基于多模态融合技术的用户画像方法

张壮[1]，冯小年[2]，钱铁云[1,*]
1 武汉大学计算机学学院，*通信作者
2 中国电力财务有限公司

摘要 针对当前用户画像工作中各模态信息不能充分利用的问题，本文提出一种跨模态学习思想，进而设计了一种基于多模态融合的用户画像模型。主要思想包括：1）利用 Stacking 集成方法融合多种跨模态学习联合表示网络，对相应的模型组合进行学习；2）引入注意力机制，使得模型能够学习不同模态的表示对预测结果的贡献差异性。本文改进之后的模型具有精心设计的网络结构和目标函数，能够生成一个由特征级融合和决策级融合组成的联合特征表示，从而合并不同模态的相关特征。本文在真实数据集上进行了大量的实验，结果表明本文提出的模型优于当前最好的基线方法。

关键词 用户画像；模型组合；stacking；跨模态学习联合表示；多层多级模型融合
中图分类号 TN914

User Profiling Based on Multimodal Fusion Technology

Zhuang Zhang[1], Xiaonian Feng[2], Tieyun Qian[1,*]

1 School of Computer Science, Wuhan University，* Contact author

2 China Power Finance Co., Ltd.

Abstract Existing studies in user profiling are unable to fully utilize the multimodal information. This paper presents a cross-modal joint representation learning network, and develop a multi-modal fusion model. The main ideas are two-fold: 1) we adopt a stacking method to learn the joint representation network which fuse the cross-modal information, and 2) we introduce attention mechanism to automatically learn the contribution of different modal to the prediction task. Our model has a well defined loss function and network structure, which enables combining the related features in various models by learning the joint representations after feature-level and decision-level fusion. We conduct extensive experiments on real data sets. Experiments show that our model outperforms the baselines.
Key words User profiling; model combination; stacking; cross-modal learning joint representation; multi-layer and multi-level model fusion

近年来，社交媒体用户画像研究吸引了越来越多研究人员的关注。在现有用户画像建模技术中，如何融合多个用户数据源或模态以获得更精确的用户画像的工作相当有限。目前的方法普遍存在如下缺陷：一方面，部分用户画像研究工作仅就单一的模态进行研究，难以全面刻画用户的特征；另一方面，大多数利用多模态的相关工作仅在特征级或决策级集成数据源[1-3]，即使少数工作能够进行两个级别的融合[4,5]，但仍然对多个数据源的深度融合缺乏探索。

本文提出一种基于多模态融合的用户画像模型 UMF-SCF（**U**ser Profiling based on **M**ultimodal **F**usion，其中 SCF 表示 **S**tacking, model **C**ombination, and **F**usion），用于解决用户画像建模中各模态不能深入交互的问题。

1 UMF-SCF 模型

1.1 UMF-SCF 模型总体架构

本文任务为用户画像，主要对用户性别和年龄属性进行分类。用户特征集 D 包含两个部分——有标签用户特征集 D^A（即训练集）和无标签用户特征集 D^U（即测试集），显然满足 $D = D^A \bigcup D^U$。用户性别和年龄的分类即是给定一系列训练集用户 $D^A \in D$ 进行训练，继而为无标签用户 $D^U \in D$ 打标签的过程，标签集为 C。其中，用户的信息包括用户的特征词、关注关系抽取的链接、用户生成的文本以及微博数量、发布时间等统计信息。

我们将一个用户所有分词后的微博合并作为该用户的文本信息，通过计算 tf-idf 值和卡方检验的方法得到用户的特征词嵌入，使用网络嵌入方法 LINE 获得关系网络中用户节点的嵌入，使用 word2vec 训练出词向量，并对用户词向量采用 tf-idf 加权得到用户的文本嵌入，最后手工提取了用户发布微博、评论、转发、关注者、发微博时间、表情符等统计类特征，利用 LIWC 心理学词典提取了用户的心理学特征，将它们作为手工特征嵌入。可以将数据集中的所有用户用一个用户特征集 $D = (F, L, T, S)$ 来表示，特征集的每个用户拥有特征词 F、文本特征 T、手工特征 S 以及连接用户彼此的链接特征 L 四种模态的特征。本文

利用这四种模态的信息来进行集成，为了包含数据源共享与不共享的表示，本文使用了模型组合的策略，并针对特定的模型组合构建相应的跨模态学习联合表示网络。模型组合一共分为（*F*，*L*，*T*，*S*）、（*F*，*L*，*T*）、（*F*，*L*，*S*）、（*F*，*T*，*S*）、（*L*，*T*，*S*）、（*F*，*L*）、（*F*，*T*）、（*F*，*S*）、（*L*，*T*）、（*L*，*S*）和（*T*，*S*）共十一个组合。其中，（*F*，*L*，*T*，*S*）对应的跨模态学习联合表示网络我们称之为 FusionFLTS，其它组合以此类推。

