第 2 章

图表征学习

Peng Cui、Lingfei Wu、Jian Pei、Liang Zhao 和 Xiao Wang[①]

摘要

图表征学习（也称图表示学习）的目的是将图中的节点嵌入低维的表征并有效地保留图的结构信息。最近，人们在这一新兴的图分析范式方面已经取得大量的成果。在本章中，我们将首先总结图表征学习的动机。接下来，我们将系统并全面地介绍大量的图嵌入方法，包括传统图嵌入方法、现代图嵌入方法和图神经网络。

2.1 导读

许多复杂的系统具有图的形式，如社交网络、生物网络和信息网络。众所周知，由于图数据往往是复杂的，因此处理起来极具挑战性。为了有效地处理图数据，第一个关键的挑战是找到有效的图数据表征方法，也就是如何简洁地表征图，以便在时间和空间上有效地进行高级的分析任务，如模式识别、分析和预测。传统上，我们通常将一个图表征为 $\mathcal{G}=(\mathcal{V},\mathcal{E})$，其中，$\mathcal{V}$ 是一个节点集合，$\mathcal{E}$ 是一个边集合。对于大型图来说，比如那些有数十亿个节点的图，传统的图表征在图的处理和分析上面临着一些挑战。

（1）高计算复杂性。这些由边集合 $\mathcal{E}$ 编码的关系使得大多数的图处理或分析算法采用了一些迭代或组合的计算步骤。例如，一种流行的方法是使用两个节点之间的最短或平均路径长度来表示它们的距离。为了用传统图表征计算这样的距离，我们必须列举两个节点之间许多可能的路径，这在本质上是一个组合的问题。由于这种方法会导致高计算复杂性，因此不适用于现实世界的大规模图。

（2）低可并行性。并行和分布式计算是处理和分析大规模数据的事实上的方法。然而，以传统方式表征的图数据给并行和分布式算法的设计与实现带来了严重困难。瓶颈在于，图中节点之间的耦合是由 $\mathcal{E}$ 显式反映的。因此，将不同的节点分布在不同的分片或服务器

① Peng Cui
Department of Computer Science，Tsinghua University，E-mail：cuip@tsinghua.edu.cn
Lingfei Wu
Pinterest，E-mail：lwu@email.wm.edu
Jian Pei
Department of Computer Science，Simon Fraser University，E-mail：jpei@cs.sfu.ca
Liang Zhao
Department of Computer Science，Emory University，E-mail：liang.zhao@emory.edu
Xiao Wang
Department of Computer Science，Beijing University of Posts and Telecommunications，E-mail：xiaowang@bupt.edu.cn

上，往往会导致服务器之间的通信成本过高并降低加速率。

（3）机器学习方法的不适用性。最近，机器学习方法，特别是深度学习，在很多领域都发挥了强大的功能。然而，对于以传统方式表征的图数据，大多数现有的机器学习方法可能并不适用。这些方法通常假设数据样本可以用向量空间中的独立向量来表示，而图数据中的样本（即节点）在某种程度上是相互依赖的，由 $\mathscr{E}$ 中的边相互连接在一起。虽然我们可以简单地用图的邻接矩阵中相应的行向量来表示一个节点，但在一个有许多节点的大图中，这种表征的维度非常高，会增加后续图处理和分析的难度。

为了应对这些挑战，人们致力于开发新的图表征学习，如针对节点学习密集和连续的低维向量表征，这样可以减少噪声或冗余信息，并保留内在的结构信息。节点之间的关系原来是用图中的边或其他高阶拓扑度量来表征的，可由向量空间中节点之间的距离捕获，节点的结构特征则被编码到该节点的表征向量中。

基本上，为了使表征空间很好地支持图分析任务，图表征学习有两个目标。首先，原始图结构可以从学习到的表征向量中重建。具体原理是，如果两个节点之间有一条边或关系，那么这两个节点在表征空间中的距离应该相对较小。其次，学习到的表征空间可以有效地支持图推理，如预测未见的链接、识别重要的节点以及推断节点标签等。应该注意的是，仅以图重建为目标的图表征对图推理来说是不够的。在得到表征后，还需要根据这些表征来处理下游任务，如节点分类、节点聚类、图的可视化和链接预测。总的来说，图表征学习方法主要有三类——传统图嵌入方法、现代图嵌入方法和图神经网络。接下来我们将分别介绍它们。

