

融合不完整多视图的异质信息网络嵌入方法

郑苏苏 关东海 袁伟伟

南京航空航天大学计算机科学与技术学院 南京 211106

软件新技术与产业化协同创新中心 南京 211106

(zhengsusu@nuaa.edu.cn)

摘要 异质信息网络(Heterogeneous Information Network, HIN)嵌入将复杂的异质信息映射到低维稠密的向量空间,有利于网络数据的计算和存储。现有的基于多视图的 HIN 嵌入方法考虑了节点之间的多种语义关系,但忽略了视图的不完整性。大多数视图存在数据缺失,直接融合多个不完整的视图会导致嵌入效果不佳。为此,文中提出了一种融合不完整多视图的 HIN 嵌入方法(Incomplete Multi-view Fusion Based HIN Embedding, IMHE)。IMHE 的关键思想是聚合其他视图的邻居以重建不完整的视图。由于不同的单视图描述的是同一个网络,因此其他视图中的邻居可以一定程度上恢复不完整视图的结构信息。IMHE 首先在不同视图中生成节点序列,并利用多头注意力方法学习单视图嵌入。对于每个不完整视图,IMHE 在其他视图中找到缺失节点的 k 阶邻居,然后将不完整视图中邻居的单视图嵌入聚合在一起,为缺失节点生成新的嵌入。最后使用多视图典型相关性分析方法获得节点的统一嵌入,同时提取多个视图的隐藏语义关系。在 3 个真实数据集上的实验结果表明,相比现有研究,该方法的嵌入性能有显著提升。

关键词: 不完整视图;多视图融合;异质信息网络;网络嵌入

中图分类号 TP391

Heterogeneous Information Network Embedding with Incomplete Multi-view Fusion

ZHENG Su-su, GUAN Dong-hai and YUAN Wei-wei

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211106, China

Abstract Heterogeneous information network (HIN) embedding maps complex heterogeneous information to a low-dimensional dense vector space, which is conducive to the calculation and storage of network data. Most existing multi-view-based HIN embedding methods consider multiple semantic relationships between nodes, but ignore the incompleteness of the view. Most of views are incomplete and directly fusing multiple incomplete views will affect the performances of the embedding model. To address this problem, we propose a novel HIN embedding model with incomplete multi-view fusion, named IMHE. The key idea of IMHE is to aggregate neighbors of other views to reconstruct the incomplete views. Since different views describe the same HIN, neighbors in other views can restore the structure information of the missing nodes. The IMHE model first generates nodes sequences in different views, and leverages the multi-head self-attention method to obtain single-view embedding. For each incomplete view, IMHE finds the k -order neighbors of the missing nodes in other views, then aggregates the embeddings of neighbors in the incomplete view to generate new embeddings for missing nodes. IMHE finally uses the multi-view canonical correlation analysis method to obtain the joint embedding of nodes, thereby simultaneously extracting the hidden semantic relationship of multiple views. Experiment results on three real-world datasets show that the proposed method is superior to the state-of-the-art methods.

Keywords Incomplete view, Multi-view fusion, Heterogeneous information network, Network embedding

1 引言

基于多视图的 HIN 嵌入方法是一种用于建模异质信息网络的新兴方法,其主要思想是将 HIN 的多个视图嵌入到统

一的低维向量空间,并尽可能保留原始网络的结构和语义信息^[1-2]。不同的元路径可以将 HIN 划分为多个代表不同语义空间的子网络,因此多视图方法可以融合多个语义空间以学习网络嵌入^[3]。HIN 包含多种类型的节点和边,通常将它们

到稿日期:2021-05-28 返修日期:2021-06-24

基金项目:江苏省重点研发计划项目(BE2019012);国家自然科学基金委员会—中国民用航空局民航联合研究基金项目(U2033202)

This work was supported by the Key Research and Development Program of Jiangsu Province(BE2019012) and Joint Fund of National Natural Science Foundation of China and Civil Aviation Administration of China(U2033202).

通信作者:袁伟伟(yuanweiwei@nuaa.edu.cn)

存储为高维向量,直接对其进行分析很容易导致维数灾难^[4]。由于低维向量易通过机器学习算法进行处理,因此多视图 HIN 嵌入方法可以节省存储空间并提高网络分析任务的效率^[5],例如节点分类、链接预测、社区发现和推荐。

现有的基于多视图的 HIN 嵌入方法通常假设每个视图都是完整的并直接融合这些视图,这导致无法解决数据稀疏问题并且难以提取 HIN 的隐藏语义关系^[6]。在现实世界的应用中,由于隐私问题、数据缺失或其他原因,只能观察到网络中部分的节点和边,从而导致单个视图的不完整。由于单个视图的不完整性,不同视图中的节点不能一一对应,即节点在某些视图中存在,但在其他视图中不存在。近年来,部分学者注意到了网络嵌入中多视图数据不完整的情况,并提出了基于多视图的同质信息网络嵌入方法,在不完整的网络上学习节点嵌入^[7]。但这些方法不能直接应用在结构更为复杂的异质信息网络上,忽视多视图不完整的问题将对异质信息网络嵌入带来以下问题和挑战^[8]。首先,不考虑多个视图的互补性而直接融合不完整的视图难以学习 HIN 的隐藏语义信息。其次,不完整视图的信息缺失使模型无法解决数据稀疏带来的一系列问题。这些都将直接影响 HIN 的嵌入性能。

图 1 给出了 DBLP 数据集上不完整多视图的示例,图 1 (b)中的每个视图代表一个特定的语义空间。此示例由 4 种类型的节点(作者、论文、场所和主题)和 3 种类型的关系(撰写、发表和涉及)组成。3 个视图分别是合著者视图、相同场所视图和相同主题视图。合著者视图意味着作者可以共同发表同一篇论文。相同场所视图表示不同作者撰写的论文在同一场所发表。相同主题视图表示不同作者撰写的论文属于同一主题。从示例中发现,不同的语义空间无法覆盖所有节点,因此每个视图是不完整的。在单视图网络(例如合著者视图)中,作者 a_4 与其他作者没有任何关系,这使得无法在此语义空间中学习节点 a_4 的向量表示。幸运的是,作者 a_4 在其他两个视图中具有许多其他类型的连接。在这种情况下,我们可以通过学习相同场所视图和相同主题视图的语义和结构信息来补全合著者视图中的节点信息,增强整个网络的嵌入。

