融合异常检测与随机森林的微博转发行为预测方法

周先亭[1] 黄文明[2] 邓珍荣[2]
(桂林电子科技大学计算机与信息安全学院 桂林 541004)[1]
(桂林电子科技大学广西可信软件重点实验室 桂林 541004)[2]

摘 要 针对目前微博转发行为预测具有的特征选择任意性、准确率不高的问题,提出了融合异常检测与随机森林的微博转发行为预测方法。首先,提取用户基本特征、博文基本特征、博文内容主题特征,并基于相对熵计算用户活跃度、博文影响力;其次,通过结合过滤式与封装式特征选择方法筛选出关键特征组;最后,融合异常检测与随机森林算法,依据筛选后的关键特征组进行微博转发行为预测,并利用袋外数据误差估计设置随机森林中的决策树和特征数。在真实新浪微博数据集上与基于逻辑回归、决策树、朴素贝叶斯、随机森林等算法的微博转发行为预测方法进行实验对比,结果表明所提方法的预测准确率(90.5%)高于基准方法中最优的随机森林方法的预测准确率,同时验证了特征筛选方法的有效性。

关键词 转发预测,随机森林,异常检测,特征筛选,相对熵
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.07.034

Micro-blog Retweet Behavior Prediction Algorithm Based on Anomaly Detection and Random Forest

ZHOU Xian-ting[1] HUANG Wen-ming[2] DENG Zhen-rong[2]
(School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China)[1]
(Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China)[2]

Abstract Aiming to solve the issue that the accuracy of micro-blog retweet behavior prediction is not good enough and features are selected with an arbitrary choice, a new method using anomaly detection and random forest algorithms to predict micro-blog retweet behavior was proposed. Firstly, the basic features of the user, the basic characteristics of blog and blog content theme features are extracted, and the user activity and blog influence are calculated based on relative entropy. Secondly, the best feature set are selected by combining the filter and wrapper feature selection method. Finally, anomaly detection and random forest algorithms are fused to predict micro-blog retweet behavior based on selected features. The algorithm parameters of random forest are selected by analyzing the error estimation of out of bag data. By contrasting with Logistic Regression, Decision Tree, Naive Bias and Random Forest algorithms, which are used in the analysis for micro-blog retweet behavior, the prediction accuracy of the proposed method is higher than that of the optimal random forest method on real data, and reaches 90.5%. Meanwhile, the validity of feature selection method is verified.

Keywords Retweet prediction, Random forest, Anomaly detection, Feature filter, Relative entropy

1 引言

微博[1]是一种以广播的方式实时分享信息的社交网络平台,因具背对脸、便捷性、草根性和内容原创性的特点而受到广大用户的青睐。中国互联网络信息中心发布的《第37次中国互联网络发展状况统计报告》[2]显示,社交应用类中的微博用户已超2亿,且以微博为代表的社交媒体已经建立了与社会网络完全不同的类型区隔,该社交媒体属性逐步得到客户市场和用户市场的认可,并逐渐成长为社交媒体领域最具备营销传播效果的社会化媒体平台。在微博网络中,信息的传播主要是通过微博的转发来实现的,通过用户的转发行为,微博信息实现了持续性传播且信息量在互联网中急速膨胀。因此,对微博转发行为的研究对于微博信息在用户之间的推荐、微博用户行为和兴趣、突发事件预测、敏感信息控制、网络舆情监控以及产品营销等方面都有重要的研究价值和实际意义。

到稿日期:2016-06-19 返修日期:2016-09-22 本文受广西科技攻关项目(桂科攻1598019-6)资助。
周先亭(1990—),男,硕士生,主要研究方向为数据挖掘、机器学习;**黄文明**(1963—),男,教授,主要研究方向为大数据处理、图形图像处理、软件工程,E-mail:wmhuang@guet.edu.cn(通信作者);**邓珍荣**(1977—),女,硕士,研究员,主要研究方向为计算机软件架构、数据挖掘。

近年来，针对微博转发行为预测准确率不高以及特征选择任意性的问题，国内外学者展开了广泛且深入的研究。Petrovic 等人[3]通过人工实验证明了微博转发预测的可行性，然后利用改进的 passive-aggressive 算法预测转发，但正确率仅为 46.6%。Morchid 等人[4]的研究表明，选取的特征如果有较高的辨别能力，预测算法的性能则会得到有效的提高。Yang 等人[5]通过半监督因素图模型进行了转发预测，虽然使用统计学方法分析了对转发行为有影响的因素，但在预测时没有剔除对预测准确率影响较大的冗余甚至无效特征。Romero 等人[6]分析了博文主题和传播机制，发现博文主题是微博被转发与否的重要因素。张旸等人[7]提出了一种微博转发行为预测方法，即首先将微博转发转换为二元分类问题，然后使用支持向量机(Support Vector Machine，SVM)算法对加权后的各特征进行训练。但该模型未考虑微博主题特征，且仅有 85.9%的总体命中率。李英乐等人[8]使用 SVM 算法对提取的 5 种特征进行训练，并构建了相应的预测转发规模的模型。但该模型未考虑微博主题的相关特征；模型中特征计算相对复杂，导致其不适合大数据量的预测；并且对数据采集有严格的时间要求，不具有一般性。赵煜等人[9]使用过采样技术来平衡数据集，再利用随机森林算法进行转发预测，但预测准确率反而下降了 2%。

