中国科学: 信息科学 2017 年 第 47 卷 第 8 期: 1095–1108
SCIENTIA SINICA Informationis
从大数据到大知识工程专刊 · 论文

混合分类/回归模型的用户年龄识别方法

陈敬, 李寿山*, 王晶晶, 周国栋

苏州大学计算机科学与技术学院, 苏州 215006
* 通信作者. E-mail: lishoushan@suda.edu.cn

收稿日期: 2017–03–30; 接受日期: 2017–06–08; 网络出版日期: 2017–08–16
国家自然科学基金 (批准号: 61331011, 61375073, 61672366) 资助项目

摘要 年龄分类方法和年龄回归方法是年龄识别任务的主要方法. 这两种方法各自具有其自身的优越性. 例如: 年龄分类方法能够灵活利用机器学习中的区分式模型, 而年龄回归方法的主要优势是能够捕捉不同年龄之间的联系. 为了能同时利用分类方法和回归方法的优势, 本文提出了一种混合分类/回归模型 (混合模型) 用于用户年龄识别. 具体而言, 我们首先基于长短时记忆 (long short-term memory, LSTM) 模型分别构建年龄分类模型和回归模型用于年龄识别; 在此基础上, 将年龄分类结果与年龄回归结果进行线性融合作为年龄识别的最终结果. 实验结果表明本文提出的混合模型能够有效提升年龄识别任务的性能.

关键词 年龄分类, 年龄回归, 混合模型, 长短时记忆, 自然语言处理

1 引言

随着社交网络的迅猛发展, 各种类型的社交网站相继产生, 如 Twitter、Facebook、新浪微博及腾讯微博等. 该类网站中用户产生的海量数据为社交网络分析提供了丰富的信息支持. 在众多的社交网络分析任务中, 用户年龄属性识别是其中一项基本任务. 该任务旨在通过用户在网上发布的文本信息或者其社交信息来确定该用户准确的年龄或者所属的年龄段. 图 1 给出了某社交媒体网站中的一个用户信息. 从图 1 中可以看出, 当在用户属性信息中获取不到关于用户年龄的信息时, 我们可以从他发布的文本信息 “我 23 岁了哎 (I’m 23)” 准确推断出该用户的年龄是 23. 年龄识别在很多社交应用中已经成为不可或缺的工具, 例如, 智能营销 [1]、在线广告 [2] 以及个性化分析 [3] 等.

基于全监督的机器学习方法是目前年龄识别研究的主流方法, 即从已标注样本数据中挖掘有效特征训练用户年龄识别模型, 例如, 抽取图 1 中的用户发布的文本信息特征训练年龄识别模型. 年龄分类和年龄回归是年龄识别的两个基本任务, 不同于年龄分类将用户分类到几个年龄组中 [4,5], 年龄回归利用连续变量预测用户年龄, 该连续变量代表一个确切的年龄数 [6,7]. 到目前为止, 大多数研究主要

引用格式: 陈敬, 李寿山, 王晶晶, 等. 混合分类/回归模型的用户年龄识别方法. 中国科学: 信息科学, 2017, 47: 1095–1108, doi: 10.1360/N112016-00278
Chen J, Li S S, Wang J J, et al. User age prediction by combining classification and regression (in Chinese). Sci Sin Inform, 2017, 47: 1095–1108, doi: 10.1360/N112016-00278

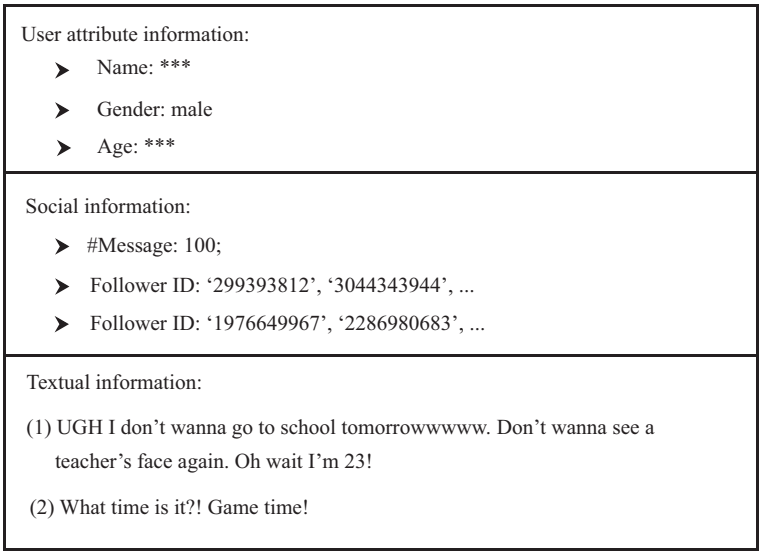

图 1　社交媒体中一个用户实例

Figure 1　A user example in a social media

将年龄识别任务建模为年龄分类任务进行研究, 还有少数研究将年龄识别任务建模为年龄回归任务进行研究. 这两个模型各自具有其鲜明的特点与优势, 例如, 年龄分类模型能够灵活利用机器学习中的区分式模型, 从而可以获得更好的年龄分类效果; 而年龄回归模型能够捕捉不同年龄之间的联系, 同时对年龄分类标签有更好的建模.

