文章编号:1000－2367(2018)02－0111－06 **DOI**:10.16366/j.cnki.1000－2367.2018.02.018

基于集成学习的微博用户转发行为预测

张效尉, 王伟, 秦东霞

(周口师范学院 网络工程学院,河南 周口 466001)

摘 要:为了提高微博用户转发行为预测的精度,提出一种有效的基于集成学习的微博用户转发行为预测算法.首先,对影响用户转发的各种特征进行综合分析,提取出用户属性、社交关系、微博内容等影响用户转发行为的特征;然后,采用 Logistic 回归、支持向量机与 BP(BackPropagation)神经网络等机器学习算法对用户转发行为进行预测;最后,利用"加权投票法"的集成学习方法对多个预测结果进行融合.实验结果表明,相对于 BP 神经网络算法,在综合评价性能的 F_1 度量值上,集成学习算法有 1.5%的性能提升.

关键词:新浪微博;转发行为预测;集成学习;社交关系

中图分类号:TP391.1 **文献标志码**:A

新浪微博作为国内最大的社交媒体,拥有庞大的用户数量,目前注册用户已达到 5 亿,用户每天产生的微博内容超过 1 亿条.用户在使用新浪微博过程中产生的数据即包含用户属性(如性别、年龄)、微博内容(如文本、图片)等基本信息,又包括点赞、评论、转发等行为信息.在对这些数据进行收集和分析的基础上,有效地预测用户的转发行为,不但可以构建微博中由用户转发引起的消息扩散模型以进行信息传播规律分析,而且可将结果应用于好友和话题的个性化推荐、用户情感分析等领域,使之能更好地服务于人们的生活、企业的发展与政府部门的决策.

针对微博用户转发行为预测的研究起源于 2008 年 Twitter 作为社交媒体的流行,早期的研究成果侧重于对引起用户转发行为的原因、转发方式和内容的理论分析;接着研究者提出用因子图、支持向量机、神经网络等机器学习算法构建模型,将用户的转发行为作为分类问题进行预测,并且国内学者将研究的对象从 Twitter 转向新浪微博.上述研究成果,大多是针对用户转发行为特征的分析与结果统计,采用机器学习算法给出用户转发行为预测结果,这些方法特征提取不全面,模型简单.本文采用的集成学习方法是近年来机器学习研究领域的热门方向,该方法将多个机器学习算法得出的结果按照某种规则进行整合以获得比单个学习器更好的预测效果.因此,本文采用基于"加权投票法"的集成学习算法预测微博用户的转发行为,对 Logistic 回归、支持向量机与 BP 神经网络等机器学习算法得出的用户转发行为结果进行融合,提高了预测性能.

1 相关工作

在微博社交网络中,各种信息通过用户的转发操作快速传播,因而,用户的转发行为对信息的传播起着关键性的作用,也是个性化推荐、社区发现等相关领域的研究基础.

在相关工作中,文献[1]以 Twitter 为研究对象,针对人们使用 Twitter 时的转发方式、转发动机以及转发内容的主题倾向进行了分析.文献[2]使用主成分分析方法研究了 URL、标签、关注和粉丝人数等影响用户转发的主要因素,并应用广义线性模型分析了影响因素与转发行为的关系.文献[3]针对用户转发行为缺少

收稿日期:2017-08-08;**修回日期**:2017-12-25.

基金项目:国家自然科学基金(U1504602);河南省科技攻关项目(172102210089;162102210396);河南省自然科学基金研究项目(152300410129);河南省高等学校重点科研项目(15A520125;17A520019;15A520114).

作者简介(通信作者):张效尉(1982－),男,河南开封人,周口师范学院讲师,主要研究方向为数据挖掘、社交网络,E-mail:252303648@qq.com.

系统性研究的问题,考虑转发活跃性、转发时间规律性、内容的重要性、用户兴趣等影响用户转发行为的因素,提出因子图模型,有效的预测了用户的转发行为.文献[4]以 Twitter 为例,预测一条微博是否被转发,研究微博中的转发行为,同时针对微博中不同特征的重要性,提出基于特征加权的预测模型.文献[5]以新浪微博为研究对象,从数据集中提取各种影响用户转发行为的特征,使用机器学习分类方法,预测用户的转发行为,利用概率级联模型对给定微博的转发路径进行预测,有效的分析了新浪微博的信息转发机制与传播特征.文献[6]面向社交网站中用户评论行为,预测用户是否参与讨论,采用基于特征的机器学习方法,引入参数控制数据集的不平衡性,采用豆瓣小组的真实数据,有效提高用户评论行为的预测效果.

