基于特征蒸馏的变分编码器交通流预测模型

欧阳毅,汤文燕,黎晏伶
(浙江工商大学管理科学与电子商务学院,浙江杭州 310000)

摘 要: 针对交通流数据高维非线性和时空依赖性复杂,本文构建了基于特征蒸馏的变分贝叶斯编码器交通流预测模型. 对每段时间序列对应的时间窗口特征,构建了基于多模态时间槽和空间槽的交通流特征提取模型. 以时空槽特征提取模型作为特征知识蒸馏架构的输入. 通过知识蒸馏结构提取的时空特征结晶体,利用教师模型指导学生模型的学习过程,从而提高学生模型的泛化能力. 变分贝叶斯编码器对交通流时空特征结晶编码获取交通流数据的隐变量,根据隐变量的生成采样,利用解码器将其解码重构成新的预测值. 实验结果表明,本文提出的模型预测性能显著提升,且中长期预测中鲁棒性更优.

关键词: 特征蒸馏;多模态时间槽;空间槽;变分贝叶斯;生成式模型;变分推断

基金项目: 浙江工商大学“数字+学科建设项目”(No.SZJ2022C004);浙江工商大学 2023 年度省级及以上教学平台自主设立校级教学项目(No.1310XJ0521036)

中图分类号: TN911.7;U491.1 **文献标识码:** A **文章编号:** 0372-2112(2024)06-1938-07

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230230

Traffic Flow Prediction Model Based on Spatio-Temporal Feature Distillation Variational Autoencoder

OUYANG Yi, TANG Wen-yan, LI Yan-ling
(*School of Management Science and E-Commerce*, *Zhejiang Gongshang University*, *Hangzhou*, *Zhejiang* 310000, *China*)

Abstract: To improve the accuracy of traffic flow prediction and to solve the problems of high-dimensional nonlinearity and spatio-temporal dependence of traffic flow, a combined feature distillation and variational Bayes encoders traffic flow forecasting model (ST-DVBE) is proposed. First, to extract the time window characteristics corresponding to each time series, the multi-modal time slots and spatial slots are constructed. Second, with spatio-temporal slot feature extraction model as the input of feature knowledge distillation architecture, and space-time feature crystallization extracted by knowledge distillation structure, the learning process of student model is guided by teacher model, so as to improve the generalization ability of student model. Finally, the variational Bayesian encoder is employed to capture the latent variables of traffic flow data by encoding the crystallization of spatiotemporal features. Utilizing the generated latent variables, the decoder reconstructs them into new predicted values. Experimental results demonstrate a significant enhancement in predictive performance with the proposed model, especially with better robustness in mid- and long-term forecasting.

Key words: feature distillation; multimodal temporal slots; spatial slots; variational Bayes; generative model; variational inference

Foundation Item(s): Zhejiang Gongshang University “Digital+Discipline Construction Project” (No.SZJ2022C004); Zhejiang Gongshang University 2023 Provincial-level Teaching Platform Teaching Project (No.1310XJ0521036)

1 引言

交通流预测[1~27]是智能交通管理系统的核心任务,通过分析历史特征预测未来交通流信息,使交通管理部门能够及时采取信号控制和调控措施,以提高交通效率、缓解拥堵、增强道路安全,最终推动智慧交通系统的发展. 然而,这些大规模的交通数据呈现复杂的非线性多模态特性,交通数据的特征提取方式也会对模型的精确度产生影响. 此外,实际模型的部署对轻量化

收稿日期:2023-03-10;修回日期:2023-12-06;责任编辑:李勇锋

的要求更高,如何将训练模型更具实用性是一个具有挑战性的问题.

早期交通流预测模型依赖于时间序列等统计模型[2],这类模型缺乏对交通流数据复杂变化模式建模,预测准确性具有局限性. 之后,传感器技术进步推动了决策树[3]、随机森林[4]及支持向量机[5]等机器学习分类算法的发展,从而被用于捕捉交通和气象等数据的非线性关系. 但交通流数据的多样性和动态性导致该模型参数调整需要人工干预,限制了模型的灵活性. 为了提高预测模型的泛化能力. 随着人工智能技术的发展,当前研究正利用神经网络相关技术解决交通流预测问题,如图卷积神经网络(Graph Convolutional Network, GCN)[6],图注意力网络(Graph Attention Network, GAT)[7],长短期记忆神经网络(Long Short-Term Memory, LSTM)[8]和基于门控递归单元(Gate Recurrent Unit, GRU)[9]等技术的使用. 虽然神经网络相较于早期模型,预测性能已有显著提升,但不同传感器、分辨率和来源渠道的交通流数据具有多模态性,全局预测并不完善. 研究将交通数据的时间特征和空间特征进行融合,能够增强模型对不同数据源的泛化能力,同时提高模型对噪声或异常数据的鲁棒性.

交通预测模型最终需要应用于实际生活,而实际部署交通预测模型需要考虑低延迟、高效率、低计算资源消耗等特性. 与训练模型相比,部署模型面临资源限制,因此模型通常需要压缩. 为了不损失模型的预测性能,本文结合知识蒸馏的思想对预测模型进行轻量化升级.

