# 微博水军识别结题报告

**摘要：**截至目前，作为国内最为流行、使用范围最为广泛、注册用户最多的社交媒体，微博每天都在为处理越来越多的水军付出代价。微博水军可以通过发布微博达到他们的恶意目的，比如散播谣言、传播恶意软件、托管控制僵尸账号或是进行其他非法地下活动，尤其是通过大量发布微博为微博事件制造舆论效应，造成了严重的影响。现有的大部分水军识别方法都是使用机器学习的方法进行水军识别。然而目前微博水军也逐渐演化到试图规避现有的识别策略的阶段，包括试图通过进行一些常规的微博行为来避免被侦测。所以在本文中，我们首先在互联网上抓取了大量微博数据，其中包含部分水军数据。在此基础上我们对现有微博水军使用的规避策略进行了综合分析，并利用现有的效果较好的方法进行了实验分析。相应的我们进一步提出了新的面向新浪微博的水军识别特征来进行水军检测。最后为了深入研究使用机器学习特征进行水军检测效果，我们针对特征集合的选取以及特征规避的难度进行了实验分析。通过我们的实验，我们设计的系统相比现有的分析方法在准确率与误分率上都有了显著地提升。

**关键词：**微博，水军，机器学习

## 数据采集

为了达到研究目的，我们需要在互联网上采集用户的微博个人信息及相应的微博数据来建立一个庞大的实验数据集，并在对其按微博事件进行分类，对其中的部分用户是否为水军用户进行标注。

为了获得用户的信息，我们编写了一个面向新浪微博的爬虫程序来爬取用户在微博中公开的用户信息以及相应的微博。以可能出现水军的特定微博事件为基础，我们在采集数据的过程中设置了关键词和发布时间作为过滤条件，爬取在指定时间段内发布的微博内容中含有相应关键词的微博作为某个微博事件的微博数据。在获取事件微博数据之后，我们以参与到事件当中的用户为种子用户，对用户截止事件起始前两个月内发布的微博进行采集，作为用户的个人数据集。通过这种方法，我们选取了10个可能含有水军的微博事件，总共采集了约85万微博用户的信息及其发布的微博。采集的数据如下图所示：

|  |  |
| --- | --- |
| 类别 | 数量 |
| 用户总数 | 853,041 |
| 微博总数 | 142,304,427 |
| 用户粉丝总数 | 130,334,187 |
| 用户关注总数 | 115,675,345 |

下一步，我们需要对数据集中的用户是否为水军进行定义和标注，在我们的研究中我们主要关注的水军类型为通过大量转发评论进行灌水的用户，这种类型的水军用户在微博中分布最为广泛影响也最大。传统的水军识别研究中，通常使用的标注方法是首先定义弱规则对用户数据进行过滤，比如用户微博中水军微博的比例等，在此基础上再进行抽样结合人工鉴定的方法进行数据的标注。这种方法的缺陷就是在大规模的数据集中使用这种取样方法会包含一定的偏差，而且这种偏差以及人工鉴别的差异性都是很难避免的。因此我们使用一种更为简单更为准确地标注方法。

我们获取标注的水军用户的途径主要通过在互联网上购买。我们通过在淘宝网等购物网站上购买水军账户，或是雇佣水军来发布灌水微博来收集标注的用户。具体实施过程中，首先我们在互联网上购买了约5万个左右的用来发布水军微博的微博账号，通过爬虫爬取各用户相应的微博信息。在另一方面我们在每个事件中随机选取10%的用户进行人工鉴定来进行真实用户的标注，相比于水军用户的人工标注，真实用户的人工标注难度相对较低，主观因素的影响也相对较小。标注过程结束之后我们便获得了总用户量约为4万的标注数据集，我们在接下来的研究部分将以此数据集或是其子集来建立训练集和测试集进行相关的实验与分析工作。

在完成对数据的标注之后，需要建立相应的训练集和测试集，对现有的多种水军识别算法进行测试。在接下来的第四部分，我们将首先对现有的水军规避进行分析，并使用采集的数据进行相应的实验。在此基础上，在第五部分我们将介绍我们根据之前的分析结合特定微博事件的场景提出的新的特征集合。在第六部分我们对新的特征集合进行实验与评估，并对特征集合中各类别特征的识别效果进行了实验分析，对各类特征进行评估并分析其对检测效果的影响。最后，在第七部分我们分析了现有研究的不足并对今后的研究工作进行了展望，在第八部分对本文进行了概括性的总结。

## 2. 规避策略的分析

在这一部分中，我们将对水军用户用来规避基于机器学习的水军识别方法的策略进行分析。在对这些策略进行深入了解之后，我们将通过实际案例与使用现有数据进行实验相结合的方法来对这些策略进行评估。

### 2.1 规避策略

在目前阶段，水军账号的持有者为了保证账号的长久使用通常都会采取一定的规避策略来躲避水军检测的识别。但是受到规避成本的限制，通常不会使用过于复杂的规避策略。现有的微博领域的规避策略主要有以下几类：1.通过修改账号的基本信息进行伪装。2.通过模拟真实用户的行为进行伪装。此外，还有部分通过对水军微博的内容进行修改来进行伪装的策略，由于出现频率较少本研究中便不做考虑。

#### 2.1.1 伪装账户信息的规避策略

通过用户的基本信息比如用户的粉丝数、是否认证、注册时间等等信息来判别用户是否为水军用户是最为常用的水军判别依据之一，因为这些信息通常直接体现了一个微博账户的声望或是活跃度，而且由于这类特征十分容易提取所以经常被用在基于机器学习的水军识别算法中。目前常用的基本信息特征规避策略包括增加粉丝数或关注数和修改用户昵称。

