ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 2,Mar. 2018,pp. 334－342
DOI:10.16337/j.1004-9037.2018.02.016

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于多任务融合模型的用户属性推断

赵　宇[1]　李佳艺[1]　王　莉[2]

(1 太原理工大学计算机科学与技术学院,晋中,030600;2 太原理工大学大数据学院,晋中,030600)

摘　要:传统的用户属性推断方法主要基于机器学习及统计学习,其推断方法忽略了用户的整体表征及任务之间的相关性。本文提出一种基于多任务融合模型的用户属性推断方法,利用 doc2vec 独特的结构特性,加入文档向量以实现用户整体表征,避免人工提取特征的局限性。为实现用户多属性推断任务,本文提出基于关联学习的多任务融合推断框架,即在分别识别用户多个属性基础上赋予单用户多属性表征,在增强用户整体表征能力的同时,建立多个属性间的关联关系,提高单任务学习的区分度;然后采用模型融合技术,完成属性间关联学习,提高学习准确率及模型泛化能力,同时使用尽可能少的模型进行融合,提高模型运行效率。经实验比对,本文在多个数据集上的实验结果较其他算法有一定优势。

关键词:自然语言处理;doc2vec;多任务融合

中图分类号:TP391　　**文献标志码**:A

Demographic Attributes Inference Based on Multi-task Ensemble Model

Zhao Yu[1], Li Jiayi[1], Wang Li[2]

(1. Department of Computer Science and Technology, Taiyuan University of Technology, Jinzhong, 030600, China;
2. College of Big Data, Taiyuan University of Technology, Jinzhong, 030600, China)

Abstract: Traditional user attribute inference method is mainly based on machine learning and statistical learning methods, and its inference method ignores the user's overall representation and the correlation between tasks. A user attribute inference method based on multitasking ensemble model is proposed, which uses doc2vec unique structural characteristics and adds document vector to achieve the overall representation of the user, thus avoiding the limitations of artificial features extraction. In order to realize the multi-attribute inference task, a multi-task ensemble framework based on association learning is proposed, which is to identify multiple attributes of a user individually and give the multi-attribute representation of a single user. It enhances the overall representation of user. The relationship between multiple attributes is established at the same time, so as to improve the distinguishing degree of single-task learning. Then, this paper uses the model ensemble technology to complete the inter-attribute learning, improves the accuracy of learning and model generalization ability, and uses as few models as possible to improve the model operation efficiency. Experimental comparison on several data sets shows some advantages over other algorithms.

Key words: natural language processing; doc2vec; multitasking ensemble

---

基金项目:国家高技术研究发展计划("八六三"计划)(2014AA015204)资助项目;山西省自然科学基金(201703D421013)资助项目;中科院计算技术研究所网络数据科学重点实验课题(CASNDST20140X)资助项目。
收稿日期:2017-11-13;**修订日期**:2017-12-15

引　　言

互联网及移动技术的快速发展,不仅改变着人们的生活方式,同时也产生了海量数据资源。如何从繁杂无序的文本中挖掘出有价值的用户信息,已成为业界广泛关注的问题,因此用户属性推断应运而生。用户属性推断,旨在通过一段时间内用户的已有数据(如搜索信息、购物信息、地理位置以及移动通信等)推断用户属性,具体属性包括:性别、年龄和受教育程度等。对于企业而言,了解自身产品受众的属性有助于设计营销方案及广告宣传策略、实现产品定位及线上线下的市场推广;对刑侦类工作而言,掌握嫌疑人的基本属性有利于排除干扰因素,进一步缩小侦查范围。因此属性推断能力的强弱直接关系到相关应用的准确性。在用户属性推断研究方面,虽然近几年得到了产业界和学术界的关注,但仍然面临着巨大的挑战,主要表现在:(1)现有的用户属性推断方法大多是为用户提取基于经验所得的特征,却忽略了用户的整体表征对用户属性推断的贡献,这在一定程度上导致用户属性间的关联关系难以发挥作用,限制了用户属性推断的可参考范围;(2)大多数用户属性推断问题多归为分类问题,即将属性划分为不同的阈值,转化成二分类或多分类问题。但使用多分类器融合技术较单一分类方法而言具有较强的泛化能力。针对以上两点,本文提出基于多任务融合模型的用户属性推断方法。从语义及语法两个维度实现用户整体表征;并在单模型训练基础上建立多个属性间的关联关系,提高单任务学习的区分度;最后采用模型融合技术,完成属性间的关联性学习,进一步提高推断准确率,增强泛化能力。

