Journal of Computer Applications
计算机应用,2015,35(7):1959 - 1964
ISSN 1001-9081
CODEN JYIIDU
2015-07-10
http://www.joca.cn

文章编号:1001-9081(2015)07-1959-06
doi:10.11772/j.issn.1001-9081.2015.07.1959

面向不平衡微博数据集的转发行为预测方法

赵　煜*,邵必林,边根庆,宋　丹
(西安建筑科技大学 管理学院,西安 710055)
(*通信作者电子邮箱 zhaoyu_xauat_mail@163.com)

摘　要:针对微博转发预测方法研究中的数据集不平衡问题,提出了一种融合过采样技术和随机森林(RF)算法的微博转发行为预测方法。首先,定义了个体信息、社交关系和微博主题3类与微博转发行为相关的特征,并基于信息增益算法实现了关键特征选取;其次,综合微博特征数据的特点来改进少数类样本合成过采样技术(SMOTE),对原始数据集进行非参数概率分布估计,并根据近似概率分布对数据集进行过采样处理,从而使正反例数据量达到平衡;最后,利用随机森林算法,依据微博转发关键特征进行分类器训练,并利用袋外(OOB)数据误差估计来分析和设置随机森林算法的相关参数。通过与基于决策树(DT)、支持向量机(SVM)、朴素贝叶斯(NB)和随机森林等算法的微博转发预测方法进行对比,所提方法整体性能优于基准方法中性能最优的SVM方法,召回率提高了8%,*F*值提高了5%。实验结果表明,所提方法在实际应用中能够有效提高微博转发行为预测的准确率。

关键词:微博;转发预测;不均匀数据集;过采样;随机森林
中图分类号:TP391.1　　**文献标志码**:A

Prediction of retweeting behavior for imbalanced dataset in microblogs

ZHAO Yu*, SHAO Bilin, BIAN Genqing, SONG Dan
(*School of Management, Xi'an University of Architecture and Technology, Xi'an Shaanxi* 710055, *China*)

Abstract: Focusing on the issue that imbalanced dataset influencing the effect of prediction for retweeting behavior in microblogs, a novel predicting algorithm based on oversampling techniques and Random Forest (RF) algorithm was proposed. Firstly, the retweeting-related features, including individual information, social relationships and topic information, were defined. The key feature selection method was implemented based on information gain algorithm. Secondly, by considering the characteristics of the microblogs feature data, an improved algorithm for oversampling based on Synthetic Minority Oversampling Technique (SMOTE) was proposed. In the course of this algorithm, the probability distribution of the original dataset was estimated based on nonparametric distribution estimation. In order to ensure a balanced number of positive examples and negative examples, an oversampling method was executed based on the improved SMOTE method, according to approximate probability distribution of the original dataset. Finally, a classifier based on random forest algorithm was trained, according to retweeting-related key features. The algorithm parameters of random forest were selected by analyzing the error estimation of Out Of Bag (OOB) data. By comparison with Decision Tree (DT), Support Vector Machine (SVM), Naive Bayesian (NB) and RF algorithms, which were used in the analysis for microblog retweeting behavior, the overall performance of the proposed method is superior to the method based on SVM, which obtains optimal results in all the baseline methods. The recall rate and *F*-measure of the proposed method are improved by 8%, 5% respectively. The experimental results show that the proposed method can effectively improve the prediction accuracy of microblog retweeting behavior analysis in practical application.

Key words: microblog; retweet prediction; imbalanced dataset; oversampling; Random Forest (RF)

0　引言

作为一种基于用户关系的互联网信息传播媒介,微博传播具有时效性、随机性、自主性等特点,目前已成为互联网舆情扩散的主要方式,是网络大数据研究领域的焦点[1]。以腾讯微博为例,截止到2012年底,注册用户数量已达到5.4亿人次,全年的热门微博创建数达2 000万[2]。转发是构成微博网络的重要功能,用户通过转发可以把其他用户的微博内容共享于自己的微博上,同时,当一些社会问题被意见领袖在微博上转发,其影响会以几何级数方式在社会中扩散,在短时间内引起用户的共鸣,形成强大的舆论,对社会稳定造成极大影响,因此对微博中的转发行为进行建模,对于准确挖掘敏感微博的传播特征,开展网络舆情监控,干预、控制敏感信息传播的范围具有重要研究价值。

近些年,针对微博转发预测问题,国内外学者开展了广泛的研究,研究内容主要集中在预测特征选取和预测算法2个方面。在微博转发行为特征选取方面,Suh等[3]以Twitter为研究对象,将微博转发率的影响因素分为直接、间接和无关因

