软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):967−980 [doi: 10.13328/j.cnki.jos.005921]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于动静态表征的众筹协同预测方法*

张 凯[1], 赵洪科[2], 刘 淇[1], 潘 镇[1], 陈恩红[1]

[1](大数据分析与应用安徽省重点实验室(中国科学技术大学),安徽 合肥 230027)

[2](天津大学 管理与经济学部,天津 300072)

通讯作者: 刘淇, E-mail: qiliuql@ustc.edu.cn

摘 要: 众筹是一个新兴的互联网金融平台,项目的发起者可以通过使用互联网,征求大量平台用户的资金来资助他们的项目.但是由于众筹平台所具有的独特规则,只有在特定时间内收集了足够的资金,项目的筹资才会成功进行交易.为了防止项目发起者和投资者在可能失败的项目上浪费时间和精力,动态追踪众筹项目的筹资过程以及估算其融资成功概率便极为重要.然而,现有的一些工作既没有针对动态预测跟踪机制的研究,也没有考虑平台上的项目发起者和投资者之间的动态行为交互.为了解决这些问题,基于长短期记忆网络设计了一种新颖的动静态协同预测模型.该模型着重分析了用户行为,包括评论的情绪倾向以及融资过程中的动态增量信息,从而将融资项目与投资人之间的交互行为进行深度挖掘分析.首先,针对平台上的静态特征和动态用户行为数据,通过不同的 Embedding 方法得到他们的深度表征.在此基础上,进一步设计了基于注意力机制的协同预测模型,以便了解项目融资的时序信息对最终结果的影响程度.最后,在真实的众筹数据集上进行的大量实验结果表明,所提出的动静态表征预测方法相比其他预测方法更为有效.

关键词: 动态追踪;用户行为分析;深度语义表征;注意力机制;长短期记忆网络

中图法分类号: TP181

中文引用格式: 张凯,赵洪科,刘淇,潘镇,陈恩红.基于动静态表征的众筹协同预测方法.软件学报,2020,31(4):967−980. http://www.jos.org.cn/1000-9825/5921.htm

英文引用格式: Zhang K, Zhao HK, Liu Q, Pan Z, Chen EH. Cooperative prediction method based on dynamic and static representation for crowdfunding. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):967−980 (in Chinese). http://www.jos.org.cn/1000-9825/5921.htm

Cooperative Prediction Method Based on Dynamic and Static Representation for Crowdfunding

ZHANG Kai[1], ZHAO Hong-Ke[2], LIU Qi[1], PAN Zhen[1], CHEN En-Hong[1]

[1](Anhui Province Key Laboratory of Big Data Analysis and Application (University of Science and Technology of China), Hefei 230027, China)

[2](College of Management and Economics, Tianjin University, Tianjin 300072, China)

Abstract: Crowdfunding is an emerging finance platform for creators to fund their efforts by soliciting relatively small contributions from a large number of individuals using the Internet. Due to the unique rules, a campaign succeeds in trading only when it collects adequate funds in a given time. To prevent creators and backers from wasting time and efforts on failing campaigns, dynamically estimating the success probability of a campaign is very important. However, existing crowdfunding systems neither have the mechanism

* 基金项目: 国家自然科学基金(61672483, 71790594, U1605251); 中国科学院青年创新促进会优秀会员专项(2014299)

Foundation item: National Natural Science Foundation of China (61672483, 71790594, U1605251); Special Fund for the Member of Youth Innovation Promotion Association of CAS (2014299)

本文由“非经典条件下的机器学习方法”专题特约编辑高新波教授、黎铭教授、李天瑞教授推荐.

收稿时间: 2019-05-27; 修改时间: 2019-07-29; 采用时间: 2019-09-20; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-14 09:53:16, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.0953.006.html

of dynamic predictive tracking, nor consider the dynamic interaction between project sponsors and investors on the platform. To address these issues, a novel dynamic and static collaborative prediction model is designed based on long and short-term memory network. This model focuses on user behavior, including the emotional tendency of reviews and the dynamic incremental information in the financing process, so as to deeply mine and analyze the interaction between financing projects and investors. Firstly, for the static features and dynamic user behavior data on the platform, their deep characterization is obtained by different embedding methods. On this basis, a collaborative prediction model based on attention mechanism is further designed to understand the impact of timing information of project financing on the final results. Finally, experiments on real crowdfunding datasets show that the proposed dynamic and static representation prediction method is more effective than other prediction methods.

Key words: dynamic tracking; user behavior analysis; deep semantic representation; attention mechanism; LSTM network

众筹[1]是通过筹集人们的小额资金来资助一个项目或企业的机制或行为.近年来,许多众筹平台得到了快速发展,并极大地提高了金融市场的资本活力.然而,尽管众筹行业在金融领域取得了巨大的发展,但成功达到预期融资金额目标的众筹项目比例仅为总数的 40%左右[2].而根据多数众筹平台上的“All or Nothing”原则[3],一旦项目的筹资金额达不到预设的目标值便会宣告筹资失败,项目的发起者将失去之前筹集的所有资金,并且在此项目上投入的所有努力都将付诸东流.同时,如果项目筹资失败,参与该项目的支持者虽然可以收回投资的资金,但可能会浪费大量的时间和机会成本.由此可见,当前众筹领域发展面临着两方面的挑战.首先,众筹作为一种高效而便捷的新型融资方式逐步受到重视并正在快速发展;其次,众筹行业目前的整体筹资完成率不高,融资效率较低,融资成功率亟待提高.因此,深度探究和挖掘影响众筹融资成功的关键因素[4],构建行之有效的成功率预测与解释模型,对众筹平台提高项目融资成功率,促进众筹行业的发展具有重大意义.

幸运的是,众筹平台在常年的发展过程中积累下了大量的数据,其不仅包含众多项目初始的静态属性,例如:项目的初始特征,包括标题(title)、发起人(user)、项目介绍(intro)等,也包含了大量的动态时序数据,例如:用户投资行为数据以及用户的评论数据,项目的信息展示如图 1 所示.因此,本文将分析用户在平台上的海量历史行为数据,跟踪项目融资过程中的用户动态行为属性,并将其与静态属性联合建模,对项目融资结果进行预测.其结果不仅可以帮助项目发起者更好、更及时地调整项目的设计,也会带来更好的项目融资环境,为项目的发起者提供更准确的用户反馈,并最终从根本上提高项目融资成功的可能性以及众筹行业整体的融资效率,促进互联网金融众筹行业的健康发展.然而,尽管这些用户动态行为数据给项目预测带来了新的突破口,但是如何对这些数据进行深层次的挖掘与分析仍然存在着诸多挑战.

