* 文章编号:2096-398X(2020)05-0165-08

基于卷积网络与支持向量机的云资源预测模型

杨 云, 闫振国

(陕西科技大学 电子信息与人工智能学院, 陕西 西安 710021)

摘 要:云原生容器生态系统的快速发展,推动了更多应用程序在云端落地.容器云作为承载业务的分布式系统支撑平台,需要进行及时准确的资源分配与调度管理.为了提升容器云面对负载变化的弹性应对能力,提出一种基于卷积网络与支持向量机的云资源预测模型,依据需求预测为资源管理提供前置响应.模型利用卷积网络深度捕获时序数据的特征信息,通过遗传算法与组合核函数优化支持向量回归进行预测.在 Google 云计算中心数据集的实验表明,该模型的预测精度与稳定性优于现有云资源预测方法.

关键词:云资源预测; 卷积网络; 支持向量机

中图分类号:TP311 **文献标志码**: A

DOI:10.19481/j.cnki.issn2096-398x.2020.05.026

Cloud resource prediction model based on convolutional network and support vector machine

YANG Yun, YAN Zhen-guo

(School of Electronic Information and Artificial Intelligence, Shaanxi University of Science & Technology, Xi'an 710021, China)

Abstract: The rapid development of the cloud-native container ecosystem has driven more applications to land on the cloud. As a distributed system support platform for carrying services, container cloud needs timely and accurate resource allocation and scheduling management. In order to improve the resilience of the container cloud in response to load changes, a cloud resource prediction model based on convolutional networks and support vector machines is proposed to provide a pre-response for resource management based on demand prediction. The model uses convolutional networks to deeply capture the feature information of time series data, and uses genetic algorithms and combined kernel functions to optimize support vector regression for prediction. Experiments on the Google Cloud Computing Center data set show that the prediction accuracy and stability of the model are better than the existing cloud resource prediction methods.

Key words: cloud resource forecast; convolutional network; support vector machines

* 收稿日期:2020-04-27

基金项目:国家自然科学基金项目(61601271); 陕西省科技厅重点研发计划项目(202021711); 陕西省教育厅专项科研计划项目(15JK1086)

作者简介:杨 云(1965-),女,山东青岛人,教授,博士,研究方向:智能信息处理、数据挖掘

0 引言

云原生容器技术的生态系统正在快速发展，以 Kubernetes 为代表的容器云成为新的分布式架构解决方案.容器云封装整个软件运行时环境，为开发者提供用于构建、发布和运行分布式应用的平台[1].云资源的调度管理与部署是影响系统性能的重要因素.资源供应不足会导致服务水平协议(Service-Level Agreement，SLA)违约和服务质量(Quality of Service，Qos)下降，而过度供应又会带来资源浪费，增加网络、冷却和维护成本[2].高效的资源管理需要与应用需求动态贴合.但是新资源从请求、调度、部署到启动存在一定的时间延迟，这意味着，当系统达到某个负载阈值才请求新资源的反应性技术，在业务繁忙期会增大系统运行压力.而应对高流量的临时系统扩容又是必不可少的，包括系统规格扩增、服务扩展以及后端扩容.所以在系统过载之前预测未来时间窗口的云资源请求，提前进行调度分配与编排部署是应对该问题的有效解决办法.

云资源需求的波动不是一个随机游走过程，而是前后关联的.其相似的形态模式随业务规律会差异性复现.因为云资源序列与时间的高度相关，现有研究将其作为时序问题开展分析.针对云资源预测，早期的方法有基于统计学的自回归移动平均法(ARIMA)[3]、指数平滑法[4]等.ARIMA 要求时序数据经过差分化之后是稳定的，本质上只能捕捉线性关系.指数平滑法对数据进行非等权处理，给予近期数据较大权值，但对转折点缺乏鉴别能力，长期效果较差.后来传统机器学习方法得到了广泛发展，Zia 等[5]利用自回归神经网络(AR-NN)组合模型预测实时资源使用情况；Jitendra Kumar 等[6]提出神经网络与自适应差分进化的结合方法，在精度上优于反向传播网络；Gopal 等[7]利用贝叶斯模型预测内存密集型应用的资源需求.赵莉[8]采用支持向量机(Support Vector Machine，SVM)结合混沌分析方法对云资源序列进行处理，仿真实验对比神经网络大幅提高了预测精度；Wei 等[9]将小波变换与支持向量机的优点结合，提出一种基于加权小波支持向量机的云负载预测模型，为不同样本赋予不同等级权重，同时利用改进的粒子群算法优化参数组合，进一步提升了预测效率.以上模型中神经网络具备良好的非线性映射能力，但随复杂度的提高阈值和权重参数成倍增加，训练结果容易过拟合或陷入局部最优，贝叶斯模型只适合特定预测场景.而相比其他学习算法，支持向量机同时考虑经验风险和结构风险最小化，使用核方法进行非线性学习，其作为二次凸优化问题同时避免了神经网络的局部最优缺陷，取得了更为理想的效果.

