基于深度学习的跨社交网络用户匹配方法

马 强 戴 军*

(西南科技大学信息工程学院 绵阳 621010)

摘 要：现有基于时空信息的跨社交网络用户匹配方案，存在着难以耦合时空信息、特征提取困难问题，导致匹配精度下降。该文提出一种基于深度学习的跨社交网络用户匹配方法(DLUMCN)，首先对用户签到数据进行时空尺度的网格映射，生成包含用户特征的签到矩阵集合，对其归一化后构成用户签到图。然后采用卷积从签到图中生成高维度的时空特征图，利用深度可分离卷积对特征图权重变换和特征融合，对特征图1维展开获得特征向量。最后利用全连接前馈网络构建分类器并输出用户匹配评分。通过在两组真实社交网络的数据集上进行实验验证，实验结果表明，与现有相关算法相比，所提算法在匹配的准确率以及F1-值均得到提升，验证了所提算法的有效性。

关键词：跨社交网络；用户匹配；深度学习；签到相似度

中图分类号：TN915; TP391 **文献标识码**：A **文章编号**：1009-5896(2023)07-2650-09

DOI: 10.11999/JEIT220702

User Matching Method for Cross Social Networks Based on Deep Learning

MA Qiang DAI Jun

(*School of Information Engineering*, *Southwest University of Science and Technology*, *Mianyang* 621010, *China*)

Abstract: The existing spatio-temporal information based user matching schemes for cross social networks have problems of spatio-temporal information decoupling and feature extraction difficulties, which result in a decrease in matching accuracy. A Deep Learning based User Matching method for Cross social Networks (DLUMCN) is proposed. Firstly, grid mapping at the spatio-temporal scale is carried out on the user sign-in data. The sign-in matrix set is generated, which contains user characteristics. User sign-in map is formed after normalization. Secondly, the convolution is used to generate high-dimensional spatio-temporal feature maps from the user sign-in map. The weight transformation and feature fusion of feature maps are carried out by deep separable convolution. The feature vector is obtained by one-dimensional expansion of feature maps. Finally, the fully connected feedforward network is used to build a classifier and output the user matching score. Experimental results on two sets of datasets of real social networks show that the proposed method has improved matching accuracy and F1-value, compared with the existing related methods. The effectiveness of the proposed method is demonstrated.

Key words: Cross social networks; User matching; Deep learning; Sign-in similarity

1 引言

现代社交网络的用户数量呈现爆炸式增长，由于社交平台的差异性，用户会在多个平台注册账号。出于隐私保护，目前缺乏链接各平台用户身份的用户画像，跨社交网络的用户匹配逐渐成为一个热门研究领域。通过相关算法准确链接用户，以衍生例如广告推送、恶意用户识别等实际应用[1]。由于位置服务和实时定位的移动设备的出现，社交网络具备海量用户签到数据，这提高了基于签到的用户匹配方案的可行性，并促进相关技术迅速发展。相较于档案、好友网络等其他类型的用户数据，签到数据结构简洁，时空特征丰富，具备高真实性和难模仿性，这些特点使基于签到的用户匹配方案具有很大的研究潜力[2]。

相关研究者从不同技术角度出发，提出了各种

收稿日期：2022-05-31；改回日期：2022-08-27；网络出版：2022-09-09
*通信作者： 戴军 arturia_pendragon@163.com
基金项目：国家自然科学基金 (62071170, 62072158)，河南省杰出青年科学基金(222300420006)，河南省高校科技创新团队支持计划(21IRTSTHN015)，西南科技大学博士基金(17zx7158)

Foundation Items: The National Natural Science Foundation of China (62071170，62072158), Henan Science Foundation for Distinguished Young Scholars (222300420006), Henan Support Plan for Science and Technology Innovation Team of Universities (21IRTSTHN015), The Doctoral Foundation of Southwest University of Science and Technology (17zx7158)

方案。张树森等人[3]从用户的多媒体内容分析社交行为，从而识别个人和组织用户，为跨社交网络用户匹配提供了新的思路。Hao等人[4]从网络和物理空间对用户建模，将网络空间的数据转换为网格，通过物理空间建模丰富用户数据提高准确率，但没有充分利用用户时空信息。Chen等人[5]基于密度聚类和高斯混合模型提取轨迹时空特征，实现了高精度的用户数据建模，但忽略了时空特征的耦合关系。有研究表明[6]用户轨迹匹配准确率的主要决定因素是轨迹中共现节点，基于这一理论，Hao等人[7]将用户IP映射到坐标，和线下设备定位建立匹配框架，通过共现位置匹配用户轨迹。该方案的缺点是依赖轨迹数据量。为了解决海量数据下轨迹聚类的计算开销问题，He等人[8]基于局部敏感哈希将轨迹映射到BLSH(Binary-Search-Based Locality-Sensitive Hash)桶，通过桶内搜索降低匹配规模，快速聚类相似轨迹。该方案面向用户轨迹，需要保证相似轨迹映射到相同的桶，对哈希方法有很高的要求。用户的多源社交轨迹具有差异性，为解决此问题，王前东[9]基于最长公共子序列理论提出一种鲁棒的轨迹相似度量方法，以实现差异轨迹的相似性度量。Qi等人[10]构建签到分布最频繁TOP-N区域，将用户相似转换成区域相似，该方案降低了计算开销但忽略了用户时序特征。Han等人[11]基于图论，通过构建3部图，将用户匹配转换为图的最优划分问题。类似的研究还有，冯朔等人[12]通过最大公共子图对网络对齐以匹配用户。陈鸿昶等人[13]针对用户签到的时序特征，引入PV-DM模型[14]构建3种签到序列训练Doc2Vec模型，该方案增强了对轨迹向量化的泛化能力，但是没有考虑地理位置的语义信息。针对此问题，Xiao等人[15]利用位置的语义信息对用户签到建模，基于最大行程匹配算法计算用户相似度。Wang等人[16]考虑用户签到的关联性，使用语义序列对用户签到建模，通过公共序列估计用户相似度，说明了语义序列可以有效表征用户的转移模式。Li等人[17]基于核密度估算求解用户相似度并根据冲突签到校正，该方案提高了匹配准确率，但签到的权重计算复杂，不适用于大规模的匹配任务。张伟等人[18]引入循环神经网络和图神经网络从空间、时间、社交3个维度刻画用户，该方案引入用户网络拓扑，有效提高了匹配准确率，同时需要更多用户数据支持。除上述方案外，还有研究采用LDA主题模型[19]展开工作，Han等人[20]通过位置语义文本化用户签到，构建LDA模型获取用户的主题分布，采用KL散度得到用户相似度。Chen等人[21]通过分析用户签到时空共现频率，计算用户签到记录在频率下的分布相似性。Zhou等人[22]提出一种基于联合聚类的框架，时空维度的聚类同步进行并相互增强，并在真实数据集上进行实验证明了所提出框架的可行性。