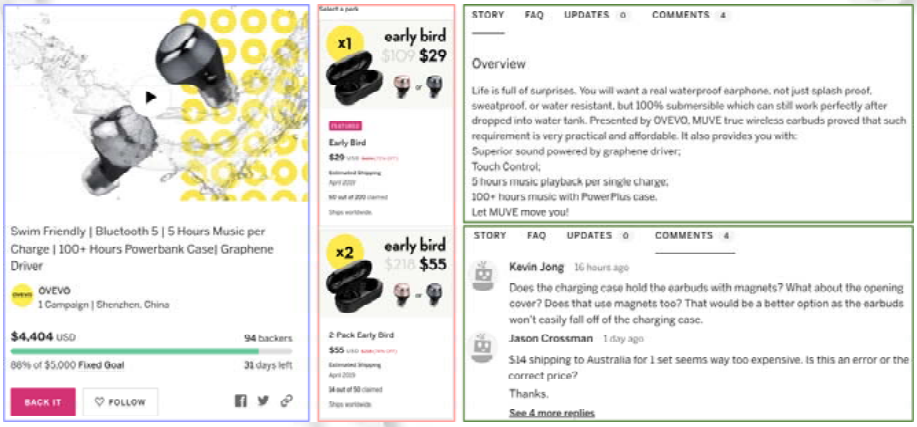

Fig.1 An example of crowdfunding campaign

图 1 项目众筹示例

1) 首先,平台上项目的数据,包括静态数据和动态数据,存在很大的异构性[5],即使对于同一个项目,在不同阶段的融资动态也有很大的不同.如何全面而准确地分析这些异构的特征,并能够将其各阶段整合起来进行准

确的预测不是一个简单的问题;

2) 第二,根据大量的数据分析可以发现,捐赠者的行为,特别是其主观行为,包括对项目的投资金额以及对项目的评论与平台上项目的融资成败有着极为密切的关联.而如何建模这些用户行为数据和项目融资成败之间的关系,并且进一步地利用这种关系进行项目融资结果的预测也是一个有待研究的问题;

3) 最后,即便能够将静态数据与动态数据进行良好的联合建模,也找到了充分挖掘用户行为信息的有效方法,但是,与一些主流的平台不同,众筹平台上的项目评论包含大量的无标签数据,即支持者的评论数据缺乏情感标签.例如:在淘宝与京东等一些主流电子商务平台上,用户对于某件产品进行评论时,往往会相应地给其打分(通常分为 1~5 星评价),而因为有这些比较量化的“星级”指标,对于用户的观点分析就显得容易许多.同样的举措也体现在一些电影业务平台上,将用户对于电影的观影评论划分为 10 个星标等级,以便用户表达他们对于影片的情感态度倾向,得到的数据标签充足,分析起来十分方便高效.而众筹属于一种新兴的互联网金融平台,由于其发展成熟度不高,平台上仅提供评论的输入功能,并不存在较为完善的评分机制,导致用户的评论数据缺乏情感标签,传统的有监督学习方法无法适用,致使对用户评论行为的挖掘形成了更为严峻的挑战.

针对以上问题,本文提出一种基于长短期记忆网络(long short-term memory network,简称 LSTM)[6]动静态表征的众筹项目协同预测方法(dynamic and static collaborative prediction,简称 DSCP).该方法根据众筹平台上用户行为数据的特殊性,首先采用迁移学习[7]的方法,并利用其良好的领域自适应特性,对无标签的用户评论信息进行情感分类.同时,本文还将用户的动态特征属性按时间序列划分,并利用 LSTM 对这些时序数据进行建模,将时间序列性的文本以及投资序列映射为低维的特征向量,使其能够精准地刻画评论所表达的语义情感信息以及资金动态所包含的融资进程信息.特别地,由于之前已有研究表明[8],项目融资到不同的时间段而被用户关注的程度应该是不同的.因此,本文创新性地使用注意力机制(attention mechanism)[9]来确保模型在更新过程中,可以通过训练得到每个阶段的注意力权重,由此来判断项目在融资过程中的各个时间段对于项目最终融资结果的影响程度.在此基础上,注意力机制也可以进一步增强模型的可解释性,有助于分析影响众筹融资成败与否的关键性因素.最后,对于项目的静态数据和动态数据的融合,由于其静态特征在形式上是异构的,本文对其进行一致化处理后与时序的动态数据进行全连接,从而将给定某个项目的动静态特征进行充分挖掘利用.而对于不同时间段的融资量信息以及评论文本信息,通过之后的注意力机制对其进行解释.本文设计了多种评价指标以及案例对所提方法进行了充分验证.实验结果表明,这种基于长短期记忆网络的深度预测模型能够较准确地预测项目的融资成败,也证明了其对项目融资过程中动态属性表征的合理性以及高效性.

本文的主要贡献如下.

1) 提出了一种基于用户动态行为表征的众筹项目预测方法,通过分析用户的投资行为以及用户评论文本信息,挖掘平台用户对于众筹项目的观点倾向变化,并进行项目的融资结果预测;

2) 通过对项目的动态融资过程以及静态属性信息进行协同建模,将动态的特征划分为不同时序阶段,利用长短期记忆网络以及注意力机制来捕捉项目融资进程中的关键性影响因素;

3) 在情感分析的基础上采用迁移学习方法对平台上出现的无标签评论进行深度语义表征、情感倾向分析,解决了用户评论缺少标签难以利用的难题,并通过可视化分析了用户行为对项目的动态影响;

4) 在真实众筹数据集上进行了大量的实验,通过与当前几种最先进的方法进行比较,验证了本文所提方法的有效性.

1 相关工作

目前与该领域相关的工作可分为如下几类,即众筹相关研究、时间序列分析、深度表征学习以及项目融资预测这 4 个方面.

1.1 众筹相关研究

由于众筹仍然是一个新兴的互联网金融平台,因此该领域的大部分工作都相对较新,很多研究内容尚处于探索阶段.当前最流行的众筹形式是以奖励为基础,即个人为项目提供资金以换取各种奖励的形式.如国内流行

的众筹网、京东众筹和淘宝众筹,美国的 Indiegogo、Kickstarter、Donors Choose 等众筹平台.而在这些平台中,Indiegogo 已成为全世界范围内最受欢迎的奖励型众筹平台之一.2014 年,众筹平台上的项目获得了高达 5.29 亿美元的众筹金额和 22 252 个成功资助的项目.在众筹平台上,投资者称为支持者,项目的发起人称为融资者.融资者通过众筹平台发布有关其项目的详细描述来展示他们的想法或创新.通常,描述信息包含解释项目新颖性的视频、图像或文本信息.除此之外,融资者还提供详细的项目规划、资金目标以及不同类别的奖励.然而,尽管 Indiegogo 平台的发展非常迅速,但与诸多众筹平台一样,项目的融资成功率并不乐观.最近的统计数据显示[3],平台上的项目总体成功率不到 40%.而由于该领域相对较新,发展速度较快,尽管有一些相关研究[10,11],但还较少有研究从数据挖掘的角度来探讨这一领域存在的问题[12,13].

1.2 时间序列分析

在科研文献中,时间序列分析是一个已被广泛研究的问题[14,15].时间序列有很多种形式,可以刻画出不同的随机过程.例如,自回归和自向量回归都是假设当前的变量按照先前变量的线性规律进行变化.然而,传统的自回归模型只能处理反应变量,不能处理输入的外生变量.为了处理输入变量之间的非线性关系以及其他外部变量,一些其他的非线性模型也被应用到时间序列预测上,例如:支持向量机[16]和神经网络[17,18].

随着近些年来深度学习的发展,很多研究发现循环神经网络(recurrent neural network,简称 RNN)[19]在处理时间序列的数据上有着独特而令人满意的效果,其优秀的序列数据处理能力使其被广泛应用于诸多自然语言处理任务上.如机器翻译、情感分析、关系实体抽取等.特别地,长短期记忆网络 LSTM 是一种特殊的 RNN,主要是为了解决长序列数据在训练过程中出现的梯度消失和梯度爆炸问题,即相比普通的循环神经网络,LSTM 能够在更长的时间序列数据中有更好的表现.然而,虽然这些前沿的方法已被广泛地应用于自然语言处理和计算机视觉任务上,但却很少有将其应用于众筹项目预测上的研究,这也导致了目前的对众筹这个领域的研究始终处于及格线的标准,项目融资结果预测的表现并不是十分优秀.此外,目前的深度学习模型大多都是黑盒形式,缺少令人信服的解释.本文在神经网络输出层的基础上,使用注意力机制,对 LSTM 输出的隐向量进行权重运算,从而得到不同时间段序列对项目融资成败的影响力大小,极大地增强了模型的可解释性.

