DOI:10.13718/j.cnki.xsxb.2017.05.003

基于非负矩阵分解的 OD 矩阵预测①

张竣伟, 高 超, 张自力

西南大学 计算机与信息科学学院,重庆 400715

摘要:提出了一种非负矩阵分解-自回归模型,并用该模型对居民出行流量进行预测. 该模型首先利用非负矩阵分解方法挖掘城市区域内的居民出行特征,而后在非负矩阵分解获得的特征矩阵和系数矩阵基础上对时序系数矩阵建立自回归模型,进而对起讫矩阵进行预测. 以北京市出租车数据为基础,与时空权重 K 近邻、传统 K 近邻、反向神经网络、朴素贝叶斯、随机森林和 C4.5 决策树回归模型对比,实验结果表明,该模型的预测准确率有显著提升.

关 键 词:OD 矩阵;非负矩阵分解;自回归;区域流量预测

中图分类号:TP311 **文献标志码**:A **文章编号**:1000-5471(2017)05-0017-05

在智能交通系统的调度管理中,起讫(Origin-Destination, OD)矩阵提供了居民从一个区域到另一个区域的出行次数,亦反映了车辆出行流量等信息. OD 矩阵是交通规划、设计和管理的关键输入信息[1-2],而实时的、精确的、可靠的 OD 矩阵预测,可为交通使用者提供可靠的出行信息,帮助交通管理者优化交通信号和应急调度[3]. 因此,OD 矩阵预测问题成为交通领域日益关注的重点.

近年来,针对 OD 矩阵预测问题,已经提出许多模型和方法. 鉴于不同的 OD 数据获取方法,OD 矩阵预测模型大致可分为两类:静态 OD 矩阵预测和动态 OD 矩阵预测. 由于传统的 OD 数据来源于耗时耗力的统计调查方法,同时也丧失了信息的时效性,因此使得 OD 信息研究局限于静态模型的分析[4]. 静态 OD 矩阵预测模型中最近的研究报告表明,多种多样的模型已经被提出,包括信息最小化、熵最大化、最大似然法、贝叶斯推理和广义最小二乘[3]. 智能交通系统的发展为获取实时有效的交通信息提供了帮助,进而使得动态 OD 矩阵成为智能交通预测模型的主要输入之一. 基于动态 OD 矩阵的时序特征和稀疏特性,Berlaire 提出了一种扩展的最小二乘法线性模型,并且 Barceló 提出卡尔曼滤波模型[5-6]. 随着城市 GPS 轨迹数据规模的爆炸式增长,高维度的 OD 矩阵预测问题成为交通领域日益关注的热点. 虽然 Djukic 用主成分分析方法将高维 OD 矩阵转化为低维空间上的 OD 矩阵,且提出了有色噪声的卡尔曼滤波模型[7],但是主成分分析法会使得分解还原后的 OD 矩阵出现负数现象,并且减少了 OD 矩阵的初始信息,违背了 OD 矩阵本身的实际物理意义,即要求出行次数非负.

综上所述,这些 OD 矩阵预测方法虽然为道路使用者提供实时的交通流量信息,但是无法满足交通行业从事者的载客需求. 因此,鉴于 OD 矩阵的非负物理意义,本文提出了一种基于非负矩阵分解(nonnegative matrix factorization, NMF)算法的自回归(auto regression, AR)模型对 OD 矩阵进行预测.

1 NMF-AR 模型

为了对未来几个小时或某个时段的居民出行流量进行预测,本文提出了基于非负矩阵分解的自回归预

① 收稿日期:2017-01-16
基金项目:国家自然科学基金项目(61403315, 61402379).
作者简介:张竣伟(1992-),男,重庆北碚人,硕士研究生,主要从事大数据研究.
通信作者:张自力,教授.

测模型. 首先，预处理从智能交通系统获取的出租车GPS轨迹数据，并构建每天各个时段的OD矩阵；其次，利用矩阵变形规则构建具有时序特征的出行信息矩阵，并用NMF算法挖掘信息矩阵中城市居民出行特征，提取基底矩阵和具有时序特征的系数矩阵；然后，针对时序系数矩阵建立自回归模型，预测某个时段的系数向量；最后，利用NMF算法的分解还原原理建立预测函数，并获取该时段的OD矩阵预测结果.