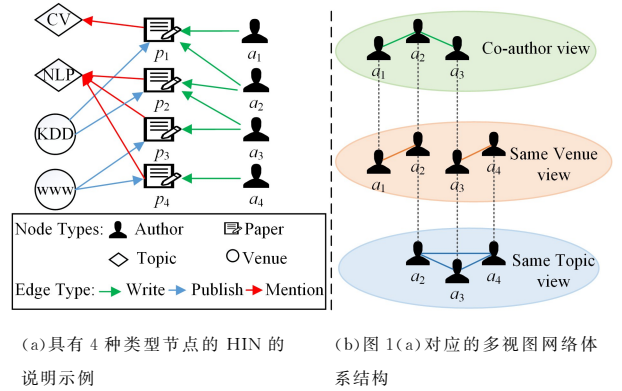


图 1 DBLP 数据集上不完整多视图的示例

Fig. 1 Example of incomplete multi-view on DBLP dataset

为了解决现有工作中存在的问题,本文提出了一种基于不完全多视图融合的 HIN 嵌入方法,称为 IMHE。IMHE 的关键思想是通过聚合其他视图的邻居信息来补全不完整视

图。尽管不同的视图具有不同的语义,但是它们都描述相同的 HIN,存在多视图之间的一致性。因此,聚合其他视图的邻居可以在一定程度上恢复缺失节点的结构信息。为了保留丰富的语义信息,IMHE 方法首先基于多头注意力模型学习单视图嵌入。根据不同视图的元路径生成节点序列,并通过在单个视图中挖掘节点序列的相关性来学习节点嵌入。然后,单视图补全通过聚合其他视图中节点的邻居来获得缺失节点的嵌入。最后是多视图融合,本文提出了多视图典型相关性分析(Multi-view CCA)方法,通过融合多视图来学习统一的 HIN 嵌入,从丰富的多视图信息中进一步提高学习性能。

本文的贡献如下:

- (1)提出的 IMHE 方法可以解决视图不完整的问题。通过邻域聚合模型来补全单视图信息,并且可以在多视图中提取隐藏的语义信息,从而增强 HIN 嵌入性能。
- (2)通过融合 HIN 中多个不同的语义视图,降低了数据稀疏性的影响。改进的 Multi-view CCA 方法融合了互补的多视图结构信息,并保留了 HIN 的原始异质语义信息。

2 相关工作

异质网络嵌入模型可以分为两类。

- (1)基于随机游走^[9]的 HIN 嵌入通过提取网络的异质节点获得节点序列,然后根据序列中节点之间的关系来学习低维向量空间中的节点表示。Metapath2vec^[10]方法使用基于元路径的随机游走对 HIN 中的节点进行采样,然后使用自然语言处理模型 skip-gram^[11]来学习节点嵌入。Esim^[12]方法同样使用基于元路径的随机游走方法来获取节点序列。与 Metapath2vec 不同,Esim 考虑了多个元路径,并使用一组权重组合来融合多个元路径的信息。MAGNN^[13]通过 3 个步骤生成节点嵌入:节点内容转换、元路径内的聚合和元路径间的聚合。与基于给定元路径的上述方法不同,Hin2vec^[14]模型根据节点之间是否存在关系来生成随机游走。通过建立神经网络模型,Hin2vec 不仅可以学习网络中节点的表示,而且可以学习关系的表示。上述基于随机游走的嵌入方法均有其局限性。Metapath2vec 仅考虑一个元路径;Esim 很难找到元路径的最佳权重组合;Hin2vec 仅考虑节点之间的低阶关系,不能充分保留 HIN 中的复杂语义信息。尽管 MAGNN 考虑到了不同视图的融合,但并未考虑视图的不完整性,这可能会导致过度拟合的问题^[15]。

(2)基于元路径分解的 HIN 嵌入

基于元路径分解的 HIN 嵌入方法首先根据不同的元路径将异质信息网络分解为多个同质网络,然后学习每个同质网络上的节点表示,最后将节点表示与不同的元路径进行深度合并以获得最终表示。节点的 PTE^[16]方法首先将异质文本网络划分为 3 个两部分的同质子网络,然后总结每个子网络中的损失函数,最后在相同的低维向量空间中获得节点嵌入。EOE^[17]方法将不同元路径下的节点之间的关系建模为多个同质子网络,学习多个同质子网络中的节点嵌入,并使用协调嵌入矩阵将不同子网络的节点合并,并将其映射嵌入到同一空间。它们之间的区别在于,PTE 通过联合损失函数学

习多个同质子网络的统一节点嵌入,而 EOE 首先学习不同同质子网络的节点嵌入,然后将其映射到同一空间。基于深度学习的模型^[18]可以更好地建模异质网络中的非线性关系。HAN 方法^[19]使用语义级别的注意力机制和节点级别的注意力机制来学习了解元路径和节点邻居的重要性,并汇总多个元路径的节点表示以获得最终的节点表示。这些基于元路径分解的 HIN 嵌入方法是直观的,但是它们不能有效地融合不同的同质子网络,也不能区分同一节点在不同子网络中的重要性,这会影响节点嵌入的效果。

由于隐私保护和其他原因,大多数视图是不完整的,从而导致节点在不同视图中不能一一对应,即某些节点存在于某些视图中,但在其他视图中没有与该节点有关的信息。不考虑视图不完整的问题,直接融合不完整的多视图信息可能导致过度拟合,学习到的节点嵌入不能充分利用 HIN 的多视图信息,从而影响后续应用的效果。本文使用基于元路径的随机游走方法来学习节点嵌入,补全不完整的单视图嵌入,并使用多视图方法来融合多个元路径学习到的节点表示。

3 相关定义

定义 1 异质信息网络定义为一个有向图 $G=(V, E, T, R)$, 每个节点 $v_i \in V$ 属于节点类型集合 T 中的一个节点类型, 每个边 $e_i \in E$ 属于边类型集合 R 中的一个边类型。它们的类型映射函数分别表示为: $V \rightarrow T$ 和 $E \rightarrow R$ 。

定义 2 元路径 P 定义为 $T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \dots \xrightarrow{R_{l-1}} T_l$ 的路径, 简称为 $T_1 T_2 \dots T_l$ 。它描述了节点类型 T_1 和 T_l 之间的关系 $R=R_1 \circ R_2 \circ \dots \circ R_{l-1}$, 其中 $\circ$ 表示关系间的组合运算符。