这些方案已经取得一定成果，但相较于GPS轨迹，社交网络签到具备稀疏性和前后轨迹序列的弱关联性，耦合时空信息和时序特征提取困难，使目前方案面临一定的局限性。为解决此问题，本文提出基于深度学习的跨社交网络用户匹配方法(Deep Learning based User Matching method for Cross social Networks, DLUMCN)，通过从时空维度构建签到图对用户数据建模，基于卷积神经网络和前馈网络提取、融合特征信息，学习特征数据和用户匹配的潜在关联，最后输出匹配度评分。在两组真实数据集上的实验结果表明，与现有相关算法相比，所提方法在匹配准确率以及F1-值下均获得最佳的结果，证明了所提方法的有效性。

2　构建用户签到图

2.1　相关定义

定义 1　签到：用户在网络中生成具备时空信息的个人数据，例如打卡记录、登录信息等，这些数据被称为签到，用$s=(\text{lon}, \text{lat}, t)$表示。lon和lat分别为签到位置的经纬度，$t$为签到时间。用户所有签到按时间顺序构成签到集$S=\{s_1, s_2,\cdots, s_n, \text{id}\}$，其中$s_n$为第$n$次签到，id为用户身份标识号。

定义 2　用户匹配：给定不同网络签到集S_1和S_2，其中$\text{id}_1\in S_1$，$\text{id}_2\in S_2$。若S_1和S_2来自同一用户，则id_1和id_2构成匹配用户对。本文的目的是通过签到集准确匹配用户，挖掘来自同一用户的网络用户。

2.2　签到网格映射

同一时空中的签到在特征上具备共性，将其同一化以简化数据同时保留特征信息，便于数据处理并降低计算开销。通过网格映射能够用网格表示用户签到。根据映射规则的不同，本文提出两种网格映射方法。

方法 1　空间网格映射。给定签到$s=(\text{lon}, \text{lat}, t)$和空间域$(\text{lon}_{\min}, \text{lon}_{\max}, \text{lat}_{\min}, \text{lat}_{\max})$，其中$\text{lon}_{\min}$和$\text{lon}_{\max}$分别为空间域经度的最小和最大值，$\text{lat}_{\min}$和$\text{lat}_{\max}$分别为纬度的最小和最大值。签到的空间网格表示$g_s=(x_s, y_s)$，$x_s$和$y_s$分别为网格的水平和垂直序号，计算如式(1)

$$\left.\begin{aligned} x_s &= f\left(k\times\frac{\text{lon}-\text{lon}_{\min}+\det}{\text{lon}_{\max}-\text{lon}_{\min}+2\times\det}\right)\\ y_s &= f\left(k\times\frac{\text{lat}-\text{lat}_{\min}+\det}{\text{lat}_{\max}-\text{lat}_{\min}+2\times\det}\right)\end{aligned}\right\} \tag{1}$$

其中，f为取整函数，k为网格密度系数，它将映射空间划为k^2个网格，det为调节因子。根据空间域划定的尺度，分为全局和局部空间域，前者包含待匹配用户对的签到位置，后者仅包含待单一用户的签到位置。

方法 2 时间网格映射。给定签到s=(lon, lat, t)和时间域(t_{s}, t_{e})，其中t_{s}和t_{e}为时间域的起始时间戳和终止时间戳。签到的时间网格表示g_t=(x_t, y_t)，网格序号计算如式(2)

$$\left.\begin{aligned} x_t &= f\left(k\times\frac{|t-t_{\mathrm{s}}|+\det}{|t_{\mathrm{e}}-t_{\mathrm{s}}|+2\times\det}\right)\\ y_t &= f\left(k^2\times\frac{t}{|t_{\mathrm{e}}-t_{\mathrm{s}}|+2\times\det}-k\times x_t\right)\end{aligned}\right\} \tag{2}$$

其中，$|t-t_{\mathrm{s}}|$表示两个时间戳间隔时长，其他参数定义同式(1)。根据时间域划定的尺度，分为全局和局部时间域，前者包含待匹配用户对的签到时间，后者仅包含单个用户的签到时间。

2.3 签到矩阵填充

为了多尺度下的特征提取，构建签到矩阵集SMS = {$\boldsymbol{A}_{\mathrm{s1}}$, $\boldsymbol{A}_{\mathrm{s2}}$, $\boldsymbol{A}_{\mathrm{t1}}$, $\boldsymbol{A}_{\mathrm{t2}}$}。其中$\boldsymbol{A}_{\mathrm{s1}}$和$\boldsymbol{A}_{\mathrm{s2}}$分别为全局和局部空间矩阵，$\boldsymbol{A}_{\mathrm{t1}}$和$\boldsymbol{A}_{\mathrm{t2}}$分别为全局和局部时间矩阵，通过网格的序号链接签到和矩阵中的元素。签到具有多维关联特性，由签到组成的轨迹中包含用户的行为特征，对于用户轨迹中的关键节点，应该赋予更高的权重并突出关联特征。为了体现方案的有效性，提两种签到矩阵填充算法，作为不同方案进行对照分析。