上述研究侧重于用户转发行为的理论研究,或用机器学习算法对用户行为的统计分析与预测,没有充分考虑社交网络结构中用户之间的相互影响和兴趣相似度.文献[7]针对当前对微博转发行为预测主要是对所有微博用户的历史数据进行学习得到转发模型,进而完成对所有用户转发行为的全局预测,存在同质性且无法对具体用户进行个性化预测的缺陷,提出基于多任务学习的个性化微博转发行为预测算法.文献[8]针对社交网络中以当前用户为中心的局部区域内,其他用户转发行为对当前用户转发行为的影响进行研究,发现当前用户转发行为易受其直接关注的其他用户的转发行为影响,利用逻辑回归验证了上述结论.文献[9]提出了社交网络中信息扩散时,由于用户之间行为的相互影响而造成信息扩散范围最大化的问题,指出了在独立级联模型中,这类问题 NP-hard 和子模的性质,进而给出解决方法,取得良好的效果.

与以上算法不同,本文在对影响用户转发行为的特征进行综合分析的基础上,采用集成学习策略融合了 Logistic 回归、支持向量机、BP 神经网络等多种机器学习方法以构建更可靠的用户转发行为预测模型.实验结果表明,本文算法可有效提高用户转发行为预测精度,整体上具有较高的性能.

2 问题定义

根据新浪微博社交网络的特点,用户转发行为预测问题的定义如下:对于社交网络 $G=\{V,E\}$,(V 为用户的集合,E 为用户之间关注关系的集合),在已知用户发布或转发微博历史数据的情况下,如果用户 $u_o \in V$ 发布或转发一条微博 T,如何可靠地预测其粉丝是否转发微博 T,或转发微博 T 的概率?

3 影响用户转发行为特征分析

在用户转发微博的过程中,用户属性、社交关系与微博内容等三类特征是影响其转发行为的关键因素.

3.1 用户属性特征

用户属性特征仅与用户个体有关,可以从用户个人信息中直接提取或计算获得,具体包括用户发布微博数、转发活跃度、用户的 PageRank 值、是否认证、是否成为达人等.其中,用户发布微博数越多,表示用户更愿意在社交网络上表达自己的观点和分享自己的生活动态,遇到感兴趣的微博时,往往会表现出主动转发的趋向,用户转发活跃度 f_r 定义为:

$$f_r = T_r / T, \tag{1}$$

其中,T 为用户发布(原创或转发)的微博总数,T_r 为用户转发的微博总数.

从(1)式可知,f_r 越大,则表明用户转发微博的积极性越高,更可能转发感兴趣的微博[10].此外,用户的 PageRank 值、是否为认证用户和是否成为达人等特征一定程度上体现了用户在社交网络上的影响力,对用户的转发行为也有着重要影响.

3.2 社交关系特征

社交关系特征体现了社交网络中用户之间的关系特性,包括用户的关注数、粉丝数、与上游用户的交互强度等.其中,用户关注数越多,表明用户通过社交网络交友和获取信息的愿望越强烈,主观上转发所关注用户发布微博的可能性越大;用户粉丝数越多,则其主观上更可能转发当前微博以扩大个人影响力;用户与其关注用户的交互强度是影响用户转发行为的重要因素,具体定义为:

$$f_{uv} = T_{uv} / T_u, \tag{2}$$

其中,U 与 V 分别为当前用户及其关注用户,T_{uv} 表示用户 V 的微博在用户 U 的转发微博中出现的次数,T_u

表示用户 U 的转发微博总次数.从(2)式可知,f_{uv} 的值越大,用户 U 与用户 V 的交互强度越大,则用户 U 转发用户 V 的微博的可能性也越大.