定义 3 单视图 HIN 嵌入学习 HIN 在低维向量空间中的节点嵌入, 保留特定语义空间的单视图信息。对于给定的 $G=(V, E, T, R)$, 用于单视图 HIN 嵌入的语义空间表示为 $M_i (i=1, 2, \dots, n)$ 。单视图 HIN 嵌入表示为映射 $f': V \rightarrow R^d$, 将每个节点 v_i 投影到 d 维空间中的向量, 并保留在单个视图中的语义信息。

定义 4 多视图 HIN 嵌入学习 HIN 在低维向量空间中的节点嵌入, 保留了不同语义空间的多视图信息。对于给定的 $G=(V, E, T, R)$, 用于多视图 HIN 嵌入的语义空间集合表示为 $M=\{M_1, M_2, \dots, M_n\}$, $M_i (i=1, 2, \dots, n)$ 是用于单视图 HIN 嵌入的特定语义空间。多视图 HIN 嵌入表示为 $f: M \rightarrow V \in R^d$ 的映射, 将每个节点 v_i 投影到由多个单视图 HIN 嵌入融合而成的 d 维空间 R^d 中。

4 融合不完整多视图的 HIN 嵌入方法

为了解决 HIN 嵌入中多视图信息不完整的问题, 本文提出了一种灵活的模型来融合多个不同视图的语义信息。提出的用于不完整多视图 HIN 嵌入的新模型称为 IMHE, 具体包括 3 个步骤: 1) 单视图嵌入; 2) 单视图补全; 3) 多视图融合。图 2 给出了所提方法的框架。图 2 中, 圆圈和正方形表示节点和节点嵌入, 白色圆圈和正方形表示缺失的节点和节点嵌入。

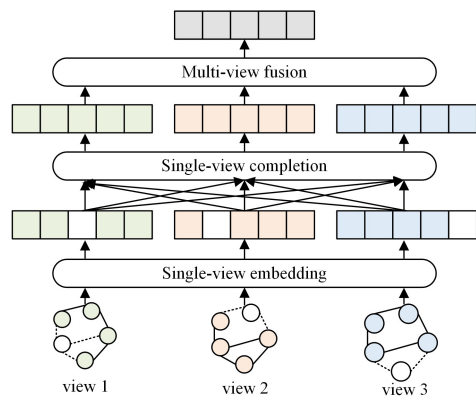


图 2 IMHE 框架的体系结构(电子版为彩色)

Fig. 2 Architecture of IMHE framework

4.1 单视图嵌入

单视图 HIN 嵌入将每个视图的语义和结构信息保留在 $M=\{M_1, M_2, \dots, M_n\}$ 中, 其中 $M_i (i=1, 2, \dots, n)$ 是 HIN 中存在的单视图子网, n 是视图数。每个元路径代表特定的语义空间, 可以用作划分不同视图的基础。给定一个元路径 P_i , 其对应的语义空间由视图 M_i 表示。基于元路径的随机游走通常通过保留节点的语义接近度来学习 HIN 嵌入^[20]。现有的大多数工作都使用 Skip-Gram 模型进行随机游走的序列学习, 例如 Metapath2vec^[10] 和 MARU^[21]。这些模型仅学习节点的共现关系, 而忽略了序列中的顺序信息和节点之间关系的权重。为了解决这个问题, 本文方法在单视图嵌入中引入了自注意力^[22-23]模型。自注意力可以学习节点序列中节点之间的权重, 并根据节点和相应的权重获得节点嵌入。单视图嵌入由 3 层组成, 即嵌入层、多头注意力层和均值合并层, 如图 3 所示。图 3 中, 彩色圆圈和正方形表示节点和节点嵌入, 而灰色正方形表示位置嵌入。

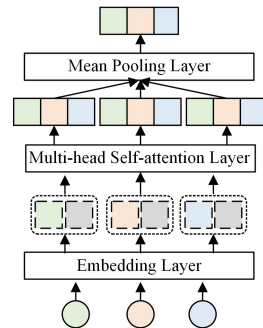


图 3 基于多头注意力的单视图 HIN 嵌入(电子版为彩色)

Fig. 3 Multi-head self-attention based Single-view HIN embedding

第一层是嵌入层, 它将所有节点嵌入到固定大小的向量中。在不同视图中按照给定的元路径进行随机游走获得节点序列, 并将其用作嵌入层的输入。嵌入层的关键思想是将序列中的每个节点映射到相应的嵌入向量。为了在嵌入学习的过程中捕获节点序列的顺序信息, 我们首先将两个嵌入矩阵 H_V 和 H_P 随机初始化, 其中 $H_V \in R^{|V| \times d}$ 是节点嵌入矩阵, $H_P \in R^{L \times d}$ 是位置嵌入矩阵, L 为序列长度。位置嵌入用于表示序列中节点的位置, 然后将 H_V 和 H_P 连接起来以更好地表示节点嵌入: $H=H_V+H_P$ 。其中, H 是融合位置信息后的节点嵌入矩阵, 并且作为多头注意力层的输入。

第二层是多头注意力层,它在网络嵌入中引入了多个自注意力机制^[24],以捕获节点之间的不同隐藏关系。对于序列中的每个节点,自注意力机制根据节点和其他节点的重要性来学习注意力分数,并根据注意力分数聚合其他节点的嵌入。为了提取 HIN 中的隐藏语义关系,我们计算节点序列的 k 头注意力:

$$Attention(Q, K, V) = Common(\parallel_{i \in [1, k]} Att-Head^i(Q^i, K^i, V^i)) \quad (1)$$

其中, Q, K, V 是通过嵌入节点 H 的不同线性变换获得的矩阵。 Q 表示查询, K 表示键, V 表示值。每个自注意力机制定义为:

$$Att-Head^i(Q^i, K^i, V^i) = \text{Softmax}\left(\frac{Q^i K^{iT}}{\sqrt{d}}\right) V^i \quad (2)$$

其中, d 为缩放参数,防止结果变得太大。然后,通过值向量的加权求和获得每个节点的注意力输出。

最后一层是均值合并层,在均值合并层中输入 k 个自注意力向量以获得每个节点的嵌入向量。

$$H = \text{Mean}(\parallel_{i \in [1, k]} Att-Head^i(Q^i, K^i, V^i)) \quad (3)$$

