中文微博用户性别分类方法研究*

王晶晶，李寿山，黄磊

（苏州大学计算机科学与技术学院自然语言处理实验室, 江苏 苏州 215006）

摘 要：本文旨在研究中文微博用户的性别分类问题，即根据微博提供的中文文本信息对注册用户的性别进行识别。虽然基于微博的性别分类已经有一定研究，但是针对中文的性别分类工作还很缺乏。本文首先提出分别利用用户名和微博文本构建两个分类器对用户的性别类型进行判别，并对不同的特征（例如：字特征、词特征等）进行了研究分析；其次，在针对用户名和微博文本的两个分类器的基础上，使用贝叶斯融合方法进行分类器融合，从而达到采用这两种文本分类信息同时对用户性别进行性别判断。实验结果表明本文的方法可以达到较高的识别准确率，并且分类器融合的方法明显优于仅利用用户名或者微博文本的分类方法。

关键词：性别分类；新浪微博；文本分类；社交网络；

中图分类号：TP391 **文献标识码：**A

User Gender Classification in Chinese Micro-blog

WANG Jingjing, LI Shoushan, HUANG Lei

（NLP Lab, School of Computer Science and Technology, Soochow University, Suzhou 215006）

Abstract: This paper focused on classifying the users into male and female with the information provided by Chinese Micro-blog. Although some researchers have devoted their efforts on gender classification, there is still a lack of researches in Chinese gender classification. In this paper, firstly, a classification method using user names or messages (sent by the users) to recognize male and female was proposed. Different types of features (e.g., character and word features) were investigated to perform the classification; Secondly, on the basis of the two classifiers trained with user names and messages, Bayes rule was employed to combine the two classifiers so as to make the prediction with classification knowledge from both the user names and messages. Experimental results demonstrate that the proposed approach yields a nice performance to gender classification, and the combination method outperforms the individual classifier trained with only user names or messages.

Key words: Gender Classification; Sina-Weibo; Text Classification; Social Media;

1 引言

近几年来，随着社交网络的迅猛发展，各种类型的微博即微型博客（Micro-blog）备受用户的青睐，例如 Twitter、Facebook 等。新浪微博是国内知名的微博网站，截止到 2012 年 12 月，新浪微博注册用户突破 5.03 亿，用户每日发博量超过 1 亿条。由于微博既具有媒体传播特性，又具有社交网络特性。因此，吸引了众多研究人员对微博数据进行分析研究 [1][2]。例如：利用微博进行在线用户潜在的人口特征识别[3]。

* **收稿日期：** **定稿日期：**

基金项目：国家自然科学基金(61375073)

在微博数据研究中，性别特征的研究渐渐收到了越来越多的关注 [4-6]。获得的用户的性别信息可以被应用在很多领域，例如市场推广、广告宣传和法律侦查等[7]。虽然基于微博的性别分类已经有一定研究，但是针对中文的性别分类工作还很缺乏。而且，中文微博的用户性别识别同英文微博的用户性别识别存在一定的差异，例如：中文的用户名的尾字包含丰富的性别区分信息（如“*刘雯雯*”、“*黄晓萌*”中的“*雯*”和“*萌*”）。

通过观察中文微博信息，我们发现，对于用户性别识别，一种简单有效的分类方法可能是利用用户名文本进行判别。例如，“*谢娜*”、“*曲婉婷*”等包含“*娜*”、“*婷*”等女性化字词的用户名更有可能是女性用户，而“*王刚*”、“*毕福剑*”等包含“*刚*”、“*剑*”等男性化字词的用户名更有可能是男性用户。

此外，我们还可以利用微博提供的与用户有关的其他信息来识别用户性别。较常见的信息是用户发表的微博。这些微博内容可能会有效帮助识别该用户的性别。例如，“*天哪！一双高跟鞋和一件连衣裙加起来才一千不到，老娘果断打劫了*”这条微博包含了“*高跟鞋*”、“*连衣裙*”、“*老娘*”这些女性使用比较频繁的词。因此，发表该微博的用户更可能是一位女性。像“我靠，防水的*剃须刀，就是牛掰*”包含了“*我靠*”、“*剃须刀*”等词的微博，更可能来自一位男性。