##### 2.1.1.1 增加粉丝数或关注数

通常情况下我们在浏览一个微博用户的主页时首先关注到的就是用户的粉丝数和关注数，粉丝数代表了一个账户的受欢迎程度或是受信任的程度，而关注数则代表了用户的兴趣范围同时还一定程度上体现用户的微博活跃程度。相应的，一个用户的的粉丝数或关注数越高，这个用户是水军用户的可能性就越小。由于考虑到水军账号的低微博活跃度，许多账号的粉丝数或关注数都是0，所以在之前的研究中比如：粉丝数、关注数、用户声望值等特征就被用来识别水军账号，或是被用来计算间接特征。 为了避免被这种策略侦测到，水军通常采取的做法有两种。第一种做法较为普遍，由于互联网水军除了提供向微博中灌水的服务之外通常还会提供为用户增加粉丝的服务，所以水军可以在自己控制的微博账户之间进行关注，这样可以快速增加账号的粉丝数和关注数；第二种方法就是在注册账号的时候关注一定数量的用户来增加用户的关注数，然后使用水军账号来关注需要增加粉丝数的账号，这种方法虽然比起第一种方法要复杂，但是由于在注册时关注的都是真实的正常用户，这样可以同时规避根据用户关注关系建立的检测规则。

##### 2.1.1.2 修改用户昵称

由于水军账号通常都是处于无人使用或是由及其操控的状态，所以判断一个账号是否由真人使用是一个常用的水军判别思路。而微博账号的昵称，通常是一个账号使用情况的直接体现，比如当我们使用手机新注册一个新浪微博账号的时候，我们的账号昵称通常会被默认设置为一段数字，需要我们进行进一步的设置来完成修改。如果是真人使用的话通常会使用自己喜欢的昵称，而如果是被注册来发布水军微博的账号则不会，并且这一区别在浏览用户的微博主页时可以很明显的被发现，所以这一特征在水军研究的初期常被用来作为识别水军账号的条件之一。而为了规避这一特征，水军使用者在注册微博账号时通常会给注册的账号加上随机生成的昵称，甚至连用户生日、博客地址等非必填资料中也填写随机生成的资料来避免被检测到。比如，下图展示了部分我们购买的部分水军中使用的随机生成的用户昵称：



图 4.1

#### 2.1.2 模仿真实用户行为的规避策略

另外一种常用的疑似水军披露方法就是对水军账号的行为规律特征。就像在引言部分的介绍一样，本文研究的微博水军账号通常是通过在微博中发布大量的灌水微博来获取利益，所需要费用的多少基本上取决于需要发布的灌水微博的数量，而灌水微博的数量越多取得的效果也越明显。从这个角度分析水军微博账号的行为通常都具有十分明显的特征，比如许多时候需要一个账号进行频繁的转发评论等等。因此许多的研究便针对这些特征提出了识别策略，包括计算发布微博内容的相似度、发布重复的微博的数量以及一些类似的特征。为了避免这一类的特征，水军通常采取的策略有避免同一账号多次转发和增加日常微博行为两种。

##### 2.1.2.1 避免同账号多次转发

由于水军被雇佣向微博中灌水通常都是为了制造舆论效应、扩大言论影响范围，所以需要进行水军账号发布的微博的数量相当巨大，甚至有时为了在短时间内发布大量微博制造热门话题需要在几小时内发布上百万条微博，所以为了节约时间成本通常会使用一个微博账号发布多条同样的微博。所以在许多的水军识别算法中都使用了诸如微博相似度、重复微博数量等检测特征。而且随着这种微博灌水的方式越来越被人们所熟知，类似的检测特征不仅是在关于社交媒体水军检测的有关研究中频繁被使用，更是已经被微博运营商作为基本的用户守则，普通用户是无法算时间内连续发布大量重复微博的。为了避免这种情况，水军采取应对策略是通过注册更多的水军账号来尽量减少每个微博账号发布的额微博数量。

##### 2.1.2.2 增加日常微博行为

根据微博用户的日常活动信息也是一种常见的水军检测依据。在早期对微博水军的研究过程中可以发现，常规的微博账号在没有灌水任务的时候通常处于一种闲置的状态，而真实用户的微博账号则会有一个相对稳定有规律的发帖行为。所以可以针对这种特点对微博账号的日常微博行为进行检测，来判断是否是水军账号，比如平均发帖间隔时间等特征。对于这种检测策略，水军们采取的对策是在进行灌水的间隙加入正常的微博活动，比如如果水军账号不需要进行灌水，则每隔一定时间段便进行转发或发布原创微博来模拟真实用户的行为。通过这种方法水军账号可以进一步向真实用户的行为靠拢。

### 2.2 规避策略的评估

在这一部分我们将对上一部分所介绍的水军用户规避检测的策略进行评估，我们将通过实例展示水军规避检测的实际案例并对水军实施规避策略的有关技术或网络服务进行介绍。在此基础上，我们选择了几种现有的水军检测方法进行了实现，并在我们的数据集中进行了实验。通过分析实验结果中没有被识别出的水军，来说明实际应用中确实已经有一部分水军成功的伪装成了真实账号避开了水军特征的检测。

**增加粉丝数或关注数：**在4.1.1.1节中我们介绍了水军通过增加水军账号粉丝关注数的伪装方法。实际上在互联网上已经有很多直接购买粉丝数或关注数的途径，其中一部分甚至本身就是水军服务提供者所提供的另一项服务。但是不同网站收取的费用有所差别，对最低购买数量的要求也各有不同。在下表中我们展示了不同网站对增加粉丝数或关注数的价格及最低购买数量：

|  |  |  |
| --- | --- | --- |
| **网站** | **每个粉丝（关注）的价格** | **最低购买数量** |
| [**艺眸设计**](https://yimou.taobao.com/) | ¥0.10 | 1000 |
| [**手机验证注册店**](https://sssaaaaa.taobao.com/) | ¥0.08 | 400 |
| [**xasdgasdgfasdf**](https://shop102673088.taobao.com/) | ¥0.07 | 500 |
| [**文芳阁新闻发布**](https://shop112550002.taobao.com/) | ¥0.05 | 400 |
| 淘我吧1996 | ¥0.03 | 100 |