1.1 GPS数据预处理

从智能交通系统获取的出租车GPS轨迹数据，由于信息量大、数量多且冗余信息量大，需要对原始出租车GPS轨迹数据做预处理工作. 基于部分轨迹数据字段(触发事件、GPS经纬度和GPS时间)进行合理的预处理，本文构建每天各个时段的OD矩阵，并定义第 i 天第 j 时段的OD矩阵为 $\boldsymbol{X}_j^i=[x_{i'j'}]_{n\times n}$，其中 $i=1,2,\cdots,m$ ，$j=1,2,\cdots,h$. 预处理包含以下3个关键步骤：① 划分每天出租车上下车点作为城市居民出行OD信息；② 基于时间的交通管制信息制定时段划分策略；③ 基于制定的划分策略，构建每天各个时段的OD矩阵. 基于上述不同划分策略得到的各类OD矩阵，是本文进一步分析和挖掘居民出行规律的数据支撑，且在各OD矩阵数据的非负特性和区域间出行次数的非负物理意义基础上，本文利用非负矩阵分解算法分析和挖掘居民出行意愿特征.

1.2 非负矩阵分解算法

日常交通系统的各个时段内，居民的出行状况有着不同的数据特征. 例如，在工作日会出现早晚出行高峰. 因此，本文将从数据整体特征上分析和挖掘居民的基本出行模式且进一步获取基底特征和系数变化，描述居民的出行规律. 首先，在数据预处理部分获取的城市居民出行OD矩阵 $\boldsymbol{X}_j^i$ 和矩阵变形方法的基础上，构建初始信息矩阵 $\boldsymbol{S}=[s_{ij}]_{n^2\times(m*h+h)}$，且矩阵变形方法如公式(1)所示. 因此，初始信息矩阵可用公式(2)描述，其反映居民的实时出行量，为挖掘出行特征提供了数据基础.

$$\boldsymbol{s}_{i*h+j}=\begin{pmatrix}(\boldsymbol{x}_j^i(1))^{\mathrm{T}}\\(\boldsymbol{x}_j^i(2))^{\mathrm{T}}\\\vdots\\(\boldsymbol{x}_j^i(n))^{\mathrm{T}}\end{pmatrix} \tag{1}$$

$$\boldsymbol{S}_{n^2\times(m*h+h)}=(\boldsymbol{s}_1\cdots\boldsymbol{s}_{i*h+j}\cdots\boldsymbol{s}_{m*h+h}) \tag{2}$$

$$\boldsymbol{S}=\boldsymbol{BP} \tag{3}$$

而后，在初始出行信息中挖掘出行特征基底矩阵 $\boldsymbol{B}=[b_{ij}]_{n^2\times k}(i=1,2,\cdots,n^2,\ j=1,2,\cdots,k)$ 和出行时序系数矩阵 $\boldsymbol{P}=[p_{ij}]_{k\times(m*h+h)}(i=1,2,\cdots,k,\ j=1,2,\cdots,m*h+h)$ 来刻画居民不同时段和区域的出行规律，其形式化表述如公式(3)所示，其中 $\boldsymbol{B},\boldsymbol{P}$ 都是未知的. 基于矩阵 $\boldsymbol{B}$ 和 $\boldsymbol{P}$ 的非负物理意义，本文采用NMF($\boldsymbol{S}$, k)算法对初始出行信息矩阵进行特征挖掘和分解. 在已知初始信息矩阵 $\boldsymbol{S}$ 和正整数 $k<\min\{m,n\}$ 的条件下，求解非负的两个未知矩阵 $\boldsymbol{B}$ 和 $\boldsymbol{P}$ 的问题可转为求解非负矩阵分解最小化问题[8]，并且该最小化求解问题形式化表述如公式(4)所示.

$$f(\boldsymbol{B},\boldsymbol{P})=\arg\min_{\{\boldsymbol{B},\boldsymbol{P}\}}\|\boldsymbol{S}-\boldsymbol{BP}\|^2 \tag{4}$$

求解该最小化问题的算法包括以下3个步骤：① 初始化非负矩阵 $\boldsymbol{B}_{n\times j}$ 和 $\boldsymbol{P}_{j\times m}$，并保证矩阵在迭代中的非负性；② 根据公式(5)所示成本函数与公式(6)和(7)所示乘性更新法则对 $\boldsymbol{B},\boldsymbol{P}$ 进行更新；③ 最终迭代求解公式(4)的最小值. 若求解值是最小值，则结束迭代；否则，重复步骤②.

$$\|\boldsymbol{S}-\boldsymbol{BP}\|^2=\sum_{i,j}(\boldsymbol{S}_{ij}-(\boldsymbol{BP})_{ij})^2 \tag{5}$$

$$\boldsymbol{B}_{n\times j}\sim\boldsymbol{B}_{n\times j}\frac{(\boldsymbol{SP}^{\mathrm{T}})_{n\times j}}{(\boldsymbol{BPP}^{\mathrm{T}})_{n\times j}} \tag{6}$$

$$\boldsymbol{P}_{j\times m}\sim\boldsymbol{P}_{j\times m}\frac{(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{S})_{j\times m}}{(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{BP})_{j\times m}} \tag{7}$$

下文将对基于上述分解算法得到的系数矩阵 $\boldsymbol{P}=[p_{ij}]_{k\times(m*h+h)}$ 进行时间序列分析.

