一种用于微博谣言检测的半监督学习算法*

路同强[1,2]，石 冰[1]，闫中敏[1]，周 珮[1]
(1. 山东大学 计算机科学与技术学院，济南 250101；2. 中国人民解放军 61516 部队，北京 100094)

摘 要： 在微博谣言检测中，对微博谣言进行正确标注需要耗费大量的人力和时间，同时数据类别的不平衡也影响了微博谣言的正确识别。为了解决该问题，提出一种基于 Co-Forest 算法针对不平衡数据集的改进方法，利用 SMOTE 算法和分层抽样平衡数据分布，并通过代价敏感的加权投票法来提高对未标记样本预测的正确率。该方法只需要对少量训练数据实例进行谣言类别标注即可有效检测谣言。10 组 UCI 测试数据和 2 组微博谣言的实证实验证明了算法有效性。

关键词： 微博；谣言检测；不平衡数据；半监督学习；Co-Forest 算法；SMOTE；代价敏感
中图分类号： TP181；TP301.6 **文献标志码：** A **文章编号：** 1001-3695(2016)03-0744-05
doi:10.3969/j.issn.1001-3695.2016.03.024

Semi-supervised learning algorithm applied to microblog rumors detection

Lu Tongqiang[1,2], Shi Bing[1], Yan Zhongmin[1], Zhou Pei[1]
(1. *School of Computer Science & Technology, Shandong University, Jinan* 250101, *China*; 2. *Troop* 61516, *PLA, Beijing* 100094, *China*)

Abstract: In microblog rumor detection, labeling microblog rumors correctly requires a huge amount of manpower and time. At the same time, imbalanced data category also affects the correct recognition of microblog rumors. To resolve this problem, this paper proposed an improved method based on Co-Forest algorithm, which could be used for imbalanced dataset. This method used SMOTE algorithm and stratified sampling to balance the data's distribution. Besides, it improved the correct rate of unlabeled sample through the cost-sensitive weighted voting method. This method required only a small amount of training data instances which labeled a rumor category, and could be used to detect rumors effectively. Experiment results on 10 UCI data sets and 2 microblog rumors prove that the algorithm is effective.

Key words: microblog; rumor detection; imbalanced data; semi-supervised learning; Co-Forest algorithm; SMOTE; cost sensitive

0 引言

谣言检测属于互联网信息可信度[1]研究范畴，是互联网信息可信度研究的新方向。国外学者对社交网络和微博尤其是 Twitter 可信度作了大量的研究[2~6]。其研究工作首先从 Twitter 上抓取数据，去除与特定事件话题无关的 tweets 作为样本数据；接着选取 tweets 文本内容包含的元素统计特征(如标签个数、链接个数等)或者浅层的内容信息(如情感词个数、名词个数等)作为特征，辅以用户信息和传播特征等，运用分类(如 naïve Bayes、support vector machine、decision tree 等)[2~6]或排序[7,8]的方法对微博信息可信度进行评估。国内学者[10~12]对微博谣言检测的研究与国外学者基本相似。这些研究基本上采用基于监督学习的方法，均需要训练数据被正确地标注；同时，为了提高分类的精度，需要训练样本足够大，需要耗费大量的人力和时间。然而在微博平台中存在着海量数据，并且每天都会产生大量的数据。以目前国内最广泛使用的新浪微博为例，其每天发布的内容就超过了 1 亿条，对这些海量数据进行内容真伪的判别并正确地标注是一项艰巨的任务，需要耗费大量的人力和时间，这也增加了训练样本集的成本。

半监督学习是利用未标记示例的主流学习技术之一，可大大减小数据标注的高昂代价。Li 等人[13]在 Co-training 基础上提出一种不需要充分冗余属性子集的半监督学习算法 Co-Forest，能使用大量的未标注数据迭代优化在标注数据上学得的假设，在实际应用(计算机辅助诊断)中取得了良好效果。微博中谣言的数量远少于非谣言，同时准确识别谣言比识别非谣言价值更高，因此本文认为微博谣言检测是一个不平衡数据的二分类问题。但由于 Co-Forest 算法采用传统的随机森林(random forest)算法来保证各分类器之间的差异性，使得 Co-Forest 算法在不平衡数据集上不能较好地识别少数类，不能直接应用到微博谣言检测中。