表 4.1

上表所示展示了部分在线直接购买粉丝数或关注数的途径。从中我们可以看出增加粉丝数或关注数的成本是很低的；而且表中部分网站本身就是依托微博水军来提供增加粉丝数和关注数的服务，因此可以在进行伪装的同时进一步降低成本。

**修改用户昵称：**为了验证确实有水军通过伪装用户ID的方法来规避水军的检测，我们在我们数据中随机选取了10000名水军用户，对这10000名用户的用户昵称进行检测，结果如下表：

|  |  |
| --- | --- |
| **昵称类型** | **数量** |
| **全部由中文组成** | **3617** |
| **全部由英文组成** | **1322** |
| **全部由数字组成** | **144** |
| **多种文字符号混合** | **4917** |

表 4.2

从表中我们可以看出只有很小比例的用户昵称为只有数字的昵称，而占比例最多的则是多文字符号组合的昵称，这导致利用昵称来识别的水军账号的难度大大增加，也说明了水军用户开始注意到利用账号昵称来伪装成真实用户这一点。

**避免账号多次转发：**在4.1.2.1中我们介绍了水军账号为了模仿真实用户而在逐渐避免多次转发同一微博，为了证明这种方法已经广泛被应用在现存的水军团体中，我们随机选取了5000名水军用户，我们定义了每个用户的平均微博转发次数，假定用户的转发微博数为，用户的转发的原微博数为，则平均微博转发次数为：

我们对这5000名用户的平均微博转发数进行了统计，如下图所示：

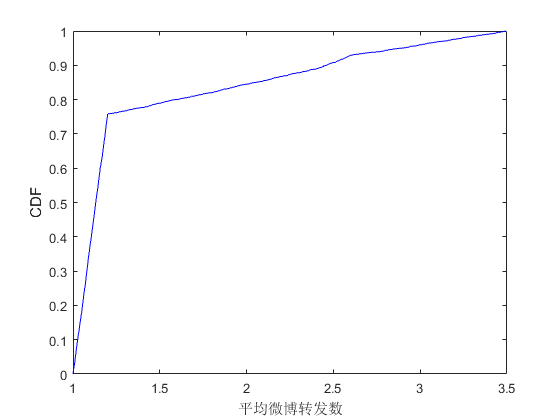


图 4.2

从图中我们可以看出水军的平均转发数大多数都在1左右，高平均微博转发次数的水军用户只占了很小的比例，说明现阶段大部分水军都已经采取了减小重复转发的策略。

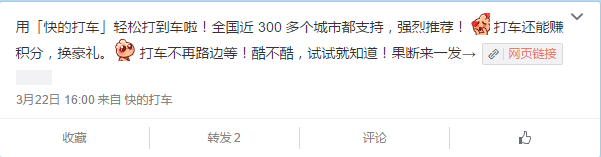
但是考虑到实际应用中常常需要水军在一到两天左右的时间内完成上百万的转发或评论，每个水军账号只转发或评论一次的所需要的账号数量是很难满足的。所以我们通过实际购买水军的方法来侦测迫使水军账号进行多次转发所要求的总转发量（时限设为24小时），对水军拥有的账号数量进行了大概的估测，以A,B,C,D代表其中四家不同的水军供应商，如下表所示：

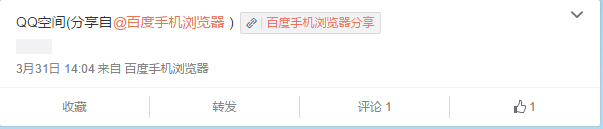
|  |  |
| --- | --- |
| 水军提供商 | 出现多次转发情况的总转发量需求 |
| A | 约为50,000 |
| B | 约为70,000 |
| C | 约为70,000 |
| D | 约为100,000 |

表 4.3

如果忽略对重复转发的时间间隔的定义，我们这可以粗略的认为水军提供商至少拥有50000左右的水军账号，现阶段新浪微博中常见的热门微博通常的转发量约为2000-8000，所以水军提供商控制的水军账号是非常可观的，足以起到推动舆论导向的效果。

**增加微博日常行为：**基于我们对水军的观察我们发现水军账号在闲置期间会间歇的发布一些微博来保持微博活跃度，下图中展示了一部分我们从获取的水军账号中截取的微博截图：





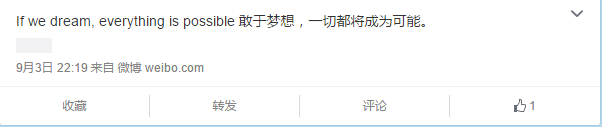


图 4.3

从图中我们可以看出水军账号在闲置期间，主要发布的是一些无用的原创微博，这些微博个性化程度较低，微博内容也并非毫无意义。这样可以做到最大程度的模仿真实用户的发帖习惯，所以对于实际应用中不能过分依赖微博内容检测，要结合其他特征的辅助来进行精确的检测。

## 3. 特征集合的选择

在这一部分我们将首先对特定微博事件进行定义，并阐述真毒特定微博事件进行水军用户检测的需求和意义。在此基础上，我们提出了一系列水军检测特征，其中包括我们针对特定微博事件新提出的特征，也包含已有研究中被广泛使用的检测特征。具体的特征包含属性特征、行为特征、内容特征三类。在接下来的部分我们将对这三类特征进行详细的描述与解释。

### 3.1 行为特征

行为特征，从一定意义上来说描述的是一个用户对微博的使用习惯，比如用户习惯在每天的特定时间登录、习惯每隔一个相对稳定的时间发布一条微博、习惯使用某种手机客户端手机或是某种浏览器等等。而对于水军来说，发布大量的灌水微博很明显属于一种十分反常的微博行为，不符合普遍的微博用户使用习惯，所以我们可以针对水军的一些行为特点提取特征进行检测。此外为了对特征的选取辅助说明，我们随机选取了某微博事件中标注的1000名真实用户与随机选取的1200名水军用户组成了辅助数据集，来完成对特征显著性的部分说明。