1.3　自回归模型

基于上文 NMF($\boldsymbol{S}$, k) 算法分解得到的时序系数矩阵 $\boldsymbol{P}$，本文针对该矩阵的时序特征建立自回归模型 AR(λ)，并提出预测函数对居民出行次数进行预测. 首先，基于自回归基础模型如公式(8) 所示，

$$\boldsymbol{z}_t = \boldsymbol{\psi}_t * \boldsymbol{H}_t + a_t \tag{8}$$

同时鉴于时序系数矩阵 $\boldsymbol{P} = [p_{ij}]_{k\times(m*h+h)}$ 中 $k \geqslant 1$，本文分别对 k 维系数矩阵分别建立如公式(9) 所示的自回归模型，

$$\begin{pmatrix} p_{1t} \\ p_{2t} \\ \vdots \\ p_{it} \\ \vdots \\ p_{kt} \end{pmatrix} = \begin{pmatrix} \varphi_{11}p_{1t-1} + \varphi_{12}p_{1t-2} + \cdots + \varphi_{1\lambda}p_{1t-\lambda} + a_{1t} \\ \varphi_{21}p_{2t-1} + \varphi_{22}p_{2t-2} + \cdots + \varphi_{2\lambda}p_{2t-\lambda} + a_{2t} \\ \vdots \\ \varphi_{i1}p_{it-1} + \varphi_{i2}p_{it-2} + \cdots + \varphi_{i\lambda}p_{it-\lambda} + a_{it} \\ \vdots \\ \varphi_{k1}p_{kt-1} + \varphi_{k2}p_{kt-2} + \cdots + \varphi_{k\lambda}p_{kt-\lambda} + a_{kt} \end{pmatrix} \tag{9}$$

并且根据自回归模型的预测方法建立第 $t+1$ 时段的自回归预测方法如公式(10) 所示

$$\begin{pmatrix} p_{1t+1} \\ p_{2t+1} \\ \vdots \\ p_{it+1} \\ \vdots \\ p_{kt+1} \end{pmatrix} = \begin{pmatrix} \varphi_{11}p_{1t} + \varphi_{12}p_{1t-1} + \cdots + \varphi_{1\lambda}p_{1t-\lambda+1} + a_{1t+1} \\ \varphi_{21}p_{2t} + \varphi_{22}p_{2t-1} + \cdots + \varphi_{2\lambda}p_{2t-\lambda+1} + a_{2t+1} \\ \vdots \\ \varphi_{i1}p_{it} + \varphi_{i2}p_{it-1} + \cdots + \varphi_{i\lambda}p_{it-\lambda+1} + a_{it+1} \\ \vdots \\ \varphi_{k1}p_{kt} + \varphi_{k2}p_{kt-1} + \cdots + \varphi_{k\lambda}p_{kt-\lambda+1} + a_{kt+1} \end{pmatrix} \tag{10}$$

其中：$\boldsymbol{z}_t$ 表示第 t 时刻的观测值；$\boldsymbol{H}_t$ 表示第 t 时刻 λ 阶自变量，即第 $t-1$ 时刻到第 $t-\lambda$ 时刻的观测变量；$\boldsymbol{\psi}_t$ 表示第 t 时刻观测值的系数，即第 $t-1$ 时刻到第 $t-\lambda$ 时刻的系数；a_t 是服从正态分布的常数.

通过公式(10) 计算得到第 $t+1$ 时段的系数向量 $\boldsymbol{P}_{t+1}$，且基于 NMF($\boldsymbol{S}$, k) 算法得到基底矩阵 $\boldsymbol{B}$，从而得到第 $t+1$ 时段的 OD 矩阵 $\boldsymbol{S}_{n^2\times(t+1)}$，其预测函数如公式(11) 所示

$$\boldsymbol{S}_{n^2\times(t+1)} = \boldsymbol{B}\boldsymbol{P}_{t+1} \tag{11}$$

最后，通过上述 NMF-AR 模型得到对某时段的 OD 矩阵预测值.