为了能够同时利用年龄分类模型和年龄回归模型的优势, 本文提出了一种混合分类/回归模型 (混合模型) 用于用户年龄识别任务, 即将年龄分类器与年龄回归器进行混合. 具体而言, 将基于同一特征信息训练得到的年龄分类器的分类结果与年龄回归器的回归结果进行线性融合 (年龄分类器和回归器所占的权重可调) 作为年龄识别的最终结果. 实验结果表明, 本文提出的混合模型能够有效提升年龄识别性能.

此外, 已有的年龄识别方法, 不管是分类方法还是回归方法, 基本都是基于浅层机器学习方法实现. 然而, 近几年的自然语言处理 (NLP) 研究已经表明基于深度学习的机器学习方法能够在很多应用, 如情感分类 [8]、机器翻译 [9] 等, 获得比浅层机器学习方法更好的性能. 因此, 本文将探索利用深度学习方法, 即长短时记忆 (long short-term memory, LSTM) 神经网络进行年龄分类或回归实现, 充分利用 LSTM 善于学习输入值长相关联系的特性, 从而提升用户年龄识别性能. 实验结果表明, 本文提出的混合基于 LSTM 的分类和回归模型能够有效提升年龄识别性能.

本文的其余部分安排如下: 第 2 节介绍年龄分类和年龄回归的相关研究工作; 第 3 节介绍相关背景知识; 第 4 节介绍年龄识别方法; 第 5 节给出实验结果与分析; 最后一节给出本文的结论和下一步工作介绍.

2　相关工作

2.1　基于分类模型的用户年龄识别方法研究

在过去十年中, 大部分研究将年龄预测看作是分类问题, 在博客领域和社交媒体领域取得了一定

的成果.

在博客领域, Schler 等 [4] 从博客文本中提取文本特征 (如词的上下文特征) 和写作风格特征 (如词性特征) 进行年龄分类. Burger 和 Henderson [5] 探究与博主年龄相关的社交特征, 如发博时间信息、地理位置信息、博客链接信息与图片信息、朋友信息以及用户兴趣信息, 进一步提高年龄分类的性能. Ikeda 等 [10] 利用文本特征训练多个子分类器进行半监督年龄预测研究. Rosenthal 和 McKeown [11] 同时利用文本特征和社交特征进行年龄分类研究.

在社交媒体领域, Mackinnon 和 Warren [12] 探究了多种特征, 如用户间的关联信息, 通过社交网络中朋友提供的信息来预测用户的年龄及居住地. Peersman 等 [13] 基于文本特征利用文本分类方法进行年龄分类研究. 最近, Marquardt 等 [14] 基于文本特征和写作风格特征提出了多标签分类方法预测用户的性别及年龄.

2.2 基于回归模型的用户年龄识别方法研究

相比年龄分类, 年龄回归的相关研究工作较少.

Nguyen 等 [6] 探究词的 Unigram 特征、词性的 Unigram 和 Bigram 特征以及性别特征等文本特征, 并且利用线性回归模型进行年龄回归实验探究. 他们的实验研究发现, 文本特征以及写作风格特征是预测用户年龄的强有力的特征, 甚至词的 Unigram 特征就已经能获得比较理想的年龄识别性能, 同时词性特征对于预测年龄较大的用户优势明显.

Nguyen 等 [7] 通过 3 种不同的方式, 分别是预测年龄类别、生命的不同阶段以及用户的实际年龄, 并利用线性回归模型进一步探究了 Twitter 上用户的年龄预测. 他们发现自动系统可以获得比人为识别更好的年龄预测性能, 并且自动系统识别用户年龄的速度明显优于人为识别.

Chen 等 [15] 探究文本特征和社交特征, 通过支持向量机回归器 (SVR) 进行年龄回归试验. 他们利用主动学习方法减少用户年龄预测中语料的标注代价, 回归问题中的置信度问题通过随机特征子空间方法进行解决.

与以上研究不同的是, 本文引入深度学习方法, 并混合年龄分类/回归模型进行用户年龄识别方法研究.

3 背景工作

3.1 语料收集与概述

本文实验数据来自新浪微博[1)]. 我们抓取的用户数据包含用户的个人属性信息 (如姓名、年龄、性别及认证类型) 以及用户发表的微博信息. 数据采集过程从一个随机的用户开始, 之后反复抓取所选用户的关注者和粉丝的数据. 我们过滤了认证类型为组织机构的用户, 因为这些用户的年龄属性没有意义. 此外虽然用户发布的微博文本是用来预测用户年龄的重要因素, 但是存在部分用户只发表过极少的微博. 为了保证数据的可靠性, 我们删除了发布少于 50 条微博的用户. 最终得到了规模约为 12000 的用户微博数据集.

图 2 显示了不同年龄的用户分布. 从图中可以看到, 用户年龄的分布情况是极其不平衡的. 其中, 绝大多数的用户为年龄处于 19 至 28 岁的年轻人.

1) http://weibo.com/.