#### 3.1.1 发帖频率

根据我们对水军运营者的调研，由于水军账号在没有灌水任务的时候通常都是处于闲置状态，或是为了保持活跃度以每周一条或每两周的一条的频率发布微博来维持活跃度，这一行为特点与真实活跃用户存在着明显的差异。真实用户随着微博使用时间的增长，发布微博的频率会趋近于稳定在一个符合用户微博习惯的频率，尤其是现阶段越来越多的水军倾向增加转发账号减少单位账号转发数量，真实用户与水军用户的发帖频率的差距也越来越明显。

在实际计算过程中为了保证数据的时效性，我们统计用户最近2个月内发布微博的平均每小时的发帖数，所以用户的发帖平率可以表示为：

其中表示计算发帖频率的时间段（以小时为单位），表示在指定时间段内发布的微博数。为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：

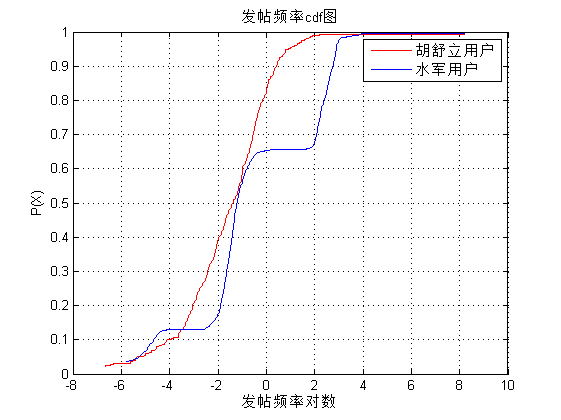


图 5.1

如图所示真实用户与水军用户的发帖频率有着明显的差异，不仅是真实用户的发帖频率要相对高于水军用户。而且从曲线中看出水军用户的曲线呈有规律的阶梯状，这表示水军用户的发帖频率集中在某个较小的区间；而真实用户则是平滑的曲线，进一步说明了两类用户在发帖频率上的区别。

#### 3.1.2 上网方式数与转发微博上网方式数

随着新浪微博等微博在生活中的日益普及，人们在浏览微博的方式上也有了更多的选择，无论是PC端的各式浏览器、微博客户端还是移动端的微博app，比如下图展示的常见的微博访问客户端：







图 5.2

在这里我们定义用户访问并发布微博的途径为用户上网方式。对于普通的用户来说，更换访问微博的方式是十分常见的，比如在外出行会使用手机客户端而如果在电脑跟前则会直接通过网页访问。相比之下，水军账号由于通常都是通过转发器或是微博api大规模发布微博，所以水军用户与真实用户使用的上网方式种类存在明显差异。

基于以上分析我们提取用户的上网方式数和转发微博上网方式数两种检测特征。之所以定义两种相似的检测特征，主要是为了应对以下两种情况：

第一种情况较为常见，由于部分水军在处于闲置期时可能会定期发布微博维持活跃度，由于发布的微博数量很少所以水军雇佣者可以不考虑切换上网方式的成本来使用不常用的上网方式发布原创微博进行伪装。为了应对这种规避方式，我们选择定义了转发微博上网方式数，由于在进行大量转发灌水的同时进行上网方式的技术难度与时间成本都是极大的，所以水军账号的转发上网方式通常会比较稳定。

第二种情况则是水军并不在日常维护的过程中切换上网方式，而是准备某几种上网方式或是几种伪装方法，在每次大规模发布灌水微博的过程中伪装不同的上网方式，比如在本周转发微博A的时候伪装为使用某浏览器作为上网方式，在下一周转发微博B时则使用手机app访问作为伪装。对于这一类型的水军，情况与上一类型正相反，在日常维护的过程中发布原创微博并不会切换上网方式。所以可以使用全部微博的上网方式数来进行检测。

为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：

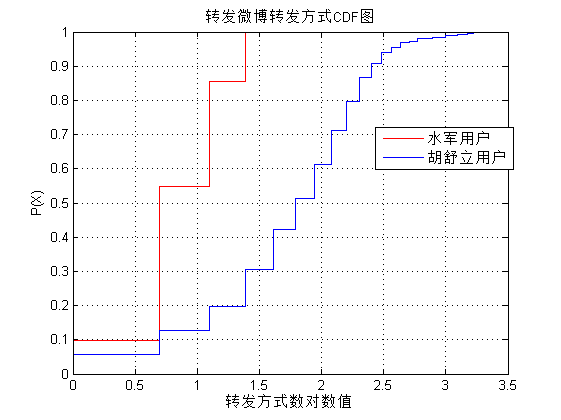
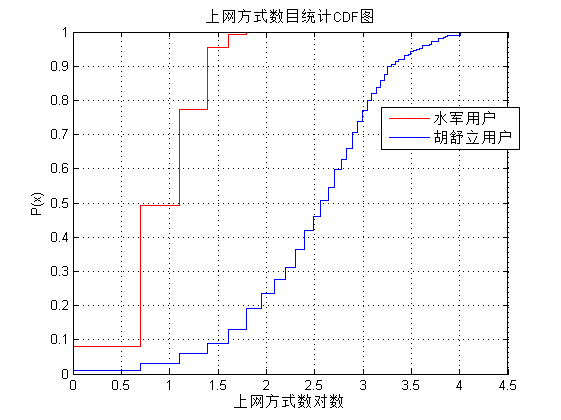


图 5.3(a) , 5.3 (b)