收稿日期:2015-01-21;**修回日期**:2015-03-18。　**基金项目**:国家自然科学基金资助项目(61272458)。
作者简介:赵煜(1981 -),男,陕西西安人,博士研究生,CCF会员,主要研究方向:数据挖掘、大数据处理;　邵必林(1965 -),男,云南腾冲人,教授,博士生导师,主要研究方向:云计算、大数据处理;　边根庆(1968 -),男,浙江浦江人,副教授,硕士,主要研究方向:云计算、云存储;宋丹(1991 -),女,陕西汉中人,硕士研究生,主要研究方向:数据挖掘、大数据处理。

素 3 类,并基于此建立微博转发次数预测模型;Xu 等[4]通过对社交、内容、tweet 相关以及作者等 4 类属性进行实验分析,验证了转发预测算法中的关键因素;Romero 等[5]通过分析微博主题与传播机制之间的关系,发现微博内容所属主题是决定微博转发概率大小的重要因素;Weng 等[6]关注微博用户影响力分析问题,并将关注关系、话题相似度引入分析方法中;Welch 等[7]则指出用户之间的转发关系比关注关系在微博影响力分析中的作用更显著;Morchid 等[8]利用主成分分析法进行转发属性选择,并在短时间大量转发行为检测中取得最佳分类效果。

国内外学者采用的预测算法主要包括条件随机场、支持向量机(Support Vector Machine, SVM)、贝叶斯模型和随机森林(Random Forest, RF)方法等 4 类:Peng 等[9]同时考虑了局部个体因素和全局网络因素,并结合随机场模型和网络划分方法提出了针对特定网络的转发预测算法。张旸等[10]分析了影响 Twitter 转发的特征因素,利用 SVM 建立了特征加权的预测模型,该研究未涉及用户之间的关系特征以及微博自身的主题特征。李英乐等[11]通过分析影响用户转发行为的诸多因素,提出了基于 SVM 的新浪微博转发行为预测方法。谢婧等[12]结合主题和用户特征,提出了一种用户转发行为预测方法。该方法首先利用互信息理论从已转发微博内容中提取特征,并依据微博内容分析了主题与转发行为之间的关系;再选取与微博转发行为相关的用户基本特征;最后利用贝叶斯模型,提出了面向特定用户的微博转发概率预测方法。Luo 等[13-14]研究了社会纽带关系对新浪微博转发行为的影响,并基于随机森林集成分类方法解决了微博转发预测问题,该研究仅关注了用户和用户间的关系属性,未考虑微博主题相关属性,因而是一种粗粒度的转发行为预测方法[14],不适用于现实微博舆情监控。

分析已有研究成果,学者们均未考虑数据集的平衡对预测算法的影响问题,即假设正反例所含的样本数大致相等。然而,对于已有分类算法,文献[15-16]指出在数据集不平衡的情况下,以总体分类精度为目标的算法会过多地关注多数类,而导致少数类样本的分类性能下降,同时,在微博信息的实际抽取中,由于受到返回结果数量、应用程序编程接口(Application Programming Interface, API)调用频率等因素的限制,难以获取全面的数据,存在不平衡数据集的可能性很大[17],因此,本文提出了面向不平衡数据集的微博转发行预测算法,首先求取少数类样本集的近似概率估计,并基于少数类样本合成过采样技术(Synthetic Minority Over-sampling Technique, SMOTE)算法的思想随机构建少数类伪样本,上述样本构造流程既保证伪样本的随机性,又保持伪样本与原样本集概率分布近似一致;其次,以个体信息、社交关系和微博主题等 3 类特征为依据,提出基于随机森林分类器的转发行为预测算法;最后,通过对比实验证实了该方法的有效性和可行性。

1 问题定义

目前,关于微博网络建模通常是基于图假设,即将整个微博用户网络描述为有向无权图 $G = \langle U, E \rangle$。其中:U 表示节点集合,用于描述微博网络中的用户;E 表示边集合,($E = \{e = \langle u_1, u_2 \rangle \mid u_1, u_2 \in U, u_1 \neq u_2\}$),用于描述微博网络的拓扑结构。在研究微博信息的转发行为时,将边集合 E 定义为用户之间的 follow(关注)关系,其方向表示信息传播的方向。

基于以上关于微博网络的描述,本文的研究问题可描述为:对于关注边 $u_a \rightarrow u_b$,即用户 b 关注用户 a。若用户 a 发布了一条微博信息 m,预测关注者 b 是否会转发微博信息 m,其形式化表示如式(1)所示,这是一个二分类问题,预测特征包括 3 类:个体信息(u_a, u_b)、社交关系($u_a \rightarrow u_b$)和微博内容(m)。

$$y = f(u_a, u_b, u_a \rightarrow u_b, m);\ y \in \{1,0\} \tag{1}$$

2 微博转发行为预测的特征选取

文献[8]通过实验证明选取具有充分辨别能力的特征,会有效提高转发行为预测算法的性能,因此,对微博转发行为相关特征的分析与选取是本文的首要步骤。依据问题的定义,预测特征包括个体信息、社交关系和微博主题等 3 类。