随着深度学习在时序预测领域的发展[10]，研究者开始利用卷积神经网络(Convolutional Neural Networks，CNN)、循环神经网络(Recurrent Neural Network，RNN)及其变体长短时记忆网络(Long Short-Term Memory，LSTM)等深度模型对序列进行处理.Yonghua Zhu 等[11]使用循环神经网络提出基于注意力机制的 LSTM 云负载预测方法；Omer 等[12]使用卷积神经网络(CNN)将时间序列转换为二维图像，根据原始时间序列起伏标注图像特征；Alberto Mozo 等[13]利用 CNN 预测本地数据中心网络流量负载的短期变化，证实了 CNN 可以捕捉具有高度非线性规律的 TCP 连接数量；S Chan 等[14]使用 CNN-SVM 的混合建模技术对非平稳的多元时间序列进行了短期负荷预测，相比传统时序方法在精度上取得了显著优势.

在云场景中，资源需求波动是一个逐渐演变的过程，存在时间维上局部相关性的同时，受到空间维拓扑结构的影响，各结点的负载波动会链式带动其他结点的计算资源管理弹性变化[15].例如 Kubernetes 云平台依据服务发布 Pod 的负载情况动态调整外部存储对于业务的扩缩容响应.卷积神经网络并行化效率高，结构稳定，能够从局部特征关联聚合得到整体，从各维度挖掘潜在模式，相比其他方法更能满足云场景的时空维建模需求.现有方法中，通常使用单一预测模型或者是使用组合模型对不同预测结果进行权重集成，虽然有一定性能提升，但并未根本解决较差模型的短板影响.基于此，本文提出一种使用卷积网络与支持向量机相结合的预测算法，将卷积网络的特征学习能力与支持向量机的回归拟合能力结合，利用图卷积建模拓扑结构的空间维特征，同时利用遗传算法改善支持向量机的参数优化.从特征建模与拟合回归两方面提升现有方法预测能力.

1 问题定义

定义 1(拓扑映射)容器云平台应用集群的资源负载随业务逻辑调用及时序演变呈非线性动态变化，将结点间的拓扑关联映射为图结构数据，用无向图 $G=(V,E)$ 表示，V 是结点集，结点个数 $|V|=N$，E 为边集，代表结点间拓扑连接.

定义 2(时序数据)结点监测 M 个云资源指标(包含 CPU、Memory 等)的时间序列数据，即结点在特定时间切片产生 M 维特征向量.用三元组

$x(v,m,\tau)$表示结点 v 的 m 维指标在时刻 τ 的监测记录,其中 v 为结点标识,$m\in M$ 为特征标识,τ 为时间标识.用 $t=[t_{start},t_{end})$表示某个时间区间,$\Delta t=t_{end}-t_{start}$为区间长度,则在 t 时段内,所有结点的第 m 维指标监测序列记录表示为张量 $X^m=|\{x(v,m,\tau)|x.v\in N\wedge x.m=m\wedge x.\tau\in t\}|$.用 $X_\tau=|\{x(v,m,\tau)|x.v\in N\wedge x.m\in M\wedge x.\tau=\tau\}|$ 表示所有结点的所有监测指标在时刻 τ 的记录值.

定义 3(滑动窗口)利用滑动窗口采集历史时间序列记录,将 $moveGAP[t_{start},\Delta t]=X^m(t_{start}$ 为起始记录,Δt 为窗口长度)称为 m 维指标的 Δt 滑动采集窗口.

定义 4(问题定义)本文的预测任务为,依据应用集群历史时序片段的资源使用记录值,预测未来时间窗口 T_P 的资源使用需求.记当前时刻为 t_0,用 $X=(X_{t_0-T_h+1},X_{t_0-T_h+2},\cdots,X_{t_0})\in\mathbb{R}^{T_h\times N\times M}$ 表示所有结点在历史区间 T_h 的记录值,用 $Y=(X_{t_0+1},X_{t_0+2},\cdots,X_{t_0+T_p})\in\mathbb{R}^{T_p\times N\times M}$ 表示所有结点在未来区间 T_p 的待预测值,则本文的目标是学习如公式(1)所示的映射模型 $f(\cdot)$.

$$(X_{t_0-T_h+1},X_{t_0-T_h+2},\cdots,X_{t_0})\xrightarrow{f(\cdot)}(X_{t_0+1},X_{t_0+2},\cdots,X_{t_0+T_p}) \tag{1}$$

数据准备:云资源时序数据的待预测时段随集群业务负载变化而与其近期、日周期、周周期片段产生关联,滑动窗口依据规律特性从三个维度采集序列数据.假设结点监测的采样频率为每天 l 次,用 χ_r 表示从近期片段采集的时序数据,即与预测窗口直接相邻的时序片段,则有 $\chi_r=moveGAP[t_{start},\Delta t](t_{start}=t_0-T_r)$,其中 $\Delta t=T_r$ 为近期片段长度;用 χ_d 表示按日周期规律采集的时序数据,即与预测窗口相邻日期的同时段数据,则有 $\chi_d=moveGAP[t_{start},\Delta t](t_{start}=t_0-(T_d/T_p)\times l,t_0-(T_d/T_p-1)\times l,\cdots,t_0-l)$,其中 $\Delta t=T_p$,T_d 为日周期片段总长,采集步长为 l;用 χ_w 表示按周周期规律采集的时序数据,即在预测窗口相邻周内同星期且同时段属性的数据,则有 $\chi_w=moveGAP[t_{start},\Delta t](t_{start}=t_0-(T_w/T_p)\times 7\times l,t_0-(T_w/T_p-1)\times 7\times l,\cdots,t_0-7\times l)$,其中 $\Delta t=T_p$,T_w 为周周期片段总长,采集步长为 $7\times l$.T_r、T_d 及 T_w 均为预测窗口 T_p 的整数倍.将三个维度的采集片段作为原始输入,确保模型充分捕获时间维特征.则模型的输入数据 X 及历史区间 T_h 可由公式(2)、(3)表示.

