浙江理工大学学报，2020，43(1)：86-93
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673－3851(n).2020.01.013

基于集成学习框架的用户画像方法

陈巧红，凌明杰，孙　麒，贾宇波
（浙江理工大学信息学院，杭州 310018）

摘　要：针对已有算法中特征构建效果不佳以及泛化能力不足的问题，提出一种基于集成学习框架的用户画像方法。该方法将整体架构分为集成学习模块与语义编码模块，并在决策时加入了投票机制。集成学习模块采用两层 Stacking 完成特征构建以及模型融合；语义编码模块使用 BERT 模型对文本进行编码，提取深层语义信息；然后对两个模块的输出结果进行投票，从而产生最终结果。对两组数据进行实验，结果显示：该方法与基于单模型的方法对比，在用户查询词数据集上，用户性别、年龄、学历标签分类准确率平均提高了 1.27%、3.52%、3.42%；在微博用户数据集上，用户性别、年龄、学历标签的分类准确率平均分别提高了 5.61%、6.49%、5.96%。这表明该方法对于用户画像任务有较好的效果，并且对不同形式的文本具有很好的适应性。

关键词：集成学习；Stacking；BERT；用户画像；机器学习
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1673-3851 (2020) 01-0086-08

User profile method based on ensemble learning framework

CHEN Qiaohong, *LING Mingjie*, *SUN Qi*, *JIA Yubo*
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Aiming at the problem of poor feature construction effect and insufficient generalization ability of existing algorithms, a user profile method based on ensemble learning is proposed. This method divides the whole architecture into ensemble learning module and semantic coding module, and adds the voting mechanism to decision-making. In ensemble learning module, two-layer stacking is employed to construct features and integrate models. Semantic encoding module applies BERT model to encode texts and extract deep semantic information. Then, the output results of the two modules are voted to get the final result. The experimental results show that compared with the method based on the single mode, the classification accuracy of this method in the user gender, age and education increases by 1.27%, 3.52% and 3.42% respectively on average on user query word dataset. The classification accuracy of user gender, age and education increases by 5.61%, 6.49% and 5.96% respectively on average on microblog user dataset. This shows that the proposed method has a good effect on user profile and has a good adaptability to different forms of texts.

Key words: ensemble learning; Stacking; BERT; user profile; machine learning

0　引　言

步入 21 世纪，随着中国互联网行业的蓬勃发展，互联网网民数量急速增长，由此产生了大量的与用户行为信息相关的数据。如何让这些海量的用户信息产生价值，是科研工作者急需解决的问题，用户画像技术由此产生。用户画像最初由交互设计之父 Alan Cooper 提出，他认为用户画像是真实用户的

收稿日期：2019－06－30　　网络出版日期：2019－10－08
基金项目：国家自然科学基金项目(51775513)
作者简介：陈巧红(1978－　)，女，浙江临海人，副教授，博士，主要从事计算机辅助设计及机器学习技术方面的研究。

虚拟表示，是基于用户的真实数据挖掘和构建的目标用户模型[1-2]。企业能够根据用户一系列的行为数据为其建立用户画像，从而加深对用户的了解，进一步实现个性化服务、精准营销、提高经济效益等目的[3-4]。

