一个基于随机森林的微博转发预测算法

罗知林[1] 陈 挺[2] 蔡皖东[1]

(西北工业大学计算机学院 西安 710129)[1] (北京邮电大学计算机学院 北京 100876)[2]

摘 要 转发(Retweet)是微博中一个重要的信息传播机制,用户可以将其关注者(Follower)的有趣微博转发到自身平台,分享给他的粉丝(Fan),快速地实现微博信息的传播。主要对微博转发预测进行了研究,首先提取了重要特征,比如用户间的微网络结构、权重比率、用户个人信息等,以研究用户微博转发行为,然后基于以上特征提出了一个随机森林微博转发预测算法(RFMR)。实验结果表明,RFMR算法优于其他分类算法,可以有效地用来预测微博转发。

关键词 社会网络,机器学习,随机森林,微博,转发

中图法分类号 TP391.4 **文献标识码** A

Microblogging Retweet Prediction Algorithm Based on Random Forest

LUO Zhi-lin[1] CHEN Ting[2] CAI Wan-dong[1]

(School of Computer, Northwestern Polytechnic University, Xi'an 710129, China)[1]

(School of Computer, Beijing University of Posts and Telecommunications, Beijing 100084, China)[2]

Abstract Retweet is an important information diffusion mechanism in Microblogging, in which users can forward their followers' blogging and share it to their fans, and it can quickly diffuse. This paper studied Microblogging retweet prediction. First we analyzed and extracted the important features, such as micro network structure, weight ratio, user profiles, and then proposed a new prediction algorithm based on random forest (RFMR). The experiment shows that, compared to other classifications, RFMR has better performance. It can effectively predict user retweet behavior.

Keywords Social network, Machine learning, Random forest, Microblogging, Retweet

1 引言

随着 Web2.0 的发展,社交网络已经成为互联网流行的信息共享和分发平台。社交网站非常流行,比如国外的 Facebook、Twitter 以及 Flicker 等,国内的 QQ、新浪微博、人人网等。其中,微博作为一个新兴的社会媒体,近来发展迅速,据最新新闻报道,新浪微博用户总数已达 2.498 亿,是国内最大社交平台。

微博(Microblogging)是一个基于用户关系的分享、传播以及获取信息的平台,用户可以通过 Web、WAP 等各种客户端组建个人社区,以 140 字左右的文字更新信息,并实现即时分享。

转发(Retweet)是微博的一个重要的机制,用户可以将其关注者(Follower)的有趣微博转发到自身平台,分享给他的粉丝(Fan)。

转发是微博网络中信息传播最重要的功能,对研究信息在微博中传播比如微博用户行为和兴趣、突发事件预测、网络舆情监控以及用户推荐等方面具有重要意义,因而越来越受到研究者的重视。

本文主要研究了如何预测用户转发行为,首先分析了用户的微型网络结构、权重以及其它特征,然后基于这些特征提出了一种新的基于随机森林的预测算法。实验结果表明该算法优于其他算法。

2 相关研究

近些年,微博中的信息传播得到了学者的广泛关注。H. Kwak 等[2]分析了 Twitter 网络,研究结果表明 Twitter 网络是一个社会媒体与社交网络的混合体,但更倾向于社会媒体网络而不是社交网络。Romero 等[3]分析了 Twitter 不同类型的主题微博传播机制,发现不同主题微博的曝光次数将不一样,含有政治内容的微博曝光次数越多越容易传播,而含有一些新兴的词汇的微博曝光次数增多会导致传播下降。Z. Luo 等[4]研究了新浪微博转发行为,研究表明微博更加容易在朋友、亲戚等强连接以及权威人士连接关系中转发,另外研究表明女性用户更容易转发来自于朋友等强连接的用户,而男性用户则相反。Z. Yang 等[5]基于 Twitter 微博转发的特征,提出了一个因子图预测模型。

本文对微博转发行为的预测进行研究,首先提取一些相关重要特征,比如用户间的微网络结构、用户权重以及其它特征,然后基于这些特征提出一个新的基于随机森林的微博转发预测算法。