$$T_h=T_r+T_d+T_w \tag{2}$$

$$X=[\chi_r,\chi_d,\chi_w]\in\mathbb{R}^{T_h\times N\times M} \tag{3}$$

2　卷积网络与支持向量机模型

2.1　卷积网络

本文通过图卷积建模云资源拓扑结构的空间维特征,刻画邻近结点的信息聚合,再沿时间轴卷积,捕获序列数据的时间维依赖,通过多个图卷积与时间维卷积的堆叠网络学习特征表示.

现有的图卷积神经网络主要有两类:谱图方法和空间方法.本文通过谱图方法在谱域定义图卷积.谱图方法概括说就是利用图的拉普拉斯矩阵的特征值和特征向量研究图的性质,通过对谱空间的信号做傅里叶变换实现卷积操作,而傅里叶变换的定义依赖于拉普拉斯矩阵[16].无向图 G 的拉普拉斯矩阵的一般定义为 $L=D-A$,其中 D 是顶点的度矩阵,为对角矩阵,对角线上的元素依次为各个顶点的度,A 是邻接矩阵.

本文采用对称规范化拉普拉斯矩阵,$L=I_N-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\in\mathbb{R}^{N\times N}$.图卷积的核心是基于拉普拉斯矩阵的谱分解,即特征分解.因为拉普拉斯矩阵是半正定对称矩阵,构成其的 n 个线性无关的特征向量相互正交,可作为构成空间的一组正交基,所以其线性组合可表示图 G 中的任意向量.将拉普拉斯矩阵进行谱分解形式为 $L=U\Lambda U^T$,$\Lambda=\mathrm{diag}([\lambda_0,\cdots,\lambda_{N-1}])\in\mathbb{R}^{N\times N}$ 是 L 的 n 个特征值构成的对角阵,特征向量矩阵 U 即是傅里叶基.

将图 G 中各结点的云资源数据 $x(v,m,\tau)$ 作为图信号进行处理,则对 x 进行傅里叶变换的矩阵形式为 $\hat{x}=U^Tx$,相应的,其傅里叶逆变换的形式为 $x=U\hat{x}$.因为卷积定理中函数卷积的傅里叶变换是函数傅里叶变换的乘积,由此推广到图结构,对于图信号的卷积操作是其与卷积核傅里叶变换乘积的逆变换.用卷积核 g_θ 对图 G 进行卷积操作[17]:

$$g_\theta*_Gx=g_\theta(L)x=g_\theta(U\Lambda U^T)x=Ug_\theta(\Lambda)U^Tx \tag{4}$$

$$y_{output}=\sigma(Ug_\theta(\Lambda)U^Tx) \tag{5}$$

其中,$g_\theta(\Lambda)=\mathrm{diag}([\Theta_0,\cdots,\Theta_{N-1}])$,卷积核参数通过初始化赋值然后利用反向传播进行调整.因为非线性变换在非欧空间数据的图结构中作用有限,使得图卷积操作发挥作用的是每一层的特征传播机制[18],所以本文放弃层之间的非线性变换,即将特征传播融合到一个层内,进行维度变换后再由激活函数作用在聚合结果上.$\sigma(\cdot)$ 是激活函数,本文使用线性修正单元 $ReLU$.y_{output} 即为卷积层输出,刻画邻近节点的信息聚合.但是由于每一次前向传播都要进行谱分解及计算大规模的矩阵乘积,代价较高,所以本文采用切比雪夫多项式进行近似

求解[17]：

$$g_\theta *_G x = g_\theta(L)x \approx \sum_{k=0}^{K-1}\theta_k T_k(\tilde{L})x \tag{6}$$

其中，$\tilde{L}=\frac{2}{\lambda_{\max}}L-I_N$，$\lambda_{\max}$ 即为 L 的最大特征值.切比雪夫多项式的递归定义为 $T_k(x)=2xT_{k-1}(x)-T_{k-2}(x)$，$T_0(x)=1$，$T_1(x)=x$.$K=3$ 即为对图 G 各结点提取融合其 0 ～ 2 阶邻居信息.上式卷积操作即为：

$$g_\theta *_G x \approx \theta_0 x+\theta_1(L-I_N)x+\theta_2(2(L-I_N)^2-I_N)x \tag{7}$$

简化计算以 $K=2$ 为例，卷积操作表示为：

$$g_\theta *_G x \approx \theta_0 x+\theta_1(L-I_N)x = \theta_0 x+\theta_1(-D^{-\frac{1}{2}}AD^{-\frac{1}{2}})x \tag{8}$$

因为建模图卷积的初衷是为了刻画结点各异的局部结构，而当 K 取值较大时，模型不满足局部性；而取值过小，两结点存在 K 跳不可达，此时 $L=0$，特征聚合点的信息更新不一定来自邻近节点.文献[19]依照主观经验的取参方式不具普适性，本文通过多轮取值校验对 K 进行全局选优.为了避免梯度消失，令 $\tilde{A}=A+I_N$，$\tilde{D}_{ii}=\sum_j \tilde{A}_{ij}$，并用 θ 替代 θ_0 与 $-\theta_1$ 共享权系数，卷积操作即得化简式[20]：

$$g_\theta *_G x = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}x\theta \tag{9}$$

其中，以 $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ 作为定义结点相关性的聚合函数，通过聚合函数定义结点特征传播.输出结果即为图中各结点在聚合函数作用下与邻居结点加权的新表达.