本文对微博转发行为的预测进行研究。由于可提取的特征较多，因此根据经验首先分析了用户活跃度、博文主题、博文影响力以及其它(如粉丝数量与用户性别等)基本特征；其次，因为弱辨别能力的特征会导致设计的预测模型性能低下，所以通过结合过滤式与封装式两种特征选择算法进行实验分析，从而得到关键特征组；最后，基于筛选后的关键特征，提出一个融合异常检测与随机森林的微博转发行为预测算法，较好地实现了对微博转发行为的预测。与基于逻辑回归、随机森林、决策树、朴素贝叶斯等算法的微博转发行为预测方法的对比表明，所提方法效果最优，能够对微博转发行为进行有效的预测。

2 问题描述

定义 1 已知微博网络 $G=\langle V,E\rangle$，其中 V 表示用户集合，E 表示关注边集合($E=\{e=\langle v_a,v_b\rangle \mid v_a,v_b\in V,v_a\neq v_b\}$)，若用户 v_b 关注 v_a，则关注边定义为：$v_b\rightarrow v_a$。

基于以上定义描述，研究的主要问题可表示为：若用户 b 关注用户 a，即 $v_b\rightarrow v_a$，当用户 a 发布一条博文信息 m 时，预测该博文信息 m 是否会被关注者 b 转发。用 $y=f(v_a,v_b,m)$ ($y\in\{1,0\}$)来形式化表示，其中 $y=1$ 表示转发，$y=0$ 表示不转发，原问题转化为二分类问题。同时，对特征的提取与筛选进行研究。

3 特征提取

微博的转发行为使得原始信息向新用户传播，并以指数级逐级传播。因此，原始信息所包含的一些价值或属性与转发行为有密切的联系。

Suh 等人[10]使用主成分分析与广义线性分析研究了对转发行为有影响的各类微博特征。通过实验发现，用户注册时间对转发行为的影响较小；博文内容是否包含话题和是否包含超链接(Uniform Resource Locator，URL)、微博用户关注他人数量和拥有的粉丝数量等对转发行为有积极影响；博文中是否含有“@”符号以及用户历史微博数量则对转发行为有消极影响。Hu[11]通过在 Twitter 数据上选取特定的话题进行试验，发现一天之内各时间段所发布的微博具有不同的特点，例如早上八九点的微博更容易获得关注且博文较长，而微博的发布量在下午一点时最大。

综合前人经验，选择从用户特征、博文特征、内容主题特征这 3 个方面提取特征。

3.1 用户特征

不同类型的用户具有不同的用户特征，而不同的用户特征对微博信息的传播具有不同的影响。

3.1.1 用户基本特征

用户基本特征反映了微博用户的一些个人信息以及社会信息。例如，不同的粉丝数量可以区分微博用户类型，而不同类型用户的博文得到转发的可能性也是不同的，比如明星用户的博文相较于普通用户更有可能得到转发。用户性别的差异对某类微博的转发起到了预示作用，比如女性用户更倾向于转发美容美发与娱乐八卦等信息，而男性用户则偏向于转发体育与数码科技类的博文。

选择提取的用户基本特征为：是否使用昵称、用户关注他人数量、用户性别、拥有粉丝的数量、教育经历、工作经历、用户博文平均被点赞数量、用户博文平均被转发数量、用户博文平均被评论数量、用户个性标签数量、日均发博数。

3.1.2 用户活跃度

用户活跃度主要体现了用户在微博类社交网络上的活动状态。用户的活动行为对扩大其影响力具有积极作用，如添加新的关注、发布新的信息等。将用户关注他人数量、博文总量、粉丝数量特征综合考虑，按照不同的权重计算得到用户活跃度。

由于不同数据产生的方式不同，需要根据数据的特点进行相应处理。对于用户博文数量，使用式(1)计算其日均发博数量；对于用户关注数、用户粉丝数，则使用式(2)进行取对数处理。

$$x_i=\frac{X_i}{T_{last,i}-T_{first,i}} \tag{1}$$

其中，x_i 表示用户 i 的日均发博数量，X_i 表示获取到的用户 i 的博文总数，$T_{last,i}$ 表示获取到的用户 i 的最新发博日期，$T_{first,i}$ 表示获取到的用户 i 的最早发博日期。

$$x_{i,j}=\log(X_{i,j}+1) \tag{2}$$

其中，$X_{i,j}$ 是第 j 类特征的第 i 个数据。由于不同用户的粉丝数量和用户关注数量差别很大，因此使用式(2)将不同数量级的差别调整到合适的范围并进行预处理。

为了定量评测各个特征的重要性，采用了比其他特征算法更简洁、有效的相对熵来计算，它在特征选择中被广泛使

用[12]。特征越重要，其相对熵越大，该特征在接下来的加权模型中就会发挥更大的作用。对于某特征 j_i，设它取值为 $x_0, x_1, \cdots, x_n$，则该特征相对熵的计算公式为式(3)，其中 c_l 代表类别，m 代表类别数目。

$$D(j_i)=\sum_{k=1}^{n}(P(j_i=x_k)\sum_{l=1}^{m}P(c_l|j_i=x_k)\log\frac{P(c_l|j_i=x_k)}{P(c_l)}) \tag{3}$$

