基于双层采样主动学习的社交网络虚假用户检测方法

谭 侃 [1,2] 高 旻 [1,2] 李文涛 [1,3] 田仁丽 [1,4] 文俊浩 [1,2] 熊庆宇 [1,2]

摘 要 社交网络的飞速发展给用户带来了便捷, 但是社交网络开放性的特点使得其容易受到虚假用户的影响. 虚假用户借用社交网络传播虚假信息达到自身的目的, 这种行为严重影响着社交网络的安全性和稳定性. 目前社交网络虚假用户的检测方法主要通过用户的行为、文本和网络关系等特征对用户进行分类, 由于人工标注用户数据需要的代价较大, 导致分类器能够使用的标签样本不足. 为解决此问题, 本文提出一种基于双层采样主动学习的社交网络虚假用户检测方法, 该方法使用样本不确定性、代表性和多样性 3 个指标评估未标记样本的价值, 并使用排序和聚类相结合的双层采样算法对未标记样本进行筛选, 选出最有价值的样本给专家标注, 用于对分类模型的训练. 在 Twitter、Apontador 和 Youtube 数据集上的实验说明本文所提方法在标签样本数量不足的情况下, 只使用少量有标签样本就可以达到与有监督学习接近的检测效果; 并且, 对比其他主动学习方法, 本文方法具有更高的准确率和召回率, 需要的标签样本数量更少.

关键词 社交网络, 虚假用户, 主动学习, 样本多样性

引用格式 谭侃, 高旻, 李文涛, 田仁丽, 文俊浩, 熊庆宇. 基于双层采样主动学习的社交网络虚假用户检测方法. 自动化学报, 2017, **43**(3): 448−461

DOI 10.16383/j.aas.2017.c160308

Two-layer Sampling Active Learning Algorithm for Social Spammer Detection

TAN Kan[1,2] GAO Min[1,2] LI Wen-Tao[1,3] TIAN Ren-Li[1,4] WEN Jun-Hao[1,2] XIONG Qing-Yu[1,2]

Abstract With the rapid development of social network, more and more people join in social network to make friends and share their views. However, social network is always suffering from fake accounts due to its openness. Fake accounts, also called spammers, always spread spam information to achieve their own purpose, which have destroyed the security and reliability of social network. Existing detection methods extract behaviour, text and relationship features of users, and then use machine learning algorithms to identify social spammers. But machine learning algorithms often suffer from insufficiently labeled training data. Aiming to solve this problem, we propose an efficient algorithm, called two-layer sampling active learning, to construct an accurate classifier with minimum labeled samples. We present three criteria (uncertainty, representative and diversity) to quantity the value of unlabeled samples, using the combination of sorting and clustering to actively select samples with max uncertainty, max representative and max diversity. Experimental results on Twitter, Apontador, and Youtube datasets prove the efficiency of our approach, and better precision and recall of our approach than other active learning methods.

Key words Social network, spammer, active learning, diversity of samples

Citation Tan Kan, Gao Min, Li Wen-Tao, Tian Ren-Li, Wen Jun-Hao, Xiong Qing-Yu. Two-layer sampling active learning algorithm for social spammer detection. *Acta Automatica Sinica*, 2017, **43**(3): 448−461

收稿日期 2016-04-05 录用日期 2016-08-08
Manuscript received April 5, 2016; accepted August 8, 2016
国家重点基础研究发展计划 (973 计划) (2013CB328903), 重庆市基础与前沿研究计划 (cstc2015jcyjA40049), 国家自然科学基金 (71102065), 国家科技支撑计划 (2015BAF05B03), 中央高校基础研究基金 (106112014CDJZR095502) 资助
Supported by National Key Basic Research Program of China (973 Program) (2013CB328903), Basic and advanced research projects in Chongqing (cstc2015jcyjA40049), National Natural Science Foundation of China (71102065), National Science and Technology Ministry (2015BAF05B03), and Fundamental Research Funds for the Central Universities (106112014CDJZR095502)
本文责任编委 周志华
Recommended by Associate Editor ZHOU Zhi-Hua
1. 信息物理社会可信服务计算教育部重点实验室 重庆 400044 中国 2. 重庆大学软件学院 重庆 400044 中国 3. 悉尼科技大学工程与信息技术学院量子计算与智能系统研究中心 悉尼 NSW 2007 澳大利亚 4. 广州博冠信息科技有限公司 广州 501665 中国
1. Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing 400044, China 2. School of Software Engineering, Chongqing University, Chongqing 400044, China 3. Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia 4. Guangzhou Boguan Telecommunication Technology Limited, Guangzhou 501665, China

Twitter、Facebook、Renren 等社交媒体网络发展迅速, 已成为人们交流、分享的热门平台[1−2]. 然而, 在线社交网络蓬勃发展的同时, 也面临着虚假信息和虚假用户泛滥的问题[3]. 虚假用户, 也叫做垃圾用户, 是指在网络中传播垃圾信息和错误信息来达到某种目的的一类用户, 如推送广告、网络钓鱼、传播恶意软件、损坏他人信誉等[3−4]. 这种行为严重影响着社交网络的安全性和稳定性. 据统计, Facebook 上有 8 300 万 (8.7 %) 的虚假用

户[1], Twitter 中最受欢迎的 10 个用户的追随者中有 27 % 以上的虚假用户[2]. 为了逃避社交网络中的防御系统, 虚假用户采取各种策略进行伪装, 如窃取正常用户的用户概貌, 模仿正常用户的各种行为等. 由于社交网络本身具有开放性, 社交网络中的用户不仅与现实世界中的亲人、朋友建立关系, 同时还与其他陌生用户建立关系、分享信息. 在 Twitter、新浪微博等有向社交网络中, 虚假用户更是能够轻易地与其他用户建立关系, 虚假用户之间还能通过互相关注提高双方在网络中的可信度[5]. 另外也有一些正常用户出于礼貌, 对他的所有追随者同样予以关注[6]. 社交网络的这些特征使得虚假用户的检测比其他网络更难.

尽管社交网络虚假用户检测存在诸多挑战, 国内外仍有不少研究者对此提出了有效的解决办法. 监督型的检测方法旨在从大量的已标记的正常用户和虚假用户集中提取能够有效区分两者的特征, 例如文本特征[7−9]、网络结构特征[10−12]、关系图特征[13] 等, 并利用机器学习分类方法对未标记的用户进行分类. 然而, 这种监督型检测方法需要大量已标记数据, 人工标注大量数据又是耗时耗力的, 因而难以实施. 无监督的检测方法则存在假阳性高, 并且鲁棒性差的问题[14]. 针对已有方法存在的问题, 本文提出使用主动学习方法检测虚假用户. 主动学习[15] 使用极少量的标签样本训练初始分类器, 迭代选择信息量最大的未标记样本加入训练集, 以此提高分类器的泛化能力. 然而, 文献 [16] 表明传统的主动学习采样方法容易选出样本中的离群点, 并且存在信息冗余问题.

为了解决离群点问题, 本文提出基于不确定性和代表性的采样算法选择未标记数据集中的样本. 样本的代表性, 是指样本与其他未标记样本或近邻样本的平均相似度, 通常使用样本密度进行衡量. 一般来说, 离群点样本与其他样本的距离较远, 因而样本密度较非离群点小, 根据这一特性, 可将样本密度与不确定性线性组合, 从而选出不确定性高的非离群点样本. 除此之外, 代表性强的样本能够有效提高训练集对整体样本空间的覆盖率, 更好地提升分类器的泛化能力. 传统主动学习方法的另一个问题是信息冗余. 主动学习的每一次迭代都将选择多个新的样本加入训练集, 由于评估样本价值的标准和方式相同, 容易造成新样本彼此之间的相似度过高, 产生信息冗余. 为了解决信息冗余问题, 本文提出基于多样性的采样算法, 通过评估多个样本点间的平均距离来保证新样本的整体多样性. 因此, 本文方法使用双层采样选择未标记数据集中的样本. 第一层采样结合样本不确定性和代表性, 选择出不确定性高且代表性高的样本, 剔除未标记样本集中的离群点; 第二层通过整体多样性考量来减少信息冗余, 选择出整体多样性最大的一组样本集合作为最终的选择样本.

本文的组织结构如下: 第 1 节介绍社交网络虚假用户检测的研究现状和常用的主动学习策略; 第 2 节详细描述本文提出的检测框架和算法; 第 3 节则对本文实验过程和实验结果进行阐述与分析; 最后, 在第 4 节总结本文的工作, 并对后续研究进行展望.

1　相关研究

社交网络虚假用户检测从本质上来说是一个二类分类问题[13], 本节分析了基于监督型、无监督和半监督学习的检测方法. 针对已有方法的不足, 本文提出使用主动学习方法来检测虚假用户. 因此, 本节还对主动学习中常用的选择策略进行了介绍.

1.1　社交网络虚假用户检测的研究现状

近年来, 社交网络中虚假用户的检测研究已经获得了国内外研究者的关注, 研究者们提出了多种针对不同的社交网络平台上的虚假用户检测研究, 包括 Twitter[17−18]、Facebook[3]、Youtube[19] 和 MySpace[20]. 已有的检测算法可以分为有监督、无监督和半监督 3 类.