本文综合考虑上述几方面的问题，提出一种基于 Co-Forest 算法针对不平衡数据集的改进方法——ImCo-Forest 算法(semi-supervised learning algorithm from imbalanced data based on Co-Forest)，利用 SMOTE 算法和分层抽样平衡数据分布，并通过引入代价因子来提高对未标记样本预测的正确率。实验表明，该方法在处理类别不平衡分类问题时较 Co-Forest 算法优越，能有效提升少数类样本的识别准确率。最后，本文还将

收稿日期：2014-12-15；修回日期：2015-01-28 基金项目：国家自然科学基金资助项目(61303005)

作者简介：路同强(1986-)，男，助理工程师，硕士，主要研究方向为机器学习、数据挖掘(lutongq@163.com)；石冰(1957-)，男，教授，硕士，主要研究方向为数据库、数据挖掘、机器学习；闫中敏(1977-)，女，副教授，博士，主要研究方向为 Web 数据集成、数据库；周珮(1990-)，女，硕士，主要研究方向为数据挖掘.

该算法应用到微博谣言检测中,进一步验证算法的有效性和可行性。

1　相关基础

1.1　不平衡数据分类

目前,解决不平衡数据分类问题主要有数据层和算法层两种策略[14]。数据层的方法又称为重抽样,它通过对正(少数)类数据重复采样或随机地从多数(负)类中删除元组,使得结果训练集包含相同个数的正元组和负元组,以此改变训练集中的元组分布。算法层的方法主要对分类模型结构进行改变,即修改已有的分类算法,通过调节不同类样本之间的成本函数、改变概率密度、调整分类边界、调整分类决策等措施使其更有利于少数类的分类。

在数据层面,Chawla 等人[15]提出的 SMOTE 算法是一种简单有效的过采样方法。其主要思想是在近邻少数类样本之间进行线性差值,通过人工合成新的少数类样本来降低类别的不平衡。具体做法是: 首先为每个少类样本 x 随机选出其 k 个最邻近少数类样本 $x_i(i=1,\cdots,k)$; 然后根据给定的向上采样倍率 K,随机选择 K 个样本,记为 $y_1,y_2,\cdots,y_K$; 接下来在 x 与 y 之间使用式(1)构造新的少数类样本 x_{new}; 重复以上步骤,直至所有少数类样本均处理完为止。该方法使分类器的分类平面向多数类的空间伸展,同时可有效地避免随机向上采样的过学习问题。

$$x_{new}=x+\text{random}(0,1)\times(x_i-x) \tag{1}$$

其中: $i=1,2,\cdots,k$; random(0,1)表示产生一个 0 ~ 1 的随机数。

在算法层面,一种比较有效的方法是对学习算法进行一定的修改,使其转变为代价敏感学习方法(cost sensitive learning, CSL)[16],即考虑不平衡类数据中正确识别少数类比多数类具有更大价值的事实,在传统分类算法中引入代价因子,对少数类样本赋予较高的代价,多数类样本赋予较小的代价,迫使最终分类器对正类样本有更高的识别率。Ting[17]将代价敏感学习引入随机决策树的训练中,提出了代价敏感的决策树算法,能有效提升不平衡数据中少数类的识别精度。Chen 等人[18]利用代价敏感思想,在构建决策树的属性分裂过程中引入权重因素,并通过 ROC 反馈调整权重因子来获得最优权值,构造出权重随机森林算法(WRF)。

1.2　Co-Forest 算法

Co-Forest 算法是基于半监督学习算法中协同训练算法的改进算法。其基本思想是用 N 个分类器来代替协同训练算法中的两个分类器。当为分类器集合中的一个组件 $h_i(i=1,\cdots,N)$ 确定最确定的标记实例时,使用不包含分类器 h_i 的集成学习分类器组合 H_i 来计算未标记实例的置信度。如果置信度超过预先设置的阈值 θ,将其标记并加入到新的标记数据集合 L'_i 中。再用这个标记数据集和原有的标记数据集对分类器 h_i 进行优化。通过使用这种方法,Co-Forest 首先在已标记数据集上训练一个分类器集合,然后使用伴随分类器选取的未标记数据优化每个元分类器。

