doi:10.11959/j.issn.1000-436x.2017006

基于贝叶斯模型的微博网络水军识别算法研究

张艳梅[1]，黄莹莹[1]，甘世杰[1]，丁熠[2]，马志龙[3]

(1. 中央财经大学信息学院，北京 100081；
2. 电子科技大学网络与数据安全四川省重点实验室，四川 成都 610054；
3. 新疆财经大学计算机科学与工程学院，新疆 乌鲁木齐 830012)

摘 要：为了能够有效地识别水军，在以往相关研究基础上，设置粉丝关注比、平均发布微博数、互相关注数、综合质量评价、收藏数和阳光信用这 6 个特征属性来设计微博水军识别分类器，并基于贝叶斯模型和遗传智能优化算法实现了水军识别算法。利用新浪微博真实数据对算法性能进行了验证，实验结果表明，提出的贝叶斯水军识别算法能够在不牺牲非水军识别率的情况下，保证水军识别的准确率，而且提出的阈值优化算法能显著提升水军识别的准确率。

关键词：网络水军；水军识别；微博；贝叶斯模型；遗传算法

中图分类号：TP393 **文献标识码**：A

Weibo spammers' identification algorithm based on Bayesian model

ZHANG Yan-mei[1], HUANG Ying-ying[1], GAN Shi-jie[1], DING Yi[2], MA Zhi-long[3]

(1. Information School, Central University of Finance and Economics, Beijing 100081, China;
2. Network and Data Security Key Laboratory of Sichuan Province，University of Electronic Science and Technology of China, Chengdu 610054, China;
3. Computer Science and Engineering School, Xinjiang University of Finance and Economics, Urumqi 830012, China)

Abstract: In order to distinguish the spammers efficiently, a classifier based on the behavior characteristics was established. By analyzing the previous research, the ratio of followers, total number of blog posts, the number of friends, comprehensive quality evaluation and favorites according to latest data set, the Weibo spammers' identification algorithm was realized based on Bayesian model and genetic algorithm. The experiment result based on the real-time data of Sina Weibo verify that the Bayesian model recognition algorithm can ensure spammers recognition accuracy without sacrificing recognition rate of non-spammers, and the proposed threshold value matrix proposed optimization can significantly improve recognition accuracy navy.

Key words: network spammer, spammer identification, Weibo, Bayesian model, genetic algorithm

1 引言

随着移动互联网的广泛运用，微博成为民众获取信息、发表言论的一种重要途径，而近年来微博上涌现出大量水军，肆意发表带有倾向性的言论，扰乱民众视听，混淆信息的真实性。微博网络水军是指受雇于网络公关或营销公司，有组织、有计划地在微博平台上集中炒作某个事件，以达成宣传、推销或诋毁效果的大量专业或非专业技术团体。水军引导了微博平台上的舆论，使民众无法发出真实声音，从而形成言论“一边倒”的形式。网络水军影响了了网络秩序、公民利益、社会稳定及国家政治。因此，如何准确识别并着力阻击微博等社交平台上的网络水军成为近年来的研究热点。现有识别

收稿日期：2016-07-12；**修回日期**：2016-09-26

基金项目：国家自然科学基金资助项目（No.61602536, No.61273293, No.61309029）；北京市社会科学重点基金资助项目（No.16YJA001）；网络与数据安全四川省重点实验室开放课题基金资助项目（No.NDSMS201605）；中央财经大学学科建设基金资助项目

Foundation Items: The National Natural Science Foundation of China (No.61602536, No.61273293, No.61309029), Beijing Municipal Social Science Foundation (No.16YJA001), The Open Project of Network and Data Security Key Laboratory of Sichuan Province (No.NDSMS201605), The Discipline Construction Foundation of the Central University of Finance and Economics

微博水军的方法主要有 3 种：基于文本与情感分析法、概率图分析法和关系网络分析法。其中，基于文本与情感的分析需要花费较长的时间进行观察提取水军特征属性，且只能简单依据发布的文本信息对其进行判别，然而水军的快速发展使其发布信息的内容与普通用户趋同，使这种方法适用性能降低。而基于关系网络交互的方法依据水军用户的节点特征和边特征，虽然可以提高水军识别的准确率，但数据的获取方面存在很大的困难。为了避免上述 2 种识别方法的缺点，同时也为了能够构建一个高准确率且便于使用实时数据更新阈值矩阵的二元分类器，使其有较好的现实适应性，本文提出基于贝叶斯模型的水军识别算法，并设计了相应的遗传算法来获取最佳阈值矩阵，从而弥补了传统概率图无法更新适应性阈值矩阵的不足。

2　国内外研究现状

基于对微博网络水军的分析，当下研究可以分为基于内容特征分析、基于用户行为特征分析以及基于用户关系特征的分析方法（交互分析关系网络等）。早期研究中，对网络水军的检测主要以基于内容特征为主，其涉及机器学习中的自然语言处理分支，包括文本分析[1]、倾向性分析[2]和情感分析[3]等方面，使用的算法如文本分析、关键词分类法、B-Tree 索引等，主要依据评论内容相似度和评论文本的倾向性进行识别。由于水军的自我隐蔽意识越来越强，单纯进行文本内容的分析常常会漏掉大部分使用正常文本特征进行假象传播的水军，传统的单一基于内容特征的识别方法实用性较低。目前，基于行为特征的识别分析方法包括贝叶斯算法、决策树分类、*k*-means 聚类算法和逻辑回归算法等。虽然使用了部分水军特征作为属性，但由于现在的网络水军有善于伪装隐藏和变化多端的特性，已有的属性不再能精准地识别出水军用户，除此之外，无法使用一个明确的阈值来区分水军和非水军类别也是目前这些识别算法的一个不足。还有基于用户关系特征的识别探测方法，如神经网络分类法、贝叶斯网络等。这类方法虽然能结合用户属性行为节点和关系边综合考察，但建立网络本身需要大量具有复杂关系的数据进行训练，操作起来并不简便。

