基于随机森林的微博互动特征分析

于　澍,曹　琦,刘　涛
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘　要:微博凭借其开放性、低门槛已成为最常用的社交媒体平台之一,其海量数据背后蕴藏着巨大的价值亟待研究。而准确地判断微博的传播趋势,降低不良微博带来的影响已成为当前面临的主要问题。文中以新浪微博为研究对象,将随机森林算法与数据分析处理相结合,对微博的博文发布一周后的转评赞行为进行预测,将数据特征分为三类并分析了每类特征对预测结果的影响。首先,简述了决策树及随机森林算法的原理;其次,对微博数据进行分析,将提取的特征分为用户特征、时间特征和文本类特征三类;最后,通过三组对比实验验证了随机森林算法在微博互动预测上的可行性,并分析了三类特征对预测结果的影响。实验结果表明,用户特征对预测准确率的影响较大。

关键词:数据挖掘;随机森林;机器学习;数据分析;决策树
中图分类号:TP181　　**文献标识码**:A　　**文章编号**:1673-629X(2019)10-0051-04
doi:10.3969/j.issn.1673-629X.2019.10.011

Analysis of Interactive Characteristics of Weibo Based on Random Forest

YU Shu,CAO Qi,LIU Tao
(School of Computer and Information Technology,Northeast Petroleum University,Daqing 163318,China)

Abstract:Weibo has become one of the most commonly used social media platforms due to its openness and low threshold,and the huge value behind its massive data needs to be studied. To accurately judge the spread trend of Weibo and reduce the impact of bad Weibo has become the main problem. Taking Sina Weibo as the research object,we combine random forest algorithm with data analysis and processing to predict the behavior of the review and praise of Weibo after one week of blog post release. We divide data features into three categories and analyze the influence of each type of features on the predicted results. Firstly,the principle of decision tree and random forest algorithm is briefly described. Secondly,the microblog data is analyzed,and the extracted features are divided into three categories:user feature,time feature and text class feature. Finally,three sets of contrast experiments are verified. The feasibility of the random forest algorithm in the interactive prediction of Weibo,and the influence of the three types of features on the prediction results are analyzed. The experiment shows that the user feature has a greater impact on the accuracy of prediction.

Key words:data mining;random forest;machine learning;data analysis;decision tree

0　引　言

随着移动通讯技术的日趋完善,一大批社交媒体平台不断涌现[1],已经成为人们沟通交流、获取信息的重要平台,影响着人们的工作和生活。微博凭借其传播速度快、内容覆盖领域广和低门槛等特性近年来迅速发展成为网民结交好友、获取新闻时事、自我分享及表达的重要社交媒体[2-3]。国内目前存在大量的微博网站,例如新浪微博、腾讯文博、搜狐微博等,其中新浪微博最受广大用户喜爱,也是人们目前最为常用的社交媒体平台。

大量的活跃用户在微博上产生大量的行为信息,海量数据背后蕴藏着巨大的学术研究价值。对于企业而言,通过观测微博用户的在线行为可以了解用户的兴趣爱好和上网习惯,以有效指导企业调整更新产品,为大众提供更好的服务;对于政府部门而言,可以通过用户行为及时了解大众关注的焦点以及对待热点问题的态度,准确判断舆论走向,以便及时采取科学的引导和有效的控制。因此提前预测微博的互动情况,对于企业和社会而言有着重要的意义。

机器学习是一门涉及到多个领域的交叉学科,致

收稿日期:2018-12-18　　**修回日期**:2019-04-24　　**网络出版时间**:2019-06-26
基金项目:国家自然科学基金面上项目(51774090);黑龙江省自然科学基金面上项目(F2015020);黑龙江省教育科研专项引导性创新基金项目(2017YDL-12);黑龙江省教育规划重大课题(GJ20170006)
作者简介:于　澍(1992-),女,硕士研究生,研究方向为机器学习、数据分析。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190626.0823.014.html

力于模拟人类利用经验做出有效决策的学习行为,使计算机能够利用经验不断改善系统本身的性能,以获取人类观测不到的新的知识。文中将机器学习中的随机森林算法应用到微博数据上,以新浪微博为研究对象,对微博数据进行处理及分析,并预测用户发布微博一周后的转发数、评论数和点赞数。