基于监督学习的检测方法通过提取社交网络用户的各种特征来训练分类模型. Benevenuto 等[19] 基于用户特征和视频特征对 Youtube 的用户进行建模. 类似地, Lee 等[20] 通过向网络中加入诱捕节点引诱虚假用户主动关注, 分析出虚假用户不同于正常用户的行为特征, 并据此提出一种基于诱捕系统的检测框架识别 MySpace 和 Twitter 中的虚假用户. 除了行为特征和文本特征, 许多研究者也从社交网络图入手, 提出了不同的检测方法. Song 等[21] 监督用户发送消息的行为轨迹, 通过分析发送者和接收者之间的距离和连通性来识别虚假用户. Hu 等[11] 则提出一种结合文本特征和网络结构的框架用于检测 Twitter 网络中的虚假用户. 程晓涛等[13] 运用 Web 信息挖掘技术[22] 挖掘社交网络中的关系图特征, 对检测虚假用户提出新的方案. 尽管监督型的检测方法准确率高, 但其需要大量的样本标签, 人工标注大量样本又是耗时耗力的, 导致实施成本过高. 于是又有研究者提出了非监督的监测模型.

非监督的检测模型主要利用社交网络的拓扑关系识别网络中的异常点. Gao 等[3] 提出利用文本和 URLs 相似度对 Facebook 中的帖子进行聚类的方法, 其假设虚假用户和正常用户之间不存在相关关系. Tan 等[23] 对这种假设提出了异议,

[1] http://www.bbc.com/news/technology-19093078
[2] http://www.webpronews.com/

他们认为虚假用户需要通过和正常用户建立关系来提高自身的可信度, 为此 Tan 等提出了一种结合社交关系图和用户链接图的非监督检测方法. 该方法先通过社交关系图确定部分正常用户, 然后根据这些用户的链接图识别虚假用户. Zhao 等[24] 则从 Twitter 中的推文内容入手, 基于动态查询扩展的方法构造推文图, 并使用异常检测的方法识别虚假推文, 从而检测虚假用户. 相比监督型的检测方法, 非监督的方法尽管不需要人工标注数据, 但假阳性率高, 并且鲁棒性没有监督型算法好[14].

最近, Li 等[14] 结合监督型和非监督型的方法, 提出一种结合信任传播的半监督检测框架. 与该框架中利用 PageRank 传播样本标签的方法不同, 本文提出的双层采样主动学习方法通过评估样本的不确定性、代表性和多样性从未标记样本中选择少量样本, 并对选择出的样本进行人工标注.

1.2 主动学习策略概述

主动学习是为了解决现实中标签数据不足, 标注数据耗时耗力的问题而提出的. 与传统的被动学习不同, 主动学习能够控制分类器对输入样本的选择, 从未标记样本中选择出信息量最高的数据, 从而获得更好的学习效率与学习效果.

一个主动学习模型可以建模成 5 个部分[16]:

$$A = (C, Q, O, L, U) \tag{1}$$

其中 C 为一个或一组分类器, Q 为选择引擎, 用来查询未标记样本中的高信息量样本, L、U 分别表示已标记数据集和未标记数据集, O 为专家, 负责确定未标记样本的标签. 主动学习的过程如图 1 所示.

主动学习根据选择未标记样本方式的不同, 可以分为成员查询综合 (Membership query synthesis)、基于流 (Stream-based) 的主动学习和基于池 (Pool-based) 的主动学习[25]. 其中, 基于池的主动学习是当前应用最广泛的采样策略. 根据选择未标记样例的标准不同, 基于池的采样策略又可以分为以下几种: 基于不确定性的采样策略、基于版本空间缩减的采样策略、基于模型改变期望的采样策略以及基于误差缩减的采样策略.

基于不确定性的采样 (Uncertainty sampling) 使用概率型分类器直接估计未标记样本的后验概率值, 选择最难被分类器区分的样本. 信息熵[26] 是不确定性采样中最常用的度量样本不确定性的方法, 计算公式如式 (2):

$$H(x) = -\sum_{y^* \in Y} P(y^*|x) \log P(y^*|x) \tag{2}$$

其中, y^* 表示样本 x 的预测标签, $P(y^*|x)$ 表示样本 x 被预测为某一特定标签的概率.

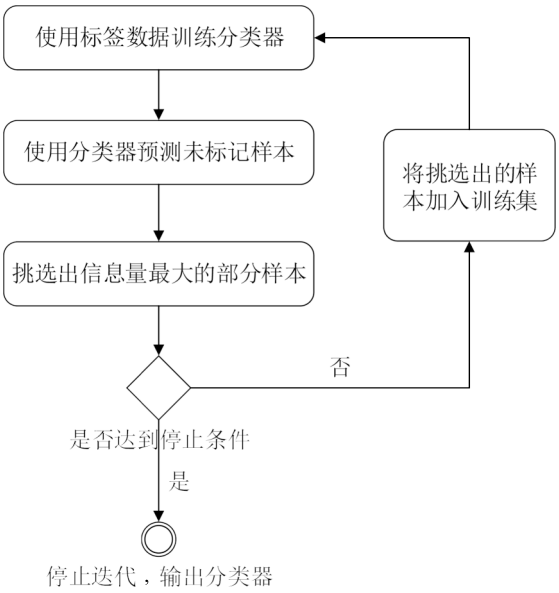


图 1 主动学习整体流程

Fig. 1 The whole flow of active learning

基于版本空间缩减的采样策略的主导思想是选择能最大程度缩减版本空间的未标记样本进行标记. 委员会投票选择算法 (Query-by-committee, QBC)[26] 是这类采样策略中应用最广泛的算法. 该算法主要分为 2 个步骤: 首先是构建多个学习模型组成委员会, 然后由委员会中的学习模型分别对未标记样本的预测标签进行投票, 选出投票最不一致的样本. Tuia 等[27] 提出的 EQB (Entropy query-by-bagging) 算法利用式 (3) 评估这种不一致性.

$$H_{bag}(x) = -\sum_{\omega=1}^{N} P(y^* = \omega|x) \log P(y^* = \omega|x) \tag{3}$$

其中, N 表示委员会预测样本 x 的标签个数, $P(y^* = \omega|x)$ 表示样本 x 被预测为某一特定标签的投票概率.

模型改变期望 (Expected model change)[28] 的主动学习框架基于决策理论, 通过评估未标记样本对当前模型的影响程度, 选择出对模型影响最大的样本进行标记. 代表性方法是 EGL (Expected gradient length), 该方法可应用于任意基于梯度下降方法的训练模型. 令 f_θ 为基于梯度的学习模型, L 表示原本的训练数据集, $\nabla f_\theta(L)$ 代表学习模型在参数 θ 时的梯度, 则 $\nabla f_\theta(L \cup (x, y))$ 代表加入新的标记样本后学习模型的梯度. 该算法使用式 (4) 对样本进行选择, $\|\cdot\|$ 表示梯度向量在欧氏空间中的长度.

$$x^*_{EGL} = \arg\max_x \sum_{y \in Y} P_\theta(y|x) \|\nabla f_\theta(L \cup (x, y))\| \tag{4}$$

基于误差缩减的采样策略 (Expected error reduction)[28] 通过减少分类器误差提高学习算法的泛化能力. 算法的思路是将每一个未标记样本标记后加入训练集, 并重新训练分类器, 计算分类器的误差变化结果, 最终选择能够最大程度缩减分类器误差的样本. 令 L 表示原本的训练数据集, $L^+ = L + (x, y)$ 表示加入样本 (x, y) 后的数据集. 由于样本 x 的标签是未知的, 所以我们需要对 y 的所有取值进行概率化. 式 (5)、(6) 分别描述使用 0/1 误差和 log 误差函数下的样本选择标准:

$$x^*_{0/1} = \arg\min_x \sum_{y\in Y} P_L \left(\sum_{u\in U} 1 - P_{L^+}\left(y|x^{(u)}\right) \right) \tag{5}$$

$$x^*_{log} = \arg\min_x \sum_{y\in Y} P_L \left(-\sum_{u\in U}\sum_{y\in Y} P_{L^+}\left(y|x^{(u)}\right) \times \log P_{L^+}\left(y|x^{(u)}\right) \right) \tag{6}$$

由于这种直接估计模型误差的方法复杂度高、计算量大, 因而部分研究者提出了一些替代的标准. Zhang 等[29] 引入 Fisher 信息函数来计算每一个样本的 Fisher 得分并构造 Fisher 矩阵, 判别模型中样本标记对模型误差的影响程度. Roy 等[30] 通过比较概率分布函数的变化来估计样本的未来期望误差, 并在朴素贝叶斯模型下大幅度提高了分类器的精度.