2.1 个体信息特征

1)用户的基本特征包括 8 个属性,即{省份 ID,城市 ID,性别,关注数,粉丝数,微博数,收藏数,加 V 用户}。

2)发布者影响力。采用 PageRank 算法来评价用户的影响力,计算方法如式(2)所示:

$$p(u_a) = (1-s) + s * \sum_{i \in S_{af}} \frac{p(u_i)}{C_{f_i}} \tag{2}$$

其中:$p(u_a)$ 表示节点 u_a 的影响力;$p(u_i)$ 表示节点 u_i 的影响力;s 表示阻尼系数,取值区间为[0,1];S_{af} 表示用户 a 关注的用户集合;C_{f_i} 表示用户 i 的粉丝数。

3)接收者行为活跃度。与其他用户比较,转发或原创微博次数越多的用户,其再次转发微博的可能性越大;同时,转发次数与原创次数比较,转发次数占用户微博行为(即转发+原创)次数的比重越大,用户转发微博的概率也越大。因此,接收者行为活跃度计算时应考虑以上两种因素,计算用户 i 活跃度的方法如式(3)所示:

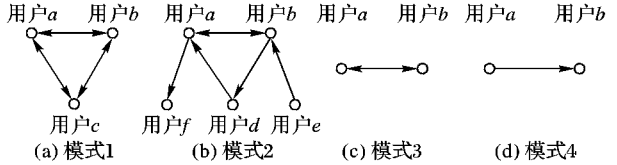
$$A_i = \left(nr_i / \sum_{i \in all} nr_i + np_i / \sum_{i \in all} np_i\right) * \frac{nr_i}{(nr_i + np_i)} \tag{3}$$

其中:nr_i 为数据抽样时间内,用户转发微博次数;np_i 为数据抽样时间内,用户原创微博次数。

2.2 社交关系特征

1)用户之间的微关联结构。

针对关注边 $u_a \rightarrow u_b$ 定义 4 种不同类型网络结构,在模式 1 中,用户 a 和 b 相互关注,且至少存在一个用户同时与他们相互关注,即存在强连通的三角关系;在模式 2 中,用户 a 和 b 相互关注,但没有其他用户与他们存在相互关注关系,仅存在单向关注关系;在模式 3 中,用户 a 和 b 仅相互关注,不存在其他用户与之存在关注关系。在模式 4 中,用户 a 和用户 b 中仅存在单向关注关系。用户间微关联结构如图 1 所示。显而易见,从模式 1 到模式 4,用户之间的关联关系越来越弱,与之对应的转发概率也越来越小。



(a) 模式1 (b) 模式2 (c) 模式3 (d) 模式4

图 1 用户之间的微关联结构

2)权威比率。

对于关注边 $u_a \rightarrow u_b$,权威比率描述了两用户不同的社会

权重关系。其中用户在微博中的权重可以用其粉丝数来标记,则关注边 $u_a \rightarrow u_b$ 的权威比率 ρ 如式(4)所示:

$$\rho = C_{fa}/C_{fb} \tag{4}$$

其中:C_{fa} 和 C_{fb} 分别表示用户 a 和用户 b 的粉丝数,ρ 值描述了用户 a 微博影响力高于用户 b 的程度,显然微博影响力差距越大,用户 a 对用户 b 行为的影响力越大。

3)用户间性别关系。

针对关注边 $u_a \rightarrow u_b$ 定义 4 种关系,即 MM(Male and Male)、MF(Male and Female)、FM(Female and Male)、FF(Female and Female),其中 MM 表示关注者与被关注均为男性,其他关系同理。

2.3　微博主题特征

微博主题精炼地描述了微博信息,构成了传播信息的关键特征,同时,用户通常仅关心特定主题的微博,主题关注程度是转发行为发生与否的重要影响因素,一定时间内用户发布微博的主题词集合构成了用户当前的微博关注空间,因此,通过提取微博主题词和数据抽取时段内用户发表微博的主题词集合,计算词汇间相似度便可以获得特定用户对特定微博的关注度,具体流程如图 2 所示。