1 相关工作

国内外有大量的学者对在线社交网络的信息传播行为进行了研究。文献[3]提出了三类综合特征,使用机器学习中的分类方法对给定微博的用户转发行为进行预测。Liben-Nowell 等[4]研究了一系列有关信息在真实社会网络中传播的特征,得出精确预测出信息的传播路径在当前技术发展的情况下是非常困难的。还有一些文献[5-10]在情感分析等方面进行了研究。

1.1 决策树算法

决策树算法是一种有监督的机器学习算法,常用于分类预测等诸多领域[11]。决策树是一个树形结构,其每个非叶节点表示一个特征属性上的测试,每个分支代表该特征属性在某个值域上的输出,而每个叶节点存放一个类别,最终产生一棵泛化能力强的决策树。其中节点分裂特征的选择为构造一棵决策树的关键。根据不同的划分标准,相关学者提出了不同的决策树算法,如基于信息熵的 ID3 算法、基于增益率的 C4.5 算法和基于基尼指数 CART 决策树算法。文中主要介绍 CART 决策树算法。

CART 分类回归树是一种二叉决策树,既可处理连续型数据又可处理离散型数据。分类树根据基尼值来度量数据集 S 的纯度,即决策树的分支节点尽可能包含同一类别的样本,其值越小,数据集的纯度越高。基尼值表示为:

$$\mathrm{Gini}(S) = 1 - \sum_{k=1}^{K} p_k^2 \tag{1}$$

其中,p_k 为当前样本集合 S 中第 k 类样本所占的比例。

然后计算属性集 A 中每个属性 a 的基尼指数,从中选择基尼指数最小的属性作为最优的划分属性。基尼指数表示为:

$$\mathrm{Gini\_index}(S,a) = \sum_{v=1}^{V} \frac{|S_v|}{|S|}\mathrm{Gini}(S_v) \tag{2}$$

其中,S_v 表示 S 中属性 a 上取值为 v 样本子集。

1.2 随机森林算法

集成学习是通过对每个基学习器的学习结果进行组合的方式将多个学习器聚集起来,形成具有更好性能的学习器。集成学习可以有效地提高学习系统的泛化能力[12]。其中一类就是以 Bagging 和随机森林为代表的。

随机森林就是在构建 Bagging 集成的基础上将决策树作为基学习器[13-14],与传统的决策树不同的是,随机森林在选择划分属性时,是从全部的特征中均匀随机地抽取一个特征子集,然后再从这个子集中选择一个最优的分裂特征。随机森林构造了多棵决策树,一般来说,分类问题由每棵决策树投票决定其最终分类,回归问题则取其平均值作为最终结果。

1.3 中文分词

一个中文词语由两个以上的汉字组成,所以,对中文的文本进行分析时,计算机很难区分词语、成语或谚语。中文分词又称中文切词,是指把一条中文语句切分成若干个有意义的词语[15]。中文分词技术属于自然语言处理的范畴,目前中文分词工具有很多,已有相关研究采用 540 篇分别来自新闻、微博、汽车之家和大众点评的数据对常见分词工具进行测试[16],测试结果见表 1。

表 1 中文分词工具对比

分词工具	分词粒度	接口	出错情况	词性标注
IK Analyzer	多模式	jar 包	无	
NLPIR	多模式	多语言接口	有	√
SCWS	多模式	PHP 库/命令行工具	无	√
JIEBA	多模式	Python 库	无	√
盘古分词	多模式	无	无	
庖丁解牛	多模式	jar 包	无	
搜狗分词	小	支持上传文档但失败率高	有	√
腾讯文智	小	REST API	有	√
新浪云	大	REST API	无	√
语音云	适中	REST API	无	√

实验部分采用 Python 语言来处理微博数据中的中文文本,用于统计博文内容的词频,选取高频词作为特征。Jieba 分词适用于 Python 环境,因此文中选择 Jieba 分词工具。

2　实　验

实验将提取的特征分为三类,应用随机森林算法预测微博在发表一周后的互动情况,分析随机森林模型的应用效果,并对比三类特征对预测结果的影响。实验数据来自于天池大数据竞赛:新浪微博互动预测,训练数据约为 122 万条微博数据,部分数据如表 2 所示。预测数据约为 17 万条微博数据。实验工具采用 anaconda3,编程语言为 Python3.7。