本节介绍 UMF-SCF 模型的总体架构——多层多级融合模型。“多层”一方面是利用 stacking 产生的双层融合模型，另一方面，stacking 第一层所使用的跨模态学习联合表示网络包含嵌入层、交互层、决策层三层，模型级融合就发生在其中的交互层。“多级”是指融合的模式同时包含特征级和决策级融合。

具体地，UMF-SCF 模型第一层运用跨模态学习联合表示网络，对模型组合进行学习，可以生成一个由特征级融合和决策级融合特征组成的联合或共享的特征表示，属于模态级融合的方法。该方法通过跨模态学习联合表示，使不同模态的特征产生深入的交互，允许模型在特征级融合，并使用注意力机制对提取的特征使用模态级的加权评分，以此来表明不同模态的重要性差别，同时也保持了特征级和决策级融合的优点。第二层是决策级融合，将第一层产生的预测概率输入到神经网络，得到最后的预测结果。

下面详细介绍模型组合的方法。设 $DS=\{D_1,D_2,\cdots,D_k\}$ 是本文要集成的 k 个数据源的有限集合。论文[6]提出了幂集组合的方法，本文采用的模型组合方法去掉了空集和单个模态的子集（不会产生模态交互）。为了合并特性和数据源之间的相关性，本文使用了跨模态学习联合表示的方法，对所有模型组合进行模型级融合。然后，本文将它们的预测结果与 ensemble 方法相结合，进行决策级融合。因此，本文的 UMF-SCF 模型是一种混合的数据集成模型。给定 k 个数据源的集合 DS，我们排除掉空集和单个模态元素的子集，所考虑的 DS 子集的个数为 2^k-k-1。对于得到的模态组合总集合中的每一个组合，我们为它建立相应的跨模态学习联合表示网络。

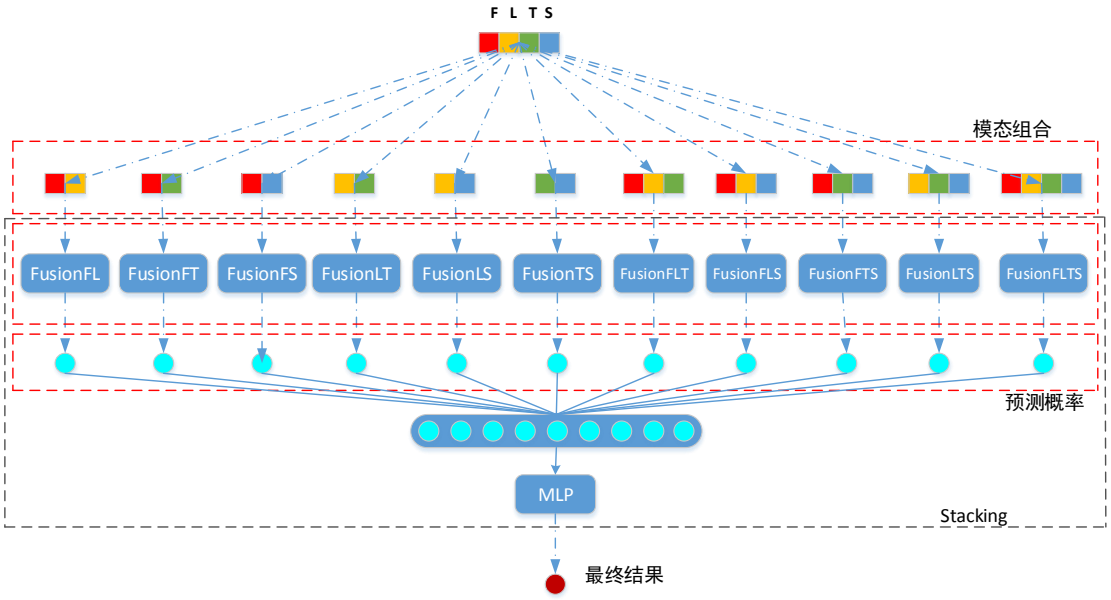


图 1 UMF-SCF 整体架构图

Fig 1 Overall architecture diagram of UMF-SCF

模型图如图 1 所示，首先学习到用户的四种数据源嵌入，然后将上述四种数据源嵌入进行模型组合，产生十一种组合形式，所有模型组合输入到多模态融合模型中，每一种模型组合使用相应的跨模态学习联合表示网络来学习，输出类别的预测概率，将上述预测概率拼接得到 stacking 第二层的训练样本和测试样本，最后输入到一个神经网络做分类。

本文选择跨模态学习联合表示网络来实现用户画像模型，其原因在于：首先，通过使用模态之间的共享表示，可以很容易地组合各种数据源。其次，将数据源与非线性函数相结合，已被证明可以增强学习过程。第三，跨模态学习联合表示能够使各模态表示深入交互，有助于各模态彼此学到其它模态的互补信息。