知识蒸馏(Knowledge Distillation, KD)[10]核心思想是将性能较好的教师模型中无监督信息,通过蒸馏的方式,提取训练小模型. 主要包含3要素:知识、蒸馏算法和师生模型架构. 目前知识蒸馏算法包括对抗知识蒸馏算法[11]、多教师蒸馏[12]、多模态知识蒸馏[13]以及量化蒸馏[14]等. 考虑到每种蒸馏方式有其独特的适用场景,本文结合交通流的时空特征,构建基于时空槽的特征的时空特征知识蒸馏结构,在此基础上提出了一种基于时空槽特征的变分贝叶斯(Variational Bayes Encoders, VBE)推断思想.

图1为变分贝叶斯编码器交通流预测模型(Spatio-Temporal Distillation Variational Bayesian Encoder for Traffic Flow Forecasting Model, ST-DVBE)的系统框图,X为交通流历史数据,Y为预测交通流数据,Local STD为时空特征蒸馏模块,该模块的输出为时空特征结晶STD,经过变分贝叶斯编码器输出隐变量,再经由解码器和全链接层生成预测数据. 预测数据和标签数据进行损失计算以更新模型参数. 另一方面在获取时空特征结晶时,也利用标签数据对教师模型和学生模型进行训练学习. 教师模型具有较深层数,指导学生模型进行进一步优化. 在预测过程中,仅使用学生模型产生时空特征结晶,该模型相较于教师模型可防止过拟合,且更具鲁棒性.

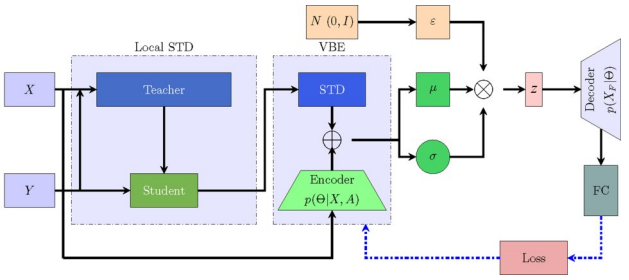

图1　ST-DVBE结构框架图

2　时空特征蒸馏

交通路网可定义为无向图网络$G=(V,E)$,其中$(v_i,v_j)\in V$代表由N个车辆检测器组成的节点集合,$e_{v_i,v_j}\in E$表示检测器节点v_i和v_j的连边. 在t时刻所观察的交通状态可以表示为$D\in \mathrm{R}^{N\times M}$,其中,$M$表示每个车辆检测器节点的交通特征维度. 交通流预测问题就是根据前T个历史时间步长的观测数据$D(t)=\{d^{t-T+1},d^{t-T+2},\cdots,d^{t}\}$,使用相应的学习函数$f$,预测未来$\tilde{T}$个时间步长的交通信息$\tilde{D}(t)=\{d^{t+1},d^{t+2},\cdots,d^{t+\tilde{T}}\}$,如式(1)所示:

$$\tilde{D}(t)=f(D(t)) \tag{1}$$

2.1　多模态时间槽特征提取单元

定义$\hat{T}=\{t_1,t_2,\cdots,t_{\hat{T}}\}$个交通流时间序列向量均有相应的交通特征矩阵$\boldsymbol{d}^{t_i}$,因此将交通流时序数据定义. 采用滑动窗口按交叉间隔Δt构造时间槽,本文的多模态时间槽如式(2):

$$\boldsymbol{T}^{u}_{\text{slots}}=\begin{bmatrix}\boldsymbol{d}^{T_1} & \boldsymbol{d}^{T_1+\Delta t} & \cdots & \boldsymbol{d}^{T_1+\eta\Delta t}\\ \boldsymbol{d}^{T_2} & \boldsymbol{d}^{T_2+\Delta t} & \cdots & \boldsymbol{d}^{T_2+\eta\Delta t}\\ & & \ddots & \\ \boldsymbol{d}^{T_u} & \boldsymbol{d}^{T_u+\Delta t} & \cdots & \boldsymbol{d}^{T_u+\eta\Delta t}\end{bmatrix} \tag{2}$$

其中,$\eta=\left\lfloor\dfrac{D_T}{\Delta t}\right\rfloor$为数据集中时间槽的样本个数,$u$为时间槽数(若采用周时间槽,数字1~7则表示是周一至周日时间子槽,$\boldsymbol{d}^{T_1}$表示数据集中周一的第一组数据). 采用多层感知器(Multilayer Perceptron, MLP)对时间槽输入特征进行建模,以此获取交通流节点时间特征信息. 对于每层神经元的输出相应如式(3)~(6)定义:

$$\boldsymbol{h}_j^1=\sigma(\boldsymbol{w}_1\boldsymbol{T}_{\text{slot}}+\boldsymbol{w}_2\boldsymbol{x}) \tag{3}$$

$$\boldsymbol{h}_j^2=\sigma(\boldsymbol{w}_{jj'}\boldsymbol{h}_1+\boldsymbol{w}_{jj'}\boldsymbol{h}_2+\cdots+\boldsymbol{w}_{jj'}\boldsymbol{h}_m) \tag{4}$$

$$\boldsymbol{y}_k=\boldsymbol{w}_{j'k}\boldsymbol{h}_{1'}+\boldsymbol{w}_{j'k}\boldsymbol{h}_{2'}+\cdots+\boldsymbol{w}_{j'k}\boldsymbol{h}_{m'} \tag{5}$$

$$\boldsymbol{X}_T=\Big\|\left(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_k\right) \tag{6}$$