表 2　训练数据

微博 ID	用户 ID	发送时间	转	评	赞	博文内容
7d45833d9865727a88	d38e9bed5d98110dc24	2015-02-23 17:41:29	0	0	0	丽江旅游(sz002033)#股票##炒股##财经#推荐包赢
68cd0258c31c2c525f9	da534fe87e7a52777be	2015-03-31 13:58:06	3	7	9	#xx 的红包#二十三,糖瓜儿粘,抢个红包乐翻天!
00b9f86b4915aedb7d	e06a22b7e065e559a1f	2015-06-11 20:39:57	0	4	8	如此平凡的日常一幕,还能够再积累多少呢…
…	…	…	…	…	…	…
f359a74cb4ac6150a3a	fbe6c953632e1b3dda6	2015-07-09 17:22:50	6	8	5	卖水果老人因没住处夜宿酒店门口被车碾死 http://t.cn/…

实验部分包括三组对比实验,实验流程如下:

Step1:观察并分析数据,提取三类特征,分别为用户特征、时间特征、博文内容特征,如表 3 所示。其中高频词提取结果如图 1 所示,取前 20 个为高频词汇。根据提取的特征建立预测的训练集与测试集。

表 3　特征表

类别	特征
用户类特征	平均转发数
	平均评论数
	平均点赞数
	关注度
	活跃度
时间类特征	是否是工作日
	是否是工作时间
文本类特征	是否有链接
	是否有标题
	是否有表情
	是否有@
	高频词
	文本长度

Step2:进行三组对比实验。

实验一:无特征简单平均法、三类特征决策树模型、三类特征 RF 模型以及线性回归模型的对比实验。

实验二:基于 RF 的三类特征对比实验。分别训练无用户特征、无时间特征、无博文内容特征的 RF 模型。

实验三:基于 RF 的博文特征对比实验。对比文本类特征中的六项特征对预测结果的影响。

Step3:将训练好的模型应用于测试集,得出预测结果后根据评估标准验证模型的预测准确率。

Step4:分析三组实验的结果,得出实验结论。

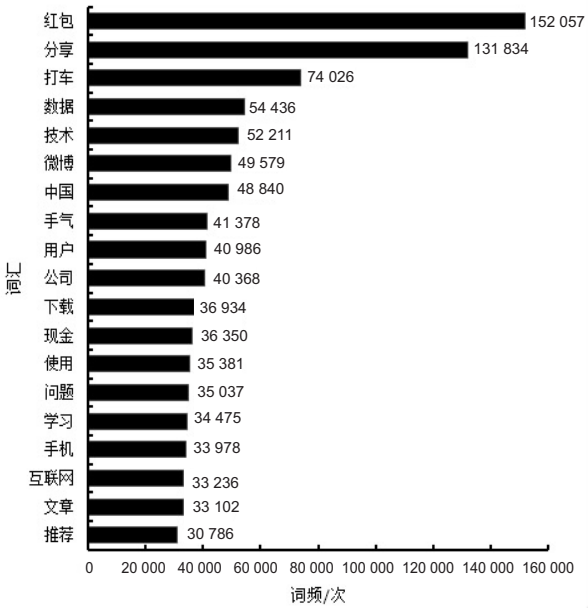

图 1　词频统计

2.1　评估标准

计算第 i 篇博文的准确率 p_i:

$$p_i = 1 - 0.5 * d_f - 0.25 * d_c - 0.25 * d_l \quad (3)$$

其中,d_f、d_c、d_l 分别表示转发偏差、评论偏差和点赞偏差,分别表示如下:

$$d_f = \frac{c_{fp} - c_{fr}}{c_{fr} + 5} \quad (4)$$

$$d_c = \frac{c_{cp} - c_{cr}}{c_{cr} + 3} \tag{5}$$

$$d_l = \frac{c_{lp} - c_{lr}}{c_{lr} + 3} \tag{6}$$

其中，c_{fp} 为预测的转发数，c_{fr} 为实际的转发数；c_{cp} 为预测的评论数，c_{cr} 为实际的评论数；c_{lp} 为预测的点赞数，c_{lr} 为实际的点赞数。