在众多用户画像模型构建的方法中，最主要的方法是通过给用户贴标签，用非抽象的词语来描述用户的行为和特征，并用不同的机器学习模型和规则分析挖掘得到高度精炼的关键词，以此构建用户画像。Cha 等[5]提出了一种为用户计算影响力的方法，利用微博的用户关注数、被转发次数、点赞和转发的微博等数据分析用户的行为。该方法依靠行为特征训练机器学习模型，从而实现用户画像；其缺点是，由于同一用户行为的复杂性以及微博中大量的混淆信息，因而导致仅依靠行为特征训练模型效果不理想。Räbiger 等[6]提出了一种使用交叉验证对多种特征进行组合的方法，利用贝叶斯网络(Bayesian network，BN)对多组交叉特征进行训练。但这种方法仅使用有限的已标注的训练数据，因此制约了学习模型的泛化能力，并且该方法采用单一模型，难以应对多用户多环境出现的不稳定。陈姝等[7]提出了一种基于用户综合行为构建用户画像的方法，从用户态度和主观因素 2 个方面对微博用户的行为进行分析，通过潜在狄利克雷分布(Latent Dirichlet allocation，LDA)模型训练微博文本与主题分布的概率，并利用逻辑回归(Logistic regression，LR)模型进行验证。该方法从多角度分析用户信息并提取了相关特征，但所提取的微博文本语义和用户特征的精确性和一致性还有待提高。费鹏等[8]提出了一种多视角融合框架，利用双通道对不同用户分别建模，针对多种类型的特征构建多元特征，并基于极端梯度提升(Extreme gradient boosting，Xgboost)算法构建多视角融合模型。该框架虽然从多角度构建了多元特征，但对于各数据中最为重要的文本数据，其提取方式并没有考虑文本语义关联性等深层次信息。李恒超等[9]提出一种二级融合算法框架以用于预测用户多维标签。该框架的第一级将神经网络模型与表示学习相结合，针对用户数据的三个标签来抽取查询的语义关联信息。第二级框架使用 Stacking 将多个 Xgboost 模型相融合。实验结果表明，在利用用户查询词文本构建用户画像的任务中，使用该框架取得了优异的效果，缺点是在框架的第二级中，仅对多个相同的模型进行融合，其泛化能力仍有提升空间，而且该框架仅通过浅层学习对如微博这种含有强语义关联性的文本的适用性有限。本文由此得到启发，改进李恒超等[9]提出的二级融合算法框架，使其能适应不同表达方式的文本。

本文对李恒超等[9]提出的二级融合算法框架进行改进，提出了一种基于集成学习框架的用户画像方法，进一步提升方法的泛化能力，并能够适应不同表达方式的文本。本文提出的方法由集成学习模块与语义编码模块两部分组成。语义编码模块采用两层 Stacking 结构，首先从用户用词习惯和用户关键词信息这两方面构建特征，再进行特征矩阵大小匹配与模型融合。语义编码模块采用表示高层语义编码的 BERT(Bidirectional encoder representation from transformers)模型提取高维稀疏特征，并进行训练。最后，两个模块各自所输出的结果进行投票决策，得到最终的分类结果。

1　方法设计

1.1　整体流程

本文设计的用户画像方法的整体流程如图 1 所示。首先输入用户数据集，然后将用户数据集进行预处理，进行分词以及去停用词操作之后，再输入集成学习模块与语义编码模块。集成学习模块由两层的 Stacking 构成。在第一层中采用表示词条重要程度的词频-逆文本频率指数(Term frequency-inverse document frequency，TFIDF)算法与表示文本关键词信息的 TextRank 算法，提取初步特征；再将这两种特征分别输入 LR 模型、支持向量机(Support vector machine，SVM)模型与随机森林(Random forest，RF)模型进行训练，得到用户用词习惯特征矩阵 $\boldsymbol{M}$ 及用户关键词特征矩阵 $\boldsymbol{H}$。在第二层中将 $\boldsymbol{M}$ 与 $\boldsymbol{H}$ 进行融合，得到矩阵 $\boldsymbol{U}$，再将 $\boldsymbol{U}$ 输入梯度提升决策树(Gradient boosting decision tree，GBDT)模型进行训练。在语义编码模块中，通过 BERT 模型对文本进行编码从而提取深层语义信息，再通过 Softmax 回归输出结果，该结果与集成学习模块中输出的结果进行投票决策，从而得到最终分类结果。

1.2　集成学习模块

集成学习模块由两层的 Stacking[10-11]构成。在第一层中完成特征的构建以及多个基本分类器的训练，第二层负责将第一层中所获得的特征进行融合[12]，再进一步输入到 GBDT 进行训练，提升泛化能力。