式中$\boldsymbol{x}$和$\boldsymbol{T}_{\text{slot}}$代表时间槽的交通特征和相对应的星期特征，$\boldsymbol{w}_1$和$\boldsymbol{w}_2$分别为两个全连接层的神经网络权重，$\boldsymbol{w}_{ij}$、$\boldsymbol{w}_{jj'}$和$\boldsymbol{w}_{j'k}$是MLP的神经元权重．$\boldsymbol{h}_j^1$、$\boldsymbol{h}_{j'}^2$是第一层隐藏层和第二层隐藏层的输出，$\boldsymbol{y}_k$是输出层的输出值，通过‖进行拼接之后可得到具有时间特性的新交通特征值，σ为ReLU非线性激活函数．

2.2 空间槽特征提取单元

空间槽的输入向量定义为：$\boldsymbol{X}_S=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$，其中$\boldsymbol{x}_i\in\mathbf{R}^{B\times N\times T\times M}$，$B$为Batch维度，$N$为传感器维度，$T$为时间窗口维度，$M$为特征维度．空间槽提取单元的输出向量为$\bar{\boldsymbol{x}}_i\in\mathbf{R}^{B\times N\times T\times F'}$，$F'$是节点特征提取输出维度．采用8个注意力头的双层GAT实现空间槽的交通特征提取，其注意力系数计算过程为

$$\alpha_{v_i,v_j}=\vec{a}[(\boldsymbol{W}_{v_i}\times\boldsymbol{W}_{v_j}^{\mathrm{T}})\odot\boldsymbol{A}] \tag{7}$$

$$\alpha_{ij}=\frac{\exp(\mathrm{LeakyReLU}(\vec{a}[(\boldsymbol{W}_{v_i}\times\boldsymbol{W}_{v_j}^{\mathrm{T}})\odot\boldsymbol{A}]))}{\sum\limits_{n\in N_i}\exp(\mathrm{LeakyReLU}(\vec{a}[(\boldsymbol{W}_{v_i}\times\boldsymbol{W}_{v_n}^{\mathrm{T}})\odot\boldsymbol{A}]))} \tag{8}$$

式中，$\boldsymbol{A}$是归一化的邻接矩阵，$\boldsymbol{W}_{v_i}$是经权值矩阵线性变换后节点v_i的特征向量，T指转置操作，$\vec{a}$是一个作为掩码的单层神经网络．×和⊙分别为矩阵相乘和哈达马积（即矩阵对应位置相乘）．N_i表示节点v_i的邻域．模块中每组注意力机制的计算和空间槽特征向量输出分别如式(9)和式(10)所示：

$$\vec{\boldsymbol{x}}'_{v_i}=S(\sum_{j\in N_i}\alpha_{ij}^c\boldsymbol{W}_{v_j}^c+b) \tag{9}$$

$$\boldsymbol{X}_S=S(\|_{c=1}^{C}(\sum_{j\in N_i}\alpha_{ij}^c\boldsymbol{W}_{v_j}^c+b)) \tag{10}$$

其中，S为softmax函数，b为偏置项，c为注意力机制头的数量，α_{ij}^c表示第c组注意力机制产生的注意力系数，$\boldsymbol{W}_{v_j}^c$为第c组注意力机制所具有的权值矩阵，‖表示拼接．

2.3 基于时空槽的特征嵌入和蒸馏

交通流时间特征$\boldsymbol{X}_T$和空间特征$\boldsymbol{X}_S$经过拼接形成初融合特征$\boldsymbol{X}_F$，将$t-1$时刻的初融合特征$\boldsymbol{X}_F^{t-1}$和邻接矩阵$\boldsymbol{A}$共同作为时空特征融合模块的输入，得到包含上一时刻交通流信息状态的当前时刻新交通特征$\bar{\boldsymbol{X}}_F^t$．时空特征融合过程的数学表达如式(11)~(13)：

$$\bar{\boldsymbol{X}}_F^t=\otimes(\boldsymbol{X}_F^t)\oplus\mathrm{sig}(\otimes(\boldsymbol{X}_F^{t-1})) \tag{11}$$

$$\hat{\boldsymbol{X}}_F^t=\otimes(\boldsymbol{X}_F^t) \tag{12}$$

$$\bar{\boldsymbol{X}}_F=\mathrm{BN}(\boldsymbol{\sigma}(\bar{\boldsymbol{X}}_F^t\oplus\hat{\boldsymbol{X}}_F^t)) \tag{13}$$