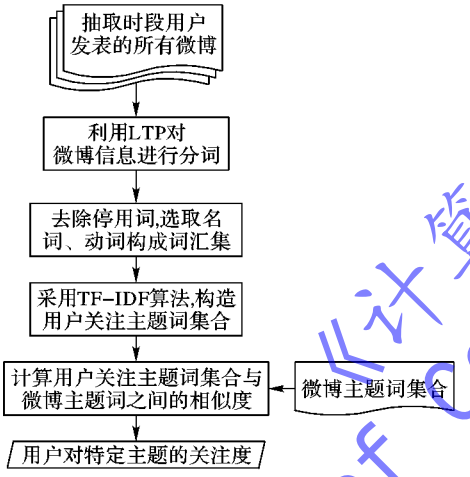


图 2　微博主题关注度计算流程

本文采用哈尔滨工业大学信息检索研究中心的语言技术平台(Language Technology Platform, LTP)对数据集进行分词。关注主题词集合与微博主题词集合之间的相似度计算采用向量之间余弦值的求解方法,其中关注主题词集合、微博主题词集合均利用词汇的 TF-IDF(Term Frequency-Inverse Document Frequency)计算结果构建。微博主题词集合的构造过程包括 5 个步骤:

步骤 1　采用人工方式选取少量与特定话题相关联的主题词,构成初始微博主题词集合;

步骤 2　利用步骤 1 选定的主题词,从微博中爬取相关的微博内容数据;

步骤 3　对爬取到的数据进行分词、去除停用词、选取名/动词等处理,并基于 TF-IDF 计算结果构造微博主题词集合;

步骤 4　采用人工方式来最终确定步骤 3 构建的微博主题词集合;

步骤 5　针对一篇微博,选择与其最相似的微博主题词集合用于微博主题关注度计算,相似度计算中采用向量之间余弦值的度量方法。

3　基于概率估计的过采样算法

针对不平衡数据集,主要采用过采样和欠采样 2 类处理方法。过采样方法依据少数类数据集,创建与少数类数据集近似一致的伪样本。欠采样方法通过消除多数类数据集中的冗余数据,从而实现数据集的平衡。由于欠采样涉及的冗余数据清除过程具有较高的时间复杂度,且容易造成有价值信息的丢失,因此微博数据集平衡过程宜利用过采样方法。

随机过采样是平衡数据集大小的最简单方式,其原理是复制少数类样本使数据量趋于一致。虽然该方法计算复杂度低,但可能产生过拟合的分类器,效果不佳。SMOTE 是一种智能过采样方法,由于其采用近邻插值策略,从而明显改善了分类过拟合现象。SMOTE 方法的主要步骤如下:1)针对任意少数类样本,在少数类数据集中寻找 k 个近邻样本;2)从 k 个近邻样本中随机选取 N 个样本;3)将该样本分别与 N 个近邻样本进行随机线性组合,从而构造 N 个伪少数类样本。

由于通过新浪 API 抽取的用户及微博内容数据量庞大,经处理后获得的数据类型多样,既包含离散数据又包含连续数据,采用寻找 k 近邻的插值策略将增加算法的计算复杂度,影响数据分类的效果。鉴于此,本文提出一种利用概率分布近似估计的随机插值方法。

该方法采用非参数的概率分布估计方法,针对连续型特征,首先将该特征的取值范围进行离散化,即将其划分为多个等长且无交集的取值区间;针对离散型特征,若特征取值范围较小(如微关联结构、性别关系、位置关系),则不进行任何处理;若特征取值范围较大(如关注数、粉丝数、微博数、收藏数),则首先对特征值进行排序,分别计算特征值最大差值和相邻特征值之差以及二者之间的比值,通过与阈值比较,将特征值集合划分为多个非交子集(离散化区间等宽且宽度为 0.1,算法性能最优[16]);最后,利用数据子集中样本频次估计少数类数据集的近似概率分布。针对连续型和离散广取值型特征的离散化方法如图 3 所示。通过分析图 3 过程可知,若离散阈值过小,则每个区间内的原始样本数量太少,伪样本与原始样本过于接近,样本构造过程缺乏足够的随机性;当阈值过大时,伪样本又无法与原始概率分布保持一致,会影响分类的性能。通过实验分析发现,当阈值为 0.1 时,二者之间基本上达到平衡。

根据特征数据的离散化方法,分别就连续型和离散广取值型特征的离散过程举例如下。

1)针对连续型特征,假设 10 个用户对“世界杯”话题的关注度分别为:{0.1,0.37,0.41,0.85,0.63,0.66,0.34,0.33,0.5,0.76},离散化后的概率估计如图 4 所示。

2)针对离散广取值型,假设 10 个用户的微博数分别为:{1,1,1,2,0,1,0,9,4,21}离散化后的概率估计如图 5 所示。

根据图 4 和图 5,可以得到少数类样本特征值的近似概率分布情况,过采样方法生成不同样本的概率应与原数据集的概率分布近似一致。伪样本生成过程如下。

1)由原数据集的概率分布估计值,分别计算各取值区间对应的随机数区间。以图 4 为例,取值区间[0.3,0.4)的概率分布为 0.3,故计算可得该区间对应的随机数区间为(0.1,0.4];以图 5 为例,由于集合{0,1,2}的概率密度区间为 0.7,故其对应的随机数区间为(0,0.7]。依此类推,取值区间概率密度越大,则伪样本产生于该区间的概率也越大,二值呈正