1.3 深度表征学习

深度表征学习是机器学习的一个分支或者子领域,使用了多层次的非线性结构对信息进行处理和抽象,转变为更高维的、更抽象的特征,用于有监督或无监督的特征学习、表征、分类和识别.表征学习旨在将研究对象的语义信息表示为低维稠密实值向量.表征学习得到的低维向量表示是一种分布式表示,孤立地看向量中的每一维都没有明确对应的含义;而综合各维形成一个向量时则能够表示对象所蕴含的语义信息.与简单的独热(one-hot)表示方法相比,表征学习的向量维度较低,有助于提高计算效率,同时能够充分利用对象间的语义信息,从而有效缓解数据稀疏问题.由于表征学习的这些优点,最近出现了大量关于文本、图像和社会网络的表征学习的研究.

近年来,以深度学习为代表的表征学习技术在语音识别、计算机视觉(CV)以及自然语言处理(NLP)领域获得了广泛关注,有着极其普遍的应用.2006 年,Hinton 提出了深度学习[20]这一概念,随后与其团队提出了深度学习模型深度信念网络[21],并给出了一种高效的半监督算法:逐层贪心算法,以训练深度信念网络的参数,打破了长期以来深度网络难以训练的僵局,开启了深度学习的热潮.随着数据量以及算力的提升,深度学习在文本分析、图像处理、语音识别以及视频处理等领域均取得了重大进展.例如,2009 年在 ImageNet 发布之后,各种深度网络模型的出现大大促进了计算机视觉领域乃至整个深度学习研究的发展,其在视觉识别挑战赛(ILSVRC)上的表现甚至已超越人类的辨别水平,并在人脸识别、物体检测等现实场景下得到了有效的落实.在自然语言理解领域,2013 年 Google 团队发表了基于语言模型获取词向量的 Word2vec 工具[22],其核心思想是通过词的上下文得到词的向量化表示,具体包括 CBOW(通过附近词预测中心词)和 Skip-gram(通过中心词预测附近词)两种方法,以及负采样和层次 Softmax 两种近似训练法.Word2vec 的词向量可以较好地表达不同词之间的相似和类比关系,自提出后被广泛应用在各种 NLP 任务中.

1.4 项目融资预测

众筹源于众包的概念,其中不同的投资者为某一个项目的解决方案做出资金贡献.在众筹中,参与者扮演推动者和投资者的角色,投资者可以贡献不同金额以支持某个众筹项目,而项目发起者可以根据支持者贡献金额的大小给予其不同的回报.在众筹平台中,项目的信息包含诸如项目目标金额、项目类别、奖励数量、奖励描述、项目持续时间、项目详情、视频是否存在、社交网络中的朋友数量、奖励水平以及项目描述中的句子数量等.而传统的一些机器学习方法仅将这些特征归一化后送到机器学习分类器,如支持向量机(SVM)和随机森林(RF),以预测众筹项目是否成功.文献[23]假设发起人对项目的描述中存在有说服力的短语,会对投资者的投资行为产生一定的影响.因此其从项目描述信息中提取出不同的主题表征,之后与其他的特征相结合作为项目的最终表示,并用来预测众筹项目成功率.此外,文献[24]提出应用项目的一些时间特征(持续时间,金额的投资序列)和机器学习方法,如 K 近邻(KNN)、马尔可夫链和支持向量机(SVM)来预测项目成功概率.在文献[25]中,作者研究了 Kickstarter 领域的融资进程动态,并基于融资动态进行项目成功率的追踪预测.此外,一些研究致力于探索激励用户投资众筹项目的因素,文献[26,27]便是基于此目的而对众筹平台进行的实际分析.文献[26]致力于分析人们为什么会在众筹平台上发布融资项目以及哪些原因会对众筹项目产生影响.文献[27]提出了一个最大熵分布模型和显示团队的行为在 Kiva.org 领域的影响力以及对项目的发起者的指引作用.文献[28]创造性地利用了平台外部的数据,该工作描述了社交网络对 Kickstarter 项目的影响,即利用社交网络特征来对项目在社交网络上的影响力进行估计.在他们的工作中,作者利用基于社交网络的功能,例如:Twitter 中的宣传活动、弱连接组件、网络直径以及三元封闭等的影响,来预测支持者的数量和项目的最终融资金额.具体来说,其将项目在 Twitter 等社交媒体上分享的频次与项目的其他特征联合起来,分析该项目的传播影响力大小,并用来预测支持者的数量和筹款成功的概率.文献[29]提出了最大熵分布模型,并揭示了团队行为在众筹领域中的影响.此外,在探索互联网对微观融资和点对点借贷交易的影响方面,也存在一些研究工作[30].关于微决策的研究发现,支持者不仅倾向于对自身相似的融资者提供融资帮助,而且还倾向于在触发积极情绪反应的情况下向项目提供资金支持[31].虽然该研究很有意思,并有可能带来巨大的融资影响,但令人惊讶的是,并没有真实的研究能将用户的情绪、融资金额时序信息与项目的关系刻画出来.

2 问题定义与方法

本文尝试从动态和静态的角度协同预测项目是否会获得足够的资金来实现其筹款目标.对项目成功率的早期预测显然对支持者、融资者以及众筹服务平台都有着积极的意义.支持者可以将注意力转向那些显示出更有成功迹象的项目,从而降低时间和金钱的浪费程度;融资者可以参考项目的用户评论、融资金额序列以及项目预测结果对项目进行及时调整;平台亦可以通过项目的预测情况扩展当前的推荐服务以提高未来项目融资的成功率,还可以改善众筹平台的项目质量.在本节中,将主要介绍基于深度学习的动静态协同预测模型,其充分融合了用户的动态行为数据以及项目的静态属性信息,采用双向长短期记忆网络对众筹项目进行建模,并使用注意力机制来捕获项目融资过程中局部时间序列特征对于项目最终融资结果的影响.

2.1 基本定义

众筹项目成功率预测与其属性高度相关,本文首先将问题进行形式化.假设每个项目可以用一个三元组 $(P_i,\boldsymbol{X}_i,Y_i)$表示,$P_i$代表第 i 个项目,$\boldsymbol{X}_i$表示项目的属性特征向量,Y_i表示项目是否众筹成功.最终的目标就是训练一个有效的模型,使得对于一个新的项目 P_j,由其特征向量 $\boldsymbol{X}_j$ 去估算该项目融资成功的概率.值得注意的是,本文将项目所有特征的最终表征向量形式化为 $\boldsymbol{X}_j$,它由 3 部分信息通过不同的技术来表征:即用户信息评论的深度表征向量、动态融资金额序列表征向量以及静态信息的表征向量层层运算得到.

定义 1(用户评论信息深度表征). 之前的多数研究并没有考虑使用用户评论信息,有些也仅将评论信息进行简单的特征映射处理.而深度表征则关注于挖掘用户评论行为所表达的深度语义信息.具体来说,给定一个项目 P_i 的评论集合 $R=\{R_1,R_2,\ldots,R_n\}$,需要得到项目每条评论$\{R_1,R_2,\ldots,R_n\}$相应的情感向量$\{S_1,S_2,\ldots,S_n\}$,然后将用

户对项目的情感倾向用于建模项目的融资进程.在用户评论表征的过程中,需要训练一个良好的分类器来将众筹项目的评论进行情感分类,这里存在如下一些挑战.

1) 众筹平台上的项目评论大多是无标签的评论,传统的监督学习方法无法适用.而由于数据量较大,人工标注又将会耗费巨大的人力资源成本;

2) 目前的深度情感分类模型的普适性较低,不具备良好的跨领域分类效果,不足以提取出众筹领域用户评论的深层次情感特征.