上述主动学习算法, 除了基于误差缩减的采样之外, 都存在离群点问题[16]. 为解决离群点问题, Settles 等提出了结合信息密度 (Information density, ID) 的主动学习算法[31], 如式 (7) 所示. 该算法根据样本间的相似度评估样本密度, 结合不确定性对样本价值进行评估, 如式 (7) 所示:

$$x^*_{ID} = \arg\max_x H(x) \times \left(\frac{1}{U}\sum_{u=1}^{U} sim\left(x, x^{(u)}\right) \right)^{\beta} \tag{7}$$

其中, $H(x)$ 表示样本 x 的信息量, 可根据不确定性采样和委员会投票选择算法求得. $sim\left(x, x^{(u)}\right)$ 表示样本 x 与其他未标记样本的平均相似度, 即样本密度. 参数 β 用来调节式中不确定性和密度的权重.

Zhu 等[32] 在此基础上提出一种基于密度的重排序算法 (Density-based re-ranking, DBRR), 该算法分为两层. 第一层根据样本的不确定性对未标记样本进行排序, 选择出部分不确定性高的候选样本; 第二层使用样本密度对候选样本排序, 选择密度大的样本作为新样本进行标注. Xu 等[33] 提出一种基于代表性的采样算法 (Representative sampling, RS). 该算法首先使用 K-means 聚类算法对未标记数据进行划分, 然后选择簇中心点作为新的训练样本进行标注.

以上三种方法都是围绕样本代表性进行采样, 目的在于解决离群点问题. 而本文的方法除了考虑离群点之外, 也针对信息冗余问题对主动学习采样方式进行了改进.

2　基于双层采样主动学习的虚假用户检测模型

本文提出的基于双层采样主动学习的社交网络虚假用户检测方法的整体框架结构如图 2 所示. 首先, 使用少量的标签用户作为训练集, 学习一个初始分类器; 接着用该分类器预测未标记数据集中的用户标签, 并使用双层采样算法对无标签用户进行选择, 选择出 "价值" 最大的多个用户; 然后, 借助人工标注的手段进行标签标注, 并将其加入训练集重新学习一个新的分类器; 重复上述过程, 直到分类器性能不再提升.

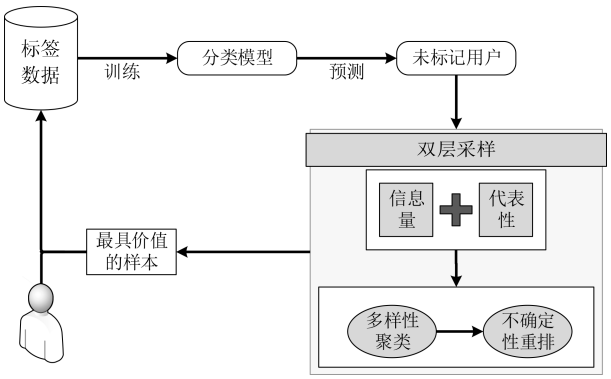


图 2　双层采样主动学习检测框架

Fig. 2　Detection framework based on active learning with two-layer sampling

双层采样主动学习的第一层采样将样本的不确定性和代表性作为样本价值的评估标准, 样本不确定性可采用任一简单的主动学习算法计算得到, 样本代表性采用 Zhu 等提出的样本密度[32] 进行评估. 第二层采样基于第一层采样的结果对候选样本进行聚类, 然后以簇为单位对样本的不确定性进行重排序, 从而选择出 "价值" 最大的未标记样本.

2.1　符号定义

本文使用的符号定义见表 1. 使用用户特征向量表示网络中的用户, 形如 $X = \{x_i\}_{i=1}^n$, 其中 x_i 表示用户 i 的特征向量, n 表示网络中的用户总数. 有标签的用户集合表示为 $L = \{(x_i, y_i)\}_{i=1}^{\ell}$, $U =$

$X-L$ 为无标签的用户集合, 其中 $y_i \in \{+1,-1\}$ 表示用户的标签 ($+1$ 为虚假用户, -1 为真实用户), ℓ 为有标签的用户的数目, $\mu = n-\ell$ 为无标签的用户数目, $\ell \ll n$. 主动学习中涉及到计算样本的不确定性, 表示为 $H(x)$, 样本的代表性表示为 $AS(x)$, 第一层采样选择出的候选用户集合表示为 $U_{candidates}$.

表 1 符号定义

Table 1 Definition of symbols

符号	含义
x	一个样本, 表示用户特征向量
y	用户标签
L, U	已标记样本集, 未标记样本集
ℓ, μ	已标记样本数目, 未标记样本数目
$H(x)$	样本 x 的信息熵
$AS(x)$	样本 x 的代表性
$U_{candidates}$	候选样本集合
k	每轮迭代选择的样本数, $k \ll \mu$
ΔL	每轮迭代选择的样本集合

2.2 用户特征

提取用户特征对于训练检测模型十分重要, 社交网络中的用户具有高维的特征, 但并不是每一个特征都能对虚假用户检测问题产生影响. 本文将社交网络用户的特征分成以下 5 类[14, 34]:

属性特征 (Profile-based feature, F_P): 属性特征是直接从用户概貌中提取的特征, 通常包括关注数、粉丝数、账户年龄、发表文章数目等.

内容特征 (Content-based feature, F_C): 内容特征即文本特征, 是从用户发表的文本中提取的, 包括单词数、URLs 数、垃圾词汇数、评论数、特殊字符数等.

行为特征 (Activity-based features, F_A): 行为特征衡量用户在社交网络中的活跃程度, 通常较为活跃的用户被检测成虚假用户的概率很小, 而活跃程度低或者行为周期化的用户有较大的概率被检测成虚假用户.

邻居特征 (Neighbor-based features, F_N): 从用户的邻居用户中提取的特征, 通常包括邻居用户的粉丝数、关注数、邻居用户发表的文章数目、评论数目等.

关系图特征 (Graph-based features, F_G): 从社交网络关系图中提取的特征, 包括聚类系数、节点核数、双向关注比、PageRank 值等.

本文参照其他研究者的研究成果[13, 35], 分别赋予 5 类用户特征不同的权重占比. 处理后的用户特征见式 (8).

$$x = [0.1F_P, 0.1F_C, 0.25F_A, 0.25F_N, 0.3F_G] \quad (8)$$

F_P, F_C, F_A, F_N, F_G 分别代表上述 5 类特征. 关系图特征 F_G 在检测虚假用户问题上具有最高的鲁棒性和准确性[13], 因而赋予最高的权重. 而属性特征 F_P 和内容特征 F_C 不能有效地区分正常用户和虚假用户[35], 所以权重最低.

用户特征向量经过上述处理后将用于训练分类器模型, 并通过双层采样算法对分类器进行优化, 最终用于虚假用户的检测.

2.3 基于不确定性和代表性的加权采样

针对传统主动学习可能选出离群点的问题, 本文考虑将样本的代表性加入样本价值的度量中, 并使用加权算法将其与样本的不确定性相结合, 形成一种基于不确定性和代表性的采样 (Sampling by uncertainty and representativeness, SUR), 形式化定义如式 (9):

$$SUR(x) = \alpha \times H(x) + (1-\alpha) \times AS(x) \quad (9)$$

其中 $H(x)$ 为样本 x 的信息熵, $AS(x)$ 为样本 x 的代表性, $SUR(x)$ 表示不确定性和代表性的加权值, α 为加权系数, $\alpha \in (0,1)$, $\alpha = 1$ 时算法变为仅考虑样本代表性的主动学习采样, $\alpha = 0$ 时即为基于不确定性的采样算法.

样本代表性反映该样本与样本集合中的其他样本的整体相似度, 本文采用 Zhu 等提出的样本密度[32] 衡量样本的代表性. 样本间的相似度采用标准化的皮尔森相关系数计算 (如式 (10)), 样本代表性形式化定义为式 (11).

$$sim(x_i, x_j) = 0.5 + 0.5 \times r_p(x_i, x_j) \quad (10)$$

$$AS(x) = \frac{\sum_{s_i \in S(x)} sim(x, s_i)}{K} \quad (11)$$

式中, $r_p(x_i, x_j)$ 表示两个向量的皮尔逊相关系数, $sim(x_i, x_j)$ 表示标准化到 $[0,1]$ 区间的用户相似度. $S(x) = \{s_1, s_2, \cdots, s_K\}$ 表示与用户 x 相似度最高的 K 个样本.

2.4 基于多样性和代表性的双层采样

SUR 算法可以有效地避免离群点, 但是仍然不能解决信息冗余问题. 为此, 本文提出一种基于多样性和代表性的双层采样算法 (Two-layer sampling based on diversity and density, DDTLS).

DDTLS 算法的第一层采样与上述 SUR 算法步骤一致, 根据 $SUR(x)$ 的大小对未标记样本进行排序, 从中选择 Top-N 的未标记样本作为第二层采样的候选样本, 如式 (12) 所示.

$$U_{candidates} \leftarrow \left\{ \arg\max_{x \in U} SUR(x) \right\}^N \quad (12)$$

得到 N 个候选样本之后, 算法的第二层采样使用多样性聚类和不确定性重排两个步骤保证样本集合的多样性:

步骤 1. 使用 K-means 方法对候选样本集合进行聚类, 得到 k 个不同的簇, $C=\{c_1,c_2,\cdots,c_k\}$. 根据聚类的性质, 可以认为聚成同一类的样本距离小, 不同类间的样本距离大[36]. 这样能够有效地保证被选样本集合整体多样性大.