Co-Forest 算法采用随机森林来保证各分类器之间的差异性。然而由于随机森林在构建过程中使用装袋(bagging)随机选取训练集,如果训练集中的少数类数据量较少,就会使得 N 个随机选取的训练集中含有的少数类的数量比原有的数据集更少或者没有,从而加剧数据集的不平衡性,使得基于此数据集训练出来的决策树的规则不能很好地体现少数类的特点。

2　基于 Co-Forest 的改进算法

本文基于 Co-Forest 算法,提出了新的半监督学习算法 ImCo-Forest。该算法旨在通过改善训练集中少数类的分布来提高基分类器对少数类的识别能力,从而使得集成后的分类器对少数类有效识别。设 $L=\{(x_1,y_1),\cdots,(x_l,y_c)\}$ 表示已标记样本,$y_i\in\{1,\cdots,c\}$,$U=\{(x_1,y_u),\cdots,(x_j,y_u)\}$ 表示未标记样本,并且 $l<<j$。算法的目的是预测未标记样本的类标签 y_u。

2.1　代价引入及终止条件

在对未标注数据进行标注时,为了提高少数类识别的准确率,引入代价敏感的思想。给定代价矩阵 Cost(i, j),如表 1 所示,其中 Cost(i, j) 表示将类 j 错误分类为 i 的代价。

表 1　代价矩阵

	预测正类	预测负类
实际正类	Cost(1,1) = 0	Cost(0,1)
实际负类	Cost(1,0)	Cost(0,0) = 0

对实例 x 进行类别预测时,H_i 中每棵决策树首先得出训练集中类别 j 的概率估计 $P(j|x)$,通过最小化式(2)给出属于类别 j 的预测,并对最终类别采用多数投票决定。

$$H(x)=\underset{i}{\text{argmin}}\left(\sum P(j|x)\,C(i,j)\right) \tag{2}$$

其中: $P(j|x)$ 是把实例 x 分类为类别 j 的后验概率。

文献[13]指出,过大的自动标注数据可能影响所学假设的性能。为保证元分类器的多样性,只需要对部分数据进行标注。与 Co-Forest 算法相同,ImCo-Forest 算法中元分类器 h_i 在 m_0 大小的初始已标记数据集 L 和新标记 $m_{i,t}$ 大小的新标记数据集 $L'_{i,t}$ 的集合上优化自己。定义 H_i 在 $L'_{i,t}$ 上的错误率为 $e_{i,t}$,错误率通过袋外(out of bag, OOB)数据估计。H_i 在 $L'_{i,t}$ 被误分的实例加权为 $e_{i,t}W_{i,t}$,其中 $W_{i,t}=\sum_{j=0}^{m_{i,t}}\omega_{i,t,j}$,$\omega_{i,t,j}$ 为 H_i 在 $L'_{i,t}$ 对实例 x_j 的预测置信度。迭代更新终止条件通过式(3)判断:

$$0<\frac{e_{i,t}}{e_{i,t-1}}<\frac{W_{i,t-1}}{W_{i,t}}<1 \tag{3}$$

2.2　算法步骤

输入标记数据集 L,未标记数据集 U,ImCo-Forest 算法按照以下步骤执行:

a) 初始化

(a) 用 SMOTE 方法对 L 样本集中的少数类样本使用式(1)进行采样,使不平衡数据趋于平衡,新的样本集为 L_s。

(b) 对平衡后数据 L_s 进行自助抽样,产生 N 个训练子集分别作为训练集,构造包含 N 个随机决策树的随机森林。

b) 未标记样本预测过程

第 t 轮迭代,对于分类器 h_i:

(a) 若本轮错误率小于上一轮,对未标记样本集 U 进行抽样产生样本数量大小小于 $e_{i,t-1}W_{i,t-1}/e_{i,t}$ 的抽样子集 $U'_{i,t}$。