本文旨在提出一种基于中文微博的用户性别分类方法。具体而言，分别利用用户名及用户发表的微博文本信息对微博用户进行性别分类；在此基础上，我们进一步提出将用户名及用户发表的微博两种文本信息进行融合的分类方法，从而达到更好的分类效果。

本文其他部分组织如下：第二节介绍微博中用户性别研究的相关工作；第三节介绍微博语料的收集；第四节介绍本文提出的基于用户名和微博文本的分类方法；第五节给出实验设置及结果分析；第六节给出结论，并对下一步工作进行展望。

2 相关工作

近十年来，自然语言处理领域的研究人员针对性别分类分别在博客、电子邮件、微博等平台上面进行了一定研究。

首先，较多的性别分类研究工作是利用博客文本识别用户性别。例如：Schler 等[4]利用男女用户的博客文本在写作风格和内容上的不同，来确定一个未知用户的性别类型；Yan and Yan[5]利用朴素贝叶斯（Na ïve Bayes）分类方法来识别博客用户的性别；Nowson 等人[8]利用博客文本构建了一个自动识别性别的特征集。类似地，其他研究人员进行了进一步的研究去发现更有效的特征来提高分类性能。例如：Mukherjee 等人[6]、Peersman 等人[9]和 Gianfortoni 等[10]。值得注意的是，Ikeda 等[11]提出了一个半监督学习方法，有效地利用未标注样本来提升性别分类性能。

其次，一些性别分类研究工作是利用电子邮件文本识别用户性别。例如：Corney 等[12]从 Email 文本中抽取了一个与内容无关的特征集来进行性别分类；Mohammad 等[13]发现不同性别的用户在他们工作邮件中使用的情感类词汇存在明显差异。

最近，随着社交网络的发展，一些研究人员开始把他们的目光转向微博文本。例如：Burger 等[3]利用 Twitter 用户的 tweets 和个人资料（账户名、全名、个人描述）以及他们的结合作为特征来识别用户的性别；Miller 等人[7]使用 *n* 元特征的感知器（Perceptron）和朴素贝叶斯（Na ïve Bayes）算法去识别 Twitter 用户的性别；Ciot 等人[14]第一次使用非英文文

本来识别 Twitter 用户的性别；Alowibdi 等人[15]在 Twitter 文本基础上探索独立于语言的性别分类。

我们的研究也是在微博文本基础上进行的。与以往微博研究不同的是：首先，本文针对中文微博的用户性别分类，目前关于中文微博的用户性别分类方法研究还比较缺乏；其次，本文将组合分类方法应用到解决微博用户性别识别问题中，用于组合分别基于用户名特征和微博文本特征的两个基分类器。实验结果表明该方法要明显优于已有的特征叠加方法例如：Burger 等[3]。

3 语料描述

我们利用新浪微博[1]提供的开放 API 接口获取用户数据。数据包含用户的个人信息（包含用户名、性别、及其认证类型等）及用户近期发表的微博。具体而言，我们首先随机选择一个用户，获取其用户信息包括他的关注者、粉丝 ID。然后再收集其关注者、粉丝的微博信息。最后重复以上操作直到收集工作结束。需要说明的是，在获取语料的过程中考虑到存储空间的局限性及为了提高抓取的速度，我们限制每个用户的微博数目不超过 500 条。并且为了保证实验结果的精确性，我们过滤掉了发表微博次数小于 3 的用户。