算法1单点填充算法：通过网格链接并填充矩阵，忽略签到关联性，填充相互独立。伪代码如算法1。

算法1 单点填充算法

输入： 用户签到集S，与S待匹配签到集S_{match}，网格密度系数k
输出：用户签到矩阵集SMS
初始化： 初始化k维的零矩阵集SMS={$\boldsymbol{A}_{\mathrm{s1}}$, $\boldsymbol{A}_{\mathrm{s2}}$, $\boldsymbol{A}_{\mathrm{t1}}$, $\boldsymbol{A}_{\mathrm{t2}}$}，通过$S$和$S_{\mathrm{match}}$设定全局时空域
(1) 遍历签到集S:
(2) 获取时空网格：g_s=(x_s, y_s)，g_t=(x_t, y_t)
(3) 链接并填充矩阵：$\boldsymbol{A}_{\mathrm{s1}}[x_s, y_s]$ += 1；$\boldsymbol{A}_{\mathrm{t1}}[x_t, y_t]$ += 1
(4) 通过S和S_{match}设定局部时空域，将$\boldsymbol{A}_{\mathrm{s1}}$和$\boldsymbol{A}_{\mathrm{t1}}$替换为$\boldsymbol{A}_{\mathrm{s2}}$和$\boldsymbol{A}_{\mathrm{t2}}$，重复执行步骤1～步骤3
(5) 输出SMS

算法2关联填充算法：考虑签到关联性，签到在时空维度越接近关联性越强。令网格g_1=(x_1, y_1)，g_2=(x_2, y_2)，网格密度系数为k，用网x格距离$d(g_1,g_2)$表示签到的接近程度，计算为

算法2 关联填充算法

输入：用户签到集S，与S待匹配签到集S_{match}，网格密度系数k，关联系数s，填充系数p
输出：用户签到矩阵集SMS
(1) 通过单点填充算法获得SMS={$\boldsymbol{A}_{\mathrm{s1}}$, $\boldsymbol{A}_{\mathrm{s2}}$, $\boldsymbol{A}_{\mathrm{t1}}$, $\boldsymbol{A}_{\mathrm{t2}}$}，定义空集合$A$, B
(2) 通过S和S_{match}设定全局时空域，选定$\boldsymbol{A}_s = \boldsymbol{A}_{\mathrm{s1}}$ $\boldsymbol{A}_{\mathrm{t}} = \boldsymbol{A}_{\mathrm{t1}}$,
(3) 遍历签到集S中任意签到sign：
(4) 获取sign的时空网格表示g_s=(x_s, y_s)和g_t=(x_t, y_t)
(5) 将满足$d(g, g_s)$ <= s的网格g，将其添加到临时集合A_t
(6) 将满足$d(g, g_t)$ <= s的网格g，将其添加到临时集合B_t
(7) 对任意网格g=(x, y)：
(8) 若$g \in A_t \cap A$：$\boldsymbol{A}_s[x, y]$+=p
(9) 若$g \in B_t \cap B$：$\boldsymbol{A}_t[x, y]$+=p
(10) 更新集合A和B：A=A_t，B=B_t
(11) 若当前映射空间为全局时空域：通过S和S_{match}设定局部时空域，选定$\boldsymbol{A}_s = \boldsymbol{A}_{\mathrm{s2}}$ $\boldsymbol{A}_t = \boldsymbol{A}_{\mathrm{t2}}$，重复执行步骤3～步骤10
(12) 否则：输出SMS

$$d(g_1,g_2)=\begin{cases}|x_1-x_2|+|y_1-y_2|, & g_1和g_2为空间网格\\ |k\times x_1+y_1-k\times x_2+y_2|, & g_1和g_2为时间网格\end{cases} \tag{3}$$

关联填充算法通过关联网格链接并填充签到矩阵的元素，定义关联系数s作为关联性的判断阈值，填充系数p作为填充值，关联填充算法的伪代码如算法2。

第(3)～(10)步在全局时空域下填充签到矩阵，其中第(4)～(6)步获取时空关联网格，第(7)～(9)步根据前后连续签到重叠的关联网格填充签到矩阵，第(10)步更新集合A和B。第(11)步在局部时空域下执行操作，完成局部时空矩阵的填充。关联填充算法的思想是假设当前签到和后续签到具有一定的关联性，在遍历签到时前后两个签到的关联网格产生重叠时，视作假设成立，重叠的网格反映用户签到之间的关联性，通过网格链接签到矩阵中的元素并填充p，以调整元素的权重。通过构建签到图MAP={SMS_1, SMS_2}完成待匹配用户对的数据建模，为保持数据一致便于后续计算，需要对签到矩阵进行归一化，将元素值限定在0到1之间。

3 用户签到匹配模型

基于卷积神经网络强大的特征提取能力，提出图1所示的匹配模型：

模型的输入为原始签到集，在不同时空尺度下生成用户签到图。从签到图中分离空间通道和时间通道，前者包含空间矩阵，后者包含时间矩阵。在

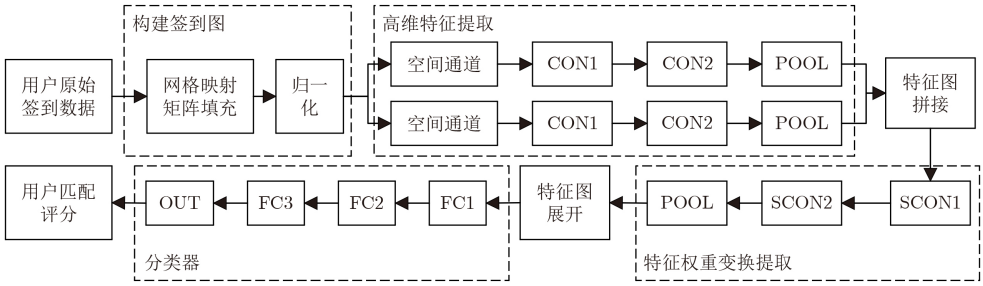


图 1 匹配模型

特征提取阶段，利用并联的网络通路分别提取出包含时空特征的高维特征图。将特征图拼接后，利用深度可分离卷积对特征图进一步处理，利用逐通道卷积和逐点卷积实现权重组合与特征融合。1维展开特征图，获得特征向量，构建前馈网络作为分类器，输入特征向量学习特征和匹配的潜在关联，最后输出匹配评分。当评分大于匹配阈值时，认为用户匹配，否则不匹配。

