基于机理嵌入的交通流图预测神经网络方法解释与延伸

hy
北京航空航天大学

lxd
北京航空航天大学

日期：2023 年 6 月 2 日

摘　要

本文对交通流图预测问题进行了简要介绍，并就《DIFFUSION CONVOLUTIONAL RECURRENT NEURAL NETWORK: DATA-DRIVEN TRAFFIC FORECASTING》[1] 这篇论文提出的经典 DCRNN 方法进行分析，得出其机理嵌入的过程及原理。DCRNN 通过引入马尔科夫过程、图论中邻接矩阵与两点路径的关系等数学机理，提高了模型的可解释性，并在效果上超过了 state-of-the-art baselines($SOTA$) 的 12%-15%。此外本文针对其应用进行介绍——以机场延误预测为例[2]，描述了此网络在当今的热门应用场景。最终我们将其与注意力网络相结合，尝试提出一种参数灵活度更高的模型结构。

关键词： 机理嵌入，深度学习，图论，马尔科夫过程

1　交通流图预测网络问题

交通流图预测网络问题是深度学习中的一个重要研究方向，旨在利用神经网络模型来预测交通流量和交通拥堵等情况。

在城市交通系统中，交通流图是描述交通网络中车辆等交通工具流动情况的图结构表示。此预测网络问题的目标是根据历史的交通流图数据，预测未来一段时间内的交通流量、交通速度或交通拥堵情况。这对于交通管理、智能交通系统和交通规划等领域具有重要意义。

为了解决交通流图预测问题，深度学习方法被广泛应用。通常，基于深度学习的交通流图预测网络模型会接收历史的交通流图数据作为输入，并通过神经网络的层次化结构进行特征学习和表示学习。训练过程中，模型通常使用监督学习方法，通过最小化预测值和真实流量之间的差异来调整网络参数。为了提高预测性能，可以采用各种技术和策略，如数据增强、特征选择、模型融合等。通过对历史数据的学习，模型可以学习到交通流量的时间和空间相关性，从而进行准确的流量预测。

交通流图预测网络的应用范围广泛，可以用于交通拥堵预警、交通优化调度、路径规划、智能交通信号控制等方面。通过准确预测交通流量和拥堵情况，可以提高交通系统的效率和安全性，减少交通拥堵和出行时间，为城市居民提供更好的出行体验。

目前此解决措施主要以 DCRNN[1] 以及 STGCN[3] 两种基本模型及各自变形为主。本篇主要介绍关于 DCRNN 的相关实现方法以及机理嵌入过程。

2　DCRNN 相关机理嵌入

2.1　马尔科夫过程

马尔可夫过程就是**未来只与现在有关，与过去无关**。设有一个随机过程 $X(t)$，如果对于下一个任一时间序列 $t_1 < t_2 < ...... < t_n$，在给定随机变量 $X(t_1) = x_1, X(t_2) = x_2, ..., X(t_{n-1}) = x_{n-1}$ 的条件下，$X(t_n) = x_n$ 的分布可以表示为

$$F_{t_n,t_1,t_2,......,t_n}(x_n|x_1, x_2, ..., x_{n-1}) = F_{t_n,t_{n-1}}(x_n|x_{n-1}) \tag{1}$$

则称 $X(t)$ 为马尔科夫过程，简称马氏过程。

这种“**下一时刻的状态只与当前状态有关，而与上一时刻的状态无关**”的性质，成为**无后效性**或者**马尔科夫性**，具有这种性质的过程就称为**马尔科夫过程**。

马尔科夫过程中有两个比较重要的概念：**转移分布函数**、**转移概率**。

马氏过程 $X(t)$，称条件概率 $F_{s,t}=P\{X(t)\le y|X(s)=x\}$ 为过程的**转移分布函数**。其条件概率为 $f_{t_n|t_{n-1}}(x_n|x_{n-1})$ 为转移概率密度，称 $P(X(t_n)=x_n|X(t_{n-1})=x_{n-1})$ 为**转移概率**。