在社交网络平台中，根据用户的行为特征可以对用户进行准确的划分。Parameswaran 等[4]的研究发现网络水军的行为方式并不稳定，可以长期监控他们的行为，并将发现的网络水军移到“黑名单”中。Gargari 等[5]利用网络水军使用资源模式趋同的特征来提高网络水军识别准确率。莫倩等[6]认为，在社交网络中，用户通过交互行为逐渐形成以用户为中心的社交圈，而水军不具有正常的社会关系，其形成的关系网络结构特殊，具有极其不平衡的关注粉丝比。Krestel 等[7]利用网络图模型的传播发现含有链接的转发或评论的网络水军，通过给定一些种子节点的可疑度，依据种子节点向外传播可疑度的特点，从而计算发现所有可疑的水军节点。Gayo-Avello 等[8]依据图论，利用 Twitter 中水军会花费不正常的时间长度来关注目标用户或等待目标用户回粉的特点来挖掘水军，还提出了 Twitter 中水军关注的话题排序，他们还利用用户影响力特征来提高网络水军识别的准确率。韩忠明等[9]依据异常用户在关系图中有异常关系结构，建立用户的加权关系图模型来定位隐藏的电子商务网络水军团体。张良等[10]采用累计分布函数，对新浪微博用户行为和账号属性建立逻辑回归分类模型来检测水军。程晓涛等[12]根据水军关系图结构的稳定性，提出一种基于关系图特征的水军账号识别方法。陈侃等[13]根据水军信息交互模式具有一定的特点，从而建立决策树方法对其进行检测。对于分类器的构建，代表性文献的研究方法如表 1 所示。

从表 1 可见，不少研究从单一的特征方面进行水军的识别分析，在逐渐复杂化的水军行为中，算法研究所覆盖的特征不够全面，易于形成识别漏洞；而且往往容易错误收集一些特征不明显的正常用户，虽然可以具有良好的召回率和准确率，但依旧无法全面分析网络水军行为特征。

3　基于贝叶斯模型的微博网络水军识别算法

3.1　基于贝叶斯模型水军识别算法框架

目前的网络水军善于伪装和隐藏，变化多端，因此，为了能够识别具有这些特征的水军，采用粉丝关注比、平均发布微博数、互相关注数、综合质量评价、收藏数和阳光信用这 6 个相互独立的属性来设计水军识别分类器，并提出一种基于贝叶斯模型的微博网络水军识别算法。这里采用朴素贝叶斯理论完成水军识别模型的优点在于：首先，该水军识别分类器满足了朴素贝叶斯理论的一大假设前提，即计算先验概率时目标值之间具有相互独立

表 1 国内外相关研究情况

文献	属性	主要算法
文献[10]	URL 率和文本自相似度以及好友数、粉丝数、博文数等	逻辑回归算法
文献[11]	评论时间、评论的 ID、来自何客户端和粉丝数等	SVM 原理、simhash 算法
文献[12]	昵称、关注用户列表、微博文本、评论等	关系图结合、朴素贝叶斯、贝叶斯网络或决策树
文献[13]	设置了关注者–传播者、发布者–传播者、传播者–传播者 3 种类型来区分传播特征	决策树
文献[14]	网页特征码	文本分析、B-Tree 索引
文献[15]	综合指数、信息熵值	计算综合指数、熵值法
文献[16]	注册时间、昵称、活跃时间	*k*-means 聚类、深度优先搜索
文献[17]	发布时间、转发数、评论数、转发者 ID 等和用户 ID 等	支持向量机
文献[18]	用户活跃度、用户类别、粉丝值、好友值等	概率图
文献[19]	账户关注度、互粉比例、@比例等	决策树
文献[20]	好友请求率、URL 率、文本相似性等	honey-profiles
文献[21]	互粉关注比、收藏数等	trust-based 矩阵、PageRank 算法
文献[22]	消息、事件、评论等	词袋模型
文献[23]	互粉关注比、收藏数、每日增加好友数等	SVM 算法、重复增量修枝算法
文献[24]	总出度（如发出的消息）、总入度、总环数等	OCTracker 算法

性；其次，由于朴素贝叶斯理论可以依据先验概率同时计算出水军与非水军的分类，这是决策树和逻辑回归等方法所不具有的特点；最后，朴素贝叶斯模型具有较好的可扩展性，当特征属性集有所优化时，该模型依然有效。基于贝叶斯模型的水军识别过程如下：首先，运用遗传算法逐代优化得到判别水军和非水军的阈值矩阵，然后利用贝叶斯模型取得水军与非水军针对同一组特征属性在不同阈值内的概率矩阵，最后采取混淆矩阵针对水军与非水军样本数量不匹配的问题进行优化并得出准确率。具体细节如图 1 所示。

为了便于描述问题，采用表 2 所示的标识符号。

3.2 确定属性

通过研究已有文献的观点，结合主流研究与近几年水军特征的变化，本文选取了粉丝数、关注数、微博数、创建微博后的时间天数、好友数、综合质量评价、收藏数和阳光信用等用户行为特征，并依据一定的规律将其组合成 6 个能够区分水军与非水军的重要属性。首先，由于水军用户一般会带有特

