基于双通道 LSTM 模型的用户性别分类方法研究

王礼敏 严 倩 李寿山 周国栋

(苏州大学计算机科学与技术学院 江苏 苏州 215006)

摘 要 微博用户性别分类旨在根据用户信息进行用户性别的识别。目前性别分类的相关研究主要针对单一类型的特征(文本特征或者社交特征)进行性别分类。与以往研究不同,文中提出了一种双通道 LSTM(Long-Short Term Memory)模型,以充分结合文本特征(用户发表的微博文本)和社交特征(用户关注者的信息)进行用户性别分类方法的研究。首先,利用单通道 LSTM 模型分别学习两组文本特征,得到两种特征表示;然后,在神经网络中加入 Merge 层,结合两种特征表示进行集成学习,以充分学习文本特征和社交特征之间的联系。实验结果表明,相对于传统的分类算法,双通道 LSTM 模型分类算法能够获得更好的用户性别分类效果。

关键词 性别分类,新浪微博,双通道 LSTM

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.021

User Gender Classification with Dual-channel LSTM

WANG Li-min YAN Qian LI Shou-shan ZHOU Guo-dong

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangshu 215006, China)

Abstract User gender classification aims at classifying the users into male and female with the provided information. Previous studies on gender classification mainly focus on a single type of features (i.e., textual features or social features). Different from previous research, this paper proposed a new approach named dual-channel LSTM by making full use of the relationship between textual features (the text which user publishes) and social features (the followers which user concerns). Specifically, this paper first got two kinds of features using single-channel LSTM respectively. Then, it proposed a joint learning method to integrate the features. Lastly, it got the final classification results by the dual-channel LSTM. Empirical studies show that the dual-channel LSTM model achieves effective results for gender classification compared with traditional classification algorithms.

Keywords Gender classification, Sina weibo, Dual-channel LSTM

1 引言

进入 21 世纪后,网络在人们的日常生活中扮演着越来越重要的角色,人们已经习惯于在各种社交平台(如微博、微信、Facebook 等)上发表自己的见闻及观点。作为国内知名的微博网站,新浪微博受到了广大用户的喜爱。截至 2015 年 9 月,微博月活跃用户数(MAU)已经达到 2.12 亿,日均活跃用户数(DAU)达到 1 亿。微博由于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[1-3],如对微博用户进行情感分析、性别分类、社交群体挖掘等。其中,获取用户的性别信息是一项基本任务,该任务可以被应用到很多领域,如市场推广、广告宣传和法律侦查等[4]。

已有的基于微博用户性别的研究多是通过对用户发表的微博内容进行分析,进而判断微博用户的性别。通过分析微博语料发现,用户发表的微博文本信息以及社交信息能够清晰地体现微博用户的性别。如图 1 所示,用户 A 的微博中包含了"老子""篮球""科比"这些男性频繁使用的词,而且从关注的信息来看,他关注了篮球类、游戏类、财经信息类等微博,因此我们倾向于认为用户 A 是一名男性。而对于用户 B,其微博内容包含了"姑娘""高跟鞋""指甲油"等词,且其关注了娱乐八卦类、明星、各种时尚品牌类微博,因此认为该微博用户更可能是位女性。从这些例子可以看出,用户发表的微博及用户的社交信息(例如关注对象)能够有效识别用户的性别信息。

本文结合用户微博文本信息以及社交文本两类信息,提出一种基于双通道 LSTM 的微博用户性别分类方法。具体而言,在神经网络中加入一个 Merge 层,将 LSTM 分别产生的文本特征表示和关注者特征表示相结合以进行联合学习,从而充分学习两者之间的关系。实验结果表明,所提方法能

到稿日期:2017-05-18 返修日期:2017-06-26 本文受国家自然科学基金(61672366)资助。

王礼敏(1993—),女,硕士,主要研究方向为自然语言处理,E-mail:lmwang@stu.suda.edu.cn;**严 倩**(1993—),女,硕士,主要研究方向为自然语言处理;**李寿山**(1980—),男,博士,教授,主要研究方向为自然语言处理,E-mail:lishoushan@suda.edu.cn(通信作者)。

