观测微博 i 在发布后 $[t_0, t_0+T)$ 时间内的转发情况，将 $[t_0, t_0+T)$ 观测到的流行度记作 P_T^i，微博最终的流行度定义为 P_∞^i。微博未来流行度增量预测问题就是，利用特定的模型预测微博 i 的流行度增量 ΔP_T^i，即：

$$\Delta P_T^i = P_\infty^i - P_T^i \tag{1}$$

本文的算法流程如图 1 所示。

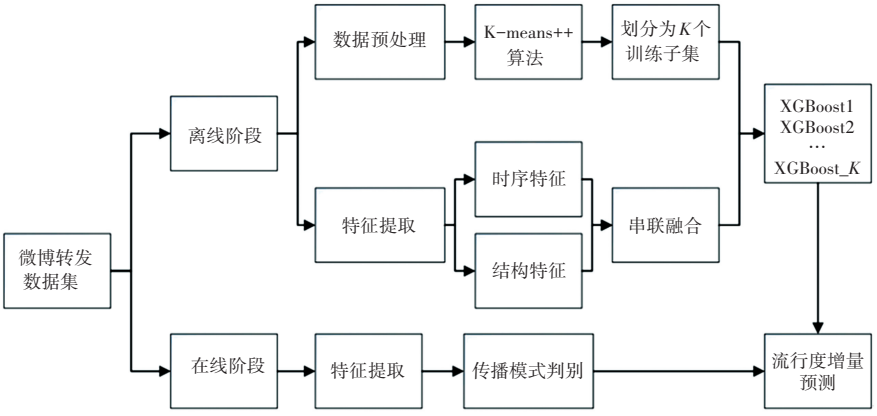

图 1　微博流行度预测算法流程图

Fig. 1　Flowchart of Weibo popularity prediction algorithm

本文提出的流行度预测算法在离线阶段分为 3 个主要部分：数据预处理、特征提取、基于 XGBoost 的离线训练，下面对各个部分进行研究论述。

1.1　数据预处理

在数据预处理阶段，本文使用传播数据的时间序列特征，利用 K-means++算法对微博的传播模式进行聚类。

时间序列 $(x_1, x_2, \cdots, x_n)$ 通常被用来将信息的传播数据转化为定长的向量化表示，其中 n 表示划分的等长时间间隔的个数，x_i 表示在第 i 个时间间隔内流行度的增长量，利用这样的表示方式反映信息的流行度变化趋势[16]。本文使用观察时间窗口内间隔固定时间的转发增量序列，作为传播模式聚类的特征。

在传播模式聚类使用的算法选择上，本文使用 K-means++作为聚类的方法。David 等学者[17]对初始化簇中心的方法做出了改进，提出了 K-means++算法。K-means++是一种高效可靠的聚类方法，计算复杂度较低，适用于大规模数据聚类。K-means++算法在初始化簇中心时的步骤具体如下：

(1)从数据集 $\mathcal{X}$ 中随机选取一个样本点作为第一个初始聚类中心 c_i。

(2)计算每个样本与当前已有聚类中心之间的最短距离，用 $D(x)$ 表示，计算每个样本点选为下一个聚类中心的概率 $P(x)$，最后选择最大概率值对应的样本点作为一个簇中心，概率 $P(x)$ 的计算公式如下：

$$P(x) = \frac{D(x)^2}{\sum_{x\in\mathcal{X}} D(x)^2} \tag{2}$$

(3)重复第(2)步，直到选择出 K 个聚类中心。

1.2　特征提取

参考已有的基于特征的社交网络流行度预测研究工作[18]，从微博的传播级联数据中构造和提取特征，这些特征已经得到广泛验证，能够较好地反映传播级联的时序演化情况和拓扑结构。