步骤 2. 以簇为单位, 对每一个簇中的样本按照不确定性进行排序, 从每个簇中选择出不确定性最大的一个样本组成最终的样本集合, 如式 (13) 所示.

$$\Delta L=\bigcup_{c\in C}\left(\arg\max_{CL(u_i)=c}H(u_i)\right) \tag{13}$$

其中, $CL(u_i)=c$ 表示样本 u_i 聚类后属于簇 c, $H(u_i)$ 表示样本的信息熵或投票熵, 可由式 (2)、(3) 计算得到. ΔL 即为最终选择的样本集合. 由于第一层采样已剔除了离群点, 所以不确定性重排能够在保证样本集合多样性的基础上, 尽可能多地提高新样本的信息量.

基于多样性和代表性的双层采样主动学习方法的整体思路见算法 1. 双层采样算法可以避免选择出样本中的离群点, 因为算法在第一层采样中考虑了样本的代表性, 如果该样本是离群点, 代表性过小, 即使不确定性足够大也不会被选择为候选样本. 除此之外, 双层采样算法还能保证被选择出来的样本集合的整体多样性较大, 因为在第二层采样中, 候选样本被聚类之后, 不同类之间的样本距离较大, 则选出的样本整体的平均距离也较大, 从而样本间的相似度较低, 减少了信息冗余.

算法 1. 基于双层采样的主动学习检测算法

输入: 用户特征向量集合 U, 已标记数据集 L

输出: 分类模型 f

初始化:

　　$f\leftarrow clf.learn(L)$

循环:

　　$f.predict(U)$

　　$SUR(x)\leftarrow\alpha\times H(x)+(1-\alpha)\times AS(x)$

　　$U_{candidates}\leftarrow\{\arg\max_{x\in U}SUR(x)\}^N$

　　$C\leftarrow KMeans.clustering(U_{candidates})$

　　for c **in** C **do**

　　　　$\Delta L.append\left(\arg\max_{CL(u_i)=c}H(u_i)\right)$

　　$L\leftarrow L\cup\Delta L$

　　$f\leftarrow clf.learn(L)$

　　if $criterion$ **do**

　　　　return f

3　实验结果及分析

本文的实验分为对比实验和参数评估实验 2 个部分. 对比实验首先使用监督型机器学习算法与本文方法进行对比, 验证本文提出的方法是否能够使用少量标签样本达到和监督型算法近似的效果. 然后, 将本文提出的方法与其他主动学习方法进行比较, 验证 DDTLS 算法和 SUR 算法是否比其他基于代表性的主动学习算法性能更优. 并对 DDTLS 算法和 SUR 算法进行比较, 验证双层采样对分类器性能的影响是否大于单层采样. 参数评估实验则是对本文提出的算法进行参数敏感性分析, 根据参数对实验效果的影响, 给出每一个参数的最优取值.

3.1　实验设计

3.1.1　数据集

本文将使用 4 个数据集进行实验, 具体描述如下:

Apontador: 分为 2 个数据集, Apontador_33[37] 和 Apontador_49[38]. 其中 Apontador_33 有 2 762 条 tip 记录, 正常 tip 和虚假 tip 各有 1 381 个, 每条 tip 包含 33 个特征; Apontador_49 数据集有 7 076 条 tip 记录, 正常 tip 和虚假 tip 各有 3 538 个, 每条 tip 包含 61 个字段, 49 个特征, 包含全部 Apontador_33 的特征. Apontador 数据集的特征具体描述见表 2, 其中粗体表示 Apontador_49 包含而 Apontador_33 不包含的特征.

表 2　Apontador 数据集用户特征 (粗体表示 Apontador_49 包含而 Apontador_33 不包含的特征)

Table 2　The user features of Apontador data set (Bold show features only in Apontador_49.)

特征类型	具体描述
属性特征	粉丝数、关注数
行为特征	注册地点个数、发表 tip 数、照片数、评论地点的总距离、发表过 tip 的地点数、地点点击数、tip 数、评分、“Thumbs up” 和 “Thumbs down” 点击数
内容特征	**某用户所有 tip 中的 1-gram、2-gram 和 3-gram、某用户每一个 1-gram、2-gram、3-gram 在该用户的所有的** tip **中出现的比例**、文本中的垃圾词汇数量、大写字母的个数、数字字符的个数、出现电话号码的次数、出现邮箱地址的次数、URLs 的个数、出现联系信息的次数、单词数、所有字母都是大写的单词数、攻击性词汇数、是否出现攻击性词汇
邻居特征	该用户的粉丝的关注数、该用户的关注者的地点个数
图特征	聚类系数、双向关注比、节点相关性、**节点度/他的邻居节点平均度、节点出入度比、节点度、节点中心性、pagerank 值**

Twitter[17]: 包含 1 065 个用户, 每个用户都有一个表示用户类型 (真实用户或虚假用户) 的标签, 其中有 710 个真实用户, 355 个虚假用户. 每个用户包含 62 个字段, 26 个不同的特征. 用户特征的具体

描述见表 3.

表 3 Twitter 数据集用户特征

Table 3 The user features of Twitter data set

特征类型	具体描述
属性特征	昵称中是否存在垃圾词汇、关注数、粉丝数、账户年龄
行为特征	发表的推文数、被他人 @ 的次数、被他人回复的次数、回复他人的次数、发表推文的时间间隔、每天发表推文的数目、每周发表推文的数目、推文被回复的比例、每篇推文的转发数
内容特征	含有垃圾词汇的推文占总推文的比例、含有 URLs 的推文占总推文的比例、'#' 符号在每篇推文中所占的比重、URLs 在每篇推文中所占的比重、每篇推文的字符数、每篇推文包含 '#' 符号的数目、每篇推文中包含符号 '@' 的数目、每篇推文中包含数字的数目、每篇推文中包含 URLs 的数量、每篇推文的单词数
邻居特征	该用户的粉丝的关注数、该用户的关注者的推文数
图特征	双向关注比

Youtube[39]: 包含 829 个用户, 每个用户都有一个表示用户类型 (真实用户或虚假用户) 的标签, 其中有 641 个真实用户, 188 个虚假用户. 每个用户包含 62 个不同的特征. 用户特征的具体描述见表 4.

表 4 Youtube 数据集用户特征

Table 4 The user features of Youtube data set

特征类型	具体描述
属性特征	朋友个数、订阅者数、订阅数
行为特征	发表的请求数、接收到的请求数、观看的视频数、下载的视频数、喜爱的视频数、24 小时内最大视频下载量、下载视频的平均时长
内容特征	用户相关的视频 (下载、评分、收藏) 的视频的总观看量、平均观看量、总下载时间、平均下载时间、总观看时间、平均观看时间、总评分数、平均评分数、总评论数、平均评论数、总收藏数、平均收藏数
邻居特征	该用户的粉丝的关注数、该用户的关注者的推文数
图特征	聚类系数、节点相关性、节点出入度比、节点中心性、Pagerank 值

3.1.2 Baseline

本文首先使用监督型机器学习算法与本文方法进行比较, 说明本文方法仅使用少量的标签样本能够达到和监督型算法相近甚至更优的实验效果. 另外, 将本文方法与以下 3 种 Baseline 方法进行对比, 说明 SUR 算法和 DDTLS 算法的合理性和有效性. 并对 SUR 和 DDTLS 算法进行了比较, 说明双层采样算法对分类器性能的影响优于单层采样算法.

结合信息密度的主动学习算法[31]: Settles 等为解决离群点问题提出的主动学习采样方法, 使用样本平均相似度代表样本的密度, 具体公式见式 (7). 本文的实验中, 令 $\beta = 1$.

基于密度的重排序算法[32]: Zhu 等在样本密度的基础上提出的改进方法, 首先基于样本的不确定性确定候选样本集合, 然后选出样本密度大的样本作为新样本进行标注.

基于聚类的代表性采样算法[33]: Xu 等提出一种基于聚类的主动学习算法, 该算法首先使用 K-means 聚类算法对未标记数据集进行划分, 然后选择簇中心点作为新的训练样本进行标注.

本文提出的 SUR 算法和 DDTLS 算法都涉及样本不确定性的计算, 根据所选的计算方式的不同, 本文方法可以与不确定采样主动学习结合形成 SUR_UNC 和 DDTLS_UNC, 同样的, 结合委员会投票选择算法形成 SUR_QBC 和 DDTLS_QBC.

实验使用 4 种不同的机器学习算法, 分别是支持向量机、决策树、逻辑回归和朴素贝叶斯. 使用 5 折交叉验证, 主动学习模型做 50 次随机试验.

3.1.3 评价指标

实验使用准确率、召回率、F 值作为评估分类器性能标准, 计算公式分别如式 (14) ∼ (16). 其中, $truePositives$ 表示被正确预测为虚假用户的用户个数, $falsePositives$ 表示真实用户被错误预测为虚假用户的个数, $falseNegatives$ 表示虚假用户被错误预测为真实用户的个数.

$$precision = \frac{truePositives}{truePositives + falsePositives} \tag{14}$$

$$recall = \frac{truePositives}{truePositives + falseNegatives} \tag{15}$$

$$F = \frac{2\,(precision \times recall)}{precision + recall} \tag{16}$$

3.2 对比实验结果分析

3.2.1 与监督型机器学习算法的比较

本文先使用监督型机器学习算法与本文提出的 SUR、DDTLS 两种算法进行比较. Twitter 和 Youtube 数据集上的实验结果见表 5, Apontador 数据集上的实验结果见表 6. Twitter 和 Youtube 数据集上的实验最终所用的训练数据为初始未标记样本总数的 20 %, Apontador 数据集上的实验最终所用的训练数据为初始未标记样本总数的 10 %.

