基于社会化表示的用户性别识别

朱裴松 钱铁云 吴闽泉
(武汉大学软件工程国家重点实验室 武汉 430072)

摘 要 由于具有针对性的广告投放和个性化搜索等潜在应用,性别预测引起了巨大的研究兴趣。现有的大多数研究依赖于文本内容,而文本信息有时较难获取,从而使得文本特征很难被提取。对此,提出了一个新框架,该框架仅使用用户ID来对性别进行预测。该框架的关键在于在嵌入式连接空间中表示用户。提出两种策略来修改词嵌入技术,使其应用到用户嵌入当中。这两种策略分别是:1)序列化用户ID以获得社会关系的顺序;2)将用户嵌入大的上下文滑动窗口。在两个真实的新浪微博数据集上进行了广泛的实验,实验结果表明该方法显著优于目前最好的图形嵌入基线方法,其准确率也高于基于内容的方法。

关键词 性别预测,社交媒体用户,社交关系,社会化表示
中图法分类号 TP391 **文献标识码** A

Identifying Users' Gender via Social Representations

ZHU Pei-song QIAN Tie-yun WU Min-quan
(State Key Lab of Software Engineering, Wuhan University, Wuhan 430072, China)

Abstract Gender prediction has evoked great research interests due to its potential applications, like targeted advertisement and personalized search. Most of existing studies rely on the content texts. However, the text information is hard to access. This makes it difficult to extract text features. In this paper, we proposed a novel frame-work which only involves the users' ID for gender prediction. The key idea is to represent users in the embedding connection space. We presented two strategies to modify the word embedding technique for user embedding. The first is to sequentialize users' ID to get the order of social context. The second is to embed users into a large-sized sliding window of contexts. We conducted extensive experiments on two real data sets from Sina Weibo. Results show that our method is significantly better than the state-of-the-art graph embedding baselines. Its accuracy also outperforms that of the content based approaches.

Keywords Gender prediction, Users in social media, Social contexts, Social representations

1 绪论

近年来,性别预测因为其潜在的应用引起了广泛的关注[5,7,10,14-16,23],如定向广告和个性化服务等。所有现有的方法几乎都使用文本内容来构建特征向量以进行分类[5,7,10,16]。由于网站的限制,访问微博或评论等文本信息通常很困难。更重要的是,社交媒体中存在大量用户仅注册浏览但未发布任何内容。例如,来自中国新浪微博的100万用户的样本显示,大约7.4%的用户不会发布任何消息,对于这样的用户,我们无法获得任何可用于分析的文本特征。

本文提出一种不使用内容特征的性别分类的新方法,其关键思想是从用户之间的关系中学习用户的社交表示。在社交媒体中有两种基本的社交关系,即关注和被关注,为了简单起见,使用"朋友"来表示这两种关系,并代表该用户相连的所有用户,包括家庭成员、同学等。例如,"2:8 6 4 10"的序列表示用户2与其他4个用户8,6,4和10均是朋友关系。

现有的用于表示社交关系的方法包括传统的基于图形的表示方法(TGR)[9]和最近提出的图形嵌入方法(GE)[18,22]。基本上每个朋友关系在图中都被表示为一条边。因此,对于上述示例,则有2-8,2-6,2-4和2-10这4条边。我们认为TGR和GE缺少一些重要信息。需要注意的是,除了将显性的社会关系作为直接的朋友相连之外,朋友之间也存在着隐性的社会关系。例如,用户2关注了用户6和10,假设他们是他/她的同学,因此可推测出用户6和10也很有可能是同学关系。如果用户6和10未互相关注,则在图中他们之间无边。

本文提出的方法可以同时捕捉隐性和显性的关系,该想法基于词嵌入技术[4,12-13]。例如,假设窗口大小为2,当对句子"这本书是非常有趣的",即对一系列单词(这本,书,是,非常,有趣的)进行建模时,每个单词被视为当前词,左右窗口中的单词作为上下文,则有五组上下文(书,是),(这本,是,非常),(这本,书,非常,有趣的),(书,是,有趣的)和(是,非常)。显然,与只捕捉显性关系的图形表示相比,词嵌入能编码出更加丰富的信息,因为它反映了一个句子中词共现的隐性关系。

本文受国家自然科学基金项目:社会媒体中的垃圾用户集团识别方法研究(61572376)资助。
朱裴松 男,硕士生,主要研究方向为数据挖掘、自然语言处理,E-mail:qty@whu.edu.cn(通信作者);**钱铁云** 博士,教授,主要研究方向为数据挖掘、自然语言处理;**吴闽泉** 男,硕士,高级工程师,主要研究方向为数据管理,E-mail:rjgc@whu.edu.cn。