(1)时序特征 $feature_t$。为捕捉微博传播随着时间变化的动态，本文选取如下的特征作为微博转发数据的时序特征。

① 转发的平均时间间隔 $feature_{t1}$[18]。设时间窗口内微博共有 N 次转发行为，第 j 次转发距离微博首发的时间为 $t_j(\ j=1,2,\cdots,N)$，则平均时间间隔可以表示为：

$$feature_{t1} = \frac{1}{N}\sum_{j=1}^{N} t_j,\ j=1,2,\cdots,N \tag{3}$$

② 始发到第一次转发的时间间隔 $feature_{t2}$[18]。N 次转发中距离微博首发时间中最小的时间间隔，即：

$$feature_{t2} = \min\{t_j\},\ j=1,2,\cdots,N \tag{4}$$

③ 等时间间隔的累积转发序列 $feature_{t3}$[19]。在观察时间窗口分别为 1 h、2 h 和 3 h 的情况下，将时间间隔分别设置为 200 s、400 s 和 600 s。假设各个时间间隔内的转发量为 $x_j(j=1,2,\cdots,n)$，计算公式为：

$$feature_{t3} = \{x_1, x_1 + x_2, \cdots, \sum_{j=1}^{n} x_j\} \quad (5)$$

④ 等时间间隔的转发增量序列 $feature_{t4}$[19]。表示观察窗口内间隔一定时间的转发增量序列。在观察时间窗口分别为 1 h、2 h 和 3 h 的情况下,时间间隔分别设置为 200 s、400 s 和 600 s。继而推得:

$$feature_{t4} = \{x_1, x_2, \cdots, x_n\} \quad (6)$$

因此,将以上 4 类时序特征串联,总的时序特征 $feature_t$ 表示为:

$$feature_t = feature_{t1} \oplus feature_{t2} \oplus feature_{t3} \oplus feature_{t4} \quad (7)$$

(2)结构特征 $feature_s$。为获取微博数据中的社交网络结构,本文提取了如下的微博转发数据结构特征。

① 一阶邻居节点数 $feature_{s1}$[20]。表示原始微博发送者的一阶邻居节点数。假设原始微博发送用户 *user_start*,在观察窗口内共有 N 个用户,设各节点到 *user* 的距离为 d_j,一阶邻居节点数即到 *user_start* 的距离为 1 的用户节点个数,即:

$$feature_{s1} = count\{j \mid d_j = 1\},\ j = 1,2,\cdots,N \quad (8)$$

② 叶子节点数 $feature_{s2}$[20]。表示观察时间窗口内传播路径中的叶子节点数。设各个节点的出度为 $degree_{out_j}(j = 1,2,\cdots,N)$,传播路径的叶子节点数即出度为 0 的节点数,推得的公式为:

$$feature_{s2} = count\{j \mid degree_{out_j} = 0\},\ j = 1,2,\cdots,N \quad (9)$$

③ 观察时间窗口内的转发量 $feature_{s3}$[5]。可由如下公式计算求得:

$$feature_{s3} = P_T^i = N \quad (10)$$

④ 传播路径的平均深度 $feature_{s4}$[7]。假设微博在观察时间窗口 T 内共有 l 条传播路径,第 j 条传播路径的深度为 $depth_j(j = 1,2,\cdots,l)$,平均深度表示为:

$$feature_{s4} = \frac{1}{l}\sum_{j=1}^{l} depth_j \quad (11)$$

总的结构特征 $Feature_s$ 表示为:

$$feature_s = feature_{s1} \oplus feature_{s2} \oplus feature_{s3} \oplus feature_{s4} \quad (12)$$

(3)特征融合。提取出上面 2 类特征之后,将其进行串联融合得到总的特征,即:

$$features = feature_t \oplus feature_s \quad (13)$$

1.3 XGBoost 训练

除了有效特征的挖掘,选择恰当的模型也有助于提高未来流行度增量预测的性能。极端梯度提升决策树(eXtreme Gradient Boosting, XGBoost)本质上属于梯度提升决策树算法(Gradient Boosting Decision Tree, GBDT),但在算法精度、速度和泛化能力上均要优于传统的 GBDT 算法。从算法精度上看,XGBoost 通过将损失函数展开到二阶导数,使得其能更好逼近真实损失;从算法速度上看,XGBoost 使用了加权分位数 *sketch* 和稀疏感知算法这 2 个技巧,通过缓存优化和模型并行来提高模型速度;从模型泛化能力上来看,通过对损失函数加入正则化项、加性模型中设置缩减率和列抽样等方法,来防止模型过拟合。XGBoost 通过迭代生成一棵棵树,将多个性能较差的弱学习器集成为一个强学习器,在 GBDT 的基础上将速度和效率都发挥到了极致[21-22]。