(b) 对任意给定的未标记数据,采用 $H_i(i=1,\cdots,N)$ 按照代价敏感的加权投票法进行标记。选择满足条件 $L_i=\{x|x\in U, H_j(x)=H_k(x)\}$ 且置信度大于给定阈值 θ 的样本生成新的标记样本 $L'_{i,t}$。

c) 迭代优化过程

对分类器 h_i，若本轮错误率小于上一轮且满足式(3)中所示条件，利用基于正负类的分层抽样方法抽样从 $L \cup L'_{i,t}$ 中抽取训练子集，更新分类器 h_i，具体地：

(a) 统计 $L \cup L'_{i,t}$ 中少数类样本数量 $n_{i,t}$；

(b) 根据 $n_{i,t}$ 大小从 $L \cup L'_{i,t}$ 中抽取相同数量的多数类，并与少数类样本组合形成子集 $L^{\text{sub}} \cup L'^{\text{sub}}_{i,t}$；

(c) 使用 $L^{\text{sub}} \cup L'^{\text{sub}}_{i,t}$ 优化。

d) 采用多数投票 $H^* = \arg\max_{y \in \text{label}} \sum_{i: h_i(x)=y} 1$ 决定类别。

2.3 ImCo-Forest 算法分析

与 Co-Forest 算法相比，改进的算法 ImCo-Forest 有两方面不同之处：a) 在初始已标注数据集和加入新标注数据的数据集上分别使用 SMOTE 算法和分层抽样平衡数据；b) 在未标注数据置信度预测时，引入代价因子，对少数类样本赋予较高的代价，多数类样本赋予较小的代价。ImCo-Forest 算法一方面通过 SMOTE 算法和分层抽样人为地将少数类的数量加大，平衡了数据分布，使得随机森林算法更稳定地发挥其优越性；另一方面通过加大少数类的误分类代价，使得分类器更加关注少数类。

具体地：

a) 对已标注数据集 L 采用 SMOTE 算法进行类别平衡，使得在计算未标记数据可信度时避免了决策树学习假设的偏斜；

b) 基于代价敏感的加权投票法计算未标记数据集合 U 中的每一个未标记数据的标记可信度，提高了少数类识别的准确率；

c) 对加入新标记数据的数据集 $L \cup L'_i$，采用基于正负类的分层抽样方法抽样，进一步保证了类别平衡，避免了因样本选取不当而造成分类性能恶化的问题。

3 实验与讨论

3.1 实验设置

采用 UCI 数据集进行算法验证是机器学习研究常用的办法，本文实验所用的测试数据集是在研究不平衡数据分类时常用的 10 个公开数据集，它们都是从 UCI(http://archive.ics.uci.edu/ml/)的机器学习数据库中获得的。其中，yeast、page-blocks 等数据集为多类数据集，本文将其中最少类作为少数类，其他类合并为多数类。具体信息见表 2。

表 2 UCI 测试数据集

数据集	属性数	总样本	正例	负例	目标类	不平衡度
cleveland	13	303	139	164	1	1.18
pima	8	768	268	500	1	1.87
iris0	4	150	50	100	1	2.0
haberman	3	306	81	225	2	2.78
vehicle3	18	846	212	634	3	2.99
cmc	9	1 473	333	1 140	2	3.42
yeast3	8	1 949	163	1 321	4	8.10
page-blocks0	10	5 472	559	4 913	2,3,4,5	8.79
page-blocks1	10	472	28	444	1 ,4	15.86
yeast4	8	1 484	244	1 240	3	28.10

对于每一个数据集，用十折交叉验证来评价。对每一折，训练数据根据给定的已标记率 μ 随机地分成已标记数据集合 L 和未标注集 U。例如，如果一个训练集包含 100 个实例，根据 20% 的标记率可以分为一组 20 个已标记实例和一组 80 个未标记实例的训练集。为了模拟不同量的未标记的数据，这里考察三种不同的已标记率，如 10%、20%、40%。注意到 L 和 U 中类分布与原始数据集相同。