够获得更佳的性别分类效果。

用户 A
姓名：*** 年龄：*** 性别：*
社交信息(用户个人信息)
发布的微博数：324 关注者：NBA，财经网，Dota2，……
文本信息(发布的微博)
“老子爱上篮球就是科比，纪念，感恩！”
用户 B
姓名：*** 年龄：*** 性别：*
社交信息(用户个人信息)
发布的微博数：286 关注者：Dior，八卦，周杰伦，……
文本信息(发布的微博)
“姑娘我 18 了，可以穿高跟鞋，涂指甲油了！”

图 1 新浪微博中的用户实例

Fig. 1 User example in Sina weibo

本文第 2 节介绍微博中用户性别研究的相关工作；第 3 节介绍提出的基于 LSTM 的用户性别分类方法；第 4 节给出实验设置及结果分析；最后总结全文并对下一步工作进行展望。

2 相关工作

近年来，自然语言处理领域的研究人员对博客、微博等社交媒体用户的性别属性预测进行了诸多研究。

Morgane 等[5]针对 Twitter 上使用非英文(如法语、日语等)的用户进行性别预测；Conover 等[6]研究预测了 Twitter 用户的政治主张；Liu 等[7]在 Twitter 上以用户的名字为特征来预测性别；Maarten 等[8]提出在社交媒体上创建年龄和性别预测词表；Miller 等[4]使用 n 元特征的感知器和朴素贝叶斯算法来识别 Twitter 用户的性别；Marquardt 等[9]基于文本提出了多标签的分类方法来预测用户的性别及年龄；王晶晶等[3]利用用户名和微博文本分别训练两个基分类器，并根据贝叶斯规则融合两个分类器来对微博用户的性别进行分类。

本文研究与已有研究具有明显的不同：首先，本文针对的是中文微博用户的分类研究，目前这方面的相关研究工作还比较匮乏；其次，本文结合微博文本和关注者这两组特征，采用基于长短时记忆(Long-Short Term Memory)模型的神经网络方法来解决微博用户的性别识别问题，获得了比传统模型更好的分类效果。

3 基于 LSTM 的用户性别分类方法

本文结合用户发表的微博文本以及用户社交信息，分别将这两种信息作为 LSTM 模型的输入，然后通过双通道 LSTM 模型融合两者的输出。

3.1 基于单通道 LSTM 的性别分类方法

Hochreiter 和 Schmidhuber[10] 于 1997 年提出了一种 RNN 的特殊类型——长短时记忆模型。在这个模型中，常规的神经元(即一个将 S 型激活应用于其输入线性组合的单元)被存储单元所代替。每个存储单元(Memory Cell)与一个输入门(Input Gate)、一个输出门(Output Gate)和一个跨越时间步骤无干扰送入自身内部状态的单元相关联。随后，Graves[11]改良推广了该模型，在存储单元中又加入了一个新的结构忘记门(Forget Gate)。LSTM 单元的大致结构如图 2 所示。

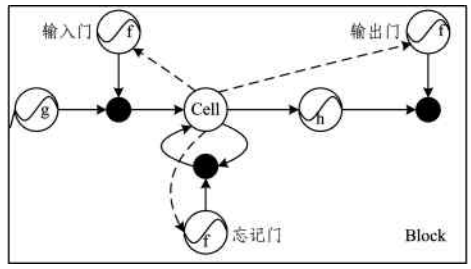

图 2 LSTM 单元

Fig. 2 LSTM cell

每个记忆单元在不同的时间点完成更新的具体过程如下。假设 x_t 表示在时间 t 下的输入；W_i，W_f，W_c，W_o，U_i，U_f，U_c，U_o 和 V_o 表示权重矩阵；b_i，b_f，b_c 和 b_o 是偏置向量。