4.2 单视图补全

对于每个不完整视图,单视图补全通过聚合重建邻居的嵌入来填充缺失节点的嵌入,如图 4 所示。

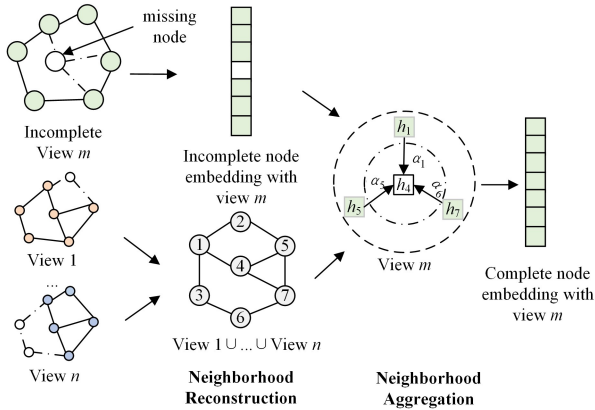


图 4 基于邻域聚合的单视图补全模型

Fig. 4 Neighborhood aggregation based single-view completion

算法 1 给出了单视图补全的详细信息。尽管不同的视图具有不同的语义,但是它们都描述相同的 HIN,不同的视图之间存在结构上的一致性。当某个视图中的节点缺失时,首先在其他视图中找到该节点的邻居。由于当前的不完整视图并不包含所有的邻居节点,因此通过这些邻居节点与不完整视图的交集来重建缺失节点的邻域。为了深入提取不完整视图中的隐藏语义信息,本文方法在不完整视图中聚合重构邻居的单视图表示,以学习缺失节点的嵌入:

$$h_i^m = \text{AGGREGATE}(\{h_j^m, \forall v_j \in N_i^m\}) \quad (4)$$

其中, $h_i^m \in H^m$ 是不完整视图 M_m 中缺失节点 v_i^m 的嵌入, $v_j \in M_i^m$ 是节点 v_i^m 的重建邻居, $h_j^m \in H_m$ 是视图 M_m 中节点 v_j 的节点嵌入。由于不同的邻居对缺失的节点具有不同的关系强度,因此在聚合邻居时应区分邻居节点的重要性。本文方法同时考虑节点的一阶邻居和高阶邻居,一阶邻居的重要性根据节点的邻接矩阵进行归一化计算,高阶邻居的重要性根据

马尔可夫链的转移概率来计算,即网络上随机游走概率较高的邻居将具有较高的权重^[25]。对于所有节点,其邻居的权重矩阵为:

$$P = \begin{cases} D^{-1}A, & \text{if } k=1 \\ P^k, & \text{if } k>1 \end{cases} \quad (5)$$

其中, D 是度矩阵, A 是邻接矩阵, P^k 是 k 阶邻居的嵌入权重矩阵。通过将邻居的嵌入从 1 阶到 k 阶进行聚合,来填充不完整视图中的所有缺失节点,以得到一个完整的节点嵌入矩阵。

$$\hat{H}_m = \sigma(\sum_{i=1}^k P^i H_m) \quad (6)$$

算法 1 基于邻域聚合的单视图补全算法

输入:多视图集合 $M = \{M_1, M_2, \dots, M_n\}$, 多视图节点集合 $V = \{V_1, V_2, \dots, V_n\}$, 不完整单视图嵌入 $H = \{h_1, h_2, \dots, h_n\}$, 邻居节点阶数 k

输出:补全后的单视图嵌入 $\hat{H}_m$

1. 计算不完整视图 M_m 中的缺失节点集合 $V_s \leftarrow V/V_m$

2. for $v_i \in V_s$ do

3. 计算缺失节点在其他视图中的邻居集合

$$N_i \leftarrow N_i^1 \cup N_i^{m-1} \cup \dots \cup N_i^{m+1} \cup N_i^n$$

4. 重建不完整视图 M_m 中缺失节点的邻居 N_i^m

$$N_i^m \leftarrow N_i \cap V_m$$

5. 计算邻居 $v_j \in N_i^m$ 的权重

6. if $k=1$ then

$$p_j = P_{ij} = \frac{1}{\text{degree}_i}$$

8. else

$$p_j = P_{ij}^{(k)} = \sum_{l \in V} P_{ij}^{(n)} P_{ij}^{(k-n)}$$

10. end if

11. 聚合不同邻居的节点嵌入

$$h_i^m = \sigma(\sum_{j \in N_i^m} p_j \cdot h_j^m)$$

$$\hat{H}_m \leftarrow H_m \cup h_i^m$$

12. end for

13. return $\hat{H}_m$

4.3 多视图融合

多视图融合通过邻域聚合完成的单视图嵌入来学习统一的 HIN 嵌入。这些单视图嵌入可用于多视图融合的输入: $H = \{H_1, H_2, \dots, H_n\}$ 。多视图融合学习一个公共的隐藏空间,所有视图空间对该公共隐藏空间的投影都是近似的。典型相关性分析(CCA)^[26]方法是用于挖掘数据关联关系的常用算法之一。由于 CCA 方法只能集成两个视图,不能应用于 HIN 中有两个以上视图的情况,因此本文提出 Multi-view CCA 来融合两个或更多视图。令 S 代表隐藏空间, F 代表从 S 到潜在向量空间的映射函数 $F_i: S \rightarrow R^d$ 。对于节点 v_i ,其在 S 中的对应嵌入应满足:

$$s_i = F_1^{-1}(H_1(v_i)) = F_2^{-1}(H_2(v_i)) = \dots = F_m^{-1}(H_m(v_i)) \quad (7)$$

其中, $v_i \in V$ 是 HIN 的节点, s_i 是 v_i 在 S 中的对应表示, H_1

$(v_i), \mathbf{H}_2(v_i), \dots, \mathbf{H}_n(v_i)$ 是在视图 $M = \{M_1, M_2, \dots, M_m\}$ 的单视图嵌入。Multi-view CCA 使用 CCA 模型融合成对的单视图嵌入,并最大化两个向量空间的投影之间的相关性。

$$F_1^{-1}, F_2^{-1}, \dots, F_m^{-1} = \max_{F_1^{-1}, F_2^{-1}, \dots, F_m^{-1}} \sum_{i=1}^m \sum_{j=i+1}^m \rho(F_i^{-1}(\mathbf{H}_i(V)), F_j^{-1}(\mathbf{H}_j(V))) \quad (8)$$