鉴于以上两方面的挑战,本文从众筹领域出发,分析其领域评论数据的独特性,采取深度迁移学习模型,引入亚马逊商品评论数据集,与众筹评论的数据联合建模,训练出一个在众筹领域表现优异的情感分类器.具体来说,本文将两个领域的评论数据进行词嵌入之后,输入迁移学习模型中进行联合训练得到情感分类模型.最后将众筹项目的评论经该模型处理后,输出得到用户评论信息的深度情感表征向量.

定义 2(融资金额时间序列表征). 上一节介绍了众筹平台融资动态追踪的重要性.其不仅可以帮助分析众筹过程中不同时间序列段对项目投资的影响情况,也可以帮助找出融资过程中项目的变化规律.对于众筹项目 P_i,其包含支持者投资金额以及具体投资时间、项目对应的每条评论的时间节点,经处理后得到项目的融资金额和评论序列 $T=(t_1,t_2,\ldots,t_n)$.n 表示预设置的项目持续时间(n=30 或 60),其中,$t_i(1\leqslant i\leqslant n)$表示项目每天收到的投资金额以及用户评论.本文将项目众筹时间划分为不同的阶段,例如每天为一个时间序列段,则每个时间段的融资时间序列可表示为 $T_1=t_1,T_2=t_2,\ldots,T_n=t_n$,并将这些融资时间序列进行初步表征之后,输入双向长短期记忆网络(bi-directional long short-term memory network,简称 Bi-LSTM)进行深度特征抽取,学习其每个融资时间序列段的隐层表示状态 $\boldsymbol{h}=\{h_1,h_2,\ldots,h_n\}$.

定义 3(项目静态属性表征). 将基于用户行为的评论以及融资金额进行深度表征之后,需要进一步对项目的其他一些属性进行表征,包括项目的基本属性以及融资者的基本属性信息.为了统一表示,本文把所有的特征表示成数值型变量,具体来说,

1) 对于时间序列数据,例如项目发起时间、每笔投资时间等,项目将原始变量转化为一个序列日期数字,表示固定的预设时间之后的天数;

2) 对于类别型特征,例如项目类别、发起人性别以及货币类型等,将一个含有 $n(n<10)$中类别的变量转换为 n 维的二进制向量,只有相应类别变量中的位置被置为 1,其他位置为 0,即 one-hot 编码方法.而对超过 10 种类别的变量,则用出现频率代替变量来表示;

3) 类似于以上方法,总体来说,本文总共使用了 20 种特征(其中包括 13 种静态属性、6 种动态属性以及本文重点研究分析的 2 种用户动态行为属性,即用户评论和投资额序列).

2.2 动态表征预测模型

如前所述,众筹平台上的每一个项目包含有 20 种属性特征,而其中多数的属性是历史的、静态的,例如项目在创建时设置的描述信息、目标金额、奖励回报等.为了方便将这些属性进行分类及表征,本文将其划分为 4 类,见表 1,分别为:项目发起者属性、投资者属性、奖励回报属性以及其他属性信息.而将这些属性划分为 4 个类别之后,对它们进行静态建模.此外,本文还对基于用户行为的评论数据以及融资金额序列进行动态建模,挖掘其深度语义以及时序表征.下面将主要介绍用于众筹环境下的动静态协同表征预测模型.具体来说,本文首先介绍历史静态数据建模技术,然后概述用户动态行为数据建模方式,并分别介绍建模过程的技术细节.

1) 静态建模

如图 2 所示,项目的静态属性可分为 4 类,即发起人(owner)、投资人(backers)、奖励设置(perks)和项目其他(others)信息.这些特征是异构的[32],其中包括数值变量、类别变量以及文本特征.为了数据的一致性,需要将所有的属性变量转换为数值形式或者数字变量,以便输入模型进行训练.具体操作如下:保留数值变量的属性;将类别变量进行 one-hot 编码成向量形式;对于文本数据,如项目的标题、描述以及奖励的描述等信息,采取 Word2vec 进行 Embedding,得到一定程度上的语义表征向量.最后,将这 4 部分的表征向量经全链接层进行降维处理.具体来说,本文将发起人、投资人、奖励设置和项目其他信息形式化为 X_{owner}、X_{backer}、X_{perk}、X_{other}.通过以上所述的

方法分别将其转化为特征向量表示($\boldsymbol{V}_{owner}$,$\boldsymbol{V}_{backer}$,$\boldsymbol{V}_{perk}$,$\boldsymbol{V}_{other}$)后,首先对它们进行拼接,然后利用全连接网络(full-connect net)进行特征降维,并学习所有静态属性特征的完整表示 $\boldsymbol{H}_{static}$,其公式表达为

$$V_{static} = V_{owner} \oplus V_{backer} \oplus V_{perk} \oplus V_{other} \tag{1}$$

$$H_{satatic} = LeakyRelu(W_1 \cdot V_{static} + b_1) \tag{2}$$

其中,符号 ⊕ 代表简单拼接,*LeakyRelu* 表示一种激活函数.

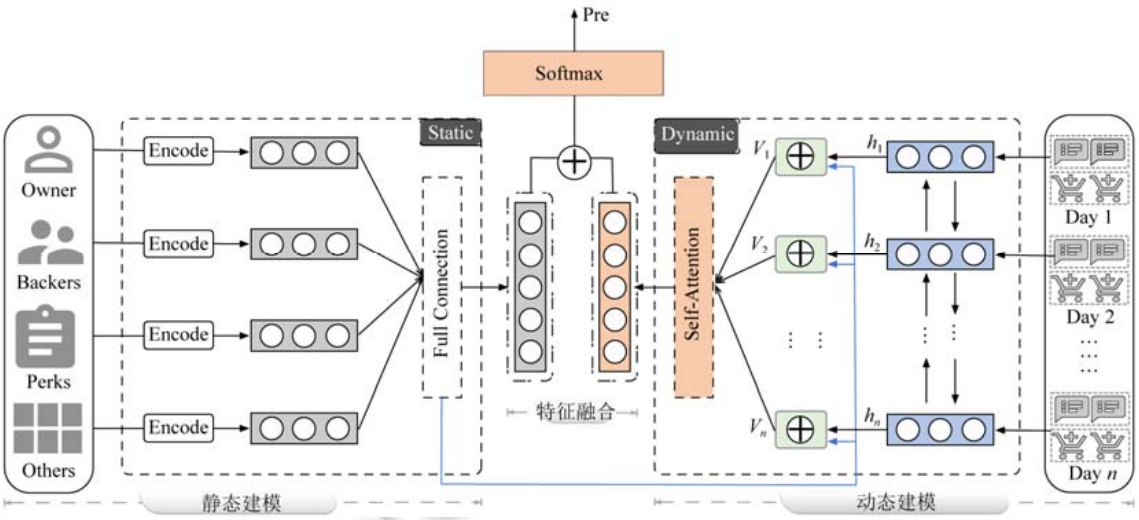

Fig.2 Dynamic and static collaborative prediction model

图 2 动态和静态协同预测模型

由此,本文完成了对静态的项目属性信息建模,并且得到了其表征向量 $\boldsymbol{H}_{static}$.而为了更加充分地利用项目的静态属性,在动态建模部分,本文也将 $\boldsymbol{H}_{static}$ 与时间序列的特征进行联合建模,以更好地对整个项目进行全面的表征.