同样地，也可以通过这种方法来捕捉同学 6 和 10 之间对应于“是”和“有趣的”两个词之间隐含的社交关系。

上面的例子说明了词嵌入技术对图形表示的改进。然而，由于语言和社会关系之间的巨大差异，词嵌入在用于编码社交关系时具有一定的局限性。在语言中，语法控制句子中单词的顺序。相比之下，在社交关系中用户 ID 缺少排序，例如，两位同学 6 和 10 可以出现在社交关系的任意位置。此外，还有很多短语或习语作为句子中固定结构的一部分，因此如 5 或 10 这样的小窗口通常足以捕获词嵌入中的固定结构表达。然而，关系相近的用户在社交网络中可能相距较远。为了解决这两个问题，提出两种修改方案：1）节点序列化；2）大尺寸滑动窗口。节点序列化是将用户的 ID 映射成固定的顺序，以消除邻域中的随机性。大尺寸滑动窗口旨在将间隔距离很远但处于相同社区的用户集中到同一个社交环境中。

据笔者所知，采用词嵌入方法来获取社交关系并用于社交表示是首创的。提出的方法具有以下关键的特点：

1）它提出了一种新模式来编码仅涉及用户 ID 的社交关系，而大多数现有方法依赖于内容来构建特征向量；

2）它捕获了用户之间的各种社会关系，而图形嵌入技术仅考虑两个用户的显性关系。

本文第 2 节回顾了相关工作；第 3 节介绍了用于性别预测的获取社交表示的方法；第 4 节介绍了用于评估的数据集；第 5 节进行了实验并对实验结果进行了分析；最后总结全文。

2　相关工作

本节将从特征集、词嵌入、图形嵌入技术以及分类算法等方面来回顾相关工作。

2.1　特征集

在性别分类方面，词或 n-gram 词袋是使用得最为广泛的特征[5-7,10,16]，还有一些从文本内容中提取的特征，包括标点符号、大写字母、特殊词[10]、俚语[11]、单词或句子长度[10-11]、概念类[5]和词性标注（POS）序列[14]等。

几乎现有的所有研究都依赖于文本内容信息，唯一例外的是直接利用网络信息的邻居向量表示方法（NVR），本文将其作为基线方法之一来显示用户嵌入方法对分类结果的提升效果。

2.2　字嵌入和图嵌入

随着深层神经网络的发展，词嵌入技术在许多自然语言处理（NLP）任务上取得了不错的效果，典型的技术包括 NNLM[4]，LBL[13]，CBOW 和 SkipGram[12]。由于 SkipGram 的效果优于 CBOW，并且它显著加快了 NNLM 和 LBL 的训练过程，因此采用 SkipGram 算法训练词向量。

图形嵌入是一个经典的问题。传统的方法（如 Isomap 和 Laplacian EigenMap）对节点数量具有平方级时间复杂度。近年来，研究人员相继提出了利用随机梯度下降法的 GF 算法[1]、利用边采样建模一阶关系和二阶关系的 LINE 算法[22]、使用随机游走的 DeepWalk 算法[18]等对大规模网络进行嵌入式表示学习。其中，LINE 的分类效果是最好的，因此将其作为基线方法之一。

2.3　分类算法

目前已经有许多机器学习方法被用来解决性别预测问题，如 SVM[7,14,16,19]、决策树[2,7,17]、NaïveBayes[2,11,14,22]、逻辑回归[3,5]、Winnow 算法[6,20]以及最大熵[10]。

分类算法不是本文的重点，由于逻辑回归不像 SVM 那样对参数敏感，并且也有很好的分类效果，因此本文实验选择使用逻辑回归作为分类器。

3　社交表示学习

在社交媒体中，用户通常与家人、朋友、同学、同事或者有同样兴趣的人联系，所有这些连接（邻居）共同形成社会关系环境。而且，拥有共同朋友的用户可能属于同一个社区，如用户 a 关注了她的同学 b 和 c，那么 b 和 c 就是用户 a 的社会关系环境，用户 a，b 和 c 组成一个社区“同学”。从而可以进一步推断，如果这个社区中还有另一个用户 d，那么 d 的社会关系环境中很可能有 a，b，c 中的一个或多个。用户在同一社会关系环境中出现的次数越多，表明这些用户之间的关系越紧密，它们属于同一个社区的概率也越大。受上述分析的启发，本文借鉴最近提出的一种用于捕获词语之间的语义和句法关系的词嵌入技术 SkipGram[12]模型中的思想。