3 问题描述

定义 1 已知微博网络 $G=(V,E)$,其中 V 为用户的集

到稿日期:2013-05-18 返修日期:2013-07-08 本文受国家 863 计划项目(2009AA01Z424)资助。

罗知林(1984—),博士生,主要研究方向为社会网络、数据挖掘,E-mail:lzluo007@gmail.com;陈 挺(1989—),主要研究方向为社会网络、数据挖掘;蔡皖东(1956—),博士生导师,主要研究方向为网络安全、社会网络。

合，E 为有向关注边的集合，如果用户 $u_1 \in V, u_2 \in V$，且 u_1 关注 u_2，则关注边定义为：

$$u_1 \rightarrow u_2 \tag{1}$$

本文主要研究的问题是，对于任何关注边 $u_1 \rightarrow u_2$，如何准确地预测用户 u_1 是否转发用户 u_2 的微博。

在网络中关注边 $u_1 \rightarrow u_2$ 实际上定义用户 u_1 与其关注者用户 u_2 的关系，而用户 u_1 的被关注关系并未给出，但是通过转化可以生成关注边，比如用户 u_1 被用户 u_3 关注，可以等价地转化为用户 u_3 关注用户 u_1，即 $u_3 \rightarrow u_1$，因而，微博中的所有关系都可以视为关注边。为了简化问题，本文主要研究关注边的微博转发的预测。

图 1 给出微博用户关注网络图。对于 Bob 而言，Bob 关注了 Greg、Harry、Fred 以及 Carol 等，同时被 Dave、Alice、Greg、以及 Harry 等所关注。注意存在一部分用户子集与 Bob 相互关注，比如 Harry 与 Fred。对于 Bob 而言，本文将研究关注边上的用户的微博，即 Greg、Harry、Fred 以及 Carol 等是否会被 Bob 转发。

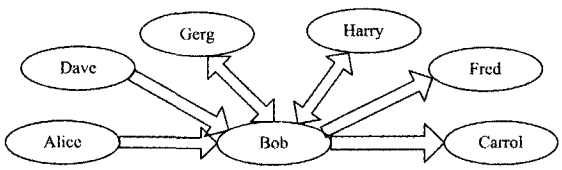

图 1　微博用户关注实例图

4　RFMR 算法

4.1　随机森林

随机森林是由 Leo Breiman 和 Adele Cutler 提出的分类预测算法，最早可追溯到贝尔实验室 Tin Kam Ho 提出的随机决策森林（random decision forests）算法，结合了 Breiman 的“Bootstrap aggregating”想法以及 Ho 的“random subspace method”方法。

定义 2　随机森林是一个由一组决策树分类器（$h(X, \theta_K), k=1 \cdots K$）组成的集成分类器，其中 $\{\theta_k\}$ 是服从独立同分布的随机变量，K 表示随机森林中决策树的个数，在给定自变量 X 的情况下，每个决策树分类器通过投票来决定最优的分类结果。

随机森林算法一般构造过程如下：

1. 对于给定训练样本，随机可重复取样，形成新的子样本数据。

2. 对新的子样本数据中 M 个特征变量，随机方法抽取 m（$m<M$）个特征，构造完整的决策树。

3. 重复步骤 1、2，得到 K 个决策树，形成随机森林。

4. 每个决策树投票，选出最优的分类。

图 2 给出随机森林的构建过程。

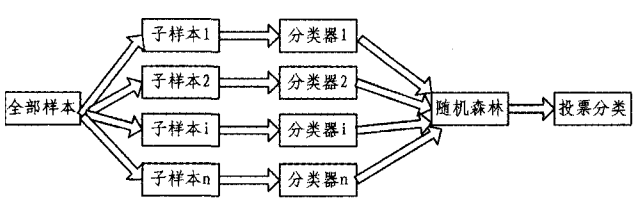

图 2　随机森林构建过程图

在随机森林中，决策树的个数 K 以及特征个数 m 影响着分类器的性能，以下研究将如何选取优参数 K 和 m。

定义 3　给定一组分类器 $h_1(X), h_2(X), \cdots, h_K(X)$，每个分类器的训练集都是从原始的服从随机分布的数据集 (Y, X) 中随机抽样所得，边际函数（margin function）为：