经过图卷积建立结点数据间的空间相关性，再由标准卷积提取隐藏的时间维特征信息，组成一个时间图卷积模块，由多个时间图卷积结构堆叠形成卷积网络层，如图 1 所示.

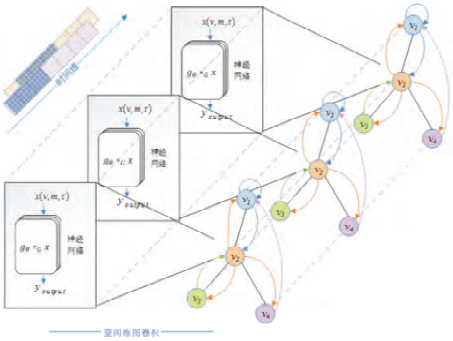

图 1 时间图卷积结构

以 $x\in\mathbb{R}^{N\times H_i\times T_i}$ 代表一层时间图卷积模块的输入，H_i 代表结点的输入特征维度，T_i 为输入时间维长度，用 $y'_{output}\in\mathbb{R}^{N\times H_o\times T_o}$ 表示该层时间图卷积的输出，也即为下一层的输入，H_o 与 T_o 分别为输出特征维度与时间维长度，则一个时间图卷积模块的形式化定义为：

$$y'_{output}=\sigma(\Gamma *(\mathrm{ReLU}(g_\theta *_G x))) \tag{10}$$

其中，$\sigma(\cdot)$ 是时间维卷积单元的激活函数，本文使用修正线性单元 $ReLU$.$\Gamma *$ 为时间维卷积操作.$g_\theta\in\mathbb{R}^{K\times H_i\times H_o}$ 为待学习的图卷积核参数.结点特征由 H_i 维转化到 H_o 维.时间维卷积示意图如图 2 所示.使用训练好的时间图卷积模型对原始输入数据进行特征提取，进行维度转换后从全连接层提取特征数据输入支持向量回归机(Support Vector Regression，SVR)完成回归训练.

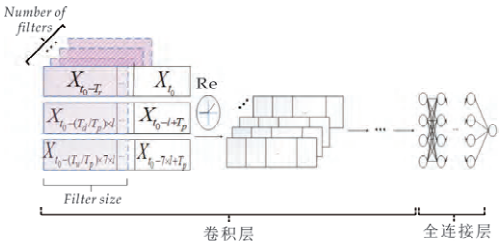

图 2 时间维卷积示意图

2.2 支持向量回归

将经过多层时间图卷积获得的特征向量送入支持向量回归机(SVR)训练，在过拟合时利用主成分分析降低特征向量维数.实验中发现特征提取相比 SVR 调参对预测结果的影响更大，而时间图卷积模块确保了特征提取不会成为提高准确率的瓶颈.

相比传统回归模型，SVR 以超出偏差间隔带 ϵ 的样本计算损失.引入间隔带两侧的松弛变量 ξ_i 和 $\hat{\xi}_i$ 替代 ϵ 不敏感损失函数，以表征该样本不满足约束的程度.求解问题可形式化定义为[21]：

$$\min_{\omega,b,\xi_i,\hat{\xi}_i}\ \frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}(\xi_i+\hat{\xi}_i) \tag{11}$$

$$\text{s.t.}\ \omega\cdot\varphi(x_i)+b_i-y_i\leqslant\epsilon+\xi_i,$$

$$y_i-(\omega\cdot\varphi(x_i)+b_i)\leqslant\epsilon+\hat{\xi}_i$$

$$\xi_i\geqslant 0,\hat{\xi}_i\geqslant 0,i=1,2,\cdots,n.$$

ω，b 是待确定的目标模型参数，分别为回归模型中的权重向量与偏置量，$\varphi(x)$ 为映射函数，C 为正则化常数，对结构风险和经验风险进行折中，$\omega\cdot\varphi(x_i)+b_i$ 为模型估计，y_i 为真实输出.面对该类二次规划问题，通过引入拉格朗日乘子求得上式对应拉格朗日函数，继而代入模型参数与松弛变量的偏导等式得其对偶问题，依据 Karush-Kuhn-Tucker 条件，解得拉格朗日乘子，即可通过下式解得模型偏置量 b，可取多个样本得其平均值：

$$b^*=y_i+\epsilon-\sum_{i=1}^{n}(\hat{\vartheta}_i-\vartheta_i)x_i^T x \tag{12}$$

$\vartheta_i\geqslant 0$，$\hat{\vartheta}_i\geqslant 0$ 为拉格朗日乘子.为避免计算高维向量内积的开销，引入核函数替代样本在高维

特征空间的内积操作.受文献[8]启发,单一核函数存在适用场景局限,此处选用在其他预测领域效果较为理想的高斯核(Radial Basis Function ,RBF)与多项式核(Polynomial kernel function ,POLY)组合核函数:

$$\kappa(x,x_i)=\mu\kappa_{RBF}(x,x_i)+(1-\mu)\kappa_{POLY}(x,x_i) \tag{13}$$

$$\kappa_{RBF}(x,x_i)=\exp\left(-\frac{\|x-x_i\|^2}{2\rho^2}\right) \tag{14}$$

$$\kappa_{POLY}(x,x_i)=(x^Tx_i)^d \tag{15}$$

其中,μ 为权值系数,$d\geqslant 1$ 为多项式次数,ρ 为带宽.SVR 的决策回归函数为:

$$f(x)=\sum_{i=1}^{n}(\hat{\vartheta}_i-\vartheta_i)\kappa(x,x_i)+b \tag{16}$$

$(\hat{\vartheta}_i-\vartheta_i)\neq 0$ 的样本即为支持向量.