3.1　SkipGram

SkipGram 的目标是最大化一个句子中单词之间出现在一个窗口内的共现概率。更正式地，将目标函数定义为：

$$L=\sum_{w\in C}\log p(Comtext(w)\mid w)=\sum_{w\in C}\log\prod_{u\in Context(w)}p(u\mid w)$$

其中，C 是用于训练的语料库。通过应用基于哈夫曼树的层次 softmax[12]，可以将 $p(u\mid w)$ 重写为：

$$p(u\mid w)=\prod_{j=2}^{l^u}p(d_j^u\mid v(w),\theta_{j-1}^u)$$

$$p(d_j^u\mid v(w),\theta_{j-1}^u)=[\sigma(v(w)^{\mathrm{T}}\theta_{j-1}^u)]^{1-d_j^u}[1-\sigma(v(w)^{\mathrm{T}}\theta_{j-1}^u)]^{d_j^u}$$

其中，$v(w)$ 是中心词 w 的 d 维向量，d_j 和 θ_j 分别是路径 p 中的第 j 个节点的霍夫曼编码（0 或 1）和其 d 维向量。该目标函数可以使用梯度下降技术来进行优化。该过程如算法 1 所示。

算法 1　SkipGram

SkipGram (Φ, b, Context (u_i), s, η)

1. for each $v_j\in$ Context(u_i)
2. for each u $\in$ Context(u_i)[b, b + s]
3. $L(\Phi)=\log\prod_{u\in Context(u_i)}p(u\mid\Phi(v_j))$
4. $\Phi=\Phi-\eta\frac{\partial L}{\partial\Phi}$
5. end for
6. end for

3.2　基于 SkipGram 的用户嵌入

SkipGram 专为词嵌入而设计。使用 Skip-Gram 的典型场景是训练语料库，其中每个单词自然地嵌入在段落或文档中。但是社交媒体用户不具有这样的特征，此外，具有局部结构的用户或社区可能彼此远离。因此，提出了两种策略使 SkipGram 应用到用户嵌入中，即节点序列化和大尺寸滑动窗口。

3.2.1　节点序列化

节点序列化是指为每个社会关系环境识别节点序列、创建节点邻居的过程，如控制句子中的单词序列的语法。最简单的序列化只使用 ID 的自然大小顺序，如用户 1 在用户 2 之前，并且用户 2 在用户 3 之前。该方法可以消除邻居的随机性，但是 ID 信息与网络的固有结构无关，因此提出以下基于

度数的序列化方法。

给出图 G 中任何两个节点 i 和 j 及其第 k 阶邻居 $N_k(i)$ 和 $N_k(j)$，基于度数的序列化定义了 G 中节点上的总顺序 $\succ$，它可以唯一确定序列中每个邻居节点 i 的位置 O，其定义如下：

(1)$O(i) \succ O(j)$ iff. $d(i) > d(j)$；

(2)$O(i) \succ O(j)$ iff. $d(i)=d(j)$，and $d(\succ_1(i)) > d(\succ_1(j))$；

(3)$O(i) \succ O(j)$ iff. $d(i)=d(j), d(\succ_k(i))=d(\succ_k(j))$ $(k=1\cdots m-1)$，and $d(N_m(i)) > d(N_m(j))$。

$d(i)$ 函数表示将节点 ID 映射为该节点的度数，其基本思想是按照其度数对用户进行排序。若两个节点的度数相等，则进一步比较其第 k 阶邻居节点的度数，直到给出一个完整的顺序。

3.2.2 大尺寸滑动窗

大尺寸滑动窗口策略用于处理由于用户 ID 的随机性而引起的问题。用户 ID 由系统自动分配，通常与用户注册该社交媒体的时间相关，因此用户之间的联系和用户 ID 之间并无直接的关系。当 SkipGram 应用在自然语言处理领域时，窗口大小通常设置为 5 或 10，但这并不适合用户嵌入，因为朋友的数量通常相当大，因此将 SkipGram 中的窗口大小设置为一个很大的值。

3.3 社交表示学习算法

我们可将 SkipGram 应用到上述的社会关系中来学习用户的社交表示。该算法称为用户嵌入(UE)，如算法 2 所示。

算法 2 User Embedding (UE)