首先，在记忆单元时间 t 的状态下，计算输出层将要更新的值 i_t，并创建一个新的候选状态值$\tilde{C}_t$：

$$i_t=\sigma(W_i x_t+U_i h_{t-1}+b_i) \tag{1}$$

$$\tilde{C}_t=\tanh(W_c x_t+U_c h_{t-1}+b_c) \tag{2}$$

然后，计算在时间 t 下，忘记门层输出的值 f_t：

$$f_t=\sigma(W_f x_t+U_f h_{t-1}+b_f) \tag{3}$$

以上求出了输入门层的值 i_t、忘记门层的值 f_t 和候选值向量$\tilde{C}_t$，从而可以计算在时间 t 下记忆单元的新状态量 C_t：

$$C_t=i_t * \tilde{C}_t+f_t * C_{t-1} \tag{4}$$

在记忆单元的新状态下，可以计算输出门层的状态以及 LSTM 单元的输出值：

$$o_t=\sigma(W_o x_t+U_o h_{t-1}+V_o C_t+b_o) \tag{5}$$

$$h_t=o_t * \tanh(C_t) \tag{6}$$

其中，σ 为 logistic sigmoid 函数。

单通道的 LSTM 模型只使用了一个 LSTM 层。根据更新公式，将单特征文本用词向量表示，通过 LSTM 层得到高维向量，将所有输出作为全连接层的输入。

Dense 层就是常用的全连接层，与传统多层感知机的隐藏层类似，接收来自上一层的输出，通过常用的激活函数对其加权并传播到下一层。本实验使用“relu”作为激励函数，它可以减少参数之间的依存关系，更符合生物的大脑特性[12]。

$$h^*=\phi(\theta^{\mathrm{T}} h+b) \tag{7}$$

其中，ϕ 是非线性激活函数，h 是 LSTM 层的输出。

Dropout 层被应用于前馈神经网络[13]，可以随机地让网络中某些隐含层节点的权重不工作，有效地防止网络过拟合。Dropout 层作为 LSTM 模型中的隐藏层出现：

$$g=h^* \cdot D(p) \tag{8}$$

其中，D 表示 dropout 操作符，p 是一个可调的超参(保留隐层单元的比率)。

最后，因为本文的任务是二分类，所以网络通过 sigmoid 层来预测各样本的类别。具体模型结构如图 3 所示。

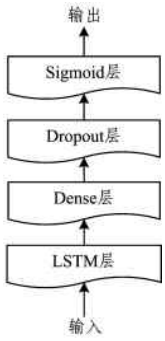

图 3　单通道 LSTM 模型

Fig. 3　One-channel LSTM

3.2　基于双通道 LSTM 的性别分类方法

为了有效区别微博文本特征和关注者特征，并充分利用这两种特征文本间的关系，提出了一种双通道 LSTM 神经网络，即通过联合学习融合两组特征。图 4 给出了双通道 LSTM 模型框架，图中的文本 LSTM 表示和社交 LSTM 表示是微博文本和关注者文本分别经过单通道 LSTM 得到的新的文本表示。

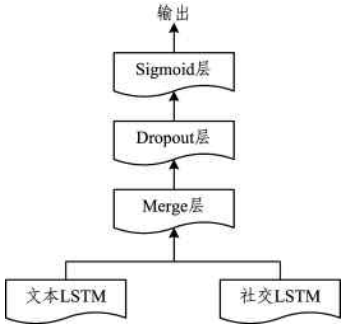

图 4　双通道 LSTM 模型

Fig. 4　Dual-channel LSTM

Merge 层将两组 LSTM 模型的输出特征进行连接融合，并通过反向传播算法进行参数更新。Dropout 层将来自 Merge 层的输出作为自己的输入，具体操作与单通道的 LSTM 完全一致。Sigmoid 输出层用于文本分类。在这个模型中选择可能性最大的类别作为模型的预测标签 $label_{\text{pred}}$：

$$label_{\text{pred}}=\arg\max_i\ P(Y=i\mid x,W,U,V) \tag{9}$$

在联合学习的过程中选择的损失函数为最小化交叉熵误差，具体公式如下：

$$J=-\sum_{i=1}^{n_c}t_i\log y_i+\lambda\sum_{i=1}^{m}\Big(\sum_{\varepsilon\in\omega}\|W_{\varepsilon i}\|_F^2+\sum_{\varepsilon\in\mu}\|U_{\varepsilon i}\|_F^2+\sum_{\varepsilon\in\nu}\|V_{\varepsilon i}\|_F^2\Big) \tag{10}$$