3.1 基于卷积神经网络的高维特征图提取

构建并联的网络通路分别从签到图不同通道提取时空特征，以空间特征提取为例：先通过CON1提取中间特征图，再通过CON2提取更抽象的高维特征数据，最后通过POOL下采样特征数据，生成高维空间特征图。令CON1中卷积核集$W_{\mathrm{C1}}=\{W_1, W_2, \cdots, W_{\mathrm{kcc1}}\}$，其中$W_i=\{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \boldsymbol{w}_4\}$，偏置集$B_{\mathrm{C1}}=\{B_1, B_2, \cdots, B_{\mathrm{kcc1}}\}$，其中$B_i=\{b_1, b_2, b_3, b_4\}$，签到图空间通道MAP_SC$=\{\boldsymbol{A}_{\mathrm{s11}}, \boldsymbol{A}_{\mathrm{s12}}, \boldsymbol{A}_{\mathrm{s21}}, \boldsymbol{A}_{\mathrm{s22}}\}$，其中的矩阵分别为用户对的全局和局部空间矩阵。kcc1为CON1中卷积核数量，W_i和B_i分别为第i个卷积核的权重矩阵集以及偏置集。$\boldsymbol{w}$和b分别为该卷积核各通道的权重矩阵和偏置。定义*运算由卷积核生成特征矩阵

$$\left.\begin{aligned}&(W_i, B_i) * \mathrm{MAP\_SC}=\boldsymbol{A}_i,\ \boldsymbol{A}_i=\begin{bmatrix} a_{i,11} & a_{i,12} & \cdots & a_{i,1s} \\ a_{i,21} & a_{i,22} & \cdots & a_{i,2s} \\ \vdots & \vdots & & \vdots \\ a_{i,s1} & a_{i,s2} & \cdots & a_{i,ss} \end{bmatrix}\\ &s=k-f+2p+1\\ &a_{i,mn}=\sum_{j=1}^{4}\left(\mathrm{ReLU}\left(\sum_{x,y=1,w_j\in W_i,b_j\in B_i}^{f}\boldsymbol{w}_j[x,y]\times\boldsymbol{v}_{j,mn}[x,y]+b_j\right)\right)\end{aligned}\right\} \tag{4}$$

其中，$\boldsymbol{A}_i$为第i个卷积核提取的特征矩阵，s为特征矩阵尺寸，k为网格密度系数，f为卷积核尺寸，p为填充参数，设置p使$s=k$。ReLU为激活函数。$\boldsymbol{v}_{j,\mathrm{mn}}$为卷积核的权重矩阵$\boldsymbol{w}_j$在MAP_SC第$j$个矩阵的$m$行$n$列卷积时选定的子矩阵，定义$\Upsilon$运算表示所有卷积核生成特征矩阵集合，构成中间特征图IFM_SC

$$\left.\begin{aligned}&(W_{\mathrm{C1}}, B_{\mathrm{C1}})\otimes\mathrm{MAP\_SC}=\mathrm{IFM\_SC}\\ &\mathrm{IFM\_SC}=\{A_1, A_2, \cdots, A_i, \cdots, A_{\mathrm{kcc1}}\},\\ &A_i=(W_i, B_i)*\mathrm{MAP\_SC}\end{aligned}\right\} \tag{5}$$

通过CON2对IFM_SC进行更抽象的空间特征提取，获得高维空间特征图FM_S

$$(W_{\mathrm{C2}}, B_{\mathrm{C2}})\otimes\mathrm{IFM\_S}=\mathrm{FM\_S} \tag{6}$$

其中，W_{C2}和B_{C2}分别为CON2卷积核集合和偏置集合。最后在POOL中通过最大池化下采样FM_S，输出尺寸为$k/2$，通道数为kcc2的空间特征图。时间特征图的提取过程和上述步骤一致，分别提取空间特征图和时间特征图并将它们按通道拼接，输出完整的用户时空特征图。

3.2 特征图权重变换和特征融合

依据对匹配的重要程度对不同特征矩阵赋权，同时，时空特征存在耦合关系，还需要进行特征融合以增强时空特征的关联性。深度可分离卷积先逐通道卷积处理特征图，然后逐点卷积对特征图加权融合，最后下采样输出。令拼接后的特征图为FM$=\{\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_c\}$，$c$为通道数。SCON1中卷积核权重矩阵集$W_{\mathrm{S1}}=\{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_c, \}$，偏置集$B_{\mathrm{S1}}=\{b_1, b_2, \cdots, b_c\}$。通过逐通道卷积处理后的特征图为SFM

$$\left.\begin{aligned}&\mathrm{SFM}=\{\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_c\},\\ &\boldsymbol{V}_i=\begin{bmatrix} v_{i,11} & v_{i,12} & \cdots & v_{i,1s} \\ v_{i,21} & v_{i,22} & \cdots & v_{i,2s} \\ \vdots & \vdots & & \vdots \\ v_{i,s1} & v_{i,s2} & \cdots & v_{i,ss} \end{bmatrix}\\ &v_{i,mn}=\mathrm{ReLU}\left(\sum_{x,y=1}^{f}w_i[x,y]\times v_{i,mn}[x,y]+b_i\right)\end{aligned}\right\} \tag{7}$$

其中，f和s的定义及计算同式(5)，$\boldsymbol{v}_{i,mn}$为卷积核的权重矩阵$\boldsymbol{w}_i$在FM的第i个特征矩阵的m行n列卷积选定的子矩阵。令SCON2中卷积核集$W_{\mathrm{S2}}=\{W_1, W_2, \cdots, W_{\mathrm{kcs1}}\}$，偏置集$B_{\mathrm{S2}}=\{B_1, B_2, \cdots, B_{\mathrm{kcs2}}\}$。kcs2为卷积核数量。通过对SFM逐点卷积实现权重变换和特征融合，逐点卷积采用尺寸为1的卷积核，对SFM逐点卷积等效对SFM不同通道加权组合，构成最终的用户特征图FFM，通过式(5)定义Υ运算得到：

$$(W_{\mathrm{S2}}, B_{\mathrm{S2}}) \otimes \mathrm{SFM} = \mathrm{FFM} \tag{8}$$

FFM的通道数为kcs2，尺寸和SFM相同。下采样FFM并设置填充系数使FFM和FM在数据维度一致。

3.3 基于全连接前馈网络的匹配评分预测

1维展开特征图获得特征向量，构建前馈网络作为分类器，自动学习输入特征向量和用户匹配的潜在关联并在OUT层输出最终的用户匹配评分。令输入特征图为FFM，对其1维展开生成特征向量$\mathbf{FV}$

$$\mathbf{FV} = [v_1, v_2, \cdots, v_{c\times k^2}]^{\mathrm{T}} \tag{9}$$

其中，c和k分别为通道数和尺寸。FC1的神经元参数矩阵为$\boldsymbol{W}_{\mathrm{F1}}$，偏置向量为$\boldsymbol{B}_{\mathrm{F1}}$，输出为$\boldsymbol{Z}_{\mathrm{F1}}$

$$Z_{\mathrm{F1}} = \mathrm{ReLU}\left(\boldsymbol{W}_{\mathrm{F1}} \times \mathbf{FV} + B_{\mathrm{F1}}\right) \tag{10}$$