用户属性推断正在由基于特征工程的推断方法向基于深度学习的推断方法发展。早期一些工作试图根据语言学中写作数据推断用户属性[1]。随后,互联网发展为用户属性推断提供了新的契机,基于用户浏览历史的统计学方法应运而生[2,3]。同时,在线社交网络与移动平台的建立为用户属性推断积累了空前规模的用户量,这使得推断用户属性的可能性和迫切性进一步提升。基于移动日志[4]、网站访问流量[5]及地理位置信息[6]的用户属性推断逐渐增多,成为连接用户离线和在线生活的桥梁,为用户属性推断提供依据。可见,在传统的基于特征工程的用户属性推断研究中,用户属性推断的好坏多依赖于经验所得,特征的针对性较强,而用户整体的表征能力一定程度上被忽略,这使得用户属性间的潜在联系难以发挥作用。

近几年兴起了深度学习方法[7-17]解决用户属性推断问题,在一定程度上降低了对经验特征的依赖关系。此类方法大多通过词嵌入方式进行词语语义学习,并使用连接或平均池化作用最终形成用户内容向量,再通过分类器完成属性分类。在文本分类、命名实体识别以及关系抽取等相关研究中也取得了显著的成果。王礼敏[8]等利用单通道长短时记忆模型(Long-short term memory,LSTM)分别学习得到微博文本及社交信息的特征表示,建立双通道 LSTM 模型学习两组特征之间的关系获得最终分类结果。戴斌[9]等将 LSTM 作为分类器,通过迭代将确定性高的视图文本及其对应的其他类型文本自动标注并加入标注样本中,同样取得了较好的性别分类结果。但 LSTM 通常对于序列任务有较大的优势,对于短文本及乱序而言,其分类效果并不理想,且上述特征选取只考虑了用户文本的语义信息,而忽略了其他角度特征提取(如关键词)对用户属性推断的重要作用,且用户属性间的关联关系并未得到很好的利用。

1　多任务融合推断模型

本文提出基于多任务融合模型的用户属性推断方法,如图 1 所示。本框架主要分为两个阶段,第一阶段,单模型单特征推断,根据用户数据采用基于文本语义(doc2vec_DBOW,doc2vec_DM)及基于文本语法的去冗余关键词(TF-IDF_MR)学习方式实现用户级向量表示,然后通过模型训练出每个用户的 M 种特征分布概率;第二阶段,基于关联学习模型的多任务融合推断,即将第一阶段每种表示方式所得结果与用户表征相结合,作为用户整体表征的补充完成模型训练,输出用户多个任务的属性矩阵,再将多

个单模型训练结果的多任务属性矩阵作为用户表征向量的附加条件，进行融合学习，最终得出用户多个属性取值。

本框架主要突出了用户的整体表征及属性关联度表征。通过基于文本语义及去冗余关键词计算的单模型单特征训练，完成用户整体表征；在第二阶段基于关联学习模型的多任务融合推断中充分利用任务间的关联性（如年龄及受教育程度的关联关系），为用户属性推断提供参考依据，从而进一步增强用户属性推断的准确性。