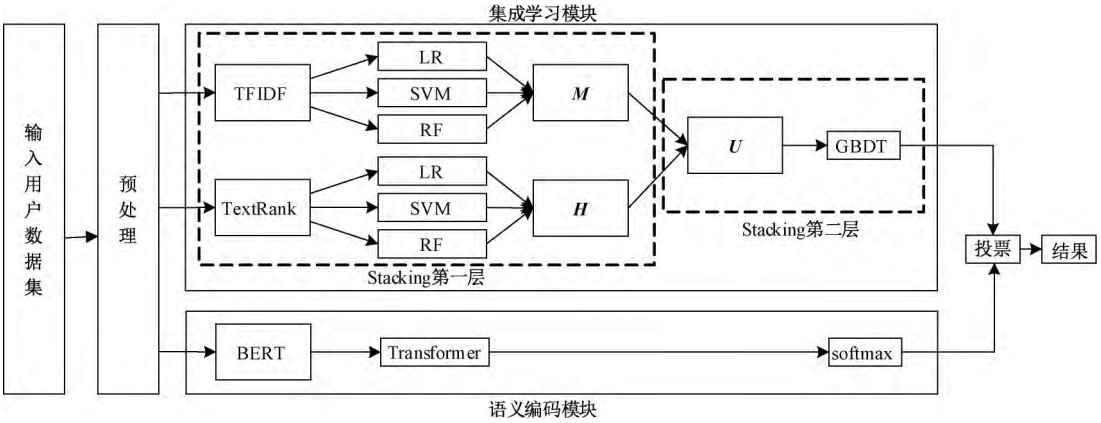


图1 用户画像方法整体流程

1.2.1 第一层 Stacking

a)用户用词习惯特征构建。为了构建文本的用户用词习惯特征,本文采用了 TFIDF 算法以及三个基本分类模型。

TFIDF 算法是一种文本权重表示方法,它衡量了各词语在文档中的重要程度,综合考虑了词语在文档中的词频和词语在整体语料库中分布的影响。某个词语的 TFIDF 值与该词在该文档中的出现次数成正相关,与该词在整个语料中的出现次数成反相关[13]。TFIDF 计算公式如式(1)所示:

$$TFIDF(w,d)=TF(w,d)\log\left(\frac{N}{DF(w)}\right) \quad (1)$$

其中:$TF(w,d)$代表词语 w 在文章 d 中出现的频率,N 代表语料库中的文档总数,$DF(w)$代表整个语料中包含词语 w 的文档数目。

假设输入的数据集大小为 $T\times K$,数据集经过第一层中的三个分类模型后,会产生 3 组与原数据集规模相同且大小为 1 的结果,然后将这 3 组结果拼接,构成了大小为 $T\times 3$ 的特征矩阵 $\boldsymbol{M}$,用来表示用户用词习惯特征。

b)用户关键词特征构建。关键词特征会极大地影响用户画像任务的结果,而 TextRank 算法对于关键词有很好的识别能力,并能够挑选出重要性高的关键词[14],因此本文采用 TextRank 算法构建特征。

TextRank 算法通过词之间的相邻关系构建网络,然后迭代计算每个节点的排名值,最终将排名值依次排序后可得到关键词。其迭代公式如式(2)所示:

$$WS(V_i)=(1-c)+c\cdot\sum_{j\in In(V_j)ejk}\frac{e_{ji}}{\sum_{V_k\in Out(V_j)}e_{jk}}WS(V_j) \quad (2)$$

其中:$WS(V_j)$表示结点 V_i 的排名值,$WS(V_j)$表示结点 V_j 的排名值,c 表示平滑系数,$In(V_j)$表示结点 V_j 的前驱结点集合,$Out(V_j)$表示结点 V_j 的后继结点集合,e_{ji}表示结点 V_j 与结点 V_j 的相似度,e_{jk}表示结点 V_j 与结点 V_k 的相似度。

在用户用词习惯特征与用户关键词特征构建的过程中,本文都采用了 LR、SVM 和 RF 这三个分类模型,于是大小为 $T\times K$ 的数据集经由 TextRank 算法以及三个分类模型训练之后,所输出的用户关键词特征矩阵 $\boldsymbol{H}$ 的大小也为 $T\times 3$。这样,Stacking 的第一层就实现了特征矩阵大小匹配的功能。

1.2.2 第二层 Stacking

Stacking 可以通过设置多层级的结构,每层放置合适的分类器簇,将新特征融合入各层之间的中间结果中。于是在本文集成学习模块的第二层中,经由第一层输出的特征矩阵 $\boldsymbol{M}$ 与 $\boldsymbol{H}$ 便能够拼接融合成矩阵 $\boldsymbol{U}$,其大小为 $T\times 6$。然后将矩阵 $\boldsymbol{U}$ 输入到 GBDT 进行训练。

1.3 语义编码模块

在中文语料中,往往存在大量的多义词,如:“我买了一部苹果手机”与“我去超市买了一斤苹果”这两个句子中,“苹果”一词所代表的含义不同。而传统的词向量,如 Word2Vec、Glove 等均无法表征词的多义性,而本文在语义编码模块中采用的 BERT 模型则能较好地解决这个问题。