不同特征对于一条微博是否会被转发有着明显不同的影响。为了得到更好的结果，需要考虑不同特征的差异，赋予各个特征不同的权重。对特征进行预处理后即可计算各类特征的权重，权重计算方法如式(4)所示。

$$w(j_i)=\sqrt{\frac{D(j_i)}{D_{MEAN}}} \tag{4}$$

其中，$w(j_i)$表示用户 i 的特征 j 的权重；$D(j_i)$表示用户 i 的特征 j 的相对熵，其计算方法如式(3)所示；D_{MEAN} 表示所有特征的平均相对熵。式中的开平方是为了缓和该特征对加权机制的影响。

定义 2　给定用户 v 的日均发博数量 X_{wb}、粉丝数量 X_{fans}、关注数量 X_{follow} 以及对应特征的权重 $w(wb)$，$w(fan)$，$w(fol)$，则用户 v 的活跃度 $ActiveValue$ 为：

$$ActiveValue(v)=w(wb)\times X_{wb}+w(fan)\times\lg(X_{fans}+1)+w(fol)\times\lg(X_{follow}+1) \tag{5}$$

3.2　微博特征

3.2.1　博文的基本特征

博文内容会影响一条微博受欢迎的程度。例如，某些用户可能因为热衷于体育赛事而转发他所见到的体育赛事的相关微博；被@的用户很有可能会转发被他提及的博文；而有些用户因为登录微博的时间习惯不同，一天内的发布时间点和星期发布时间可能很重要，在某些时间段内具有更高的转发概率。

选择提取的博文的基本特征包括：博文发布月份，发布星期，发布小时，是否分享图片，是否为分享，是否为收藏，是否为转发，是否包含关键字“红包”，“抽奖”，“转发”，“教程”，包含话题的数量，“@”他人的数量，包含外链的数量，博文长度。

3.2.2　博文影响力

博文影响力体现了用户博文在微博平台的感召力与说服力，影响力大小会对转发行为产生影响。博文影响力与用户粉丝数量、博文平均被评论数量、被转发数量、被点赞数量有着密切的关系，将这些特征按照不同权重计算得到博文影响力。

首先，需要根据其特点对数据进行相应的预处理。通过式(2)对用户粉丝数量、博文平均被评论数量、被转发数量、被点赞数量进行处理，将较大数量级的差别调整到一个合适的范围内。

基于相对熵，可以定量分析出各特征关键性的区分度。对于不同的特征，需要采用不同的权重，以更好地区分其关键性。通过式(3)计算相对熵，通过式(4)计算出不同特征的权重。

定义 3　给定用户 v 的粉丝数量 X_{fans}、博文的平均被评论数量 X_{comm}、博文的平均被转发数量 $X_{retweet}$、博文的平均被点赞数量 X_{like} 及对应特征的权重 $w(fan)$，$w(comm)$，$w(ret)$，$w(like)$，则用户 v 的博文影响力 $InfluenceValue$ 为：

$$InfluenceValue(v)=w(fan)\times\lg(X_{fans}+1)+w(comm)\times\lg(X_{comm}+1)+w(ret)\times\lg(X_{retweet}+1)+w(like)\times\lg(X_{like}+1) \tag{6}$$

3.3　微博内容主题特征

微博主题构成了信息传播的关键特征，精炼地描述了微博信息。用户通常只关心自己感兴趣的博文主题，其关注程度是转发行为发生与否的重要影响因素。

隐含狄利克雷分布(Latent Dirichlet Allocation，LDA)是由 Blei 等人在《Journal of Machine Learning Research》上发表的基于统计的概率语言模型。LDA 模型对文本主题信息进行建模，是一种典型的有向概率模型[13]，如图 1 所示。为了高效地处理大规模的文档数据集，LDA 模型只保留文档的本质统计信息，对文档进行简短的描述。它由三层生成式贝叶斯网络构成，包含文档、主题、单词这 3 层。该模型基于一个前提假设：文档是通过一些隐含主题构成的，而文本中某些特定词汇构成了这些主题，且忽略了文档中的词语顺序和句法结构[14]。

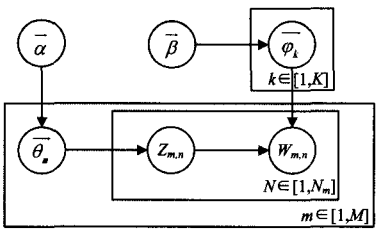

图 1　LDA 模型有向概率图

图 1 中矩形框代表生成过程的重复次数，通常 $\vec{\alpha}$ 和 $\vec{\beta}$ 是固定值并且是对称分布的。$\vec{\alpha}$ 和 $\vec{\beta}$ 是先验超参数，$\vec{\alpha}$ 代表文档数据集中隐含主题间的相对强弱，$\vec{\beta}$ 代表了隐含主题的自身概率分布。$\vec{\varphi}_k$ 和 $\vec{\theta}_m$ 分别用来生成主题与单词，$\vec{\varphi}_k$ 代表主题 k 中的词项概率分布，$\vec{\theta}_m$ 代表第 m 篇文档的主题分布。k 是潜在主题数量，M 是给定集合的文档数量，N_m 表示第 m 篇文档的长度，$W_{m,n}$ 和 $Z_{m,n}$ 分别表示第 m 篇文档中第 n 个单词及其主题。后文算法 1 中的参数含义与此处相同。