2.2 传统图嵌入方法

传统图嵌入方法最初是作为降维技术进行研究的。图通常是从特征表示的数据集中构建出来的，如图像数据集。如前所述，图嵌入通常有两个目标——重建原始图结构和支持图推理。传统图嵌入方法的目标函数主要针对图的重建。

具体来说，首先，Tenenbaum et al（2000）使用 K 近邻（KNN）等连接算法构建了一个邻接图 $\mathscr{S}$。其次，基于 $\mathscr{S}$ 可以计算出不同数据之间的最短路径。因此，对于数据集中的 N 个数据条目，我们有一个图距离矩阵。最后，将经典多维尺度变换（Multi-Dimensional Scaling，MDS）应用于该矩阵，以获得坐标向量。我们通过 Isomap 学习的表征近似地保留了低维空间中节点间的地理距离。Isomap 的关键问题在于其高复杂性，因为需要计算成对的最短路径。随后，局部线性嵌入（Locally Linear Embedding，LLE）方法（Roweis and Saul，2000）被提出来，用于减少估计相距甚远的节点之间距离的需要。LLE 假设每个节点及其邻居节点都位于或接近一个局部的线性流体。为了描述局部几何特征，每个节点都可以通过其邻居节点来重建。最后，在低维空间中，LLE 在局部线性重建的基础上构造了一个邻域保留映射。拉普拉斯特征映射（Laplacian Eigenmap，LE）（Belkin and Niyogi，2002）也是首先通过 ε 邻域或 K 近邻构建一个图，然后利用热核（Berline et al，2003）来选择图中两个节点的权重，最后通过基于拉普拉斯矩阵的正则化得到节点表征。此外，人们还提出了局部保持投影（Locality Preserving Projection，LPP）（Berline et al，2003），这是一种针对非线性 LE 的线性近似算法。

在丰富的图嵌入文献中，根据构建的图的不同特征，这些方法得到了不同的扩展（Fu and Ma，2012）。我们发现，传统图嵌入方法大多适用于从特征表示的数据集中构建出来的图，

其中，由边权重编码的节点之间的接近度在原始特征空间中有很好的定义。与此形成对比的是，2.3 节将要介绍的现代图嵌入方法主要工作在自然形成的网络上，如社交网络、生物网络和电子商务网络。在这些网络中，节点之间的接近度并没有明确或直接的定义。例如，两个节点之间的边通常只是意味着它们之间存在某种关系，但无法表明具体的接近度。另外，即使两个节点之间没有边，我们也不能说这两个节点之间的接近度为零。节点接近度的定义取决于具体的分析任务和应用场景。因此，现代图嵌入通常包含丰富的信息，如网络结构、属性、侧面信息和高级信息，以促进解决不同的问题和应用。现代图嵌入方法需要同时针对前面提到的两个目标。鉴于此，传统图嵌入方法可以看作现代图嵌入方法的特例，而现代图嵌入的最新研究进展则更加关注网络推理。

2.3 现代图嵌入方法

为了更好地支持图推理，现代图嵌入学习考虑了图中更丰富的信息。根据图表征学习中所保留信息的类型，现代图嵌入方法可以分为三类：（1）保留图结构和属性的图表征学习；（2）带有侧面信息的图表征学习；（3）保留高级信息的图表征学习。在技术方面，不同的模型可以用来纳入不同类型的信息或针对不同的目标。常用的模型包括矩阵分解、随机行走、深度神经网络及其变体等。

2.3.1 保留图结构和属性的图表征学习

在图中编码的所有信息中，图的结构和属性是在很大程度上影响图推理的两个关键因素。因此，图表征学习的一个基本要求就是适当地保留图的结构并捕捉图的属性。通常，图结构包括一阶结构和高阶结构（如二阶结构和群落结构）。不同类型的图有不同的属性。例如，有向图具有非对称传递性。结构平衡理论常见于符号图的处理中。