BERT 模型是 Devlin 等[15]在 2018 年提出的,它采用 Transformer 模型作为编码器来预测词语上下文。Transformer 编码器舍弃了长短项记忆网络(Long short-term memory, LSTM)中的循环式网络结构,而完全采用注意力机制对文本进行建模。Transformer 模型如图 2 所示,每个位置上的词语经过词嵌入(Word embedding)转化为词语向量,然

后经过 Self-Attention 层提取特征。在对应的每个位置上，Self-Attention 层都会输出等长度的向量，再输入到前馈神经网络中。Self-Attention 层的核心思想是计算句子中每个词语在这个句子中所有词之间的关联性及重要程度，从而获得每个词语在全局关系上的表达向量[16]，因此，对同一个词语在不同语境下的表达有较好的区分能力。

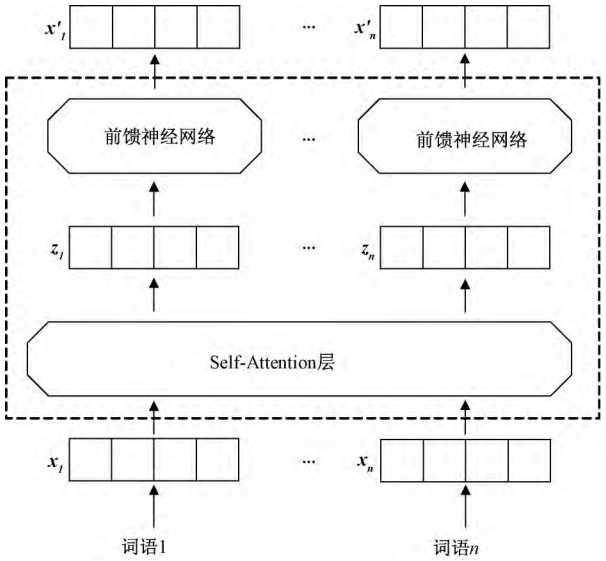


图 2　Transformer 模型

图 2 中：x_1，x_2，…，x_n，表示一段文本中的 n 个词语经过词嵌入转化后的向量，z_1，z_2，…，z_n 表 Self-Attention 层输出的向量，x'_1，x'_2，…，x'_n 表示 Transformer 编码器输出的向量。

Self-Attention 层中每一个词语向量 x_i 均生成 3 个新的向量 q_i、k_i、v_i，这些向量可以用矩阵的形式表示成 $\boldsymbol{X}$：$(x_1, x_2, \cdots, x_n)$，$\boldsymbol{Q}$：$(q_1, q_2, \cdots, q_n)$，$\boldsymbol{K}$：$(k_1, k_2, \cdots, k_n)$，$\boldsymbol{V}$：$(v_1, v_2, \cdots, v_n)$。输出后的向量用矩阵的形式表示成 $\boldsymbol{Z}$：$(z_1, z_2, \cdots, z_n)$。$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$、$\boldsymbol{Z}$ 计算过程如式(3)—(6)所示：

$$\boldsymbol{Q} = \boldsymbol{X} \cdot \boldsymbol{W} \tag{3}$$

$$\boldsymbol{K} = \boldsymbol{X} \cdot \boldsymbol{W} \tag{4}$$

$$\boldsymbol{V} = \boldsymbol{X} \cdot \boldsymbol{W} \tag{5}$$

$$\boldsymbol{Z} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right) \cdot \boldsymbol{V} \tag{6}$$

其中：W_Q、W_K、W_V 为模型训练时的权值矩阵，d_k 为输入向量维度，K^{T} 表示 K 矩阵的转置。

BERT 自问世以来，在众多自然语言处理任务中表现优异，现如今已经是最热门的自然语言处理模型之一。因此，本文选用 BERT 模型对文本进行深层次的分析。BERT 模型通过多层的 Transformer 编码器进行训练，从而输出每个对应词语位置上的表示向量。BERT 模型训练过程如图 3 所示。图 3 中：E_1，E_2，…，E_n 表示输入文本中的所有词语，Trm 代表上述的 Transformer 编码器，T_1，T_2，…，T_n 表示经过多层 Transformer 编码器输出的对应位置的向量。