根据隐含狄利克雷分布主题模型的生成过程，给定一篇文档集合，文档 m 中的第 n 个单词 $W_{m,n}$ 的生成概率为：

$$P(w_{m,n}=t|\vec{\theta}_m,\vec{\varphi}_k)=\sum_{k=1}^{K}P(w_{m,n}=t|\vec{\varphi}_k)P(z_{m,n}=k|\vec{\varphi}_k) \tag{7}$$

而 LDA 模型生成文档 m 即产生全部 N_m 个单词的生成概率为：

$$P(\vec{w}_m,\vec{z}_m,\vec{\theta}_m,\vec{\varphi}_k|\vec{\alpha},\vec{\beta})=\prod_{n=1}^{N_m}P(w_{m,n}|\overrightarrow{\varphi_{z_{m,n}}},\vec{\theta}_m)\cdot P(\vec{\theta}_m|\vec{\alpha})\cdot P(\vec{\varphi}_k|\vec{\beta}) \tag{8}$$

多篇文档共同组成语料库，其似然计算如下：

$$P(W|\vec{\alpha},\vec{\beta})=\prod_{m=1}^{M}P(w_m|\vec{\alpha},\vec{\beta}) \tag{9}$$

算法 1　LDA 主题模型生成算法

```
# TOPICS
for topic in topics: # topics=k, k∈[1,k]
```

Do sample mixture components $\vec{\varphi_k} \sim Dir(\vec{\beta})$

#DOCUMENTS

for document in documents:

#documents=m,m∈[1,M]

Do sample mixture components $\vec{\theta_m} \sim Dir(\vec{\alpha})$

Do sample document length $N_m \sim Poiss(\xi)$

#WORDS

for word in words:

#words=n,n∈[1,N_m]

Do sample topic index $z_{m,n} \sim Mult(\vec{\theta_m})$

Do sample term for word $w_{m,n} \sim Mult(\overrightarrow{\varphi_{z_{m,n}}})$

#V:vocabulary size

以每个用户发布或转发的博文为文本语料，将语料进行预处理后，即可用来训练博文主题模型。预处理过程：首先通过对中文分词效果较好的“结巴分词”对语料进行分词处理，其次通过停用词字典去掉停用词，并去掉标点符号，接着将英语单词词干化，最后去掉低频词汇。令超参数 $\alpha=\frac{50}{K}$，$\beta=0.1$，话题数 $K=50$，构建主题模型，并根据构建好的主题模型推断博文主题分布概率，将其作为转发预测模型特征之一。

4 转发预测

通过编写 Python 爬虫程序，在随机选择某用户后根据其关注用户逐层爬取新浪微博用户和博文数据，获取了 2015 年 9 月到 12 月的部分新浪微博有效数据 841443 条，用于本文实验。

第 2 节中提取了各类用户特征、博文特征、主题特征；本节将进行特征筛选，并融合异常检测与随机森林算法进行微博转发预测。整个预测主流程如图 2 所示。

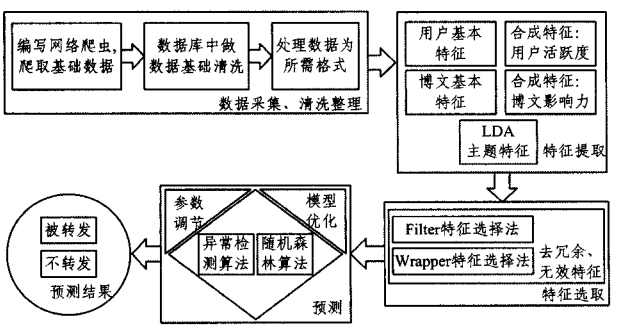

图 2 微博转发预测主流程

4.1 特征筛选

数据集中存在本身与预测无关的特征以及冗余特征，如果选择了几乎不具辨别能力的特征，将导致设计的预测模型的性能低下，但若选择的特征具有充分的辨别能力，则会极大地提高预测模型的预测精度。因此，特征选取过程至关重要。

特征选择的目的是寻找一个能够有效识别目标的最小特征子集，通常从分类正确率及类分布的角度考虑。其框架如图 3 所示。从图 3 所示的基本框架可以看出，特征选择方法首先进行候选特征子集的生成，然后对子集进行评价，若符合停止准则，则进行结果验证，否则重新进行子集生成。

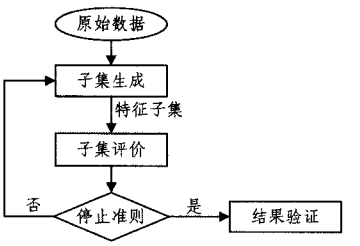

图 3 特征选择的基本框架

按照特征子集的形成过程，搜索策略可分为全局最优、随机搜索和启发式搜索。特征选择方法包括独立于后续学习算法的过滤式(Filter)特征选择方法和需要纳入后续学习算法的封装式(Wrapper)特征选择方法。

为了增强特征与类的相关性并削弱特征间的相关性，Filter 特征选择方法主要使用一致性度量、距离度量、信息度量、依赖性度量 4 种评价准则。Wrapper 特征选择方法的特征重要程度的评价标准是分类性能，它的依据是选择子集最终被用于构造后续模型。该方法准确率高，但时间复杂度也较高，并且泛化能力弱。

