ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 2,Mar. 2022,pp. 383－395
DOI:10. 16337/j. 1004-9037. 2022. 02. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于XGBoost的微博流行度预测算法

任敏捷[1,2]，靳国庆[1]，王晓雯[2]，陈睿东[2]，袁运新[2]，聂为之[2]，刘安安[2]

(1. 人民网传播内容认知国家重点实验室,北京 100733; 2. 天津大学电气自动化与信息工程学院,天津 300072)

摘　要：随着全媒体时代的到来和社交网络的发展,流行度预测在舆情监测和数据话语权的争夺上开始发挥重要的作用。现有的流行度预测研究多集中于外文媒体,对以微博为代表的国内主流媒体进行流行度预测是一个新兴且具有挑战的方向。本文针对微博这一国内社交媒体平台进行研究,通过对微博内容及微博用户的特征分析,设计了多种流行度预测方案,同时,提出了一种基于XGBoost的微博流行度预测算法,将流行度预测问题转换为互动值档位分类问题,在分类式框架下将提取融合后的特征用于模型训练,可以较为准确地对有用户信息的微博的流行度情况进行预测。本文的算法在微博流行度预测数据集中得到验证,并且取得了准确率高达85.69％的优越效果。

关键词：社交媒体预测;XGBoost;特征提取;特征融合;微博流行度

中图分类号：TP391　　**文献标志码**:A

Microblog Popularity Prediction Algorithm Based on XGBoost

REN Minjie[1,2], JIN Guoqing[1], WANG Xiaowen[2], CHEN Ruidong[2], YUAN Yunxin[2], NIE Weizhi[2], LIU An'an[2]

(1. State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing 100733, China; 2. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: With the advent of the all-media era and the development of social networks, the popularity prediction begins to play an important role in the monitoring of public opinion and the competition of data discourse power. The existing popularity prediction researches mostly focuse on foreign media, and it is an emerging and challenging direction to predict the popularity of domestic mainstream media such as microblog. In this paper, we conduct the research on microblog, a domestic social media platform, through the analysis of microblog's content and users, and design a variety of popularity prediction schemes. Meanwhile, we propose a microblog popularity prediction algorithm based on XGBoost, which converts the popluarity prediction problem into an interactive value file classification problem, and use the extracted and fused features for model training under the categorical framework, which can predict the popularity of microblog with user information more accurately. The proposed algorithm is verified in the microblog popularity prediction dataset, whose accuracy rate can achieve as high as 85.69％.

Key words: social media prediction; XGBoost; feature extraction; feature fusion; microblog popularity

---

基金项目:传播内容认知国家重点实验室开放基金(20K04)。
收稿日期:2021-03-01;**修订日期**:2022-03-01

引 言

随着互联网的普及和媒体融合建设的推进,主流社交媒体的流行度预测是全媒体时代下备受瞩目的研究课题[1],可以广泛应用于舆情监测和数据话语权争夺的领域中,具有相当可观的现实意义。在我国,微博是一个影响力较广的主流社交媒体,对微博流行度预测问题进行研究有助于计算信息未来的热度、发现热点话题和提取信息传播的规律,进而广泛应用于信息检索、舆情研判和企业营销等领域[2]。

流行度预测指的是对由用户发布的信息未来所获得的关注程度进行预测[3]。而流行度的定义往往取决于社交媒体的平台,不同的网络平台有不同的数值指标度量。当前许多研究仅使用单一评价指标,例如,Pinto 等[4]将流行度定义为 YouTube 上在线视频的浏览数,提出通过训练多元线性模型(Multivariate linear model, ML Model)和多元径向基模型(Multivariate radial basis functions model, MRBF Model)来预测视频的未来指定时刻的浏览数;孔庆超等[5]基于动态演化的论坛的讨论帖展开流行度预测,认为相较于帖子的浏览数,将流行度的度量定义为讨论帖的评论数更加能够反映用户的关注情况。Hong 等[6]将给定时刻 Twitter 的转发数作为 Twitter 的流行度,Gao 等[7]同样将转发数作为 Twitter 和微博的流行度度量,但这种度量没有将微博的评论数和点赞数考虑在内,对受欢迎程度的指标范围覆盖不够全面,因此,为了使流行度的评价指标更具有代表性和普遍性,本文同时将微博的转发数、评论数、点赞数和三者之和定义为互动值来作为微博流行度度量的标准。