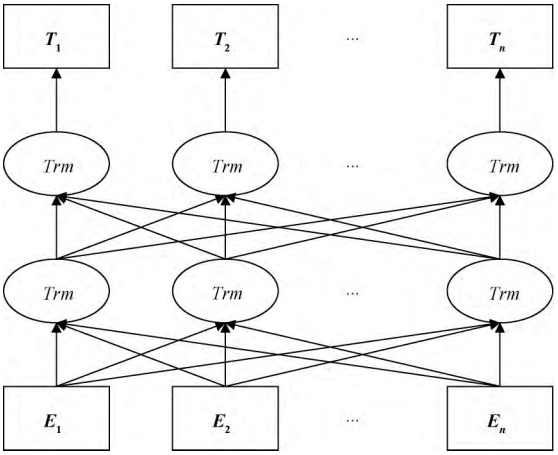


图 3　BERT 模型训练过程

2　实验及结果分析

2.1　实验数据

本文使用 2016 CCF 大数据与计算智能大赛提供的用户查询数据进行实验一，训练集与测试集数据量均为 100000 条；使用爬虫技术在新浪微博上爬取了一部分用户微博文本数据进行实验二，训练集与测试集数据量也均为 100000 条。为了与实验一数据尽可能一致，实验二的数据爬取了与实验一所用数据相同的标签，即用户的性别、年龄、学历。

实验数据举例如表 1 所示。从表 1 的数据描述举例可以看到，在实验一所采用的用户查询数据中，样本由多个关键词组成，如“国庆 旅游 杭州”，而“国庆”、“旅游”、“杭州”这三个词语之间并没有太强的关联性。实验二的微博用户评论数据则与之相反，由表述连贯的时间序列组成。

表 1　实验数据举例

实验	数据集	数据描述举例	标签		
			性别	年龄/岁	学历
实验一	用户查询数据	国庆，旅游，杭州	女	22	本科
实验二	微博用户评论数据	这个时代中精神与物质的关系	男	38	硕士

2.2　衡量指标

本文的用户数据包含三个标签：性别、年龄、学历，这三个标签对应的类别数目分别是二分类、七分类、七分类。其中对年龄标签采取数据分箱技术[17]，即对不同年龄段设置不同的类别。本文使用

准确率作为衡量标准,并用混淆矩阵解释准确率,混淆矩阵见表2。

表2 分类结果混淆矩阵

预测结果	属于该类别样本数	不属于该类别样本数
预测属于该类别样本	TP	FP
预测不属于该类别样本	FN	TN

准确率表示预测该类别正确的数量占总样本数目的比例,可用式(7)计算:

$$P/\% = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (7)$$

其中:P 表示准确率,TP 表示将正类预测为正类的样本数,FP 表示将负类预测为正类的样本数,FN 表示将正类预测为负类的样本数,TN 表示将负类预测为负类的样本数。

2.3 结果分析

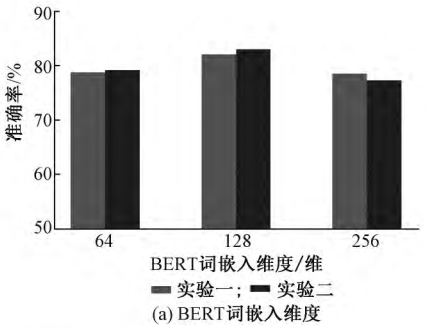
为了验证所提出方法的有效性,本文分别使用LR模型、SVM模型、RF模型、GBDT模型、BERT模型和本文方法进行对比实验。

在实验过程中,为了达到各个模型的最佳分类效果及参数调优,本文对模型参数进行选择,每选择一个参数时,都控制其他参数保持不变进行测试,如表3所示。

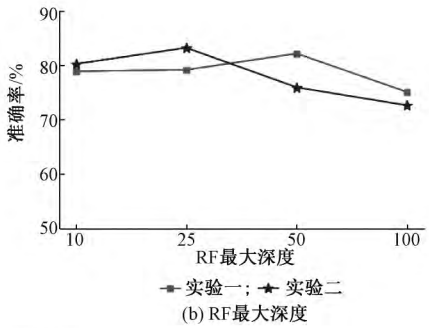
模型参数取不同的值会得到不同的实验结果,实验一与实验二中模型参数选取不同值对平均分类准确率的影响如图4(a)—(h)所示。