2.3.1.1 保留图结构的图表征学习

图的结构可以分为不同的类别，而且不同类别拥有不同粒度的图表征。在图表征学习中，经常用到的图结构是邻域结构、高阶接近度和群落结构。

如何定义图中的邻域结构是第一个挑战。基于短时随机行走中出现的节点分布与自然语言中单词分布相似的发现，DeepWalk（Perozzi et al，2014）采用了随机行走来捕捉邻域结构，然后对于随机行走产生的每个行走序列，按照 Skip-Gram 模型，最大化行走序列中邻居节点出现的概率。node2vec 定义了一个灵活的节点图邻域概念，并设计了一种二阶随机行走策略来对邻域节点进行抽样，从而在广度优先抽样（Breadth-First Sampling，BFS）和深度优先抽样（Depth-First Sampling，DFS）之间平稳插值。除邻域结构以外，LINE（Tang et al，2015b）被提出用于大规模的网络嵌入，LINE 可以保留一阶接近度和二阶接近度。一阶接近度指的是观察到的两个节点之间成对节点的接近度。二阶接近度是由两个节点的“环境”（邻居节点）的相似性决定的。在衡量两个节点之间的关系方面，它们两者都很重要。从本质上说，由于 LINE 是基于浅层模型的，因此其表现能力有限。SDNE（Wang et al，2016）是一个用于网络嵌入的深度模型，其目的也是捕捉一阶接近度和二阶接近度。SDNE 使用具有多个非线性层的深度自编码器架构来保留二阶接近度。为了保留一阶接近度，SDNE 采用了拉普拉斯特征映射的思想（Belkin and Niyogi，2002）。Wang et al（2017g）提出了一个用于图表征学习的模

块化非负矩阵因子化（M-NMF）模型，旨在同时保留微观结构（即节点的一阶接近度和二阶接近度）以及中观群落结构（Girvan and Newman，2002）。他们首先采用 NMF 模型（Févotte and Idier，2011）来保留微观结构，同时通过模块化来最大化检测群落结构（Newman，2006a）。然后，他们引入了一个辅助的群落表征矩阵来连接节点的表征和群落结构。通过这种方式，学习到的节点表征将同时受到微观结构和群落结构的制约。

总之，许多网络嵌入方法的目的是在潜在的低维空间中保留节点的局部结构，包括邻域结构、高阶接近度以及群落结构。通过在线性和非线性模型中进行尝试，深度模型在网络嵌入方面具有巨大潜力。

2.3.1.2 保留图属性的图表征学习

目前，现有的保留属性的图表征学习方法大多数侧重于保留所有类型图的传递性以及有符号图的结构平衡性。

图常常存在传递性，同时我们也发现，保留这样的属性并不难。这是因为在度量空间中，不同数据之间的距离天然地满足三角形不等式。然而，这在现实世界中并不总是对的。Ou et al（2015）想要通过潜在的相似性组件来保留图的非传递属性。非传递属性的内容是，对于图中的节点 v_1、v_2 和 v_3，其中的（v_1; v_2）和（v_2; v_3）是相似对，但（v_1; v_3）可能是一个不相似对。例如，在社交网络中，一名学生可能与家人和同学有紧密联系，但这名学生的同学和家人可能彼此并不熟悉。上述方法的主要思想是，首先学习多个节点的嵌入表征，然后根据多个相似性而不是一个相似性来比较不同的节点接近度。通过观察可以发现，如果两个节点有很大的语义相似性，那么它们至少有一种嵌入表征的相似性很大，否则所有表征的相似性都很小。有向图通常具有非对称传递性。非对称传递性表明，如果有一条从节点 i 到节点 j 的有向边以及一条从节点 j 到节点 v 的有向边，则很可能存在一条从节点 i 到节点 v 的有向边，但不存在从节点 v 到节点 i 的有向边。为了测量这种高阶接近度，HOPE（Ou et al，2016）总结了 4 种测量方法，然后利用广义 SVD 问题对高阶接近度进行了因子化（Paige and Saunders，1981），这样 HOPE 的时间复杂度便大大降低了，这意味着 HOPE 对于大规模的网络是可扩展的。在一个既有正边又有负边的符号图中，社交理论〔如结构平衡理论（Cartwright and Harary，1956；Cygan et al，2012）〕与在无符号图中的区别非常大。结构平衡理论表明，在有签名的社交网络中，用户应该能够让他们的“朋友”比他们的“敌人”更亲密。为了给结构平衡现象建模，SiNE（Wang et al，2017f）提出了由两个具有非线性函数的深度图组成的深度学习模型。