新浪微博根据用户的认证类型将用户分为五大类：“黄 V 用户”、“蓝 V 用户”、“微博女郎”、“达人用户”和“普通用户”。其中“蓝 V 用户”是一些企业性质的用户而非个人用户，“普通用户”是指没有经过新浪微博官方认证的用户(包含个人用户和企业性质的用户)。为了避免大量的人工标注及保证实验结果的精确性，本次实验中我们抓取了除这两者以外的其他经过官方认证的个人用户，总共 4190 个。其中男性用户 1880 个，女性用户 2310 个。各种认证类型的用户数目如表 1 所示。

表 1 不同认证类型的用户介绍及实例

Tab.1 Introduction and Examples of Different Verification Types

认证类型	用户数	男	女	用户名例子
“黄 V 用户”	2360	1001	1359	*“邓紫棋”、“张杰”;*
“微博女郎”	30	0	30	*“青钶姑娘”、“白静”;*
“达人用户”	1800	879	921	*“行者靠谱”、“黑木崖之東方不敗”。*

4 微博用户的性别分类方法

本文采用基于机器学习的方法进行用户性别分类。首先，我们从两种文本（即：用户名和用户发表的微博文本）中抽取特征；然后，基于这两种文本分别构建两个不同的分类器；最后，采用组合分类器方法对这两个分类器进行融合，以便获得更佳的分类性能。下面我们将具体描述这两种文本的特征及组合分类器方法。

4.1 特征分析

用户名文本：根据用户名文本，我们发现女性用户的用户名尾字普遍包含偏女性化字

[1] http://weibo.com

眼。例如，“*刘雯雯*”、“*黄晓萌*”、“*休眠中的王小姐*”等。并且像“*妞妞*”、“*女王不淡定叻*”、及“*珊珊*”这些用户名，根据其首字判断其为女性用户的可能性较大；此外，男性用户的用户名也有类似的特点。例如“*赵英俊是潇洒哥*”、“*林书豪*”、“*天行者-王猛*”及“*好好先生李大爷*”这几个用户名，根据其尾字判断其为男性的可能性较大。因此，我们考虑将用户名的首尾字特征加入特征空间，即在首尾字特征后加入特定符号（例如‘_f’、‘_l’等）以示区别。

表 2 文本字特征介绍及样例

Tab.2 Introduction and Examples of Character Features

字特征	用户名特征举例 （“刘雯雯”）	微博文本特征举例 （“今天我笑到黄瓜菜都凉了”）
Unigram	‘刘’、‘雯’、‘雯’	‘今’、‘天’ … ‘凉’、‘了’
Unigram+首尾字	**‘刘_f’**、‘刘’、‘雯’、‘雯’、**‘雯_l’**	**‘今_f’**、‘今’、‘天’ … 凉’、‘了’、**‘了_l’**
Unigram+Bigram	‘刘’、‘雯’、‘雯’、‘刘雯’、‘雯雯’	‘今’、‘天 … ‘凉’、‘了’、‘今天’、‘天我’ … ‘都凉’、‘凉了’
Unigram+Bigram+首尾字	**‘刘_f’**、‘刘’、‘雯’、‘雯’、**‘雯_l’**、**‘刘雯_f’**、‘刘雯’、‘雯雯’、**‘雯雯_l’**	**‘今_f’**、‘今’、‘天’ … ‘凉’、‘了’、**‘了_l’**、**‘今天_f’**、‘今天’、‘天我’ … ‘都凉’、‘凉了’、**‘凉了_l’**

表 3 文本词特征介绍及样例

Tab.3 Introduction and Examples of Word Features

词特征	用户名特征举例 （“刘雯雯”）	微博文本特征举例 （“今天 我 笑 到 黄瓜 菜 都 凉 了”）
Unigram	‘刘’、‘雯雯’	‘今天’、‘我’ … ‘凉’、‘了’
Unigram+首尾字	**‘刘_f’**、‘刘’、‘雯雯’、**‘雯雯_l’**	**‘今天_f’**、‘今天’、‘我’ … ‘凉’、‘了’、**‘了_l’**
Unigram+Bigram	‘刘’、‘雯雯’、‘刘雯雯’	‘今天’、‘我’ … ‘了’、‘今天我’、‘我笑’ … ‘都凉’、‘凉了’
Unigram+Bigram+首尾字	**‘刘_f’**、‘刘’、‘雯雯’、‘刘雯雯’、**‘雯雯_l’**、‘**刘雯雯_l**’	**‘今天_f’**、‘今天’、‘我’ … ‘了’、‘**今天我_f**’、‘今天我’、‘我笑’ … ‘都凉’、‘凉了’、**‘了_l’**、**‘凉了_l’**