$$mg(X,Y)=av_K I(h(X)=Y)-\max_{j \neq Y} av_k I(h(X)=j) \tag{2}$$

其中，$I(\cdot)$ 是示性函数，$av_k(\cdot)$ 表示取平均值。

边际函数度量了正确分类 Y 中 X 的平均得票数超过其他类的平均选票数量的程度。该值越大，表明分类器的置信度就越高。

定义 4　分类器的泛化误差定义为：

$$PE=P_{X,Y}(mg(X,Y)<0) \tag{3}$$

其中，下标 X, Y 表明了概率覆盖的定义空间。在随机森林 $h_k(X)=h(X,\theta_k)$ 中，当决策树分类器足够多时，它服从强大数定律。

定理 1　随着决策树数目增加，对于所有序列 $\theta_1, \theta_2, \cdots, \theta_n$，PE 几乎处处收敛于：

$$P_{X,Y}(P_\theta(h(X,\theta)=y)-\max_{j \neq Y} P_\theta(h(X,\theta)=j)<0) \tag{4}$$

其中，θ 是对应单个决策树的随机向量，$h(X,\theta_k)$ 为基于 X 和 θ 的输出。

定理 1 表明随着决策树分类器个数 k 的增加，随机森林不会出现过分拟合（overfit）问题，但是会产生有限的泛化误差。

定义 5　随机森林的边际函数：

$m_r(x,y)=P_\theta(h(x,\theta)=y)-\max_{j=y} p_\theta(h(x,\theta)=j)$，则组分类器的分类效能定义为：

$$s=E_{x,y} m(x,y) \tag{5}$$

假设 $s \geqslant 0$，根据切比雪夫不等式可得到

$$PE \leqslant \mathrm{var}(m_r)/s^2 \tag{6}$$

m_r 的方差还有另外一个更明显的表达式，可以描述如下：

$$\hat{y}(x,y)=\operatorname{argmax}_{j \neq y} p_\theta(h(x,\theta)=j) \tag{7}$$

则有

$$m_r(x,y)=E_\theta[I(h(x,\theta)=y)-I(h(x,\theta)=\hat{y}(x,y))] \tag{8}$$

由于

$$\mathrm{var}(m_r)=\bar{\rho}(E_\theta sd(\theta))^2 \leqslant \bar{\rho}(E_\theta \mathrm{var}(\theta)) \tag{9}$$

其中，$\bar{\rho}$ 是相关系数的均值。

因此

$$E_\theta \mathrm{var}(\theta) \leqslant E_\theta(E_{X,Y} rmg(\theta,X,Y))^2 - s^2 \leqslant 1-s^2 \tag{10}$$

由式(8)－式(10)可推导如下定理：

定理 2　泛化误差的上界为：

$$PE \leqslant \bar{\rho}(1-s^2)/s^2 \tag{11}$$

其中，$\bar{\rho}$ 是分类器相关性均值，s 是分类器效能强度。

由定理 2 可知，随机森林的泛化误差上界可以由 $\bar{\rho}$ 和 s 两个参数推导出来，其中与 $\bar{\rho}$ 成正比，与 s^2 成反比，因此，比值 $\bar{\rho}/s^2$ 越小越好。

对于特征值选取的个数 m，Breiman 等研究表明，随着 m 值增加，分类器的相关性 $\bar{\rho}$ 和效能强度 s^2 也相应地增加，反之亦然。因此，m 值会产生一定的泛化误差，当 m 值在某一区间时，泛化误差上界将处于最低。

4.2　微博特征提取

在随机森林分类预测中，另外一个重要的任务是寻找相

关的重要特征。一方面，一个原因是数据集中许多特征本身对分类预测无关，另一个原因是有些特征可能是冗余的。如果选择的特征几乎不具有辨别能力，那么接下来设计的分类器性能也将是很差的。另一方面，如果选择了具有充分辨别能力的特征，那么极大地提高了分类器的预测精度，因此特征选取这个过程至关重要。本文基于网络结构、权重比率等不同方面提取 28 个特征值。