XGBoost 可以表示为:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (14)$$

根据前向分步算法,根据第 t 次迭代的基模型为 $f_t(x)$,有:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} \hat{y}_i^{(t-1)} + f_t(x_i) \quad (15)$$

XGBoost 的损失函数基本形式由检验损失项和正则化项构成,具体公式如下:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \quad (16)$$

其中,$l(y_i, \hat{y}_i)$ 为经验损失项,表示训练数据预测值与真实值之间的损失;$\sum_{i=1}^{t} \Omega(f_i)$ 为正则化项,表示全部 t 棵树的复杂度之和,用于防止模型过拟合。

根据前向分步算法,以第 t 步模型为例,假设模型对第 i 个样本 x_i 的预测值为:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (17)$$

其中,$\hat{y}_i^{(t-1)}$ 是由第 $t-1$ 步的模型给出的预测值,且作为一个已知常量存在,$f_t(x_i)$ 为第 t 步树模型的预测值。因而目标函数可以改写为:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_i) + C \quad (18)$$

针对式(18)前半部分,使用二阶泰勒展开式,相应的损失函数经验损失项可以改写为:

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) = l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \quad (19)$$

其中,g_i 表示损失函数一阶导数,h_i 表示损失函数二阶导数。将式(19)的损失函数二阶泰勒展开式带入公式计算,可得损失函数的近似表达式:

$$L^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (20)$$

本文在挖掘微博信息时间特征与结构特征的基础上，使用 XGBoost 算法对微博信息中提取的特征进行训练并预测未来流行度。

1.4　在线阶段

在离线阶段，利用 K-means++方法，使用聚类特征对训练样本进行聚类，得到 K 个聚类中心，将样本划分为 K 个样本子集。在这 K 个样本子集上分别训练 XGBoost 得到 K 个模型用于预测。在线阶段的步骤如下：

(1)从在线阶段的样本中同样提取出用于判别的特征。

(2)计算各个样本到各聚类中心的欧氏距离，并将样本判别为欧氏距离最小的一类。设第 i 个样本的特征为 $(f_{i1}, f_{i2}, \cdots, f_{im})$，第 j 个聚类中为 $(f_{j1}, f_{j2}, \cdots, f_{jm})$，则第 i 个样本到第 j 个聚类中的欧氏距离为：

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (f_{ik} - f_{jk})^2} \quad (21)$$

(3)使用该类别训练出的 XGBoost 模型进行微博未来流行度的预测。

2　实验及结果分析

2.1　实验参数

本文实验平台使用 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30 GHz，运行内存 16 GB。实验使用的 Python 版本为 3.9.7。

本文在新浪微博转发数据集[12]上进行实验，该数据集收集了 2016 年 6 月 1 日新浪微博平台上所有的原创微博，并收集了这些原创微博发送 24 h 内的所有转发微博。

由于微博上用户的转发活动具有明显的周期性[13]，对每个小时的原始微博发布量进行统计，不同时间段的用户活跃度具有较大的差异。为避免这样的影响，考虑发布时间在 8:00 到 18:00 之间的原始微博，将这一时段的微博及其转发微博作为考察对象，从而每条原始微博在发布后有 6 h 的活跃期，进行充分的转发。

实验中分别选取微博转发观察时间窗口 T 为 1 h、2 h 和 3 h。为了防止较为极端的样本对最终的预测结果产生影响，本文选取在观察时间窗口内的转发量在 10~1 000 之间的原始微博。经过过滤，将数据集按照 85%、15%的比例划分，分别用于算法的离线阶段和在线阶段。数据集的基本描述见表 1。

表 1　经过筛选的微博数据集基本描述

Table 1　Basic description of the filtered Weibo dataset

观察时间窗口/h	离线阶段	在线阶段
1	35 859	6 327
2	42 989	7 586
3	46 842	8 266

XGBoost 训练时的主要参数设置如下：objective 为 reg:squarederror，eval_metric 为 rmse，*gamma* 为 0.1，*learning_rate* 为 0.06，*min_child_weight* 为 3，*max_depth* 为 7。

2.2　评价指标

在本文的实验中，使用均方对数误差(Mean Square Log-transformed Error，*MSLE*)以及对数平方误差的中位数(median of Square Log-transformed Error，*mSLE*)来衡量模型的性能。