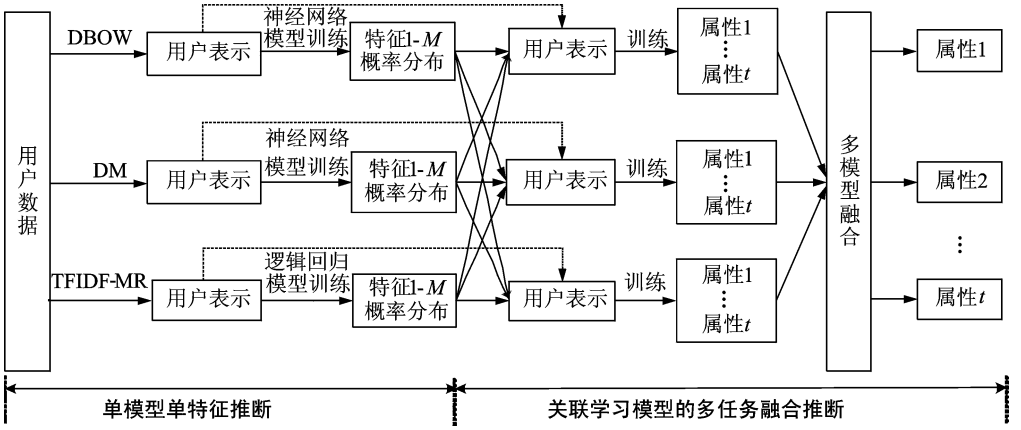

图 1 多任务融合推断模型框架

Fig. 1 Multi-task ensemble inference model framework

2 单模型单特征推断

本文的单模型单特征推断主要采用两类模型：基于 doc2vec 模型的单特征推断及基于 TF-IDF_MR 模型的单特征推断。采用 doc2vec 中的 DM 及 DBOW 从语义层面实现用户的整体表征，并使用改进后的去冗余关键词策略 TF-IDF_MR 算法从词频角度实现用户文本数据的重要性度量。

2.1 基于 doc2vec 模型的单特征推断

自然语言处理困难之处在于语义复杂、难以表征，通常需要将语言数学化，而向量化便是一种很好的方法。One-hot representation 是较为常见的词语表示方法之一，但该方法在受到维数灾难困扰的同时并不能很好地刻划词语之间的相似性。

Google 提出了一种开源的可对词语实现低维实数向量表示的工具包 word2vec，其使用的训练方法有两种：CBOW（Continuous bag-of-words）和 Skip-gram。该向量表示基于神经语言模型，通过对大规模数据的训练，将词的上下文信息表示在向量中，实现文本内容向量化，但其只针对词语语义表征。doc2ve 在 word2vec 词向量表征思想基础上，实现了分布式文档表征。作为一个处理可变长度文本的方法，该方法与 word2vec 的最大区别在于加入了一个新的与单词维度相等的向量作为句子表征、段落表征或者文档表征。故本文将 doc2vec 的文档表示应用于用户级别，依据用户文档，对每个用户进行很好的语义总结，提取出反映用户属性的用户文档向量。

在 doc2vec 中有两种模型可生成用户向量：DM（Distributed memory model）和 DBOW（Distributed bag of words）。图 2 为 DM 训练过程。其中 $\boldsymbol{D}$ 为用户文档矩阵，$\boldsymbol{W}$ 为单词矩阵。在该模型中，使用具有 3 个词的上下文的池化或连接来预测第 4 个词，用户文档量表示当前上下文中缺少的信息，可以作为用户文档主题的记忆。在 word2vec 模型 CBOW 基础上新增文档 ID 作为训练语料中每个文档的唯一标

识,即本文的用户文档 ID 标识,与其他单词 $\boldsymbol{W}$ 一样映射成一个维度相同的向量,然后在固定窗口内将其与词向量进行池化或连接用于下一个单词的预测。在训练过程中使用随机梯度下降及反向传播方式更新用户文档向量、单词向量及其所对应的权值。在每一次的训练中,用户文档向量共享,即用户文档向量会作为一个的固定的向量参与到整个训练过程,不因窗口滑动而忽略。DBOW 训练方式与 word2vec 中的 Skip-gram 类似,不同之处也在于用户文档向量的添加,即每一次单词概率的训练都有用户全局文档语义的参与。训练完成后再通过分类模型进一步提升 DBOW 及 DM 所得模型在相应的单任务预测结果上的准确率。