从表 5 和表 6 中可以看出, 无论是结合不确定性采样还是委员会投票选择, 本文提出的 SUR 算法和 DDTLS 算法的准确率都非常逼近监督型机器学习方法, 甚至更高. 例如, 表 5 中朴素贝叶斯模型下的 DDTLS_QBC 在 Twitter 数据集上能够达到 89.43 % 的准确率, 而监督学习只能达到 83.82 %.

表 5　Twitter 和 Youtube 数据集上的实验结果 (%)

Table 5　Experimental results on Twitter and Youtube data set (%)

分类模型	算法	Supervised		SUR_UNC		SUR_QBC		DDTLS_UNC		DDTLS_QBC	
		Twitter	Youtube	Twitter	Youtube	Twitter	Youtube	Twitter	Youtube	Twitter	Youtube
支持向量机	准确率	90.58	77.48	86.27	75.69	85.54	73.82	85.45	74.53	88.00	77.05
	召回率	70.42	62.23	66.08	**65.61**	**73.80**	**69.57**	**70.99**	**62.74**	72.56	70.40
	F 值	79.26	69.03	74.74	**69.94**	79.15	**71.63**	77.36	67.78	**79.49**	**73.58**
朴素贝叶斯	准确率	83.82	32.06	82.37	**41.15**	**94.96**	**54.59**	83.30	**47.35**	**89.43**	**39.60**
	召回率	72.96	93.62	64.51	86.16	72.11	80.30	67.15	**94.51**	66.25	89.33
	F 值	78.01	47.76	72.26	**55.66**	**81.97**	**63.42**	74.36	**63.09**	76.12	**53.37**
决策树	准确率	87.20	74.55	83.03	69.07	86.21	**78.52**	82.10	**80.04**	**87.30**	73.32
	召回率	70.99	65.43	67.61	**66.33**	70.70	62.66	68.08	62.64	69.39	61.12
	F 值	79.25	69.69	74.78	66.82	78.01	68.70	74.56	69.54	77.45	65.85
逻辑回归	准确率	88.81	75.50	87.81	**76.62**	87.52	**79.63**	86.77	**77.05**	85.85	**77.09**
	召回率	71.55	60.64	69.29	**71.26**	**74.36**	**67.04**	**73.01**	**67.06**	**72.30**	**66.49**
	F 值	79.24	62.23	77.46	**73.60**	78.01	**72.69**	79.10	**71.48**	78.58	**71.03**

表 6　Apontador 数据集上的实验结果 (%)

Table 6　Experimental results on Apontador data set (%)

分类模型	算法	Supervised		SUR_UNC		SUR_QBC		DDTLS_UNC		DDTLS_QBC	
		_33	_49	_33	_49	_33	_49	_33	_49	_33	_49
支持向量机	准确率	87.70	89.18	83.73	86.45	86.22	88.34	83.14	86.26	**89.50**	87.27
	召回率	75.38	79.88	74.22	70.12	70.45	72.80	**76.63**	**81.89**	**76.76**	**80.55**
	F 值	81.07	84.27	78.52	77.43	77.54	79.82	79.75	83.50	82.64	**84.46**
朴素贝叶斯	准确率	76.24	87.83	**77.47**	84.18	64.88	**92.96**	**84.51**	**97.14**	**80.65**	**91. 15**
	召回率	60.17	81.18	**70.16**	74.51	**72.27**	73.51	**64.73**	51.39	**66.18**	45.90
	F 值	67.26	84.37	**73.08**	79.16	64.11	84.12	**72.46**	67.04	**68.80**	60.65
决策树	准确率	82.49	87.20	**84.48**	**96.85**	**95.20**	**91.23**	**94.88**	**89.62**	**96.70**	**99.29**
	召回率	81.17	70.99	63.14	49.96	52.93	**80.46**	55.48	65.92	52.85	68.48
	F 值	81.82	79.25	67.59	66.92	68.00	**85.51**	69.44	74.25	68.17	**80.86**
逻辑回归	准确率	85.25	87.83	82.33	87.67	**86.59**	87.52	**86.53**	86.26	**85.53**	87.04
	召回率	75.31	81.18	**76.18**	79.93	**77.62**	74.36	74.29	**82.11**	**77.55**	**82.25**
	F 值	79.97	84.37	79.04	83.46	**81.79**	78.01	79.80	84.10	**81.31**	**84.56**

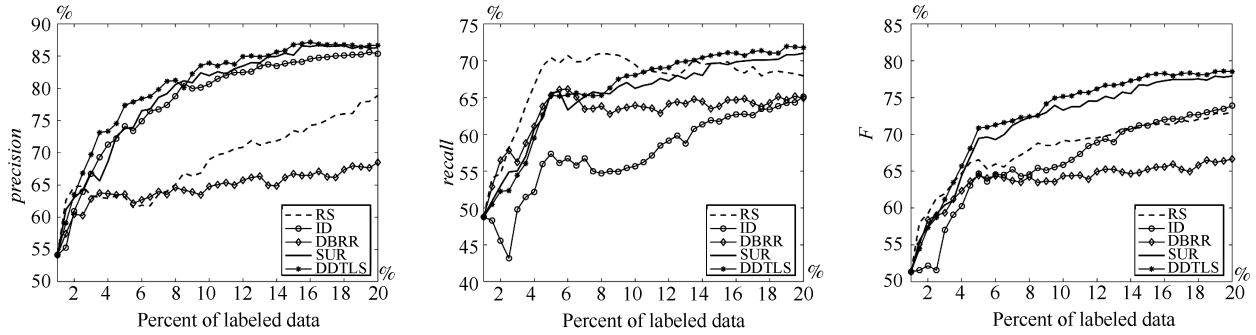
Youtube 数据集上使用逻辑回归模型作为基本分类算法的情况下, 本文方法的 3 项指标都优于监督学习.

3.2.2　与多种主动学习算法的比较

SUR 算法和 DDTLS 算法对比 Baseline 算法的实验结果显示在图 3～6 中. 实验使用支持向量机模型作为基础的分类算法, 初始标签样本数设定为样本总数的 1%, 每次迭代选择的新样本数为总数的 1%. Twitter 和 Youtube 数据集上的结果 (图 3、图 4) 显示训练数据集达到总样本的 15% 之后分类器的各项性能就趋于稳定, Apontador 数据集上的结果 (图 5、图 6) 显示训练数据集达到 6% 之后分类器的各项指标就趋于基本稳定状态, 只出现了小幅的波动. 这说明将主动学习算法应用到社交网络领域中能够很好地解决标签用户不足的问题.

对比图 3～6 中不同主动学习方法的准确率和召回率的增加速度和稳定值可以发现, 本文提出的 SUR 算法和 DDTLS 算法比其他主动学习方法对分类器性能的影响更快且更好. 如图 3 所示, SUR 和 DDTLS 算法的准确率 (图 3 (a)) 和 F 值 (图 3 (c)) 增加较其他算法更快, 并且最终稳定在一个较高的值. 图 3 (b) 中虽然 RS 算法的召回率上升速度快, 但随着训练样本的增加 RS 的召回率出现小幅下降, 最终的召回率反而比 SUR 和 DDTLS 小.

相较于单层的采样算法 SUR, 双层采样算法 DDTLS 在 3 项指标上都有所领先, 特别是在召回率指标上, DDTLS 的提升尤为明显. 这说明第二层的聚类算法能够有效地保证样本集合的多样性, 减少信息冗余, 从而提升分类器的泛化能力.