本文采用混合选择的方式，首先通过 Filter 方法剔除无关特征或噪声特征，以期有效缩减后续特征搜索规模；然后将选取的特征通过 Wrapper 方法继续进行优化选取。首先采用奇异值分解(Singular Value Decomposition，SVD)方法对数据降维去噪，然后将随机森林分类器的分类准确率作为特征可分性判据，基于随机森林算法本身的变量重要性度量进行特征重要性排序，利用可回溯的贪婪搜索扩张及最好优先原则选取特征子集。

除 50 个博文主题特征外，本文提取了 28 个基础特征。特征全集为：是否使用昵称，用户关注他人数量，用户性别，拥有粉丝的数量，教育经历，工作经历，用户博文平均被点赞数量，用户博文平均被转发数量，用户博文平均被评论数量，用户个性标签数量，日均发博数，博文发布月份，发布星期，发布小时，是否分享图片，是否为分享，是否为收藏，是否为转发，是否包含关键字“红包”、“抽奖”、“转发”、“教程”，包含话题的数量，“@”他人的数量，包含外链的数量，博文长度，博文影响力，用户活跃度，博文主题特征。

经过等频离散化后的博文影响力数据如图 4 所示，从该结果可以看出基于相对熵计算的博文影响力特征有较高的预测能力。最终，除 LDA 主题以外，通过上述方法筛选出的特征组如图 5 所示。选定特征组的相对熵如图 6 所示，该结果可以对特征重要性提供辅助参考，同时从图 6 中可以看出用户活跃度与博文影响力对后续预测有积极作用。现定义上文筛选出的图 5 所示的特征组及主题特征为子集 1；特征全集除去博文影响力、用户活跃度、博文主题后的特征组为子集 2；随机选择特征为子集 3－子集 6。使用这 6 个特征集和特征全集，以随机森林作基分类器进行对比实验。如图 7 所示，选择不同特征集会显著影响分类的准确率，在基础特征中加入影响力、活跃度、主题特征后预测准率得到提高，说明本文提取的特征是有效的；而使用筛选后的特征子集 1 的预测准确率高于使用特征全集和其他随机特征子集的预测准确率，说明特征筛选方法是有效的。后续实验所采用的筛选后的特

征即为此处的特征子集 1。

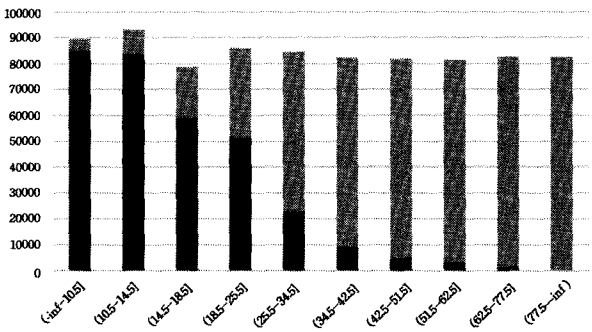

图 4　博文影响力等频离散化后的直方图

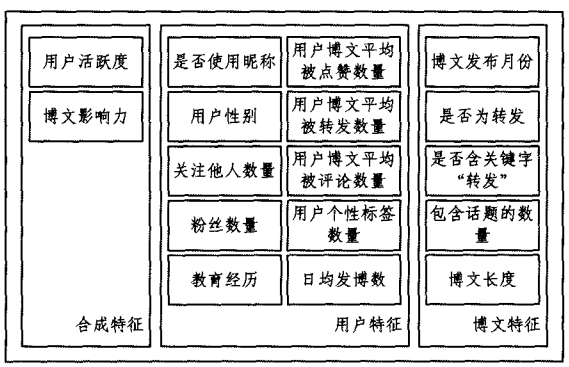

图 5　选定特征组

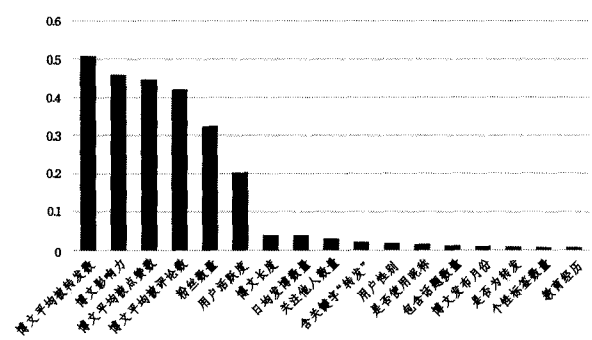

图 6　选定特征组的相对熵

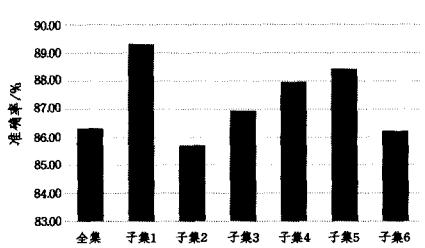

图 7　不同特征集对预测准确率的影响

4.2　异常检测

异常检测是一个非监督学习算法。本文通过高斯分布异常检测来提升随机森林算法在微博转发预测方面的性能。多元高斯分布虽然能够自动捕捉特征间的相关性，但是其计算代价高且要求协方差必须可逆，因此最终选择原高斯分布模型。原高斯分布模型虽然不能捕捉特征间的相关性，但可以通过组合特征线性来解决此问题，并且其计算代价低，适应大规模特征，训练集较小时也适用。