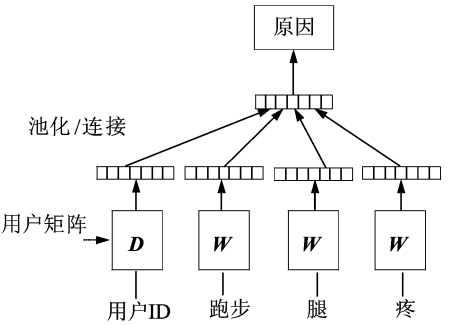

图 2　DM 训练过程

Fig. 2　DM training process

2.2　基于 TF-IDF_MR 模型的单特征推断

最大化特征与分类变量之间的相关度是特征选择中较为常见的一种方式。但由于提取的特征之间存在较高相关性,使得提取出的前 N 个较好的特征组合并不一定能提升最终分类效果。因此本文提出 TF-IDF(Term frequency-inverse document frequency)_MR 联合策略进行关键词提取。即在 TF-IDF 提取的关键词基础上,采用互信息概念通过最小冗余(MR)标准消除关键词之间的冗余。

TF-IDF 算法是一种简单而高效的关键词提取方法,用于衡量单词对文本的重要程度。设文档集为 D,d 为 D 其中一个文档,N 表示 D 中的全部文档数目,计算公式为

$$\boldsymbol{W}_{\text{TF-IDF}} = TF \times \text{IDF}$$
$$\text{IDF} = \log(N/n) \tag{1}$$

式中:TF 计算单词 t 占文档 d 所有词的比重,即在文档 d 中的出现频率;IDF 表示文档的逆文档频率,指该词在其他文档中分布的稀疏性,n 表示在 D 中包含 t 的文档数。

对于用户文档而言,TF-IDF 可以很好地提取用户关键词信息,但其选择过程没有考虑关键词之间的相关性,即选择的关键词之间可能存在冗余。MR 算法可通过计算特征间互信息的方式判断冗余度的大小。对于两个高相关度的单词来说,去掉其中某个单词并不会影响最终分类结果。所以对关键词子集 S($S \subset \text{tfidf}$,tfidf 为通过 TF-IDF 方式选取得到的关键词集),使用最小冗余标准,如式(2)所示。

$$\min R(S), R = \frac{1}{|S|^2} \sum_{w_i, w_j \in S} I(w_i, w_j) \tag{2}$$

式中 $I(w_i, w_j)$ 表示两个单词之间的互信息。

结合上述约束,产生关键词集为

$$\max\sigma(\text{tfidf}, R), \sigma = \text{tfidf} - R \tag{3}$$

基于此,最后采用相对快速的逻辑回归方式进行用户分类。

3　基于关联学习的多任务融合推断

因为数据本身和最终分类属性间的联系复杂,所以数据间的关联性一定程度上有利于提高属性推断的准确率。上述单任务模型虽然训练有效,但对多分类断而言效果并不理想。从另一个角度讲,单任务学习会忽略用户属性之间的关联关系,但这种关联关系却可以作为相关任务的分类依据,所以在传统单模型单任务训练基础上,加入相关属性特征,共同作为单模型、多任务用户属性推断的输入。同时,使用不同模型分类的侧重点有所不同,本文采用 Stacking 融合机制,借助单模型、多任务训练结果并将其组合起来,从而达到比单模型训练结果更好的分类效果,降低模型过拟合的可能性。

基于 Stacking 的多模型融合框架中，将关联学习中单模型、单任务学习结果组合作为单模型、多任务学习的输入，通过关联学习针对各属性得到多个分类模型；在融合阶段，将上一阶段各单模型、多任务训练结果中的各属性组合形成新的数据集，使用支持向量机方法在新的数据集上进行训练得到新的分类模型，用于最终用户属性的推断。

