# 基于LSM-RN模型的流量预测

路网流量预测是指给定路网历史流量数据，预测未来的流量数据。随着手机的普及和基站等基础设施的发展，我们可以容易地获得高精度高频率的手机信令数据。并可以通过一定的算法，从信令数据中获得路网上的车速、流量等信息。现有的路网信息预测模型中大多数是通过历史路网中的边信息预测未来路网中的边信息，如自回归滑动平均模型（ARIMA）[1],支持向量回归模型[2]，和高斯过程[3], 也有一些模型考虑了路网节点的拓扑关系，将相邻节点的历史信息用于对节点信息的预测，如HMM[4], 这些模型将时间信息和空间信息结合，取得了更好的效果。然而一些模型的计算复杂性高，如高斯过程和HMM，预测结果依赖于长时间的线下训练，再实际应用中很难做到实时。基于这些现状，我们采用一种基于隐空间模型的路网预测模型进行预测，即LSM-RN（Latent Space Modeling for Road Networks）。

LSM-RN将路网流量信息嵌入表示到隐空间，同时考虑了时间和空间的影响因素。认为路网流量信息在隐空间中，相邻节点的流量信息具有相似性，且节点的相邻时间片上流量信息具有一定的转化规律，并且可以通过矩阵描述。通过对路网历史流量信息的学习，挖掘出路网所对应的隐空间属性以及路网流量信息在隐空间中的变化规律。利用学习到的规律对未来几个时间片进行预测。此外，LSM-RN可以实现大规模的实时计算。

## 问题定义

道路根据沿路的基站经纬度划分成不同的路段，如图[]。每个路段对应一个基站，该路段的流量信息通过所对应基站的4G手机信令记录的数据根据文献[]中的算法得到，如下图中编号为1的路段流量数据由1号基站的数据得到。

（补图）

经过路段划分，道路网络可以被看作是一个有向图 , 其中是图的节点集，是边集。道路网络中的节点即为不同路段的起点和终点，边即为连接起始点的有向路段，每条边的数据对应该路段上的流量值。因此路网的邻接矩阵的元素对应值代表节点所连接边的权重，在本文中即为流量数据。举例说明，下图中展示了某个时间片下一个有七个节点和10条有向边的路网图

(补图)

流量信息的采样率可以在手机信令数据的采样率基础上调整。如手机信令数据的采样率为 可以设定流量采样间隔 。根据所设定的时间间隔 span， 我们对每个时间片都可以形成一个路网的快照信息图 。道路网络的流量动态变化可以表示成。在某一时间片下的路网流量快照中，不排除出现由于设备问题出现数据丢失等情况。

根据以上描述，路网流量信息预测可以定义为：给定一个动态道路网络，预测个时间片以后的路网,其中是预测步长，如若,则我们预测，即下一个时间片的路网信息。

## 隐空间模型

隐空间模型假设一个图的节点可以嵌入表示到一个可以描述节点隐层属性的隐层空间，并认为相似节点在隐层属性上相似度高。例如，在社交网络的社团发现[5]、话题发现[6]等问题中，认为在某些隐形属性上（如社团、话题等）距离近的用户更容易聚集为一个社团，或者更容易形成一个话题。在道路网络中，假设路网中的节点拥有很多隐层属性，且一条边的起始点和终止点不同的隐层属性间可能会相互影响，通过从历史数据中学习这种隐空间表示和变化规律，预测出未来路网的流量数据[7]。

隐空间模型实现原理是通过最小化实际值和根据隐空间属性表示得到的估计值间的差别，差别描述通常采用KL散度、最小平方误差等指标。具体的实现方法有很多种，最常用的是非负矩阵分解(Non-negative Matrix Factorization)。非负矩阵分解是将非负矩阵分解成两个非负矩阵和,其中

非负矩阵分解可以通过乘子迭代法[8]优化目标函数，得到局部最优解。

## LSM-RN模型

### LSM模型

一个拥有个节点的道路网络在时间的流量数据可表示成

对路网中的每一个顶点，假设其具有某些隐空间特征，设顶点在隐空间上的表示为,其中,为隐空间属性个数。不同的隐空间属性会影响道路的流量情况，例如交叉路口比非交叉路口流量大，城内道路比城际道路流量大等。举例说明如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 交叉路口 | 城市内 | 居民区 | 商业区 |  |
|  | 0.9 | 0.1 | 0.8 | 0.2 |  |