选择的特征需要近似服从高斯分布，如果明显不服从高斯分布，可以做适当的转换，例如 $\log(x)$，$\log(x+c)$，$\sqrt{x}$，$x^{1/3}$ 等。将随机森林不能很好预测转发行为的部分数据作为异常对待，通过异常检测的方式来提高该部分的转发预测准确率，过程如算法 2 所示，使用 $\log(x+1)$来处理不服从高斯分布的数据。通过实验令阈值 $\varepsilon=0.03$，图 8 所示的随机森林预测转发概率与错误数示意图反映出概率低于 0.6 的预测易出错，因此将随机森林算法预测转发或不转发概率小于 0.6 的数据取出作为待检测数据集。异常数据极易被随机森林算法错误预测，将其预测结果反转，以期修正。

算法 2　高斯分布异常检测提升预测结果性能的算法

Step 1　Choose features x_i that might be indicative of anomalous examples. And converting data to fit Gauss distribution by

$$X=\log(x+1)$$

Step 2　Training set of random forest which will be right predicted is used as the training set to fit parameters $\mu_1,\cdots,\mu_n,\sigma_1^2,\cdots,\sigma_n^2$ by

$$\mu_j=\frac{1}{m}\sum_{i=1}^{m}\chi_j^{(i)}$$

$$\sigma_j^2=\frac{1}{m}\sum_{i=1}^{m}(\chi_j^{(i)}-\mu_j)^2$$

And then construct the $p(\chi)$ function.

Step 3　Extract data which Random forest algorithm predicted the retweet probability below 0.6, they can be used as the data set to be detected.

Step 4　Data obtained from step 3 detected by Gauss anomaly detection. Compute

$$p(\chi)=\prod_{j=1}^{n}p(\chi_j;\mu_j;\sigma_j^2)=\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi\sigma_j}}\exp\left(-\frac{(\chi_j-\mu_j)^2}{2\sigma_j^2}\right)$$

Anomaly if $p(\chi)<\varepsilon$

Step 5　If the data is determined to be anomaly, the prediction results of the random forest need to be reversed.

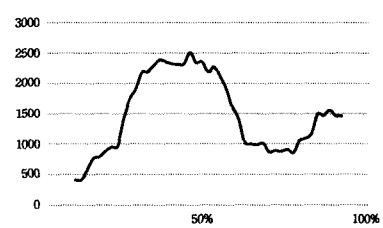

图 8　随机森林算法预测转发概率与错误数示意

4.3　转发预测实验结果与分析

目前随机森林算法通常利用袋外(Out of Bag，OOB)误差估计确定随机森林中的决策树和特征数。通过分别固定特征数量、决策树数量对 OOB 误差估计进行观察，最终选择 6 个特征和 45 个决策树。通过 Python 编写爬虫程序，获取了新浪微博的 841443 条有效数据，使用上文提取的特征，利用著名数据挖掘平台 Weka 进行实验，采用十折交叉验证方式验证算法性能。整个实验在 CPU 为 i5-4200U 1.6GHz，内存为 8GB，运行 Windows10 的 64 位系统电脑上进行。

4.3.1　预测准确率分析

使用不同算法的对比实验结果如表 1 所列，其中 Basic 特征是指特征全集；Filtered 特征是特征全集经筛选后的特征组，即上文中的特征子集 1。

表 1 微博转发行为预测结果

Algorithm	Feature	Precision /%	Recall /%	F-Measure /%	ROC Area/%
Naive Bayes	Basic	80.0	80.2	80.1	86.7
Logistic	Basic	84.2	84.2	84.2	91.7
J48	Basic	87.3	87.3	87.3	93.0
Random Forest	Basic	86.3	86.4	86.2	93.6
Naive Bayes	Filtered	87.2	86.7	86.8	93.8
Logistic	Filtered	88.4	88.4	88.4	94.8
J48	Filtered	89.1	89.1	89.1	94.1
Random Forest	Filtered	89.4	89.3	89.3	95.9
ADRF	Filtered	90.5	90.5	90.5	96.8

通过计算分类算法常用的度量标准 Precision、Recall、F-Measure和 ROC Area 来对预测结果进行度量。部分算法在不同交叉验证折数下的准确率比较结果如图 9 所示。

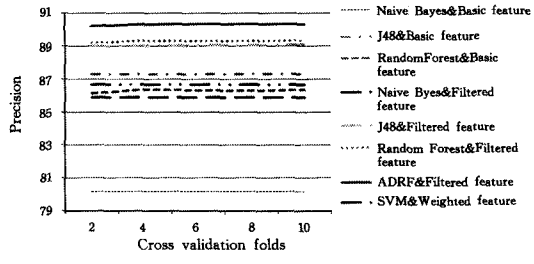

图 9 预测准确率对比图

由表 1 及图 9 得出，随机森林算法优于其他基本算法，说明选取的基准预测算法是有效的；同时，筛选后多特征的使用效果也好于使用未筛选特征的效果；在不同的交叉验证折数下，算法的运行效果也比较稳定。融合异常检测与随机森林的预测方法(Anomaly Detection and Random Forest，ADRF)优于在微博转发行为预测中常用的逻辑回归、决策树、朴素贝叶斯、随机森林等算法，并且预测准确率优于文献[3]的 46.6%和文献[7]基于支持向量机特征加权模型的 85.9%，也不会出现如文献[9]导致准确率反而下降的情况。特征的计算也没有文献[8]那样复杂，加入用户活跃度、博文影响力和 LDA 主题特征并进行特征筛选后，即便使用微博转发行为预测常用的基本算法也能取得较好的预测效果，说明特征不应随意选择，对特征进行筛选是非常必要的。