3.3　微博内容特征

微博内容包含了用户发布微博的方式和内容,具体包括微博长度、微博发布时间、微博是否包含 URL、是否包含主题 hashtag、是否@某个用户、转发数、评论数、点赞数、微博中蕴含的用户兴趣相似度.其中,兴趣相似度表示当前微博内容与用户兴趣偏好的相近程度,而用户的兴趣偏好可以通过分析用户的历史转发微博获取.一般认为,当前微博内容与用户的兴趣偏好越相近,则其被转发的可能性越大.

为了度量用户 U 对微博的内容感兴趣的程度,本文将其历史原创与转发的微博汇集成文档 D,然后采用 LDA(Latent Dirichlet Allocation)模型分别计算文档 D 与微博 T 在预定的 50 个主题(如教育、军事等)上的概率分布,最后利用余弦距离确定相应的主题相似度[11],即:$L(D,T)=\dfrac{L_{D,A}(D)\cdot L_{D,A}(T)}{\|L_{D,A}(D)\|\ \|L_{D,A}(T)\|}$.

综上所述,本文共提取了用户属性、社交关系与微博内容等 15 个特征对用户的转发行为进行预测,特征数据见表 1.

表 1　影响用户转发行为的特征

特征序号	特征类别	特征名称	特征序号	特征类别	特征名称
1	用户属性	微博数	9	微博内容	转发数
2	用户属性	转发活跃度	10	微博内容	微博长度
3	用户属性	PageRank 值	11	微博内容	微博发布时间
4	用户属性	是否认证	12	微博内容	微博内容中是否包含 URL
5	用户属性	是否成为达人	13	微博内容	是否包含主题 hashtag
6	社交关系	粉丝数	14	微博内容	是否@某个用户
7	社交关系	关注数	15	微博内容	微博内容与用户的兴趣相似度
8	社交关系	与关注用户的交互强度			

4　用户转发行为预测算法

集成学习作为目前流行的机器学习方法,它本身不是一个单独的机器学习算法,而是通过构建并结合多个机器学习器来完成学习任务,即博采众长来达到更好的学习效果.集成学习算法根据作为完成预测任务的个体学习器是同类的或不全是同类的分为同质和异质算法,其中,同质算法依据个体学习器之间是否存在依赖关系又可以分为具有强依赖关系串行生成的 boosting 系列算法,或不存在强依赖关系并行生成的 bagging 和随机森林系列算法;异质算法则将多个不同的个体学习器通过某种组合策略融合在一起.不同个体学习器的组合策略分为平均法、投票法和学习法,其中投票法因简单且可解释性强易于理解被广泛使用.投票法一般分为绝对投票法、相对投票法和加权投票法等,加权投票法是比绝对投票法和相对投票法更加复杂的投票方法,它对每个弱分类器的分类票数乘以一个权重,最终将各个类别的加权票数求和,最大值对应的类别为最终类别.相对于前两种投票法,加权投票法考虑了不同异质分类器的性能差异,组合时赋予它们不同的权重,以充分发挥强分类器在预测时的作用.

为了提高用户转发行为预测的精度,见图 1,本文在 Logistic 回归、支持向量机和 BP 神经网络等机器学习算法的基础上,采用集成学习中被广泛使用的"加权投票法"的组合策略将多个异质学习器的分类结果进行融合,以构成新的用户转发行为预测模型(简称 RBPEL, Retweet Behavior Prediction based on Ensemble Learning).

具体而言,对于微博用户转发行为预测的分类任务,学习器 h_i 将从类别标记集合 $\{c_1,c_2,\cdots,c_N\}$ 中预测出一个标记,本文 N 取值为 2,c_1 代表转发,c_2 代表不转发.RBPEL 采用以下"加权投票法"组合策略对多种机器学习算法进行融合:$H(x)=c_{\underset{j}{\operatorname{argmax}}\sum_{i=1}^{T}w_i h_i^j(x)}$,其中,$h_i$ 表示 Logistic、支持向量机和 BP 神经网络等分类

模型；$h_i^j(x)$ 表示 h_i 在类别 c_j 上的输出；w_i 是 h_i 的权重，通常情况下，$w_i \geq 0$，$\sum_{i=1}^{T} w_i = 1$. 本文经过实验对比发现，相对于 Logistic 回归和支持向量机，BP 神经网络对微博用户转发行为具有更好的预测性能，特别当权重 w_1、w_2 与 w_3 分别设置为 0.2、0.3 与 0.5 时，将 Logistic 回归、支持向量机与 BP 神经网络集成后的算法具有最好的预测性能.