从图中可以明显看出水军用户与真实用户在常用上网方式上的区别，整体上真实用户使用的上网方式要丰富很多，而水军用户使用的相对单调。

#### 3.1.3 是否参与热门微博

对于通过频繁转发或评论来进行灌水地微博水军来说，需要为他们指定一条或几条目标微博，而他们需要做的就是对这些指定的微博进行转发及评论。根据我们的调研，现阶段互联网上提供的水军水军转发或评论服务都有所谓的最低消费标准，也就是购买的转发货评论数不能小于最低消费标准。所以综合我们对水军服务提供者的调研，我们定义转发数或评论数大于100的微博为热门微博。相应的，以100次转发或评论为分界点，对非热门微博进行转发或评论的用户是水军用户的可能性通常很小。

### 3.2 内容特征

内容特征是最为常用的水军识别特征，因为无论水军如何伪装，最终都要通过发布的微博内容来传达相应的信息，所以水军发布的内容进行研究是识别水军的必备特征。但是考虑到微博中水军有出现的情况日益复杂，比如某一条微博被水军大量转发，有可能是为了对其中的产品进行推销，也有可能是为了扩大博主的影响力，所以我们放弃了对微博内容的语意研究。我们选择对用户的微博内容进行整体的研究，选取了以下三个特征对内容进行描述，避免了对内容进行语义分析的高计算复杂度。

#### 3.2.1 原创微博比例与转发微博比例

经过我们研究发现，对于水军账号发布的微博，其中原创微博和转发微博的比例具有明显的特点。目前，新浪微博中的微博类型按信息来源可以分为原创微博、转发微博和评论微博三种。从水军微博的行为分析我们可以发现，大部分的水军都是以转发或评论他人的微博为主，而原创微博的比例则相对很低。出现这种情况的原因主要有一下几点：1.水军微博长期处于无人使用的状态，需要使用时也是通过程序进行集体控制，所以很少会浪费成本发布原创微博，2.根据我们同水军持有者的交流得知，相比于单纯的对原微博进行转发，每次生成语意相同但文本表示不同的微博进行原创发布需要的时间成本会很高，而且效果并不会有明显的改变。

根据之前的描述，我们设用户发布的总微博数为，其中原创微博数为，转发微博数为，在此基础上可以计算一个微博用户的原创微博比例和转发微博比例：

为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：

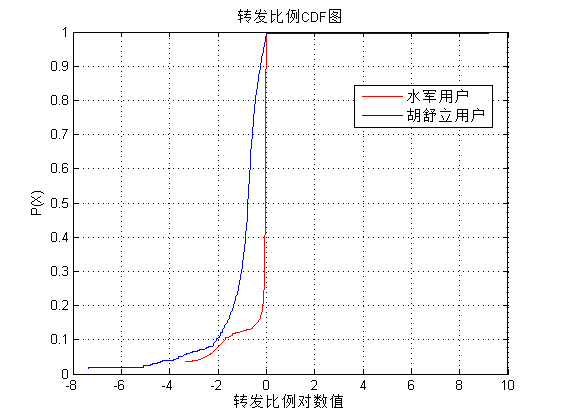
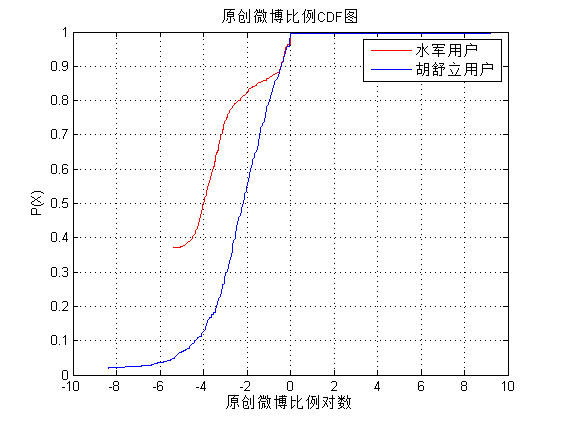


图 5.4(a) , 5.4 (b)

#### 3.2.2 转发微博平均“@”符号数

##### 3.2.2.1 “@”符号的含义

符号“@” 在新浪微博等社交媒体中是一个十分常用并且十分重要的符号，这衣服好主要起到以下两种作用：

1. @代表at，意思是“对某人说”或者“需要引起某人的注意”。想在自己的微博向某人“喊话”，或者在某条微博与某人有关，在发表的同时也想给TA一个消息提醒，让TA可以查看到时，可以使用@。这里的某人可以是在新浪微博上有帐号的任何人，也就是说，除了亲友、粉丝，你也可以公开向明星、公众人物“喊话”，他们能看到，但并不一定会回复。（“@”与“私信”功能的区别就是你可以公开向任何人喊话，却只能对关注你的人说悄悄话，除非某人设置开放私信功能给任何人。）

【格式】 @+微博用户昵称(即ID)+空格或标点，作为整体格式，可以插入到整条微博的任何位置，例如：@酒红一隅 是谁呢？



图 5.5

如上图所示，通过点击@字段可以直接跳转到用户主页，每个用户都可以使用这种方法在发布微博的时候提到其他用户，试图与提到的用户进行互动。

2. 与“//”一同出现使用。//一般是由新浪系统自动添加的，出现在再一次转发已转发并带有评论的微博时，主要起分隔针对同一微博的多人多次评论的作用，体现的是微博排队围观、众口纷纭的精神。

【表现形式】我的评论 //@微博用户昵称（第N个转发人）:TA的评论 // …… //@微博用户昵称（第1个转发人）:TA的评论。具体表现如下图所示：



图 5.6

通过上图我们可以看出“//@”表示的是一条微博的转发过程，@符号后边标注出了参与转发的用户。

##### 3.2.2.2 平均“@”符号数的含义

对“@”符号的用法有了一个全面了解之后，我们认为可以使用“@”符号来描述用户的多次转发特征。根据我们的对水军服务提供者的调研，水军进行灌水地一般过程是用户指定或自己发布目标微博，水军账号再对目标微博进行转发。为了节约时间和操作成本，不存在通过不同的用户对原文进行间接转发或二次转发的情况。也就是说，水军用户在转发时包含附加信息中的“@”符号数通常为0，所以我们针对这个特点提出了用户转发微博中平均出现的“@”符号这一特征。假定用户的转发平均“@”符号数可以表示为，则可以通过如下公式进行计算：