其中，$t\in\mathbb{R}^{n_c}$ 是用“one- hot”表示的实际标签，$y\in\mathbb{R}^{n_c}$ 是 sigmoid 层输出的每个类别的概率（n_c 是目标类别的数量），$\|\cdot\|_F$ 表示 Frobeniu 范数，$\omega=\{i,f,o,c\}$，$\mu=\{i,f,o,c\}$ 和 $\nu=\{i,f,o\}$ 表示不同门的集合（分别为 W,U,V），λ 是用来指定惩罚权重的超参。

4　实验

4.1　语料概述

本文的语料是通过新浪微博提供的开放 API 接口来获取的，数据包含用户的个人信息（包括用户 ID、关注者 ID、注册时间、性别、认证类型等）以及用户近期发表的动态。收集语料的具体做法是：首先随机选择一个用户，获取其关注者以及粉丝 ID，然后获取其关注者以及粉丝 ID 的微博信息。重复上述步骤，直到收集工作结束。需要说明的是，为了节省抓取时间及存储空间，只保留发博量不少于 3 条且不多于 200 条的用户的信息。最终，一共抓取了 2967 条女性用户信息和 1978 条男性用户信息。

4.2　实验设置

1）语料：男、女用户各 1800 个样本，选取其中的 80% 作为训练样本，另 20% 作为测试样本。

2）基本分类算法：在实验中，分别采用最大熵方法（Maximum Entropy，ME）和 LSTM 神经网络方法作为分类算法。其中，ME 使用的是 MALLET 机器学习工具[1)]，所有训练参数都设置为默认值；LSTM 模型的具体参数设置如表 1 所列。

3）特征：分类特征采用词特征。采用复旦大学自然语言处理实验室开发的分词软件 FudanNLP[2)] 对文本进行分词操作，实验所用特征为词袋特征（BOW）。

4）评价准则：采用准确率 Acc（Accuracy）作为分类结果的评价指标。

表 1　LSTM 中的参数设置

Table 1　Parameters of LSTM

参数及描述	设置
词的 unigram 特征总规模	40000
LSTM 层输出维度	128
全连接层输出维度	64
Dropout 速率	0.5
迭代次数	15

4.3　实验结果与分析

4.3.1　基于分类算法的结果比较

在本实验中，基于不同特征比较了以下几种性别分类方法的效果。

1）SVM：分别基于用户发表的微博状态特征、社交特征以及两种特征的融合；分类算法使用支持向量机算法。

2）MaxEnt：分别基于用户发表的微博状态特征、社交特征以及两种特征的融合；分类算法使用最大熵算法。

3）LSTM：分别基于用户发表的微博状态特征、社交特征以及两种特征的融合；分类算法使用单通道 LSTM 模型。

表 2 列出了分别基于文本特征、关注者特征以及两者特征的融合，采用最大熵分类算法以及 LSTM 分类算法的实验结果。从表中可以看出：1）将文本特征和社交特征进行融合后的分类结果优于仅使用单个特征进行分类的结果，性能提高了 2 个百分点左右；2）LSTM 模型分类算法明显优于最大熵分类算法，基于单特征时 LSTM 的分类结果相对于最大熵的分类结果提高了 1～2 个百分点，相对于支持向量机算法的分类结果提高了 3 个百分点左右。在后续的实验中，将采用

1) http://mallet.cs.umass.edu/

2) https://code.google.com/p/fudannlp/

LSTM 算法作为基本的性别分类算法。

表 2 支持向量机、最大熵与 LSTM 的分类结果比较

Table 2 Comparison of classification accuracy between SUM, MaxEnt and LSTM

算法	文本特征	社交特征	联合特征
SVM	0.812	0.839	0.85
MaxEnt	0.825	0.853	0.864
LSTM	0.843	0.865	0.889

4.3.2 单通道 LSTM 和双通道 LSTM 的分类结果比较

表 3 列出了训练样本数从 20%变化到 80%时，分别采用单通道 LSTM 进行特征学习和利用双通道 LSTM 算法进行联合特征性别分类的结果。

1)LSTM-文本：采用用户发表的微博文本特征，使用 LSTM 模型作为分类算法。

2)LSTM-社交：采用用户微博的社交特征，使用 LSTM 模型作为分类算法。

3)LSTM-文本＋社交：混合用户文本特征和社交特征，使用 LSTM 模型作为分类算法。

4)D-LSTM：采用双通道 LSTM 模型对两组特征产生的 LSTM 表示进行分类。

表 3 单通道 LSTM 和双通道 LSTM 的分类结果比较

Table 3 Comparison of classification results of one-channel LSTM and dual-channel LSTM