表3 模型参数设计

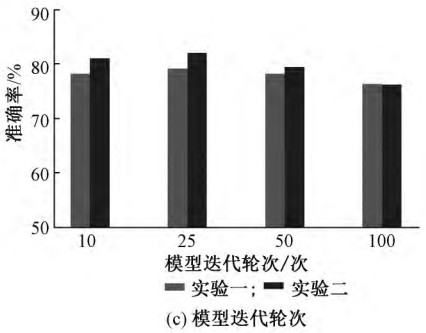
模型参数	实验一模型参数最优取值	实验二模型参数最优取值	模型参数取值范围
BERT词嵌入维度	128	128	64,128,256,512
RF最大深度	50	25	10,25,50,100
GBDT最大深度	10	25	10,25,50,100
SVM惩罚系数	1.0	1.0	1.0,2.0,3.0
LR惩罚系数	1.0	1.0	1.0,2.0,3.0
模型迭代轮次	25	25	10, 25, 50,100
学习率	0.001	0.001	0.001,0.010,0.050,0.100
每批训练集数据大小	64	64	8, 16, 32, 64



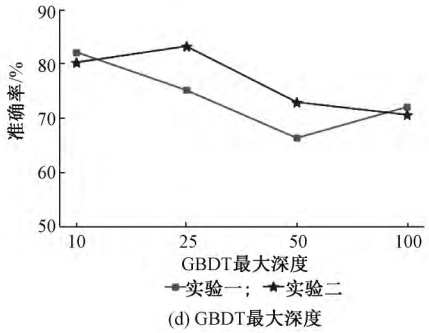
(a) BERT词嵌入维度



(b) RF最大深度



(c) 模型迭代轮次



(d) GBDT最大深度

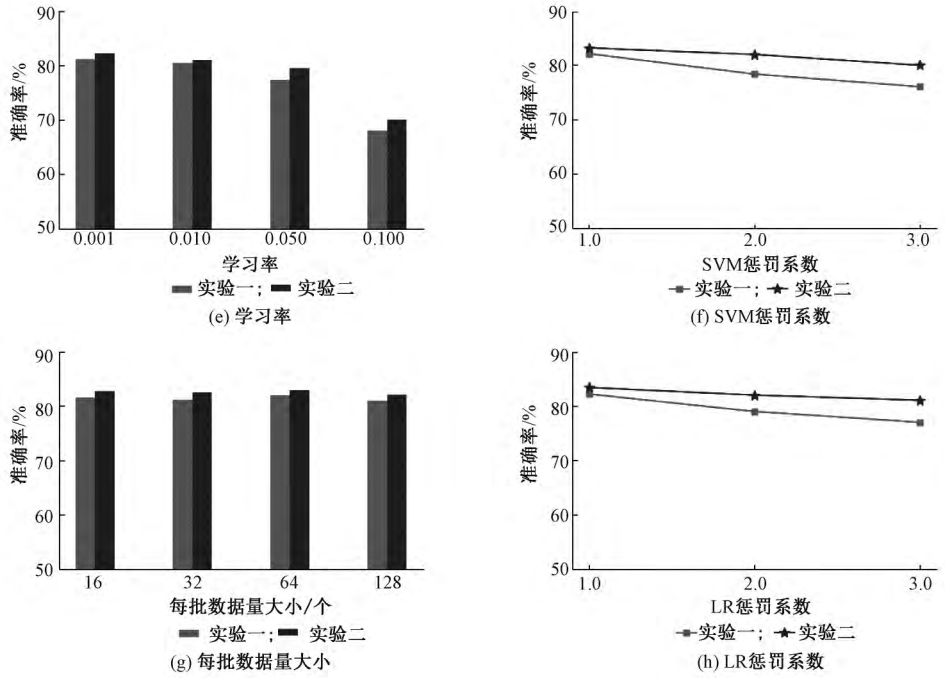


图 4　不同模型参数下的平均分类准确率

图 4(a)—(h)中每个子图表示实验一与实验二的各参数取值结果。图 4(a)表示，实验一与实验二中 BERT 词嵌入维度均选择 128 时平均分类准确率最高。当样本数目较多时，可以通过调整 RF 和 GBDT 的最大深度以减小过拟合程度。从图 4(b)和图 4(d)中可以看出，对于实验一，RF 和 GBDT 的最大深度分别选取 50 和 10 时，平均分类准确率最高；而对于实验二，RF 和 GBDT 的最大深度均为 25 时，平均分类准确率最高。SVM 和 LR 的惩罚系数也可以有效减小过拟合的程度，当惩罚系数过大时，表示分类模型不允许出现分类误差样本，则会造成过拟合现象；而当惩罚系数过小时，则模型不容易收敛。图 4(f)和(h)的实验结果表明，实验一和实验二的 SVM 和 LR 惩罚系数均选取 1.0 时效果最佳。由于两个实验的样本数一致，且总词量也基本一致，因而两个实验中模型的迭代次数到达 25 时，两个实验的平均分类准确率均达到峰值，当选取 50 时，开始发生过拟合现象。每批数据量的大小对于两个实验的平均分类准确率影响都不大，故都选取 64。学习率应尽量选低，因为过大的学习率会导致模型在接近局部最低点时难以收敛，图 4(e)的结果显示，学习率选 0.001 时，实验一与实验二的分类结果最优。

在对比实验的过程中，本文分别对性别、年龄、学历三个标签在测试集上做分类预测，并统计三个标签的平均分类准确率，实验结果如表 4—表 5 所示。

表 4　实验一(用户查询词数据)不同模型的分类准确率　%

标签	分类模型					
	LR	SVM	RF	GBDT	BERT	本文方法