1.2 FusionFLTS 模型

本文工作受跨模态学习联合表示模型 CrossTL[7]启发，原始 CrossTL 模型用于学习文本与链接之间联合表示。CrossTL 采用单层的交互模式，对模态的融合比较有限。为了达到了更深层次交互，本文改进了 CrossTL 模型，将其设计为双层的交叉融合结构，因此第一交互层的模态表示包含了其它模态原始的信息，

第二交互层的模态表示不仅包含了其它模态第一交互层的信息，还包含了原始的信息。此外，我们简化了 CrossTL 的模型参数，并将其扩展到学习 4 个模态的联合表示 FusionFLTS。

图 2 展示了本文的多模态融合模型 FusionFLT 的架构。为简洁起见，我们仅展示了应用于三个模态的 FusionFLT 模型。可以看到，FusionFLT 模型与原始模态的表示是相互独立的，此外，多模态间产生了深层的交互，从而有助于每个模态学习其它模态补充的信息。

FusionFLTS 模型架构：

为了充分利用其他模态表示中的互补信息，本文设计了一个三层跨视图学习体系结构 FusionFLTS。FusionFLTS 模型顶部是嵌入层，中间是交互层，底部是决策层，用于将表示转换为预测的类别概率。为了加强模型融合的效果，使模型特征更充分交互，本文将位于模型中间的交互层设计为两层结构。注：因两层交互层算法一致，为避免模型图过于复杂，图 2 仅展示一层交互层的模型。

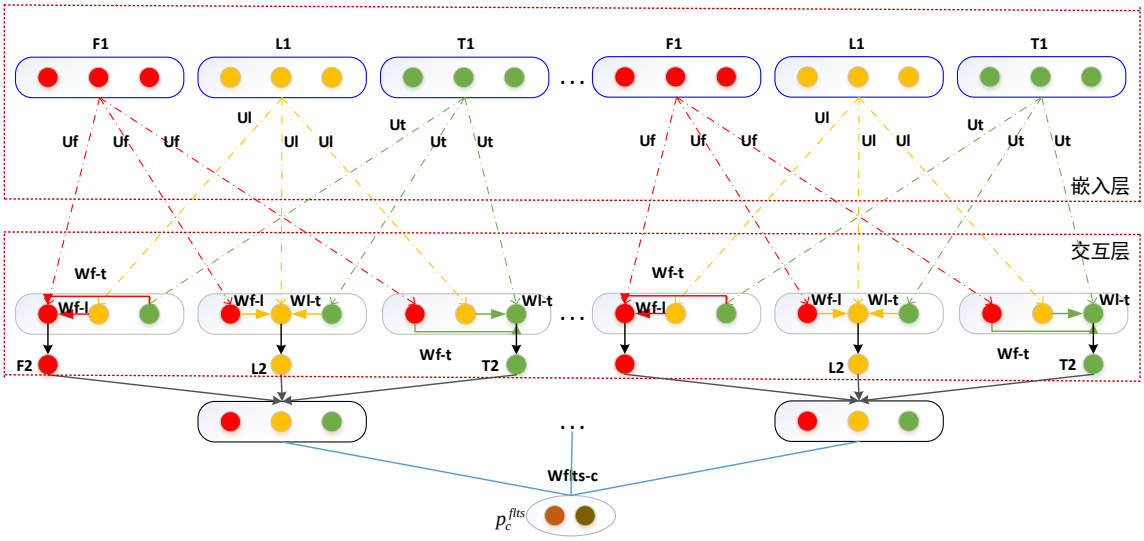


图 2 FusionFLT 网络架构图

Fig 2 Network architecture diagram of FusionFLT

各层详细介绍如下：

- 嵌入层：根据现有工作的理论和实验经验，先对每个输入进行预训练，从而得到不同模态的嵌入是有效的方法。本文嵌入层包含用户特征词的嵌入、链接源的图形嵌入、文本源的文本嵌入和手工特征嵌入。指定一个输入样本的特征词嵌入、图形嵌入、文本嵌入与手工特征嵌入分别为 $F_1(a)$、$L_1(a)$、$T_1(a)$ 和 $S_1(a)$，可以依次简写为 F_1、L_1、T_1 和 S_1。

- 交互层：现有学习联合表示的工作主要是直接拼接各表示。然而，各模态拥有各自的语义，通过简单的拼接无法将其语义有效混合。为了解决此类问题，我们引入了交互层，能够有效减少不同模态间的语义鸿沟，同时充分利用其它模态的互补信息。

首先，使用隐藏层将各模态表示进行转换。其次，针对每个已经转换过的模态表示，补充了其与其它转换过的模态嵌入的关联表示。由此，可以得到第一（二）层交互层之后的特征词表示 F_2，它包含了与其它模态对应的交互信息。相似地，可以得到其它模态经过第一（二）层交互层的表示。得到各模态在交互层之后的表示如公式（7）所示：

$$M^1{}_b = tanh\left(U_{m^1} \cdot M^1{}_a + W_{m^1m^2} \cdot U_{m^2} \cdot M^2{}_a + W_{m^1m^3} \cdot U_{m^3} \cdot M^3{}_a + W_{m^1m^4} \cdot U_{m^4} \cdot M^4{}_a\right), \quad (1)$$