其中表示用户的转发微博总数，表示用户转发微博附加内容中的“@”符号数。为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：

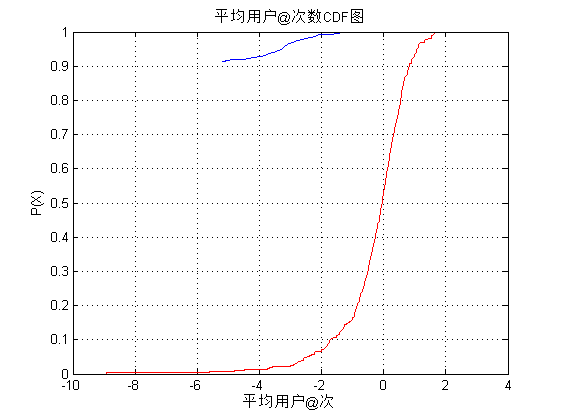


图 5.7

### 3.3 属性特征

尽管越来越多的水军持有者已经意识到了对水军账号的基本信息进行伪装的重要性，但是我们还是将属性特征作为一类重要的特征加入到了我们的特征集合当中，这样的做法主要是出于以下几方面的考虑：1.并非所有的水军持有者都会去进行帐号个人信息的伪装，尤其是小型的水军团体通常还是会使用比较原始的灌水方式和未经过伪装的账号，这样可以一定程度上提高系统的召回率；2.考虑在实际的应用过程中属性特征的计算复杂度更低更加便于提取和使用。

#### 3.3.1 粉丝数与关注数

在之前的部分我们也已经介绍过了用户的粉丝数（关注数）这一概念，这一特征可以体现用户在微博中的声望或受关注程度，也是用户活跃度的一个侧面体现。粉丝数（关注数）这一特征可以直接使用微博中给出的数字，计算复杂度低。为了确定这一特征的加入是否具有意义，我们在我们标注的数据中随机抽取真实用户与水军用户共2000人，对这两类用的粉丝数（关注数）分布进行了研究如下图所示：

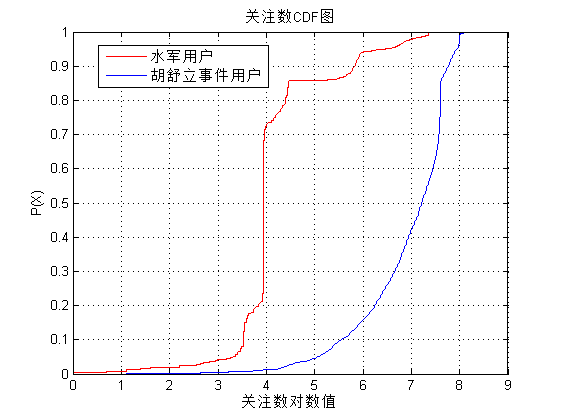
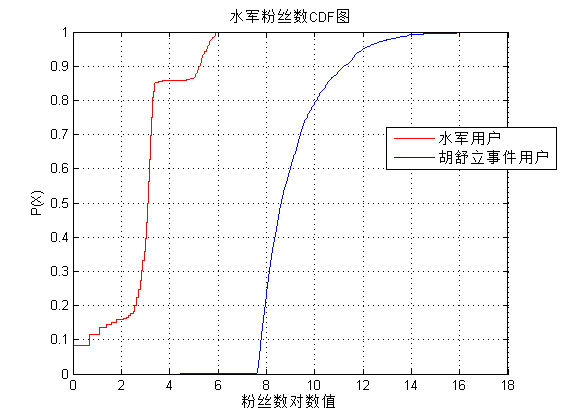


图 5.8(a) , 5.8 (b)

从图中我们可以看出，尽管水军持有者可以增加他们持有的水军账号的粉丝数（关注数），但是水军账号的粉丝数（关注数）的整体分布仍然与真实用户有所区别，这说明增加粉丝数（关注数）的策略仅仅规避了基于数值的检测条件，并不能完全规避属性特征的检测。

#### 3.3.2 粉丝关注比

为了计算这一特征，首先我们要提取用户的粉丝数和关注数，这里我们这两个特征表示为和。用户的粉丝关注比代表的就是用户的粉丝数与关注数的比值，所以用户的粉丝关注比可以通过以下公式计算获得：

选择这个特征的原因是因为在我们调研的过程中发现，大部分提供水军服务的团体或个人通常都不仅限于提供微博水军，而是包含多种有关新浪微博的业务，其中一个重要的部分就是为用户增加粉丝，也正是由于这个原因大部分的水军账号的粉丝数是要远远小于关注数的。相比之下普通用户的粉丝数和关注数比较接近，甚至对于一些大v用户粉丝数要远远大于关注数。为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：

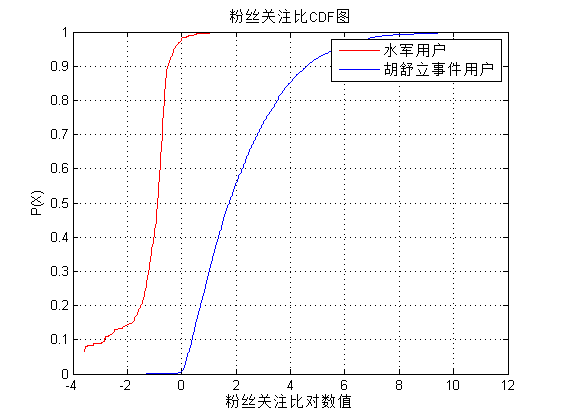


图 5.9

对于水军检测的特征，我们定义了一个特征的可规避度，一个有效的检测特征应该具有可规避度低的特点，即这种特征的伪装很难或者要付出很大的成本来伪装：伪装难度大表示水军要规避这种特征需要从根本上改变在微博中灌水地方式；付出很大的成本则是指要进行相应的伪装需要付出大量的金钱、时间或其他资源作为成本。