数据在后续层中传递计算和FC1一致。FC3的输出为$\boldsymbol{Z}_{\mathrm{F3}}$，OUT激活输出用户匹配评分score

$$\mathrm{score} = \mathrm{Sigmoid}\left(\boldsymbol{W}_{\mathrm{OUT}} \times Z_{\mathrm{F3}} + b\right) \tag{11}$$

其中，Sigmoid为OUT激活函数，$\boldsymbol{W}_{\mathrm{OUT}}$为OUT神经元参数矩阵，$b$为OUT偏置。

3.4 匹配模型的参数优化

模型的参数可表示为model$\{W_{\mathrm{C1}}, \cdots, W_{\mathrm{F1}}, W_{\mathrm{OUT}}, B_{\mathrm{C1}}, \cdots, B_{\mathrm{F1}}, b\}$，通过反向传播算法计算损失函数在各层参数上的微分，基于梯度下降算法，最小化损失函数迭代更新模型参数。训练过程中，正向传播输出匹配评分，反向传播更新模型参数。各层的前向传播计算为

$$\mathrm{AF}\left((W, B) \otimes \boldsymbol{V}_{\mathrm{in}}\right) = \boldsymbol{V}_{\mathrm{out}} \tag{12}$$

其中，$\boldsymbol{V}_{\mathrm{in}}$，$\boldsymbol{V}_{\mathrm{out}}$分别为输入和输出，$W$, B分别为权重集和偏置集，AF为激活函数。损失函数Loss

$$\mathrm{Loss} = -\sum_{i=1}^{\mathrm{bs}} l_i \times \lg s_i + (1-l_i) \times \lg(1-s_i) \tag{13}$$

其中，bs为批次训练样本量，l_i和s_i分别为第i个样本的标签和预测值。计算各层反向传播

$$\delta \boldsymbol{V}_{\mathrm{OUT}} = \frac{\partial \mathrm{Loss}}{\partial \boldsymbol{V}_{\mathrm{OUT}}} \times \mathrm{Sigmoid}'\left(V_{\mathrm{OUT}}\right) \tag{14}$$

$$\left.\begin{aligned} \delta \boldsymbol{V}_{\mathrm{in}} &= W^{\mathrm{T}} \times \delta \boldsymbol{V}_{\mathrm{out}} \odot \mathrm{ReLU}'\left(\boldsymbol{V}_{\mathrm{in}}\right) \\ \delta \boldsymbol{V}_{\mathrm{in}} &= \delta \boldsymbol{V}_{\mathrm{out}} \times \mathrm{rot180}\left(W\right) \odot \mathrm{ReLU}'\left(\boldsymbol{V}_{\mathrm{in}}\right) \end{aligned}\right\} \tag{15}$$

其中，$\odot$为Hadamard积，W^{T}表示集合内转置矩阵，rot180表示旋转集合内矩阵半周。各层参数的梯度

$$\left.\begin{aligned} \delta W &= \delta \boldsymbol{V}_{\mathrm{out}} \times [\boldsymbol{V}_{\mathrm{in}}]^{\mathrm{T}}, \delta \boldsymbol{B} = \delta \boldsymbol{V}_{\mathrm{out}} \\ \delta W &= \boldsymbol{V}_{\mathrm{in}} \times \delta \boldsymbol{V}_{\mathrm{out}}, \delta \boldsymbol{B} = \mathrm{sum}\left(\delta \boldsymbol{V}_{\mathrm{out}}\right) \end{aligned}\right\} \tag{16}$$

其中，sum表示对矩阵元素求和。模型的优化算法如算法3。

算法3 模型优化算法

输入：训练样本集train_data，迭代轮数epoch，批次尺寸bs，学习率α

输出：model

(1) 随机初始化model$\{W_{\mathrm{C1}}, \cdots, W_{\mathrm{F1}}, W_{\mathrm{OUT}}, B_{\mathrm{C1}}, \cdots, B_{\mathrm{F1}}, b\}$
(2) for i =1 to epoch:
(3) for batch_data in train_data: #按批次遍历整个训练样本集
(4) for sample in batch_data: #按样本遍历单个批次
(5) 构建样本签到图MAP
(6) 根据式(12)进行前向传播，计算模型各层输出
(7) 根据式(13)—式(16)计算模型预测的损失和各层参数的梯度
(8) 更新模型各层参数：W-= $\alpha\times\delta W$，B-= $\alpha\times\delta B$
(9) 输出 model

4 实验分析

4.1 数据集和评价指标

实验采用斯坦福大学的社交网络数据集Brightkite和Gowalla，他们由不同社交网络的公共API收集，签到数据由用户id、时间、经纬度和位置id字段组成，将数据集随机划分为a和b两个部分，表示两个社交网络的关联签到。划分策略：以相同概率将相同用户id的签到划分到a或b，并保证划分结束时a和b都至少拥有一条签到。构建正负例数据时的构建策略：随机选择50%签到通过用户id链接构建正例，剩下的构建负例。在构建负例时，分别在a和b选择不同用户u和u^*，将其组成负例并标记u^*的id，后续仅在b中选择未标记id，每个签到仅出现一次以确保数据的唯一性。考虑真实用户签到的多样性，不对签到数据做清洗、筛选等操作以保证实验的客观性。数据集的80%作训练集，20%作测试集。数据集概况如表1。

实验采用准确率acc、精确率pre、召回率rec以及f1 4种指标。计算为

$$\left.\begin{aligned}\text{acc} &= \frac{\text{tp}+\text{tn}}{\text{tp}+\text{fp}+\text{tn}+\text{fn}}\\ \text{pre} &= \frac{\text{tp}}{\text{tp}+\text{fp}}\\ \text{rec} &= \frac{\text{tp}}{\text{tp}+\text{fn}}\\ \text{f1} &= \frac{2\times\text{pre}\times\text{rec}}{\text{pre}+\text{rec}}\end{aligned}\right\} \tag{17}$$