其中，M^* 可分别表示四种模态嵌入 F、L、T、S，若公式表示的是第一层交互层得到的表示，则 a 为 1、b 为 2，否则 a 为 2、b 为 3，$tanh$ 是神经网络激活函数，U_* 是第 a 层交互层隐藏层中相应模态嵌入对应的权重，W_{**} 分别是第 a 层交互层中相应两模态嵌入的关联权重矩阵。

- 决策层：经过交互层，对于每个用户都可以得到 5 个高层次的表示，分别为 F_3、L_3、T_3、S_3 和 $F_3 \oplus L_3 \oplus T_3 \oplus S_3$。虽然上述表示都包含来自特征词、链接、文本和手工特征的融合信息，但每种表示仍然有其独特的聚焦点。例如，T_3 主要是用户级文本表示，同时带有附加的特征词、链接和手工特征信息，其它表示同理。因此，将上述五种高层次、跨模态学习到的表示应用于用户性别、年龄的分类任务中。在多模态融合模型的最后，本文使用了一个线性层来将五种表示映射到它们的标签类别空间，F_3、L_3、T_3、S_3 和 $F_3 \oplus L_3 \oplus T_3 \oplus S_3$ 对应的标签类别空间分别是 C^f、C^l、C^t、C^s 和 C^{flts}。然后，使用一个 softmax 层来获

得类别 c 的概率分布。

线性映射定义如公式 **错误!未找到引用源。** 所示：

$$C^* = W_{*-C} \cdot A_3 + b_{*-C}\,, \tag{2}$$

$$p_c^* = \frac{exp\left(C_c^*\right)}{\sum_{m=1}^{M} exp\left(C_m^*\right)}\,, \tag{3}$$

其中，$A \in \left\{F_3, L_3, T_3, S_3, F_3 \oplus L_3 \oplus T_3 \oplus S_3\right\}$，$C^*$ 表示 A 对应的标签类别空间，W_{*-C} 表示 A 对应的线性层权重矩阵，b_{*-C} 是相应的偏置值。

softmax 层定义如公式 **错误!未找到引用源。** 所示：

其中，p_c^* 表示 F_3、L_3、T_3、S_3 和 $F_3 \oplus L_3 \oplus T_3 \oplus S_3$ 预测类别 c 的概率，M 是类别个数，C_c^* 表示标签类别空间的第 c 个类别。

FusionFLTS 目标函数：

为了充分利用 FusionFLTS 学到的表示，我们设计了由多种损失组成的目标函数，主要包括三个部分：各模态表示的损失 J_{mod}、联合表示的损失 J_{flts} 和模态表示间的一致性损失 J_{dif}。其中 J_{mod} 包括各模态表示的损失以及联合表示的损失，J_{dif} 包含任意两种模态组合的一致性损失。值得一提的是，J_{flts} 强调各模态的联合效果，并且 J_{dif} 被用来平衡损失函数各部分的比例。

为了方便公式的表达，我们使用 $P_{mod} = \left\{p_c^f, p_c^l, p_c^t, p_c^s\right\}$ 表示计算 J_{mod} 的预测概率的集合，使用 $P_{dif} = \left\{\left(p_c^f, p_c^l\right), \left(p_c^f, p_c^t\right), \left(p_c^f, p_c^s\right), \left(p_c^l, p_c^t\right), \left(p_c^l, p_c^s\right), \left(p_c^t, p_c^s\right)\right\}$ 表示计算 J_{dif} 的预测概率的集合。使用交叉熵作为损失，则三个部分的损失函数可以被定义为公式 **错误!未找到引用源。** - **错误!未找到引用源。** 所示。

$$J_{mod} = -\sum_{p \in P_{mod}} \sum_{a \in D^A} \sum_{c=1}^{C} p_c^g(a) \cdot log\left(p(a)\right)\,, \tag{4}$$

$$J_{flts} = -\sum_{a \in D^A} \sum_{c=1}^{C} p_c^g(a) \cdot log\left(p_c^{flts}\left(F_3(a) \oplus L_3(a) \oplus T_3(a) \oplus S_3(a)\right)\right)\,, \tag{5}$$

$$J_{dif} = -\sum_{(p_1, p_2) \in P_{dif}} \sum_{a \in D^A} \sum_{c=1}^{C} p_1(a) \cdot log\left(p_2(a)\right)\,, \tag{6}$$

$$\begin{aligned} J_{FLTS} = & \sum_{i=0}^{3} w_s[i] \cdot List_J^{mod}[i] + w_s[4] \cdot J_{flts} \\ & + \sum_{j=5}^{10} w_s[j] \cdot List_J^{dif}[j-5] + w_s[11] \cdot reg_loss \end{aligned}\,, \tag{7}$$