4.3.2 预测耗时分析

使用不同预测算法及不同预测数据量的耗时对比如表 2 所列，其中 Basic 特征指特征全集；Filtered 特征是特征全集经筛选后的特征组，即上文中的特征子集 1。表中 1/3，2/3，3/3 表示本次预测使用的数据量占总数据的比例，时间单位为 s。

表 2 微博转发行为预测耗时对比

Algorithm	Feature	1/3	2/3	3/3
Naive Bayes	Basic	13	23	35
Logistic	Basic	5809	7615	9828
J48	Basic	329	487	755
Random Forest	Basic	504	798	1272
Naive Bayes	Filtered	5	11	17
Logistic	Filtered	2731	4289	6289
J48	Filtered	52	90	138
Random Forest	Filtered	264	426	708
ADRF	Filtered	317	508	813

将表 2 绘制成曲线，以便观察，如图 10 所示。

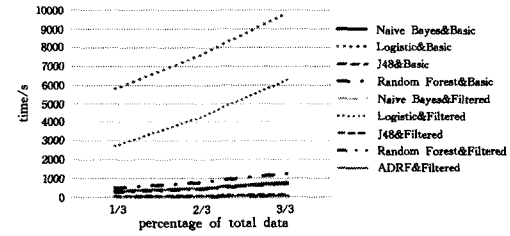

图 10 预测耗时对比图

由表 2 及图 10 可以看出，逻辑回归预测方法非常耗时，而朴素贝叶斯预测方法非常快速，但通过上节可知两者的预测精度相较于其他方法偏低。ADRF 方法的耗时略高于随机森林方法，但其准确率、ROC 曲线等各项指标均好于随机森林方法，因此 ADRF 方法的耗时略升高是可接受的。而特征的筛选与否对各预测方法的耗时均有明显的影响，说明特征筛选是有必要的。

结束语 在微博网络中，信息的传播主要是通过微博的转发实现的。为了解决微博转发行为预测准确率不高及特征选择任意性的问题，本文首先对各基本特征、博文主题特征、用户活跃度、博文影响力进行了研究；其次，通过融合两种特征选择方法，对特征进行筛选；最后，基于筛选出的最佳特征组，融合异常检测与随机森林算法对微博转发预测效果进行提升。在采集的新浪微博数据上的对比实验表明，提出的预测方法与特征选择方法效果显著，能够对微博转发行为进行有效的预测。后续将在本研究基础上对微博转发规模的预测做相关研究。

参 考 文 献

[1] CAO J X, WU J L, SHI W, et al. Sina microblog information diffusion analysis and prediction[J]. Chinese Journal of Computers, 2014, 37(4): 779-790. (in Chinese)
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.

[2] 中国互联网络信息中心. 第 37 次中国互联网络发展状况统计报告[R]. 北京：中国互联网络信息中心，2016.

[3] PETROVIC S, OSBORNE M, LAVRENKO V. RT to win! Predicting message propagation in Twitter[C]// Proceedings of the Fifth International Conference on Weblogs and Social Media. Barcelonia, Spain, 2011.

[4] MORCHID M, DUFOUR R, BOUSQUET P M, et al. Feature selection using Principal Component Analysis for massive retweet detection[J]. Pattern Recognition Letters, 2014, 49: 33-39.

[5] YANG Z, GUO J, CAI K, et al. Understanding retweeting behaviors in social networks[C]// Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010). Toronto, Ontario, Canada, 2010: 1633-1636.

[6] ROMERO D M, MEEDER B, KLEINBERG J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter[C]// Proceedings of the 20th International Conference on World Wide Web (WWW 2011). Hyderabad, India, 2011: 695-704.

(下转第 220 页)

Aspects of Contemporary Intelligent Computing Techniques. 2007:115-124.

[7] DONG W, LEI M, YU R. A New Evolutionary Algorithms for Global Numerical Optimization Based on Ito Process[M]// Computational Intelligence and Intelligent Systems:5th International Symposium(ISICA 2010). Wuhan, China, Springer, Berlin, Heidelberg, 2010.

[8] DONG W, YU R, LEI M. Merging the Ranking and Selection into ITO Algorithm for Simulation Optimization[C]// 5th International Symposium Computational Intelligence and Intelligent Systems(ISICA 2010). Wuhan, China, Springer, 2010.

[9] DONG W Y, ZHANG W S, YU R G. Convergence and Runtime Analysis of ITO Algorithm for One Class of Combinatorial Optimization[J]. Chinese Journal of Computers, 2011, 34(4): 636-646. (in Chinese)
董文永,张文生,于瑞国. 求解组合优化问题伊藤算法的收敛性和期望收敛速度分析[J]. 计算机学报,2011,34(4):636-646.