参数的选择对模型的预测精准度与泛化能力有着重要影响,相对于文献[8]的主观取参方式,本文结合云资源序列的非线性与复杂性特点,利用遗传算法(Genetic Algorithm,GA)进行参数寻优.

GA 寻优的 SVR 参数包括:惩罚参数 C、核函数系数 ρ、多项式核最高次项次数 d、核函数组合权值系数 μ .算法步骤描述如下:

(1)初始化种群.随机生成初始群体,以 C、ρ、d、μ 参数作为基因构建 n 个染色体,采用多参数级联拼接方式,惩罚参数 C 和核函数带宽 ρ 使用间接二进制编码,去除小数点与符号位,避免交叉变异结果出现一串基因两位符号位或小数点.次数 d 和权值系数 μ 采用格雷码编码.指定参数寻优区间 $C\in[2^{-8},2^8]$,$\rho\in[2^{-8},2^8]$,$d\in[1,3]$,$\mu\in[0.1,0.9]$.当前迭代次数 $t\leftarrow 1$,最大迭代次数 $T=200$.

(2)依据各条染色体基因编码中的参数在训练集上训练 SVR 模型,采用式(17)的均方误差作为适应度函数进行评估并排序,淘汰后 10%个体,依据适应度大小按梯度选择复制,保持个体总数不变.

$$\gamma=\sqrt{\frac{1}{s}\sum_{i=1}^{S}(y_i-f_i)^2)} \tag{17}$$

式(17)中:S 为评估样本数目,f_i 为模型预测值,y_i 为真实值,γ 为适应度.

(3)从种群随机选择 2 个染色体作为亲代染色体,每个染色体的选择概率依据式(18)计算.

$$p_1(i)=\gamma_i/\sum_{k=1}^{n}\gamma_k \tag{18}$$

式(18)中:$p_1(i)$ 表示被选为亲代染色体的概率,γ_i 为适应度值.

(4)将亲代染色体进行基因交叉重组,交叉算子如下:

$$b_f^*=p_2b_f+(1-p_2)b_m \tag{19}$$

$$b_m^*=(1-p_2)b_f+p_2b_m \tag{20}$$

式(19)、(20)中:以 b_f^* 与 b_f 分别代表父染色体更新前后编码,b_m^* 与 b_m 同理为母染色体更新前后编码,$p_2\in(0,1)$ 为交叉率,以式(21)计算:

$$p_2(i)=\frac{\beta_1\mid\gamma_{\max}-\gamma'\mid}{\mid\gamma_{\max}-\hat{\gamma}\mid} \tag{21}$$

式(21)中:$\gamma_{\max}$ 为当前代的最佳适应度,$\hat{\gamma}$ 为平均适应度,γ' 为交叉重组的亲代染色体中的较大适应度,β_1 为调节因子.

(5)基因突变.将交叉重组后的基因编码 b_f^* 与 b_m^* 进行取补变异,突变率以式(22)计算:

$$p_3(i)=\frac{\beta_2\mid\gamma_{\max}-\gamma_i\mid}{\mid\gamma_{\max}-\hat{\gamma}\mid} \tag{22}$$

式(22)中:γ_i 为突变个体适应度值,β_2 为调节因子.

$t\leftarrow t+1$,若 $t\leqslant T$,则执行步骤 2;否则执行步骤 6.

(6)将当前种群中最佳适应度染色体的基因编码作为最优解解码输出,即得 C、ρ、d、μ 的最优参数值,依此在训练集训练最终的 SVR 预测模型.

遗传算法具备自适应和自学习性,其从问题解域搜索的并行化实现使得在参数寻优中避免了陷入局部最优的风险.

3　实验与结果分析

为了验证本文模型的有效性,本节介绍实验设置以及与其他模型的对比分析.

3.1　数据集及实验参数

本文使用 Google 云公开数据集,数据来自 Google 云计算中心 Borg 集群计算单元的工作负载.CusterData2019 数据集提供了 2019 年 5 月 8 个 cells 跨度 30 天的资源请求、调度与任务记录,单个 cell 包含 12 500 台机器,672 000 个作业,去除了终端用户对存储系统的访问模式等额外数据.数据集中的 CPU 使用信息依据 5 m 一个周期的频率采样汇总,同时包含带时间戳的内存、带宽等多维特征信息.本文筛选计算节点的数据子集作为需求序列信息,以前 25 天作为训练集,后 5 天作为测试集.

对输入数据进行预处理.利用线性插值法填补空缺值,为了使模型性能不受大规模输入样本影

响，再通过中心化和归一化得到均值为 0 的规范化输入.模型的训练和测试采用深度学习框架 Pytorch.卷积网络参数设置如表 1 所示.