公式 **错误!未找到引用源。** - **错误!未找到引用源。** 中，a 是有标签数据 D^A 中的一个用户，p_c^g 是类别 c 的真实标签。

如果简单地将上述损失相加来构成最终的损失，则不能揭示不同模态的实际重要性，其原因在于：相同的模态对不同的用户属性分类任务可能有不同的贡献。例如，通过对用户微博文本数据进行分析，发现手工提取的各种统计类特征对于预测用户的年龄更有利，不同年龄段的用户发微博的时间、活跃程度、使用的表情符等, 上述手工特征都是具有分辨性的特征，而文本内容对用户性别的分类更重要。即使是同一任务，不同的模态也可能对预测的结果产生不同的影响。因此本文引入注意力机制，即引入权重系数列表 *w_s*，总体损失是以上损失函数的线性加权合并，如公式 **错误!未找到引用源。** 所示。

公式 **错误!未找到引用源。** 中 *reg_loss* 为 L2 正则化损失，$List_J^{mod}$ 表示 J_{mod} 各项损失的列表，$List_J^{dif}$ 表示 J_{dif} 各项损失的列表。权重系数列表 *w_s* 用来平衡各模态表示和它们的联合表示以及一致性。如果将 *w_s* 中的权重均设为 1，则代表各模态表示、它们的联合表示以及一致性简单相加。应用于三个和两个模态的跨模态学习联合表示模型（FusionFLT、FusionFL 等）的算法同理。

FusionFLTS 从以下三方面提高了学习的效果：

- 从输入的层面讲，FusionFLTS 可以容纳任何类型的模态嵌入。本文选择特征词、网络、文本和手工特征表示，它们很好地捕获数据属性作为输入。
- 从结构的层面讲，FusionFLTS 使用了神经网络架构，可以循环地学习每个模态和对应的其它模态的表示，以全面和准确地描述数据。

- 从目标函数层面讲，FusionFLTS 通过包含多种类型的损失来限制错误率，应用于四种模态的 FusionFLTS 共包含 11 种损失函数，各模态两两组合间的差异性损失都加入损失函数。此外，本文在目标函数各模态损失及一致性损失上使用了注意力机制，以确保不同预测效果的模态具有不同的重要程度。

2 实验及分析

本节将详细展现本文的实验结果和进一步分析，并且回答以下几个关键研究问题。

- 与其他优秀的相关工作相比，本文提出的 UMF-SCF 模型预测结果如何？
- 使用多模态融合的方法来进行用户画像是否确实有益于所研究的问题？
- UMF-SCF 的每个组件如何对预测性能做出贡献？

为了解答上述问题，本节首先提供实验设置的细节，然后，我们从模型对比效果、模型简化分析、参数分析等方面依次回答了上述三个问题。

2.1 实验设置

为了评测本文所提出的 UMF-SCF 模型的性能，我们在 SMP CUP 2016 微博用户画像数据集上对模型进行了训练和测试。由于数据集微博文本数据比较复杂，包含很多对于推测用户属性无用的信息，我们对原始数据进行了预处理。主要预处理过程包括分词、去停用词、缺失值补零、去除用户重复微博等。

在本文中定义了四种数据源嵌入：用户特征词表示、关系网络中的用户节点表示、用户级文本表示和手工特征表示。本文在上述数据集上获得了以上数据源嵌入。

在本文预测用户性别、年龄任务中，为了与 SMP CUP 2016 竞赛评估指标一致，本文也选取了准确率 (Accuracy) 作为性别和年龄预测任务的评测指标。此外，本文还选用了分类任务中常用的精确率 (Precision) 、召回率 (Recall) 和 F1 作为评测指标。对于三分类的年龄预测任务，采用宏平均的 F1 作为评价指标。

为了验证本文提出的模型 UMF-SCF 的有效性，以及回答本节提出的前两个问题，我们选取了一些基线方法作为对比，具体介绍如下：

1、**特征词**：考虑到卡方检验的低频词缺陷，本文计算出用户微博文本与平台文本中所有词的 tf-idf 值，根据该值计算每个词语与类别的卡方值，选取微博文本中卡方值最高的 5000 个词与平台文本中卡方值最高的 300 个词，作为用户文本的特征词，使用 SVD 降维的方法将特征词嵌入降维至 300 维，即为用户特征词表示。

2、**LINE**[8]：该基线方法是本实验中学习用户关注关系网络的用户节点表示时所使用的方法。

3、**Word2Vec**[9]**+TF-IDF**：使用 word2vec 预训练的词向量作为模型中的词向量矩阵，通过计算用户文本所有词（包含微博文本与平台文本）的 tf-idf 值为词向量赋予权重，得到用户的 300 维文本表示。