相关关系。

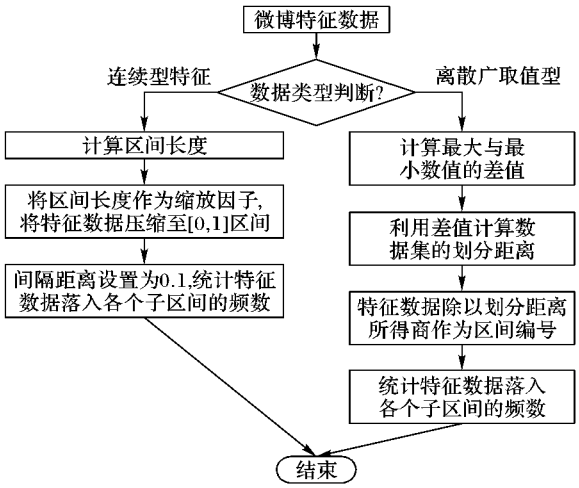


图 3 特征数据的离散化流程

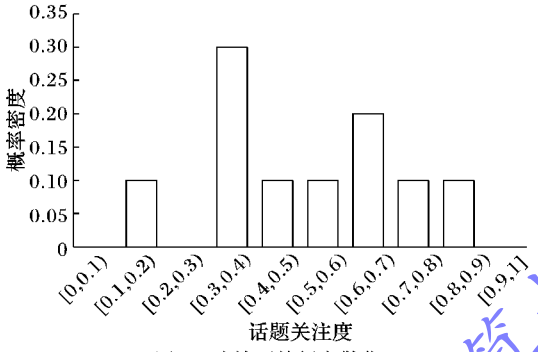


图 4 连续型特征离散化

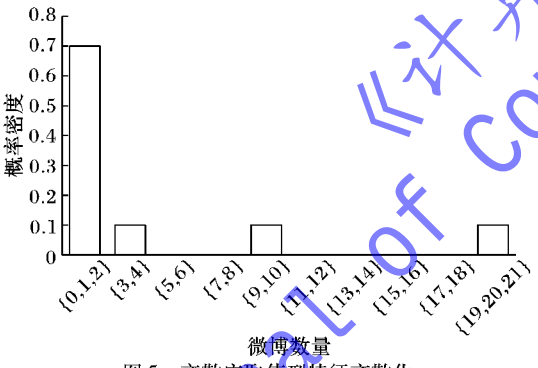


图 5 离散广取值型特征离散化

2)随机生成 0 到 1 之间的插值参数。

3)根据包含插值参数的随机数区间,查找其所对应的特征取值区,并从中随机选择某个特征数值作为伪特征。

4)合并所有伪特征构造伪样本。

这种数据采样方法既保证了伪样本的随机性,又使新样本与少数类数据集在概率分布上近似一致,新数据集更加趋于合理。

4 基于随机森林算法的微博转发预测

Breiman[18]首先提出了随机森林算法,并从理论上证明该算法不会过度拟合于训练数据,且对未知实例具有很强的泛化能力,同时,随机森林采用基于投票的分类方式,也对不平衡数据所带来的分类影响有一定消减作用。另外,基于随机森林方法的分类过程会对特征进行重要性评估和选择,对于存在噪声和未知特征的环境具有很好的鲁棒性。目前,随机森林算法以其良好的分类性能和集成学习的特点已应用于文本[15]、生物信息[16]、图像等领域中,且在微博分析中取得了初步的成果[14],这就是本文选择该方法的原因所在。

随机森林算法是利用一组决策树分类器构造的集成分类算法,所有决策树分类算法是独立、并行执行,决策树分类器的投票结果决定集成分类算法的最优解,随机森林算法的形式化描述如式(5)所示:

$$H(x_i) = \arg\max_{y_j} \sum_k I(f_k(x_i) = y_j) \tag{5}$$

其中:$f_k(\ )$ 为第 k 个决策树分类器,$H(\ )$ 为集成分类结果,x_i 为第 i 个待预测数据。随机森林算法运行过程中,数据采样、投票机制、参数设置是 3 个关键问题。

1)数据采样包括行采样和列采样。行采样是指决策树分类器对输入的数据进行采样。目前常用的方法是 bootstrap 方法,这是一种有放回的采样方法。由于 bootstrap 方法应用于大数据集中效果不佳,因此宜采用 K 折交叉验证方法对微博数据集进行随机采样。该采样方法首先将数据集随机地划分为 K 个大小一致且不相交的子集,第 i 次的测试数据为第 i 个子集。