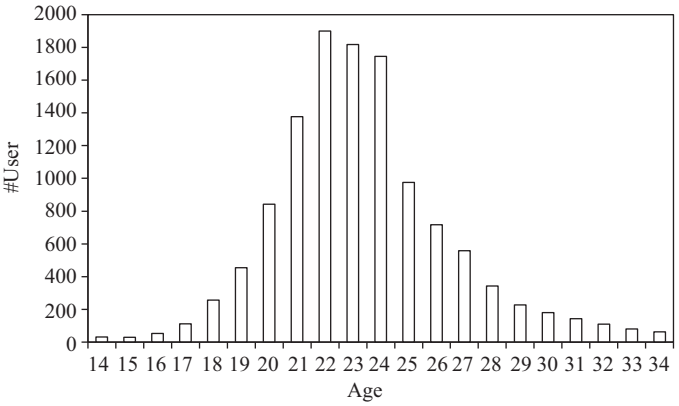

图 2　不同年龄的用户分布

Figure 2　User distribution in different ages

表 1　年龄识别中所用的文本特征和社交特征

Table 1　Textual and social features in age prediction

	Feature	Remarks	Examples
Textual features	BOW	Word unigrams in the user-generated messages	don't, wanna, ⋯
	POS patterns	Top trigrams of the POS tag in user-generated messages	DT_SP_PU, PU_VV_VV, ⋯
Social features	Statistics	# of Messages, # of Comments , # of Followers, # of Followings	78, 128, 189, 108
	Time	Probability distribution of the user posts messages over 24 hours (00–23)	[0.1, 0.1, 0, ..., 0.2]
	Follower list	All IDs of the followers	2806602710, 3807796763, ⋯
	Following list	All IDs of the followings	3841580965, 3843099598, ⋯

3.2　文本与社交特征

每个用户均被表示为一个特征向量: $x_i \in \mathbb{R}^d$, 作为年龄识别模型的输入. 在已有文献中, 多种形式的特征如词的 Unigram 特征及社交特征已被成功应用于年龄识别 [6]. 本文将特征分为两组: 文本特征和社交特征. 前者包含用户发表微博文本词的 Unigram 特征和 POS (part of speech) 词性序列特征, 后者主要包括用户的在线社交行为信息, 如发博时间、关注者及粉丝等. 表 1 详细描述了这两类特征的所有特征.

首先, 对于文本特征 (textual features), 文本的词特征可以反映出用户在新浪微博上所关注的话题, 为我们识别用户年龄提供了有效信息. 其中, 词袋 (bag of words, BOW) 特征是年龄识别任务中用的最多的一类特征, 并且其有效性也已经被多项研究验证. 同时, POS 词性序列特征 (POS patterns) 也是文本特征中较为流行的一类特征, 此类特征可以获取用户写作风格信息, 同样可以为准确识别用户年龄提供帮助.

其次, 对于社交特征 (social features), 我们划分为 4 类特征: 统计信息 (statistics)、发博时间 (time)、粉丝列表 (follower list) 及关注者列表 (following list). 其中, 统计信息特征包含用户的微博数、

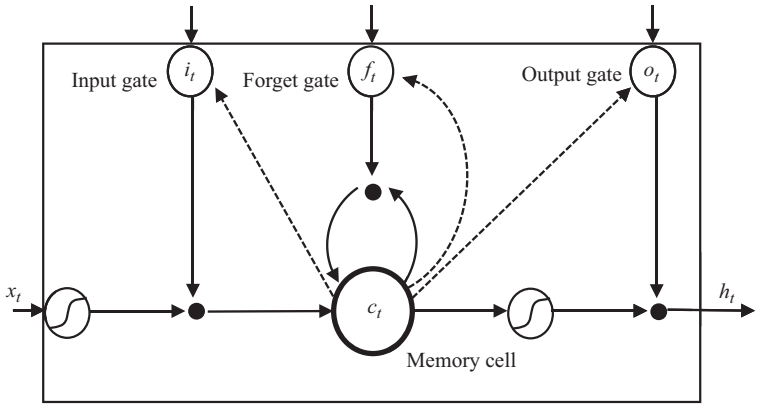

图 3 长短时记忆神经网络模型

Figure 3 A long short-term memory unit

评论数、粉丝数及关注者数等信息, 该类特征主要用于获取用户的在线行为特征; 发博时间特征是用户各时间发表的微博比例, 用以反映用户发布微博的习惯. 例如, 年龄处于 20 岁至 24 岁之间的用户更有可能在晚上甚至凌晨发布微博. 相比之下, 年龄较大的用户则更习惯早晨或上午等上班时间发微博; 用户的关注者列表和粉丝列表可以反映出用户感兴趣的人群, 例如年轻的用户可能喜欢打游戏, 关注明星和时尚; 而年龄较大的用户会对投资理财、养生比较感兴趣, 因此该类信息也可以为识别用户年龄提供帮助.

最后, 我们将文本特征与社交特征进行特征融合, 具体实现是将社交特征和文本特征进行叠加组成联合特征 (joint features).

4 年龄识别方法

长短时记忆神经网络是由 Hochreiter 和 Schmidhuber 提出的 [16], 近期 Alex Graves 对 LSTM 进行了改良和推广 [17], 使得 LSTM 在很多问题都取得相当巨大的成功, 并得到了广泛的使用. LSTM 使用记忆单元 (memory cell), 避免梯度在反向传播中遇到爆炸和衰减问题, 与此同时 LSTM 可以学习长期依赖关系即建立输入值之间的长相关联系.

如图 3 所示, LSTM 由输入门 i、输出门 o、遗忘门 f 和记忆单元 c 组成, 3 个门元素的取值在 $[0, 1]$ 之间, 其中输入门、输出门和遗忘门是控制记忆单元的读、写和丢失操作的控制器, 利用形式化语言, t 时刻 LSTM 的更新方式如下:

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

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1}), \tag{2}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_{t-1}), \tag{3}$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1}), \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{5}$$

$$h_t = o_t \odot \tanh(c_t), \tag{6}$$

其中, x_t 表示当前 t 时刻的输入, σ 是激活函数; $\odot$ 是点乘运算; W_*, U_* 和 V_* 表示系数矩阵; i_t 控制每个内存单元加入多少新的信息, f_t 控制每一个内存单元需要遗忘多少信息, o_t 控制每一个内存单元输出多少信息; c_t 表示 t 时刻记忆单元的计算方法, h_t 为 t 时刻 LSTM 单元的输出信息. 由图 3 可以看出输入门、输出门和遗忘门的输出分别连接到一个乘法单元上, 从而分别控制网络的输入、输出以及记忆单元状态.