4、**手工特征**：本实验提取了用户关注者数目、评论数目、发帖总数等统计信息，以及表情符特征、用户心理学特征等，使用卡方检验的方法提取了与用户性别、年龄预测最相关的 300 个手工特征。

针对基线方法 1-4，为了与 Concat+SVM 和 Concat+MLP 作对比，最后本文使用支持向量机来分类。我们也尝试了使用多层感知机分类，分类效果与使用支持向量机类似。

5、**Concat+SVM**：该方法是将通过数据元嵌入方法得到的各模态嵌入拼接，然后输入到支持向量机分类器进行分类。

6、**Concat+MLP**：该方法是将通过数据源嵌入得到的各模态嵌入拼接，然后输入到多层感知机 (MLP, MultiLayer Perceptron) 进行分类。

7、**UHAN**[10]：为了研究涉及视觉和文本模态的非结构化特性，并进一步考虑从流行度预测的多模态中学习有效的表示，论文作者提出了一种基于用户引导的分层注意力网络 (UHAN) 模型，该模型不仅能够学习每种模态的有效表示，而且能够融合它们，在用户嵌入的指导下得到一个完整的多模态表示。

8、**UDMF**[6]：作者提出的混合用户画像框架利用了模态之间的共享表示，在特征层集成了三个数据源，并结合了在决策层操作每个数据源组合的独立网络的决策。

9、**SMP CUP 2016 竞赛第一名（SMP CUP 1st）**：作者利用了用户特征词特征、关系网络特征、文本特征三种模态，但做法上与本文方法略有不同。在学习用户特征词表示时，作者为了降低噪声（广告词、分享、新闻等）的影响，每个维度使用了 0/1 表示，而本文使用了 tf-idf 表示，可以更精确地表征用户。在得到三种特征词特征、关系网络特征、用户文本特征三种模态之后，作者对三种表示 stacking 进行集成，并输出最终预测结果。

以上的基线方法中，对于并没有使用本文数据集的方法，如 UDMF、UHAN 等，我们在本文数据集上的实现了上述方法，并保证实现的所有基线方法与原文中的参数设置相同。

2.2 结果分析

为验证本文提出的 UMF-SCF 模型的性能，我们设置了对比实验来验证，此外，我们设置了模型简化

实验，通过对实验的分析来回答本节前言部分提出的三个关于本文模型和用户画像研究的关键问题。

2.2.1 对比实验

在 SMP CUP 2016 微博用户画像数据集上各项基线方法准确率的实验结果见表 1、表 2 所示。

关于第一个问题：由表 1 和表 2 可以看出，本文的 UMF-SCF 模型在实验数据集的用户性别、年龄预测的任务上取得了领先的效果。

关于第二个问题：特征词、LINE、W2V+TF-IDF 和手工特征的方法是单个模态嵌入的方法，对比 Concat+MLP、UHAN、UDMF、SMP CUP 1st 以及本文的模型 UMF-SCF 等多模态融合的方法，可以发现多模态融合的方法明显优于单模态嵌入的方法。由此可见使用多模态融合的方法来进行用户画像确实有益于所研究的问题。

此外，LINE 方法效果差的原因在于 SMP CUP 2016 微博用户画像数据集中训练集、测试集有部分用户的链接表示缺失。在实验中，对于没有在关注关系网络中的用户，我们采用特征补零作为用户表示的策略，该策略在一定程度上使得链接表示的效果不能发挥到最好。

表 1 微博用户画像数据集上各基线方法的年龄预测结果

Table 1 The age prediction results of each baseline method in the weibo user profiling data set

基线方法	Accuracy	Precision	Recall	F1
特征词	0.622	0.569	0.449	0.457
LINE	0.574	0.191	0.333	0.243
W2V+TF-IDF	0.613	0.530	0.478	0.489
手工特征	0.620	0.536	0.426	0.417
Concat+SVM	0.606	0.530	0.514	0.522
Concat+MLP	0.645	0.597	**0.526**	0.559
UHAN	0.643	0.581	0.471	0.520
UDMF	0.659	0.589	0.491	0.536
SMP CUP 1st	0.648	0.609	0.483	0.539
UMF-SCF	**0.681**	**0.676**	0.521	**0.589**

表 2 微博用户画像数据集上各基线方法的性别预测结果

Table 2 The gender prediction results of each baseline method in the weibo user profiling data set

基线方法	Accuracy	Precision	Recall	F1
特征词	0.855	0.879	0.946	0.912
LINE	0.822	0.841	0.948	0.891
W2V+TF-IDF	0.869	**0.911**	0.922	0.916
手工特征	0.802	0.826	0.946	0.882
Concat+SVM	0.811	0.897	0.860	0.878
Concat+MLP	0.866	0.879	**0.963**	0.919
UHAN	0.856	0.903	0.916	0.909
UDMF	0.870	0.888	0.955	0.920
SMP CUP 1st	0.883	0.901	0.960	**0.930**
UMF-SCF	**0.887**	0.907	0.955	**0.930**