表 1 实验参数

参数	值
时间维卷积核	3
填充	same
步长	2
学习率	0.000 75
滤波器数量	32
层数	6
切比雪夫多项式系数 K	3
批处理大小	64
优化器	SGD
迭代次数	200

实验中对切比雪夫多项式系数 $K\in\{1,2,3\}$ 分别取值测试，预测误差随 K 值增大而减小，邻居节点超过 3 阶后卷积聚合不满足局部性要求.图卷积与时间维卷积均采用 32 个滤波器进行局部特征提取，以 3 个时间维周期采样片段作为 3 个通道(channel)，通过调整步长改变序列长度.学习率在对数尺度进行取样.优化算法采用小批量梯度下降(mini-batch gradient descent)，实验中批处理大小(batch_size)取值过小其损失值随迭代会震荡式下降，故取样测试后在 2 的幂次方尺度以 64 为最佳.遗传算法获取的 SVR 最佳超参数分别为 $C=105.32$，$\rho=0.89$，$d=2$，$\mu=0.6$.寻优过程中的适应度函数变化曲线如图 3 所示.

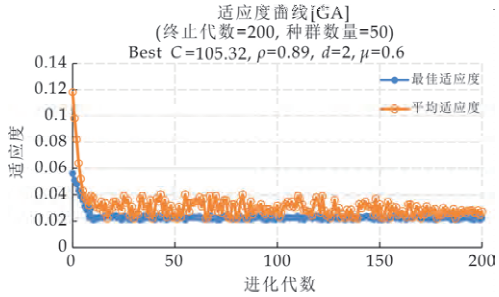

图 3 遗传算法优化参数适应度曲线变化图

图中平均适应度在种群进化后期收敛到值域为[0.02，0.03]的小幅波动区间，最佳适应度在 20 代之后趋于稳定.

3.2 实验结果及分析

本文针对云资源预测模型的改进主要立足于支持向量回归与卷积网络的结合，具体创新点为融合图卷积的空间维建模以及对 SVR 的 GA 寻参优化.因此，本文基于支持向量回归和卷积神经网络分别选择对比模型，包括：本文提出的新的预测模型；本文未融合图卷积网络结构的(CNN-SVM)；文献[8]未进行 GA 寻优的支持向量机(SVM)模型；以及进行泛化对比的经典时序预测算法 ARIMA、LSTM.

因为容器云平台的计算资源管理主要针对 CPU 与 Memory 进行，超过使用阈值即进行扩缩容响应，所以对比试验针对 CPU 利用率与 Memory 使用量序列进行预测研究.使用均方根误差(RMSE)、平均绝对误差(MAE)、平均绝对值百分比误差(MAPE)、平均均方误差(MSE)评估对比模型，计算式如下：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|(y_i-f_i)| \tag{23}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-f_i)^2} \tag{24}$$

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-f_i}{y_i}\right| \tag{25}$$

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-f_i)^2 \tag{26}$$

式(23)～(26)中：n 为预测样本数，y_i 为实际值，f_i 为预测值.

各模型在测试数据集上对未来 2 h 每隔 5 min 的单步预测结果对比见图 4 和图 5 所示，以 original 表示原始数据，ours 为本文模型.

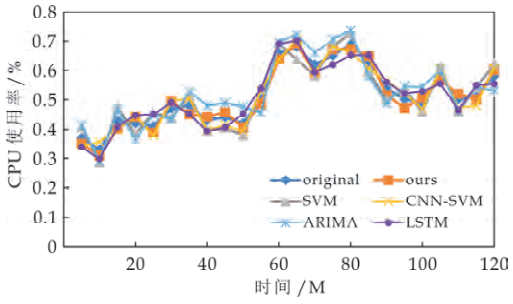

图 4 各模型 CPU 预测结果对比

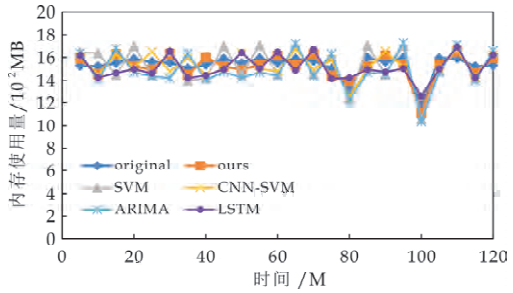

图 5 各模型内存预测结果对比

从图 4 和图 5 的实验结果可以观察到，所有算法与原始序列的趋势基本一致，本文模型相比其他算法，取得了更加理想的拟合效果.同时也可以看到，CPU 序列数据起伏波动较大，各算法相比内存序列在转折点处预测偏差更为明显.为了进一步

量化对比结果,各模型具体误差统计见图 6 和图 7 所示.