图 3 给出了单模型、多任务关联学习过程。即对于单个模型而言，分别将 DBOW，DM 及 TF-IDF_MR 单任务、单模型分类所得的属性特征矩阵加入用户文档向量，用于用户的联合表征，并通过分类器训练完成用户属性推断。此后，还需与其他单模型多任务训练结果融合训练，从而训练出具有较强泛化能力的分类模型。

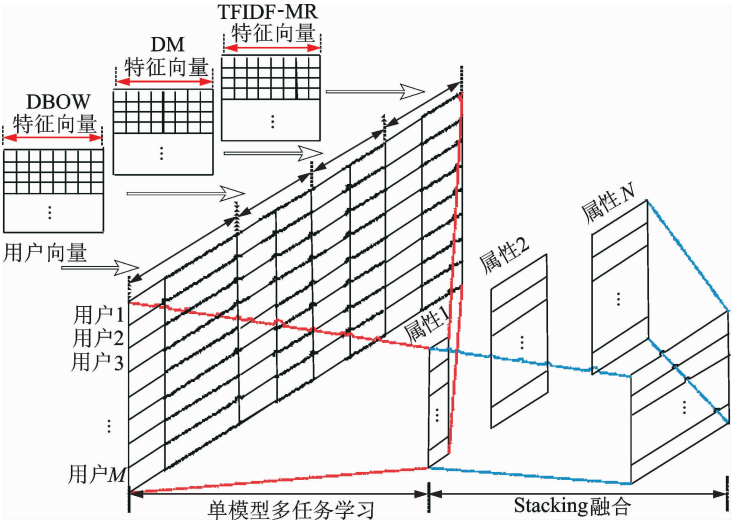

图 3 多任务关联学习模型

Fig. 3 Multitasking association learning model

4 实验结果与分析

4.1 实验数据集及实验环境

数据集 1 本实验采用 2016 年由中国计算机学会举办的大数据竞赛《大数据精准营销中搜狗用户画像挖掘》的比赛数据为实验数据集，其中有 10 万条训练数据，如表 1 所示。

表 1 数据集 1 格式说明

Tab. 1 Format description of data set 1

字 段	说 明
ID	加密后的 ID
Age	0：未知年龄；1：0～18 岁；2：19～23 岁；3：24～30 岁；4：31～40 岁；5：41～50 岁；6：51～999 岁
Gender	0：未知；1：男性；2：女性
Education	0：未知学历；1：博士；2：硕士；3：大学生；4：高中；5：初中 6：小学
Query LIst	搜索词列表

数据集 2 采用大型超市的零售数据集(2012～2013 年)。经过前期预处理后数据集中包含 7 万个用户及其属性和购买记录，平均每个用户购买商品约 225 件。其用户属性包括：性别、婚姻、所在年龄

段、收入以及受教育程度。本文仅选取性别、所在年龄段和受教育程度 3 个属性作为推断属性。

实验环境:实验在服务器上进行,实验环境如下:处理器: Intel(R) Xeon(R) 4 颗 8 核;内存为 128 GB;操作系统为 Red Hat Enterprise Linux Server Release 6.5 操作系统。

4.2　评价指标

评价指标为用户各属性推断结果的准确率。其中,性别、所在年龄段和受教育程度分别计算准确率,最终以平均准确率作为评判依据。

评价指标准确率为

$$P=\frac{1}{N}\sum_{i=0}^{N}I(\hat{y}_i=y_i)$$

式中:$\hat{y}_i$表示第 i 个用户的推断属性,y_i 表示第 i 个用户的真实属性。$I(\cdot)$函数为指标函数,当推断结果与真实结果完全相同时,输出为 1,否则为 0。

平均准确率计算公式为

$$\overline{P}=\frac{P_{\text{gender}}+P_{\text{age}}+P_{\text{education}}}{3}$$

4.3　数据预分析

本文的假设基础是:用户属性之间存在关联关系,故本文针对数据集 1,进行了年龄段与受教育程度关联度分析:实验数据采用数据集 1 中随机抽样的 2 万个用户数据,通过 weka 平台建立散点图,如图 4 所在年龄段与受教育程度关联关系分布。由图可知,所在年龄段与受教育程度近似呈线性相关,即年龄较小的用户其受教育程度较低的可能性较大。