目前社交媒体流行度预测的主流方法是基于特征的模型预测,即先进行有效特征的挖掘,再进行模型的构造用以训练学习,最后得到流行度的各项指标。有效特征的挖掘立足于社交媒体平台信息特点的分析,Wu 等[8]在研究社交媒体流行度时针对 Flickr 平台进行了考察,认为 Flickr 平台上照片和帖子的时空信息对于最后流行度的影响十分重要。Mazloom 等[9]针对 Instagram 上的帖子进行研究,发现其帖子的分类特征对流行度的准确预测大有益处。Vilares 等[10]在研究 Twitter 上的信息时关注更多的是文本特征,流行度预测基于 Twitter 信息的词汇和句法处理。这些方法都立足于所研究社交媒体的特点,表明特征的提取依赖于社交媒体的特性分析。而关于微博流行度的特点,有研究表明微博信息的流行度呈现幂律分布[11]。这种现象的出现源于微博社会网络中的信息过载导致的用户注意力稀缺[12],即微博信息的流行度与用户密不可分。张旸等[13]采用信息增益法分析多种发帖用户特征的重要性,证实了用户影响力之于帖子流行度的重要地位,Jiang 等[14]发现在影响微博信息流行度的重要因素包括该信息内容对相关用户的提及率。可以看出,以上发现多基于单类影响因素重要性的分析,没有综合考虑多种影响因素,特征利用不够全面,同时,不同社交媒体的特点具有独特性,现有的基于其他社交媒体平台的相关工作不能直接应用于微博的研究。

针对上述问题,本文对微博这一社交媒体平台进行分析,针对其特点提出和构造了对应的特征,设计了多种流行度预测方案。考虑到 XGBoost 可以有效地对所提特征进行联合利用[15-16],本文着重提出了一种基于 XGBoost 的微博流行度预测算法。所提出的算法能够从多方面充分考虑与微博流行度密切相关的影响因素,将涉及的相关特征进行提取和融合。首先,基于对原始数据分析,分别从博文信息、话题信息和用户信息 3 方面提取特征。在博文特征中,重点构造了博文内容数值化特征和博文时间特征,并基于博文特征衍生出话题特征。在用户特征中,将用户的影响力具象化,同时从统计学的角度对用户的档位分布特征进行比例计算,作为新的用户特征。本文算法采用了分类式框架,多类特征融合之后,提前对流行度的档位进行划分,使用 XGBoost 作为分类模型对微博的流行度档位进行预测,将

流行度预测问题转换为流行度分类问题。最后,对用户特征进行再构造,基于新的用户特征,将微博的流行度进行分类输出,得到需要的微博转发数、评论数、点赞数和互动值。

总而言之,本文的创新点可以归纳为以下3个方面:

(1) 针对国内社交媒体流行度预测工作匮乏的情况,对微博这一国内主流社交媒体平台的流行趋势特点进行分析和建模,着重挖掘了发博用户、发博时间、博文话题等信息与博文流行度的关联并构造了对应的多种特征。

(2) 基于提取和构造的多种特征,设计了多种微博流行度预测方案,在实验部分进行了性能比较。

(3) 着重提出了一种基于XGBoost的微博流行度预测算法,该算法采用了分类式框架,综合考虑了点赞数、评论数和转发数3个指标,将提取好的博文特征、话题特征和用户特征融合起来,对流行度进行分档预测,在微博流行度预测数据集上取得准确率高达85.69%的良好效果。

1 相关技术

1.1 特征提取和特征融合

本文运用特征提取(Feature extraction)和特征融合(Feature fusion)的思想。特征提取指的是对初始的某一模式的未处理数据进行变换,建立非冗余的能够提供该模式有代表性信息的派生值,即特征,以便后续学习与泛化,特征提取被广泛应用于模式识别和机器学习中,提取出特征的好坏与泛化能力密切相关[17]。

特征融合,是指对同一模式抽取不同的特征矢量进行优化组合[18]。根据融合时间的不同,特征融合又可分为两大类,一类为前期融合(Early fusion),即在模型训练前就将不同的特征融合,融合后的特征用于训练和学习,经典的特征融合方法有串联拼接(Concat)和并行策略(Add)。另一类为后期融合(Late fusion),这一类在特征未完全融合之前就进行模型训练,根据结果改进后多次训练后融合。后期融合典型的方法有Single shot multibox detector (SSD)[19],Multi-scale CNN(MS-CNN)[20]和Feature pyramid network(FPN)[21]等。基于前期融合在社交媒体领域流行度预测的良好表现[22],本文算法采用的是前期融合中的串联拼接方法。

1.2 机器学习模型的应用

社交媒体的流行度预测还依赖于良好模型的构建。极端梯度提升决策树(eXtreme gradient boosting, XGBoost)是在梯度提升决策树(Gradient boosting decision tree, GBDT)的基础上将速度和效率发挥到极致的机器学习模型[15,23],其核心思想是根据样本的特征,从零开始,每一次迭代都在现有基础上增加一棵树,即分类器,去拟合上一次迭代中预测值和真实值的残差,训练完成得到所有分类器的值相加,即为最终的预测结果。在整个迭代的过程中,需要定义一个目标函数,使整个树群的预测值尽可能靠近真实值,同时保障有较大的泛化能力。

本文算法将采用在残差学习中,表现比GBDT更好的XGBoost[15-16]对微博信息的多模态特征进行训练,在充分挖掘和构造有效特征的基础上,利用机器学习的模型提高算法的性能。

1.3 深度神经网络原理

在下文的对比实验中,采用深度神经网络(Deep neural networks, DNN)[22]结构设计了基于深度学习框架的流行度预测方法,与本文算法进行性能对比。