其中，tp, fp, tn, fn分别表示正确预测正例、错误预测正例、正确预测负例、错误预测负例的样本数。

4.2 模型参数调节

为评估模型参数优化时不同训练轮数epoch的效果，从训练集中随机分离5%作验证集，训练时每迭代完成1次，对验证集进行1次验证。设置学习率为0.001，模型的训练曲线如图2所示。

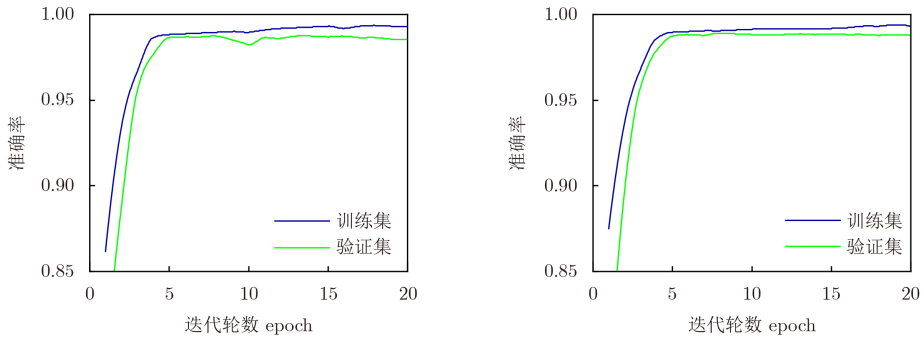
结果表示epoch超过5之后，模型在验证集上的损失和准确率趋于稳定。Brightkite的曲线波动幅度要大于Gowalla，原因是Gowalla中的签到数据整体分布更加均匀，所以训练曲线较平缓，模型在Brightkite和Gowalla的验证准确率分别达到98.70%和98.85%左右，整体相差约0.15%。

为验证网格映射对特征提取的作用以及关联填充算法的有效性，在不同条件测试准确率，用S表示仅提取空间特征，T表示仅提取时间特征，TS表示同时提取时空特征。A和P分别表示用关联填充算法和单点填充算法生成签到图，分别用实线和虚线表示。测试结果如图3所示。

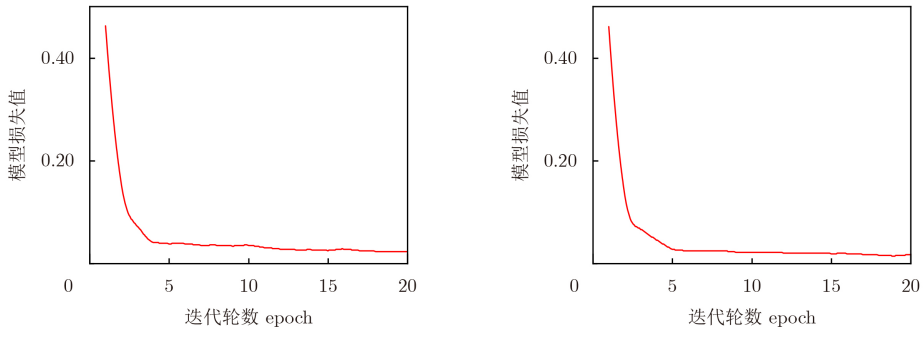
随着k增加，签到图的精度增加，准确率呈上升趋势，当k超过40之后，准确率趋于稳定。同时利用时空特征能够最大限度分析用户签到行为并有效耦合时空特征，提升匹配准确率，模型性能最

表 1　数据集概况

	数据集			
	Brightkite(样本量50 686)		Gowalla(样本量107 092)	
划分部分	a	b	a	b
平均签到数	111	111	75	75
起始时间	2008–03–22 06:34:37	2008–03–21 20:36:21	2009–02–05 06:27:43	2009–02–04 05:17:38
终止时间	2010–10–18 18:34:01	2010–10–18 18:39:58	2010–10–23 05:22:06	2010–10–23 05:22:06
经度范围	–163.193～151.198	–163.193～151.198	–90.011～105.659	–90.011～105.625
纬度范围	–179.824～179.999	–179.824～179.999	–176.309～177.463	–166.525～177.453

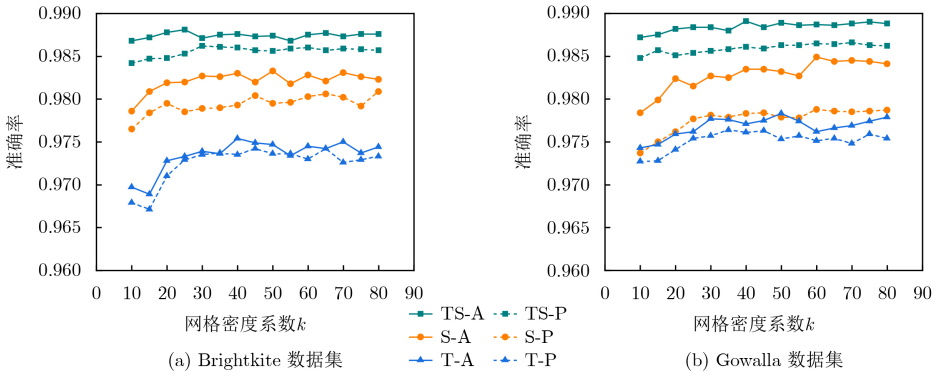


(a) Brightkite 数据集准确率曲线　　(b) Gowalla 数据集准确率曲线



(c) Brightkite 数据集损失曲线　　(d) Gowalla 数据集损失曲线

图 2 模型的准确率和损失曲线



(a) Brightkite 数据集 (b) Gowalla 数据集

图 3 模型在不同条件下的准确率

优。此时模型在两个数据集上的准确率分别为98.62%和98.65%。关联填充算法在不同条件下均促进了模型性能提升，仅利用空间特征时，提成幅度最大，提升约0.24%和0.61%。由于签到的时序特征较弱，仅利用时间特征时提升有限。实验结果表明关联填充算法能够有效加强模型对签到关联特征的提取以提高匹配性能。

关联系数s和填充系数p影响数据建模的有效性，当s和p设置过小，签到的关联特征体现不充分，设置过大则会对签到矩阵引入噪声。为了探究合理值，设置多组实验进行对照，实验结果如图4所示。