(a) 准确率随标签样本数变化

(a) The change of *precision* as the number of labeled data varies

(b) 召回率随标签样本数变化

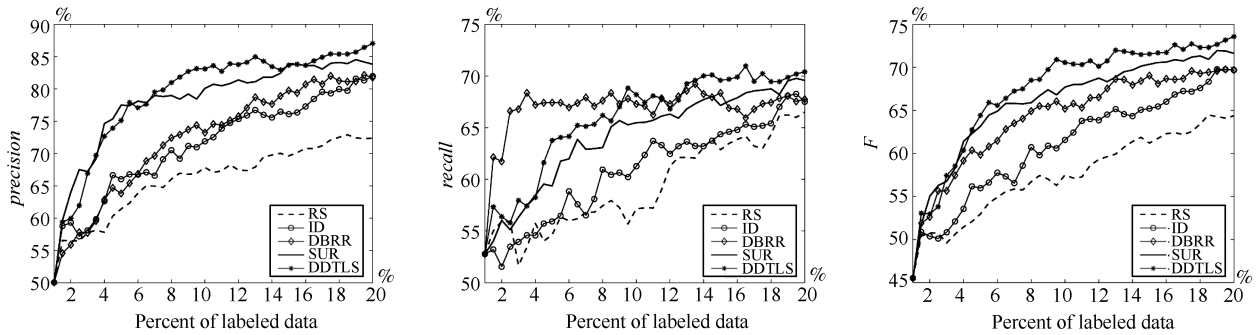
(b) The change of *recall* as the number of labeled data varies

(c) F 值随标签样本数变化

(c) The change of F value as the number of labeled data varies

图 3 Twitter 数据集的实验结果

Fig. 3 Experimental results on Twitter data set



(a) 准确率随标签样本数变化

(a) The change of *precision* as the number of labeled data varies

(b) 召回率随标签样本数变化

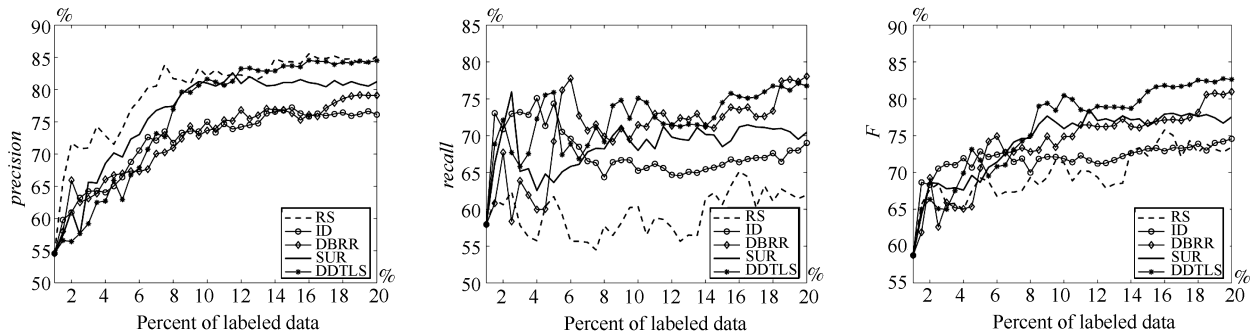
(b) The change of *recall* as the number of labeled data varies

(c) F 值随标签样本数变化

(c) The change of F value as the number of labeled data varies

图 4 Youtube 数据集的实验结果

Fig. 4 Experimental results on Youtube data set



(a) 准确率随标签样本数变化

(a) The change of *precision* as the number of labeled data varies

(b) 召回率随标签样本数变化

(b) The change of *recall* as the number of labeled data varies

(c) F 值随标签样本数变化

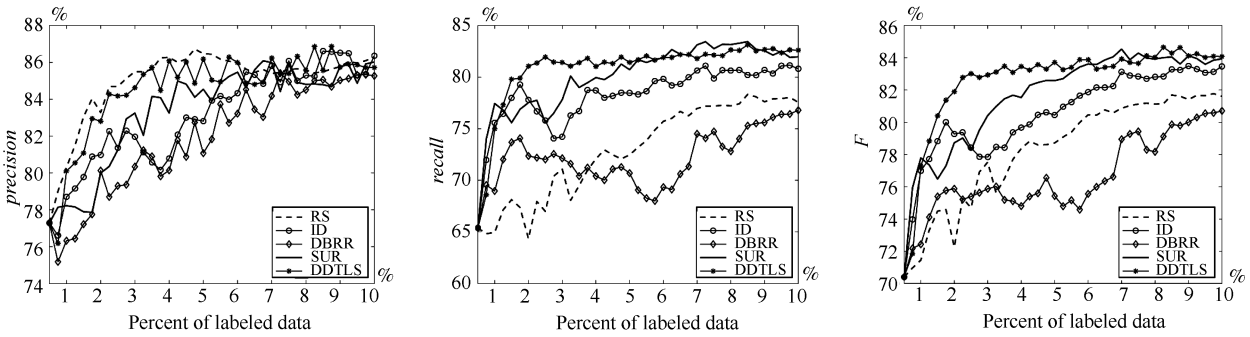
(c) The change of F value as the number of labeled data varies

图 5 Apontador_33 数据集的实验结果

Fig. 5 Experimental results on Apontador_33 data set

3.3 参数敏感性分析

这一部分分析双层采样模型的参数敏感性, 模型中待确定的参数有加权系数 α、近邻个数 K、候选样本数 N 以及初始标签样本数. 参数敏感性分析实验使用信息熵 (式 (2)) 计算样本的不确定性, 用逻辑回归模型作为基础的分类算法, 在 Twitter 数据集上对 DDTLS 算法进行单因素的敏感分析. 图 7 分别展示了上述 4 种参数对实验准确率、召回率和



(a) 准确率随标签样本数变化

(a) The change of *precision* as the number of labeled data varies

(b) 召回率随标签样本数变化

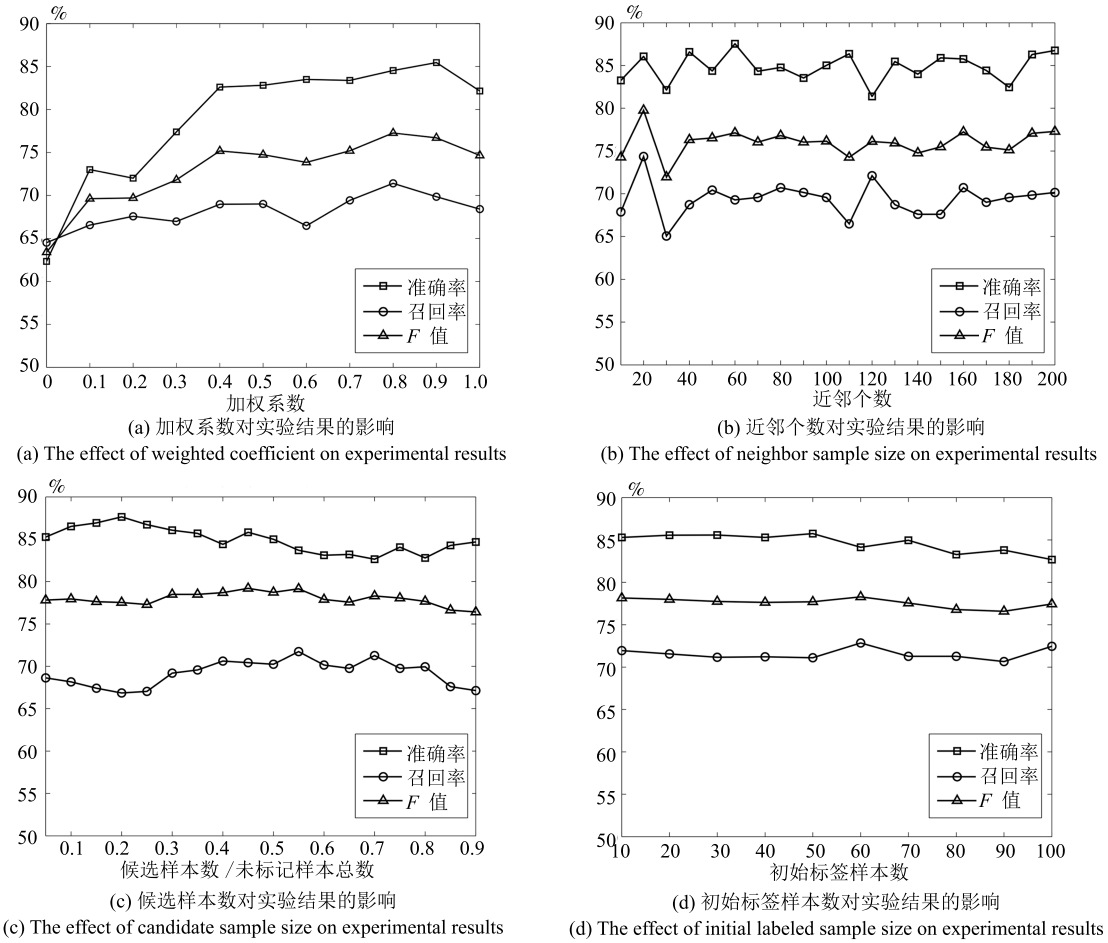
(b) The change of *recall* as the number of labeled data varies

(c) F 值随标签样本数变化

(c) The change of F value as the number of labeled data varies

图 6　Apontador_49 数据集的实验结果

Fig. 6　Experimental results on Apontador_49 data set



(a) 加权系数对实验结果的影响

(a) The effect of weighted coefficient on experimental results

(b) 近邻个数对实验结果的影响

(b) The effect of neighbor sample size on experimental results

(c) 候选样本数对实验结果的影响

(c) The effect of candidate sample size on experimental results

(d) 初始标签样本数对实验结果的影响

(d) The effect of initial labeled sample size on experimental results

图 7　参数敏感性实验结果

Fig. 7　Experimental results of parameter sensitivity analysis

F 值的影响.