式中，sig代表Sigmoid激活函数，BN是批次归一化．⊗为神经网络二维卷积操作，$\hat{\boldsymbol{X}}_F^t$表示经过卷积的原始的t时刻初融合特征$\boldsymbol{X}_F^t$．

特征蒸馏部分学生模型采用LeakyReLU(0.1)为激活函数和MLP相同深度的时空特征提取模块层数，而教师模型的深度是学生模型的5倍．教师模型的复杂结构可提供更强的表示能力，从而更能学习到训练数据中的多模态非线性特性．而学生模型的简化结构有助于降低过拟合风险，使其更容易泛化到整体数据空间．将t时刻的交通流分别送入教师模型和学生模型，可得到两个模型的输出值．本文定义教师模型的输出值称为时空结晶特征，通过该特征来监督学生模型的蒸馏特征．将这个输出值与真实标签进行损失计算，以最佳损失的输出值作为教师模型软结晶体的标签，从而调试教师模型的参数．性能最优的参数产生的时空结晶特征则作为学生模型的监督信息．根据软结晶体和学生模型的蒸馏信息之间的损失差异对学生模型进行梯度更新，从而得到具有更高性能和泛化性能更优的学生模型，具体蒸馏过程如图2.

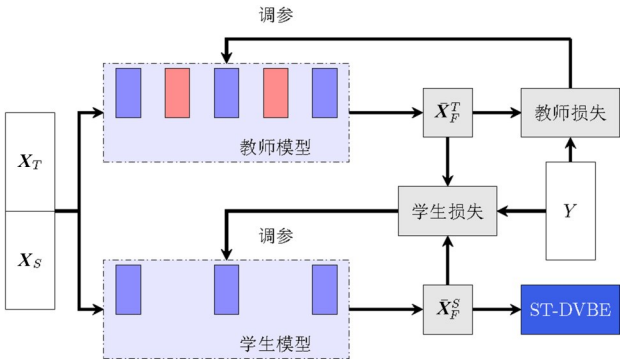

图2 时空特征蒸馏流程图

3 基于时空特征蒸馏变分自编码器的交通流预测模型

3.1 交通流概率建模问题定义

本文提出一种深度生成式模型ST-DVBE，采用时空蒸馏特征$\boldsymbol{X}_F$，学习数据$\boldsymbol{X}$和隐变量$z\in\mathrm{R}^M$的联合概率分布．交通预测数据$\hat{y}$可使用隐变量z通过产生器生成．因此目标函数为最大化预测数据在隐变量z分布下的边缘或然率的期望．对于给定的观察数据x，使用隐变量的后验概率$q_\phi(z|x)$来逼近$p_\theta(z)p_\theta(y|z)$，使用参数ϕ来建模x和z之间的非线性联系．令隐变量的先验概率为高斯分布$p(z)=N(0,I)$，可得有约束的优化问题，如式(14)：

$$\begin{aligned}&\max_{\varphi,\theta}\ E_{x\sim D}[E_{q_\phi(z|x)}[\log p_\theta(x|z)]],\\&\text{s.t.}\quad D_{\mathrm{KL}}(q_\phi(z|x)\|p(z))<\delta\end{aligned} \tag{14}$$

在满足KKT条件下，该式的拉格朗日对偶式为式(15)：

$$L(\theta,\varphi,\lambda;\ x,z)=E_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]-\lambda(D_{\mathrm{KL}}(q\|p(z))-\delta) \tag{15}$$

因此可以得到下界 $\underline{L}=\ell(\theta,\varphi,\lambda;\ x,z)<L(\theta,\varphi,\lambda;\ x,z)$

$$\underline{L}=E_{q_\phi(z|x)}[\log p_\theta(x|z)]-\lambda D_{\mathrm{KL}}(q\|p(z)) \tag{16}$$

$$q^*(\Theta)=\arg\min_q \mathrm{KL}(q(\Theta)\|p(\Theta|y)) \tag{17}$$

对于 $q^*(\Theta)$ 分布，设其服从均值场，则有 $q(\Theta)=\prod_{i=1}^{M}q_i(\Theta_i)_i$. 其中每个成分通过最小化KL差异获取，可得如下的因式分解形式为式(18)：

$$q(\Theta)=q_X(x)q_Z(z)q_\Gamma(\gamma) \tag{18}$$

根据贝叶斯公式，未知参数的后验概率分布 $\Theta=(x,z,\gamma)$ 在给定观察数据 y 条件下，可定义为式(19)：

$$p(\Theta|y)\propto p(y|x,z)p(x|\gamma)p(z|\pi)p(\gamma) \tag{19}$$

本文给出一种基于通过提升变分下界来近似求解数据分布的变分贝叶斯推断的方法，定义为式(20).

$$p(\Theta|y)\geqslant F(\Theta|y;\ \lambda) \tag{20}$$

变分贝叶斯(Variational Bayes，VB)是用于估计各种概率模型参数的通用统计推断方法，结合自编码器和概率图模型思想的生成式模型. 将VBE引入交通流预测模型，训练过程中会采用变分推断的方式衡量计算模型的最大化似然函数，逼近真实观测数据的后验分布. 设交通流观测数据，其涵盖了某路网区域 T 段时间内的车流量、车速及占空比等路况信息. 交通流预测未来 $\tilde{T}$ 段的交通流信息为式(21)：

$$\arg\max_{\Theta^{T+i}}\sum_{i=1}^{\tilde{T}}\log p(\tilde{D}^{T+i}|D^T) \tag{21}$$

交通流预测目标是根据给定的历史步长为 T 的观测数据情况下，找出一组参数 Θ^{T+i} 使得观测数据 $\boldsymbol{d}^T$ 出现的概率越大.