将预测的未来流行度增量记作 $\Delta\hat{P}_T^i$，观察到的真实流行度增量为 ΔP_T^i，*MSLE* 用于衡量所有传播级联上预测结果和真实流行度之间的平均误差，*MSLE* 的定义公式如下：

$$MSLE = \frac{1}{N} \sum_{i=1}^{N} (\log \Delta\hat{P}_T^i - \log \Delta P_T^i)^2 \quad (22)$$

mSLE 的定义如下式所示：

$$mSLE = median(\{(\log \Delta\hat{P}_T^i - \log \Delta P_T^i)^2\}, i = 1, 2, \cdots, N) \quad (23)$$

2.3　算法性能描述

2.3.1　基于聚类指标的 K 值初步估计

对于 K-means++算法而言，聚类数 K 是一个重要的超参数。关于聚类数的选择，通过考察聚类指标 *SSE* 的变化情况，对 K 值进行初步的估计。在观察时间窗口分别为 1 h、2 h 和 3 h 的情况下，实验发现，当聚类数设置为 60 及以上时，超出了样本所能划分的最多的聚类数。将聚类数设置为 2~60，绘制出在 3 个时间观察窗口下，各类样本到各自聚类中心的误差平方和(*SSE*)随着聚类数 K 的变化情况，如图 2 所示。

由图 2 可以看到，随着聚类数 K 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，误差平方和 *SSE* 逐渐变小。当 K 小于真实聚类数时，由于 K 的增大会大幅增加每个簇的聚合程度，故 *SSE* 的下降幅度会很大；当 K 到达真实聚类数时，再增加 K 所得到的聚合程度结果值会迅速变小，所以 *SSE* 的下降幅度会骤减，然后随着 K 值的继续增大而趋于平缓。综上可知，K 的选取范围应该在 10~30 之间。

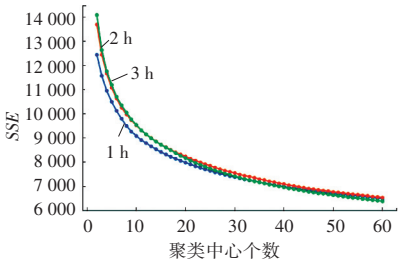

图 2 1 h、2 h、3 h 时间窗口下误差平方和随聚类数 *K* 变化图

Fig. 2 The sum of squared errors changes with the number of clusters *K* in the 1 h, 2 h, and 3 h time windows

2.3.2 基于预测性能指标的精确估计

基于 2.3.1 节对 *K* 范围的初步估计,在不同传播模式聚类数 *K* 下进行预测实验。每种参数 *K* 下的 *MSLE* 和 *mSLE* 整体变化情况,如图 3~图 5 所示。

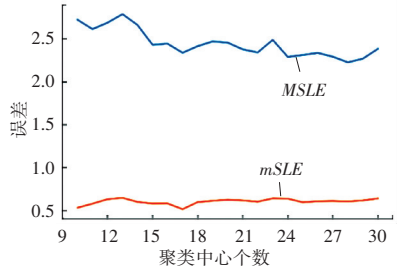

图 3 1 h 时间窗口下不同传播模式聚类数的预测效果

Fig. 3 Prediction effect of cluster number of different propagation patterns in 1 h time window

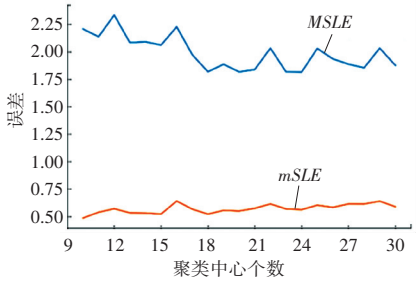

图 4 2 h 时间窗口下不同传播模式聚类数的预测效果

Fig. 4 Prediction effect of cluster number of different propagation patterns in 2 h time window

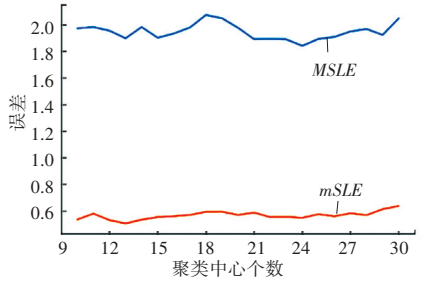

图 5 3 h 时间窗口下不同传播模式聚类数的预测效果

Fig. 5 Prediction effect of cluster number of different propagation patterns in 3 h time window