表 4 男女用户微博文本中前 10 个 IG 值最高的词特征

Tab.4 Highest IG Features of Weibo in Male and Female

男性	*‘兄弟’、‘哥’、‘打篮球’、‘日’、‘比赛’、‘足球’、‘酒’、‘蛋疼’、‘球员’、‘老婆’*
女性	*‘亲亲’、‘委屈’、‘可爱’、‘老娘’、‘蛋糕’、‘亲爱’、‘礼物’、‘宝宝’、‘丈夫’、‘害羞’*

发表的微博文本：对于用户发表的微博文本，我们利用一种标准的特征选择方法信息增益（IG）[16]计算了其每个词特征的性别信息量。并且对男性和女性用户的微博文本分别选取了前 10 个 IG 值最高的词特征，如表 4。从表 4 中可以看出，用户发表的微博文本中包含了许多与性别相关的字眼和称谓，而这些特征能给我们预测用户的性别提供很好的线索。例如，‘*哥*’、‘*老婆*’这两个特征更可能来自男性用户，而‘*亲亲*’、‘*老娘*’等更像女

性用户发表的。

表 2 和表 3 给出了一些例子来描述我们选择用户名和微博文本的具体特征的方法。其中，Unigram 即一元特征，Unigram+首尾字特征是指在文本一元特征的基础上加入首尾一元特征 ，Uni+Bigram 即一、二元的组合特征，Uni+Bigram+首尾字特征是指在一、二元特征的基础上加入首尾一、二元特征。

4.2 分类器融合

组合分类方法是融合多个分类器的结果从而得到一个新的分类结果作为最终的分类决定[17]。组合分类方法是模式识别以及机器学习理论研究领域里面的一个重要的研究方向。

图 1 给出了组合分类器方法的系统框架图。从图中可以看出，我们的组合分类器方法主要分为两步：

(1) **训练基分类器**：我们通过训练两个不同的语料产生两个不同的基分类器。两个基分类器分别为基于用户名文本和基于微博文本训练的分类器。

(2) **分类器融合**：我们利用融合算法将两个基分类器结果融合得到最终分类结果。假设有 R 个参加组合的分类器 $f_k(k=1,....,R)$，这些分类器给样本 x 的分类结果为 L_k $(L_k=c_1,...,c_m)$．另外，他们提供出了属于每个类别的概率信息：$P_k=[p(c_1\mid d_k),\ ...,\ p(c_m\mid d_k)]^t$，其中 $p(c_i\mid d_k)$ 表示样本 d_k 属于类别 c_i 的概率。如果样本 d_k 属于类别 c_j，在不同的融合算法中需要满足不一样的条件。本文中我们采用的是一种常见的固定融合算法，贝叶斯规则的条件：

$$\textbf{贝叶斯规则：}\quad \begin{aligned} &assign \quad y\to c_j \\ &j=\arg\max_i p(c_i)\prod_{k=1}^{R} p(c_i\mid d_k) \end{aligned} \tag{1}$$