2) 动态建模

在本小节中,我们将展示建模用户行为数据(即融资金额序列数据和评论数据)的过程.如上一节所述,首先通过迁移学习方法用情感标签来标记项目评论信息的情感倾向.这里,本文选择已有的领域自适应方法,该方法基于交互式注意力迁移网络(interactive attention transfer network,简称 IATN)[33],以亚马逊评论数据集作为源域,众筹领域的项目评论数据集作为目标域进行跨域的情感分类训练,得到跨领域的模型(M_{iatn}).之后对众筹领域下的项目评论进行情感倾向分析,最终获得所有用户评论的情感表示.该问题可以形式化为:假设某项目包有含一条有 n 个单词的评论信息,表示为 $\boldsymbol{r}=\{w_1,w_2,\ldots,w_n\}$,首先将其使用 Word2vec 进行向量表征,得到 $\boldsymbol{e}_r=\{e_1,e_2,\ldots,e_n\}$.然后通过预训练的模型 M_{iatn} 进行情感表征后得到其深度情感表征向量 $\boldsymbol{V}_{sen}$.这里值得注意的是,通过分析发现,在约 85%的评论中包含的单词数在 50 个之内,故将 n 的大小设为 50,过长则截取,过短则补 0.最后,通过评论的时间戳信息,将项目评论的情感特征向量进行聚合得到项目每日的评论特征表示.

类似地,本文使用相同的汇总方法来处理每日的融资金额序列,并通过独热(one-hot)方法将它们映射成特征向量 $\boldsymbol{V}_{fund}$.最后,将每日评论特征表示和资金特征表示连接为项目的每日动态特征 $\boldsymbol{V}_t$,其中,$1 \leqslant t \leqslant n$,$n$ 表示预设置的项目持续时间(n=30 或 60),公式化为

$$V_t = Day(V_{sen}) \oplus V_{fund} \tag{3}$$

其中,函数 $Day(\cdot)$代表将每日的特征拼接.

由上述步骤,可以得到项目在融资期间每天的融资金额以及评论特征向量表示.而由于这些动态特征具有较强的时间序列特征,本文采用双向长短期记忆网络(bi-LSTM)的深度学习方法来学习特征序列的隐向量,因为

它在学习时序的长期依赖性方面表现良好,并且可以有效地解决梯度消失和爆炸问题.此外,相比 LSTM,其在双向序列信息特征表示方面会有更好的表现.具体来说,给定众筹项目的每日特征向量表示(例如,$\boldsymbol{V}_1,\boldsymbol{V}_2,\ldots,\boldsymbol{V}_n$)作为输入,正向 LSTM 从 t=1 到 n(反向从 t=n 到 1)不断更新记忆细胞序列 $\boldsymbol{c}_t$ 和隐状态 $\boldsymbol{h}_t$.在初始化之后,第 t 次迭代的步骤中,每次迭代的隐状态 $\boldsymbol{h}_t$ 由前一个隐状态 $\boldsymbol{h}_{t-1}$ 和当前的特征向量 $\boldsymbol{V}_t$ 共同决定,以正向长短期记忆网络为例,更新具体过程为

$$\left.\begin{aligned} i_t &= \delta(W_{ei}\cdot V_t + W_{hi}\cdot \vec{h}_{t-1} + b_i),\\ f_t &= \delta(W_{ef}\cdot V_t + W_{hf}\cdot \vec{h}_{t-1} + b_f),\\ c_t &= f_t\cdot c_{t-1} + V_t\cdot \tau(W_{ec}\cdot V_t + W_{hc}\cdot \vec{h}_{t-1} + b_c),\\ o_t &= \delta(W_{eo}\cdot V_t + W_{ho}\cdot \vec{h}_{t-1} + b_o),\\ \vec{h}_t &= o_t\cdot \tanh(c_t) \end{aligned}\right\} \tag{4}$$

其中,i_t,f_t 和 o_t 分别是第 t 次迭代的输入、遗忘和输出门.$\boldsymbol{V}_t$ 是特征嵌入向量.c_t 是记忆单元,$\vec{h}_t$ 是 Bi-LSTM 的前向隐向量输出.$\delta(\cdot)$是非线性激活函数,在本文中使用的为 sigmoid 函数.标点符号表示向量之间的元素乘法.$\boldsymbol{W}_*$ 表示权重矩阵,$\boldsymbol{b}_*$是偏置向量.通过类似的方法,可以得到 Bi-LSTM 的后向隐向量输出 $\overleftarrow{h}_t$.最后,通过向量拼接,可以得到用户动态时序特征的隐向量表示 $\boldsymbol{h}$={$h_1,h_2,\ldots,h_n$},其中,

$$h_t = \left[\vec{h}_t \oplus \overleftarrow{h}_t\right], t\in[1,n] \tag{5}$$

3) 特征融合

在 Bi-LSTM 层之后,每日项目评论和融资金额将由基础的向量表示转换为蕴含丰富时序特征的隐状态向量,即 $\boldsymbol{h}$={$h_1,h_2,...,h_n$}.为了将静态特征和动态特征更高效及协作地加以利用,本文将静态特征向量 $\boldsymbol{H}_{static}$ 与动态特征的每个隐状态向量 $\boldsymbol{h}_t$ 相连接,并将结果 V'_t 作为 t 时刻的特征向量表示,公式表示如下:

$$V'_t = [H_{static} \oplus h_t] \tag{6}$$

如前所述,在对动态时序数据进行分析的时候,不同时刻的隐状态对项目成功与否的影响力是不同的,其被关注的程度也应该是不同的.因此,项目每天的动态增量数据对融资结果预测应该也有着不同的影响[34].例如,若第 10 天项目的大多数评估是消极的,那么对于最终预测,该天的影响可能更大,因为相比正面情绪,人们对待事物的态度更容易受到负面情绪的影响.因此,在运算过程中,本文使用注意力机制,通过在特征表示的每个步骤中为编码向量分配权重来突出显示输入的不同部分的影响力,训练得到每个时刻(每天的特征)的注意力权重,并将该权重和其对应的隐状态向量 $\boldsymbol{V}_t$ 相乘来强化对应时刻隐状态对于整个项目序列的影响程度.其权重计算如下:

$$\alpha_t = \frac{\exp(\tan h(V'_t\cdot \boldsymbol{W}_s + \boldsymbol{b}_s))}{\sum_{t=1}^{n}\exp(\tan h(V'_t\cdot \boldsymbol{W}_s + \boldsymbol{b}_s))} \tag{7}$$

其中,$\boldsymbol{W}_s$ 表示权重矩阵,$\boldsymbol{b}_s$ 是偏置矩阵.tan$h$ 是非线性激活函数.在计算得到每天的特征的权重表示之后,将其与隐向量进行相乘,得到动态特征的最终表示如下:

$$H_{dynamic} = \sum_{t=1}^{n}\alpha_t\cdot V'_t \tag{8}$$

最后,将项目的历史静态数据特征表示 $\boldsymbol{H}_{static}$ 与最终动态时序数据表示 $\boldsymbol{H}_{dynamic}$ 结合起来,并通过 Softmax 层处理它们,以对项目 i 成功与否的标签 Y'_i 进行预测,即:

$$Y'_i = \text{soft}\max([H_{static} \oplus H_{dynamic}] + b) \tag{9}$$

3 实验验证

3.1 数据集介绍

为了验证方法的有效性和实用性,实验数据集选取真实的众筹项目数据,其源于一个在全世界范围内被广

泛关注和使用的众筹平台 Indiegogo(https://www.indiegogo.com)上的真实项目数据.本文将所有的数据经预处理满足独立同分布之后,随机划分为 80%训练集、10%的验证集和 10%的测试集,训练数据集用于训练深度表征模型,验证数据集用来验证模型在训练过程中是否过拟合,测试数据集用于验证模型效果及质量.数据集中项目的主要属性见表 1,数据集大小见表 2.