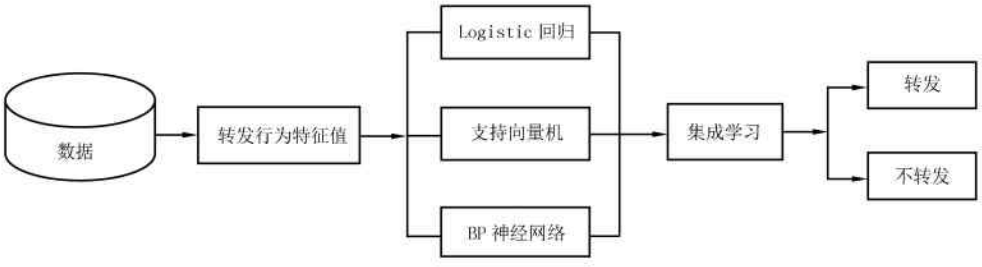

图 1 基于集成学习的用户转发行为预测模型

5 实验结果与分析

5.1 数据集描述

本文采用了文献[8]中公开的数据集对算法的可行性进行了验证.新浪微博提供了可以获取用户属性、好友列表、微博帖子内容等方面信息的各种 API 开放接口.由于新浪微博对没有通过审核的开发账户有每小时仅 150 次访问权限的访问频次限制，且一直在缩小 API 对未经用户授权情况下用户数据的获取范围，为获取海量微博用户数据，该数据集采用对新浪微博网页分布式抓取并解析的方式采集用户数据，抓取过程中，不仅对新浪微博服务器进行模拟浏览器登录操作，而且为提高抓取效率采用基于 Actor 模型的分布式爬虫.爬虫程序首先从全网中随机选取 100 个用户，然后抓取他们的关注者以及关注者的关注者，接着对抓取到的每个用户收集其个人用户属性信息，同时抓取其最近发表的 1 000 条微博相关信息.最终收集了 2012 年 8 月 28 日至 2012 年 9 月 29 日一个月时间抓取用户的数据，共包含 1 776 950 个用户和 308 489 739 条关注关系，以及从用户发帖中随机挑选了转发次数较多的 300 000 个原帖及其 23 755 810 个转帖.

针对上述数据统计分析发现，每个用户发表原帖的平均转发次数为 80 次；拥有超过 200 万个以上粉丝的用户只有 0.05%，这与以前发现的二八定理的规律相吻合，即 20%的用户控制了 80%的粉丝和信息传播，用户的平均粉丝为 1 006 人；新浪微博对用户关注人数有 3 000 人的上限要求，除了一小部分达到上限，用户关注人数平均为 161 人，满足 150 定律，即著名的“邓巴数字”，人们受限于时间和精力，能维持稳定关系的好友数量为 150 人左右.

在预测过程中，采用十折交叉验证，将全部数据划分为 10 个大小相似的互斥子集，每次用 9 个子集的并集作为训练集，余下的那个子集作为测试集，这样可以获得 10 组训练/测试集，从而进行 10 次训练和测试，最后返回 10 个测试结果的平均值.

5.2 特征值归一化

对于影响用户转发行为的特征向量，为了消除不同特征之间数值类型(如离散与连接型)与取值范围的差异，本文对其进行了规范化处理，即：$x^* = \dfrac{x - x_{\min}}{x_{\max} - x_{\min}}$，其中，$x$ 与 x^* 分别为初始特征与归一化后的特征，$x_{\min}$ 与 $x_{\max}$ 分别为所有用户当前特征的最小值与最大值.

5.3 评测指标

为了评价转发行为预测模型的有效性，本文采用信息检索中的查全率、准确率与 F_1 度量等标准对预测结果进行评估.其中，准确率为被正确预测为转发微博的数量占全部被预测为转发微博数量的比例；查全率为被正确预测为转发微博的数量占实际转发微博中全部微博数量的比例；而 F_1 度量为是一个综合性指标，

相应的计算公式为 $F_1=2\times$正确率$\times$查全率/(正确率+查全率).此外,本文还采用 ROC 曲线分析了各类特征对预测结果的影响程度.ROC 曲线以"真正确率"为纵坐标,以"假正确率"为横坐标.如果预测算法对应的 ROC 曲线偏向左上角,则表明该算法预测结果越好.