Concat+SVM 效果较差，虽然利用了多模态的信息，但是简单拼接并不能使得多个模态间产生交互，使用支持向量机分类达到的效果反而差于特征词、W2V+TF-IDF 等单模态嵌入的方法，说明不能合理使用多模态融合，反而会削弱其所利用的单模态原始的较优效果。

Concat+MLP 虽然也只是简单拼接多模态嵌入，但达到的效果远超 Concat+SVM。由此可见由于使用了神经网络，Concat+MLP 使多种模态的数据之间产生了交互，在一定程度上发挥了多模态融合的优势。

UHAN 模型在本实验中效果一般。考虑本文实验数据集的关系网络中约 45%的用户节点缺失的问题，该模型引入用户向量的方法并不能在本文使用的数据集上奏效。

UDMF 模型与本文的 UMF-SCF 模型相比，由于只采用了简单的神经网络来融合各个数据源，在多模态表示的深度交互方面有所欠缺。同时 UDMF 模型采用的多任务学习方式在本实验上并没有效果的提升。经分析，我们认为是由于性别、年龄两个任务不具有很大的相关性导致。

本文提出的 UMF-SCF 模型由于设计了多层多级模型融合，使多模态数据可以在第一层完成模态级融

合。跨模态学习联合表示允许模型在特征级融合，并使用注意力机制对提取的特征使用模态级的加权评分，以此来表明不同模态的重要性差别，同时也保持了特征级和决策级融合的优点。由于使用了 stacking 的集成方法，模型可以在第二层进一步进行决策级融合，更增加了模型间的交互。

2.2.2 模型简化学习

为了回答本节前言部分提出的第三个问题，本文对 UMF-SCF 模型进行了模型简化学习，即在去掉或替换模型的关键组件的情况下，通过观察分析简化的模型的效果，来说明模型各组件的真实作用。本文考虑 UMF-SCF 模型的四个主要的模块，来检测它们对于最终预测的贡献，它们分别是：跨模态学习联合表示网络、stacking 决策级集成方法、模型组合以及注意力机制。具体地，本文设计了以下四个简化模型实验：

表 3 微博用户画像数据集上各项简化模型的准确率

Table 3 Accuracy of simplified models in weibo user profiling data set

方法	年龄	性别
Concat+Stacking+DNN	0.656	0.872
Fusions+Attention	0.658	0.874
FusionFLTS+Stacking+Attention	0.665	0.878
除去注意力机制	0.666	0.881
全部架构	**0.681**	**0.887**

1、**Concat+Stacking+DNN**：与 UMF-SCF 模型全部架构相比，相当于将模型中使用的跨模态学习联合表示网络替换为普通的神经网络。

2、**Fusions+Attention**：与 UMF-SCF 模型全部架构相比，相当于去除了 stacking 集成的模块，从融合架构角度讲，缺失了第二层决策级融合。

3、**FusionFLTS+Stacking+Attention**：与 UMF-SCF 模型全部架构相比，相当于去除了除 FusionFLTS 外的其它跨模态学习联合表示网络。

4、**除去注意力机制**：该方法将 UMF-SCF 模型的目标函数各项损失系数全置为 1，即将各模态和各一致性损失的重要程度一视同仁。

表 3 展示了在 SMP CUP 2016 微博用户画像数据集上各项简化模型准确率的实验结果。

关于第三个问题：由表 3 展示的模型简化学习的结果，可以看到 UMF-SCF 模型各个组件都发挥着不可或缺的作用。UMF-SCF 模型全部架构与 Concat+Stacking+DNN 相比，充分表明了本文的跨模态学习联合表示网络相对于普通神经网络的优越性。Fusions+Attention 的效果表明使用 stacking 集成的作用同样明显，可见第二层决策级别的数据源集成对于模型效果的提高至关重要，该层的集成也是本文设计的多层多级模型融合的优势所在。相比于前两者，Fusion+Stacking+Attention 与除去注意力机制两种简化的模型的影响相对稍小一些，在性别预测的任务上与 UMF-SCF 全部架构已经非常接近，但在年龄预测的任务上依然存在差距。可见相对于只使用 stacking 集成一种组合方式进行分类的方法，集成多种模型组合的方法更加有利于模型的性能，注意力机制亦然。