人们已充分认识到在网络嵌入空间中保持图属性的重要性，特别是那些在很大程度上影响网络演化和形成的属性。关键的挑战是如何解决原始网络空间和嵌入矢量空间在属性层面的差异和不均匀性。一般来说，大多数结构和属性保护方法都考虑了节点的高阶接近度，这表明了在图嵌入中预先服务高阶接近度结构的重要性，区别在于获得高阶接近度结构的策略。一些方法通过假设从一个节点到其邻居节点的生成机制来隐含地保留高阶接近度结构，而另一些方法则通过在嵌入空间中明确地逼近高阶接近度来实现。由于拓扑结构是图数据最明显的特征，因此很大一部分文献介绍了保留拓扑结构的方法。相对而言，可以保留属性的图嵌入方法是一个相对较新的研究课题，目前只有比较浅显的研究。图属性由于通常驱动着图的形成和演化，因此它们在未来的研究和应用中具有巨大的潜力。

2.3.2 带有侧面信息的图表征学习

除图结构以外，侧面信息是图表征学习的另一个重要信息源。在图表征学习中，侧面信息可以分为两类——节点内容以及节点和边的类型，它们的区别在于整合网络结构和侧面信息的方式。

带有节点内容的图表征学习。在某些类型的图（如信息网络）中，节点伴随着丰富的信息，如节点标签、属性甚至语义描述。如何在图表征学习中把它们与网络拓扑结构结合起来？这引发了人们相当大的研究兴趣。Tu et al（2016）通过利用节点的标签信息，提出了一种半监督的图嵌入算法——MMDW。MMDW 同样基于 DeepWalk 衍生的矩阵分解，采用支持向量机（Support Vector Machine，SVM）（Hearst et al，1998）并结合标签信息来找到最佳分类边界。Yang et al（2015b）提出了 TADW——TADW 在学习节点的低维表征时会考虑与节点相关的丰富信息（如文本）。Pan et al（2016）提出了一个耦合的深度模型，旨在将图结构、节点属性和节点标签纳入图嵌入方法。虽然不同的方法采用不同的策略来整合节点内容和网络拓扑结构，但它们都认为节点内容提供了额外的接近度信息来约束节点的表征。

异质图表征学习。与带有节点内容的图不同，异质图由不同类型的节点和边组成。如何在图嵌入方法中统一异质类型的节点和边？这也是一个有趣但具有挑战性的问题。Jacob et al（2014）提出了一种用于分类节点的异质社交图表征学习算法，该算法将在一个共同的向量空间中学习所有类型节点的表征，并在这个空间中进行推理。Chang et al（2015）提出了一种针对异质图（其中的节点可以是图像、文本等类型）的深度图表征学习算法，图像和文本的非线性嵌入方法可以分别由 CNN 模型和全连接层学习到。Huang and Mamoulis（2017）提出了一种保留元路径相似性的异质信息图表征学习算法。为了对一个特定的关系进行建模，元路径（Sun et al，2011）需要是一个带有边类型的对象类型的序列。

在保留侧面信息的方法中，侧面信息引入了附加的接近度度量，这样可以更全面地学习节点之间的关系。这些方法的区别在于整合网络结构和侧面信息的方式，它们中的许多是由保留图结构的网络嵌入方法自然延伸出来的。

2.3.3 保留高级信息的图表征学习

与侧面信息不同，高级信息是指特定任务中的监督或伪监督信息。保留高级信息的网络嵌入通常包括两部分：一部分是保留网络结构，以便学习节点表征；另一部分是建立节点表征和目标任务之间的联系。高级信息和网络嵌入技术的结合使得网络的表征学习成为可能。

信息扩散。信息扩散（Guille et al，2013）是网络上无处不在的现象，尤其是在社交网络中。Bourigault et al（2014）提出了一种用于预测社交网络中信息扩散的图表征学习算法。该算法的目标是学习潜在空间中的节点表征，使得扩散核能够更好地解释训练集中的级联。该算法的基本思想是将观察到的信息扩散过程映射为连续空间中的扩散核所模拟的热扩散过程。扩散核的扩散原理是，潜在空间中的一个节点离源节点越近，这个节点就会越早被源节点的信息感染。这里的级联预测问题被定义为预测给定时间间隔后的级联规模增量（Li et al，2017a）。Li et al（2017a）认为，关于级联预测的前期工作依赖手动制作的特征袋来表征级联和图结构。作为替代，他们提出了一个端到端的深度学习模型，旨在利用图嵌入方

法的思想来解决这个问题。整个过程能够以端到端的方式学习级联图的表征。

异常检测。异常检测在以前的工作中得到了广泛研究（Akoglu et al，2015）。图中的异常检测旨在推断结构上的不一致，也就是检测连接到各种具有影响力群落的异常节点（Hu et al，2016；Burt，2004）。Hu et al（2016）提出了一种基于图嵌入的异常检测方法，他们假设两个链接节点的群落成员身份应该是相似的。异常节点是指连接到一组不同群落的节点。由于学习到的节点嵌入方法捕捉了节点和群落之间的关联性，基于该节点嵌入方法，他们提出了一个新的度量来表明节点的异常程度。度量值越大，节点成为异常节点的概率就越高。

