图上的机器学习系列- 聊聊MetaPath2vec

原创 AaronLou 享受编程的乐趣 2020-04-14

前言

本篇继续Graph Embedding这个话题。我们以前已经聊过DeepWalk、Node2Vec、LINE、GraphSAGE（详见公众号历史文章），它们都是面对的同质图问题，而今天将讨论的MetaPath2Vec则是要回答**如何对于异质图也能进行低维的空间向量表示。**

结合的论文为《metapath2vec: Scalable Representation Learning for Heterogeneous Networks》，作者提出了两个框架模型：metapath2vec、metapath2vec++。

算法原理消化

问题定义

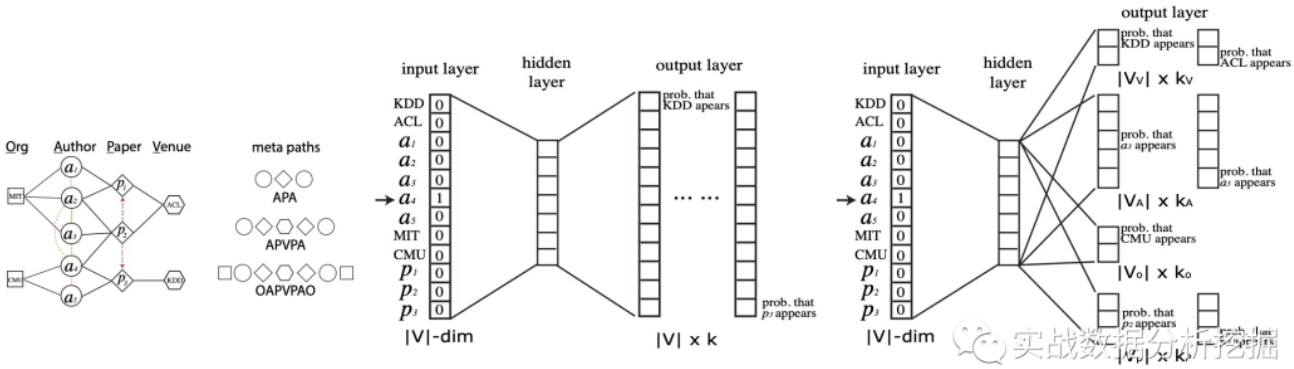
以往我们提到图嵌入，目标会表述成将节点进行向量化表达，并且能保持了节点之间的相似性，即在图上距离较近的节点，在新的向量空间中，也具有较近的距离。但本文除了要保持这个网络结构的特征之外，还提到了semantic relations，这一点在类似DeepWalk的方法中是未曾见过的。

> PROBLEM 1. ***Heterogeneous Network Representation Learning:*** *Given a heterogeneous network G, the task is to learn the d-dimensional latent representations $\mathbf{X} \in \mathbb{R}^{|V| \times d}$, $d \ll |V|$ that are able to capture the structural and semantic relations among them.*

解题思路

当我们面临一个新的问题时，一种有效的思路往往是先思考一下有没有处理过相似的问题，然后将其中使用的方法进行迁移改造以适配新的问题。猜想本文的作者也是类似这样来思考的。我们回顾一下DeepWalk与Node2Vec，它们虽然采用了不同的方法来生成节点序列，但都是为了将数据预处理成Word2Vec可以使用的格式，然后再直接利用SkipGram解决问题，而SkipGram本身是一个简单的神经网络（只有一个隐藏层）。带着这些旧知识，我们来看MetaPath2Vec怎样举一反三，它的核心逻辑其实是极像deepwalk，无非是把应用场景拓展到了异质图，需要做一些方法改造而已。

先给出论文中的一张信息含量极高的图，可以在脑海里存档一下，然后接下来我们逐步拆解Metapath2vec的算法思想。



(a) An academic network (b) Skip-gram in *metapath2vec*, node2vec, & DeepWalk (c) Skip-gram in *metapath2vec++*

此时此刻，我们的脑海里可能会浮现出很多问题，不妨带着问题来找答案，这样有利于更主动地进行学习。

在异质图上，每一个节点的领域怎么表示?

我们知道deepwalk是从每个节点的领域中随机选择下一个游走的对象，那么在异质图中的节点领域怎么理解?

我们首先在论文描述Heterogeneous Skip-Gram的目标函数时，看到了一处对于领域的讨论：

network $G = (V, E, T)$ with $|T_V| > 1$ by maximizing the probability of having the heterogeneous context $N_t(v), t \in T_V$ given a node v:

$$\arg\max_{\theta} \sum_{v \in V} \sum_{t \in T_V} \sum_{c_t \in N_t(v)} \log p(c_t | v; \theta) \qquad (2)$$

where $N_t(v)$ denotes v's neighborhood with the t^{th} type of nodes

这里是把节点v的邻居按照节点的类型进行了划分，然后Nt(v) 单独表示某一类的领域。

在讨论异质图上的随机游走时，又看到对于第i步的节点，第i+1步只能从元路径中的下一个节点范围中选择，因此这里的领域可以基于元路径上的节点顺序来理解。

所以领域是一个相对的概念，在不同的场景讨论里有不同的理解角度。

在异质图上怎么生成随机游走的节点序列?