Input: The set of users $U=\{u_i\}$ and the friends of users $\{u_i, F(u_i)=\{f_{i1},\cdots,f_{in}\}\}$

Parameters: the window size s, the dimensionality of user embedding d, and the learning ratio η

Output: matrix of user representations $\Phi \in R^{|U|\times d}$

Steps:

1. for each user u_i
2. $A(u_i)=\{u_i\}\cup F(u_i)$
3. end for
4. Sequentializing ids in $F(u_i)$ and $A(u_i)$
5. for each user u_i
6. for the jth friend (j = i−1⋯i−n) in $F(u_i)$
7. SkipGram(Φ, j, $F(u_i)$, s, η)
8. end for
9. SkipGram(Φ, 1, $A(u_i)$, $|A(u_i)|$, η)
10. end for

算法 2 中，第 1－3 行初始化显性的社会关系环境，即用户的朋友序列；第 4 行序列化用户朋友 ID；第 5－10 行构建用户嵌入式表示；第 6－8 行获取隐性的社会关系；第 9 行获取显性的社会关系。

在 UE 算法中有 3 个可调参数：学习率 η、嵌入式表示向量的维度 d 以及上下文的大小(即窗口大小)。其中 η 与训练速度相关，本文使用默认设置。后文将研究 d 和 s 对实验的影响。

4 数据集

4.1 数据收集

实验数据来自中国最大的微博服务之一——新浪微博。每个用户在新浪微博上都有一个简介，其包括几个字段，如用户名、昵称、性别、标签、描述、粉丝数量、关注数量和消息数量等。表 1 列出了新浪微博上一位名人的简介。

表 1 新浪微博上的用户信息示例

用户 ID	1749127163
用户名	雷军
性别	男
微博数	4353
粉丝数	11979274
关注数	921
标签	天使投资、小米手机、我们爱米聊
简介	小米 CEO，金山 CEO

用户简介中的大多数字段是选填的，许多用户选择不填，也有许多用户不会发布任何微博。鉴于社交媒体的本质，用户通常会与他人联系，因此用户之间会存在许多相连的边，通过新浪提供的 API 获取这些朋友的用户名，这些用户之间的联系可用于本文实验。

从一个公共的数据集开始，该数据集包括 100 万个用户的个人资料，从中构建了两个数据集：1)沉默名人用户数据集；2)其他普通用户数据集。

4.2 沉默名人

沉默名人数据集(简称 MC)包含 1280 位用户(640 名女性和 640 名男性)，这些用户最多发布 5 条状态。基于以下原因，建立了这样一个数据集：

1)起初准备从原始的公共数据中选择真正的沉默名人，而满足该要求的只有 21 位，这个样本数量对于实验而言太少。因此，本文通过放宽条件来扩大数据集，将微博发布数量扩大到 5 条状态。

2)选择沉默名人而不是沉默的普通用户的原因在于，对于沉默的普通用户而言，如果没有潜在用户的其他信息(例如微博、照片)，将难以检查他们的性别。相比之下，名人的性别已被新浪验证，因此无需人工检查性别。该数据集用于验证所提方法对沉默用户的有效性，这也是本文研究的初始目标。

4.3 普通用户

普通用户(OU)的数据集包含 400 个用户(200 个女性和 200 个男性)，其中每人至少发布了 10 条微博状态。普通用户的数量远远少于沉默名人的数量，因为人工检查会耗费大量的时间。为了验证该数据集中的用户性别，我们招聘 3 名本科生进行人工检查数据，以确保：1)帐户不是垃圾用户或开微店的隐含集团用户等；2)用户性别是真实的。我们将采用两个及以上学生一致的标注结果来保证用户性别的真实性。

该数据集用于评估所提方法在普通用户上的效果。此外，我们希望使用从微博中提取的文本特征来将本文的用户嵌入方法与基于内容的方法进行比较，这也是将该数据集微博动态的最小数量设置为 10 条的原因。

对于 MC 和 OU 数据集，通过新浪提供的 API 来为每位用户下载 500 个朋友，通过这些朋友用户来构建隐性和显性的社交关系环境，以训练用户的嵌入式社交表示。这些用户形成第一层关系图。我们进一步收集二阶关系朋友(邻居的邻居)，使用相同的方法来构建第二层关系图。这部分额外的数据是为了开展基于图形的基线方法所必需的(见下文)。表 2 列出了两个数据集的统计量，其中第一层、第二层关系图的边和度分别用上标 1 和上标 2 表示。