其中, ρ 是两组投影 $\mathbf{H}_i(V)$ 和 $\mathbf{H}_j(V)$ 之间的相关性。Multi-view CCA 的原理是首先使用 CCA 模型对成对的单视图嵌入进行融合,融合后的嵌入用作下一轮融合的新向量,并进行迭代直到剩下一个向量空间为止。最终得到的嵌入向量是多视图嵌入。当视图数为奇数时,只需取任一单视图向量空间并将其放入融合过程中。算法 2 给出了多视图融合的详细信息。

算法 2 基于多视图 CCA 的融合嵌入算法

输入:单视图补全后的节点嵌入集合 $\mathbf{H} = \{\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_n\}$

输出:多视图嵌入 $\mathbf{H}^*$

```
1.  $\mathbf{H}_0 \leftarrow \mathbf{H}$ 
2.  $\mathbf{H}_1 \leftarrow \text{empty}$ 
3.  $c = 0$ 
4. while  $|\mathbf{H}_0| > 1$  do
5.   if  $|\mathbf{H}_0|$  为奇数 then
6.      $c = (|\mathbf{H}_0| - 1) / 2$ 
7.      $\mathbf{H}_1 \leftarrow \mathbf{H}_1 \cup \mathbf{H}_0[c : |\mathbf{H}_0|]$ 
8.   else
9.      $c = |\mathbf{H}_0| / 2$ 
10.  end if
11. for  $i = 0, 1, 2, \dots, c$  do
12.    $S_1 \leftarrow \mathbf{H}_0[2i]$ 
13.    $S_2 \leftarrow \mathbf{H}_0[2i+1]$ 
14.    $(F_1^*, F_2^*) = \max_{\rho}(F_1^{-1} \cdot S_1, F_2^{-1} \cdot S_2)$ 
15.    $S \leftarrow (F_1^* \cdot S_1, F_2^* \cdot S_2) / 2$ 
16.    $\mathbf{H}_1 \leftarrow \mathbf{H}_1 \cup S$ 
17. end for
18.  $\mathbf{H}_0 \leftarrow \mathbf{H}_1$ 
19. end while
20.  $\mathbf{H}^* \leftarrow \mathbf{H}_0[0]$ 
21. return  $\mathbf{H}^*$ 
```

5 实验验证

本文在 3 个真实的数据集上验证了所提出的 IMHE 方法的性能。将 IMHE 和其他最新的对比方法学习的节点嵌入用于 3 个数据挖掘任务,包括节点分类、节点聚类 and 链接预测。此外,实验还分析了所提方法的参数对嵌入性能的影响。

5.1 数据集

为了验证所提模型的性能,实验选取如下 3 个数据集。1) DBLP 数据集,是在线书目网络 DBLP 的子集,有 4 种类型的节点:作者(A)、论文(P)、主题(T)、场所(V)。网络中的边类型包括作者身份(P-A)、出版场所(P-V)、论文主题(P-T)和引用关系(P-P)。2) Yelp 数据集,是点评网站 Yelp 的在线子集,有 4 种类型的节点:用户(U)、商户(B)、城市(Ci)、称赞

(C)。网络中的边类型包括评估关系(U-B)、用户喜好(U-C)和营业地点(B-Ci)。3) Last. fm 数据集,是在线音乐网络 Last. fm 的子集。网络中有 3 种类型的节点:用户(U)、艺术家(Ar)、标签(Ta)。网络中的边类型包括“聆听关系”(U-Ar)和“艺术家的标签”(Ar-Ta)。为了在确保数据有效性的同时提高实验效率,实验随机采样了 30% 的原始网络,并过滤了度较小的节点(设置在 DBLP, Yelp 和 Last. fm 中的节点度小于 2, 3, 2)。表 1 列出了数据集的统计信息。由于 Last. fm 数据集没有标签信息,因此实验仅将其用于链接预测任务。

表 1 3 个 HIN 数据集的基本统计数据
Table 1 Basic statistics of three HIN datasets

Datasets	V	E	# labels	views
DBLP	10 240	75 374	4	APA, APTA, APVPA
Yelp	12 800	130 838	8	BUB, BCiB, BUCUB
Last. fm	7 466	91 236	/	ArUAR, ArTaAr

5.2 对比方法

为了验证 IMHE 的性能,本文将 IMHE 与多种类型的网络嵌入方法进行了比较,包括同质信息网络嵌入方法、异质信息网络嵌入方法和多视图网络嵌入方法。1) 同质信息网络嵌入方法。Node2vec^[27] 通过深度优先和广度优先的随机游走生成多个节点序列,并在 NLP 中使用 word2vec 算法学习节点嵌入。Line^[28] 通过保留节点之间的一阶和二阶邻近度来学习节点嵌入。2) 异质信息网络嵌入方法。Metapath2vec^[10] 基于一个给定的元路径生成由元路径指导的随机游走,然后将其输入到 SkipGram 模型中学习节点嵌入。Esim^[12] 通过权重组合来捕获多个元路径的语义信息。Hin2vec^[14] 是一个神经网络模型,它同时学习网络中节点的嵌入和关系的嵌入。3) 多视图嵌入方法。MNE^[29] 将高维嵌入和低维节点附加嵌入相结合,将多个关系信息表示到一个统一的嵌入空间中。

对比方法的参数设置为在其原始论文中的默认值。对于 Metapath2vec 方法,我们对每个元路径进行实验,并选择性能最佳的元路径。对于每个数据集,实验中在 IMHE, Metapath2vec 和 Esim 中设置了相同的元路径,如表 1 所列。本实验设置注意力头的数量为 5,元路径实例的长度为 20,节点的嵌入维数为 64。

5.3 节点分类

对于节点分类任务,实验将 IMHE 和其他对比方法学习到的嵌入用作节点特征,并将节点分类为不同类别。目标分类节点根据不同数据集设置为带有标签的节点(DBLP 中的作者节点和 Yelp 中的商户节点)。随机选择一组节点作为训练节点,其余节点用于测试。根据训练节点的嵌入,对逻辑回归分类器进行训练,以预测测试节点最可能的标签。每个网络数据集都分为训练集和测试集,训练集的比例从 10% 到 50% 依次增加。进行 10 次重复实验,平均 Micro-F1 和 Macro-F1 得分如图 5 所示。此外,为了验证本文方法针对不完整的多视图聚合的有效性,我们直接对多个单视图学习的节点嵌入进行平均,将结果作为不考虑视图不完整性的对比方法(AHE),并在节点分类任务中将其与 IMHE 进行比较。IMHE 和 AHE 方法的 Micro-F1 得分和 Macro-F1 得分如图 6 所示。