基于感知机的扩展，DNN可以被理解为含有多层隐藏层的神经网络，其内部可分为输入层、隐藏层和输出层3类。见图1，使用的DNN网络包含两个隐藏层，最左边一层是输入层，中间两层是隐藏层，分别为256和128维（此处分别用4和2个神经元代替表达），最终输出层为1维的输出。输入层即为融合后的特征输入。

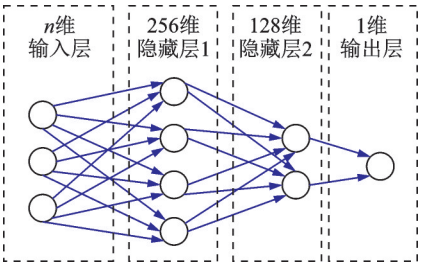


图1 DNN算法结构

Fig.1 Algorithm structure of DNN

2 基于XGBoost的微博流行度预测算法

本文提出了一种基于XGBoost的微博流行度预测算法（图2）。在算法架构中主要包括数据分析、特征的提取与融合以及XGBoost训练3个模块。

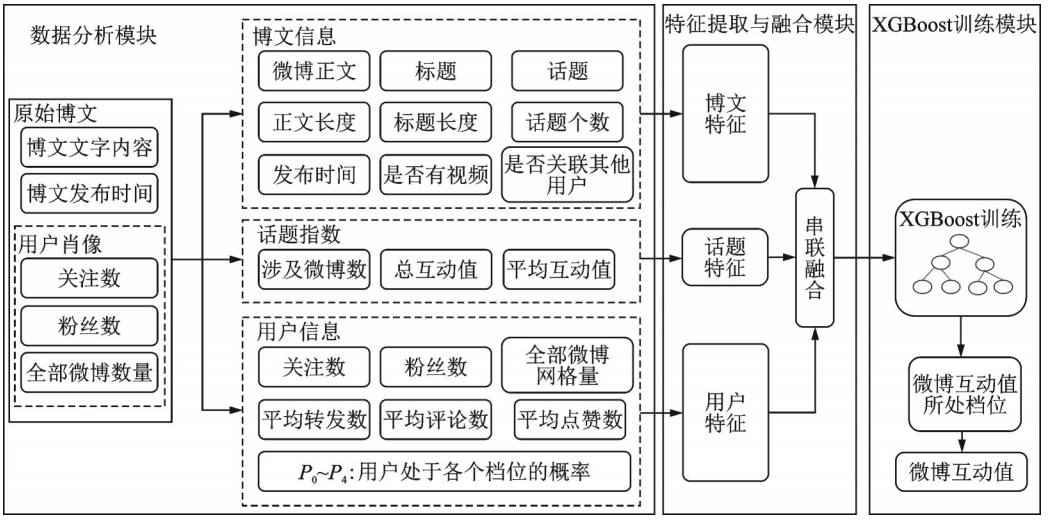


图2 基于XGBoost的微博流行度预测算法架构

Fig.2 Framework of microblog popularity prediction algorithm based on XGBoost

2.1 数据分析

微博流行度预测数据集中的用户肖像信息和微博信息分别见表1和表2。对以上数据进行分析，选取影响流行度预测的关键因素汇总，见表3。从表1~3可以看出，原始数据主要可分为博文信息和用户肖像信息两大类，博文信息有博文文字内容和博文发布时间，用户肖像信息包含用户的全部微博数量、该用户的关注数和粉丝数。这两类数据信息与流行度预测密切相关，都是微博流行度的重要影响因素[2]。但两类数据包含的特征重要性各有差别，在此基础上分别进行特征的提取与构造是十分必要的。

表1 用户肖像信息

Table 1 User portrait information

数据名称	描述
UsertId	帖子Id(抽样&字段加密)
Intro	用户简介
Verified	微博认证
AllWeibo	全部微博数量
Follow	关注数
Follower	粉丝数
DateCrawl	抽取日期

在进一步特征提取与构造的过程中，本算法将有效特征分为3类：博文特征、话题特征和用户特征。

表2 微博信息

Table 2 Microblog information

数据名称	描述
PostId	帖子 Id(抽样&字段加密)
UserId	用户 Id(抽样&字段加密)
Content	博文文字内容
DatePost	博文发布时间
Forward_count	博文在抽取日期时的转发数
Comment_count	博文在抽取日期时的评论数
Like_count	博文在抽取日期时的点赞数

表3 微博流行度预测相关的原始数据

Table 3 Original data related to microblog popularity prediction

数据名称	描述
PostId	帖子 Id(抽样&字段加密)
UserId	用户 Id(抽样&字段加密)
Content	博文文字内容
DatePost	博文发布时间
AllWeibo	全部微博数量
Follow	关注数
Follower	粉丝数

2.2 特征提取与融合

2.2.1 博文特征提取

将原始的博文文字内容和博文发布时间进一步分析可以提取和构造出如表4所示的博文特征。

表4 博文特征

Table 4 Blog features

特征名称	描述	特征来源
Content	博文文字内容	原始
DatePost	博文发布时间	原始
Time	博文发布时刻	构造
Content	微博正文	构造
Title	标题	构造
Tags	话题	构造
Content_Len	正文长度	构造
Title_Len	标题长度	构造
Tags_Count	话题个数	构造
Video	是否有视频	构造
At	是否关联其他用户	构造