5.4　结果分析

下面将介绍不同机器学习算法对用户转发行为的预测结果,描述 RBPEL 集成学习算法在不同权重取值下的预测结果,分析不同特征对用户转发行为的重要程度,通过实验结果,可以看出本文提取的特征与构建的模型,在预测微博用户转发行为方面具有较高的性能.

5.4.1　用户转发行为预测

根据表 1 所示的特征,本文分别对 Logistic 回归、支持向量机(SVM)、BP 神经网络以及集成学习算法 RBPEL 进行了实验部分对比.表 2 所示实验结果可以看出,相对于 Logistic 回归、支持向量机(SVM)、BP 神经网络等机器学习算法,集成学习算法 RBPEL 在 F_1 度量的综合评价上获得了较好的结果.

表 2　用户转发行为预测结果

算法	准确率	查全率	F_1 度量
Logistic 回归	0.972 5	0.764 1	0.855 8
支持向量机	0.968 3	0.852 0	0.906 4
BP 神经网络	0.971 5	0.892 7	0.930 5
RBPEL 算法	0.979 1	0.913 9	0.945 4

5.4.2　RBPEL 算法不同权重取值对预测精度的影响

表 3 列出了集成学习算法在 w_1,w_2,w_3 取不同权重值时的预测结果.其中,BP 神经网络相对于 Logistic 回归和支持向量机具有较高的 F_1 度量综合评价值,在 RBPEL 算法集成学习过程中,应具有较高的权重,才能发挥 BP 神经网络在集成时的作用,整体提高预测的性能.从表 3 可以看出 RBPEL 算法在 w_1,w_2,w_3 分别取值为 0.2、0.3、0.5 时具有最好的预测效果.

5.4.3　不同特征对用户转发行为的影响程度分析

为了进一步考查不同特征对用户转发行为预测精度的影响程度,如表 4、表 5 与表 6 所示,本文也给出了分别采用用户属性特征、社交关系特征和微博内容特征对用户的转发行为进行预测的结果.

表 3　RBPEL 在取不同值时的预测结果

类别	结果			
	(0.2,0.3,0.5)	**(0.1,0.2,0.7)**	**(0.2,0.4,0.4)**	**(0.7,0.2,0.1)**
准确率	0.979 1	0.971 5	0.970 1	0.972 5
召回率	0.913 9	0.892 7	0.842 9	0.764 1
F_1 度量	0.945 4	0.930 5	0.902 0	0.855 8

表 4　用户属性特征预测结果

学习方法	准确率	查全率	F_1 度量
Logistic 回归	0.947 4	0.657 4	0.776 2
支持向量机	0.953 3	0.669 9	0.786 9
BP 神经网络	0.942 8	0.722 8	0.818 2

表 5　社交关系特征预测结果

学习方法	准确率	查全率	F_1 度量
Logistic 回归	0.965 4	0.649 3	0.776 4
支持向量机	0.968 6	0.598 3	0.739 7
BP 神经网络	0.956 9	0.852 8	0.901 8

表 6　微博文本特征预测结果

学习方法	准确率	查全率	F_1 度量
Logistic 回归	0.899 7	0.449 6	0.599 5
支持向量机	0.924 4	0.377 0	0.535 6
BP 神经网络	0.900 3	0.576 5	0.702 9

图 2 显示了用户属性特征、社交关系特征和微博内容特征对用户转发行为预测结果影响的 ROC 曲线.实验结果表明,在以上 3 种特征中,对用户转发行为影响最大的是社交关系特征,微博内容特征影响较小.其主要原因在于:在微博社交网络中,用户更多以社交为需求分享自己关注的感兴趣用户的微博,而对微博本身内容则关注较少.