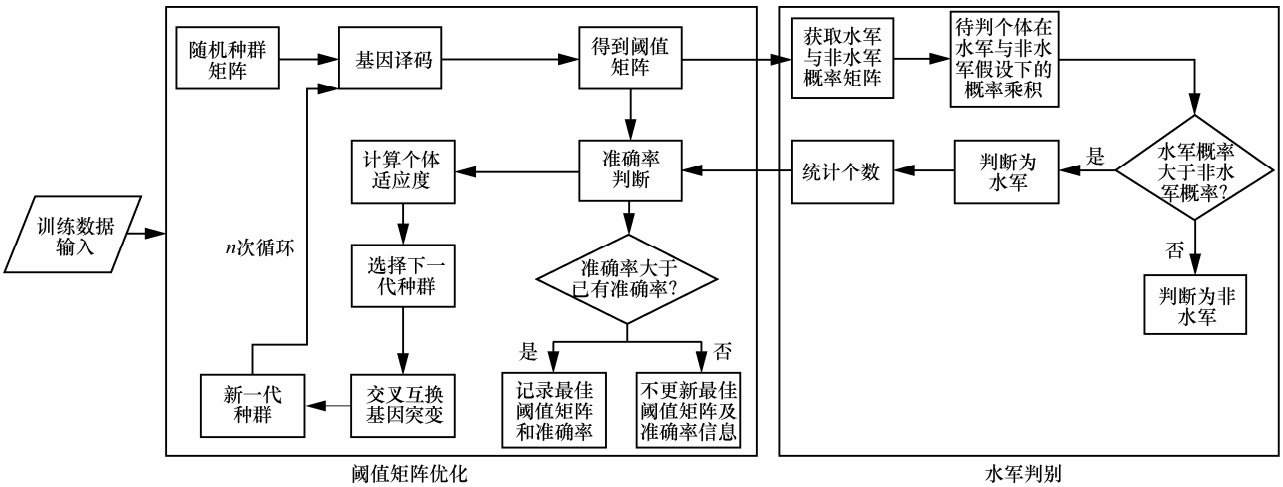


图 1 分类器框架

殊的目的关注其他用户，而正常用户只会有选择地去关注他们熟悉或感兴趣的用户。又因其发布的消息通常不具有价值和吸引力，导致微博水军在具有较高关注数的同时却具有较少的粉丝数。因此可以认为，如果一个用户具有反差较大的关注数和粉丝数（关注数居多），则该用户很有可能为水军。同理，由于水军用户几乎不会有好友（即互相关注数），水军用户的粉丝中互关数应比正常用户低。其次，正常用户在建立微博后，一天内发布微博的数量与其空闲时间成比例关系，即平均发布微博的数量不会过大；且一个正常用户通常只会维护登录一个账号。而水军用户一般为了达到某种宣传的目的，会在某一聚集的时间段发布大量的微博，并会在短期内申请一定数量的账号以完成其发布数量的任务。因此，如果某一用户在平均时间内发布的微博数量过高，则该用户是水军的可能性也会增大。水军用户所发布的微博中被收藏的数量也应该较正常用户的少。由于水军发布的微博通常没有用户留言评论，且一般缺乏完善的简介，也不会花费时间为水军的账号建立认证、登录微博刷新等级数等，从这些方面都可以很直观地将水军用户与正常用户分类。而阳光信用属性为 2015 年末新浪微博为了评价用户信用度，仅针对手机客户端新增的一个特征值，根据官方的定义，这一属性能够很好地评价用户的发言历史、活跃度、违规记录和社交关系等，可以认为水军用户的阳光信用属性应低于平均水平，对于区分水军具有关键性的作用，本文认为这是一个可以增强水军识别能力的重要属性，而且迄今尚未发现有文献用过这个属性。因此，结合 PC 端接口，再加上“阳光信用”这一重要属性，

表 2　　**标识符号**

标识	解释
FF	粉丝关注比
AW	平均发布微博数
IF	互相关注数
QE	综合质量评价
C	收藏数
I	矩阵行数，这里表示代表属性个数
J	矩阵列数
$\boldsymbol{M}$	非水军阈值矩阵
$\boldsymbol{T}$	非水军概率矩阵
$\boldsymbol{N}$	水军阈值矩阵
$\boldsymbol{S}$	水军概率矩阵
$x=\{a_1,a_2,\cdots,a_{m-1},a_m\}$	未分类的数据，每个 a 为 x 的每个属性的值
$B=\{y_1,y_2\}$	类别集，y_1 表示此条数据代表非水军，y_2 表示此条数据代表水军
var	阈值矩阵（通用）
population	种群矩阵，每行代表一个个体，每 4 列表示个体的一个属性的基因值
TP(true positive)	水军样本被预测为水军的个数
TN(true negative)	非水军样本被预测为非水军的个数
FN(false negative)	预测错误的实际水军类样本数目
FP(false positive)	预测错误的实际非水军类样本数目
acc^+	分类器对水军类样本的分类准确率
acc^-	分类器对非水军类样本的分类准确率
g	数据集整体的平均分类性能（即总体分类准确率）
SR(spammer recall)	水军召回率
LR(legitimate recall)	非水军召回率

由此得到 6 个彼此间相互独立属性构成水军识别模型。综上所述，将 6 个属性具体定义如下。

定义 1 粉丝关注比（*FF*）。每个用户的粉丝数与关注数的比值，计算方法如式(1)所示。

$$FF=\frac{Fans}{Followers} \tag{1}$$

定义 2 互相关注数（*IF*）。在用户的关注人里，与用户是互相关注关系的比例，计算方法如式(2)所示。

$$IF=\frac{Interact}{Followers} \tag{2}$$

定义 3 平均发布微博数（*AW*）。一个用户在其创建微博后的时间内一共发布多少条微博，计算方法如式(3)所示。

$$AW=\frac{Weibo\_numbers}{Days} \tag{3}$$

定义 4 收藏数（*C*）。即微博中被收藏的数量。

定义 5 综合质量评价（*QE*）。涉及多个属性变量，分别为是否有评论（*E*）、是否有简介（*R*）、是否进行认证（*I*）和等级数（*A*），给各个属性的权重分别为 0.2、0.3、0.3、0.2，计算方式如式(4)所示。

$$QE=0.2E+0.3R+0.3I+0.2A \tag{4}$$

在测试中，为使模型数据更加简便和使此项属性更有区分度，采用前 5 条微博中是否有评论来进行数据采集。

定义 6 阳光信用。此项属性是微博中近期加入的属性（即一种通过直观清晰的数据，结合用户的发言历史、活跃度、违规记录、商业记录、实名以及社交关系等行为，作为微博用户在网络上阳光讨论、积极表达、理性交流的衡量标尺），分为极低、较低、一般、较高、极高共 5 个等级，在实验中分别使用数值 1~5 来表示。