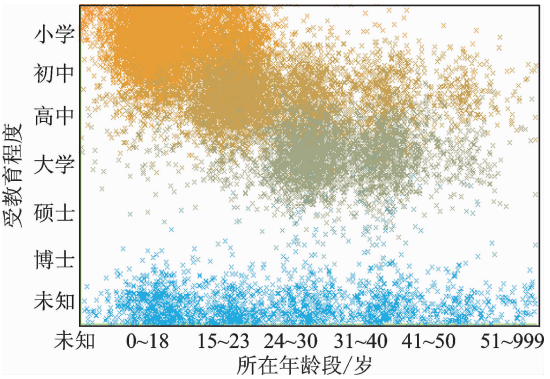

图 4　所在年龄段与受教育程度关联关系

Fig. 4　Relationship between age and education

4.4　基于关键词策略的属性推断结果对比

在实验中发现,本文所提取的用户关键词中存在一定的冗余。所以,本文在基于 TF-IDF 的关键词策略基础上,为去除关键词冗余,采用互信息概念,加入最小冗余机制,共同作为用户属性推断依据。实验结果如表 2 所示。

表 2　TF-IDF 与 TF-IDF_MR 结果对比

Tab. 2　Comparison of results between TF-IDF and TF-IDF_MR

评价指标	数据集 1				数据集 2			
	$\overline{P}$	P_{gender}	P_{age}	$P_{\text{education}}$	$\overline{P}$	P_{gender}	P_{age}	$P_{\text{education}}$
TF-IDF	0.699	0.836	0.602	0.659	0.552	0.597	0.520	0.540
TF-IDF_MR	0.701	0.838	0.605	0.660	0.553	0.599	0.521	0.540

由表 2 可知,加入最小冗余机制的 TF-IDF_MR 分类准确率略高于传统的关键词提取算法。由于传统的 TF-IDF 算法所提取的前 k 个关键词可能存在较强的冗余性,导致用户的片面表征,丧失了整体属性特质。而去除冗余后的前 k 个关键词具有较全面的用户表征能力,故在实验中用户属性推断准确率有所提升。

4.5 基于关联学习的多任务推断结果对比

本文使用 doc2vec 中的 DBOW 模型及 DM 模型在数据集上进行多次试验，试验结果表明，由于属性间的相互联系，加入预测属性特征的单模型训练可以赋予用户更丰富的表征，有助于更好地单任务预测。实验结果如表 3，4 所示。其中，Multi-DBOW 与 Multi-DM 为多任务模型。

表 3 单任务与多任务 DBOW 模型结果对比

Tab. 3 Comparison of single task and multi-task results based on DBOW model

评价指标	数据集 1				数据集 2			
	$\overline{P}$	P_{gender}	P_{age}	$P_{\text{education}}$	$\overline{P}$	P_{gender}	P_{age}	$P_{\text{education}}$
DBOW	0.701	0.842	0.611	0.650	0.542	0.589	0.516	0.520
Multi-DBOW	0.719	0.849	0.631	0.677	0.562	0.608	0.535	0.543

表 4 单任务与多任务 DM 模型结果对比

Tab. 4 Comparison of single task and multi-task results based on DM model

评价指标	数据集 1				数据集 2			
	$\overline{P}$	P_{gender}	P_{age}	$P_{\text{education}}$	$\overline{P}$	P_{gender}	P_{age}	$P_{\text{education}}$
DM	0.627	0.781	0.530	0.571	0.530	0.578	0.502	0.510
Multi-DM	0.660	0.809	0.567	0.604	0.556	0.602	0.527	0.539