4.1 基于 LSTM 的年龄分类方法

如图 4 所示, 基于 LSTM 的年龄分类模型包含一个 LSTM 层, 模型的输入既可以是社交特征表示也可以是文本特征表示. 根据以上 LSTM 的更新方式, 输入传播到 LSTM 层后返回高维度的向量. 随后向量传播到全连接层, 全连接层类似于传统多层感知机中的隐藏层, 接收来自上一层的输出, 通过常用的激活函数加权并传播到 Dropout 层. Dropout 已经成功地运用在前馈网络中 [18], 在训练过程中, 通过阻止网络中特征检测器的作用, 获得更少相互依赖的网络单元, 进而实现更好的性能.

本文采用 “Softmax” 激活函数输出年龄分类结果, 激活函数如式 (7) 所示, 输出后验概率最大的预测标签:

$$p = \text{softmax}(W^d h^d + b^d), \tag{7}$$

其中, p 表示年龄分类预测概率的集合, W^d 表示需要学习的权重向量, b^d 表示偏置.

针对年龄分类问题, 我们的训练目标是最小化交叉熵误差, 损失函数定义如式 (8) 所示:

$$\text{loss}_\text{C} = -\frac{1}{m} \cdot \sum_{i=1}^{m} \sum_{j=1}^{l} y_{ij} \cdot \log p_{ij}, \tag{8}$$

其中, loss_C 表示年龄分类的损失函数, m 表示样本的总数目, l 表示年龄类别数, y_{ij} 表示第 i 个样例是否属于第 j 个年龄类别, p_{ij} 指的是预测概率.

4.2 基于 LSTM 的年龄回归方法

图 5 介绍了年龄回归的模型框架, 与年龄分类模型框架相比较, 年龄回归模型中大部分的学习层包括 LSTM 层、全连接层、Dropout 层与年龄分类模型基本一致. 只是在最后输出结果时, 不同于分类任务使用 “Softmax” 激活函数, 回归任务采用 “Linear” 激活函数输出年龄回归结果, 激活函数如式 (9) 所示:

$$f = W^d h^d + b^d, \tag{9}$$

其中, W^d 与 b^d 与年龄分类交叉熵公式意义一致, 分别代表权重向量与偏置, f 表示预测的年龄值, 该年龄值是一个连续值.

针对回归问题, 我们选择常用的 “Mean squared error” 作为损失函数, 具体损失函数定义如下:

$$\text{loss}_\text{R} = \frac{1}{2m} \cdot \sum_{i=1}^{m} ||f_i - y_i||^2, \tag{10}$$

其中, loss_R 表示年龄回归的损失函数, y_i 表示第 i 个样本的真实标签, f_i 表示第 i 个样本的年龄预测值, m 表示训练样本的总数目. 无论是分类问题还是回归问题, 在训练模型的时候使用基于梯度下降的方法以及反向传播算法 [19] 来学习模型参数.

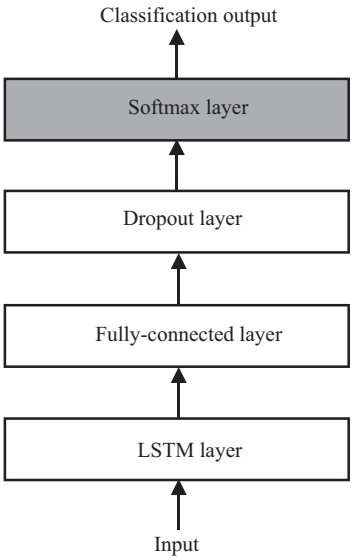

图 4 基于 LSTM 的年龄分类

Figure 4 Age classification with LSTM

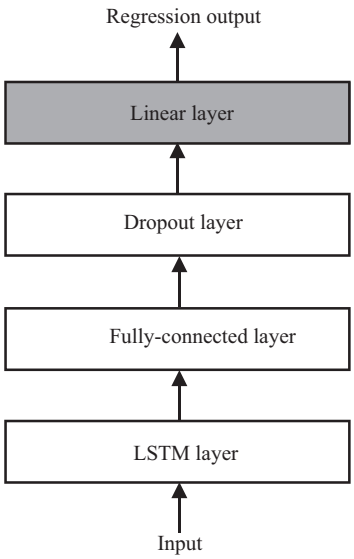

图 5 基于 LSTM 的年龄回归

Figure 5 Age regression with LSTM

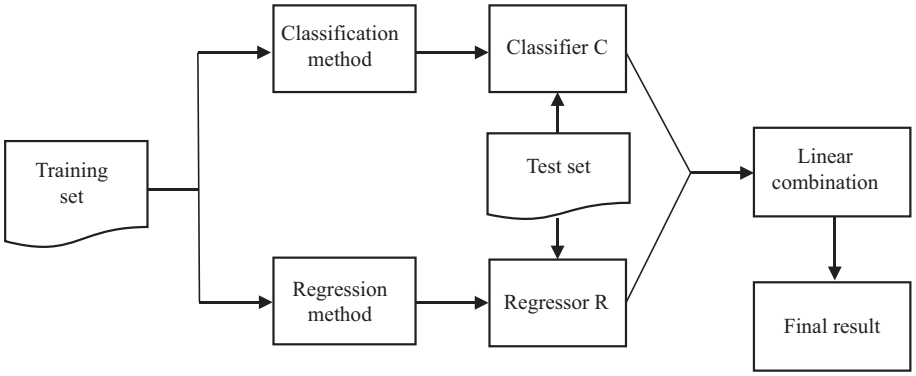

图 6 基于混合模型的年龄识别方法框架

Figure 6 The framework of age prediction with hybrid model

4.3 混合分类/回归模型的年龄识别方法

基于混合模型的年龄识别方法利用年龄分类器的分类结果与年龄回归器的回归结果进行线性融合作为年龄识别的最终结果, 并且年龄分类和年龄回归所用的特征信息一致.