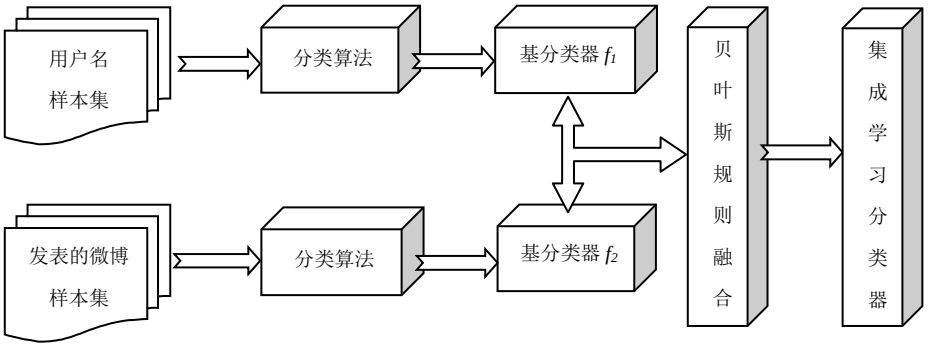

图 1 组合分类器系统的框架架构

Fig.1 Framework of Combining Classifier

5 实验

5.1 实验设置

本实验中，我们利用新浪微博 API 收集了 4190 个经过官方认证的个人用户的信息及其发表的微博文本。其中男性用户有 1880 个（正类样本），女性用户有 2310 个（负类样本）。本实验的任务是利用用户的用户名和微博文本来识别用户的性别。在实验过程中，我们采用最大熵方法（Maximum Entropy，ME）作为分类算法，其中 ME 使用的是 MALLET 机器学习工具包[2]。在使用个工具包的时候，所有的参数都设置为它们的默认值。此外，我们采用用户名和微博文本的字和词等特征构建分类器，具体的特征描述可参考 4.1 节。需要说明的是，我们采用复旦大学自然语言处理实验室开发的分词软件 FudanNLP[3]对文本进行分词操作以便选取词特征。从搜集的样本中，我们选取 2800 个样本作为训练样本，200 个样本作为测试样本。

5.2 实验结果与分析

首先，图 2 给出了分别使用用户名文本和微博文本的不同特征时的分类结果。从结果可以看出：

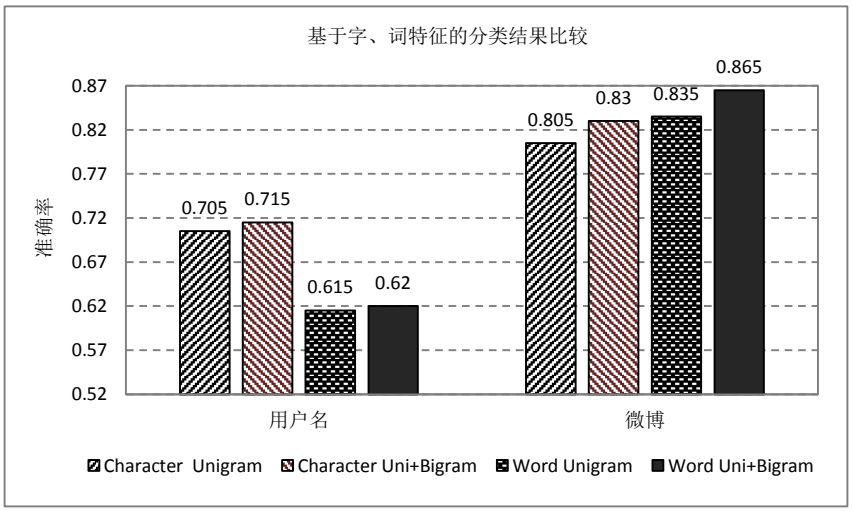

图 2 基于字和词特征的分类结果比较

Fig.2 Performance comparison of using character and word features

1) 利用用户名文本进行分类时，其字特征要明显优于词特征，高出的幅度超过 9 个百分点。该结果主要是因为分词操作对用户名文本进行切分时，会将某些用户名作为一个整体，导致训练样本特征空间非常稀疏。此外，用户名中的某些字，如，称呼（‘*姐*’、‘*哥*’等）对性别分类是有很大帮助的；而利用微博文本分类时，词特征要好于字特征。在 Uni+Bigram 下尤其突出，前者比后者高出了 3.5 个百分点。