对于博文的文字内容,关注到其包含着微博正文、标题和话题等重要内容特征,且内容结构工整,格式较为统一。例如,标题一般由“【】”进行标注,话题存在于“##”之间,微博正文是剩余的文字内容。基于以上特性,对原始的博文进行第一次数据清洗,得到了标题、话题和第一版微博正文。

针对所得第一版微博正文,部分博文带有特殊符号@和网址,分别表示关联其他用户和存在视频链接。基于该发现,对第一版微博正文进行第二次数据清洗,得到了最终的微博正文和是否有视频及是否关联其他用户的布尔类型的附加特征,其中1代表有,0代表无。

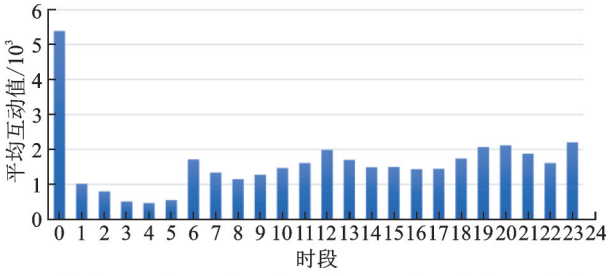
得到微博正文、标题和话题等文本化的内容特征并不能完全满足模型训练的要求,在此基础上进一步对其进行数值化的构造,得到正文长度、标题长度和话题个数3个新的数值化特征。

对于博文的发布时间,基于一天中不同时刻社交媒体的流量存在高低峰差异,见图3,横轴为一天24小时中不同的时段,纵轴表示数据集中所有博文在一天中该时段的平均或总互动值,可以看出不同时刻互动值相差较大,反映了流行度的时间敏感性,故重点关注博文的发布时刻,并从原始发布时间中提取出这个时间特征,认为其可以作为需要关注的有效博文特征。

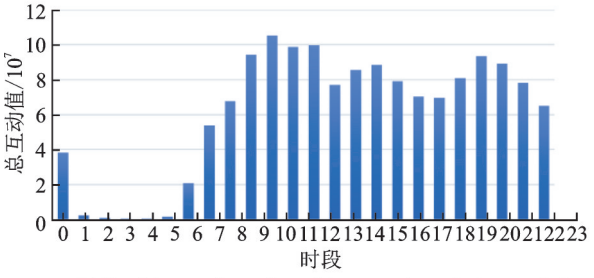
2.2.2 话题特征提取

话题特征是从博文特征中的话题衍生出来的新的特征,主要反映某话题的影响力,即话题指数。

表5主要构造了3个话题特征:话题涉及的微博数、话题的总互动值和话题平均互动值。其中互动



(a) Average interaction value of the posts in each time period



(b) Total interaction value of the posts in each time period

图3 博文发布时段与互动值关系

Fig.3 Relationship between publishing period and interactive value

值表示为微博的转赞评之和,总互动值是话题涉及微博的所有转赞评之和,而平均互动值是话题涉及微博的单条微博的平均转赞评之和。一般认为,话题指数越高,该话题的影响力越大。

2.2.3 用户特征提取

如表4所示,用户特征是从用户肖像信息中提取和构造得来。原始的用户肖像特征如用户全部微博数量、关注数和粉丝数可以大致反映用户的影响力,但还不够细致和全面。

在此基础上,对用户的影响力进一步具象化,主要反映在用户的每条微博的平均转发数(Avg_Repost)、平均评论数(Avg_Comment)、平均点赞数(Avg_Like)和平均互动值(Avg_Total)这4个新构造的用户特征。他们之间存在如下关系

$$\text{Avg\_total}=\text{Avg\_repost}+\text{Avg\_comment}+\text{Avg\_like} \tag{1}$$

除此之外,还创造性地从统计的角度对用户的微博在不同档位的概率进行了计算。档位划分见表6。

用户的微博在不同档位的比例计算公式为

$$P_i=\frac{\text{Count}_{i+1}}{\text{AllWeibo}} \quad i\in\{0,1,2,3,4\} \tag{2}$$

式中Count_{i+1}表示用户在$i+1$档位的微博数。最终,提取的用户特征总结见表7。

特征工程的最后一步,是特征的融合,如图2所示,将提取和构造得到的博文特征f_{post}、话题特

表5 话题特征

Table 5 Topic features

特征名称	描述	特征来源
Index_Count	涉及微博数	构造
Index_Total	总互动值	构造
Index_Avg	平均互动值	构造

表6 档位划分

Table 6 Division of gears

档位	转赞评之和
1	0～10
2	11～50
3	51～150
4	151～300
5	＞300

征f_{tag}和用户特征f_{user}进行串联合并得到f_{all},即$f_{all}=f_{post}\oplus f_{tag}\oplus f_{user}$,作为XGBoost模型的输入特征。