2.2.3参数分析

本小节探讨 UMF-SCF 模型中的两个对模型效果有重要影响的参数，分别是正则化参数和学习率。

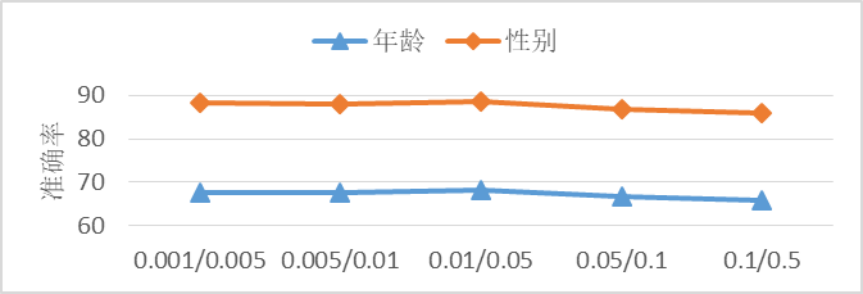


图 3 正则化参数对分类结果的影响

Fig 3 The influence of regularization parameters on classification results

正则化参数的影响：由于 UMF-SCF 模型采用了多层多级模型融合的方法，网络参数较多，模型较为复杂，为防止模型产生过拟合现象，采用合理的正则化方法十分必要。针对不同模型组合，本文在设置正

则化参数方面略有不同。图 3 给出了对应不同正则化参数 λ 的 UMF-SCF 模型的准确率，单位为%。其中，0.001/0.005 表示用于融合两个模态的跨模态学习联合表示网络使用的正则化参数是 0.001，用于融合三个和四个模态的跨模态学习联合表示网络使用的正则化参数是 0.005，其它同理。

由图 3 可知，正则化参数 λ 取 0.01/0.05 时 UMF-SCF 模型可以达到最佳效果，在本文的实验中采取的也是上述参数组合。

学习率的影响：学习率是神经网络训练中的一个重要的超参数，它控制着神经网络基于损失梯度调整网络权重的速度。我们选取了 0.0001、0.001、0.01 和 0.1 四个学习率的设置来实验，实验结果如图 4 所示，单位为%。可以看到，当学习率为 0.01 时模型的效果是最佳的，所以本文所使用的所有神经网络都采用了该学习率。

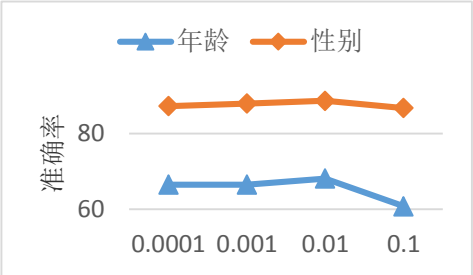


图 4 学习率对分类结果的影响

Fig 4 The influence of learning rate on classification results

3 结语

用户属性预测是用户画像研究的一项重要课题。本文提出了一种跨模态学习的思想，通过学习多模态联合表示来融合多种数据源，我们称之为 UMF-SCF 模型。该模型是多层多级的融合模型，包含三个简单而有效的方法，即模型组合策略、跨模态学习联合表示网络和 stacking 集成方法，可以通过将不同数据源的用户数据整合到联合表示来完成多模态融合，从而更好地实现用户属性预测。UMF-SCF 模型在特征级和决策级融合不同的模态，根据用户生成的内容和社会关系内容，准确预测社交媒体用户的多个属性。

本文在来自新浪微博的超过四千个用户的用户建模任务上对 UMF-SCF 体系结构进行了评估，构建了四个数据源。通过对新浪微博用户的年龄、性别进行预测，得出用户的用户画像。结果表明，本文的 UMF-SCF 模型提高了数据源组合的学习能力。本文的实验取得了领先于其它基线方法的结果，其中年龄预测任务准确率得分超过 0.68，性别预测任务准确率得分超过 0.88。在进一步的工作中，我们将与 GCN 下得到用户画像的构建方法相比较。

致谢： 本文受国家自然科学基金项目(61572376)资助。

参考文献

[1] Xiao J, Hao Y, He X, et al. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks[J], 2017.

[2] Wei H, Zhang F, Yuan N J, et al. Beyond the Words: Predicting User Personality from Heterogeneous Information[C]. Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017.

[3] Wollmer M, Weninger F, Knaup T, et al. YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context[J], 2013, 28(3): 46-53.

[4] Gu Y, Yang K, Fu S, et al. Hybrid Attention based Multimodal Network for Spoken Language Classification[C]. international conference on computational linguistics, 2018: 2379-2390.

[5] Gu Y, Chen S, Marsic I. Deep Mul Timodal Learning for Emotion Recognition in Spoken Language[C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018: 5079-5083.

[6] Farnadi G, Jie T, Cock M D, et al. User Profiling through Deep Multimodal Fusion[C], 2018.

[7] You Z, Qian T. Learning a Joint Representation for Classification of Networked Documents[C]. international conference on neural information processing, 2018: 199-209.

[8] Tang J, Qu M, Wang M, et al. LINE: Large-scale Information Network Embedding[J], 2015: 1067-1077.

[9] Mikolov T, Chen K, Corrado G S, et al. Efficient Estimation of Word Representations in Vector Space[J], 2013.

[10] Zhang W, Wang W, Wang J, et al. User-guided hierarchical attention network for multi-modal social image popularity

prediction[C]. Proceedings of the 2018 World Wide Web Conference on World Wide Web, 2018: 1277-1286.