Graph Embedding Techniques Summary

Graph Embedding目前有五种主流的技术，包括矩阵分解技术，深度学习技术，边重构技术，图内核技术，以及generative model技术。

1. 矩阵分解技术

在graph embedding任务中，存在着（1）拉普拉斯特征图分解和（2）结点邻接矩阵分解两种。其**效果取决于结点间距离的计算方法以及采用何种目标函数**。

文章[85]中指出了graph embedding的总体框架，包括linearization，kernelization,tensorzation三个主要步骤，解释了一系列的降维算法，表明结点的向量表示，可以从有约束图的拉普拉斯矩阵的特征值所对应的特征向量中获取。

计算距离方面，文章【1】采用了传统的欧氏距离进行计算。【2】优化了之前由迪杰斯特拉算法计算的测地距离，用在高维空间中的欧氏距离衡量点之间的连接性，并且指出当点的邻近结点非常多或者数据存在噪声时，利用原有的方式会使得embedding结果出现很大偏差。文章【3】通过寻找Laplace Beltrami operator的最优线性逼近特征函数，获得LPP（Locality Preserving Projections）。LPP可以用于数据降维，以及寻找一个符合目前图结构的投影。进一步，文章【4】介绍了一种锚图（anchor graph），显著提升了LPP的效率。

目标函数的选择方面，文章【5】介绍了SR（Spectral Regression）模型，通过构造一种包含了带标签与不标签的knn图，采用半监督学习方法，其目标函数是LDA和LPP目标函数的线性拓展。前者关注于处理带标签集合，使得不同标签的点尽可能分离；后者关注全图数据的几何结构性质。另有文章【6】证明了，当图具有对称性质时，SR和LPP方法时等价的。文章【7】中利用聚类操作配合降维进行graph embedding，其目标函数组件包括PCA，K-means以及图的拉普拉斯正则化。另有文章认为，结点对间的关系不能直接用于图的构建，图的生成需要经过学习。如【8】中通过学习得到一个矩阵，将图中不相连的结点的距离最大化。

针对结点邻接矩阵的分解，目标函数是明确的：使得对于结点距离估计损失最小化。目标函数的解有【9】中提出的正则化高斯矩阵分解（regularized Gaussian），【10】中的添加正则项以便加强图上的约束

1. Deep Learning技术

DL技术用于graph embedding，按照是否采用随机游走技术采样进行分类。

DeepWalk【11】首次引入深度学习技术，通过模拟流式的短距离随机游走，学习图中结点的social representation，可以捕获到邻居的相似性以及community membership;并且可以捕获到人类语言的语义和句法结构，进行逻辑类比等。随机游走的方式就是从一个随结点出发，随机选择它的一个邻居，以此类推，形成一个结点序列。当图发生变化，我们只需要更改一些random walks，重新表征这个图，设计出的算法应该是亚线性的。文章node2vec【12】提出了一种服从指定分布的的随机游走策略：给定出发点和游走长度，从当前点出发，下一节点的选择服从由转移概率和标准化参数构成的分布。进一步考虑了网络结构和搜索不同的邻域后，增加了两个参数对于随机游走的方式进行了约束：当我们选择出继续游走的下一节点后，计算在访问结点的可能性，尽量保证在两步内采样已访问过的节点的可能性比较低，避免拘于拘束。另一个参数用于表征我们的随机游走和BFS或者DFS的相似程度，但是可以保证时空复杂度优于前面二者

对于不采用random walk采样策略作为核心的DL技术，有相当一部分采取在同构图中对于大小可变的子图进行建模的方式。例如，文章【13】中提出的struct2vec，考虑到一些场景下，两个不是近邻的顶点也可能拥有很高的相似性，而诸如deepwalk，node2vec等方法难以捕捉到。主要思想是1.忽略节点和边缘属性以及它们在网络中的位置来评估节点之间的结构相似性；2.建立层次结构来度量节点在不同尺度上的结构相似性，允许对结构相似的含义有更加严格的概念；3.为结点生成随机上下文，这些结点是通过加权有偏随机游走遍历一个多层图观察到的结构相似的节点序列。进一步GraphSAGE【14】在原有的GCN【15】基础上，改变了GCN难以支持动态结点embedding计算的困难（因为需要重新训练），转而直接学习一种节点的表示方法。去学习一个节点的信息是怎么通过其邻居节点的特征聚合而来的。 学习到了这样的“聚合函数”，而本身就已知各个节点的特征和邻居关系，就可以很方便地得到一个新节点的表示了。

另一种方式是在异构图的不同类型的结点之间，采用interaction的学习方式，具有较高的可拓展性。文章【16】对知识图谱中的多元关系数据进行建模，在不引入额外知识的情况下，高效的实现知识补全，关系预测。模型包含数量较少的参数，并且可以扩展到非常大的知识库。核心思想是关系作为向量空间转变的桥梁：如果三元组(h,l,t)成立，则头实体embedding和关系embedding相加约等于尾实体的embedding。利用空间传递不变性，找到一个实体和向量空间，使得整关系三元组之间的势能差值最小。