3.3 基于贝叶斯的水军判别模型

模型的具体工作过程如下。建立一个 i 行 j 列非水军阈值矩阵 $\boldsymbol{M}$(i 为属性个数，j 代表划分对应属性的阈值)，根据人工标注识别出非水军用户的准确集合，集合中非水军用户个数为 n_1。建立一个 i 行 j 列的全 0 非水军概率矩阵 $\boldsymbol{T}$，非水军概率矩阵中 T_{ij} 代表着非水军数据中第 i 个属性落在 M_{ij} $M_{i(j+1)}$ 之间的概率。概率统计通过计数统计法获得。同理，人工标注判断的水军用户的准确集合中，水军用户个数为 n_2，得到水军阈值矩阵 $\boldsymbol{N}$ 后，也通过计数得到水军概率矩阵 $\boldsymbol{S}$。水军概率矩阵中 S_{ij} 代表非水军数据中第 i 个属性落在 N_{ij} 和 $N_{i(j+1)}$ 之间的概率。

数据 $x=\{a_1,a_2,\cdots,a_{m-1},a_m\}$ 为一个未分类的数据，而每个 a 为 x 的每个属性的值，类别集 $B=\{y_1,y_2\}$，y_1 表示此条数据代表非水军，y_2 表示此条数据代表水军，$P(y_1\,|\,x)$、$P(y_2\,|\,x)$ 中较大的一个即为判定的结果。

根据贝叶斯定理，得

$$P(y_i\,|\,x)=\frac{P(x\,|\,y_i)P(y_i)}{P(x)} \tag{5}$$

由于各特征属性相互独立，分母为常数，且测试数据集中水军和非水军的数量明确，$P(y_i)$ 值确定，即当 $P(x_i\,|\,y)$ 取得最大值时，$P(y_i\,|\,x)$ 达到最大值，即

$$\begin{aligned}&P(x\,|\,y_i)P(y_i)=P(a_j\,|\,y_i)P(a_j\,|\,y_i)\cdots\\&P(a_j\,|\,y_i)P(y_i)=P(y_i)\prod\nolimits_{j=1}^{m}P(a_j\,|\,y_i)\end{aligned} \tag{6}$$

其中，$P(a_j\,|\,y_i)$ 为对应概率矩阵中第 j 个属性中对应 a_j 的概率。

通过计数，若判定为非水军，则不计数，若判定为水军，则在计数变量上加 1，最终得到判定为水军的数据个数。

3.4 阈值优化算法

目前，相关研究对于水军分类的阈值尚未发现权威的选取方法，也没有找到阈值的具体数值。然而，要构建一个准确性较高的分类器，进行阈值优化非常重要。因此，本文在对获取数据的属性特征和贝叶斯的数学原理的进行综合分析基础上，决定采用遗传算法对初始的阈值矩阵进行优化，以期优化后的阈值能够有效地提高分类模型的准确性。

在优化模型中，首先创建一个 i 行 4 列的阈值矩阵 ***var***，i 表示属性个数，属性取值范围被 4 个值分为 3 段，$var_{i,1}=0$，$var_{i,3}$、$var_{i,4}$ 分别是此属性的水军最大值和非水军最大值中的最小值、最大值。$var_{i,2}$ 是本文要优化的重点内容。

随机生成 10 行 16 列，值为 0~1 的矩阵表示的种群矩阵 ***population***（每行代表一个个体，每 4 列表示个体的一个属性的基因值）。对第 i 行进行阈值获取，将每个个体的第 i 个属性值通过二进制转换成十进制的数 s。

$$var_{i,2}=\frac{s+1}{16var_{i,3}} \tag{7}$$

将得到的阈值矩阵 ***var***，代入水军判别模型中进行判断。种群中 10 个个体将按照准确率高低在（0,1）区间内分得相应的长度，使用（0,1）之间的随机数矩阵 ***num***，***num*** 选中的个体可以进入到 ***population*** 中继续下一次的选择。因此，准确率较高的个体将优秀的基因遗传给下一代的可能性会提高。随后，将新得到的 ***population*** 矩阵进行随机的 2 行之间任意长度的交换，并进行随机的基因变异，得到最终的 ***population*** 矩阵进入下一次循环，记录产生的最佳阈值矩阵和最高准确率。

在经过 *n* 次循环后，阈值矩阵与准确率维持在一个相对稳定的水平，即为模型的输出。

3.5　算法伪代码

输入：

var：初始阈值矩阵

n_1：非水军用户的数量

n_2：水军用户的数量

population：种群随机矩阵

输出：

Best_Var: 最佳阈值矩阵

Spammer_Probability: 最佳水平概率矩阵

LegitimatUser_Probability: 最佳非水军概率矩阵

Spammer_Number: 识别的水军数量

1) Initialize ***Best_Var*** by random number, ***Spammer_Probability***, ***LegitimatUser_Probability*** by zero

//用随机数初始化最佳阈值矩阵；将水军与非水军的概率矩阵初始化为 0 矩阵

2) for generation, do

3)　for number, do

4) Get_new_matrix_of_threshold (***var***, ***population***) //获取阈值矩阵

5) Get_probability_of_threshold_for_spammers (n_1, ***var***)//获取水军概率矩阵

6) Get_probability_of_threshold_for_legitimateusers (n_2, ***var***)//获取非水军概率矩阵

7) Return ***Spammer_Probability*** and ***LegitimatUser_Probability***

8) Get_exact_number_of_spammers (n_1, ***var***, const1, const2)

9) Get_exact_number_of_legitimateusers (n_2, ***var***, const1,const2)