4.2.1 用户间微网络结构

对于关注边 $A\rightarrow B$，定义 3 类微网络结构。如图 3 所示，在 Pattern I 中，用户 A 和 B 相互关注，同时至少存在第 3 个用户与用户 A、B 都相互关注，即强连通的三角关系；在 Pattern II 中，用户 A 和 B 相互关注，但是不存在强连通的三角关系；在 Pattern III 中，用户 A 与 B 只是单向关注关系。

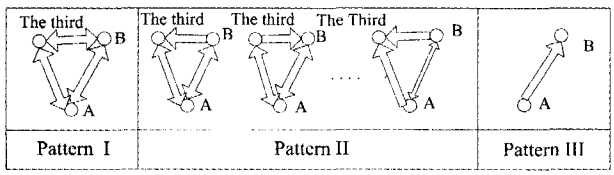

图 3 3 类不同微结构 Pattern

4.2.2 权重比率

对于关注边 $A\rightarrow B$，用户 A、B 有不同的社会权重。用户社会权重用其粉丝数目来标记，则关注边 $A\rightarrow B$ 的权重比率定义为：

$$p_{A\rightarrow B}=\frac{\#Followers(B)}{\#Followers(A)} \tag{12}$$

它反映了用户 A 与 B 之间的社会权重差值。比率越大表明用户 B 的社会地位越高于用户 A，反之亦然。

4.2.3 用户间距离关系

距离关系是指用户间的地理位置关系。针对关注边 $A\rightarrow B$，定义 3 类关系：Near、Medium 和 Far，分别表示两用户在同一城市、两用户在不同城市但在同一省份以及两用户在不同省份。

4.2.4 用户间性别关系

对于关注边 $A\rightarrow B$，定义 4 类关系，分别为 MM、MF、FM、FF，其中 MM 表示男性用户 A 关注男性用户 B，其他类似。

4.2.5 用户自身特征

对于关注边 $A\rightarrow B$，用户自身特征主要包括省份、城市、性别、微博数、粉丝数、关注数等 12 个特征。总共得到两个用户 24 个特征值。

4.3 RFMR 算法实现

由于 M 值较少，本文采取完全随机方法选取 k 个特征，而对于结果分类预测，则采用简单多数投票法。最终的分类决策：

$$H(x)=\arg\max_{Y}\sum_{i}I(h_i(x)=Y) \tag{13}$$

其中，$H(x)$ 表示组合分类模型，$h_i(x)$ 表示单个决策树分类模型。则 RFMR 算法如下。

算法 1 随机森林微博转发预测算法(RFMR)

输入：微博训练数据集 S，预测数据集 T

随机森林决策树个数 k 以及特征个数 m

模型训练：

1. 对微博数据集 S 采用 bootstrap 方法采样，形成新的训练集 S_i。
2. 训练集 S_i 含有 M 个特征的，完全随机选取 m(m<M)个特征，形成新的训练集 S_n。
3. 对数据集 S_n 利用 CRAT 算法构造完整的决策树，不进行剪枝。
4. 循环步骤 1、2、3，直到建立 K 个决策树，随机森林构造完成。

预测：

5. 对数据集 T 的每个变量 x 的类型进行预测，k 个决策树分别对变量 x 的类型进行投票。
6. 计算所有投票数 H(x)，找出票数最高的分类就是变量 x 的分类标签。

输出：数据集 T 中变量 x 的类型

5 实验结果分析

本实验数据来自于新浪微博，自 2011 年 5 月到 2011 年 7 月，随机地采集了 171169 个用户和 702789 条关注边，其中用户包括标签、用户基本资料以及微博信息等，关注边包括了关注关系和是否转发。如果关注边存在转发，则视为正例，否则为负例，总共得到 185237 个正例和 517552 个负例。