1. Edge reconstruction

一个好的embedding应当能够在图中重构某些边，并且最大化的保留我们原有的能观测到的边。于是，现有的工作中引入了第一相似度（first-order proximity）和第二相似度(second-order proximity)，通过将这种相似性最大化来学习embedding。进而，针对每种相似性，设计了许多不同的目标函数。文章【17】采用一阶相似性，最大化点之间距离，但是并未将点划分为不同的种类。文章【18】将任务划分成事件event和对象object。其中事件利用超边表示，每条超边包含的点就是事件中涉及的对象，对象允许由多种不同的类型。进而，设计的目标函数中，对于单个事件：包括超边权重（衡量事件内容重要性），包含的对象子集，基于事件的约束以及对应的指标函数。当任务涵盖多个事件，将上述函数求和即可。文章【19】引入了三种图相关的内容去学习embedding，包括邻居内容，边内容，路径内容（利用到了随机游走），从不同方面表示了图相关的知识。最终的目标函数是上述三个方面的目标函数的加权和，然后设置学习策略学习这些参数。在学习方面，引入了注意力机制，学习不同顶点或边的代表能力。

文章【20】针对社交网络的图，关注用户和兴趣，物品的关系。目前该领域的研究都有着共同的前提，即：相似的用户有着相似的兴趣趋向。然而，该文指出，这个条件过强，因为社交网络不一定是根据相同的兴趣构建的。为了避免过度强调用户相似性，我们在一个结点和他的邻近结点间加入一个虚拟公共结点。在目标函数的计算上，将常见的两种相似加权求和，并用SGD学习参数。

也有研究将排序损失（ranking loss）和目标放在一起进行优化，以便表示更多的信息。文章【21】提出了在语义空间进行投影的模型（SSP）。在知识图谱原有的三元组表示上，同时利用结点的文本描述，分析语义相关性，提供了精确的语义性的embedding。通过这种方式，只要计算L2范数投影到语义空间后的损失（足够小），我们可以鉴别三元组的真实性（成立）。文章【22】对于实体和关系分别进行embedding，提出了解决跨语言问题的方法，使图的逻辑更加清晰合理。文章【23】提出利用推理的方法，将这种类比结构结合到目标函数中，改善多关系的嵌入问题。

参考文献：

[1].T. Hofmann and J. M. Buhmann, “Multidimensional scaling and data clustering,” in NIPS, 1994, pp. 459–466.

[2]. M. Balasubramanian and E. L. Schwartz, “The isomap algorithm and topological stability,” Science, vol. 295, no. 5552, pp. 7–7, 2002.

[3]. X. He and P. Niyogi, “Locality preserving projections,” in NIPS, 2003, pp. 153–160.

[4]. R. Jiang, W. Fu, L. Wen, S. Hao, and R. Hong, “Dimensionality reduction on anchorgraph with an efficient locality preserving projection,” Neurocomputing, vol. 187, pp. 109–118, 2016.

[5]. D. Cai, X. He, and J. Han, “Spectral regression: a unified subspace learning framework for content-based image retrieval,” in MM, 2007, pp. 403–412.

[6]. F. Nie, W. Zhu, and X. Li, “Unsupervised large graph embedding,” in AAAI, 2017, pp. 2422–2428.

[7]. K. Allab, L. Labiod, and M. Nadif, “A semi-nmf-pca unified framework for data clustering,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 1, pp. 2–16, 2017.

[8]. L. Vandenberghe and S. Boyd, “Semidefinite programming,” SIAM Rev., vol. 38, no. 1, pp. 49–95, 1996.

[9]. A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola, “Distributed large-scale natural graph factorization,” in WWW, 2013, pp. 37–48.

[10]. C. Tu, W. Zhang, Z. Liu, and M. Sun, “Max-margin deepwalk: Discriminative learning of network representation,” in IJCAI, 2016, pp. 3889–3895.

[11]. B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in KDD, 2014, pp. 701–710.

[12]. A. Grover and J. Leskovec, “Node2vec: Scalable feature learning for networks,” in KDD, 2016, pp. 855–864.

[13]. L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo, “Struc2vec: Learning node representations from structural identity,” in KDD, 2017, pp. 385–394.

[14]. W. L. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in NIPS, 2017, pp. 1025–1035.

[15]. T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in ICLR, 2017.

[16]. A. Bordes, N. Usunier, A. Garc´ıa-Dur´an, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multirelational data,” in NIPS, 2013, pp. 2787–2795.

[17]. T. Man, H. Shen, S. Liu, X. Jin, and X. Cheng, “Predict anchor links across social networks via an embedding approach,” in IJCAI, 2016, pp. 1823–1829.

[18]. H. Gui, J. Liu, F. Tao, M. Jiang, B. Norick, and J. Han, “Large-scale embedding learning in heterogeneous event data,” in ICDM, 2016, pp. 907–912.

[19]. J. Feng, M. Huang, Y. Yang, and X. Zhu, “GAKE: graph aware knowledge embedding,” in COLING, 2016, pp. 641–651.

[20]. Q. Zhang and H. Wang, “Not all links are created equal: An adaptive embedding approach for social personalized ranking,” in SIGIR, 2016, pp. 917–920.

[21]. H. Xiao, M. Huang, L. Meng, and X. Zhu, “SSP: semantic space projection for knowledge graph embedding with text descriptions,” in AAAI, 2017, pp. 3104–3110.

[22]. M. Chen, Y. Tian, M. Yang, and C. Zaniolo, “Multilingual knowledge graph embeddings for cross-lingual knowledge alignment,” in IJCAI, 2017, pp. 1511–1517. [134].