//获取数据中水军与非水军的数量

10) Get the Fitness

11) Compare with Best_Fitness and ***Best_Var***//通过比较获得最佳阈值矩阵和最佳适应度

12) end

13) Crossing_over(***population***)//基因重组（个体属性间译码交叉互换）

14) Variation(***population***)//基因突变（突变位译码取反）

15) Return ***Best_Var*** and ***Spammer_Number***

16) end

对算法的时间复杂度进行分析。算法主要分为 2 个步骤，设总的时间复杂度为 C，2 个步骤的时间复杂度为 C_1 和 C_2。首先，导入非水军(N_1 个)的属性(K 个)矩阵和水军(N_2 个)的属性(K 个)矩阵，并对种群(M 个个体)的基因(长度为 L)矩阵初始化为随机的 0~1 矩阵，其时间复杂度为 $C_1=O(N_1K)+O(N_2K)+O(ML)$。然后，通过导入的用户数据对阈值矩阵进行选择，外层循环是控制遗传代数(T)并进行基因的交叉互换和基因突变，内层循环首先要对种群的基因矩阵进行译码，然后使用用户数据训练概率矩阵，最后测试分类器的准确率并记录结果。两重循环的时间复杂度为 $C_2=O(TML)+O(TN_1K)+O(TN_2K)$。总之，算法的时间复杂度为 $C=O(TML)+O(TN_1K)+O(TN_2K)$，主要与遗传代数 T、种群规模 M、基因长度 L、用户数 N_1+N_2 和用户属性数 K 有关。

4　实验

实验运行环境为：Windows 7 操作系统，1.90 GHz 4 核处理器，4 GB 内存，算法的性能实验利用 Matlab 软件实现。

4.1　数据集的获取

为了更加准确地验证本文所提算法的有效性，准备了 3 套数据集。第 1 套数据集通过购买方式获得，在该数据集可以确切知道每个用户是否为水军；第 2 套数据集通过 R 语言编写爬虫从新浪微博 API 获取，数据量比较大；第 3 套数据集的属性信息比较完整。具体情况如下。

1) 已知分类的数据集

通过购买水军的方式，手动获取了 600 个水军数据及其相关属性；再通过亲朋好友等可信关系，

获取了 400 个非水军用户及相关属性作为实验的第 1 套数据集。该数据集可以提取到粉丝关注比、平均发布微博数、综合质量评价、收藏数和阳光信用这 5 个特征属性，但由于微博权限设置，这套数据无法计算“互相关注数”这一属性。

2) 人工标记的数据集

采用 R 语言中的 Rweibo 包，在不同时间段的数据池内随机获取 72 809 条数据。邀请了水军识别研究领域的 15 名学者对这 7 万余条数据进行人工识别，并对每一条数据的标识结果进行汇总，以认同度最高的分类结果对其进行标注。人工标记的结果中有 1 720 条水军和 71 089 条非水军数据，作为实验的第 2 套数据集。但由于 Rweibo 包尚未与微博同步更新，因此这套数据没有获取“阳光信用”这一属性。

3) 完整属性的数据集

为了增强实验结果的说服力，又进一步获取同时包含 6 个属性的数据集。但由于“阳光信用”一值是仅可见于手机客户端的特征值，即 PC 端口无法找到对应接口（新浪微博仅在手机客户端开放了这一特征值）；因此，本文根据第 2 套数据集中随机获取的 2 000 条数据的用户名，对应在手机客户端搜索用户，并记录“阳光信用”属性。其中，人工标记出水军 238 条，非水军 1 762 条。

4.2 评价指标的选取

在水军判别模型中，水军与非水军的比例常数由测试数据取得，水军与非水军在用户分布中的比例差别较大，即用于训练分类器的训练集为 2 组不平衡的数据。为了解决这个问题，提高分类器的准确性，本实验根据相关文献的研究结果[25]，建立混淆矩阵，如表 3 所示。其中，*TP*(true positive)为水军样本被预测为水军的个数，*TN*(true negative)为非水军样本被预测为非水军的个数，*FN*(false negative)、*FP*(false positive)分别是预测错误的实际水军类和非水军类样本数目。

表 3　　混淆矩阵定义

类别	实际水军类	实际非水军类
预测水军类	*TP*	*FP*
预测非水军类	*FN*	*TN*

借鉴相关研究，网络水军识别结果评估主要针对其准确性能，包括准确率、召回率、调和平均值、FPR(false positive rate)、ROC 曲线(receive operating characteristic curve)等评价指标。其中，准确率和召回率是数据挖掘领域中用来评价实验准确性的指标，调和平均值是对实验准确率和召回率综合表现的评价指标。本文采用准确率、召回率和调和平均值 *F*1 来评估分类器的准确性，各指标的定义如下。

1) 准确率

准确率的定义如式(8)所示。

$$g=\sqrt{acc^{+}\cdot acc^{-}} \tag{8}$$

其中，$acc^{+}=\dfrac{TP}{TP+FP}$ 表示分类器对水军类样本的分类准确率；$acc^{-}=\dfrac{TN}{TN+FN}$ 表示分类器对非水军类样本的分类准确率；g 表示平均分类准确率。当 acc^{+}、acc^{-}两者取值均较高时，平均准确率 g 才会高。

2) 召回率

水军和非水军的召回率定义分别如式(9)和式(10)所示。

$$SR=\frac{TP}{TP+FN} \tag{9}$$

$$LR=\frac{FP}{FP+TN} \tag{10}$$

3) 调和平均值 *F*1

实验着重观察水军分类的调和平均值，其定义如式(11)所示。

$$F1=\frac{2gSR}{g+SR} \tag{11}$$

4.3 基于贝叶斯的水军识别算法性能测试

为了使实验结果有对比性和说服力，本文分别在人工标注数据集、已知分类数据集和完整属性数据集上进行了实验。其中，人工标注数据集和已知分类数据集的实验准确率、召回率随遗传代数的变化如图 3 和图 4 所示。调和平均值 *F*1 如图 5 和图 6 所示。