Table 1 Project attribute category description

表 1 项目属性类别描述

种类	特征类型	属性	描述
Owner	静态	Name Age	发起人姓名 发起人年龄
Backer	静态	Country Times	支持者国籍 投资时间
Perks	静态	Number Perk option Describe	奖励的个数 奖励的金额 奖励的描述
Others	静态	Title –	项目标题 –
Funding-progress	动态	Reviews Funding	用户评论 筹集到的金额

Table 2 Data set of crowdfunding

表 2 众筹数据集

类别	数值
项目总数	119 113
用户总数	207 221
评论数	590 335
众筹成功项目数	30 387
众筹失败项目数	88 726

3.2 实验设置

1) 特殊数据预处理.正如先前所述,众筹平台上的每一个项目包含 20 种属性特征.在这些属性中,不可避免地会存在诸多数据问题,例如噪声、空值等.对于这种数据,本文用归一化、离散化和缺省值填充等方法对其进行数据处理,如对缺失属性过多的数据直接进行丢弃;对含有空值的数据,用众数或平均值进行填充.另外,所有的数值型数据都利用 Z-score 转换进行正则化.特别地,由于要分析用户评论信,所以对文本的处理显得尤为重要.本文使用 National Language Toolkit (nltk) tool 开源工具库对评论文本进行特殊字符清洗、拼写检查、词干提取以及词形还原等.然后使用 Word2vec[35]进行词表征(word embedding),将每个单词表示成 200 维的向量.

2) 参数设置.在实验中,所有文本中的单词都被映射为 200 维的词向量.LSTM 的时间序列为 50,隐向量的纬度设置为 50,batch size 大小设置为 64.所有权重矩阵由均匀分布μ(–0.01,0.01)随机初始化,并且所有偏置矩阵初始设置为 0.对于模型的性能调优,最终将 *Dropout* 率、l_2 归一化系数、学习率分别设置为 0.25、10^{-4} 和 10^{-3}.模型基于 Tensorflow 实现,并在具有 4 个 2.0GHz Intel Xeon E5-2620 CPU 和 Tesla K20m GPU 的 Linux 服务器上进行训练.此外,在模型的训练过程中,使用 Adam 作为优化器,并将所有的 weight initializer 置为 0.

3) 训练策略.基于定义,每个项目可以用一个三元组$(P_i,\boldsymbol{X}_i,Y_i)$表示,$P_i$ 代表第 i 个项目,$\boldsymbol{X}_i$ 表示项目所有的属性特征向量,Y_i 表示项目是否众筹成功(Ground Truth 标签).

根据前文所述,对于项目 P_i,其特征向量 $\boldsymbol{X}_i$(即动态特征向量与静态特征向量的拼接,$H^i_{dynamic} \oplus H^i_{static}$)输入模型后可得到预测的标签 Y'_i.所以在模型训练阶段,本文采用交叉熵分类损失函数来优化模型的学习过程,其公式表达为

$$L = -\frac{1}{N}\sum_{i=1}^{N} Y'_i \ln Y_i + (1-Y'_i)\ln(1-Y_i) + \lambda L_{reg} \tag{10}$$

其中,N 表示训练数据集中的项目数.正则化项 L_{reg} 用来防止训练过程中的过拟合现象,λ是正则化参数.模型的训练目标就是最小化损失函数 L.此外,训练过程中所有的参数优化采取的都是反向传播(back-propagation)算法[36].

3.3 对比方法

为了验证所提出模型方法的效果,本文选取了 6 种已有方法作为对比实验.此外,为了进一步验证项目动态时序数据的有效性,本文还设计了 3 种模型的变种来进行对比.

1) 逻辑斯蒂回归(logistic regression,简称 LR).一种广义的线性模型,用于解决二分类问题的机器学习方法,该方法用于估计某种事物的可能性.其通过 Sigmoid 函数引入非线性因素,可以轻松处理 0/1 分类问题.

2) 随机森林(random forest,简称 RF).由分类树算法改进而来,即在变量和数据的使用上采取随机化的方法生成分类树,再汇总分类树的结果,在运算量没有显著增加的前提下大大提高了预测精度.而因为随机性,其具备很好的抗噪以及防止过拟合能力.

3) 支持向量机(support vector machines,简称 SVMs).这是一种二分类模型,其基本模型是定义在特征空间上的间隔最大的线性分类器.其基本思想就是求解能够正确划分训练数据集并且几何间隔最大的分离超平面,可形式化为一个求解凸二次规划的问题.

4) 基于社交媒体(social media,简称 SM)的预测方法.该方法由文献[28]提出,提出挖掘项目的社交媒体(Twitter 分享的次数)特征,以此来分析项目在人群中的受欢迎程度以及社交媒体的影响力,并通过这种挖掘出来的特征建模预测项目的投资人数目和成功率.

5) 基于生存分析(survival analysis,简称 SA)的方法[28].将项目成功预测表示为生存分析问题,并应用删失回归方法,其中可以在存在缺失部分信息的情况下执行回归预测.其严格研究了众筹数据的项目成功时间分布,并表明逻辑和对数逻辑分布是从这些数据中学习的自然选择.

6) 基于域约束主题分布模型(domain-constraint latent Dirichlet allocation,简称 DC-LDA)[23].文献[23]指出,目前很少研究项目的文本信息来分析众筹能否成功.其研究工作的主要贡献是设计一个新的基于文本分析的框架,该框架可以从项目的文本描述中提取潜在的语义,并用随机森林方法预测项目的筹款结果.

为了进行实验效果的对比,验证用户行为数据(评论和融资序列)的有效性,在对静态属性进行表征的基础上,本文设计了如下基于两种不同动态特征的对比模型.

7) 基于静态信息的预测模型($DSCP^{none}$).为了验证动态数据表征的有效性,设计了基于静态属性的对比实验.即采用动静态协同预测模型的 Static 部分,仅仅使用静态特征对众筹项目进行建模预测.

8) 基于评论信息的动静态协同预测模型($DSCP^{review}$).使用基于 LSTM 的深度学习方法,对项目的所有静态属性特征进行深度表征,在动态数据方面使用用户评论信息进行联合建模.

9) 基于融资金额序列的动静态协同预测模型($DSCP^{funds}$).使用基于 LSTM 的深度学习方法,对项目的所有静态属性特征进行深度表征,在动态数据方面使用融资金额序列进行联合建模.

10) 动态和静态协同预测模型(DSCP).本文提出的深度表征方法,充分利用用户的动态行为数据以及项目的历史静态数据,在动态数据方面使用用户评论信息和融资金额序列进行协同建模.

4 实验结果

4.1 评价指标

如表 2 所示,在 Indiegogo 众筹数据集中成功的项目数量远远大于失败项目的数量[37],导致模型在训练过程中出现正负例数据不平衡的现象,所以此时项目预测准确率(accuracy)不能够完全反映出模型的预测效果.为了更加全面地评价这些预测算法,本文基于混淆矩阵,同时使用精度(precision)、召回率(recall)和 $F1$ 值($F1$-score)作为衡量预测模型的性能指标,其计算公式分别为

$$Precision=\frac{TP}{TP+FP} \tag{11}$$

$$Recall=\frac{TP}{TP+FN} \tag{12}$$

$$F1\text{-}Score=\frac{2\times P\times R}{P+R} \tag{13}$$

其中,*TP* 表示真正,即将成功的项目标签预测为成功(1);*TN* 表示真负,即将失败项目的标签预测为失败(0);*FP* 代表假正,即将失败的项目有标签预测为成功(1);*FN* 代表假负,即将成功项目的标签预测为失败(0).*P* 表示 Precision 的值,*R* 代表 Recall 的值.