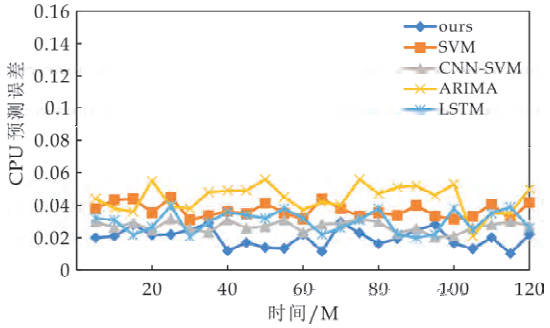

图 6　CPU 预测误差对比

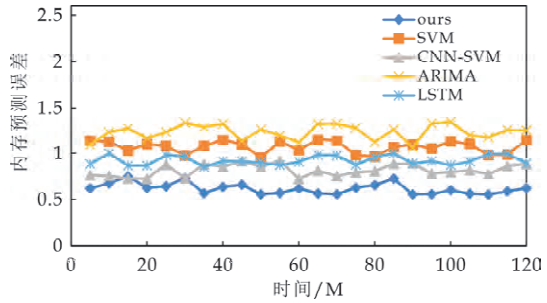

图 7　内存预测误差对比

从图 6 和图 7 可以更清楚地看到,在 CPU 和内存序列的预测中,本文模型相比其他方法具有更低的预测误差,且融合图卷积与 GA 优化的效果反映在图中表现为对 CNN-SVM 和 SVM 的明显精度优势,这说明本文的改进思路是有效的.但是也应当注意到,面对随机性更强且波动剧烈的 CPU 序列,各模型的预测稳定性显著降低,本文模型也出现了较为明显的精度波动,且相较于内存序列优势减小,这说明面对弱平稳的不规则波动,本文模型依然存在可优化空间,这也是未来的工作方向.各模型的详细误差指标如表 2 和表 3 所示.

表 2　预测性能对比(CPU)/10^{-2}

模型	MSE	RMSE	MAE	MAPE
LSTM	0.093	3.049	2.988	6.095
ARIMA	0.203	4.506	4.425	8.952
SVM	0.138	3.715	3.690	7.521
CNN-SVM	0.074	2.720	2.703	5.475
ours	0.043	2.073	2.001	4.056

表 3　预测性能对比(内存)

模型	MSE	RMSE	MAE	MAPE
LSTM	0.856	0.925	0.924	0.060
ARIMA	1.534	1.238	1.236	0.081
SVM	1.149	1.072	1.069	0.070
CNN-SVM	0.670	0.819	0.816	0.053
ours	0.386	0.621	0.618	0.041

云资源预测主要针对未来一段时间的变化趋势进行分析,单步预测结果只能描述下一时刻的资源变化情况,为了检验本文方法的泛化能力,每隔 6、12、18 个原始采集点分别求取平均,以此构建 30 min、60 min、90 min 间隔的资源序列作为新的训练样本,在测试集的多步预测结果如表 4 和表 5 所示.

表 4　不同间隔步长的预测性能(CPU)/10^{-2}

时间间隔	MSE	RMSE	MAE	MAPE
30 min	0.055	2.345	2.292	4.590
60 min	0.071	2.665	2.604	5.203
90 min	0.097	3.114	3.063	6.159

表 5　不同间隔步长的预测性能(内存)

时间间隔	MSE	RMSE	MAE	MAPE
30 min	0.510	0.714	0.713	0.046
60 min	0.723	0.850	0.849	0.055
90 min	1.112	1.054	1.052	0.068

从表中可以看到,随时间间隔增加,预测难度越来越大,但本文模型在不同步长依然保持较低的预测误差,能够稳定地描述资源的多步变化趋势,证明了模型的可靠性与稳健性,且具备一定的泛化能力.这为实际云场景中通过多步需求预测为基础平台的弹性承载力赋能提供了重要指导意义.

4　结论

本文提出一种基于卷积网络与支持向量回归的云资源预测模型,该模型结合图卷积与标准卷积提取网络拓扑下的时序数据特征,并利用遗传算法优化 SVR 预测性能.在谷歌云计算中心数据集上的实验表明,本文模型相比传统时序预测方法提高了预测精度与稳定性,提升了容器云平台的资源分配与调度效率.

参考文献

[1] 浙江大学 SEL 实验室.Docker 容器与容器云[M].2 版.北京:人民邮电出版社,2016.

[2] Maryam Amiri, Leyli Mohammad Khanli. Survey on prediction models of applications for resources provisioning in cloud[J]. Journal of Network and Computer Applications, 2017, 82: 93-113.

[3] Qi Zhang, Mohamed Faten Zhani, Raouf Boutaba, et al. Dynamic heterogeneity-aware resource provisioning in the cloud[J]. IEEE Transactions on Cloud Computing, 2014, 2(1): 510-519.

[4] 谢晓兰,张征征,王建伟,等.基于三次指数平滑法和时间卷积网络的云资源预测模型[J].通信学报,2019,40(8):143-150.

[5] Zia Ullah Qazi, Hassan Shahzad, Khan Gul Muhammad. Adaptive resource utilization prediction system for infrastructure as a service cloud[J]. Computational Intelligence and Neuroscience, 2017, 2017: 1-13.

[6] Jitendra Kumar, Ashutosh Kumar Singh. Workload prediction in cloud using artificial neural network and adaptive differential evolution[J]. Future Generation Computer Systems, 2018, 81: 41-52.