分析可知,在 1 h、2 h 时间窗口下的预测效果图中,*MSLE* 整体都呈现降低的趋势,而 3 h 时间窗口下、*MSLE* 都较小且比较稳定。3 种时间窗口下 *mSLE* 都保持在一个较小的水平,但是随着 *K* 的增大,趋势变化不明显,聚类数 *K* 对 *mSLE* 影响较小。

对数值进行比较,发现在观察窗口为 1 h 的情况下,聚类数设置为 28 比 24 的效果略好;在观察窗口为 2 h 的情况下,聚类数设置为 20 与 24 的效果相近,两者在 2 项性能上各有优势;在观察窗口为 3 h 的情况下,将聚类数设置为 24,在 2 项指标上都达到较好的性能。综合考虑,本文将传播模式聚类数设置为 24。

2.4 实验结果

为了说明本文所提方法在性能上的优越性,选择了主流的流行度预测方法,在新浪微博数据集上进行实验性能的对比。

(1)线性回归。将 1.2 节提取的特征,输入到线性回归模型中,对微博未来流行度增量进行预测。

(2)多层感知机。将 1.2 节提取的特征,输入到多层感知机中,对微博未来流行度增量进行预测。

(3)决策树。将 1.2 节提取的特征,输入到 XGBoost 的基学习器决策树中。

(4)XGBoost-t。仅使用提取的时序特征,不经过样本聚类,输入到 XGBoost 进行训练。

(5)XGBoost-s。仅使用提取的结构特征,不经过样本聚类,输入到 XGBoost 进行训练。

(6)XGBoost。将 1.2 节提取的特征,不经过样本聚类,直接输入到 XGBoost 进行训练。

在新浪微博数据集上进行各类方法的实验结果见表 2。

相较于 XGBoost-t 和 XGBoost-s 这 2 个只使用单独特征输入到 XGBoost 中的算法,除了在 1 h 下的 *mSLE* 指标,将 2 类特征串联融合后输入 XGBoost 的模型在各项性能上均达到最优,并且 1 h 下的 *mSLE* 指标结果相差不大,说明了特征融合的有效性。

将不同的算法之间进行比较,除了在 1 h 时间窗口下的 *mSLE*,2 h 时间窗口下的 *MSLE* 和 *mSLE* 弱于多层感知机。说明了使用 XGBoost 这一集成学习算法的有效性。

在新浪微博转发数据集上的实验表明,在观察窗口分别为 1 h、2 h 和 3 h 的情况下,与对比方法中的最优性能相比,在 *MSLE* 和 *mSLE* 这 2 项指标上都有显著的提升。不同的实验结果对比说明了使用聚类划分子集、特征融合以及使用 XGBoost 集成学习算法这些策略的有效性。

表 2　新浪微博数据集上各类方法预测性能

Table 2　Prediction performance of various methods on the Sina Weibo dataset

方法	1 h		2 h		3 h	
	MSLE	*mSLE*	*MSLE*	*mSLE*	*MSLE*	*mSLE*
线性回归	3.926	1.606	2.968	1.162	2.785	0.975
多层感知机	3.360	1.226	2.561	0.871	2.185	0.795
决策树	3.633	1.345	2.834	1.025	2.305	0.798
XGBoost-t	3.310	1.311	2.763	0.987	2.062	0.686
XGBoost-s	3.328	1.340	2.804	0.952	2.221	0.773
XGBoost	3.282	1.314	2.663	0.967	2.047	0.651
本文	**2.290**	**0.639**	**1.815**	**0.562**	**1.842**	**0.550**

3　结束语

本文考虑信息传播方式之间存在异质性，提出了基于传播模式聚类与 XGBoost 的算法，从传播数据中提取出结构特征与时序特征，进行微博流行度增量的预测。在新浪微博转发数据集上，与主流预测方法相比，在 *MSLE* 和 *mSLE* 两项评价指标上均有性能的提升。在观察窗口分别为 1 h、2 h 和 3 h 的情况下，与对比方法相比，性能均为最优。

虽然该模型整体预测表现较好，但仍存在一些局限。使用观察时间窗口内间隔固定时间的转发增量序列作为聚类的特征，传播模式的聚类效果会受到时间粒度的影响。此外，本文提取的特征都是在已有工作的基础上选择的，不能完全穷尽所有可能会影响未来流行度的特征。未来的工作会考虑使用深度学习方法，自动地从微博传播数据中学习特征用于预测。

参考文献

[1] 吴越，陈晓亮，蒋忠远. 微博信息流行度预测研究综述[J]. 西华大学学报(自然科学版)，2017，36(1)：1-6.