列采样是指决策树分类器对子样本的特征进行再次采样,即从数据的 M 个特征中以某种随机方法选取 m 个特征($m \ll M$)。列采样常用的方法有 2 种,分别是随机选择特征和随机选取特征变量的线性组合。与以上 2 种完全随机特征选取方法不同,本文采取基于信息增益的特征抽取方法,由于特征的信息增益值与其权重大小正相关,从而使影响微博转发行为权重大的特征更容易被抽取。

2)随机森林分类算法的投票机制主要分为 2 大类:简单投票机制和贝叶斯投票机制。

简单投票机制中,多个决策树首先进行分类预测,然后根据分类结果,并按照一票否决、多数原则或阈值表决等方法进行类别划分。贝叶斯投票机制有别于简单投票机制,该算法依据决策树已有的分类表现,利用贝叶斯定理为每个决策树设定一个权值。由于无法保证先验概率正确性假设,因此本文采用基于多数原则的简单投票机制。

3)决策树数和特征选取数的设置直接地影响随机森林算法的性能,本文利用袋外(Out Of Bag, OOB)数据误差估计来选取决策树数和特征选取数,当 OOB 估计最少时,则参数设置为最优。由于以上 2 个参数都对 OOB 误差估计有影响,需要进行 $K \times m$ 次参数组合实验对 OOB 误差估计进行比较,本文采用文献[14]方法简化实验过程。基于随机森林的微博转发行为预测算法如下。

输入:微博数据集 S、微博预测数据集 P。

输出:预测数据集 P 的分类标签。

模型训练如下:

1)利用 K 折交叉采样方法对 S 采样,得到 K 个数据子集 S_n;

2)针对数据子集 S_n,计算每个特征的信息增益,排序并排除小于设定阈值的特征;

3)在数据子集 S_n 中,从大于设定阈值的特征中,依据特征的信息增益值,随机选取 $m(m \ll M)$ 个特征,构成新的训练数据集 S_m;

4)利用 S_m,采用分类与回归树(Classification And Regression Tree, CART)算法构造决策树,且不进行剪枝;

5) 循环步骤 1) 到 4),构造 K 个决策树,生成随机森林。

转发微博预测如下:

1) 对数据集 P 中的每一个 x_i,每棵决策树独立进行预测;

2) 依据各决策树的预测结果,采用多数原则的简单投票方法计算 $H(x_i)$,即 x_i 所属类别。

5　转发预测实验结果与分析

本文实验所用数据采样自 2012 年 5 月的新浪微博,微博爬取程序利用新浪微博 API[17] 以特定用户为采样入口,爬取策略依据广度优先原则。爬取过程首先利用关注读取接口和粉丝读取接口爬取微博用户间的关注关系网;其次,利用微博读取接口提取用户发表的微博;最后,利用微博读取接口获取每一条微博的转发微博,并抽取转发用户的信息和最新微博。利用上述微博数据爬取方法,最终获得用户记录 13 087 条,提取微博数据 86 351 条。本文利用著名数据挖掘平台 Weka 进行实验。

5.1　算法参数设置

Breiman[19] 指出袋外数据可以替代测试集进行误差估计,并已证明 OOB 误差估计是无偏估计。目前已有随机森林算法通常利用 OOB 误差估计确定随机森林中决策树个数和决策树特征数。借鉴文献[14]提出的算法参数确定过程,通过依次固定决策树数量和特征数 2 个变量中的一个变量,观察 OOB 误差估计随另一个变量的变化,最终选取决策树数量为 21,决策树特征为 6 个。

本文利用信息增益算法对微博转发行为预测进行特征选取,微博转发行为预测涉及的 3 类特征的信息增益值如图 6 所示。分析图 6 可得,用户基本特征的信息增益值趋近于 0,发布者影响力、权威比、接收者行为活跃度、用户之间的微关联结构图、微博主题关注度、性别关系的信息增益值最大,说明以上 6 个特征对于微博行为预测问题具有显著作用,故选择以上 6 个特征用于预测算法。