2.3 XGBoost训练与输出

2.3.1 模型分类训练

得到特征工程输出的特征f_{all}后,利用XGBoost对f_{all}进行分类训练,用以预测微博的流行度档位,亦即互动值档位。XGBoost是一种广泛应用于分类和回归问题的决策树模型,在本文算法架构中,输入数据表示如下

$$D=\{(f_i,t_i)|i=1,2,\cdots,n,f_i\in\mathbf{R}^d,t_i\in\mathbf{Z}\} \tag{3}$$

在输入数据表示中,f_i表示第i条微博的总体特征,t_i表示该微博的互动值档位,n为数据集中的微博总数,d表示特征的维度。训练目的是得到预测的微博互互动值档位$\hat{t}_i$,定义如下

$$\hat{t}_i=\phi(f_i)=\sum_{k=1}^{K}g_k(f_i)\qquad \hat{t}_i\in\Gamma \tag{4}$$

式中Γ指的是XGBoost决策树分类树的映射空间,附加有K个激活函数,映射关系为

$$\Gamma=\{g(x)=wf(x)\}(f:R_d\rightarrow L,w\in R_L) \tag{5}$$

在上述公式中,g是由独立树结构f构成的权重为w的分类决策树,每个独立树结构f包含着L片叶子。为了使K个激活函数都得到最好的学习和训练,确定目标函数如下

$$o(\phi)=\sum l(t_i,\hat{t}_i) \tag{6}$$

式中l表示损失函数。目标函数o越小,XGBoost的训练效果越好。

在训练的过程中,还采取了丢弃过大互动值(大于等于10 000)的训练策略,以提高模型对小数值的预测能力,在该训练策略下模型对丢弃的大数值也有能将其预测到档位5的能力。

2.3.2 基于用户特征的分类输出

XGBoost的分类训练运用于微博档位的预测事实上将互动值预测问题转换为互动值的分类问题,在得到预测的档位结果后,需要基于用户的特征将微博的互动值乃至转发数、评论数和点赞数进一步计算得到。

用户在不同的档位上会有不同的互动值,构造一个新的用户特征User_label_avg,计算公式为

$$\text{User\_label\_avg}_i=\frac{\sum_{j=1}^{\text{Count}_i}(\text{Total\_count}_{ij})}{\text{Count}_i}\qquad i\in\{1,2,3,4,5\} \tag{7}$$

式中Total_count$_{ij}$表示在i档位的该用户第j条微博的互动值。

此外,还构造了用户的转赞评比例分布特征Label_distribution用以进一步确定微博的转发数、点赞数和评论数。

$$\text{Label\_distribution}=[\text{Label\_repost},\ \text{Label\_comment},\ \text{Label\_like}] \tag{8}$$

表7 用户特征

Table 7 User features

特征名称	描述	特征来源
AllWeibo	全部微博数量	原始
Follow	关注数	原始
Follower	粉丝数	原始
Avg_repost	平均转发数	构造
Avg_comment	平均评论数	构造
Avg_like	平均点赞数	构造
Avg_total	平均互动值	构造
P_0	用户的微博在第1档位的比例	构造
P_1	用户的微博在第2档位的比例	构造
P_2	用户的微博在第3档位的比例	构造
P_3	用户的微博在第4档位的比例	构造
P_4	用户的微博在第5档位的比例	构造

$$\text{Label\_repost}=\frac{\text{Avg\_repost}}{\text{Avg\_total}} \tag{9}$$

$$\text{Label\_comment}=\frac{\text{Avg\_comment}}{\text{Avg\_total}} \tag{10}$$

$$\text{Label\_like}=\frac{\text{Avg\_like}}{\text{Avg\_total}} \tag{11}$$

式中：Label_repost表示用户的转发比例，Label_comment表示评论比例，Label_like表示点赞比例，3者共同构成了label_distribution这一新的用户特征。

3 实验结果

3.1 数据集

微博流行度预测数据集由随机抽取的500个主流价值观微博用户数据，以及这500个用户于抽取日期前发布的共100万条原创博文数据所构成。实验取每个用户随机90%博文内容数据形成训练集，而每个用户剩下10%数据为测试集。在训练过程中，随机选取训练集的80%用于模型训练，剩下的20%用于算法验证。

数据集中的用户数据包含用户Id(抽样&字段加密)、用户简介、微博认证、全部微博数量、关注数、粉丝数和抽取日期这些用户肖像信息。训练集中原创博文数据包含帖子Id(抽样&字段加密)、用户Id(抽样&字段加密)、博文文字内容、博文发布时间、博文在抽取日期时的转发数、评论数和点赞数。测试集的博文转发数、评论数和点赞数不公开。

3.2 评价指标

实验的评价指标按照分档规则，将每条微博的互动值(转赞评之和)划分为5档，0～10为1档，11～50为2档，51～150为3档，151～300为4档，大于300为5档。每个档位对应的权重见表8。