2.1.1 马尔科夫链

马尔科夫链（Markov）是最简单的马氏过程，即**时间和状态过程的取值参数都是离散的**马氏过程，其时间和状态的取值都是离散值。

假设在每一个时刻 $t_n(n=1,2,...), X_n=X(t_n)$ 所有可能的状态的集合 S 是可数的，也即可以表示为 $S=0,1,2...$ 对应于时间序列 $t_1,t_2,...t_n,...$，马氏链的状态序列为 $i_1,i_2...,i_n,...$。

对于马尔科夫链，若转移概率 $P\{X(t_n)=x_n|X(t_{n-1})=x_{n-1}\}$ 与 n 无关（即与哪一次转移无关，仅与转移前后的状态有关），则该马氏链为**齐次马氏链**；否则称为**非齐次马氏链**。接下来我们仅讨论齐次马氏链。

对于齐次马氏链，转移概率为 $P_{ij}=P\{Xn=j|X_{n-1}=i\}$，成为马氏链的一步转移概率，并且满足条件：$P_{ij}\ge 0,\sum_{j=0}^{\infty}P_{ij}=1, j=0,1,......$

$$P=\begin{bmatrix} P_{00} & P_{01} & P_{02} & ... \\ P_{10} & P_{11} & P_{12} & ... \\ & ...... & & \\ P_{i0} & P_{i1} & P_{i2} & ... \\ & ...... & & \end{bmatrix}$$

马尔科夫链的 n 步转移概率为 $P_{ij}^{n}=P\{X_{n+m}=j|X_m=i\}$，表示当前（第 m 步）的状态为 i，经过 n 步转移后（第 $n+m$ 步）系统的状态为 j 的概率。

马尔科夫链的 n 步转移过程可以先经过 m_1 步由状态 i 转移到状态 k，再经过 m_2 步由状态 k 转移到状态 j：

$$P_{ij}^{m_1+m_2}=\sum_{k=0}^{\infty}P_{ik}^{m_1}P_{kj}^{m_2} \tag{2}$$

2.1.2 马尔科夫链的稳定分布

若 $p_j=\sum_{j=0}^{\infty}p_iP_{ij}$，$j=0,1,...$，则称概率分布 $\{p_j|j\ge 0\}$ 是马氏链的稳态分布。对于稳态概率分布，存在 $\sum_{j=0}^{\infty}p_j=1$。

稳态概率反映了系统达到稳态后，系统处于某一状态的可能性（概率）。

稳态分布可以表示为：

$$p_j=\lim_{n\to\infty}P\{X_n=j|X_0=i\},\ i=0,1,... \tag{3}$$

即过程从初始状态 $X_0=i$ 出发，最终转移到状态 $X_n=j$ 的概率，并且与初始状态 $X_0=i$ 无关。

也可以表示为：

$$p_j=\lim_{k\to\infty}\frac{\text{前}k\text{次转移中访问状态}j\text{的次数}}{k} \tag{4}$$

其中，p_j 表示该过程中访问状态 j 的时间比例或频率，且与初始状态无关。

2.1.3 马尔科夫链的全局平衡方程

在马尔可夫链在稳态情况下从一个状态出发总会转移到一个状态，所以

$$p_j \sum_{i=0}^{\infty} P_{ji} = \sum_{i=0}^{\infty} p_i P_{ij},\ j = 0, 1, ... \tag{5}$$

上式称为全局**平衡方程**，它表示在稳态情况下，从一个状态 j 转移出去的频率等于转移进入状态 j 的频率，全局平衡方程是一种典型的求解概率分布的方法。

2.2 图的邻接矩阵

2.2.1 基本含义

设 $G =< V, E >$ 是**有向图**，$|V| = n$，矩阵 $A_{n\times n} = [a_{ij}]$，若边 $(v_i, v_j) \in E$，则 $a_{ij} = d_{ij}$，否则 $a_{ij} = 0$，称矩阵 $A_{n\times n}$ 是 G 的邻接矩阵。

设 $G =< V, E >$ 是**无向图**，$|V| = n$，矩阵 $A_{n\times n} = [a_{ij}]$，若边 $(v_i, v_j) \in E$，则 $a_{ij} = d_{ij}$ 且 $a_{ji} = d_{ij}$，否则 $a_{ij} = 0$ 且 $a_{ji} = 0$，称矩阵 $A_{n\times n}$ 是 G 的邻接矩阵。