3.2　基于时空特征蒸馏变分自编码器的交通流预测模型

面对高维的交通流数据，每个时间节点的交通流量或速度数据往往都具有其独特复杂性的分布，针对这些高维数据的隐变量分布建模通常比较困难，并且计算其真实的后验分布时可能伴随着复杂的高维积分或优化过程. 因此，VBE[15]在其编码器结构中引入适当的近似后验分布 $q(\Theta|X)$，以变分推断的方式学习原数据的隐变量的真实后验分布 $P(\Theta|X)$. 在本文构建的时空槽特征编码器结构中，采用了三层全连接层构成的MLP对输入特征编码学习. 每个全连接层的神经元均和上一层的所有神经元相连接，每条连接均赋予一个权重. 并且每层的节点输入都是上一层节点的输出. 另外，神经网络对近似后验分布 $q(\Theta|X)$ 建模产生的参数为 φ(为简化符号，后文用 Q_φ 表示 $q_\phi(\Theta|X)$). 为了使 $\int Q_\varphi \mathrm{d}\Theta=1$ 和 $P(\Theta|X)$ 分布之间的差异最小化，采用KL散度量化VBE的两种分布，KL散度的定义如式(22)所示：

$$D_{\mathrm{KL}}(Q_\varphi\|p(\Theta|X))=\int Q_\varphi\log\frac{Q_\varphi}{p(\Theta|X)}\mathrm{d}\Theta \tag{22}$$

通过联合概率公式 $p(\Theta,X)=p(\Theta|X)p(\Theta)$、积分 $\int Q_\varphi \mathrm{d}\Theta=1$ 和整理符号，对式(23)进行如下变换：

$$D_{\mathrm{KL}}(Q_\phi\|p(\Theta|X))=\log p(X)+D_{\mathrm{KL}}(Q_\phi\|p(\Theta))-E_{\Theta\sim Q_\phi}\log p(X|\Theta) \tag{23}$$

$$D_{\mathrm{KL}}(Q_\phi\|p(\Theta|X))=\log p(X)-\underline{L} \tag{24}$$

其中，$\underline{L}=-D_{\mathrm{KL}}(Q_\phi\|p(\Theta))+E_{\Theta\sim Q_\phi}\log p(X|\Theta)$，VBE的目标是最小化 Q_φ 和 $p(\Theta|X)$ 分布，即最小化 $D_{\mathrm{KL}}(Q_\varphi\|p(\Theta|X))$，那么根据式(24)，即最大化 $\underline{L}$. 因为KL散度是非负数，由此可得到式(25)：

$$\log p(X)\geqslant\underline{L} \tag{25}$$

$\underline{L}$ 即为 $\log p(X)$ 的下界，也是目标函数的证据下界(Evidence Lower BOund，ELBO). 最小化散度 $D_{\mathrm{KL}}(Q_f\|p(\Theta|X))$，即等价于最大化ELBO. 以往ELBO的固定权重系数倾向于生成较为平滑的数据，限制了生成样本的多样性和隐式特征的连续性. 本文在ELBO引入一个可调节的时空特征因子 λ，用来平衡重建项和KL散度项. 当因子 $\lambda=1$ 时，时空槽特征编码器就和最初的VBE结构一致(实际上是将原问题转换为拉格朗日对偶问题)；当 $\lambda>1$ 时，就会增加KL散度项在损失函数的重要性，增加隐变量的约束强度，促使其更加稳定；当 $\lambda<1$ 时，就会提高重构项在损失函数中的影响力，能够促使生成样本更加多样化. 因此本模型的ELBO公式为

$$\arg\max_{\phi,\Theta}\{E_{\Theta\sim q_\phi(\Theta|\bar{X}_F,A)}\log p(\bar{\boldsymbol{X}}_F,\boldsymbol{A}|\Theta)-\lambda D_{\mathrm{KL}}(q_\phi(\Theta|\bar{\boldsymbol{X}}_F,\boldsymbol{A})\|p(\Theta))\} \tag{26}$$

其中，$E_{\Theta\sim q_\phi(\Theta|\bar{X}_F,A)}\log p(\bar{\boldsymbol{X}}_F,\boldsymbol{A}|\Theta)$ 是解码器的重构损失，$D_{\mathrm{KL}}(q_\phi(\Theta|\bar{\boldsymbol{X}}_F,\boldsymbol{A})\|p(\Theta))$ 是编码器的隐变量分布和隐变量先验之间的损失，隐变量先验假定为是服从 $N(0,I)$ 的正态分布. 最大化ELBO，则隐变量分布和隐变量先验就需要尽可能靠近. 对于ELBO的重构项，本文采用均方误差度量预测值和观测值之间的损失差异. 最终目标损失函数如式(27)：

$$\mathrm{Loss}=\mathrm{MSE}+\lambda D_{\mathrm{KL}}(q_\phi(\Theta|\bar{\boldsymbol{X}}_F,\boldsymbol{A})\|p(\Theta)) \tag{27}$$