图 6 给出了基于混合模型的年龄识别方法的系统框架图, 从图中可以看出, 本文提出的基于混合模型的年龄识别方法主要分为两步:

- **训练分类器与回归器:** 基于同一特征, 采用不同年龄识别方法, 即分类方法和回归方法产生年龄分类器和年龄回归器.
- **混合分类器与回归器结果:** 本文采用加权求和规则来线性融合年龄分类器与年龄回归器的输出结

表 2　平衡数据集与不平衡数据集中每一个年龄类别的用户规模

Table 2　Numbers of users in each age in the balanced and imbalanced data sets

Age	19	20	21	22	23	24	25	26	27	28	ALL
Balanced data set	200	200	200	200	200	200	200	200	200	200	2000
Imbalanced data set	305	469	603	821	1491	1550	1611	1163	677	360	9050

表 3　LSTM 中的参数设置

Table 3　Parameter setting in LSTM

Parameter description	Value
Size of total unigram features in balanced data set	30000
Size of total unigram features in imbalanced data set	80000
Dimension of the LSTM layer output	128
Dimension of the full-connected layer output	64
Dropout probability	0.25
Epochs of iteration	15

果, 具体实现如下所示:

$$y_{\text{last}} = \lambda \cdot y + (1 - \lambda) \cdot \text{label}_{\text{pred}}, \tag{11}$$

其中, y 表示年龄回归器的输出值, $\text{label}_{\text{pred}}$ 为年龄分类器的输出值, 计算如式 (11) 所示, y_{last} 为混合模型下年龄识别方法的输出值. λ 为回归器所占的权重, 通过训练样本中的十倍交叉实验结果选取最佳权重 (λ 值) 为 0.8.

5　实验设计与分析

5.1　实验设置

数据设置: 数据收集在第 3 节已经详细介绍. 本文选择 19 至 28 年龄段的用户作为实验数据, 并分别使用平衡数据集和不平衡数据集进行两组实验. 每一个年龄类别中所使用的用户数如表 2 所示. 具体实验中, 我们选取 70% 的数据作为训练集, 10% 的数据作为验证集, 其余 20% 作为测试集.

特征选择: 实验所用特征如表 1 所示, 联合特征指文本特征与社交特征的融合. 实验所用特征为词袋特征 (BOW).

参数设置: LSTM 参数设置如表 3 所示.

评价准则: 实验中采用确定性系数 R^2 作为年龄识别性能的评价标准, R^2 也被称之为拟合优度, 表示自变量对因变量的解释程度. R^2 的值处于 0∼1 之间, 越接近 1 代表训练模型的预测值与实际观测值拟合程度越高 [20].

5.2　实验结果

本实验实现了以下几种年龄分类方法和回归方法:

表 4 使用不同特征时的 SVR 年龄识别结果 (平衡数据集)

Table 4 SVR performances with different kinds of features (balanced data set)

	Feature	# of features	R^2
Textual features	BOW	274461	0.382
	POS patterns	13153	0.084
	ALL	287614	0.376
Social features	Statistics	4	0.030
	Time	24	0.050
	Follower list	395066	0.198
	Following list	238327	0.308
	ALL	633121	0.351
Joint features	Textual+Social	920735	**0.489**

- **年龄分类方法**

  SVM: 使用 libSVM 工具包[2)] 提供的 SVM 分类算法实现年龄分类, 所有参数设置为默认值.

  ME: 使用 Mallet 工具包[3)]提供的最大熵分类方法实现年龄分类, 所有参数设置为默认值.

  C_CNN: 利用卷积神经网络进行年龄分类, Bow_CNN 模型由 Johnson 提出 [21].

  C_LSTM: 利用 4.1 小节介绍的 LSTM 分类模型实现年龄分类, 参数设置如表 3 所示.

- **年龄回归方法**

  SVR: 采用 libSVM 工具包提供的 SVM 回归算法及线性核函数, 所有参数设置为默认值.

  MLP: 采用全连接层, 激励层以及 Dropout 层实现多层感知机模型实现年龄回归方法.

  R_CNN: 模型结构与 C_CNN 类似, 唯一不同的是最后的输出层是 Linear 层.

  R_LSTM: 利用 4.2 小节介绍的 LSTM 回归模型实现年龄回归, 参数设置如表 3 所示.

5.2.1 基于平衡数据集的实验结果

表 4 给出了当使用平衡数据集时基于不同特征的 SVR 年龄识别结果. 从表中结果可以看出, 在文本特征中, BOW 特征比较有效. 结合词性特征的文本特征并不能提升用户年龄识别性能. 在社交特征中, 关注者特征最为有效. 结合所有社交特征能够获得比仅仅使用关注者特征更好的年龄识别性能. 使用所有文本特征, SVR 方法获得了 0.376 的识别结果. 该结果优于使用所有社交特征的结果 (0.351). 当使用联合特征时, 即同时使用文本特征与社交特征时, 用户年龄回归效果最佳, 达到 0.489.