Table 3 Experimental results of the DSCP model on the Indiegogo dataset

表 3 DSCP 模型在 Indiegogo 数据集上的实验结果

基线	准确率	精度	召回率	*F*1-值
LR	0.630 9	0.628 5	0.655 6	0.641 7
RF	0.647 5	0.657 2	0.661 4	0.659 3
SVM	0.700 6	0.674 3	0.709 8	0.691 5
SM (social)	0.742 2	0.721 5	0.678 0	0.700 2
SA (survival analysis)	0.784 3	0.723 3	0.726 2	0.724 7
DC-LDA	0.770 0	0.736 8	0.742 0	0.739 2
$DSCP^{none}$	0.767 7	0.725 6	0.730 7	0.728 1
$DSCP^{review}$	0.796 2	0.781 3	0.770 0	0.776 0
$DSCP^{funds}$	0.804 1	0.785 5	**0.784 3**	0.784 9
DSCP	**0.818 2**	**0.807 7**	0.773 3	**0.790 1**

4.2 实验结果

实验结果见表 3,动态和静态协同预测模型从各个评测指标上都取得了较好的效果.在实验中,LR、RF、SVM 这 3 种传统机器学习方法由于不具备对项目属性的深度特征学习能力,所以最终在预测表现上一般,其中最优的支持向量机方法的准确率为 70.06%,*F*1 值为 0.691 5.而引入了社交媒体信息的预测方法 SM 充分挖掘了项目在社交网络上的影响力,拓展了项目属性的纬度,其 *F*1 值表现较大幅度地提升到了 0.700 2.基于生存分析的 SA 方法以及项目描述信息的 DC-LDA 方法较之前效果也有了一定的提升,*F*1 值分别达到了 0.724 7,0.739 2.

实验结果表明,本文提出的用于众筹项目融资结果预测的协同预测模型是有效的.更具体地,通过将所提出的 DSCP 表征模型应用于众筹项目的用户评论信息和融资金额序列,从中挖掘用户偏好倾向,获得了语义上丰富的情感特征以及时序特征.这些语义丰富的特征与基本的项目静态特征相结合,以构建一个特征集,Bi-LSTM 以及 Attention 机制利用该特征集来增强众筹成功的预测.根据从 Indiegogo 众筹网站收集的真实数据,实验结果表明,本文所提出框架的表现最优,可以达到 *F*1 值为 0.790 1,相较于目前已有的方法,提升了约 5.10%.此外,基于深度表征方法设计的基于评论信息的动态和静态协同预测模型($DSCP^{review}$)和基于融资金额序列的动态和静态协同预测模型($DSCP^{review}$),其表现也分别提升了 3.68%和 4.57%,均远高于仅使用静态属性信息的预测方法($DSCP^{none}$),因而更进一步地验证了动态信息对于项目融资结果的重要性.总的来说,众筹项目融资过程中获得的用户动态行为特征可以反映用户偏好,从而成为影响项目融资成功与否的一个重要决定因素.这些特征不仅有助于预测筹款成功,而且还有助于解释众筹项目过程中重要的时间点信息[38](例如,用户通常喜欢在项目即将成功的时候,为其追加投资).通过挖掘项目的评论信息,发起人可以更好地追踪筹款过程中支持者对于项目的态度,从而可以通过后续的改善更新、更好地推进他们的项目.

4.3 案例分析

在实验中,为了进一步展示用户评论以及时序信息对于众筹项目的影响,我们对实验的中间结果及数据进行了分析与可视化.具体来说,本文对评论预测的情绪倾向进行统计分析.如图 3 所示,横轴代表天数,纵轴代表评论的次数,红色柱状图表示使用迁移学习方法预测的正面评论的数量,灰色表示负面评论的数量.通过这个案例,我们可以直观地观察到项目每天收到用户的正、负面评论数量与时间的关系.例如,在活动的第 6 天收到的

积极评论的数量是10条,消极的评论数是6条.随着项目投放时间的增加,平台用户的新颖感下降导致其收到的关注度降低.而在项目临近结束时,其关注度有了相应的回升,不仅说明了用户的行为受项目融资进度(时间)的影响,也佐证了项目众筹成功与否和时间序列信息极大的相关性.

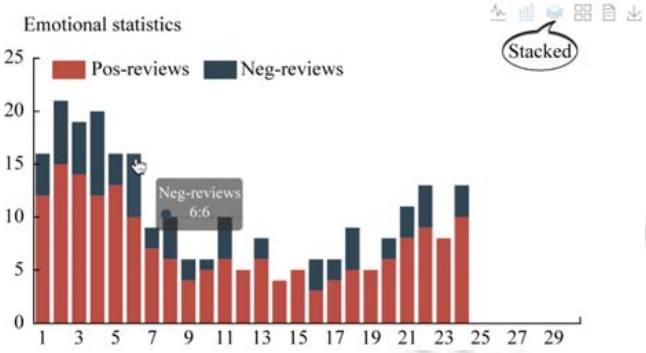

Fig. 3 Analysis of emotional tendency of user comments

图3 用户评论情感倾向分析

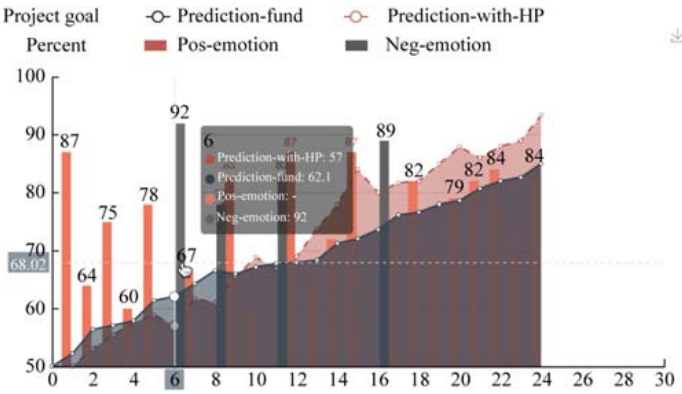

Fig.4 Demonstration of projects success rate prediction

图4 项目成功率预测展示

通过以上的介绍,我们可以得到评论的情感信息,以及每天用户对项目的正面和负面评价的数量.之后,我们将进一步挖掘这些信息,并将它们结合起来进行动态时序分析.具体如图 4 所示:红色直方图表示当天总体评价为正面的概率值.相反,灰色柱状图表示项目为负面情绪的概率(我们只显示当天预测概率超过 50%的结果),也就是说,如果在图表的当天显示红色柱状图,则该天的公众意见通常是正面的,灰色则代表相反意见.值得注意的是,这个结果不是通过简单的统计得到的,而是通过模型预测得到的.除了直方图之外,我们还使用一个折线图来表示项目的成功率动态预测结果.灰色线图表示仅使用资金的项目预测结果.如图 4 所示,灰色线图的趋势总是缓慢而平稳地上升,这意味着,随着资金的增加,项目的成功率也在增加.最后,我们用红色折线图来表示用资金和评论共同参与预测的项目成功率结果.我们可以观察到,曲线在总体趋势上上下下变化,但在一些日子里,例如第 6 天,当项目总体评论呈负面倾向时,其趋势总是下降的,这意味着评论以及时序信息在项目的动态成功率预测中起着至关重要的作用.具体到现实的场景中,若众筹平台上的一个项目在发起的第6天收到很多的负面评论,则说明该项目的新颖、可行性等在平台用户心中并没有受到广泛认可,甚至可能会被怀疑.因此,项目在该天的风评走低就会导致其总体的众筹成功率下降,而这与我们模型预测的结果相符合,从而佐证了模型的有效性.

5 结 论

针对众筹平台上的数据异构、隐含语义难以表征、动静态数据难以处理的问题,本文提出一种基于双向长短期记忆网络(Bi-LSTM)的动静态协同表征的众筹项目预测方法.该方法根据众筹平台上用户行为数据的特殊性,对无标签的评论文本采用了迁移学习模型进行情感分类.同时,还将用户的动态行为特征按时间序列划分,并利用注意力机制对这些时序数据进行协同建模,从而进一步分析了不同时间段序列对项目最终结果的影响力.