4　实验分析

本文使用PeMS-BAY、PeMS04和PeMS08三个公开数据集，分别验证ST-DVBE模型在车速和流量预测方

面的有效性. 通过与多个基准模型的量化评价指标进行对比,以证明ST-DVBE模型的性能优越性. 数据集按6:2:2的比例划分为训练、验证和测试集.

4.1 评价指标与参数设置

本文ST-DVBE及基准模型的预测性能用平均绝对误差(Mean Absolute Error, MAE)、均方根误差(Root Mean Squared Error, RMSE)和平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)进行评估[17,26,27]. 在Python3.6版本PyTorch1.8的环境下完成本文预测模型实验. 由于交通流的采集时间是5分钟/次,本模型的历史时间步长为12,表示历史1小时的数据量. 未来时间步长分别设置为3、6和12,预测未来15、30和60分钟的数据结果. 时空特征融合模块和时空槽变分预测模块均采用Adam优化器,学习率分别为1×10^{-4}和1×10^{-3}. 每次送入模型的批次量$B=128$. 训练的epoch为300.

4.2 实验对比分析

首先,PeMS-BAY车速数据集被用于进行15、30和60 min不同时间步长的ST-DVBE模型训练和预测. 在模型迭代中,关注了隐变量Z和均值μ的演变. 图3呈现了15 min车速预测训练过程中均值μ的热度图,选择前325行和前32列进行性能比较,其中325是传感器个数,32是数据记录行数. 观察到左边在100次迭代后,μ值的分布相对随机且差异较大;而右边经过1 000次迭代后,μ值整体趋于平滑. 因此,通过多次迭代均值μ优化,可以减少控制变量预测输出的差异,为模型性能提供改进的可能性.

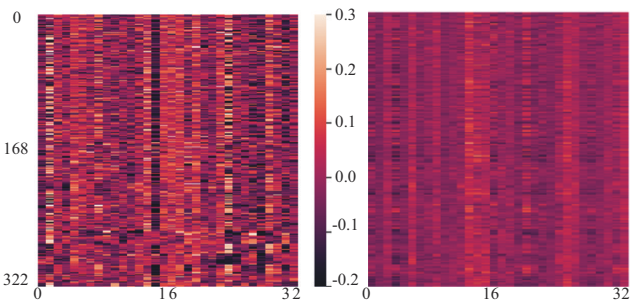

图3 针对15分钟交通车速预测的训练过程均值μ热图

本文还在PeMS-BAY上进行传感器预测消融实验. 将Batch设为64,观测单个传感器上两种MAE值的差异:仅经过Local STD处理的预测与真实样本的MAE,以及Local STD输出送往VBE后的预测与真实样本的MAE. 实验结果显示,VBE能有效维持误差在较小范围,对噪声数据有抑制作用.

图4展示了教师模型、学生模型和MLP分别进行时间特征提取时多轮迭代的MSE损失情况,3种模型曲线均是在LeakyReLU(0.1)为激活函数情况下产生的. 蓝线和红线分别表示教师模型与学生模型,而绿线是系统仅采用MLP进行时间相关性特征提取的MSE损失值情况. 从图中可以发现,MLP的收敛速度明显弱于师生模型,这是因为作为一种简单的神经网络结构,MLP对于复杂的时间序列模式的学习存在一定的局限. 而师生架构通常会结合更深层次的网络,具有更强的学习能力和表示能力,能更好地捕捉时间序列中的复杂模式和关联,因此收敛速度可能会更快. 学生模型在多轮迭代后,MSE存在波荡性,主要由于教师模型会比学生模型更复杂、更强大,能够更好地拟合训练数据,可能更容易更好地捕捉数据中的模式和关联,从而在预测过程中更加稳定. 为减少学生模型的波荡性,后续可增加训练数据量、改善参数初始化、调整模型复杂度、调整损失函数或者通过模型集成等方式来提升预测稳定性.

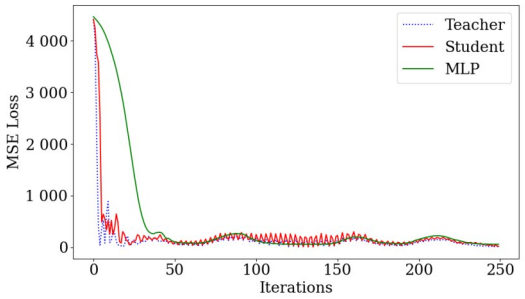

图4 LeakyReLU激活函数下特征蒸馏消融性实验

表1展示了各模型在PeMS-BAY数据集中速度预测对比情况. 从表中可以看出ST-DVBE模型的在预测步长为15 min的预测中,MAE、RMSE和MAPE三种度量指标数值略优于较多基准模型. 在30 min和60 min的预测中,三种度量指标也明显优于其他模型. 此外,所有度量指标都保持在一个较小的级别. 这不仅说明ST-DVBE具有较好的稳健性,也体现了该模型在中长期数据预测方面的优势.