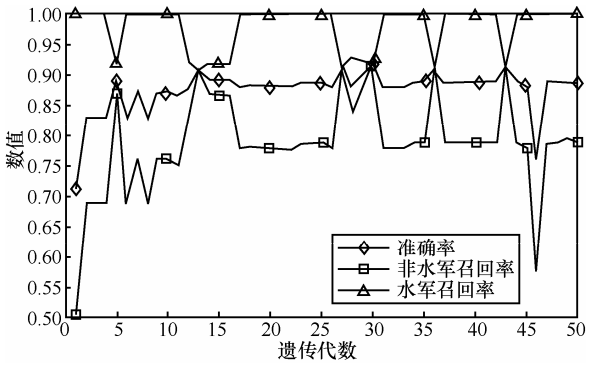


图 3　人工标记数据集的准确性测试结果

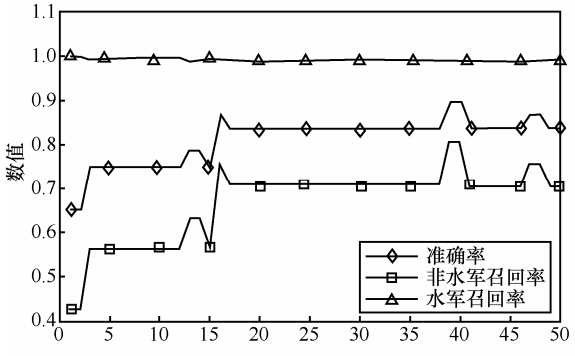


图 4　已知分类数据集准确性测试结果

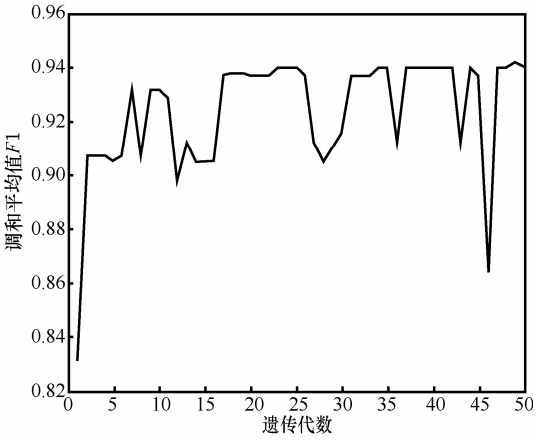


图 5　人工标记数据集的水军识别 $F1$ 值

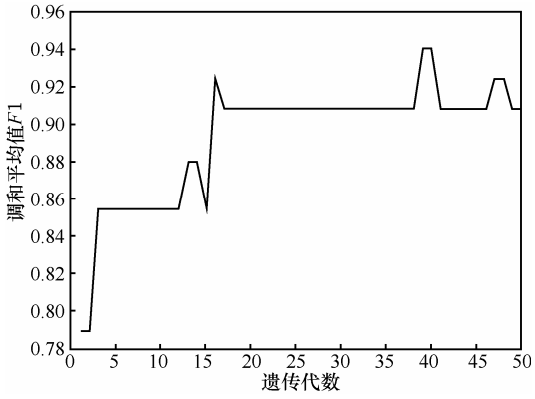


图 6　已知分类数据集的水军识别 $F1$ 值

图 3~图 6 的测试结果表明：2 个数据集上调和平均值 $F1$ 都在 90%以上，已知分类数据集上的召回率和调和平均值 $F1$ 略低于人工标记数据集，而准确率高于人工标记数据集。在此基础上对比完整属性数据集，其准确率召回率随遗传代数的变化如图 7 所示，调和平均值随遗传代数的变化如图 8 所示。

对比图 3~图 6，可以发现，完善阳光信用属性后，实验准确性能有一定幅度的提高，其中，准确率最高可达 97%，非水军召回率不低于 95%，调和平均值也高于 96%。