加权系数 α 用于第一层采样 (基于不确定性和代表性的加权采样) 中计算样本的 SUR 值, 通过调节 α 值可以得到不同偏重的加权组合. 根据式 (9), 加权系数越大, 样本的不确定性对第一层采样的影响越大, 反之, 则第一层采样会优先选择代表性高的样本. 从图 7 (a) 中可以看出, $\alpha < 0.5$ 的情况下, 准确率、召回率都随着 α 值单调递增. 当 $\alpha > 0.5$, 准

确率和召回率趋向相对稳定状态, F 在 $\alpha = 0.8$ 时取得最优. 这说明尽管样本代表性可以有效地剔除离群点样本, 但是样本代表性对分类器性能的影响比样本不确定性低, 因此在 SUR 计算中应使不确定性的权重高于代表性. 此外, 当 $\alpha = 0$ 或 1.0 时, 该方法变成仅考虑代表性或不确定性的采样.

图 8 描述了当 α 取值分别为 0、0.8 和 1.0 时, 分类器的准确率和召回率随主动学习迭代过程的变化. 从图 8 (a) 中可以看出, $\alpha = 0$ 时分类器的准确率非常低, 这说明单纯基于代表性的采样不能有效地提高分类器的精度. 同样地, 图 8 (b) 显示 $\alpha = 0$ 时分类器的召回率也较低, 并且提升缓慢, 当标签样本接近 20 % 时召回率才与其他两种 α 取值时相近. 对比 $\alpha = 0.8$ 和 $\alpha = 1.0$ 的情况, $\alpha = 0.8$ 时准确率和召回率明显较好, 这一结果说明结合代表性的加权采样是有效的, 对分类器的精度和泛化能力都有提升作用.

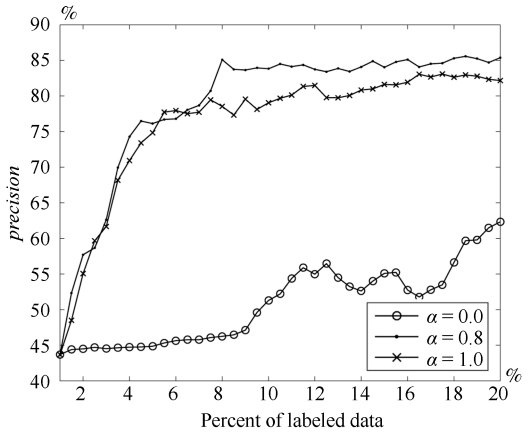
近邻个数 K 对双层采样算法的影响如图 7 (b) 所示. 近邻个数为 20 时, 双层采样算法的召回率和 F 值最大, 并且这个最大值远高于其他近邻个数下的结果. 这是因为, 过大或者过小的近邻个数值都不能准确地反映样本在未标记样本空间中的代表性, 近邻个数过小, 容易选出离群点, 近邻个数过大, 则所有样本的代表性值十分相近. 所以, 本文使用 20 个近邻的平均相似度计算样本密度.

候选样本数对实验结果的影响如图 7 (c) 所示, 候选样本数指的是第一层采样选择出的样本个数. 这个值同样不宜过大或过小, 如果候选样本数过大, 则第一层采样的作用将被减少, 无法筛选掉不确定性和代表性都不高的样本, 并且过大的候选样本量将会增加算法的时间复杂度. 相反, 如果候选样本的数量过小, 容易漏掉有价值的样本, 使得最终选择出的样本信息量小, 不能快速提升分类器的性能. 图 7 (c) 中的结果验证了这一结论. 值得注意的是, 算法的准确率在候选样本数为 200 时, 即未标记样本总数的 20 % 时, 取得最优值, 之后随着候选样本的增大, 准确率呈下降趋势. 而召回率则是在候选样本为 400, 也就是未标记样本总数的 40 % 开始取得较高值. 可能的原因是, 加权系数设定为 0.8, 导致第一层采样会先选择出不确定性高的一批样本, 不确定性高的样本全部选出来之后, 再根据样本的代表性选择另外的样本, 而这类代表性高的样本加入训练集后, 有利于提高分类器的召回率, 同时也会适当降低分类器的准确率. 因此, 本文的实验选择 $N = 300$, 即在第一层采样中剔除 2/3 的无价值样本, 保留 1/3 价值较高的样本.

从图 7 (d) 中, 明显可以看出初始样本数对实验结果没有影响. 为了减少人工标注的成本, 实验选择尽可能少的初始样本, 即初始样本数为 10 (样本总数的 1 %).

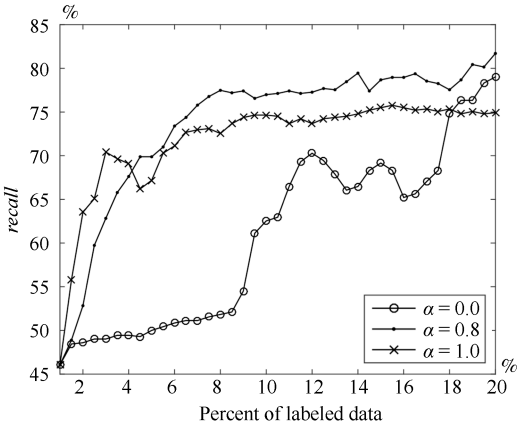
4 结论

本文针对社交网络中标签用户量少, 并且人工标注大量标签用户费时又费力的问题, 提出了一种基于双层采样主动学习的虚假用户检测框架. 该框架的第一层, 从未标记样本空间中挑选出不确定性大且代表性高的部分样本作为候选样本; 第二层则使用聚类重排序的方法选择出样本集合整体多样性高的样本进行人工标注, 标注后的样本用于新一轮的训练.



(a) 不同加权系数取值下准确率随标签样本数的变化

(a) In the case of different weighted coefficients, the change of *precision* as the number of labeled data varies



(b) 不同加权系数取值下召回率随标签样本数的变化

(b) In the case of different weighted coefficients, the change of *recall* as the number of labeled data varies

图 8 加权系数对双层采样算法效果的影响

Fig. 8 Influence of the weighted coefficient on experimental results

本文在 4 个数据集上的对比实验说明基于主动学习的虚假用户检测方法只需要少量的标签用户即可达到和监督型机器学习方法接近的, 甚至更好的检测效果. 并且, 相较于其他考虑样本代表性和多样性的主动学习方法, 本文提出的基于双层采样主动学习方法具有更优的检测精度, 所需要的标签用户数量更少.

然而, 本文方法仍然存在可以改进的地方. 1) 在实验中我们发现 SUR 和 DDTLS 的准确率和召回率可能会出现振荡. 出现这一现象的可能原因是, 基于近邻相似度计算的样本代表性有时不能有效地反映样本对未标记样本集合的整体覆盖率, 导致这一类代表性高的样本加入训练集后, 不能提供分类器更多的有效信息. 因此, 在以后的研究中, 可以就如何评估样本的整体代表性方面对算法进行改进, 从而进一步提高分类器的泛化能力. 2) 本文方法仍然需要人工对未标记样本进行标注, 针对这个问题, 后续的研究将会考虑主动学习和半监督的结合, 一方面可以利用半监督算法代替主动学习的人工标注步骤, 另一方面结合半监督可以充分利用未标记数据集中的最大量价值, 提高检测的准确率和召回率.

References

1 Huang Zhen-Hua, Zhang Jia-Wen, Tian Chun-Qi, Sun Sheng-Li, Xiang Yang. Survey on learning-to-rank based recommendatio algorithms. *Journal of Software*, 2016, **27**(3): 691−713
(黄震华, 张佳雯, 田春岐, 孙圣力, 向阳. 基于排序学习的推荐算法研究综述. 软件学报, 2016, **27**(3): 691−713)

2 Xin Yu, Yang Jing, Xie Zhi-Qiang. An overlapping semantic community structure detecting algorithm by label propagation. *Acta Automatica Sinica*, 2014, **40**(10): 2262−2275
(辛宇, 杨静, 谢志强. 基于标签传播的语义重叠社区发现算法. 自动化学报, 2014, **40**(10): 2262−2275)

3 Gao H Y, Hu J, Wilson C, Li Z C, Chen Y, Zhao B Y. Detecting and characterizing social spam campaigns. In: Proceedings of the 17th ACM Conference on Computer and Communications Security. Chicago, Illinois, USA: ACM, 2010. 681−683

4 Wang G, Wilson C, Zhao X H, Zhu Y B, Mohanlal M, Zheng H T, Zhao B Y. Serf and turf: crowdturfing for fun and profit. In: Proceedings of the 21st International Conference on World Wide Web. Lyon, France: ACM, 2012. 679−688

5 Stringhinl G, Wang G, Fgele M, Kruegel C, Vigna G, Zheng H T, Zhao B Y. Follow the green: growth and dynamics in Twitter follower markets. In: Proceedings of the 2013 Conference on Internet Measurement Conference. Barcelona, Spain: ACM, 2013. 163−176

6 Ghosh S, Viswanath B, Kooti F, Sharma N K, Korlam G, Benevenuto F, Ganguly N, Gummadi K P. Understanding and combating link farming in the Twitter social network. In: Proceedings of the 21st International Conference on World Wide Web. Lyon, France: ACM, 2012. 61−70

7 Gupta A, Kumaraguru P, Castillo C, Meier P. TweetCred: real-time credibility assessment of content on Twitter. In: Proceedings of the 6th International Conference on Social Informatics. Barcelona, Spain: Springer, 2014. 228−243