图对齐。图对齐的目标是建立两个图中节点之间的对应关系，即预测两个图之间的锚链接。不同社交网络共享的相同用户自然形成了锚链接，这些锚链接是不同图之间的桥梁。锚链接预测的问题可以定义为给定源图和目标图以及一组观察到的锚链接，识别两个图中的隐藏锚链接。Man et al（2016）提出了一种图表征学习算法来解决这个问题。学习到的表征可以保留图的结构并重视观察到的锚链接。

保留高级信息的图嵌入通常包括两部分：一部分是保留图的结构，以便学习节点表征；另一部分是建立节点表征和目标任务之间的联系。前者类似于保留结构和属性的网络嵌入，后者则通常需要考虑特定任务的领域知识。对领域知识这种高级信息的编码使得开发图应用的端到端模型成为可能。与手动提取的网络特征（如众多的图中心度量）相比，高级信息和图嵌入技术的结合使图的表征学习成为可能。许多图应用可以从这种新模式中获益。

2.4 图神经网络

在过去的 10 年中，深度学习已经成为人工智能和机器学习的“皇冠上的明珠”，在声学、图像和自然语言处理等方面具有卓越的表现。尽管众所周知，图在现实世界中无处不在，但利用深度学习方法来分析图数据仍非常具有挑战性。具体表现在：（1）图的不规则结构。与图像、音频、文本有明确的网格结构不同，图有不规则的结构，这使得一些基本的数学运算很难推广到图上。例如，为图数据定义卷积和池化操作（这是卷积神经网络中的基本操作）并不简单。（2）图的异质性和多样性。图本身可能很复杂，包含不同的类型和属性。针对这些不同的类型、属性和任务，解决具体问题时需要利用不同的模型结构。（3）大规模图。在大数据时代，现实中的图可以很容易拥有数量达到数百万或数十亿的节点和边。如何设计可扩展的模型（最好的情况是模型的时间复杂度相对于图的大小具有线性关系）是一个关键问题。（4）纳入跨学科知识。图经常与其他学科相联系，如生物学、化学和社会科学等。这种跨学科的性质使得机会和挑战并存：领域知识可以用来解决特定的问题，但整合领域知识也会使得模型设计更为复杂。

图神经网络在过去几年中得到了大量的研究与关注，所采用的架构和训练策略千差万别，从监督到非监督，从卷积到循环，包括图循环神经网络（Graph RNN）、图卷积网络（GCN）、图自编码器（GAE）、图强化学习（Graph RL）和图对抗方法等。具体来说，Graroperty h RNN 通过在节点级或图级进行状态建模来捕捉图的循环和顺序模式；GCN 则在不规则的图结构上定义卷积和读取（readout）操作，以捕捉常见的局部和全局结构模式；GAE 假设低秩图结构并采用无监督的方法进行节点表征学习；图强化学习定义了基于图的动作和奖励，以便在遵循约束条件的同时获得图任务的反馈；图对抗方法采用对抗训练技术来提高

图模型的泛化能力，并通过对抗攻击测试其鲁棒性。

另外，许多正在进行的或未来的研究方向也值得进一步关注，包括针对未研究过图结构的新模型、现有模型的组合性、动态图、可解释性和鲁棒性等。总的来说，图深度学习是一个很有前途且快速发展的研究领域，它既提供了令人兴奋的机会，也带来了许多挑战。对图深度学习进行研究是关系数据建模的一个关键构件，也是迈向未来更好的机器学习和人工智能技术的重要一步。

2.5 小结

在本章中，我们首先介绍了图表征学习的动机。其次，在 2.2 节中讨论了传统图嵌入方法，并在 2.3 节中介绍了现代图嵌入方法。基本上，保留结构和属性的图表征学习是基础。如果不能很好地保留图结构并在表征空间中保留重要的图属性，就会存在严重的信息损失并损害下游的分析任务。基于保留结构和属性的图表征学习，人们可以应用现成的机器学习方法。如果有一些额外信息，那么可以将它们纳入图表征学习。此外，可以考虑将一些特定应用的领域知识作为高级信息。