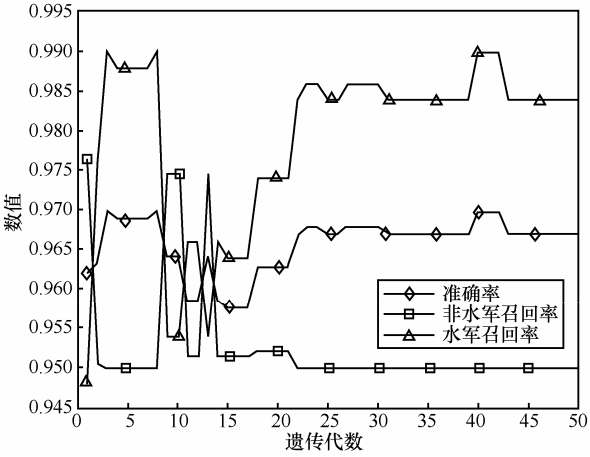


图 7　完整数据集的准确性测试结果

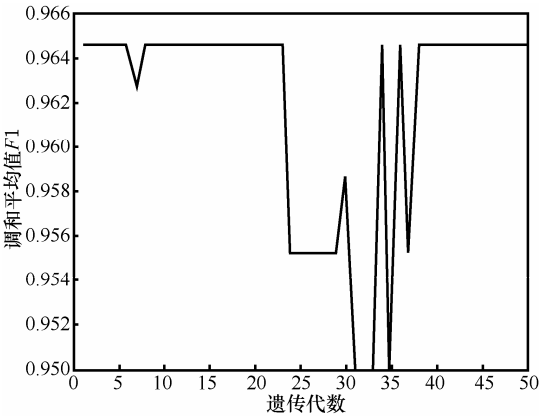


图 8　完整数据集的水军识别 $F1$ 值

为了更加有效地测试贝叶斯算法的性能，在 4.4 节的对比实验中取 3 个数据集上的测试结果的平均值。

4.4　对比实验与结果分析

由于本文所获取的数据是结构化的数据，而数据间不存在双向关系，所以没有选取基于网络分析的方法作为对照算法，而是选取逻辑回归[9]和决策树[13]算法作为基准算法，计算出准确率和召回率 2 个评价指标，结果如图 9 所示。本文所提的贝叶斯算法非水军识别率高于逻辑回归算法和决策树算法，水军识别率与逻辑回归算法相近，低于决策树算法。3 种算法原理不同导致了识别结果的差异。逻辑回归算法是在线性回归的基础上，套用了一个逻辑函数，然而针对属性较多的情况时，会存在线性关系不明显的现象，所以误差相对较大；决策树算法通过一系列条件来对用户数据进行判断，对于特征明显的分类识别率较高，因此，在准确划分阈值的前提下能准确分出特征明显的水军，但由于非水军特征不明显，该算法的非水军分辨能力较差，非水军召回率较低。本文所提的贝叶斯算法是基于

全局视角，通过综合衡量水军与非水军的概率大小比较，对一个用户所属类别进行判断，在此基础上，加入阈值矩阵优化算法，在提升了非水军识别率的情况下，保证了较高的水军识别率，从而保证算法准确性的均衡。

另外，本文所提阈值矩阵优化算法可以提高分类器的识别准确性。因此，本文将优化的阈值参数代入决策树方法，并与相关文献的传统决策树方法进行性能对比实验，结果如图 10 所示。决策树算法在采用本文的优化阈值参数后准确性能显著提升，说明本文设计的阈值优化模型能够有效地提高分类器的识别准确性。

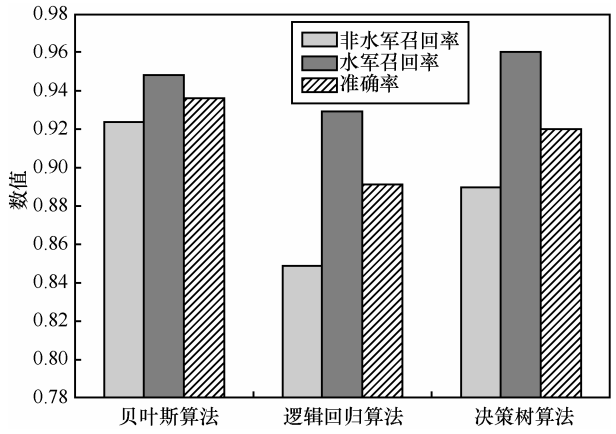


图 9 不同算法准确性对比

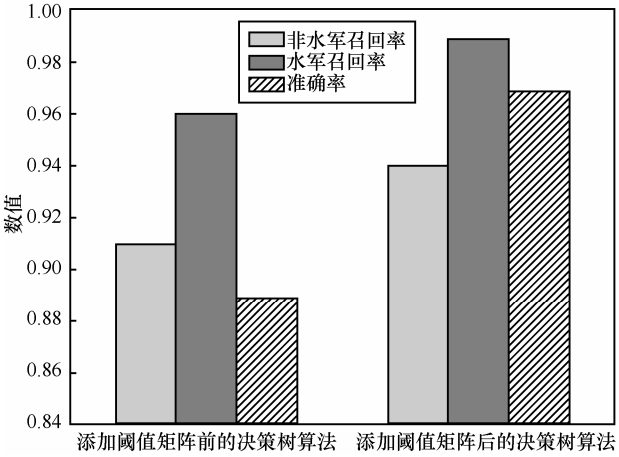


图 10 添加阈值优化算法前后决策树算法的准确性对比

5 结束语

网络水军的存在极大混淆了信息的真实性，给人们的工作和生活带来负面影响。而且由于现在的网络水军有善于伪装隐藏和变化多端的特性，相关研究往往通过牺牲非水军识别率来提升水军识别率。因而，迫切需要一个可以准确和高效地识别水军或虚假用户的方法。通过深入分析相关文献，并结合可靠经验，提取所有数据的可用特征作为分类参数，使用遗传算法确定各属性阈值，得到条件概率，最终建立识别微博水军的二元分类器。本文将贝叶斯模型和遗传算法结合，使用可靠性较强的阈值矩阵训练优化分类器，通过得到的概率矩阵较准确地识别水军用户。根据结果可以发现，分类器具有较好的学习能力，能够有效地识别微博中的水军，并且在一定数量的样本学习后，分类器可以在实际应用中对微博用户进行分类，且分类的准确性较高。

参考文献：

[1] SRIRAM B, FUHRY D, DEMIR E, et al. Short text classification in Twitter to improve information filtering[C]//33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2010). New York: ACM Press, 2010: 841-842.

[2] LIU B. Sentiment analysis and subjectivity[M]//Handbook of Natural Language Processing. Boca Raton: CRC Press, 2010: 627-666.

[3] ZHAO Y Y, QIN B, LIU T. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834-1848.

[4] PARAMESWARAN M, RUI H, SAYIN S. A game theoretic model and empirical analysis of spammer strategies[C]//7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. 2010: 1-7.

[5] GARGARI S M, OGUDUCU S G. A novel framework for spammer detection in social bookmarking systems[C]//IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM 2012). 2012: 827-834.

[6] 莫倩, 杨珂. 网络水军识别研究[J]. 软件学报, 2014, 25(7): 1505-1526.
MO Q,YANG K. Overview of Web spammer detection[J]. Journal of Software, 2014, 25(7): 1505-1526.

[7] KRESTEL R, CHEN L. Using co-occurrence of tags and resources to identify spammers[C]//Discovery Challenge Workshop at the European Conf on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2008). 2008: 38-46.

[8] GAYO-AVELLO D, BRENES D J. Overcoming spammers in Twitter—a tale of five algorithms[C]//Spanish Conf. on Information Retrieval (CERI 2010). 2010: 41-52.