表 2　数据集统计信息

数据集	OU	MC
用户数[1]	400	1280
边数[1]	66686	125340
平均度[1]	166.72	97.92
用户数[2]	51152	87202
边数[2]	1184586	2722699
平均度[2]	23.16	31.22

从表 2 中可以看出，在第一层关系图中 MC 的平均度数远小于 OU 的平均度数，可能的原因在于沉默名人用户和普通用户之间的区别。沉默用户不但在绝大多数时间不发布微博状态，而且在社交网络中也不活跃。另外，第二层关系图的平均度数明显小于第一层关系图，其原因在于资源限制，我们只保留属于第一层关系的节点以及该部分节点的二阶朋友关系节点，其余既不属于始 400 和 1280 个用户节点又不属于其二阶朋友关系的节点不在我们的考虑范围内。

5　实验

本文在第 4 节中介绍的两个真实数据集上进行了实验。将用户随机分为 5 份。在数据集上进行 5 折交叉验证，并取 5 折上的平均值作为最终的结果。由于该数据集为平衡数据集，因此使用准确率作为评价指标。

5.1　窗口大小的影响

窗口大小和维度是用户嵌入中的两个关键参数。本文先研究窗口大小对结果的影响，如表 3 所列。

表 3　窗口大小的影响/%

窗口大小	OU 数据集	MC 数据集
5	52.50	50.25
10	50.00	63.67
50	80.00	81.33
100	82.25	81.33
150	**82.50**	81.80
200	80.50	81.80
250	79.75	**82.03**
300	79.50	**82.03**
400	79.50	81.72
500	80.00	81.95

由表 3 可知，OU 和 MC 数据集分别在窗口大小为 150 和 250 时取得了最好的结果，比语言模型中通常使用的 5 或 10 的窗口大小更大。另外，当窗口大小为 5 时，两个数据集上的准确率分别为 52.00%和 50.25%，两者都是最差的，这表明大尺寸窗口策略适用于用户嵌入问题。本文使用最佳窗口大小作为后文实验中的默认值。

5.2　维度的影响

将维度从 50 逐渐调整至 300 来研究维度对实验结果的影响，如表 4 所列。

表 4　维度大小的影响/%

维度大小	OU 数据集	MC 数据集
50	79.75	81.88
100	82.50	**82.03**
150	82.00	81.48
200	82.50	81.95
250	**83.00**	81.25
300	82.50	81.88

从表 4 可以看出，在 MC 数据集上准确率的波动不如 OU 数据集上的明显，最大的变化仅为 0.78%，表明该数据集上的结果对维度较不敏感。另外，OU 和 MC 分别在维度为 250 和 100 时取得了最好的结果，分别为 83.00%和 82.03%。但是，为了与词嵌入和图嵌入方法进行公平比较，将在以下实验中将维数设置为 100，这与文献[12,22]中的设置相同。

5.3　序列化的影响

本节将评估 3 种类型的节点序列化方法（分别是随机排序、自然排序和度排序）对实验结果的影响，结果如表 5 所列。对于每一种排序方法，所有的结果都是在各自最佳窗口大小的设置下得到的。可以看出，当随机排序时，MC 上的准确率高于 OU。这与自然排序和度排序上的结果相反，表明随机排序不是一种稳定的方法。另外，度排序的结果是最好的，其原因可能在于在社交网络中度较大的节点倾向于相互连接，这也被称为富人俱乐部现象[8]。使用度排序有助于找到社交网络中的固有结构，从而提高准确率。

表 5　维度大小的影响/%

排序方法	OU 数据集	MC 数据集
随机序列	78.00	79.61
自然序列	81.00	79.53
按度排序	**82.50**	**82.03**

5.4　对比实验

在 6 个基线方法上进行了广泛的对比实验，这 6 种基线方法包括基于内容、基于图形、图形嵌入和用户嵌入方法。基线方法的具体描述如下。

1）词频（WF）：该方法将微博内容作为特征。微博中的每个单词被表示为对应维度上的向量，如 $x:y$，其中 x 是词的 ID，y 表示该词出现的频度。这是在性别分类任务中被广泛使用的表示方法。

2）词嵌入（WE）：该方法使用用户和他们的朋友的微博文本作为语料，训练出词的分布式表示。按照文献[12]中的做法，将维度设为 200，并使用层次 softmax 来加速训练过程。