Weka 作为一个公共免费的数据挖掘与分析平台，集合了大量数据挖掘算法，包括数据预处理、分类、回归、聚类、关联规则等。本文所有实验都将在 Weka 上运行。

5.1 参数优化

在随机森林的 bootstrap 采样过程中，约有 37%数据不在样本中，这些数据称为袋外数据(Out of Bag，OOB)，使用袋外数据估计模型性能称为 OOB 估计。对于每一棵决策树，我们都可以得到一个 OOB 误差估计，将所有决策树的 OOB 误差估计取平均，即可得到随机森林的泛化误差估计。Breiman 通过实验证明，OOB 误差是无偏估计。

本文将以 OOB 误差估计为指标来选取决策树的个数 k 以及特征选取的个数 m，当 OOB 估计最少时，k 和 m 参数最优。由于 k 和 m 两参数都对 OOB 估计产生影响，参数组合将需要试验 $k*m$ 次。为了简化试验过程，本实验在评估一个参数之前固定另一个参数，这样只需 $k+m$ 次试验。

图 4 给出了当 $m=2$ 时，决策树个数 k 与 OOB 的曲线变化。由图可见，随着 k 值增加，OOB 误差在减少，但是下降趋势不相同。当 k 处于 5－17 区间时，在开始阶段下降趋势较快，然后逐步减弱，当 k 处于 17－20 区间时，下降趋势再次减弱，逐步趋于平稳，接近收敛。而随着 k 值的增加，算法的运行时间也在逐步增加，由图 5 可见，随着 k 值增大，运行时间近乎直线上升。综合考虑 OOB 误差和时间因素，$k=18$ 为最优。

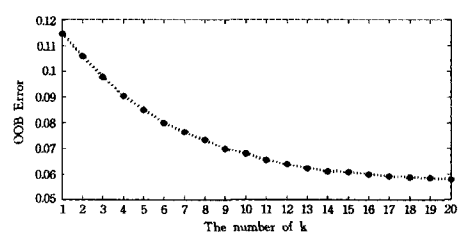

图 4 随机森林决策树个数 k 与 OOB 误差

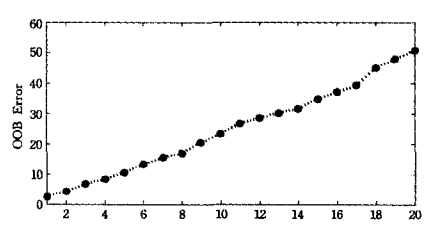

图 5 随机森林决策树个数 k 与运行时间

(下转第 74 页)

[2] Jai M, Sharma P, Banerjee S. QoS-Guaranteed Path Selection Algorithm for Service Composition [C] // IEEE IWQoS, 2006. 2006: 288-289

[3] Tang C, McKinley P K. On the cost-quality tradeoff in topology-aware overlay path probing[C] // Proceedings of the 11th IEEE International Conference on Network Protocols (ICNP), (Atlanta, Georgia). November 2003: 268-279

[4] Samimi F A, McKinley P K. Dynamis: Dynamic Overlay Service Composition for Distributed Stream Processing [C] // SEKE 2008. 2008: 881-886

[5] Gu Xiao-hui. Spidernet: a Quality-Aware Service Composition Middleware [D]. University of Illinois at Urbana-Champaign, 2004

[6] Deb K. Multi-objective evolutionary algorithms: Introducingbias among Pareto-optimal solutions[C] // Ghosh A, Tsutsui S, eds. Advances in Evolutionary Computing: Theory and Applications. London: Springer-Verlag, 2003: 263-292

[7] 夏亚梅，孟祥武，陈俊亮，等. 基于改进蚁群算法的服务组合优化[J]. 计算机学报，2012，35(2)：207-281

[8] 赵欣，沈立炜，彭鑫，等. P MOEA：一种多目标决策辅助遗传算法用于服务组合 QoS 优化[J]. 中国科学：信息科学，2013，43(1)：73-89

[9] 李俊，郑小林，陈松涛，等. 一种高效的服务组合优化算法[J]. 中国科学：信息科学，2012，42：280-289

(上接第 64 页)