[9] 韩忠明, 杨珂, 谭旭升. 利用加权用户关系图的谱分析探测大规模电子商务水军团体[J/OL]. http://www.cnki.net/kcms/detail/11.1826.TP.20160809.1459.006.html.
HAN Z M,YANG K,TAN X S. Analyzing spectrum features of weight user relation graph to identify large spammer groups in online shopping websites[J/OL]. http://www.cnki.net/kcms/detail/11.1826.TP.20160809.1459.006.html.

[10] 张良, 朱湘, 李爱平, 等. 一种基于逻辑回归算法的水军识别方法[J]. 信息安全与技术, 2015(4): 57-62.
ZHANG L, ZHU X, LI A P, et al. The Spammer detection based on logistic regression[J]. Information Security and Technology, 2015 (4): 57-62.

[11] 叶施仁, 孙宁. 基于 SVM 的新浪微博营销类水帖识别研究[J]. 湘潭大学自然科学学报, 2015, 37(4): 70-74.
YE S R, SUN N. Research on Sina microblogging marketing spam review detection based on support vector machine[J]. Natural Science Journal of Xiangtan University, 2015, 37(4): 70-74.

[12] 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[J]. 自动化学报, 2015, 41(9): 1533-1541.
CHENG X T, LIU C X, LIU S X. Graph-based features for identifying spammers in microblog networks[J].Acta Automatica Sinica, 2015, 41(9): 1533-1541.

[13] 陈侃, 陈亮, 朱培栋, 等. 基于交互行为的在线社会网络水军检测方法[J]. 通信学报, 2015, 36(7): 120-127.
CHEN K, CHEN L, ZHU P D, et al. Interaction based on method for spam detection in online social networks[J]. Journal on Communications, 2015, 36(7): 120-127.

[14] 杨长春, 徐小松, 叶施仁, 等. 基于文本相似度的微博网络水军发现算法[J]. 微电子学与计算机, 2014, 31(3): 82-85.
YANG C C, XU X S, YE S R, et al. A method to find water armies in weibo based on text similarity[J]. Microelectronics & Computer, 2014, 31(3): 82-85.

[15] 袁旭萍, 王仁武, 翟伯荫. 基于综合指数和熵值法的微博水军自动识别[J]. 情报杂志, 2014, 33(7): 176-179.
YUAN X P, WANG R W, ZHAI B Y. Automatic recognition of micro-blog water army based on multi-index comprehensive index method and entropy method[J]. Journal of Intelligence, 2014, 33(7): 176-179.

[16] 倪平, 张玉清, 闻观行, 等. 基于群体特征的社交僵尸网络检测方法[J]. 中国科学院大学学报, 2015, 31(5): 691-700.
NI P, ZHANG Y Q,WEN G X, et al. Detection of socialbot networks based on population characteristics[J]. Journal of University of Chinese Academy of Sciences, 2015, 31(5): 691-700.

[17] 董雨辰, 刘琰, 罗军勇, 等. 基于支持向量机的炒作微博识别方法[J]. 计算机工程, 2015, 41(3): 7-14.
DONG Y C,LIU Y,LUO J Y, et al. Hype microblog recognition method based on support vector machine[J].Computer Engineering, 2015, 41(3): 7-14.

[18] 韩忠明, 许峰敏, 段大高. 面向微博的概率图水军识别模型[J]. 计算机研究与发展, 2013, 50(S2): 180-186.
HAN Z M, XU F M, DUAN D G. Probabilistic graphical model for identifying water army in microblogging system[J]. Journal of Computer Research and Development, 2013, 50(S2): 180-186.

[19] 刘勘, 袁蕴英, 刘萍. 基于随机森林分类的微博机器用户识别研究[J]. 北京大学学报, 2015, 52(2): 290-300.
LIU K, YUAN Y Y, LIU P. A Weibo bot-users indentification model based on random forest[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2015, 52(2): 290-300.

[20] STRINGHINI G, KRUEGEL C, VIGNA G. Detecting spammers on social networks[C]//26th Annual Computer Security Applications Conf. (ACSAC 2010). 2010: 1-9.

[21] MURMANN A J. Enhancing spammer detection in online social networks with trust-based metrics[D]. San Jose: San Jose State University, 2009.

[22] SRIRAM B, FUHRY D, DEMIR E, et al. Short text classification in Twitter to improve information filtering[C]//33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2010). 2010: 841-842.

[23] MOH T S, MURMANN A J. Can you judge a man by his friends? Enhancing spammer detection on the Twitter microblogging platform using friends and followers[C]//Int'l Conf. on Information Systems and Technology Management (ICISTM 2010). 2010: 210-220.

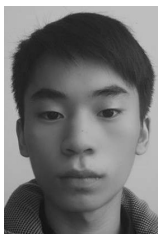
[24] BHAT S Y, ABULAISH M. Community-based features for identifying spammers in online social networks[C]//2013 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM 2013). 2013: 100-107.

[25] 潘正茂. 不平衡数据分类问题研究[D]. 西安: 西安建筑科技大学, 2012: 2-49.
PAN Z M. Research on classification for imbalanced dataset[D]. Xi'an: Xi'an University of Architecture and Technology, 2012: 2-49.

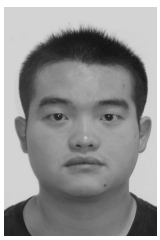
作者简介：

张艳梅（1976-），女，吉林省吉林市人，博士，中央财经大学副教授，主要研究方向为智能数据分析和服务计算。

黄莹莹（1995-），女，海南海口人，主要研究方向为智能数据分析。



甘世杰（1997-），男，四川邻水人，主要研究方向为信息安全、智能数据分析。



丁熠（1985-），男，四川宜宾人，博士，电子科技大学副教授，主要研究方向为医学图像处理、模式识别。

马志龙（1983-），男，新疆裕民人，新疆财经大学讲师，主要研究方向为信息安全。