[2] http://mallet.cs.umass.edu/
[3] https://code.google.com/p/fudannlp/

2) 在 Unigram 基础上加入 Bigram 特征后，微博文本在字特征和词特征下分类性能都获得了提升，其中使用词特征时提高了 3 个百分点。然而，在使用用户名文本时，性能提高并不明显。

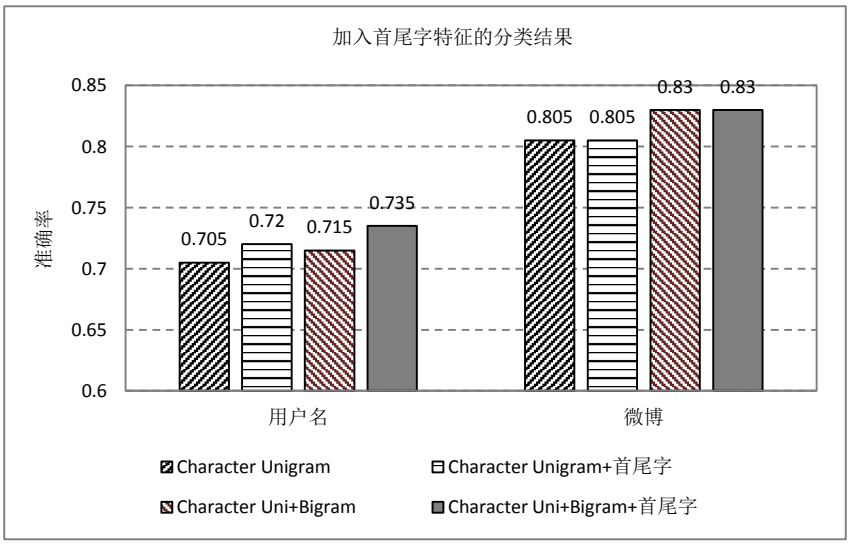

图 3 加入首末字特征的分类结果

Fig.3 Performance Comparison of Classifiers with Initial and End Character Features

其次，考虑到首尾字对用户名分类的重要性，我们尝试将用户名的首尾字作为特征加入到特征空间中，具体来讲，在首尾字后追加特殊符号例如‘_f’、‘_l’以示区别，实验结果如图 3 所示。由图 3 可以看出，对于用户名，加入首尾字特征后分类性能获得了提升。其中字的 Uni+Bigram 特征下最明显，提高了 2 个百分点；对于微博文本，加入首尾字特征对分类结果没有影响。这是由于用户名如 4.1 节特征分析中描述的一样，其首尾字对性别的区分有帮助，而微博文本的首尾字对性别分类并没有特殊意义，并且每条微博训练样本太长，分类器的特征空间太大，导致加入的两类用户名特征对分类结果没有影响。

值得注意的是，从图 2 和图 3 可以看出，微博文本的分类效果始终要明显优于用户名的分类效果。在使用字特征的时候，前者仍比后者平均高出 10 个百分点。为此，我们特地随机抽取 200 样本（男女各 100），采用人为识别的方式对用户名进行性别分类，结果显示在表 5 中。如表 5 所示，分类器的分类效果（使用最好特征集合）只比人工标注的低 4.8 个百分点。即使通过人工来识别用户名，其分类效果仍比通过微博文本识别差 8.2 个百分点。其主要原因是某些用户的用户名不包含明确的性别特征，例如，“*思想聚焦*”、“*爱吃鱼的列*”、“*幸运素数*”等。

表 5 用户名性别识别的人工识别效果和自动识别效果比较

Tab.5 Performance Comparison between Manual and Automatic Recognition

人工识别正确的数目	人工识别错误的数目	人工无法识别的数目	人工识别正确率	分类器自动识别正确率
124	13	63	0.783	0.735

综合以上结果分析得出：对于用户名文本，利用其字的 Uni+Bigram+首尾字特征分类效果最佳。而对于微博文本，其词的 Uni+Bigram 特征表现最好。因此，我们使用这两种特征作为两种文本的基分类器的特征实现，用于进一步的分类器融合。