性别	83.77	84.03	83.85	83.99	84.01	85.20
年龄	74.89	74.18	75.21	77.03	76.81	79.14
学历	78.65	77.16	77.84	78.94	78.89	81.72
平均	79.10	78.45	78.97	79.99	79.90	82.02

表 5　实验二(微博用户文本数据)不同模型的分类准确率　%

标签	各分类模型					
	LR	SVM	RF	GBDT	BERT	本文方法
性别	81.04	84.03	78.98	77.69	86.19	87.20
年龄	67.80	74.18	73.14	76.87	78.81	80.65
学历	72.15	73.26	72.84	76.94	80.89	81.18
平均	70.33	73.26	72.65	76.17	81.96	83.01

表4显示了在用户查询词数据上各个分类模型的详细比较。从第6列与2—5列的结果对比中可以看出,当数据集是用户查询词这种大多数以关键词构成的文本时,BERT模型与LR、SVM、RF、GBDT这些传统机器学习模型在准确率上相比并没有显著的提升,这是因为这种文本只由几个关键词组成,关键词之间没有统一的顺序,并且前后语义关联度不大,采用深度学习的方法不但难以提取到有用的信息,反而会因为其深层次的网络结构导致过拟合现象。而从第7列的结果可以看出,本文所提出的使用改进的集成学习框架的用户画像方法在各个标签上的分类准确率均表现最优,这验证了本文提出的方法适用于多用户的复杂环境,可以提高用户画像的准确率。从第7列与1—6列的对比中,可以看出本文的集成方法对于年龄和学历这两个多分类标签上的准确率提升要比二分类的性别标签上的准确率提升更为显著,这是由于进行多分类任务时,本文方法中集成学习模块的Stacking第一层能够输出每一个类别的概率,以便第二层能够在其之上做阈值判断、相互校验的操作。此外,各个标签之间往往存在一定的关联性,如用户的年龄会极大程度影响到其学历,而Stacking的第二层对这样的特征关系有较好的学习能力,这就导致了多分类任务的特征融合效果更好,进一步验证了本文提出的集成方法可以充分融合特征。

表5显示了在微博用户文本数据上各分类模型的详细比较。从2—5列的结果可以看出,LR、SVM、RF和GBDT这四个传统机器学习模型的效果不佳,因为微博文本具有时序性及深层语义关联性等特点,而针对浅层学习的传统机器学习模型对这种文本的学习能力有限。而第6列的结果显示了使用BERT模型后,各个标签的分类准确率均提升显著,说明BERT模型能够充分挖掘与分析文本内部的深层语义信息。第7列的结果显示本文提出的集成方法进一步提升了分类准确率,说明本文方法能够结合BERT模型与各传统机器学习模型的优点,进一步提升了泛化能力。