图 6 给出了 $k=15$ 时特征个数 m 与 OOB 的曲线变化。由图可见，随着 m 值增加，OOB 误差值先迅速地下降，当 $m=3$ 时，处于最低值，然后逐步回升。因而，当 $m=3$ 时，OOB 误差最小，故 $m=3$ 为最优。最终我们选定随机森林 $k=18$，$m=3$。

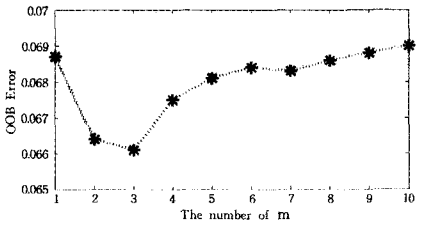

图 6 随机森林特征选择 m 与 OOB 误差

5.2 算法比较

针对 RFMR 算法，本实验将与逻辑回归(LR)、决策树(DT)、Adaboost(Ada)、朴素贝叶斯(NB)和多层感知器(MP)等经典分类算法进行比较。

由表 1 可见，在正例预测中，不同算法效果相差明显。在 Precision 中，RFMR 算法达到了 92.9%，相比于 DT 的 71.4%，提高了 21.5%，其他算法表现更加一般，NB 算法效果很差，不到 36%。总体来说，各个算法的 Recall 值较低，其原因是正负例的比例不均衡，大量被错误分类的负例降低了 Recall 值。但是 RFMR 算法的 Recall 值最大，达到了 68%，依然比 DT 算法高 4.1%，而其他算法表现很差，不到 30%。另外从 F-Measure、ROC 等指标来看，在正例预测中，RFMR 算法明显优于其他算法。

表 1 微博正例算法性能比较

	Precision	Recall	F-Measure	ROC
LR	0.645	0.133	0.221	0.79
NB	0.359	0.282	0.316	0.74
DT	0.714	0.629	0.669	0.852
MP	0.709	0.254	0.374	0.816
Ada	0.615	0.142	0.231	0.794
RFMR	0.929	0.68	0.785	0.917

表 2 给出了负例预测中各个算法的结果。相比于正例，各个算法的性能差距较小，同时各个算法的 Precision 和 Recall 值都比较高，说明这些算法预测负例的效果都比较好。但是 RFMR 算法在各项指标中最高，这说明在负例预测中它优于其他算法。

表 2 微博负例算法性能比较

	Precision	Recall	F-Measure	ROC
LR	0.883	0.989	0.933	0.79
NB	0.895	0.924	0.909	0.74
DT	0.945	0.962	0.953	0.852
MP	0.897	0.984	0.939	0.816
Ada	0.883	0.986	0.932	0.794
RFMR	0.953	0.992	0.977	0.917

综合比较正负例中各个算法的性能，可以看出 FMR 算法是最优的。

结束语 在微博网络中，用户间的微网络结构、权重比率等特征，对微博的转发行为有着显著的作用。为了研究用户的微博转发行为，本文首先分析提取相关特征，然后基于这些特征，提出了一个 RFMR 算法。实验结果表明，相比于其他分类方法，RFMR 算法性能最优。

参 考 文 献

[1] Granovetter M. The strength of weak ties[J]. The American Journal of Sociology, 1973, 78(6): 1360-1380

[2] Kam H T. Random Decision Forest[C] // Proceedings of the 3rd International Conference on Document Analysis and Recognition. 1995: 278-28

[3] Romero D M, Meeder B, Kleinberg J. Differencesin 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. 2011: 695-704

[4] Luo Z, Wu X, Cai W, et al. Examining Multi-factor Interactions in Microblogging based on Log-linear Modeling [C] // Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Istanbul, Turkey, August 2012: 6

[5] Yang Z, Guo J, Cai K, et al. Understanding Retweeting Behaviors in Social Networks [C] // Proceedings of the Nineteenth Conference on Information and Knowledge Management. 2010: 1633-1636

[6] Kwak H, Lee C, Park H, et al. What is twitter, a social network or a news media? [C] // Proceedings of the 19th International Conference on World Wide Web. ACM, 2010: 591-600

[7] Leo B. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32