本文利用PeMS04和PeMS08数据集中的流量特征,利用60 min历史数据,对未来15 min预测实验,以验证ST-DVBE模型的泛化性能. 表2中的数据表明,传统SVR性能较差,而FC-LSTM虽有改进,但未处理空间特征,尚有提升空间. STG2Seq和DCRNN考虑了时空相关性,因此在性能上取得了更显著的进展. ASTGCN和Graph WaveNet进一步改进了时空特征提取的方法. ST-DVBE模型综合考虑了时空特征蒸馏和时间周期性,在各项性能指标上表现更出色,尤其是在PeMS04数据集中RMSE值上. 其优势在于变分自适应编码器的解码器模块,能够更好地拟合历史趋势,因此对异常值更具鲁棒性. 具体预测结果详见表2.

表 1　PeMS-BAY 速度预测对比　　单位:mile/h

Methods	15 min			30 min			60 min		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
ARMIA[2]	1.62	3.30	3.50%	2.33	4.76	5.40%	3.38	6.50	8.30%
STGCN[16]	1.36	2.96	2.90%	1.81	4.27	4.17%	2.49	5.69	5.79%
DCRNN[17]	1.38	2.95	2.90%	1.74	3.97	3.90%	2.07	4.74	4.90%
GMAN[18]	1.34	2.82	2.81%	1.62	3.72	3.63%	1.86	4.32	4.31%
GWNET-COV[19]	1.30	2.73	2.69%	1.62	3.67	3.59%	1.91	4.40	4.47%
STTN[20]	1.36	1.67	1.95%	2.89	3.78	4.58%	2.87	3.79	4.50%
AGC-net[21]	1.18	2.31	2.30%	1.48	3.14	3.10%	1.85	3.90	4.20%
ASTTN[22]	1.32	2.70	2.78%	1.58	3.72	3.64%	1.72	4.02	3.98%
ST-DVBE	**1.29**	**2.30**	**2.52%**	**1.42**	**2.79**	**2.78%**	**1.84**	**3.68**	**3.76%**

表 2　PeMS04 和 PeMS08 数据集流量预测对比

Methods	Pems04			Pems08		
	MAE/(车辆数/15 min)	RMSE/(车辆数/15 min)	MAPE/%	MAE/(车辆数/15 min)	RMSE/(车辆数/15 min)	MAPE/%
SVR[23]	28.70	44.56	19.20	23.25	36.16	14.64
FC-LSTM[24]	27.14	41.59	18.20	22.20	34.06	14.20
STG2Seq[25]	25.20	38.48	18.77	20.61	31.23	18.32
DCRNN[17]	24.70	38.12	17.12	18.29	27.45	11.45
ASTGCN[26]	22.93	35.22	16.56	18.61	28.16	13.08
Graph WaveNet[27]	25.45	39.70	17.29	19.13	31.05	12.68
ST-DVBE	**22.36**	**34.46**	**16.72**	**19.03**	**28.45**	**15.92**

5　总结

本研究提出了基于时空槽的交通流特征提取蒸馏模型(ST-DVBE),通过多模态时间槽和空间槽的特征提取构建,以增强模型泛化性能. 在特征蒸馏中,以时空槽特征提取模型为主导,通过知识蒸馏提高学生模型的泛化能力. 虽然实验证明ST-DVBE在交通数据预测上性能显著提升,但未来研究需考虑引入复杂因素如突发事件、事故、天气,同时改进时间依赖性提取,考虑采用内部记忆机制的深度学习模型. 知识蒸馏将继续在未来研究中发挥关键作用,推动模型进一步改进.

参考文献

[1] 马君, 刘小冬, 孟颖. 基于神经网络的城市交通流预测研究[J]. 电子学报, 2009, 37(5): 1092-1094.
MA J, LIU X D, MENG Y. Research of urban traffic flow forecasting based on neural network[J]. Acta Electronica Sinica, 2009, 37(5): 1092-1094. (in Chinese)

[2] MAKRIDAKIS S, HIBON M. ARMA models and the box-Jenkins methodology[J]. Journal of Forecasting, 1997, 16 (3): 147-163.

[3] MIRCHANDANI P, HEAD L. A real-time traffic signal control system: Architecture, algorithms, and analysis[J]. Transportation Research Part C: Emerging Technologies, 2001, 9(6): 415-432.

[4] LESHEM G, RITOV Y. Traffic flow prediction using adaboost algorithm with random forests as a weak learner[J]. International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering, 2007, 1(1): 1-6.

[5] TANG J, CHEN X, HU Z, et al. Traffic flow prediction based on combination of support vector machine and data denoising schemes[J]. Physica A: Statistical Mechanics and Its Applications, 2019, 534: 120642.