所有节点的隐空间表示组合一起可以构成隐空间矩阵 。

起始点和终止点不同的隐空间属性值会对流量值产生影响，例如两个交叉口间的路段相较其他普通路段会更大。这种隐空间属性之间的影响可以通过矩阵来描述，其中。节点和节点之间的流量可以看作隐空间属性的线性加和，如下：

路网中所有节点在隐空间上的映射表示可以表示成矩阵，我们希望通过隐空间表示矩阵和隐空间属性矩阵还原路网矩阵G，求解和的过程可以看作非负矩阵分解的过程，目标函数定义如下：

### LSM-RN 模型

道路网络具有稀疏性，道路网络流量变化具有平滑性，不会出现跳变现象，因此根据路网本身的特点，可以对基本LSM模型进行完善，形成LSM-RN模型，获得更好的表示结果。

#### 路网数据的稀疏性

在实际的路网数据中，道路网络是非常稀疏的，节点的平均度数非常小，且可能会出现数据丢失的情况，路网矩阵流量数据为0的路段有可能是两节点之间并无连接，也有可能是数据的丢失，但两种情况是无法区分的。因此为了克服矩阵的稀疏性，只考虑观测到的有值的部分，即流量值的元素。定义指示矩阵，其中

目标函数变为

其中，表示矩阵的哈达马积，满足。

实际路网数据中可能存在数据丢失，路网中相邻路段间的流量不可能发生跳变，因此我们可以合理假设相邻路段间的流量变化是平滑的，可以对目标函数加入惩罚因子如下：

对该式子进行化简如下：

其中矩阵为图的拉普拉斯矩阵，即，其中表示图的权重矩阵，即相似度矩阵，在路网中则可看成路网的邻接矩阵。

因此，目标函数化为

其中

为正则项系数。

#### 路网数据的时间相关性

每个时间片下路网会有对应的流量图,这些路网数据快照在隐层空间分别映射成对应的隐层空间图表示矩阵,同时有对应的指示矩阵，同时隐层属性交互矩阵应属于路网的固有属性，因此我们认为矩阵与时间无关，同理路网的拉普拉斯矩阵与时间也无关。则目标函数可改写如下：

路网流量数据变化是平滑的，不会出现跳变，可以合理假设相邻时间片的流量数据在隐空间模型上的变化可以通过转移矩阵来刻画，即，其中,通过历史流量数据的变化，可以学习流量转移矩阵，转移矩阵刻画了在时间到时间之间，隐层属性之间相互转化的可能性，如属性经过一个时间片将会有多大可能性转化为属性。在目标函数上添加转移矩阵的惩罚项，目标函数变为

其中为正则项系数。

#### 路网流量数据的预测

通过优化目标函数

其中为正则项系数，可以根据路网的历史流量数据学习到隐层属性矩阵、转移矩阵以及每个时间片路网对应的隐层属性表示矩阵，设预测步长为，则有以下式子

### 模型求解

非负矩阵分解的常见求解方式是迭代乘子法[8]，模型输入参数为时间到中的路网流量数据快照,指示矩阵,路网相似度矩阵等，学习到的变量有各个时间片对应的隐空间图表示矩阵，隐空间属性交互矩阵,转移矩阵,模型求解伪代码如下：

输入：$G\_1,G\_2,\ldots,G\_T$，$Y\_1,Y\_2,\ldots,Y\_T$，$W$  
输出：$U\_t(1\le{t}\le{T})$,A,B  
  
1. 初始化$U\_t(1\le{t}\le{T})$,A,B；  
2. while 不收敛 do  
3. for t=1 to T do  
4. 更新$U\_t$  
5. 更新 $B$  
6. 更新 $A$

各变量的更新规则[7]如下：

根据文献[8–10],求解过程正确性和收敛得证。

### 实验

#### 实验数据说明

本文实验采用福建省七条高速公路沿途基站2016年10月1日至2016年10月31日期间所收集的手机信令数据和基站经纬度数据，由中国移动（CMCC）提供，根据基站经纬度将道路划分为若干路段，道路信息如下：

高速公路信息表{#tbl:road\_info}

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 线路ID | 线路名称 | 路段个数 |
| 1 | G15 | 沈海高速公路 | 741 |
| 2 | S35 | 沈海高速复线 | 118 |
| 3 | G25 | 长深高速公路 | 305 |
| 4 | G70 | 福银高速公路 | 213 |
| 5 | G72 | 泉南高速公路 | 225 |
| 6 | G76 | 厦蓉高速公路 | 234 |
| 7 | G1501 | 福州绕城高速公路 | 68 |