实验基于 WEKA(http://www.cs.waikato.ca.nz/ml/weka/)平台进行，实验中参照文献[13]的参数设置，将 Random Forest 个数 N 设为 6，置信度阈值设为 0.75，其他参数为 WEKA 默认参数。算法中 SMOTE 采样倍率 $K = \text{round}(\text{IL}) - 1$。其中：round(IL)表示对 IL 四舍五入得到的数值，IL 指不平衡度，最近邻 $k = 5$。多数类误分代价固定为 $\text{Cost}(1,0) = 1.0$，少数类误分代价固定为 $\text{Cost}(0,1) = 2.0$。为便于对样本进行分类，首先对特征集合中特征向量采用线性归一化方法进行处理，使每个特征处于同一量纲之下。

3.2 评价指标

在不平衡数据分类任务中，衡量分类器的性能指标也与平时有所差异，常用评价标准有 F-measure 以及 G-mean 等。在两类别情形下，将训练样例少但具有高识别重要性的少数类视为正类，多数类视为负类。经过分类过程后，训练样例可以分为表 3 所示混淆矩阵中所表示的四种情况。

表 3 二分类问题混淆矩阵

类别	预测正类	预测负类
实际正类	*TP*	*FP*
实际负类	*FP*	*TN*

F-measure 指标是一种综合考虑查全率和查准率的分类评价指标。其定义为

$$\text{F-measure} = \frac{(1+\beta^2) \times \text{recall} \times \text{precision}}{\beta^2 \times \text{recall} + \text{precision}} \tag{4}$$

对于少数类来说，查全率 $\text{recall} = \frac{TP}{TP+FN}$，查准率 $\text{precision} = \frac{TP}{TP+FP}$。实验中 β 取值为 1，即当查全率和查准率都比较大时，F-measure 才可取得较大的值。

G-mean 表示的是少数类分类精度和多数类分类精度的几何平均值。只有在多数类和少数类的分类精度同时都高的情况下，G-mean 的值才最大。

$$\text{G-mean} = \sqrt{\text{sensitivity} \times \text{specificity}} \tag{5}$$

其中：$\text{sensitivity} = \frac{TP}{TP+FN}$，$\text{specificity} = \frac{TN}{FP+TN}$。

同时，为了与改进前算法作比较，采用机器学习中最常用的指标分类正确率 Accuracy 来衡量总体正确率。

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \tag{6}$$

3.3 结果与分析

图 1～3 分别显示了在不同已标注率下，改进后的 ImCo-Forest 与 Co-Forest 算法的性能对比。

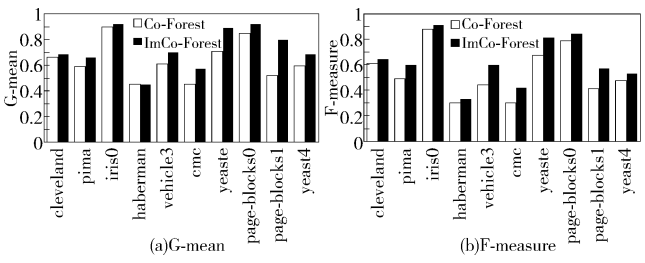
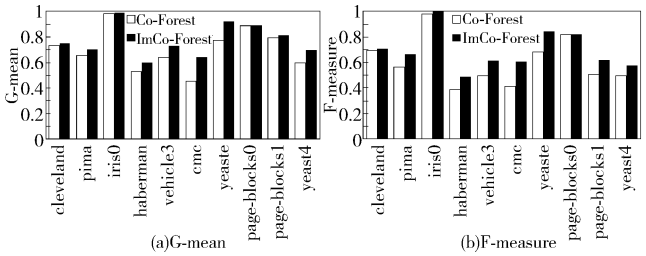
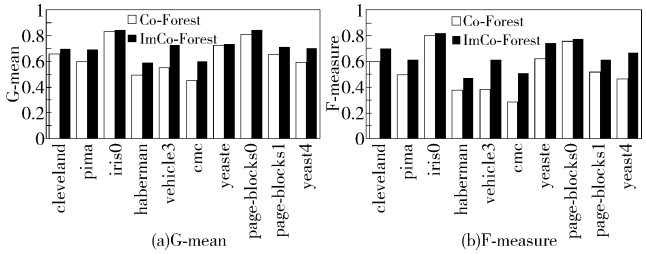