其中 d_{ij} 代表边的权重。

对于无向图 G 的邻接矩阵 A：

1. A 是对称阵。
2. A 的对角线元素全部为 0。
3. A 的每一行元素之和 = 对应顶点的度数。

2.2.2 矩阵 A^n 的含义

定理（有向图）：设 $G =< V, E >$ 是有向图，矩阵 $A_{n\times n}$ 是 G 的邻接矩阵，则矩阵 A^k 中第 i 行第 j 列元素的值 $a_{ij}^{(k)}$，恰好等于从顶点 i 到顶点 j 的、长度为 k 的通路的数量。

推论（有向图）：$a_{ii}^{(k)}$ 等于长度为 k 的、经过顶点 i 的回路数量。

定理（无向图）：设 $G =< V, E >$ 是无向图，矩阵 $A_{n\times n}$ 是 G 的邻接矩阵，则矩阵 A^k 中第 i 行第 j 列元素的值 $a_{ij}^{(k)}$，恰好等于从顶点 i 到顶点 j 的、长度为 k 的链的数量。

推论（无向图）：

1. $a_{ii}^{(k)}$ 等于长度为 k 的、经过顶点 i 的闭合链数量。
2. 如果有 $a_{ij}^{(1)}, a_{ij}^{(2)}, ..., a_{ij}^{(k-1)}$ 均为 0，而 $a_{ij}^{(k)}$ 不为 0，则顶点 i 与顶点 j 之间的距离为 k。
3. 如果有 $a_{ij}^{(1)}, a_{ij}^{(2)}, ..., a_{ij}^{(n-1)}$ 均为 0，则顶点 i 与顶点 j 分别属于不同的连通分支。

2.3 从理论基础到图预测模型

2.3.1 交通预测问题

交通预测的目标是预测未来的交通速度，给定先前从道路网络上的 N 个相关传感器观察到的交通流量。我们可以将传感器网络表示为一个加权有向图 $G = (V, E, W)$，其中 V 是一组节点 $|V| = N$，E 是一组边，$W \in R^{N\times N}$ 是一个加权邻接矩阵，表示节点的邻近度（例如，它们的路网距离的函数）。将 G 上观察到的交通流量表示为图形信号 $X \in R^{N\times P}$，其中 P 是每个节点的特征数量（例如速度、体积）。假设 $X(t)$ 表示在时间 T 观察到的图形信号，交通预测问题旨在学习一个函数 $h(\cdot)$，该函数将 T 历史图形信号映射到未来的 T 图形信号，给定一个图形 G：

$$h([X^{(t-T'+1)}, ..., X^{(t)}; g]) \rightarrow [X^{(t+1)}, ..., X^{(t+T)}] \tag{6}$$

2.3.2 空间依赖模型

通过将交通流与扩散过程相关联来对空间依赖性进行建模，这明确地利用了交通动力学的随机性质。这个扩散过程的特征是一个**重启概率为** $\alpha \in [0,1]$ 的随机游走和一个**状态转移矩阵** $D_0^{-1}W$。这里 $D_0 = diag(WI)$ 是外度对角矩阵，$I \in R$ 表示全一向量。

经过许多时间步长后，这种马尔可夫过程**收敛到一个平稳分布** $P \in R^{N\times N}$ ，行 $P_{i,:} \in R^N$ 表示从节点 $v_i \in V$ 扩散的可能性，因此接近节点 v_i。下面的扩散卷积引理为平稳分布提供了一个封闭形式的解：

引理：扩散过程的平稳分布可以表示为图上无限随机游动的加权组合，并以**封闭形式**计算：

$$P = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k (D_0^{-1}W)^k \tag{7}$$

α：停机概率，即不再扩散。

$1-\alpha$：继续扩散的概率。

$D^{-1}W(i,j)$：点 i 转移到 j 的概率，其中 W 是带权矩阵，D 是度矩阵。

W 经过操作

$$W_{ij} = e^{-\frac{dist(v_i,v_j)^2}{\sigma^2}} \quad if\ dist(v_i,v_j) \le \kappa, otherwise\ 0 \tag{8}$$