准确率随着s的增加呈现先增后降的趋势，增大s会增强数据关联性，准确率上升，当s过大时，由于引入的噪声影响使准确率下降。p值较小时引入噪声的影响被降低，下降趋势靠后。增大s会增加数据建模的复杂度，综合考虑对s和p分别设置为2和0.6。模型的其他参数的设置如表2所示。

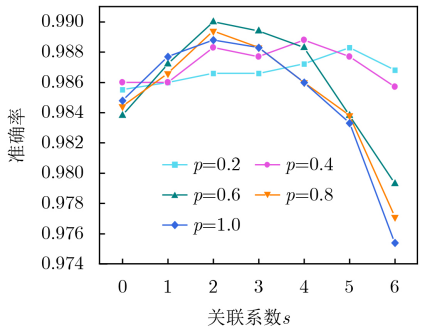


图 4 关联系数s和填充系数p的确定

4.3 算法比较

实验中UNICORN[4], STUL[5], CDTraj2vec[13]以及UIDwST[17]为基准算法。UNICORN和CDTraj2vec向量化用户签到，用向量相似度表示用户相似度；STUL基于密度聚类和高斯混合模型提取用户时空特征；UIDwST基于核密度估计，利用冲突签到校正相似度评分。本文算法为DLUMCN，基于时空特征和关联填充算法构建匹配模型。令用户对数为N，用户平均签到数为S，网格密度系数为k，不同算法的时空复杂度和在两组数据集上的测试结果如表3、表4所示。

通过时空网格映射和关联填充算法，DLUMCN能够针对签到关联性构建签到图，提取并融合用户时空特征，在两组数据集上均表现最佳。同时从实验结果中不难发现，模型的精确率大于召回率，分析原因是实验保留原始签到数据集中签到数量极少的正例用户，这部分的用户被误判为负例，导致算法召回率下降。对比基准算法，DLUMCN在两组数据集上的准确率acc及f1值均为最佳，验证了所提算法的有效性。

5 结束语

针对社交网络用户签到和传统轨迹数据的差异性，用户签到数据时空信息难以耦合，特征提取困

表 2 匹配模型的其他参数

模型参数	设定值	说明
k	65(Brightkite) / 75(Gowalla)	网格密度系数
det/mt	1×10^{-4}/0.5	调节因子/匹配阈值
batch-size，α	256, 1×10^{-3}	训练的批次样本量，学习率
kcc1(fc1), kcc2(fc2)	16(3), 32(3)	CON1, CON2卷积核数 (尺寸)
kcs1(fs1), kcs2(fs2)	1(2), 64(1)	SCON1, SCON2卷积核数(尺寸)
n_{F1}, n_{F2}, n_{F3}	16, 12, 8	前馈网络隐藏层神经元数量

表 3　不同算法的复杂度以及耗时(s)

算法	时间复杂度	空间复杂度	(Brightkite数据集)		(Gowalla数据集)	
			训练时间	测试时间	训练时间	测试时间
UNICORN	$O(N^2{\cdot}S{\cdot}k^2)$	$O(N{\cdot}S)$	–	11.85	–	25.12
STUL	$O(N{\cdot}S^3)$	$O(N{\cdot}S)$	–	451.663	–	1394.156
CDTraj2vec	$O(N{\cdot}S^2)$	$O(N{\cdot}S^2)$	375.833	11.85	587.174	26.385
UIDwST	$O\ (N^2{\cdot}S^2)$	$O\ (N{\cdot}S)$	–	217.775	–	631.105
DLUMCN	$O(N{\cdot}S{\cdot}k^2)$	$O(N{\cdot}k^2)$	69.678	0.475	200.066	1.311

表 4　不同算法在两组数据集上的测试结果

算法	(Brightkite数据集)				(Gowalla数据集)			
	acc	pre	rec	f1	acc	pre	rec	f1
UNICORN	0.8801	0.8085	0.9960	0.8925	0.8847	0.8194	0.9870	0.8954
STUL	0.9253	0.9039	0.9516	0.9271	0.9281	0.9094	0.9447	0.9267
CDTraj2vec	0.9529	0.9833	0.9212	0.9512	0.9567	0.9686	0.9360	0.9520
UIDwST	0.9740	0.9594	0.9897	0.9743	0.9773	0.9619	0.9870	0.9742
DLUMCN	0.9881	0.9949	0.9814	0.9881	0.9890	0.9969	0.9812	0.9889

难的问题，本文提出一种基于深度学习的跨社交网络用户匹配方法。从时空维度进行网格映射，提出关联填充算法生成用户签到图，最大限度保留签到信息，同时加强对签到关联特征提取。利用卷积神经网络生成时空特征图，并基于深度可分离卷积实现特征的权重变换和融合。最后构造前馈网络分类器，学习特征数据和匹配的潜在关系，输出用户对匹配评分。通过在两组真实社交网络用户签到数据集上的实验表明，本文所提算法在不同的指标上均具有优势。真实社交网络用户签到的多样性会产生一部分签到数据量极少的用户，针对这类用户，模型的召回率下降导致匹配性能下降。后续研究可以聚焦于如何改进对用户签到数据建模以减少稀疏数据的影响，提高模型的召回率以增强匹配性能。

参 考 文 献

[1] DENG Kaikai, XING Ling, ZHENG Longshui, *et al*. A user identification algorithm based on user behavior analysis in social networks[J]. *IEEE Access*, 2019, 7: 47114–47123. doi: 10.1109/ACCESS.2019.2909089.

[2] 邢玲, 邓凯凯, 吴红海, 等. 复杂网络视角下跨社交网络用户身份识别研究综述[J]. 电子科技大学学报, 2020, 49(6): 905–917. doi: 10.12178/1001-0548.2019182.
XING Ling, DENG Kaikai, WU Honghai, *et al*. Review of user identification across social networks: The complex network approach[J]. *Journal of University of Electronic Science and Technology of China*, 2020, 49(6): 905–917. doi: 10.12178/1001-0548.2019182.

[3] 张树森, 梁循, 弭宝瞳, 等. 基于内容的社交网络用户身份识别方法[J]. 计算机学报, 2019, 42(8): 1739–1754. doi: 10.11897/SP.J.1016.2019.01739.
ZHANG Shusen, LIANG Xun, MI Baotong, *et al*. Content-based social network user identification methods[J]. *Chinese Journal of Computers*, 2019, 42(8): 1739–1754. doi: 10.11897/SP.J.1016.2019.01739.