在上述的特征研究的基础上，我们给出下面四种分类方法的分类结果进行比较：

- **用户名**：利用用户名文本的字的 Uni+Bigram+首尾字特征训练基分类器；
- **发表的微博**：利用用户所发微博文本的词的 Uni+Bigram 特征训练基分类器；
- **特征叠加**：将上述两种方法中的两种特征叠加之后作为训练集训练分类器；
- **融合**：训练上述用户名和微博文本的两个基分类器，然后根据贝叶斯融合规则将两个基分类器融合，得到最终分类结果；

图 4 给出训练样本规模从 900 变化到 2800 时这四种方法的分类准确率。从图中可以看出：1）将用户名和微博文本进行特征叠加并不能明显提高分类性能。该结果的可能原因是微博的特征空间远远高于用户名的特征空间，导致用户名特征不足以对分类结果产生影响。相比而言，分类器融合方法获得的分类性能明显优于其他三种方法，比用户名分类器平均提高出 18 个百分点，比微博分类器平均提高 3 个百分点。2）训练样本从 900 变化到 2800 时，融合的结果稳定提高，且始终高于其他三种方法。该结果表明我们的融合方法有较好的稳定性。

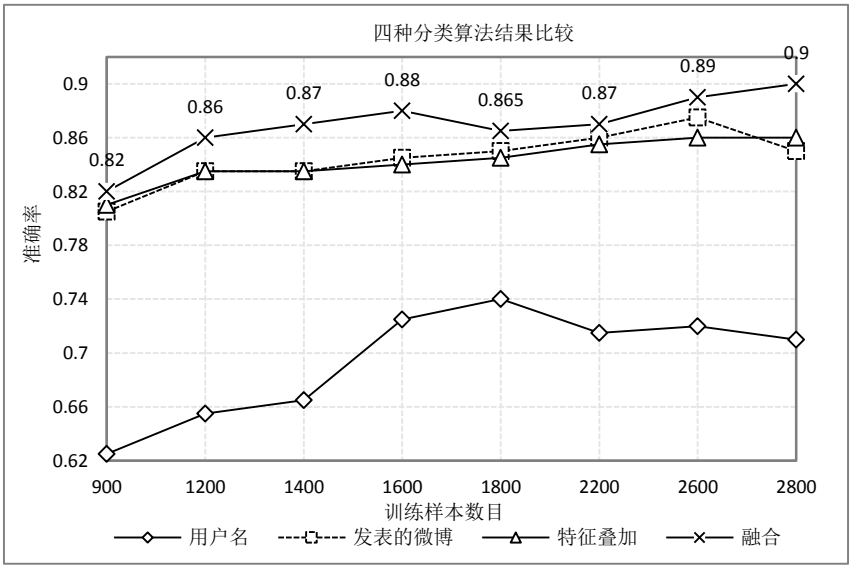

图 4 四种分类方法分类结果比较

Fig.4 Performance Comparison of Classifiers with Different Methods

最后，我们对分类结果进行分析，发现分类错误主要存在以下几点原因：1）测试样本中仍然存在某些特征未在训练样本中出现，导致分类器无法学习到该部分信息；2）存在部分用户由于发表的微博文本较少而且其用户名属于男性还是女性并不明显，即使人工查看也无法区分其类别，以至于分类方法无法对其进行分类，例如，“*思想聚焦*”、“*爱吃鱼的列*”、“*幸运素数*”等。