表 5 给出了在使用平衡数据集和不同特征集合时不同年龄分类方法的实验结果. 从表中结果可以看出, 不管使用何种特征, ME 分类方法明显优于 SVM 分类方法; 当使用文本特征时, C_CNN 分类方法明显好于 ME 分类方法; 当使用社交特征或联合特征时, ME 分类方法优于 C_CNN 分类方法. 在众多方法中, C_LSTM 分类方法表现最佳.

表 6 给出了在使用平衡数据集和不同特征集合时不同年龄回归方法的实验结果. 从表中结果可以看出, 不管使用何种特征, MLP 回归方法优于 SVR 回归方法; 当使用文本特征时, R_CNN 回归方法优于 SVR 回归方法; 然而, 当使用社交特征或联合特征时, MLP, SVR 回归方法好于 R_CNN 回归方

2) http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.

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

表 5 使用不同特征和分类方法的年龄识别结果 (平衡数据集)

Table 5 Performances with different kinds of features and classification approaches (balanced data set)

Classification method	Textual features	Social features	Joint features
SVM	0.251	0.250	0.334
ME	0.285	0.304	0.389
C_CNN	0.323	0.298	0.375
C_LSTM	**0.359**	**0.330**	**0.421**

表 6 使用不同特征和回归方法的年龄识别结果 (平衡数据集)

Table 6 Performances with different kinds of features and regression approaches (balanced data set)

Regression method	Textual features	Social features	Joint features
SVR	0.376	0.351	0.489
MLP	0.429	0.383	0.499
R_CNN	0.383	0.312	0.475
R_LSTM	**0.454**	**0.405**	**0.535**

表 7 使用不同特征和混合方法的年龄识别结果 (平衡数据集)

Table 7 Performances with different kinds of features and a hybrid approach (balanced data set)

Age identification method	Textual features	Social features	Joint features
C_LSTM	0.359	0.330	0.421
R_LSTM	0.454	0.405	0.535
C_LSTM+R_LSTM	**0.466**	**0.429**	**0.552**

法; 在众多方法中, R_LSTM 回归方法表现最佳. 以上结果验证了 LSTM 实现年龄分类与回归的有效性. 后续的模型混合实验主要是混合基于 C_LSTM 分类方法和 R_LSTM 回归方法的年龄识别结果.

表 7 给出了使用平衡数据集和不同特征集合时, 混合 C_LSTM 分类方法和 R_LSTM 回归方法的年龄识别结果. 表格中 C_LSTM+R_LSTM 代表混合模型. 需要注意的是, 基于 LSTM 的分类模型和回归模型是完全独立的两个网络, 没有参数共享. 从该表结果中可以看出: 使用文本特征时, C_LSTM+R_LSTM 混合方法比 R_LSTM 回归方法和 C_LSTM 分类方法分别提高了约 1.2% 和 10.7%; 使用社交特征时, C_LSTM+R_LSTM 混合方法比 R_LSTM 回归方法和 C_LSTM 分类方法分别提高了约 2.4% 和 9.9%; 使用联合特征时, C_LSTM+R_LSTM 混合方法比 R_LSTM 回归方法和 C_LSTM 分类方法分别提高了约 1.7% 和 13.1%. 综合而言, 在使用平衡数据的情况下, 混合基于 LSTM 的分类和回归方法能够有效提升年龄识别性能.

5.2.2 基于不平衡数据集的实验结果

表 8 给出了当使用不平衡数据集时基于不同特征的 SVR 年龄识别结果. 从表中结果可以看出, 在文本特征中, BOW 特征比较有效. 结合词性特征的文本特征并不能提升年龄识别性能. 在社交特征中, 关注者特征最为有效. 结合所有社交特征能够获得比仅仅使用关注者特征更好的识别性能. 这些结果同基于平衡数据集的实验结果基本一致. 然而, 同基于平衡数据集的实验结果不同的是, 使用所有文本特征 SVR 方法获得结果 (0.309) 弱于使用所有社交特征的结果 (0.363). 当使用联合特征时, 即

表 8 使用不同特征时的 SVR 年龄识别结果 (不平衡数据集)

Table 8 SVR performances with different kinds of features (imbalanced data set)

	Feature	# of features	R^2
Textual features	BOW	920440	0.317
	POS patterns	74652	0.065
	ALL	995092	0.309
Social features	Statistics	4	0.015
	Time	24	0.024
	Follower list	1861859	0.241
	Following list	1029664	0.353
	ALL	2891551	0.363
Joint features	Textual+Social	3886643	**0.442**

表 9 使用不同特征和分类方法的年龄识别结果 (不平衡数据集)

Table 9 Performances with different kinds of features and classification approaches (imbalanced data set)

Classification method	Textual features	Social features	Joint features
SVM	0.166	0.148	0.188
ME	0.254	0.263	0.295
C_CNN	0.215	0.276	0.301
C_LSTM	**0.265**	**0.349**	**0.353**

同时使用文本特征与社交特征时, 用户年龄回归效果最佳, 达到 0.442.