表8 档位权重

Table 8 Weight of gears

档位	转赞评之和	权重
1	0～10	1
2	11～50	10
3	51～150	50
4	151～300	100
5	＞300	300

在这个分档规则下，将对于每一条博文抽取日期时的互动值(转赞评之和)的预测准确率进行评测，准确率(Accuracy)计算公式为

$$\text{Accuracy}=\frac{\sum_{i=1}^{5}(\text{Weight}_i*\text{Count\_r}_i)}{\sum_{i=1}^{5}(\text{Weight}_i*\text{Count}_i)} \tag{12}$$

式中：Weight_i为第i个档位的权重，Count_r_i为第i个档位预测正确的博文数量，Count_i为第i个档位的博文数量。

3.3 对比实验

除了上文提出的基于XGBoost的分类式流行度预测算法，本文还提出了基于深度学习框架的方法、基于XGBoost的预测式流行度预测算法和用户匹配方法3类不同的设计方案与本文算法进行性能比较。

具体的实验细节为XGBoost训练时采用5次交叉验证,主要参数设置如下:“n_estimators”设为500,“ base_score”设为0.5,“gamma”设为0.1,“learning_rate”设为0.02,“min_child_weight”设为3,“max_depth”设为7。基于深度学习框架的方法在PyTorch环境下训练,学习速率为10^{-4},训练验证次数设为10。

3.3.1 与深度学习方法对比

表9给出了深度学习方法和本文方法在微博流行度预测数据集上的性能对比实验。

表9 与基于深度学习框架的方法对比实验

Table 9 Comparative experiments with methods based on deep learning framework

方法		Accuracy
模型	数据清洗方式	
DNN[22]	全数据集训练,数值特征标准化	0.391 0
	丢弃无需预测用户数据	0.452 2
	丢弃过大数据,对label归一化	0.461 0
	丢弃无需预测数据,丢弃过大数据,对label归一化	0.488 9
	对用户进行分类	0.685 7
	对用户进行分类,丢弃过大数据,对label归一化	0.749 2
本文方法		**0.856 9**

深度学习的方法即基于前文介绍的DNN结构,见图4,在此将所有特征分为两大类,文本特征使用BERT模型处理,数字特征进行串联拼接处理,最后将所有特征融合,送入DNN中训练学习。

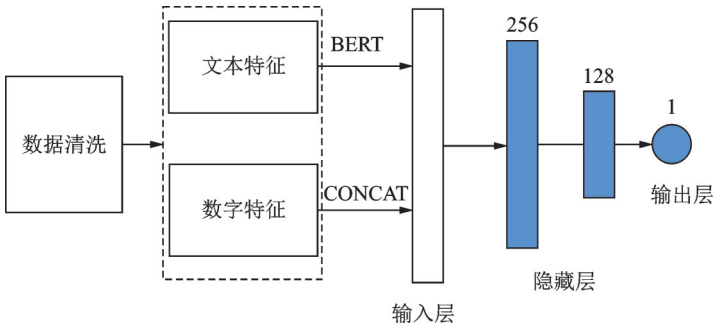


图4 深度学习方法框架

Fig.4 Framework of deep learning methods

如表9所示,在深度学习的各种方法中,发现仅基于DNN模型[22]进行全数据集的训练准确率只有39.10%。

而丢弃无需预测的用户数据,即数据集内其所发微博有95%以上处于某一固定分类,在测试计算准确率时,将这部分权重单独计算,取测试集比例折合到训练结果中,和丢弃互动值过大(≥10 000)的数据,分别提高了6.12%和7%的准确率。同时将两种数据丢弃并将预测目标归一化用于DNN训练,则可以将准确率提高到48.89%。另一种数据处理的方案是对用户进行分类(图5),不同的用户类别赋予不同权重,分类依据如下:

(1)用户类别1:95%以上博文全属于某一分段的用户(0,1,2,3,4);

(2)用户类别2:90%以上博文属于两个相邻分段的用户(01,12,23,34);

(3)用户类别3:90%以上博文属于3个相邻分段的用户(012,123,234);