[6] 金苍宏, 董腾然, 陈天翼, 等. 融合序列分解与时空卷积的时序预测算法[J]. 电子学报, 2021, 49(2): 233-238.
JIN C H, DONG T R, CHEN T Y, et al. Spatio-temporal convolutional forecasting based on time-series decomposition strategy[J]. Acta Electronica Sinica, 2021, 49(2): 233-238. (in Chinese)

[7] 徐兴荣, 刘聪, 李婷, 等. 基于双向准循环神经网络和注意力机制的业务流程剩余时间预测方法[J]. 电子学报, 2022, 50 (8): 1975-1984.
XU X R, LIU C, LI T, et al. Business process remaining time prediction: An approach based on bidirectional quasi recurrent neural network with attention[J]. Acta Electronica

Sinica, 2022, 50(8): 1975-1984. (in Chinese)
[8] LU H, GE Z, SONG Y, et al. A temporal-aware LSTM enhanced by loss-switch mechanism for traffic flow forecasting[J]. Neurocomputing, 2021, 427: 169-178.
[9] HUSSAIN B, AFZAL M K, AHMAD S, et al. Intelligent traffic flow prediction using optimized GRU model[J]. IEEE Access, 2021, 9: 100736-100746.
[10] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. [2023]. http://arxiv.org/abs/1503.02531.pdf.
[11] 钱亚冠, 马骏, 何念念, 等. 面向边缘智能的两阶段对抗知识迁移方法[J]. 软件学报, 2022, 33(12): 4504-4516.
QIAN Y G, MA J, HE N N, et al. Two-stage adversarial knowledge transfer method for edge intelligence [J]. Journal of Software, 2022, 33(12): 4504-4516. (in Chinese)
[12] LIU Y, ZHANG W, WANG J. Adaptive multi-teacher multi-level knowledge distillation[J]. Neurocomputing, 2020, 415: 106-113.
[13] DAI W, HOU L, SHANG L, et al. Enabling multimodal generation on CLIP via vision-language knowledge distillation[C]//Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg: ACL, 2022: 2383-2395.
[14] HUANG Y, HAO Y, XU J, et al. Compressing speaker extraction model with ultra-low precision quantization and knowledge distillation[J]. Neural Networks, 2022, 154: 13-21.
[15] KINGMA D P, WELLING M. Auto-encoding variational Bayes[C]//International Conference on Learning Representations (ICLR2014). Canada: ICLR. 2014:1-14.
[16] YU B, YIN H, ZHU Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York: ACM, 2018: 3634-3640.
[17] LI Y, YU R, SHAHABI C, et al. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting [EB/OL]. [2023]. http://arxiv.org/abs/1707.01926.pdf.
[18] ZHENG C, FAN X, WANG C, et al. GMAN: A graph multi-attention network for traffic prediction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 1234-1241.
[19] YOO B, LEE J, JU J, et al. Conditional temporal neural processes with covariance loss[C]//International Conference on Machine Learning. PMLR, 2021: 12051-12061.
[20] XU M, DAI W, LIU C, et al. Spatial-temporal transformer networks for traffic flow forecasting[EB/OL]. [2023]. http://arxiv.org/abs/2001.02908.pdf.
[21] LI Z, LI W, HWANG K. Adaptive graph convolution networks for traffic flow forecasting[EB/OL]. [2023]. http://arxiv.org/abs/2307.05517.pdf.
[22] FENG A, TASSIULAS L. Adaptive graph spatial-temporal transformer network for traffic flow forecasting[EB/OL]. [2023]. http://arxiv.org/abs/2207.05064.pdf.
[23] WILLIAMS B M, DURVASULA P K, BROWN D E. Urban freeway traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing models[J]. Transportation Research Record, 1998, 1644(1): 132-141.
[24] LIU B, TANG X, et al. Traffic flow combination forecasting method based on improved LSTM and ARIMA[J]. International Journal of Embedded Systems, 2020, 12(1): 22-30.
[25] BAI L, YAO L, KANHERE S, et al. Stg2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. New York: ACM, 2019:1981-1987.
[26] BAI L, YAO L, LI C, et al. Adaptive graph convolutional recurrent network for traffic forecasting[J]. Advances in Neural Information Processing Systems, 2020, 33(1): 17804-17815.
[27] WU Z, PAN S, LONG G, et al. Graph wavenet for deep spatial-temporal graph modeling[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. New York: ACM, 2019: 1907-1913.

作者简介

欧阳毅 男,1975年10月出生于贵州省贵阳市. 现为浙江工商大学管理工程与电子商务学院副教授、硕导. 2005年在浙江大学计算机科学与技术学院获工学博士学位,2018年在美国Oakland大学作访问学者,获Academic EXCELLCE奖项. 在国内外发表学术论文30余篇.
E-mail: oyy@mail.zjgsu.edu.cn

汤文燕 女,1998年7月13日出生于江苏省泰州市兴化市. 浙江工商大学管理科学与电子商务学院硕士研究生. 研究方向为智慧物流与智能决策.
E-mail: twy15298547169@163.com