## 4. 实验与评估

在这一部分我们将通过实验的方法对选取的特征集合的有效性进行验证。基于机器学习的相关技术，我们通过将特征集合应用于训练分类模型来对检测效果进行评估。为了尽可能的模拟真实情况，考虑到水军检测为类类不平衡问题，我们在已标注数据中随机抽取了720名标注的真实用户和480名标注的水军用户来构成实验数据集。为了对检测效果有一个全面的评估，我们将实验分为以下几个部分：检测模型的评估，不同检测策略的效果比较，特征集合的评估。

### 4.1 检测模型的评估

通过这一部分的实验我们主要对两方面的内容进行研究：首先是通过一些评价分类器性能的常用指标来对我们构建的检测模型进行评估，其次就是分析不同的分类模型对实验结果的影响。我们共选取了4种分类模型，包括：Logistic Regression，SMO，ADtree和Random Forest，这四种分类算法根据各自特点在解决分类问题方面都有广泛的应用。对于每一个机器学习分类器，我们使用10交叉验证来进行训练与评估，在下表中我们列出了一些常用的评价指标：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 分类器 | Accuracy | Recall | F-Measure | Precision | FP-rate | ROC Area |
| SimpleLogistic | 84.83% | 0.804 | 0.809 | 0.814 | 0.122 | 0.804 |
| ADTree | 88.08% | 0.844 | 0.861 | 0.879 | 0.078 | 0.953 |
| SMO | 80.50% | 0.931 | 0.793 | 0.690 | 0.279 | 0.826 |
| RandomForest | 90.08% | 0.856 | 0.874 | 0.892 | 0.069 | 0.962 |

表 6.1

通过上面的表格我们可以对对我们的水军检测系统有一个整体的认识，使用不同的分类器都可以保证至少80%的识别准确率，召回率也都是在80%以上，从数据指标的角度上已经基本满足了我们的对于水军检测的需求。同时我们也发现，不同的分类算法之间在某些评判指标上有着较大的差距，所以我们又选取了部分指标绘制了下图：

表 6.2

通过上图我们可以对不同的算法进行比较，得到以下结论：

1. 我们可以看出smo算法无论是在直观展示检测性能的指标准确率上还是在显示分类器整体性能的roc面积上都相对较低，说明基于smo的分类模型并不能很好的完成区分两类用户的任务。但是同时我们可以看到smo模型拥有比其他三种模型更高的找汇率，所以如果在实际使用中对区分能力要求不高，而是要求找到尽可能多的水军的话则可以考虑使用smo算法。
2. 通过对比我们可以发现，ADTree和RandomForest两种基于决策树的算法在在识别准确率和和其他大部分评价指标中都优于其它两种算法，说明使用我们的特征集合基于决策树的算法可以去的相对较好的效果。
3. 由于在实际的微博用户群中水军用户的比例相对较低，所以水军检测问题可以归类为分类问题中的类不平衡问题。阈值移动与组合技术是两种常用的提高类不平衡问题准确率的方法，而我们选取的SimpleLogistic和RandomForest分别属于阈值移动方法和组合技术方法，实验结果也可以看出这两种方法也取得了相对较高的准确率。

为了进一步分析不同分类算法的检测结果，我们对研究了各个分类器分类结果的交叉关系，并绘制了下面的图表：

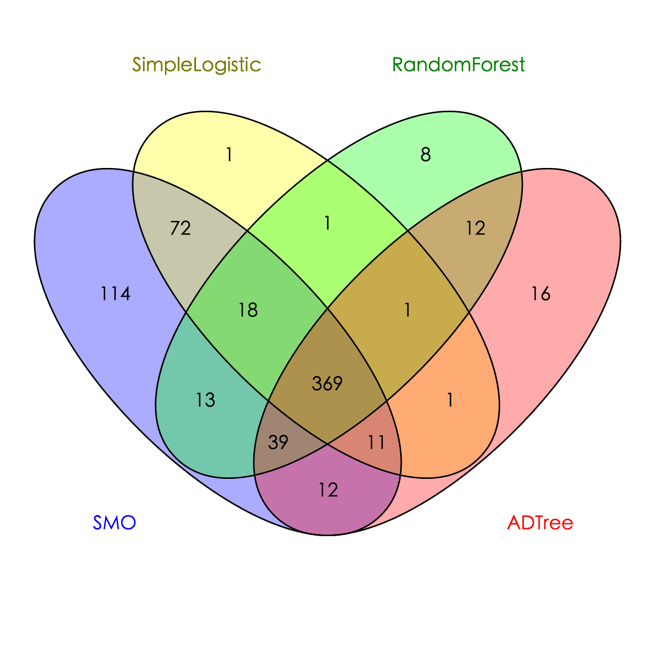


图 6.1

|  |  |
| --- | --- |
| 算法模型 | 公共部分比例 |
| SMO | 56.94% |
| SimpleLogistic | 77.84% |
| RandomForest | 80.04% |
| ADTree | 80.04% |

表6.3

上图将不同算法的结果形象地展示出来，我们可以直观的看出smo算法与其它三类算法在识别结果上的差距。经过统计四种算法的公共部分比例如表所示，除去效果较差的smo算法之后进行研究我们可以发现令外三种算法的结果重合度均在80%左右，进一步说明了结果的可信度。

### 4.2 不同检测策略的比较

这一部分中我们将对比我们的检测方法与两种现有研究方法。在这里我们选择了这两种方法中也用到的simplelogistic方法来进行实验。我们将我们的方法标记为A，方法一为B，方法二为C。对比结果如图所示：