原始4G信令数据记录字段包括：用户标识（加密）、日期时间、位置区码(LAC)，小区号(Cell Id)，信令类型等字段组成，其中位置区码和小区号唯一确定一个基站。示例数据如下：

|  |  |
| --- | --- |
| 字段 | 值（示例） |
| 用户标识（加密） | 4B2DD9C6843D678410647A4375A2A1E2 |
| 日期时间 | 2017-09-06 20:54:30 |
| 位置区码(LAC) | 6074 |
| 小区号(Cell Id) | 67FF510 |
| 信令类型 | 12 |

根据文献[@]中的算法，将4G手机信令数据处理得到各路段的车速和相对流量值，示例数据如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 公路标号 | 时间戳 | 路段编号 | 方向 | 车速 | 相对流量 |
| S35 | 2016/10/01 00:05:00 | 4 | 1 | 71.6 | 3 |

#### 数据预处理

选定一路段，将该路段一天之内的流量数据可视化，数据有明显的抖动，为了增加模型稳定性，对数据进行滤波预处理。利用滑动平均滤波后，效果如下：

#### 实验设置与结果

本实验以公路G1501为例训练模型并预测，对模型中所涉及的不同变量，如训练数据时间跨度，预测步长，高峰与非高峰，各正则项参数等进行多组实验，并分别将实验结果与滑动平均所得预测值进行对比，评估本模型的有效性。

为了全面评估实验效果，我们采用平均绝对百分比误差(Mean Absolute Percent Error,MAPE)、平均绝对误差(Mean Absolute Error,MAE)、均方根误差(Root Mean Square Error, RMSE)、归一化的平均绝对误差(Normalized Mean Absolute Error,NMAE)衡量，计算公式如下：

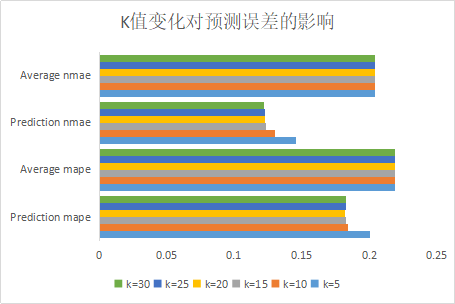
|  |  |
| --- | --- |
| 评估指标 | 公式 |
| 平均绝对误差(MAE) |  |
| 平均绝对百分比误差(MAPE) |  |
| 均方根误差(RMSE) |  |
| 归一化的平均绝对误差(NMAE) |  |

##### 参数敏感性测试

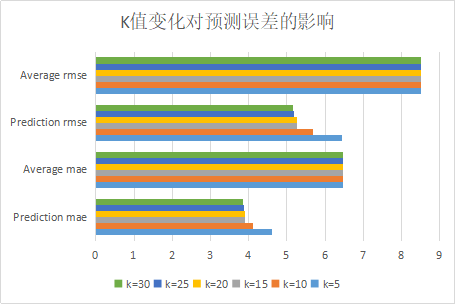
本模型中涉及等常数项，为了探讨各个常数项的取值对模型效果的影响，设计了三组对比实验。本模型中采用随机值初始化等变量，为了避免不同随机值给实验造成的误差，需要固定初始随机值。此外，我们假设对模型的影响是独立的，对比实验将其中两个变量设置为默认值，观察另一个变量对模型的影响。对比实验各参数设置情况如下：

|  |  |  |
| --- | --- | --- |
| 参数 | 数值设置 | 默认值 |
|  | 5,10,15,20,25,30 | 20 |
|  | 0.5,1,2,4,8,16 | 1 |
|  | 0.5,1,2,4,8,16 | 1 |