结果是距离越大，权重越小，转移到的概率越小，因此是合理的。

$(1-a)^k$ 代表第 k 次还继续扩散的概率，$(D^{-1}W)^k$ 代表转移 k 次到各点的概率（图论结论）。

（1）扩散卷积——双向扩散

对图信号 $X \in R^{NP}$ 和一个滤波器 f_θ 的扩散卷积运算定义为：

$$X_{:,p\star g}\ f_\theta = \sum_{k=0}^{K-1} (\theta_{k,1}(D_0^{-1}W)^k + \theta_{k,2}(D_I^{-1}W^T)^k)X_{:,p} \tag{9}$$

其中 p 代表 $feature$ 维度，可能不只有一个特征，我们需要对每个特征分别进行操作：

1. 分别计算点到其它各点的概率，同时计算各点到该点的概率，进行加和处理。

2. $D_0^{-1}W$ 可以理解成列上的点到行上点的概率（因为是对行求的概率）。

3. $D_0^{-1}W^T$ 可以理解成行上的点到列上点的概率（因为求了转置）。

因此相加后就能得到点 a 到点 b 和点 b 到点 a 双向概率。θ 可以理解对 $a(1-a)^k$ 的替代，以**降低公式复杂度**，且形式简单。

（2）扩散卷积——升维

$$H_{:,q} = a(\sum_{p=1}^{P} X_{:,p\star g} f_{\Theta_{q,p,:,:}}) \ \ for\ q \in \{1,...,Q\} \tag{10}$$

相当于对每一个数据都进行了 q 次卷积，起到升维的作用。

2.3.3 时间依赖性

利用循环神经网络 (RNN) 对时间依赖性进行建模——**使用门控循环单元（GRU）**，这是一种简单但功能强大的 LSTM 网络变体，文章将 GRU 中的矩阵乘法替换为扩散卷积，从而产生了所提出的扩散卷积门控循环单元 (DCGRU)。

$$\begin{aligned} r^{(t)} &= \sigma(\Theta_{r\star g}[X^{(t)}, H^{(t-1)}] + b_r) & u^{(t)} &= \sigma(\Theta_{u\star g}[X^{(t)}, H^{(t-1)}] + b_u) \\ C^{(t)} &= tanh(\Theta_{C\star g}[X^{(t)}, (r^{(t)} \odot H^{(t-1)})] + b_c) & H^{(t)} &= u^{(t)} \odot H^{(t-1)} + (1-u^{(t)}) \odot C^{(t)} \end{aligned} \tag{11}$$

2.4 机理嵌入汇总

1. 马尔科夫过程到扩散过程平稳分布：$P = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k (D_0^{-1}W)^k$。
2. 通过 A^n 求解从 i 经过 k 次转移后到达 j 点的路径权重。
3. 矩阵求转置即为 2 的逆运动过程。
4. 通过多次扩散卷积升维，提高特征提取数量。

3 图预测应用——以机场延误预测为例

3.1 机场延误问题

航空运输在商业和旅游业中具有重要意义，2019 年全球旅客吞吐量近 45 亿人次，创历史新高。据同年统计，欧洲平均每次航班延误 13.1 分钟，美国为 12.9 分钟，中国为 14 分钟左右。这些延误导致了不可避免的后果，例如不愉快的乘客体验，以及随之而来的相关空域用户的经济损失。据估计，航班延误每年给全球经济造成的损失为 2019 年 500 亿。如此高的损失促使工业界和学术界对空中交通延误进行分析，并开发更先进的航班延误预测方法。

3.2 问题建模

在航班延误预测中，数学模拟和数据驱动方法是目前具有代表性的两类工作。

数学模拟方法使用数学工具对空中交通运行进行建模，通常需要大量的计算资源，而且因为存在一些假设和简化而导致其预测精度大大下降。

现有的许多数据驱动方法仅仅只关注单一机场的场景，而忽略了隐藏在机场网络中的动态空间相互作用。事实上，由于大量相互关联的资源（例如，飞机、机组人员、乘客和基础设施），航班延误可以在航空运输系统内随机发生，并且很容易在整个紧密连接的机场网络中传播。因此，有必要关注连接机场之间的动态空间交互，即图结构信息。