(4)用户类别4:剩余用户。

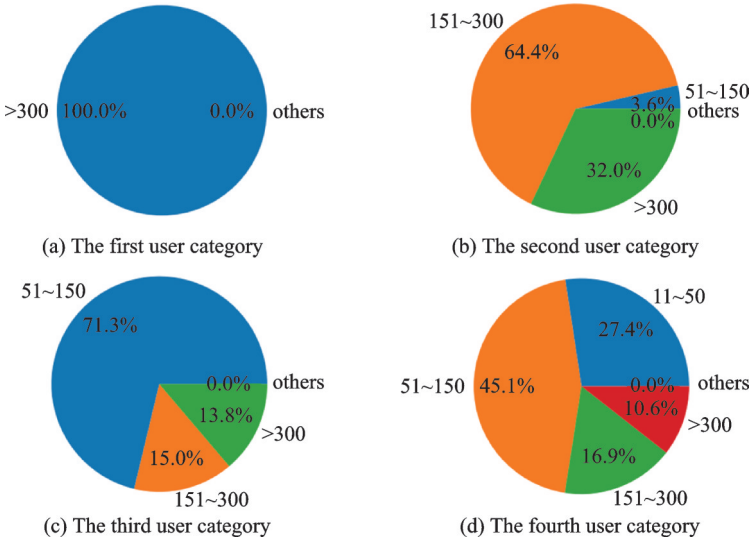


图5 用户分类示例

Fig.5 Example of user classification

如表9所示,对训练集进行用户分类处理后,准确率达到68.57%,性能得到大幅度提升,基于此,将所有对性能提升有益的方案加以融合,最终的准确率达到了74.92%。

尽管深度学习的方法对数据的特征利用率已经很高,但是最终的性能并没有超过本文方法。这是由于与深度学习的方法相比,XGBoost模型对数字特征的敏感性较强,在本文微博流行度预测的情境下,结合用户信息博文信息数字特征占比很大的情况,本文方法所使用的XGBoost模型能更好地利用这些数字特征的关联信息,具有一定的优势,因此优于深度学习算法。

3.3.2 预测与分类对比

本文的方法基于机器学习下的XGBoost模型构建,其中预测对象是区别预测类型的重要关注项,预测对象为互动值或转发数、评论数和点赞数的方法,视作预测方法;预测对象为互动值所处档位的方法,则称之为分类方法。表10列出了基于XGBoost的预测与分类方法的不同数据划分方案的对比实验。

如表10所示,考虑到互动值过大的数据对预测结果的不良影响以及其导致的训练样本分布不均衡的问题,在机器学习的所有方法中,训练时均丢弃了互动值过大的数据。对比发现,分类方法的性能明显高于预测方法,由于分类方法在流行度预测过程中对档位边缘的数据具体互动值容错率较高,提高了其档位预测的准确性。进一步分析可知,在预测方法中,对用户进行分类较其他数据划分方案性能更好,而预测方法中对互动值的预测较转发数、评论数和点赞数的预测更为准确,准确率高了1.53%。在分类方法中,对全数据集的训练性能要略高于丢弃部分用户数据,这表明在分类方法中,即使存在不同用户微博互动值差异过大的损失,提高数据量即使用所有的用户数据对提升互动值档位分类准确度仍有贡献。

表10 基于XGBoost的预测与分类方法对比实验

Table 10 Comparative experiments with prediction and classification methods based on XGBoost

类型	预测对象	数据划分	Accuracy
预测	互动值	全数据集训练	0.772 2
		丢弃无需预测用户数据	0.786 4
		对用户进行分类	0.816 3
	转发数、评论数、点赞数	全数据集训练	0.777 7
		丢弃无需预测用户数据	0.786 7
		对用户进行分类	0.801 0
分类	互动值所处档位	全数据集训练	0.856 9
		丢弃无需预测用户数据	0.856 3

3.3.3 与用户匹配方法对比

表11展示了本文方法和用户匹配等其他方法的综合对比实验。

表11 与用户匹配方法的对比实验

Table 11 Comparative experiments with user matching methods

类别	方法	Accuracy
用户匹配	按用户最大权重档位匹配	0.805 5
	按用户对应时段最大权重档位匹配	0.817 5
机器学习	本文方法	0.856 9

用户匹配方法是一类基于用户特征的,不依靠于任何模型训练的方法。这类方法依赖于已知的用户微博流行度统计信息,将微博流行度情况与微博用户紧密联系在一起。用户匹配方法分为按用户最大权重档位匹配方法和按用户对应时段最大权重档位匹配方法。

按用户最大权重档位匹配方法的匹配策略是通过计算出用户微博各档位总权重分布,将某档位博文数量乘该档位权重。按用户对应时段最大权重档位匹配方法的匹配策略是先按用户一天中各个时段(0~23)统计所发博文获得最大权重的档位,然后按用户和时间进行匹配,若测试集中出现训练集中未出现的时段,匹配两侧相邻时段中权重大的那一个档位。见表11,实验结果表明,用户匹配方法准确率最高可达81.75%,这体现出用户特征在分类处具有显著作用,尽管用户匹配的方法对用户特征的挖掘十分全面,但是缺少模型的支持,在准确率上仍未超过本文方法。

所有实验结果表明,本文方法在评价指标上优于用户匹配和深度学习的所有方法。

4 结束语

针对全媒体时代下,社交媒体流行度预测在信息处理领域的重要性和必要性,本文提出了一种基于XGBoost的微博流行度预测算法。首先,通过对微博数据的特点进行分析,梳理提炼出需要重点考虑的数据;其次,算法运用特征工程的思想详尽地挖掘、提取和构造了与微博博文及微博用户相关的包括博文特征、话题特征和用户特征在内的有效特征,并将有效特征进行融合;最后,将融合后的特征与XGBoost模型进行结合用于训练学习,对用户特征进行二次构造利用,构建一个分类式的流行度预测架构,实现对微博的流行度预测。本算法证实了用户特征在流行度预测上的高影响力,并且在微博流

行度预测的数据集上取得了优越效果,进一步验证了本文算法的合理性和优越性。在实际的应用中,本文算法可用于揭示社交媒体中个人偏好和公众关注,有助于预判社会舆情趋势,提前做出应对决策。同时,高准确率的流行度预测还可以提高用户体验和服务效率,并有利于广泛的应用,如内容推荐、在线广告和信息检索等,具有巨大的商业价值。但基于算法与用户的高度关联性,对缺乏用户信息的流行度预测或存在一定的局限性。

参考文献:

[1] WU B, CHENG W H, ZHANG Y, et al. Sequential prediction of social media popularity with deep temporal context networks [C]//Proceedings of Twenty-Sixth International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan, 2017.