6　结　论

为了提高微博社交网络中的用户转发行为预测精度,本文对影响用户转发行为的特征进行了综合分析,

并在 Logistic 回归、支持向量机、BP 神经网络的基础上采用集成学习方法构建了更可靠的用户转发行为预测模型.实验结果显示,本文算法可以有效提高用户转发行为预测精度,整体上具有较高的性能.当前,本文算法的缺点及主要改进之处在于:①用于预测的特征数量与特征组合为人工确定,如何自动地选择特征及最优组合是一个需要深入研究的问题;②在采用集成学习组合策略融合多种机器学习方法时,针对微博这种由大量用户数据引起的大数据问题,可以采用"学习法"中的 Stacking 方法,在多个初级学习器的基础上,构建一个预测性能更高的次级学习器.在下一步的工作中,拟采用深度学习方法解决特征选择问题,同时通过融入更多数据类型(如用户所发图片)以进一步提高用户转发行为预测的精度.

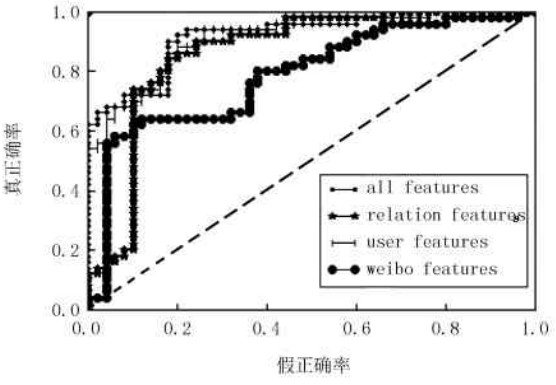

图 2 不同特征的 ROC 曲线

参 考 文 献

[1] Boyd D,Golder S,Lotan G.Tweet,tweet,retweet:Conversational aspects of retweeting on Twitter[C].Hawaii:Proceedings of the 43rd Hawaii International Conference on System Sciences,2010:1-10.

[2] Suh B,Hong L,Pirolli P,et al.Want to be retweeted large scale analytics on factors impacting retweet in twitter network[C].Minneapolis:IEEE second International Conference on Social Computing, 2010:177-184.

[3] Yang Z,Guo J Y,CaiK K,et al. Understanding retweeting behaviors in social networks[C].Toronto: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010:1633-1636.

[4] 张砀,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.

[5] 曹玖新,吴江林,石伟,等.新浪微博网信息传播分析与预测[J].计算机学报,2014,37(4):780-790.

[6] 孔庆超,毛文吉,张育浩.社交网站中用户评论行为预测[J].智能系统学报,2015,10(3):1-5.

[7] Tang X,Miao Q G,Quan Y N,et al.Predicting individual retweet behavior by user similarity: A Multi-Task Learning Approach [J].Knowledge-Based Systems,2015,89:681-688.

[8] Zhang J, Tang J,Li J Z,et al.Who influenced you? predicting retweet via social influence locality[J].ACM Transactions on Knowledge Discovery from Data,2015,9(3):1-26.

[9] Wang Z F,Yang Y,Pei J,et al. Activity maximization by effective information diffusion in social networks[J].IEEE Transactions on Knowledge and Data Engineering,2017,29(11):2374-2387.

[10] 刘立波,任静,周杰,等.基于句型分类的清真食品评论倾向性判别[J].河南师范大学学报(自然科学版),2017,45(4):97-102.

[11] 王冲,纪仙慧.基于用户兴趣与主题相关的 PageRank 算法改进研究[J].计算机科学,2016,43(3):275-277.

Predicting microblog user retweet behaviors based on ensemble learning

Zhang Xiaowei, Wang Wei, Qin Dongxia

(School of Network Engineering,Zhoukou Normal University,Zhoukou 466001,China)

Abstract: In order to improve the accuracy of predicting userretweet behaviors in a micorblog social network, the paper proposes an effective method based on Ensemble Learning. Firstly, the papercomprehensively analyzes the performances of various features that affect user retweetbehaviors, such asuser attributes, social relationships and microblog contents,et al. Based on the extracted features, the propsed method respectively predictsuser retweetbehaviors with Logistic regression, SVM (Support Vector Machine) and BP (Back Propagation) neural network, and incorporates the corresponding results in a weighted voting manner based on Ensemble Learning.The experimental results show that the propsed method has a performance improvement of 1.5% on F_1 metric of the overall evaluation,compared with the BP neural network.

Keywords: sinamicroblog; retweetbehavior prediction; ensemble learning; socialrelation

[责任编校 陈留院]