而这恰好与文章提到的 DCRNN 问题相对应，是交通流预测的变型。

因此做如下问题建模：

空运网络可以抽象为图 $G = (V, E, W)$

$V = \{v_1, v_2, ..., v_n\}$：代表空运网络的所有机场。

E：代表航线的边集。

$W_{n\times n}$：带权邻接矩阵，w_{ij} 代表 $t-1 \sim t$ 时刻，$< v_i, v_j >$ 之间的航班数量。

$Y = \{y_1, y_2, ..., y_n\}$：$t$ 时刻 n 各机场的观察向量。

y_i 记录 $t-1 \sim t$ 时刻，机场 v_i 经历的延误，可以由以下公式进行计算：

$$y_i^{(t)} = \frac{m_i^{(t)} + \rho * c_i^{(t)}}{a_i^{(t)}} \tag{12}$$

其中各个参数表示的含义：

m_i：所有 $t-1 \sim t$ 时刻延误航班的延误时间和；

c_i：取消航班的数量；

a_i：计划航班总数；

ρ：表示取消航班对应延误时间权值的常量，论文[2] 中被定为 180。

即航班延误不仅考虑了真实的延误时间，同时也将航班取消纳入考虑。

因此，输入的为飞机机场与航班的有向图，特征为延误时间，这样便可以用 DCRNN 进行求解，实现飞机航班延误时间的预测，从而提供实时的、准确的延误信息，帮助航空公司、机场管理部门和旅客做出相应的决策和应对措施，减小损失。

4 模型的改造——与注意力网络相结合

4.1 图注意力网络 GAT

GAT[4] 的主要目标是对图中的**每个节点进行表示学习**，同时**考虑其与邻居节点之间的关系**。在 GAT 中，每个节点都与其邻居节点相连，通过学习节点对邻居节点的重要性权重，可以对每个节点进行更精确的表示学习。注意力权重是通过对节点之间的特征进行计算和归一化得到的，其中计算过程使用了**多层感知机（MLP）来学习权重系数**。核心权重公式为：

$$\alpha_{ij} = \mathrm{softmax}\left(\sigma\left(\vec{a}^T\left[W\overrightarrow{h_i}\|W\overrightarrow{h_j}\right]\right)\right) \tag{13}$$

首先权重矩阵 W，是一个 $A * B$ 尺寸的矩阵。A 表示输入节点特征的维数，而 B 表示该层输出节点的维数。h 来表示隐藏层特征 hidden feature，其尺寸经常为 $1 * A$。中间 $\|$ 代表张量粘合，再乘上注意力核 $\vec{a}^T$(尺寸为 $1 * 2B$)，最终得到 1×1 的数字，这个数就是**未加工的 attention 系数**。

GAT 模型的核心是通过**自适应地计算**注意力权重来融合邻居节点的信息。这使得模型能够对不同节点之间的关系进行灵活的建模，更好地捕捉图结构中的局部和全局信息。其优势在于能够通过注意力机制对不同节点之间的关系进行个性化建模，从而提高了节点表示的质量和表达能力。此外，GAT 还可以处理具有不同数量和连接性的节点的图数据，并**具有一定的鲁棒性和泛化能力**。

4.2 核心公式的变化

$$P = \sum_{k=0}^{\infty} \alpha(1-\alpha)^k A^k \tag{14}$$

当 $A = D_0^{-1}W$ 时，就是上文提到的 DCRNN 的扩散公式。有以下几个问题值得考虑：

1. 用距离的远近表示概率，相同距离的效果一致，这是否合理。
2. 随机扩散与现实交通流是否相符，即是否具有物理意义。
3. 双向传播是否有效。

能够发现，DCRNN 方法尽管理论充足，但在物理意义方面的可解释性仍有所欠缺。

我们的方法为，将此处的矩阵 A 设置为注意力矩阵。注意到扩散矩阵应为无量纲矩阵，但 DCRNN 的 A 矩阵依旧**有长度量刚的影响**，而注意力矩阵则**消除了量刚影响**。此外根据注意力计算公式：

$$\alpha_{ij} = \mathrm{softmax}\left(\sigma\left(\vec{a}^T\left[W\overrightarrow{h_i}\|W\overrightarrow{h_j}\right]\right)\right) \tag{15}$$