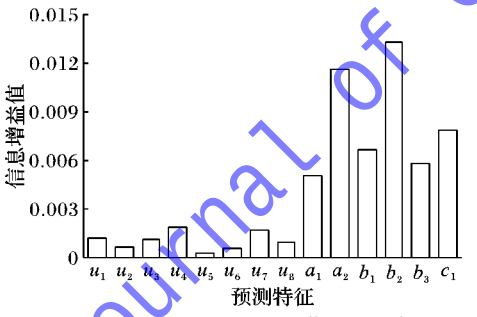


图 6　转发行为相关特征的信息增益值

5.2　实验结果与分析

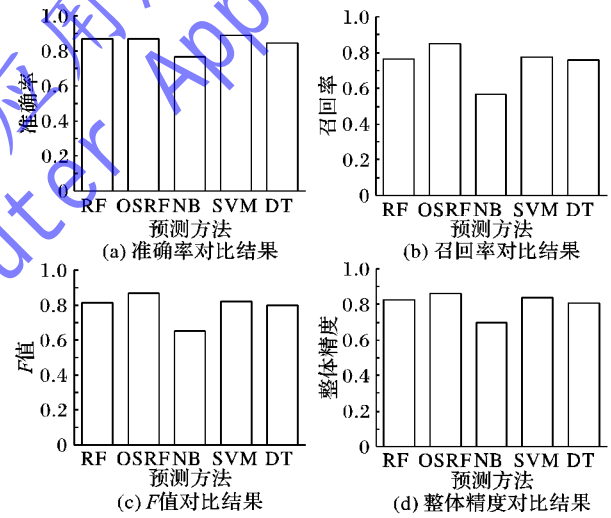
利用微博实验数据,本文构造一个不平衡数据集,其中微博转发样本数为 10 000,微博未转发样本数为 20 000,数据不平衡度为 2。为了验证本文 OSRF(Over Sampling Random Forest)方法的有效性,将实验分为两个部分。实验 1 的目的是验证基于概率估计的过采样方法是否对转发预测效果具有提高作用,因此,实验中分类方法保持一致,均采用随机森林(RF)算法。在微博转发行为预测和不平衡数据集分类研究中,文献[10-11,16]采用基于混淆矩阵的分类器验证方法,本文亦采用该评价方法,实验结果的混淆矩阵如表 1 所示。

由表 1 可见,由于采用了基于概率估计的过采样方法,本文预测方法较基本随机森林预测方法,转发样本被正确预测的比例显著提高,达 8.8%,说明基于概率估计的过采样方法所构造的伪样本与原始数据集近似一致,验证了该过采样方法的有效性。虽然,未转发样本被正确预测的比例略有下降,却说明了分类器更加趋近于真实分类边界。

表 1　过采样方法验证实验结果　%

事实	RF		OSRF	
	转发	未转发	转发	未转发
转发	76.3	23.7	85.1	14.9
未转发	11.5	88.5	12.8	87.2

实验 2 中,本文将微博转发行为预测常用的决策树(Decision Tree, DT)、SVM、朴素贝叶斯(Naive Bayesian, NB)分类器、RF 与本文算法进行比较,用于验证 OSRF 算法的性能。面向不平衡数据集,通常不采用整体分类指标来评价分类结果[16],因此验证指标的计算仍然利用分类结果混淆矩阵。在微博转发行为预测和不平衡数据集分类研究中,文献[10-12,14,16]均将正类(文中指转发类)的分类准确率、召回率、F 值作为算法性能的评价指标,本文也采用上述评价指标,同时还将衡量分类算法的整体精度,对比实验结果如图 7 所示。



(a) 准确率对比结果　(b) 召回率对比结果

(c) F值对比结果　(d) 整体精度对比结果

图 7　对比实验结果

由图 7 可得,针对转发样本类预测问题,OSRF 算法的整体性能优于对比方法中分类效果最佳的 SVM 分类结果,其中召回率提高了 8%,F 值提高了 5%。虽然在准确率方面低于 SVM 方法 2%,但分析实验数据可知,其中主因是由 SVM 方法在非转发类预测中召回率达到最高而形成的,实际意义小于转发类预测结果,因此,实验 2 表明,本文算法处理非平衡微博数据集是可行的,并能够有效提高微博转发行为预测的准确率。

6　结语

微博转发行为是一种需要重点研究的信息传播机制。本文首先分析了用户间的个人特征、社交信息、微博主题等 3 类重要特征对微博转发行为的影响;其次,面向不平衡数据的预测问题,提出了利用过采样思想、根据样本近似概率估计随机构造伪样本的方法;最后,融合与微博转发行为显著相关的用户和微博特征,提出了基于随机森林分类算法的转发行为预测方法。通过与已有主流分类方法进行对比实验得到如下结论。

1) 基于过采样思想,并利用近似概率估计随机构造的伪

样本既具有一定的随机性,又保持了原始样本的概率分布,从而有效提升了数据质量,解决了数据不平衡现象严重影响小类预测精度的问题。

2)融合用户信息、社交关系和主题信息的集成学习方法,明显提高了微博转发行为预测的精度,证明本文方法在解决面向主题的微博行为预测问题时是有效而且可行的。

参考文献:

[1] WANG Y, JIN X, CHENG X. Network big data: present and future [J]. Chinese Journal of Computers, 2013, 36(6): 1125 – 1138. (王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望[J]. 计算机学报, 2013, 36(6): 1125 – 1138.)

[2] HUANG Y, SUN X, LIU Z, *et al.* The microblog retweeting prediction evaluation system and performance comparation [J]. Journal of Harbin University of Science and Technology, 2013, 18(4): 52 – 57. (黄英来, 孙晓芳, 刘镇波, 等. 微博转发预测算法评测系统的建立及性能比较[J]. 哈尔滨理工大学学报, 2013, 18(4): 52 – 57.)