表 9 给出了在使用不平衡数据集和不同特征集合时不同年龄分类方法的实验结果. 从表中结果可以看出, 不管使用何种特征, ME 分类方法明显优于 SVM 分类方法; 当使用文本特征时, C_CNN 分类方法明显弱于 ME 分类方法; 当使用社交特征或联合特征时, C_CNN 分类方法优于 ME 分类方法. 在众多方法中, C_LSTM 分类方法表现最佳.

表 10 给出了在使用不平衡数据集和不同特征集合时不同年龄回归方法的实验结果. 从表中结果可以看出, 当使用文本特征时, MLP 回归方法明显优于 SVR 回归方法; 然而, 当使用社交特征或联合特征时, MLP 回归方法弱于 SVR 回归方法. 当使用文本特征时, R_CNN 回归方法优于 SVR 回归方法, 但是弱于 MLP 回归方法; 当使用社交特征和联合特征时, MLP 和 SVR 回归方法优于 R_CNN 回归方法. 在众多方法中, R_LSTM 回归方法表现最佳. 以上结果验证了 LSTM 实现年龄分类与回归的有效性. 后续的模型混合实验主要是混合基于 C_LSTM 分类方法和 R_LSTM 回归方法的年龄识别结果.

表 11 给出了使用不平衡数据集和不同特征集合时, 混合 C_LSTM 分类方法和 R_LSTM 回归方法的年龄识别结果. 需要注意的是, 基于 LSTM 的分类模型和回归模型是完全独立的两个网络, 没有参数共享. 从该表结果中可以看出: 使用文本特征时, C_LSTM+R_LSTM 混合方法比 R_LSTM 回归方法和 C_LSTM 分类方法分别提高了约 1.2% 和 12.9%; 使用社交特征时, C_LSTM+R_LSTM 混合方法比 R_LSTM 回归方法和 C_LSTM 分类方法分别提高了约 2.7% 和 7.6%; 使用联合特征时, C_LSTM+R_LSTM 混合方法比 R_LSTM 回归方法和 C_LSTM 分类方法分别提高了约 1.3% 和 13.4%. 综合而言, 在使用不平衡数据的情况下, 混合基于 LSTM 的分类和回归方法能够有效提升年龄识别性能.

表 10 使用不同特征和回归方法的年龄识别结果 (不平衡数据集)

Table 10 Performances with different kinds of features and regression approaches (imbalanced data set)

Regression method	Textual features	Social features	Joint features
SVR	0.309	0.363	0.442
MLP	0.361	0.358	0.435
R_CNN	0.314	0.336	0.417
R_LSTM	**0.382**	**0.398**	**0.474**

表 11 使用不同特征和混合方法的年龄识别结果 (不平衡数据集)

Table 11 Performances with different kinds of features and a hybrid approach (imbalanced data set)

Age identification method	Textual features	Social features	Joint features
C_LSTM	0.265	0.349	0.353
R_LSTM	0.382	0.398	0.474
C_LSTM+R_LSTM	**0.394**	**0.425**	**0.487**

5.3 LSTM 模型参数分析

LSTM 层的输出维度、全连接层的输出维度和 Dropout 概率是 LSTM 模型中几个重要参数, 下面将分别分析这 3 个参数对于 R_LSTM 回归方法性能的影响.

首先, 考察 LSTM 层的输出维度对于 R_LSTM 回归方法性能的影响. 保持其他参数不变, 将 LSTM 层的输出维度设置为 128, 256 和 512 进行年龄回归实验. 实验结果表明, LSTM 层的输出维度对于 R_LSTM 回归方法性能的影响较小, 不同参数设置的 R^2 结果之间的差距小于 0.003.

其次, 考察全连接层的输出维度对于 R_LSTM 回归方法性能的影响. 保持其他参数不变, 将全连接层的输出维度设置为 16, 32 和 64 进行年龄回归实验. 实验结果表明, 全连接层的输出维度对 R_LSTM 回归方法性能的影响较小, 不同参数设置的 R^2 结果之间的差距小于 0.005.

最后, 考察 Dropout 概率对于 R_LSTM 回归方法性能的影响. 保持其他参数不变, 将 Dropout 概率设置为 0.15, 0.2, 0.25, 0.3, 0.35, 0.4 进行年龄回归实验. 实验结果表明, Dropout 概率在 0.2 到 0.3 范围内能够取得比较稳定的结果. 在此范围内, 不同参数设置的 R^2 结果之间的差距小于 0.005.

6 本文结论和下一步工作介绍

本文提出了一种基于混合模型的用户年龄识别方法. 该方法将年龄分类结果与年龄回归结果进行线性融合作为最终的年龄识别结果. 此外, 我们采用了 3 种不同的特征, 即文本特征、社交特征以及联合特征探究基于 LSTM 模型的年龄识别性能. 实验结果表明基于 LSTM 模型的深度学习年龄分类或回归方法能够获得比传统的浅层学习方法及另外一种基于 CNN 的深度学习方法更好的性能. 此外, 实验结果表明基于混合模型的年龄识别方法与混合之前的年龄分类方法和年龄回归方法相比均取得更好的性能, 这充分说明了混合模型对于年龄识别任务的有效性.

下一步工作中, 可以从以下几个方面扩展本文的研究. 首先, 将探索更多的特征, 如用户发表文本的主题特征, 用于提高年龄识别任务的性能. 其次, 可以尝试提出更好的融合方法. 例如, 可以在分类模型和回归模型的中间表示层进行融合, 而不是在最后的结果进行融合, 这样融合可以学习更多的信息. 最后, 尝试将本文方法应用到其他可能同时使用分类模型和回归模型的自然语言处理或社交网络

分析任务中, 例如在情感分析研究中, 分类模型和回归模型都可以应用于评论评分任务中 [22, 23].