实验结果表明、在单任务单模型训练基础上加入多任务因素，有利于提高单模型训练结果的准确率。同时，由表 3 及表 4 可以看出，加入多任务后，两个数据集中对于所在年龄段和受教育程度的预测结果提升较为明显。数据集 1 中，所在年龄段分别提升 2.0% 和 3.7%，受教育程度分别提升 2.7% 和 3.3%；数据集 2 中，所在年龄段分别提升 1.9% 和 2.5%，受教育程度分别提升 2.3% 和 2.8%，分析原因在于：所在年龄段和受教育程度在单任务中的训练本身较为困难，然而受教育程度与所在年龄段存在着较强的关联性，多任务关联学习中二者有明显提高。

在两个数据集中比较表 3 与表 4 可以看出，基于 DM 的多任务训练平均提升值(3.3%，2.6%)要高于基于 DBOW 的多任务训练的平均提升值(1.8%，2.0%)，原因可能在于 DM 的训练方式是从句子中随机抽取词语进行训练，一定程度上忽略了词语之间的次序关系，所以当对用户向量进行非词语语义向量扩充时，其效果更明显。其所在年龄段的提升同理。

4.6 多模型融合推断结果比较

本文分别选择了改进后基于 Multi-DBOW，Multi-DM 和 TF-IDF_MR 作为 3 个单独模型，与使用 Stacking 机制的融合模型进行对比。对比结果如表 5 所示。由表 5 可知，实验集 1 与实验集 2 经过模型融合后准确率较单模型都有很大程度的提升，可见多模型融合在用户属性推断中起着重要作用。且实验中发现，随着模型数量的不断增加，训练结果准确率会逐步提高。尤其是当加入 doc2vec 模型后，提升效果较为显著，说明该模型的加入有助于整体的学习和分类。与此同时，较多模型的加入会直接影响模型融合效率，故选择合适的且具有针对性的模型对融合而言至关重要。

表 5　单模型与多模型融合推断结果比较

Tab. 5　Comparison of inference results between single-model and multi-model ensemble

评价指标	数据集 1				数据集 2			
	$\overline{P}$	P_{gender}	P_{age}	$P_{\text{education}}$	$\overline{P}$	P_{gender}	P_{age}	$P_{\text{education}}$
Multi-DBOW 模型	0.719	0.849	0.631	0.677	0.562	0.608	0.535	0.543
Multi-DM 模型	0.660	0.809	0.567	0.604	0.556	0.602	0.527	0.539
TF-IDF_MR 模型	0.701	0.838	0.605	0.660	0.553	0.599	0.521	0.540
Multi-DBOW,Multi-DM, TF-IDF_MR 融合模型	0.740	0.862	0.656	0.703	0.582	0.633	0.549	0.564

4.7　与 CCF2016 测评结果比较

表 6 中给出了在数据集 1 上,利用本文方法在使用较少模型的情况下所得评价结果与该参评系统前 10 名平均水平和最高测评结果的对比。由表 6 可见,本文方法较测评结果准确率有所提高。

表 6　测评结果比较

Tab. 6　Comparison of evaluation results

参评系统	$\overline{P}$
前 10 名均值	0.713
第 1 名	0.727
本文结果	0.740

5　结束语

本文讨论了使用多任务融合模型的用户属性推断方法研究。通过考虑用户属性之间的关联关系,在单任务模型的基础上增加关联学习,从而利用属性间的隐性关联,更好地表征用户,提高单任务模型分类的准确率。同时,结合 stacking 多模型融合技术,进一步提高模型推断结果。接下来的工作将进一步优化分词效果,将 doc2vec 中采用到的 word2vec 训练词向量与现有训练好的词向量结合,丰富语义表征;发掘更好的预测模型,尝试使用多种深度学习框架进行多任务间的属性推断;将用户属性推断与异构数据结合,从而更好地完成属性推断任务。

参考文献:

[1] Schler J M,Koppel M, Arfamon S, et al. Effects of age and gender on blogging[J]. Frontiers of Information Technology & Electronic Engineering,2006,274(S 1/2):199-205.

[2] Hu J, Zeng H J, Li H, et al. Demographic prediction based on user's browsing behavior[C]// International Conference on World Wide Web. Banff, Alberta, Canada: DBLP, 2007:151-160.