其中 j 号节点为 i 号节点的邻居。

能够发现其也考虑了邻居节点的关系（一阶邻居）。

我们将节点与节点之间的特征关系看做特征的“流动”，信息流在不同节点间的流动性不同。而信息流的“流动”在交通流问题中恰好能够**看做扩散过程**，即有 $A = [a_{ij}]_{n\times n}$，其中 a_{ij} **描述了从 i 到 j 的流动率**。因此 DCRNN 的扩散过程中的 $D_0^{-1}W$ 通过注意力系数矩阵 A 替代。

值得注意的一点，A 已经代表了流动过程，因此停机概率在 A 是可变训练矩阵的条件下便可以与其合并，最终采取 $softmax$ 归一化后，得到最终的转移概率计算方法：

$$\begin{aligned}
&P = \mathrm{softmax}(\sum_{k=0}^{K} A^k) \\
&A = [a_{ij}]_{n\times n} \\
&\alpha_{ij} = \mathrm{softmax}\left(\sigma\left(\vec{a}^T\left[W\overrightarrow{h_i}\|W\overrightarrow{h_j}\right]\right)\right)
\end{aligned} \tag{16}$$

回顾整体的思路过程，从起初的**马尔科夫过程开始**得到了**随机扩散概率计算方法**，而后通过改变扩散相关的系数矩阵——**从邻接矩阵到注意力矩阵**，具体了参数的意义，再由 A 的可训练性**将 α 吸收**，最

终经过**归一化**得到我们所需的特征。尽管最初共识与改变后的有较大区别，但是整个过程通过一步步推到简化，保留了其原有的物理意义。

此外在论文中[4] 作者**提到了** $multi - head$ **方法**，能够进一步提高注意力网络的性能。

最终我们的方法具有以下优势：

1. 给予了扩散和注意力模型“流动”的特性，具有了扩散的物理意义。

2. 参数少于 DCRNN 网络。

3. 马尔科夫过程的 k 参数代表了每个节点与 k 阶邻居节点能够产生关系，计算过程简单，覆盖节点多。

5 总结

深度学习模型的数据驱动方法依赖于大量的训练数据来学习复杂的模式和特征。然而，当**数据量有限、存在噪声或者任务复杂时**，仅仅依靠数据可能存在一些限制，因此通过引入机理使得模型能够更深入的**接触问题核心**，这对深度模型的训练是重要的。它可以帮助模型克服数据不足、噪声和过拟合等问题，提高模型的泛化能力和性能。通过引入已有数学、物理等机理，深度学习模型可以更好地学习并适应复杂的任务和领域。

本篇文章主尝试对 DCRNN 网络实现过程中的机理嵌入部分进行了解释——包括马尔科夫过程、图的邻接矩阵含义等等，并就注意力神经网络进行了拓展尝试，旨在进一步提高模型的泛化能力，提升模型性能。

通过本次研究，让我们认识到了机理嵌入在模型效果以及可解释性方面的贡献，为我们今后的研究方向提供了一种选择。未来进行模型构建时，我们也会尝试引入机理的概念，使得模型更加合理。

参考文献

[1] Yaguang Li et al. “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting”. In: *arXiv preprint arXiv:1707.01926* (2017).

[2] Kaiquan Cai et al. “A deep learning approach for flight delay prediction through time-evolving graphs”. In: *IEEE Transactions on Intelligent Transportation Systems* 23.8 (2021), pp. 11397–11407.

[3] Bing Yu, Haoteng Yin, and Zhanxing Zhu. “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting”. In: *arXiv preprint arXiv:1709.04875* (2017).

[4] Petar Veličković et al. “Graph attention networks”. In: *arXiv preprint arXiv:1710.10903* (2017).