区域需求预测

明岭峰

January 7, 2020

1 摘要

智能交通系统的一个重要阻碍是乘客需求的预测，一个精准的模型可以帮助中心调度提前分配资源，以满足乘客出行需求，减少空车巡航时间和乘客的等待时间，同时避免交通堵塞，减少能源浪费。

基于历史打车数据，我们提出GCN+LSTM模型解决乘客需求预测问题，该框架主要基于两层：(1) 考虑地理近邻与语义近邻关系，构建GCN层以提取时空依赖关系；(2) 构建LSTM层提取时序依赖关系。接下来，考虑距离因素，根据重力模型得出边的权重存在方向性，改进后的GCN可以提高模型精度。我们的模型在真实数据上展示出更好的性能并在后续地调度算法中得到验证。

2 简介

交通出行影响着人们的日常生活，可以说是一个城市的脉搏，而构建一个高效的智能交通系统的主要障碍是精准地预测乘客地出行需求。出租需求量预测模型越准，提前调度资源以满足出行需求可以更好地减少空车巡航时间和乘客等待时间。

我们关注地问题是基于历史数据，预测某区域在未来时间片的出租需求数目。其由以下因素影响：

1. **空间依赖**：地理相邻的区域总是存在相互影响，距离较远的区域也可能存在某种联系，比如A,B两个区域同样有学校和医院，那么它们的需求量比较相似；
2. **时序依赖**：区域的需求量既受最近也受较远时间段的影响。比如某区域上午8点的交通堵塞会影响到9点；除此之外，在工作日某区域上午的出行需求基本相似，并且可能随着月份的变化也会有所变化，因为气候的变化，人们作息时间也会变化，比如中国实行的夏季与冬季作息；

3. **额外因素**：一些额外的因素也会影响出行需求量，比如：天气、空气质量、节假日、大型事件等等。

解决上述问题，我们基于GCN+LSTM预测模型的主要创新点在于：

- 我们提出的端到端深度学习模型，不仅适合于基于网格划分的规则区域，也适合于基于路段划分的不规则区域，取决于应用需求，同时综合考虑了空间、时序、额外因素的影响；
- 基础模型时，考虑区域之间的近邻关系不仅包括地理近邻，也包括语义近邻，如果区域近邻，那么存在空间依赖关系；
- 在基础模型上，根据重力模型，考虑距离与本身节点信息，我们认为区域之间的相对影响是不同，具有方向性，对此分别构建出度与入度邻接矩阵，使用GCN提取区域的近邻关系；
- 我们的模型在New York City的数据集上得到验证，通过多组对比实验展示出该模型的优势，并在后续的调度算法中再次得到很好效果的验证。

3 现有工作总结

对于预测区域需求的问题，主要有基于CNN与RNN结合的方式和基于GCN与RNN结合的模型。

由于基于CNN的模型主要存在以下问题:

- 无法捕捉需求模式相同但距离较远区域的依赖关系，即无法捕捉语义近邻；
- 只能用于欧几里得数据（规则的数据格式）

所以倾向于用基于GCN的模型解决此类问题，因为GCN能很好地提取网络拓扑特征。而此前基于GCN的模型有同时考虑邻居、功能相同、连通性等多图融合的方式，此模型比较复杂，计算量大，且存在冗余关系；最新的研究有仅仅考虑区域之间的语义近邻，认为与地理近邻无关，此方法简化了模型，但是达到了同时考虑功能相同和连通性的效果。

但是以上的基于GCN的工作均认为邻接区域间的影响是一样，也没有考虑区域间的距离因素；我们综合以上两种方法，考虑邻居近邻与语义近邻，同时考虑区域间的距离因素，采用重力模型，即认为两区域间的影响是不一样的，构建GCN提取近邻关系，再结合LSTM提取时序信息与额外因素，预测区域的需求量。