[2] SHEN Huawei, WANG Dashun, SONG Chaoming, et al. Modeling and predicting popularity dynamics via reinforced poisson processes[J]. arXiv preprint arXiv，1401.0778，2014.

[3] BAO Peng, SHEN Huawei, JIN Xiaolong, et al. Modeling and predicting popularity dynamics of microblogs using self-excited hawkes processes[J]. arXiv preprint arXiv，1503.02754，2015.

[4] 高金华，刘悦，程学旗. 去中心化的微博传播动力学建模[J]. 中国科学：信息科学，2018，48(11)：1575-1588.

[5] SZABO G, HUBERMAN B A. Predicting the popularity of online content[J]. Communications of the ACM，2010，53(8)：80-88.

[6] BAKSHY E, HOFMAN J M, MASON W A, et al. Everyone's an influencer: Quantifying influence on Twitter[C]//Proceedings of the 4th International Conference on Web Search and Web Data Mining. New York：ACM, 2011：65-74.

[7] BAO Peng, SHEN Huawei, HUANG Junming, et al. Popularity prediction in microblogging network: A case study on Sina Weibo [C]// Proceedings of the 22nd International World Wide Web Conference. New York：ACM，2013：177-178.

[8] CHENG J, ADAMIC L A, DOW P A, et al. Can cascades be predicted? [J]. arXiv preprint arXiv, 1403.4608，2014.

[9] 任敏捷，靳国庆，王晓雯，等. 基于 XGBoost 的微博流行度预测算法[J]. 数据采集与处理，2022，37(2)：383-395.

[10] 曹骞，沈华伟，高金华，等. 基于深度学习的流行度预测研究综述[J]. 中文信息学报，2021，35(2)：1-18.

[11] LI Cheng, MA Jiaqi, GUO Xiaoxiao, et al. DeepCas: An end-to-end predictor of information cascades[J]. arXiv preprint arXiv, 1611.05373，2016.

[12] CAO Qi, SHEN Huawei, CEN K, et al. DeepHawkes: Bridging the gap between prediction and understanding of information cascades[C]//Conference on Information and Knowledge Management. New York：ACM, 2017：1149-1158.

[13] 朱海龙，云晓春，韩志帅. 基于传播加速度的微博流行度预测方法[J]. 计算机研究与发展，2018，55(6)：1282-1293.

[14] 王巍，李锐光，周渊，等. 基于用户与节点规模的微博突发话题传播预测算法[J]. 通信学报，2013，34(S1)：84-91.

[15] CHEN Tianqi, GUESTRIN C. XGBoost: A scalable tree boosting system[J]. arXiv preprint arXiv，1603.02754，2016.

[16] 高金华，沈华伟，程学旗，等. 基于相似消息的流行度预测方法[J]. 中文信息学报，2018，32(11)：79-85.

[17] DAVID A, VASSILVITSKII S. K-means++: The advantages of careful seeding[C]//Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete algorithms (SODA'07). New York：ACM, 2007：1027-1035.

[18] SHULMAN B, SHARMA A, COSLEY D. Predictability of popularity: Gaps between prediction and understanding[J]. arXiv preprint arXiv, 1603.09436，2016.

[19] PINTO H, ALMEIDA J M, GONÇALVES M A. Using early view patterns to predict the popularity of YouTube videos[C]// Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York: ACM, 2013：365-374.

[20] JOHAN U, LARS B, CAMERON M, et al. Structural diversity in social contagion[J]. Proceedings of the National Academy of Sciences of the United States of America，2012，109(16)：5962-5966.

[21] MARTIN T, HOFMAN M J, SHARMA A, et al. Exploring limits to prediction in complex social systems[J]. arXiv preprint arXiv，1602.01013，2016.

[22] 鲁伟. 机器学习：公式推导与代码实现[M]. 北京：人民邮电出版社，2022.