根据每篇博文的准确率 p_i 计算最终预测的准确率 P：

$$P = \frac{\sum_i^N (C_i + 1) * \mathrm{sgn}(P_i - 0.8)}{\sum_i^N (C_i + 1)} \tag{7}$$

其中，sgn(x) 为第 i 篇博文的总转发、评论、点赞之和，当 C_i >100 时，取值为 100。

2.2 实验结果分析

由实验一的结果表明，随机森林算法可以应用到微博数据上，且随机森林算法预测模型的准确率比决策树等算法的相对准确率稍高，实验结果如表 4 所示。

表 4 四种算法预测结果对比

算法	P/%
简易平均法	27.35
决策树	26.24
线性回归	27.27
随机森林	28.39

通过实验二的结果可得出，三类特征中用户特征对预测结果的影响较大，其次是时间特征，文本类特征影响较小，实验结果如图 2 所示。通过实验三得出，博文内容特征中的“是否含有链接”会对预测结果产生负影响，即导致预测结果的准确率下降。其他博文内容特征对预测结果的影响较小。

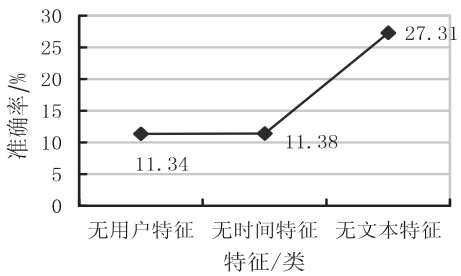

图 2 预测结果

3 结束语

文中将随机森林算法应用到微博数据上，从数据中提取了部分特征并将特征分为三类，由实验结果分析了这三类特征对预测结果的影响，并对其中的文本数据进行了进一步的分析和实验，分析并对比了文本类特征对预测结果的影响。预测结果与数据及特征有关，文中特征是由人工提取，具有一定的局限性，还有待进一步的完善。

参考文献：

[1] 丁兆云，贾 焰，周 斌. 微博数据挖掘研究综述[J]. 计算机研究与发展，2014，51(4)：691-706.

[2] 李 洋，陈毅恒，刘 挺. 微博信息传播预测研究综述[J]. 软件学报，2016，27(2)：247-263.

[3] 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)：779-790.

[4] LIBEN-NOWELL D，KLEINBERG J. Tracing information flow on a global scale using Internet chain-letter data[J]. Proceedings of the National Academy of Sciences of the United States of America，2008，105(12)：4633-4638.

[5] BOYD D，GOLDER S，LOTAN G. Tweet，tweet，retweet：conversational aspects of retweeting on twitter[C]//Hawaii international conference on system sciences. [s. l.]：IEEE，2010：1-10.

[6] ZAMAN T，FOX E B，BRADLOW E T. A Bayesian approach for predicting the popularity of tweets[J]. Annals of Applied Statistics，2014，8(3)：1583-1611.

[7] PANG Bo，LEE L. Opinion mining and sentiment analysis[J]. Foundations & Trends in Information Retrieval，2008，2(1-2)：1-135.

[8] CHANG P S，TING I H，WANG S L. Towards social recommendation system based on the data from microblogs[C]// International conference on advances in social networks analysis and mining. Kaohsiung，Taiwan：IEEE，2011：672-677.

[9] 张 华. 基于优化 BP 神经网络的微博舆情预测模型研究[D]. 武汉：华中师范大学，2014.

[10] 周胜臣，瞿文婷，石英子，等. 中文微博情感分析研究综述[J]. 计算机应用与软件，2013，30(3)：161-164.

[11] 梁 循. 数据挖掘算法与应用[M]. 北京：北京大学出版社，2006.

[12] 于 玲，吴铁军. 集成学习：Boosting 算法综述[J]. 模式识别与人工智能，2004，17(1)：52-59.

[13] 方匡南，吴见彬，朱建平，等. 随机森林方法研究综述[J]. 统计与信息论坛，2011，26(3)：32-38.

[14] 刘 凯，郑山红，蒋 权，等. 基于随机森林的自适应特征选择算法[J]. 计算机技术与发展，2018，28(9)：101-104，111.

[15] 龙树全，赵正文，唐 华. 中文分词算法概述[J]. 电脑知识与技术，2009，5(10)：2605-2607.

[16] 王婧雅. 微博数据挖掘可视化系统的设计与实现[D]. 长春：吉林大学，2017.