[7] Gopal Kirshna Shyam, Sunilkumar S Manvi. Virtual resource prediction in cloud environment: A bayesian ap-

proach[J]. Journal of Network and Computer Applications,2016,65:144-154.
[8] 赵 莉.基于支持向量机的云计算资源负载预测模型[J].南京理工大学学报,2018,42(6):687-692.
[9] Wei Zhong, Yi Zhuang, Jian Sun, et al. A load prediction model for cloud computing using PSO-based weighted wavelet support vector machine[J]. Applied Intelligence, 2018,48(11):4 072-4 083.
[10] Takashi Kuremoto, Shinsuke Kimura, Kunikazu Kobayashi, et al. Time series forecasting using a deep belief network with restricted Boltzmann machines[J]. Neurocomputing,2014,137:47-56.
[11] Yonghua Zhu, Weilin Zhang, Yihai Chen, et al. A novel approach to workload prediction using attention-based LSTM encoder-decoder network in cloud environment [J].EURASIP Journal on Wireless Communications and Networking,2019(1):7-18.
[12] Omer Berat Sezer, Ahmet Murat Ozbayoglu. Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach[J]. Applied Soft Computing Journal,2018,70:525-538.
[13] Alberto Mozo, Bruno Ordozgoiti, Sandra Gómez Canaval. Forecasting short-term data center network traffic load with convolutional neural networks[J]. Plos One, 2018, 13(2):e0 191 939.
[14] S Chan, I Oktavianti, V Puspita. A deep learning CNN and AI-tuned SVM for electricity consumption forecasting: Multivariate time series data[C]//2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). Vancouver, BC, Canada: IEEE, 2019:0 488-0 494.
[15] 王喜军.云计算中网络节点流量输出效率预测研究[J].计算机仿真,2018,35(8):393-396.
[16] Bruna Joan, Zaremba Wojciech, Szlam Arthur, et al. Spectral networks and locally connected networks on graphs [DB/OL]. https://arxiv. org/abs/1312. 6203, 2014-05-21.
[17] Michaël Defferrard, Xavier Bresson, Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Neural Information Processing Systems 2016. Barce Iona: Neural Information Processing Systems,2016:3 844-3 852.
[18] Klicpera Johannes, Aleksandar Bojchevski, Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank[DB/OL]. https://arxiv.org/abs/1810.05997,2019-02-27.
[19] 冯 宁,郭晟楠,宋 超,等.面向交通流量预测的多组件时空图卷积网络[J].软件学报,2019,30(3):759-769.
[20] Kipf Thomas N, Welling Max. Semi-supervised classification with graph convolutional networks[DB/OL]. https://arxiv.org/abs/1609.02907,2017-02-22.
[21] Harris Drucker, Chris J C Burges, Linda Kaufman, et al. Support vector regression machines[J]. Advances in Neural Information Processing Systems, 1997, 28(7): 779-784.

【责任编辑:蒋亚儒】

(上接第 156 页)

[14] M S M Soliman, Eihab M Abdel Rahman, Ehab El Saadany, et al. A wideband vibration-based energy harvester [J]. Journal of Micromechanics and Microengineering, 2008,18(11):115 021.
[15] M S M Soliman, Eihab M Abdel Rahman, Ehab El Saadany, et al. A design procedure for wideband micropower generators[J]. Journal of Microelectromechanical Systems,2009,18(6):1 288-1 299.
[16] Miah A Halim, S Khym, J Y Park. Frequency up-converted wide bandwidth piezoelectric energy harvester using mechanical impact[J]. Journal of Applied Physics, 2013, 114(4):0 449 021-0 449 024.
[17] Miah A Halim, Jae Y Park. Piezoceramic based wideband energy harvester using impact-enhanced dynamic magnifier for low frequency vibration[J]. Science Direct, 2015, 41:S702-S707.
[18] Miah Abdul Halim, Dae Heum Kim, Jae Yeong Park. Low frequency vibration energy harvester using stopper-engaged dynamic magnifier for increased power and wide bandwidth[J]. Journal of Electrical Engineering and Technology,2016,11(3):707-714.
[19] 王 辰.低频宽带多稳态升频能量采集器的设计及动力特性研究[D].天津:天津大学,2017.
[20] Chen Wang, Qichang Zhang, Wei Wang. Low-frequency wideband vibration energy harvesting by using frequency up-conversion and quin-stable nonlinearity[J]. Journal of Sound and Vibration,2017,399:169-181.
[21] 程千驹.分段线性压电能量收集器的宽频俘能特性研究[D].哈尔滨:哈尔滨工程大学,2017.
[22] Shaogang Liu, Qianju Cheng, Dan Zhao, et al. Theoretical modeling and analysis of two-degree-of-freedom piezoelectric energy harvester with stopper[J]. Sensors and Actuators A: Physical, 2016, 245:97-105.
[23] Liya Zhao, Yaowen Yang. An impact-based broadband aeroelastic energy harvester for concurrent wind and base vibration energy harvesting[J]. Applied Energy, 2018,212:233-243.
[24] Jinhui Zhang, Lifeng Qin. A tunable frequency up-conversion wideband piezoelectric vibration energy harvester for low-frequency variable environment using a novel impact- and rope-driven hybrid mechanism[J]. Applied Energy, 2019, 240:26-34.
[25] Kangqi Fan, Qinxue Tan, Haiyan Liu, et al. Improved energy harvesting from low-frequency small vibrations through a monostable piezoelectric energy harvester[J]. Mechanical Systems and Signal Processing, 2019, 117: 594-608.
[26] K Zhou, H L Dai, A Abdelkefi, et al. Theoretical modeling and nonlinear analysis of piezoelectric energy harvesters with different stoppers[J]. International Journal of Mechanical Sciences, 2020, 166:105 233.

【责任编辑:蒋亚儒】