4 问题定义

将城市划分为N个小区域（网格划分或者基于路网划分都可）：$r_1, r_2, ..., r_N$，将一天划分为T个时间片，所以$D_t(r_i)$表示第i个区域在第t个时间片上的需求量，根据历史前t时刻数据（需求量与额外因素）:$\{D_0, D_1, ..., D_t\}$与$\{E_0, E_1, ..., E_t\}$，预测$T+1$时刻区域的需求量D_{t+1}，即学习一个函数，使得:

$$D_{t+1} = F(D_0, D_1, ..., D_t, E_0, E_1, ..., E_t)$$

5 模型

我们的预测模型框架由图卷积层和长短期网络层组成，其中图卷积层用于空间拓扑特征，再将空间特征、时序与额外因素信息，输入到长短期网络层提取时序依赖关系，用于预测下个时间片的需求量。

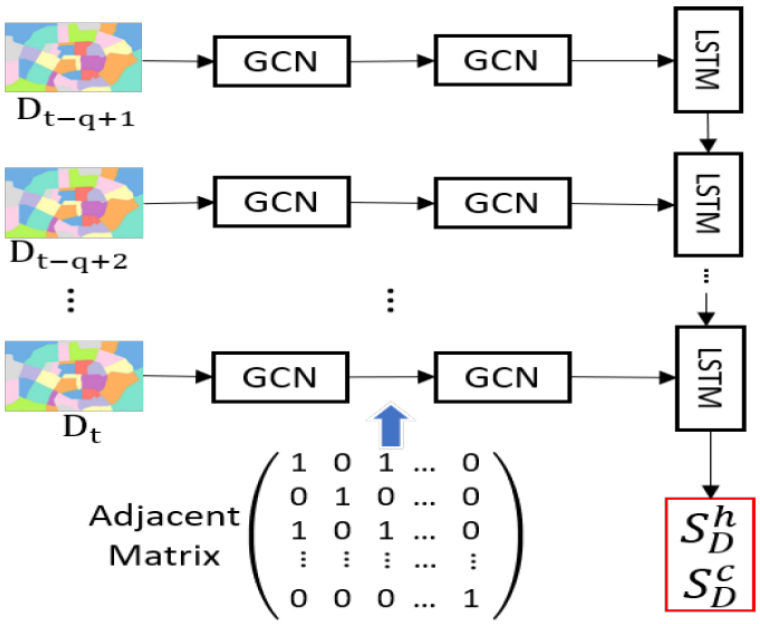

Figure 1: 图卷积网络

5.1 基础模型

同时考虑地理近邻与语义近邻作为节点之间存在边的依据，即如果两个节点地理近邻或者语义近邻，则两节点存在边，用二进制表示。

5.1.1 图卷积层

对于图可表示为$G = \{V, A\}$，其中V：节点集合，A：邻接矩阵；而图卷积的核心思想为利用边的信息对节点信息进行聚合，从而生成新的节点表示，所以邻接矩阵的构建直接影响GCN网络的特征提取。

基础模型的邻接矩阵如下：

$$A_{ij} = \begin{cases} 1, & \text{如果 } (i,j)\text{近邻} \\ 0, & \text{otherwise} \end{cases}$$

其中近邻关系包含地理近邻与语义近邻。地理近邻表示空间地理位置的相邻关系，语义近邻指需求模式的相似，认为需求模式相似为：$Similarity(i,j) > \epsilon$,其中

$$Similarity(i,j) = Pearson(D_{0-t}(r_i), D_{0-t}(r_j))$$

$D_{0-t}(r_i)$表示区域i的前t时间的训练数据需求量组成的序列。

一般图卷积网络的输入为$G_i \in R^{N\times K}$，输出为$G_o \in R^{N\times K^{'}}$，其中$N$为节点数量，$K$为节点特征维度。但由于我们与LSTM结合，所以我们输入GCN层增加一个步长q维度，即图卷积网络的输入为前q时间片的数据$G_i \in R^{q\times N\times K}$，输出为$G_o \in R^{q\times N\times K^{'}}$，本问题中节点的特征指需求量，所以$K = K^{'} = 1$。