在众筹数据集上的实验结果表明,本文提出的模型充分挖掘了平台上的用户行为信息,提高了众筹项目最终融资成功与否的预测效果,同时还充分说明了时序信息对于项目融资进程的影响.总体而言,本文的研究提高了对众筹模式下项目融资成败因素的了解,进而提高了众筹项目的融资效率.在当前的互联网金融背景下,不仅有助于众筹项目在筹资过程中的结构调整,也极大地改善了整个众筹平台内的融资环境.

References:

[1] Belleflamme P, Lambert T, Schwienbacher A. Crowdfunding: Tapping the right crowd. Journal of Business Venturing, 2014, 29(5):585–609.

[2] Liu Q, Wang G, Zhao H, *et al*. Enhancing campaign design in crowdfunding: A product supply optimization perspective. In: Proc. of the IJCAI. 2017. 695–702.

[3] Wang G, Zhao H, Liu C, *et al*. Product supply optimization for crowdfunding campaigns. IEEE Trans. on Big Data, 2018.

[4] Li Y, Rakesh V, Reddy CK. Project success prediction in crowdfunding environments. In: Proc. of the 9th ACM Int'l Conf. on Web Search and Data Mining. ACM, 2016. 247–256.

[5] Zhang H, Zhao H, Liu Q, *et al*. Finding potential lenders in P2P lending: A hybrid random walk approach. Information Sciences, 2018,432:376–391.

[6] Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation, 2000,12(10): 2451–2471.

[7] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2010,22(10):1345–1359.

[8] Zhao H, Liu Q, Zhu H, *et al*. A sequential approach to market state modeling and analysis in online P2P lending. IEEE Trans. on Systems, Man, and Cybernetics: Systems, 2018,48(1):21–33.

[9] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017. 5998–6008.

[10] Huang JQ, Huang XF, Yin GP. Influencing factors and forecasting model for successful crowdfunding projects. China Soft Science, 2017,(7) (in Chinese with English abstract).

[11] Li YT, Zuo WM. Prediction model of crowdfunding project financing results. Statistics and Decision, 2016,(2):86–89 (in Chinese with English abstract).

[12] Rakesh V, Lee WC, Reddy CK. Probabilistic group recommendation model for crowdfunding domains. In: Proc. of the 9th ACM Int'l Conf. on Web Search and Data Mining. ACM, 2016. 257–266.

[13] Etter V, Grossglauser M, Thiran P. Launch hard or go home!: Predicting the success of Kickstarter campaigns. In: Proc. of the 1st ACM Conf. on Online Social Networks. ACM, 2013. 177–182.

[14] Box GEP, Jenkins GM, Reinsel GC, *et al*. Time Series Analysis: Forecasting and Control. John Wiley & Sons, 2015.

[15] Brockwell PJ, Davis RA, Calder MV. Introduction to Time Series and Forecasting. New York: Springer-Verlag, 2002.

[16] Cao LJ, Tay FEH. Support vector machine with adaptive parameters in financial time series forecasting. IEEE Trans. on Neural Networks, 2003,14(6):1506–1518.

[17] Zhang G, Patuwo BE, Hu MY. Forecasting with artificial neural networks: The state of the art. Int'l Journal of Forecasting, 1998,14(1):35–62.

[18] Yan W. Toward automatic time-series forecasting using neural networks. IEEE Trans. on Neural Networks and Learning Systems, 2012,23(7):1028–1039.

[19] Mikolov T, Karafiát M, Burget L, *et al*. Recurrent neural network based language model. In: Proc. of the 11th Annual Conf. of the Int'l Speech Communication Association. 2010. 1045–1048.

[20] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015,521(7553):436.

[21] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006,18(7):1527–1554.

[22] Goldberg Y, Levy O. word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv Preprint arXiv:1402.3722, 2014.

[23] Yuan H, Lau R YK, Xu W. The determinants of crowdfunding success: A semantic text analytics approach. Decision Support Systems, 2016,91:67–76.

[24] Mitra T, Gilbert E. The language that gets people to give: Phrases that predict success on Kickstarter. In: Proc. of the 17th ACM Conf. on Computer Supported Cooperative Work & Social Computing. ACM, 201. 49–61.

[25] Kuppuswamy V, Bayus BL. Crowdfunding creative ideas: The dynamics of project backers in Kickstarter. In: Hornuf L, Cumming D, eds., The Economics of Crowdfunding: Startups, Portals, and Investor Behavior. 2017.

[26] Gerber EM, Hui JS, Kuo PY. Crowdfunding: Why people are motivated to post and fund projects on crowdfunding platforms. In: Proc. of the Int'l Workshop on Design, Influence, and Social Technologies: Techniques, Impacts and Ethics. 2012,2(11):10.

[27] Hui JS, Greenberg MD, Gerber EM. Understanding the role of community in crowdfunding work. In: Proc. of the 17th ACM Conf. on Computer Supported Cooperative Work & Social Computing. ACM, 2014. 62–74.

[28] Lu CT, Xie S, Kong X, *et al*. Inferring the impacts of social media on crowdfunding. In: Proc. of the 7th ACM Int'l Conf. on Web Search and Data Mining. ACM, 2014. 573–582.

[29] Choo J, Lee D, Dilkina B, *et al*. To gather together for a better world: Understanding and leveraging communities in micro-lending recommendation. In: Proc. of the 23rd Int'l Conf. on World Wide Web. ACM, 2014. 249–260.

[30] Ashta A, Assadi D. Do social cause and social technology meet? impact of Web 2.0 technologies on peer-to-peer lending transactions. Cahiers du CEREN, 2009,29:177–192.

[31] Stephen AT, Galak J. The effects of traditional and social earned media on sales: A study of a microlending marketplace. Journal of Marketing Research, 2012,49(5):624–639.

[32] Zhao H, Jin B, Liu Q, *et al*. Voice of charity: Prospecting the donation recurrence & donor retention in crowdfunding. IEEE Trans. on Knowledge and Data Engineering, 2019.

[33] Zhang K, Zhang H, Liu Q, *et al*. Interactive attention transfer network for cross-domain sentiment classification. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. 2019,33(1):5773–5780.

[34] Jin B, Zhao H, Chen E, *et al*. Estimating the days to success of campaigns in crowdfunding: A deep survival perspective. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence.2019,33(1):4023–4030.

[35] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). 2014. 1532–1543.

[36] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. arXiv Preprint arXiv:1409.7495, 2014.

[37] Zhao H, Ge Y, Liu Q, *et al*. P2P lending survey: platforms, recent advances and prospects. ACM Trans. on Intelligent Systems and Technology (TIST), 2017,8(6):72.

[38] Zhao H, Zhang H, Ge Y, *et al*. Tracking the dynamics in crowdfunding. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2017. 625–634.

附中文参考文献:

[10] 黄健青,黄晓凤,殷国鹏.众筹项目融资成功的影响因素及预测模型研究.中国软科学,2017,(7).

[11] 李雅婷,左文明.众筹项目筹资结果预测模型.统计与决策,2016,(2):86–89.

张凯(1993－),男,安徽亳州人,博士生,CCF 学生会员,主要研究领域为数据挖掘,自然语言处理,个性化推荐.

潘镇(1988－),男,博士生,主要研究领域为数据挖掘,社交网络.

赵洪科(1989－),男,博士,讲师,CCF 专业会员,主要研究领域为数据挖掘,商务分析,互联网金融.

陈恩红(1968－),男,博士,博士生导师,CCF 会士,主要研究领域为数据挖掘,社交网络,推荐系统.

刘淇(1986－),男,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为数据挖掘,社交网络,个性化推荐.