[4] HAO Tianyi, ZHOU Jingbo, CHENG Yunsheng, *et al*. User identification in cyber-physical space: A case study on mobile query logs and trajectories[C]. The 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Burlingame, USA, 2016: 71. doi: 10.1145/2996913.2997017.

[5] CHEN Wei, YIN Hongzhi, WANG Weiqing, *et al*. Exploiting spatio-temporal user behaviors for user linkage[C]. The ACM International Conference on Information and Knowledge Management, Singapore, 2017: 517–526. doi: 10.1145/3132847.3132898.

[6] KONDOR D, HASHEMIAN B, DE MONTJOYE Y A, *et al*. Towards matching user mobility traces in large-scale datasets[J]. *IEEE Transactions on Big Data*, 2020, 6(4): 714–726. doi: 10.1109/TBDATA.2018.2871693.

[7] HAO Tianyi, ZHOU Jingbo, CHENG Yunsheng, *et al*. A unified framework for user identification across online and offline data[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 34(4): 1562–1575. doi: 10.1109/TKDE.2020.3000287.

[8] HE Wenqiang, LI Yongjun, ZHANG Yinyin, *et al*. A binary-search-based locality-sensitive hashing method for cross-site user identification[J]. *IEEE Transactions on Computational Social Systems*, 2022, 10(2): 480–491. doi: 10.1109/TCSS.2022.3158664.

[9] 王前东. 经典轨迹的鲁棒相似度量算法[J]. 电子与信息学报, 2020, 42(8): 1999–2005. doi: 10.11999/JEIT190550.
WANG Qiandong. A robust trajectory similarity measure method for classical trajectory[J]. *Journal of Electronics & Information Technology*, 2020, 42(8): 1999–2005. doi: 10.11999/JEIT190550.

[10] QI Mengjun, WANG Zhongyuan, HE Zheng, *et al*. User identification across asynchronous mobility trajectories[J]. *Sensors*, 2019, 19(9): 2102. doi: 10.3390/s19092102.

[11] HAN Xiaohui, WANG Lianhai, XU Lijuan, *et al*. Social Media account linkage using user-generated geo-location data[C]. IEEE Conference on Intelligence and Security Informatics, Tucson, USA, 2016: 157–162. doi: 10.1109/ISI.2016.7745460.

[12] 冯朔, 申德荣, 聂铁铮, 等. 一种基于最大公共子图的社交网络对齐方法[J]. 软件学报, 2019, 30(7): 2175–2187. doi: 10.13328/j.cnki.jos.005831.
FENG Shuo, SHEN Derong, NIE Tiezheng, *et al*. Maximum common subgraph based social network alignment method[J]. *Journal of Software*, 2019, 30(7): 2175–2187. doi: 10.13328/j.cnki.jos.005831.

[13] 陈鸿昶, 徐乾, 黄瑞阳, 等. 一种基于用户轨迹的跨社交网络用户身份识别算法[J]. 电子与信息学报, 2018, 40(11): 2758–2764. doi: 10.11999/JEIT180130.
CHEN Hongchang, XU Qian, HUANG Ruiyang, *et al*. User identification across social networks based on user trajectory[J]. *Journal of Electronics & Information Technology*, 2018, 40(11): 2758–2764. doi: 10.11999/JEIT180130.

[14] MA Jiangtao, QIAO Yaqiong, HU Guangwu, *et al*. Social account linking via weighted bipartite graph matching[J]. *International Journal of Communication Systems*, 2018, 31(7): e3471. doi: 10.1002/dac.3471.

[15] XIAO Xiangye, ZHENG Yu, LUO Qiong, *et al*. Inferring social ties between users with human location history[J]. *Journal of Ambient Intelligence and Humanized Computing*, 2014, 5(1): 3–19. doi: 10.1007/s12652-012-0117-z.

[16] WANG Fengzi, ZHU Xinning, and MIAO Jiansong. Semantic trajectories-based social relationships discovery using WiFi monitors[J]. *Personal and Ubiquitous Computing*, 2017, 21(1): 85–96. doi: 10.1007/s00779-016-0983-z.

[17] LI Yongjun, JI Wenli, GAO Xing, *et al*. Matching user accounts with spatio-temporal awareness across social networks[J]. *Information Sciences*, 2021, 570: 1–15. doi: 10.1016/j.ins.2021.04.030.

[18] 张伟, 李扬, 张吉, 等. 融合时空行为与社交关系的用户轨迹识别模型[J]. 计算机学报, 2021, 44(11): 2173–2188. doi: 10.11897/SP.J.1016.2021.02173.
ZHANG Wei, LI Yang, ZHANG Ji, *et al*. A user trajectory identification model with fusion of spatio-temporal behavior and social relation[J]. *Chinese Journal of Computers*, 2021, 44(11): 2173–2188. doi: 10.11897/SP.J.1016.2021.02173.

[19] 沈佳琪, 周国民. 跨社交网络的同一用户识别算法[J]. 电子技术应用, 2022, 48(1): 109–114. doi: 10.16157/j.issn.0258-7998.211518.
SHEN Jiaqi and ZHOU Guomin. User alignment across social networks[J]. *Application of Electronic Technique*, 2022, 48(1): 109–114. doi: 10.16157/j.issn.0258-7998.211518.

[20] HAN Xiaohui, WANG Lianhai, XU Shujiang, *et al*. Linking social network accounts by modeling user spatiotemporal habits[C]. 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 2017: 19–24. doi: 10.1109/ISI.2017.8004868.

[21] CHEN Wei, WANG Weiqing, YIN Hongzhi, *et al*. User account linkage across multiple platforms with location data[J]. *Journal of Computer Science and Technology*, 2020, 35(4): 751–768. doi: 10.1007/s11390-020-0250-7.

[22] ZHOU Xueyan and YANG Jing. Matching user accounts based on location verification across social networks[J]. *Revista Internacional de Métodos Numéricos para Cálculo y Diseñ o en Ingeniería*, 2020, 36(1): 8. doi: 10.23967/j.rimni.2019.12.001.

马 强：男，副教授，研究方向为智能数据处理、社交媒体计算.
戴 军：男，硕士生，研究方向为跨社交媒体计算、用户账号匹配.

责任编辑：马秀强