回想deepwalk的随机游走机制，不加区分地选择下一步前进的方向，但在异质图中，因为节点类型可能有多种，如果某种类型的节点数量特别多，游走出来的路径中这类节点的占比就会更高，易引入统计偏差。所以作者借用了一个概念叫"元路径"(meta-path)，定义如下：

Formally, a meta-path scheme $\mathcal{P}$ is defined as a path that is denoted in the form of $V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \cdots V_t \xrightarrow{R_t} V_{t+1} \cdots \xrightarrow{R_{l-1}} V_l$, wherein $R = R_1 \circ R_2 \circ \cdots \circ R_{l-1}$ defines the composite relations between node types V_1 and V_l [25]. Take Figure 2(a) as an example, a meta-path "APA" represents the coauthor relationships on a paper (P) between two authors (A), and "APVPA" represents two authors (A) publish papers (P) in the same venue (V). Previous work has shown that many data mining tasks in heterogeneous information networks can benefit from the modeling of meta-paths [6, 25, 27].

给给定一种元路径的模式后，随机游走的下一个节点就被限制到了元路径上的节点范围，然后以等概率来随机抽样（类似于deepwalk在邻居中随机抽样）。

$$p(v^{i+1}|v_t^i, \mathcal{P}) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) = t+1 \\ 0 & (v^{i+1}, v_t^i) \in E, \phi(v^{i+1}) \neq t+1 \\ 0 & (v^{i+1}, v_t^i) \notin E \end{cases}$$

得到节点序列后，怎么利用Word2Vec中的SkipGram来进行向量化表示?

普通的SkipGram选择的优化目标函数如下，它的目标就是要在给定v的情况下，使得其邻居单词出现的联合概率值最大。

$$\arg\max_{\theta} \prod_{v \in V} \prod_{c \in N(v)} p(c|v;\theta)$$

对于求极值问题而言，加一个单调函数（比如log）并不影响极值求解，所以对于连续相乘的计算，套一个log, 会变成连续相加的计算。

Heterogeneous Skip-Gram. In *metapath2vec*, we enable skip-gram to learn effective node representations for a heterogeneous network $G = (V, E, T)$ with $|T_V| > 1$ by maximizing the probability of having the heterogeneous context $N_t(v), t \in T_V$ given a node v:

$$\arg\max_{\theta} \sum_{v \in V} \sum_{t \in T_V} \sum_{c_t \in N_t(v)} \log p(c_t | v; \theta) \quad (2)$$

where $N_t(v)$ denotes v's neighborhood with the t^{th} type of nodes and $p(c_t|v;\theta)$ is commonly defined as a softmax function [3, 7, 18, 24], that is: $p(c_t|v;\theta) = \frac{e^{X_{c_t} \cdot X_v}}{\sum_{u \in V} e^{X_u \cdot X_v}}$, where X_v is the v^{th} row of $\mathbf{X}$, representing the embedding vector for node v. For illustration, consider the academic network in Figure 2(a), the neighborhood of one author node a_4 can be structurally close to other authors (e.g., a_2, a_3 & a_5), venues (e.g., ACL & KDD), organizations (CMU & MIT), as well as papers (e.g., p_2 & p_3).

对于metapath2vec而言，通过对上述目标函数的求解即可得到最终的向量化表达。但是，注意上述对于概率p的定义中，其实是在所有节点上进行的标准化；metapath2vec++则改动了一下：

$$p(c_t|v;\theta) = \frac{e^{X_{c_t} \cdot X_v}}{\sum_{u_t \in V_t} e^{X_{u_t} \cdot X_v}} \quad (5)$$

where V_t is the node set of type t in the network. In doing so, *metapath2vec++* specifies one set of multinomial distributions for each type of neighborhood in the output layer of the skip-gram model. Recall that in *metapath2vec* and node2vec / DeepWalk, the

这就是metapath2vec、metapath2vec++的关键差别啦！

异质图中有不同类型的节点，那它们生成的向量表示的长度一样吗？

答案是一样的。论文中有这样一段话：

> although there are different types of nodes in V , their representations are mapped into the same latent space.

小结

介绍了异质图上的图嵌入表示Metapath2Vec。它的核心思路仍然是类似于deepwalk的，即先进行随机游走得到节点序列，再借用skipgram方法进行向量化求解，这里面重点在于结合了异质图的特点对原方法进行改造，在随机游走中利用了元路径的概念，并可以定义出不同的节点领域，用于目标函数的优化求解。

参考资料

1. 提出MetaPath2vec的论文：
https://dl.acm.org/doi/pdf/10.1145/3097983.3098036
2. https://ericdongyx.github.io/papers/KDD17-dong-chawla-swami-metapath2vec-poster.pdf
3. 讲 metapath2vec 的一份slides，https://ericdongyx.github.io/papers/KDD17-dong-chawla-swami-metapath2vec-slides.pdf
4. 论文《Meta-Path-Based Search and Mining in Heterogeneous Information Networks》