3）原始图形表示（TGR）：该方法是一种传统的独热表示方法，向量的每一维分别表示为 $x:y$，其中 x 为用户 ID，y 的取值为 1 或 0，用来表示 x 是否出现在该用户的朋友中。

4）邻居向量表示（NVR）：该方法用邻居向量表示每个用户，向量的每一维分别为 $x:y$，其中 x 为二阶邻居节点，y 代表其粉丝中与其它每个邻居是朋友关系的比例。本文实验将严格采用文献[9]中的设置。

5）图形嵌入（LINE）：该方法包含两个子基线方法 LINE_1 和 LINE_{1+2}，分别使用用户的一阶、二阶朋友构建图形嵌入网络。在实验中，将使用文献[22]中的默认设置，即维度为 100，负样本数 $K=5$，样本总数 $T=100$ 亿。

与基线方法的对比结果如表 6 所列。

表 6　与基线方法的对比/%

基线方法		OU 数据集	MC 数据集
Content based	WF	76.00	*
	WE	74.75	*
Graph based	TGR	79.53	76.00
	NVR	66.00	58.44
Graph embedding	LINE_1	73.00	74.06
	LINE_{1+2}	71.50	68.44
User embedding	UE	**82.50**	**82.03**

从表 6 中可以发现：

(1)所提的 UE 方法是所有方法中最有效的。与其他基线方法相比，本方法在置信度为 0.95 的显著性检验下有显著的提升。首先，相较于基于图形的 NVR 基线，UE 在 OU 和 MC 数据集上分别提升了 16.50%和 23.59%，提升幅度十分显著。其次，UE 的结果也大幅优于两种图形嵌入方法。OU 上的准确率从 73.00%($LINE_1$)提升至 82.50%。$LINE_{1+2}$在 OU 和 MC 上的表现也比 UE 差得多。这清楚地表明，UE 比目前最好的图形嵌入方法 LINE 更有效，资源利用效率更高。最后，UE 明显好于基于内容的方法 WF 和 WE。这一点强有力地指出了 UE 的潜在应用场景。对于分类任务，不管是否有文本信息，UE 都是一个较好的选择。

(2)传统的 one-hot 表示 TGR 的结果是次好的。这表明用户性别与他/她的朋友高度相关。该发现表明了女性用户的社交网络中大多是女性，同样的结论也适用于男性用户。此外，TGR 也优于图形嵌入方法 LINE，这表明图形嵌入可能存在一定程度上的信息损失。相比之下，所提的用户嵌入方法 UE 同时对隐性和显性两种社会关系建模，而不只对显性的连接关系建模，因此所提方法可以得到更好的用户表示。结果表明另一种基于图形的方法 NVR 是最差的，原因可能在于 NVR 只能保持间接用户之间的关系。

(3)图形嵌入方法 $LINE_{1+2}$的性能比 $LINE_1$ 更差，这与文献[22]中的结果不符。原因可能在于实验中去除了那些节点不在原始用户集中的二阶关系边。因此，第二层关系图的平均度远小于第一层关系图，特别是在 OC 数据集上。该发现很重要，它表明如果没有足够多的节点和边，二阶关系可能会降低实验结果。然而，补充足够多的节点和边又需要额外的开销。通过实验还发现，WF 的效果比 WE 更好，表明基于内容的方法并未从词嵌入技术中获益，这可能是由微博主题和微博词汇的多样性造成的。

结束语 本文提出一种新的社交表示方法来捕捉社交媒体中用户之间的显性和隐性关系，通过修改词嵌入技术以有效利用社会关系语境，在性别分类方面取得了非常准确的结果。在两个真实的新浪微博数据集上进行了广泛的实验，实验结果表明本文的用户嵌入方法明显优于传统的基于图形的方法和目前最好的图形嵌入算法。很显然，当文本难以获取时所提方法非常有效。此外，所提方法也优于基于内容的方法，从而有力地证明了所提方法对于各个类型的用户都有普遍的应用价值。

未来将在 Twitter 或 Facebook 数据集上进行更多的实验，进一步研究如何将所提方法应用到其他用户分析的任务。

参 考 文 献

[1] AHMED A, SHERVASHIDZE N, Narayanamurthy S, et al. Distributed large-scale natural graph factorization[C]// International Conference on World Wide Web. ACM, 2013: 37-48.

[2] ALOWIBDI J S, BUY U A, YU P. Empirical Evalu-ation of Profile Characteristics for Gender Classification on Twitter[C]// International Conference on Machine Learning and Applications. 2013: 365-369.