图 1 10% 已标注比例算法对比



(a)G-mean　(b)F-measure

图 2　20% 已标注比例算法对比



(a)G-mean　(b)F-measure

图 3　40% 已标注比例算法对比

从图中可以看出，对于 10 个不平衡数据集，与 Co-Forest 算法相比，除 10% 已标注比例下 haberman 数据集上本文算法的平均 G-mean 有所下降以外，本文算法的平均 G-mean、平均 F-measure 均有所显著提升，表明本文算法能较好地处理不平衡数据。

表 4 显示了不同已标注数据比例下算法的分类正确率。可以看出，在 20% 已标记数据比例下，算法结果整体指标最好；在 40% 已标注数据比例下，不管是 Co-Forest 算法还是改进后的 ImCo-Forest 算法，其性能均有所下降，这是因为这两种算法性能的提升主要得益于未标记数据，当未标注训练集较小时，通过自主抽样获得的初始多样性很有限，结果学习过程中的多样性快速下降，集成器的性能也相应下降，这也与文献[13]的实验结果相符。同时，从表 4 中可以看出，改进后的算法只有在少数数据集上分类正确率较 Co-Forest 算法有所小幅度下降，说明本文算法在提升少数类识别率的同时并没有降低对多数类的识别精度。

表 4　不同已标注比例下算法 Acc 比较　/%

数据集	Co-Forest			ImCo-Forest		
	10	20	40	10	20	40
cleveland	69.64	77.59	71.34	72.60	78.81	72.61
pima	70.31	72.52	70.05	72.00	75.12	72.15
iris0	94.00	98.90	89.33	96.67	98.90	89.50
haberman	72.23	71.55	73.13	72.22*	70.25*	73.25
vehicle3	69.15	74.58	66.56	82.35	77.21	77.21
cmc	76.10	75.49	73.18	75.67*	76.41	75.45
yeast3	93.67	94.07	91.55	95.97	97.10	91.10*
page-blocks0	95.98	96.44	95.56	96.71	96.55	96.10
page-blocks1	94.49	94.28	94.28	93.75*	94.28	95.12
yeast4	86.05	86.92	89.19	84.56*	84.92*	89.19

注：加* 表示相同标注比例时性能下降。

以上实验结果表明，本文所提算法在处理类别不平衡分类问题时较 Co-Forest 算法优越，能有效提升少数类样本的识别准确率。

4　微博谣言检测验证

本文以新浪微博为例进行谣言检测实证实验。实验首先从新浪微博抓取了两个热点事件相关的微博作为语料，以新浪微博官方的辟谣信息为依据，选取共 5 894 条微博进行人工标注。为了保证标注精度，对抓取的微博语料，采用两个组员分别独立地对语料进行人工标注，并通过计算 Kappa 系数来保证微博数据标注的一致性。实验采用文献[11,12]提出的微博文本内容特征、用户属性信息和微博传播特征三类基本特征中 16 个特征作为分类属性，这些特征已经被证实能有效检测微博谣言。检测数据集描述如表 5 所示。

表 5　微博语料数据集

谣言事件	属性数	总样本	谣言数	非谣言数	不平衡率
蓟县大火	16	2 909	849	2 060	2.42
少女遭毁容	16	2 985	312	2 673	8.57

通过 UCI 数据集上算法性能对比，可以看出 ImCo-Forest 算法主要通过未标记数据来提升性能。同时，本文的主要目的是减少人工数据标注的代价，因此，只考虑在少量已标注数据的情况下算法对微博谣言的检测性能。为了与其他已有工作比较，本文在 $L \cup U$ 上，当 $\mu = 0\%$ 情况下的数据集上训练 SVM、Bayes 及 J48 分类器和已标注比例 10% 的情况下对 Co-Forest 和 ImCo-Forest 算法性能进行了比较。实验采用本文 2.2 节中评价指标。实验结果如表 6 所示。