对于由多个关键词组成用户查询词文本数据,由于其语义不连贯、表述不全面、词条关联度不高的特点导致深度神经网络模型表现不佳。而对于表述连贯的微博文本等数据,BERT模型表现十分优异。故本文采用两个模块分别从浅层与深层分析文本数据,再通过投票机制结合两者的优点,使本文的方法能够对两种不同形式的文本都有较好的学习能力和泛化能力。表4和表5表明了本文提出的方法在微博用户文本数据集和用户查询文本数据集上,均验证了其有效性与适应性。

综上可知,本文提出的基于集成学习框架的用户画像方法有效,并且对不同形式的文本具有很好的适应性。

3 结 论

本文提出了一种基于集成学习框架的用户画像方法。该方法设计了集成学习模块与语义编码模块这两个模块。在集成学习模块中通过TFIDF算法与TextRank算法构建特征,并通过两层的Stacking结构融合多个基本机器学习模型后输出结果。在语义编码模块中使用BERT模型提取深层语义信息,再得到所输出的结果。最后两个模块所输出的结果共同投票得出最终分类结果。实验结果显示,本文方法比传统的机器学习模型在用户画像的准确率上有比较高的提升,表明使用本文提出的方法可以有效地融合特征,并能够结合多个机器学习模型的优点。

本文虽然从多角度分析和挖掘用户文本数据,从而构建用户画像,但是没有深入挖掘用户之间的信息。在社交网络中,单个用户的行为往往会受到相邻用户影响,因此用户关联性与相似度等因素在一定程度上会制约着用户画像的效果。随着神经网络的发展,图卷积、图网络等工具的出现使得社交网络人群的关联性与相似度的计算效果相比以往有了较大提升,可将这些工具应用到以后的研究中,进一步提升用户画像的效果。

参考文献:

[1] Cooper A. The Inmates are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity[M]. Indianapolis: Sams, 2004: 2.

[2] Pruitt J, Adlin T. The Persona Lifecycle: Keeping People in Mind Throughout Product Design[M]. San Francisco: Morgan Kaufmann, 2006: 15.

[3] 饶元, 吴连伟, 王一鸣, 等. 基于语义分析的情感计算技术研究进展[J]. 软件学报, 2018, 29(8): 2397-2426.

[4] 弭宝瞳, 梁循, 张树森. 社交物联网研究综述[J]. 计算机学报, 2018, 41(7): 1448-1475.

[5] Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in twitter: The million follower fallacy

[C]// Fourth International AAAI Conference on Weblogs and Social Media. Palo Alto: AAAI Press, 2010:10-17.

[6] Räbiger S, Spiliopoulou M. A framework for validating the merit of properties that predict the influence of a twitter user[J]. Expert Systems with Applications, 2015, 42(5): 2824-2834.

[7] 陈姝，窦永香，张青杰. 基于理性行为理论的微博用户转发行为影响因素研究[J]. 情报杂志，2017,36(11): 147-152.

[8] 费鹏，林鸿飞，杨亮，等.一种用于构建用户画像的多视角融合框架[J]. 计算机科学，2018，45(1)：179-182.

[9] 李恒超，林鸿飞，杨亮，等. 一种用于构建用户画像的二级融合算法框架[J]. 计算机科学，2018，45(1)：157-161.

[10] Jahrer M, Töscher A, Legenstein R. Combining predictions for accurate recommender systems[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. New York: ACM, 2010: 693-702.

[11] Mesnil G, Mikolov T, Ranzato M A, et al. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews[EB/OL]. (2015-05-27)[2019-09-15]. https://arxiv.org/abs/1412.5335.

[12] 马朋辉，王雪宁，李勇，等. 用户画像构建研究[J]. 现代信息科技，2019,3(6)：17-18.

[13] 郑霖，徐德华.基于改进 TFIDF 算法的文本分类研究[J]. 计算机与现代化，2014(9)：6-94.

[14] 刘竹辰，陈浩，于艳华，等. 词位置分布加权 TextRank 的关键词提取[J]. 数据分析与知识发现，2018(9)：74-79.

[15] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2019-09-12]. https://arxiv.org/abs/1810.04805.

[16] 杨飘，董文永.基于 BERT 嵌入的中文命名实体识别方法[J/OL].计算机工程：1-7[2019-06-20].https://doi.org/10.19678/j.issn.1000－3428 0054272.

[17] 傅涛，孙文静，孙亚民. 基于分箱统计的 FCM 算法及其在网络入侵检测中的应用[J]. 计算机科学，2008，35(4)：36-39.

（责任编辑：康　锋）