表 6.4

由上图中可看出我们的检测方法效果要更好。在召回率方面分别有由2%和3%的提升，而在整体区分能力上的提升更为明显，准确率分别提高了12%和7%。通过以上分析可以发现我们提出的水军检测策略相比之前的研究效果上有了一定的提升。我们认为主要原因有以下几点：

1.B和C使用的特征集合的特征维度较低，导致识别效果不是很理想。

2.B中使用的***互粉比例***这一特征可以被水军使用在4.1.1.1节中提到策略以很低的成本进行规避，所以会导致实验中的效果受到影响。

3.C的***最大转发数***特征可以被4.1.2.1节中提到的规避策略所规避。

### 4.3 特征集合的分析与评估

在对我们提出的水军检测方法进行了整体评估之后，为了对选取的特征与机器学习算法之间的内在关系，在这一部分我们将对选取的特征集合进行拆分与组合，来分析不同种类特征对检测效果的影响。

在这一部分我们按照最初的特征种类将特征集合分为三个子集，如下表所示：

|  |  |  |
| --- | --- | --- |
| 行为特征 | 内容特征 | 属性特征 |
| 发帖频率(PF) | 微博平均“@”数**(MFF)** | 粉丝数**(FAN)** |
| 上网方式数(TPWN) | 转发微博比例**(RMR)** | 关注数**(FON)** |
| 转发上网方式数(TRWN) | 原创微博比例**(OMR)** | 粉丝关注比**(FDF)** |
| 是否参与人们微博(PI100) |  |  |

表 6.5

对表中的三类特征，我们定义行为特征为特征集合A ，内容特征为特征集合B，属性特征为特征集合C。在接下来的实验中，我们针对六种特征集合：A，B，C，A+B，B+C，A+C进行建模，使用SimpleLogistic分类器来进行水军检测效果的评估，结果如下表所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 特征组合 | Accrucy | Fp-rate | Precision | Recall | F-measure |
| A | 81.08% | 0.192 | 0.739 | 0.815 | 0.775 |
| B | 77.25% | 0.268 | 0.675 | 0.833 | 0.746 |
| C | 73.50% | 0.335 | 0.626 | 0.840 | 0.717 |
| A+B | 83.25% | 0.156 | 0.777 | 0.815 | 0.796 |
| A+C | 81.75% | 0.175 | 0.754 | 0.806 | 0.779 |
| B+C | 81.41% | 0.228 | 0.720 | 0.877 | 0.791 |

表 6.6

#### 4.3.1 单种类特征集合

为了跟方便进行比对，我们使用柱状图对不同特征组合的识别效果进行对比：

表 6.7

从表中我们可以看出检测模型的准确率按照行为特征、内容特征、属性特征的顺序依次下降，并且下降的幅度较大均在4%左右，所以在区分真实用户与水军用户的能力上使用行为特征的效果要更好。个别情况下我们需要尽可能多的找到疑似水军用户，这时酰胺召回率对我们更为重要，我们从表中可以看出使用属性特征可以获得更好的召回率。但是我们注意到在获得1%左右的召回率提升的同时误分率也增长了约8%，这样1%的提升就显得并不十分重要了。同时为了平衡精度与召回率，我们还对不同特征集合的f值进行了比对，从表中我们可以看到同准确率一样，特征集合的f值也是按照行为特征、内容特征、属性特征的顺序的顺序依次下降。综合以上分析，我们认为总体上对水军的检测效果为：行为特征>内容特征>属性特征。

#### 4.3.2 两种类特征集合

同样为了方便进行比对，我们仍使用柱状图对不同特征组合的识别效果进行对比：

表 6.8

通过对比表1与表2可以看出，使用两种特征组合训练的分类模型在分类准确率上有了明显的提升，同时召回率的也只有行为特征与属性特征的组合有约2%左右的损失；从另一个角度分析使用两种特征组合训练的分类模型在f值上也都有了明显提升，因此可以证明使用多种特征的组合来训练检测模型是有其必要性的。

将使用全部特征集合的评价指标与表2进行对比可以发现，加入A特征集合后检测效果的提升最为明显，加入C特征集合的提升效果最差，这也从一个侧面验证了上一节相关结论的正确性。

### 4.4 新特征的提升

将为了进一步验证我们提出的新特征的正确性，我们对两个特征集合的检测效果进行了测试。第一个集合包含我们使用的特征集合中在之前的一些研究中使用过的特征，即：转发微博比例，原创微博比例，粉丝数，关注数，粉丝关注比；第二个集合中包含我们使用的全部特征，实验结果如下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 不包含我们提出的特征 | | | 包含我们提出的特征 | | |
| 分类算法 | 准确率 | 误分率 | F值 | 准确率 | 误分率 | F值 |
| SimpleLogistic | 73.75% | 0.356 | 0.728 | 84.83% | 0.122 | 0.809 |
| ADTree | 83.66% | 0.116 | 0.834 | 89.08% | 0.078 | 0.861 |
| SMO | 63.08% | 0.161 | 0.409 | 80.50% | 0.279 | 0.793 |
| RandomForest | 85.5% | 0.095 | 0.843 | 90.08% | 0.069 | 0.874 |

表 6.9

从表中可以看出在加入了我们新设计的分类特征之后，每种算法的准确率都都有至少5%的提升，最多可以提升约17%的准确率，并且同时获得更低的误分率。以上的数据表明水军检测效果的提升确实是因为加入了我们新设计的检测特征而非是依靠对一些现有特征的组合。

## 5. 结论

在本文中，我们基于对新浪微博相关数据的深入研究以及对水军账号规避检测的策略的分析建立了新的微博水军发现方法。我们通过使用在互联网上收集大量水军的数据并对现有方法进行深入的研究，同时对水军账号的规避策略进行了分类与总结，从而提出了新的水军识别特征集合。最终，经过我们的评估，使用四种不同的机器学习模型，我们的特征集合识别准确率和误分率比现有方法都获得了较大的提升，并且可以根据不同的需求进行相应的调整。