8 Amleshwaram A A, Reddy N, Yadav S, Gu G F, Yang C. CATS: characterizing automation of Twitter spammers. In: Proceedings of the 5th International Conference on Communication Systems and Networks (COMSNETS). Bangalore, India: IEEE, 2013. 1−10

9 Prasetyo P K, Lo D, Achananuparp P, Tian Y, Lim E P. Automatic classification of software related microblogs. In: Proceedings of the 28th IEEE International Conference on Software Maintenance (ICSM). Trento, Italy: IEEE, 2012. 596−599

10 Hu X, Tang J L, Liu H. Online social spammer detection. In: Proceedings of the 28th Conference on Artificial Intelligence. Québec, Canada: AAAI, 2014. 59−65

11 Hu X, Tang J L, Zhang Y C, Liu H. Social spammer detection in microblogging. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China: AAAI, 2013. 2633−2639

12 Ravikumar S, Talamadupula K, Balakrishnan R, Kambhampati S. RAProp: ranking tweets by exploiting the tweet/user/web ecosystem and inter-tweet agreement. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco, CA, USA: ACM, 2013. 2345−2350

13 Cheng Xiao-Tao, Liu Cai-Xia, Liu Shu-Xin. Graph-based features for identifying spammers in microblog networks. *Acta Automatica Sinica*, 2015, **41**(9): 1533−1541
(程晓涛, 刘彩霞, 刘树新. 基于关系图特征的微博水军发现方法. 自动化学报, 2015, **41**(9): 1533−1541)

14 Li Z X, Zhang X C, Shen H, Liang W X, He Z Y. A semi-supervised framework for social spammer detection. In: Proceedings of the 19th Pacific-Asia Conference in Knowledge Discovery and Data Mining (PAKDD). Ho Chi Minh City, Vietnam: Springer, 2015. 177−188

15 Cohn D A, Ghahramani Z, Jordan M I. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 1996, **4**: 129−145

16 Hajmohammadi M S, Ibrahim R, Selamat A, Fujita H. Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. *Information Sciences*, 2015, **317**: 67−77

17 Benevenuto F, Magno G, Rodrigues T, Almeida V. Detecting spammers on Twitter. In: Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference. Redmond, Washington, USA, 2010.

18 Jeong S, Noh G, Oh H, Kim C K. Follow spam detection based on cascaded social information. *Information Sciences*, 2016, **369**: 481−499

19 Benevenuto F, Rodrigues T, Almeida V, Almeida J, Goçalves M A. Detecting spammers and content promoters in online video social networks. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston, MA, USA: ACM, 2009. 620−627

20 Lee K, Caverlee J, Webb S. Uncovering social spammers: social honeypots + machine learning. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva, Switzerland: ACM, 2010. 435−442

21 Song J, Lee S, Kim J. Spam filtering in Twitter using sender-receiver relationship. In: Proceedings of the 14th International Symposium, Recent Advances in Intrusion Detection. Menlo Park, CA, USA: Springer, 2011. 301−317

22 Cao Jian-Ping, Wang Hui, Xia You-Qing, Qiao Feng-Cai, Zhang Xin. Bi-path evolution model for online topic model based on LDA. *Acta Automatica Sinica*, 2014, **40**(12): 2877−2886
(曹建平, 王晖, 夏友清, 乔凤才, 张鑫. 基于 LDA 的双通道在线主题演化模型. 自动化学报, 2014, **40**(12): 2877−2886)

23 Tan E H, Guo L, Chen S Q, Zhang X D, Zhao Y H. UNIK: unsupervised social network spam detection. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco, USA: ACM, 2013. 479−488

24 Zhao L, Chen F, Dai J, Hua T, Lu C T, Ramakrishnan N. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PLoS One*, 2014, **9**(10): e110206

25 Wu Jian, Sheng Sheng-Li, Zhao Peng-Peng, Cui Zhi-Ming. Minimal difference sampling for active learning image classification. *Journal on Communications*, 2014, **35**(1): 107−114
(吴健, 盛胜利, 赵朋朋, 崔志明. 最小差异采样的主动学习图像分类方法. 通信学报, 2014, **35**(1): 107−114)

26 Zhu J B, Ma M. Uncertainty-based active learning with instability estimation for text classification. *ACM Transactions on Speech and Language Processing*, 2012, **8**(4): Article No. 5

27 Tuia D, Ratle F, Pacifici F, Kanevski M F, Emery W J. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2009, **47**(7): 2218−2232

28 Settles B. Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, Wisconsin, 2010.

29 Zhang T, Oles F J. A probability analysis on the value of unlabeled data for classification problems. In: Proceedings of the 17th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann, 2000. 1191−1198

30 Roy N, McCallum A. Toward optimal active learning through sampling estimation of error reduction. In: Proceedings of the 18th International Conference on Machine Learning. Williams College, Williamstown, MA, USA: Morgan Kaufmann Publishers Inc., 2001. 441−448

31 Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii, USA: ACL, 2008. 1070−1079

32 Zhu J B, Wang H Z, Tsou B K, Ma M. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, **18**(6): 1323−1331

33 Xu Z, Yu K, Tresp V, Xu X W, Wang J Z. Representative sampling for text classification using support vector machines. In: Proceedings of the 25th European Conference on IR Research. Pisa, Italy: Springer, 2003. 393−407

34 Yang C, Harkreader R, Gu G F. Empirical evaluation and new design for fighting evolving Twitter spammers. *IEEE Transactions on Information Forensics and Security*, 2013, **8**(8): 1280−1293

35 Freitas C, Benevenuto F, Ghosh S, Veloso A. Reverse engineering socialbot infiltration strategies in Twitter. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris, France: ACM, 2015. 25−32

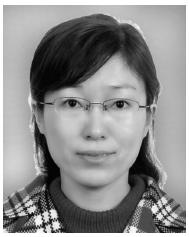
36 Chen Jin-Yin, He Hui-Hao. Research on density-based clustering algorithm for mixed data with determine cluster centers automatically. *Acta Automatica Sinica*, 2015, **41**(10): 1798−1813
(陈晋音, 何辉豪. 基于密度的聚类中心自动确定的混合属性数据聚类算法研究. 自动化学报, 2015, **41**(10): 1798−1813)

37 Costa H, Benevenuto F, Merschmann C D. Detecting tip spam in location-based social networks. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing. Coimbra, Portugal: ACM, 2013. 724−729

38 Costa H, Merschmann L H C, Barth F, Benevenuto F. Pollution, bad-mouthing, and local marketing: the underground of location-based social networks. *Information Sciences*, 2014, **279**: 123−137

39 Benevenuto F, Rodrigues T, Almeida J, Goncalves M, Almeida V. Detecting spammers and content promoters in online video social networks. In: Proceedings of the 28th IEEE International Conference on Computer Communications Workshops. Rio de Janeiro, Brazil: IEEE, 2009. 337−338

谭 侃 重庆大学软件学院硕士研究生. 主要研究方向为虚假用户检测, 数据挖掘. E-mail: 188313135@163.com
(**TAN Kan** Master student at the School of Software Engineering, Chongqing University. Her research interest covers spammer detection and data mining.)



高 旻 重庆大学软件学院副教授. 分别于 2005 年和 2010 年获得重庆大学计算机学院硕士和博士学位. 雷丁大学商学院访问学者. 主要研究方向为推荐系统, 服务计算, 数据挖掘. 本文通信作者.
E-mail: gaomin@cqu.edu.cn
(**GAO Min** Associate professor at the School of Software Engineering, Chongqing University. She received the master and Ph. D. degrees in computer science from Chongqing University in 2005 and 2010, respectively. She was a visiting researcher at the School of Business, University of Reading. Her research

interest covers recommendation system, service computing, and data mining. Corresponding author of this paper.)

李文涛 悉尼科技大学工程与信息技术学院量子计算与智能系统研究中心博士研究生. 主要研究方向为社交媒体挖掘与大图处理.
E-mail: wentao.li@student.uts.edu.au
(**LI Wen-Tao** Ph. D. candidate at the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney. His research interest covers social media mining and big graph processing.)

田仁丽 广州博冠信息科技有限公司运营数据分析师. 主要研究方向为个性化推荐与数据挖掘.
E-mail: tianrenli1@163.com
(**TIAN Ren-Li** Operating data analyst at Guangzhou Boguan Telecommunication Technology Limited. Her research interest covers personal recommendation and data mining.)

文俊浩 重庆大学软件学院教授. 主要研究方向为计算智能与推荐系统.
E-mail: jhwen@cqu.edu.cn
(**WEN Jun-Hao** Professor at the School of Software Engineering, Chongqing University. His research interest covers computing and recommendation systems.)

熊庆宇 重庆大学软件学院教授. 分别于 1986 年和 1991 年获得重庆大学学士和硕士学位. 2002 年获得日本九州大学博士学位. 主要研究方向为神经网络及其应用.
E-mail: xiong03@cqu.edu.cn
(**XIONG Qing-Yu** Professor at the School of Software Engineering, Chongqing University. He received his bachelor and master degrees from the School of Automation, Chongqing University in 1986 and 1991, respectively, and the Ph. D. degree from Kyushu University of Japan in 2002. His research interest covers neural networks and their application.)