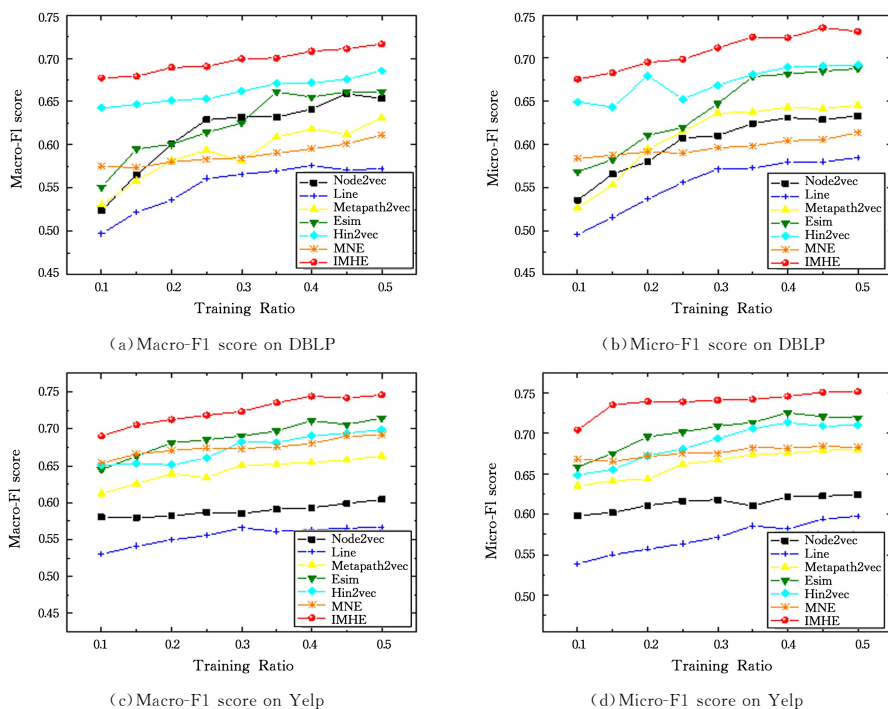


图 5 不同方法的节点分类性能

Fig. 5 Node classification performance of different methods

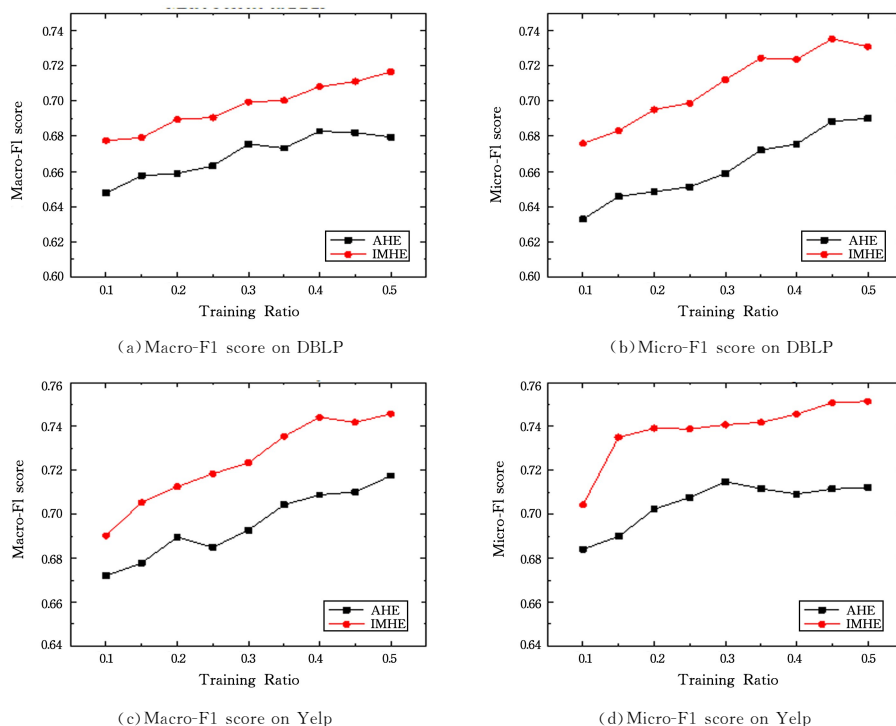


图 6 IMHE 和 AHE 的节点分类性能比较

Fig. 6 Comparison of node classification performance between IMHE and AHE

图 5 给出了本文提出的 IMHE 模型的节点分类性能。结果表明:1) IMHE 的节点分类性能优于对比方法。在所有 3 个 HIN 数据集上,IMHE 的 Micro-F1 和 Macro-F1 得分均高于对比方法。2) 在大多数情况下,异质方法优于同类方法。这是因为同质方法难以处理相同两个节点之间的不同语义关系。3) 在较低的训练集比例下,多视图方法优于单视图方法。例如,当训练比例为 0.1 时,多视图方法 IMHE 和

MNE 的 Micro-F1 得分和 Macro-F1 得分高于其他比较方法。但是,随着训练集比例变大,MNE 方法的性能落后于 Esim 和 Hin2vec 方法。这是因为 MNE 方法没有学习到 HIN 的异质语义信息,并且不能区分 HIN 中的复杂语义关系。

如图 6 所示,在所有情况下,IMHE 方法的节点分类性能均优于 AHE 方法。与 AHE 方法相比,通过 IMHE 方法获得的 Micro-F1 和 Macro-F1 分数在 DBLP 数据集上分别高出

约 4% 和 6%，在 Yelp 数据集上分别高出约 3% 和 4%。这是因为 IMHE 通过补全多个不完整视图以保存 HIN 的隐藏语义关系，并通过融合不完整视图解决了数据稀疏问题。本文的 IMHE 方法可以更有效地捕获 HIN 中的结构和语义信息，并提高 HIN 嵌入的性能。