[3] SUH B, HONG L C, PIROLLI P, *et al.* Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network [C]// Proceedings of the 2010 IEEE International Conference on Social Computing. Piscataway: IEEE, 2010: 177 – 184.

[4] XU Z, YANG Q. Analyzing user retweet behavior on twitter [C]// Proceedings of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Piscataway: IEEE, 2012: 46 – 50.

[5] 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. New York: ACM, 2011: 695 – 740.

[6] WENG J, LIM E P, JIANG J. TwitterRank: finding topic-sensitive influential twitterers [C]// Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 261 – 270.

[7] WELCH M J, SCHONFELD U, HE D. Topical semantics of twitter links [C]// Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM, 2011: 327 – 336.

[8] MORCHID M, DUFOUR R, LINARES G, *et al.* Feature selection using principal component analysis for massive retweet detection [J]. Pattern Recognition Letters, 2014, 49(11): 33 – 39.

[9] PENG H, ZHU J, PIAO D Z, *et al.* Retweet modeling using conditional random fields [C]// ICDMW'11: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. Washington, DC: IEEE Computer Society, 2011: 336 – 343.

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

[11] LI Y, YU H, LIU L. Predict algorithm of micro-blog retweet scale based on SVM [J]. Application Research of Computers, 2013, 30(9): 2594 – 2597. (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.)

[12] XIE J, LIU G, SU B, *et al.* Prediction of user's retweet behavior in social network [J]. Journal of Shanghai Jiaotong University, 2013, 47(4): 584 – 588. (谢婧, 刘功申, 苏波, 等. 社交网络中的用户转发行为预测[J]. 上海交通大学学报, 2013, 47(4): 584 – 588.)

[13] LUO Z, WU X, CAI W, *et al.* Examining multi-factor interactions in microblogging based on log-linear modeling [C]// Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York: ACM, 2012: 189 – 193.

[14] LUO Z, CAI W, CHEN T. Microblogging retweet prediction algorithm based on random forest [J]. Computer Science, 2014, 41(4): 62 – 64. (罗知林, 蔡皖东, 陈挺. 一种基于随机森林的微博转发预测算法[J]. 计算机科学, 2014, 41(4): 62 – 64.)

[15] FANG X, ZHANG H, GAO S. Web spam detection based on SMOTE and random forests [J]. Journal of Shandong University: Engineering Science, 2013, 43(1): 22 – 26. (房晓南, 张化祥, 高爽. 基于 SMOTE 和随机森林的 Web spam 检测[J]. 山东大学学报: 工学版, 2013, 43(1): 22 – 26.)

[16] YU H, GAO S, ZHAO J, *et al.* Classification for imbalanced microarray data based on oversampling technology and random forest [J]. Computer Science, 2012, 39(5): 190 – 194. (于化龙, 高尚, 赵靖, 等. 基于过采样技术和随机森林的不平衡微阵列数据分类方法研究[J]. 计算机科学, 2012, 39(5): 190 – 194.)

[17] LIAN J, ZHOU X, CAO W, *et al.* SINA microblog data retrieval [J]. Journal of Tsinghua University, 2011, 51(10): 1300 – 1305. (廉捷, 周欣, 曹伟, 等. 新浪微博数据挖掘方案[J]. 清华大学学报: 自然科学版, 2011, 51(10): 1300 – 1305.)

[18] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5 – 32.

[19] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123 – 140.

(上接第 1958 页)

[10] LIU C, LI D, DU Y, *et al.* Statistics on atomized feature of normal cloud model [J]. Journal of Beijing University of Aeronautics and Astronautics, 2005, 34(2): 236 – 239. (刘常昱, 李德毅, 杜鹢, 等. 正态云模型的统计分析[J]. 信息与控制, 2005, 34(2): 236 – 239.)

[11] LI D, DU Y. Artificial intelligence with uncertainty [M]. Beijing: National Defense Industry Press, 2005: 237 – 248. (李德毅, 杜鹢. 不确定性人工智能[M]. 北京: 国防工业出版社, 2005: 237 – 248.)

[12] PHAM D, CHAN A. Control chart pattern recognition using a new type of self organizing neural network [J]. Journal of Systems and Control Engineering, 1988, 212(2): 115 – 127.

[13] KANNAN S R, RAMATHILAGAM S, CHUNG P C. Effective fuzzy *c*-means clustering algorithms for data clustering problems [J]. Expert Systems with Applications, 2012, 39(7): 6292 – 6300.

[14] BELHAOUARI S. Fast and accuracy control chart pattern recognition using a new cluster-*k*-nearest neighbor [J]. International Scholarly and Scientific Research & Innovation, 2009, 3(1): 970 – 974.