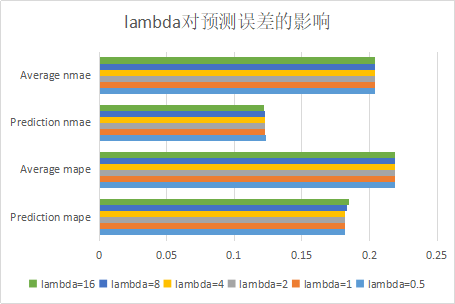
实验结果如下：



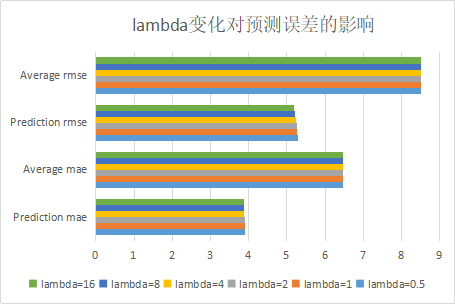
k值变化对预测误差的影响1



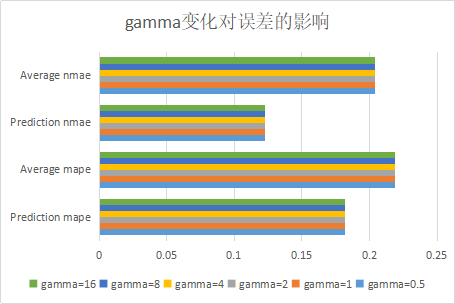
k值变化对预测误差的影响2



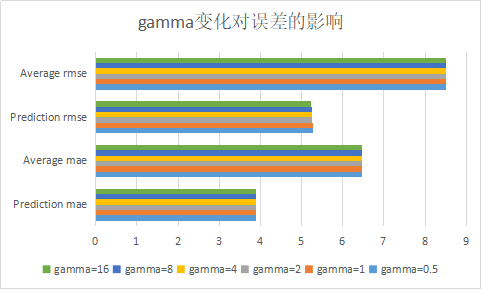
lambda值变化对预测误差的影响1



lambda值变化对预测误差的影响2



gamma值变化对预测误差的影响1



gamma值变化对预测误差的影响2

由图 **???** **???** 可观察到随着隐空间维度值的增大，预测误差逐渐减小，这是因为隐空间维度越高，对原始数据的还原能力越高，损失信息越少，因而更容易通过训练数据学习到流量变化的规律。由图 **???** **???** 可观察到，正则项的变化对结果影响非常小，同理由图 **???** , **???** 可观察到正则项对结果影响也非常小。根据对参数

##### 训练数据时间跨度

训练数据时间跨度会对隐空间特征的学习产生影响，如根据1小时前的数据预测与根据1天前的数据预测结果会不同，因此本组实验设置训练数据的时间跨度 分别为，实验结果如下：

|  |  |  |
| --- | --- | --- |
| 参数 | LSM-RN模型MAPE | 滑动平均MAPE |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

（实验结果分析：待补充）

##### 预测步长

本实验设置预测步长，对模型分别进行训练和测试，实验结果如下：

|  |  |  |
| --- | --- | --- |
| 步长 | LSM-RN模型MAPE | 滑动平均MAPE |
|  |  |  |
|  |  |  |
|  |  |  |

(实验结果分析：待补充。步长增多带来的累积误差影响)

##### 高峰期和非高峰期

（这块儿还得再详细改改） 路段流量变化具有周期性，高峰期和非高峰期的变化规律有差别，如：上午8：00-10：00是高峰期，晚上则流量稀少为非高峰期，此外假期期间的流量与平时比也会有异常。本实验分别选取高峰期和非高峰期的流量数据进行分析，测试本模型在不同情况下的性能。实验结果如下：

|  |  |  |
| --- | --- | --- |
| 时段 | LSM-RN模型MAPE | 滑动平均MAPE |
|  |  |  |
|  |  |  |

参考文献

[1] PAN B, DEMIRYUREK U, SHAHABI C. Utilizing real-world transportation data for accurate traffic prediction[C]//Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012: 595–604.

[2] RISTANOSKI G, LIU W, BAILEY J. Time series forecasting using distribution enhanced linear regression[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2013: 484–495.

[3] ZHOU J, TUNG A K. Smiler: A semi-lazy time series prediction system for sensors[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015: 1871–1886.

[4] KWON J, MURPHY K. Modeling freeway traffic with coupled HMMs[R]. Technical report, Univ. California, Berkeley, 2000.

[5] WANG F, LI T, WANG X, 等. Community discovery using nonnegative matrix factorization[J]. Data Mining and Knowledge Discovery, Springer, 2011, 22(3): 493–521.

[6] SAHA A, SINDHWANI V. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization[C]//Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012: 693–702.

[7] DENG D, SHAHABI C, DEMIRYUREK U, 等. Latent space model for road networks to predict time-varying traffic[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 1525–1534.

[8] LEE D D, SEUNG H S. Algorithms for non-negative matrix factorization[C]//Advances in neural information processing systems. 2001: 556–562.

[9] CAI D, HE X, HAN J, 等. Graph regularized nonnegative matrix factorization for data representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2011, 33(8): 1548–1560.

[10] ZHU L, GALSTYAN A, CHENG J, 等. Tripartite graph clustering for dynamic sentiment analysis on social media[C]//Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014: 1531–1542.