[3] BAMMAN D, EISENTEIN J, SCHNOEBELEN T. Gen-der identity and lexical variation in social media[J]. Journal of Sociolinguistics, 2014, 18(2): 135-160.

[4] BENGIO Y, SCHWENK H, SENÉCAL J, et al. Neural Probabilistic Language Models[J]. Journal of Machine Learning Research, 2001, 3(6): 1137-1155.

[5] BERGSMA S, DURME B V. Using Conceptual Class Attributes to Characterize Social Media Users[C]// Meeting of the Association for Computational Linguistics. 2013: 710-720.

[6] BURGER J D, HENDERSON J, KIM G, et al. Discrimi-nating gender on Twitter[C]// Conference on Empirical Methods in Natural Language Processing (EMNLP 2011). 2011: 1301-1309.

[7] CHENG N, Chen HEN, Chandramouli R, et al. Gender identification from E-mails[C]// IEEE Symposium on Computational Intelligence and Data Mining, 2009 (CIDM '09). IEEE, 2009: 154-158.

[8] COLIZZA V, FLAMMINI A, SERRANO rrano M A, et al. Detecting rich-club ordering in complex networks[J]. Nature Physics, 2006, 2(3): 110-115.

[9] CULOTTA J C A, KUMAR N R. Predicting the demographics of Twitter users from website traffic data[C]// Proc. 29th Conf. on AI. 2015: 72-78.

[10] FILIPPOVA K. User demographics and language in an implicit social network[C]// Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 1478-1488.

[11] GOSWAMI S, SARKAR S, RUSTAGI M. Stylometric Analysis of Bloggers' Age and Gende[C]// International Conference on Weblogs and Social Media (ICWSM 2009). San Jose, California, Usa, DBLP, 2009.

[12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distri-buted Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[13] MNIH A, HINTON G. Three new graphical models for statistical language modelling[C]// Machine Learning, Proceedings of the Twenty-Fourth International Conference. DBLP, 2007: 641-648.

[14] MUKHERJEE A, LIU B. Improving gender classi-fication of blog authors[C]// Conference on Empirical Methods in Natural Language Processing (EMNLP 2010). 2010: 207-217.

[15] OTTERBACHER J. Inferring gender of movie re-viewers: exploiting writing style, content and meta-data[C]// ACM Conference on Information and Knowledge Management (CIKM 2010). Toronto, Ontario, Canada, DBLP, 2010: 369-378.

[16] PEERSMAN C, DAELEMANS W, VAERENBERGH L V. Predicting age and gender in online social net-works[C]// International CIKM Workshop on Search and Mining User-Generated Contents (Smuc 2011). Glasgow, United Kingdom, DBLP, 2011: 37-44.

[17] PENNACCHIOTTI M, POPESCU A M. A Machine Learning Approach to Twitter User Classification [C] // International Conference on Weblogs and Social Media. Barcelona, Catalonia, Spain, DBLP, 2011.

[18] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 701-710.

[19] RAO D, YAROWSKY D, SHREEVATS A, et al. Classi-fying latent user attributes in twitter[C]// Interna-tional Workshop on Search and Mining User-Generated Contents. ACM, 2010: 37-44.

[20] SCHLER J, KOPPEL M, ARGAMON S, et al. Effects of Age and Gender on Blogging[J]. Frontiers of Information Technology & Electronic Engineering, 2006, 274(s1/2): 199-205.

[21] TANG C, ROSS K, SAXENA N, et al. What's in a name: a study of names, gender inference, and gender behavior in face book[M]// Database Systems for Adanced Applications. Springer Berlin Heidelberg, 2011: 344-356.

[22] TANG J, QU M, WANG M, et al. LINE: Large-scale Information Network Embedding [C] // International Conference on World Wide Web. ACM, 2015: 1067-1077.

[23] XIAO C, ZHOU F, WU Y. Predicting audience gender in online content-sharing social networks[J]. Journal of the Association for Information Science and Technology, 2013, 64(6): 1284-1297.