参考文献

1 Preotiuc-Pietro D, Lampos V, Aletras N. An analysis of the user occupational class through twitter content. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Pennsylvania: Association for Computational Linguistics, 2015. 1754–1764

2 Volkova S, Wilson T, Yarowsky D. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Pennsylvania: Association for Computational Linguistics, 2013. 1815–1827

3 O'Connor B, Balasubramanyan R, Routledge B R, et al. From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the 4th International Conference on Weblogs and Social Media. California: AAAI Press, 2010. 1842–1850

4 Schler J, Koppel M, Argamon S, et al. Effects of age and gender on blogging. Front Inform Tech Electron Eng, 2006, 274: 199–205

5 Burger J D, Henderson J C. An exploration of observable features related to blogger age. In: Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs. California: AAAI Press, 2006. 15–20

6 Nguyen D, Smith N A, Rose C. Author age prediction from text using liner regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Pennsylvania: Association for Computational Linguistics, 2011. 115–123

7 Nguyen D, Gravel R, Trieschnigg D, et al. "How old do you think I am?": a study of language and age in twitter. In: Proceedings of the 7th International Conference on Weblogs and Social Media. California: AAAI Press, 2013. 439–448

8 Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Pennsylvania: Association for Computational Linguistics, 2016. 214–224

9 Barone A V M, Attardi G. Non-projective dependency-based pre-reordering with recurrent neural network for machine translation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Pennsylvania: Association for Computational Linguistics, 2015. 846–856

10 Ikeda D, Takamura H, Okumura M. Semi-supervised learning for blog classification. In: Proceedings of the 23rd AAAI Conference on Artificial intelligence. California: AAAI Press, 2008. 1156–1164

11 Rosenthal S, McKeown K. Age prediction in blogs: a study of style, content, and online behavior in pre- and post-social media generations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, 2011. 763–772

12 Mackinnon I, Warren R H. Statistical Network Analysis: Models, Issues, and New Directions. Berlin: Springer, 2006

13 Peersman C, Daelemans W, Vaerenbergh L V. Predicting age and gender in online social networks. In: Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. New York: ACM, 2011. 37–44

14 Marquardt J, Farnadi G, Vasudevan G, et al. Age and gender identification in social media. In: Proceedings of the 5th Conference and Labs of the Evaluation Forum (CLEF 2014), Sheffield, 2014. 1129–1136

15 Chen J, Li S S, Dai B, et al. Active learning for age regression in social media. In: Proceedings of China National Conference on Chinese Computational Linguistics. Berlin: Springer, 2016. 351–362

16 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780

17 Graves A. Generating sequences with recurrent neural networks. arXiv:1308.0850, 2013

18 Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors. Comput Sci, 2012, 3: 212–223

19 LeCun Y A, Bottou L, Orr G B, et al. Efficient backprop. Neur Net Tricks Trade, 2012, 1524: 9–50

20 Cameron A C, Windmeijer F A G. R-squared measures for count data regression models with applications to health-care utilization. J Bus Econ Stat, 1996, 14: 209–220

21 Johnson R, Zhang T. Effective use of word order for text categorization with convolutional neural networks. In:

Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics. Pennsylvania: Association for Computational Linguistics, 2015. 103–112

22 Agarwal B, Sharma V K, Mittal N. Sentiment classification of review documents using phrase patterns. In: Proceedings of International Conference on Advances in Computing, Communications and Informatics. New York: IEEE, 2013. 1577–1580

23 Elkouri A. Predicting the sentiment polarity and rating of yelp reviews. arXiv:1512.06303, 2015

User age prediction by combining classification and regression

Jing CHEN, Shoushan LI*, Jingjing WANG & Guodong ZHOU

School of Computer Science and Technology, Soochow University, Suzhou 215006*, China*
* Corresponding author. E-mail: lishoushan@suda.edu.cn

Abstract Age classification and age regression are two main approaches of age prediction, and both approaches have their respective advantages. For example, age classification can flexibly utilize distinguished model in machine learning while the main advantage of age regression is its ability to capture the relationship between different ages. In order to utilize advantages of age classification and age regression simultaneously, we propose a hybrid age prediction approach that combines classification and regression. First, we build the long short-term memory (LSTM) models of age regression and age classification respectively for age prediction. Then, we linearly combine the results of the age classifier and age regressor as the final result of age prediction. Empirical evaluations demonstrate that the proposed hybrid model effectively improves the performance.

Keywords age classification, age regression, hybrid model, long short-term memory, natural language processing

Jing CHEN was born in 1992. He is a graduate student at Soochow University, majoring in computer science and technology. His research interests include sentiment analysis, social computing, and natural language processing.

Shoushan LI was born in 1980. He received his Ph.D. from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, in 2008. He is currently a professor in the School of Computer Science and Technology, Soochow University. His research interests include sentiment analysis, social computing, and natural language processing.

Jingjing WANG was born in 1990. He received his M.S. degree from the School of Computer Science and Technology, Soochow University, Suzhou, in 2015. He is currently a Ph.D. student at Soochow University, majoring in computer science and technology. His research interests include sentiment analysis, social computing, and natural language processing.

Guodong ZHOU was born in 1967. He received his Ph.D. from the National University of Singapore, Singapore, in 1999. He is currently a professor at the School of Computer Science and Technology, Soochow University. His research interests include natural language understanding, information extraction, statistical machine translation, and machine learning.