[3] Torres S D, Weber I. What and how children search on the web[C]// ACM International Conference on Information and Knowledge Management. [S. l.]:ACM, 2011:393-402.

[4] Zhong E, Tan B, Mo K, et al. User demographics prediction based on mobile data[J]. Pervasive & Mobile Computing,2013, 9(6):823-837.

[5] Culotta A, Ravi N K, Cutler J. Predicting the demographics of Twitter users from website traffic data[C]// Twenty-Ninth AAAI Conference on Artificial Intelligence. [S. l.]:AAAI Press, 2015:72-78.

[6] Zhong Y, Yuan N J, Zhong W, et al. You are where you go:Inferring demographic attributes from location check-ins[C]// ACM International Conference on Information and Knowledge Management. [S. l.]:ACM, 2015:295304.

[7] Evgeniou T, Pontil M. Regularized multi-task learning[C]// Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l.]: ACM, 2004: 109-117.

[8] 王礼敏,严倩,李寿山,等. 基于双通道 LSTM 模型的用户性别分类方法研究[J]. 计算机科学,2018,45(2):121-124.
Wang Limin, Yan Qian, Li Shoushan, et al. User gender clssification with dual-channel LSTM[J]. Computer Science, 2018, 45(2): 121-124.

[9] 戴斌, 李寿山, 贡正仙,等. 基于多类型文本的半监督性别分类方法研究[J]. 山西大学学报(自然科学版), 2017, 40(1): 14-20.
Dai Bin, Li Shoushan, Gong Zhengxian, et al. Semi-supervied gender classification with multiple types of text[J]. Journal of Shanxi University (Natural Science Edition), 2017, 40(1): 14-20.

[10] Wang P, Guo J, Lan Y, et al. Your cart tells you: Inferring demographic attributes from purchase data[J]. Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, 2016, 1: 173-182.

[11] Mesnil G, Mikolov T, Ranzato M A, et al. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews[J]. Lightwave Technology Journal of, 2014, 32(17): 3043-3060.

[12] Liu Y, Liu Z, Chua T S, et al. Topical word embeddings[C]// Twenty-Ninth AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2015: 2418-2424.

[13] Wang P, Guo J, Lan Y, et al. Multi-task representation learning for demographic prediction[C]// European Conference on Information Retrieval. [S. l.]: Springer, 2016: 88-99.

[14] Sun F, Guo J, Lan Y, et al. Sparse word embeddings using regularized online learning[C]// International Joint Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2016: 2915-2921.

[15] Xiang L, Sang J, Xu C. Demographic attribute inference from social multimedia behaviors: A cross-OSN approach[C]// International Conference on Multimedia Modeling. [S. l.]: Springer, 2017: 515-526.

[16] 赵军,王红,朱华方. 一种改进的融合关联词典的微博倾向性分析方法[J]. 数据采集与处理, 2016, 31(6): 1220-1227.
Zhao Jun, Wang Hong, Zhu Huafang. Improved method for analyzing microblog orientation based on association lexicon[J]. Journal of Data Acquisition and Processing, 2016, 31(6): 1220-1227.

[17] 张文艳,李存华,仲兆满,等. 结合规则与语义的中文人称代词指代消解[J]. 数据采集与处理, 2017, 32(1): 149-156.
Zhang Wenyan, Li Cunhua, Zhong Zhaoman, et al. Coreference resolution of Chinese personal pronouns with combination of semantics and rules[J]. Journal of Data Acquisition and Processing, 2017, 32(1): 149-156.

作者简介:

赵宇(1992-),女,硕士研究生,研究方向:网络媒体大数据挖掘和自然语言处理,E-mail: 411975417 @ qq. com。

李佳艺(1995-),女,硕士研究生,研究方向:网络媒体大数据挖掘, E-mail: 1057339032@qq. com。

王莉(1971-),女,博士,教授,研究方向:网络媒体大数据挖掘和人工智能,E-mail: l_lwang@126. com。

(编辑:刘彦东)