表 6　算法性能比较

算法	性能							
	蓟县大火				少女遭毁容			
	Acc	Precision	G-mean	F-measure	Acc	Precision	G-mean	F-measure
SVM	74.67	0.705	0.259	0.124	90.03	0.501	0.130	0.500
Bayes	61.93	0.327	0.538	0.367	86.01	0.250	0.422	0.216
J48	72.61	0.452	0.409	0.259	89.23	0.143	0.109	0.023
Co-Forest	65.05	0.388	0.530	0.378	89.05	0.330	0.345	0.179
ImCo-Forest	70.21	0.694	**0.601**	**0.633**	87.90	**0.620**	**0.650**	**0.874**

注：加粗表示性能表现最好。

从表 6 中可以看出，ImCo-Forest 算法在两个语料数据集上 G-mean 和 F-measure 指标均最好，说明其在处理非平衡数据问题时较其他算法有优势。值得注意的是，SVM 算法虽然在总体正确率上比较高，甚至在语料 2 上达到了 90.23% 的高正确率，但是 G-mean 和 F-value 都较低，说明对少数类的识别性能较差，表明其并不能准确识别谣言。

同时，实验对选取的三种监督学习算法的训练数据是全部标注的理想数据集，其结果仅 SVM 和 J48 算法在总体正确率指标上优于 ImCo-Forest，表明要达到较高的正确率，需要比 ImCo-Forest 算法更多的标注数据，这无疑中减弱了其在实际应用中的可行性。

以上实验结果表明，本文算法能在少量标注数据下较好地检测谣言，可在微博谣言识别中大大减小数据标注的代价。

5　结束语

本文算法是在 Co-Forest 算法基础上改进的，继承了半监督学习算法的优点，利用少量的已标记样本和大量的未标记样本进行分类。同时考虑了数据分布不平衡的问题，增强算法对少数类的识别性能，使得能在微博谣言检测中应用。对于下一步工作，将考虑对算法的抽样阶段进行改进，以避免由于样本

复制而导致的算法复杂度的提升，以期达到更好的效果。另外，对于不同的错分代价如何影响算法性能的研究，也是下一步将要做的工作。

参考文献：

[1] Metzger M J. Making sense of credibility on the Web: models for evaluating online information and recommendations for future research [J]. Journal of the American Society for Information Science and Technology, 2007, 58(13): 2078-2091.

[2] Wang A H. Don't follow me: spam detection in twitter [C] //Proc of International Conference on Security and Cryptography. [S. l.]: IEEE Press, 2010: 142-151.

[3] Qazvinian V, Rosengren E, Radev D R, *et al*. Rumor has it: identifying misinformation in microblogs [C] //Proc of Conference on Empirical Methods in Natural Language Processing. [S. l.]: ACL, 2011: 1589-1599.

[4] Castillo C, Mendoza M, Poblete B. Information credibility on twitter [C] // Proc of the 20th International Conference on World Wide Web. New York: ACM Press, 2011: 675-684.

[5] Suzuki Y. A credibility assessment for message streams on microblogs [C] //Proc of International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. [S. l.]: IEEE Press, 2010: 527-530.

[6] Mendoza M, Poblete B, Castillo C. Twitter under crisis: can we trust what we RT? [C] //Proc of the 1st Workshop on Social Media Analytics. New York: ACM Press, 2010: 71-79.

[7] Canini K R, Suh B, Pirolli P L. Finding credible information sources in social networks based on content and social structure [C] //Proc of the 3rd IEEE International Conference on Social Computing. Boston, MA: IEEE Press, 2011: 1-8.

[8] Gupta A, Kumaraguru P. Credibility ranking of tweets during high impact events [C] //Proc of the 1st Workshop on Privacy and Security in Online Social Media. New York: ACM Press, 2012.

[9] Gupta M, Zhao Peixiang, Han Jiawei. Evaluating event credibility on twitter [C] //Proc of SIAM International Conference on Data Mining. 2012: 153-164.

[10] 程亮，邱云飞，孙鲁. 微博谣言检测方法研究 [J]. 计算机应用与软件，2013，30(2)：226-228.

[11] Yang Fan, Yu Xiaohui, Liu Yang, *et al*. Automatic detection of rumor on Sina Weibo [C] //Proc of ACM SIGKDD Workshop on Mining Data Semantics. New York: ACM Press, 2012: 1-7.