5.1.2 长短期网络层

从图卷积层提取的空间信息、额外信息喂入长短期网络层，以提取时序依赖关系。值得注意的是，我们使用的是前q步长作为GCN层的输入，即GCN的输出为$G_e \in R^{q\times N\times K^{'}}$，其转化为LSTM的输入为$L_g \in R^{q\times NK^{'}}$，与额外信息$E \in R^{q\times m}$，（m为额外因素特征维度）合并为$X \in R^{q\times (NK^{'}+m)}$，LSTM层后加上密集连接层，最后输出为$Y \in R^{N\times 1}$。

5.1.3 损失函数

采用均方根误差作为损失函数，即

$$Loss(\theta) = \frac{1}{m}\sum_{k=1}^{m}(y_k - \bar{y}_k)^2$$

其中θ是需要学习的参数。

5.2 重力图卷积模型

认为节点之间的边存在权重，考虑节点之间距离因素的影响，认为权重不仅与距离有关，还与节点本身信息有关，即边的权重存在方向性。

5.2.1 重力模型

首先，对于邻接矩阵中边的确定：认为如果两个区域i,j是近邻的，则认为区域i,j之间存在边关系。

根据万有引力定理: 两个物体之间的相互作用力大小为：$F=G\frac{m_1m_2}{r^2}$，相互作用力的大小与它们质量的乘积成正比与它们距离的平方成反比；又根据牛顿第二定律:$a=\frac{F}{m}$，可以得到两物体之间由于引力作用而产生的加速度为$a_1=\frac{F}{m_1}=G\frac{m_2}{r^2}$,$a_2=\frac{F}{m_2}=G\frac{m_1}{r^2}$，由此可知两物体间的作用力大小相同，但若两物体的质量不同，则由于引力产生的加速度大小就不同。

将此万有引力定理运用到GCN模型中，认为邻接矩阵中边的权重具有方向性，即两个节点之间的权重 不仅与两节点的空间距离呈负相关，还与本节点信息有关，可以将其拆看为出度邻接矩阵与入度邻接矩阵，其边的权重关系定义分别如下：

- 出度矩阵

$$A_{ij}=\begin{cases}\frac{\log(mean(R_j))}{e^{\alpha*dist(i,j)}}, & \text{如果 } (i,j)\text{近邻}\\ 0, & \text{otherwise}\end{cases}$$

- 入度矩阵

$$A_{ij}=\begin{cases}\frac{log(mean(R_i))}{e^{\alpha*dist(i,j)}}, & \text{如果 } (i,j)\text{近邻}\\ 0, & \text{otherwise}\end{cases}$$

其中$mean(R_j)$表示区域j的平均需求量，α为常系数，$dist(i,j)$为区域(i,j)之间的空间距离。

5.2.2 模型输入输出

此时在原来的基础上，仅在模型输入上有一点小的改变，现在对出度与入度矩阵分别构建GCN，形成两个图，在输入LSTM时合并为 $L_g\in R^{q\times 2NK'}$，与额外信息$E\in R^{q\times m}$，（m为额外因素特征维度）合并为$X\in R^{q\times(2NK'+m)}$，LSTM层后加上密集连接层，最后输出为$Y\in R^{N\times 1}$。

6 实验

实验数据为New York 2016/06订单数据，其中最后5天作为测试数据，采用$Mean - Std$标准化处理需求量数据。考虑的额外因素主要有气象信息与时序信息，主要特征有天气，空气质量，time of day，day of week，前q时间步长内的平均需求量，采用$Max - Min$标准化处理额外信息数据。

6.1 度量

度量标准有平均百分误差(MAPE)、平均完全误差(MAE)和均方根误差(RMSE):

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\frac{|y_{i+1} - \bar{y}_{i+1}|}{y_{i+1}}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_{i+1} - \bar{y}_{i+1}|$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}|y_{i+1} - \bar{y}_{i+1}|^2}$$

6.2 方法对比

对比方法有

- 均值估计(HA)：使用相同时间间隔中给定区域上的先前历史平均值需求来预测需求
- LSTM：不加GCN层，只用LSTM模型
- GCN+LSTM(base)：不考虑距离的GCN+LSTM模型