[10] YI Y F, CAI Y L, DONG W Y, et al. Improved ITO Algorithm for Multiobjective Real-time Vehicle Routing Problem with Customers' Satisfaction[J]. Acta Electronica Sinica, 2015, 43(10): 2053-2061. (in Chinese)
易云飞,蔡永乐,董文永,等. 求解带用户满意度的多目标实时车辆路径问题的改进伊藤算法[J]. 电子学报,2015,43(10):2053-2061.

[11] YI Y F, DONG W Y, LIN X D, et al. The Improved ITO Algorithm to Solve the Vehicle Routing Problem with Soft Time Windows and Its Convergence Analysis[J]. Acta Electronica Sinica, 2015, 43(4): 658-664. (in Chinese)
易云飞,董文永,林晓东,等. 求解带软时间窗车辆路径问题的改进伊藤算法及其收敛性分析[J]. 电子学报,2015,43(4):658-664.

[12] WANG H G, YU S M. Improved ITO Algorithm for Solving VRP[J]. Computer Science, 2015, 42(9): 253-256. (in Chinese)
王浩光,余世明. 求解车辆路径问题的改进伊藤算法[J]. 计算机科学,2015,42(9):253-256

[13] ZHAO Z Y, LI Y X, YU F. Artificial Bee Colony Algorithm Based on ITO Algorithm[J]. Computer Science, 2014, 41(6A): 29-32. (in Chinese)
赵志勇,李元香,喻飞. 基于伊藤算法的改进人工蜂群算法[J]. 计算机科学,2014,41(6A):29-32.

[14] YI Y F, CAI Y L, DONG W Y, et al. Improved ITO Algorithm for Solving the CVRP[J]. Computer Science, 2013, 40(5): 213-216. (in Chinese)
易云飞,蔡永乐,董文永,等. 求解带容量约束的车辆路径问题的改进伊藤算法[J]. 计算机科学,2013,40(5):213-216.

[15] MUNEMOTO M, TAKAI Y, SATO Y. A Migration Scheme for the Genetic Adaptive Routing Algorithm[C]// IEEE International Conference on Systems, Man, and Cybernetics. 1998: 2774-2779.

[16] INAGAKI J, HASEYAMA M, KITAJIMA H. A Genetic Algorithm for Determining Multiple Routes and Its Applications[C]// Proceedings of IEEE International Symposium on Circuits and Systems. 1999: 137-140.

[17] AHN C W, RAMAKRISHNA R S. A Genetic Algorithm for Shortest Path Routing Problem and the Sizing of Populations [J]. IEEE Transactions on Evolutionary Computation, 2002, 6 (6): 566-579.

[18] SUN B L, LI L Y, CHEN H. Shortest Path Routing Optimization Algorithms Based on Genetic Algorithms[J]. Computer Engineering, 2005, 31(6): 142-144. (in Chinese)
孙宝林,李腊元,陈华. 基于遗传算法的最短路径路由优化算法[J]. 计算机工程,2005,31(6):142-144.

(上接第 196 页)

[7] ZHANG Y, LU R, YANG Q. Predicting retweeting in microblogs[J]. Journal of Chinese Information Processing, 2012, 26 (4): 109-114. (in Chinese)
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.

[8] LI Y L, YU H T, LIU L X. Predict algorithm of micro-blog retweet scale based on SVM[J]. Application Research of Computers, 2013, 30(9): 2594-2597. (in Chinese)
李英乐,于洪涛,刘力雄. 基于 SVM 的微博转发规模预测方法[J]. 计算机应用研究,2013,30(9):2594-2597.

[9] ZHAO Y, SHAO B L, BIAN G Q, et al. Prediction of retweeting behavior for imbalanced dataset in microblogs[J]. Journal of Computer Applications, 2015, 35(7): 1959-1964. (in Chinese)
赵煜,邵必林,边根庆,等. 面向不平衡微博数据集的转发行为预测方法[J]. 计算机应用,2015,35(7):1959-1964.

[10] SUH B, HONG L, PIROLLI P, et al. Want to be Retweeted? Large scale analytics on factors impacting retweet in Twitter network[C]// 2010 IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust. IEEE, 2010: 177-184.

[11] HU W. Real-time Twitter sentiment toward midterm exams[J]. Sociology Mind, 2012, 2(2): 177-184.

[12] WU J H, ZUO K Z, JIE B, et al. New discriminative feature selection method[J]. Journal of Computer Applications, 2015, 35 (10): 2752-2756. (in Chinese)
吴锦华,左开中,接标,等. 新颖的判别性特征选择方法[J]. 计算机应用,2015,35(10):2752-2756.

[13] KAR M, NUNES S, RIBEIRO C. Summarization of changes in dynamic text collections using Latent Dirichlet Allocation model [J]. Information Processing & Management, 2015, 51(6): 809-833.

[14] LIU S P, YIN J, OUYANG J, et al. Topic mining from microblogs based on MB-HDP model[J]. Chinese Journal of Computers, 2015(7): 1408-1419. (in Chinese)
刘少鹏,印鉴,欧阳佳,等. 基于 MB-HDP 模型的微博主题挖掘[J]. 计算机学报,2015(7):1408-1419.