（上接第 135 页）

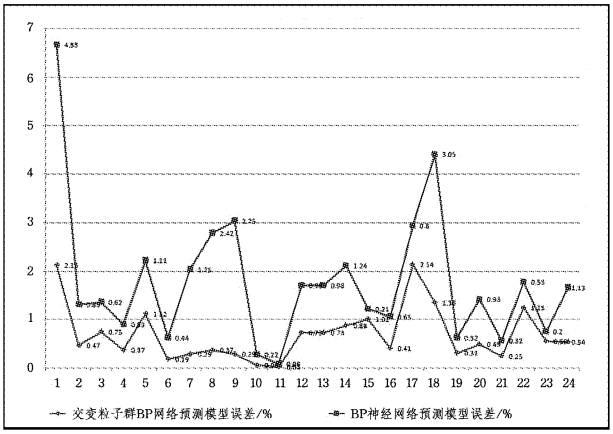

图 3　两种电荷预测模型的预测数据误差对比图

结束语　粒子群优化算法(PSO)是一种模拟鸟群行为的优化算法，该算法简单易实现且参数少，被广泛应用于数值优化、参数调优等多种领域。但该算法也存在早收敛及收敛过慢、易求局部最优值而不是全局最优值等缺点。本文针对粒子群算法的缺点，引入交叉操作来更新粒子的位置，利用变异操作以及设定粒子能量值和阈值来保持粒子的多样性，保证算法不过早收敛和陷入局部最优；并提出交变粒子群算法，利用该算法优化 BP 网络预测模型并应用于短期负荷预测，采用电力部门历史负荷数据来验证该预测模型的有效性及精度。实验结果表明采用本文提出的预测模型预测的短期电力负荷的精度较高。

参 考 文 献

[1] 刘晨晖. 电力系统负荷预报理论和方法[M]. 哈尔滨：哈尔滨工业大学出版社，1987.

[2] XIE K G, LI C Y, ZHOU J Q. Research of the combination forecasting model for load based on artificial neural network [J]. Proceedings of the CSEE, 2002, 22(7): 85-89.

[3] YOU Y, SHENG W X, WANG S A. Short-term load forecasting using artificial immune network[J]. Proceedings of the CSEE, 2003, 23(3): 26-30.

[4] ZHENG G, LIU B, ZHOU Y, et al. Short-term load forecasting based on neural network[J]. Journal of Xi'an University of Technology, 2002, 18(2): 126-130.

[5] TAI N L, HOU Z J. New short-term load forecasting principle with the wavelet transform fuzzy neural network for the power systems[J]. Proceedings of the CSEE, 2004, 24(1): 24-29.

[6] YOU Y, SHENG W X, WANG S A. The study and application of the electric power system short-term load forecasting using a new model[J]. Proceedings of the CSEE, 2002, 22(9): 15-18.

[7] JIANG Y. Short-term load forecasting using a neural network based on fuzzy clustering[J]. Power System Technology, 2003, 27(2): 45-49.

[8] YAO L X, YAO J X, LI B Q, et al. Short-term load forecasting using neural network based on competitive learning classification [J]. Power System Technology, 2004, 28(10): 45-48.

[9] MCCLELLAND J L, RUMELHART D E. Parallel distributed processing [M]. Cambridge: MIT Press, 1986.

[10] LARSEN E V, MILLER N W, NILSSON S L, et al. Benefits of GTO-based compensation systems for electric utility applications[J]. IEEE Trans on Power Delivery, 1992, 7(4): 2056-2061.

[11] CEN W H, LEI Y K, XIE H. The application of the electric power system short-term load forecasting using artificial network and genetic algorithm[J]. Automation of Electric Power Systems, 1997, 21(3): 29-32.

[12] DING J W, SUN Y M. Short-term load forecasting using a neural network based on chaos learning algorithm[J]. Automation of ElectricPower Systems, 2000, 24(2): 32-35.

[13] YANGQ Y, SUN J G, ZHANG J Y, et al. A Hybrid Discrete Particle Swarm Algorithm for Open-Shop Problems[C]// Proceedings of the 6th International Conference on Simulated Evolution And Learning (SEAL 2006). Hefei, China, LNCS 4247, 2006: 158-165.

[14] VAN DEN BERGH F. An Analysis of Particle Swarm Optimizers[D]. Department of Computer Science, University of Pretoria, Pretoria, South Africa, 2002.

[15] RAMESHKUMAR K, SURESH R K, Mohanasundaram K M. Discrete Particle Swarm Optimization (DPSO) Algorithm for Permutation Flowshop Scheduling to Minimize Makspan[C]// Proc. ICNC 2005. LNCS 3612, 2005: 572-581.

[16] LIAN Z G, GU X S, JIAO B. A similar particle swarm optimization algorithm for permutation flowshop scheduling to minimize makespan[J]. Applied Mathematics and Computation, 2006, 175(1): 773-785.