5.4 节点聚类

聚类任务也用于验证 DBLP 和 Yelp 数据集上不同嵌入方法的性能。为了进行评估，将学习到的每种方法的嵌入信息输入到聚类模型中，并使用 K -Means^[30] 算法对其进行聚类。与节点分类任务相同，实验关注标记的节点 (DBLP 中的作者和 Yelp 中的业务)。10 次重复实验的标准化互信息 (NMI) 和调整兰德指数 (ARI) 的得分如图 7 所示。

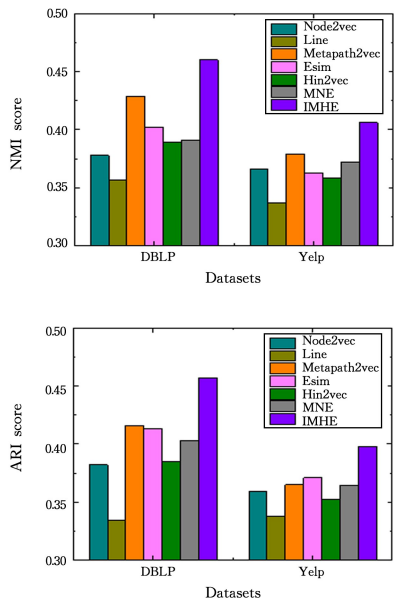


图 7 不同方法的节点聚类性能

Fig. 7 Node clustering performance of different methods

图 7 中，在节点聚类任务中，IMHE 优于所有对比方法。与最佳对比方法相比，通过 IMHE 方法获得的 NMI 和 ARI 分数在 DBLP 数据集上约提高 3%，并且在 Yelp 数据集上提高了 4%。与节点分类任务相似，同质网络嵌入方法在所有数据集上的执行效果均较差。与节点分类任务不同，Metapath2vec 和 Esim 在两个数据集上的所有对比中表现最佳，但 Hin2vec 在所有数据集上的表现不佳。可能的原因是，手动设置的元路径可以比基于关系的随机游走更多地捕获 HIN 中的重要语义信息。通过手动设置元路径学习的节点嵌入更适合于节点聚类任务。

5.5 链接预测

链接预测旨在预测网络中缺失的链接。实验使用 3 个数据集来比较 IMHE 和对比方法学习的节点嵌入的链接预测性能。在此任务中，我们预测 DBLP 中的作者-论文 (A-P) 链接、Yelp 中的用户-商户 (U-B) 链接和 Last.fm 的用户-艺术家 (U-Ar) 链接。首先将数据集的边随机分为训练集和测试集，将 75% 的边设置为训练集，将其余边设置为测试集。然后，将节点嵌入用作分类器的特征，以训练训练集并验证测试集上的链接预测性能。AUC 和 AP 的得分用作反

映节点嵌入性能的指标。

表 2 列出了 3 个数据集上不同方法的链接预测性能。与节点分类任务类似，IMHE 方法在 3 个数据集上的性能明显优于所有对比方法。链接预测任务的结果表明，IMHE 方法可以为异质信息网络保留更丰富的语义信息。与节点分类任务不同，Hin2vec 和 Metapath2vec 在对比方法中表现最佳，但在所有异质网络嵌入方法中，Metapath2vec 的链接预测性能最差。这可能是由于异质信息网络中存在多个有意义的元路径，并且在 Metapath2vec 模型学习的节点嵌入中缺少了一部分结构和语义信息。总的来说，本文提出的 IMHE 模型在所有实验任务中始终可以实现稳定的性能。

表 2 链接预测性能评估

Table 2 Performance evaluation of link prediction

Methods	DBLP		Yelp		Last. fm	
	AUC	AP	AUC	AP	AUC	AP
Node2vec	0.7992	0.8053	0.7025	0.6601	0.8111	0.8094
Line	0.7209	0.7147	0.5831	0.5807	0.7553	0.7389
Metapath2vec	0.8054	0.8216	0.7879	0.7632	0.8222	0.8206
Esim	0.8328	0.8305	0.8060	0.8730	0.8451	0.8333
Hin2vec	0.8574	0.8644	0.8502	0.8408	0.8517	0.8673
MNE	0.8221	0.8337	0.7514	0.7221	0.8399	0.8452
IMHE	0.8890	0.8981	0.8322	0.8393	0.8978	0.9017

5.6 参数分析

IMHE 中的参数设置会影响嵌入性能。在实验过程中，我们发现两个参数对所提方法的性能有很大影响：嵌入维数 (d) 和多头注意力模型中的头数 (k)。本文在节点聚类任务中对 DBLP 和 Yelp 数据集使用不同的 d 和 k 进行了实验。实验设置 d 的取值为 16~256， k 为 1~8。参数分析结果如图 8 和图 9 所示。

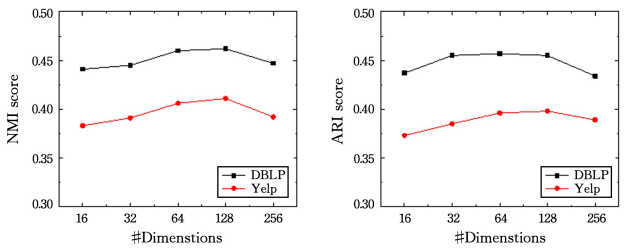


图 8 嵌入维数对网络嵌入性能的影响

Fig. 8 Effect of embedding dimension on network embedding performance

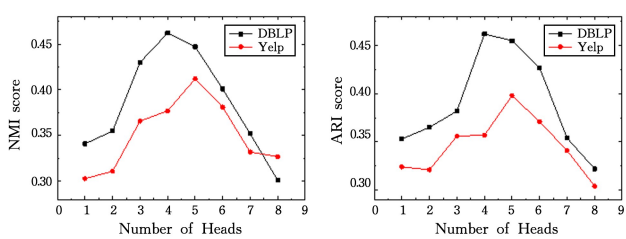


图 9 注意力头数对网络嵌入性能的影响

Fig. 9 Effect of the number of attention heads on network embedding performance

如图 8 所示，合适的嵌入维数为 64 或 128。当 Yelp 数据集上的 d 设置为 128，而 DBLP 上的 d 设置为 64 和 128 时，NMI 和 ARI 的得分最高。聚类性能随嵌入维数的增加而增