[2] 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.
WU Yue, CHEN Xiaoliang, JIANG Zhongyuan. Survey on predicting popularity of information in microblogs[J]. Journal of Xihua University(Natural Science Edition), 2017, 36(1): 1-6.

[3] 艾擎,张凤荔,陈学勤,等.在线社交网络信息流行度预测综述[J].计算机应用研究,2020,37(S1): 1-5.
AI Qing, ZHANG Fengli, CHEN Xueqin, et al. Survey of information popularity prediction in online social networks[J]. Application Research of Computers, 2020, 37(S1): 1-5.

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

[5] 孔庆超,毛文吉.基于动态演化的讨论帖流行度预测[J].软件学报,2014,25(12): 2767-2776.
KONG Qingchao, MAO Wenji. Predicting popularity of forum threads based on dynamic evolution[J]. Journal of Software, 2014, 25(12): 2767-2776.

[6] HONG L, DAN O, DAVISON B D. Predicting popular messages in twitter[C]//Proceedings of the 20th International Conference Companion on World Wide Web. New York, NY, USA: ACM, 2011: 57-58.

[7] GAO S, MA J, CHEN Z. Modeling and predicting retweeting dynamics on microblogging platforms[C]//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. New York, NY, USA: ACM, 2015: 107-116.

[8] WU B, CHENG W H, ZHANG Y, et al. Time matters: Multi-scale temporalization of social media popularity[C]//Proceedings of the 24th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2016: 1336-1344.

[9] MAZLOOM M, PAPPI I, WORRING M. Category specific post popularity prediction[C]//Proceedings of International Conference on Multimedia Modeling. Switzerland, German: Springer, 2018: 594-607.

[10] VILARES D, ALONSO M A, GÓMEZ-RODRÍGUEZ C. On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages[J]. Journal of the Association for Information Science and Technology, 2015, 66(9): 1799-1816.

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

[12] DAVENPORT T H, VÖLPEL S C. The rise of knowledge towards attention management[J]. Journal of Knowledge Management, 2001, 5(3): 212.

[13] 张旸,路荣,杨青.微博客中转发行为的预测研究[J].中文信息学报,2012,26(4): 109-114, 121.
ZHANG Yang, LU Rong, YANG Qing. Predicting retweeting in microblogs[J]. Journal of Chinese Information Processing, 2012, 26(4): 109-114, 121.

[14] JIANG Y, COUNTS S. Predicting the speed, scale, and range of information diffusion in twitter[C]//Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Palo Alto, CA, USA: AAAI, 2010.

[15] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM Sigkdd International

Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2016: 785-794.

[16] LI L, SITU R, GAO J, et al. A hybrid model combining convolutional neural network with xgboost for predicting social media popularity[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2017: 1912-1917.

[17] 周志华.机器学习[M].北京:清华大学出版社,2015: 114-115.
ZHOU Zhihua. Machine learning[M]. Beijing: Tsinghua University Press, 2015: 114-115.

[18] 全国科学技术名词审定委员会.计算机科学技术名词[M]. 3版.北京:科学出版社, 2018.
China National Committee for Terms in Sciences and Technologies. Chinese terms in computer science and technology[M]. 3rd ed. Beijing: Science Press, 2018.

[19] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//Proceedings of European Conference on Computer Vision. Switzerland, German: Springer, 2016: 21-37.

[20] CAI Z, FAN Q, FE RIS R S, et al. A unified multi-scale deep convolutional neural network for fast object detection[C]//Proceedings of European Conference on Computer Vision. Switzerland, German: Springer, 2016: 354-370.

[21] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 936-944.

[22] DING K, WANG R, WANG S. Social media popularity prediction: A multiple feature fusion approach with deep neural networks[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2019: 2682-2686.

[23] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. Annals of Statistics, 2001(4): 1189-1232.

作者简介:

任敏捷(1995-),女,博士研究生,研究方向:自然语言处理、情感计算、多媒体内容分析,E-mail: renminjie@tju.edu.cn。

靳国庆(1988-),男,教授,研究方向:多媒体内容分析与理解、多媒体内容安全、机器学习。

王晓雯(2000-),女,硕士研究生,研究方向:机器学习。

陈睿东(1998-),男,硕士研究生,研究方向:机器学习。

袁运新(1996-),男,硕士研究生,研究方向:自然语言处理。

聂为之(1987-),**通信作者**,男,副教授,博士生导师,研究方向:计算机视觉、多媒体信息分析、跨媒体信息检索、人工智能,E-mail: weizhinie@tju.edu.cn。

刘安安(1982-),男,教授,博士生导师,研究方向:计算机视觉、机器学习、三维模型检索、多媒体信息处理。

(编辑:夏道家)