	LSTM-文本特征	LSTM-社交特征	LSTM-联合特征	D-LSTM
20%	0.816	0.832	0.855	0.845
40%	0.828	0.847	0.868	0.89
60%	0.840	0.84	0.883	0.89
80%	0.843	0.865	0.889	**0.903**

从表 3 可以看出，当训练样本较少时，使用双通道 LSTM 算法的分类效果相对于单通道 LSTM 的分类算法降低了 1 个百分点，主要原因在于深度学习算法适合在大规模标注样本下学习。而在训练样本为 20%时，标注样本数目较少，深度学习算法无法充分学习。在样本数量较多时，使用联合特征进行学习的效果比分别使用文本特征和关注者特征进行学习的效果更好，大概提升了 2 个百分点；而相对于基于联合特征的单通道 LSTM，本文提出的基于双通道 LSTM 的性别分类的结果提升了 1～2 个百分点。

结束语 本文利用微博文本信息和关注者信息进行用户性别分类。为了充分利用文本和社交两组特征信息，提出一种基于双通道 LSTM 的性别分类方法。实验结果表明，基于 LSTM 分类算法的性别分类方法明显优于传统分类算法。同时，本文提出的双通道 LSTM 模型能够获得最佳的分类效果，明显优于仅利用单类型特征的 LSTM 分类模型或者仅混合两组特征的 LSTM 分类模型。

除了用户发表的微博文本外，微博中还包含了一些其他重要信息，包括粉丝、转发、评论等，这些信息可能对性别分类有较大帮助。因此，下一步将考虑加入更多的用户社交信息，以提升微博用户的性别分类性能。

参 考 文 献

[1] WEN K M, XU S, LI R X, et al. Survey of Microblog and Chinese Microblog Information Processing[J]. Journal of Chinese Information Processing, 2012, 26(6): 28-36. (in Chinese)
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] ZHANG J F, XIA Y Q, YAO J M. A Review towards Microtext Processing[J]. Journal of Chinese Information Processing, 2012, 26(4): 21-27. (in Chinese)
张剑锋，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] WANG J J, LI S S, HUANG L. User Gender Classification in Chinese Microblog[J]. Journal of Chinese Information Processing, 2014, 28(6): 150-155. (in Chinese)
王晶晶，李寿山，黄磊. 中文微博用户性别分类方法研究[J]. 中文信息学报，2014，28(6)：150-155.

[4] DICKINSON M B, HU W. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features[J]. Proceedings of International Journal of Intelligences Science, 2012, 2(4): 143-148.

[5] MORGAN M S, DEREK R. Gender Inference of Twitter Users in Non-English Contexts[C]// Proceedings of EMNLP. 2013: 1136-1145.

[6] GONCALVES C B, RATIKIEWICZ J, FLAMMINI A, et al. Predicting the political alignment of Twitter user[C]// Proceedings of the International Conference on Social Computing. 2011.

[7] LIU, RUTHS D. What's in a name? Using first names as features for gender inference in Twitter[C]// Analyzing Microtext: 2013 AAAI Spring Symposium. 2013.

[8] EICHSTAEDT M C, KERN L, et al. Developing Age and Gender Predictive Lexica over Social Media[C]// Proceedings of EMNLP. 2014: 1146-1151.

[9] FARNADI M G, VASUDEVAN G, DAVALOS S, et al. Age and gender identification in social media[C]// Proceedings of CLEF 2014 Evaluation Labs pages. 2014: 1129-1136.

[10] HOCHREITER, JURGEN S. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[11] GRAVES A. Generating Sequences With Recurrent Neural Networks[J]. arXiv preprint arXiv: 1308.0850, 2013.

[12] ANTOINE X B, YOSHUA B. Deep Sparse Rectifier Neural Networks[C]// Proceedings of AISTATS. 2011: 315-323.

[13] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors[J]. Computer Science, 2012, 3(4): 212-223.