加。但是,当嵌入维数超过最佳 d 值时,聚类性能会随着嵌入维数的增加而降低。 d 太小无法充分获取语义,而 d 太大则可能导致过度拟合的问题。考虑到后续应用的性能和效率,建议将嵌入维度设置为 64。

从图 9 可以看出,对于 DBLP 和 Yelp 数据集,当将注意力头数分别设置为 4 和 5 时,聚类性能最佳。聚类性能随着注意力头数量的增加而提高。但是,当超出合理数量的注意力头时,聚类性能会随着头数量的增加而降低。一个可能的原因是网络中的节点之间存在多个隐藏的语义关系,但是过多的注意力头可能会导致拟合不足。因此,本文建议根据不同数据集中语义空间的数量来设置注意力头的数量。

结束语 本文提出了一种用于融合不完整多视图的异质信息网络嵌入模型。所提方法融合了 HIN 隐藏空间的多个视图,并保留了节点的各种语义关系。为了学习节点之间的权重并引入随机游走序列中节点的顺序信息,将多头注意力模型引入到单视图 HIN 嵌入中。对于每个不完整的视图,提出了一种邻居聚合方法,聚合缺失节点的 k 阶邻居信息,以生成该节点的新嵌入。本文基于 CCA 方法提出了 Multi-view CCA 方法来融合多个单视图嵌入,该方法可以融合两个以上的视图来生成一个多视图统一嵌入。所提方法在节点分类、节点聚类 and 链接预测任务上的性能优于现有的 HIN 嵌入模型,尤其是在训练样本数量非常有限的情况下。为了改进所提方法,未来的工作将扩展所提方法以解决更复杂的网络上的节点表示问题,例如属性异质网络、符号异质网络和动态异质网络。

参 考 文 献

- [1] DONG Y, HU Z, WANG K, et al. Heterogeneous network representation learning[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence. 2020:4861-4867.
- [2] WANG X, BO D, SHI C, et al. A Survey on Heterogeneous Graph Embedding: Methods, Techniques, Applications and Sources[J]. arXiv:2011.14867, 2020.
- [3] DENG H, HAN J, ZHAO B, et al. Probabilistic topic models with biased propagation on heterogeneous information networks [C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011:1271-1279.
- [4] LI Z, JIANG J Y, SUN Y, et al. Personalized question routing via heterogeneous network embedding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019,33:192-199.
- [5] HU X T, SHA C F, LIU Y J. Post-processing Network Embedding Algorithm with Random Projection and Principal Component Analysis[J]. Computer Science, 2021,48(5):124-129.
- [6] ZHAO K, BAI T, WU B, et al. Deep Adversarial Completion for Sparse Heterogeneous Information Network Embedding[C]//Proceedings of The Web Conference 2020. 2020:508-518.
- [7] YANG D J, WANG S Z, LI C Z, et al. From Properties to Links: Deep Network Embedding on Incomplete Graphs[C]//CIKM. 2017:367-376.
- [8] LIN Y, GOU Y, LIU Z, et al. COMPLETER: Incomplete multi-view clustering via contrastive prediction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:11174-11183.
- [9] HE Y, SONG Y, LI J, et al. HeteSpaceyWalk: a heterogeneous spacey random walk for heterogeneous information network embedding[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019:639-648.
- [10] DONG Y, CHAWLA N V, SWAMI A. metapath2vec: Scalable representation learning for heterogeneous networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017:135-144.
- [11] HOSSEINI A, CHEN T, WU W, et al. Heteromed: Heterogeneous information network for medical diagnosis[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018:763-772.
- [12] HU B, SHI C, ZHAO W X, et al. Leveraging meta-path based context for top-n recommendation with a neural co-attention model[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018:1531-1540.
- [13] FU X, ZHANG J, MENG Z, et al. MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding[C]//Proceedings of The Web Conference 2020. 2020:2331-2341.
- [14] FU T, LEE W C, LEI Z. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017:1797-1806.
- [15] SHI R, LIANG T, PENG H, et al. HEAM: Heterogeneous Network Embedding with Automatic Meta-path Construction[C]//International Conference on Knowledge Science, Engineering and Management. Springer, Cham, 2020:304-315.
- [16] TANG J, QU M, MEI Q. Pte: Predictive text embedding through large-scale heterogeneous text networks[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015:1165-1174.
- [17] XU L, WEI X, CAO J, et al. Embedding of embedding (EOE) joint embedding for coupled heterogeneous networks[C]//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 2017:741-749.
- [18] SHAPIRA T, SHAVITT Y. A Deep Learning Approach for IP Hijack Detection Based on ASN Embedding[C]//Proceedings of the Workshop on Network Meets AI & ML. 2020:35-41.
- [19] WANG X, JI H, SHI C, et al. Heterogeneous graph attention network[C]//The World Wide Web Conference. 2019:2022-2032.
- [20] WANG L, GAO C, HUANG C, et al. Embedding heterogeneous networks into hyperbolic space without meta-path[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021.
- [21] JIANG J Y, LI Z, JU C J T, et al. MARU: Meta-context Aware Random Walks for Heterogeneous Network Representation Learning[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020:575-584.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv:1706.03762, 2017.

[23] YU W J, DING S F. Conditional Generative Adversarial Network Based on Self-attention Mechanism[J]. Computer Science, 2021, 48(1):241-246.

[24] VOITA E, TALBOT D, MOISEEV F, et al. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned[J]. arXiv:1905.09418, 2019.

[25] BANSAL T, JUAN D C, RAVI S, et al. A2N: attending to neighbors for knowledge graph inference[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:4387-4392.

[26] SUN Z, SARMA P, SETHARES W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:8992-8999.

[27] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:855-864.

[28] TANG J, QU M, WANG M, et al. Line: Large-scale information

network embedding[C]// Proceedings of the 24th International Conference on World Wide Web. 2015:1067-1077.

[29] ZHANG H, QIU L, YI L, et al. Scalable Multiplex Network Embedding[C]// IJCAI. 2018, 18:3082-3088.

[30] YUAN C, YANG H. Research on K-value selection method of K-means clustering algorithm[J]. J-Multidisciplinary Scientific Journal, 2019, 2(2):226-235.



ZHENG Su-su, born in 1993, postgraduate. Her main research interests include network representation, data mining and complex network analysis.



YUAN Wei-wei, born in 1981, Ph.D, associate professor. Her main research interests include machine learning, pattern recognition, social computing and recommender systems.