[12] 贺刚，吕学强，李卓，等. 微博谣言识别研究 [J]. 图书情报工作，2013，57(23)：114-120.

[13] Li Ming, Zhou Zhihua. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples [J]. IEEE Trans on Systems, Man and Cybernetics, Part A: Systems and Humans, 2007, 37(6): 1088-1098.

[14] 叶志飞，文益民，吕宝粮. 不平衡分类问题研究综述 [J]. 智能系统学报，2009，4(2)：148-156.

[15] Chawla N V, Bowyer K W, Hall L O, *et al*. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[16] Elkan C. The foundations of cost-sensitive learning [C] //Proc of the 17th International Joint Conference on Artificial Intelligence. 2001: 973-978.

[17] Ting Kaiming. An instance-weighting method to induce cost-sensitive trees [J]. IEEE Trans on Knowledge and Data Engineering, 2002, 14(3): 659-665.

[18] Chen Chao, Liaw A, Breiman L. Using random forest to learn imbalanced data [R]. Berkeley: University of California, 2004: 1-12.

(上接第 743 页) 目标相距较近时依然可以分辨目标的数目并实现定位；但当多个目标相距太近时，多目标位置信息场定位法也可能无法分辨目标，将多个目标判为一个目标，出现漏警现象。

5 结束语

本文从最大似然估计定位法出发，分析了在单目标定位过程中存在测量误差较大或错误参数时最大似然估计定位法性能急剧下降的问题，引出了目标位置信息场定位法，通过加入代价函数使得该方法能够剔除错误数据，降低了误差大的数据对定位结果估计的影响，性能稳定。在对不可区分的多目标进行定位时，最大似然估计定位法只能得到一个多目标平均位置的估计，而目标位置信息场定位法可以同时得到目标的数目和位置。在后续的处理中，可以此目标数目和位置为基础，对测量参数按目标区分，转换为单目标定位问题用经典方法处理。然而文中对多目标位置信息场定位法的研究还不够深入，如多于两个目标的定位问题、关于各个目标测量参数信息不均衡条件下定位问题、定位的漏警虚警问题等，都将有待进一步研究，以便算法在工程中实际使用。

参考文献：

[1] Lee J Y, Hudson R E, Yao K. Acoustic DOA estimation: an approximate maximum likelihood approach [J]. IEEE Systems Journal, 2014, 8(1): 131-141.

[2] Shen Junyang, Molisch A, Salmi J. Accurate passive location estimation using TOA measurements [J]. IEEE Trans on Wireless Communications, 2012, 11(6): 2182-2192.

[3] 袁罡，陈鲸. 三站时差定位模糊问题解决方法 [J]. 中国电子科学研究院学报，2014，9(1)：89-92.

[4] Picard J S, Weiss A J. Time difference localization in the presence of outliers [J]. Signal Processing, 2012, 92(10): 2432-2443.

[5] Wang Zhi, Luo Ji'an, Zhang Xiaoping. A novel location-penalized maximum likelihood estimator for bearing-only target localization [J]. IEEE Trans on Signal Processing, 2012, 60(12): 6166-6181.

[6] 白晶，王国宏，王娜，等. 测向交叉定位系统中的最优交会角研究 [J]. 航空学报，2009，30(2)：298-304.

[7] 蔡晶晶，鲍丹，李鹏，等. 强约束优化降维 MUSIC 二维 DOA 估计 [J]. 电子与信息学报，2014，36(5)：1113-1118.

[8] Chen Chen, Zhang Xiaofei. A RD-ESPRIT algorithm for coherent DOA estimation in monostatic MIMO radar using a single pulse [J]. International Journal of Electronics, 2014, 101(8): 1074-1085.

[9] 张敏，郭福成，周一宇，等. 时变长基线 2 维干涉仪测向方法 [J]. 电子与信息学报，2013，35(12)：2882-2888.

[10] 韩月涛，潘伟萍，吴嗣亮，等. 干涉仪解模糊异常值检测及纠错方法 [J]. 北京理工大学学报，2012，34(8)：849-854.

[11] 马贤同，罗景青，张奎. 面向 DOA 测量的多目标位置信息场定位法 [J]. 信号处理，2013，29(1)：121-126.