6 结论

本文提出了一种基于用户名和微博文本的分类器融合方法来对用户性别进行分类。具体来讲，首先利用用户名和微博文本分别训练两个基分类器，然后根据贝叶斯规则对分类结果进行融合。实验结果表明（1）在中文微博性别分类中，用户名文本里面的字特征具有一定的分类性能（但分类性能有限）；（2）使用微博文本能够获得比使用用户名文本更好的分类性能；（3）我们的分类器融合方法对用户性别的识别能取得最佳的分类性能，并且分类效果明显优于利用用户名文本、微博文本或两者文本特征叠加的分类方法。

除了用户名和微博文本外，微博中往往还包含了其他与用户性别相关的信息，例如，关注者、粉丝及转发等信息。在下一步工作中，我们考虑将更多用户信息加入到中文微博用户性别识别任务中。

参考文献

[1] 文坤梅，徐帅，李瑞轩，等. 微博及中文微博信息处理研究综述[J]. 中文信息学报，2012，26（6）：28-36.

[2] 张剑峰，夏云庆，姚建民. 微博文本处理研究综述[J]. 中文信息学报，2012，26（4）：21-27.

[3] Burger J. and J. Henderson and G. Kim and G. Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of EMNLP-11*, pp. 1301–1309.

[4] Schler J., M. Koppel, S. Argamon and J. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *Proceedings of AAAI-06*, pp. 199-205.

[5] Yan X. and L. Yan. 2006. Gender Classification of Weblog Authors. In *Proceedings of AAAI-06*, pp. 228-230.

[6] Mukherjee A. and B. Liu. 2010. Improving Gender Classification of Blog Authors. In *Proceedings of EMNLP-10*, pp. 207-217.

[7] Miller Z., B. Dickinson and W. Hu. 2012. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. In *Proceedings of International Journal of Intelligence Science, Vol.2, No.4*, pp.143-148.

[8] Nowson S. and J. Oberlander. 2006. The Identity of Bloggers: Openness and Gender in Personal Weblogs. In *Proceeding of AAAI-06*, pp. 163-167.

[9] Peersman C., W. Daelemans, L. Van Vaerenbergh. 2011. Predicting Age and Gender in Online Social Networks. In *Proceedings of SMUC-11*, pp. 37-44.

[10] Gianfortoni P., D. Adamson and C. Rosé 2011. Modeling of Stylistic Variation in Social Media with Stretchy Patterns. In *Proceedings of EMNLP-11*, pp. 49–59.

[11] Ikeda D., H. Takamura and M. Okumura. 2008. Semi-Supervised Learning for Blog Classification. In *Proceedings of AAAI-08*, pp.1156-1161.

[12] Corney M., O. Vel, A. Anderson and G. Mohay. 2002. Gender-Preferential Text Mining of E-mail Discourse. In *Proceedings of ACSAC-02*, pp. 282-289.

[13] Mohammad S. and T. Yang. 2011. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis-11*, pp.70-79.

[14] Ciot M., M. Sonderegger and D. Ruths. 2013. Gender Inference of Twitter Users in Non-English Contexts. In *Proceedings of EMNLP-13*, pp. 1136–1145.

[15] Alowibdi J., U. Buy and P. Yu.2013. Language Independent Gender Classification on Twitter. In *Proceedings of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 739-743.

[16] Li S., R. Xia, C. Zong, and C. Huang. 2009. A Framework of Feature Selection Methods for Text Categorization. In *Proceedings of ACL-IJCNLP-09*, pp. 692-700.

[17] Kittler J., M. Hatef, R. Duin and J. Matas. 1998. On Combining Classifiers. In *Proceedings of IEEE-98*, pp. 226-239.

作者简介：王晶晶（1990--），男，江苏南通人，硕士研究生，研究方向为自然语言处理，Email:djingwang@gmail.com；李寿山（1980--），男，教授，博士后，研究方向为自然语言处理，Email:shoushan.li@gmail.com；黄磊（1989--），男，江苏徐州人，硕士研究生，研究方向为自然语言处理方向，Email:lei.huang2013@gmail.com。

王晶晶

李寿山

黄磊