2　实验分析

基于“数据堂”(http：//www.datatang.com/data/44502)公开的北京市 12000 辆出租车在 2012 年 11 月内的 GPS 轨迹数据和北京市早晚高峰限行通知(http://www.bjjtgl.gov.cn/zhuanti/20140328wr.html)时间段划分构建的 OD 矩阵数据，本文将 NMF-AR 模型与其他 6 种熟知的预测模型在最佳参数情况下进行预测精度对比，包括：时空权重 K 近邻(Spatial-Temporal Weighted K-Nearest Neighbor，SWT-KNN(21，0.4，0.4，0.9))、传统 K 近邻(K-Nearest Neighbor，KNN(15))、反向神经网络(Back Propagation Neural Network，BP(2，10，10))、朴素贝叶斯(Naïve Bayesian，NB)、随机森林(Random Forest，RF(50))和 C4.5 决策树回归模型[9]. 并且本文利用文献[9]中提出的 MOEs 测试指标集合进行评估，其对比结果见图 1，其中(a)，(b)，(c)，(d)分别表示 MOEs 评估指标集合的 MAPE，RMSE，MAE，ME. 由于箱线图存在最大值和最小值的随机性，准确率的比较主要基于分位数的对比，其中箱线图的框底和顶部表示$\frac{1}{4}$分位数和$\frac{3}{4}$分位数，框内的线段表示中位数，且框内的小正方形环表示平均值. 在关键评估指标 MAPE 上，NMF-AR 模型的第一分位数、第三分位数、中位数和平均值都比其他算法的指标低，且该模型的箱线框也比其他模型的框短，且 MAPE 指标值平均分布于 3%～20%的范围内，且图中 NMF-AR 模型的最大值和最小值也比其他模型的低. 在图 1 的(b)和(c)中，NMF-AR 模型的 RMSE 和 MAE 指标值的各个方面的值都比其他模型的值低. 因此，实验结果表明 NMF-AR 模型具有较好的预测能力.

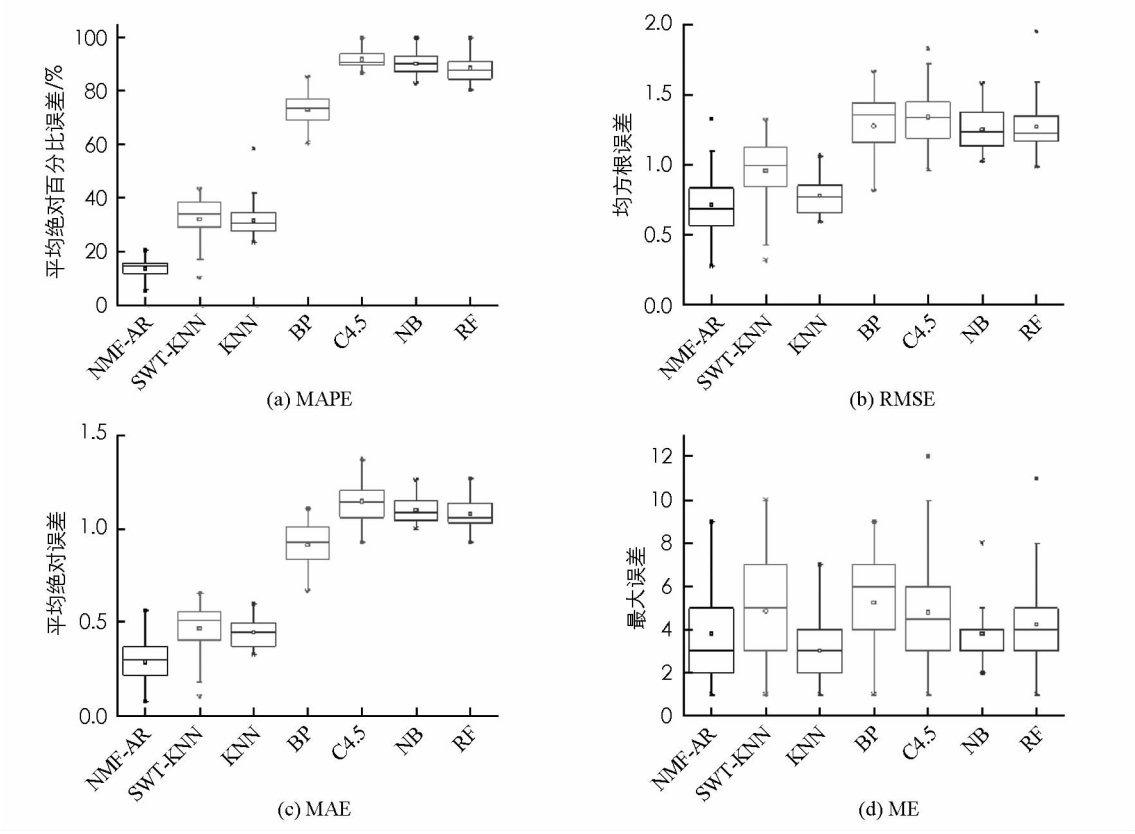

图1 各模型OD矩阵预测评估指标平均值对比

3 结 论

基于道路使用者的需求分析，本文利用OD数据对居民出行进行预测分析. 而且针对OD矩阵预测问题，本文将短时交通流的非参数化模型，包括SWT-KNN,KNN,BP,NB,RF和C4.5，应用于该问题求解，并与本文提出的NMF-AR模型进行实验对比，结果表明，本文提出的NMF-AR模型比其他模型具有更好的预测能力，并能为交通使用者提供有效、实时的居民出行信息，帮助交通使用者提高运营效率.

参考文献：

[1] TOLEDO T, KOLECHKINA T. Estimation of Dynamic Origin-Destination Matrices Using Linear Assignment Matrix Approximations [J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(2): 618－626.

[2] LOU Y, YIN Y. A Decomposition Scheme for Estimating Dynamic Origin-Destination Flows on Actuation-Controlled Signalized Arterials [J]. Transportation Research Part C Emerging Technologies, 2010, 18(5): 643－655.

[3] CHENG Y, YE X, WANG Z. A Forecasting Model of the Proportion of Peak-Period Boardings for Urban Mass Transit System: A Case Study of Osaka Prefecture [C]//Transportation Research Board 95th Annual Meeting. Washington D C: Transportation Research Board, 2016.

[4] TANAKA M, KIMATA T, ARAI T. Estimation of Passenger Origin-Destination Matrices and Efficiency Evaluation of Public Transportation [C]// Iiai International Congress on Advanced Applied Informatics. New York: IEEE Press, 2016: 1146－1150.

[5] BIERLAIRE M, CRITTIN F. An Efficient Algorithm for Real-Time Estimation and Prediction of Dynamic OD Tables [J]. Operations Research, 2004, 52(1): 116－127.

[6] BUGEDA J B, MERCADÉ L M, MARQUÉS L, et al. A Kalman-Filter Approach for Dynamic OD Estimation in Corridors Based on Bluetooth and Wi-Fi Data Collection [J]. Annals of Botany, 2010, 103(2): 377－386.

[7] DJUKIC T, FLOTTEROD G, VAN LINT H, et al. Efficient Real Time OD Matrix Estimation Based on Principal Com-

ponent Analysis [C]//International IEEE Conference on Intelligent Transportation Systems. New York: IEEE Press, 2012: 115－121.

[8] BERRY M W, BROWN M, LANGVILL A N, et al. Algorithms and Applications for Approximate Nonnegative Matrix Factorization [J]. Computational Statistics & Data Analysis, 2007, 52(1): 155－173.

[9] XIA D, WANG B, LI H, et al. A Distributed Spatial-Temporal Weighted Model on MapReduce for Short-Term Traffic Flow Forecasting [J]. Neurocomputing, 2015, 179: 246－263.

On Estimation of OD Matrix Based on Nonnegative Matrix Factorization

ZHANG Jun-wei, GAO Chao, ZHANG Zi-li

School of Computer and Information Science, Southwest University, Chongqing 400715, China

Abstract: In this paper, a simple and effective model has been proposed for predicting the Origin-Destination matrix, which combines the Nonnegative Matrix Factorization algorithm and the Autoregressive model, named NMF-AR. In details, the NMF algorithm has been first used to extract the base patterns of the resident trip characteristics. Then the AR model is applied to model the nonlinear time series coefficient matrix based on the results returned by the NMF algorithm. Finally, based on the taxi GPS data in Beijing, the proposed model has been compared with some famous predicting models, including Spatial-Temporal Weighted K-Nearest Neighbor (SWT-KNN), conventional K-Nearest Neighbor (KNN), Back Propagation Neural Network (BP), Naïve Bayesian (NB), Random Forest (RF) and C4.5. And these experimental results show that the predictive performance of the NMF-AR model is better than these of the compared models under the typical conditions.

Key words: OD matrix; nonnegative matrix factorization; autoregressive; taxi GPS data

责任编辑　张　枸